ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Behavioural pricing revisited: An improved matching algorithm to test psychological pricing effects in Airbnb listings

Student: Megan van der Ham

Student ID number: 483271

Supervisor: N.M. Almeida Camacho

Second assessor: A. Alfons

January 9, 2022

Table of contents

Executive Summary

Common wisdom suggests that "nine-ending prices" work, i.e., consumers are more likely to purchase products or services when prices end on a 9-digit instead of on a round number (e.g., €49.99 vs. €50). While there is substantial evidence in the literature supporting this psychological pricing effect, the vast majority of existing studies use primary data – typically laboratory experiments – to test such effect. The reliance on experimental data is logical since the most important advantage of experimental research is its reliance on randomization to estimate causal relationships. However, several researchers have also raised concerns with the external validity of laboratory experiments, i.e., whether the results obtained in controlled but often unnatural settings would replicate in the "real world".

In this thesis, the goal is to use evidence "from the field" to test whether it can replicate the psychological pricing effect in an important emerging industry: the sharing economy. In contrast with experiments, secondary data does not rely on randomization and, therefore, is often criticized by its lack of internal validity (i.e., its capacity to establish causal relationships). This happens because price changes observed in secondary data may suffer from *treatment selection bias*, i.e., the hosts who opt for a nine-ending price may differ in important characteristics from the hosts who opt for a rounded-number price, limiting researchers' capacity to estimate a causal effect. Therefore, in order to ensure that the results have both internal and external validity, the data requires to be pre-processed to ensure having a "quasi-experiment". In other words, this thesis needs to control for pre-existing differences between groups (treatment vs. control), in order to tease out the casual effect from other confounding factors.

Matching methods are often used in an attempt to ensure that treatment and control groups are "as equal as possible" (often measured through the extent they are able to balance these groups in a series of covariates, i.e., *covariate balance*). Unfortunately, classical matching is often criticized by not achieving sufficient covariate balance. Therefore, in this thesis an approach that attempts to optimize standard matching is proposed. Specifically, two matching approaches are compared: (1) a "Standard Matching" approach, i.e. matching observations (i.e., Airbnb properties) through a classic propensity score matching approach, versus (2) a "Cluster-then-Matching" approach, i.e., first clustering the data to find comparable 'types' of properties, followed by the standard matching approach employed on the clustered data. Thus, this thesis offers two contributions to the literature. First, it contributes to the marketing literature by testing the generalizability of one of the most important effects in behavioural pricing. Second, it contributes to the literature on causality with nonrandom data, namely the matching literature.

The results show that clustering before matching improves the matching results (i.e., increases covariate balance). After pre-processing the data in the two different manners, a DID analysis is performed which showed insignificant results for the causal effect of the treatment, leading to the impossibility of drawing conclusions about a psychological pricing effect using secondary data. It cannot be stated with certainty that these studies cannot be replicated with secondary data, since there could be other reasons for insignificancy. Examples of these limitations are the existing time trend due to the COVID-19 pandemic, but also research related limitations such as the sample size, the sector used to find this effect (sharing economy), and geography. This thesis can therefore be viewed as a pre-review and further research is encouraged to really exclude evidence for psychological pricing effects in secondary data, as well as prove that clustering does improve matching results.

Keywords: Psychological Pricing; Causality; Non-random Data; Quasi-experimental Methods; Propensity Score Matching; Clustering; Matching Algorithms.

1. Introduction

1.1 Psychological pricing

Are you more likely to buy a product when it is priced as €4.99 instead of €5.00? If the answer is yes, you are a dupe of a *psychological pricing* strategy, which is the case for most people. As the name implies, this strategy is based on prices and offers having a psychological impact. This particular psychological pricing strategy is also known as *nine-ending pricing*, *charm pricing*, *odd(-ending) pricing* or *just-below pricing*, however, in this thesis the term 'psychological pricing' is used (unless another term is more appropriate at that moment). Psychological pricing is a pricing strategy that makes use of the perception of the customers, by changing the price to prices ending on a 9-digit instead of on a round number (e.g., €0.99 instead of €1.00, or even €89 instead of €90). It is proven that this leads to the product being more attractive on price level.

The merits of a psychological pricing strategy have been studied extensively in the consumer behaviour literature, typically using primary data collected either via lab experiments or survey data (see, e.g., Table 1 and 2 in Section 2.2). Thomas and Morwitz (2005) found that the *anchoring heuristic* plays a role in this phenomenon. The anchoring heuristic is found by Tversky and Kahneman (1974). A heuristic is a mental shortcut which enables people to put minimum effort into decision making (Myers, 2010). The anchoring heuristic works as follows: people tend to have an initial judgement, called their *anchor*, and from there on their judgement is adjusted (Lieder, Griffiths, Huys and Goodman, 2018). For pricing decisions, the anchoring heuristics looks as follows: when the price is considered in the mind, the numerical differences are anchored on the left digit, causing €1.99 to be more attractive than €2.00. Here, the first named price is associated more with €1.00 than with €2.00, causing people to see it as cheaper than it actually is even though the difference is only as little as one cent, which is also known as the *left-digit effect* (Thomas and Morwitz, 2005). This pricing strategy is globally used in all industries.

Research about psychological pricing has been done for a substantive amount of time. Thomas and Morwitz (2005) examined the perception of psychological prices and found that indeed these prices are perceived to be relatively smaller than a price one cent higher. In addition, they find that this 9-ending effect is not only restricted to prices, but holds for other entities as well. Guéguen and Jacob (2005) also found that psychological prices indeed led to an increase in the amount of purchases. Another research performed by Bizer and Schindler (2005), where respondents were asked to estimate how many products can be purchased with a specific amount of money, found that respondents thought they were able to purchase significantly more products priced with 9-endings compared to products with comparable prices containing 0-endings. In 2009, Manning and Sprott also found that once cent differences affects choice towards 9-ending prices.

Similar to the previously discussed papers, this thesis investigates the causal effect of psychological pricing. This pricing strategy is an important topic and has been studied multiple times. However, what can be noticed in these previous discussed papers, is the methods used to analyse the effect, which is typically by the means of surveys and experiments where primary data is gathered. In contrast to the previously discussed papers, this paper uses secondary data to find whether a causal effect of psychological pricing exists, leading to the following main research question:

"Is there evidence for psychological pricing effects in secondary data?"

To investigate the (causal) effect of psychological pricing strategy using secondary data, data with sufficient price variation is needed. In other words, data from an industry where pricing decisions are frequently made and are an important determinant of customers' decisions. For this research, the focus lies on the sharing economy, and in particular on a dataset with hosts' pricing decisions and subsequent booking decisions by guests made in the platform Airbnb. For Airbnb, the price of an accommodation is one of the most important aspects since it is one of the main things a person considers when booking. One of the advantages here, is that the hosts can change their price very easily, meaning that the price of Airbnb accommodations is very volatile. This makes evaluating the causal effect of pricing more feasible. Second, this sector demands cognitive effort leading to a high reliance on heuristics (Kleinsasser and Wagner, 2011). Lastly, there are no other factors, such as discounts, that influence the pricing in this sector, meaning that the data can be seen as the "laboratory" for pricing in general.

For this main research question, the first thing that needs to be analysed is the trade-off between external and internal validity. Since the goal is to generalize the findings to all fields, the external validity should be high. However, the goal is also to find the causal effect and therefore, it is important that it is not influenced by other factors, meaning that the internal validity needs to be high as well. Thus, it is important to find a possible way that considers both external and internal validity in the most optimal form.

1.2 Optimizing data pre-processing: The importance of matching

Matching algorithms, which are often used in both economics and the healthcare field, are critical in causal analyses of secondary data. Matching offers a solution for typical challenges faced by researchers seeking to understand causal effects in observational studies, where randomly allocating the treatment is either not ethical or not possible (Rubin, 1973). Through matching a researcher "pairs" observations with similar observable characteristics but, simultaneously, exposed to different "treatments", creating a usable counterfactual. This improves a researcher's capacity to properly estimate the causal effect of such treatment, despite the absence of a randomized allocation of subjects to treatment and control groups. Hence, matching methods are often used for estimating causal effects between two different treatment groups (Scotina and Gutman, 2019).

Matching is often performed on the raw data. However, a key goal of this thesis is to investigate if an extra step – a machine learning data pre-processing step done before matching – improves the final results of matching and, therefore and purportedly, the capacity to make causal inferences from a secondary dataset (in this case applied to the important question of whether psychological pricing strategies pay off). The extra data pre-processing step that will be analysed is a cluster analysis, leading to the following two alternative approaches:

(i) *"Standard Matching":* Perform a matching algorithm on the raw data using propensity score.

(ii) *"Cluster-then-Matching*": Cluster the raw data first, followed by performing the same matching algorithm on the clustered data.

The matching algorithm that is used is *Propensity Score Matching* (PSM), and the clustering method that is applied in the Cluster-then-Matching approach is *Hierarchical Clustering* (HC), for the following reasons. PSM is one of the most popular matching methods, however, it has its critics as well. King and Nielsen (2019) show that PSM often increases imbalance, bias and model dependence, which is disadvantageous for the results since the goal is to balance the data as well as possible to be able to find unbiased causal estimates that are externally valid too. The popularity of PSM remains high despite these limitations, especially in fields that are less related to *Data Science,* due to its simplicity. However, the main claim in this thesis is that clustering the data and then performing the matching algorithm on the clustered data will improve the matching results because, by definition, the within-cluster variation in the data will be lowered (when compared with the non-clustered data), which "simplifies" the matching task[1]. This thesis relies on HC because it is a powerful (clustering) technique, and in comparison to other possible clustering alternatives such as K-means, no advanced knowledge about the required number of clusters is needed beforehand. HC is a *machine learning algorithm* and the machine learning "workflow" is often considered as circular since it can learn from experience (Kwartler, 2017). Machine learning methods can identify patterns and make predictions about unobserved data (Murdoch, Singh, Kumbier, Abbasi-Asl and Yu, 2019).

However, it cannot easily be implied that the more advanced a method is, the better it will perform. Hence, both approaches – Standard Matching and Cluster-then-Matching – are compared in this paper to find if adding a cluster analysis improves matching results and how, with the end goal of estimating

[1] Cluster analysis (or clustering) is used to group observations in a cluster in which they are more similar to each other than to observations in other clusters. This eventually leads to matching within a selection of observations that are already quite similar instead of matching using the entire dataset.

the causal effect of psychological pricing. Meaning that, before being able to answer the main research question regarding whether there is an evidence for psychological pricing effects in secondary data, a secondary research question needs to be answered first, which is the following:

"Does adding a cluster analysis before matching the observations improve the matching results?"

Answering this secondary research question helps us understand and evaluate the validity of the answer to the main substantive research question of this thesis. Moreover, it is possible to show if with this particular topic (psychological pricing), data pre-processing influences matching quality and the robustness of secondary data analysis results, which is a contribution to empirical pricing research in and of itself. Finding out if there is evidence for psychological pricing effects in secondary data is done by performing the *Difference in Differences* (DID) method. In order to perform this, it is of importance for this thesis to know whether matching solely or a combination of matching and clustering performs best in preparing the data for the DID analysis.

1.3 Research contributions

1.3.1 Substantive academic contribution

As briefly discussed before, previous literature has shown the effectiveness of psychological pricing. However, this is mostly done with experiments and surveys, leading to primary data. What this paper contributes is the fact that it proves whether or not there is evidence for psychological pricing effects in secondary data.

1.3.2 Methodological academic contribution

As discussed, both internal and external validity should be high in this paper. Internal validity is important because of the causal effect that is measured, and external validity is important since the main research question aims to generalize the findings of psychological pricing effects to other fields as well. To find the balance between these two requirements, the methodology has to be as optimal as possible. Therefore, it is decided to compare two different approaches, Standard Matching and Cluster-then-Matching, which adds an extra machine learning method as data pre-processing step. As mentioned, it cannot easily be implied that the more advanced a method is, the better it will perform. Therefore, this paper demonstrates if adding clustering (a machine learning method) as an extra data pre-processing step before matching will improve the matching results, which is something that is not studied in previous literature yet.

1.3.3 Managerial contributions

The sector used to demonstrate the findings is the sharing economy, with the focus on the company Airbnb. Discovering the effectiveness of psychological pricing for Airbnb listings can help the owners who list their accommodation, understand the importance of pricing and possibly manage their pricing strategy in a way that proves to be more successful. In turn ensuring Airbnb's success.

1.4 Research structure

First, theory about psychological pricing is discussed to draw a line between what has already been done and what this paper will contribute. Next, the methodology is emphasised since it clarifies the different approaches and corresponding methods, namely (1) a "Standard Matching" approach, i.e., matching observations (i.e., Airbnb properties) through a classic propensity score matching approach, versus (2) a "Cluster-then-Matching" approach, i.e., first clustering the data to find comparable 'types' of properties, followed by the standard matching approach employed on the clustered data. After that, data of Airbnb listings were collected online. However, this data does not include all the necessary variables, and therefore, the variables *occupation rate* and *treatment* were created manually. Next, the results are discussed including pre-processing the data using the different approaches. With the help of several performance metrics – which are discussed later in this paper – the matching quality can be evaluated and it can be concluded if adding a clustering method to the data pre-processing part before matching the data improves or does not improve matching results. Lastly, the matched dataset is used to answer the primary (substantive) research question. This is done by performing a differences-in-differences (DID) analysis. In short, relying on the dataset created using the 'winning' data pre-processing approach, a DID analysis is performed that exploits variation in properties' occupancy rates before-versus-after pricing changes, leading to an answer to the main research question.

1.5 Preview of research findings

The main findings are as follows. When evaluating the matching results of both approaches, it can be concluded that clustering does improve matching results in this thesis. Clustering improves the balance in binary variables to total balance, i.e., the Standardized Bias (SB) for each binary covariate is reduced to 0, which was impossible to do with the matching algorithm solely. After clustering is performed, the matching algorithm could improve balance in the other covariates and thus, find more complete matches. Using this 'winning' approach, a DID analysis is performed which showed insignificant results for the causal effect of the treatment, leading to the impossibility of drawing conclusions about a psychological pricing effect using secondary data. It cannot be stated with certainty that these studies cannot be replicated with secondary data, since there could be other reasons for insignificancy that can be found in the results of the DID analysis. They show a significant negative treatment group specific effect, however, this estimate is biased due to an additional significant negative time trend. This means that the insignificancy of the psychological pricing effect (e.g. the pure causal effect) is most likely due to a predominant existing time trend (probably the COVID-19 crisis), which proves to have a more powerful effect on the occupancy rate than the treatment does.

