Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science & Marketing Analytics

Drivers behind the success of vegetarian-friendly and non-vegetarian-friendly restaurants.

Jet van Dis (647498)

Supervisor:	dr. C.S. Bellet
Second assessor:	prof.dr. P.J.F. Groenen
Date final version:	27th October 2023

Abstract

In recent years, the rise of electronic Word-Of-Mouth (eWOM) has transformed how customers express their opinions on products and services. It has become a powerful source of information for consumers, thereby shaping their decision-making process. Another recent trend is the substantial growth of the market for vegetarian and vegan diets. Individuals adopt these diets for reasons spanning health, ethics, community, and sustainability. Given this evolving landscape, it is crucial to understand the drivers of success within the restaurant industry. This research bridges a critical gap by advancing our comprehension of eWOM dynamics and the role of social networks within the unique context of vegetarian and non-vegetarian-friendly restaurants. Therefore, this study aims to answer the following research question: What are the key drivers of success for vegetarian and non-vegetarian-friendly restaurants? Leveraging a data set from Yelp, our analysis employs a diverse set of methodologies to unravel the key drivers of a restaurant's success. These methodologies include 1:1 nearest neighbor propensity score matching, the Latent Dirichlet Allocation (LDA) algorithm, sentiment analysis, an Ordinary Least Squares (OLS) regression with Fixed Time Effects, a Least Absolute Shrinkage and Selection Operator (LASSO) regression with Fixed Time Effects, a Random Forest model and a Support Vector Regression. We conclude that the key drivers involve topics related to price, food quality, service quality and ambience, and the social networks of reviewers. Moreover, we find evidence for differences in topic importance among restaurant segments.

Contents

1 Introduction

1.1 Situation and Research Question

In recent years, the rise of electronic Word-Of-Mouth (eWOM) has transformed how customers express their opinions on products and services, including restaurants (Nilashi et al., 2021). The emergence of Information and Communication Technology (ICT) and social media platforms has provided consumers with a convenient and influential platform to share their experiences and recommendations. Following that, it has become a powerful source of information for consumers, shaping their decision-making process. Nilashi et al. (2021) argue that the impact of eWOM is particularly significant for intangible products and services, such as restaurant experiences, as they are challenging to evaluate before utilization (Nilashi et al., 2021). Therefore, it has become crucial for the restaurant industry to understand the drivers of success in this context.

Furthermore, the market for vegetarian and vegan diets has experienced substantial growth in recent decades (Godfray et al., 2018). Hargreaves et al. (2021) argue that individuals opt for vegetarian or vegan diets due to various factors. Firstly, a vegetarian or vegan diet is associated with improved physical health, promoting well-being by consuming nutritious plant-based foods. Secondly, adopting a vegetarian diet often evokes a sense of moral correctness, resulting in positive emotions associated with making ethical choices. Thirdly, embracing vegetarianism can foster a sense of belonging as individuals connect with the vegetarian community, sharing common values and dietary preferences. Lastly, vegetarianism helps reduce environmental impact by minimizing reliance on animal products, which promotes sustainability.

Previous research has explored various drivers of restaurant success. Bufquin et al. (2017) identify food quality, price, service quality, social aspects, and healthfulness. Additionally, Gan et al. (2017) added the attribute context and ambience to the previous list. However, these studies did not differentiate between restaurant segments and customer preferences. Choi et al. (2022) addressed this gap by investigating the factors influencing vegetarian customers and considering the availability of vegetarian menu options. Moreover, Harrington et al. (2012) examined the influence of various attributes on customers from Generation Y across three distinct restaurant segments, arguing that customers have different needs and wants based on the segment they are dining in. Drawing from these studies, the research question driving this study is:

> What are the key drivers of success for vegetarian and non-vegetarian-friendly restaurants?

1.2 Research Context and Design

This study employs a literature review and analysis focusing on electronic Word-Of-Mouth (eWOM) in the form of online restaurant reviews. To investigate the key drivers of success for vegetarian and non-vegetarian-friendly restaurants, the data set from Yelp is utilized. Before conducting the analysis, several steps are taken.

The Yelp business data set is used to extract restaurant information. To ensure comparability between vegetarian and non-vegetarian-friendly restaurants, we employ a 1:1 nearest neighbor

propensity score matching approach. This matching process aims to create pairs of restaurants with similar characteristics, such as the number of reviews, days since first check-in, average star rating, and various categories, states, and cities.

Once a subset of matched vegetarian and non-vegetarian-friendly restaurants with comparable characteristics is established, the corresponding review data from the Yelp review data set is prepared for further analysis. The reviews are cleaned and stemmed, and the Latent Dirichlet Allocation (LDA) algorithm is applied to extract different topics from the reviews. Additionally, sentiment analysis is performed on the cleaned and stemmed reviews to capture the sentiment expressed in the text.

Furthermore, various variables related to social networks and social environment are collected. These variables include the number of friends of the reviewer as a measure of social network. Moreover, Google Trends search terms and the ratio of vegetarian-friendly restaurants to non-vegetarian-friendly restaurants in a state as a measure of the social environment.

The analysis is based on a cross-sectional time series data set, with each variable included at the restaurant and month level. The dependent variable is the number of reviews, which serves as a proxy for restaurant success. This follows from previous studies showing a significant positive effect of the number of reviews on sales (Chevalier & Mayzlin, 2006; Dellarocas, Zhang & Awad, 2007).

In this study, we compare the performance of several models for our analysis, including an Ordinary Least Squares (OLS) regression with Fixed Time Effects, a Least Absolute Shrinkage and Selection Operator (LASSO) regression with Fixed Time Effects, a Random Forest model and a Support Vector Regression. These models are compared based on three performance measures: the Residual Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the R-squared. We find that the Random Forest model is the best fit for our data. Therefore, the Random Forest model extracts conclusions regarding the main drivers of vegetarian and non-vegetarian restaurant success. Finally, we employ another regression with time and restaurant fixed effects using the most highly predictive variables uncovered in the Random Forest model. We interact these with a dummy variable for being a vegetarian-friendly restaurant. This final approach provides a more robust causal identification, as it leverages within-restaurant variation in product reviews to compare success drivers between vegetarian and non-vegetarian-friendly restaurants.

By employing this research design, the study aims to uncover the factors that drive the success of vegetarian and non-vegetarian-friendly restaurants based on an analysis of online reviews and associated variables.

1.3 Relevance

1.3.1 Academic Relevance

This thesis contributes to academia in multiple ways. Firstly, it adds significant value to the extensive body of existing literature on electronic Word of Mouth (eWOM) by expanding the current knowledge base. Through the application of Latent Dirichlet Allocation (LDA) topic modeling, this study unveils hidden patterns and insights in the eWOM of restaurants. By examining the influence of various factors on eWOM volume across different restaurant segments,

the research extends our understanding of the dynamics and drivers of eWOM.

Secondly, the study expands on existing research by investigating the impact of reviewers' online social networks on eWOM volume, explicitly focusing on the context of vegetarian-friendly and non-vegetarian-friendly restaurants. By employing the 'diffusion of innovation' theory (Rogers, 1962), the research provides a novel perspective and contributes to our understanding of how social connections influence eWOM dynamics in this specific restaurant segment.

1.3.2 Managerial Relevance

These findings also have practical implications for restaurant managers who rely on eWOM to promote their restaurants. Firstly, this study uncovers previously unknown factors influencing eWOM volume, providing a more objective view of existing reviews. Restaurant managers can benefit from this knowledge by gaining nuanced insights into quality and performance. This helps them understand which aspects of their restaurant drive more reviews, which they can use to make informed decisions to enhance customer experience and improve their offerings. In line with this, Nilashi et al. (2021) argue that restaurants can improve their service quality and marketing strategy by detecting customer preferences through online reviews.

Secondly, this study reveals potential variations in the influence of these factors across different restaurant segments. Managers can leverage these findings by developing tailored strategies that align with the preferences and expectations within their specific segment. By enhancing the quality of topics that hold significance for their restaurant segment, managers can effectively boost eWOM volume, which ultimately contributes to the overall success of a restaurant.

In summary, this thesis contributes to academia by expanding our understanding of eWOM dynamics and uncovering hidden patterns in the eWOM of restaurants. Additionally, it offers practical implications for restaurant managers, providing valuable insights to improve their restaurant's reputation, customer satisfaction, and overall performance in the restaurants' eWOM.

1.4 Research Outline

In the remainder of this thesis, Section 2 provides a comprehensive overview of the relevant literature on electronic Word-Of-Mouth (eWOM), drivers behind the success of restaurants, social networks, social environment, and sentiment. Next, Section 3 proposes five hypotheses, along with their sub-hypotheses, to address the research question. Furthermore, Section 4 outlines the methodology employed in this study. It describes the steps taken to subset the restaurants, clean the reviews, extract the topics, analyze the sentiment, and conduct the data analysis. After that, Section 5 presents the data utilized for the analysis, including descriptive statistics of the data set. This section provides an overview of the characteristics and key variables used in the study. Moreover, Section 6 presents the results obtained from the analysis. The conclusions drawn from the obtained results are outlined in Section 7. This section synthesizes the findings and implications, addressing the research question and hypotheses. Finally, Section 8 discusses the study's limitations and suggests directions for future research.

2 Literature Review

As discussed in the introduction, this study explores the drivers behind the success or failure of vegetarian-friendly and non-vegetarian-friendly restaurants. This study utilizes the number of reviews as a proxy for the success or failure of a restaurant. This was derived from a range of studies showing that the number of reviews positively affects sales (Chevalier & Mayzlin, 2006; Dellarocas et al., 2007). In the remainder of this section, we will explore why many scholars are interested in the drivers behind online reviews (Subsection 2.1), why we are segmenting the restaurants based on their vegetarian-friendliness (Subsection 2.2), why we expect sentiment to influence the success of a restaurant (Subsection 2.4) and why we expect social network and environment to influence the success of a restaurant (Subsection 2.3).

2.1 Emergence of electronic Word-Of-Mouth (eWOM)

In recent years, customers have increasingly been utilizing diverse social media platforms to express their views on the quality of different products and services, including restaurants (Nilashi et al., 2021). With the emergence of Information and Communication Technology (ICT) in the last decades, it has become increasingly useful for consumers to leave reviews about products and services. The definition of the concept of Word-Of-Mouth (WOM) dates back to a study from Buttle (1998), who describe it as face-to-face communication related to products, goods, and services. While WOM was already recognized as a significant source of information that exerts a strong influence on customer behavior and decision-making (Jeong & Jang, 2011), the emergence of ICT, and thereby the World Wide Web and e-commerce, has introduced the concept of electronic Word-Of-Mouth (eWOM). Also regarding eWOM, research has already indicated that it contributes significantly to the sales and overall value of a company (Babić Rosario et al., 2016; You et al., 2015).

Moreover, Nilashi et al. (2021) argues that eWOM communication between customers is primarily relevant for intangible products and services that cannot or are difficult to evaluate before utilization, such as tourism and restaurant services. Therefore, it is important for these industries to find the drivers for customer satisfaction, as this may lead to customers revisiting and recommending the service to other people.

Consequently, marketing professionals are increasingly trying to understand how to effectively stimulate and manage online reviews, while researchers have been investigating the underlying factors that drive electronic Word Of Mouth (eWOM). The availability of big data collected from these social media platforms offers a valuable resource to enhance restaurant success based on customers' online reviews. Online reviews are regarded as credible and dependable sources that assist consumers in assessing the quality of a restaurant (Nilashi et al., 2021).

2.2 Drivers Behind Restaurant Success

As a result of the emergence of eWOM, there has been a growing interest among researchers and scholars to study the underlying factors driving eWOM. Specifically, researchers have focused on understanding the attributes that influence customers' satisfaction and behavioral intentions,

such as their willingness to pay more and their revisit intentions. In this subsection, we will provide an overview of the literature that explores these attributes.

To start, Bufquin et al. (2017) argue that when consumers choose among restaurants, they consider various attributes, such as food quality, price, service quality, social aspects, and healthfulness. They found that among these attributes, only food quality and healthfulness have a positive impact on customer satisfaction and their behavioral intentions. This suggests that customers prioritize the quality of the food and perceive healthfulness as an important factor in their dining experience. Building upon the importance of healthfulness (Bufquin et al., 2017), Hargreaves et al. (2021) conducted research that demonstrates the positive correlation between physical health and the adoption of a vegetarian diet. Therefore, we aim at enhancing our understanding of the relationship between healthfulness and the success of vegetarian-friendly restaurants.

In addition to these attributes, another study by Gan et al. (2017) highlights the importance of context in evaluating restaurants. They argue that factors such as food, service, ambience, and price play a significant role in shaping customers' overall ratings of restaurants. Food quality again emerges as the most critical factor since it is the primary product of a restaurant. If the food fails to meet customers' expectations, it often leads to a negative review. In addition, service quality is identified as a crucial determinant of customer satisfaction and loyalty (Gan et al., 2017). Gan et al. (2017) also found evidence supporting the significant effect of context, ambience, and price on a customer's overall rating.

