Master Thesis
Data Science and Marketing Analytics

A unifying framework of Airbnb's performance predicting characteristics

Lars van Bodengraven
(457546)

May, 2022

Erasmus University
Rotterdam,
Erasmus School of Economics

Burgemeester Oudlaan 50
3062 PA Rotterdam

Supervisor:
Dr. N.M. Almeida Camacho
Second reader:
Dr. C. Cavicchia

Acknowledgements

I would like to express my sincere appreciation and gratitude to my friends and family. Likewise, I would like to thank my supervisor Dr. Nuno Almeida Camacho for his extensive guidance, support and feedback during this research. In addition, I would like to thank dr. Carlo Cavicchia for serving as co-reader. Moreover, this research would not have been possible without the data provided by Inside Airbnb.

Abstract

By unifying the effects found in previous literature this research has investigated the drivers of Airbnb's listing performance. This is done by investigating the effects of listing' characteristics (Lancasterian perspective), reviews and ratings (Social Learning perspective) and the presentation of both the host and listing (Presentation perspective). The presentation variables were extracted by means of text and image analysis. The influences of these three perspectives were examined by performing a linear regression, conditional inference tree and random forest model. The results of this research suggest that all three perspectives contain valuable features which have the ability to drive listings performance. This research especially found significant results for having more bedroom, additional and safety amenities, being a superhost, a higher brightness of the listing' main-image and a younger looking estimated age of the host' profile picture. This study contributes to academia by simultaneously investigating the effects of all three perspectives on the listing' performance. In addition, this is the first research to investigate both the host and listing presentation factors simultaneously.

Contents

Chapter 1

Introduction

The peer-to-peer (P2P) sharing economy business has made it possible to match capacity-constrained assets and resources with consumer demand (Wirtz et al., 2019). By using the P2P platforms, people across the world are now able to engage in a transaction with each other. P2P platforms such as Uber and Airbnb in particular demonstrate the possibilities and impact of these technologies. Uber, a mobile software platform, allows drivers to share a ride with customers and Airbnb platform enables a host to share their home with guests. The sharing economy is expected to grow to $ 335 billion by 2025, taken the growth of Airbnb and Uber as an indicative (Wyonch, 2017). That Airbnb is part of such an explosive market comes as no surprise considering headlines such as "Airbnb is now bigger than the world's top five hotel brands put together" (Wood, 2017) or "Airbnb is worth more than the 3 largest hotel chains combined" (Sonnemaker, 2020).

A crucial element for the growth of these P2P companies is trust, which ensures a stranger to share services without being cheated (Li, 2021). As most of Airbnb's income comes from the transaction fees charged with each booking, it is essential that the host and guest trust each other to engage in transaction. Trust in the service ensures that two parties who are not acquainted with each other are willing to do business with each other (Li, 2021). There have been incidents of host cancelling at the last minute, falsely accusing them of destroying property and delivering poorly maintained properties (Parker et al., 2016). To ensure that the customer and provider can transact with each other in good faith, Airbnb has created various means of trust.

To facilitate trust users are able to leave reviews, ratings, communicate with the Airbnb chat function, and present oneself by means of a profile. Furthermore, the hosts can exhibit their accommodation by describing the accommodations characteristics, include images, and add a text description. In addition, the host are encouraged to have a profile picture and description of themselves. These communication tools can be used by hosts and guests to judge whether they want to engage in a trade. The importance of such social factors suggests that how hosts present themselves and their listing may be of great importance for their booking demand (Lee et al., 2015, Tang and Sangani, 2015).

In addition to the social factors previous literature has also found various presentation factors to be important. Guests their perceived trust and purchase intention are influenced by the characteristics from a host's profile photo (Ert et al., 2016, Fagerstrøm et al., 2017, Wu et al., 2017, Zhang and Luo, 2018). as well as the characteristics of the listings image(s) (Zhang et al., 2017, 2018, 2019). In addition, both the self-description from the host and the listing-description have a positive influence on the Airbnb's performance ((Zhang et al., 2020, Liang et al., 2020). The majority of researchers have focused on a specific component, photo or text, with some focusing on several. To date, the existing literature has yet to analyze all the presentation factors together. Therefore, this research contributes by unifying all these presentation factors together into one research.

Besides the aforementioned two perspectives, a different strand of literature has considered the influence of various listing characteristics on its performance. This stream of literature has found the number of amenities, (Cheng and Jin, 2019), room type and size (Kwok and Xie, 2019), location (Sthapit and Jimenez-Barreto, 2018), and, superhost badge (Abrahao et al., 2017) to influence the demand of a listing. this research will investigate these listings characteristics while simultaneously controlling for the effects of the social and presentation features.

In summary, the existing literature can be divided into three strands of literature. The first strand of literature mainly focuses itself upon the influences of a listings characteristics which corresponds to Lancaster and Demand (1971) new approach to consumer theory. The second strand of literature focuses primarily on the effects of different social cues. This stream of literature will be examined under the Social Learning perspective. The third strand of literature mainly concerns itself with the presentation of the listing. This stream of literature will be examined under the Presentation perspective.

By investigating these three perspectives this research provides a better understanding of their underlying relations with a listings performance. In order to investigate these relations this research has formulated the following research question:

RQ "How can Airbnb Hosts present their accommodation the best in order to achieve a higher listing performance on the Airbnb platform?"

To answer this question, this research contrasts and synthesizes the three theoretical perspectives on the drivers of performance and tests them simultaneously on a sample of active Airbnb listings in Amsterdam. This research design is inspired by the approach of (Stremersch et al., 2007). Applying their approach made it possible to unify all perspectives into one research.

This paper is structured as follows: In Chapter 2 the previous literature is reviewed and the existing relationships are evaluated. Chapter 3 gives an overview of the data and variables. Chapter 4 introduces the text and image mining techniques that are used to create new variables which are obtained in order to do the analysis. Chapter 5 introduces the linear regression, conditional inference tree and random forest model methodology. These models are used to analyze the hypothesized relationships. In Chapter 6 the results are presented. Lastly, the implications and limitation of this research are given in Chapter 7.

Chapter 2

Literature Review

Before delving deeper into what has already been researched about and around Airbnb, it is helpful to take a quick look at the market in which it operates. After laying this general foundation, I will go into more detail about what has been found for P2P platform companies in the accommodation business. From this general accommodation P2P literature, the specific effects around Airbnb will be explained after which the conceptual framework is shaped and hypothesis will be drawn.

2.1 What is the Peer-2-Peer platform market?

The Peer-2-Peer (P2P) platform market derives its existence from the Internet and more specifically the Web 2.0 (Hall and Williams, 2020). The P2P market, also known as the gig- and sharing economy, has been growing at a remarkable rate worldwide (Hou, 2018). The P2P market has been defined by Richardson (2015) as 'forms of exchange facilitated through online platforms, encompassing a diversity of for-profit and non- profit activities that all broadly aim to open access to under-utilized resources through what is termed "sharing"' (p. 1). In other words, P2P platforms enable their users to trade their resources such as, a spare room (Airbnb), their car (Uber) or knowledge (Upwork), with the help of the internet. Doing so these companies enable service seekers and providers to find one another in new ways.

Removing these transaction barriers is considered one of the main drivers of the P2P market. Millions of transactions that take place now would not have taken place in the past as the transaction cost were too high (Benkler, 2004). Besides the lower transaction costs there is also a significant economic benefit for both parties. The lender / provider does not need the product during the lending period and the borrower gains access to the product (Frenken and Schor, 2019) often for some financial compensation.

2.2 P2P in the accommodation industry

The accommodation industry is the largest sub-sector within the tourism industry (Sharpley, 2000). The companies that operate within this sector enable their users to rent, share or temporarily trade their (spare) room or house. Examples from such companies are Airbnb, Couch-Surfing and Home exchange. These companies operate in more or less the same market with different ideologies, i.e., profit, non-profit and reciprocal, respectively. These P2P accommodation platforms give the user an authentic experience. They do so by providing a homely ambience, social interaction, meeting locals, and low-cost accommodation (Kulshreshtha and Kulshrestha, 2019).

An essential aspect for companies that operate in the P2P accommodation market is trust with some going as far as calling it the 'Achilles Heel' for P2P platforms (de Luna et al., 2019). As trust is a key factor for P2P platforms (Geiger et al., 2018, Gibbs et al., 2018), companies have introduced various mechanisms that help facilitate it. The trust enhancing mechanisms are found to drive consumer's booking decision and so the eventual performance (Abramova et al., 2017). This principle has led to several studies using customer evaluation as a proxy for performance (Ke, 2017, Lee et al., 2015, Liang et al., 2017).

2.3 The case of Airbnb

To get the highest performance Airbnb deliberately designs itself for trust (Gebbia, 2016). In a Ted Talk Joe Gebbia, co-founder of Airbnb, explains that "[they are] aiming to build Olympic trust between people who had never met. ... Is it possible to design for trust?" (min 4:34). As it turns out, yes, it is possible. As Joe Gebbia continuous Airbnb designs trust by making sure that the two parties are properly introduced and their reputation is communicated. This is done by obligating the host to create a profile and allowing hosts and guests to review and rate one another.

2.3.1 Airbnb's trust mechanisms

Following Joe Gebbias' trust design the literature concerning the performance of an Airbnb listing can be categorized within three perspectives. The first one being the Lanasterican perspective. According to Lancaster and Demand (1971), goods possess several characteristics that determine consumer choice. The utility that is obtained by the consumers' choice is assumed to be determined by the characteristics the goods possess. Therefore, the characteristics of an Airbnb Listings, such as amenities and location, determine the eventual preference from a consumer for an Airbnb.

The second is the Social Learning perspective, which claims that consumers on the Airbnb platform take lessons from previous guests. Social Learning has attracted significant academic interest. As Staff (2007) his survey shows, about 43% of consumers are reinforced of their original purchase decision by reviews, and 43% of consumers changed their intentions about which product to buy, and 9% of consumers even abandoned their purchase decisions because of negative reviews.

The third is the Presentation perspective. This strand of literature claims that the Presentation of both the host and the listing influence the eventual performance. Research has found positive influences of both text and image elements (Fagerstrøm et al., 2017, Ma et al., 2017, Zhang et al., 2017, 2019, 2020). This stream of literature has investigated this perspective using theories such as signaling theory (Janssens et al., 2021), uncertainty reduction theory (Ma et al., 2017) and social presence theory (Deng and Ravichandran, 2020). These three perspectives are summarized in Figure 2.1. Here listing performance is defined as the amount of days a listing is occupied, which is considered a key performance indicator within the hotel industry (Agarwal et al., 2003).

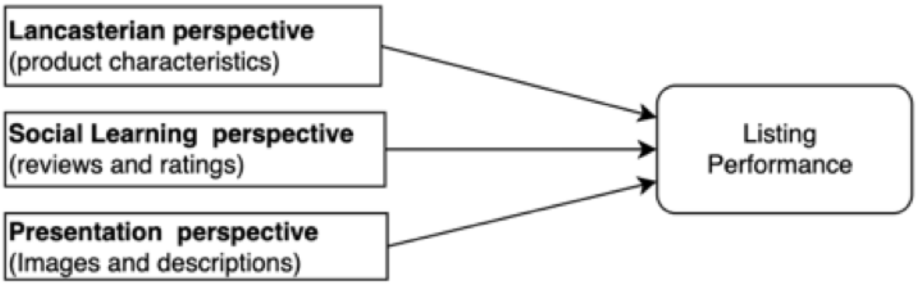

Lancasterian perspective
(product characteristics)
Social Learning perspective
(reviews and ratings)
Presentation perspective
(Images and descriptions)
Listing Performance

Figure 2.1: *Visual representation of the relation between the three perspectives and the dependent variable*

To provide a general idea of the state of the present literature, Table 2.1 provides an overview of key publications for this study. In this table, the four Presentation variables are shown to highlight that to date there has not been a paper that has analyzed all the Presentation factors together.

It is further worth noting that not all topics have been studied alike. The literature review of Dann et al. (2018) found that research has focused itself on text reviews and self-description (12.5% - 21 studies), profile images (5.36% - 9 studies). In a more recent literature review regarding Airbnb, Hati et al. (2021) mentions the research that has investigated the listing description, 5 studies. This is all included within the theme ‘how host market listings’. However, within their literature review Hati et al. (2021) does not mention any research that has investigated the effects of the listings image on any performance measure. This is likely because there is relatively little research investigating this topic, what makes it all the more interesting investigating the influences listing images.

2.4 Hypothesis development

Before delving deeper into what the literature has found concerning Airbnb a small outline of what comes next. First, the effects regarding the Lancasterian perspective will be examined and subsequently those of the Social Learning and Presentation perspective.

2.4.1 Lancasterian perspective

According to Lancaster's new approach to consumer theory (1971), goods possess, or give rise to, multiple characteristics in fixed proportions. These characteristics and not the goods themselves are that which determine the eventual choice of a consumer. One assumption of Lancaster's new approach is that the characteristics of goods are the same for all consumers. This assumption holds well for Airbnb as all consumers are shown the same characteristics for each accommodation on the website from Airbnb. In their paper Lancaster and Demand (1971) continue to explain that a good will possess multiple characteristics and that many characteristics will be shared by more than one good. This means that listings can be different in some ways, e.g., amenities, but have common characteristics like the number of bathrooms. Within this research we will look at the characteristics that have proven to be influential on the performance of a listing in previous research

Lancasterian perspective – Amenities

One of the characteristics that has been proven to influence accommodations performance are the amenities. When a host wants to list an accommodation he or she is asked to provide which amenities are provided. According to Airbnb's own research the top 10 amenities are a pet-friendly space, wifi, Free parking, a pool, a jacuzzi, a kitchen, air conditioning, heating, dishwasher, TV (Airbnb, 2022a). Some of these amenities are luxurious, i.e., jacuzzi and pool while other can be considered more essential, i.e., heating and a kitchen. On their website Airbnb, distinct six high level amenity categories which are bedroom, safety, accessibility, family, remote work and additional amenities. This list is provided within the Appendix A.

Authors	Host-Self-description	Host - Profile Image	Listing - Description	Listing - Image	Key findings	Dependent variable – Calculated by:
Le Zhang, Qiang Yan, Leihan Zhang (2020)	X				Readability and perspective taking in the host self-description help build trust. Sentiment has an inverted U-shaped relation with trust.	Trust - Calculated with a RNN
Deng, C., & Ravichandran, T. R. (2020)		X			Face disclosure in the profile picture of a host has a positive influence on the property demand. In addition, having a positive facial expression has a positive influence on the property demand.	Demand -Nr. of days reserved/ Nr. of days available
LeZhang, Qiang Yan, Leihan Zhang (2018)	X	X			Hosts who organize their self-description by concentrating on interactions, services, and the familiarity with nearby places rather than profession, personality, and hobbies, are perceived as more trustworthy. Expressing positive sentiment within both the host self-description and profile picture increases trust.	Trust-Nr. of reviews / Days listing is online
Janssens, B., Bogaert, M., & Van den Poel, D. (2021)			X		Topics in listing description significantly impact demand, with both the potential to enhance and decrease demand. The usage of multiple simultaneously used topics does not decrease demand, thereby allowing hosts to combine multiple successful strategies.	Demand - Min. Night * Avg. Reviews p/Mth
Zhang, S., Mehta, N., Singh, P. V., & Srinivasan, K. (2019)				X	The direct effect of image quality on demand is always positive, in the short and long term for all types of properties	Demand – Nr. of days booked/ Nr. of days bookable
Shunyuan Zhang, Dokyun Lee, Param Vir Singh, Kannan Srinivasan (2021)				X	Verified photos have an 8.98% higher occupancy rate. Higher quality pictures drives booking demand.	Demand – Nr. of days booked/ Nr. of days bookable

Table 2.1: *Existing literature review.*

Amenities are found to rank among the most popular topics discussed within Airbnb reviews (Cheng and Jin, 2019), which indicates their importance for the consumers. Household amenities found to be one of the top motivators for guests to choose for Airbnb (Guttentag, 2015)). The importance of the amenities does not retain itself to Airbnb but is also found within the hotel industry. Dev et al. (2018) found that amenities can increase the number of stays and enhance customer retention. In the case of Airbnb Guttentag (2015), Chattopadhyay and Mitra (2019), and Lyu et al. (2019) have found that Airbnb listings with amenities such as wireless Internet, free parking, kitchen, and laundry services had a higher performance as compared to those that lacked these amenities. It is therefore believed that a similar effect will be found for the number of amenities, that are described within the six high level amenity categories, of an Airbnb accommodation.

Lancasterian perspective – Location

Furthermore, the location of an accommodation has been found to influence the demand. Sthapit and Jimenez-Barreto (2018) interviewed Airbnb users all over the world and found that location is one of the primary drivers for using Airbnb. This is also reflected within the reviews that the guests leave behind as location is one of the top ranked topics (Cheng and Jin, 2019). This sentiment does not limit itself to reviews. In their research to reviews and tweets concerning Airbnb von Hoffen et al. found that guests particularly value a central and quite location. Masiero et al. (2015) found that also hotel demand is depended on how central the accommodation is located. When location is mentioned within Airbnb reviews it is associated with a higher rating (Tussyadiah, 2016). Location is thus considered an important characteristic for Airbnb consumers

Lancasterian perspective – House characteristics

Another type of often mentioned characteristics are the characteristics of the property, i.e., room type, number of bedrooms / bathrooms and bed type. The type of place the guest will be staying in can be categorized in four by Airbnb categorized options. These are an entire place, shared- , private- and hotel-room (Airbnb, 2022b). In addition, a host can communicate how many bed and bathrooms the accommodation has as well as the type of bed they will be sleeping on. As these features differentiate Airbnb from hotels, the accommodation characteristics have been found to be a motivator for consumers to use Airbnb (Guttentag et al., 2018).

Liang et al. (2020) found that shared homes with fewer rooms but more beds receive a high performance. Also, Biswas et al. (2020) found that the number of bedrooms is inversely correlated with listing performance. Biswas et al. (2020) elaborates that in their research this is possibly due to data selection bias, as in London private rooms are preferred for business travelers. In contrast, Kwok and Xie (2019) found that guests in general prefer accommodation that are large with more bedrooms and amenities. Their research also illuminates that the bed type matters for the performance of a listing. Providing a real bed or pull out sofa is desirable and would increase the odds of being booked by 28.55%.

In summary, the attributes of an accommodation have been found to influence the performance. This research therefore expects room type, number of bed / bathrooms, and bed type to influence the performance of a listing.

Lancasterian perspective – Superhost Badge

Another product characteristic that has been found to influence listing performance is the 'Superhost Badge'. The Superhost status is a badge of honor given to Airbnb hosts that meet certain strict criteria. These criteria are,

- Completed at least 10 trips or 3 reservations that total at least 100 nights
- Maintained a 90% response rate or higher
- Maintained a 1% cancellation rate (1 cancellation per 100 reservations) or lower, with exceptions made for those that fall under our Extenuating Circumstances policy
- Maintained a 4.8 overall rating (based on the date the guest left a review, not the date they checked out, over the past 365 days) (Airbnb, 2022c).

Once a host has reached these criteria outlined by Airbnb, they are awarded with the badge shown in Figure 2.2 below. This badge will be visible for potential guest on their listing and personal profile

Figure 2.2: *Airbnb Superhost badge (Airbnb, 2022c)*

Having achieved the superhost badge is associated with a boost in revenue (Abrahao et al., 2017), increase property demand (Xie and Mao, 2017), higher booking volume, and guest are willing to pay a premium for these accommodations (Liang et al., 2017). Hosts who have achieved the superhost badge are expected to have listings which perform better.

Lancasterian perspective – Identity Verified

Another characteristic of the listing is whether or not the host has a verified identity. Airbnb entrusts hosts and guests to decide whether they want to validate their identity or not. If they decide to do so, Airbnb verifies their identity by checking an ID card or passport. The ID verification is introduced by Airbnb to built trust between the two parties and ensure that they can take responsible decision (Labouisse, 2013). In line with Airbnb's intentions the identity verification has been found to increase guests' perceived trust (Kim et al., 2015, Abramova et al., 2017, Jung and Lee, 2017, Zhang et al., 2018). Because of the trust enhancing power of an ID verification this research expects to find that listings from which the host' ID is verified to perform better.

Lancasterian perspective – Licence

In addition, the city of Amsterdam requires listings to have a licence. In order to intervene with the overcrowded effect of short term rental the license was introduced to counteract this problem. Host have to obtain a license in order to legally rent out their house. In order to obtain a license hosts have to pay a fee of €45.- and they must meet certain requirements, one of which is that the house has taken fire safety measures. If a host does not have a license they risk a fine of €20.750.-. (Amsterdam, 2022a). Requiring a license is not limited to the city of Amsterdam (Guttentag, 2015, Miller, 2014). However, research has not provided the effects of having a license, yet. Since having a license is an acknowledgement of a city or government that an accommodation is licensed to be rented out, this research expects that having a license has a similar effect to having a verified ID.

Lancasterian perspective – Price

Lastly, research has found that consumers choose Airbnb for its economically advantageous aspects (Tussyadiah, 2016). Visser et al. (2017) found that the price of an accommodation is very important for guests when choosing a listing. Also Xie and Mao (2017) found that the price impacts a listings demand. Since price is a determining factor for guests whether or they settle on a listing, price is expected to inversely influence listing performance.

H1a: The more central the Airbnb is located the better the performance of an Airbnb listing.

H1b: Private Airbnb rooms perform better than other Airbnb room types.

H1c: The superhost badge has a positive influence on listing performance.

H1d: The number of amenities positively influences listing performance.

H1e: The host's verified ID has a positively influences listing performance.

H1f: If a listing has a license this has a positively influences listing performance.

H1g: The price is negatively related to a listings performance.

H1h: The number of guests a listing accommodates positively influences listing performance.

2.4.2 Social Learning perspective

Ratings and reviews are a collection of sharing economy service tools (Möhlmann and Teubner, 2020). These tools encourage consumers to share product feedback and communicate their own personal experiences and opinions with other consumers. As a result of the Internet's ubiquity and abundance of user-generated-content, a growing number of consumers are consulting online product reviews before making purchasing decisions (Staff, 2007). This is reflected in research from Staff (2007) where they used a survey to analyze the impact of online product reviews on the purchasing decision of consumers. According to their analysis, product reviews reinforced 43 % of consumers' original purchase intent, however 43% changed their minds about which product to buy, and a stunning 9% of consumers abandoned their purchase plan after reading the reviews.

Social Learning

In the research of Staff (2007) participants engaged in what is called a Social Learning process. According to Vaccari et al. (2018), this online Social Learning process consist out of consumers sharing their experiences and opinions about a product with other consumers by means of ratings and reviews. While more reviews amass, consumers will be able to obtain better assessments of the quality of the products, enabling them to make a better and well-considered purchasing decision. In other words, consumers are able to use online reviews and ratings to evaluate which products are more trustworthy and so help them make their purchase decision (Jung et al., 2016). Taking lessons from these social cues – reviews and ratings – is what is understood in this research as Social Learning.

Number of Reviews

As Airbnb consumers do no fully rely on the review content (Bae and Koo, 2018), likely due to them being mostly positive (Ke, 2017, Cheng and Jin, 2019), they also make use of alternatives such as review quantity (Abrahao et al., 2017), (Bae and Koo, 2018). Review quantity has been found to influences demand (Xie and Mao, 2017), increases listing revenue (Abrahao et al., 2017), increase the trust from the consumer (Abrahao et al., 2017) and the popularity of the listing (Mauri et al., 2018).The influences of reviews have been found throughout all generations (Chen and Chang, 2018). The effect of review quantity also became apparent in the research from Kwok and Xie (2019) where 5000 people participated in an online trust game. The results show that going from 4 to 5 stars is equivalent to having 10 more reviews. However, when the listings are less comparable the effect also diminishes.

Rating Score

In addition to the review quantity Airbnb guests also focus on the accompanying star ratings. Airbnb enables guests to leave a star rating, on a scale from 0 (bad) to 5 (good), with their reviews. The rating scores have been found increase the perceived value of an accommodation (Chen and Chang, 2018). Similarly, research has found the rating score to be significantly related to a listings perceived trustworthiness (Chen and Xie, 2017, Zhang et al., 2017, Jaeger et al., 2019). However, the ratings did not increase purchase intention (Chen and Chang, 2018). Nevertheless, this research therefore expects the rating score to positively influence a listings performance.
In summary, consumers on the Airbnb platform utilize the ratings and reviews left by previous guests to evaluate their preferences. As guests do no fully trust the review content (Bae and Koo, 2018), likely due to them being mostly positive (Ke, 2017, Cheng and Jin, 2019), they also divert to proxies such as review quantity (Bae and Koo, 2018). This research therefore hypothesizes that Airbnb guest learn from previous guest by a Social Learning process, i.e., review count and ratings. These variables are therefore expected to influence the performance of an Airbnb listing.

H2a: Review quantity positively influences Airbnb performance.

H2b: Airbnb ratings positively influence Airbnb performance.

2.4.3 Presentation perspective

Another stream of literature has investigated the effects of the Presentation of both the host and their listing. On Airbnb host are able and encouraged to present themselves and their listing in the best way to attract as many guests as possible. Airbnb encourages host to do so by providing listing-description examples, providing photographers and a personal profile page where they can present themselves to potential guests. Research has investigated the effects of these textual and visual Presentation variables by analyzing both the text-description from the listing and host as well as their accompanying pictures. First the written elements will be examined after which the visual elements will be explored.

Host self-description

On the Airbnb platform guests and hosts are empowered to give a self-description on their profile. The self-description contains important trust related-cues for both parties as this is the first place where they can introduce themselves to the other and in doing so develop trust (Zhang and Luo, 2018). The seller's self-description in the form of a well-written, high-quality texts can be used to communicate personal traits like perceived social capital, ability, and integrity, all of which can have a substantial impact on the perceived trust by the other party (Malinen and Ojala, 2011, Norcie et al., 2013, Chen et al., 2015). In addition, the seller's self-descriptions and text reviews are both used commonly to reduce perceived risks and to protect users from ambiguities and false expectations (Jung and Lee, 2017).

Besides investigating the effects of the host's self-description on trust indicators research has also analyzed these texts to identify and differentiate clusters of hosts (Tussyadiah, 2016). This indicates that Airbnb hosts can be differentiated by their writing style. Similarly, Ma et al. (2017) found that the host's self-description usually refers to eight different themes; origin/residence 69%; work/education 60%; interests and tastes 58%; hospitality 53%; travel 48%; relationships 28%; personality 27%; life motto and values 8%. More importantly, Ma et al. (2017) found that the credibility of a host increases when they include information referring to their; origin; hospitality; occupation and personal interest.

More recently, Liang et al. (2020) found that presenting rich and extensive information in both the accommodation and host description improves review volume, which in turns indicates higher booking performances. Similarly, Mauri et al. (2018) found that the popularity of a listing was driven substantially by personal narrative story telling in the host's self-description.