2. Literature review

Substantive literature is available about psychological pricing, since research draws back to at least the 1930s by Bader and Weinland (Schindler and Kibarian, 1996). This pricing manner occurred during the late 19th century and there are several speculations regarding how it is developed: either because advertisers lowered prices with one cent so that the one cent priced Chicago Daily News could be sold more easily, or to control employee theft (Landsburg, 2012). Several papers have been written about the most-used end digits, showing that *"there is a definite bias in favour of odd price endings"* (Holdershaw, Gendall and Garland, 1997), especially towards 5 and 9. The final numbers can be debated, however, it shows that the companies do strategically think about their price's last digit(s). Even now, when some countries such as the Netherlands, Finland and Australia decided to eliminate the 1 and 2 cents, it is still common practice to price products with a .99 ending. This shows that this pricing strategy is effective, most likely for other reasons than monetary ones as well. This chapter is divided into two parts, both discussing a central topic: (1) the first part discusses why psychological pricing works (i.e., the cognitive and emotional triggers at play when a customer gets in touch with psychological pricing), and (2) the second part discusses how and how much psychological pricing influences consumer behaviour (i.e., empirical evidence documenting the magnitude of its effects on key variables of interest such as quality perception and the underestimation of purchasing power).

2.1 Behavioural mechanisms behind psychological pricing

The classic price theory assumes that consumers behave rationally during their purchasing process, however, this is often not the case. With a behavioural pricing strategy, the pricing depends on the behaviour of (potential) customers, who are most likely to behave irrationally. Nützenadel and Hartmann (2021) describe this strategy as follows: *"Behavioural pricing assumes that consumers receive an objective stimulus, which leads to a subjective evaluation, which then provokes individual behaviour (response)".* This subjective evaluation exists, inter alia, of price perception, consumer's numerical processing capabilities and the memory of prices (Cheng and Monroe, 2013; Homburg and Koschate, 2005). One way of evaluating price is by comparing it to a "price threshold" – this indicates a price point at which the consumer's price response changes disproportionately (Monroe, 1990). Comparing odd and even prices is the best-known example of relative price thresholds.

It has been proven often enough that psychological pricing is effective, but what exactly are the causes of this? The main cause for the effectiveness of psychological pricing is the fact that consumers associate prices with 9-endings with a lower price level than it actually is. As briefly introduced in the Introduction, this phenomenon is caused by heuristics – also known as mental shortcuts to decrease the effort of decision making (Myers, 2010) – and in particular by the anchoring heuristic (Thomas and Morwitz, 2005). This anchoring heuristic is found by Tversky and Kahneman in 1974 and indicates that people tend to have an initial judgement, which is also known as the anchor, and from there on adjust their judgement to establish their final (often incomplete) judgement (Lieder, Griffiths, Huys and Goodman, 2018). Tversky and Kahneman (1974) explain in their paper that reliance on heuristics, such as the anchoring heuristic, leads to cognitive biases and that these biases are like perceptual errors. In the case of psychological prices, it is the consumers' reliance on heuristics that leads to the fact that they see psychological prices as being cheaper than they actually are. In 2005, Thomas and Morwitz found the left-digit effect, which *"refers to the observation that using a nine ending versus a zero ending, for example, $2.99 versus $3.00, changes the leftmost digit (i.e., the dollar digit changes from three to two) and that it is this change in the left digit, rather than the one cent drop, that affects the magnitude perception".* This means that psychological pricing is only effective when the leftmost digit changes, which became an important finding in the field of behavioural pricing.

Previous literature shows that respondents were convinced that they could purchase significantly more products with 9-ending prices compared to products with round number prices (Bizer and Schindler, 2005; Schindler and Kirby, 1997). This can be seen as evidence that the *drop-off mechanism*, which is the possibility that consumers either ignore or give little attention to the ending digits when considering a price, exists in price information processing (Bizer and Schindler, 2005). One of the reasons for relying on heuristics during this processing is because consumers are exposed to a continuous flow of information on prices (Brenner and Brenner, 1982). Since information processing must be done in a very short interval and an additional step should be taken when rounding the number upward, it is often chosen to only store the first digit of a number which is perceived to be the most valuable (Brenner and Brenner, 1982). Since this information processing of psychological prices exists of tedious mental calculations, consumers rather ignore the rightmost digit and as a result, it influences decisions on how much they buy (Schindler and Kibarian, 1996).

Thus, the psychological prices are often underestimated and either perceived as being equal to the leftmost digits during left-to-right processing, or sometimes consumers even perceive a psychological price as a round price from which they get a small amount back (Schindler and Kirby, 1997). In addition, psychological prices are less likely to be recalled accurately compared to rounded prices, because the focus that is put on the leftmost digit leads to price underestimation (Schindler and Wiman, 1989). To conclude, the combination of information abundance and the urge to simplify the price computation by the means of heuristics, leads to perceiving psychological prices to be lower than they actually are.

2.2 Empirical evidence of psychological pricing effects

Even though the psychological pricing strategy emerged for reasons that had nothing to do with influencing consumer's behaviour into buying the product, it is nowadays a frequently used pricing strategy to increase profits. In this part, the empirical evidence of psychological pricing effect is examined. Literature proving the effectiveness of this pricing strategy is discussed, separated in lower-priced and higher-priced goods, both being complemented by a table (Table 1 and 2) detailing the used methodology and most important findings.

2.2.1. Psychological pricing in retail sector

In 1973, Holloway concluded that despite this topic being studied for some decades, there is only little knowledge about the actual effectiveness on sales. After that, Schindler and Kibarian (1996) and Guégen and Jacob (2005) proved that the use of psychological pricing can increase the purchase amount spent, since a large proportion of people ignores the rightmost 9 digit(s), which results in a substantive potential underestimation of price (Schindler and Kibarian, 1996). Schindler and Kibarian (1996) even find an increase in sales to be as high as 8%. However, consumers who are made aware of the difference between psychological and round prices are no longer influenced by it, since this effect *"is mediated by guilt reduction"* (Choi, Li, Rangan, Chatterjee and Singh, 2014). It is concluded that for low-priced items specifically, psychological pricing can lead to a greater choice share than round pricing (Manning and Sprott, 2009). Macé (2012) confirmed this finding by stating that *"a nine-ending price is more effective for increasing sales of small brands (e.g., low market-share, low price, and new items) that belong to weaker categories (e.g., low price, low budget-share)".* In addition, she concluded that psychological prices can even lead to sales losses for premium brands.

Continuing with the low-priced goods, in the retail sector it can be concluded that the price perception influences the quality image, and specifically that psychological prices have a negative effect on the quality image (Schindler and Kibarian, 2001; Schindler, 2006; Kreul, 1982). However, the effect of psychological pricing can also be generalized to other business situations (Guéguen and Jacob, 2005). Just as when purchasing lower-priced goods, consumers purchasing higher-priced goods might be influenced by price endings as well (Kleinsasser and Wagner, 2011). Therefore, the next section will discuss psychological pricing effects for higher-priced products, and how it affects the quality image.

2.2.2 Psychological pricing in tourism

Given that in this thesis the focus lies on the pricing of Airbnb accommodations, which are typically used by tourists, the existing psychological pricing literature in the tourism sector is reviewed as well. Psychological pricing is often used in the online environment where price comparison is easy (Jeong and Crompton, 2017). Nowadays trips and other stays are booked online where there is a large supply of places to stay, leading to hotels, motels, B&B's, etc. having to differentiate their listings in a way that it catches peoples' eye in a few seconds. One of the possibilities to do this, is by choosing the right pricing strategy. In addition to the retail sector, which overall represents the lower-priced goods, the effectiveness of psychological pricing in tourism, which can be seen as representing the higher-priced goods, is important as well. However, only little research about this has been done. According to Kleinsasser and Wagner (2011), this sector is an appropriate study context for psychological pricing, since purchasing decisions in 'the hotel business' demand cognitive effort leading to consumers often relying on purchase decision heuristics, which makes them vulnerable to psychological pricing effects. The few existing studies have shown that the left-digit effect also exists for these higher-priced goods.

When personal involvement is high, it leads to a stronger reaction to odd-ending versus round number prices compared to when personal involvement is lower (Kleinsasser and Wagner, 2011). In the case of booking a stay, personal involvement is perceived to be higher and therefore, price endings might play a big role. In addition, vacations can be seen as hedonic purchases, and it is proven that psychological prices are more effective when it comes to hedonic purchases compared to utilitarian purchases (Choi, Li, Rangan, Chatterjee and Singh, 2014). As explained, psychological prices affect the quality image. When looking at tourism, there exists a relationship between the rating of the rooms and their price-ending strategies: when the average rating of a room decreases, the price-ending strategy changes from round number digits to psychological prices (Collins and Parsa, 2006). This shows that not only from the consumer's perspective but also from the business' perspective, 9-ending prices are associated with lower quality. The study of Kim, Cui, Choi, Lee and Marshall (2020) proves that participants that are exposed to psychological prices prefer the lower priced/quality hotels to the higher priced/quality ones, because of the price perception. In addition, Zou and Petrick (2020) proved that psychological pricing is *"an effective tactic to increase purchase intentions for low-priced hotels"*. This means that the general takeaway of psychological pricing in tourism is that it plays a big role, however, it is correlated with a lower quality perception of the room. This may not always be optimal for hotels, etc. and therefore, using this pricing strategy should be well thought through. When it comes to Airbnb, price does affect the consumer's satisfaction of an accommodation as well (Hamari, Sjöklint and Ukkonen, 2016; Luchs, Naylor, Rose, Catlin, Gau, Kapitan and Simpson, 2011) since the company is known for being an economical alternative (Zervas, Proserpio and Byers, 2017). However, nowadays Airbnb also provides luxury accommodations, leading to the consideration whether psychological pricing is effective for all accommodations.

Table 1. Literature overview table of studies on the effectiveness of psychological pricing in retail.

Study	**Methodology**	**Data type**	**Main findings**
Bizer, G.Y., & Schindler, R.M. (2005) *Psychology & Marketing*	Respondents participating in the experiment were presented with hypothetical prices of items and were asked how many of these products could be purchased with a specific amount of money.	Primary: Lab Experiment Data.	Evidence for the existence of the drop-off mechanism in price information processing is provided. This paper showed that consumers were convinced they could purchase more products with a psychological price compared to products with round number prices.
Brenner, G.A., & Brenner, R. (1982) *Journal of Business*	Creating a model that is based on Becker's model of demand for basic commodities.	-	They state that one of the reasons for relying on heuristics is because consumers are exposed to a continuous flow of information on prices. Since information processing must be done in a very short interval and an additional step should be taken when rounding the number upward, it is often chosen to only store the first digit of a number which is the most valuable of this whole number.
Choi, J., Li, Y.J., Rangan, P., Chatterjee, P., & Singh, S.N. (2014) *Journal of the Academy of Marketing Science*	54 students participated in an experiment where they had to choose between two almost identical laptops.	Primary: Lab Experiment Data.	Their findings state that consumers who are made aware of the difference between 9-endings and round prices are no longer influenced by it, since this effect *"is mediated by guilt reduction"*.

Table 1 (continued). Literature overview table of studies on the effectiveness of psychological pricing in retail.

Study	**Methodology**	**Data type**	**Main findings**
Gúeguen, N., & Jacobs, C. (2005) *Journal of Applied Sciences*	A two-day experiment with 241 mostly women-customers in the cheese department in a grocery store.	Primary: Lab Experiment Data.	They confirmed the possibility of generalizing the positive effect of psychological prices on the mean purchase amount spent by customers to another business situation.
Macé, S. (2012) *Journal of Retailing*	Uses store-level scanner data of grocery goods with prices that include two digits after the decimal, across 83 stores for 399 weeks.	Secondary: Panel Data.	(1) She found that psychological prices are more effective when wanting to increase sales of small brands (e.g., low market-share, low price, and new items) that belong to weaker categories (e.g., low price, low budget-share). (2) In addition, she concluded that psychological prices can even lead to sales losses for premium brands.
Manning, K.C., & Sprott, D.E. (2009) *Journal of Consumer Research*	Three experiments were performed containing different price conditions for pens, gifts for friends and acquaintances and for themselves.	Primary: Lab Experiment Data.	Their main finding is that for low-priced items, psychological pricing can lead to a greater choice share than round pricing.
Schindler, R.M., & Kibarian, T.M. (1996) *Journal of Retailing*	(1) Experiment on two differently priced 169-item clearance catalogs of women’s clothing. (2) 90.000 customers received three differently priced catalogs.	Primary: Lab Experiment Data.	(1) A large proportion of participants ignored the rightmost 9 digits, resulting in a potential underestimation of at least 5% of the price, going up to 20% or more, and confirming that 9-ending pricing can increase the purchase amount spent. The reason is the requirement of tedious mental calculations when processing the price information. (2) The 99-ending catalog led to 8% more purchases and larger amounts per purchase than the round digit catalog.

Table 1 (continued). Literature overview table of studies on the effectiveness of psychological pricing in retail.

Study	**Methodology**	**Data type**	**Main findings**
Schindler, R.M., & Kirby, P.N. (1997) *Journal of Consumer Research*	Sampled selling prices of advertisements in newspapers.	Primary: Lab Experiment Data.	(1) Consumers tend to perceive psychological prices as round-number prices but in addition they get a small amount given back. (2) Consumers tend to underestimate a psychological price because they perceive the price being the first round number during incomplete left-to-right processing.
Schindler, R.M., & Wiman, A.R. (1989) *Journal of Business Research*	198 college undergraduates were showed a set of pictures with prices. Two days later, the same picture was shown to the same students, however, this time without prices and the students were asked to recall the prices.	Primary: Lab Experiment Data.	It was found that odd-ending prices are less likely than even-ending prices to be recalled accurately and that expressing a price as an psychological price increases the likelihood that it will be underestimated when it is recalled.
Thomas, M., & Morwitz, V. (2005) *Journal of Consumer Research*	Five experiments in which 27 up to 154 undergraduate students participated.	Primary: Lab Experiment Data.	(1) They found the *left-digit effect*, meaning that psychological pricing is only effective when the leftmost digit changes, which became an important finding in the field of behavioural pricing. (2) This effect depends on both the numerical and psychological distances between two prices and the closer these prices are to each other, the more likely is the left-digit effect.

Note: Only the findings interesting for this paper are displayed here.

Table 2. Literature overview table of studies on the effectiveness of psychological pricing in tourism.

Study	Methodology	Data type	Main findings
Collins, M., & Parsa, H.G. (2006) *Hospitality Management*	(1) Qualitative interviews with hotel executives. (2) Comparison of the online ratings of rooms.	Primary: Lab Experiment Data.	Findings show that there is a relationship between the ratings of a room and their price-ending strategy: when the average rating of a room decreases, the price-ending strategy changes from round number digits to psychological prices.
Kim, J., Cui, Y., Choi, C., Lee, S.J., & Marshall, R. (2020) *Tourism Management*	5 experimental studies where the preciseness of the prices is manipulated (odd-ending vs. round) and 1 analysis using secondary data.	Primary: Lab Experiment Data.	When exposed to psychological prices, participants prefer the relatively cheaper vacation spots and lower priced/quality hotels compared to when exposed to round prices.
Kleinsasser, S., & Wagner, U. (2011) *Journal of Retailing and Consumer Services*	Survey questionnaire using conjoint measurement approach with price as of the attributes.	Primary: Lab Experiment Data.	(1) Price perception of higher-priced goods is influenced by price endings, just as for lower-priced goods. (2) Utility decreases when prices increase from an odd to an even price. (3) Personal involvement and price interest are moderating effects of price perceptions leading to a stronger reaction to psychological prices.
Zou, S., & Petrick, J.F. (2020) *Journal of Travel Research*	A hypothetical scenario-based experiment based on hotel reservations.	Primary: Lab Experiment Data.	(1) Psychological pricing increases purchase intentions for low-priced hotels. (2) Psychological pricing moderates the relation between perceived value and purchase intentions of hotel rooms.

