ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Economics and Business Economics

# A Machine Learning Analysis on Sustainable Influencer Marketing

Name Student: Ryan Feenstra

Student ID Number: 561424

Supervisor: Maximilian Beichert

Second assessor:

Date Final Version: 24/07/2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

## Table of Contents

**Table of Contents........................................................................................................... 2**
**Abstract.........................................................................................................................4**
**Chapter 1: Introduction................................................................................................. 5**
1.1 Relevance.................................................................................................................. 6
1.2 Research Question....................................................................................................7
To what extent can machine learning models promote sustainable influencer marketing?.........................................................................................................7
1.3 Ethical Research Issues.............................................................................................7
1.4 Research Limitations..................................................................................................7
1.5 Chapter Overview.......................................................................................................8
**Chapter 2: Literature Review......................................................................................... 9**
Drivers for Product Returns:..........................................................................................9
Influencer Marketing:...................................................................................................10
Sustainability:.............................................................................................................. 13
Predicting Returns & Text Analytics:.......................................................................... 15
2.5 Summary of Key Findings........................................................................................ 17
2.6 Sub Questions..........................................................................................................18
2.7 Hypotheses.............................................................................................................. 19
Hypothesis 1: Mega Influencers will have the lowest effect on returns..................... 19
Hypothesis 2: The return reason with the highest importance in predictions will be "did not meet expectations".................................................................................20
Hypothesis 3: The sentiment of the "open" return reasons will relate to product characteristics.................................................................................................... 20
Hypothesis 4: The amount of followers an influencer has will be the most important influencer feature in predicting return reasons...................................................21
2.8 Conceptual Framework............................................................................................ 21
**Chapter 3: Research Methodology...............................................................................22**
3.1 Data Collection.........................................................................................................22
3.2 Research Sample & Variables...................................................................................23
3.3 Data Analysis Methods.............................................................................................25
3.4 Possible Research Bias............................................................................................ 26
**Chapter 4: Results......................................................................................................... 26**
4.1 Results..................................................................................................................... 26
General Metrics.........................................................................................................26
XGBoost Model......................................................................................................... 31
Text Analytics............................................................................................................ 39
**Chapter 5: Conclusions and Recommendations........................................................ 44**

5.1 Comparison of Key Findings........................................................................ 44
5.2: Hypothesis Review.................................................................................45
Hypothesis 1: Mega Influencers will have the lowest effect on returns...................... 46
Hypothesis 2: The return reason with the highest importance in predictions will be “did not meet expectations”......................................................................... 46
Hypothesis 3: The sentiment of the “open” return reasons will be product centred... 46
Hypothesis 4: The amount of followers an influencer has will be the most important influencer feature in predicting return reasons.......................................................... 47
5.3 Central Research Question........................................................................ 47
5.4 Recommendations for Future Research.................................................... 49
5.5 Limitations................................................................................................ 49
**Appendix A: Works Cited.................................................................................... 50**
**Appendix B: University of Mannheim Data NDA..............................................56**
**Appendix C: R Code........................................................................................... 57**

## Abstract

The topic of sustainability is one of importance to many companies that aim to adapt their business practices to align with the green transition. One key factor relating to sustainability that has a lot of room for improvement is the frequency of product returns in the retail industry. Product returns are not only detrimental in relation to profits with returns costing an average of 59% of the original sales price (Chaturvedi, 2022), they also leave a significant impact on the environment with product returns accounting for 27 million tons of CO2 emissions yearly (Chaturvedi, 2022). In order to effectively analyse the reasons behind product returns, it is important to investigate how companies promote their products to customers. Many businesses make use of social media influencers in order to promote their products, therefore it is of interest to investigate the impacts these social media influencers have on the reasons for product returns, and how this information can be utilised to develop a sustainability framework. In order to investigate the impact of influencers on product returns, data on purchases from an online e-commerce platform was linked to promotional discount codes unique to different influencers. The e-commerce influencer data was then put through an XGBoost machine learning model and text analytics techniques with the goal of understanding the various aspects of influencers that lead to a product return. The central research question was: To what extent can machine learning models promote sustainable influencer marketing. The results from the XGBoost machine learning model indicated that the amount of followers, following, and posts that an influencer has are important in determining the return reason of an item, while the text analytical models highlighted the importance of value, quality, and discounts in determining the return reason. The results of the XGBoost and text analytics models were combined into a sustainability framework that can aid companies in choosing influencers to promote their products which will lead to the lowest amount of product returns and subsequently reduce the environmental footprint of the company.

## Chapter 1: Introduction

In our ever growing society, the issue of sustainability is one of the most spoken about and prevalent issues of our time.  Sustainability affects or will affect every individual and business in some way, and in order to promote a sustainable future, businesses will need to adapt and make changes to their business model. Improving sustainability can be tackled from a multitude of different angles, however, one approach that is not looked at often is how social media plays a role in the issue of sustainability. Social media alongside sustainability is also one of the most prevalent aspects of our life and social media will only continue to grow and have greater importance. As of data in 2020 there are around four billion social media users (Kemp, 2020), while social media has such a broad reach that many have conducted studies on the addictive nature of social media (Cheng et al, 2021). Hence, social media is a prime channel for companies to promote their products by allowing influencers to provide reviews or discount codes for certain products. Social media and sustainability can be studied conjointly when looking at the retail industry. Many companies approach "influencers" on social media as a means of advertising, however, a lot of products that are sold end up being returned for a variety of reasons (De Veirman et al., 2016). These product returns are often handled in an unsustainable manner or are excessive and end up leading to extra waste and costs for the company involved (Cullinane et al, 2019). Many retail firms have frameworks in place to deal with product returns, however, there is limited literature on the topic and  a lack of consensus on how to properly handle a large volume of  product returns (Zhang et al, 2023). Therefore, a data driven approach is necessary in order to provide insights on the reasons why products bought from influencer promotion are being returned and what features of an influencer lead to higher rates of returns. The results from the data driven approach can then be adapted into a framework on how to handle returns that are purchased from influencer advertising. A data driven approach will aim to utilise machine learning techniques composed of an XGBoost model and text analytics in order to highlight key insights on why a product is returned.

### 1.1 Relevance

The paper firstly is scientifically relevant as the amount of literature on machine learning within sustainability and product returns is limited. It is challenging to find literature that aims to utilise real-world data and machine learning to aid in  solving the problem of product returns. Therefore, this paper will aim to provide a use-case scenario where machine learning can be applied to real data, and the insights from machine learning can be adapted into a format that can aid companies in dealing with their product returns. Furthermore, while studies have been conducted on how influencers may impact the effectiveness of marketing campaigns, the use of machine learning is scarce, while there is also limited research on which exact factors of an influencer are most important in determining whether a product is returned or not. The paper is also of practical relevance as it will provide a data driven framework for managing product returns that can be adapted by firms. By providing a data driven product return framework, companies will be able to visualise key factors that affect their product returns and be able to adapt the findings of the thesis into their own product return frameworks, with the goal of lowering the impact of their returns on sustainability. Furthermore, the thesis will also provide the needed attention on the issue of product returns and how they can negatively impact sustainability. As mentioned earlier, the impact of product returns on sustainability is one that has not been studied at a great depth, meaning that the consequences of product returns are not fully understood. Therefore, this thesis is relevant as it can elicit a reaction from retail firms to have a deeper look into their product returns and the current framework that is in use, and to evaluate their overall environmental footprint. In addition, companies will be more aware of the costs/benefits of making use of influencers for product advertisement and the overall impact of these influencers on sustainability. Having more information on the impact of influencers on sustainability will allow for companies to make better choices regarding which influencers they decide to partner with and to cut ties with certain influencers that may have an overwhelmingly negative impact on the environment due to related product returns. Overall, the thesis will add to the scientific literature surrounding the topic of product returns and how influencers may affect product returns, while the thesis also provides valuable business insights for developing product return frameworks.

## 1.2 Research Question

The research question that the thesis aims to answer is:

**To what extent can machine learning models promote sustainable influencer marketing?**

The research questions aims to explore the possibilities of machine learning in designing a framework that aims to aid companies in achieving sustainable influencer marketing. By having models that are able to predict return reasons and insights from text analytics, companies are better able to understand why a product is returned and what characteristics of influencers may have an impact on the reason of the return. With the information above companies can make better choices regarding sustainable influencer marketing and have a stronger grip on their product returns.

## 1.3 Ethical Research Issues

There are few ethical issues that could be present in the research. All potential personal information about an individual in the data has been removed, which include the first names, zip codes, and ages of customers. The only location based data that has been kept is the city and the country that the customer placed the order from. All customers in the data have a unique customer id, however, this customer id cannot be traced to a specific customer, therefore, maintaining anonymity in the dataset. Lastly, consent has been received for using the customer purchase data.

## 1.4 Research Limitations

Due to the nature of the research, a few limitations are present. The first issue is the balance of the data. A lot of customers tend to not decide to put in a reason when they return their product. By not providing a reason for their product return, it leads to a majority of the data having no reason for a product return, making it harder for a machine learning model to make accurate predictions. However, the issue of unbalanced data can be mitigated by using certain techniques and sophisticated machine learning models that are designed to handle unbalanced data. Another limitation in the research is the open field for writing the reason for a product return. In the open field a customer is able to write a personalised reason for why the product is going to be returned. The issue with the open field is that it is hard to validate if the reason mentioned is genuine, furthermore, it is hard to categorise open field reviews as customers may leave multiple reasons for the product return, making it difficult to determine which reason will be fed into the machine learning model. The last limitation is that one of the product return options is “Other reason”. The “other reason” category is not informative as it does not provide any information as to why a customer decided to return an item, and companies do not gain any value if a machine learning model predicts a return reason to be “Other reason”.

### 1.5 Chapter Overview

The structure of the thesis will be split into 5 chapters. The chapters are Introduction, literature review, research methodology, results, and conclusion & recommendations. The literature review will define key terms related to the research and will investigate how these terms have been discussed in other literature. The key findings of the literature will be discussed for each term. Within the literature review chapter the sub-questions and hypotheses will be introduced. The research methodology chapter will illuminate the quantitative machine learning methods that will be employed in the thesis, and will explain the data collection and the sample of data used. The results chapter will dive into the key results from the analysis and how the results link back to the research question. Lastly, the conclusions and recommendations chapter will compare the results of the thesis to the results from other literature, while also providing recommendations to companies on how sustainable influencer marketing can be achieved. Also, the hypotheses will be accepted or rejected, and the central research question will be answered.

## Chapter 2: Literature Review

Firstly, some key terms will have to be defined. The literature review will then be grouped by these key terms. The first key term is “drivers for product returns” which can be defined as factors that cause people to return their products. Influencer Marketing which can be defined as products being promoted on the instagram page of an influencer. The next key term is sustainability, which will be seen in the sense of lowering the amount of product returns in order to lead to more sustainable marketing. Lastly, predicting returns can be defined as using machine learning models in order to classify if a product is returned or not.

### Drivers for Product Returns:

One of the features of social media that may drive products to be returned is Word-of-mouth (WOM). Stephen (2016), looks into the idea of online WOM in product reviews and how it affects purchase chance. The idea can be converted to Instagram comments to investigate how WOM from Instagram comments on influencer posts may drive certain reasons for product returns. Shahbaznezhad et al. (2018) show that user comments and the sentiment within the comments affect the content strategies of companies, highlighting the importance of WOM in a social media platform. Another potential driver for product returns is the follower count of influencers and the amount of accounts the influencer follows. De Veirman et al. (2016) show that the amount of followers an influencer has does have a significant relationship with that influencer being an opinion leader. Therefore, if the influencer is an opinion leader, it adds a layer of trust in the influencer which may be a driver to lower the amount of product returns, hence, an influencer with a very low follower count may lead to an increased amount of product returns. However, De Veirman et al. (2016) also indicate that if an influencer themself follow only a few people, a larger follower count may lower the likeability of an influencer. Another potential factor that drives product returns is the gender of the followers. Powers & Jack (2013) and Powers & Jack (2015) investigate the ideas of product and emotional dissonance and show that males differ from females as males had a low relationship for the consideration of liberal return policies and emotional dissonance. Powers & Jack (2013) indicate the possibility that the gender of a follower could have a significant impact on whether a product is returned and the reason for the return. Griffis et al. (2012) highlight that the loyalty and return rate of a customer can have an implication on the frequency of future returns, and that these factors should be taken into account when a retailer determines its product returns framework. Pei & Paswan (2018) illuminate the fact that customers have two distinct reasons for returning a product, which are opportunistic and legitimate return reasons. Pei & Paswan (2018) describe legitimate return reasons as when the consumer makes impulsive purchases, product compatibility, and social influence, while opportunistic reasons relate to immoral reasons for product returns. Lastly, Lv & Liu (2022) show that "information overload" can be a significant driver in product returns. Considering the fact that such a wide variety of products and retailers are endorsed by influencer marketing, there could be a significant chance of information overload having an impact on the return rate.

The main key finding is that there are a plethora of factors that may influence the return rate of a customer. Factors like word of mouth, individual characteristics of customers and influencers, and information overloads are all shown to be important in determining the return rate (Stephen, 2016), (De Veirman et al., 2016), (Powers & Jack , 2013), (Griffis et al., 2012). While factors about consumers are also important, like the gender Powers and Jack (2013), and the loyalty and previous return rate (Griffis et al., 2012). Factors regarding influencers also play an important role (De Veirman et al., 2016). The consequence of these results is that potential machine learning models aiming to analyse the product return rate need to incorporate a wide range of these factors in the analysis, to provide a robust model that can be utilised effectively for developing a product returns framework.

**Influencer Marketing:**

The rise of social media has allowed for the formation of mini-celebrity "influencers" (Booth & Matic, 2011). However, the topic of influencer marketing is currently lacking a lot of academic papers surrounding the subject (Martinez-Lopez et al, 2020). These influencers have significant sway on how consumers perceive  a brand and act as opinion leaders that garner consumer's trust (Booth & Matic, 2011). Bush et al. (2004) presents an example of celebrities having significant influence over consumers, highlighting that famous sportspeople get looked up at as role models and are crucial in promoting a whole range of products. Influencers promoting products can be called “Influencer Marketing” and revolves around companies using individuals that have a large reach on social media to promote their brand / product (Cheng et al, 2024). The platform that enables the presence of influencers is social media, Yuchi et al. (2017) investigate the general relationship between social media and consumer shopping activities. Yuchi et al. (2017) look into a year's worth of data on customer social media activity and the use of social media, finding that increased time on social media leads to more shopping. Yuchi et al. (2017) highlight that the effect of increased shopping only takes effect after a certain period of time as immediately after browsing social media it seems that shopping activity decreases. Yuchi et al. (2017) therefore indicate that social media may have lagged effects on customer purchasing behaviour which has important practical applications as companies may need to wait a certain amount of time before evaluating the effects of their influencer marketing. Lou & Yuan (2019) look at the relationship between influencers and consumers, by utilising surveys sent to people who follow influencers and subsequently rating influencers on their trustworthiness. Lou & Yuan (2019) use a partial least squares model path model, which indicated that factors like trustworthiness, attractiveness, and informative value of influencer posts increase the trust among followers relating to branded posts, subsequently leading to higher brand awareness. The insights from Lou & Yuan (2019) indicate that companies need to be rigorous in their selections of influencers and need to ensure that they pick influencers with informative content and an already established sense of trustworthiness. McCormick (2016) expands on the idea of influencer credibility, by investigating consumer reactions to unfamiliar celebrities. Consumers tend to react unfavourably to unfamiliar celebrities and are less likely to purchase a product that is endorsed by an unfamiliar celebrity (McCormick, 2016). Trivedi & Sama (2019) utilise surveys to test the moderation effect between brand admiration and attitude for influencer marketing and online purchase intentions. Trivedi & Sama (2019) highlight that companies should choose expert influencers instead of celebrity influencers when promoting a product, as consumers are more likely to trust influencers that are experts in the field of the product that they are buying, while they are not that interested in messages from well-known celebrities. Kim et al. (2014)

expand on the idea of influencer credibility by illuminating the fact that trust in a celebrity can be transferred to a product or service. Kim et al. (2014) use the example of a hotel in Korea, where perceptions of trust in influencers were transferred to trust in the hotel. Ultimately, Kim et al. (2014) conclude that the choice of influencer has to match the market you are promoting too, and that multiple influencer promoters might be needed to appeal to a wide audience. Wang et al. (2017) study the airline industry and aim to explore the effectiveness of influencer credibility. Wang et al. (2017) conclude that using a credible influencer for promoting a brand through advertisements leads to an increase in brand reputation, purchase intentions and brand attitudes. On the other hand, celebrity influencers can also have an adverse effect on the promotion of a brand (Erfgen et al,  2015). Erfgen at al. (2015) discusses the "vampire effect" which is when a celebrity overshadows the brand they are meant to be promoting. The vampire effect can lead to lower brand recall and companies need to ensure they choose influencers that suit the nature and size of the company in order to avoid the negative repercussions of the vampire effect (Erfgen at al, 2015). In addition to the vampire effect, the idea of negative publicity is important in determining the effectiveness of a celebrity as an influencer (Zhou & Witla, 2013). Negative publicity surrounding an influencer evokes the morality of a consumer, and causes an endorsement campaign to perform worse (Zhou & Witla, 2013).

The key insight from influencer marketing is the choice of the type of influencer to promote a product. The choice of an influencer is very important as choosing an influencer that is too well-known can lead to lowering brand recall (Erfen et al, 2015), while choosing an influencer that fits the brand well can lead to positive awareness and higher purchase intentions (Wang et al, 2017). Hence, the idea of influencer credibility is of paramount importance (Kim et al, 2014), and companies should choose influencers who are experts in the field of the product they are selling (Trivedi & Sama, 2019). Ignoring credibility and popularity and using lesser known celebrities can lead to lower purchase intentions (McCormick, 2016).

**Sustainability:**

Sustainability and focusing on the implementation of sustainability drivers is currently a very important target that organisations must implement (Lozano & von Haartman, 2017). However, currently the amount of academic research on sustainability within retail organisations is quite low (Wiese et al, 2012).  Product returns are not only a major problem for retailers in terms of profitability (Jack et al, 2019), but product returns also have a significant impact on sustainability as excessive product returns have a negative impact on the environment (Cullinane et al, 2019). Furthermore, Aydin et al. (2018) utilise simulations to indicate that having uncertainty surrounding the return rate of products leads to an adverse reaction for not only the profitability of a product but also the impact on the environment that the product will have. Looking at the specifics of how product returns may impact the environment,  Zhang and Frei (2023) investigate the ecological impact of product returns, and how to develop infrastructure as a company to effectively manage the product returns you have. Zhang and Frei (2023) indicate that the effects of product returns on sustainability are widespread, as returns may lead to increased packaging and transportation costs, and a lot of returns may be discarded instead of re-sold. Zhang and Frei (2023) find that the financial impact of returns is very clear to the retailers, however, many retailers are not aware of the inherent risks to sustainability that product returns present. Furthermore, retailers indicate that handling product returns is a very complex task.  Frei et al. (2020) look into the effects of eCommerce on sustainability. Frei et al. (2020) find that the consumption of goods online is at an all-time high, which leads to an increased amount of product returns. In line with other papers, Frei at al. (2020) finds that companies are not prepared to deal with the environmental impact of product returns and that the actual impact on the environment is not well understood. Having to return products involves having to transport the goods back to retailers or distribution centres, Edwards et al. (2010) finds that car trips to pick up or return shopping items lead to unnecessary and harmful levels of C02 to be released, hurting sustainability as a whole. The current most environmentally friendly standard for dealing with product returns is a term called “reverse logistics”, which is the idea of using delivery vehicles to pick up packages that need to be returned (Bernon & Cullen, 2007).  Reverse logistics still contributes to an overall rise in emissions, hence, hurting sustainability (Bertrand & Chi, 2018). Having looked at the impact of product returns, another important area to discuss is the current state of methods that organisations use to assess the impact of the environment on supply chains. Currently there is very limited literature on environmental assessment methods that incorporate the idea of product returns (Zhang et al, 2023). One of the current main environmental assessment methods is life-cycle assessment, which involves gathering data on environmental issues and restructuring supply chains based on the data collected in order to reduce the environmental impact of the supply chain (Hagelaar, 2001). Hagelaar (2001) also mentions that the integration of LCA into a supply chain does not have well established guidelines making it hard for organisations to currently make use of the method. Hagelaar (2001), suggests that organisations need to differ between different types of LCA's and tune the structure of a supply chain to fit a specific LCA, in order to properly integrate it. Daniel et al. (2004) describes an extension to the LCA model, which is the LCA polygon. The LCA polygon aims to find "impact categories", that take a value between 0 and 1, these impact categories are then placed on an n-sided polygon to visualise the effect of each impact category on the environment (Daniel et al, 2004). Another relevant environmental analysis method is "Material Flow Analysis" (MFA), which is used to evaluate the environmental impact of material flows (Laner, 2014). While MFA is able to quantify the associated environmental impacts of product returns at each stage of the returns stage (Zhang et al, 2023), MFA is still clouded with a lot of uncertainty, due to the assumptions and number of sources needed to accurately construct the method (Laner, 2014).

As seen above, product returns are not only a financial trouble point for businesses, they also have a profound impact on the environment due to the logistics required in the returns process (Cullinane et al, 2019).  The key takeaway is that businesses currently do not adequately account for product returns, due to the environmental takeaways not being clearly understood in comparison to the financial burden, which is currently well understood (Zhang & Frei, 2023). In addition, many of the environmental assessment methods that companies currently employ, do not investigate the impact of product returns to a high enough degree (Hagelaar, 2001), or are clouded with uncertainty which makes the models not fully accurate in estimating the environmental costs of a product return (Laner, 2014).