Moreover, Choi et al. (2022) conducted a study focusing on the vegetarian customer segment. They examined the effect of different attributes, including food quality, service quality, atmosphere, convenience, price, and vegetarian menu options, on customer satisfaction and behavioral intentions. They conclude that price and vegetarian menu options positively influence the satisfaction of vegetarian customers. However, Choi et al. (2022) also conclude that only offering vegetarian alternatives on the menu does not necessarily suggest the success of a restaurant, as this depends on different factors that affect customer attitude, customer satisfaction, and the behavioral intentions of general restaurant customers. This is in line with Garnett et al. (2019), who argue that increasing the availability of vegetarian meals on the menu, significantly increases vegetarian sales. However, they also found that serving more vegetarian meals has little impact on the overall sales of a restaurant.

In addition to Choi et al. (2022), who examine the effect of diverse attributes on one customer segment, Harrington et al. (2012) examined the influence of different attributes on customers from generation Y. They find that the attributes that positively influence these customers' experiences are the food quality, service quality, friendliness of staff, the atmosphere of the restaurant, and the speed of the service. In contrast, the attributes that most negatively influence these customers' experiences are service quality, speed of service, quality of food/drinks, friendliness of staff, and cleanliness. This study confirms an essential topic within marketing research, which is that customer segmentation affects the drivers behind restaurant success. Another interesting aspect from the study of Harrington et al. (2012), is that it segments across three types of restaurants, which are quick-service restaurants, casual dining, and fine dining. They argue, based on a study of Knutson (2000), that customers may have different needs and wants based on the

segment they are dining in. This essentially is what we are studying in this report, as we aim to understand the different drivers between two segments of restaurants: vegetarian-friendly and non-vegetarian-friendly.

Finally, Nilashi et al. (2021) propose a method to use customers' online reviews for customer segmentation and use this for predicting customer preferences for vegetarian-friendly restaurants in Bangkok. They argue that restaurants can improve their service quality and marketing strategy by detecting customer preferences through online reviews. These customer preferences are segmented based on four main attributes, which include food, value, atmosphere, and service (Nilashi et al., 2021). Using the Latent Dirichlet Allocation (LDA) algorithm, they find that it is possible to effectively discover satisfaction dimensions from online reviews on the service quality of the restaurant. An important difference between this study and the study from Choi et al. (2022), is that they segment the 'general' consumers based on only vegetarian-friendly restaurants, whereas Choi et al. (2022) study one type of consumer (vegetarian consumers) based on a sample of 'general' restaurants. Moreover, Nilashi et al. (2021) argue that future research may use the proposed methodology to analyze online reviews and focus on discovering other important features. As an example, they argue that for evaluating vegetarian-friendly restaurants, features related to healthy food or vegetarian food may be considered. Finally, they argue that customer satisfaction should be a priority for restaurant owners, as this results in loyal customers. This is also supported by Choi et al. (2022) who find that increased customers satisfaction, leads to a higher willingness to pay and a higher intention to revisit the restaurants.

To summarize, the literature on the attributes influencing customers' satisfaction and behavioral intentions in the context of eWOM and restaurant choices reveals several key findings. Food quality, service quality, price, and ambience consistently emerge as critical drivers of customer satisfaction and behavior. Moreover, some literature includes social aspects (Bufquin et al., 2017), healthfulness (Bufquin et al., 2017), convenience (Choi et al., 2022), and the availability of vegetarian menu options (Choi et al., 2022; Garnett et al., 2019) as significant drivers customer satisfaction and behavior. Moreover, an important difference between the studies is that some do not segment the target customers and restaurants (Bufquin et al., 2017; Gan et al., 2017), some segment only the customers (Choi et al., 2022; Nilashi et al., 2021), and some segment both customers and restaurant (Harrington et al., 2022).

2.3 Effect of Social Network and Environment on Online Reviews

Based on the studies above, we argue that the drivers behind restaurant success differ depending on whether the restaurant is vegetarian-friendly or not. To elaborate on this, we study the Diffusion of Innovation Theory, which was developed by Rogers (1962). This theory explains how new ideas or products spread through a social system, and identifies factors that influence the rate and extent of adoption. The theory proposes that a new idea or technology spreads through a social system with five key stages based on different characteristics. The five stages include innovation, early adoption, early majority, late majority, and laggards. Where the innovation stage involves the creation of a new idea or technology, it evolves further into the social system to more people who are adopting the innovation. Therefore, the last stage, the laggards, are a small group of people who resist change and thus do not adopt the new idea or

technology until necessary.

Based on these stages, Díaz (2017) argues that a vegan diet is an innovation that is not (yet) diffused and is therefore in the early adoption phase. The Diffusion of Innovation Theory suggests that early adopters adopt the innovation out of appraisal for the innovations' attributes and include the 'opinion leaders' (Dearing, 2009). Following this stage, the late adopters adopt the innovation because they are often influenced by their social networks, including friends, family, and colleagues. This suggests that social factors, such as word-of-mouth recommendations, social norms, social identity, and the presence or absence of influential early adopters may influence the success of a vegetarian-friendly restaurant. Research by Janssen et al. (2016) supports this idea, finding that social norms and social identity play an important role in the adoption of a vegetarian diet. Therefore, customers may be more likely to visit and leave positive reviews for a vegetarian-friendly restaurant if they identify with a social group that values vegetarianism or if they feel that it aligns with their personal values. In addition, customers may be more likely to recommend a vegetarian-friendly restaurant to their social networks if they perceive it as socially responsible or if it aligns with their own identity.

This would indicate that restaurants have to rely on people with high prestige or many ties in the online community, which would encourage chatter and advance its direction in a way that favors the restaurant (Waite & Perez-Vega, 2017). However, Zhang and Liu (2019) argue that only a few researches focus on how a reviewer's online network influences a business's online review performance. Therefore, they have studied the impact of online social networks of reviewers on the volume and valence of a business's eWOM and they conclude that the number of ties that a reviewer has on Yelp positively affects the review volume of the restaurant, but also that it negatively affects the review valence of the restaurant in the next period.

Furthermore, we find literature arguing that a vegan or vegetarian diet can be considered as a social innovation, which is interesting because Phills et al. (2008) suggest that social innovation is the best way to obtain lasting social change. As discussed before, we find that the social environment of consumers influences the adoption of an innovation. Ploll et al. (2020) suggest that vegetarian or vegan diets can be considered as a social innovation. This is based on the definition of Phills et al. (2008) for a social innovation: "A novel solution to a social problem that is more effective, efficient, sustainable, or just than existing solutions and for which the value created accrues primarily to society as a whole rather than private individuals". They argue that vegetarianism and veganism present a novel solution to a social problem, referring to climate change, animal protection, land use, and industrialized food production. Although these diets are not new, they argue that these diets have only recently become mainstream diets (Ploll et al., 2020). Based on the definition of Phills et al. (2008), Ploll et al. (2020) also argue that veganism and vegetarianism promote more sustainable practices and that the value that is created rather benefits the society, than only private individuals. Moreover, Ploll et al. (2020) consider the increase in vegetarianism and veganism as part of a social movement, which is a movement that typically involves collective actions and behaviors toward the same goal.

Finally, Ploll et al. (2020) conclude that the convictions and motives of vegetarians and vegans contribute to qualifying these diets as social innovations. They argue most of these consumers are motivated by wide social and environmental problems and animal concerns (Ploll

et al., 2020). As a result of these increasing concerns regarding the negative effects of meat consumption, the number of people adopting a vegetarian diet is increasing. The most common concerns include animal welfare, human health, and the environment (Godfray et al., 2018). Literature suggests that group values, norms, and behavior have a great influence on the adoption of consumer attitudes and behaviors (Raggiotto et al., 2018).

This is in line with the beliefs of Gonera et al. (2021). They segment customers based on combining the concept of social innovation with the Diffusion of Innovation Theory. Their segments include Flexitarians, Open to vegetarian foods, Piscivores, Processed food eaters, Omnivores, Conservatives, and Carnivores. Besides having different eating patterns and attitudes towards food, these segments show significant differences in demographic variables such as sex, age, education, household, and geographical affiliation. In line with Harrington et al. (2012), who argued that the effect of different restaurant attributes on customer satisfaction is affected by the type of customer, this study finds that the different customer segments selected different attributes as drivers for food purchase (Gonera et al., 2021). Although all segments choose taste as the most important driver, some segments are opposites regarding their values in food purchases. For example, where Flexitarians highly value 'animal welfare' and 'environment/climate', the Carnivores highly value 'price', 'easy and fast preparation', and a 'familiar product'. This study is interesting because it demonstrates how these different segments are at different stages of the innovation adoption curve (Gonera et al., 2021).

2.4 Effect of Sentiment on Restaurant Success

In addition to social networks, environmental factors, and various restaurant attributes, the impact of a review's sentiment on the success of a business has been extensively studied. It is theorized that textual reviews not only provide cognitive information for prospective consumers but also can influence their emotional states through the transmission of emotions from previous customers Lia, Wu and Mai (2019). This dual role emphasizes the significance of reviews as a powerful tool beyond providing information, impacting both cognitive and emotional aspects of consumer decision-making processes. Furthermore, (Lia et al., 2019) argue that the sentiment expressed in a review reveals consumers' deep thoughts, which are essential in forming perceptions and credibility of online reviews. Considering the importance of sentiment as a variable, many studies have investigated its effect on sales prediction. For example, (Yuan, Xu, Li & Lau, 2018) provide empirical evidence supporting that sentiment is a crucial predictor of sales. Similarly, a study by (Hu, Koh & Reddy, 2014) reveals that the effect of ratings on sales is mostly indirect through sentiments, whereas the effect of sentiments on sales is primarily direct.

3 Conceptual Framework and Hypotheses

Following the previously discussed literature, this section formulates our conceptual framework in Figure 1, which schematically highlights the research hypotheses.

3.1 Topics Discussed in Reviews

Building upon the literature discussed in Section 2.2, this study acknowledges the influence of different customer segments on consumer behavior. Furthermore, considering the variations in customer preferences based on the dining segment, as highlighted by Knutson (2000), the research focuses on identifying the key drivers of success for two distinct restaurant segments: vegetarian-friendly restaurants and non-vegetarian-friendly restaurants. By employing a Latent Dirichlet Allocation (LDA) topic model, the study examines the topics discussed in the reviews of these restaurants. The hypothesis is formulated as follows:

> **Hypothesis 1 (H_1):** The topics (such as price, service, food, location, and ambience) influencing the number of reviews received by restaurants differ between vegetarian and non-vegetarian restaurant segments.

3.2 Sentiment of Reviews

Considering the previously discussed literature, which captures that the sentiment of reviews is a significant predictor of sales Hu et al. (2014); Yuan et al. (2018), we aim to investigate whether the influence of sentiment differs per restaurant segment. Therefore, our second hypothesis is formulated as follows:

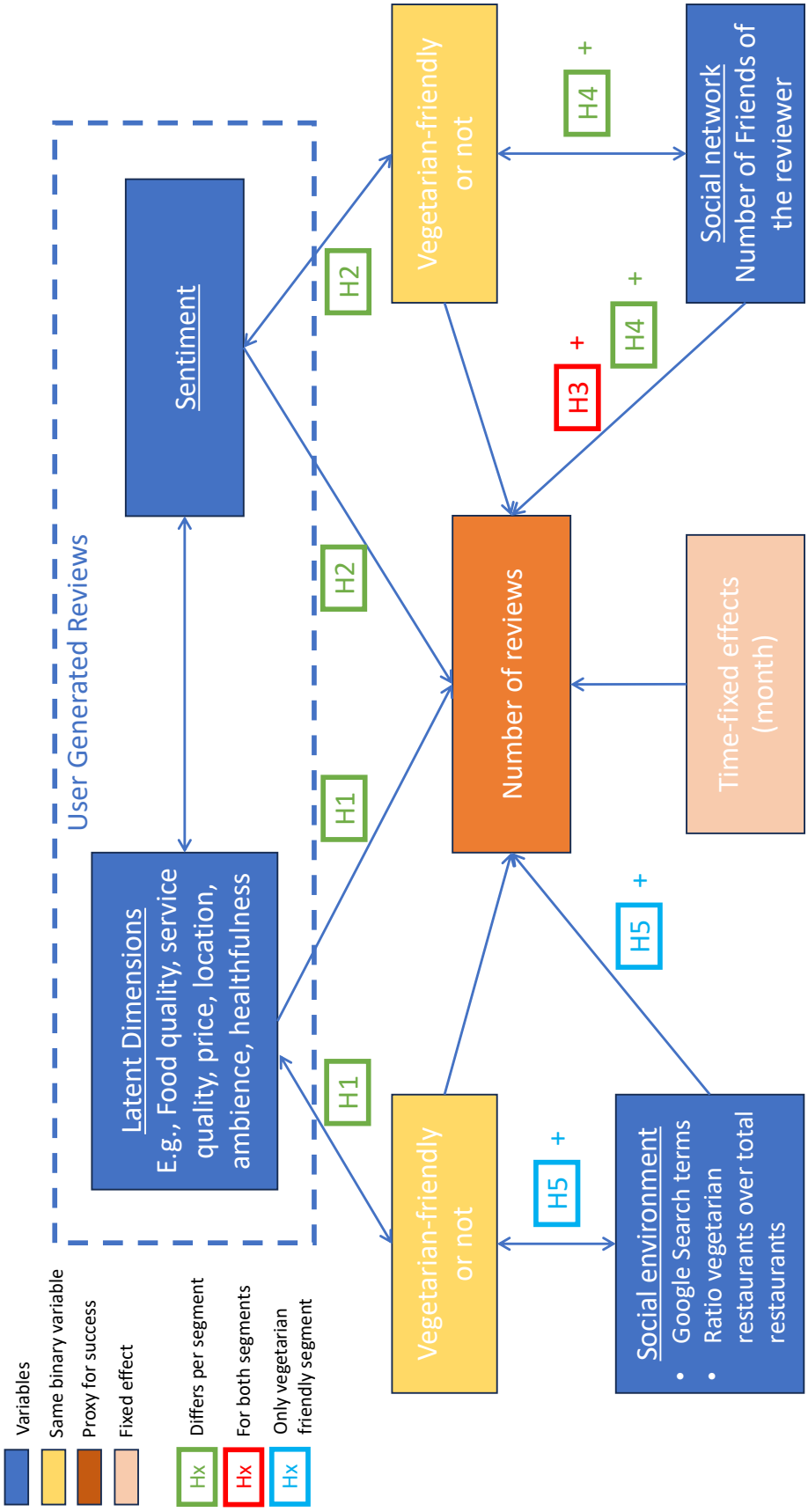
> **Hypothesis 2 (H_2):** The effect of sentiment on the number of reviews received by restaurants differs between vegetarian and non-vegetarian restaurant segments.