Note: Only the findings interesting for this paper are displayed here.

2.2.3 Conclusion of psychological pricing studies

The reason behind the effectiveness of psychological pricing is that consumers use heuristics, in particular the anchoring heuristic (Thomas and Morwitz, 2005). Relying on these heuristics leads to cognitive biases (Tversky and Kahneman, 1974) which in turn lead to perceiving prices with 9-endings to be lower than they actually are. The explanation for using these heuristics is that consumers are exposed to a continuous flow of information on prices and since information processing existing of tedious mental calculations, must be done in a very short interval, it is often chosen to only store the first digit of a number which is perceived to be the most valuable (Brenner and Brenner, 1982). An important finding in the field of behavioural pricing is the left-digit effect by Thomas and Morwitz (2005), which proves that psychological pricing is only effective when the leftmost digit changes. Thus, the behavioural mechanisms behind psychological pricing can be summarized as the combination of information abundance and the urge to simplify the price computation by the means of heuristics, which lead to perceiving psychological prices to be lower than they actually are.

As can be established from this chapter, psychological pricing is an effective pricing strategy and empirical evidence of this effect is found for both lower-priced and higher-priced goods. It can increase the purchase amount spent with even up to 8% (Schindler and Kibarian, 1996; Guégen and Jacob, 2005), since these 9-endings result in an underestimation of price (Schindler and Kibarian, 1996) leading to the persuasion that it is possible to purchase more products with a psychological price compared to products with round number prices (Bizer and Schindler, 2005). Besides the effectiveness on sales, psychological pricing has an influence on the quality image of both lower-priced and higher-priced products as well. Thus, for both retail and tourism it can be said that psychological pricing is most effective when it comes to lower valued and lower quality goods or accommodations. In addition, personal involvement and products being hedonic or not also play a role in the effectiveness of psychological pricing (Kleinsasser and Wagner, 2011; Choi, Li, Rangan, Chatterjee and Singh, 2014). These studies show that there is substantial evidence for psychological pricing effects, which can be generalized over different fields.

2.3 Hypothesis: effectiveness of psychological pricing

An important aspect of these studies is the methodology used leading to the findings, which is presented in Table 1 and 2. All but one paper made use of experiments or surveys to measure effectiveness. This means that, broadly speaking, primary data is used to find the effectiveness of this behavioural pricing strategy. This is not a bad thing since experimental research has its advantages, with the most important one being: demonstrating causality. However, experiments and surveys have a few important disadvantages as well. Specifically, critics of (especially laboratory) experiments typically raise doubts about the reliability and (external) validity of experiments.

First, in terms of reliability, some scholars have expressed concerns that experimental subjects may not always be reliable and, thus, results retrieved from laboratory[2] experiments could depend on the participants and their behaviour in controlled settings. For instance, Paolacci, Chandler and Ipeirotis (2010) have addressed concerns about the participants of the Amazon Mechanical Turk study, since this might be an unrepresentative sample, and the experiment could possibly have not be taken seriously. In addition, the *Hawthorne effect* (French, 1953) which implies that participants behave differently in experiments as they are aware they are being observed (Franke and Kaul, 1978), can also play a role in experiments. Lastly, laboratory experiments can suffer from observer bias (Fitzpatrick, Preisser, Ellison and Elkinton, 2009). In addition, survey studies suffer from similarly problematic threats to respondent reliability, such as response bias (Baumgartner and Steenkamp, 2001), selection biases and common method variance (Rindfleisch, Malter, Ganesan and Moorman, 2008).

Several authors have questioned the external validity of surveys and especially laboratory experiments due to unusually artificial environment in which such experiments are conducted (e.g., Galizzi and Navarro-Martinez, 2017). Specifically, participants of laboratory experiments are placed in an artificial environment, containing artificial restrictions, limited choices and time horizons which may lead to observing different behaviour than that what would have occurred in the real world where, possibly, choices are limitless (Levitt and List, 2007). When it comes to surveys, they may not retrieve the most accurate and honest answers, because respondents are scared to give their honest opinion, they might lack knowledge or their answer might not be in the options (Malhotra and Nunan, 2019).

As mentioned, one of the papers that is reviewed (being Macé, 2012) uses secondary data to examine psychological pricing and its impact on consumer decisions. Specifically, the authors examined the causal effect of factors such as brand, category and store, on psychological pricing. However, the authors do not examine the causal effect of psychological pricing on a certain observed variable of interest, as is done in this thesis. Since – to the best of my knowledge – there is furthermore no clear evidence for the causal effect of psychological pricing using secondary data in existing literature, this thesis is trying to contribute to the literature by examining if it is possible to find these effects in secondary data, which might lead to the opportunity of being able to generalize this effectiveness to multiple fields. Because previous literature using primary data, substantively shows that there is evidence for psychological pricing effects, in this thesis the baseline expectation, and hypothesis, is to also find evidence for psychological pricing effects when using secondary data, and so:

Hypothesis 1: Using secondary data, evidence for psychological pricing effects on occupancy rate of Airbnb listings can be found.

[2] Here, the emphasis lies on laboratory experiments since natural or field experiments typically suffer much less from these biases (Harrison and List, 2004).

3. Causal Inferences in Secondary Data: Matching Algorithms

The effectiveness of psychological pricing using primary data is substantively showed in the previous chapter. However, no evidence exists for this effect in secondary data and therefore, the main research question in this thesis aims to find out if this holds for secondary data as well by examining a causal effect. This chapter theoretically explains what a causal effect is and how to find it using secondary data. In order to do this, the data should be matched, however, this comes with some drawbacks. In addition, it is explained how these drawbacks can be solved.

3.1 Causal inferences

Chambliss (2006) defined causal effects as something where *"variation in the independent variable is followed by variation in the dependent variable, when all other things are equal (ceteris paribus)"*, so shortly: one thing causes another. Causal effects are often measured by the means of a so-called 'treatment', which is the independent variable. The effect of this particular treatment on the dependent variable is what is represented by the causal effect. The goal of causal inferences are to find *"an unbiased estimation of the treatment effect"* (Sizemore and Alkurdi, 2019). The easiest way to explain it is with the help of testing a medicine. To find out if a certain medicine for, for example, headaches works, it is examined whether taking the medicine leads to less headaches. Thus, the causal effect of the medicine (treatment/independent variable) on headaches (dependent variable) is examined.

Particularly, the second part of the definition, "*ceteris paribus"*, is an important part of measuring causal effects. As Pearl (2003) states *"causal analysis deals with changing conditions"* such as treatments, and therefore, experimental control is needed to be able to verify causal assumptions (Pearl, 2003). In case of the previous example, it could be that other things influence the decrease in headaches as well, such as a decline in stress, no longer having a cold, etc. The goal is to keep the rest of the variables (the control variables) equal, however, (1) it is often impossible to make sure all other things are equal except for the variation in the independent variable (King, Keohane and Verba, 1994); and (2) it is impossible to go back in time to give or not give the treatment (Rubin, 1974).

The solution for this is taking a baseline into account when measuring a causal effect, also known as a counterfactual. Therefore, a causal effect is measured by taking the difference between the potential outcome if the observation was allocated to the treatment group and the potential outcome if this observation was allocated to the control group, representing the baseline (Hanck, Arnold, Gerber and Schmelzer, 2019). Since an observation cannot be allocated to both the treatment and control group, only one of the potential outcomes can be observed for every observation (Holland, 1986). This allocation should be randomly to make sure both groups will follow the same distribution eventually. Here, the naming "observation" is chosen explicitly, since the treatment does not necessarily have to be allocated to an individual, which does often happen to be the case. It could also be that the treatment is allocated to or found on materialistic objects, which is the case in this thesis.

3.2 Causality in secondary data

The optimal way to estimate causal effects is by allocating treatment and no treatment randomly within an experiment, with the goal that the treatment and control group eventually follow the same distribution such that one group can be seen as the counterfactual of the other group. Therefore, it is important to find well matching samples in both the treatment and control group (Stuart, 2010). However, it could be the case that the treatment is not ethical to allocate, such as smoking or car accidents, or the data is collected already and therefore, no influence can be exerted anymore. When this is the case, researchers have to use observational data, which is data where the independent variable has not been under control of the researcher. In other words, the counterfactual should be constructed from naturally occurring data. Previous literature criticised that only (properly) randomized experiments are able to show useful causal effects, however, Rubin (1974) discussed that controlled nonrandomized data can be used as well to find causal effects. Holland (1986) agrees by stating that experiments are not the only right framework for discussing causality, even though it is the simplest. However, when it comes to real world data, it is never possible to estimate the pure causal effect because such data lacks randomization, and thus, effects may be “polluted” by treatment selection bias (i.e., unobservables correlated with the treatment variable). Therefore, whenever causal effect is mentioned throughout this thesis, an approximate causal effect or, alternatively, a mimic of it is meant.

When performing a nonrandomized study, there often is a reasonable suspicion that external variables which affect the independent variable and also differ per treatment and control group, exist. However, this could be the case for randomized experiments too and therefore, when performing nonrandomized studies they could be ignored when irrelevant, just as would be done in randomized experiments (Rubin, 1974). Another issue is the fact that with nonrandomized studies, the treatment is not randomly assigned, which may lead to the treatment and control group not being directly comparable (Rosenbaum, 1984). This is a problem when wanting to estimate a causal effect, since it is needed that treatment and control observations which are most alike, are compared with each other in order to find the most significant effect. In this case, matching algorithms can be used to group treatment and control observations so both groups follow approximately the same distribution.

As introduced in the Introduction, matching algorithms offer a solution for researchers seeking to understand causal effects in observational studies, where randomly allocating the treatment is either not ethical or not possible (Rubin, 1973). Since assigning the treatment was not under control of the researcher when it comes to observational data, observations in the treatment and control group may quite differ (Rosenbaum, 1982). What these matching algorithms do is trying to replicate a randomized experiment by eliminating variation due to observed covariates, to make sure that the covariates' distributions are similar (Lopez and Gutman, 2017) with the result that direct comparison is more meaningful (Rosenbaum and Rubin, 1983). The final goal of matching observational data is pre-processing the data to reduce bias in estimating the treatment effect (Stuart, 2010). The most used matching algorithm for causal analysis in observational studies is propensity score matching (Pearl, 2010), which is used in this thesis as well.

3.3 Solving drawbacks through better data pre-processing

Even though propensity score matching is the general method used in these scenarios, it has its critics as well. King and Nielsen (2019) show that PSM often increases imbalance, model dependence and bias. In their paper, they discuss this by showing the *Propensity Score Paradox* (also *PSM Paradox*). This paradox implies that when matching is performed, the worst matched observations are pruned based on their propensity score distance, however, after finding a subset with approximately constant propensity scores, pruning is continued at a random manner. Random pruning increases imbalance and model dependence which in turn leads to biased estimates (King and Nielsen, 2019). In addition, Pearl (2010) also discussed the combination of Propensity Score and bias, concluding that it *"tends to amplify bias"*. He states that when the *back-door criterion*, which can be defined as the requirement that none of the control variables should influence the dependent and independent variable, is not met, propensity score cannot reduce bias (Pearl, 2008). However, no matter how well the observations are matched, it is always possible to find variables that systematically differ between the treatment and control group and thus might influence the causal effect, in both randomized experiments and nonrandomized studies (Rubin, 1974).

King and Nielsen (2019) prove that PSM matches globally, resulting in nonrobust outcomes since data space and analysis space are not the same, leading to violating the congruence principle. When matching is performed locally, it comes with an additional degree of robustness (King and Nielsen, 2019). Furthermore, matched pairs that have a larger distance than a chosen caliper are pruned (Rosenbaum and Rubin, 1985). When matches are made locally, the distance will be smaller and therefore, less pruning is needed. A possible solution to make PSM perform more locally is to perform a cluster analysis before starting the matching process. With the cluster analysis, the data is split into clusters (groups) of observations that are more alike to the other observations within the same cluster than to observations in other clusters. When PSM is performed within each cluster, the matching will be done more locally in comparison to data without clusters, possibly leading to a higher level of robustness.

One possible way of clustering is by finding pairs of clusters that lead to a minimal increase in total within-cluster variance when merged, this is called Ward's Method (Ward, 1963) and is used here. Variance represents the spread of the data, so when there is small variance, it means that the observations are relatively similar. As mentioned, this method clusters the observations in a way that the variance increases in the smallest way possible. Therefore, the distance between the observations is not taken into account, which could lead to the possibility that two clusters with the smallest distance will not be merged together because the increase in variance becomes too high compared to other merges. On the contrary, matching aims to find treatment and control observations with the same covariate distribution. The matching method used in this thesis makes use of the nearest neighbor algorithm, meaning that it looks at observations that are closest to each other in terms of propensity score. To conclude, when starting with clustering, the observations are grouped in a way that minimizes within-cluster variance. When performing matching within each created cluster, the observations that can be matched already belong to a group with the smallest variance possible and afterwards will be matched based on minimum distance between propensity scores within the clusters. By combining two different manners of finding comparable observations, it may lead to a better covariate balance compared to using just one. This approach will show if there is a necessary trade-off between minimizing variance of multiple observations and finding matched pairs leading to covariate balance between these observations.

In addition, matching only uses the covariates to find similar observations, meaning that the dependent variable plays no role in the matching process. It could be the case that this dependent variable is valuable as well in evaluating how similar observations are. Clustering can be done on the "dependent" variable in this thesis, since it is an unsupervised learning technique, meaning that there is no dependent variable. This means that when taking the dependent variable into account when clustering, the matching process afterwards can find different pairs of observation than with the non-clustered data which do not take into account the dependent variable.

Lastly, Li, Zaslavsky and Landrum (2013) found that when cluster level covariates are important, and the unconfoundedness assumption (which is explained in Table A2 in Appendix A) does not hold, a bias occurs. However, exploiting this multilevel structure, thus accounting for the clusters, while performing propensity score matching can reduce this bias.

3.4 Hypothesis: optimizing matching results

As discussed above, PSM often increases imbalance, model dependence and bias, because of the PSM Paradox and the additional random pruning. PSM matches on a global level, however, when adjusting the data pre-processing in such a way that PSM matches more locally, robustness increases. In addition, bias can occur when the assumptions are not met. One way to obtain this effect could be by starting with a cluster analysis and performing PSM within the clusters afterwards, and so

Hypothesis 2: Adding a cluster analysis as extra data pre-processing step before matching will improve the matching results (i.e., better covariate balance).

4. Methodology

The secondary research question aims to optimize the data pre-processing steps in order to obtain the best matching results that in turn can be used to better approximate a causal effect. Now, two key methodological contributions of this thesis are discussed. First will be discussed the proposed data pre-processing approaches to improve matching quality, i.e., a comparison of the "Standard Matching" with "Cluster-then-Matching" approach, including how to evaluate the matching quality to demonstrate the superiority of the latter approach. Second, this chapter discusses theory on how the causal effect can be found.