**Predicting Returns & Text Analytics:**

The section on predicting returns and text analytics will look into the literature surrounding the use of machine learning and text analytics models in influencer marketing and product returns. Cui et al. (2020) find that a Least Absolute Shrinkage and Selection Operator (LASSO) model is the most accurate in modelling the volume of product returns. The use of a LASSO model allows for the introduction of interaction terms and provides a deeper insight on which particular features may interact to have a certain effect on product returns. However, while classification is possible in LASSO models, usually the use of more complex machine learning models yields more accurate classification results. Cui et al. (2020) utilise a variety of other data-driven models which include machine learning models like Random Forest and Gradient Boosting, while both of the machine learning models result in very high accuracy, the prediction accuracy on the test set ranks lower compared to the LASSO model, so Cui et al. (2020) decide that the LASSO model is most suited for the data. Cui et al. (2020) provide a robust framework on how machine learning can be deployed on product returns data, the paper focuses on predicting the volume of product returns, while the thesis will aim to classify return reasons, as having the reason for a product return will give companies more detailed insights on their returns and how to address them. Joshi et al. (2018) employ a two-step model that combines machine learning and network science in order to model if an apparel/garment product will be returned, by analysing past purchase and return history of a customer. The major findings are that apparel and garments are the product category that are returned the most and that one of the prominent reasons for returning apparel is due to the size or fit not being correct. The model used by Joshi et al. (2018) employs a variety of machine learning techniques like clustering and Support Vector Machines (SVM) that analyse return data where customers have returned apparel/garments with the reason being that the fit is incorrect. The model used by Joshi et al. (2018) is intricate and robust while producing extremely accurate predictions, however, the scope of the model is limited to customers that returned their products based on a size or fit issue, while in the thesis, the focus will be placed on predicting the reason for an item being returned more generally.

Furthermore, Purba & Tan (2023) utilises a random forest model to find the ideal time to post a promotional post, suggesting that a ratio of one regular post for 5.4 promotional posts is ideal. Li et al., (2018) implement machine learning models to evaluate the risk of a customer returning an item based on the composition of items in the online shopping basket, while also suggesting the reason why a product will be returned. Li et al., (2018) also indicate that the method they employ is able to distinguish risk of return at the basket and product level, giving webshops actionable insights on both levels. Using the findings from Li et al., (2018), a webshop can then take measures for customers that have a basket flagged for high risk of return. Jungmok & Harrison (2016) highlight the predictive power of machine learning models, as they predict the volume of returns using a case study of reusable bottles. Jungmok & Harrison (2016) conclude that using their predictive model selection algorithm leads to more accurate predictions of returns volume compared to other models that are employed like distributed lag models (DLM's). Looking past predicting product returns, another method of examining why customers return products is by analysing  reviews / reasons for returns that customers have posted using text analytics methods. Cheng et al., (2024) look at product reviews from Amazon and utilise various text analytics methods to predict the reason of a product return. Cheng et al., (2024) make use of a baseline word embeddings model and expand the model by introducing global vectors (GloVe) and Bidirectional Encoder Representations from Transformers (BERT). Cheng et al., (2024) finds that the model enriched with BERT scored the highest on measures like accuracy, recall, and precision, while customer reviews beats out product, merchant, time, and customer features on the same metrics, showing the importance and predictive power of product reviews. Another key aspect of text analytics is the use of sentiment analysis. Gallagher et al., (2019) use sentiment analysis on a set of product reviews, and assign sentiment scores (positive, negative, or neutral) to a review. Using the insight from the sentiment scores businesses can gauge how customers are feeling about products and what changes need to be made if the sentiment around products is neutral or negative (Gallagher et al., 2019).

The key takeaway from the section above revolves around the effectiveness of machine learning models in predicting return volumes and return reasons. Methods like Random Forests, SVM, and

LASSO are all very accurate in predicting return reasons and volume (Joshi et al, 2018), (Li et al, 2018), Cui et al, (2020), Purba & Tan (2023),  and highlight that machine learning is an appropriate and insightful way to analyse the impact of product returns. Furthermore, text analytics is also an important aspect of machine learning that can aid in explaining the reasons behind product returns. Text analytics models like GloVe and BERT are very accurate in predicting the reason behind a product return, while in some cases these text analytics models are even more important in prediction than customer and merchant features (Cheng et al., 2024).

### 2.5 Summary of Key Findings

In summation, product returns present themselves as not only issues on a financial level, but are a key detrimental factor in overall sustainability (Cullinane et al, 2019). Looking at the reasons behind product returns leads to a plethora of factors. Word of mouth, Individual characteristics of customers and influencers, and information overloads are all factors that can affect the return rate of a product (Stephen, 2016), (De Veirman et al., 2016), (Powers & Jack , 2013), (Griffis et al., 2012). All these factors mentioned above can be integrated into a machine learning model to investigate the predictive power of each factor. Businesses commonly use influencers to promote their products, and these products that are purchased can be returned for a variety of reasons, and one of the reasons is the choice of the influencer. The first important consideration in the choice of an influencer is credibility (Lou & Yuan, 2019). Credibility is important as consumers would rather purchase products that are promoted by influencers that are experts in the field that relate to the product bought rather than celebrities that have no relation to the product at all (Trivedi & Sama, 2019). On the other hand, choosing influencers or celebrities that are not well known lead to a significant detrimental effect on the purchase intention of customers (McCormick, 2016). Instead, companies should choose influencers that suit the brand image and have a strong relation to the products that the brand is selling. (Wang et al, 2017). The core issue with influencer marketing in relation to product returns is the negative impact it has on sustainability (Cullinane et al, 2019). The product returns process leads to a combination of factors like packaging, transportation, and disposal that have profound negative effects on the environment (Zhang and Frei, 2023), (Edwards et al., 2010), (Bertrand & Chi, 2018). However, these detrimental effects to the environment are currently not well understood by companies in relation to the clear financial burden that product returns pose (Frei at al., 2020). Furthermore, current environmental assessment methods including LCA and MFA do not adequately incorporate the issue of product returns into their framework (Hagelaar, 2001), (Laner, 2014). The tools that will be employed to visualise the return reasons will be a collection of machine learning models. Models like SVM's, LASSO, and Random Forest are proven to be very accurate in predicting return reasons (Joshi et al, 2018), (Li et al, 2018), Cui et al, (2020), Purba & Tan (2023). Text analytics methods have also proven to be extremely useful as GloVe, BERT, and sentiment models are able to very accurately predict the return reason of a product and in some cases these text analytic models prove to yield better results than models regarding the customer and influencer features.

### 2.6 Sub Questions

The first sub-question is: **What is the average return rate per type of influencer?**
Influencers can be grouped into different categories based on their follower count. These options are nano, micro, macro, and mega influencer. Knowing the return rate per type of influencer is important as the amount of followers an influencer has is an important determinant of whether the influencer is seen as an opinion leader (De Veirman et al, 2016), which could have an impact on the rate of return. The subquestion will also investigate if the "vampire" effect will kick in for influencers that have an extremely large following, as a very large amount of followers may cause the influencer to overshadow the product they are promoting (Erfgen at al, 2015).

The second sub-question is: **What is the most frequently mentioned return reason?**
The sub question arises due to the multitude of factors that may cause a consumer to return a product. Consumers may return items due to characteristics about themselves like loyalty and previous return rate, characteristics about the influencer that promotes the product like amount of followers and following, characteristics about the product itself and amount of information surrounding a product

(Stephen, 2016), (De Veirman et al., 2016), (Powers & Jack , 2013), (Griffis et al., 2012. Therefore, it will be investigated which of these reasons is most prevalent in the data sample used for this research. Knowing the most frequent return reasons will allow organisations to refine their products / influencer choice and investigate areas of a product which may be leading to a lot of product returns.

The third sub-question is: **How accurate is the XGBoost model in predicting the return reason of a consumer**

The main goal of the thesis is to use machine learning to aid in creating a framework to manage sustainable product returns. One of the machine learning applications that will be used is the XGBoost model for predicting return reasons. Previous studies that employed machine learning predictor methods to predict product return reasons find that the models are quite accurate in their predictions (Joshi et al, 2018), (Li et al, 2018), Cui et al, (2020), Purba & Tan (2023).

The last sub-question is: **What is the overall sentiment of the "open answer" review reasons?**

When consumers return their product, they get the choice between pre-defined return reasons, however, there is an option for consumers to fill in their own custom return reason. These return reasons will be analysed using the text analytics tool of sentiment analysis. Sentiment analysis has been shown to be an effective tool to evaluate product reviews (Gallagher et al., 2019), and will be used to determine how customers feel regarding the products they have bought and why they are returning them.

### 2.7 Hypotheses

Based on the literature review; hypotheses will be created, which aim to address the key findings of the literature review. These hypotheses will be investigated in Chapter 4: Results.

**Hypothesis 1: Mega Influencers will have the lowest effect on returns.**

McCormick (2016) highlights that consumers react unfavourably to products that are promoted by influencers with a lack of credibility. Lou & Yuan (2019) further show that the credibility of an influencer is important in determining purchase intentions. Based on the available literature we can hypothesise that consumers who purchase goods from established well known influencers, are likely purchasing from companies that make quality products that suit the consumers needs as Wang et al. (2017) shows that companies that use well-established influencers usually also enjoy higher brand reputation and loyalty. Based on the factors above, the return rate of a customer will be lower for products endorsed by well-established influencers as the companies behind the products most likely enjoy established brand loyalty and the consumer is buying products that suit their needs and would not need to be returned as quickly as other products.

**Hypothesis 2: The return reason with the highest importance in predictions will be "did not meet expectations"**

Powers & Jack (2015) highlight the fact that there is a positive relationship between emotional dissonance resulting from product dissatisfaction and the amount of product returns. Furthermore, Powers & Jack (2015) indicate that females experience the emotional dissonance effect more strongly compared to males. Furthermore, Powers & Jack (2015) indicate that the two main reasons for a customer to want to return their product is the product not meeting expectations and the customer finding a better priced product. Therefore, it can be expected that the return reason that will have the highest importance in predictions will be the "did not meet expectations" return reason, as it relates to a customer's dissatisfaction with a product.

**Hypothesis 3: The sentiment of the "open" return reasons will relate to product characteristics**

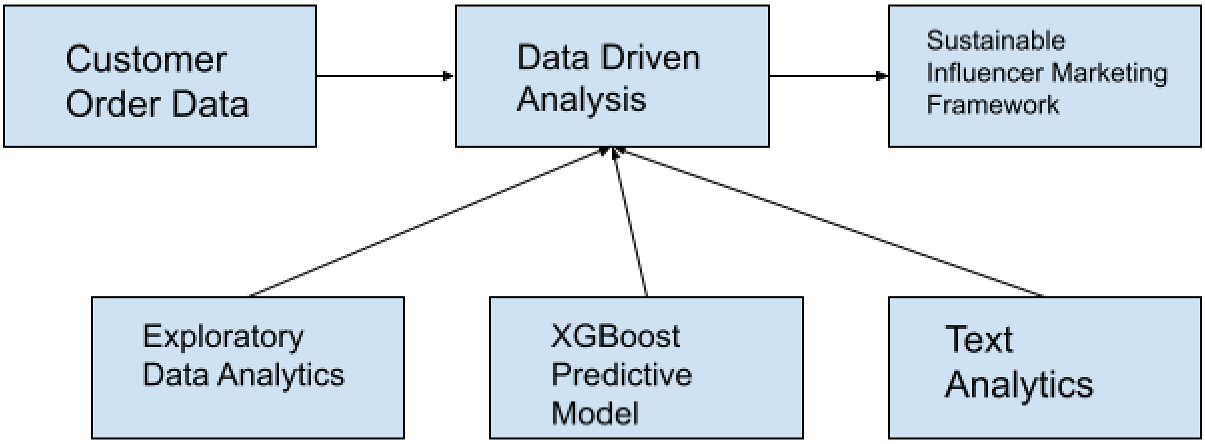
Due to the open return reasons allowing for customers to fill in their own return reason, it can be expected that customers will fill in this custom return reason when the reason of return does not match the standard options presented. Therefore, it can be expected that the "open" return reason will be utilised when the customer has specific views of the products bought that are driving them to return the product. It is less likely that the open return reasons will focus on areas of the returns process like incorrect orders or consumers changing their mind, as these return reasons are already present in the standard options.

**Hypothesis 4: The amount of followers an influencer has will be the most important influencer feature in predicting return reasons**

Lou & Yuan (2019) and McCormick (2016) both indicate the importance of influencer credibility and how it may affect purchase intentions. Furthermore, De Veirman et al, (2016) highlight that the amount of followers that an influencer has is an important factor in making the influencer an opinion-leader. Therefore it can be hypothesised that the amount of followers will be a key consideration for consumers when they purchase and return a product. If the influencer has a lot of followers and is branded as an opinion leader, consumers can be influenced to purchase specific goods from the influencer and are less likely to return the goods, while goods bought from influencers with a low amount of followers may be returned much quicker.

### 2.8 Conceptual Framework

The goal of the thesis will be to ultimately answer the research question by providing a product returns framework that uses machine learning to take the impact of influencers into account. Based on the data and research question, the following conceptual framework is employed.



The conceptual framework highlights the journey the thesis will be taking. The goal of the thesis is to apply data driven analysis on customer order data in order to assess the impact of influencers on product returns with the goal of developing a sustainable influencer marketing framework. The upper branch of the conceptual framework  shows the journey described. The moderating features of the data driven analysis include the various data based models that will be used to produce insights. These models include exploratory data analysis, using an XGBoost predictive machine learning model, and Text analytics.

## Chapter 3: Research Methodology

The research will make use of quantitative machine learning models in order to assess the impact of influencers on product returns. The choice for a quantitative method was made due to the nature of the research question revolving around how machine learning models can be applied and how the results of machine learning models can be transformed into insights. Since the research question is designed around the use of machine learning, no qualitative research methods will be utilised, as qualitative research methods would be more suited for research where the research question can be approached from a more subjective manner and does not have a clear analytical answer (Mulisa, 2021). A mixed approach would also not be appropriate as a mixed approach is used for research where the central question needs a balance between objectivity and subjectivity (Mulisa, 2021).

### 3.1 Data Collection

The data originates from a research project conducted at the University of Mannheim, and has been approved to be used for this thesis. The data revolves around orders that have been placed on online webshops from a Swedish retail company and contains information on the order value, the product that has been ordered, and if a voucher was used when purchasing the order. In addition, the data includes the value of orders that have been returned. The product returns data also includes a reason why the item was returned, which can be a value from a list of options provided by the retailer, or can be filled in manually by a customer, allowing for an open field. The data from the return orders is linked to the general information about an order by a unique “order id”. The next set of data included is information surrounding customers. Customers can be linked to a specific order by a unique “customer id” and the information available about a customer includes the birth date, the city and country the customer lives in, the gender, the zip code, and lastly the first name of the customer. As mentioned earlier in the ethical research issues, the zip code and first names of customers have been omitted from the data in order for none of the results to be traced back to an individual. Lastly, the dataset includes information about influencers. There is information on the username of the influencer, the amount of posts made, and how many followers and following a specific influencer has. All the influencer data is taken from the social media platform Instagram, and is linked to a specific order based on the voucher code that is used when a customer places an order. All the data is combined into one query in order to facilitate further analysis. The data was collected in a multitude of countries, hence all the return reason data has been passed through a google translator R package that has translated all return reasons to English in order to facilitate analysis.

### 3.2 Research Sample & Variables

The research sample includes information on 267,000 return orders, where 3655 orders are directly linked to an influencer. There is a sample of 357,000 customers in the data. Customers reside in a multitude of countries around the world, allowing for the research to not be limited to the behaviour of customers in a certain region of the world. In addition, 72,000 unique influencers are included in the dataset who have followers in the range from the low thousands to the multiple millions. Having a broad range of influencers is important as it allows for analysis on how specific types of influencers (eg. mega, micro, macro, nano, and nano-nano) impact product returns.

The following variables will be used in the analysis:

| Variable Name | Variable Origin |
|---|---|

| | |
|---|---|
| Quantity | Originates from the Return Table dataset.<br>Indicates the quantity of items returned.<br>Type: Numeric |
| Value | Originates from the Return Table dataset.<br>Indicates the value of the item returned.<br>Type: Numeric |
| Price (discount) | Originates from the All Orders dataset. Indicates the price of an order with discounts applied.<br>Type: Numeric |
| Price | Originates from the All Orders dataset. Indicates the price of an order without discounts<br>Type: Numeric |
| Gender | Originates from the All Customers dataset.<br>Indicates the gender of a customer.<br>Type: Factor with 4 levels (Man, Woman, Unknown, Non-Binary) |
| Number of Orders | Variable created in the All Customers dataset.<br>Indicates the number of orders a customer has placed.<br>Type: Numeric |
| Returns | Variable created in the All Customers dataset.<br>Indicates the amount of orders a customer has returned.<br>Type: Numeric |
| Return Rate | Variable created in the All Customers dataset.<br>Indicates the amount of returned orders divided by the amount of total orders.<br>Type: Numeric |
| Posts | Variable originates from Influencer dataset.<br>Indicates the amount of posts an influencer has posted.<br>Type: Numeric |
| Follower | Variable originates from the Influencer dataset.<br>Indicates the amount of followers the influencer has on social media.<br>Type: Numeric |
| Following | Variable originates from the Influencer dataset.<br>Indicates the amount of accounts the influencer is following on social media.<br>Type: Numeric |
| Influencer Class | Variable originates from the Influencer dataset.<br>Indicates the class of an influencer based on the amount of followers the influencer has.<br>Type: Factor with 5 levels (macro-influencer, |

| | |
|---|---|
| | mega-influencer, micro-influencer, nano-influencer, nano-nano-influencer). |
| **Return Reason** | Target variable. Originates from the Return Table dataset. Highlights the reason why a product was returned.<br>Type: Factor with 6 levels (incorrect order, does not match description, no return reason mentioned, changed mind, another reason, does not meet expectations) |

### 3.3 Data Analysis Methods

The initial analysis of the data will include simple descriptive statistics that will aim to highlight key factors of the data like the type of influencers, and value of product returns. After the initial descriptives machine learning methods will be applied based on the nature of the data. Due to the structure of the “return reason” column, data can take the form of a structured return reason that was listed by a retailer, or it can take the unstructured form which is an open field where the customer can type out a personalised return reason. Based on the type of data found in the return reason column, two distinct machine learning methods will be used. All the data where a selectable return reason has been clicked is separated from the “open field” returns and will be analysed using an XGBoost model. XGBoost was chosen as it is able to handle unbalanced data better than other machine learning models (Su et al,. 2018), while also maintaining strong levels of accuracy (Su et al,. 2018). Using the XGBoost model, the return reason will be predicted using a train and test sample of the data. The accuracy of the XGBoost prediction will be evaluated through the use of confusion matrices. Lastly, variable importance will be applied to attempt to uncover which factors of a customer or influencer are most important in predicting a return reason. Due to the “black box” nature of the XGBoost model, SHAP variable importance will be utilised in order to add more interpretability to the XGBoost model. When the return reason is unstructured, text analytics in combination with machine learning will be used. First, a PCA analysis will be performed to identify principal components within the text. The next step will be to perform sentiment analysis on the return reasons in order to gauge the overall sentiment within the reviews. The last step will be to use a Latent Dirichlet Allocation

(LDA) in order to perform topic modelling which will identify latent topics within the text. The results that the models yield will be transformed into insights and a product return framework, which will ultimately answer the central research question of the thesis.

### 3.4 Possible Research Bias

Due to the data being fully anonymized, there are no avenues of possible researcher bias.

## Chapter 4: Results

### 4.1 Results

### General Metrics

The first results that will be investigated are the general exploratory data tables. These results are meant to provide an overview of the current situation surrounding the product returns and influencers.

Table 1 General Metrics describing the amounts and value of orders and returns

| **Metric** | **Value** |
|---|---|
| Total Order Value | 2,591,580,948 |
| Total Returns Value | 102,321,368 |
| Share of Value Attributed to Returns | 3.95% |
| Amount of Total Orders | 11,370,505 |
| Amount of Total Returns | 572,274 |
| Return Rate | 5.03% |

Table 1 shows the general metrics revolving around the value and share of returns compared to the total order value. As seen from table 1, returns make up around 3,95% of the total order value, totalling at around 102,000,000 Euros. Furthermore, the total amount of product returns are 572,274 orders which represent a 5.03% return rate based on 11,370,505 total orders. Table 1 provides an overview of how the ‘company’ is performing in relation to product returns, with the ‘company’ performing relatively well with product returns, as reports by Capital One (2024) highlight that the current retail return rate sits at around 16.6%.

Table 2 Amount of Influencers per Influencer Class

| **Influencer Class** | **Number of Influencers** |
|---|---|
| Mega Influencer | 225 |
| Macro Influencer | 3154 |
| Micro Influencer | 18420 |
| Nano Influencer | 44116 |
| Nano-Nano Influencer | 6513 |

Table 2 indicates the number of influencers per influencer class. The influencer class with the most number of influencers is the nano influencer class, which considers influencers with a following of lower than a 1000 people. As the bar of a 1000 followers is quite low it is expected that the Nano influencer class is the most populated. The least populated influencer class is the “Mega” influencer class, which contains influencers with 1 million or more followers.