3.3 Social Networks

Based on the approach used by Liu and Zhang (2019) to measure reviewer centrality in online communities, we propose to examine the impact of the number of friends that the reviewer has within the online community on the number of reviews received by restaurants. Liu and Zhang (2019) argue that this is an effective way of measuring a person's centrality in an online community for two reasons. First, a reviewer who has a high number of friends is connected to numerous other reviewers, making their reviews visible to a larger network, Secondly, a reviewer with a larger network of friends tends to enjoy higher visibility and a more positive reputation (Liu and Zhang, 2019). Consequently, their reviews capture more attention and greater awareness. Therefore, a restaurant that receives reviews from reviewers with a larger network of friends is likely to attract more attention and visits, and subsequently increase the number of reviews. This leads to our third hypothesis, which is formulated as follows:

> **Hypothesis 3 (H_3):** The higher the average number of friends of reviewers who have reviewed a restaurant on Yelp, the higher the number of reviews of the restaurant.

Figure 1: Conceptual Framework



Additionally, we aim to explore the potential influence of the 'diffusion of innovation' theory on the success of vegetarian-friendly restaurants. This theory suggests that late adopters adopt the innovation because their social networks influence them. Drawing from Díaz's (2017) argument that a vegan diet is an innovation that is not yet fully diffused, we suggest that social factors may influence the success of vegetarian-friendly restaurants. Reviewers are more likely to recommend a vegetarian-friendly restaurant to their social networks if they perceive it as socially responsible or if it aligns with their own identity. To investigate this, we propose the following hypothesis:

> **Hypothesis 4 (H_4):** The effect of the average number of friends on the number of reviews received by restaurants differs between vegetarian and non-vegetarian restaurant segments.

3.4 Social Environment

Based on the literature suggesting that vegan and vegetarian diets can be considered social innovations, we propose to investigate the influence of the social environment on the number of reviews received by restaurants. To capture this effect, we introduce five additional variables that measure different aspects of the social environment. The first variable is the ratio between vegetarian-friendly restaurants and the total number of restaurants in a state, which reflects the prevalence of vegetarian options in the local dining scene. The remaining four variables are based on Google Trends data and represent the search interest for specific terms related to vegetarianism and veganism in each state. Based on these considerations, we hypothesize the following:

> **Hypothesis 5 (H_5):** When the social environment is more focused on the vegetarian and vegan diet, the number of reviews of vegetarian-friendly restaurants increases.
>
> **Hypothesis 5.1 ($H_{5.1}$):** The relationship between the ratio of vegetarian-friendly restaurants over total restaurants and the number of reviews received by vegetarian-friendly restaurants is positive.
>
> **Hypothesis 5.2 ($H_{5.2}$):** The relationship between the number of searches in a state for the term 'Vegetarian' and the number of reviews received by vegetarian-friendly restaurants is positive.
>
> **Hypothesis 5.3 ($H_{5.3}$):** The relationship between the number of searches in a state for the term 'Vegetarianism' and the number of reviews received by vegetarian-friendly restaurants is positive.
>
> **Hypothesis 5.4 ($H_{5.4}$):** The relationship between the number of searches in a state for the term 'Vegan' and the number of reviews received by vegetarian-friendly restaurants is positive.
>
> **Hypothesis 5.5 ($H_{5.5}$):** The relationship between the number of searches in a state for the term 'Veganism' and the number of reviews received by vegetarian-friendly restaurants is positive.

4 Methodology

In this section, we discuss the methods employed in our analysis. In Figure 2, the methods employed are schematically highlighted. Moreover, we discuss the method used for matching the vegetarian and non-vegetarian restaurants in Subsection 4.1. In Subsection 4.2, we discuss how we prepared and cleaned the reviews for analysis. In Subsection 4.3, we discuss the algorithm used to extract the topics, the parameters that need to be tuned for this LDA topic model, and how we label the topics. In Subsection 4.4, we discuss how we extract the sentiment scores per review. Finally, in Subsection 4.5 we discuss the models used to examine the factors influencing restaurant success and the metrics used to compare these models.

4.1 Restaurant Data

To ensure comparability between vegetarian and non-vegetarian-friendly restaurants in our study, we utilize a 1:1 nearest neighbor propensity score matching approach. This method helps to address potential selection bias and confounding effects that may arise in observational studies (Rubin, Donald B., 1973). Therefore, it is commonly used to balance the distribution of covariates between treatment and control groups in observational studies.

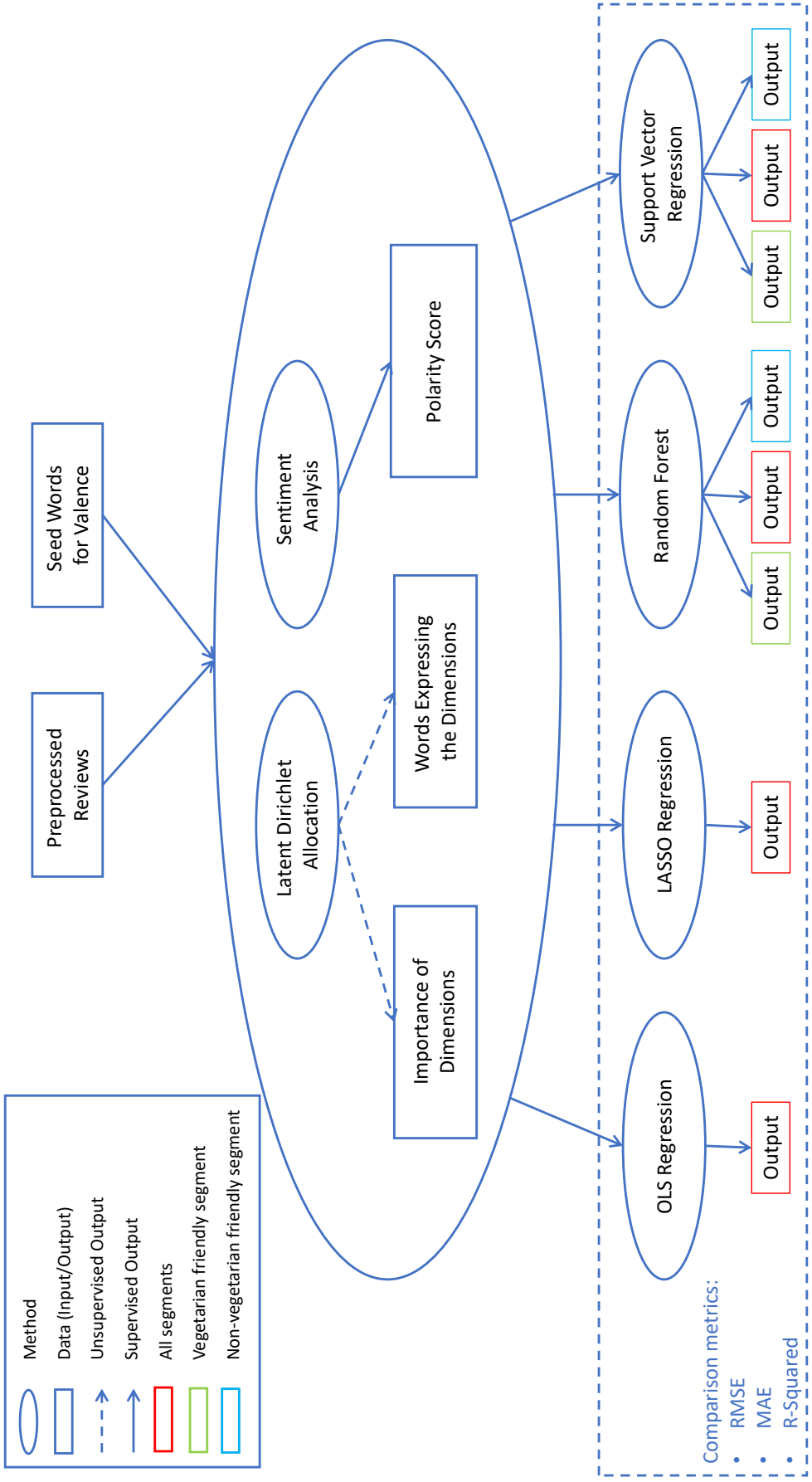
The first step in this approach involves estimating the propensity scores. For this, a logistic regression model is used to estimate the likelihood of a restaurant being vegetarian based on relevant covariates. We apply a three-step approach to select these relevant covariates. Firstly, we obtain binary variables from the categories, the state, and the city that the restaurant is in. Secondly, we only consider covariates that occurred in more than 4% of the restaurants. Lastly, we perform a t-test to compare the means of these covariates between the treatment (vegetarian restaurants) and control (non-vegetarian restaurants) groups. Only covariates with a statistically significant difference in means were included in the propensity score estimation.

After obtaining the propensity scores, the second step involves performing 1:1 nearest-neighbor matching in which each vegetarian-friendly restaurant is paired with a non-vegetarian-friendly restaurant with the closest propensity score. This matching process aims to create balanced pairs by ensuring similarity in propensity scores between the matched restaurants. To verify the effectiveness of the matching, we performed another t-test of the means of the covariates to assess the similarity between the matched groups.

To determine the nearest neighbors, a caliper distance measure was applied. The caliper represents a threshold that limits the acceptable difference in propensity scores between matched pairs. For this study, we set the caliper value to 0.005, ensuring that only closely matched pairs are included in the analysis. By employing a caliper, we aim to enhance the matches' quality and minimize the potential bias.

To conclude, the matching process resulted in a subset of paired vegetarian and non-vegetarian-friendly restaurants with comparable characteristics, reducing the impact of confounding factors. This subset was used for subsequent analyses and comparisons between the two groups.

Figure 2: Methodological Framework



4.2 Textual Data Preparation

After obtaining the subset of restaurants, the next step is to prepare the corresponding review data from the Yelp review data set for further analysis. In this section, we describe the process of cleaning and transforming the review text to ensure meaningful statistical analysis.

Analyzing the text within reviews presents challenges due to the unstructured nature of User Generated Content (UGC) and the need for standardization. To address these challenges, several steps are followed to prepare the review data:

1. *Conversion of uppercase letters to lowercase letters.* This helps to ensure consistency in text representation.

2. *Removal of excess white spaces and punctuation marks.* This helps to improve the readability and standardization of the text.

3. *Removal of stop words and special characters, including (most) smileys.* This was done because they do not carry significant meaning for analysis.

4. *Substitution of numbers, the smileys ':)' and ':(', and currency symbols for words.* This was done to ensure the analysis focuses on meaningful textual content.

5. *Removal of the least frequent words (occurring <1% of the time).* This helps to eliminate rare terms that may not provide significant insights.

6. *Stemming of the reviews using the 'SnowballC'-package.* Stemming is used to reduce words to their base or root form. This helps to simplify the analysis and improve the accuracy of text-based measurements.

By implementing these text-cleaning steps, the review data is prepared for further analysis. This process enhanced the data quality and relevance of the textual information, enabling subsequent statistical exploration, such as word frequency analysis, topic modeling, and sentiment analysis.

4.3 Topic Extraction

4.3.1 The LDA algorithm

After obtaining the cleaned reviews, the next step is to extract the different topics discussed in the reviews. Extracting qualitative topics from User Generated Content (UGC) presents several challenges (Tirunillai & Tellis, 2014). Firstly, consumers have a diverse range of ways to express their opinions about the different attributes of a restaurant. This leads to a large and skewed corpus of words, with thousands of unique terms. This characteristic poses a challenge known as the 'curse of dimensionality' (Bellman & Kalaba, 1959). Secondly, consumers tend to emphasize specific topics relevant to their experiences, meaning that not all topics are discussed in every review. As a result, when extracting topics from the reviews, the matrix representing the reviews by words becomes very large while most cells remain empty. Lastly, sentiments and adjectives used in reviews are context-specific and dependent on the attributes that are being evaluated. These challenges make it unreliable to employ traditional factor-analytic methods

due to convergence issues and overfitting (Blei, Ng & Jordan, 2001; Buntine & Jakulin, 2006). Therefore, a specialized methodology is required that can handle the high-dimensional and sparse nature of the data, as well as account for the context-specific nature of sentiments and adjectives across different product attributes within the same market.