Hence, the personal description of the host influences the final performance of a listing. In addition to the information and topics it covers, the sentiment it contains is also important. The sentiment in the host's self-description has been found to be positively correlated with the performance of their listing(s) (Ma et al., 2017, Zhang et al., 2018). In their research Zhang et al. (2020) however note that there is an inverse U-shaped relation between occupancy rate and sentiment. Expressing too much sentiment can thus be detrimental for performance. In order to investigate the relation between the host self-description and listing performance, this research will look at the sentiment of the host's self-description.

Listing description

Apart from a description about themselves, hosts are asked to provide a description about their accommodation. Airbnb encourages hosts to provide an honest description of the accommodation but also to tell a story, remain authentic, and honest (Airbnb, 2022d). Host can use the listing description to promote the accommodation for various different target groups. (Lutz and Newlands, 2018) found that the listing description of entire homes and shared space rentals differed from each other. Entire home host explicitly target "older guests, couples, business travelers and high-income professionals, while highlighting professional-level cleanliness and ensuring privacy. In contrast, shared room hosts targeted younger and frugal guests, did not boast about cleanliness and assured social interaction was a part of the experience" (Lutz and Newlands, 2018).

Considering that the listing description can be used to attract various consumer groups the writing style, and information provided influence listing performance. Research from Liang et al. (2020) confirms that providing comprehensive and detailed listing descriptions positively influences the performance. In addition, what information is included also matters. Janssens et al. (2021) found that the topics that are addressed within the description can both enhance and decrease listing demand, depending on the topic. Descriptions that write about topics such as enthusiastic home experience and neighborhood touring are able to enhance the performance of a listing. In contrast, descriptions which have a 'hotel description', meaning giving a hotel like feel, decreases listing performance. They furthermore found that guests in San Francisco preferred to read what a neighborhood has to offer instead of the special amenities a listing has. In addition, they found that hosts can use multiple successful strategic topics without diminishing the demand.

In similar research on San Francisco and Oakland, Chung and Sarnikar (2021) also found that the listings description has a significant impact on listing performance. However, they found that there is a mismatch between the frequently emphasized topics and the most beneficial ones. Using a deep-learning model, Chung and Sarnikar (2021) found that including more information about the kitchen enhances performance for entire homes and shared rooms. Addressing the interior style enhances private room performance. Different topics can thus influence the performance of some room types while decreasing performance for another.

The existing literature has thus investigated the listing' description thoroughly. Nevertheless, there is only one paper that has addressed the sentiment embedded in the listing' description. Martinez et al. (2017) investigated the sentiment of New York Airbnb listings using a lexicon based approach. Since this is the only paper so far that has examined listings sentiment and since the technology for sentiment analysis has improved it is interesting to see how this relationship holds up now. Doing so, this research will not only give a more recent view of the relation between sentiment and performance it will also expand this knowledge to another geological location namely, Amsterdam. In summary, the description of the listing has been found to influence the eventual performance of it. Therefore, this research expects to find a significant effect from a listing description on performance by investigating the listings description sentiment.

Host profile image

Not only textual information is important for establishing trust but also visual aspects. In their research Bae and Koo (2018) found that Airbnb guests did not fully trust the review content, and therefore used, in particular, the accompanying images. Similarly, Ert et al. (2016) found that, Airbnb consumers not only rely on textual information when choosing an accommodation but are also influenced by the host profile picture. In their research, they found that the perceived trustworthiness of a host's profile picture increases the booking probability.

When choosing an accommodation, Airbnb guests will thus also consider the visual elements on the website in determining their decision. The visual elements that are provided on the Airbnb listing page are images of the concerned accommodation and a profile picture of the host (see Figure 2.1). As just touched upon the host's profile picture influences the trustworthiness of an

Airbnb listing and therewith the eventual booking results. Fagerstrøm et al. (2017) investigated this effect by looking at the host's facial expressions. By manipulating the facial expression, they found that even when accompanied by low prices and positive ratings, neutral and positive facial expressions increased booking tendencies, whereas negative facial expressions and a lack of host pictures reduced booking tendencies. Additionally, the amount of self-disclosure in the profile picture positively influences the trustworthiness and booking intentions of the guests (Broeder and Crijns, 2019).

The effect of images on consumer decision making can also occur unconsciously (Todorov, 2008, Todorov et al., 2009). Within as little as 100 milliseconds the brain has already established a primary opinion about the trustworthiness of a person's face. Zhang et al. (2020) also enforces this idea by stating that among four kinds of factors, self-disclosure probably contains the most abundant information. They do so because the profile photo can directly reveal a person's face, gender, and even age. Also, Barnes (2021) underlines this idea and suggest that it would be interesting to look at the effect of gender, and age differences between host's. This research therefore expects that having a profile picture with self-disclosure, positive facial expression and the estimated age can increase booking performances.

Listing main-image

As Figure 2.3 shows, the host's profile picture is only a part of the visual information that is provided on Airbnb (bottom right). When guests search for an Airbnb accommodation, they are initially shown the accommodation main image, title characteristics, review score and price (list left). It is only when they click on the property that they are directed to the page shown on the right side. On the left side, the search page for all the listings are shown. Zervas et al. (2017) analyzed the ratings of over 600,000 accommodations globally and discovered that 94 percent had a rating of >4.5 (out of 5) stars, with almost none having a rating of less than 3.5. Because of this abundance of positive information Airbnb guests have to rely on additional information such as visual information (Zhang et al., 2019).

As over half of the human brain is used to process graphical information, this visual aspect is extremely important in marketing (Marieb and Hoehn, 2007). Especially since images have been found to be most effective in capturing attention (Wedel and Pieters, 2007). Using pictures as a marketing strategy is found to be particularly important in highly intangible businesses such as tourism (Tijana Rakić, 2010). Pictures help consumers to reduce the complexity of the information allowing them to compare more of accommodation and also compare them more thoroughly (Pan and Zhang, 2016).

Even though photos are excellent communication tools - a picture tells more than a thousand words - a large proportion of hosts do not use the professional photo service of Airbnb. The reason for this according to Zhang et al. (2019) is that higher quality photos also come with higher expectations, which translates into poorer experiences and the absence or receipt of negative reviews. In the long run, this could lead to a decrease in demand. However, Zhang et al. (2019) always found a positive effect of the photo quality on the review quantity, which in addition was used as a proxy for demand. These findings were consistent in a follow-up study of Zhang et al. (2021) where they again found that higher quality photos positively affect demand. This research therefore expects the quality of the listing image to positively influence the performance of an Airbnb listing.

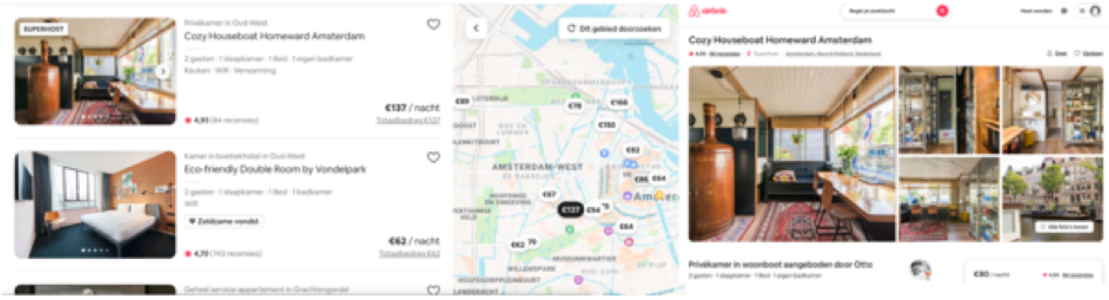

SUPERHOST
Privékamer in Oud-West
Cozy Houseboat Homeward Amsterdam
€137 / nacht
Eco-friendly Double Room by Vondelpark
Zeldzame vondst
€62 / nacht
Dit gebied doorzoeken
AMSTERDAM-WEST
€137
airbnb
Cozy Houseboat Homeward Amsterdam
Privékamer in woonboot aangeboden door Otto

Figure 2.3: *Airbnb listing overview (left) and listing landing page (right) (Airbnb)*

Hypothesis Presentation Perspective

In short, much research has been done on the various factors that belong to the Presentation perspective. Researchers have found positive effects for both text and picture variables. It is worth noting here that, as shown in Table 2.1, to date there has not been a single research that has considered all of the host's presentation factors combined in one study.

The literature cited so far shows that the self-description of the host influences the final performance of a listing. What kind of information, topics, and sentiment is included can determine the final performance. Because of the found positive effects of sentiment this research will investigate the effects thereof on the listings performance. Similarly, previous research found that the description accompanying the listings affects a listings performance. The effects of sentiment are in this case however understudied. This research will therefore investigate the effects listing description sentiment on its performance.

Furthermore, the profile picture of both the listing and host have been found to influence performance. Research has found that having a profile picture increases performance and especially when the host smiles or shows a positive emotion. Therefore, this research investigates the effects of the emotion shown in the profile picture. In addition, the main image of the listing will be investigated. Research has shown that higher quality images lead to better listing performance. However, Airbnb host are wary about the expectations they set by doing so.

Nevertheless, research has found that higher quality leads to better performance. From be the presentation literature the following hypothesis are therefore formulated.

H3a: The host's self-description has a positive influence on the listing performance.

H3b: The listings description has a positive influence on the listing performance.

H3c: The host's profile picture characteristics have a positive influence on listing performance.

H3d: The listings main-image picture quality positively influences listing performance.

2.5 Conceptual framework

In summary, this research expects to find positive influences of the three perspectives namely, Lancasterian- Social Learning- and Presentation- perspective on the performance of an Airbnb listing. Here performance is defined as the amount of days a listing is occupied, which is a key performance indicator within the hotel industry (Agarwal et al., 2003). The specific operations of the variable will be discussed within the methodology section. These expectations are based on the previously discussed literature. The expected relations and corresponding hypothesis are depicted in the following Figure 2.4.

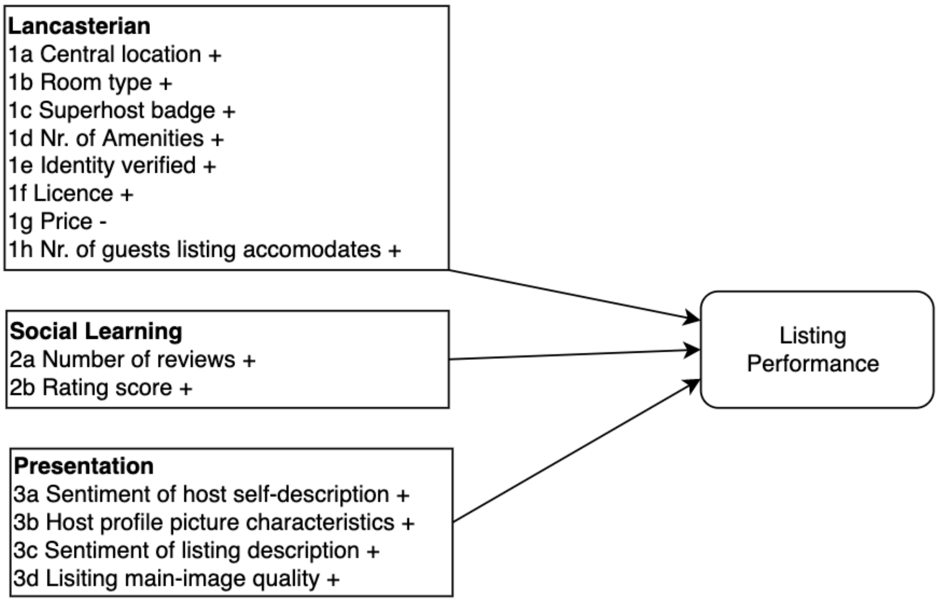


Figure 2.4: *Conceptual framework*

Chapter 3

Data

This section explains the data that is used and why this particular time frame was chosen. Furthermore, the dependent and independent variables are introduced. Chapter 4 will further explain how this list of independent variables is expanded through the use of text and image analysis.

3.1 Data collection

To study the effects of text and images on booking performances this research will make use of the publicly available information from InsideAirbnb (2022). InsideAirbnb is an activist project who provides data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect cities from the impacts of short-term rentals (InsideAirbnb, 2022). In this research, the focus will be on Amsterdam as the researcher has knowledge about the city's language, culture and legislation.

The data sets that will be used are scraped on the third of June 2021 and the fifth of December 2021. The variables included within dataset from June third are used for the analysis and the data from December fifth is used for calculating the occupancy rate. Before delving deeper into the specifics of why this time-frame has been chose I will first elaborate on the specifics of the occupancy rate.

3.1.1 Performance measure

Since the Airbnb dataset does not include a variable to directly measure the listing performance, as shown by making a booking, a proxy variable will be used. Before explaining the formula used for this proxy first some information on how it is constituted. As this research is not the first to use the dataset provided by InsideAirbnb (2022) I will turn to the solutions provided by previous literature. Using the occupancy rate as performance measure has been proven to be a key performance indicator within the hotel industry (Agarwal et al., 2003) as well as online sales (Ye et al., 2009). Also within the Airbnb literature the occupancy rate has been used by various researchers. Zhang et al. (2019) and Zhang et al. (2021) for example used the number of days bookable divided by the number of days available as performance metric. They however had access to a richer data set. While using InsideAirbnb' 2019 dataset Janssens et al. (2021) came up with a clever performance metric. They multiplied the average number of reviews per month with the minimum nights a guest should at least stay. Doing so, they calculated the demand for a specific listing. By combining the performance metrics of Janssens et al. (2021), Zhang et al. (2019), and, Zhang et al. (2021) this research will use the following formula to calculate listing' performance.

$$\text{Performance } = \frac{(\# \text{ Reviews } * 2) * \text{ Minimum nights } (l_i)}{200} \tag{3.1}$$

Since guests are not required to leave a review the numbers of reviews are multiplied by two. This is found to be appropriate as Marqusee (2015) found that the rate in which customer leave reviews is 72%. In contrast, Brousseau et al. (2015) estimated that this review rate is equal to 31.5%. By evening these out this research adapts a review rate of 50% meaning that 1 review equals two visitors, which is in line with InsideAirbnb (2022) . Hence, this calculated number of visitors is multiplied by the amount of nights a visitor stays in an accommodation. The average amount of nights a guest stays in Amsterdam is estimated to be 3.5 nights (Briene et al., 2021)). If the minimum nights of listing is lower than 3 the estimated average has been taken since guests are expected to stay more than one night on average. The total number of nights the accommodation is booked are divided by 200 days. This is because there are 200 days between the time frame the occupancy rate is calculated on – more on the time frame shortly. In summary, listing performance is measured by the minimum nights, weighted over two hundred days and capped at one.

3.1.2 Independent variables

The independent variables consist out of variables which are collected by Inside Airbnb such as, the price and the amenities of a listing. In addition to these variables this research will utilize text mining techniques to extract the sentiment from both the listing' and host' description. This will be explained in chapter Variable Extraction Methodology. Furthermore, the picture URL and listing URL will be used to extract these pictures characteristics. This will be explained in section Extracting variables from host-profile picture and Extracting variables form listing-profile picture. An overview of the variables that are currently in the dataset can be found in table 3.1

Variable	Type of variable	Description	Role in the study
id	Numeric	Airbnb's unique identifier for the listing	Observation identifier
Occupancy_rate	occupancy_rate	Calculated performance measure using the formula as explained in section	Booking performance measure
Lancasterian perspective			
Price	Numeric	Daily price in euro's	Investigating listing' characteristics
Accommodates	Numeric	The maximum capacity of the listing	Investigating listing' characteristics
Bedroom amenities	Numeric	The number of bedroom amenities a listing as described in appendix	Investigating listing' characteristics
Additional amenities	Numeric	The number of additional amenities a listing as described in appendix	Investigating listing' characteristics
Family amenities	Numeric	The number of family amenities a listing as described in appendix	Investigating listing' characteristics
Safety amenities	Numeric	The number of safety amenities a listing as described in appendix	Investigating listing' characteristics
Top amenities	Numeric	The number of top amenities a listing as described in appendix	Investigating listing' characteristics
Host Identity is verified	Binary	Binary indicator whether the host is verified, a 1 being yes	Investigating listing' characteristics
Number of verification's	Numeric	The number of verification a host has, i.e., email, phone, government id, Facebook etc.	Investigating listing' characteristics
Bathrooms	Numeric	The number of bathroom the listing has	Investigating listing' characteristics
Host is a superhost	Binary	Binary indicator whether the host is a superhost, a 1 being yes	Investigating listing' characteristics
Host has a license	Binary	Binary indicator whether the host has a license which are mandatory in Amsterdam, a 1 being yes	Investigating listing' characteristics
Room type	Categorical	All homes are grouped into the following three room types i.e., private room, hotel room, entire home / apartment	Investigating listing' characteristics
Neighbourhood	Categorical	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	Investigating listing' characteristics
Social learning perspective			
Number of Reviews	Numeric	The total number of reviews a listing has accumulated since it is available on Airbnb	Investigating listing' social factors
Rating score	Numeric	The average rating scores guests have left by their reviews, a 0 being bad and a 5 good	Investigating listing' social factors
Presentation perspective			
Host has description	Binary	Binary indicator whether the host has written a self-description on their personal profile	Investigating listing' presentation factors
Picture URL	character	Contains a link to the host' profile picture, this link will, as explained in section , be used to extract the photo characteristics	Used to extract image characteristics
Listing URL	character	Contains a link to the Listing' main-image, this link will, as explained in section , be used to extract the photo characteristics	Used to extract image characteristics

Table 3.1: *Table summary of variables in data set*

3.1.3 Time frame

Due to the Covid-19 restrictions, Amsterdam was not very accessible for both national and international tourism. However, since May 19, step 2 of the reopening plan went into effect. From then on swimming pools, gyms, amusement parks and open-air museums were allowed to open again under certain conditions. Terraces were also allowed to open longer and music and dance schools were also allowed to teach indoors under certain conditions. Furthermore, step 3 of the opening plan took effect on June 5. This means that the Netherlands, and therefore Amsterdam, went from 'closed unless' to 'open unless'. This means that virtually all locations were open again. This includes the arts and culture sector, hospitality industry, sports, saunas and tanning studios, indoor spaces, amusement parks and zoos, canal boats, community centers and casinos (van Algemene Zaken, 2022).

Since step 2 of the reopening plan led to more opportunities and this was continued two weeks later with step 3, this period can be seen as the reopening of Amsterdam. For that reason, it was chosen to look at how the effects of the three different perspectives influence the performance of the listing from this period onwards. This timeframe is of course sup-optimal but just as covid-19 has applied restrictions to society these residual restrictions are also found within this research. Nonetheless, national tourism was back to pre-Covid-19 levels in 2021 (NBTC, 2022). However, due to international restrictions there were still fewer tourists from outside the Netherlands, -71% compared to pre-corona year 2019 (NBTC, 2022).

3.1.4 Listing selection

In this analysis only active listings are taken into account. Active listings are listings that are bookable for thirty or more days a year. In addition, these listings should be active for the period preceding as well as following may nineteenth. This is because accommodations are not automatically deleted from Airbnb when they are inactive. Hence, taking listings that are active from 2020 till December 2021 are considered in this analysis while excluding all others.

3.2 Pre-processing and data cleaning

As the data from InsideAirbnb.com is web scraped data some pre-processing steps are necessary to ensure that de data is interpretable. InsideAirbnb.com provides multiple data sets from which the Listing information and Review information datasets will be used. These two datasets are matches based on their respective Id. The initial dataset form June third, 2021 consist out of 16.973 listings and contains 74 variables concerning the characteristics of the listings, availability, reviews, ratings, and corresponding information, identifier number for both the host and each listing, and URL's for both the profile picture of the host and the main listing image. From these sixteen thousand listings 3.296 listings have had guests between the 19th of May and December

5th. Of these, 1964 listings have been active since 2020.

Furthermore, as there are four room types, i.e., entire home/apartment, hotel room, private room, and shared room, and shared room type only contained four observations these were excluded from the data. In addition, 474 listings did not provide any results for the host' profile-picture and 42 not for the listing' main-image. These were observations were therefore excluded from the analysis resulting in 1421 remaining listings.

In addition, while examining the text description it became apparent that some descriptions were incomplete. Descriptions such as 'Hi' or 'I am' are considered incomplete and the value for these descriptions were transformed to missing. Also, descriptions that were not written in English are transformed to missing for the sentiment analysis, which I will explain shortly. However, a binary variable was included to denote whether there was a host description present as it can be an important factor for guests (Liang et al., 2020). Since the included observations all contained a listing description the binary variable only refers to the host' self-description. Furthermore, it became apparent that there seems to be a maximum amount of words that can be scraped from Airbnb. The listing descriptions contain a maximum of 1000 characters which results in some abrupt endings.

In addition, the amenities were categorized by the groups shown in Appendix A. Since not all amenities are listed in this document, another variable was created with the total number of amenities.

Chapter 4

Variable Extraction Methodology

This chapter explains how this research will make use of multiple machine learning techniques before doing the eventual analysis. To analyze the multitude of textual information – self-descriptions and listing descriptions – this research will turn to the use of natural language processing or NLP. NLP is becoming more and more prevalent in marketing and “has the potential to shed light on consumer, firm, and market behavior, as well as society more generally” (Berger et al., 2020, p. 1). In order to do so, I will use DistilBERT a distilled version of Google’s BERT. In addition, to extract the image characteristics from both the host’ profile picture and listing’ main-image this research will use the deep learning models of Face ++ and Sight engine. These will be explained after the sentiment analysis methodology.

4.1 Sentiment analysis

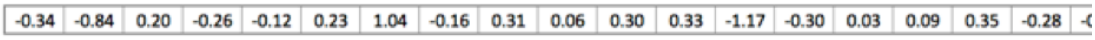
In order to analyze both the host and listing descriptions I will use sentiment analysis. Sentiment analysis is a form of text analytics, which can be used to analyze people their “opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” (Liu, 2012, p. 7). In order to do so it uses natural language processing and machine learning to identify the underlining sentiment, e.g., positive, neutral, or negative, from a word, sentence or text. By means of text analytics this research is able to analyze the enormous amount of written information and its underlying sentiment(Mayya et al., 2020).

To do the sentiment analysis I will use a version of the BERT algorithm which stands for Bidirectional Encoder Representation from Transformers to do the sentiment analysis. BERT, developed by Google researcher Devlin et al. (2018), has achieved state-of-the-art performances on sentiment analysis performances. It outperformed all the existing sentiment algorithms in predicting The

Stanford Sentiment Treebank, which is a movie review database with human annotations of their sentiment. More recently, a BERT algorithm has been found to outperforms the lexicon-based methods (Kotelnikova et al., 2021). Because of the outstanding performance of BERT, a version of this algorithm will be used to accurately understand the meaning of words and their context and so predict their sentiment. To understand how BERT works I will first explain how the predecessors of BERT worked as BERT 'builds on top of a number of clever ideas that have been bubbling up in the NLP community' Alammar (2022). Therefore, I will first explain what a word-embedding is, after which the original BERT algorithm will be explained. Lastly, I will introduce the distilled version of BERT, DistilBERT, which will be used to do the eventual sentiment analysis.

4.1.1 Word-embedding

A machine learning model cannot process words in their alphabetic shape. To process these words the model needs some form of numeric representation. To enable machine learning models to process words, word-embedding's have been invented. A word-embedding is thus a numerical representation of a word. The numerical representation for the word 'stick' for example is shown in Figure 4.1. The vector for stick consist out of 200 numbers, as represented by GloVe a pre-trained word embedding package.



-0.34	-0.84	0.20	-0.26	-0.12	0.23	1.04	-0.16	0.31	0.06	0.30	0.33	-1.17	-0.30	0.03	0.09	0.35	-0.28	-(

Figure 4.1: *Representation of the word embedding of the word Stick by GloVe*

The vectorisation of words also enables us to perform mathematical equations on it. One example of this is: King − man + woman = Queen. However, a shortcoming of this approach is that a word can have different meanings depending on its context. Taking the example of 'stick' once more, it becomes obvious that "to 'stick' with someone" or "throwing a 'stick'" represent two very different meanings of the word 'stick'. In order to overcome this shortcoming researchers have advanced to more complex text analytics models. One of which is BERT.

4.1.2 Who is BERT?

According to creators, BERT is "designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" (Devlin et al., 2018, p. 1). The algorithm is pre-trained on two unsupervised tasks namely, masked language modeling and next sentence prediction which I will elaborate on shortly. Because of the fact that BERT allows a document to simultaneously belong to multiple classes it is an advanced as well as a more realistic language processing technique (Alaparthi and Mishra, 2021). Since BERT is a state-of-the-art language processing algorithm that has achieved exceptional results in eleven natural language understanding tasks (Devlin et al., 2018), this is considered to be an appropriate method to analyze the text of both the listing and the host. The superiority of BERT becomes once again evident by the sheer fact that Google announced that it has been using BERT for all of its English queries since 2020 (Raghaven, min. 7:57).

4.1.3 How does BERT work

To understand how BERT works it is necessary to understand a crucial element of BERT namely, the encoder. The encoder takes the text description of the listing or host as an input and simultaneously creates word-embeddings for each word. These word-embeddings are just as shown earlier a vector of numbers only, in this case the word surroundings are taken into account. Therefore, each word has a different word-embedding depending on its context however, similar words have closer numbers in their vector than non-similar words. By taking the context of each word into account the encoder learns the meaning of words and its context. The left part of Figure 4.2 schematically illustrates what such an encoder looks like. The encoder first compares the encoding of one word to the encoding of other words in the multi-headed-attention layer. During this process, it controls whether a word encoding can be improved in order to better understand the context of a word, and so the entire sentence. Secondly the feed-forward layer contains all the weights which are trained during training process.

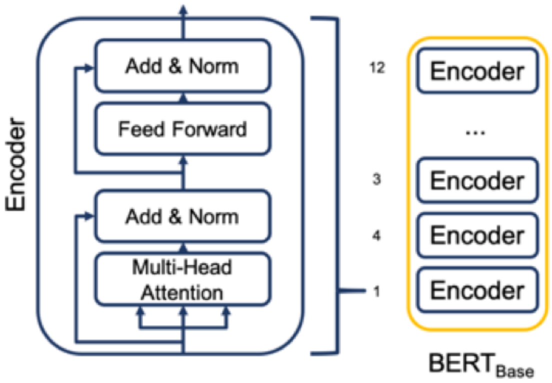


Figure 4.2: *Schematic representation of one Encoder (left) and a stack of Encoders (right) which are used within the BERT model (Evtimov et al., 2020)*

If we pile several of these encoders up we get the Bidirectional Encoder Representation from Transformers or BERT, as seen on the right of Figure 4.2. In other words, Bert uses the encoder's linguistic understanding to be able to conduct things like sentiment analysis, text summarizing and question answering. For BERT to gain any knowledge of language, however, the model must first be trained on large amounts of data. It is only after BERT has a knowledge of the language that the model can be fine-tuned to learn a specific task.