Table 3 Average Return Rates per Influencer Class

| Influencer Class | Average Return Rate |
|---|---|
| Mega-influencer | 4.70% |
| Macro-influencer | 8.00% |
| Micro-influencer | 5.82% |
| Nano-influencer | 6.19% |
| Nano-nano-influencer | 6.68% |

Table 3 shows the average return rates per influencer class. Table 1.3 indicates that the class with the lowest rate of return is the “mega” influencer class. Mega influencers having the lowest return rate can relate to the ideas brought forward by McCormick (2016), Lou & Yuan (2019), and Wang et al. (2017) of influencer credibility being instrumental in consumers' purchasing behaviour. People buying from credible influencers are most likely buying products with a higher quality from more reputable sources, these products are therefore returned much less compared to products that may be promoted by other less credible influencers, leading to overall lower return rates. Mega influencers having the lowest return rate allows for the first hypothesis to be accepted. Further connecting to the idea of credibility, Table 3 highlights the fact that the Nano-nano influencers have the 2nd highest return rate of all influencers. Nano-nano influencers are influencers with a tiny following, which does not allow for credibility to be established, which in turn would turn off larger brands from promoting their products with nano-nano influencers. Another interesting observation from Table 1.3 is the 8% return rate for macro influencers. It could be expected that the return rate of macro influencers would be closer to the return rate of the mega influencers due to the idea of credibility being very important, however, the return rate for macro influencers is the highest of any of the influencer classes. A potential reason for the high return rate for macro influencers is the misalignment of influencers with company values and potential “vampire” effects. Firstly, compared to the 225 mega influencers there are relatively many more macro influencers as they number 3154. The larger number of macro influencers means that there are many more chances for the influencer to not match the ideals of the company. Wang et al, (2017) highlight that it is important for the brand to match the influencer as this will lead to a positive view of the brand and stronger purchasing editions. Furthermore, looking at the first sub-question, the “vampire” effect is mentioned. Erfgen at al, (2015), explains the vampire effect as when an influencer overshadows the company they are promoting for, which ends up in lower brand recall which could subsequently lead to more people returning products that they have purchased from the brand. Due to the fact that macro influencers still have a large following it is possible that some influencers in the upper end of followers may be overshadowing the brands they are promoting for and lead to the vampire effect upping the return rate.

The next analysis will look at a simple linear regression with **return rate** being the dependent variable.

Table 4 Linear Regression Results

| **Coefficients** | **Estimate** | **Pr(>\|t\|)** |
|---|---|---|
| Intercept | 7.663e-01 | <2e-16 *** |
| Quantity | 4.623e-02 | <2e-16 *** |
| Value | -7.875e-06 | 0.70 |
| Price | -2.205e-04 | 4.11e-09*** |
| Price Discount | 1.756e-04 | 1.31e-06*** |
| Gender: Female | 4.797e-02 | 5.73e-07 *** |
| Gender: Male | 2.710e-02 | 0.28 |
| Gender: Unknown | 9.724e-02 | 5.92e-05*** |
| Number of Orders | -2.016e-02 | <2e-16 *** |
| Posts | 1.984e-07 | 0.95 |
| Follower | -4.643e-08 | 0.001** |
| Following | -2.705e-05 | 0.0003*** |
| Influencer Class: Mega Influencer | 6.461e-03 | 0.76 |
| Influencer Class: Micro Influencer | -9.667e-02 | 2.30e-14*** |
| Influencer Class: Nano Influencer | -6.084e-02 | 0.0001*** |
| Influencer Class: Nano-Nano Influencer | -6.861e-02 | 0.009** |
| Return Reason: Changed Mind | 4.384e-02 | 0.10 |
| Return Reason: Does Not Match Description | -7.540e-02 | 0.007** |
| Return Reason: Does Not Meet Expectations | 1.086e-01 | 0.03* |
| Return Reason: Incorrect Order | -8.202e-03 | 0.72 |

| | | |
|---|---|---|
| Return Reason: No Reason Mentioned | 1.032e-01 | 8.12e-08*** |

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 4 shows the results of the linear regression on return rate. The regression results can be interpreted by looking at the estimate of columns where the result is significant. An example is the “Value” column, if the value of a product went up by 1 unit, the return rate would drop by 7.875e-06, keeping all other columns constant. Looking at Table 4 in more detail,  Griffis et al. (2012) mentions that the loyalty of a customer is a determining factor for returns, therefore, if “number of orders” is taken as a proxy for customer loyalty, the regression results in table 4 confirm that loyalty is a significant determinant of the return rate, with an increase in the number of orders lowering the return rate. Furthemore, an interesting result is that being Female is significant at a higher level compared to being Male. The discrepancy in the results for the genders links back to the work by Powers & Jack (2013) who state that gender plays an important role in product returns and state that males have a weaker relationship with liberal return policies compared to women. A surprising result is that the number of posts an influencer has is not significant while the amount of followers and following of the influencer are significant. The regression results seem to not align fully with the findings from De Veirman et al. (2016) which indicated that the amount of posts, followers, and following are all very important dynamics for an influencer and are significant drivers for product returns, while Table 4 only highlights follower and following to be significant. Looking at the influencer class, the classes that are significant include Micro, Nano, and Nano-Nano influencers; a potential reason for these classes being most significant in determining returns is the fact that most influencers in the data fall into these influencer categories. Lastly, looking at the return reasons, the three reasons that are significant are: does not meet description / expectation, and no return reason mentioned.  A potential explanation for “does not meet description / expectation” to be significant compared to the rest of the reasons relates to the work by  Lv & Liu (2022) regarding “information overload”. Information overload can be linked to the return reasons of “does not meet description / expectations” as consumers may be overloaded by information and may therefore not properly read or understand the description of a product and may have warped expectations of the product they are buying. The misalignment of descriptions and expectations may then lead to consumers returning their purchased products.

**XGBoost Model**

The next avenue of analysis involves an XGBoost model. XGBoost stands for Extreme Gradient Boosting, and is a supervised machine learning model that employs the concept of boosting to transform weak learners into strong learners. The influencer data is grouped into a testing and training set, while 5 fold cross-validation was conducted on a 100 rounds in order to determine the hyperparameters to be used in the model. The target variable (Return Reason) is converted to a numerical variable with the following classification:

Table 5 Target Variable Classification

| | |
|---|---|
| “Another Reason” | Class 1 |
| “Changed Mind” | Class 2 |
| “Does Not Match Description” | Class 3 |
| “Does Not Meet Expectations” | Class 4 |
| “Incorrect Order” | Class 5 |
| “No Return Reason Mentioned” | Class 6 |

To answer Subquestion 2, The most frequently mentioned class excluding “No Return Reason Mentioned” is “Incorrect Order” followed by “Another Reason”, “Changed Mind”, “Does Not Match Description” and lastly “Does not meet expectations”.

Running the XGBoost model resulted in the following accuracy:

Table 6: XGBoost Metrics

| | |
|---|---|
| Accuracy | 0.88 |
| 95% Confidence Interval | (0.86, 0.91) |
| No Information Rate | 0.78 |

| | |
|---|---|
| P-Value [Acc > NIR] | 2.6e-12 |

Table 6 highlights that the XGBoost model is performing strongly. Firstly, the accuracy of the model sits at 88% which implies that the model predicted the correct class 88% of the time. The 95% confidence interval indicates the range of the true accuracy value with a confidence of 95%. However, in order to benchmark the accuracy value it needs to be compared to the No Information Rate (NIR). The NIR indicates the accuracy of the model if all cases were classified as the most frequent class, which in the case of the research would be “No Return Reason Mentioned”. The NIR lies at 78%, which illuminates the fact that the XGBoost model performs 10 percentage points higher than the NIR. A t-test is conducted to test if the greater accuracy of the XGBoost model is statistically significant. The t-test results in a P-value of 2.6e-12, indicating that the result is indeed significant, meaning that the XGBoost model is a significant improvement over the NIR. Overall, Table 6 shows that the XGBoost model is able to classify return reasons with high accuracy and is a capable model to employ when dealing with the data, therefore answering the third subquestion.

The XGBoost model is a black box model which can be expanded upon in order to aid interpretation of the model results. The first set of techniques that are utilised are the variable importance graphs. The variable importance graphs indicate which variables are the most important in classifying a return reason.

Figure 1: Full Model Importance

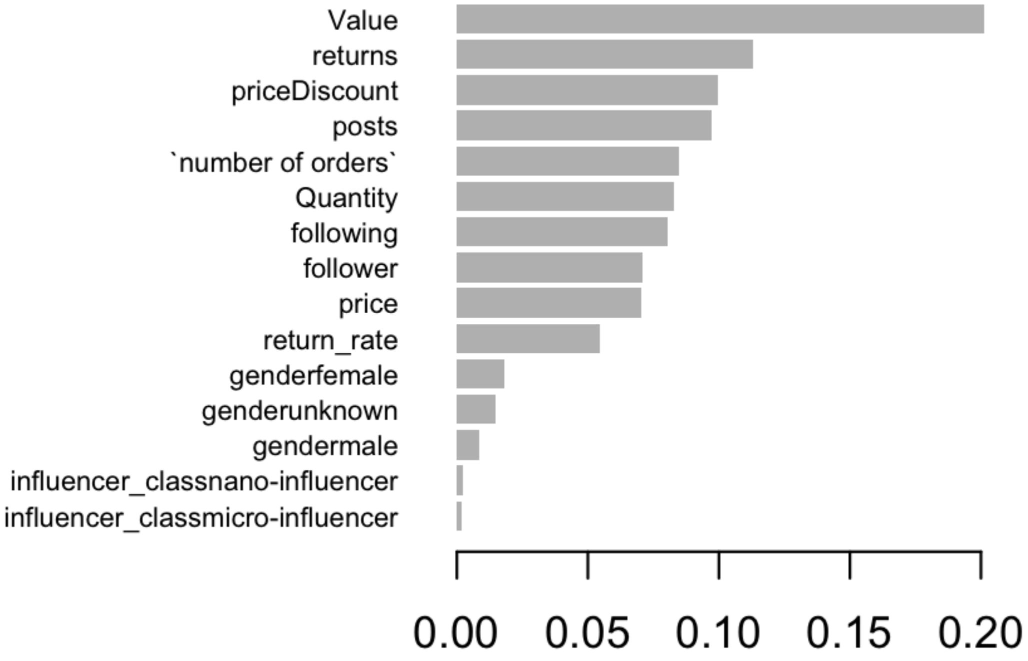


Figure 1 highlights the variable importance for each variable in the XGBoost model. Figure 1 is constructed using the built-in importance function in the XGBoost package. The XGBoost package calculates variable importance based on "gain", which represents the gain in accuracy when the feature is included in the model. Based on Figure 1, the most important two variables are the "Value" and the amount of returns. The model indicates that the value of an item is the most important in determining returns which is an important insight, as companies will need to take the value of a product into account when determining their returns policy. The importance of the Value is also almost double the size of the next most important variable showing that it is the undisputed most important variable in the model. The amount of returns is the next most important variable in predicting the reason of return, which links back to the work by Griffis et al. (2012) which states that having returned a lot of items previously can be an indicator for future returns. An interesting finding in Figure 1 is that the amount of posts an influencer has is more important in predicting the return reason than the amount of people followed by the influencer or the amount of followers that the influencer has. The importance of the posts variable can be an indicator to the fact that the credibility of an influencer is represented to customers based on the amount of posts an influencer has instead of the amount of followers that the influencer has. The amount of posts could be a proxy for how involved the influencer is with their follower base and how active the influencer interacts with social media. Hence, posts can then be a more appropriate measure of trust towards an influencer as consumers may see an influencer with a lot of posts as more trustworthy than an influencer who has a huge fan base, but does not interact with the fanbase and posts infrequently. Figure 1 also indicates that the influencer class does not seem to play an important role in predicting a return reason. The only influencer classes that show up in Figure 1 are “nano” and “micro” influencers, and the importance of these variables is minimal compared to the other variables in the model. Influencer classes having low importance is an interesting result, as the amount of followers has a significantly higher importance value, while influencer class is based on the number of followers an influencer has. A potential reason for the wide gap between importance of influencer class and followers can be that the amount of followers is a measure that captures the actual amount of followers an influencer has and therefore captures differences in follower count that the influencer class may not pick up on, as two influencers with a similar number of followers will be grouped into the same class.

Due to the return reason compromise of six different classes, the variable importance interpretation can be expanded to include the variable importance of each variable for every respective return reason. In order to highlight the variable importance per return reason, SHAP importance will be employed.

Figure 2: SHAP Value Importance per Class

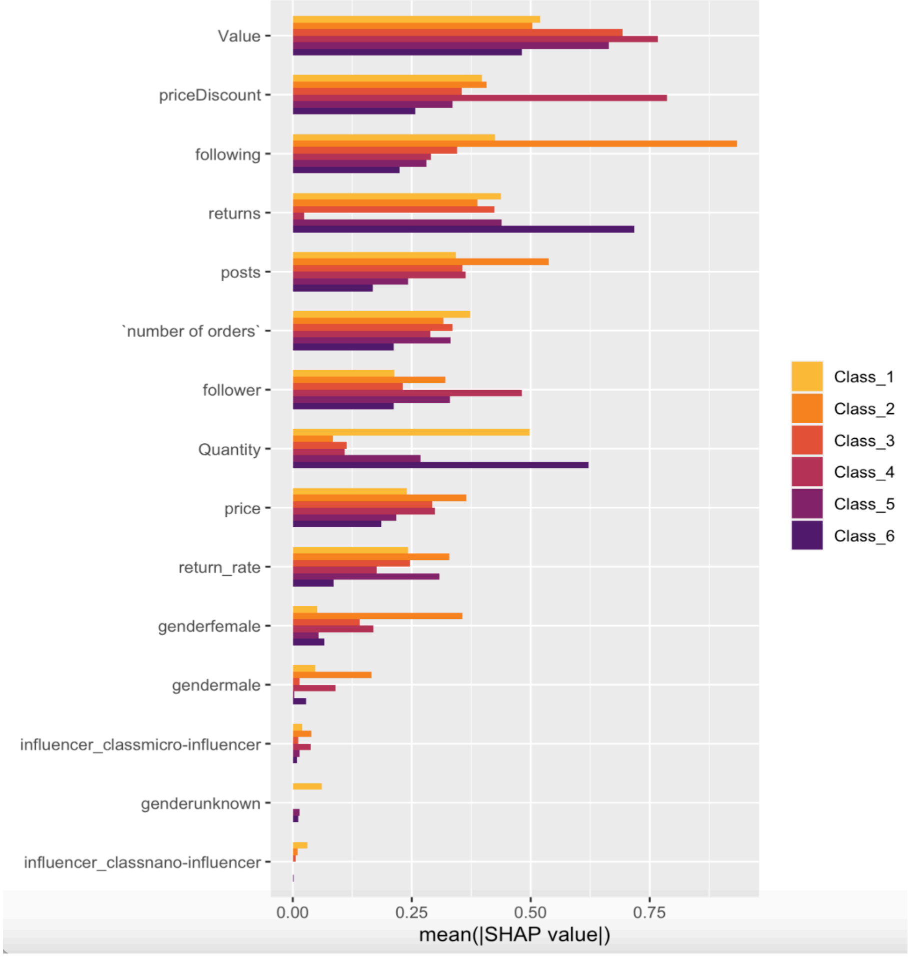


Figure 2 showcases the variable importance for each of the six return reasons. The variable importance in Figure 2 is calculated based on the mean of the absolute value of the SHAP value. The SHAP value is computed by utilising game theory to determine how much each variable contributes to the prediction that the model outputs. Therefore, Figure 2 is able to illuminate the differences between outcomes and the different ways that variables may impact a return reason. Firstly, looking at the importance of “Value”, which has the highest level of importance for the return reasons of “Does Not Match Description”, “Does not Match Expectations", and “Incorrect Order”. The results of the importance for Value do seem to make logical sense, as a consumer would find the value of an item quite important in the decision to return it if they had purchased the order without understanding the description or expectations, or if they had incorrectly purchased the item. The importance of the "Price Discount" variable also stands out, as the importance is driven by the "Does not meet expectations" return reason. The reasoning for the importance to be driven by "Does not meet expectations" could be due to the fact that many consumers may have only purchased the product because of the discount that had been placed on the price, and these consumers have realised that the product actually does not match their original expectations. The insight for retail companies should be to exercise caution on which products are discounted, as it may cause a significant amount of said products to be returned due to consumers only looking at the price and then being disappointed with the product. Looking at the "following" variable, another interesting observation can be seen as the importance is driven by the "Changed Mind" return reason. A potential explanation for the "following" variable importance is that the amount of people an influencer follows is a factor that determines a consumer's trust regarding the influencer, hence, the customer may be swayed by the perceived trust of the influencer causing them to change their mind about their purchase. Next, the variable "returns (amount)" has very low importance regarding the "Does not meet expectations" class. The reason for the low importance regarding amount of returns, could be that no amount of previous returns will affect the customer if the product is unable to meet the expectations of the customer. It is therefore of paramount importance that a retail company is able to have a gauge on customer expectations in order to reduce the amount of returns. Sticking with the "returns (amount)" variable, the importance is highest for the "No Return Reason Mentioned" reason, which implies that consumers who have returned a lot in the past may not be incentivized to indicate a return reason due to the amount of times they have returned an item. The next variable of interest is the "posts" variable which is the most important when predicting the "Changed Mind" return reason. The "posts" variable could again be a measure of influencer credibility,implying that a perceived lack of credibility can cause consumers to change their mind about a product that they have bought due to influencer promotion, as they do not invoke enough trust in the credibility of the influencer, subsequently not trusting the product as the right purchase. The results of the "posts" and "following" variables in

Figure 2 further reinforce the importance for companies to choose credible influencers that will lead to consumers trusting the products they are purchasing and not change their mind about a purchase and return it. The next variable of interest is the "follower" variable which has the most importance for the "Does Not meet Expectations" return reason. A potential explanation for the importance of the "follower" variable is that consumers may expect a lot from products that are endorsed by high profile influencers that have a large amount of followers. However, once these products are purchased, the consumers realise that the product does not match the expectations that may have been perceived when the item was promoted by the influencer. Due to the discrepancy between expectations, consumers may want to return the product. Lasty, the "Quantity" variable is most important in predicting the "Another Reason" and "No Return Reason Mentioned", which could imply that consumers that return a large quantity of items do not put down a specific reason for why the product was returned. A reason why consumers may not put down specific reasons for returns with a high quantity could be that multiple reasons are actually at play as they are returning more than a single product. Retail companies should therefore review the returns process and investigate how to optimise the return reason selection when multiple products are returned.

The last interpretation technique that will be utilised will be zooming in on the decision trees that make up the XGBoost model. XGBoost is a combination of decision trees, however in order to aid interpretation of the model individual decision trees can be shown to highlight how the model is able to predict a result.