In this study, we employ a probabilistic topic model called Latent Dirichlet Allocation (LDA) to discover topics within the reviews. LDA is an unsupervised machine-learning technique used in text mining to reveal hidden structures in large collections of text documents. By identifying patterns in the words used and grouping them into topics, LDA allows for a representation of the overall content of the documents (Blei et al., 2001).

The LDA algorithm consists of three main entities: the words, the documents (the restaurant reviews), and the corpus of documents. A *document* is a sequence of N words, denoted by $w = (w_1, w_2, ..., w_N)$. Each document is associated with a topic, and each word is associated with a topic. Following from this, the algorithm operates on a corpus of documents D, which is a collection of M documents and a diverse range of words, denoted by $(D = (w_{1,1}, w_{2,2}, ..., w_{N,M}))$. In addition, z_n denotes the topic from which the n-th word in the document w was generated. $\emptyset_j^{(w_N)}$ represents the conditional probability of observing word w_N given topic $z = j$: $= p(w|z = j)$. Moreover, $\theta_j^{(D)}$ represents the conditional probability of topic $z = j$ given document D: $p(z = j)$ (Andrzejewski & Zhu, 2009).

The LDA model utilizes the Dirichlet distribution to model the topic mixtures (θ) and the probability distributions of words within topics ($\emptyset$). By estimating the parameters of the Dirichlet distributions, the algorithm infers the underlying topics and their word distributions in the corpus. The Dirichlet distribution is a useful probability distribution defined on the simplex. It belongs to the exponential family, possesses finite-dimensional sufficient statistics, and is conjugate to the multinomial distribution (Blei et al., 2001). To summarize this, the LDA model involves the following relationships:

$$\theta \sim Dirichlet(\alpha), \tag{1}$$

where α denotes the hyperparameter for the document-topic Dirichlet distribution. This hyperparameter controls for the sparsity of the document-topic distribution, influencing the likelihood of topics being assigned to documents. This parameter will be tuned as discussed in Section 4.3.2.

Consequently, each document is associated with a topic z_n, and the following relationship is established:

$$z_n|\theta^{(D_n)} \sim Multinomial(\theta^{(D_n)}), \tag{2}$$

which expresses that the topic assignment for each word in a document depends on the document's topic proportions.

In a similar fashion, this relationship exists for words:

$$\emptyset \sim Dirichlet(\beta) \tag{3}$$

where β denotes the hyperparameter for the topic-word Dirichlet distribution. Therefore, β controls the sparsity of the topic-word distribution.

Following this, we have the following equation that suggests that each word is associated with a topic z_n:

$$w_n|z_n, \emptyset \sim Multinomial(\emptyset_{z_n}) \tag{4}$$

which expresses that the word generation process depends on the topic and the distribution of words associated with that topic.

To provide a more comprehensive insight into the LDA model and its underlying probability distributions, Equations 5, 6, 7 and 8 are introduced. Equation 5 describes the probability density of a k-dimensional Dirichlet random variable θ on the (k-1)-simplex. The (k-1)-simplex is a geometric object where a k-vector θ lies if $\theta_i \geq 0$ for all i and the sum of all θ_i equals 1 (Blei et al., 2001).

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{1=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}, \tag{5}$$

where α is a k-vector with components $\alpha_i > 0$ and where $\Gamma(x)$ is the Gamma function (Blei et al., 2001).

Given the parameters α and β, as previously described in Equations 1, 2, 3 and 4, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{6}$$

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$.

Moreover, Equation 7 (Blei et al., 2001) calculates the marginal distribution of a document by integrating over θ and summing over the topic assignments z.

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) (\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)) d\theta. \tag{7}$$

Finally, the probability of a corpus is obtained by taking the product of the marginal probabilities of single documents:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) (\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|zdn, \beta)) d\theta_d. \tag{8}$$

To conclude, we use the LDA algorithm because of its capability to unveil hidden patterns in the data, thereby eliminating the need for prior knowledge about the topics (Liu, Tang, Dong, Yao & Zhou, 2016). Moreover, it derives the sentiment or valence of words by considering the contextual usage of the words, recognizing that the same word can possess different meanings within different markets (Tirunillai & Tellis, 2014). Besides that, topic modeling is known for being easily interpretable. However, it is essential to note that the complexity and computational demands of the algorithm can pose challenges during implementation.

4.3.2 Tuning Parameters for the LDA Model

Before applying the LDA algorithm to our cleaned review data set, we need to tune two important parameters to ensure robust and informative results for our analysis. The first parameter is the number of topics (k). We define k as the total number of topics representing consumers' expressed experiences across all D reviews (Tirunillai & Tellis, 2014). For this, we employ a methodology based on the analysis of metrics. This approach involves finding the extremum values that optimize the selection of k. We consider both minimization and maximization techniques that are outlined in various papers. For minimization, we refer to the study by Arun, Suresh, Madhavan and Murty (2010) and Cao, Xia, Li, Zhang and Tang (2009), which provide a simple approach to identifying the optimal number of topics by minimizing the metrics. For maximization, we refer to the work of Deveaud, Sanjuan and Bellot (2014) and Griffiths and Steyvers (2004). These studies propose methods for maximizing metrics to determine the optimal number of topics or dimensions. By leveraging these approaches, we aim to identify the most suitable number of dimensions (k) for our analysis, ensuring a robust and informative representation of the underlying structure in the data.

The second parameter to tune is alpha (α), which influences the distribution of topics within a document, as explained in Section 4.3. By setting higher α priors for topics, the distribution of topics within a document becomes more uniform. In contrast, a lower α value encourages the inference process to allocate probability mass to a smaller number of topics in each document. Considering our study objective of identifying dominant or distinctive topics within each restaurant review, we choose a lower α ($\alpha = 0.1$). This leads to a sparser distribution of topics, in which each review is more likely to be associated with a few prominent topics. This approach helps highlight the most salient aspects or specific themes discussed in the reviews, which allows for a more focused analysis.

4.3.3 Labeling the Topics

In this subsection, we discuss two crucial tasks that are important after the dimensions are extracted. The first task is the selection of words that effectively distinguish the reviews associated with each dimension. This selection process ensures that we identify words prevalent in the corpus of reviews discussing a specific dimension while being infrequently mentioned in reviews that do not discuss that particular dimension (Tirunillai & Tellis, 2014). It is a crucial criterion as the methodology relies on the frequency of word occurrences in the reviews. By identifying these distinguishing words, we can accurately capture the essence of each dimension.

The second task involves assigning a suitable label to each dimension, reflecting the topic of discussion expressed across all the reviews related to that dimension. The labeling process is closely connected to selecting important words for each dimension. The chosen words play a significant role in determining the appropriate label or providing guidance for labeling the dimension effectively. These two tasks are interconnected, as selecting relevant words for a given dimension directly influences its label assignment (Tirunillai & Tellis, 2014). By carefully considering and carrying out both tasks, we ensure that the topics are well-defined, representing the specific topics of discussion present in the reviews expressing each dimension. To gain insights into the inferred topics, we use the term probabilities β, which was discussed in Section 4.3, as

an indicator for which words represent each topic. Specifically, we identify the 15 most likely terms within the distribution of each inferred topic. These terms provide valuable information about the dominant words associated with each topic. Finally, we use the 'Rank-1' method to re-rank the topics. This method counts how often a topic appears as the primary topic in a review. By employing the Rank-1 method to sort topics, we find that topics with distinct thematic coherence are positioned higher than others.

4.4 Sentiment Analysis

In addition to analyzing the topics in our review data, we aim to extract the sentiment expressed in the text. Sentiment analysis is a valuable technique for understanding and quantifying the sentiment expressed in textual data. For this analysis, we employ the polarity algorithm from the 'qdap'-package in R, which assigns numerical scores to indicate the sentiment polarity of a review. These scores range from -1 to 1, where -1 indicates a negative sentiment and 1 indicates a positive sentiment. The polarity algorithm follows a series of steps to determine the sentiment polarity of a review:

1. Adjust for negations, amplifiers, and length of a review.
2. Find words that appear in the polarity dictionary.
3. Consider the four preceding and two following words in the review.
4. Assign a score of +1 for a positive polarity word and a score of -1 for a negative polarity word.
5. Flip the sign of the score when a negation word is found in the preceding or following words.
6. Add (subtract) 0.8 for every positive (negative) amplifier word encountered.
7. Divide the final score by $\sqrt{NumberofWords}$

By following these steps, the polarity algorithm calculates a score that represents the sentiment polarity of the entire document. It is important to note that while polarity scores provide a quantitative measure of sentiment, they may not capture the full complexity of human emotions or the nuances of context. The algorithm relies on the assumption that the sentiment of its constituent words can adequately represent the sentiment of a document.

Leveraging polarity scores allows us to gain valuable insights into the reviews of the restaurants. We will calculate the average polarity score per restaurant, which will be used as a variable for estimating restaurant success. This information can help us understand the overall sentiment associated with each restaurant and potentially uncover relationships between sentiment and the number of reviews.

4.5 Models for Examining Factors Influencing Restaurant Success

In our analysis, we employ several models to examine the factors influencing restaurant success. Each model offers a different approach to understanding the relationships between the explanatory variables and the dependent variable, which is the log transformation of the number of

reviews and is used as a proxy for restaurant success. The first model is explained in Section 4.5.1, which is the Ordinary Least Squares (OLS) regression. The second model is explained in Section 4.5.2, which is the Least Absolute Shrinkage and Selection Operator (LASSO) regression. The third model is explained in Section 4.5.3, which is a Random Forest (RF) model, and the last model is explained in Section 4.5.4, which is a Support Vector Regression (SVR). Finally, the metrics used to compare the model's performance are introduced in Section 4.5.5.

4.5.1 Ordinary Least Squares (OLS) regression with Time Fixed Effects

The first model employed to examine the factors influencing restaurant success is a log-log transformed Ordinary Least Squares (OLS) regression model with time-fixed effects. This is a widely used linear regression model that assumes a linear relationship between the dependent variable and the explanatory variables. This means that a limitation of this model is that it does not deal well with data that is not normally distributed. Our data exhibited high skewness (see Section 5.3). Therefore, log-log transformation to address this issue is required. This transformation involves taking the logarithm of both the dependent variable and the explanatory variables before fitting the regression model. This approach helps to address the skewness and capture the multiplicative effects of the predictors on the outcome variable. Since the polarity score is already at a scale between -1 and 1, we decided not to log transform this variable.

In addition to applying the log-log transformation, we included time-fixed effects in our regression model. This approach helps mitigate potential bias caused by omitted time-invariant characteristics. By incorporating fixed effects, the estimated coefficients remain unbiased. The restaurants act as the panel identifier, controlling for time-invariant restaurant characteristics. The inclusion of time-fixed effects allows for the control of both seasonal fluctuations and year-to-year variations that are consistent across different restaurants. We have chosen to use the fixed effects regression over the random effects regression because after running a Hausman test, we failed to reject the null hypothesis ($p<0.001$), indicating that the preferred model is the random effects model. Moreover, we included interaction terms between the binary variable indicating whether a restaurant is vegetarian-friendly (1) or not (0) and the other variables. Finally, we added interaction terms between the topics and the polarity score to study how topics and sentiments influence the number of reviews. Based on this, The regression specification is as follows:

$$
\begin{aligned}
\log(Q_{it}) = \alpha_i &+ \beta_1 \log(topic1_{it}) + \beta_2 \log(topic2_{it}) + \beta_3 \log(topic3_{it}) + \\
&\beta_4 \log(topic4_{it}) + \beta_5 \log(topic5_{it}) + \beta_6 \log(topic6_{it}) + \beta_7 \log(topic7_{it}) + \\
&\beta_8 \log(topic8_{it}) + \beta_9 \log(topic9_{it}) + \beta_{10} \log(topic10_{it}) + \beta_{11} \log(ratio_{it}) + \\
&\beta_{12} \log(GT1_{it}) + \beta_{13} \log(GT2_{it}) + \beta_{14} \log(GT3_{it}) + \beta_{15} \log(GT4_{it}) + \\
&\beta_{16} PS_{it} + \beta_{17} \log(F_{it}) + \beta_{18} V_i + \beta_{19} V_i \log(topic1_{it}) + \\
&\beta_{20} V_i \log(topic2_{it}) + \beta_{21} V_i \log(topic3_{it}) + \beta_{22} V_i \log(topic4_{it}) + \\
&\beta_{23} V_i \log(topic5_{it}) + \beta_{24} V_i \log(topic6_{it}) + \beta_{25} V_i \log(topic7_{it}) + \\
&\beta_{26} V_i \log(topic8_{it}) + \beta_{27} V_i \log(topic9_{it}) + \beta_{28} V_i \log(topic10_{it}) + \\
&\beta_{29} V_i \log(ratio_{it}) + \beta_{30} V_i \log(GT1_{it}) + \beta_{31} V_i \log(GT2_{it}) + \\
&\beta_{32} V_i \log(GT3_{it}) + \beta_{33} V_i \log(GT4_{it}) + \beta_{34} V_{it} PS_{it} + \\
&\beta_{35} V_{it} \log(F_{it}) + \beta_{36} PS_{it} \log(topic1_{it}) + \beta_{37} PS_{it} \log(topic2_{it}) + \\
&\beta_{38} PS_{it} \log(topic3_{it}) + \beta_{39} PS_{it} \log(topic4_{it}) + \\
&\beta_{40} PS_{it} \log(topic5_{it}) + \beta_{41} PS_{it} \log(topic6_{it}) + \beta_{42} PS_{it} \log(topic7_{it}) + \\
&\beta_{43} PS_{it} \log(topic8_{it}) + \beta_{44} PS_{it} \log(topic9_{it}) + \beta_{45} PS_{it} \log(topic10_{it}) + \rho_t + u_i, \quad (9)
\end{aligned}
$$