4.1.4 Training BERT

BERT learns what is language by training on two unsupervised tasks simultaneously. These are Masked Language Model (MLM) and Next Sentence Prediction (NSP). For MLM BERT takes in a sentence and 'masks' 15% of the words. Masking means that the words are hidden for the model. The goal from MLM is to train BERT to output the masked words. This helps BERT to get a bidirectional understanding of the words within a sentence. MLM is comparable with 'fill in the blanks' where one makes use of the words that come before as well as after the blank. In addition, BERT check whether two sentences are likely to follow each other. This helps BERT to understand context across sentences themselves. A visual representation of this process from Devlin et al. (2018) is shown in Figure 4.3.

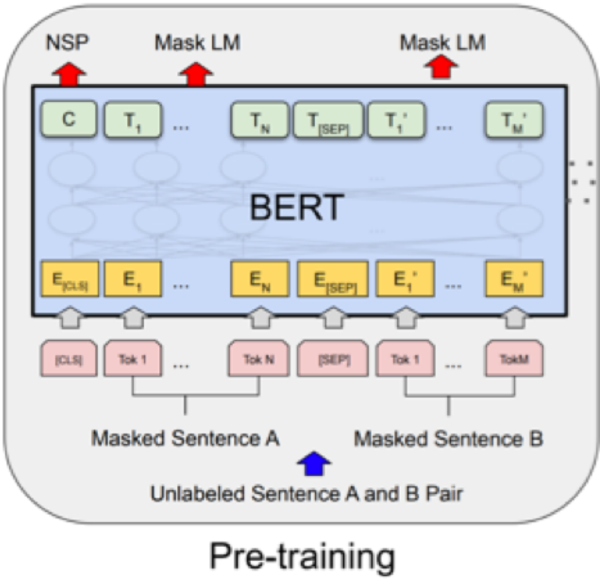


Figure 4.3: *Visual representation of training the BERT model (Devlin et al., 2018)*

In Figure 4.3, Token 1 to Token N represent the tokenized words from the input sentence. Word tokens can be used to split words into multiple pieces, and so, reducing the vocabulary size for covering every word (Wu et al., 2019). For example, the word 'sleeping' is tokenized into 'sleep' and '##ing'. This approach helps to break many unknown words into some known words. After the tokenization of words model attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. Within the stack of BERT encoders (the blue area) the words are first transformed into word-embeddings and come out as output tokens for the masked language model problem. C is a classifier which will be used classify the sentiment of both the host and listing description.

4.1.5 Softmax

The final step in the BERT algorithm is the Softmax function. Softmax is a function that is able to transform a vector of values to a value between 0 and 1. Due to this function the vectors can now be interpreted as probabilities. In the case of BERT, the Softmax function transforms the classifier vector output into a probability using the following equation where Z_i represents the input vectors.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4.1}$$

Doing so the Softmax function enables the BERT algorithm to give a probability for a possible outcome to happen. In the case of this research BERT gives a probability for the text having a positive or a negative sentiment.

4.1.6 DistilBERT

Recently researchers have created a new version of BERT. This version DistilBERT is a distilled version of the original BERT. This version reduced the size of the BERT model by 40% while retaining 97% of it language understanding and being 60% faster (Sanh et al., 2019). To create this distilled version of BERT, DistilBERT' creators disregarded the segment-embedding token and reduced the number of encoders by a factor of 2. DistilBERT has been proven to be great at extracting the sentiment from sentences (Büyüköz et al., 2020). This model will be used to extract the sentiment from the hosts self-description and listing description. By utilizing the, by Wolf et al. (2019), pre-trained DistilBERT algorithm, the probability of a sentence being negative or positive, are calculated.

By running the pre-trained DistilBERT algorithm on the description of both that of the host and of the listing, this study obtains information about the probability of these being positive or negative. This research chooses to separately analyze the sentiment of the host' self-description and listing' description since the information is found on two distinct locations. DistilBERT assigns a value between 0 and 1 to the description which indicates the sentiment. Negative sentiment contains a value between -1 and 0 and positive sentiment between 0 and 1. Adding these two variables together results in the total negative or positive sentiment of the given description.

Consequently, the final variable has a value between -1 and 1 where -1 stands for a negative sentiment and 1 represents positive sentiment.

4.2 Extracting variables from host-profile picture

The data from InsideAirbnb (2022) data set contains a lot of information, one of which is a link to the profile picture of the host. As mentioned earlier the host profile picture is found to influence the perceived trust (Ert et al., 2016, Broeder and Crijns, 2019) and booking demand (Fagerstrøm et al., 2017). Since the data set only contains a link to the profile picture this research utilizes the services of Face ++ to extract characteristics of the hosts' profile-pictures. Face ++ is a face detection software that uses deep-learning models to extract face characteristics (Fan et al., 2014). Face ++ is considered to be one of the best face detection softwares and has been extensively used in academic research (Edelman et al., 2017, Jung and Lee, 2017, Deng and Ravichandran, 2020, e.g.). Research has used this service to, among others, detect the gender, race, age, expression and emotion of a face.

So, Face ++ uses deep learning models to analyze the faces. Deep learning models are composed of numerous processing layers that get an understanding of the data with multiple levels of abstraction (LeCun et al., 2015). In their often-cited paper LeCun et al. (2015) explain that '[t]hese methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains'(p. 436). Deep learning algorithms mirror the way the brain processes information and comparably performs better the more often it has had to tackle a task. Consequently, a limitation of deep learning models is getting a large number of training data (Zhao, 2017). This in turn is one of the reasons for this research to turn to the service of Face ++.

Face ++, part of parent company MEGVII, utilizes it personally developed MegEngine algorithm to analyze and detect faces (MEGVII, 2022a). Such algorithms typically start by identifying the eyes as these are the easiest feature to detect. From this reference point the algorithm finds the alignment of a face in a photo. In addition to finding the faces, this software is trained with large amounts of data to recognize emotion, gender and age. According to Face ++, the algorithm can determine with 99% certainty the age (with a std. of 5 years) and gender (MEGVII, 2022b).

To utilize this service this research provides the host-profile URL to Face ++'s trained MegEngine algorithm which extracts the features from the profile pictures. As a deep learning models generally provides better results considering the amount of data it has been trained on, this research uses this service instead of building a model from scratch. Face++'s face detection software provides information about the amount of faces in the pictures. In addition, it provides, for each face, information about the emotion, age and gender. Face ++ returns a value between 0 and 100 for the estimated emotion. An example of a host' profile picture and Face++ estimation is given in Appendix D.

In addition, Face ++ returns the age, between 0 and 100, and whether the face belongs to a male or female. However, some pictures contain multiple faces, e.g., family picture with 5 or 6 faces. Consequently the Face ++ returns a list of characteristics for each face. After a visual inspection of the host' profile pictures the average age of all faces proved to match that of the host. This is because, despite Face++ their accuracy claims, the algorithm does not perform well with recognizing the age of younger people. The age of the parents is nevertheless estimated reasonably accurate. Therefore, taking the average age of the faces within the picture is considered to be a great approximation of the average age of the hosts. Appendix E contains two examples where there are multiple faces within the picture and how Face ++ categorizes them. Since only a fraction of the data picture consists out of more than two faces – 1.95% – the average age of the picture is considered to be a good indicator for these family pictures.

In line with research from Deng and Ravichandran (2020) this research categorizes the seven possible emotions into three categories. The dominant emotion is kept while the other two options are set to zero. Positive facial expression is kept when the average of happiness and surprise are highest among all expressions. Neutral is kept when neutral is highest among all others. Lastly, negative emotion is kept when the average of anger, disgust fear and sadness scored the highest among all facial expressions. In the example of Appendix D, the dominant emotion is happiness. The host' emotion is therefore coded positive with value 95.544.

4.3 Extracting variables form listing-profile picture

In addition, the InsideAirbnb (2022) data set also contains the an URL of the listings main-image. As mentioned in the literature review the quality of the listings image is found to influence the performance of that listing. Higher quality images result in higher listing performance (Pan and Zhang, 2016, Zhang et al., 2019, 2021). Airbnb also recognizes the importance of picture quality by encouraging their host to implement brightness and contrast in their pictures (Airbnb, 2022d) as well as having a focused photo (Airbnb, 2022e).

Brightness, contrast and focus are important attributes for determining an image quality (Ouni et al., 2011). By utilizing the software of SightEngine this research is able to extract this information from the listing' main-image. SightEngine utilizes deep convolutional neural networks to analyze the quality features of a given image (SightEngine, 2022).

A Convolutional Neural Network (CNN), is a type of neural networks that is specialized in processing data that has a grid-like topology, such as an image (Mishra, 2020). A digital image is a collection of zero's and one's that contain the values for a pixel's color and brightness. Similar to how neurons in a human brain respond to visual stimuli the neurons from a CNN also respond to specific fields of a picture. Doing so, CNN is able to distinguish simple things such as lines and curves but also more complex patterns such as faces. SightEngine's CNN utilizes 'billions of neurons [that] are arranged into successive layers' (SightEngine, 2022).

SightEngine returns a value between 0 and 1 for each image quality category. For sharpness, a higher value indicates a sharper picture whereas zero denotes a blurry picture. For brightness, a 0 indicates a dark picture and a 1 a very bright picture. For contrast a 0 means low contrast and a 1 high contrast. The examples that SightEngine gives are illustrated in Appendix C.

4.4 Summary

By now the dataset has been enriched with new variables with the help from machine learning techniques. Therefore, this is an ideal moment to pause and review what has happened. First the new variables will be summarized shortly after which an overview of all the variables will be given.

As just discussed, I used text mining (DistiBert) to create two variables: (1) host self-description sentiment: The extent to which the text the host uses to describe herself has a negative or positive sentiment, and (2) listing description sentiment: The extent to which the text the host uses to describe her property has a negative or positive sentiment. These values range from -1 to +1 where -1 represents negative sentiment and 1 represents positive sentiment. Since Airbnb hosts naturally want to portrait their listing in a positive light, it comes as no surprise that many hosts use positive sentiment in both their own and listing' description.However, this text mining analysis ensures that I can capture the extent in which they do so.

In addition, I used the services of Face ++ and SightEnginge to create variables new variables for the host profile picture and the listing main-image, respectively. From the host' profile picture I created a variable containing (1) the estimated average age of the face(s) in the picture, (2) the gender of the faces and (3) what kind of emotion these faces show. Emotion consists of seven categories namely; happiness, neutral, surprise, sadness, disgust, fear, and anger, which add up to 100. Here a 0 represents nothing of that particular emotion and 100 represents only that emotion. The emotions have been categorized following Deng and Ravichandran (2020) namely, positive facial expression is kept when the average of happiness and surprise are highest among all expressions. Neutral is kept when neutral is highest among all others. Lastly, negative emotion is kept when the average of anger, disgust fear and sadness scored the highest among all facial expressions.

Additionally, three variables are created to embody the listing' main-image quality namely, (1) the sharpness, (2) the brightness and (3) the contrast of the photo. These variables range from 0 to 1 where a 1 represents sharp, bright and high contrast photos and a 0 blurry, dark and low contrast photos (see Brightness for examples).

The variables already present within the InsideAirbnb.com dataset together with the just described extracted variables are merged together in one data frame. A summary of these variables and their corresponding scales are provided in Table 4.1

Variable	**Type of variable**	**Description**	**Role in the study**
Id	Numeric	Airbnb's unique identifier for the listing	Observarion identifier
#Occupancy rate	Numeric	Calculated performance measure using the formula as explained in section	Booking performance measure
Lancasterian perspective			
Price	Numeric	Daily price in euro's	Investigating listing' characteristics
Accommodates	Numeric	The maximum capacity of the listing	Investigating listing' characteristics
Bedroom amenities	Numeric	The number of bedroom amenities a listing as described in appendix	Investigating listing' characteristics
Additional amenities	Numeric	The number of additional amenities a listing as described in appendix	Investigating listing' characteristics
Family amenities	Numeric	The number of family amenities a listing as described in appendix	Investigating listing' characteristics
Safety amenities	Numeric	The number of safety amenities a listing as described in appendix	Investigating listing' characteristics
Top amenities	Numeric	The number of top amenities a listing as described in appendix	Investigating listing' characteristics
Host identity verified	Binary	Binary indicator whether the host is verfified, a 1 being yes	Investigating listing' characteristics
Nr. of verifications	Numeric	The number of verification a host has, i.e., email, phone, government id, facebook etc.	Investigating listing' characteristics
Bathrooms	Numeric	The number of bathroom the listing has	Investigating listing' characteristics
Host is Superhost	Binary	Binary indicator whether the host is a superhost, a 1 being yes	Investigating listing' characteristics
License	Binary	Binary indicator whether the host has a licens which are mandetory in Amsterdam, a 1 being yes	Investigating listing' characteristics
Room type	Categorical	All homes are grouped into the following three room types i.e., private room, hotel room, entire home / appartment	Investigating listing' characteristics
Neighbourhood	Categorical	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	Investigating listing' characteristics
Social learning perspective			
Number of reviews	Numeric	The total number of reviews a listing has accumulated since it is available on Airbnb	Investigating listing' social factors
Rating score	Numeric	The average rating scores guests have left by their reviews, a 0 being bad and a 5 good	Investigating listing' social factors
Presentation perspective			
Host has description	Binary	Binary indicator whether the host has written a self-description on their personal profile	Investigating listing' presentation factors
Sentiment host	Numeric	The sentiment expressed in the host' self-description, -1 being negative 1 being positive	Sentiment of host self-description
Sentiment description	Numeric	The sentiment expressed in the listing description, -1 being negative 1 being positive	Sentiment of listing description
Happy	Numeric	The amount of positive facial expression shown on the host' profile picture, 0 being none and 100 only happiness	Host profile picture
Neutral	Numeric	The amount of neutral facial expression shown on the host' profile picture, 0 being none and 100 only neutral	Host profile picture
Negative	Numeric	The amount of negative facial expression shown on the host' profile picture, 0 being none and 100 only negative	Host profile picture
Average age	Numeric	The average estimated age of the host' profile picture face(s).	Host profile picture
Count male	Numeric	The number of male faces in the host' profile picture	Host profile picture
Count female	Numeric	The number of female faces in the host' profile picture	Host profile picture
Sharpness	Numeric	The sharpness of the picture, a 0 being blurry and a 1 being sharp	Listing' image quality
Brightness	Numeric	The brightness of the picture, a 0 being dark and a 1 being bright	Listing' image quality
Contrast	Numeric	The contrast in the picture, a 0 being none and a 1 being a lot of contrast	Listing' image quality

Table 4.1: *Table summary of variables in data set*

Chapter 5

Data Analysis Methodology

This chapter will explain the models which are used to investigate the hypothesized relationships between the three perspectives and the occupancy rate. Since the objective of this research is to find the relationships between the occupancy rate and the various perspectives the primary model is a linear regression model. Linear regression is a very interpretable analysis and provides primary knowledge of the relation between the variables. The overall performance of the linear models will be evaluated by using the AIC and BIC criteria. These will be explained in more depth shortly. The second section of this chapter, describes the conditional inference tree and random forest models. These tree-based models are estimated to investigate the non-linear relationships. Lastly, a performance metrics for the comparing the models will be introduced as well as how the random forest model will be interpreted.

5.1 Linear Regression

To establish the relationships between the three perspectives and the occupancy rate this research has estimated four different linear models. In general, linear regression finds the best fit of the model by minimizing the sum of squared errors. The linear regression model provides a baseline to which the advanced models can be compared. First a linear model containing only the Lancasterian perspective variables will be conducted, as shown in equation 5.1

$$\begin{aligned}\text{occupancy rate}_i =&\beta_0 + \beta_1 price_i + \beta_2 accommodates_i + \beta_3 bedroom.amenities_i +\\ &\beta_4 additional.amenities_i + \beta_5 family.amenities_i + \beta_6 top.amenities_i +\\ &\beta_7 safety.amenities_i + \beta_8 nr.verifications_i + \beta_9 bathrooms_i +\\ &\beta_{10} superhost_i + \varepsilon_i\end{aligned} \tag{5.1}$$

Secondly a model containing both the Lancasterian and Social learning perspective will be estimated, as illustrated in equation 5.2.

$$\begin{aligned}\text{occupancy rate}_i =& \beta_0 + \beta_1 price_i + \beta_2 accommodates_i + \beta_3 bedroom.amenities_i + \\ & \beta_4 additional.amenities_i + \beta_5 family.amenities_i + \beta_6 top.amenities_i + \\ & \beta_7 safety.amenities_i + \beta_8 nr.verifications_i + \beta_9 bathrooms_i + \beta_{10} superhost_i + \\ & \beta_{11} number.of.reviews_i + \beta_{12} rating.scores_i + \varepsilon_i\end{aligned} \tag{5.2}$$

And thirdly, a model with both the Lancasterian and Presentation perspective will be estimated, as shown in equation 5.3.

$$\begin{aligned}\text{occupancy rate}_i =& \beta_0 + \beta_1 price_{i,t} + \beta_2 accommodates + \beta_3 bedroom.amenities + \\ & \beta_4 additional.amenities + \beta_5 family.amenities + \beta_6 top.amenities + \\ & \beta_7 nr.amenities + \beta_8 nr.verifications + \beta_9 bathrooms + \beta_{10} superhost + \\ & \beta_{11} host.has.descr + \beta_{12} sentiment.host + \beta_{13} sentiment.description + \\ & \beta_{14} neighbourhood.cleansed + \beta_{15} postive + \beta_{16} negative + \\ & \beta_{17} sharpness + \beta_{18} brightness + \beta_{19} average.age + \\ & \beta_{20} male + \beta_{21} female + \varepsilon_i\end{aligned} \tag{5.3}$$

Lastly, the fourth model is estimated by the formula given in equation 5.4.

$$\begin{aligned}\text{occupancy rate}_i =& \beta_0 + \beta_1 price_i + \beta_2 accommodates_i + \beta_3 bedroom.amenities_i + \\ & \beta_4 additional.amenities_i + \beta_5 family.amenities_i + \beta_6 top.amenities_i + \\ & \beta_7 safety.amenities_i + \beta_8 nr.verifications_i + \beta_9 bathrooms_i + \beta_{10} superhost_i + \\ & \beta_{11} number.of.reviews_i + \beta_{12} rating.scores_i + \\ & \beta_{13} host.has.descr_i + \beta_{14} sentiment.host_i + \beta_{15} sentiment.description_i + \\ & \beta_{16} neighbourhood.cleansed_i + \beta_{17} postive_i + \beta_{18} negative_i + \\ & \beta_{19} sharpness_i + \beta_{20} brightness_i + \beta_{21} average.age_i + \\ & \beta_{22} male_i + \beta_{23} female_i + \varepsilon_i\end{aligned} \tag{5.4}$$

5.1.1 AIC and BIC

In order to evaluate the linear models the Akaike Information Criterion (AIC) and, Bayesian Information Criterion(BIC) will be used to compare the models. Both the AIC (Akaike, 2011) and BIC (Schwarz, 1978) are a measure of goodness of fit which can be used to evaluate linear models. In contrast to the AIC, BIC penalizes stronger for model complexity. In other words, AIC generally allows for more complex models than the BIC. It is therefore considered reasonable to use both information criteria to give a balanced overview. For both the AIC and BIC measure a lower value is considered to be better. More specifically, a value that is two points lower is considered to be significantly better for the AIC (Wilcox et al., 2016) and more than ten for the BIC (Lorah and Womack, 2019). Both the AIC and BIC use the log-likelihood estimate to determine the eventual information criteria. The formula for calculating the AIC is given by equation 5.5.

$$AIC = 2K - 2\ln(L) \tag{5.5}$$

Where K is the number of independent variables used, L is the log-likelihood estimate. Similarly, the formula for BIC is shown in equation 5.6.

$$\mathrm{BIC} = K\ln(n) - 2\ln(L) \tag{5.6}$$

Where K is the number of parameters, n the number of data points and, L the likelihood estimate. The difference between the AIC and BIC criteria is the penalty term. AIC penalizes the number of parameters by 2 K whereas BIC penalizes by K $*\ln(n)$. This penalty term is dependent on the number of independent variables. Since a lower AIC and BIC score are preferred the penalty terms discourage overfitting by penalizing the number of variables.

5.2 Tree based models

In order to further investigate the non-linear relations I will use a conditional inference tree and random forest model, to get a better understanding of the relation between the three perspectives and the performance measure. Both the conditional inference tree and the random forest model are non-parametric methods which allow this study to investigate the non-linear effects. Hence, these models are able to improve our understanding of the relationship between the three perspectives and listing performance. Before I explain how the conditional inference tree and random forest will be used it is convenient to have an underlying understanding of what a regression tree analysis is. This is because the regression tree is the foundation of both the conditional inference tree and the random forest model.

5.2.1 Regression Tree

A regression tree attempts to create the best splits to predict a specific outcome. In the case of Airbnb when predicting the occupancy rate of a listing the tree might use the price variable for the first split. This split is based upon the residuals sum of squared (RSS) which is shown in equation 5.7.

$$RSS = \sum_{i=1}^{n} \left(y^i - f\left(x_i\right)\right)^2 \tag{5.7}$$

Here y^i is the real value of y and $f(x_i)$ is the predicted value of y. The algorithm searches, taken price as example, to the price where the RSS is the lowest. Splitting the data at this price point, in the soon introduced model €103.-, results in the best prediction of occupancy rate based on a higher or lower price than €103.- . The regression tree does this for all variables within the dataset and selects the best variable with corresponding split value which generates the lowest RSS.

This iterative process is called recursive binary splitting. The regression tree is built from the top-down, meaning that first of all the top node is made and only afterwards the node below. Doing so, the regression tree progressively divides the data into various predicting branches where each split result into new branches. Even though the regression tree is build top-down, the best split calculated at the top does not necessary guarantee a better tree at the end. Calculating the optimal split locally, at each node, is what is known as the greedy approach. The greedy approach has two fundamental problems 1) over-fitting and 2) variable selection issues (Hothorn et al., 2006). As the regression tree optimizes the split at each node, it tends to be biased to variables which can be separated many times, such as price or number of reviews. This selection bias is a problem for the models interpretability since, the interpretability of these trees is 'affected by the biased variable selection' (Hothorn et al., 2006, p. 652). Since the InsideAirbnb data set contains a lot of numeric variables resulting in many split points this is a point of concern. To avoid this, the following method is be used.

5.2.2 Conditional Inference Tree

The Conditional Inference tree (C.I.T.) overcomes the shortcomings of the regression tree by selecting variables based on a significance test, rather than selecting the variable that minimizes the RSS. The C.I.T. is therefore unaffected by over-fitting, unbiased towards numeric variables and additionally becomes very interpretable (Hothorn et al., 2006, Strobl et al., 2007). In order to do so, the C.I.T. recursively partitions the data by the variable, X, that is able to split the data the best so that the occupancy rate, Y, is significantly different after the split. The C.I.T. only does this if the global null hypothesis of independence among the independent variables and a listing' performance is rejected. By means of equation 5.8 the C.I.T finds the variable with the strongest association with a listing' performance rate.

$$\mathbf{T}_{j^*}^{A}\left(\mathcal{L}_n, \mathbf{w}\right) = \operatorname{vec}\left(\sum_{i=1}^{n} w_i I\left(X_{j^*i} \in A\right) h\left(\mathbf{Y}_i, \left(\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right)\right)^{\top}\right) \in \mathbb{R}^q \tag{5.8}$$

Once the C.I.T has found a variable X to split it uses equation 5.9 to find the optimal split. As mentioned earlier the C.I.T. recursively partitions the data by the variable, X, in subsets A. Using equation 5.10 the optimal split is found.

$$\mathbf{T}_{j^*}^{A}\left(\mathcal{L}_n, \mathbf{w}\right) = \operatorname{vec}\left(\sum_{i=1}^{n} w_i I\left(X_{j \cdot i} \in A\right) h\left(\mathbf{Y}_i, \left(\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right)\right)^{\top}\right) \in \mathbb{R}^q \tag{5.9}$$

$$A^* = \underset{A}{\operatorname{argmax}} c\left(\mathbf{t}_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A}\right) \tag{5.10}$$

Stated more simply, the C.I.T. follows three steps. First the C.I.T exhaustively searches all possible splits for all independent variables that reject the H_0 of independence between the X and Y variables. Secondly, the variable with the best split is taken. Lastly, the algorithm continues to repeats this process until the H_0 cannot be rejected anymore.

Doing so, the algorithm reduces the potential bias, prevents overfitting, and solves variable selection problems. A drawback of this approach is however, that it does not enhance predictive performance. There is nevertheless, a general consensus the results are highly interpretable (Schivinski, 2021). Another drawback is that the algorithm does not work well with high-categorical variables and missing values within the dataset. Fortunately, this is not the case for the Inside Airbnb dataset.

5.2.3 Random Forest Regression

As interpretable as the conditional inference tree may be a disadvantage is that is prone to having a high variance. The random forest regression – a machine learning method that introduced by Breiman (2001) – overcomes this flaw. By combining multiple – in paragraph 5.2.1 explained – low bias, high variance regression trees, random forest ensures a lower variance (Hastie et al., 2009). The combination of trees that makes up the random forest is defined by the following equation:

$$\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{5.11}$$

Here $\mathrm{T_b}(x)$ refers to a vector of regression trees. B stands for the total number of trees and b represents an individual tree. This technique is also known bootstrap aggregating, or bagging, where the algorithm draws random samples - bootstraps - from the original dataset to compute the various decision trees. Hence, the name random forest.

By averaging the predictions of all the estimated trees, random forest balances the over estimating trees with the underestimating trees which leads to a more accurate prediction (Hastie et al., 2009). There is however a difference between bagging and the random forest algorithm. Bagging uses a greedy approach where the algorithm is able to use all variables to find the optimal split. This is problematic as this makes the algorithm prone to bias towards strong predictors. In order to overcome this, the random forest algorithm is only able to use a subset of variables to calculate the optimal split-point. Doing so, the predictions of the estimated trees de-correlate which reduces the variance. Furthermore, this also prevents the algorithm of over-focusing on variables that have strong predictive power. This is especially convenient for this study as variables such as 'number of reviews' and 'number of amenities' are, as discussed earlier, been found to be strong predictors with many split-points.