Figure 3: Decision Tree (0) from XGBoost model

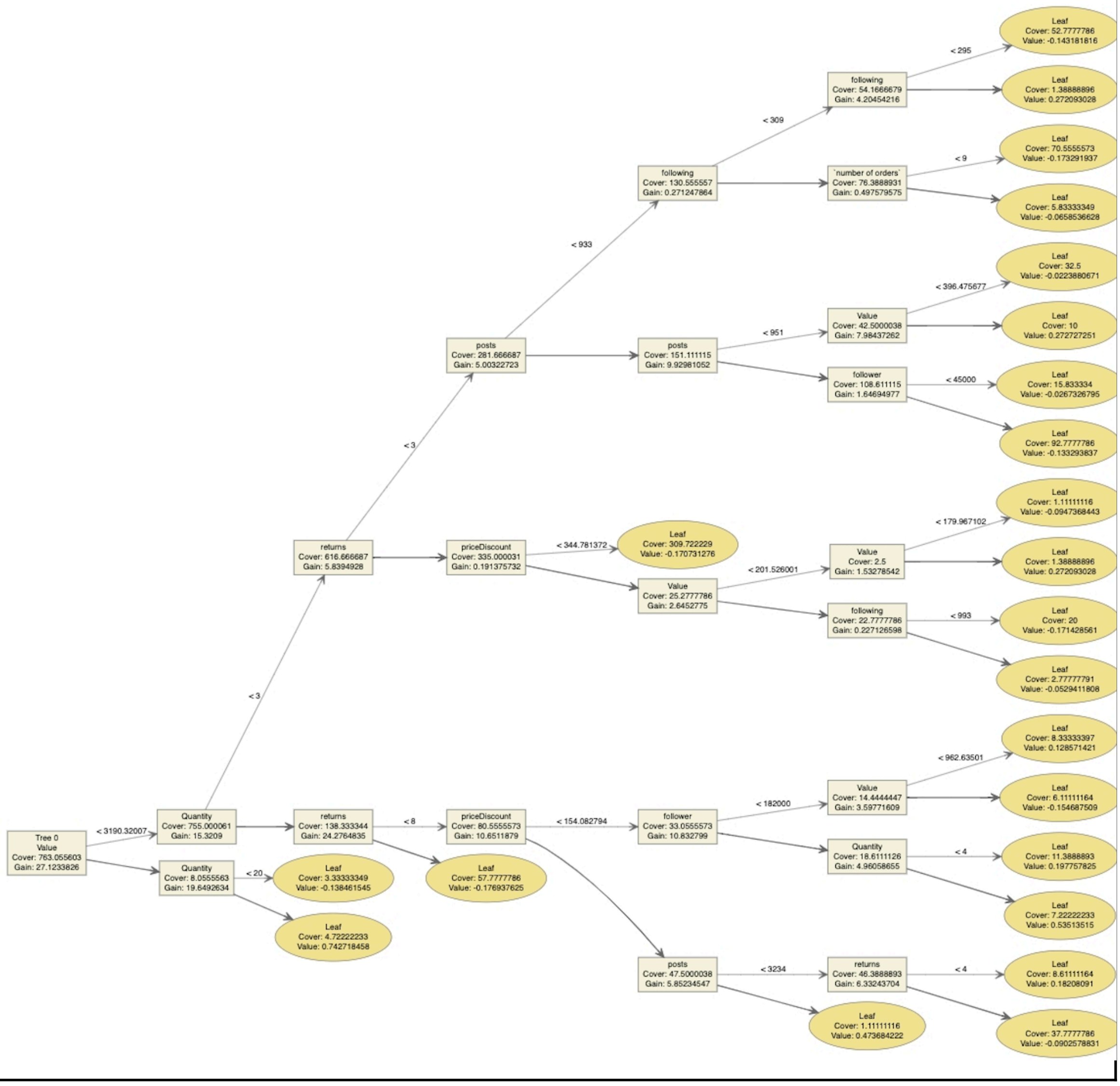


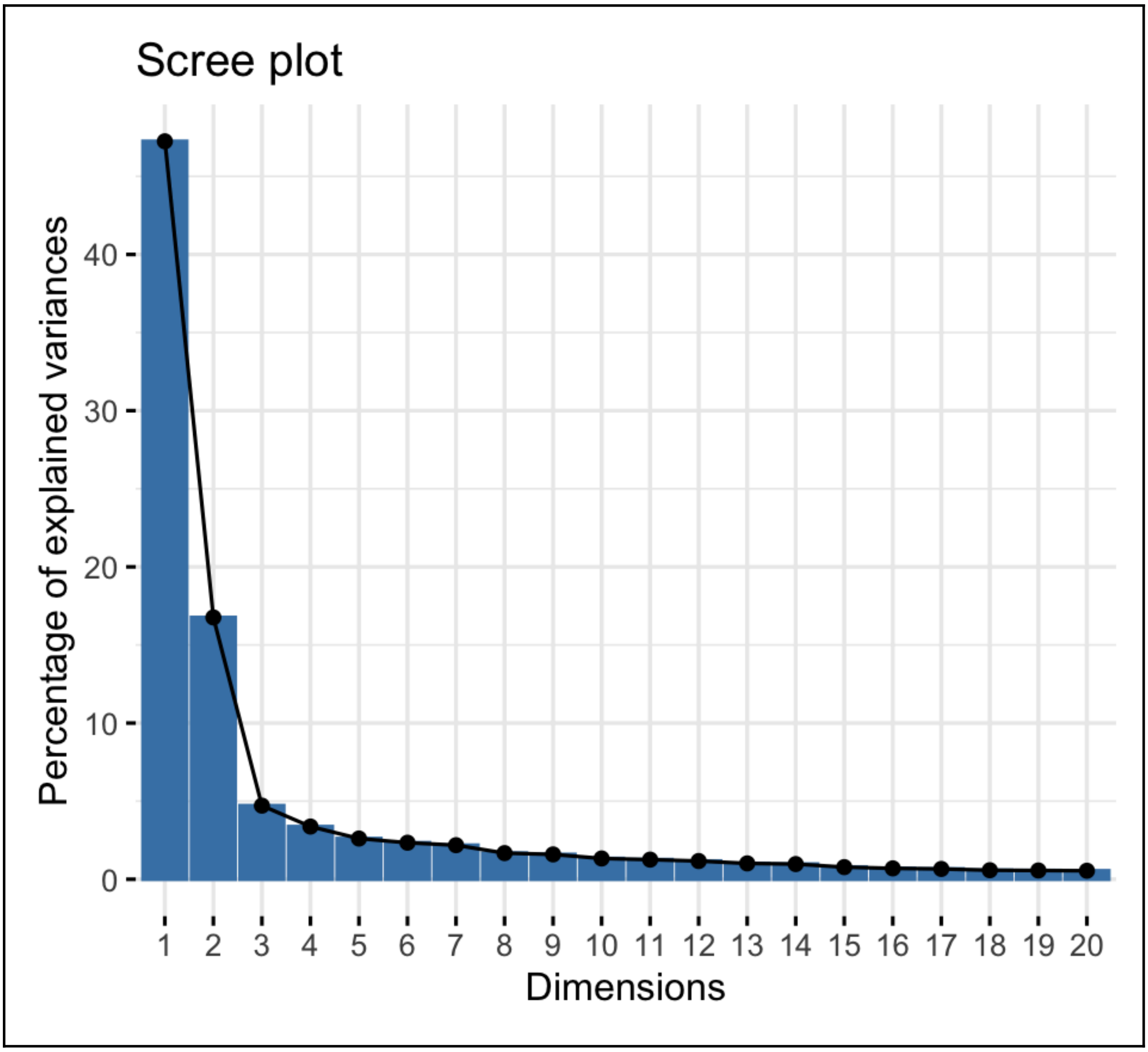
Figure 3 highlights the first decision tree used in the XGBoost model. Figure 3 aims to highlight how the XGBoost model is able to make predictions by indicating the decisions the model takes to arrive at every leaf in the tree. Every tree in the leaf has a value which indicates if the leaf increases / decreases the predicted probability of a class, and a cover value which indicates the weight of observations that fall into each leaf. The decision tree shown in Figure 3 is the “0” tree which is the first tree generated by the model.

**Text Analytics**

The following section will look into the open returns where customers were able to type out a custom reason as to why a product was returned. In total 222 reviews fall in the "open" category. The following section will apply Principal Component Analysis (PCA), Sentiment Analysis, and LDA topic modelling in order to uncover the customer sentiment surrounding the products they are returning.

The first model that will be presented is a PCA model. The first step in a PCA model is to determine the number of components that will be used in the model. In order to determine the number of components a scree plot is used.

Figure 4: Scree Plot for PCA Analysis



Based on Figure 4, the point at which the percentage of explained variance starts to drop drastically (the elbow of the graph) is at around the second dimension. Therefore, two dimensions will be selected for the PCA analysis.

Table 7: Top Words per Component

| Component 1 | | Component 2 | |
|---|---|---|---|
| Word | Loading | Word | Loading |
| Holder | 1.39 | Charge | -1.24 |
| Card | 1.39 | Doesn't | -0.74 |
| Mount | 1.38 | Card | 0.67 |

| | | | |
|---|---|---|---|
| Vent | 1.38 | Holder | 0.67 |
| Car | 1.38 | Mount | 0.66 |
| Phone | 0.7 | Vent | 0.66 |
| iPhone | 0.7 | Car | 0.66 |
| Fall | 0.69 | Charger | -0.63 |
| Attach | 0.69 | Qi | -0.62 |
| Easy | 0.69 | Match | -0.6 |

The results from Table 7 are quite mixed. Certain words like “Card”, “Holder”, and “Vent” appear in both components, however the loadings indicate two main points of interest. Firstly, component 1 seems to relate to phones and how they can be mounted in cars, while component 2 seems to load negatively on words relating to chargers. Table 7 seems to indicate that people are writing reviews that mention phone holders and chargers, which may be items that are frequently returned by customers. For the case of phone holders the issue seems to lie around the idea of mounting the phone within the holder and the fact that the phone is falling out, while for the case of chargers the issue lies in the quality of the charger which may be faulty. Overall, Table 7 allows for an overview of some of the specific reasoning that customers have when they elect to return products.

In order to gain a more detailed insight of the consumer sentiment regarding product returns, sentiment analysis will be conducted.

Figure 5: Most frequent negative and positive words based on ratio of word frequency

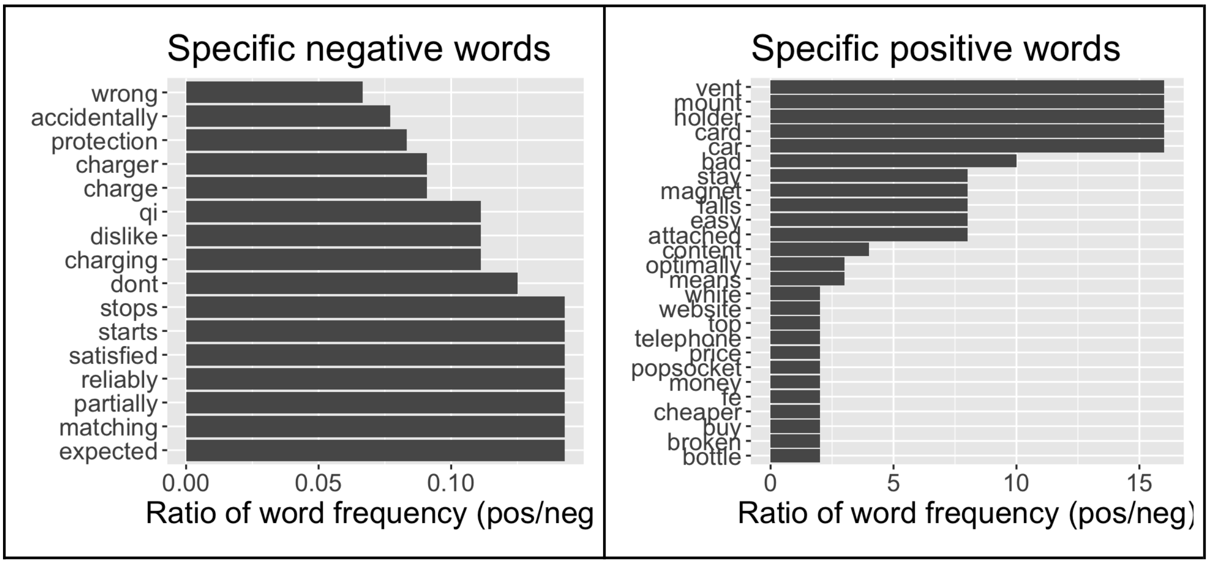


Figure 5 indicates the most frequent negative and positive words based on the ratio of the word frequency (positive / negative). The most negative words seem to relate the most to accidents / wrong orders and the issue of charging. The most positive words seem to relate to mounts and holders in cars. Some key insights that can be extracted from Figure 5 is that consumers explicitly mention when a product is not up to quality standard, which is being done with the issue of chargers. In addition the top two most frequent negative words are "wrong" and "accidentally" which could be referring to accidental orders that are being placed. It is therefore important to investigate orders that are marked as incorrect in order to have a better understanding of what is causing customers to purchase incorrect orders, in order to further lower the amount of product returns.

Lastly, to get more insights on the topics that are being mentioned in the open reviews, LDA topic modelling was performed to garner more insights on the topics being mentioned in the open reviews.

Figure 6: Top 10 Terms for LDA Topics

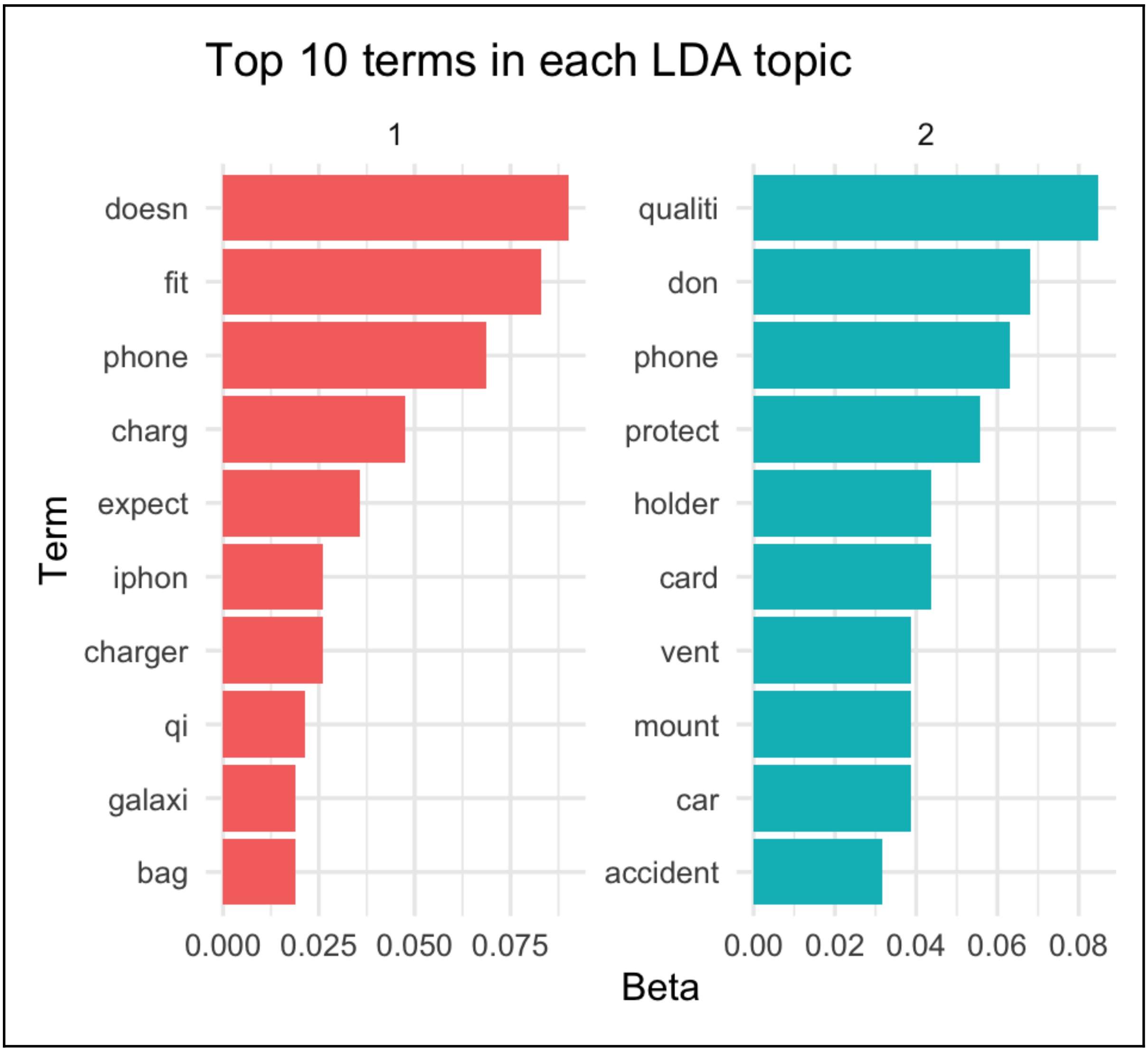


Figure 6 indicates the top 10 most frequent terms in each LDA topic, using Figure 6 the two LDA topics can be classified and described. 2 LDA topics were eventually chosen as they resulted in the most interpretable topics. Figure 6 adds more context to the previous PCA analysis as Topic 1 seems to relate to the idea of charging. However, the most frequent word is “doesn’t” which may indicate that consumers are complaining that their iPhones and Samsung Galaxy phones are not charging and do not fit with the accessories that they are buying. Topic 1 seems to relate to the return reason of “Does not meet expectations” from the XGBoost model discussed earlier, as consumers are specifically mentioning quality issues of products they are buying. Topic 2 seems to focus on the quality of phone holders. Frequent words mentioned in Topic 2 include “holder”, “mount”, “car”, and “protect” which signifies that these consumers may have returned their orders due to the lack of quality of the car holders/mounts they have bought. Once again the main issue that consumers are expressing is the functionality of products that are being bought which do not match the expectations that consumers have of these products. It is therefore important to ensure that influencers are promoting products that are of high quality and do not risk this level of discrepancy in customer expectations and the real functionality of the product. The overall sentiment of the open reviews relates to complaints consumers have with the products they are returning, answering subquestion 4.

## Chapter 5: Conclusions and Recommendations

Chapter 5 will aim to compare the key findings of the results section to the key findings of the literature review, review the hypotheses and subquestions, and ultimately answer the central research question. Lastly, further recommendations and limitations of the research will be discussed.

### 5.1 Comparison of Key Findings

Looking back at the literature review there were a selection of key findings. Firstly, machine learning methods are seen as accurate models that are appropriate to be employed for understanding customer returns data (Joshi et al, 2018), (Li et al, 2018), Cui et al, (2020), Purba & Tan (2023). While machine learning models are seen as accurate, environmental assessment methods like LCA and MFA are less sophisticated and currently do not adequately account for the impact of product returns on sustainability (Hagelaar, 2001), (Laner, 2014). Zooming in on the actual product returns, the key findings were that certain aspects of influencers affected whether or not customers would trust the product that they are purchasing. The main aspect of influencers arising from the literature is the aspect of trust, with trust determining whether or not an influencer promotion will be successful (Lou & Yuan, 2019), Trivedi & Sama (2019), McCormick (2016). While influencer credibility is a main factor, other factors like gender, follower counts, and information overload all play a part in the decision whether a customer will return a product (Stephen, 2016), (De Veirman et al., 2016), (Powers & Jack , 2013), (Griffis et al., 2012).

Next, the key findings of the results chapter will be discussed. First, looking at the results of the XGBoost model, the accuracy of the model is deemed high and significant, showing that the use of a machine learning model was an appropriate approach to predict the reason for a product return. Further zooming in on the actual return reasons, it is clear that certain aspects of influencers play an important part in determining whether a product is returned or not. According to Figure 2 it is clear that the amount of people an influencer is following seems to be important in predictions where the return reason is “Changed mind”. In addition, Figure 2 indicates that the amount of posts and followers of an influencer were also of high importance in predicting the return reasons of “Changed mind” and “Does not meet expectations”. In addition to aspects surrounding the influencers itself, factors like the value of the products being returned and whether products are discounted are of the most importance in predicting whether a product is returned. Looking at the Text analytics section of the analysis, the main point of discussion that emerges is the apparent lack of quality in certain products that are being sold. Consumers feel the will to write out custom return reasons to express the issues with the quality of products that they have bought, further emphasising the importance of the “Does not meet expectations” return reason.

Evaluating both sets of key findings uncovers various similarities. Mainly that influencer aspects like followers, following, and number of posts are important in predicting the reason for a product return. However, some of the main differences are that the gender of consumers does not play a large role in predicting return reasons, the value and discounts of an item seem to be the most important factors in determining the return reason, and that the return reason of “does not meet expectations” seems to be the most prevalent return reason.

**5.2: Hypothesis Review**

This section will aim to reject or accept the hypotheses that were constructed before the analysis stage of the research.

**Hypothesis 1: Mega Influencers will have the lowest effect on returns.**

The hypothesis can be accepted. Firstly, mega influencers had the lowest return rate of any of the influencer classes (see Table 3) highlighting that customers that purchase goods promoted by mega influencers return them less frequently than any of the other influencer classes. Furthermore, within the XGBoost model, mega influencers do not show up as a contributing factor in determining the return rate, which could indicate that there are not enough instances of returned goods being promoted by Mega influencers, and therefore they are not significant in predicting the return reason. Overall, the results of the thesis correspond to work by (Lou & Yuan, 2019), Trivedi & Sama (2019), McCormick (2016) regarding the importance of influencer credibility in keeping the amount of returns low.

**Hypothesis 2: The return reason with the highest importance in predictions will be “did not meet expectations”**

The hypothesis can be accepted. The idea behind the hypothesis was based on the work by Powers & Jack (2015), which mentions that customers feel emotional dissonance due to the products they are purchasing not matching the expectations. Based on the results of Figure 2 and 6 it seems that the return reasons of  “Changed mind” and “does not meet expectations” are the most important. Figure 2 shows that the return reasons of “changed mind” and “does not meet expectations” have  the highest importance for the most amount of variables, while Figure 6 further highlights that the open return reasons are centred around the issue of unexpected poor quality, which relates to the return reason of “does not meet expectations”. In addition, Powers & Jack (2015) highlights that the feeling of emotional dissonance is stronger for males compared to females, which can be seen in Figure 2 as the “does not meet expectations” return reason has a higher importance for females compared to males.

**Hypothesis 3: The sentiment of the "open" return reasons will be product centred**

The hypothesis can be accepted. Based on the results in the text analytics section, it is clear that the open return reasons were centred around two of the products being sold and the apparent lack of quality and compatibility these products had. The open reviews allowed for a detailed view into the consumers mind regarding the specific characteristics of products that cause them to be returned.

**Hypothesis 4: The amount of followers an influencer has will be the most important influencer feature in predicting return reasons**

The hypothesis can be rejected. While the idea of influencer credibility on the whole is shown to be a driving factor in determining the reason of a product return, the feature of influencers that acted as a proxy for credibility turned out to be the amount of people an influencer themselves follow. Looking at Figure 2, the variables of following and even posts rank higher in terms of influence on the return reason than the amount of followers the influencer has. It seems that posts signify how active the influencer is on social media which allows for higher trust in the influencer, while the amount of following the influencer has represents the amount of credibility an influencer invokes on the customers.

## 5.3 Central Research Question

Looking back, the central research question of the thesis is:

**To what extent can machine learning models promote sustainable influencer marketing**

In order to evaluate the central research question, an XGBoost model combined with the text analytical tools of PCA, Sentiment Analysis, and LDA topic modelling were employed to evaluate how machine learning can aid in promoting sustainable influencer marketing. Based on the findings from Chapter 4, the central research question can be answered. Overall, machine learning allows for a detailed overview on which factors regarding influencers are important in minimising the amount of returned products, while machine learning models within text analytics are able to uncover specific product features that drive consumers to return products. However, in order to properly make use of the results, a framework needs to be established that can guide companies to make an effective choice of influencers that promote their products, in order to lower the amount of product returns and ultimately exercise more sustainable business practices.