where Q_it is the number of reviews per restaurant i at time t, α_i is the unknown intercept per restaurant, $topic1_{it}$ is the theta of topic 1 per restaurant i and time t, $topic2_{it}$ is the theta of topic 2 per restaurant i and time t, $topic3_{it}$ is the theta of topic 3 per restaurant i and time t,$topic4_{it}$ is the theta of topic 4 per restaurant i and time t, $topic5_{it}$ is the theta of topic 5 per restaurant i and time t, $topic6_{it}$ is the theta of topic 6 per restaurant i and time t, $topic7_{it}$ is the theta of topic 7 per restaurant i and time t, $topic8_{it}$ is the theta of topic 8 per restaurant i and time t, $topic9_{it}$ is the theta of topic 9 per restaurant i and time t, $topic10_{it}$ is the theta of topic 10 per restaurant i and time t, $GT1_{it}$ is the Google Trends variable for the search term 'Vegetarianism' per restaurant i and time t, $GT2_{it}$ is the Google Trends variable for the search term 'Veganism' per restaurant i and time t, $GT3_{it}$ is the Google Trends variable for the search term 'Vegetarian' per restaurant i and time t, $GT4_{it}$ is the Google Trends variable for the search term 'Vegan' per restaurant i and time t, PS_{it} is the average polarity score of a review at restaurant-level i and time-level t, F_{it} is the average number of friends that the review user has per restaurant i and time t, ρ_t is the time fixed effects, and u_i is the error term for unobserved restaurant-specific time-varying factors.

4.5.2 Least Absolute Shrinkage and Selection Operator (LASSO) regression model

The second model employed to examine the factors influencing restaurant success is a Least Absolute Shrinkage and Selection Operator (LASSO) regression. This model addresses the concern of potentially overfitting the data due to a high number of explanatory variables used in the OLS regression. While an advantage of the OLS regression is that it is highly interpretable, the LASSO regression also offers interpretability with the added advantage of variable selection. In this subsection, we explain the LASSO model and indicate the final LASSO specification.

The LASSO regression is a penalized regression method in which a 'penalty' is introduced to the regression model for having too many variables. This results in a reduced number of variables in the regression. Besides the LASSO regression model, important penalized regression models include Ridge and Elastic Net. However, the LASSO method can force variables to go to exactly zero and therefore only includes the most significant and relevant variables in the model. Therefore, we will employ the LASSO regression model in our study. It is important to note that the LASSO regression also possesses some limitations. One limitation is that when there is high collinearity among the explanatory variables, the LASSO tends to arbitrarily select one variable from a group of highly correlated variables, which leads to potential instability in the selected variables. Additionally, the choice of the tuning parameter lambda (λ) in the LASSO regression requires careful consideration to achieve the right balance between model simplicity and predictive accuracy.

To choose this optimal lambda value for the LASSO regression, we have performed cross-validation. For this, we have used the 'cv.glmnet' function in R, which automatically splits the data into several folds, trains the model on a subset of the data, and evaluates its performance on the remaining fold. It does this for different lambda values, which helps us to estimate the predictive performance of the model and select the lambda that yields the best results. After that, the LASSO regression achieves variable selection by minimizing the sum of squared residuals subject to the constraint that the sum of the absolute values of the coefficients is less than the chosen value of lambda. This LASSO model only leaves out the two interaction terms between whether a restaurant is vegetarian friendly and the search term 'Vegan' and the search term 'Veganism'. The LASSO regression is therefore specified as follows:

$$
\begin{aligned}
\log(Q_{it}) = {} & \alpha_i + \beta_1 \log(topic1_{it}) + \beta_2 \log(topic2_{it}) + \beta_3 \log(topic3_{it}) + \beta_4 \log(topic4_{it}) + \\
& \beta_5 \log(topic5_{it}) + \beta_6 \log(topic6_{it}) + \beta_7 \log(topic7_{it}) + \beta_8 \log(topic8_{it}) + \beta_9 \log(topic9_{it}) + \\
& \beta_{10} \log(topic10_{it}) + \beta_{11} \log(ratio_{it}) + d\beta_{12} \log(GT1_{it}) + \beta_{13} \log(GT2_{it}) + \beta_{14} \log(GT3_{it}) + \\
& \beta_{15} \log(GT4_{it}) + \beta_{16} PS_{it} + \beta_{17} \log(F_{it}) + \beta_{18} V_i + \beta_{19} V_i \log(topic1_{it}) + \beta_{20} V_i \log(topic2_{it}) + \\
& \beta_{21} V_i \log(topic3_{it}) + \beta_{22} V_i \log(topic4_{it}) + \beta_{23} V_i \log(topic5_{it}) + \beta_{24} V_i \log(topic6_{it}) + \\
& \beta_{25} V_i \log(topic7_{it}) + \beta_{26} V_i \log(topic8_{it}) + \beta_{27} V_i \log(topic9_{it}) + \beta_{28} V_i \log(topic10_{it}) + \\
& \beta_{29} V_i \log(ratio_{it}) + \beta_{30} V_i \log(GT1_{it}) + \beta_{31} V_i \log(GT3_{it}) + \\
& \beta_{32} V_{it} PS_{it} + \beta_{33} V_{it} \log(F_{it}) + \beta_{34} PS_{it} \log(topic1_{it}) + \beta_{35} PS_{it} \log(topic2_{it}) + \\
& \beta_{36} PS_{it} \log(topic3_{it}) + \beta_{37} PS_{it} \log(topic4_{it}) + \beta_{38} PS_{it} \log(topic5_{it}) + \\
& \beta_{39} PS_{it} \log(topic6_{it}) + \beta_{40} PS_{it} \log(topic7_{it}) + \beta_{41} PS_{it} \log(topic8_{it}) + \\
& \beta_{42} PS_{it} \log(topic9_{it}) + \beta_{43} PS_{it} \log(topic10_{it}) + \rho_t + u_i, \qquad (10)
\end{aligned}
$$

where Q_it is the number of reviews per restaurant i at time t, α_i is the unknown intercept per restaurant, $topic1_{it}$ is the theta of topic 1 per restaurant i and time t, $topic2_{it}$ is the theta of topic 2 per restaurant i and time t, $topic3_{it}$ is the theta of topic 3 per restaurant i and time t,$topic4_{it}$ is the theta of topic 4 per restaurant i and time t, $topic5_{it}$ is the theta of topic 5 per restaurant i and time t, $topic6_{it}$ is the theta of topic 6 per restaurant i and time t, $topic7_{it}$ is

the theta of topic 7 per restaurant i and time t, $topic8_{it}$ is the theta of topic 8 per restaurant i and time t, $topic9_{it}$ is the theta of topic 9 per restaurant i and time t, $topic10_{it}$ is the theta of topic 10 per restaurant i and time t, $GT1_{it}$ is the Google Trends variable for the search term 'Vegetarianism' per restaurant i and time t, $GT2_{it}$ is the Google Trends variable for the search term 'Veganism' per restaurant i and time t, $GT3_{it}$ is the Google Trends variable for the search term 'Vegetarian' per restaurant i and time t, $GT4_{it}$ is the Google Trends variable for the search term 'Vegan' per restaurant i and time t, PS_{it} is the average polarity score of a review at restaurant-level i and time-level t, F_{it} is the average number of friends that the review user has per restaurant i and time t, ρ_t is the time fixed effects, and u_i is the error term for unobserved restaurant-specific time-varying factors.

4.5.3 Random Forest (RF) model

The third model employed to examine the factors influencing restaurant success is the Random Forest model. This is a machine-learning technique that is widely used for both regression and classification tasks. It leverages ensemble learning by combining multiple decision trees to obtain more accurate and robust predictions, thereby creating a forest that consists of several decision trees.

This collection of trees is trained using the bagging method, which means first bootstrapping, and then aggregating the data. In the first part, the individual decision trees are built and during the construction of each decision tree, a random subset of the training data is sampled with replacement (bootstrap samples), and a subset of predictors is randomly selected at each split. These randomizations introduce diversity among the trees, reducing the risk of overfitting and improving the overall predictive performance. In the second part, the individual tree predictions are combined to obtain the final prediction.

Important to note is that the Random Forest model can capture non-linear relationships between predictor variables and the dependent variable without explicitly introducing interaction terms. Additionally, there is no need for logarithmic transformations of the predictor variables in this model. The predictor variables that are used for estimating the log of the number of reviews are the average topic theta values at the restaurant and time level, the ratio of vegetarian restaurants over total restaurants in the state, the different Google Trends search terms and their corresponding values (Vegetarianism, Veganism, Vegetarian, Vegan), the average polarity score per restaurant, the average number of friends that the review user has at restaurant and time level, and the binary variable whether the restaurant is vegetarian or not. Besides that, a categorical variable for the months is included to account for the Fixed Effects.

To ensure optimal performance of the Random Forest algorithm, we will tune the two important parameters, which include the number of trees (*ntree*) and the number of variables to use at each splitting node (*mtry*). Increasing the number of trees generally improves performance, but beyond a certain threshold, the performance gains diminish while computational requirements increase significantly. Therefore, we will determine an appropriate value for *ntree* that balances performance and computational efficiency. In our model, we use 1000 trees, which is more than the default number of trees (500) and should produce a more stable model. The *mtry* parameter controls the number of randomly selected predictors at each splitting node. The trade-off is that

smaller values reduce the correlation between individual trees, but may also lead to underfitting. We employ the 'tuneRF'-function in R to find the optimal value of *mtry*. This function evaluates different values and selects the one that minimizes the mean squared error. This method obtains a value of 5 for the Random Forest model that is run on the complete restaurant data set and a value of 2 for the two Random Forest models that are run on the subsets of only vegetarian-friendly or only non-vegetarian-friendly restaurants.

4.5.4 Support Vector Regression (SVR)

The last model employed to examine the factors influencing restaurant success is a Support Vector Regression (SVR). Similar to the Random Forest model, this is a machine-learning algorithm that is widely used for both regression and classification. Both models can capture non-linear relationships between the predictor variables and the target variable. Compared to the Random Forest model, SVR is particularly effective when using small training sets. Therefore, we will also include the SVR in our model performance comparison.

To explain how SVR works, we will start with a brief explanation of linear SVR. Similar to a linear regression, the equation of the line of a linear SVR is represented as:

$$y = w \cdot x + b, \tag{11}$$

where x represents the input variables, y represents the target variable, w represents the weight vector, and b represents the bias term. However, SVR differs from traditional linear regression by introducing the concept of support vectors and a *margin of tolerance*. The straight line in Equation 12 is the *hyperplane*. *Support vectors* are the data points closest to the hyperplane on either side and are used to define the boundary line. The objective of SVR is to minimize the error between the predicted values and the actual values, while also controlling the margin of tolerance. SVR allows for a certain degree of error (ϵ) within the margin, which enables the model to handle outliers and noise in the data. For that reason, the SVR model aims to satisfy the following conditions:

$$-\epsilon < y - w * x + b < \epsilon, \tag{12}$$

However, as mentioned earlier, our data is inappropriate for linear analysis. Therefore, we can use different kernel functions, which offer SVR the ability to capture non-linear relationships between the predictor variables and the target variable. By transforming the input variables into a higher-dimensional space, SVR can learn and model complex non-linear patterns. Commonly used kernel functions include polynomial, radial basis function (RBF), and sigmoid functions.

In our analysis, we compare the performance of the models using these different kernels. Based on this, we select the RBF function for further analysis. The RBF kernel maps the data into a high-dimensional space by computing the dot products and squares of all features in the data set. It extends on the basic idea of a linear SVR to capture non-linear relationships. The RBF calculates the similarity between data points based on their distance and uses this information to perform regression analysis. The objective radial basis function of the RBF kernel is defined as:

$$K(X_1, X_2) = exp(-\frac{\parallel X_1 - X_2 \parallel^2}{2\sigma^2}, \tag{13}$$

where σ is the variance of the model and can be used to tune the equation and $\parallel X_1 - X_2 \parallel^2$ is known as the Squared Euclidean Distance. Following from this equation, we introduce the parameter $\gamma = \frac{1}{2\sigma^2}$ and the equation results in:

$$K(X_1, X_2) = exp(-\gamma \parallel X_1 - X_2 \parallel^2), \tag{14}$$

After this, our task is to tune three parameters: γ, ϵ, and Cost (C). ϵ determines the width of the decision boundary around the hyperplane, while C controls the trade-off between achieving a low training error and a low testing error. In other words, C adjusts how hard or soft the margin is by defining the weight of how much samples inside the 'soft margin' contribute to the overall error. A low value of C penalizes the samples inside the margins less than a higher C. The value of C can range between 0 and 1, where 0 indicates that the model does not penalize the samples inside the margins at all and 1 means that the model has a 'hard margin', meaning it penalizes every sample inside the margins. We utilize the 'SVR'-function from the 'fdm2id'-package in R to retrieve the optimal values of these three hyperparameters, which are $\gamma = 0.111, \epsilon = 0.3$ and $C = 4$ for both data sets.