5.2.4 Root Mean Squared Error

In order to obtain the predictive performances of the models the root means squared error (RMSE) will be used. Since the goal of this research is not to build a predictive model but rather explain the underlying mechanisms the RMSE does not have to be optimized. It is however a good indication of how well the models predict new data and therefore how well the estimated relations hold for new cases. The RMSE provides information the average distance between the predicted listing performance and the actual listing performance in the dataset. A lower RMSE thus indicates that the model has a better fit to the data. In order to calculate the RMSE the data will be divided into a train and test set. The models will be trained on the train dataset and predicted on the test set. The train data consist out of 80% of the data and the test set contain 20% of the data resulting in an 80/20 split. The RMSE can be calculated using the following equation.

$$\text{RMSE } = \sqrt{\Sigma\left(P_i - O_i\right)^2/n} \tag{5.12}$$

Here P_i is the models predicted value of listing i, O_i is the real value from listing i and n is the sample size of the test data.

5.3 SHAP

To understand the output of the black box method random-forest this research utilizes SHAP values (Shapley, 1952). SHAP stands for *SHapley Additive exPlanations* and is a tool to understand specific predictions of a model. If we only consider the models prediction capability it is not clear *why* this prediction is made. Since this research is interested in the specific relations between the dependent and independent variables investigating *why* the model predicts certain listing performances is important. SHAP explains what happen within the algorithm by calculating the local additive feature importance. This means that the algorithm explains one specific instance at the time. It does so by using the following equation.

$$\varphi_i(v) = \sum_{S \subseteq N(t)} \frac{|S|!(N-|S|-1)!}{N!}(v(S \cup \{i\}) - v(S)) \tag{5.13}$$

Here N are the number of features and S is a subset of these features. Furthermore, the marginal contribution of feature i, in subset S is calculated by $(v(\text{ S} \cup \{\hat{i}\}) - v(\text{ S}))$. The eventual contribution of each feature is determined by taking the average contribution from all permutations.

Since SHAP calculates the contribution of each feature for random subsets it could be that in predicting a certain observation the algorithm encounters a missing value. SHAP deals with this missing value by going both ways at the node with the missing value and recursively average the results. Furthermore, SHAP has found to have difficulties with highly correlated variables. If this is the case SHAP allocates all feature contribution to one variable. In this research, the SHAP algorithm will be used to calculate both the exact local and general additive attributions values from the Random Forest model. By means of the SHAP Python package (Lundberg and Lee, 2017) the SHAP values are calculated.

Chapter 6

Results

In this chapter the results of the regression models, conditional inference tree and random forest are presented. The first section assesses whether there are linear relationships between the independent and dependent variables. Doing so, this section forms the base-line to which the more advanced models can be compared. In the second section of this chapter, the non-linear relationships are investigated by means of the tree-based models. A more general and contextual interpretation of the results will be provided within the conclusion section.

Table 6.1: *Descriptive Statistics*

Descriptive Statistics					
Independent variable	**Mean**	**(SD)**	**Median**	**Min**	**Max**
Occupancy rate	0.279	0.304	0.136	0	1.0
Lancasterian perspective					
Price	158	109	130	22	1490
Accommodates	3.00	1.63	2.00	1	16
Bedroom amenities	1.11	0.971	1.00	0	3
Additional amenities	4.18	1.58	5.00	0	7
Family amenities	1.99	0.919	2.00	0	5
Safety amenities	2.34	1.09	2.00	0	4
Top amenities	3.35	1.11	4.00	0	6
Nr. verification's	5.78	2.10	6.00	1	11
Nr. of bathrooms	1.12	0.403	1.00	0	5
Host is a superhost	0.432	0.496	0	0	1
Host ID verified	0.821	0.383	1.00	0	1
License	0.643	0.479	1.00	0	1
Social Learning perspective					
Number of Reviews	84.4	107	43.0	1	825
Rating scores	4.81	0.251	4.87	1.00	5.0
Presentation perspective					
Host has description	0.706	0.456	1.00	0	1
Sentiment descr. host	0.521	0.573	0.984	-1.00	1.00
Sentiment descr. listing	0.767	0.609	0.998	-1.00	1.00
Positive	67.9	41.8	95.5	0	100
Neutral	14.7	32.6	0	0	100
Negative	6.24	21.9	0	0	99.99
Sharpness	0.963	0.0597	0.980	0.370	0.99
Brightness	0.552	0.158	0.560	0.0100	0.99
Average age	44.4	12.8	44.3	14.0	88
Count Male	0.615	0.548	1.00	0	3
Female	0.639	0.622	1.00	0	4
Categorical variables - Lancasterian perspective					
Room type	Hotel_room	Entire_home_apt	Private_room		
	29 (2.0%)	765 (53.8%)	627 (44.1%)		
Neighborhood	Other	Centrum_Oost	Bos_en_Lommer	De_Pijp_Rivierenbuurt	
	38 (2.7%)	173 (12.2%)	39 (2.7%)	129 (9.1%)	
	Noord_Oost	Centrum_West	Watergraafsmeer	Oostelijk_Havengebied_Indische_Buurt	
	37 (2.6%)	293 (20.6%)	31 (2.2%)	50 (3.5%)	
	Slotervaart	Oud_Noord	Baarsjes_Oud_West	Gaasperdam_Driemond	
	19 (1.3%)	61 (4.3%)	175 (12.3%)	17 (1.2%)	
	Zuid	Noord_West	Westerpark	De_Aker_Nieuw_Sloten	
	88 (6.2%)	41 (2.9%)	80 (5.6%)	14 (1.0%)	
	Oud_Oost	IJburg_Zeeburgereiland		Geuzenveld_Slotermeer	
	71 (5.0%)	43 (3.0%)		22 (1.5%)	

6.1 Descriptive statistics

Table 6.1 gives an overview of the descriptive statistics of the variables used. The variables are summarized by the three perspective. Table 6.1 shows that the average occupancy rate is 27,9% which translates into 55 days per two hundred days on average. The median however is 13.6% which translates into 27 days. Considering that the regulations in Amsterdam only allow a listing to be rented thirty nights per year the occupancy rate can be considered as relatively high. From Table 6.1 it is also evident that the distribution of the ratings is consistent with what was brought up in the literature review namely, the majority of the ratings are positive (Ke, 2017, Cheng and Jin, 2019).

6.2 Check Linear Regression assumptions

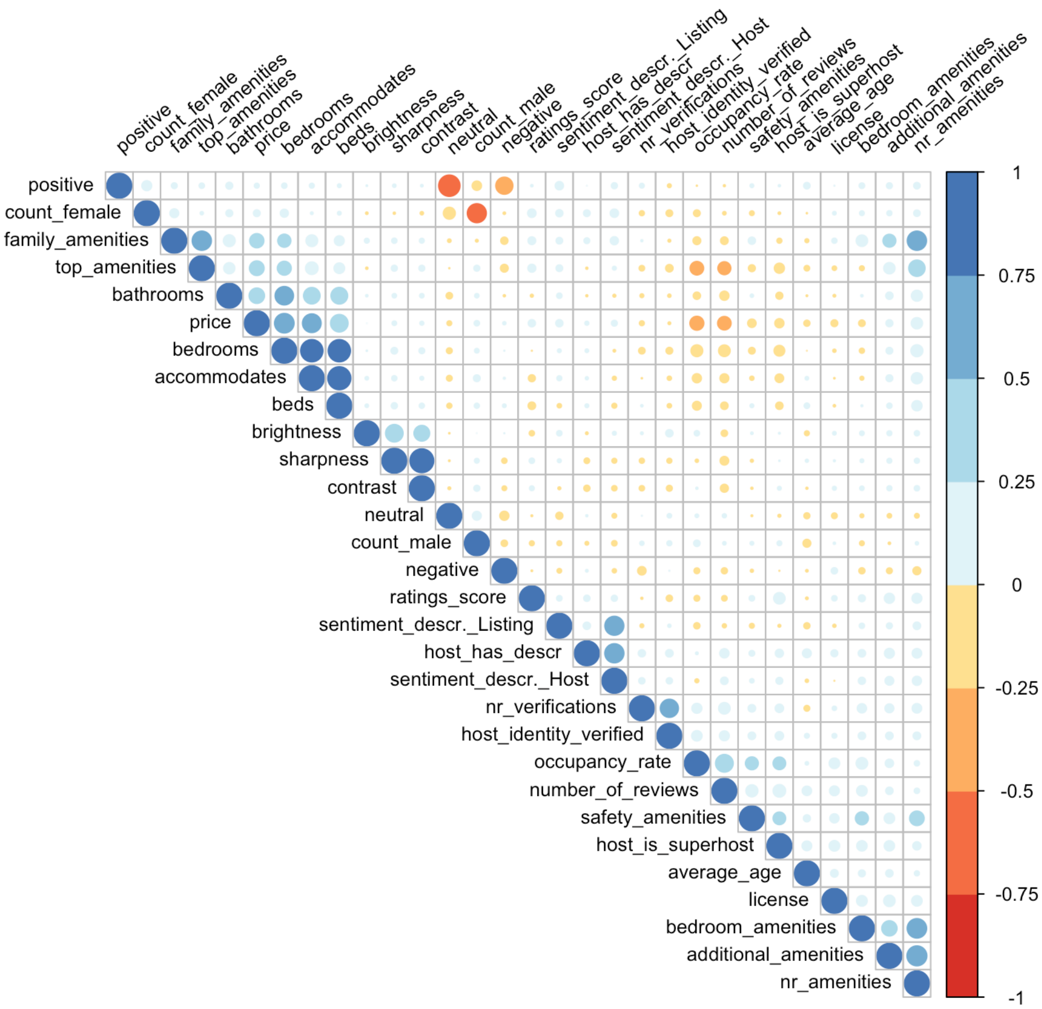


Figure 6.1: *Pearson's correlation matrix*

First of all, this research controls for the correlation between variables. This is important since multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the models (Frost, 2017). In order to do so both the correlation between variables as well as Variance Inflation Factor have been investigated. The correlation plot in Figure 6.1 shows the visualized Pearson correlation coefficients which calculated the correlation for each pairwise relationship (Lantz, 2013). The size and color of the dot represent the amount of correlation there is between two variables. Since the variables are perfectly correlated with themselves the correlation on the diagonal are one. From this correlation plot it becomes apparent the emotions positive and neutral are negatively correlated (-0.72) as well as positive and negative (-0.42). Additionally, the number of guests a listing accommodates correlates with the number of bedrooms (0.796) and beds (0.853). Lastly, the number of total amenities correlates with the number of bedroom (0.603), additional (0.634), family (0.590), top (0.421), and safety (0.334) amenities.

In order to detect whether there is multicollinearity between variables the Variance Inflation Factor (VIF) is used. For all four models, some of the squares of the generalized VIFs are larger than 4. More specifically the VIF's are higher than four for the number of amenities, host' emotions and number of beds, bedrooms and the amount of guests the listing accommodates. This implies that there is strong multicollinearity between these variables which in turns means that some of these variables have to be removed. The high amount of correlation also became apparent from Pearson's correlation matrix.

To prevent multicollinearity issues to affect this research results this study chooses to exclude the number of amenities of the analysis since it is heavily correlated with the other amenities categories. Consequently, only the amenity categories are used in the analysis. Additionally, the number of beds and bedrooms are excluded as the number of people a listing accommodates can be expected to include this information. Lastly, the neutral facial expression is excluded from the analysis.

In addition, to determine whether the linear relationship assumptions hold the residuals were plotted against the fitted values. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good. However, the plots for the four fitted models in Appendix M show that this does not hold for any of the models. As this remains the same after log transforming the dependent variable this is considered as a limitation.

6.3 Results Linear Regression analysis

Table 6.2 shows the results of the linear regression models. The first model that was estimated contains all the variables from the Lancasterian perspective. Since some of the Lancasterian perspective variables can be seen as control variables the second model that is estimated consist out of the Lancasterian and Social Learning perspective. Similarly, the third model consist out of the Lancasterian and Presentation perspective. The fourth linear model is estimated using all three perspectives. Because of compactness, the entire model including the 18 neighborhood variables is shown in Appendix F.

Table 6.2: *Results from the Linear regression models*

Linear regression models								
	Lancasterian		**Lanc.+Social**		**Lanc.+Presentation**		**Lanc.+Social+Pres.**	
Predictors	*Estimates*	*std. Error*	*Estimates*	*std. Error*	*Estimates*	*std. Error*	*Estimates*	*std. Error*
(Intercept)	0.067	0.073	0.191	0.146	0.224	0.139	0.299	0.179
Lancasterian perspective								
Price	**-0.001** ***	0.000	**-0.001** ***	0.000	**-0.001** ***	0.000	**-0.001** ***	0.000
Accommodates	**0.021** ***	0.006	**0.014** **	0.005	**0.020** ***	0.006	**0.014** **	0.005
Bedroom amenities	**0.021** **	0.008	0.017 *	0.008	**0.021** **	0.008	0.018 *	0.008
Additional amenities	0.011 *	0.005	0.008	0.005	0.012 *	0.005	0.009 *	0.005
Family amenities	0.013	0.010	0.006	0.009	0.013	0.010	0.006	0.010
Safety amenities	**0.027** ***	0.007	**0.023** ***	0.007	**0.028** ***	0.007	**0.023** ***	0.007
Top amenities	**-0.044** ***	0.009	**-0.030** ***	0.008	**-0.045** ***	0.009	**-0.031** ***	0.008
Nr. verification's	0.005	0.004	-0.001	0.004	0.004	0.004	-0.002	0.004
Nr. bathrooms	0.028	0.019	0.038 *	0.018	0.027	0.019	0.037 *	0.018
Host is Superhost	**0.055** ***	0.015	0.036 *	0.014	**0.058** ***	0.015	**0.042** **	0.014
Identity verified	**0.057** **	0.021	**0.056** **	0.020	**0.058** **	0.022	**0.057** **	0.020
License	**0.051** ***	0.015	**0.038** **	0.014	**0.052** ***	0.015	**0.038** **	0.014
Entire home/apartment	0.094	0.052	0.076	0.048	0.071	0.053	0.056	0.050
Private room	0.130 *	0.051	0.074	0.048	0.105 *	0.053	0.052	0.050
Neighborhood	*For compactness, please find in Appendix F*							
Social learning perspective								
Number of Reviews			**0.001** ***	0.000			**0.001** ***	0.000
Rating scores			-0.030	0.027			-0.026	0.027
Presentation perspective								
Host has description					0.008	0.021	-0.007	0.019
Sentiment descr. host					-0.018	0.020	-0.024	0.018
Sentiment descr. listing					0.001	0.015	0.001	0.014
Positive					-0.000	0.000	0.000	0.000
Negative					-0.000	0.000	-0.000	0.000
Sharpness					-0.122	0.115	-0.065	0.108
Brightness					0.053	0.044	0.068	0.041
Average age					-0.001 *	0.001	**-0.002** ***	0.001
Count male					0.021	0.016	0.016	0.015
Count female					0.006	0.014	0.001	0.013
Observations	1421		1421		1421		1421	
R2 / R2 adjusted	0.319 / 0.303		0.399 / 0.384		0.325 / 0.304		0.409 / 0.390	
AIC	169.992		-3.592		176.928		-8.121	
BIC	348.801		185.736		408.329		233.798	

** p<0.05 ** p<0.01 *** p<0.001*

6.3.1 Lancasterian perspective model

The first model that has been fitted only includes the variables from the Lancasterian perspective. This model has been created in order to investigate hypothesis 1a – 1e. Model 1 shows that first of all price has a negative relation with a listing' occupancy rate. The significant coefficient indicates that an increase of price by €1.- decreases the occupancy rate by 0.001, ceteris paribus. This translates in a loss of 0.2 days per euro increase. In other words, the more expensive an accommodation the lower the estimated performance, according to this model.

In addition, the number of guests a listing accommodates positively influences a listings performance. If an accommodation is able to host one additional guest the listing' occupancy rate is expected to increase by 0.021, which translates into 4 days additionally booked, ceteris paribus. Similarly, the number of bedroom amenities increases a listing' performance by 0.021. The Lancasterian model therewith shows that an increase of one bedroom amenity results into an increase of 4 occupied days. Comparably but with a stronger effect the number of safety amenities increase the occupancy rate by 0.027. This translates into an additional of 5 days the accommodation is booked. The safety amenity effect remains positive and significant for all four estimated models.

In contrast, the number of top amenities a listing contains negatively affects the occupancy rate. The negative coefficient with a value of -0.044 can be interpreted as an increase of one top amenity decreases the listings occupancy by 8 days. The negative effect of the top amenities is therewith stronger than that of the other two significant categories.

Furthermore, the host' characteristics are positively related to the occupancy rate. The strongest effects in the Lancasterian model are found for these variables. If a host has a verified identity, is a superhost and has a license this increases the occupancy rate with, 0.057, 0.055 and, 0.051 respectively, the remaining conditions kept the same. This translates in an occupancy gain of 10^+ days when a host' contains one of these binary characteristics.

Interestingly, the Lancasterian models exclusively finds the two neighborhoods within Amsterdam's city center i.e., Center-West and Centre-East, to have a positive coefficient. From these two only Centre-West contains a significant relation based on a 99% confidence interval. That the center neighborhoods are the only positive locations illustrates the importance of having a listing near the city center. These listing' are found to perform better than their more remote counterparts.

6.3.2 Lancasterian Social Learning perspective model

The second model that has been fitted includes both the variables from the Lancasterian perspective and the Social Learning perspective. This model has been created in order to investigate hypothesis 2a and 2b. Based on the AIC and BIC criteria this model significantly improves on the Lancasterian model. This second model thus contains a better fit to the model and so explains the greatest amount of variation using the fewest possible independent variables.

Secondly, the previous effects of the Lancasterian perspective remain consistent for; the price, how many guests the listings accommodate, if a host's identity is verified and whether they got a license and the number of safety- and top-amenities. From the Social Learning perspective only the number of reviews are found to have a significant positive effect on the occupancy rate. The results from this model imply that an increase of one review the occupancy rate for that listing increases by 0.001, ceteris paribus. This translates to an increase of one review ensures that the listing is booked for an additional 2 days.

Unlike the number of reviews the rating score has a negative coefficient. Even though this coefficient is insignificant the linear model does not estimate a positive effect. This coefficient remains negative and insignificant within the full model.

6.3.3 Lancasterian Presentation perspective model

The third model that has been fitted includes both the variables from the Lancasterian perspective and the Presentation perspective. This model has been created in order to investigate hypothesis 3a – 3d. The third model does not significantly improve on the Lancasterian or Lancasterian + Social Learning models according to the AIC and BIC criteria. Adding the Presentation variables to model 1 does thus not result in a better fit.

Nevertheless, the effects of the Lancasterian perspective remain roughly the same. The only difference with model 2 is that being a superhost is once again significant. Being a superhost or having a verified license have the biggest effect on a listing' performance according to this model.

From the Presentation perspective variables only the estimated average age has a moderately significant effect based on a 95% confidence interval. The results of this third model indicate that when age increases by 1 the occupancy rate diminishes by 0.001. This translates into a decrease of 0.2 nights if the average estimated age of a host' is estimated to be one year higher.

Because the characteristics of a host' profile picture could only be determined if there was a functioning URL, only listings that passed this criterion were included. As a robustness check these omitted observations have been included within the model once more while imputing them with both a 0 (see Appendix J) and with the mean (see Appendix K). These models show that the effect of average age remains robust with a significant coefficient of -0.002 in the full model and becomes significant with a coefficient of -0.001 in the Presentation perspective model. Additionally, the zero-imputed regression model (Appendix J) significantly improves upon the Lancasterian model but, the Lancaster + Social Learning model remains superior.

6.3.4 Model containing all three perspective

The fourth model contains the variables from all three perspectives. This model found to be best model based on the AIC but is not found to improve upon model 2 according to the BIC. The information criteria thus both have a preference for a different model. The AIC prefers the full model whereas the BIC prefers the model containing the Lancasterian + Social Learning perspective variables. This is likely due to the fact that BIC penalizes more for model complexity. The second model containing the Lancasterian + Social Learning perspective variables thus contains the best-fit to the data and so explains the greatest amount of variation using the fewest possible independent variables according to the BIC.

In contrast to model 3 the full model finds the average estimated age to significantly decrease a listing' performance. According to the fourth model if the average estimated age increases by 1 the occupancy rate decreases by 0.4 days. Furthermore, the effects of the Lancasterian and Social Learning perspective remain constant. In contrast to model 1 only the number of bedroom amenities does not remain significant. These results remain the same even after standardizing all numeric variables for all models, see Appendix I

In order to further investigate the underlying linear relationships a step-wise AIC and BIC model have been created which are shown in Appendix H. These step-wise models have been created using both forward and backward selection. From these step-wise models, it becomes apparent that from the Lancasterian perspective the price, number of safety and bedroom amenities, top amenities (remains negative), being a superhost, having their identity verified and having a license remains significant and are therefore considered to be important prediction features by both selection criteria. In the case of the Social Learning perspective, the number of reviews remains significant for both models while containing the same coefficient of -0.001. Similarly for the Presentation perspective, the average age appears to be important for the goodness of fit of the models. In addition, the AIC and BIC step-wise models also value the sentiment of the host' self-description which is negatively related to the occupancy rate. Encouraged by the earlier found insufficiency of heteroscedasticity this research moves on to the tree-based models which allow for possible nonlinear relationships.

6.4 Conditional Inference Tree

The conditional inference tree (C.I.T.) model is created to investigate the effects between the three perspectives and a listing' performance. Since the C.I.T. does not make any assumptions about the linearity of the relation it is able to make non-linear predictions about the data. By default, the C.I.T. creates the nodes based on the condition that the p-value should be lower than 0.05. In Figure 6.2 the results of the C.I.T. are shown. Each node represents a split based on the variable described in the circle with its accompanying p-value. The branches of the node display the values on which the split is based. It is noteworthy, that the C.I.T. contains variables from every perspective.

The C.I.T. in Figure 6.2 can be interpreted as follows: if the number of reviews is ≤ 74 and the room type is a private the predicted occupancy rate is 0.287. If the room type however is an Entire home / apartment or Hotel room and the host is a super host with 1 or less bedroom amenities the expected occupancy rate is 0.162

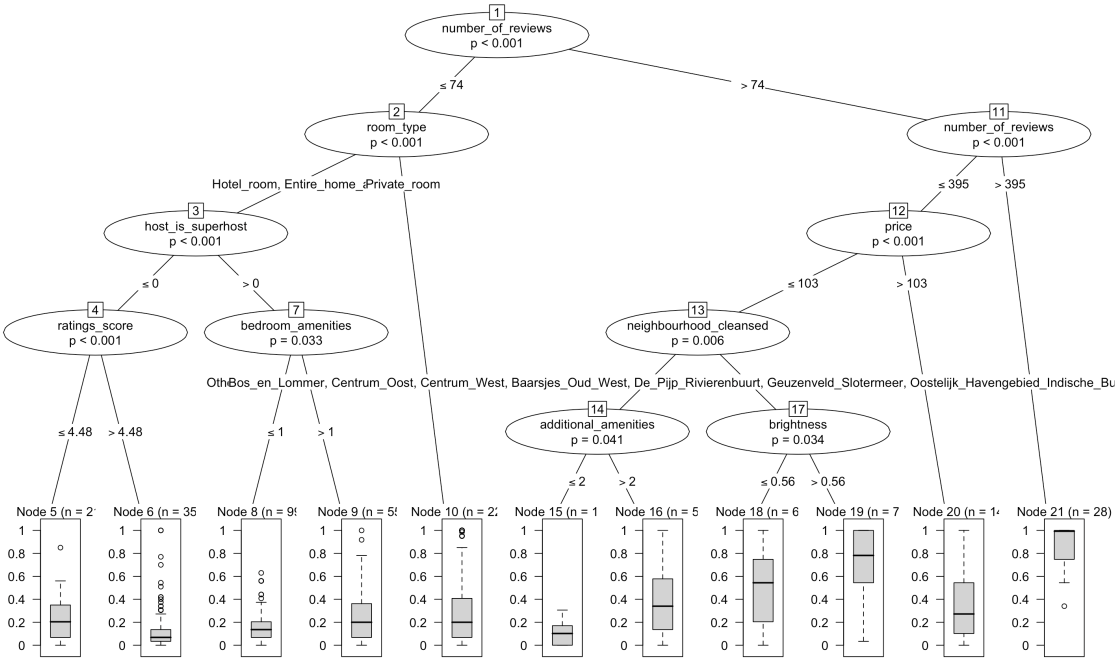


Figure 6.2: *Results from the Conditional Inference Tree plot*

What is interesting to note is that the number of reviews is responsible for two large splits. Thus, in accordance with the linear models, the number of reviews is also found to be an influential variable by the conditional inference tree. If a listing has more than 395 reviews the models expects this listing to have the highest expected occupancy rate.

Also, the total rating score is considered to contain predictive power for the occupancy rate where a listing with a rating equal to or below 4.48 is expected to have a higher occupancy rate than a listing with a rating above 4.48. Besides these Social Learning variable most of the remaining variables belong to the Lancasterian perspective. The listing' characteristics, i.e., room type, number of bedroom and additional amenities, the neighborhood, and, price, are all found to influence the occupancy rate.

Lastly, the neighborhood and amount of brightness in the listing' main image are found to influence the occupancy rate. If a listing has a review amount of 75-395 and a price of $\leq$ 103 and is situated Bos Lommer/ Centrum-Oost/ Centrum-West/ Baarsjes / Oud-West/ De Pijp / Rivierenbuurt/ Geuzenveld/ Slotermeer / Oostelijk-Havengebied/ Indische-Buurt/ Oud-Noord/ Oud-Oost and the brightness is above 0.56 the expected occupancy rate is 0.738. It is important to note here is that the split of the city' neighborhoods is based on how centrally located these neighborhoods are. In other words, more central locations are predicted to perform better by the conditional inference tree. A visual presentation hereof can be found in Appendix L.

Since the conditional inference tree main node is the number of reviews this variable can be considered as the most influential according to this model. The lower the variables are in the three the lower their ability to split the data into significant subsets. It is important to note however that only nine out of the potential twenty-seven independent variables are selected by the model to create the tree. An overview of the C.I.T. split criteria and predicted occupancy rates within the end nodes van be found in Appendix B.

6.5 Random Forest Regression

In addition to the linear models and conditional inference tree a random forest model is fitted. The random forest model is created to be able to further investigate the relations between the three perspectives and a listing' performance. One of the benefits of fitting a random forest model is that the model is not drawn to strong predictors and is capable of predicting non-linear relationships. Since the random forest model does not provide coefficients like a linear model or split points like the C.I.T. we have to resort to alternatives. One of these alternatives are the in paragraph 5.3 explained SHAP values.