Figure 7: Framework for Sustainable Influencer Product Promotion

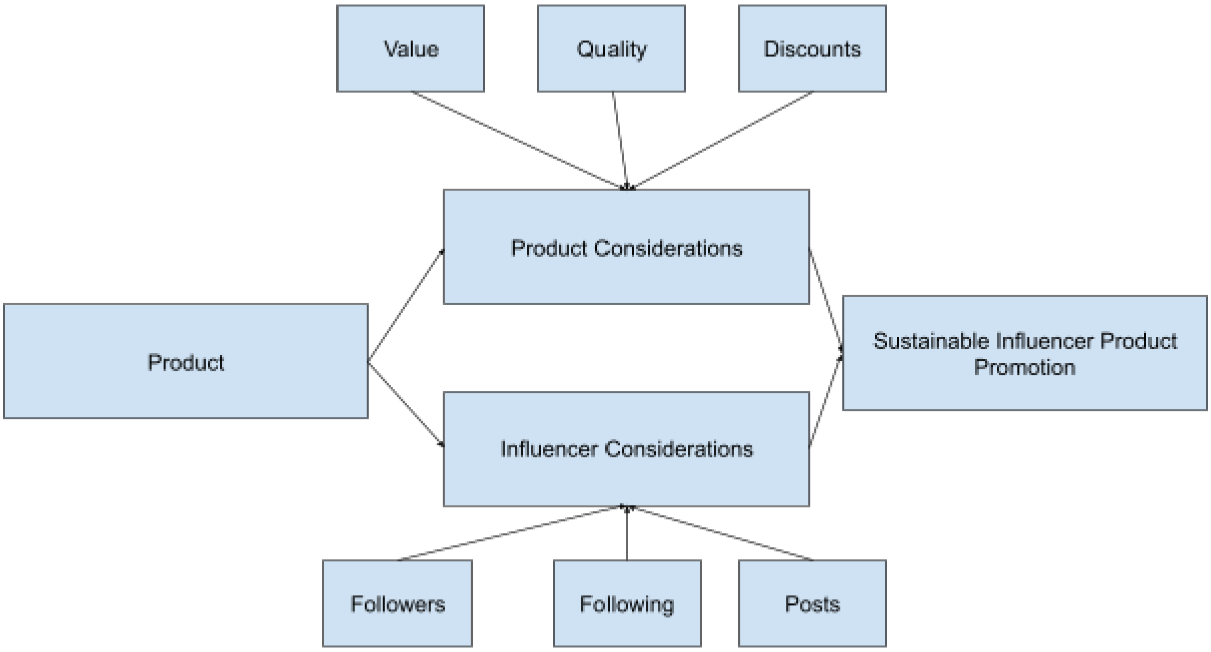


Figure 7 showcases a framework that companies can make use of in order to select influencers that will minimise the amount of product returns and lead to a more sustainable promotion of products. Firstly, the product needs to go through the product considerations which include the value, quality, and discounts. As seen from the results section the value of a product is the most important reason in predicting the return reason, which implies that companies need to decide whether high value products will be promoted by influencers as consumers may easily make the decision to return those products compared to lower value products which may not be a priority for return in the eyes of a consumer. Furthermore, the text analytics section of the results indicated that the quality of a product is very important for returns, as the open return reasons were centred around the lack of quality of the products being returned. Therefore, if a company decides to promote a product with an influencer, they need to be wary of the fact that the product having low quality can easily lead to higher returns. Lastly, discounts need to be monitored carefully. Figure 2 highlighted that discounts are very important in determining the return reason, which can imply that consumers are just purchasing the product due to the discount and then returning the product when they realise it does not suit their needs. Therefore, companies need to be careful when applying discounts as it can lead to a significant amount of future product returns. Next, the influencer considerations need to be applied. As seen from Figure 2, the influencer features of posts, follower, and following are able to represent the trust people put in an influencer, and consumers having the trust of influencers they are purchasing from aids in lowering the amount of returns. Therefore, companies need to approach influencers with a large follower base that are frequent posters on social media, as these types of influencers are going to garner the most amount of trust from consumers. Combining the product and influencer considerations will allow companies to promote products with a certain set of influencers to actively reduce the rate of product returns and achieve more sustainable business practices.

### 5.4 Recommendations for Future Research

The main avenue for future research is to zoom in on the product level and aim to determine specific product features that drive specific return reasons, and which types of products are most likely to be returned. There is a plethora of different types of products and being able to narrow down the specific products that are most likely going to result in a certain return is beneficial information for any retail company. Further analysis could include taking shopping baskets of customers and predicting the specific products that are most likely going to be returned.

### 5.5 Limitations

The primary limitation of the research related to the limited amount of observations where a customer purchased a good with a specific influencer discount code. Most of the observations did not have any entered discount codes which did not allow for the observation to be connected to a specific influencer. If more observations were present then the machine learning models could be further optimised to produce more accurate results that may have given additional insights that were not discovered by the models used in the thesis. It is likely that a significant amount of people who were actually influenced by an influencer either forgot or did not use the influencer discount code, leading to a group of consumers not incorporated into the research while they should have been.

**Appendix A: Works Cited**

Average retail return rate (2024 data): ECommerce VS in-store. Capital One Shopping. (2024, February 14). https://capitaloneshopping.com/research/average-retail-return-rate/

Aydin, R., Brown, A., Badurdeen, F., Li, W., Rouch, K. E., & Jawahir, I. S. (2018). Quantifying impacts of product return uncertainty on economic and environmental performances of product configuration design. Journal of Manufacturing Systems, 48, 3–11. https://doi.org/10.1016/j.jmsy.2018.04.009

Bernon, M., & Cullen, J. (2007). An integrated approach to managing reverse logistics. International Journal of Logistics Research and Applications, 10(1), 41–56. https://doi.org/10.1080/13675560600717763

Bertram, R. F., & Chi, T. (2017). A study of companies' business responses to fashion e-commerce's environmental impact. International Journal of Fashion Design, Technology and Education, 11(2), 254–264. https://doi.org/10.1080/17543266.2017.1406541

Booth, N., & Matic, J. A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions. Corporate Communications: An International Journal, 16(3), 184–191. https://doi.org/10.1108/13563281111156853

Bush, A. J., Martin, C. A., & Bush, V. D. (2004). Sports celebrity influence on the behavioral intentions of generation Y. Journal of Advertising Research, 44(1), 108–118. https://doi.org/10.1017/s0021849904040206

Chaturvedi, H. (2022, May 30). Product returns are wasteful for companies and the planet. Here's how to change that. Fastcompany. https://www.fastcompany.com/90756025/product-returns-are-wasteful-for-companies-and-the-planet-heres-how-to-change-that

Cheng, C., Lau, Y., Chan, L., & Luk, J. W. (2021). Prevalence of social media addiction across 32 nations: Meta-analysis with subgroup analysis of classification schemes and cultural values. Addictive Behaviors, 117, 106845. https://doi.org/10.1016/j.addbeh.2021.106845

Cheng, H.-F., Krikon, E., & Murdock, V. (2024). Why do customers return products? using customer reviews to predict product return behaviors. Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval. https://doi.org/10.1145/3627508.3638326

Cui, H., Rajagopalan, S., & Ward, A. R. (2020). Predicting product return volume using Machine Learning Methods. European Journal of Operational Research, 281(3), 612–627. https://doi.org/10.1016/j.ejor.2019.05.046

Cullinane, S., Browne, M., Karlsson, E., & Wang, Y. (2019). Retail clothing returns: A review of key issues. Contemporary Operations and Logistics, 301–322. https://doi.org/10.1007/978-3-030-14493-7_16

Daniel, S. E., Tsoulfas, G. T., Pappis, C. P., & Rachaniotis, N. P. (2004). Aggregating and evaluating the results of different environmental impact assessment methods. Ecological Indicators, 4(2), 125–138. https://doi.org/10.1016/j.ecolind.2004.01.003

De Veirman, M., Cauberghe, V., & Hudders, L. (2017). Marketing through Instagram influencers: The impact of number of followers and product divergence on brand attitude. International Journal of Advertising, 36(5), 798–828. https://doi.org/10.1080/02650487.2017.1348035

Edwards, J. B., McKinnon, A. C., & Cullinane, S. L. (2010). Comparative analysis of the Carbon Footprints of conventional and online retailing. International Journal of Physical Distribution &amp; Logistics Management, 40(1/2), 103–123. https://doi.org/10.1108/09600031011018055

Erfgen, C., Zenker, S., & Sattler, H. (2015). The vampire effect: When do celebrity endorsers harm brand recall? International Journal of Research in Marketing, 32(2), 155–163. https://doi.org/10.1016/j.ijresmar.2014.12.002

Frei, R., Jack, L., & Brown, S. (2020). Product returns: A growing problem for business, Society and Environment. International Journal of Operations &amp; Production Management, 40(10), 1613–1621. https://doi.org/10.1108/ijopm-02-2020-0083

Gallagher, C., Furey, E., & Curran, K. (2019). The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying. International Journal of Data Warehousing and Mining, 15(4), 21–47. https://doi.org/10.4018/ijdwm.2019100102

Griffis, S. E., Rao, S., Goldsby, T. J., & Niranjan, T. T. (2012). The customer consequences of returns in online retailing: An empirical analysis. Journal of Operations Management, 30(4), 282–294. https://doi.org/10.1016/j.jom.2012.02.002

Hagelaar, G. (2001). Environmental Supply Chain Management: Using life cycle assessment to structure supply chains. The International Food and Agribusiness Management Review, 4(4), 399–412. https://doi.org/10.1016/s1096-7508(02)00068-x

Jack, L., Frei, R., & Krzyzaniak, S.-A. C. (2019). The Problems & Opportunities of E-Commerce Returns: Managing Returns as a Profit Centre.  Centre for Operational Research & Logistics Faculty of Business & Law.

Joshi, T., Mukherjee, A., & Ippadi, G. (2018). One size does not fit all: Predicting product returns in e-commerce platforms. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). https://doi.org/10.1109/asonam.2018.8508486

Kemp, Simon. "Digital 2022: July global statshot report." *DataReportal. Available online November* 30 (2022): 2022.

Kim, S. S., Lee, J., & Prideaux, B. (2014). Effect of celebrity endorsement on tourists' perception of corporate image, corporate credibility and corporate loyalty. International Journal of Hospitality Management, 37, 131–145. https://doi.org/10.1016/j.ijhm.2013.11.003

Laner, D., Rechberger, H., & Astrup, T. (2014). Systematic evaluation of uncertainty in material flow analysis. Journal of Industrial Ecology, 18(6), 859–870. https://doi.org/10.1111/jiec.12143

Li, J., He, J., & Zhu, Y. (2018). E-tail product return prediction via hypergraph-based local graph cut. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining. https://doi.org/10.1145/3219819.3219829

Lou, C., & Yuan, S. (2019). Influencer marketing: How message value and credibility affect consumer trust of branded content on social media. Journal of Interactive Advertising, 19(1), 58–73. https://doi.org/10.1080/15252019.2018.1533501

Lozano, R., & von Haartman, R. (2017). Reinforcing the holistic perspective of sustainability: Analysis of the importance of sustainability drivers in organizations. Corporate Social Responsibility and Environmental Management, 25(4), 508–522. https://doi.org/10.1002/csr.1475

Ma, J., & Kim, H. M. (2016). Predictive model selection for forecasting product returns. Journal of Mechanical Design, 138(5). https://doi.org/10.1115/1.4033086

Martínez-López, F. J., Anaya-Sánchez, R., Fernández Giordano, M., & Lopez-Lopez, D. (2020). Behind influencer marketing: Key marketing decisions and their effects on followers' responses. Journal of Marketing Management, 36(7–8), 579–607. https://doi.org/10.1080/0267257x.2020.1738525

McCormick, K. (2016). Celebrity endorsements: Influence of a product-endorser match on Millennials Attitudes and Purchase Intentions. Journal of Retailing and Consumer Services, 32, 39–45. https://doi.org/10.1016/j.jretconser.2016.05.012

Mulisa, F. (2021). When does a researcher choose a quantitative, qualitative, or mixed research approach? Interchange, 53(1), 113–131. https://doi.org/10.1007/s10780-021-09447-z

Pei, Z., & Paswan, A. (2017). Consumers' legitimate and opportunistic product return behaviors: An extended abstract. Marketing at the Confluence between Entertainment and Analytics, 1405–1408. https://doi.org/10.1007/978-3-319-47331-4_278

Petersen, J. A., & Kumar, V. (2009). Are product returns a necessary evil? antecedents and consequences. Journal of Marketing, 73(3), 35–51. https://doi.org/10.1509/jmkg.73.3.035

Powers, T. L., & Jack, E. P. (2013). The influence of cognitive dissonance on retail product returns. Psychology &amp; Marketing, 30(8), 724–735. https://doi.org/10.1002/mar.20640

Purba, K. R., & Tan, Y. J. (2023). Data-driven influencer marketing strategy analysis and prediction based on social media and Google Analytics data . Applied Marketing Analytics, 8(3), 314–328.

Robertson, T. S., Hamilton, R., & Jap, S. D. (2020). Many (un)happy returns? the changing nature of retail product returns and Future Research Directions. Journal of Retailing, 96(2), 172–177. https://doi.org/10.1016/j.jretai.2020.04.001

Sahoo, N., Dellarocas, C., & Srinivasan, S. (2018). The impact of online product reviews on product returns. Information Systems Research, 29(3), 723–738. https://doi.org/10.1287/isre.2017.0736

Shahbaznezhad, H., Dolan, R., & Tripathi, A. K. (2018). The power of Facebook and Instagram fans: An exploration of fan comments and their effect on social media content strategy. Lecture Notes in Business Information Processing, 109–117. https://doi.org/10.1007/978-3-319-99936-4_10

Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. Current Opinion in Psychology, 10, 17–21. https://doi.org/10.1016/j.copsyc.2015.10.016

Su, P., Liu, Y., & Song, X. (2018). Research on intrusion detection method based on improved smote and xgboost. Proceedings of the 8th International Conference on Communication and Network Security. https://doi.org/10.1145/3290480.3290505

Trivedi, J., & Sama, R. (2019). The effect of influencer marketing on consumers' brand admiration and online purchase intentions: An emerging market perspective. Journal of Internet Commerce, 19(1), 103–124. https://doi.org/10.1080/15332861.2019.1700741

Wang, S. W., Kao, G. H.-Y., & Ngamsiriudom, W. (2017). Consumers' attitude of endorser credibility, brand and intention with respect to celebrity endorsement of the airline sector. Journal of Air Transport Management, 60, 10–17. https://doi.org/10.1016/j.jairtraman.2016.12.007

Wiese, A., Kellner, J., Lietke, B., Toporowski, W., & Zielke, S. (2012). Sustainability in retailing – a summative content analysis. International Journal of Retail &amp; Distribution Management, 40(4), 318–335. https://doi.org/10.1108/09590551211211792

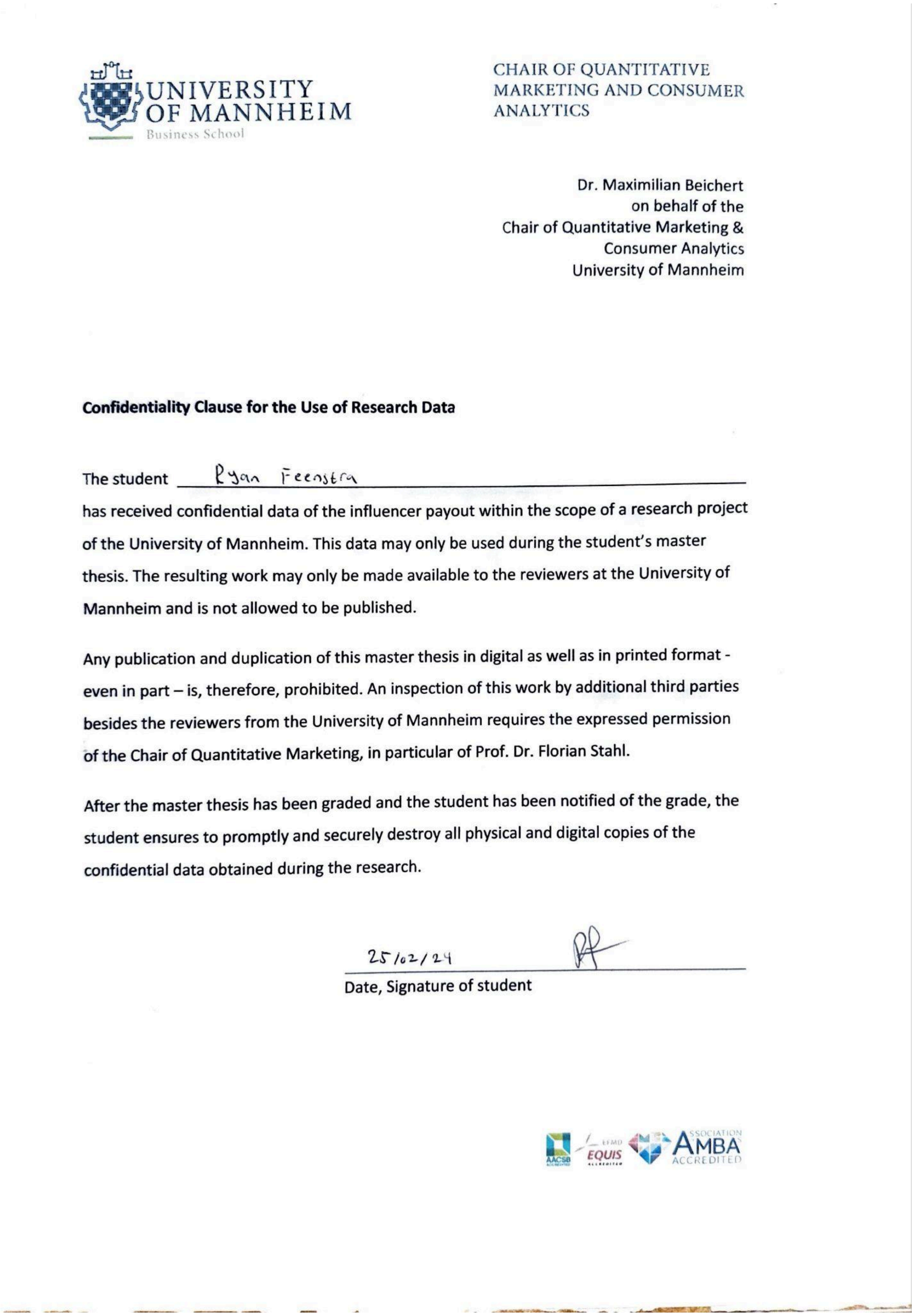
Zhang, D., Frei, R., Wills, G., Gerding, E., Bayer, S., & Senyo, P. K. (2023). Strategies and practices to reduce the ecological impact of Product returns: An environmental sustainability framework for multichannel retail. Business Strategy and the Environment, 32(7), 4636–4661. https://doi.org/10.1002/bse.3385

Zhang, Yuchi, Trusov, M., Stephen, A. T., & Jamal, Z. (2017). Online shopping and social media: Friends or foes? Journal of Marketing, 81(6), 24–41. https://doi.org/10.1509/jm.14.0344

Zhang, Yufei, Voorhees, C. M., Lin, C., Chiang, J., Hult, G. T. M., & Calantone, R. J. (2022). Information search and product returns across mobile and traditional online channels. Journal of Retailing, 98(2), 260–276. https://doi.org/10.1016/j.jretai.2021.05.001

Zhou, L., & Whitla, P. (2013). How negative celebrity publicity influences consumer attitudes: The mediating role of moral reputation. Journal of Business Research, 66(8), 1013–1020. https://doi.org/10.1016/j.jbusres.2011.12.025

## Appendix B: University of Mannheim Data NDA



UNIVERSITY
OF MANNHEIM
Business School

CHAIR OF QUANTITATIVE
MARKETING AND CONSUMER
ANALYTICS

Dr. Maximilian Beichert
on behalf of the
Chair of Quantitative Marketing &
Consumer Analytics
University of Mannheim

**Confidentiality Clause for the Use of Research Data**

The student Ryan Feenstra has received confidential data of the influencer payout within the scope of a research project of the University of Mannheim. This data may only be used during the student's master thesis. The resulting work may only be made available to the reviewers at the University of Mannheim and is not allowed to be published.

Any publication and duplication of this master thesis in digital as well as in printed format - even in part – is, therefore, prohibited. An inspection of this work by additional third parties besides the reviewers from the University of Mannheim requires the expressed permission of the Chair of Quantitative Marketing, in particular of Prof. Dr. Florian Stahl.