Similar to the LASSO regression and the OLS regression, it is necessary to normalize the predictor variables for the SVR model. Therefore, we use the logarithmic transformation of the predictor variables. The predictor variables that are used for estimating the log of the number of reviews are the logistic transformations of the following variables: the average topic theta values at the restaurant and time level, the ratio of vegetarian restaurants over total restaurants in the state, the different Google Trends search terms and its corresponding values (Vegetarianism, Veganism, Vegetarian, Vegan), the average polarity score per restaurant, the average number of friends that the review user has at restaurant and time level, and the binary variable whether the restaurant is vegetarian or not. Besides that, a categorical variable for the months is included to account for the Fixed Effects.

4.5.5 Comparing Performance of the Models

After introducing the four models employed to examine the factors influencing restaurant success, this subsection focuses on the metrics used to compare the performance of the different models. Three different metrics are utilized, including the Residual Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the R-squared.

The RMSE and MAE provide insights into the accuracy of the models by evaluating the errors between the actual value Y_i to the corresponding predicted value $\hat{Y}_i$. Lower values for both metrics indicate higher performance. To calculate the RMSE, the Mean Squared Error (MSE) is computed as the mean of the squared errors according to the following equation:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i = Y_i)^2 \tag{15}$$

Taking the square root of the MSE yields the RMSE:

$$RMSE = \sqrt{MSE} \tag{16}$$

A consequence of taking the square root of the MSE is that the RMSE gives more weight to larger errors. Therefore, it is more sensitive to outliers. Therefore, we also include the MAE in our model comparison. While the RMSE emphasizes larger errors, the MAE measures the average error term without considering their direction. One characteristic of the MAE is its linearity, which means that each individual difference contributes equally to the mean. The equation for MAE is as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{Y}_i = Y_i| \tag{17}$$

The final metric that is used for the model comparison is the R-squared. This value indicates the proportion of the variance in the dependent variable that can be explained by the predictor variables in the model. Therefore, a higher value indicates greater accuracy and overall better performance. The equation of R-squared is as follows:

$$R^2 = 1 - \frac{UnexplainedVariation}{TotalVariation} \tag{18}$$

To conclude, these metrics allow us to evaluate and compare the performance of the models in capturing the relationships between the predictor variables and the dependent variable. Lower values for RMSE and MAE indicate better performance, while higher values for R-squared signify a better performance.

Table 1: Variables contained in the *business.json* file

Variables	Description	Type
business_id	Unique business identifier	String
name	Business name	String
address	Address of the business	String
city	City that the business is in	String
state	State that the business is in	String
postal_code	Postal code of the business	String
latitude	Latitude of the business	Numeric
longitude	Longitude of the business	Numeric
stars	Average star rating	Numeric
review_count	Total review count	Integer
is_open	Whether the business still open	Integer
categories	Categories that the business has tagged	String

Table 2: Variables contained in the *review.json* file

Variables	Description	Type
review_id	Unique review identifier	String
user_id	Unique user identifier	String
business_id	Unique business identifier	String
stars	Star rating of the review	Integer
useful	Number of useful votes	Integer
funny	Number of funny votes	Integer
cool	Number of funny votes	Integer
text	The review text	String
date	The date of the review	Date YYYY-MM-DD

5 Data

The data for this study is obtained from Yelp, an online directory for discovering local businesses that provide access to its data for personal, educational, and academic purposes. The available data is divided into data sets for businesses, reviews, users, check-ins, tips, and photos. This study utilizes the businesses, reviews, check-ins, and user data sets. The variables included in these data sets are presented in Table 1, Table 2, Table 3, and 4. In the remainder of this section, we discuss the restaurant data that is used (Subsection 5.1), the review data that is used (Subsection 5.2), why we use logarithmic transformation for our analysis (Subsection 5.3), and the additional data we retrieved for our control variables (Subsection 5.5).

5.1 Restaurant Data

This section describes the steps to prepare the Yelp business data set for the analysis. Initially, only the businesses categorized as 'restaurant' are retained, resulting in a decrease from 150.346

Table 3: Variables contained in the *checkin.json* file

Variables	Description	Type
business_id	Unique business identifier	String
date	Check-ins to the business	String which is a comma-separated list of timestamps for each check-in

Table 4: Variables contained in the *user.json* file

Variables	**Description**	**Type**
user_id	Unique user identifier	String
name	The user's first name	String
review_count	The number of reviews	Integer
yelping_since	When the user joined Yelp	Date YYYY-MM-DD
friends	User's friends as user_ids	Array of strings
Useful	Number of useful votes	Integer
Funny	Number of funny votes	Integer
Cool	Number of cool votes	Integer
Fans	Number of fans	Integer
elite	The years the user was elite	Array of integers
average_stars	Average rating of all reviews	Float
compliment_hot	Number of hot compliments received by the user	Integer
compliment_more	Number of more compliments received by the user	Integer
compliment_profile	Number of profile compliments received by the user	Integer
compliment_cute	Number of cute compliments received by the user	Integer
compliment_list	Number of list compliments received by the user	Integer
compliment_note	Number of note compliments received by the user	Integer
compliment_plain	Number of plain compliments received by the user	Integer
compliment_cool	Number of cool compliments received by the user	Integer
compliment_funny	Number of funny compliments received by the user	Integer
compliment_writer	Number of writer compliments received by the user	Integer
compliment_photos	Number of hot compliments received by the user	Integer

to 35,004 businesses. Next, the Yelp check-in data is utilized to filter the restaurants based on their oldest check-in date. To measure restaurant success based on the number of reviews, it is necessary to ensure the oldest check-in date for each restaurant is similar. Therefore, only restaurants with the oldest check-in date later than January 1st, 2016, are included. Following this step, the number of restaurants declines from 35,004 to 10,621 observations. From these restaurants, 490 restaurants identify with a 'vegan' and/or 'vegetarian' category (vegetarian-friendly). The goal is to match these 490 restaurants to another restaurant from this data set that does not identify with either 'vegan' or 'vegetarian' but which does have similarities.

To test whether the vegetarian and non-vegetarian-friendly restaurants are similar, we perform a t-test on a large set of variables. The variables include the average star rating, the days since first check-in, the cumulative number of reviews, different categories, different states, and different cities the restaurant is in. As we found many variables (1421 variables), we decided only to include the variables that occur in more than 4% of the restaurants. This way, we avoid the risk of overfitting and facing the 'curse of dimensionality' (Bellman & Kalaba, 1959), in which the available data becomes sparse, making it challenging to find meaningful matches.

In Table 5, the results from this set can be found. We find that 25 variables are not similar within the two segments of restaurants. These variables include the average star rating, the cumulative number of reviews, the categories food, Sushi Bars, Japanese, American (Traditional), Italian, Bars, Nightlife, Salad, Burgers, Seafood, Mexican, Fast Food, Pizza, Cocktail Bars, Chicken Wings, Cafes, Desserts and Chinese, the cities Indianapolis, New Orleans and Tampa, and the states Indianapolis (IN) and Arizona (AZ).

Table 5: t-test for equality of means for two segments of restaurants.

Variables	**Mean vegetarian friendly restaurants**	**Mean non-vegetarian friendly restaurants**	**t-stat**
Average star rating	4.24	3.86	14.35***
Days since first check-in	1657.27	1659.04	-0.06
Cumulative number of reviews	85.35	69.71	3.18**
Category: Food	0.51	0.32	8.2***
Category: Food Trucks	0.04	0.04	-0.17
Category: Sushi Bars	0.01	0.04	-10.95***
Category: Japanese	0.01	0.05	-10.015***
Category: American (Traditional)	0.06	0.13	-7.02***
Category: Italian	0.04	0.06	-2.08*
Category: Bars	0.08	0.19	-8.37***
Category: Nightlife	0.09	0.19	-8.05***
Category: Salad	0.18	0.08	5.81***
Category: Burgers	0.05	0.09	-3.50***
Category: American (New)	0.11	0.09	1.53
Category: Coffee and Tea	0.10	0.09	0.84
Category: Breakfast and Brunch	0.14	0.14	0.47
Category: Sandwiches	0.15	0.15	-0.48
Category: Event Planning and Services	0.06	0.05	0.65
Category: Seafood	0.04	0.09	-5.28***
Category: Mexican	0.07	0.12	-3.84***
Category: Fast Food	0.06	0.09	-3.24**
Category: Pizza	0.08	0.12	-3.04**
Category: Cocktail Bars	0.02	0.06	-4.31***
Category: Chicken Wings	0.03	0.07	-4.45***
Category: Cafes	0.10	0.07	2.33*
Category: Desserts	0.08	0.05	2.19*
Category: Chinese	0.02	0.04	-3.78***
City: Philadelphia	0.12	0.10	1.30
City: Nashville	0.07	0.06	0.26
City: Indianapolis	0.03	0.05	-3.25**
City: New Orleans	0.07	0.04	2.35*
City: Tampa	0.10	0.07	2.15*
City: Edmonton	0.05	0.05	0.73
State: PA	0.24	0.22	1.01
State: TN	0.09	0.10	-0.79
State: FL	0.21	0.20	0.62
State: IN	0.06	0.08	-2.53*
State: NJ	0.05	0.06	-0.69
State: LA	0.09	0.07	1.78
State: MO	0.05	0.06	-1.24
State: AZ	0.03	0.05	-2.44*
State: AB	0.06	0.05	0.53

Significance levels: * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 6: T-test for equality of means for two segments of restaurants after PSM.

Variables	**Mean vegetarian friendly restaurants**	**Mean non-vegetarian friendly restaurants**	**t-stat**
Average star rating	4.22	4.24	-0.66
Days since first check-in	1649.08	1585.43	1.61
Cumulative number of reviews	83.31	74.43	1.30
Category: Food	0.49	0.51	-0.79
Category: Food Trucks	0.03	0.06	-2.30*
Category: Sushi Bars	0.01	0.01	-0.82
Category: Japanese	0.01	0.02	-1.52
Category: American (Traditional)	0.06	0.05	0.73
Category: Italian	0.05	0.04	0.16
Category: Bars	0.09	0.06	1.24
Category: Nightlife	0.09	0.08	0.95
Category: Salad	0.15	0.12	1.15
Category: Burgers	0.05	0.05	-0.15
Category: American (New)	0.11	0.07	2.06*
Category: Coffee and Tea	0.09	0.12	-4.61*
Category: Breakfast and Brunch	0.14	0.15	-0.37
Category: Sandwiches	0.14	0.20	-2.10*
Category: Event Planning and Services	0.05	0.06	-0.71
Category: Seafood	0.04	0.03	0.35
Category: Mexican	0.08	0.06	0.64
Category: Fast Food	0.06	0.03	1.58
Category: Pizza	0.08	0.08	0.24
Category: Cocktail Bars	0.03	0.02	0.91
Category: Chicken Wings	0.03	0.03	0.19
Category: Cafes	0.11	0.11	-0.42
Category: Desserts	0.08	0.06	1.01
Category: Chinese	0.02	0.02	0.24
City: Philadelphia	0.13	0.12	0.49
City: Nashville	0.07	0.05	1.10
City: Indianapolis	0.03	0.04	-0.72
City: New Orleans	0.06	0.05	0.87
City: Tampa	0.08	0.09	-0.69
City: Edmonton	0.05	0.02	2.67**
State: PA	0.25	0.24	0.31
State: TN	0.10	0.10	0.11
State: FL	0.20	0.22	-0.57
State: IN	0.06	0.07	-0.40
State: NJ	0.05	0.06	-0.99
State: LA	0.08	0.08	0.25
State: MO	0.05	0.07	-0.96
State: AZ	0.03	0.02	0.41
State: AB	0.05	0.02	2.59**

Significance levels: * p<0.05, ** p<0.01, *** p<0.001

To reduce the dissimilarities between the two segments of restaurants, a Propensity Score Model (PSM) (Rosenbaum & Rubin, 1983) using a 1:1 nearest neighbor matching technique with a bandwidth of 0.005 is created, including the variables that were different across the two segments of restaurants. By implementing a common support criterion, we exclude treatment observations with propensity scores exceeding the maximum or falling below the minimum propensity score of the control group. Finally, this resulted in a data set of 928 restaurants, including columns for distance, weights, and subclass. After performing another t-test on the means of the two groups within this data set, we verify that the PSM has eliminated the imbalance between the two types of restaurants. The results are reported in Table 6. We conclude that using a PSM has significantly improved the balance between the two types of restaurants, and we will use this data set in the remainder of this analysis.