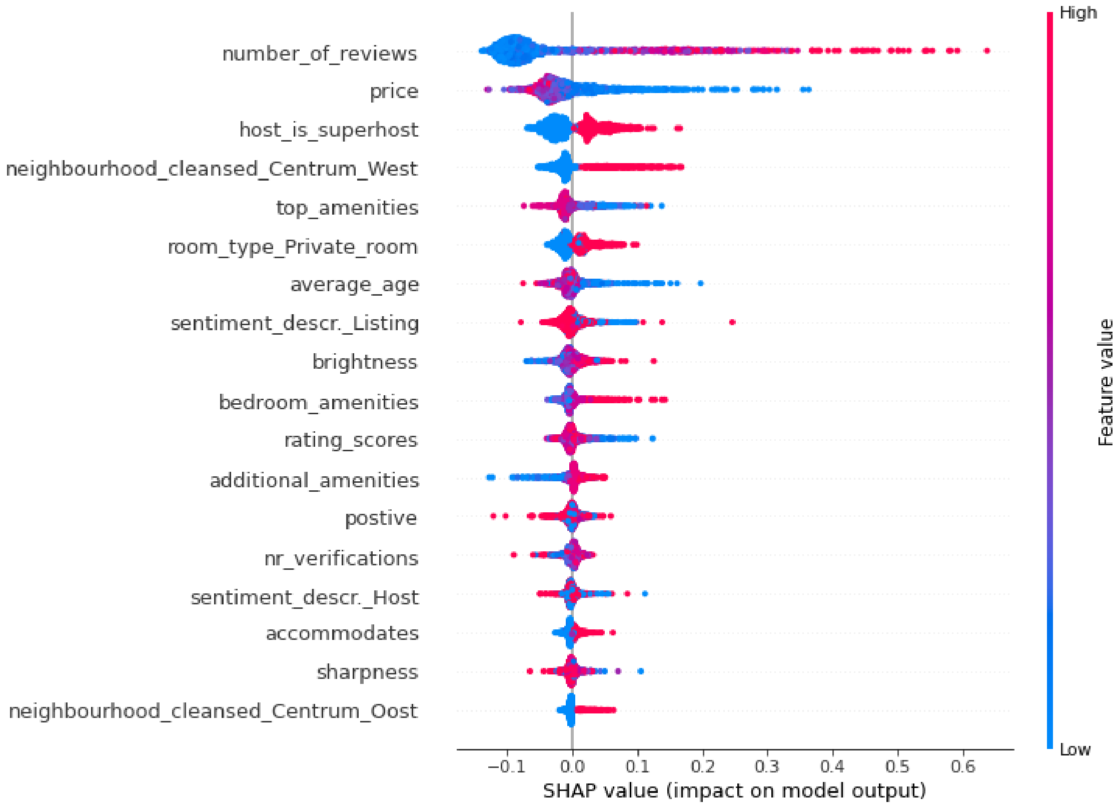


Figure 6.3: *SHAP values from the Random Forest model*

Figure 6.3 shows a summary plot of the predicted SHAP values. The summary plot provides information about many observations and separated for each feature. The color represents the actual feature value, blue is a low value and red is a high value respective to that specific features scale. Each dot is thus a predicted observation with its corresponding feature value. On the X-axis Figure 6.3 shows the SHAP values which, in this case, is the impact on the occupancy rate. Furthermore, the density of the dots represents the amount of observations with the corresponding SHAP value. The features are ordered by their mean average SHAP value which means they are arranged based on their average impact on listings performance. The features in Figure 6.3 all have a mean average SHAP value of 0.01^{+}.

6.5.1 Random Forest - Lancasterian perspective

Multiple of the Lancasterian perspective variables are included within the summary plot of the Random Forest model. The price, location, number of amenities, room type and host characteristics are found to influence a listings performance.

First of all the price. As illustrated in Figure 6.3 price is found to have a negative relation with a listings occupancy rate. If a listing has a lower price the Random Forest model predicts the listing to have higher occupancy rates. However, a low price does not necessarily mean a higher occupancy rate. As the distribution of the price illustrates there are a lot of medium priced listings which are still found to have a negative impact on the occupancy. Nevertheless, there is a clear negative relation between price and performance. This finding is in line with the both the linear regression and conditional inference tree model.

Secondly, the binary indicator of being a superhost is found to have positive relation with the occupancy rate. Figure 6.3 shows a clear distinction between being a super host (red) and not (blue). Additionally, the number of a host's verification's has an effect on the occupancy rate. For this host-characteristic variable however the relation is more ambiguous. This is likely do to the interaction effect there exist between the number of reviews and the number of a host's verification's (see Appendix N). There seems to be a negative relation between the performance of listings with a medium to high number of reviews and the number of verification's a host has.

Thirdly, the location of a listing is found to influence the amount of days a listing is occupied. Listings that are located in Amsterdam's city centre, both East and West, are predicted to have higher occupancy rates. Especially those located on the West-side of the city centre are found to boost a listings performance a lot. This result for the West-side of the city centre is found by all three models and the East-side by both tree-models.

Furthermore, the number of top amenities is found to have a negative impact on a listings performance. Listings which have less than average top amenities are found to have better performances than those with average or more amenities. In contrast to the top amenities the number of bedroom and additional amenities show to have a positive relation with a listings performance. From these two the number of bedroom amenities has a stronger positive effect. The number of additional amenities also has a positive effect but especially shows a stronger negative impact. In both cases however, more of amenities of these categories result in higher occupancy rates. Additionally, just as the results from the linear regression showed the negative impact of the top amenities is larger than the positive effect of the bedroom or additional amenities.

In addition, an accommodations characteristics are found to influence the occupancy rate. Both the room type as well as the number of guests a listing accommodates can potentially increase listing performance. If the room type is a private room the amount of days a listing is booked is predicted to be higher. Similarly the more guests a listing is able to accommodate the higher the occupancy rate according to the Random Forest model.

6.5.2 Random Forest - Social learning perspective

The variable with the most impact on a listings performance is found to be the number of reviews, according to the Random Forest model. The number of reviews is negatively correlated with a listings performance. The density of the number of reviews bar shows that there is a great amount of low number of review listings which have a negative impact of -0.1 on their performance. In contrast, how higher the number of reviews of a listing are the higher the predicted performance of that listing, with an impact ranging from 0.1-0.6^{+}. In addition also the rating scores are found to impact a listings performance. However, the Random Forest model penalizes for higher number of ratings and predicts higher performance for lower ratings. This finding is in line with the results of the Conditional inference tree.

6.5.3 Random Forest - Presentation perspective

Finally, the variables of the presentation perspective are also found to impact the occupancy rate. The characteristics from a host's profile picture are found to influence the performance. The average estimated age is negatively related to the impact on a listings performance. Hosts with younger-looking profile picture are predicted to have higher occupancy rates. Notably, Figure 6.3 shows that the average estimated age (44.4) does not negatively impact a listings performance a lot but, higher estimated average ages do. Furthermore, the host's emotion has an ambiguous relation with a listings performance. The summary plot does not show a clear pattern however, there seems to be an interaction effect between the number of reviews and the emotion positive which results in this scattered relation. The interaction plot (see Appendix N) shows the possibility that for listings with low number of reviews, a positive profile picture results in higher occupancy rates. However, listings with more reviews tend to have a negative relation with the positive expressed emotion on a host's profile picture.

Furthermore, the quality of the listings main-image is found to positively impact the performance. Listings that have a brighter picture are found to have a higher occupancy rate. In contrast, the sharpness of a picture seems to be negatively related to a listings performance. A sharper listings image reduces an accommodations performance whereas a more blurry picture increases the predicted performance.

Lastly, the sentiment within the description of both the host and the listing are found to influence listing performance. In general, the model finds listings with a more negative host self-description to perform better. However, a positive sentiment still has, for some of these listings, a beneficial impact on the occupancy rate. In addition, a listings sentiment seems to have an interaction effect with whether the host is a super host or not (see: Appendix N). The interaction effect between hosts with a superhost status and a negative listing description sentiment seems to increase a listings performance. However, since the main effects of being a superhost is positive and having a negative sentiment description is negative, the interaction effect merely diminishes the main effects. That is to say, if you are a superhost and have a negative sentiment description, the negative sentiment description affects you less. Furthermore, the sentiment of the host's self-description has an ambiguous relation with a listings performance. However, in general the plot shows that high sentiment is associated with lower listing performance whereas lower sentiment is associated with higher listing performance.

Figure 6.4 shows a bar plot with the mean average SHAP values for each feature which was also depicted in Figure 6.3. In Appendix G, a bar plot containing the mean average SHAP value for all features is included.

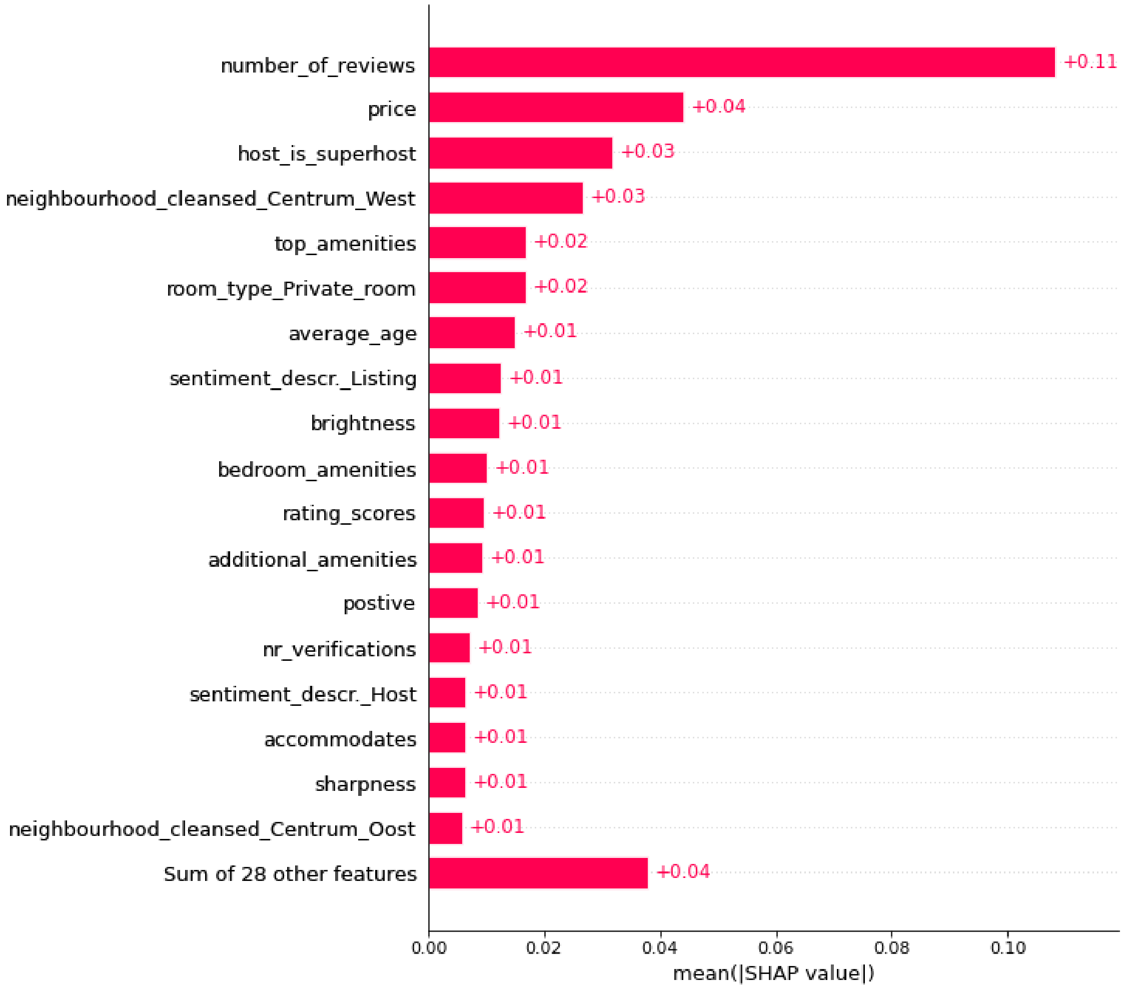


Figure 6.4: *Barplot from the Random Forest SHAP values.*

6.6 Comparing the models

In order to compare the models predictive performance this research will use the Root Mean Squared Error (RMSE). As explained in paragraph 5.2.4, the RMSE provides information about the average distance between the predicted listing performance and the actual listing performance in the dataset. A lower RMSE thus indicates that the model is better at predicting the occupancy rate. For the sake of clarity the occupancy rate is measured on a scale of 0 to 1 and has a standard deviation of 0.304.

Table 6.3 shows the estimated RMSE of the predictions from the models on the test set. In order to do so, the data was separated into an 80% train and 20% test data. First of all the RMSE shows that the best performing linear model is the full model, only outperforming the Lancasterian Social Learning perspective by 0.01. From the tree-based models the Random Forest model proofs to be able to estimate the true occupancy rate better than the conditional inference tree. With a RMSE of 0.201 the Random Forest model is also the best model at estimating the relationship between the three perspectives and a listings performance. This is possibly due to the fact that that the Random Forest model is better at estimating lower level occupancy rates (see: Appendix O). Considering the scale of the occupancy rate a RMSE of 0.201 can be considered reasonably well.

Table 6.3: *Root Means Squared Error of the estimated models*

	RMSE
Linear regression models	
Lancasterian perspective	0.247
Lancasterian & Social Learning perspective	0.238
Lancasterian & Presentation perspective	0.249
Lancasterian, Social Learning, & Presentation perspective	0.237
Tree-based models	
Conditional Inference Tree	0.259
Random Forest	0.201

Chapter 7

Conclusion and Discussion

The aim of this research was to investigate how Airbnb hosts can position their accommodation to achieve the highest possible listing performance. In order to do so this research investigated the effects from three perspectives namely the Lancasterian perspective, Social Learning perspective, and, Presentation perspective. By doing so, this research unifies all the existing literature into one research paper while simultaneously controlling the effects for all the perspectives. Additionally, this paper is the first to study all Presentation perspective variables in one research. To examine these three perspectives, linear regression, conditional inference tree, and, random forest models have been employed. The results show that each perspective contains features which influence the eventual performance of a listing.

7.1 Lancasterian perspective

The Lancasterian perspective refers to the idea that the characteristics of a product and not the product itself is that which determines the eventual choice of a consumer. In the case of Airbnb this refers to idea that characteristics of a listing, i.e., the amenities, location, size, a superhost badge and price, determine the choice from a potential Airbnb guest. In order to do so, this research investigated the listing characteristics that have been proven to be influential for a listings performance by previous research.

The hypothesis for the Lancasterian perspective state that the central location of the listing (1a), room type (1b), a superhost badge (1c), number of amenities (1d), having a verified identity (1e), having a licence (1f), and the number of guest a listing accommodates (1h) positively influence listing performance. In contrast, the listings price (1g) is hypothesized to have a negative relation with the listing performance. Hypothesis 1a is investigated by evaluating the listing' location by considering its neighborhood. All three models show that the neighborhood of a listing matters for its eventual performance. Especially listings that are situated within the west-side of Amsterdam' city centre are found to have higher occupancy rates. In addition, the Conditional Inference Tree (C.I.T.) separates the listings by their neighborhood. This node splits the higher from lower performing listings by their respective neighborhoods. In order to do so the C.I.T. separates the neighborhoods central neighborhoods from those more remote. In other words, the C.I.T. indicates that listings that are closer to the center perform better. An overview can be found in Appendix L. Similarly to, von Hoffen et al. and Masiero et al. (2015) this research finds that the more proximate a listing is to the city center the better its performance.

Secondly, the tree based models show that the room type matters for the listing performance (1b). Both tree models find that if an Airbnb accommodation is a private room the occupancy rate is higher. Interestingly, a private room is thus considered to be favorable in comparison to a hotel room or an entire home / apartment. These findings are in line with the results of Biswas et al. (2020). The preference for a private room can be partially explained by the fact that 22% of the Amsterdam's tourism in 2020 was business related (Amsterdam, 2022b). Nevertheless, this research found supporting evidence for the fact that the room type matters for a listings performance.

In addition, the number of guests a listing accommodates (1h) has been found to positively influence listing performance by the Random Forest model. The Random Forest model especially showed a clear effect of the added value from high numbers that a listing is able to accommodate. Since almost every room type that is able to host more than three people is an entire / home apartment this suggests that Amsterdam's tourist also have a preference for larger accommodations. This in combination with the above indicates that the preference for Amsterdam tourist on the Airbnb platform is 1), a private room, 2), an entire home / apartment and 3), a hotel room. This finding is in line with Kwok and Xie (2019).

Furthermore, having a superhost badge (1c), verified identity (1e), and licence (1f) are all three found to positively influence the occupancy rate. In other words, the host's characteristics are found to influence listings performance. Having a superhost badge is found to positively influence performance by all three models whereas having a verified ID and license were only found to be significant by the linear models. This implies that guests value these trust and quality features of an Airbnb listing. These findings are in line with research from Xie and Mao (2017) and Zhang and Luo (2018). The findings for having a license however, have, to this research knowledge, not been found before. It would therefore be interesting to investigate the effects hereof in different cities.

Additionally, the number of amenities have been investigated to answer hypothesis 1d. The number of amenities have been divided into six groups of amenities in accordance with Airbnb's own categorization. Since none of the investigated listings had any of the accessibility amenities this category was later excluded from the analysis. The remaining five categories, bedroom, safety, additional, top and family amenities have been investigated. The number of bedroom and additional amenities have been found to positively influence listing performance by the conditional inference tree and random forest model. Also, the number of safety amenities remained positive and significant through all four linear regression models. For these types of amenities, we can state that in general adding more of these amenities to a listing is beneficial for its performance. In contrast, the number of top amenities was found to negatively influence the listings performance by the random forest model. Since the top amenities can be considered the more luxurious amenities this relation may lay in the fact that the more luxurious segment is booked less. Listings with more than average of four top amenities are indeed more expensive, ≈ €60.-, and have a lower occupancy rate.

Lastly, as expected the price (1g) is found to have a negative relation with a listings performance. A higher price for Airbnb listings in Amsterdam is therefore found to decrease a listings performance. This finding is in line with research from Xie and Mao (2017).

7.2 Social Learning perspective

Besides the Lancasterian perspective this research also investigated the effects of the Social learning perspective. The Social Learning perspective refers to the process in which consumers share their experiences and opinions about a product with other consumers by means of ratings and reviews. As reviews amass, consumers will be able to obtain better assessments of the quality of the products, enabling them to make a better and well-considered purchasing decision. In other words, consumers are able to use online reviews and ratings to evaluate which products are more trustworthy and so help them make their purchase decision (Kim Park, 2013). Consequently, this perspective is expected to influence the performance of a listing where having more reviews (2a) and higher ratings scores (2b) were expected to drive performance.

The number of reviews (2a) have been found to positively influence the listings performance. This is in line with previous research. According to the linear models an increase of the amount of reviews by one adds 0.001 to the occupancy rate which translates into 2 nights. The conditional inference tree also associates higher review numbers with higher occupancy rates. Lastly, the random forest model finds that high review number translate into high expected occupancy rates. The random forest additionally, penalizes heavily for low amounts of review numbers. Reviews are thus found to positively influence a listing' performance in which more is better. This is in line with research from Abramova et al. (2017).

In contrast, the tree based models found higher ratings (2b) to be associated with lower occupancy rates. This finding is similar to that of the research from Chen and Chang (2018) where they did not find an effect of rating score on purchase intention, since there has to be a purchase in order to increase a listings performance. Hypothesis 2b is therefore rejected since higher ratings do not necessarily translate into higher listing performance. This could be explained by the fact that Airbnb ratings are mostly high which results in potential guests using other features to establish their consumer decision on. Another explanation could be that potential guests deem slightly lower ratings, i.e., lower than 4.48 stars as found by the conditional inference tree, more trustworthy.

7.3 Presentation perspective

The third investigated perspective is the Presentation perspective. On Airbnb host are able and encouraged to present themselves and their listing in the best way to attract as many guests as possible. This research has investigated the effects of these textual and visual presentation variables by analyzing both the text-description from the listing and host as well as their accompanying pictures. This research expected sentiment in both the host self-description (3a) and listing description (3b) to positively affect listing performance, where more sentiment is better. In addition, the characteristics of the host's profile picture (3c) and listing' main-image quality (3d) were expected to increase listing performance.

The sentiment of both the host' self-description and listing' description have been extracted using DistilBERT which is a distilled version of Google's BERT algorithm. The sentiment of the host' self-description (3a) has been found to have a negative relation with listing performance. This means that, in contrast to the expectations, a negative sentiment in the self-description results in a higher listing-performance. Similarly, the listing description (3b), influence the occupancy rate negatively according to the random forest model. Lower sentiment description is on average associated with higher listing performance. This is not to say that having sentiment in the listing description decreases listing performance since Figure 6.3 shows that having sentiment can also add to listing performance. However, having low sentiment adds relatively more to listings performance. These findings are in line with research from Zhang et al. (2020) since they found an that excessive sentiment in the description has a negative effect on performance.

One of the characteristics of the host profile picture (3c) is found to influence the listings performance. The estimated average age of a host' profile picture is found to be negatively associated with listing performance. Host which have a younger-looking profile picture are thus found to have higher predicted listings performances. Unlike previous research this study thus found the age and not the emotion to be of significant influence.

Lastly, the quality of a listings main-image (3d) was investigated. The sharpness, brightness and contrast – three classic image quality measures – were extracted from the main-image. Since contrast and sharpness were heavily correlated only the sharpness of the picture was considered.

Both the tree based models found the brightness of the picture to be positively associated with listing' performance. The conditional inference tree attributes higher occupancy rates to picture with a higher brightness than 0.56. Similarly, the results of the random forest model in Figure 6.3 show that more brightness results in higher occupancy rates. However, the sharpness of the main-image is not found to influence the performance. Thus, the quality of the listings main-image only increases performance by its brightness, where brighter is better. This is in line with Airbnb's own recommendations.

7.4 Implications

This research has uniquely combined the influence of the Lancasterian, Social Learning and Presentation perspective to analyze there relationship to a listings performance. By doing so, this study has analyzed the influences of these three perspectives both separately as well as mutually controlling for each other. Hence, this research offers meaningful insights in concerning the consumers decision making process within the sharing economy.

First of all, the results show that the number of amenities is positively associated with listings performance. Especially the number of additional-, bedroom- and safety- amenities have been found to boost performance (see: Appendix A for list of amenities). Airbnb hosts that have not included all these amenities are encouraged to do so. These can be relatively low-effort changes such as providing towels and soap or providing an emergency plan and local numbers. In contrast, the number of top amenities are found to negatively impact a listings performance. As mentioned earlier this could possibly be attributed to the fact that these amenities can, in general, be considered the more luxurious amenities. This luxurious segment is booked less on average.

Secondly, it has been found that the brightness of a picture is positively related to a listings performance. Especially for listings that can be contain the following characteristics: review amount between 75-395 and a price of $\leq$ 103, situated in one of the following neighborhoods; Bos Lommer; Centrum-Oost; Centrum-West; Baarsjes; Oud-West; De Pijp; Rivierenbuurt; Geuzenveld; Slotermeer; Oostelijk-Havengebied; Indische-Buurt; Oud-Noord; or Oud-Oost, it would be beneficial to increase the brightness of the listing' main-image, if the brightness is not already above 0.56. The effect of brightness on performance is also found for the other listings in general. Furthermore, these aforementioned neighborhoods are all located near Amsterdam's city centre. Especially the listings located within the west-side of Amsterdam's city centre are found to increase performance.

Thirdly, the results show that the average estimated age of a host's profile picture negatively impacts a listings performance. This may be because the listings of younger hosts are related to more favorable listings. Nonetheless, uploading a young-looking photo is a small effort change that can achieve beneficial results.

Lastly, the sentiment of a listings description has a complex relation with its performance. Listing description which are perceived to be negative are found to have higher occupancy rates. This does not mean that listings with high sentiment do not only, higher sentiment is more often associated with lower increases in performance or decreases in performance depending on the other characteristics. However, this implies that guests prefer listings that are not overly positive. likely because guests want to specific information from the description and to a lesser extent want the listing to express a certain emotion.

7.5 Limitations

This thesis contains some limitations which future research can improve upon. First, even though this study examines a wide range of variables there still remain variables and extraction methods that are not included or used within this research. For example, this research did not investigate the information concealed within the reviews. Additionally, this research only considered the sentiment that is found within the both the host and listing description. Future research could improve upon this study by investigating these unconsidered methods and variables.

Secondly, this research has only been able to investigate the effect of a listings main-image quality. Since a listing can contain up to 100 photos this research only considers a fraction of the true number of pictures. The main-image is nevertheless the first image a potential guest encounters and therefore determines the most whether or not a guest will consider that specific listing. This image additionally sets the standard for subsequent photographs. It is nevertheless interesting to examine how the quality effect of the subsequent photos influence the demand for that listing. Furthermore, this study only looked at three image quality measures namely sharpness, brightness and contrast. It would therefore be interesting to see if the results remain robust for different quality measures.

Thirdly, this research was only able to investigate the listings that both contained a link to the host and listing picture and in addition were also feasible to examine. Some of the pictures did provide a link but Face ++ and SightEngine were not able to examine these. Therefore, these listing were removed from the scope. Moreover, the photo analyzed may no longer correspond to the photo that was online from July to December. Working with historical data can therefore give a distorted idea of the actual relationships.

Furthermore, this research used the neighborhood as a variable to investigate the effects of the location of an accommodation on its performance. Since this is not the most accurate possible variable I urge future research to investigate the effects of listing location relative to popular attractions within a city. In the case of Amsterdam this would be possible by using the openly available data from the Central Bureau of Statistics (CBS), available at www.cbs.nl.

Lastly, due to COVID-19 this research has only been able to investigate the time period of May nineteenth till December fifth. Even tough national tourism was back to pre-Covid-19 levels in 2021 the number of international tourists were still 71% lower compared to pre-COVID-19 year 2019 (NBTC, 2022). It would therefore be interesting to conduct this study again when tourism has returned to its former level.

Bibliography

Bruno Abrahao, Paolo Parigi, Alok Gupta, and Karen S Cook. Reputation offsets trust judgments based on social biases among airbnb users. *Proceedings of the National Academy of Sciences*, 114(37):9848–9853, 2017.

Olga Abramova, Hanna Krasnova, Chee-Wee Tan, et al. How much will you pay? understanding the value of information cues in the sharing economy. In *ECIS*, page 66, 2017.

Sanjeev Agarwal, M Krishna Erramilli, and Chekitan S Dev. Market orientation and performance in service firms: role of innovation. *Journal of services marketing*, 2003.

Airbnb. Airbnb home: Vacation rentals, cabins, beach houses, unique homes amp; experiences. URL `https://www.airbnb.com/s/Amsterdam--Netherlands/homes?tab_id=home_tab&amp;refinement_paths%5B%5D=%2Fhomes&amp;flexible_trip_dates%5B%5D=may&amp;flexible_trip_lengths%5B%5D=weekend_trip&amp;date_picker_type=calendar&amp;query=Amsterdam%2C+Netherlands&amp;place_id=ChIJVXealLU_xkcRja_At0z9AGY&amp;source=structured_search_input_header&amp;search_type=autocomplete_click`.