When constructing the counterfactual from naturally occurring data – in essence, what is done in this thesis when constructing two groups from observational data – three general assumptions must be met, which are given in Table A1 in Appendix A. Some of which are unable to be tested, and therefore should be assumed to hold anyways.

4.1 Data pre-processing: matching

As previously explained, the motivation for matching observations is to estimate causal effects in secondary data (Rosenbaum and Rubin, 1983). With matching, selection bias is eventually removed from the results leading to a purer average treatment effect (Sizemore and Alkurdi, 2019). Matching is part of the data pre-processing step and therefore, cannot be seen as being part of the model itself. There are different matching methods, however, the approach stays the same. The matching process looks as follows: (1) defining a measure of distance/closeness, (2) matching the treatment and control observations based on this measure, and (3) omitting the unmatched observations (Sizemore and Alkurdi, 2019). This thesis will pre-process the data using two alternative approaches (as shown in the Introduction):

(i) *"Standard Matching":* Perform a matching algorithm on the raw data using propensity score.

(ii) *"Cluster-then-Matching*": Cluster the raw data first, followed by performing the same matching algorithm on the clustered data.

For the first approach, PSM is performed on the raw (but cleaned) data. The second approach has an additional step, namely clustering the raw data first using HC analysis, and then performing PSM on the clustered data with the end goal of potentially increasing the matching results. Below, both approaches are explained separately for the sake of structure.

4.1.1 Standard Matching: using propensity score

Propensity scores are first introduced by Rosenbaum and Rubin (1983) and are the coarsest balancing scores, which *"can be used to group treated and control units so that direct comparisons are more meaningful"*. In order to estimate the effect of treatment solely, the distribution of the covariates should be the same for both the treatment and control group, leading to covariate balance (Sizemore and Alkurdi, 2019). Therefore, the goal of PSM is to find observations that have the same covariate distribution, however, differ in treatment. First, propensity scores are calculated and used to figuratively replace the distribution of covariates with one single value. Compared to randomized experiments where the propensity score is a known function, this function is almost always unknown in nonrandomized studies (Rosenbaum and Rubin, 1983). Propensity scores estimate the probability (propensity) of being treated given the observed covariates, and can be defined as (Stuart, 2010):

$$e_i(X_i) = P(T_i = 1|X_i) \quad (1)$$

Where $e_i(X_i)$ represents the propensity score, T_i represents the treatment (where 1 is being treated and 0 is not being treated, thus being control) and X_i represents the observed covariates.

Estimating the propensity score can be done using different models, from standard to algorithmic. The most popular propensity score estimation used is Logistic Regression (Stuart, 2010), which often proves to perform well enough. However, the last decades it is proven that machine learning approaches used to estimate the propensity score have a good performance as well (McCraffy, Ridgeway and Morral, 2004; Setoguchi, Schneeweiss, Brookhart, Glynn and Cook, 2008; Lee, Lessler and Stuart, 2009). In contrast to Logistic Regressions which come with parametric assumptions about the population distribution, it is possible to choose a nonparametric machine learning approach leading to possibly higher model performance for predictions (Brownlee, 2016). As Couronné, Probst and Boulesteix (2018) proved, Random Forests measure a higher accuracy than Logistic Regressions. Therefore, this thesis uses the well-known, generally well-performing machine learning method Random Forest to estimate the propensity scores.

Random Forest is a decision tree based approach and more specifically, a form of *Bagging*, which stands for bootstrapped aggregation. Simple decision trees suffer from high variance because of the ease of overfitting, which is undesirable (James, Witten, Hastie and Tibshirani, 2013). The Bagging approach decreases variance because it uses multiple trees. The process looks as follows: create *B* different bootstrapped samples out of the dataset, construct *B* decision trees by sequentially choosing the best predictor, and finally – because this thesis presents a classification problem – take the majority vote of the resulting predictions (Breiman, 2001). Random Forest works the same, except for one major difference leading to a decrease in variance and correlation between the constructed trees, and in turn increases the accuracy (Breiman, 2001). The difference lies in the way the trees are constructed, because Random Forest only considers a random subset of *m* predictors when constructing the trees, in contrast to Bagging where the entire set of *p* predictors is considered (James, Witten, Hastie and Tibshirani, 2013). The pitfall of Random Forests is the lack of interpretability. However, in the matching process the focus lies on the accuracy of the propensity and thus, no interpretability is needed. To conclude, Random Forests provide the most accurate results of all decision tree methods, at the expense of interpretability – which is no problem in this case – and therefore, fit perfectly into the matching approach of this thesis.

In addition to choosing the right approach to estimate the propensity scores, other model specifications need to be decided upon, which are (Sizemore and Alkurdi, 2019):

1. <u>One-to-one, one-to-k or k-to-one observations</u>: how many treatment observations can be matched to how many control observations?
2. <u>With or without replacement</u>: is it allowed to use a particular control observations for multiple matches?
3. <u>Maximum caliper distance (if necessary)</u>: represents the maximum absolute difference in propensity score allowed to create a match

The model specifications are partly dependent on each other. For example, since it could be possible that for two different treatment observations one particular control observation lies closest, it is chosen to use k-to-one matching, meaning that multiple treatment observations can be matched with one control observation. In order to achieve this, it should also be allowed to use replacement, meaning that certain control observations can be chosen multiple times as match to different treatment observations. This can often decrease bias as well (Stuart, 2010). One general concern could be that unavailability of qualitatively good control observations could lead to poor matches, since it will choose inadequate control observations that have a bigger distance to the treatment observation. However, this could be solved by imposing a caliper (Stuart, 2010), which is a parameter that should be specified when performing the matching process. When a difference in propensity score between a certain treatment observation *i* and control observation *j* is higher than this parameter, no match will be realized. The previously mentioned differences in propensity scores is used to match observations and is also known as the distance. This thesis performs matching based on finding the smallest distance between the propensity scores of treatment observations and control observations, also known as the Nearest Neighbor (NN) approach. Defining the formula of nearest neighbor matching of the observation *i* and *j* looks as follows (Stuart, 2010):

$$D_{ij} = \min | e_i - e_j | \tag{2}$$

Where D_{ij} represents the distance between the propensity scores of observation *i* and *j*, and e_i and e_j represent the propensity scores of observation *i* and *j*, respectively.

How matching is performed in this thesis, using the chosen specifications, is visualized in Figure A1 in Appendix A. The assumptions that must be met for propensity score matching can be found in Table A2 in Appendix A.

4.1.2 Cluster-then-Matching

The second approach is a combination of the matching process explained above and a clustering process. The goal of this approach is to first cluster the data and afterwards perform matching within the clusters in order to find out if this will improve the final matching results.

Clustering is an unsupervised learning problem, meaning that the analysis is about finding underlying structures in the data (Kwartler, 2017) rather than making a prediction for example. The goal of clustering methods is to identify observations that are more similar to the observations within that cluster than to observations within other clusters. This method is often used for data pre-processing, however, it can also be used on its own if the end goal is finding patterns or allocating (new) observations to a certain group. The type of clustering that is used in this paper is Hierarchical Clustering, which is a distance-based clustering method that aims to minimize distances between members within a group and simultaneously maximize the distance to members from other groups (Chapman and Feit, 2015). These distances are used to create a dissimilarity matrix which eventually is used to perform clustering. Since the data consists of mixed variable types, the distance measure used here to create a dissimilarity matrix is the *Gower distance (d)* (Gower, 1971). This metric applies a standardisation to each variable and combines the sum of all variable-specific distances to create the final distance between two observations (Maechler, Rousseeuw, Struyf, Hubert and Hornik, 2021). This can be defined as (Anand, 2020):

$$d_{Gower}(x_1,x_2) = 1 - \left(\frac{1}{p}\sum_{j=1}^{p} s_j(x_1,x_2)\right) \tag{3}$$

Where x_1 and x_2 represent the observations, p represents the number of dissimilarities being calculated and $s_j(x_1,x_2)$ represents the similarity function calculating the similarity between the observations.

Murtagh and Contreras (2011) discuss that with the agglomerative criterion, the hierarchies are more balanced, which is advantageous. The agglomerative Hierarchical Clustering approach begins with all observations solely in their own cluster and combines the clusters closest to each other, based on the dissimilarity matrix created by calculating the distances, into one new cluster. Once there is more than one observation in the clusters, it should be established what type of distance (or linkage) between the two clusters will be used. This thesis uses *Ward's Method* (Ward, 1963) as linkage method, which creates clusters that minimize the total within-cluster variance (Murtagh and Legendre, 2014).

The final results of HC can be visualized using a dendrogram. As Chapman and Feit (2015) explain, *"a dendrogram is interpreted primarily by height and where observations are joined"*. This is because the height represents the dissimilarity between clusters: the higher the observations are combined, the less similar the groups are (Chapman and Feit, 2015). What a dendrogram shows is the order in which the clusters are joined together. Therefore, this clustering method does not give you the optimal number of clusters to use for potential further analysis and thus, the final (number of) clusters need to be chosen manually, depending on what works best for your analysis. Once the clusters with similar observations are created, matching can be performed within each cluster in the way that is explained above.

As referred to in Section 3.3, this approach may prove if there is a necessary trade-off between minimizing variance of multiple observations and finding matched pairs leading to covariate balance between these observations. The improvement of adding clustering is the fact that matching takes place more locally, which improves robustness. In addition, when the unconfoundedness assumption does not hold and bias increases, accounting for clusters while performing propensity score matching can reduce this bias (Li, Zaslavsky and Landrum, 2013).

4.2 Evaluating matching quality: covariate balance

To find whether (clustering before) matching improves the matching results, the quality of the matches should be evaluated. As explained, matching aims to create similar distributions of covariates for both the treatment and control group to truly measure the effectiveness of the treatment (Sizemore and Alkurdi, 2019). This is also known as the covariate balance which essentially means, with the optimal result, that the treatment is unrelated to the covariates (Stuart, 2010). This thesis will use two common balance diagnostics presented by Rosenbaum and Rubin (1985) to evaluate the quality of the matches. The first balance diagnostic used is the *standardized difference in means*, also known as the *standardized bias* (*SB*), which is compared before and after matching in order to measure the possible reduction in bias (Rosenbaum and Rubin, 1985). When a reduction in bias occurs, it can be concluded that the matching process improved the covariate distributions of the treatment and control observations compared to before the matching took place, leading to balance improvement. It can be calculated by (Stuart, 2010):

$$\text{SB} = \frac{X_T - X_C}{\sigma_T} \qquad (4)$$

Where $X_T - X_C$ represents the difference in means of each covariate between the treatment and control group and σ_T is the standard deviation of the treatment group.

The second balance diagnostic used in this thesis is the *two-sample t statistic*, which compares the means of the covariates between the treatment and control group (Li, 2013) and can be calculated by (Moore, McCabe, Alwin and Craig, 2016):

$$t = \frac{(X_T - X_C)}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}} \quad (5)$$

Where $X_T - X_C$ again represents the difference in means of each covariate between the treatment and control group, s_T and s_C represent the standard deviation of the treatment and control group, and n_T and n_C represent the sample size of the treatment and control group.

4.3 Causal effect: Difference-in-Differences

Causal effects are measured by finding the real impact of a specific X (representing a treatment) on Y. Often, it is hard to find this effect because of the confounding effect of other covariates (Rajendran, 2019). One of the manners to estimate this causal effect is by performing a randomized experiment. As mentioned, this thesis makes use of observational data and therefore, it is not possible to retrieve perfectly random observations. Thus, there is a possibility of selection bias. For this reason, matching and a combination of matching and clustering is performed before the causal effect analysis of psychological pricing. Once the data pre-processing is completed, what will be left are matches that are created in such a way that the causal effect can be estimated. The method used for this analysis is a Difference in Differences (DID) method, which is a popular tool to analyse natural experiments. The pioneer of the DID method is the First Differences method. This method looks at the difference before and after the treatment (Rajendran, 2019). However, since with this approach there is no view of the initial trend (the counterfactual), it cannot be said with confident that the whole effect can be allocated to receiving the treatment. This is where DID is an improvement. The DID method looks at the effect of receiving the treatment, the first difference, and thereby compares it to a proxy, a control group, to see how the observations would have developed without receiving the treatment, the second difference. This leads to more adequate predictions of the approximate causal effect of a treatment on Y, the outcome variable.

The following statistical framework is based on the paper of Albouy (2004). To find the causal effect of the treatment on the dependent variable, a simple linear regression can be defined by the following 'main' equation:

$$Y_i = \alpha + \beta T_i + \gamma t_i + \delta_{DID}(T_i * t_i) + \varepsilon_i \quad (6)$$

Here, Y_i represents the outcome, α represents the constant term and ε_i represents the error term. In addition, T_i represents the treatment, which takes a value of 1 when receiving the treatment and 0 when not receiving the treatment and thus being in the control group, and t_i represents the time period, which takes a value of 1 after the treatment is being allocated and 0 before the treatment is allocated.

The effect of receiving the treatment, thus the first difference, is calculated by comparing the difference in outcome Y_i before and after receiving the treatment and thereby, only considers the treatment group. This is viewed as the treatment group specific effect, and is represented by the parameter β, which on its own can be defined by:

$$\beta = \bar{Y}_1^T - \bar{Y}_0^T \quad (7)$$

The difference between the treatment and control group in outcome Y_i post-treatment, thus the second difference, ignores the pre-treatment outcomes. It can be known as the time trend common to both the treatment and control group and is represented by the parameter γ, which on its own can be defined by:

$$\gamma = \bar{Y}_1^T - \bar{Y}_1^C \quad (8)$$

Lastly, the true effect of the treatment can be established. This is calculated as the difference in outcome in the treatment group before and after receiving the treatment, minus the difference in outcome in the control group before and after the treatment. This causal effect of the treatment is represented by the parameter δ_{DID}, which on its own can be defined by:

$$\delta_{DID} = \bar{Y}_1^T - \bar{Y}_0^T - (\bar{Y}_1^C - \bar{Y}_0^C) \quad (9)$$

To conclude, finding the parameter δ_{DID} which is the coefficient of the interaction effect in the first 'main' equation, allows to find the causal effect of a certain treatment on the outcome variable. This estimator should be unbiased, and the corresponding assumptions that should be met in order for this to be the case can be found in Table A3 in Appendix A.

5. Data

To find out whether there is evidence for behavioural pricing effects in secondary data, the causal effect of psychological pricing is estimated. As previously discussed, in order to find this causal effect, a matching algorithm is performed to pre-process the data. In addition, the secondary research question asks whether performing a clustering analysis on the raw data before matching the observations increases the matching results. Therefore, the matches are created in the two previously stated approaches: (i) Standard Matching, and (ii) Cluster-then-Matching. As indicated in the two approaches, they both start with the raw but cleaned data about the listings.