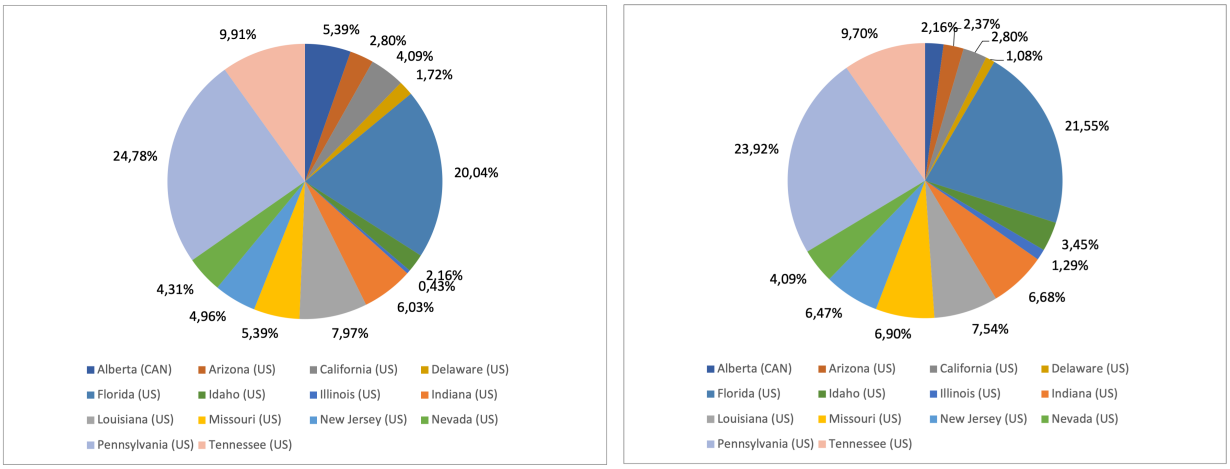
Following this, descriptive statistics are generated based on this subset of restaurants. Figure 3 presents two charts illustrating the distribution of vegetarian and non-vegetarian-friendly restaurants across different states. From this, we argue that the two data sets represent each other regarding the state distribution, as this distribution is comparable among the two types of restaurants. For example, the graphs reveal that most restaurants in both subsets are in Pennsylvania and Florida, followed by Tennessee and Louisiana. Moreover, it is essential to note that the data set includes only one Canadian state, Alberta, while the remaining states are located within the United States.

Furthermore, Figure 4 displays the most frequently represented categories, excluding the category 'food', in both restaurant segments. We notice that many categories are equally distributed over both types of restaurants. However, there are some notable differences. For example, the categories 'Salad', 'Juice Bars & Smoothies', 'American (New)', 'Gluten-Free', 'Nightlife', 'Bars', 'Specialty Food', 'Desserts', 'Mexican', 'Indian', 'American (Traditional0', 'Acai Bowls', 'Fast Food', 'Mediterranean', 'Wraps' and 'Health Markets' are represented significantly more often in the vegetarian-friendly restaurants. In contrast, the categories 'Sandwiches', 'Coffee & Tea', 'Food Trucks', 'Caterers', 'Asian Fusion', 'Delis', 'Vietnamese', 'Thai', and 'Barbeque' are represented significantly more often in non-vegetarian friendly restaurants.

In addition, Figure 5 displays the distribution of the restaurants over the most frequent cities. We find that for both vegetarian and non-vegetarian restaurants, the most common cities include Philadelphia, Tampa, New Orleans, Nashville, and Edmonton, and the other distributions are also comparable.

Finally, the number of monthly reviews for vegetarian and non-vegetarian-friendly restaurants is shown in Figure 6. From this, we suggest that the trend for both types of restaurants is quite similar. For example, both types of restaurants had a severe decline in number of reviews in April 2020, which can be explained by the COVID-19 pandemic that started around that time. Besides this similar trend, we find that vegetarian-friendly restaurants received more reviews than non-vegetarian-friendly restaurants in most months before the COVID-19 outbreak, except for the period between September, 2017, and January, 2018. However, after the COVID-19 outbreak, the graph illustrates that the number of reviews of vegetarian-friendly restaurants is very similar to that of non-vegetarian-friendly restaurants.

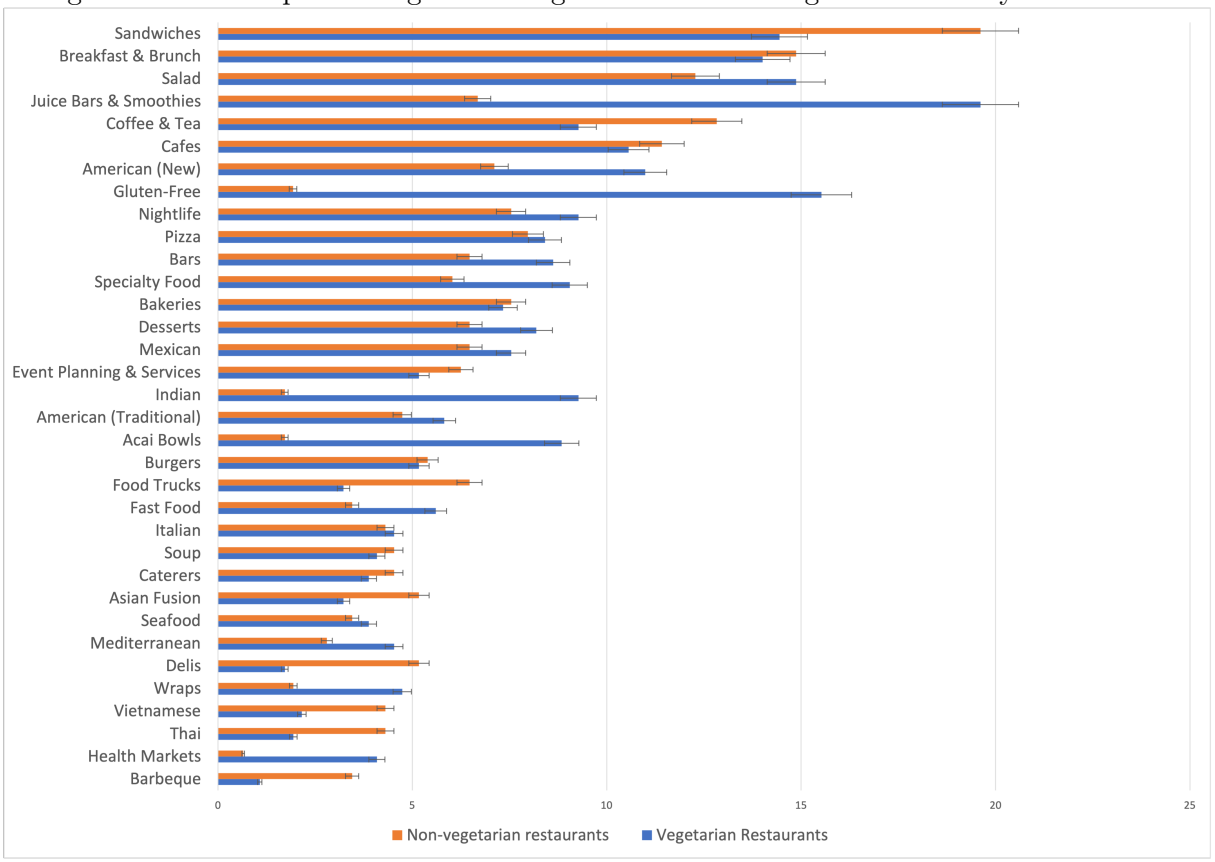
Figure 3: States distribution across restaurants.



(a) Vegetarian friendly restaurants

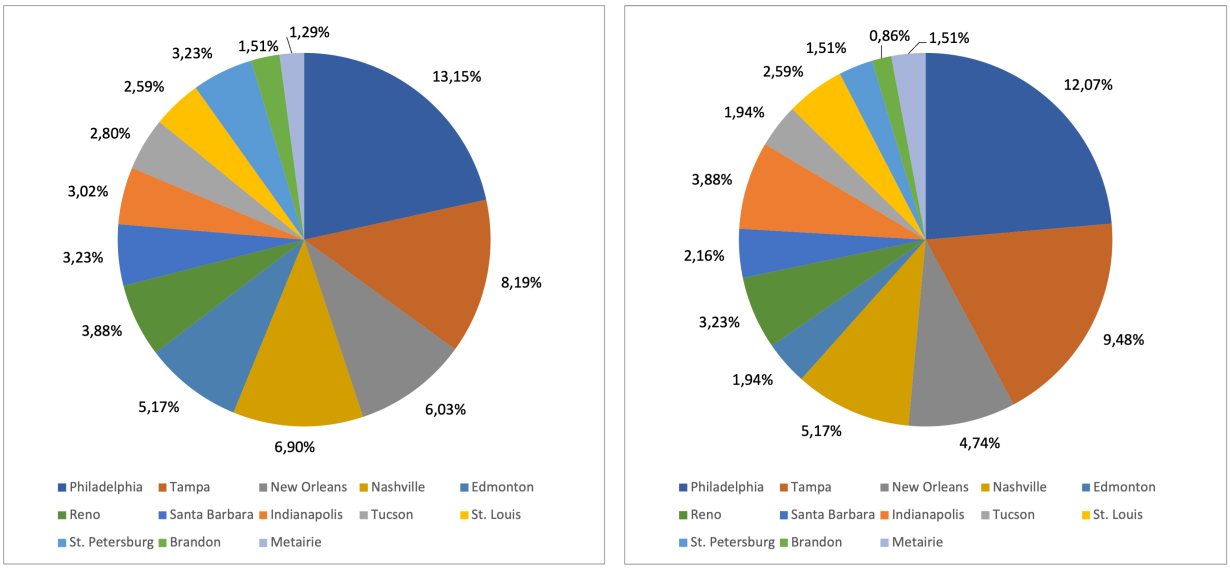
(b) Non-vegetarian friendly restaurants

Figure 4: Most frequent categories in vegetarian and non-vegetarian friendly restaurants.



Note: The error term corresponds to a range of 5% around the average value.

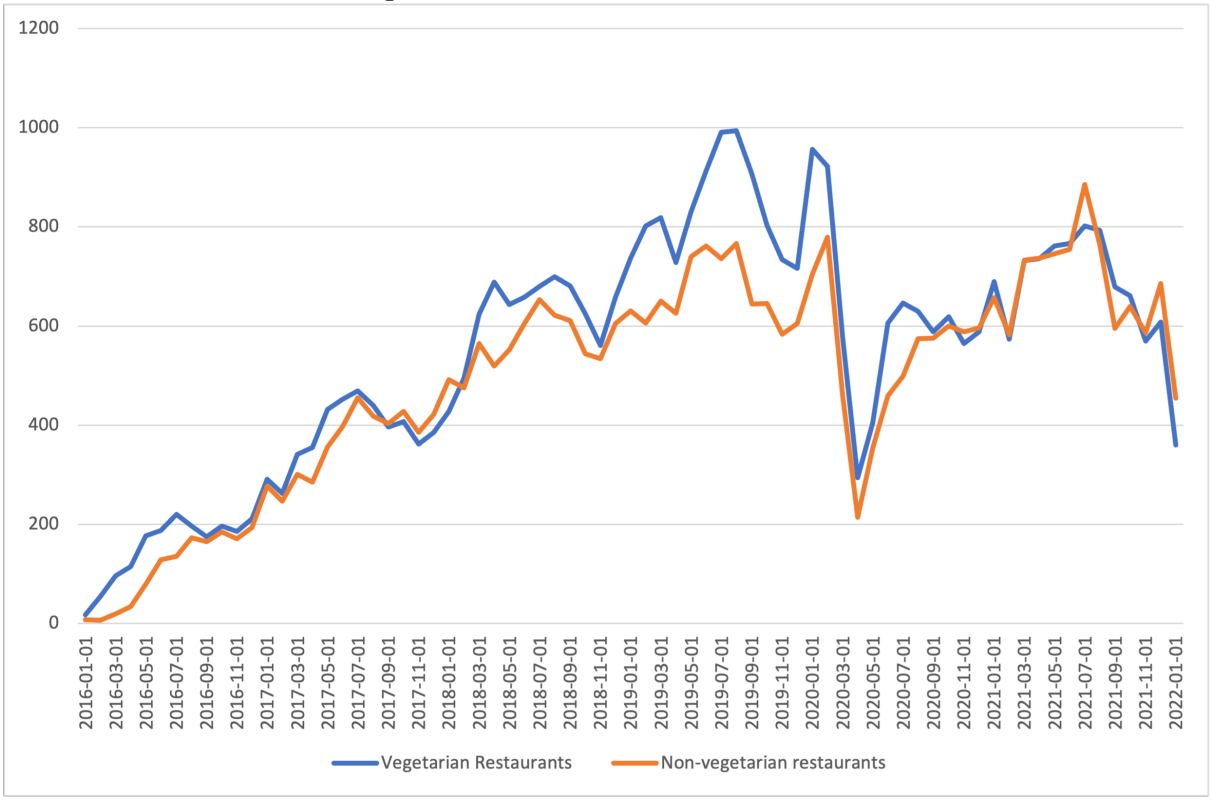
Figure 5: Cities distribution across restaurants.



(a) Vegetarian friendly restaurants

(b) Non-vegetarian friendly restaurants

Figure 6: Number of reviews over time.