Airbnb, 2022a. URL `https://www.airbnb.com/resources/hosting-homes/a/the-amenities-guests-want-25`.

Airbnb, 2022b. URL `https://www.airbnb.com/help/article/317/select-your-home-type`.

Airbnb, 2022c. URL `https://www.airbnb.com/help/article/829/how-to-become-a-superhost`.

Airbnb, 2022d. URL `https://www.airbnb.com/resources/hosting-homes/a/take-great-listing-photos-with-your-phone-14`.

Airbnb, 2022e. URL `https://www.airbnb.com/resources/hosting-homes/a/take-great-listing-photos-on-your-smartphone-221`.

Hirotugu Akaike. *Akaike's Information Criterion*, pages 25–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_110. URL `https://doi.org/10.1007/978-3-642-04898-2_110`.

Jay Alammar. The illustrated bert, elmo, and co. (how nlp cracked transfer learning), 2022. URL `https://jalammar.github.io/illustrated-bert/`.

Shivaji Alaparthi and Manit Mishra. Bert: A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9(2):118–126, 2021.

Gemeente Amsterdam. Particuliere vakantieverhuur, Apr 2022a. URL `https://www.amsterdam.nl/wonen-leefomgeving/wonen/vakantieverhuur/`.

Gemeente Amsterdam. Hoeveel toeristen komen er naar amsterdam en regio?: Website onderzoek en statistiek, 2022b. URL `https://onderzoek.amsterdam.nl/interactief/toerisme-in-amsterdam`.

Joonheui Bae and Dong-Mo Koo. Lemons problem in collaborative consumption platforms: Different decision heuristics chosen by consumers with different cognitive styles. *Internet Research*, 2018.

Stuart J Barnes. Understanding the overvaluation of facial trustworthiness in airbnb host images. *International Journal of Information Management*, 56:102265, 2021.

Yochai Benkler. Sharing nicely: On shareable goods and the emergence of sharing as a modality of economic production. *Yale Lj*, 114:273, 2004.

Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W Moe, Oded Netzer, and David A Schweidel. Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1): 1–25, 2020.

Baidyanath Biswas, Pooja Sengupta, and Dwaipayan Chatterjee. Examining the determinants of the count of customer reviews in peer-to-peer home-sharing platforms using clustering and count regression techniques. *Decision Support Systems*, 135:113324, 2020.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Michel Briene, Elvira Meurs, Daan Krins, and Nick Rundberg. Tourism in amsterdam today and tomorrow, 2021. URL `https://news.airbnb.com/wp-content/uploads/sites/4/2021/05/Airbnb-Report-on-Travel-Living.pdf`.

Peter Broeder and Kyra Crijns. Self-disclosure and trust on airbnb: a cross-cultural perspective. *Storytelling across platforms: Managing corporate and marketing communications from a storytelling perspective*, pages 160–171, 2019.

F Brousseau, J Metcalf, and M Yu. Analysis of the impacts of short term rentals on housing. *San Francisco, CA: City and County of San Francisco*, 2015.

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. Analyzing elmo and distilbert on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, 2020.

Manojit Chattopadhyay and Subrata Kumar Mitra. Do airbnb host listing attributes influence room pricing homogenously? *International Journal of Hospitality Management*, 81:54–64, 2019.

Chia-Chen Chen and Ya-Ching Chang. What drives purchase intention on airbnb? perspectives of consumer reviews, information quality, and media richness. *Telematics and Informatics*, 35 (5):1512–1523, 2018.

Dongyu Chen, Hao Lou, and Craig Van Slyke. Toward an understanding of online lending intentions: Evidence from a survey in china. *Communications of the Association for Information Systems*, 36(1):17, 2015.

Yong Chen and Karen Xie. Consumer valuation of airbnb listings: A hedonic pricing approach. *International journal of contemporary hospitality management*, 2017.

Mingming Cheng and Xin Jin. What do airbnb users care about? an analysis of online review comments. *International Journal of Hospitality Management*, 76:58–70, 2019.

Yeojin Chung and Surendra Sarnikar. Understanding host marketing strategies on airbnb and their impact on listing performance: a text analytics approach. *Information Technology & People*, 2021.

David Dann, Timm Teubner, and Christof Weinhardt. Poster child and guinea pig–insights from a structured literature review on airbnb. *International Journal of Contemporary Hospitality Management*, 2018.

Iviane Ramos de Luna, Àngels Fitó-Bertran, Josep Lladós-Masllorens, and Francisco Liébana-Cabanillas. *Sharing Economy and the Impact of Collaborative Consumption.* IGI Global, 2019.

Chaoqun Deng and T Ravi Ravichandran. To smile or not? the effect of facial expression on service demand in sharing economy platforms. 2020.

Chekitan S Dev, Rebecca W Hamilton, Roland T Rust, and Matthew V Valenti. What do hotel guests really want? anticipated versus actual use of amenities. 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics*, 9(2):1–22, 2017.

Eyal Ert, Aliza Fleischer, and Nathan Magen. Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism management*, 55:62–73, 2016.

R. Evtimov, M. Falli, and A. Maiwald. Bidirectional encoder representations from transformers (bert), Feb 2020. URL `https://humboldt-wi.github.io/blog/research/information_systems_1920/bert_blog_post/`.

Asle Fagerstrøm, Sanchit Pawar, Valdimar Sigurdsson, Gordon R Foxall, and Mirella Yani-de Soriano. That personal profile image might jeopardize your rental opportunity! on the relative impact of the seller's facial expressions upon buying behavior on airbnb™. *Computers in Human Behavior*, 72:123–131, 2017.

Haoqiang Fan, Zhimin Cao, Yuning Jiang, Qi Yin, and Chinchilla Doudou. Learning deep face representation. *arXiv preprint arXiv:1403.2802*, 2014.

Koen Frenken and Juliet Schor. Putting the sharing economy into perspective. In *A research agenda for sustainable consumption governance*. Edward Elgar Publishing, 2019.

Jim Frost. Multicollinearity in regression analysis: problems, detection, and solutions. *Statistics by Jim*, 2017.

Joe Gebbia. How airbnb designs for trust - joe gebbia ted talk, Feb 2016. URL `https://www.ted.com/search?q=How%2BAirbnb%2Bdesigns%2Bfor%2Btrust`.

Alina Geiger, Chris Horbel, and Claas Christian Germelmann. "give and take": how notions of sharing and context determine free peer-to-peer accommodation decisions. *Journal of Travel & Tourism Marketing*, 35(1):5–15, 2018.

Chris Gibbs, Daniel Guttentag, Ulrike Gretzel, Jym Morton, and Alasdair Goodwill. Pricing in the sharing economy: A hedonic pricing model applied to airbnb listings. *Journal of Travel & Tourism Marketing*, 35(1):46–56, 2018.

Daniel Guttentag. Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current issues in Tourism*, 18(12):1192–1217, 2015.

Daniel Guttentag, Stephen Smith, Luke Potwarka, and Mark Havitz. Why tourists choose airbnb: A motivation-based segmentation study. *Journal of Travel Research*, 57(3):342–359, 2018.

Michael C Hall and Allan Williams. *Tourism and innovation (2nd edn.)*. Routledge, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Random forests. In *The elements of statistical learning*, pages 587–604. Springer, 2009.

Sri Rahayu Hijrah Hati, Tengku Ezni Balqiah, Arga Hananto, and Elevita Yuliati. A decade of systematic literature review on airbnb: the sharing economy from a multiple stakeholder perspective. *Heliyon*, 7(10):e08222, 2021.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

Liyang Hou. Destructive sharing economy: A passage from status to contract. *Computer Law & Security Review*, 34(4):965–976, 2018.

InsideAirbnb. Insideairbnb, 2022. URL `http://insideairbnb.com/`.

Bastian Jaeger, Willem WA Sleegers, Anthony M Evans, Mariëlle Stel, and Ilja van Beest. The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology*, 75:102125, 2019.

Bram Janssens, Matthias Bogaert, and Dirk Van den Poel. Evaluating the influence of airbnb listings' descriptions on demand. *International Journal of Hospitality Management*, 99:103071, 2021.

Jiwon Jung and Kun-Pyo Lee. Curiosity or certainty? a qualitative, comparative analysis of couchsurfing and airbnb user behaviors. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1740–1747, 2017.

Jiwon Jung, Susik Yoon, SeungHyun Kim, SangKeun Park, Kun-Pyo Lee, and Uichin Lee. Social or financial goals? comparative analysis of user behaviors in couchsurfing and airbnb. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 2857–2863, 2016.

Qing Ke. Sharing means renting? an entire-marketplace analysis of airbnb. In *Proceedings of the 2017 ACM on web science conference*, pages 131–139, 2017.

Jeonghye Kim, Youngseog Yoon, and Hangjung Zo. Why people participate in the sharing economy: A social exchange perspective. In *19th Pacific Asia Conference on Information Systems, PACIS 2015*. Pacific Asia Conference on Information Systems, 2015.

Anastasia Kotelnikova, Danil Paschenko, Klavdiya Bochenina, and Evgeny Kotelnikov. Lexicon-based methods vs. BERT for text sentiment analysis. *CoRR*, abs/2111.10097, 2021. URL `https://arxiv.org/abs/2111.10097`.

Sandeep Kulshreshtha and Ruchika Kulshrestha. The emerging importance of "homestays" in the indian hospitality sector. *Worldwide Hospitality and Tourism Themes*, 2019.

Linchi Kwok and Karen L Xie. Pricing strategies on airbnb: Are multi-unit hosts revenue pros? *International Journal of Hospitality Management*, 82:252–259, 2019.

Monroe Labouisse. Airbnb announces "verified identification", 2013. URL `https://www.airbnb.nl/press/news/airbnb-announces-verified-identification#:~:text=%E2%80%9CIn%20a%20marketplace%20like%20Airbnb,to%20help%20them%20do%20so.%E2%80%9D`.

K Lancaster and Con-sumer Demand. A new approach. *New York-London*, 1971.

B Lantz. Evaluating model performance. *Machine Learning with R*, 2013.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Donghun Lee, Woochang Hyun, Jeongwoo Ryu, Woo Jung Lee, Wonjong Rhee, and Bongwon Suh. An analysis of social features associated with room sales of airbnb. In *Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing*, pages 219–222, 2015.

Xiangming Samuel Li. Short or long review?-text analytics and machine learning approaches to online reputation. *International Journal of Business and Management Research*, 9(1):28–40, 2021.

Sai Liang, Markus Schuckert, Rob Law, and Chih-Chien Chen. Be a "superhost": The importance of badge systems for peer-to-peer rental accommodations. *Tourism management*, 60:454–465, 2017.

Sai Liang, Markus Schuckert, Rob Law, and Chih-Chien Chen. The importance of marketer-generated content to peer-to-peer property rental platforms: evidence from airbnb. *International Journal of Hospitality Management*, 84:102329, 2020.

CiCi Siyue Liu. A couchsurfing ethnography: Traveling and connection in a commodified world. *Inquiries Journal*, 4(07), 2012.

Julie Lorah and Andrew Womack. Value of sample size for computation of the bayesian information criterion (bic) in multilevel modeling. *Behavior research methods*, 51(1):440–450, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Christoph Lutz and Gemma Newlands. Consumer segmentation within the sharing economy: The case of airbnb. *Journal of Business Research*, 88:187–196, 2018.

Jing Lyu, Mimi Li, and Rob Law. Experiencing p2p accommodations: Anecdotes from chinese customers. *International Journal of Hospitality Management*, 77:323–332, 2019.

Xiao Ma, Jeffrey T Hancock, Kenneth Lim Mingjie, and Mor Naaman. Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 2397–2409, 2017.

Sanna Malinen and Jarno Ojala. Perceptions of trust between online auction consumers. *International Journal of Web Portals (IJWP)*, 3(4):15–26, 2011.

Elaine Nicpon Marieb and Katja Hoehn. *Human anatomy & physiology*. Pearson education, 2007.

Alex Marqusee. Airbnb and san francisco: Descriptive statistics and academic research. *San Francisco Planning Department*, 2015.

Richard Diehl Martinez, Anthony Carrington, Tiffany Kuo, Lena Tarhuni, and Nour Adel Zaki Abdel-Motaal. The impact of an airbnb host's listing description'sentiment'and length on occupancy rates. *arXiv preprint arXiv:1711.09196*, 2017.

Lorenzo Masiero, Juan L Nicolau, and Rob Law. A demand-driven analysis of tourist accommodation price: A quantile regression of room bookings. *International Journal of Hospitality Management*, 50:1–8, 2015.

Aurelio G Mauri, Roberta Minazzi, Marta Nieto-García, and Giampaolo Viglia. Humanize your business. the role of personal reputation in the sharing economy. *International Journal of Hospitality Management*, 73:36–43, 2018.

Raveesh Mayya, Shun Ye, Siva Viswanathan, and Rajshree Agarwal. Who forgoes screening in online markets and why? evidence from airbnb. *Evidence from Airbnb (October 12, 2020). Mayya, Raveesh, Ye, Shun, Viswanathan, Siva and Agarwal, Rajshree*, 2020.

MEGVII. Megvii open sources proprietary deep learning framework megengine, 2022a. URL `https://en.megvii.com/news_detail/id/124`.

MEGVII. technologies face$_{r}ecognition,2022b.URL.$

Stephen R Miller. Transferable sharing rights: A theoretical model for regulating airbnb and the short-term rental market. *Available at SSRN 2514178*, 2014.

Mayank Mishra. Convolutional neural networks, explained, Sep 2020. URL `https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939`.

Mareike Möhlmann and Timm Teubner. Navigating by the stars: current challenges for ensuring trust in the sharing economy. *NIM Marketing Intelligence Review*, 12(2):22–27, 2020.

NBTC. Voorzichtige herstart: meer internationale bezoekers verwacht in 2022, Mar 2022. URL `https://www.nbtc.nl/nl/site/organisatie/actueel/voorzichtige-herstart-meer-internationale-bezoeke` `html`.

Gregory Norcie, Emiliano De Cristofaro, and Victoria Bellotti. Bootstrapping trust in online dating: Social verification of online dating profiles. In *International Conference on Financial Cryptography and Data Security*, pages 149–163. Springer, 2013.

Sonia Ouni, Ezzeddine Zagrouba, Majed Chambah, and Michel Herbin. No-reference image semantic quality approach using neural network. In *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 106–113. IEEE, 2011.

Bing Pan and Lixuan Zhang. An eyetracking study on online hotel decision making: The effects of images and umber of options. 2016.

G Parker, M Van Alstyne, and S Choudary. How networked markets are transforming the economy-and how to make them work for you, 2016.

Prabhakar Raghaven. Search on 2020 event. URL https://www.youtube.com/watch?v=ZL5x3ovujiM.

Lizzie Richardson. Performing the sharing economy. *Geoforum*, 67:121–129, 2015.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Bruno Schivinski. Eliciting brand-related social media engagement: A conditional inference tree framework. *Journal of Business Research*, 130:594–602, 2021.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

LLOYD S Shapley. Quota solutions of n-person games. Technical report, RAND CORP SANTA MONICA CA, 1952.

Richard Sharpley. The influence of the accommodation sector on tourism development: lessons from cyprus. *International journal of hospitality management*, 19(3):275–293, 2000.

SightEngine. What is the underlying technology?, 2022. URL https://sightengine.com/faq/sightengine-underlying-technology.

Tyler Sonnemaker. Airbnb is worth more than the 3 largest hotel chains combined after its stock popped 143% on its first day of trading, Dec 2020. URL https://www.businessinsider.com/airbnb-ipo-valuation-tops-three-hotel-chains-combined-opening-day-2020-12?international=true&amp;r=US&amp;IR=T.

M Staff. Most consumers read and rely on online reviews. *URL: http://www. marketingcharts.com/online/most-consumers-read-and-rely-on-onlinereviews-companies-must-adjust-2234*, 2007.

Erose Sthapit and Jano Jimenez-Barreto. Exploring tourists' memorable hospitality experiences: An airbnb perspective. *Tourism Management Perspectives*, 28:83–92, 2018.

Stefan Stremersch, Isabel Verniers, and Peter C Verhoef. The quest for citations: Drivers of article impact. *Journal of Marketing*, 71(3):171–193, 2007.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):1–21, 2007.

Emily Tang and Kunal Sangani. Neighborhood and price prediction for san francisco airbnb listings. *Departments of Computer science, Psychology, economics–Stanford University*, 2015.

Donna Chambers Tijana Rakić. Innovative techniques in tourism research: An exploration of visual methods and academic filmmaking. *International Journal of Tourism Research*, 12(4):397–389, 2010.

Alexander Todorov. Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, 1124(1):208–224, 2008.

Alexander Todorov, Manish Pakrashi, and Nikolaas N Oosterhof. Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, 27(6):813–833, 2009.

Iis P Tussyadiah. Factors of satisfaction and intention to use peer-to-peer accommodation. *International Journal of Hospitality Management*, 55:70–80, 2016.

Stefano Vaccari, Costis Maglaras, and Marco Scarsini. Social learning from online reviews with product choice. In *Proceedings of the 13th Workshop on Economics of Networks, Systems and Computation*, pages 1–1, 2018.

Ministerie van Algemene Zaken. Coronavirus tijdlijn. *Rijksoverheid.nl*, Mar 2022. URL `https://www.rijksoverheid.nl/onderwerpen/coronavirus-tijdlijn`.

Gustav Visser, Inge Erasmus, and Matthew Miller. Airbnb: The emergence of a new accommodation type in cape town, south africa. *Tourism Review International*, 21(2):151–168, 2017.

Moritz von Hoffen, Marvin Hagge, Jan Hendrik Betzing, and Friedrich Chasin. Leveraging social media to gain insights into service delivery: a study on airbnb. *Information Systems and e-Business Management*.

Michel Wedel and Rik Pieters. *Visual marketing: From attention to action*. Psychology Press, 2007.

Chris Wilcox, Nicholas J Mallos, George H Leonard, Alba Rodriguez, and Britta Denise Hardesty. Using expert elicitation to estimate the impacts of plastic pollution on marine wildlife. *Marine Policy*, 65:107–114, 2016.

Jochen Wirtz, Kevin Kam Fung So, Makarand Amrish Mody, Stephanie Q Liu, and HaeEun Helen Chun. Platforms in the peer-to-peer sharing economy. *Journal of Service Management*, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL `http://arxiv.org/abs/1910.03771`.

B Wood. Airbnb is now bigger than the world's top five hotel brands put together, 2017.

Jiang Wu, Panhao Ma, and Karen L Xie. In sharing economy we trust: The effects of host attributes on short-term rental purchases. *International Journal of Contemporary Hospitality Management*, 2017.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer, 2019.

Rosalie Wyonch. Bits, bytes, and taxes: Vat and the digital economy in canada. *CD Howe Institute Commentary*, 487, 2017.

Karen Xie and Zhenxing Mao. The impacts of quality and quantity attributes of airbnb hosts on listing performance. *International Journal of Contemporary Hospitality Management*, 2017.

Qiang Ye, Rob Law, and Bin Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.

Georgios Zervas, Davide Proserpio, and John W Byers. The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of marketing research*, 54(5): 687–705, 2017.

Le Zhang, Qiang Yan, and Leihan Zhang. A computational framework for understanding antecedents of guests' perceived trust towards hosts on airbnb. *Decision Support Systems*, 115: 105–116, 2018.

Le Zhang, Qiang Yan, and Leihan Zhang. A text analytics framework for understanding the relationships among host self-description, trust perception and purchase behavior on airbnb. *Decision Support Systems*, 133:113288, 2020.

Mengxia Zhang and Lan Luo. Can user-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Evidence from Yelp (March 1, 2018)*, 2018.

Shunyuan Zhang, Dokyun Lee, Param Vir Singh, and Kannan Srinivasan. How much is an image worth? airbnb property demand estimation leveraging large scale image analytics. *Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics (May 25, 2017)*, 2017.

Shunyuan Zhang, Nitin Mehta, Param Vir Singh, and Kannan Srinivasan. Can lower-quality images lead to greater demand on airbnb? *Work. Pap*, 2019.

Shunyuan Zhang, Dokyun Lee, Param Vir Singh, and Kannan Srinivasan. What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science*, 2021.

Wei Zhao. Research on the deep learning of the small sample data based on transfer learning. In *AIP Conference Proceedings*, volume 1864, page 020018. AIP Publishing LLC, 2017.

Appendices

Appendix A

Amenities according to Airbnb (2022a)
Bedroom amenities:

- Toilet paper
- Soap for hands and body
- One towel per guest
- Linens for each bed
- One pillow per guest
- Cleaning supplies

Top amenities quests search for:*

- A pet-friendly space
- Wifí
- Free parking
- A pool
- A jacuzzi
- A kitchen
- Air conditioning
- Heating
- A washer
- TV or cable

Safety amenities:

- Carbon monoxide alarm
- Smoke alarm
- Fire extinguisher
- First-aid kit
- Emergency plan and local numbers

Accessibility amenities:

- Step-free entryway
- Wide entrances (at least 32″)
- Wide hallways (at least 36°)
- Accessible bathroom

Additional amenities

- Extra toilet paper, linens, and towels
- Basic toiletries like shampoo and conditioner
- Dish soap and cleaning supplies
- Dining basics like a coffee maker, cooking utensils, dishes, and silverware
- Wine glasses
- Basic cooking supplies like salt, pepper, and oil
- Coffee, tea
- Light breakfast or snacks
- Hangers
- Adapters and chargers

Remote work amenities:

- Fast and reliable wifi
- Laptop-friendly workspace
- Good lighting

- Fully equipped kitchens
- Office supplies

Family amenities:

- A crib and high chair
- A bathtub
- Air conditioning
- A washer and or dryer
- Extra cleaning supplies
- Fumiture covers
- Bowls for pet food and water
- Towels to wipe off paws at the door

Appendix B

Results of the fitted conditional inference tree
The alpha was set to 0.05 by default.

Model formula:
occupancy_rate ∼ price + accommodates + bedroom_amenities + additional_amenities + family_amenities + safety_amenities + top_amenities + nr_verifications + bathrooms + host_is_superhost + host_identity_verified + license + room_type + number_of_reviews + review_scores_rating + sentiment_score_description + host_has_descr + sentiment_score_description_host + neighbourhood_cleansed + happy + neutral + sad + sharpness + brightness + average_age + male + female

Fitted party:

```
[1] root
— [2] number_of_reviews <= 74
— — [3] room type in Hotel room, Entire home apt
— — — [4] host_is_superhost <= 0
— — — — [5] review_scores_rating ≤= 4.48 : 0.247(n = 21, err = 1.0)
— — — — [6] review_scores_rating > 4.48 : 0.096(n = 355, err = 5.3)
— — — [7] host_is_superhost >0
— — — — [8] bedroom_amenities <= 1 : 0.162(n = 99, err = 2.0)]
— — — — [9] bedroom_amenities > 1 : 0.248(n = 55, err = 3.1)
— — [10] room_type in Private_room: 0.287(n = 226, err = 20.7)
— [11] number_of_reviews > 74
— — [12] number_of_reviews <= 395
— — — [13] price <= 103
— — — — [14] neighbourhood_cleansed in Other, De_Aker_Nieuw_Sloten, Gaasperdam_Driemond, IJburg_Zeeburgereiland, Noord_Oost, Noord_West, Slotervaart, Watergraafsmeer, Westerpark, Zuid
— — — — — [15] additional_amenities <= 2 : 0.105(n = 13, err = 0.2)
— — — — — [16] additional_amenities > 2 : 0.370(n = 53, err = 4.6)
— — — — [17] neighbourhood_cleansed in Bos_en_Lommer, Centrum_Oost, Centrum_West, Baarsjes_Oud_West, De_Pijp_Rivierenbuurt, Geuzenveld_Slotermeer, Oostelijk_Havengebied_Indische_Buurt, Oud_Noord, Oud_Oost
— — — — — [18] brightness <= 0.56 : 0.518(n = 67, err = 6.9)
— — — — — [19] brightness > 0.56 : 0.738(n = 74, err = 5.9)
— — — [20] price > 103 : 0.348(n = 145, err = 12.1)
— — [21] number_of reviews > 395 : 0.873(n = 28, err = 1.0)
```

Number of inner nodes: 10
Number of terminal nodes: 11

Appendix C

The examples shown are from Sight Engine's (2022) image quality detection website.