5.1 Data collection

For this research, the data of Airbnb listings in Rome is used. As the main research question indicates, the data used in this paper is secondary data. The data is collected by Inside Airbnb on the 14th of June 2020 (N = 18,416) and the 13th of April 2021 (N = 15,323), and is retrieved from www.insideairbnb.com, where Airbnb publishes scraped data of most cities almost every month. Since not all listings occurred in both datasets, the data is merged into one dataset only using the listings that included information for both time periods. The reason for choosing these two months is seasonality. Since these two months fall in approximately the same holiday season, but both do not fall in the month May in which the number of trips significantly increase, no change in occupancy rate due to seasonality has to be taken into account. Afterwards, all observations with missing values were removed.

In order to measure the effectiveness of the psychological pricing strategy, there are two important variables needed: *occupancy rate,* and whether there is psychological pricing applied, which eventually will be the *treatment*. Both variables are further explained below. Beyond these two variables, control variables (also known as the previously discussed covariates when it comes to matching) are needed as well in order to be able to perform the matching algorithm and clustering analysis on the listings. The listings' characteristics that are used as covariates are *latitude, longitude* and *room type*, and are discussed further during the descriptive statistics in Section 5.4.

5.2 Dependent variable: Occupancy rate

The occupancy rate is used as the measure of effectiveness, however, it is not included in the dataset and should be calculated manually for each observation. This is done by using Inside Airbnb's own provided Occupancy Model, also known as the San Francisco Model (Inside Airbnb, 2021a). The occupancy rate on a yearly basis can be calculated by using the following formula:

$$\text{Occupancy rate} = \text{Average length of stay} * \text{Estimated number of bookings}$$

The *average length of stay* is given by Airbnb per city and equals three nights for Rome (Inside Airbnb, 2021b). The estimated number of bookings has to be calculated using the formula:

$$\text{Estimated number of bookings} = \text{Reviews per month} * 12 * 2$$

The variable *reviews per month* is given in the dataset, however the ultimate occupancy rate should be on yearly basis. In addition, Inside Airbnb (2021a) states that a review rate of 50% can be applied. These two aspects result in the latter formula. One drawback to this calculation of the occupancy rate is that this review rate is a general assumption and therefore, the estimated number of bookings might not be accurate for all listings. This leads to the occupancy rate not being perfectly approximated. In the dataset, values of *occupancy rate* over 100% are found, which is impossible and therefore, these observations are removed. Even though an occupancy rate of (nearly) 100% seems unlikely, it cannot be assumed to be impossible and therefore, these observations are kept in the dataset.

5.3 Treatment: Psychological pricing

Here, the treatment is whether a listing has applied psychological pricing. This variable is based on the variable price per night, provided in the datasets. The difference in price between 2020 and 2021, and more specifically the most right digit for every price, is analysed. Ultimately, all observations that changed their price with €1 from 0-ending to 9-ending (e.g., €850 to €849) within the two time periods of June 2020 and April 2021, fall into the treatment group. All other listings that changed price with €1 – other than from 0-ending to 9-ending – fall into the control group, since in this manner the magnitude of the change in price for treatment and control group stays the same. This however leads to large decrease in the sample size of the dataset, which eventually consists of 253 observations, of which 85 in the treatment group and 168 in the control group (also shown in Table 3 below).

5.4 Control variables

Next to the *occupancy rate* and *treatment*, the dataset also consists of control variables. The control variables used are *latitude*, *longitude* and *room type*, which is the only categorical control variable. These variables seem to be most important and might influence how a listing is perceived. In order to measure the effectiveness of matching by applying the previously named balance diagnostics, all categorical variables should be converted to binary variables (Stuart, 2010). Therefore, the categorical variable *room type* is changed into four binary variables of all categories, taking value 1 if the room type equals that specific room type and 0 otherwise. The numeric variables *latitude* and *longitude* are geographic coordinate systems that together represent a specific position on earth. In order to perform a cluster analysis, the numeric variables need to be normalized so the algorithm will not depend on an arbitrary variable unit. Therefore, the numeric variables *latitude* and *longitude* are normalized and afterwards, both range from 0 to 1.

5.5 Descriptive statistics

The descriptive statistics of all variables are displayed in Table 3 below, however, the most important remarks are discussed here. As can be seen in Table 3, the ratio of treatment and control observations is not balanced, which can be of an issue when it affects the performance of the algorithm, e.g., when the goal is predicting or classifying. However, with clustering this does not pose a great problem since it gives descriptive results, i.e., it identifies and interprets structure within the data, and unbalanced data will therefore not bias the outcome (Brendel, 2021). In addition, since matching takes place with replacement and in the form of k-to-one, the unbalanced treatment variable is not a problem.

The mean occupancy rates for 2020 and 2021 are respectively 25.6% and 20%, and are roughly in line with Inside Airbnb’s generally estimated occupancy rate for Rome, which is 24.4% (Inside Airbnb, 2021b).

Lastly, the descriptive statistics show that none of the observations take a value of 1 for the binary variable *shared room* (which is created from the variable *room type*). This means that the room type shared room is not within this dataset and therefore, this variable will also not be used in the further analyses.

Table 3. Descriptive statistics.

	0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Treatment	168	85						
Occupancy rate 2020			0.004	0.055	0.148	0.256	0.395	0.944
Occupancy rate 2021			0.002	0.039	0.116	0.200	0.310	0.882
Room type								
Entire home	111	142						
Hotel room	247	6						
Private room	148	105						
Shared room	253	0						
Latitude			0.000	0.529	0.591	0.569	0.631	1.000
Longitude			0.000	0.416	0.459	0.472	0.530	1.000

N = 253

5.6 Model free evidence

Before starting the analyses, the visual display of the treatment effect in the current (unmatched) dataset is observed. This is displayed in Figure 1 by the means of boxplots.

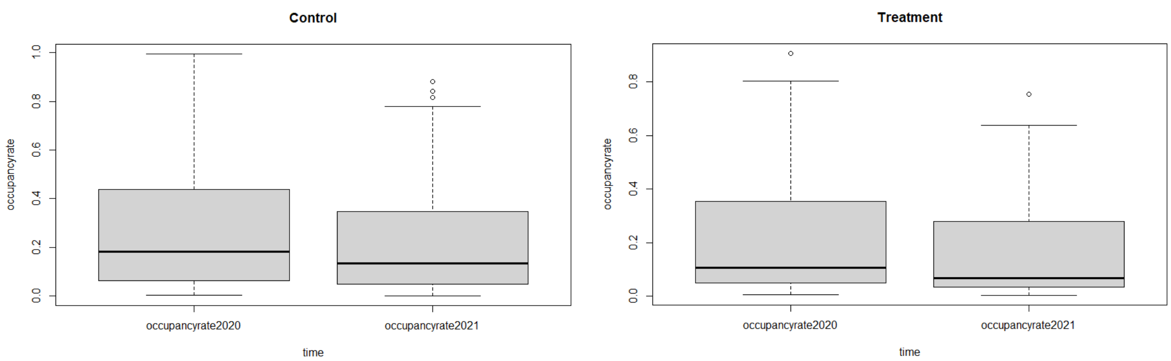

Figure 1. Boxplots showing the occupancy rates before and after psychological pricing took place for the control and treatment group.

The boxplots on the left show the true effect of time since no treatment has taken place during the observation of this group. This illustrates that the occupancy rate is already slightly lower in 2021 than in 2020, being a visualization of the time trend. The boxplot on the right shows the effect of receiving the treatment (also known as the first difference), without taking the time trend into account. This illustrates that for treatment observations, the occupancy rate is slightly lower in 2021 than in 2020 as well. When comparing it to the control group, these differences in occupancy rate over time do not differ much, visually speaking. This can already be a sign of a non-existing or rather small psychological pricing effect in the currently used dataset.

6. Research findings

In this section, the results of the analyses are discussed[3]. The first part of this section discusses the results of the two approaches used to match observations and offers an answer to the methodological research question: Does adding a cluster analysis before matching the observations improve the matching results (i.e. better covariate balance)? Afterwards, in the second part of this section, the best approach is used to test H1 and find answer to the substantive research question: Is there evidence for psychological pricing effects in secondary data (i.e., on the occupancy rate of Airbnb listings in the city of Rome)? In this thesis, a significance level of 90% is accepted since the sample size is small.

6.1 Data pre-processing: comparing Standard Matching with Cluster-then-Matching

Before performing any analyses, the unmatched data is analysed. The results of the SB and the two sample t-tests on the covariates of the unmatched data is displayed in Table B1 in Appendix B. This shows that the null hypothesis – which states: true difference in means of treatment and control group is equal to 0 – can be rejected for the variables *entire home*, *private room* and *longitude*, meaning that the means between treatment and control observations differ when it comes to these variables. This shows that both groups do differ from each other in some ways and matching could improve this.

As discussed before, this thesis uses the two approaches Standard Matching and Cluster-then-Matching to find if clustering before matching improves the matching results. After matching, the quality of the matches is evaluated by the two balance diagnostics standardized bias (SB) and a two sample t-test.

In both the Standard Matching and Cluster-then-Matching approach, no caliper is chosen. The reason for this is the fact that both approaches should be as similar as possible in order to get the best comparison, however, they deal with very different sample sizes since the second approach matches within subsections of the total dataset. Therefore, it would not be fair to use one specific caliper for both

[3] The analyses are performed in R version 4.0.2

approaches, nor to change the caliper per approach. In addition, allowing replacement also decreases the chance of bad matches, which is the goal of defining the caliper as well, since it leads to choosing the closest control observation for each treatment observation without limitations.

6.1.1 Standard Matching

Results of the Standard Matching approach are displayed in Table 4, and the descriptive statistics of the matched data using the Standard Matching approach can be found in Table C1 in Appendix C. Using this matching approach, all treatment observations were matched.

Table 4. Balance improvement measured by SB and two sample t-test using Standard Matching approach.

Covariates	**Balance improvement measured by SB (in %)**	**t-test**
entire_home0	35.5	1.704*
entire_home1	35.5	-1.704*
hotel_room0	100.0”	-0.439
hotel_room1	100.0”	0.439
private_room0	35.4	-1.566
private_room1	35.4	1.566
latitude	-286.1	0.572
longitude	33.2	2.890***

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.

As displayed in Table 4, the balance improvement measured by SB is positive in all but one case, namely *latitude*. For *hotel room*, which has an balance improvement of 100%, it means that there is total balance between the treatment and control observations. In addition, the balance for *entire home*, *private room* and *longitude* increased with approximately 35%. The balance for *latitude* decreased with 286.3% meaning that higher imbalance is created between the treatment and control observations when it comes to this covariate because of the Standard Matching. This is common when already well balanced covariates are matched, because creating an overall well balanced dataset can be at the expense of already well balanced covariates (Ford, 2018). To conclude, by solely looking at the balance improvement measured by SB, it can be concluded that overall this matching approach did increase covariate balance at the expense of one covariate.

The performed t-tests show that the null hypothesis – which states: true difference in means of treatment and control group is equal to 0 – can be rejected for the covariates *entire home* and *longitude*. This means that the means between treatment and control observations still differ when it comes to these covariates even after matching is performed. The t-test for *private room* has become insignificant, meaning that it cannot be concluded anymore that the mean of treatment and control observations differ when it comes to this covariate, which is a sign of improved covariate balance. It can be concluded that Standard Matching did improve covariate balance in comparison to the unmatched dataset, however, there is still a significant imbalance in the covariates *entire home* and *longitude*.

6.1.2 Cluster-then-Matching

As discussed in Section 4.1.2, the optimal number of clusters has to be decided upon manually during the analysis. The dendrogram displayed as Figure D1 in Appendix D visualizes how the clusters are created and includes visualization of the possible number of clusters by the horizontal blue lines crossing the branches in the dendrogram. As can be seen, 2, 3, 4 or 5 clusters can be used to perform further analysis. The reason that 5 is the maximum amount of clusters tried in this thesis is because the lower in the dendrogram, the more small clusters will be created. This is disadvantageous since it could be that certain clusters do not contain treatment observations or enough control observations, because the sample size is too small.

Table 5. Total balance improvement measured by SB and two sample t-test of total matching results per clustering possibility.

	2 clusters (N = 140)		3 clusters (N = 134)		4 clusters (N = 141)		5 clusters (N = 138)	
Covariates	**Balance improvement measured by SB (in %)**	**t-test**	**Balance improvement measured by SB (in %)**	**t-test**	**Balance improvement measured by SB (in %)**	**t-test**	**Balance improvement measured by SB (in %)**	**t-test**
entire_home0	0.0”	0.555	0.0”	0.151	0.0”	-0.325	0.0”	-0.843
entire_home1	0.0”	-0.555	0.0”	-0.151	0.0”	0.325	0.0”	0.843
hotel_room0	520.1	0.218	0.0”	-	0.0”	-	0.0”	-
hotel_room1	520.1	-0.218	0.0”	-	0.0”	-	0.0”	-
private_room0	520.1	-0.629	0.0”	-0.151	0.0”	0.325	0.0”	0.843
private_room1	520.1	0.629	0.0”	0.151	0.0”	-0.325	0.0”	-0.843
latitude	-90.0	0.796	-92.1	0.899	136.4	0.686	-122.8	0.788
longitude	13.1	2.448**	18.9	2.733***	-11.4	3.312***	-22.0	3.784***

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset. [4]
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

This approach first clusters the data and then performs matching within each cluster. This leads to matching results per cluster. However, to be able to compare the matching results of the different cluster divisions, the total balance improvement of the entire dataset (all clusters combined) is needed, which is calculated manually for each covariate[5]. This approach will be performed four times to see which division of clusters gives the best matching results. Table 5 shows the total balance improvement in the final dataset using SB and two sample t-test of the matching results per cluster division. As can be seen, the balance improvements per cluster division measured by SB for the different room types are often 0.0, meaning that matching did not lead to balance improvement. However, this is because these variables already became totally balanced for each cluster division during the clustering process (as explained in Footnote 4). This is also displayed in the descriptive statistics per cluster in Tables D1, D3, D5 and D7 in Appendix D.

From Table 5 the optimal number of clusters can be chosen, which is 4 clusters since this method has the highest overall balance improvement. Thereby, there is no difference in t-test results when it comes to all possible cluster divisions, since the covariate *longitude* proves to be significant for all options. Table D1 in Appendix D shows the descriptive statistics per cluster using 4 clusters. Tables D2 to D7 in Appendix D show both the matching results per cluster and the descriptive statistics for the other divisions of clusters separately (i.e., 2, 3 and 5 clusters).

Table 6 below shows the balance improvement measure by SB and two sample t-test per cluster using 4 clusters. The first noticeable thing is that matching could not be performed on cluster 3. This is because the sample size of this cluster was too small (see Table D1 in Appendix D). Thus, using this approach led to a total of six observation being discarded, including two unmatched treatment observations. Furthermore, as explained before, the balance improvements per cluster division measured by SB for the different room types are 0.0, however, these covariates are already totally balanced due to clustering, since in each cluster there is only one certain room type. Therefore, it can be assumed that the

[4] The balance improvement is measured using the SB before and after matching per cluster, leading to a given balance improvement based solely on the matching performance. The notation 0.0 means that matching did not improve the balance. However, adding ” means that total balance is achieved compared to unmatched data, and more precisely this is achieved during the clustering process. Thus, the notation 0.0” means that clustering solely led to a totally balanced covariate. This way of visualization shows where in the process the balancing took place.
[5] The formula used for this calculation is: $\left(\sum_{n=1}^{C} SB_{n,\ matched\ data} - \sum_{n=1}^{C} SB_{n,\ all\ data} \right) / \sum_{n=1}^{C} SB_{n,\ all\ data}$
Where C is the total number of clusters in that division, and SB is the SB per covariate in a certain cluster. This formula is the standard formula for calculating a percental change, however, adjusted for the fact that the SB’s of each cluster are aggregated.

combination of clustering and matching improved the balance in the room types to total balance, while the matching results solely (displayed in Table 4) proved not able to do this.