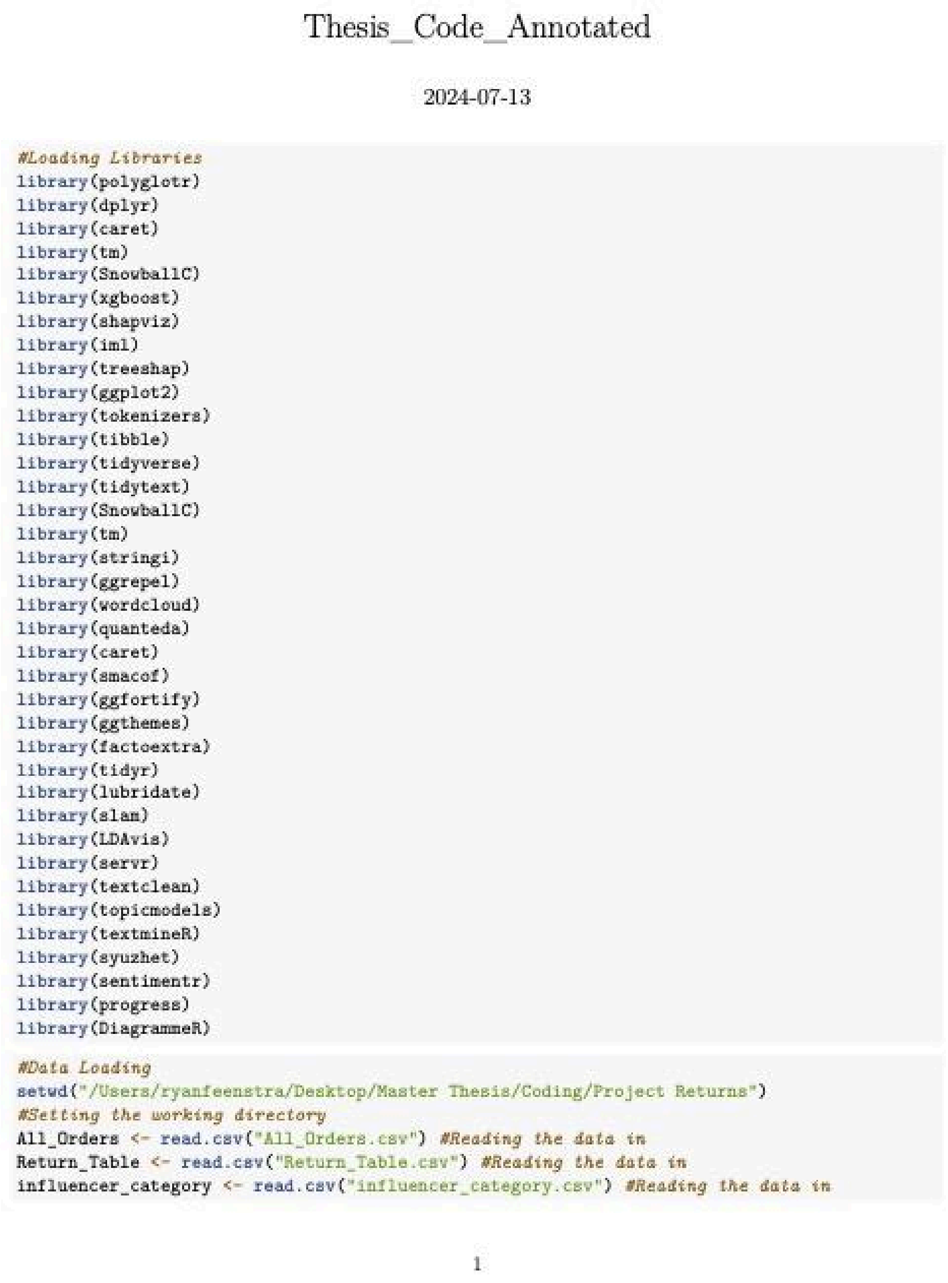
After the master thesis has been graded and the student has been notified of the grade, the student ensures to promptly and securely destroy all physical and digital copies of the confidential data obtained during the research.

25/02/24

Date, Signature of student

AACSB ACCREDITED
EFMD EQUIS ACCREDITED
ASSOCIATION AMBA ACCREDITED

**Appendix C: R Code**

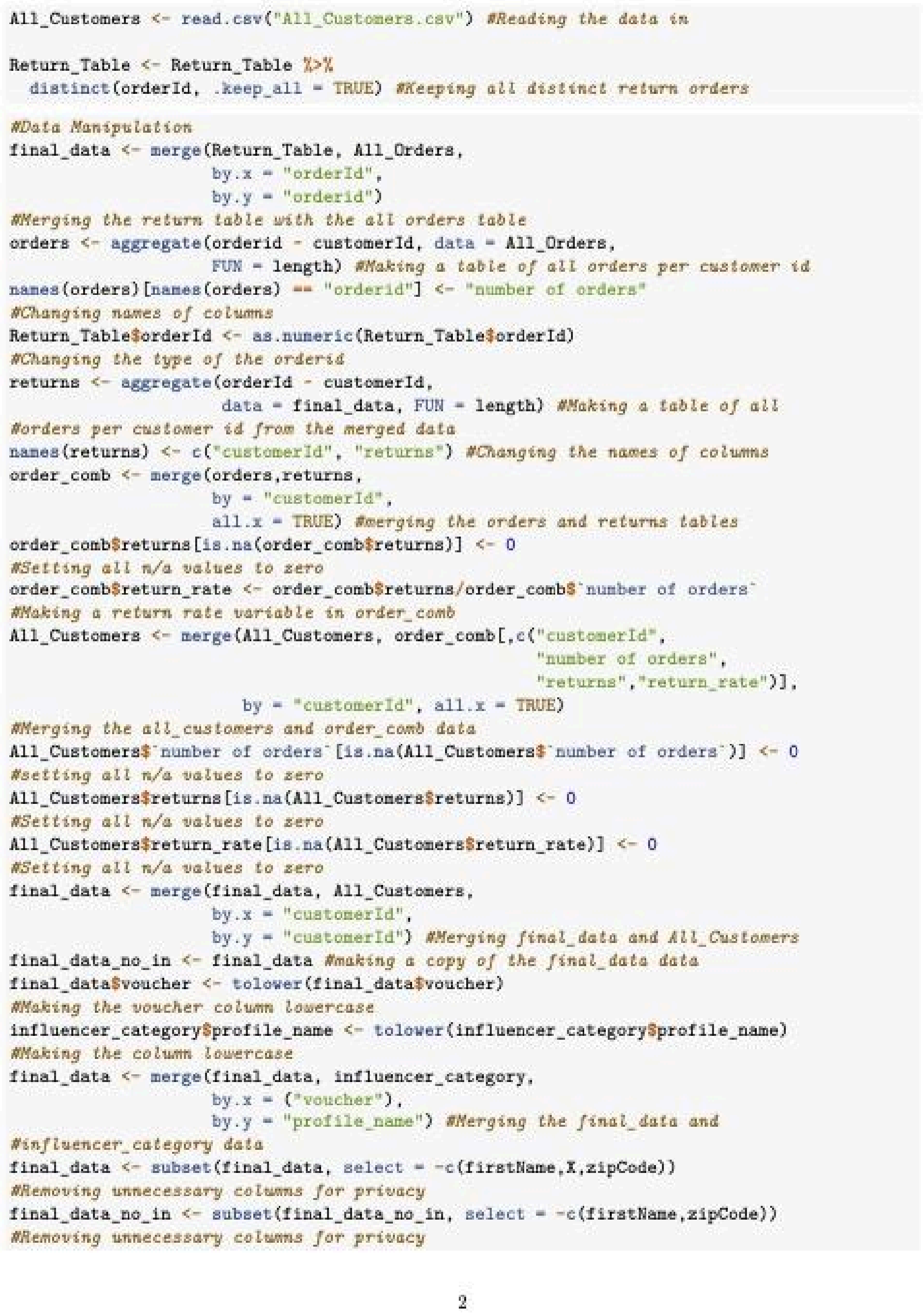


# Thesis_Code_Annotated

2024-07-13

```r
#Loading Libraries
library(polyglotr)
library(dplyr)
library(caret)
library(tm)
library(SnowballC)
library(xgboost)
library(shapviz)
library(iml)
library(treeshap)
library(ggplot2)
library(tokenizers)
library(tibble)
library(tidyverse)
library(tidytext)
library(SnowballC)
library(tm)
library(stringi)
library(ggrepel)
library(wordcloud)
library(quanteda)
library(caret)
library(smacof)
library(ggfortify)
library(ggthemes)
library(factoextra)
library(tidyr)
library(lubridate)
library(slam)
library(LDAvis)
library(servr)
library(textclean)
library(topicmodels)
library(textmineR)
library(syuzhet)
library(sentimentr)
library(progress)
library(DiagrammeR)
```

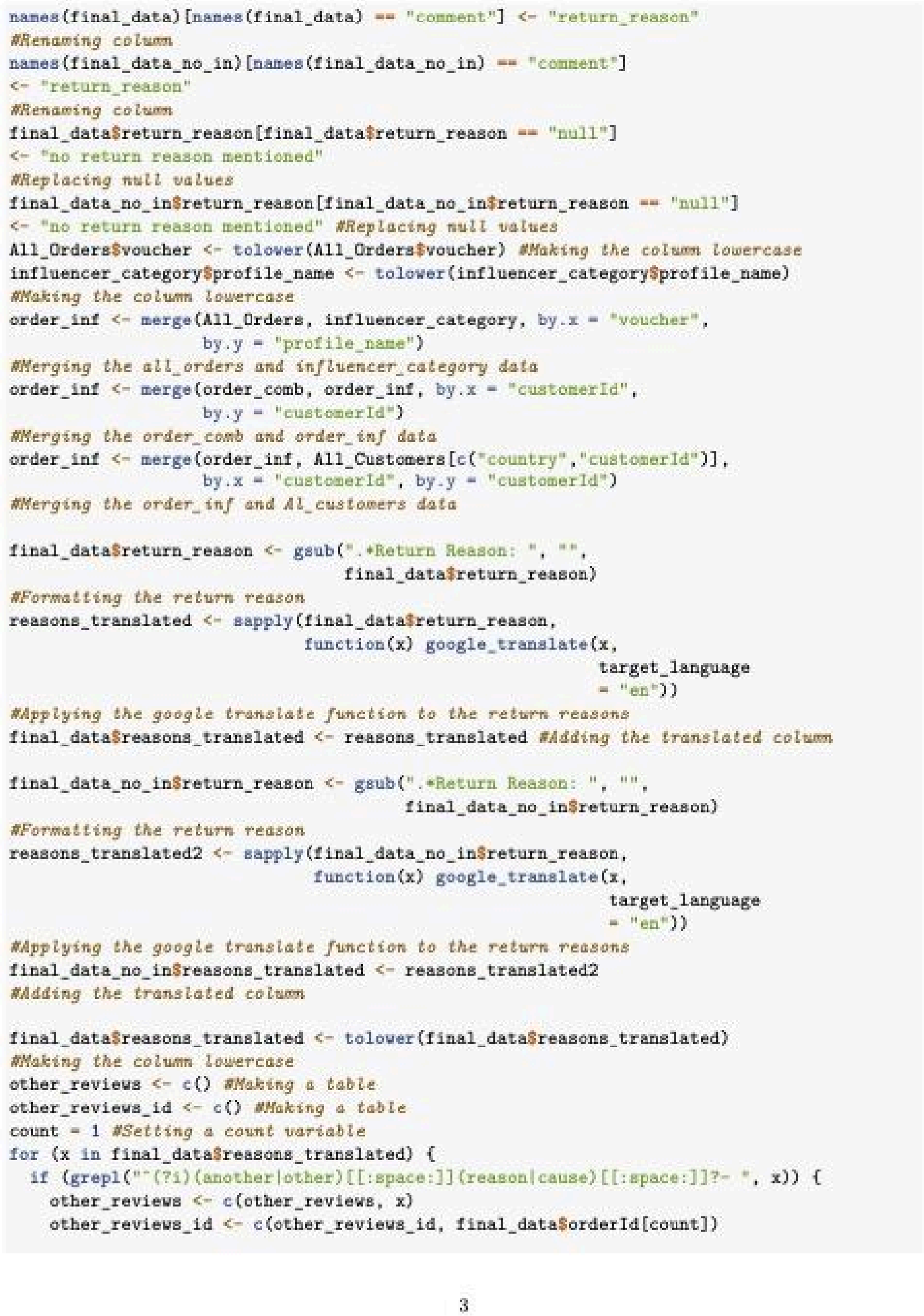
```r
#Data Loading
setwd("/Users/ryanfeenstra/Desktop/Master Thesis/Coding/Project Returns")
#Setting the working directory
All_Orders <- read.csv("All_Orders.csv") #Reading the data in
Return_Table <- read.csv("Return_Table.csv") #Reading the data in
influencer_category <- read.csv("influencer_category.csv") #Reading the data in
```



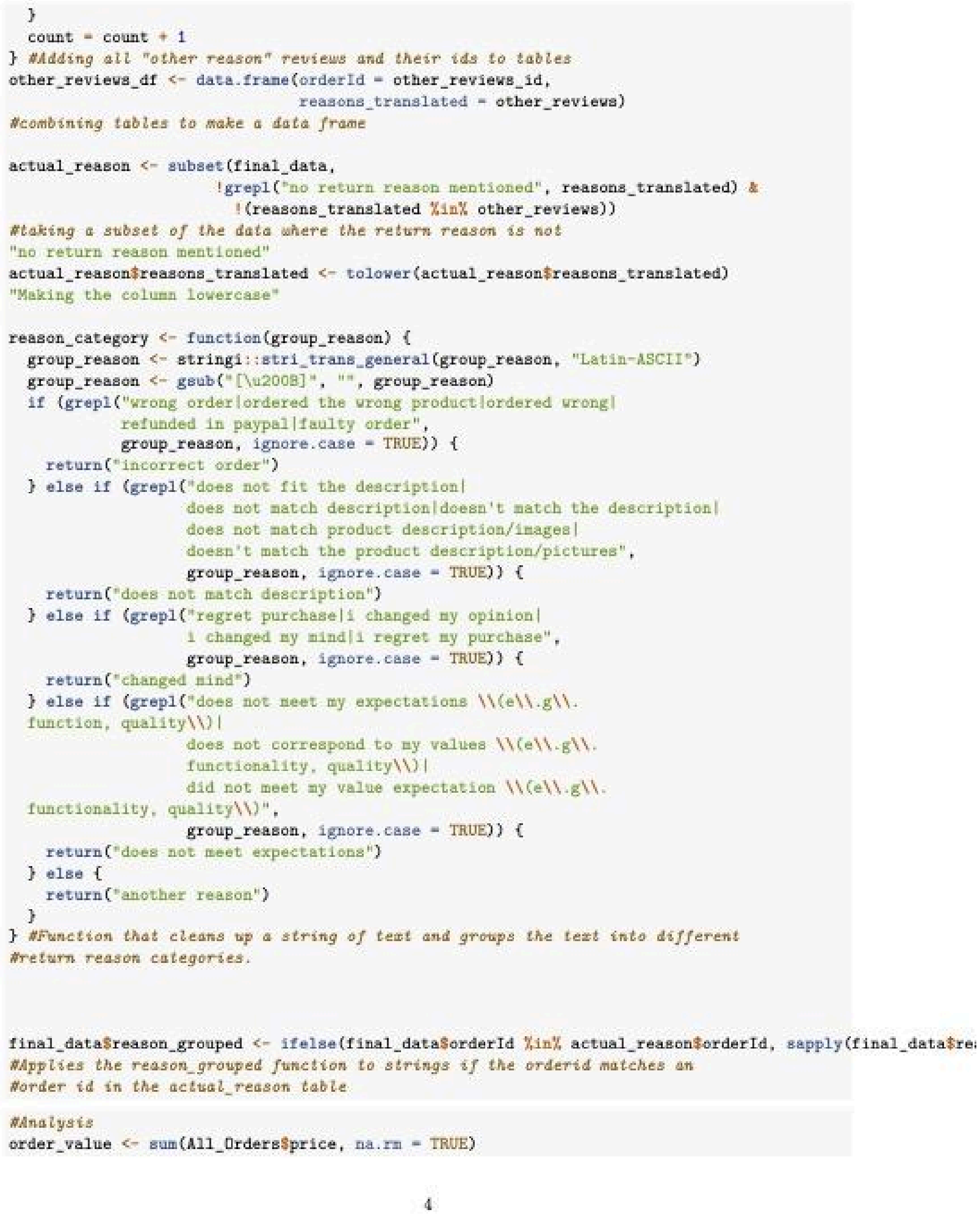
```r
All_Customers <- read.csv("All_Customers.csv") #Reading the data in

Return_Table <- Return_Table %>%
  distinct(orderId, .keep_all = TRUE) #Keeping all distinct return orders
```

```r
#Data Manipulation
final_data <- merge(Return_Table, All_Orders,
                    by.x = "orderId",
                    by.y = "orderid")
#Merging the return table with the all orders table
orders <- aggregate(orderid ~ customerId, data = All_Orders,
                    FUN = length) #Making a table of all orders per customer id
names(orders)[names(orders) == "orderid"] <- "number of orders"
#Changing names of columns
Return_Table$orderId <- as.numeric(Return_Table$orderId)
#Changing the type of the orderid
returns <- aggregate(orderId ~ customerId,
                     data = final_data, FUN = length) #Making a table of all
#orders per customer id from the merged data
names(returns) <- c("customerId", "returns") #Changing the names of columns
order_comb <- merge(orders,returns,
                    by = "customerId",
                    all.x = TRUE) #merging the orders and returns tables
order_comb$returns[is.na(order_comb$returns)] <- 0
#Setting all n/a values to zero
order_comb$return_rate <- order_comb$returns/order_comb$`number of orders`
#Making a return rate variable in order_comb
All_Customers <- merge(All_Customers, order_comb[,c("customerId",
                                                    "number of orders",
                                                    "returns","return_rate")],
                       by = "customerId", all.x = TRUE)
#Merging the all_customers and order_comb data
All_Customers$`number of orders`[is.na(All_Customers$`number of orders`)] <- 0
#setting all n/a values to zero
All_Customers$returns[is.na(All_Customers$returns)] <- 0
#Setting all n/a values to zero
All_Customers$return_rate[is.na(All_Customers$return_rate)] <- 0
#Setting all n/a values to zero
final_data <- merge(final_data, All_Customers,
                    by.x = "customerId",
                    by.y = "customerId") #Merging final_data and All_Customers
final_data_no_in <- final_data #making a copy of the final_data data
final_data$voucher <- tolower(final_data$voucher)
#Making the voucher column lowercase
influencer_category$profile_name <- tolower(influencer_category$profile_name)
#Making the column lowercase
final_data <- merge(final_data, influencer_category,
                    by.x = ("voucher"),
                    by.y = "profile_name") #Merging the final_data and
#influencer_category data
final_data <- subset(final_data, select = -c(firstName,X,zipCode))
#Removing unnecessary columns for privacy
final_data_no_in <- subset(final_data_no_in, select = -c(firstName,zipCode))
#Removing unnecessary columns for privacy
```

2

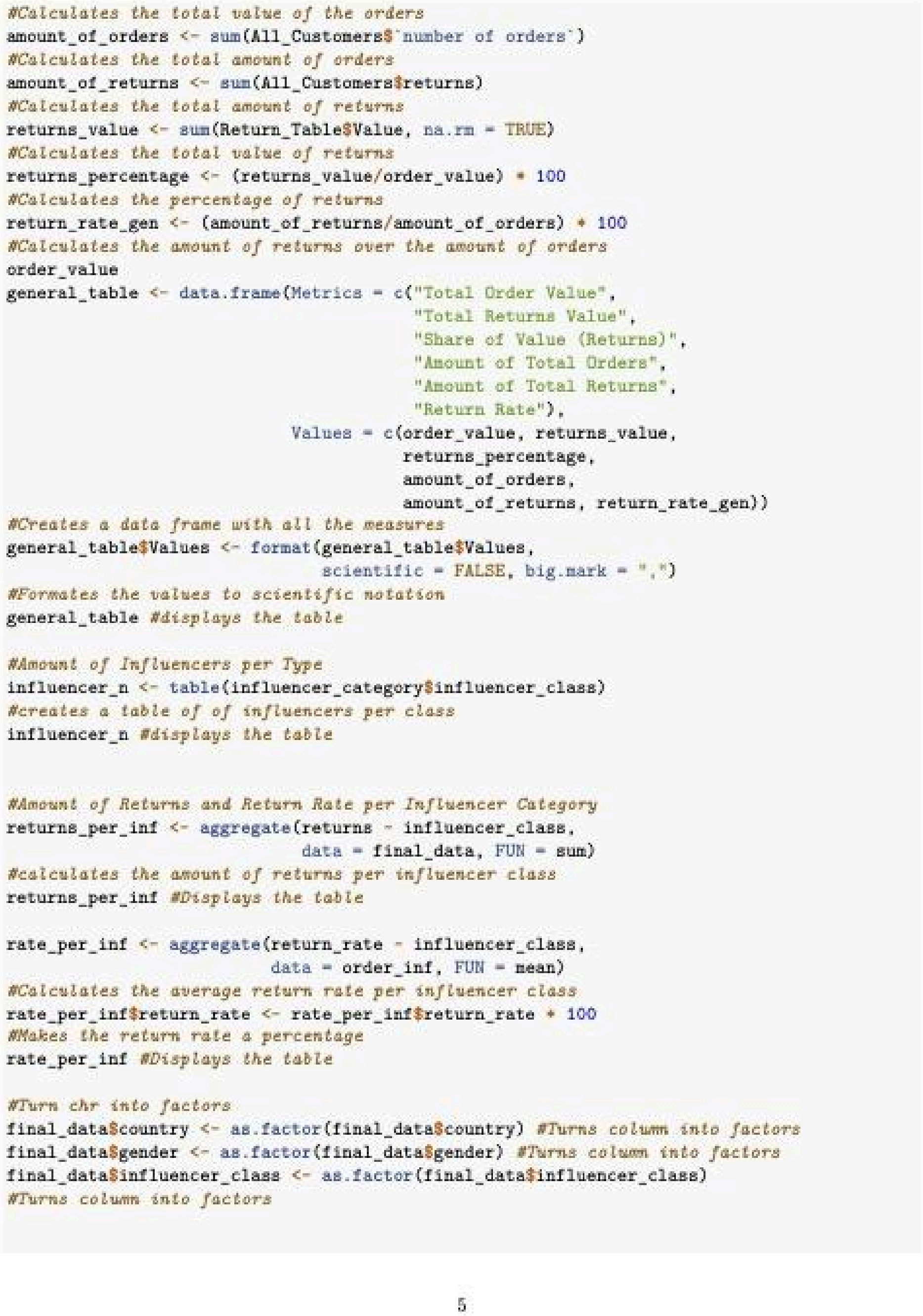


```r
names(final_data)[names(final_data) == "comment"] <- "return_reason"
#Renaming column
names(final_data_no_in)[names(final_data_no_in) == "comment"]
<- "return_reason"
#Renaming column
final_data$return_reason[final_data$return_reason == "null"]
<- "no return reason mentioned"
#Replacing null values
final_data_no_in$return_reason[final_data_no_in$return_reason == "null"]
<- "no return reason mentioned" #Replacing null values
All_Orders$voucher <- tolower(All_Orders$voucher) #Making the column lowercase
influencer_category$profile_name <- tolower(influencer_category$profile_name)
#Making the column lowercase
order_inf <- merge(All_Orders, influencer_category, by.x = "voucher",
                   by.y = "profile_name")
#Merging the all_orders and influencer_category data
order_inf <- merge(order_comb, order_inf, by.x = "customerId",
                   by.y = "customerId")
#Merging the order_comb and order_inf data
order_inf <- merge(order_inf, All_Customers[c("country","customerId")],
                   by.x = "customerId", by.y = "customerId")
#Merging the order_inf and Al_customers data

final_data$return_reason <- gsub(".*Return Reason: ", "",
                                 final_data$return_reason)
#Formatting the return reason
reasons_translated <- sapply(final_data$return_reason,
                             function(x) google_translate(x,
                                                          target_language
                                                          = "en"))
#Applying the google translate function to the return reasons
final_data$reasons_translated <- reasons_translated #Adding the translated column

final_data_no_in$return_reason <- gsub(".*Return Reason: ", "",
                                       final_data_no_in$return_reason)
#Formatting the return reason
reasons_translated2 <- sapply(final_data_no_in$return_reason,
                              function(x) google_translate(x,
                                                           target_language
                                                           = "en"))
#Applying the google translate function to the return reasons
final_data_no_in$reasons_translated <- reasons_translated2
#Adding the translated column

final_data$reasons_translated <- tolower(final_data$reasons_translated)
#Making the column lowercase
other_reviews <- c() #Making a table
other_reviews_id <- c() #Making a table
count = 1 #Setting a count variable
for (x in final_data$reasons_translated) {
  if (grepl("^(?i)(another|other)[[:space:]](reason|cause)[[:space:]]?- ", x)) {
    other_reviews <- c(other_reviews, x)
    other_reviews_id <- c(other_reviews_id, final_data$orderId[count])
```