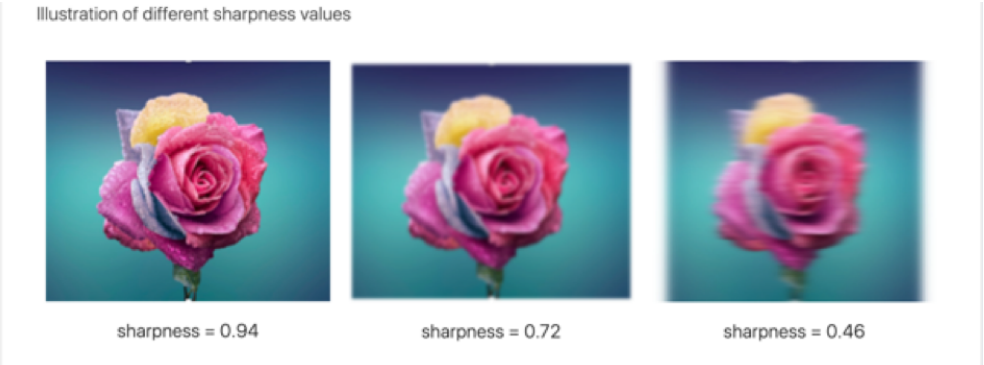
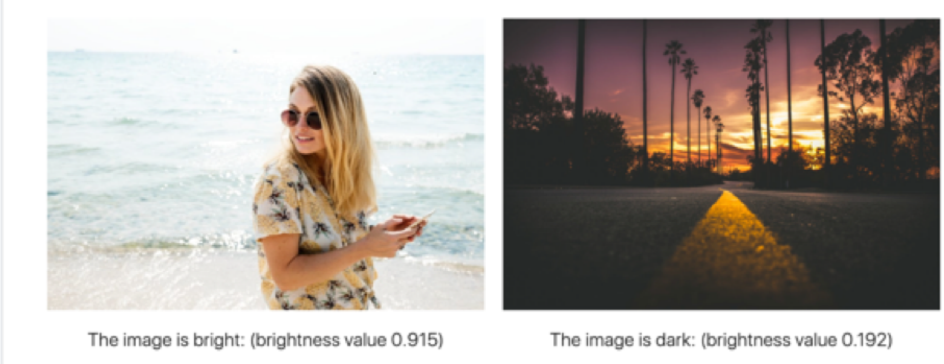


Figure C1: *Sharpness*


The image is bright: (brightness value 0.915)

The image is dark: (brightness value 0.192)

Figure C2: *Brightness*

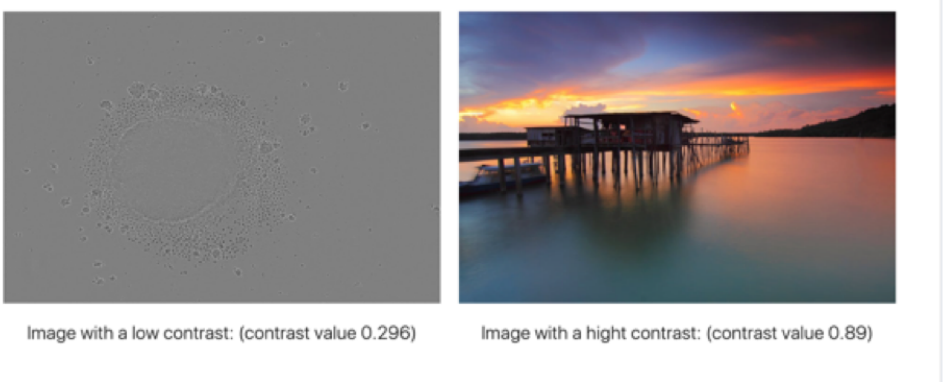


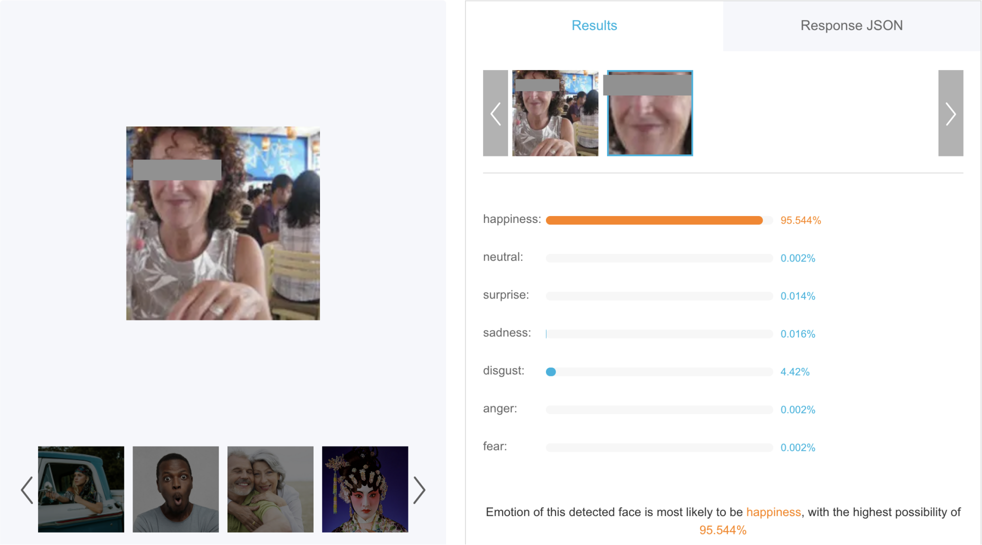
Image with a low contrast: (contrast value 0.296)

Image with a hight contrast: (contrast value 0.89)

Figure C3: *Contrast*

Appendix D

Example of the output of Face ++ face emotion analysis.
This is a real-world example where the URL of one of the hosts was provided to Face ++. On the right side the output from Face ++ is shown.



happiness: 95.544%

neutral: 0.002%

surprise: 0.014%

sadness: 0.016%

disgust: 4.42%

anger: 0.002%

fear: 0.002%

Emotion of this detected face is most likely to be happiness, with the highest possibility of 95.544%

Figure D1: *Output Face ++*

Appendix E

In the example below a picture of a family of 5 and 6 are shown. The people on the left are considered to be, from left to right – 37, 30, 44, 41, 54. The average age of the family on the left would thus be 41 what is close to the average of the parents 45. Similarly, the family on the right is considered to be from top to bottom – 22, 55, 30, 29 ,27, 47. The average age of the family on the right consequently amounts to ≈ 35.5 in this case the average age is quite far from 51. As only a fraction of the data picture consists out of more than two faces – 1.95 % – the average age of the picture is considered to be a good indicator for these family pictures.

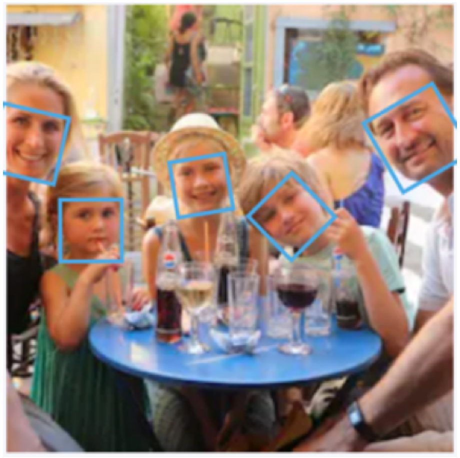


Figure E1: *Family 1*

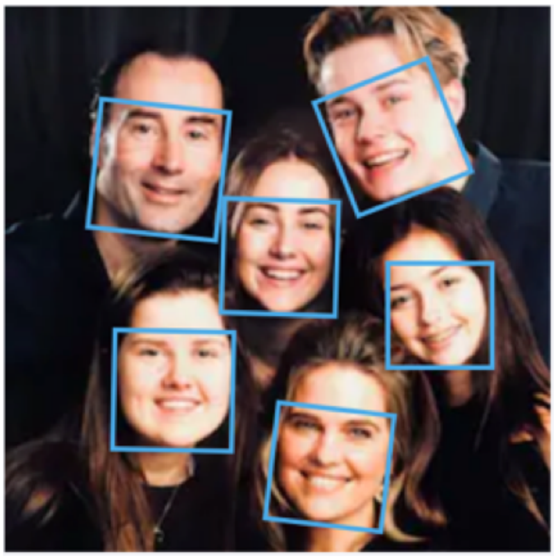


Figure E2: *Family 2*

Appendix F

Linear models including the neighborhood variable effects.

Linear regression models								
	Lancasterian		**Lanc.+Social**		**Lanc.+Presentation**		**Lanc.+Social+Pres.**	
Predictors	*Estimates*	*std. Error*	*Estimates*	*std. Error*	*Estimates*	*std. Error*	*Estimates*	*std. Error*
(Intercept)	0.067	0.073	0.191	0.146	0.224	0.139	0.299	0.179
Lancasterian perspective								
price	**-0.001** ***	0.000	**-0.001** ***	0.000	**-0.001** ***	0.000	**-0.001** ***	0.000
accommodates	**0.021** ***	0.006	**0.014** **	0.005	**0.020** ***	0.006	**0.014** **	0.005
bedroom amenities	**0.021** **	0.008	0.017 *	0.008	0.021 **	0.008	0.018 *	0.008
additional amenities	0.011 *	0.005	0.008	0.005	0.012 *	0.005	0.009 *	0.005
family amenities	0.013	0.010	0.006	0.009	0.013	0.010	0.006	0.010
safety amenities	**0.027** ***	0.007	**0.023** ***	0.007	**0.028** ***	0.007	**0.023** ***	0.007
top amenities	**-0.044** ***	0.009	**-0.030** ***	0.008	**-0.045** ***	0.009	**-0.031** ***	0.008
nr verifications	0.005	0.004	-0.001	0.004	0.004	0.004	-0.002	0.004
bathrooms	0.028	0.019	0.038 *	0.018	0.027	0.019	0.037 *	0.018
Superhost	**0.055** ***	0.015	0.036 *	0.014	**0.058** ***	0.015	**0.042** **	0.014
Identity verified	**0.057** **	0.021	**0.056** **	0.020	**0.058** **	0.022	**0.057** **	0.020
license	**0.051** ***	0.015	**0.038** **	0.014	**0.052** ***	0.015	**0.038** **	0.014
Entire home/apartment	0.094	0.052	0.076	0.048	0.071	0.053	0.056	0.050
Private room	0.130 *	0.051	0.074	0.048	0.105 *	0.053	0.052	0.050
[Bos_en_Lommer]	-0.048	0.058	-0.052	0.055	-0.048	0.058	-0.058	0.054
[Centrum_Oost]	0.087	0.046	0.075	0.043	0.089	0.046	0.073	0.043
[Centrum_West]	**0.116** **	0.044	**0.113** **	0.042	**0.116** **	0.044	**0.108** **	0.042
[De_Aker_Nieuw_Sloten]	-0.028	0.080	-0.018	0.075	-0.020	0.080	-0.007	0.075
[Baarsjes_Oud_West]	-0.023	0.046	-0.021	0.043	-0.023	0.046	-0.025	0.043
[De_Pijp_Rivierenbuurt]	-0.006	0.047	-0.013	0.044	-0.008	0.047	-0.016	0.044
[Gaasperdam_Driemond]	**-0.179** *	0.075	**-0.157** *	0.071	**-0.187** *	0.076	**-0.160** *	0.071
[Geuzenveld_Slotermeer]	-0.030	0.069	-0.019	0.065	-0.031	0.069	-0.028	0.065
[IJburg_Zeeburgereiland]	**-0.139** *	0.057	**-0.107** *	0.054	**-0.133** *	0.057	-0.103	0.054
[Noord_Oost]	-0.108	0.059	-0.089	0.055	-0.104	0.059	-0.089	0.055
[Noord_West]	**-0.122** *	0.057	**-0.113** *	0.054	**-0.124** *	0.058	**-0.114** *	0.054
[Oostelijk_Havengebied_Indische_Buurt]	-0.011	0.055	0.025	0.052	-0.016	0.055	0.016	0.052
[Oud_Noord]	-0.016	0.053	-0.016	0.050	-0.005	0.053	-0.002	0.050
[Oud_Oost]	-0.037	0.051	-0.022	0.048	-0.033	0.051	-0.022	0.048
[Slotervaart]	-0.054	0.072	-0.068	0.068	-0.052	0.072	-0.067	0.068
[Watergraafsmeer]	**-0.137** *	0.062	-0.113	0.058	**-0.129** *	0.062	-0.104	0.058
[Westerpark]	-0.043	0.050	-0.043	0.047	-0.041	0.051	-0.045	0.047
[Zuid]	-0.052	0.050	-0.064	0.047	-0.045	0.050	-0.061	0.047
Social lerarning perspective								
number_of_reviews			**0.001** ***	0.000			**0.001** ***	0.000
review_scores_rating			-0.030	0.027			-0.026	0.027
Presentation perspective								
host_has_descr					0.008	0.021	-0.007	0.019
sentiment host					-0.018	0.020	-0.024	0.018
sentiment description					0.001	0.015	0.001	0.014
happy					-0.000	0.000	0.000	0.000
sad					-0.000	0.000	-0.000	0.000
sharpness					-0.122	0.115	-0.065	0.108
brightness					0.053	0.044	0.068	0.041
average_age					-0.001 *	0.001	**-0.002** ***	0.001
male					0.021	0.016	0.016	0.015
female					0.006	0.014	0.001	0.013
Observations	1421		1421		1421		1421	
R2 / R2 adjusted	0.319 / 0.303		0.399 / 0.384		0.325 / 0.304		0.409 / 0.390	
AIC	169.992		-3.592		176.928		-8.121	
BIC	348.801		185.736		408.329		233.798	

** p<0.05 ** p<0.01 *** p<0.001*

Appendix G

Figure G1 show the Mean Average SHAP Values for all features and Figure G2

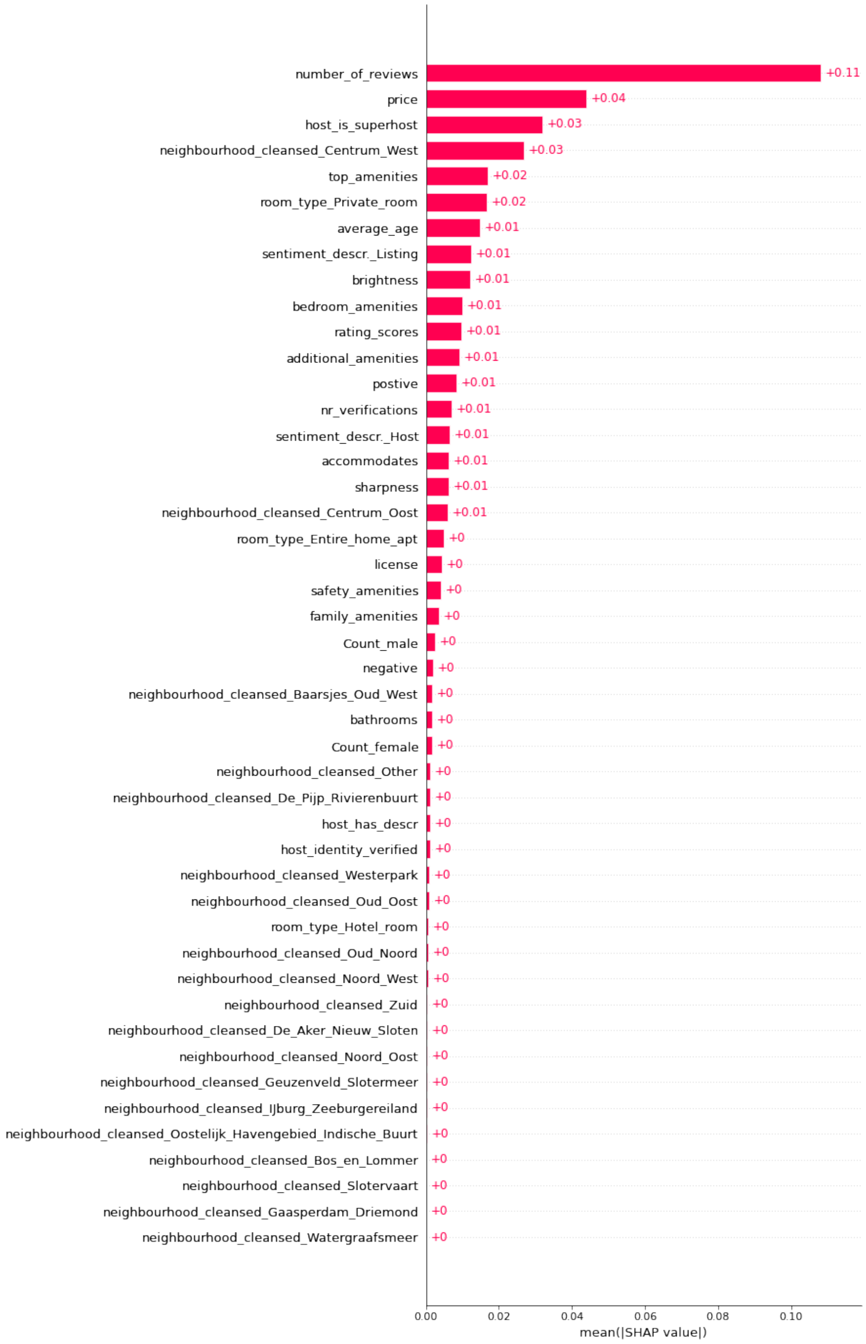


Figure G1: *Mean Average SHAP Values for all features from the Random Forest model*

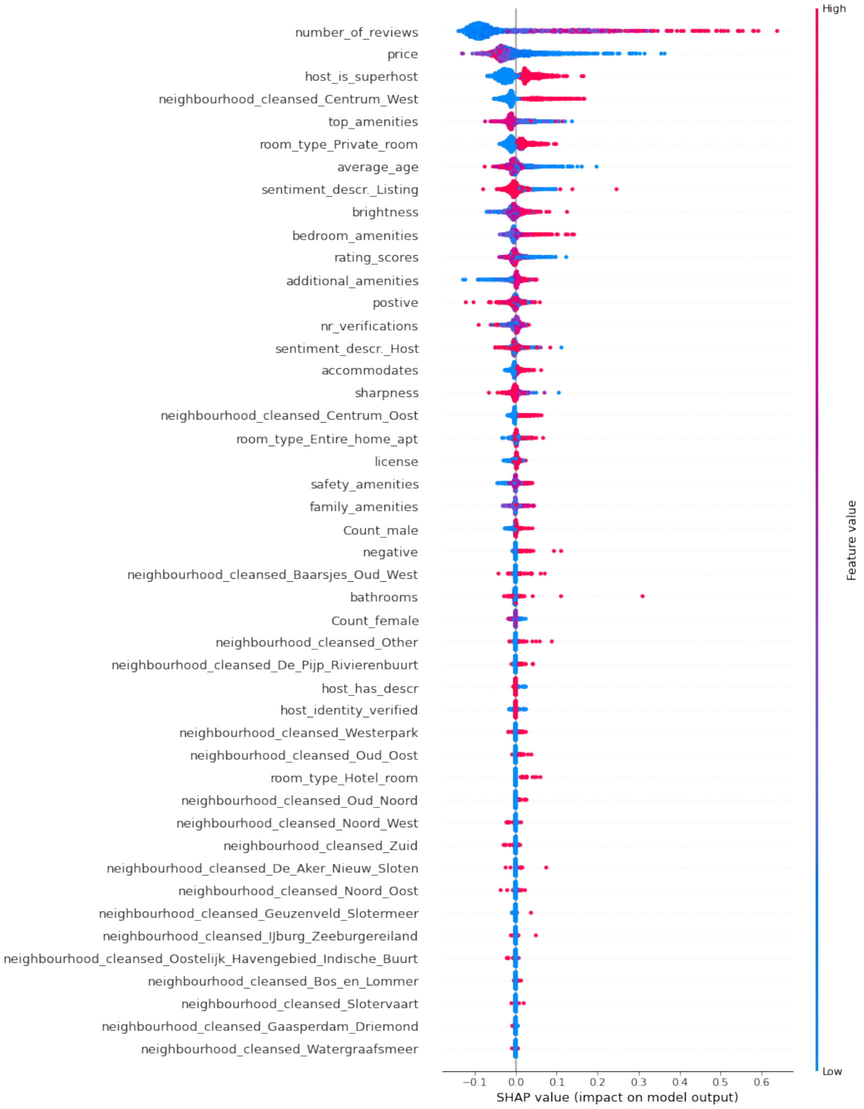


Figure G2: *Summary plot for all features of the Random Forest model*

Appendix H

Linear regression model based on simultaneous forward and backward selection. The criteria used are AIC and BIC criteria.

Both Forward and Backward model selection based on AIC en BIC						
	Stepwise AIC		**Stepwise BIC**		**Full-linear model**	
Predictors	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*
(Intercept)	**0.161** **	0.060	**0.262** ***	0.038	0.299	0.179
Lancasterian perspective						
price	**-0.001** ***	0.000	**-0.000** ***	0.000	**-0.001** ***	0.000
accommodates	**0.015** **	0.005			**0.014** **	0.005
bedroom_amenities	**0.019** **	0.007	**0.024** ***	0.007	0.018 *	0.008
additional_amenities	0.008	0.005			0.009 *	0.005
safety_amenities	**0.023** ***	0.006	**0.031** ***	0.006	**0.023** ***	0.007
family_amenities					0.006	0.010
nr_verifications					-0.002	0.004
top_amenities	**-0.026** ***	0.007	**-0.034** ***	0.006	**-0.031** ***	0.008
bathrooms	0.039 *	0.018			0.037 *	0.018
host_is_superhost	**0.039** **	0.014	**0.050** ***	0.014	**0.042** **	0.014
host_identity_verified	**0.050** **	0.017	**0.069** ***	0.017	**0.057** **	0.020
license	**0.038** **	0.014	**0.043** **	0.014	**0.038** **	0.014
Entire_home/apt.					0.056	0.050
Private_room					0.052	0.050
Social lerarning perspective						
number_of_reviews	**0.001** ***	0.000	**0.001** ***	0.000	**0.001** ***	0.000
review_scores_rating					-0.026	0.027
Presentation perspective						
sentiment_score_description_host	-0.027 *	0.011	**-0.034** **	0.012	-0.024	0.018
brightness	0.070	0.041			0.068	0.041
average_age	**-0.002** ***	0.001	**-0.002** ***	0.001	**-0.002** ***	0.001
sentiment_score_description					0.001	0.014
host_has_descr					-0.007	0.019
happy					0.000	0.000
sad					-0.000	0.000
sharpness					-0.065	0.108
male					0.016	0.015
female					0.001	0.013
Observations	1421		1421		1421	
R2 / R2 adjusted	0.407 / 0.393		0.348 / 0.343		0.409 / 0.390	
AIC	-25.911		63.876		-8.121	
BIC	152.896		126.986		233.798	

** p<0.05 ** p<0.01 *** p<0.001*

Appendix I

The table below contains the full-linear regression models with standardized numeric variables. The same variables remain significant.

Linear regression models								
	Lancasterian		**Lanc.+Social**		**Lanc.+Presentation**		**Lanc.+Social+Pres.**	
Predictors	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*
(Intercept)	0.067	0.073	0.191	0.146	0.224	0.139	0.299	0.179
Lancasterian perspective								
price	-0.001 ***	0.000	-0.001 ***	0.000	-0.001 ***	0.000	-0.001 ***	0.000
accommodates	0.021 ***	0.006	0.014 **	0.005	0.020 ***	0.006	0.014 **	0.005
bedroom_amenities	0.021 **	0.008	0.017 *	0.008	0.021 **	0.008	0.018 *	0.008
additional_amenities	0.011 *	0.005	0.008	0.005	0.012 *	0.005	0.009 *	0.005
family_amenities	0.013	0.010	0.006	0.009	0.013	0.010	0.006	0.010
safety_amenities	0.027 ***	0.007	0.023 ***	0.007	0.028 ***	0.007	0.023 ***	0.007
top_amenities	-0.044 ***	0.009	-0.030 ***	0.008	-0.045 ***	0.009	-0.031 ***	0.008
nr_verifications	0.005	0.004	-0.001	0.004	0.004	0.004	-0.002	0.004
bathrooms	0.028	0.019	0.038 *	0.018	0.027	0.019	0.037 *	0.018
host_is_superhost	0.055 ***	0.015	0.036 *	0.014	0.058 ***	0.015	0.042 **	0.014
host_identity_verified	0.057 **	0.021	0.056 **	0.020	0.058 **	0.022	0.057 **	0.020
license	0.051 ***	0.015	0.038 **	0.014	0.052 ***	0.015	0.038 **	0.014
[Entire_home_apt]	0.094	0.052	0.076	0.048	0.071	0.053	0.056	0.050
Private room	0.130 *	0.051	0.074	0.048	0.105 *	0.053	0.052	0.050
[Bos_en_Lommer]	-0.048	0.058	-0.052	0.055	-0.048	0.058	-0.058	0.054
[Centrum_Oost]	0.087	0.046	0.075	0.043	0.089	0.046	0.073	0.043
[Centrum_West]	0.116 **	0.044	0.113 **	0.042	0.116 **	0.044	0.108 **	0.042
[De_Aker_Nieuw_Sloten]	-0.028	0.080	-0.018	0.075	-0.020	0.080	-0.007	0.075
[Baarsjes_Oud_West]	-0.023	0.046	-0.021	0.043	-0.023	0.046	-0.025	0.043
[De_Pijp_Rivierenbuurt]	-0.006	0.047	-0.013	0.044	-0.008	0.047	-0.016	0.044
[Gaasperdam_Driemond]	-0.179 *	0.075	-0.157 *	0.071	-0.187 *	0.076	-0.160 *	0.071
[Geuzenveld_Slotermeer]	-0.030	0.069	-0.019	0.065	-0.031	0.069	-0.028	0.065
[IJburg_Zeeburgereiland]	-0.139 *	0.057	-0.107 *	0.054	-0.133 *	0.057	-0.103	0.054
[Noord_Oost]	-0.108	0.059	-0.089	0.055	-0.104	0.059	-0.089	0.055
[Noord_West]	-0.122 *	0.057	-0.113 *	0.054	-0.124 *	0.058	-0.114 *	0.054
[Oostelijk_Havengebied_Indische_Buurt]	-0.011	0.055	0.025	0.052	-0.016	0.055	0.016	0.052
[Oud_Noord]	-0.016	0.053	-0.016	0.050	-0.005	0.053	-0.002	0.050
[Oud_Oost]	-0.037	0.051	-0.022	0.048	-0.033	0.051	-0.022	0.048
[Slotervaart]	-0.054	0.072	-0.068	0.068	-0.052	0.072	-0.067	0.068
[Watergraafsmeer]	-0.137 *	0.062	-0.113	0.058	-0.129 *	0.062	-0.104	0.058
[Westerpark]	-0.043	0.050	-0.043	0.047	-0.041	0.051	-0.045	0.047
[Zuid]	-0.052	0.050	-0.064	0.047	-0.045	0.050	-0.061	0.047
Social Learning perspective								
number_of_reviews			0.001 ***	0.000			0.001 ***	0.000
review_scores_rating			-0.030	0.027			-0.026	0.027
Presentation perspective								
host_has_descr					0.008	0.021	-0.007	0.019
sentiment_score_description_host					-0.018	0.020	-0.024	0.018
sentiment_score_description					0.001	0.015	0.001	0.014
happy					-0.000	0.000	0.000	0.000
sad					-0.000	0.000	-0.000	0.000
sharpness					-0.122	0.115	-0.065	0.108
brightness					0.053	0.044	0.068	0.041
average_age					-0.001 *	0.001	-0.002 ***	0.001
male					0.021	0.016	0.016	0.015
female					0.006	0.014	0.001	0.013
Observations	1421		1421		1421		1421	
R2 / R2 adjusted	0.319 / 0.303		0.399 / 0.384		0.325 / 0.304		0.409 / 0.390	
AIC	169992		-3592		176928		-8121	
BIC	348802		185736		408329		233798	
** p<0.05 ** p<0.01 *** p<0.001*								

Appendix J

The table below contains the linear regression results with a zero imputed average age.