When looking at the covariate *longitude,* it can be seen that in cluster 1 and 2 there is a significance t-value, meaning that within these clusters it was not possible to balance this covariate well enough. In cluster 4 it cannot be said that the null hypothesis – which states: true difference in means of treatment and control group is equal to 0 for *longitude* – can be rejected, which is positive. However, the balance improvement for this covariate in this cluster shows that the balance deteriorates.

Table 6. Balance improvement measured by SB and two sample t-test on clustered data using 4 clusters.

	Covariates	Balance improvement measured by SB (in %)	t-test
Cluster 1	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	-181.0	-0.040
	longitude	-71.4	2.624***
Cluster 2	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	-34.3	0.998
	longitude	38.3	1.873*
Cluster 3	entire_home0	-	-
	entire_home1	-	-
	hotel_room0	-	-
	hotel_room1	-	-
	private_room0	-	-
	private_room1	-	-
	latitude	-	-
	longitude	-	-

Cluster 4	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	19.1	-0.929
	longitude	-32.8	1.036

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

Table 7 below represents the total balance improvement measured by SB and two sample t-test of the total matching results using the final dataset created with the division of 4 clusters (which is also part of Table 5). It shows that the room types are totally balanced, and for *entire home* and *private room* the t-test is insignificant. This means that the null hypothesis – which states: true difference in means of treatment and control group is equal to 0 – for these covariates cannot be rejected, which is positive, especially when comparing it to the unmatched data where both t-values were significant. When looking at the t-test results of *hotel room*, it can be seen that it was impossible to run. This is due to the fact that all observations in cluster 3 were hotel rooms (see Table D1 in Appendix D) and this cluster was impossible to match.

The balance in the covariate *latitude* increases, and the t-test shows that the null hypothesis still cannot be rejected. The balance in the covariate *longitude* deteriorates and in addition, the null hypothesis can still be rejected here. This means that when it comes to the longitude of the observations there is a significant difference between treatment and control observations. This is the same as with the Standard Matching approach.

Table 7. Total balance improvement measured by SB and two sample t-test of total matching results of the final dataset of 4 clusters.

Covariates	Balance improvement measured by SB (in %)	t-test
entire_home0	0.0”	-0.325
entire_home1	0.0”	0.325
hotel_room0	0.0”	-
hotel_room1	0.0”	-

private_room0	0.0”	0.325
private_room1	0.0”	-0.325
latitude	136.4	0.686
longitude	-11.4	3.312***

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

6.1.3 Choosing ‘winning’ approach

Table 8 is used to compare the matching results both between the two matching approaches used in this thesis, as well as with the unmatched data. From this a conclusion about the optimal data pre-processing approach can be drawn.

Table 8. Comparing balance diagnostics for unmatched data, matched data using Standard Matching approach and matched data using the optimal Cluster-then-Matching approach (i.e., 4 clusters), respectively.

	Unmatched data		Standard Matching		Cluster-then-Matching	
Covariates	**SB**	**t-test**	**Balance improvement measured by SB (in %)**	**t-test**	**Balance improvement measured by SB (in %)**	**t-test**
entire_home0	-0.3917	2.856***	35.5	1.704*	0.0”	-0.325
entire_home1	0.3917	-2.856***	35.5	-1.704*	0.0”	0.325
hotel_room0	0.0018	-0.014	100.0”	-0.439	0.0”	-
hotel_room1	-0.0018	0.014	100.0”	0.439	0.0”	-
private_room0	0.3996	-2.891***	35.4	-1.566	0.0”	0.325
private_room1	-0.3996	2.891***	35.4	1.566	0.0”	-0.325
latitude	-0.0553	0.471	-286.1	0.572	136.4	0.686
longitude	-0.5410	4.257***	33.2	2.890***	-11.4	3.312***

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

First, the balance improvement measured by SB is analysed. As can be seen in Table 8, the Cluster-then-Matching approach has the best overall balance improvement. First, this approach led to totally balanced covariates when it comes to all room types. Second, the balance improvement of *latitude* is higher with this approach compared to Standard Matching. When it comes to *longitude*, only the Standard Matching approach resulted in a balance improvement.

When looking at the t-values of the unmatched data, it can be seen that the null hypothesis – which states: true difference in means of treatment and control group is equal to 0 – can be rejected for the room types *entire home* and *private room*, and for *longitude*. When looking at the Standard Matching approach, it can be seen that the null hypothesis can be rejected for the room type *entire home*, and *longitude*. This means that the Standard Matching approach improved the balance in the covariate *private room*. Lastly, the Cluster-then-Matching approach shows that the null hypothesis can only be rejected for *longitude*, meaning that this approach improved the balance for both room types *entire home* and *private room* compared to the unmatched data.

To conclude, when analysing both balance diagnostics, it can be said that the Cluster-then-Matching approach is the best performing approach and therefore, will be used in the next analysis to find if there is evidence for psychological pricing effects in secondary data.

6.1.4 Summary of results

As proved in the section above, the "winning" approach is the Cluster-then-Matching approach with a cluster division of 4 clusters, and the matching results created with this approach are used to perform the further DID analysis on. The overall advantage of clustering before matching in this thesis proves to be the balance in room type, which are the binary variables. Often, observations within clusters all have the same room type, meaning that this covariate is already totally balanced without having to perform propensity score matching. Therefore, the matching algorithm can focus more on improving balance in the other covariates and thus, finding more complete matches. In addition, when comparing this to the improvement of covariate balance of the room types with the Standard Matching approach, it can be seen that not all room types are balanced after matching. These results confirm the findings of King and Nielsen (2019) stating that performing matching more locally increases robustness.

6.2 Psychological pricing effect: Difference in Differences

The matching results retrieved from the Cluster-then-Matching approach are used to find if there is evidence for psychological pricing effect in secondary data.

6.2.1 Interpreting DID analysis

In Table 9 below, the results of the DID analysis, obtained by performing a linear regression – noted as formula (6) in Section 4.3 – on the outcome variable *occupancy rate*, are presented. The true effect of the treatment can be found by looking at the coefficient of the interaction term. The DID estimator is insignificant, leading to the impossibility of drawing a conclusion about an existing psychological pricing effect in the secondary data used in this thesis. However, this does not necessarily imply that the psychological pricing effect is an artifact of the methods used in prior research – proven to be mostly lab experiments and surveys – which cannot be replicated with secondary data. There could be other reasons, influenced by this thesis as well, that lead to insignificant results. These are partly explained by interpreting the other significant coefficients, and the rest is discussed in the Limitations.

Table 9. Effects of treatment on occupancy rate compared to non-treated.

	Coefficient (Std. Error)
Intercept	0.309***
	(0.030)
Treatment	-0.077**
	(0.038)
Time	-0.074*
	(0.042)
Treatment * Time	0.017
	(0.054)

Significance: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

R^2: 0.042 [6]

As explained, the coefficient of *Treatment* represents the first difference, also known as the treatment group specific effect. This coefficient is significant, meaning that an effect of receiving the treatment can be found on the treatment observations, however, this effect is negative. It can be concluded that the listings that were allocated to the treatment group have a 7.7% lower occupancy rate after applying the psychological pricing strategy compared to before applying it, ceteris paribus. When comparing it to results from previous literature, one possible reason for the decrease in occupancy rate due to psychological pricing could be that the price perception influences the quality image (Schindler and Kibarian, 2001; Schindler, 2006; Kreul, 1982). Thus, it could be that the use of psychological prices has a negative effect on the quality image, leading to less bookings. However, the biggest part of this effect

[6] The low R^2 does not prove to be a problem in this case, since the regression is only used to test a relationship (not for predicting), suggesting that it is not important in this context. This low value is caused by the fact that the treatment has little or no effect on the occupancy rate, which is normal given that many other factors determine a property's occupancy rate. Therefore, a general manner to increase the R^2 is by adding control variables. However, given the matching process, those controls are not needed since the treatment effect is supposedly unbiased. In addition, only a small set of variables that are of importance when it comes to a property's occupancy rate are given in the dataset.

can be explained by taking into account the expectation of this estimator[7], which seems more logical. This shows that if a time trend exists ($\gamma \neq 0$), it is confound to being part of the treatment effect, meaning that the coefficient of the *Treatment* variable is biased (Albouy, 2004). Therefore, the explanation of the significant time trend explained below, is also partly the reason for having a significant effect in the *Treatment* as well.

The coefficient of *Time* represents the second difference, also known as the time trend. This coefficient is significant as well, meaning that time also affects the occupancy rate in this analysis. From the results it can be concluded that the listings have a decrease of 7.4% in occupancy rate in April 2021 compared to June 2020, ceteris paribus. One reason for this could be the COVID-19 pandemic, which negatively affected travelling and therefore, negatively affected the occupancy of Airbnb listings (and other stays such as hotels etc.) as well. To conclude, this insignificancy of the psychological pricing effect is most likely due to a time trend (probably the COVID-19 crisis), which proves to have a more powerful effect on the occupancy rate.

6.2.2 Robustness tests

To assess whether the results are independent of the methodology, two robustness tests are performed. First, the analysis is performed using unmatched data to compare the results to a baseline. Second, the analysis is performed using the Cluster-then-Matching approach, however, without allowing replacement while matching (one of the options discussed in Section 4.1.1), since this leads to different matches. The results are displayed in Table E1 and E2 in Appendix E, respectively.

Performing the DID analysis on unmatched data shows a significant coefficient for *Time* only, meaning that there already is a time trend present. When comparing the rest of the results to the results of the DID analysis on the data pre-processed using the Cluster-then-Matching approach, it can be seen that a combination of matching and clustering improved the significance of the *Treatment* variable, meaning that it led to a treatment group specific effect which was not existing in the unmatched data. This shows that a combination of matching and clustering does bring forward a significant difference before and after applying psychological pricing within the treatment group, which is an improvement of the current analysis.

Performing the DID analysis on the data pre-processed using the Cluster-then-Matching approach without replacement, shows a significant effect for both *Treatment* and *Time*. This matches the results from the initially used Cluster-then-Matching approach, since with both approaches the two coefficients

[7] $E[\beta] = E[\bar{Y}_1^T] - E[\bar{Y}_0^T]$
$= [\alpha + \beta + \gamma + \delta] - [\alpha + \beta]$
$= \gamma + \delta$
(Albouy, 2004)

show a negative effect on occupancy rate. The magnitudes of the coefficients however do slightly differ. The DID estimator is still insignificant, meaning that again no conclusion can be drawn about evidence of psychological pricing effects using secondary data in this thesis. This method wholly verifies the results found with the optimal data pre-processing approach, and thereby, shows that in the case of this thesis, matching with or without replacement does not give significantly different results.

7. Discussion

Writing this thesis led to the possibility of combining two very interesting but very different pillars in the economic field, namely behavioural economics and mathematical economics, which is hardly done. This led to interesting useful insights for future research as well.

7.1 Conclusion

To the best of the author's knowledge, no papers exist (yet) about studying psychological pricing effects using secondary data. While analysing literature about psychological pricing, a pattern emerged existing of the use of primary data in such studies. This is logical, since the most important advantage of experimental research is demonstrating causality. However, different disadvantages of primary data do exist as well and therefore, this thesis uses secondary data to find evidence for the causal effect of psychological pricing as well, leading to the following main research question:

"Is there evidence for psychological pricing effects in secondary data?"

To investigate this effect by the means of secondary data, data pre-processing steps need to be optimized. Matching algorithms are critical in causal analyses of secondary data, since it offers a solution to understand causal effects when randomly allocating the treatment is either not ethical or not possible (anymore). The two approaches used in this thesis are matching and a combination of matching and clustering. Evaluating this helps us understand the validity of the answer to the main substantive research question of this thesis. Thus, before being able to answer the main research question regarding whether there is evidence for psychological pricing effects in secondary data, a secondary research question needs to be answered first, which is the following:

"Does adding a cluster analysis before matching the observations improve the matching results?"

This thesis contributes to existing literature in several ways. First, it proves whether or not there is evidence for psychological pricing effects in secondary data, which is still unknown to this day. Second, it evaluates several ways to pre-process the data, including adding clustering as an extra data pre-processing step before the well-known matching approach used with secondary data. This is where the mathematical economics really crosses the behavioural economics field. Lastly, the potential evidence for psychological pricing found in the Airbnb sector may form implications for listing owners to understand the importance of pricing and possibly manage their pricing strategy in a way that proves to be more successful.

7.1.1 Methodological implications for causal inferences

This section discusses the results regarding the optimization of the methodology. Existing literature shows that PSM often increases imbalance, model dependence and bias. In addition, it is also proven that when adjusting the data pre-processing in such a way that PSM matches more locally, robustness increases. These findings in combination with the intention of combining clustering with matching, led to the following hypothesis as answer to the secondary research question:

Hypothesis 2: Adding a cluster analysis as extra data pre-processing step before matching will improve the matching results (i.e., better covariate balance).

The findings in this thesis prove that the Cluster-then-Matching approach does improve the covariate balance more than the Standard Matching approach does. This is confirmed by evaluating the standard balance diagnostics used when matching. The overall advantage of clustering before matching in this thesis proves to be the balance in room type. Clustering on its own already leads to total balance in all room types, which was impossible with matching solely. To generalise this finding, it could possibly be said that clustering before matching improves the covariate balance of binary variables, since clustering finds groups of observations that are more similar to each other than to observations in other clusters. When having a dataset consisting of mainly binary variables, it shows that the groups created are often groups that take the same value of one particular binary variable (either 0 or 1), leading to increased balance within the clusters. Afterwards, the matching algorithm can focus more on improving balance in the other covariates and thus, finding more complete matches. Therefore, looking at the key findings of this thesis, it can be concluded that this hypothesis can be accepted. Clustering before matching improves matching results, i.e., covariate balance.

7.1.2 Implications for pricing researchers

This section discusses the results regarding the causal effect of psychological pricing. Since previous literature using primary data substantively shows that there is evidence for psychological pricing effects, the hypothesis states:

Hypothesis 1: Using secondary data, evidence for psychological pricing effects on occupancy rate of Airbnb listings can be found.

The findings in this thesis prove to be insignificant, leading to the impossibility of drawing conclusions about a psychological pricing effect using secondary data. Since this hypothesis does not correspond to the findings, it is rejected for this thesis. However, it cannot be stated with certainty that these studies cannot be replicated with secondary data, since there could be other reasons for insignificancy. One of them can be retrieved from analysing the regression, since the effect of both the treatment on the treated and the time trend separately are significant. The reason for finding a significant treatment effect on the treated could mainly be because this estimator is biased, since an existing time trend is confound to being part of the treatment effect. The reason for finding a significant time trend could be the COVID-19 pandemic taking place in the timeframe in which the data is collected. To conclude, this insignificancy of the psychological pricing effect is most likely due to a time trend (probably the COVID-19 crisis), which proves to have a more powerful effect on the occupancy rate. Therefore, no managerial implications can be given.