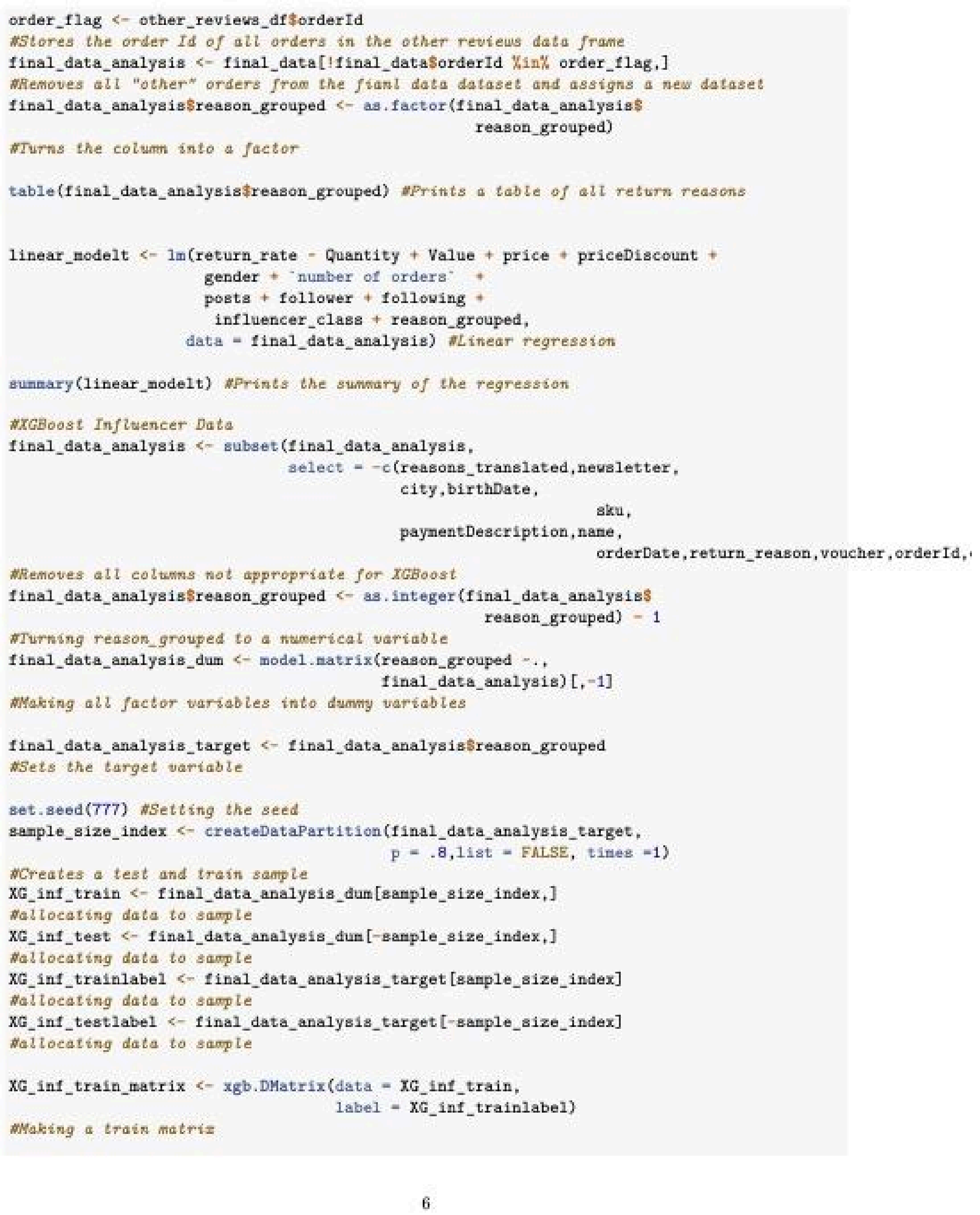


```r
  }
  count = count + 1
} #Adding all "other reason" reviews and their ids to tables
other_reviews_df <- data.frame(orderId = other_reviews_id,
                               reasons_translated = other_reviews)
#combining tables to make a data frame

actual_reason <- subset(final_data,
                        !grepl("no return reason mentioned", reasons_translated) &
                          !(reasons_translated %in% other_reviews))
#taking a subset of the data where the return reason is not
"no return reason mentioned"
actual_reason$reasons_translated <- tolower(actual_reason$reasons_translated)
"Making the column lowercase"

reason_category <- function(group_reason) {
  group_reason <- stringi::stri_trans_general(group_reason, "Latin-ASCII")
  group_reason <- gsub("[\u200B]", "", group_reason)
  if (grepl("wrong order|ordered the wrong product|ordered wrong|
            refunded in paypal|faulty order",
            group_reason, ignore.case = TRUE)) {
    return("incorrect order")
  } else if (grepl("does not fit the description|
                   does not match description|doesn't match the description|
                   does not match product description/images|
                   doesn't match the product description/pictures",
                   group_reason, ignore.case = TRUE)) {
    return("does not match description")
  } else if (grepl("regret purchase|i changed my opinion|
                   i changed my mind|i regret my purchase",
                   group_reason, ignore.case = TRUE)) {
    return("changed mind")
  } else if (grepl("does not meet my expectations \\(e\\.g\\.
  function, quality\\)|
                   does not correspond to my values \\(e\\.g\\.
                   functionality, quality\\)|
                   did not meet my value expectation \\(e\\.g\\.
  functionality, quality\\)",
                   group_reason, ignore.case = TRUE)) {
    return("does not meet expectations")
  } else {
    return("another reason")
  }
} #Function that cleans up a string of text and groups the text into different
#return reason categories.



final_data$reason_grouped <- ifelse(final_data$orderId %in% actual_reason$orderId, sapply(final_data$re
#Applies the reason_grouped function to strings if the orderid matches an
#order id in the actual_reason table
```

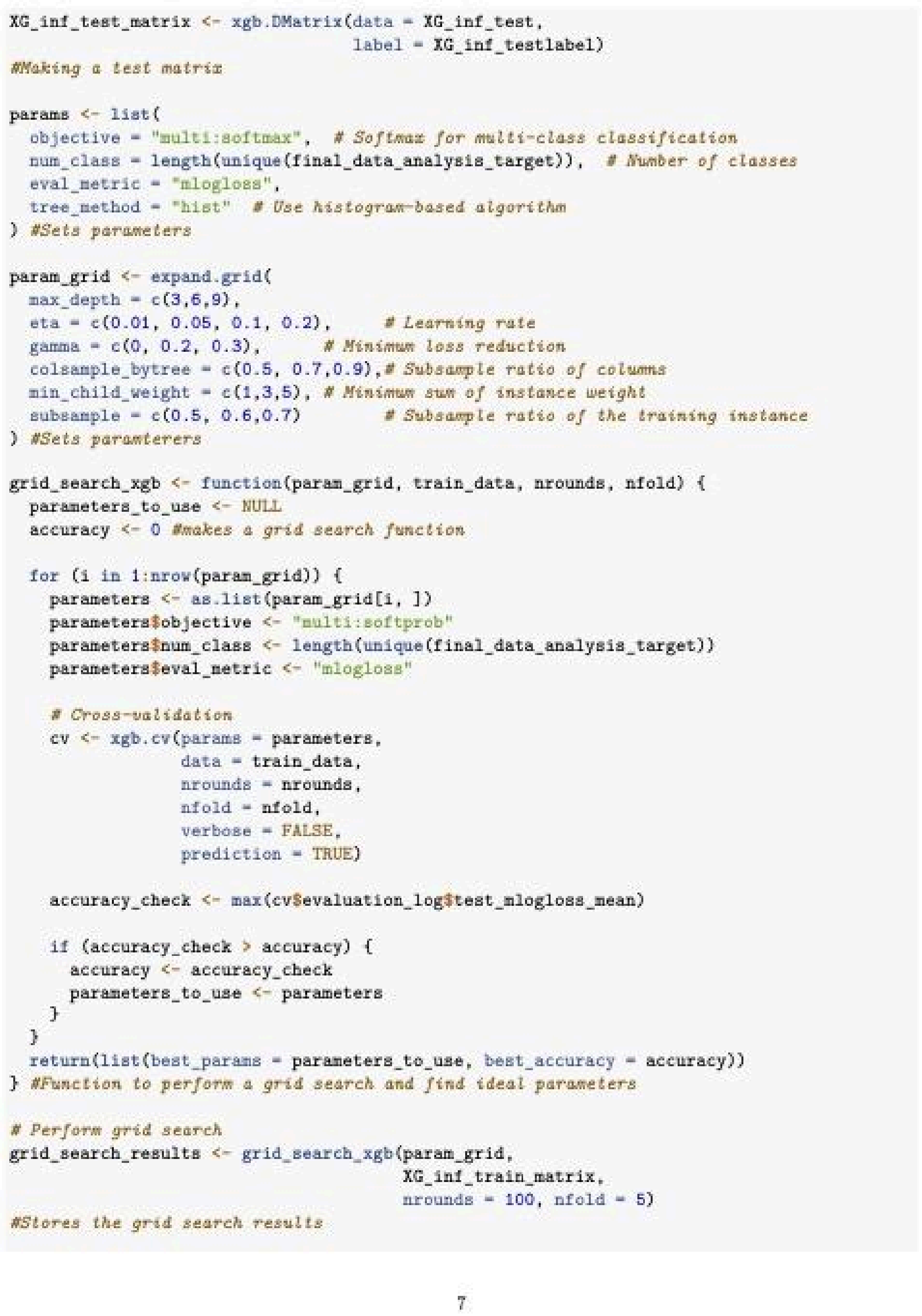
```r
#Analysis
order_value <- sum(All_Orders$price, na.rm = TRUE)
```



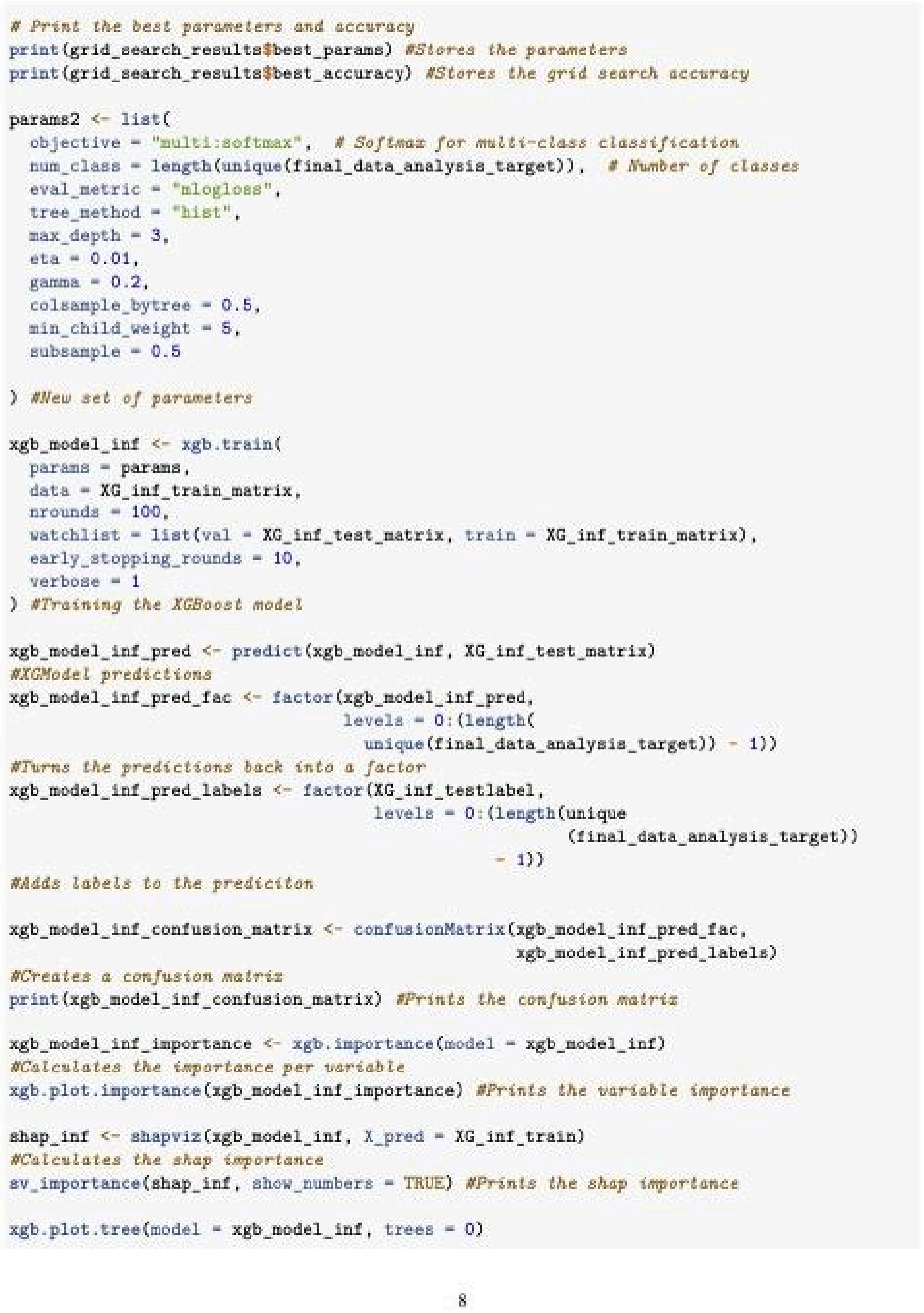
```
#Calculates the total value of the orders
amount_of_orders <- sum(All_Customers$`number of orders`)
#Calculates the total amount of orders
amount_of_returns <- sum(All_Customers$returns)
#Calculates the total amount of returns
returns_value <- sum(Return_Table$Value, na.rm = TRUE)
#Calculates the total value of returns
returns_percentage <- (returns_value/order_value) * 100
#Calculates the percentage of returns
return_rate_gen <- (amount_of_returns/amount_of_orders) * 100
#Calculates the amount of returns over the amount of orders
order_value
general_table <- data.frame(Metrics = c("Total Order Value",
                                        "Total Returns Value",
                                        "Share of Value (Returns)",
                                        "Amount of Total Orders",
                                        "Amount of Total Returns",
                                        "Return Rate"),
                            Values = c(order_value, returns_value,
                                       returns_percentage,
                                       amount_of_orders,
                                       amount_of_returns, return_rate_gen))
#Creates a data frame with all the measures
general_table$Values <- format(general_table$Values,
                               scientific = FALSE, big.mark = ",")
#Formates the values to scientific notation
general_table #displays the table

#Amount of Influencers per Type
influencer_n <- table(influencer_category$influencer_class)
#creates a table of of influencers per class
influencer_n #displays the table


#Amount of Returns and Return Rate per Influencer Category
returns_per_inf <- aggregate(returns ~ influencer_class,
                             data = final_data, FUN = sum)
#calculates the amount of returns per influencer class
returns_per_inf #Displays the table

rate_per_inf <- aggregate(return_rate ~ influencer_class,
                          data = order_inf, FUN = mean)
#Calculates the average return rate per influencer class
rate_per_inf$return_rate <- rate_per_inf$return_rate * 100
#Makes the return rate a percentage
rate_per_inf #Displays the table

#Turn chr into factors
final_data$country <- as.factor(final_data$country) #Turns column into factors
final_data$gender <- as.factor(final_data$gender) #Turns column into factors
final_data$influencer_class <- as.factor(final_data$influencer_class)
#Turns column into factors
```



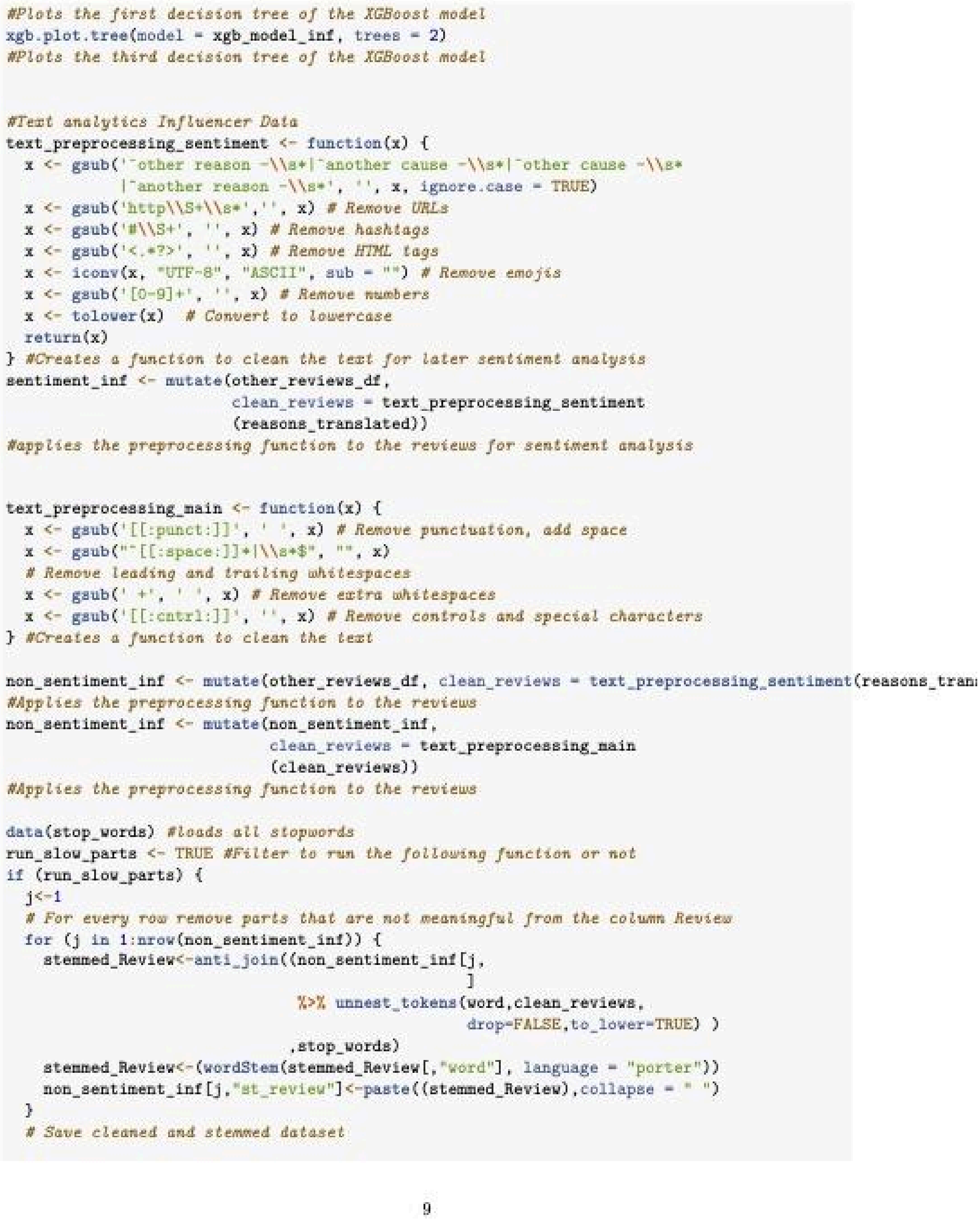
```r
order_flag <- other_reviews_df$orderId
#Stores the order Id of all orders in the other reviews data frame
final_data_analysis <- final_data[!final_data$orderId %in% order_flag,]
#Removes all "other" orders from the fianl data dataset and assigns a new dataset
final_data_analysis$reason_grouped <- as.factor(final_data_analysis$
                                                   reason_grouped)
#Turns the column into a factor

table(final_data_analysis$reason_grouped) #Prints a table of all return reasons


linear_modelt <- lm(return_rate ~ Quantity + Value + price + priceDiscount +
                     gender + `number of orders`  +
                     posts + follower + following +
                      influencer_class + reason_grouped,
                   data = final_data_analysis) #Linear regression

summary(linear_modelt) #Prints the summary of the regression

#XGBoost Influencer Data
final_data_analysis <- subset(final_data_analysis,
                              select = -c(reasons_translated,newsletter,
                                          city,birthDate,
                                                               sku,
                                          paymentDescription,name,
                                                               orderDate,return_reason,voucher,orderId,
#Removes all columns not appropriate for XGBoost
final_data_analysis$reason_grouped <- as.integer(final_data_analysis$
                                                    reason_grouped) - 1
#Turning reason_grouped to a numerical variable
final_data_analysis_dum <- model.matrix(reason_grouped ~.,
                                        final_data_analysis)[,-1]
#Making all factor variables into dummy variables

final_data_analysis_target <- final_data_analysis$reason_grouped
#Sets the target variable

set.seed(777) #Setting the seed
sample_size_index <- createDataPartition(final_data_analysis_target,
                                         p = .8,list = FALSE, times =1)
#Creates a test and train sample
XG_inf_train <- final_data_analysis_dum[sample_size_index,]
#allocating data to sample
XG_inf_test <- final_data_analysis_dum[-sample_size_index,]
#allocating data to sample
XG_inf_trainlabel <- final_data_analysis_target[sample_size_index]
#allocating data to sample
XG_inf_testlabel <- final_data_analysis_target[-sample_size_index]
#allocating data to sample

XG_inf_train_matrix <- xgb.DMatrix(data = XG_inf_train,
                                   label = XG_inf_trainlabel)
#Making a train matrix
```