	Linear regression models - Age imputed with 0							height
	Lancasterian		**Lanc.+Social**		**Lanc.+Presentation**		**Lanc.+Social+Pres.**	
Predictors	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*
(Intercept)	0.199 *	0.090	0.223	0.121	0.289 *	0.138	0.225	0.153
Lancasterian perspective								
price	-0.001 ***	0.000	-0.001 ***	0.000	-0.001 ***	0.000	-0.001 ***	0.000
accommodates	0.032 ***	0.005	0.028 ***	0.005	0.031 ***	0.005	0.026 ***	0.005
bedroom_amenities	0.043 ***	0.009	0.039 ***	0.009	0.042 ***	0.009	0.036 ***	0.009
additional_amenities	0.004	0.005	-0.002	0.005	0.007	0.005	0.002	0.005
family_amenities	0.012	0.010	0.005	0.010	0.011	0.010	0.004	0.010
top_amenities	-0.037 ***	0.008	-0.027 ***	0.008	-0.036 ***	0.008	-0.026 **	0.008
nr_amenities	-0.000	0.001	-0.000	0.001	-0.000	0.001	-0.000	0.001
nr_verifications	0.001	0.003	-0.005	0.003	0.002	0.003	-0.004	0.003
bathrooms	0.014	0.018	0.015	0.017	0.017	0.018	0.019	0.017
host_is_superhost	0.053 ***	0.014	0.028 *	0.013	0.059 ***	0.014	0.035 **	0.013
host_identity_verified	0.081 ***	0.019	0.079 ***	0.019	0.075 ***	0.020	0.071 ***	0.019
license	0.062 ***	0.014	0.054 ***	0.013	0.063 ***	0.014	0.055 ***	0.013
room_type [Hotel room]	-0.043	0.038	-0.055	0.036	-0.041	0.039	-0.062	0.037
room_type [Private room]	0.058 ***	0.017	0.027	0.016	0.055 **	0.017	0.023	0.016
[Bijlmer-Oost]	-0.027	0.138	-0.025	0.131	0.013	0.138	0.012	0.130
[Bos en Lommer]	-0.011	0.090	-0.002	0.086	-0.000	0.090	0.001	0.085
[Buitenveldert - Zuidas]	-0.128	0.108	-0.115	0.103	-0.112	0.108	-0.102	0.102
[Centrum-Oost]	0.099	0.084	0.095	0.080	0.109	0.084	0.096	0.079
[Centrum-West]	0.103	0.083	0.107	0.079	0.113	0.083	0.108	0.079
[De Aker - Nieuw Sloten]	-0.039	0.106	-0.035	0.101	-0.020	0.106	-0.018	0.100
[De Baarsjes - Oud-West]	0.009	0.084	0.019	0.080	0.020	0.084	0.022	0.080
[De Pijp - Rivierenbuurt]	-0.013	0.084	-0.006	0.080	-0.004	0.084	-0.006	0.080
[Gaasperdam - Driemond]	-0.196	0.105	-0.165	0.100	-0.177	0.105	-0.141	0.100
[Geuzenveld - Slotermeer]	-0.010	0.097	0.010	0.093	0.010	0.098	0.020	0.092
[IJburg -Zeeburgereiland]	-0.157	0.091	-0.116	0.086	-0.133	0.091	-0.098	0.086
[Noord-Oost]	-0.126	0.092	-0.103	0.088	-0.104	0.093	-0.089	0.088
[Noord-West]	-0.154	0.092	-0.138	0.087	-0.141	0.092	-0.127	0.087
[Oostelijk Havengebied - Indische Buurt]	0.030	0.088	0.058	0.084	0.037	0.088	0.055	0.084
[Osdorp]	0.024	0.113	-0.002	0.108	0.033	0.113	0.002	0.107
[Oud-Noord]	-0.064	0.087	-0.062	0.083	-0.047	0.087	-0.049	0.083
[Oud-Oost]	-0.012	0.087	-0.005	0.082	0.004	0.087	0.002	0.082
[Slotervaart]	-0.049	0.097	-0.061	0.093	-0.039	0.097	-0.059	0.092
[Watergraafsmeer]	-0.154	0.093	-0.124	0.089	-0.129	0.093	-0.103	0.089
[Westerpark]	-0.041	0.086	-0.029	0.082	-0.028	0.086	-0.023	0.082
[Zuid]	-0.051	0.086	-0.050	0.082	-0.032	0.086	-0.039	0.081
Social learning perspective								
number_of_reviews			0.001 ***	0.000			0.001 ***	0.000
review_scores_rating			-0.009	0.018			0.001	0.019
Presentation persepective								
host_has_descr					0.017	0.019	0.006	0.018
sentiment_score_description_host					-0.013	0.018	-0.017	0.017
sentiment_score_description					-0.017	0.014	-0.012	0.013
happy					-0.000	0.000	-0.000	0.000
sad					-0.001	0.000	-0.000	0.000
sharpness					-0.094	0.110	-0.033	0.104
brightness					0.036	0.041	0.039	0.038
average_age					-0.001 **	0.000	-0.002 ***	0.000
male					0.018	0.014	0.015	0.013
female					0.005	0.013	0.000	0.013
Observations	1876		1876		1876		1876	
R2 / R2 adjusted	0.285 / 0.271		0.355 / 0.342		0.295 / 0.277		0.368 / 0.352	
AIC	455.753		267.808		450.319		247.157	
BIC	660.618		483.747		710.553		518.465	

** p<0.05 ** p<0.01 *** p<0.001*

Appendix K

The table below contains the linear regression results for a mean imputed average age.

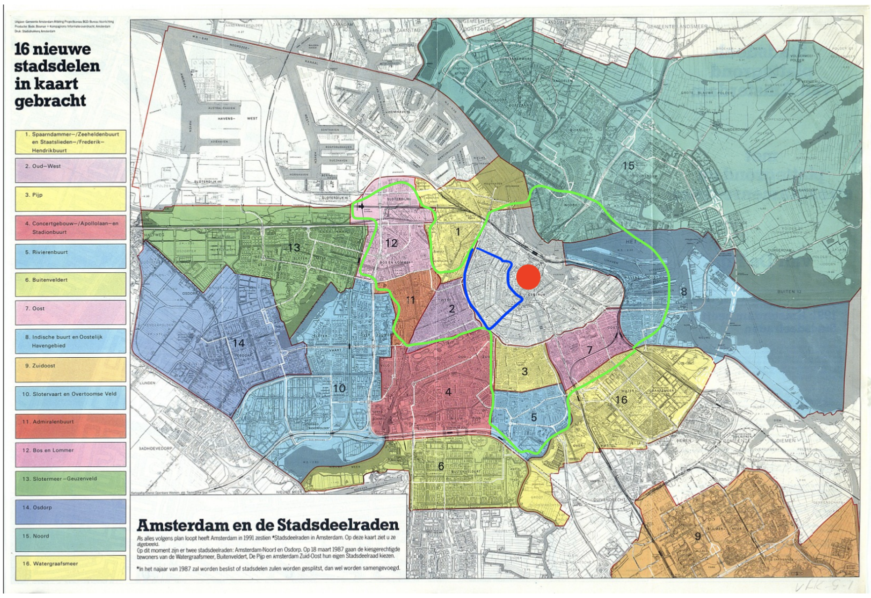
Linear regression models - Age imputed with the mean

	Lancasterian		Lanc.+Social		Lanc.+Presentation		Lanc.+Social+Pres.	
Predictors	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*	*Estimates*	*S.E.*
(Intercept)	0.199 *	0.090	0.223	0.121	0.323 *	0.140	0.286	0.155
Lancasterian perspective								
price	**-0.001 *** **	0.000	**-0.001 *** **	0.000	**-0.001 *** **	0.000	**-0.001 *** **	0.000
accommodates	**0.032 *** **	0.005	**0.028 *** **	0.005	**0.031 *** **	0.005	**0.027 *** **	0.005
bedroom_amenities	**0.043 *** **	0.009	**0.039 *** **	0.009	**0.043 *** **	0.009	**0.037 *** **	0.009
additional_amenities	0.004	0.005	-0.002	0.005	0.006	0.005	0.001	0.005
family_amenities	0.012	0.010	0.005	0.010	0.012	0.010	0.004	0.010
top_amenities	**-0.037 *** **	0.008	**-0.027 *** **	0.008	**-0.037 *** **	0.008	**-0.028 *** **	0.008
nr_amenities	-0.000	0.001	-0.000	0.001	-0.000	0.001	-0.000	0.001
nr_verifications	0.001	0.003	-0.005	0.003	0.001	0.003	-0.005	0.003
bathrooms	0.014	0.018	0.015	0.017	0.018	0.018	0.020	0.017
host_is_superhost	**0.053 *** **	0.014	0.028 *	0.013	**0.058 *** **	0.014	0.035 *	0.013
host_identity_verified	**0.081 *** **	0.019	**0.079 *** **	0.019	**0.077 *** **	0.020	**0.074 *** **	0.019
license	**0.062 *** **	0.014	**0.054 *** **	0.013	**0.064 *** **	0.014	**0.057 *** **	0.013
room_type [Hotel room]	-0.043	0.038	-0.055	0.036	-0.037	0.039	-0.055	0.037
room_type [Private room]	**0.058 *** **	0.017	0.027	0.016	0.055 **	0.017	0.023	0.016
[Bijlmer-Oost]	-0.027	0.138	-0.025	0.131	0.007	0.138	0.004	0.131
[Bos en Lommer]	-0.011	0.090	-0.002	0.086	0.003	0.090	0.005	0.085
[Buitenveldert - Zuidas]	-0.128	0.108	-0.115	0.103	-0.108	0.108	-0.096	0.102
[Centrum-Oost]	0.099	0.084	0.095	0.080	0.114	0.084	0.102	0.080
[Centrum-West]	0.103	0.083	0.107	0.079	0.116	0.083	0.112	0.079
[De Aker - Nieuw Sloten]	-0.039	0.106	-0.035	0.101	-0.017	0.106	-0.013	0.101
[De Baarsjes - Oud-West]	0.009	0.084	0.019	0.080	0.025	0.084	0.029	0.080
[De Pijp - Rivierenbuurt]	-0.013	0.084	-0.006	0.080	0.000	0.085	0.000	0.080
[Gaasperdam - Driemond]	-0.196	0.105	-0.165	0.100	-0.180	0.105	-0.145	0.100
[Geuzenveld - Slotermeer]	-0.010	0.097	0.010	0.093	0.012	0.098	0.022	0.093
[IJburg - Zeeburgereiland]	-0.157	0.091	-0.116	0.086	-0.131	0.091	-0.095	0.086
[Noord-Oost]	-0.126	0.092	-0.103	0.088	-0.103	0.093	-0.087	0.088
[Noord-West]	-0.154	0.092	-0.138	0.087	-0.138	0.092	-0.123	0.087
[Oostelijk Havengebied - Indische Buurt]	0.030	0.088	0.058	0.084	0.043	0.088	0.063	0.084
[Osdorp]	0.024	0.113	-0.002	0.108	0.036	0.113	0.005	0.107
[Oud-Noord]	-0.064	0.087	-0.062	0.083	-0.047	0.088	-0.049	0.083
[Oud-Oost]	-0.012	0.087	-0.005	0.082	0.008	0.087	0.007	0.082
[Slotervaart]	-0.049	0.097	-0.061	0.093	-0.032	0.097	-0.049	0.092
[Watergraafsmeer]	-0.154	0.093	-0.124	0.089	-0.128	0.094	-0.101	0.089
[Westerpark]	-0.041	0.086	-0.029	0.082	-0.025	0.086	-0.020	0.082
[Zuid]	-0.051	0.086	-0.050	0.082	-0.028	0.086	-0.032	0.082
Social Learning perspective								
number_of_reviews			**0.001 *** **	0.000			**0.001 *** **	0.000
review_scores_rating			-0.009	0.018			-0.000	0.019
Presentation perspective								
host_has_descr					0.016	0.019	0.006	0.018
sentiment_score_description_host					-0.012	0.018	-0.017	0.018
sentiment_score_description					-0.016	0.014	-0.011	0.013
happy					-0.000 *	0.000	-0.000	0.000
sad					-0.001 *	0.000	-0.001	0.000
sharpness					-0.094	0.110	-0.032	0.105
brightness					0.036	0.041	0.039	0.039
average_age					-0.001	0.001	**-0.002 ** **	0.001
male					0.000	0.013	-0.012	0.012
female					-0.008	0.013	-0.019	0.012
Observations	1876		1876		1876		1876	
R2 / R2 adjusted	0.285 / 0.271		0.355 / 0.342		0.293 / 0.276		0.365 / 0.349	
AIC	455.753		267.808		454.424		255.778	
BIC	660.618		483.747		714.658		527.086	

** p<0.05 ** p<0.01 *** p<0.001*

Appendix L

The drawn green line represents the neighborhoods as selected by the conditional inference tree. The drawn blue represents Centrum-West which was found to be most significant by both the linear and the tree based models. Lastly, the red-dot represents the Amsterdam's city center.


16 nieuwe stadsdelen in kaart gebracht
1. Spaarndammer–/Zeeheldenbuurt en Staatslieden–/Frederik–Hendrikbuurt
2. Oud–West
3. Pijp
4. Concertgebouw–/Apollolaan–en Stadionbuurt
5. Rivierenbuurt
6. Buitenveldert
7. Oost
8. Indische buurt en Oostelijk Havengebied
9. Zuidoost
10. Slotervaart en Overtoomse Veld
11. Admiralenbuurt
12. Bos en Lommer
13. Slotermeer–Geuzenveld
14. Osdorp
15. Noord
16. Watergraafsmeer
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Amsterdam en de Stadsdeelraden

Als alles volgens plan loopt heeft Amsterdam in 1991 zestien *Stadsdeelraden in Amsterdam. Op deze kaart ziet u ze afgebeeld.
Op dit moment zijn er twee stadsdeelraden: Amsterdam-Noord en Osdorp. Op 18 maart 1987 gaan de kiesgerechtigde bewoners van de Watergraafsmeer, Buitenveldert, De Pijp en Amsterdam Zuid-Oost hun eigen Stadsdeelraad kiezen.

*In het najaar van 1987 zal worden beslist of stadsdelen zulen worden gesplitst, dan wel worden samengevoegd.

Figure L1: *Neighborhoods of Amsterdam*

Appendix M

Plots to check the linear assumption of homoscedasticity. The scatter plots show that the error is not constant along the values of the dependent variable.

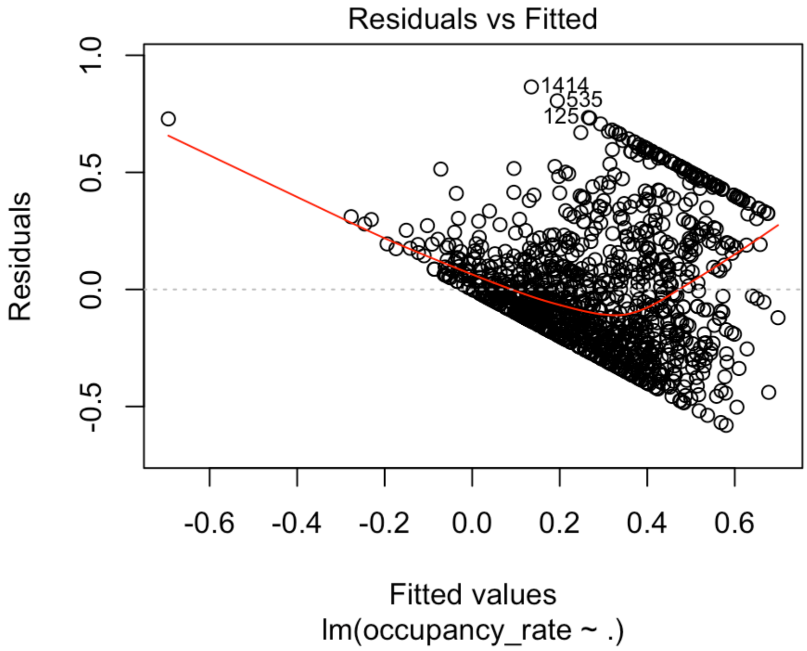


Figure L2: *Scatter-plot - residuals against the dependent variable - Lancasterian model.*

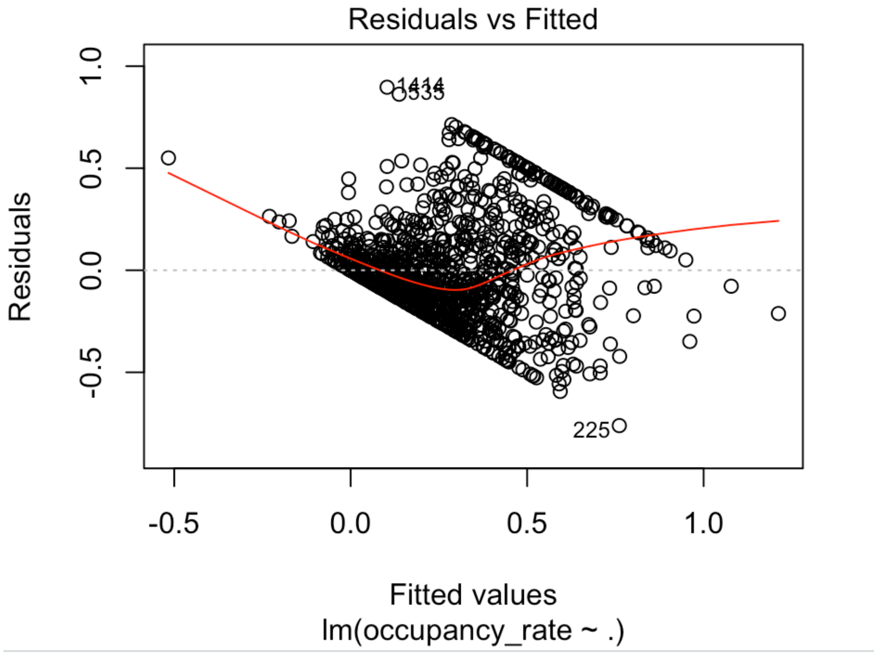


Figure L3: *Scatter-plot - residuals against the dependent variable - Lancasterian + Social Learning model.*

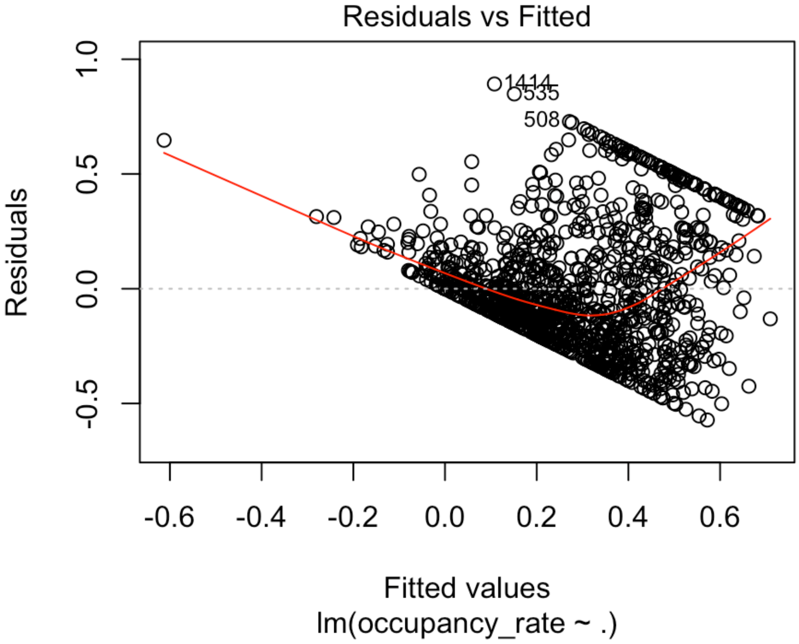


Figure L4: *Scatter-plot - residuals against the dependent variable - Lancasterian + Presentation model.*

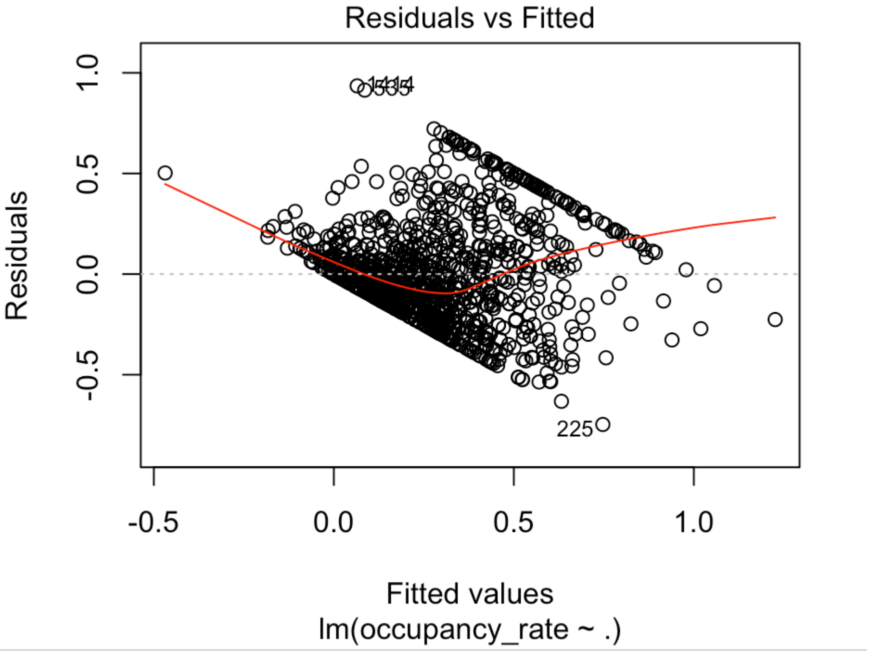


Figure L5: *Scatter-plot - residuals against the dependent variable - Full model.*

Appendix N

SHAP interaction plots.

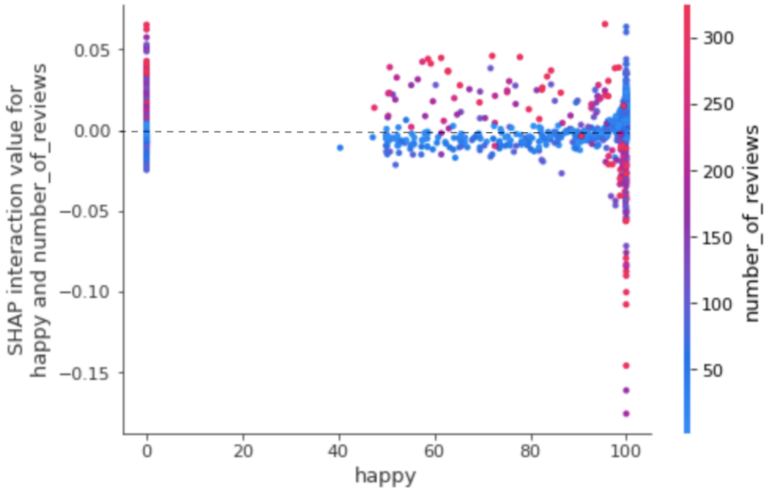


Figure L6: *SHAP interaction plot between the happy expression of a host' profile picture and the number of reviews.*

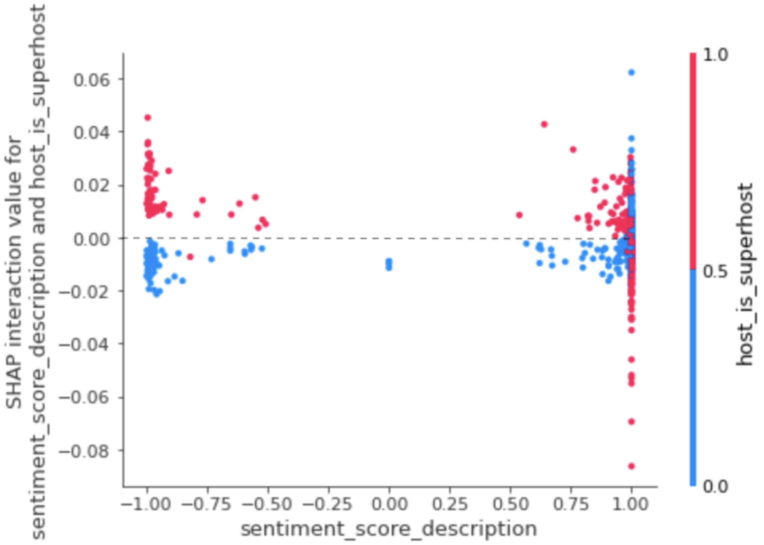


Figure L7: *SHAP interaction plot between the sentiment in the listing description and whether the host is a Superhost.*

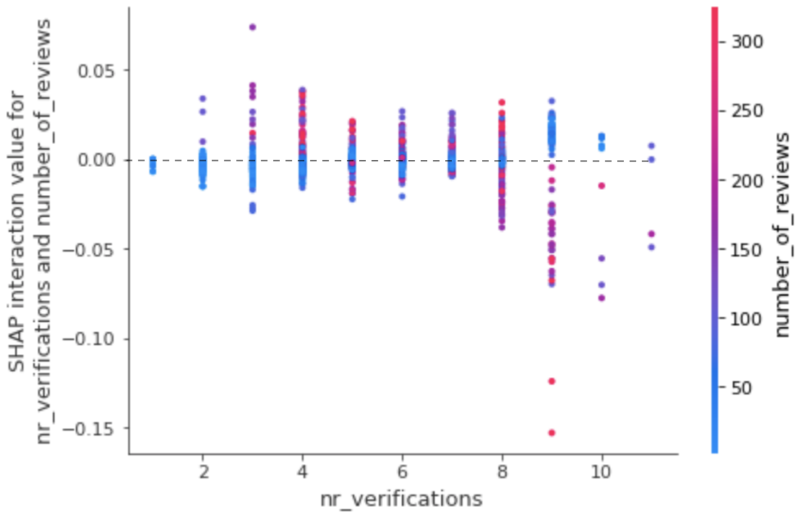
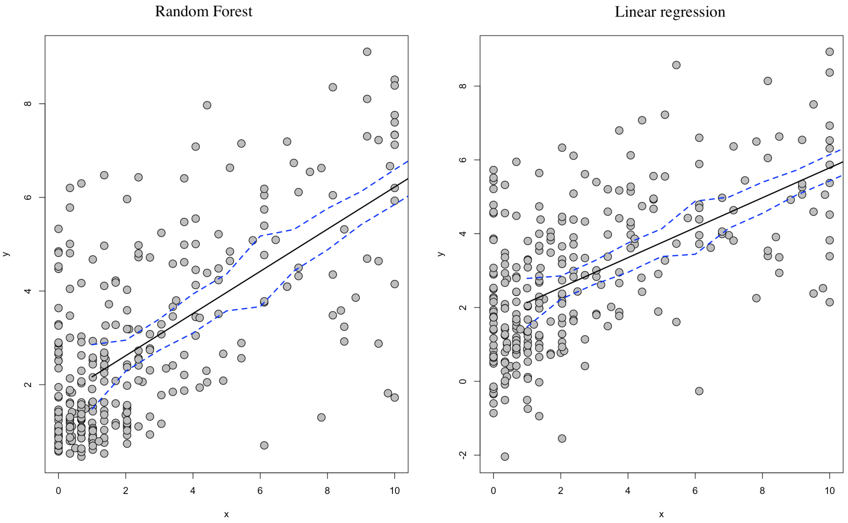


Figure L8: *SHAP interaction plot between the number of host' verification's and the number of reviews.*

Appendix O

The following plot includes the predicted (y-axis) against the observed (x-axis) values of the Random Forest model and the best performing linear regression model. The best performing linear regression model contains the variables from the Lancasterian, Social Learning and Presentation Perspective.


Random Forest

Linear regression

Figure L9: *Predicted against the observed observations of the Random Forest (left) and Linear regression (right) model.*