7.2 Limitations and implications for future research

Next to the timeframe, which included the COVID-19 pandemic, this thesis has other limitations that come with implications for future research, and are discussed in this section. The first limitation of the analysis in this thesis is the sample size, which can be seen as a power issue. Because of the specific requirement of changing prices with a €1 difference (to keep the same magnitude), the initially big dataset changed into a small subset of observations. The final dataset consisted of 85 treatment observations and 168 control observations. This did not allow for an extensive matching process, especially not when using Random Forest to estimate propensity scores, which works well with huge datasets. The sample size proved to be a limitation, since in the optimal matching approach using the optimal division of clusters, it was impossible to match observations in one particular cluster. This led to the elimination of the observations from that particular cluster in the total dataset. Therefore, for future research it is suggested to perform the same analysis using a larger sample size.

The second limitation could be the type of sector used for this analysis. A reason for insignificancy can be the fact that psychological pricing effects might show up in a much less obvious way, or even not at all, because of the use of data from the luxury sector. As discussed in the Literature Review, psychological pricing could affect the quality image, leading to a decrease in bookings of a particular room, in this case an Airbnb listing. Therefore, for future research it is suggested to either analyse both low-quality and high-quality stays to have the possibility to compare the results or even find a confounding effect with quality perception. Another more general suggestion is to perform this analysis using secondary data from supermarkets or newspapers, since this type of data is used standardly when studying psychological pricing effects. This might make it more easy to compare psychological pricing effects in secondary data with primary data, since the environment stays the same and previous literature already exists. Taking the results of this thesis into account as well, it could possibly be proven that psychological pricing might not be universal.

The last limitation is geography. This thesis used data from Rome, leading to the limitation that this effect is only studied for Rome listings. When considering only one geographical place, it could be that the results are biased. Reasons for this are: (i) the type of people booking stays in Rome can differ from the type of people booking stays in other cities; (ii) only a certain type of Airbnb owners use psychological pricing, leading to a selection of specific (potentially different) listings compared to the listings not engaging in psychological pricing; or (iii) other non-monetary reasons might play a role in booking (or not booking) a stay in Rome, for example touristic attractions, political or social turmoil, culture, or climate and nature. Therefore, for future research it is suggested to take multiple different cities, possibly even in different countries or continents, into account when analysing psychological pricing effects.

To summarize, it cannot be concluded with certainty that there is no evidence for psychological pricing effects in all secondary data. This thesis can be viewed as a prereview and further research is encouraged to really exclude evidence for psychological pricing effects in secondary data. In addition, this also holds for the data pre-processing findings. It proves that clustering before matching improves the final matching results, however, since the dataset used is not optimal, it could be the case that the results depend on the data. For this, further research on the influence of clustering on the matching results is encouraged as well.

References

Albouy, D. (2004). Program evaluation and the difference in difference estimator. *Economics, 131*, 1-4.

Anand, D. (June 17, 2020). *Gower's distance*. Medium. Retrieved from: https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553

Baumgartner, H., & Steenkamp, J.B.E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156.

Bizer, G.Y., & Schindler, R.M. (October 2005). Direct evidence of ending-digit drop-off in price information processing. *Psychology & Marketing, 22*(10), 771–783.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brendel, C. (May 9, 2021). *Clustering analyses with highly imbalanced datasets*. Retrieved from: https://towardsdatascience.com/clustering-analyses-with-highly-imbalanced-datasets-27e486cd82a4

Brenner, G.A., & Brenner, R. (1982). Memory and markets, or why are you paying $2.99 for a widget? *Journal of Business, 55*(1), 147-158.

Brownlee, J. (March 14, 2016). Parametric and nonparametric machine learning algorithms. *Machine Learning Mastery*. Retrieved from: https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/

Chambliss, D.F. (2006). *Causation and experimental design*. Retrieved from: https://www.sagepub.com/sites/default/files/upm-binaries/23639_Chapter_5___Causation_and_Experimental_Design.pdf

Chapman, C., & Feit, E. M. (2015). *R for marketing research and analytics*. New York, NY: Springer.

Cheng, L.L., & Monroe, K.B. (22 March 2013). An appraisal of behavioral price research (part 1): price as a physical stimulus. *Academy of Marketing Science, 3*, 103-129.

Choi, J., Li, Y.J., Rangan, P., Chatterjee, P., & Singh, S.N. (2014). The odd-ending price justification effect: the influence of price-endings on hedonic and utilitarian consumption. *Journal of the Academy of Marketing Science, 42*(5), 545-557.

Collins, M., & Parsa, H.G. (2006). Pricing strategies to maximize revenues in the lodging industry. *Hospitality Management, 25*, 91-107.

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics, 19*(1), 1-14.

Econometrics Academy (2013). *Propensity score matching*. Retrieved from: https://sites.google.com/site/econometricsacademy/econometrics-models/propensity-score-matching

Fitzpatrick, M.C., Preisser, E.L., Ellison, A.M., & Elkinton, J.S. (2009). Observer bias and the detection of low-density populations. *Ecological applications, 19*(7), 1673-1679.

Ford, C. (April 24, 2018). *Getting started with matching methods*. Retrieved from: https://data.library.virginia.edu/getting-started-with-matching-methods/

Galizzi, M.M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: a systematic lab-field study. *Management Science, 65*(3), 976-1002.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Gúeguen, N., & Jacobs, C. (2005). Nine-ending price and consumer behavior: an evaluation in a new context. *Journal of Applied Sciences 5*(2), 383-384.

Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology, 67*(9), 2047-2059.

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2019). *Introduction to econometrics with R*. Essen: University of Duisburg-Essen.

Harrison, G.W., & List, J.A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009-1055.

Holdershaw, J., Gendall, P., & Garland, R. (1997). The widespread use of odd pricing in the retail sector. *Marketing Bulletin, 8*, 53-58.

Holloway, R.J. (1973). Odd-even price endings (Book Review). *Journal of Retailing, 49*(1), 77-78.
Homburg, C., & Koschate, N. (2005). Behavioral pricing-forschung im überblick: teil 1: grundlagen, preisinformationsaufnahme und preisinformationsbeurteilung. *Zeitschrift für Betriebswirtschaft, 75*(4), 383-423
Inside Airbnb. (2021a). *The occupancy model*. Retrieved from: http://insideairbnb.com/about.html
Inside Airbnb. (2021b). *Airbnb in Rome*. Retrieved from http://insideairbnb.com/rome/#
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer: New York.
Jeong, J.Y., & Crompton, J.L. (2017). The use of odd-ending numbers in the pricing of five tourism services in three different cultures. *Tourism Management, 62*, 135-146.
Kim, J., Cui, Y., Choi, C., Lee, S.J., & Marshall, R. (2020). The influence of preciseness of price information on the travel option choice. *Tourism Management, 79*, 104012.
King, G., Keohane, R.O., & Verba, S. (1994). *Designing social inquiry: scientific inference in qualitative research*. Princeton university press.
King, G., & Nielsen, R.A. (2019). Why propensity scores should not be used for matching. *Political Analysis, 27*(4), 435-454.
Kleinsasser, S., & Wagner, U. (2011). Price endings and tourism consumers' price perceptions. *Journal of Retailing and Consumer Services, 18*, 58-63.
Kreul, L.M. (1982). Magic numbers: Psychological aspects of menu pricing. *Cornell Hotel and Restaurant Administration Quarterly, 23*(2), 70-75.
Kwartler, T. (2017). *Text mining in practice with R* (First Edition). John Wiley & Sons Ltd.
Landsburg, S.E. (2012). *The armchair economist: economics & everyday life* (Rev. ed.). New York: Free Press.
Lee, B.K., Lessler, J., & Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346.
Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153-174.
Li, M. (2013). Using the propensity score method to estimate causal effects: A review and practical guide. *Organizational Research Methods, 16*(2), 188-226.
Li, F., Zaslavsky, A.M., & Landrum, M.B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine, 32*(19), 3373-3387.
Lieder, F., Griffiths, T.L., Huys, Q.J., & Goodman, N.D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review, 25*(1), 322-349.
Lopez, M.J., & Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science, 32*(2), 432-454.
Luchs, M. G., Naylor, R. W., Rose, R. L., Catlin, J. R., Gau, R., Kapitan, S., & Simpson, B. (2011). Toward a sustainable marketplace: Expanding options and benefits for consumers. *Journal of Research for Consumers, 19*, 1-12
Macé, S. (2012). The impact and determinants of nine-ending pricing in grocery retailing. *Journal of Retailing, 88*(1), 115-130.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2021). *Cluster: cluster analysis basics and extensions*. R package version 2.1.2.
Malhotra, N.K., & Nunan, D. (2019). *Marketing research, an applied orientation*. Pearson Education Limited.
Manning, K.C., & Sprott, D.E. (August, 2009). Price endings, left-digit effects, and choice. *Journal of Consumer Research, 36*(2), 328-335.
Monroe, K.B. (1990). *Pricing*. McGraw-Hill, New York.
Moore, D.S., McCabe, G.P., Alwan, L.C., Craig, B.A., & Duckworth, W.M. (2016). *The practice of statistics for business and economics* (Fourth Edition). WH Freeman.
Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071-22080.
Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(1), 86-97.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification, 31*(3), 274-295.
Nützenadel, C., & Hartmann, M. (2021). From classical pricing theory to behavioural pricing. Retrieved from: http://www.bankerstalk.ch/_Resources/Persistent/d0dd27b4aa073449c4846171048fb30b1933b03f/Artikel_Behavioural-Pricing_EN_low.pdf
Pearl, J. (2003). Statistics and causal inference: A Review. *Test*, *12*(2), 101-165.
Pearl, J. (2008). Causal inference. *JMLR Workshop and Conference Proceedings, 6*, 39-58.
Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, *40*(1), 75-149.
Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making, 5*(5), 411-419.
Rajendran, P. (March 28, 2019). *Causal inference using difference in differences, causal impact, and synthetic control*. Retrieved from: https://towardsdatascience.com/causal-inference-using-difference-in-differences-causal-impact-and-synthetic-control-f8639c408268
Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research, 45*(3), 261-279.
Rosenbaum, P.R. (1982). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association, 79*(387), 565-574.
Rosenbaum, P.R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*(385), 41-48.
Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.
Rosenbaum, P.R., & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.
Rubin, D.B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159-183.
Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.
Schindler, R.M. (2006). The 99 price ending as a signal of a low-price appeal. *Journal of Retailing, 82*(1), 71-77.
Schindler, R.M., & Kibarian, T.M. (1996). Increased consumer sales response through use of 99-ending prices. *Journal of Retailing*, *72*(2), 187-199.
Schindler, R.M., & Kibarian, T.M. (2001). Image communicated by the use of 99 endings in advertised prices. *Journal of Advertising*, *30*(4), 95-99.
Schindler, R.M, & Kirby, P.N. (1997). Patterns of rightmost digits used in advertised prices: Implications for nine-ending effects. *Journal of Consumer Research, 24*, 192-201.
Schindler, R.M., & Wiman, A.R. (1989). Effects of odd pricing on price recall. *Journal of Business Research, 19*(3), 165-177.
Scotina, A.D., & Gutman, R. (2019). Matching algorithms for causal inference with multiple treatments. *Statistics in Medicine*, *38*(17), 3139-3167.
Setoguchi, S., Schneeweiss, S., Brookhart, M.A., Glynn, R.J., & Cook, E.F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546-555.
Sizemore, S. & Alkurdi, R. (August 18, 2019). *Matching methods for causal inference: a machine learning update*. Humboldt-Universität zu Berlin. Retrieved from: https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/
Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1.
Thomas, M., & Morwitz, V. (2005). Penny wise and pound foolish: the left-digit effect in price cognition. *Journal of Consumer Research*, *32*(1), 54-64.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.
Zervas, G., Proserpio, D., & Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of Marketing Research*, *54*(5), 687-705.
Zou, S., & Petrick, J.F. (2020). Left-digit effect in tourists' price evaluations: The moderating role of price level and composite price. *Journal of Travel Research, 0047287520957418.*

Appendix A. Methodology

This appendix supports the methodology by a visualization of the PSM process and the assumptions for PSM and DID.

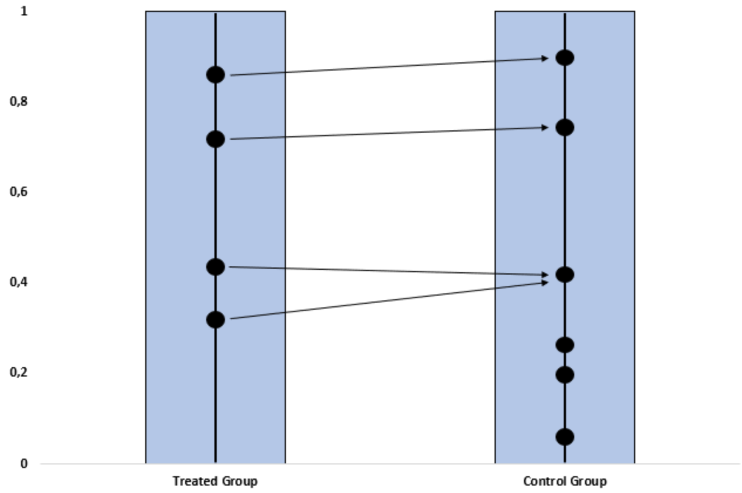

Figure A1. Visualization of the nearest neighbor matching process with the discussed model specifications (inspired by Econometrics Academy, 2013).

Table A1. General assumptions for constructing a counterfactual from naturally occurring data.

Assumption	Explanation
Treatment and control group must be the same in absence of the treatment, on average	Both groups should be comparable in terms of (un)observable characteristics.
Treatment and control group should react to the treatment in the same way	When the first assumption holds, this one should follow automatically.
Treatment and control group cannot be exposed in isolation to a third factor	Both groups should be treated in the same way or should have experienced the same thing except for the allocated treatment.

Table A2. Additional assumptions propensity score matching.

Assumption	**Explanation**
Conditional independence	The outcomes are independent of treatment, conditional on X. $y_0, y_1 \perp D \mid X$
Unconfoundedness	Conditional independence of the control group outcome and treatment. $y_0 \perp D \mid X$
Matching or overlap	For each value of X, there are both treated and control observations. $0 < prob(D = 1 \mid X) < 1$
Balancing condition	Assignment to treatment is independent of the X characteristic, given the same propensity score. $D \perp X \mid p(x)$
Equality	Unobserved characteristics are equal for treated and untreated.

Source: Econometrics Academy (2013)

Table A3. Additional assumptions difference in differences.