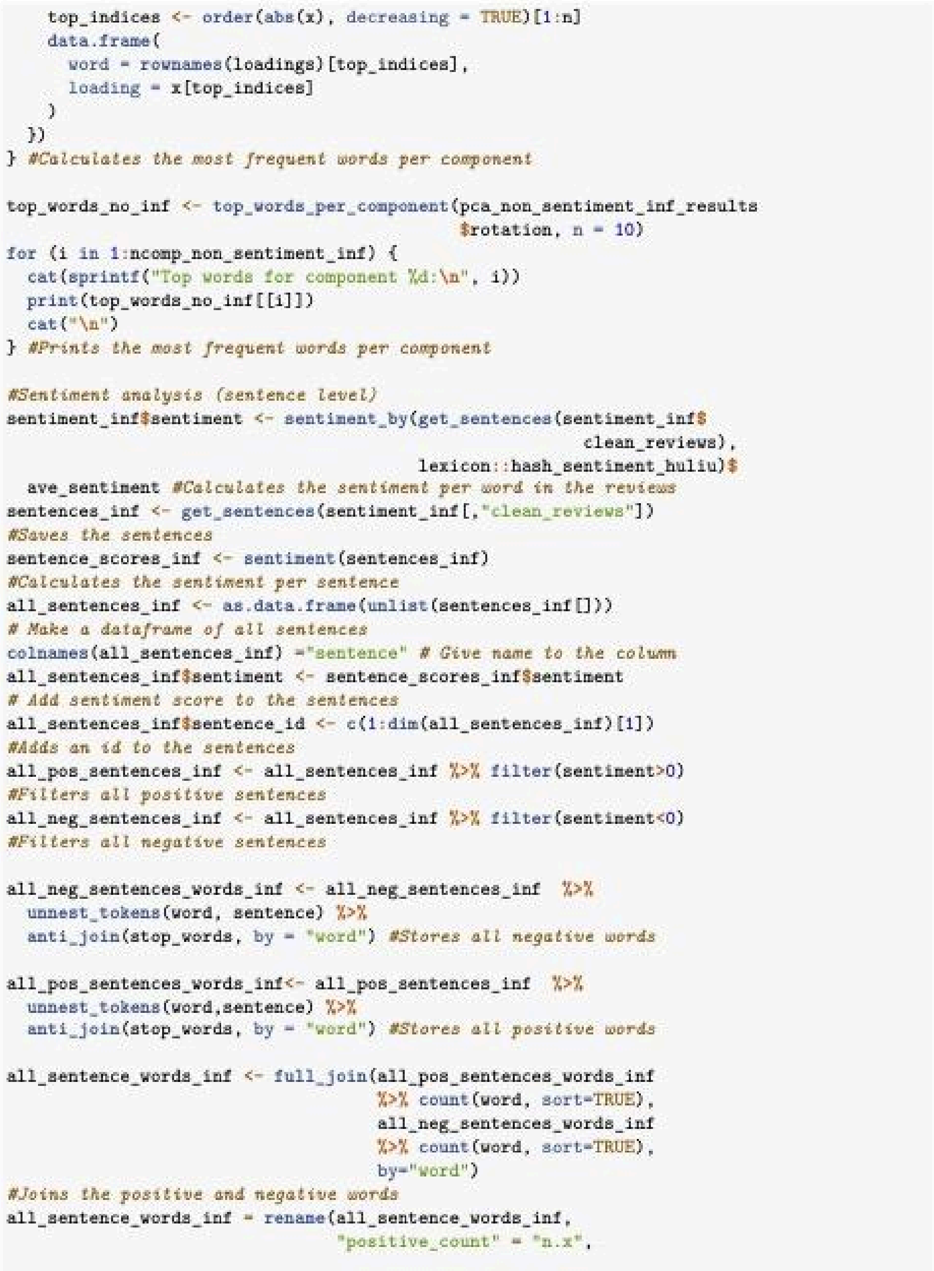
```
XG_inf_test_matrix <- xgb.DMatrix(data = XG_inf_test,
                                  label = XG_inf_testlabel)
#Making a test matrix

params <- list(
  objective = "multi:softmax",  # Softmax for multi-class classification
  num_class = length(unique(final_data_analysis_target)),  # Number of classes
  eval_metric = "mlogloss",
  tree_method = "hist"  # Use histogram-based algorithm
) #Sets parameters

param_grid <- expand.grid(
  max_depth = c(3,6,9),
  eta = c(0.01, 0.05, 0.1, 0.2),     # Learning rate
  gamma = c(0, 0.2, 0.3),       # Minimum loss reduction
  colsample_bytree = c(0.5, 0.7,0.9),# Subsample ratio of columns
  min_child_weight = c(1,3,5), # Minimum sum of instance weight
  subsample = c(0.5, 0.6,0.7)        # Subsample ratio of the training instance
) #Sets paramterers

grid_search_xgb <- function(param_grid, train_data, nrounds, nfold) {
  parameters_to_use <- NULL
  accuracy <- 0 #makes a grid search function

  for (i in 1:nrow(param_grid)) {
    parameters <- as.list(param_grid[i, ])
    parameters$objective <- "multi:softprob"
    parameters$num_class <- length(unique(final_data_analysis_target))
    parameters$eval_metric <- "mlogloss"

    # Cross-validation
    cv <- xgb.cv(params = parameters,
                 data = train_data,
                 nrounds = nrounds,
                 nfold = nfold,
                 verbose = FALSE,
                 prediction = TRUE)

    accuracy_check <- max(cv$evaluation_log$test_mlogloss_mean)

    if (accuracy_check > accuracy) {
      accuracy <- accuracy_check
      parameters_to_use <- parameters
    }
  }
  return(list(best_params = parameters_to_use, best_accuracy = accuracy))
} #Function to perform a grid search and find ideal parameters

# Perform grid search
grid_search_results <- grid_search_xgb(param_grid,
                                       XG_inf_train_matrix,
                                       nrounds = 100, nfold = 5)
#Stores the grid search results
```



```
# Print the best parameters and accuracy
print(grid_search_results$best_params) #Stores the parameters
print(grid_search_results$best_accuracy) #Stores the grid search accuracy

params2 <- list(
  objective = "multi:softmax",  # Softmax for multi-class classification
  num_class = length(unique(final_data_analysis_target)),  # Number of classes
  eval_metric = "mlogloss",
  tree_method = "hist",
  max_depth = 3,
  eta = 0.01,
  gamma = 0.2,
  colsample_bytree = 0.5,
  min_child_weight = 5,
  subsample = 0.5

) #New set of parameters

xgb_model_inf <- xgb.train(
  params = params,
  data = XG_inf_train_matrix,
  nrounds = 100,
  watchlist = list(val = XG_inf_test_matrix, train = XG_inf_train_matrix),
  early_stopping_rounds = 10,
  verbose = 1
) #Training the XGBoost model

xgb_model_inf_pred <- predict(xgb_model_inf, XG_inf_test_matrix)
#XGModel predictions
xgb_model_inf_pred_fac <- factor(xgb_model_inf_pred,
                                 levels = 0:(length(
                                   unique(final_data_analysis_target)) - 1))
#Turns the predictions back into a factor
xgb_model_inf_pred_labels <- factor(XG_inf_testlabel,
                                    levels = 0:(length(unique
                                                        (final_data_analysis_target))
                                                 - 1))
#Adds labels to the prediciton

xgb_model_inf_confusion_matrix <- confusionMatrix(xgb_model_inf_pred_fac,
                                                  xgb_model_inf_pred_labels)
#Creates a confusion matrix
print(xgb_model_inf_confusion_matrix) #Prints the confusion matrix

xgb_model_inf_importance <- xgb.importance(model = xgb_model_inf)
#Calculates the importance per variable
xgb.plot.importance(xgb_model_inf_importance) #Prints the variable importance

shap_inf <- shapviz(xgb_model_inf, X_pred = XG_inf_train)
#Calculates the shap importance
sv_importance(shap_inf, show_numbers = TRUE) #Prints the shap importance

xgb.plot.tree(model = xgb_model_inf, trees = 0)
```



```
#Plots the first decision tree of the XGBoost model
xgb.plot.tree(model = xgb_model_inf, trees = 2)
#Plots the third decision tree of the XGBoost model


#Text analytics Influencer Data
text_preprocessing_sentiment <- function(x) {
  x <- gsub('^other reason -\\s*|^another cause -\\s*|^other cause -\\s*
          |^another reason -\\s*', '', x, ignore.case = TRUE)
  x <- gsub('http\\S+\\s*','', x) # Remove URLs
  x <- gsub('#\\S+', '', x) # Remove hashtags
  x <- gsub('<.*?>', '', x) # Remove HTML tags
  x <- iconv(x, "UTF-8", "ASCII", sub = "") # Remove emojis
  x <- gsub('[0-9]+', '', x) # Remove numbers
  x <- tolower(x)  # Convert to lowercase
  return(x)
} #Creates a function to clean the text for later sentiment analysis
sentiment_inf <- mutate(other_reviews_df,
                        clean_reviews = text_preprocessing_sentiment
                        (reasons_translated))
#applies the preprocessing function to the reviews for sentiment analysis


text_preprocessing_main <- function(x) {
  x <- gsub('[[:punct:]]', ' ', x) # Remove punctuation, add space
  x <- gsub("^[[:space:]]+|\\s+$", "", x)
  # Remove leading and trailing whitespaces
  x <- gsub(' +', ' ', x) # Remove extra whitespaces
  x <- gsub('[[:cntrl:]]', '', x) # Remove controls and special characters
} #Creates a function to clean the text

non_sentiment_inf <- mutate(other_reviews_df, clean_reviews = text_preprocessing_sentiment(reasons_tran
#Applies the preprocessing function to the reviews
non_sentiment_inf <- mutate(non_sentiment_inf,
                            clean_reviews = text_preprocessing_main
                            (clean_reviews))
#Applies the preprocessing function to the reviews

data(stop_words) #loads all stopwords
run_slow_parts <- TRUE #Filter to run the following function or not
if (run_slow_parts) {
  j<-1
  # For every row remove parts that are not meaningful from the column Review
  for (j in 1:nrow(non_sentiment_inf)) {
    stemmed_Review<-anti_join((non_sentiment_inf[j,
                                                 ]
                               %>% unnest_tokens(word,clean_reviews,
                                                 drop=FALSE,to_lower=TRUE) )
                              ,stop_words)
    stemmed_Review<-(wordStem(stemmed_Review[,"word"], language = "porter"))
    non_sentiment_inf[j,"st_review"]<-paste((stemmed_Review),collapse = " ")
  }
  # Save cleaned and stemmed dataset
  save(non_sentiment_inf, file = "non_sentiment_influencer.Rdata")
} else {load(file = "non_sentiment_influencer")
} #Function to stem the words within the reviews

review_tdm <- non_sentiment_inf %>% unnest_tokens(word,st_review) %>%
  count(word,orderId,sort=TRUE) %>%ungroup()%>%cast_tdm(word,orderId,n)
#Creates a term document matrix
counts <- rowSums(as.matrix(review_tdm)) #Counts the sum of rows
sortedcount <- counts %>% sort(decreasing=TRUE) #Sorts the counts variable
nwords <- 152 #Sets the amount of words in the reviews
sortednames <- names(sortedcount[1:nwords]) #Sorts the names
review_dtm <- t(review_tdm)
#Transposes the term document matrix to a document term matrix

pca_non_sentiment_inf_results <- prcomp(review_dtm, scale = FALSE, rank. = 40)
#Performs PCA analysis
fviz_screeplot(pca_non_sentiment_inf_results,ncp=20) #Prints a screeplot
ncomp_non_sentiment_inf<-2 #Sets the amount of components

axeslist <- c(1, 2)
fviz_pca_var(pca_non_sentiment_inf_results, axes=axeslist
             ,geom.var = c("arrow", "text")
             ,col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), # colors to use
             repel = TRUE     # Avoid text overlapping
) #Prints the two components on a grpah

rawLoadings <- pca_non_sentiment_inf_results$rotation[sortednames,
                                                       1:ncomp_non_sentiment_inf]
%*%
  diag(pca_non_sentiment_inf_results$sdev,
       ncomp_non_sentiment_inf, ncomp_non_sentiment_inf) #Calculates the loadings
rotated_non_sentiment_inf <- varimax(rawLoadings) # rotate loading matrix
pca_non_sentiment_inf_results$rotation <- rotated_non_sentiment_inf$loadings
# Saves the rotated results
pca_non_sentiment_inf_results$x <- scale(pca_non_sentiment_inf_results$
                                         x[,1:ncomp_non_sentiment_inf]) %*%
  rotated_non_sentiment_inf$rotmat #Scales the rotated loadings

axeslist <- c(1, 2)
fviz_pca_var(pca_non_sentiment_inf_results, axes=axeslist
             ,geom.var = c("text", "arrow")
             ,col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#C1CAD6", "#CB429F", "red"), # colors to use
             repel = TRUE,
             xlim = c(-0.8, 0.8),
             ylim = c(-0.8, 0.8)
) #Prints the rotated loadings on a graph



top_words_per_component <- function(loadings, n = 10) {
  apply(loadings, 2, function(x) {
```



```
    top_indices <- order(abs(x), decreasing = TRUE)[1:n]
    data.frame(
      word = rownames(loadings)[top_indices],
      loading = x[top_indices]
    )
  })
} #Calculates the most frequent words per component

top_words_no_inf <- top_words_per_component(pca_non_sentiment_inf_results
                                            $rotation, n = 10)
for (i in 1:ncomp_non_sentiment_inf) {
  cat(sprintf("Top words for component %d:\n", i))
  print(top_words_no_inf[[i]])
  cat("\n")
} #Prints the most frequent words per component

#Sentiment analysis (sentence level)
sentiment_inf$sentiment <- sentiment_by(get_sentences(sentiment_inf$
                                                         clean_reviews),
                                        lexicon::hash_sentiment_huliu)$
  ave_sentiment #Calculates the sentiment per word in the reviews
sentences_inf <- get_sentences(sentiment_inf[,"clean_reviews"])
#Saves the sentences
sentence_scores_inf <- sentiment(sentences_inf)
#Calculates the sentiment per sentence
all_sentences_inf <- as.data.frame(unlist(sentences_inf[]))
# Make a dataframe of all sentences
colnames(all_sentences_inf) ="sentence" # Give name to the column
all_sentences_inf$sentiment <- sentence_scores_inf$sentiment
# Add sentiment score to the sentences
all_sentences_inf$sentence_id <- c(1:dim(all_sentences_inf)[1])
#Adds an id to the sentences
all_pos_sentences_inf <- all_sentences_inf %>% filter(sentiment>0)
#Filters all positive sentences
all_neg_sentences_inf <- all_sentences_inf %>% filter(sentiment<0)
#Filters all negative sentences

all_neg_sentences_words_inf <- all_neg_sentences_inf  %>%
  unnest_tokens(word, sentence) %>%
  anti_join(stop_words, by = "word") #Stores all negative words

all_pos_sentences_words_inf<- all_pos_sentences_inf  %>%
  unnest_tokens(word,sentence) %>%
  anti_join(stop_words, by = "word") #Stores all positive words

all_sentence_words_inf <- full_join(all_pos_sentences_words_inf
                                    %>% count(word, sort=TRUE),
                                    all_neg_sentences_words_inf
                                    %>% count(word, sort=TRUE),
                                    by="word")
#Joins the positive and negative words
all_sentence_words_inf = rename(all_sentence_words_inf,
                                 "positive_count" = "n.x",
                                "negative_count" = "n.y") #Renames Columns

all_sentence_words_inf[is.na(all_sentence_words_inf$positive_count),
                       "positive_count"]<- 0 # set missing values equal to zero
all_sentence_words_inf[is.na(all_sentence_words_inf$negative_count),
                       "negative_count"]<- 0 # set missing values equal to zero

all_sentence_words_inf$positive_count  <- all_sentence_words_inf$positive_count/sum(all_sentence_words_
#Counts all positive words
all_sentence_words_inf$negative_count  <- all_sentence_words_inf$negative_count/sum(all_sentence_words_
#Counts all negative words

all_sentence_words_inf$diff <- all_sentence_words_inf$
  positive_count-all_sentence_words_inf$negative_count
# Determine difference between ratio of positive and negative sentences for each word


all_sentence_words_inf[is.na(all_sentence_words_inf$positive_count),
                       "positive_count"] <- 1
# missing values: avoid division by 0
all_sentence_words_inf[is.na(all_sentence_words_inf$negative_count),
                       "negative_count"] <- 1
# missing values: avoid division by 0

all_sentence_words_inf$ratio <- all_sentence_words_inf$
  positive_count/all_sentence_words_inf$negative_count
# Determine ratio: positive divided by negative scores for each word

all_sentence_words_inf%>% #Only consider words with negative and positive score > 5,
  #prints top 15 words based on ratio
  mutate(word = reorder(word, -ratio)) %>%
  top_n(-15, ratio) %>%
  ggplot(aes(word,ratio)) +
  geom_col() +
  labs(x = NULL, y = "Ratio of word frequency (pos/neg)") +
  coord_flip() +
  theme(text = element_text(size = 17)) +
  ggtitle("Specific negative words")

all_sentence_words_inf%>% # Only consider words with negative and positive score
  #>5, prints top 15 words based on ratio
  mutate(word = reorder(word,ratio)) %>%
  top_n(15, ratio) %>%
  ggplot(aes(word,ratio)) +
  geom_col() +
  labs(x = NULL, y = "Ratio of word frequency (pos/neg)") +
  coord_flip() +
  theme(text = element_text(size = 17)) +
  ggtitle("Specific positive words")

#Topic modelling
LDA_dtm <- review_dtm #Sets the document term matrix for LDA
LDA_dtm <- as.DocumentTermMatrix(review_dtm) #Formats the document term matrix
```

```r
num_topics_inf <- 2 #Sets the amount of LDA topics
lda_inf <- LDA(LDA_dtm, k = num_topics_inf, control = list(seed = 777))
#Performs LDA
lda_inf_terms <- terms(lda_inf, 10) #Sets the amount of LDA terms to show
tidy_lda_inf <- tidy(lda_inf) #Terns the LDA output to the tidy format

for (i in 1:num_topics_inf) {
  cat(sprintf("Top terms for topic %d:\n", i))
  print(lda_inf_terms[, i])
  cat("\n")
} #Peints the top 10 terms per topic

top_terms_inf <- tidy_lda_inf %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta) #Arranges and prints the top 10 terms per topic
#from the tidy LDA format

ggplot(top_terms_inf, aes(reorder(term, beta), beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "Top 10 terms in each LDA topic",
       x = "Term",
       y = "Beta") +
  theme_minimal() #prints the 10 terms in each LDA topic in a plot
```