Assumption	**Explanation**
Model specification	The model in equation (outcome Y_i) is correctly specified. For example, the additive structure imposed is correct.
Average error term equals 0	The error term is on average zero. $E[\varepsilon i]=0$
Parallel-trend	The observed and unobserved characteristics should remain constant over time. Mathematically, this means that the error term is uncorrelated with the other variables in the equation: $cov(\varepsilon_i, T_i) = 0$ $cov(\varepsilon_i, t_i) = 0$ $cov(\varepsilon_i, T_i \cdot t_i) = 0$

Source: Albouy (2004)

Appendix B. Unmatched data

This appendix supports Section 6.1 of the results by displaying the SB and the results of the two sample t-test on the unmatched data. This is used as a baseline for further comparison.

Table B1. SB and two sample t-test on unmatched data.

Covariates	SB	t-test
entire_home0	-0.3917	2.856***
entire_home1	0.3917	-2.856***
hotel_room0	0.0018	-0.014
hotel_room1	-0.0018	0.014
private_room0	0.3996	-2.891***
private_room1	-0.3996	2.891***
latitude	-0.0553	0.471
longitude	-0.5410	4.257***

Significance: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Appendix C. Standard Matching

This appendix supports Section 6.1.1 of the results by displaying the descriptive statistics for the data matched by the Standard Matching approach.

Table C1. Descriptive statistics for matched data using Standard Matching approach.

	0	**1**	**Min**	**1st quantile**	**Median**	**Mean**	**3rd quantile**	**Max**
Treatment	54	85						
Occupancy rate 2020			0.006	0.055	0.112	0.232	0.371	0.905
Occupancy rate 2021			0.004	0.037	0.099	0.182	0.283	0.882
Room type								
Entire home	52	87						
Hotel room	135	4						
Private room	91	48						
Latitude			0.000	0.546	0.604	0.569	0.631	1.000
Longitude			0.000	0.403	0.450	0.447	0.527	0.892

N = 139

Appendix D. Cluster-then-Matching

This appendix supports Section 6.1.2 of the results by displaying the dendrogram used to find the possible number of clusters, and the descriptive statistics and balance improvement measures on the clustered and matched data using the number of clusters that were not optimal. This allows for matching results comparison between the different cluster options.

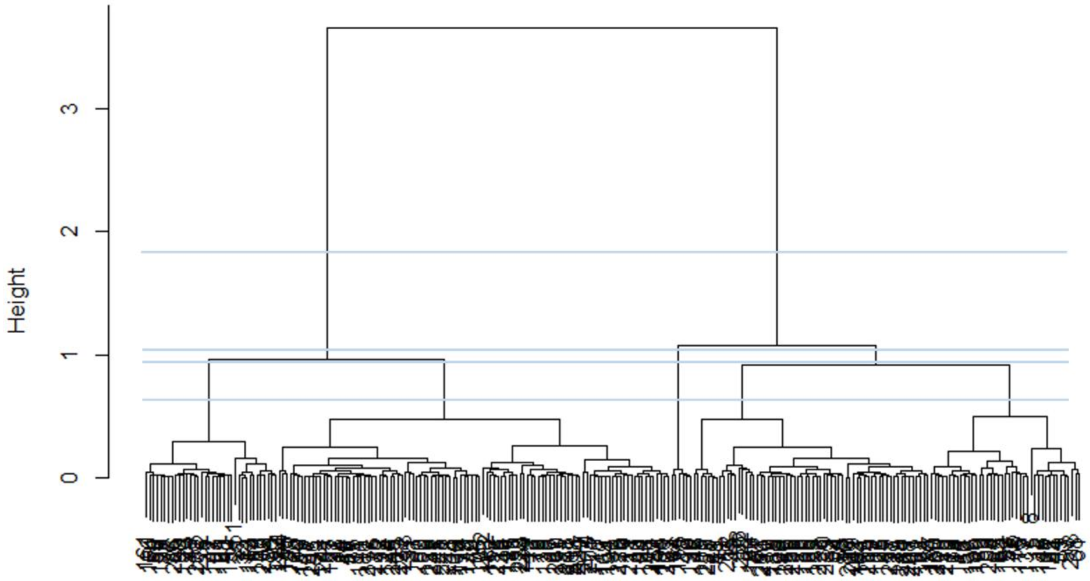

Figure D1. Dendrogram including visualization of possible number of clusters.

Table D1. Descriptive statistics per cluster using 4 clusters.

		0	**1**	**Min**	**1st quantile**	**Median**	**Mean**	**3rd quantile**	**Max**
Cluster 1	Treatment	61	45						
	Occupancy rate 2020			0.008	0.054	0.115	0.160	0.261	0.468
	Occupancy rate 2021			0.006	0.037	0.100	0.118	0.181	0.367
	Room type								
	Entire home	0	106						
	Hotel room	106	0						
	Private room	106	0						
	Latitude			0.049	0.544	0.596	0.593	0.644	1.000
	Longitude			0.000	0.404	0.452	0.448	0.498	0.787

Table D1 (continued). Descriptive statistics per cluster using 4 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 2	Treatment	80	25						
	Occupancy rate 2020			0.004	0.049	0.110	0.231	0.363	0.943
	Occupancy rate 2021			0.002	0.028	0.081	0.182	0.272	0.842
	Room type								
	Entire home	105	0						
	Hotel room	105	0						
	Private room	0	105						
	Latitude			0.000	0.502	0.575	0.536	0.627	0.852
	Longitude			0.013	0.424	0.515	0.495	0.568	1.000
Cluster 3	Treatment	4	2						
	Occupancy rate 2020			0.020	0.042	0.080	0.084	0.110	0.174
	Occupancy rate 2021			0.018	0.033	0.043	0.059	0.079	0.126
	Room type								
	Entire home	6	0						
	Hotel room	0	6						
	Private room	6	0						
	Latitude			0.545	0.591	0.606	0.607	0.634	0.658
	Longitude			0.386	0.403	0.461	0.459	0.507	0.537
Cluster 4	Treatment	23	13						
	Occupancy rate 2020			0.288	0.545	0.588	0.639	0.754	0.994
	Occupancy rate 2021			0.341	0.411	0.488	0.517	0.570	0.882
	Room type								
	Entire home	0	36						
	Hotel room	36	0						
	Private room	36	0						
	Latitude			0.483	0.579	0.597	0.591	0.608	0.719
	Longitude			0.368	0.440	0.452	0.477	0.510	0.626

$N_{cluster\ 1} = 106$
$N_{cluster\ 2} = 105$
$N_{cluster\ 3} = 6$
$N_{cluster\ 4} = 36$

Table D2. Balance improvement measured by SB and two sample t-test on clustered data using 2 clusters.

	Covariates	Balance improvement measured by SB (in %)	t-test
Cluster 1	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	-314.3	0.342
	longitude	13.9	1.982*
Cluster 2	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	520.1	0.336
	hotel_room1	520.1	-0.336
	private_room0	520.1	-0.336
	private_room1	520.1	0.336
	latitude	-26.2	0.927
	longitude	29.4	1.659

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

Table D3. Descriptive statistics per cluster using 2 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 1	Treatment	84	58						
	Occupancy rate 2020			0.008	0.069	0.198	0.281	0. 459	0.994
	Occupancy rate 2021			0.006	0.051	0.153	0.219	0.344	0.881
	Room type								
	Entire home	0	142						
	Hotel room	142	0						
	Private room	142	0						
	Latitude			0.049	0.556	0.596	0.592	0.632	1.000
	Longitude			0.000	0.412	0.452	0.455	0.500	0.787

Table D3 (continued). Descriptive statistics per cluster using 2 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 2	Treatment	84	27						
	Occupancy rate 2020			0.004	0.048	0.108	0.223	0.332	0.943
	Occupancy rate 2021			0.002	0.028	0.079	0.176	0.245	0.824
	Room type								
	Entire home	111	0						
	Hotel room	105	6						
	Private room	6	105						
	Latitude			0.000	0.503	0.584	0.540	0.628	0.852
	Longitude			0.013	0.420	0.512	0.493	0.561	1.000

$N_{cluster\ 1}$ = 142
$N_{cluster\ 2}$ = 111

Table D4. Balance improvement measured by SB and two sample t-test on clustered data using 3 clusters.

	Covariates	Balance improvement measured by SB (in %)	t-test
Cluster 1	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	514.3	0.342
	longitude	-13.9	1.982*
Cluster 2	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	-34.3	0.998
	longitude	38.3	1.873*

Table D4 (continued). Balance improvement measured by SB and two sample t-test on clustered data using 3 clusters.

	Covariates	Balance improvement measured by SB (in %)	t-test
Cluster 3	entire_home0	-	-
	entire_home1	-	-
	hotel_room0	-	-
	hotel_room1	-	-
	private_room0	-	-
	private_room1	-	-
	latitude	-	-
	longitude	-	-

Significance: * p < 0.1; ** p < 0.05; *** p < 0.01
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

Table D5. Descriptive statistics per cluster using 3 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 1	Treatment	35	58						
	Occupancy rate 2020			0.008	0.055	0.140	0.258	0.383	0.905
	Occupancy rate 2021			0.008	0.041	0.112	0.196	0.290	0.771
	Room type								
	Entire home	0	93						
	Hotel room	93	0						
	Private room	93	0						
	Latitude			0.049	0.555	0.604	0.597	0.642	1.000
	Longitude			0.000	0.408	0.450	0.444	0.497	0.658
Cluster 2	Treatment	16	25						
	Occupancy rate 2020			0.006	0.059	0.134	0.250	0.363	0.943
	Occupancy rate 2021			0.004	0.028	0.063	0.188	0.280	0.817
	Room type								
	Entire home	41	0						
	Hotel room	41	0						
	Private room	0	41						
	Latitude			0.000	0.491	0.599	0.508	0.627	0.846
	Longitude			0.013	0.407	0.459	0.431	0.528	0.796

Table D5 (continued). Descriptive statistics per cluster using 3 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 3	Treatment	1	2						
	Occupancy rate 2020			0.055	0.080	0.105	0.111	0.139	0.174
	Occupancy rate 2021			0.036	0.043	0.051	0.071	0.089	0.126
	Room type								
	Entire home	3	0						
	Hotel room	0	3						
	Private room	3	0						
	Latitude			0.545	0.579	0.612	0.599	0.627	0.641
	Longitude			0.398	0.450	0.502	0.479	0.520	0.537

$N_{cluster\ 1} = 93$
$N_{cluster\ 2} = 41$
$N_{cluster\ 3} = 3$

Table D6. Balance improvement measured by SB and two sample t-test on clustered data using 5 clusters.

	Covariates	Balance improvement measured by SB (in %)	t-test
Cluster 1	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	-181.0	-0.040
	longitude	-71.4	2.624***
Cluster 2	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	97.2	1.115
	longitude	4.8	2.811**

Table D6 (continued). Balance improvement measured by SB and two sample t-test on clustered data using 5 clusters.

	Covariates	**Balance improvement measured by SB (in %)**	**t-test**
Cluster 3	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	-71.6	-0.663
	longitude	-13.3	0.573
Cluster 4	entire_home0	-	-
	entire_home1	-	-
	hotel_room0	-	-
	hotel_room1	-	-
	private_room0	-	-
	private_room1	-	-
	latitude	-	-
	longitude	-	-
Cluster 5	entire_home0	0.0”	-
	entire_home1	0.0”	-
	hotel_room0	0.0”	-
	hotel_room1	0.0”	-
	private_room0	0.0”	-
	private_room1	0.0”	-
	latitude	19.1	-0.929
	longitude	-32.8	1.036

Significance: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$
” Covariate became totally balanced compared to the unmatched dataset.
Note: the balance improvement is relative to the unmatched dataset.
Note: ‘-’ means that it is impossible to run a SB analysis or t-test due to either a too small sample size per cluster (leading to impossibility of matching) or the data being totally balanced.

Table D7. Descriptive statistics per cluster using 5 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 1	Treatment	61	45						
	Occupancy rate 2020			0.008	0.054	0.115	0.160	0.261	0.468
	Occupancy rate 2021			0.006	0.037	0.100	0.118	0.181	0.367
	Room type								
	Entire home	0	106						
	Hotel room	106	0						
	Private room	106	0						
	Latitude			0.049	0.544	0.596	0.593	0.644	1.000
	Longitude			0.000	0.404	0.452	0.448	0.498	0.787
Cluster 2	Treatment	46	18						
	Occupancy rate 2020			0.004	0.024	0.060	0.066	0.094	0.181
	Occupancy rate 2021			0.002	0.017	0.040	0.046	0.068	0.150
	Room type								
	Entire home	64	0						
	Hotel room	64	0						
	Private room	0	64						
	Latitude			0.000	0.461	0.567	0.516	0.627	0.852
	Longitude			0.013	0.415	0.513	0.494	0.608	1.000
Cluster 3	Treatment	34	7						
	Occupancy rate 2020			0.193	0.298	0.446	0.487	0.661	0.943
	Occupancy rate 2021			0.132	0.229	0.359	0.395	0.558	0.842
	Room type								
	Entire home	41	0						
	Hotel room	41	0						
	Private room	0	41						
	Latitude			0.137	0.544	0.588	0.569	0.629	0727
	Longitude			0.071	0.447	0.517	0.496	0.537	0.686
Cluster 4	Treatment	4	2						
	Occupancy rate 2020			0.020	0.042	0.080	0.084	0.110	0.174
	Occupancy rate 2021			0.018	0.033	0.043	0.059	0.079	0.126
	Room type								
	Entire home	6	0						
	Hotel room	0	6						
	Private room	6	0						
	Latitude			0.545	0.591	0.606	0.607	0.634	0.658
	Longitude			0.386	0.403	0.461	0.459	0.507	0.537

Table D7 (continued). Descriptive statistics per cluster using 5 clusters.

		0	1	Min	1st quantile	Median	Mean	3rd quantile	Max
Cluster 5	Treatment	23	13						
	Occupancy rate 2020			0.288	0.545	0.588	0.639	0.754	0.994
	Occupancy rate 2021			0.341	0.411	0.488	0.517	0.570	0.882
	Room type								
	Entire home	0	36						
	Hotel room	36	0						
	Private room	36	0						
	Latitude			0.483	0.579	0.597	0.591	0.608	0.719
	Longitude			0.368	0.440	0.452	0.477	0.510	0.626

$N_{cluster\ 1}$ = 106

$N_{cluster\ 2}$ = 64

$N_{cluster\ 3}$ = 41

$N_{cluster\ 4}$ = 6

$N_{cluster\ 5}$ = 36

Appendix E. Difference in Differences

This appendix supports Section 6.2.2 of the results by displaying the effects of treatment on occupancy rate by running the linear regression in ways that differ from the "winning" approach. This is done as robustness check.

Table E1. Effects of treatment on occupancy rate compared to non-treated, using unmatched data.

	Coefficient (Std. Error)
Intercept	0.268***
	(0.018)
Treatment	-0.038
	(0.030)
Time	-0.055*
	(0.025)
Treatment * Time	0.002
	(0.043)

Significance: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

R^2: 0.021

Table E2. Effects of treatment on occupancy rate compared to non-treated, using Cluster-then-Matching without replacement.

	Coefficient (Std. Error)
Intercept	0.282***
	(0.017)
Treatment	-0.051*
	(0.024)
Time	-0.065**
	(0.024)
Treatment * Time	0.008
	(0.034)

Significance: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

R^2: 0.030