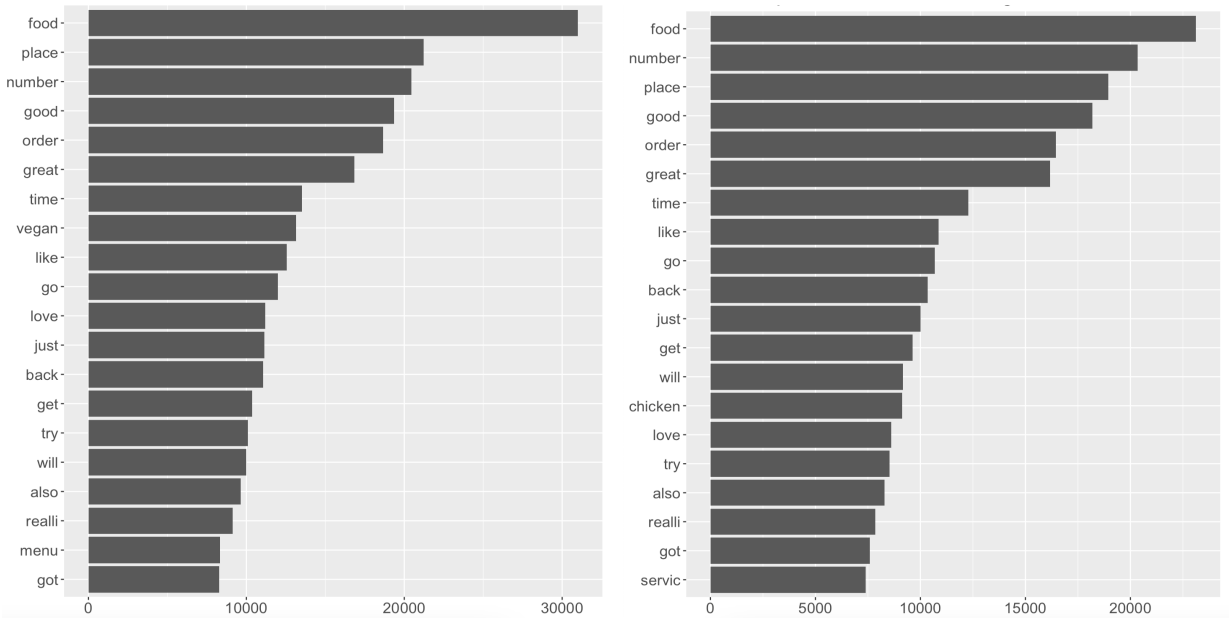
5.2 Review Data

After carefully considering the restaurants selected for our analysis, this subsection examines the reviews of this subset of restaurants. A subset of reviews specific to the 928 restaurants was extracted from the complete Yelp review data set, which consists of nearly 7 billion reviews.

The resulting review subset comprises a total of 75,664 reviews. However, although we have carefully considered the check-in date previously, some reviews are still identified that predate 2016 and subsequently removed to ensure a fair comparison of restaurant success. As a result, the total review data set includes 75,624 reviews. These reviews were posted between the 7th of January, 2016, and the 19th of January, 2022. After this, the reviews are cleaned to reduce the noise in the reviews, which is crucial for an accurate analysis. This is performed using the steps explained in Section 4.2.

After the text-cleaning process, 31,003 unique words are left for further analysis. Figure 7 displays the most frequent words for the two types of restaurants: non-vegetarian-friendly and vegetarian-friendly. We observe that the most common words are frequently used in both types of restaurants. However, a notable difference is kept in the words 'vegan' and 'menu', which are observed more frequently in vegetarian-friendly restaurants. In contrast, the words 'chicken' and 'servic' are used more frequently in non-vegetarian-friendly restaurant reviews.

Figure 7: Frequent words in reviews



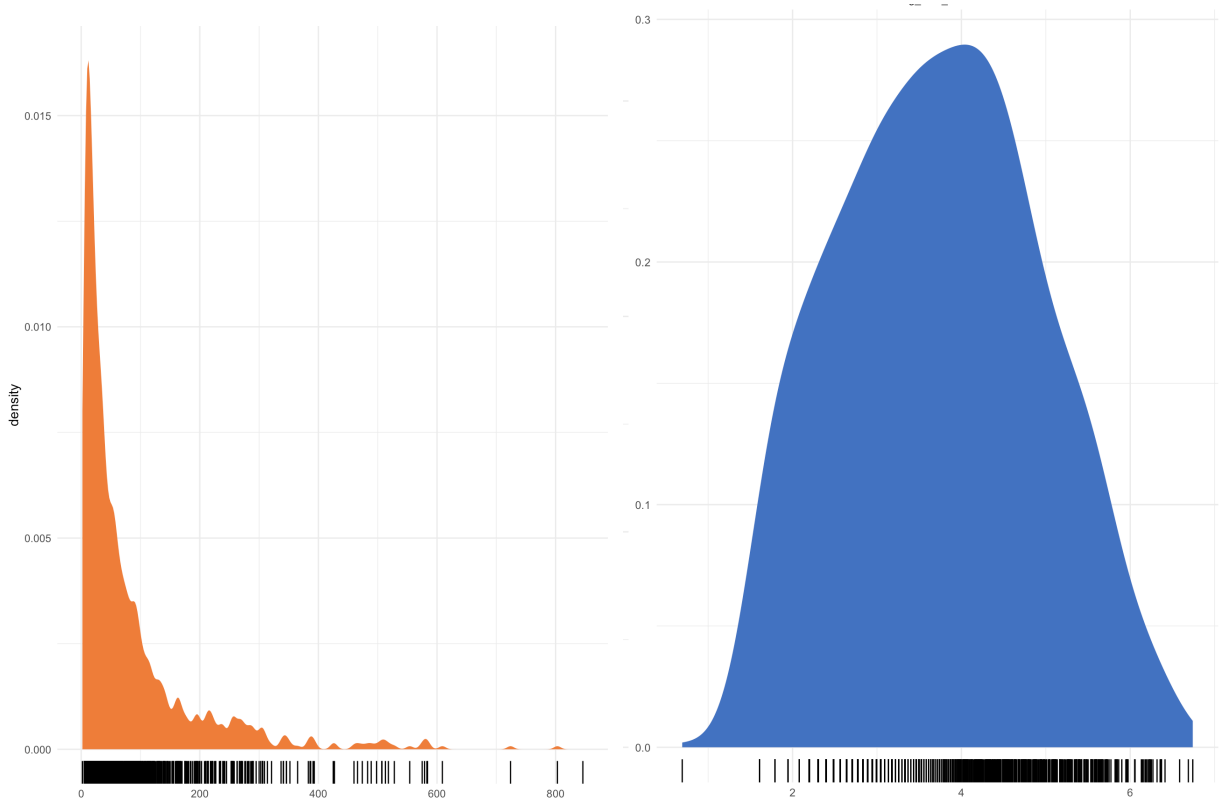
(a) Vegetarian friendly restaurants

(b) Non-vegetarian friendly restaurants

5.3 Logarithmic Transformation of Variables

In this section, we discuss why we use the logarithmic transformation of our dependent variable for all four models and why we also use this for the dependent variables in the OLS regression, LASSO regression, and the Support Vector Regression. This is because the variables exhibit a severely right-skewed distribution. Therefore, taking the logarithm can help reduce the extreme values' impact and make the distribution more symmetrical, improving the model's performance. The distribution of the dependent variable is illustrated in Figure 8, and we see that the logarithmic transformation is successful for this variable.

Figure 8: Distribution of our dependent variable: number of reviews.



(a) Number of Reviews

(b) Log(Number of Reviews)

5.4 Social Network Measure

The indicator used to measure a person's social network structure is relatively straightforward and conceptually basic. It is the count of friends that a reviewer has on Yelp, which refers to the number of individuals directly connected to them in the network (Wasserman & Faust, 1994). We first calculate the number of friends for each reviewer who has reviewed the same restaurant to obtain this measure. After that, we take all reviewers' average number of friends to get the restaurant-level measurement. The descriptive statistics of this variable can be found in Table 7.

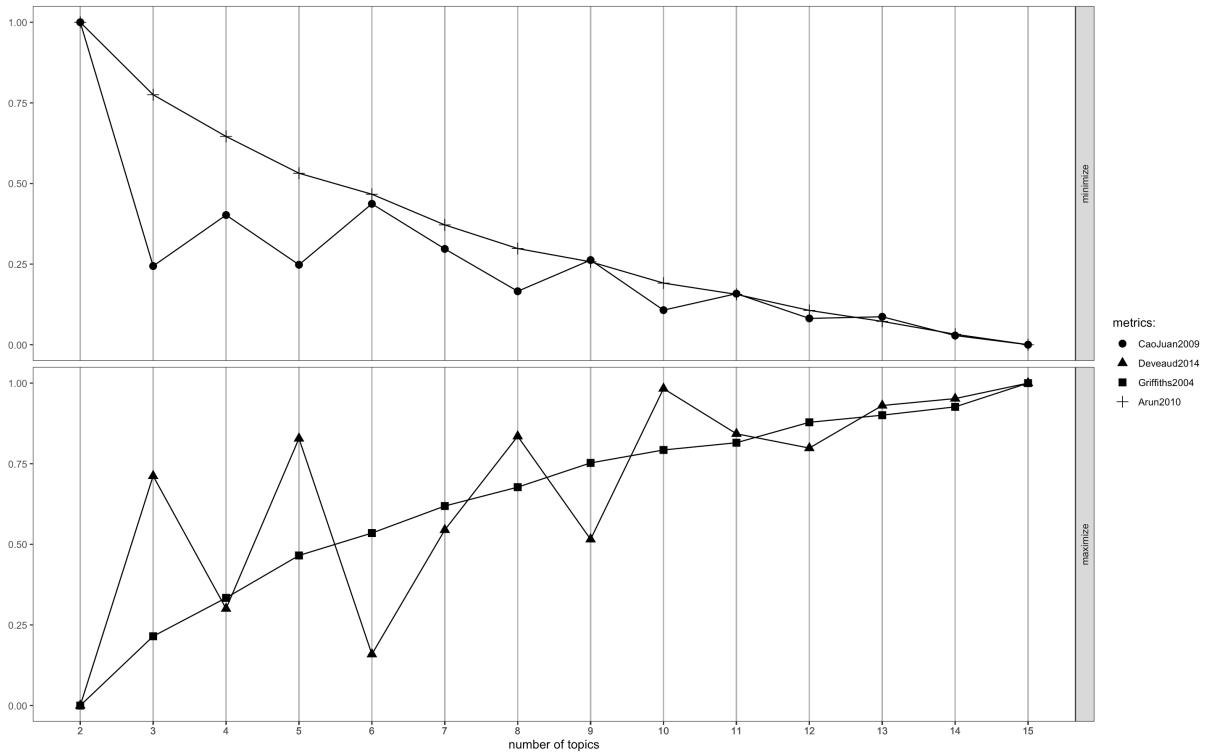
5.5 Social Environment Measures

In addition, five measures are included as proxies for the social environment, capturing the popularity of vegetarianism and veganism in the corresponding state. The first variable included is the ratio between the number of vegetarian-friendly restaurants and the total number of restaurants in a state, which reflects the prevalence of vegetarian options in the local dining scene. This ratio has been derived from all 35,004 restaurants in the Yelp data set. The remaining four variables are based on Google Trends data and represent the search interest for the terms 'Vegetarian', 'Vegetarianism', 'Vegan', and 'Veganism' in each state. The descriptive statistics from these five variables are included in Table 7.

Table 7: Descriptive statistics.

Variables	**N**	**Mean**	**Min**	**1st Quartile**	**3rd Quartile**	**Max**
Number of reviews:	75,624	81.49	2.00	17.00	96.25	846.00
Vegetarian restaurants	39,963	86.13	5.00	19.00	105.50	724.00
Non vegetarian restaurants	35,661	76.86	2.00	16.00	86.25	846.00
Average star rating:	75,624	4.22	1.25	3.93	4.62	5.00
Vegetarian restaurants	39,963	4.22	1.25	3.91	4.61	5.00
Non vegetarian restaurants	35,661	4.23	1.66	3.95	4.63	5.00
Number of Friends	21,770	116.4	1.00	8.00	127.20	6896.00
Ratio vegetarian restaurants over total restaurants per state	21,770	0.030	0.007	0.027	0.034	0.064
Google trends: Vegetarianism	21,770	48.90	35.00	48.00	51.00	63.00
Google trends: Veganism	21,770	60.90	45.00	53.00	63.00	80.00
Google trends: Vegetarian	21,770	50.70	34.00	46.00	55.00	72.00
Google trends: Vegan	21,770	54.00	39.00	44.00	62.00	80.00

Figure 9: 4 different metrics to obtain the optimal number of topics.



6 Results

In this section, we will elaborate on the results of the analysis. In Subsection 6.1, the results from our LDA model are explained, including the tuning parameters and the extracted topics. In Subsection 6.2, the results from the sentiment analysis are shown. After that, we compare the performance of our models in Subsection 6.3.1. In Subsection 6.3.2, we conclude by analyzing the results from the Random Forest model and the results from an additional regression model, including the most important features obtained from the Random Forest model.

6.1 Latent Dirichlet Allocation (LDA) Model

6.1.1 Tuning the LDA Model

Before running the LDA model, we must tune two parameters: the k number of topics and the alpha (α). As discussed in Section 4.3.2, we use two minimization and two maximization metrics to decide on an optimal number of topics k. Figure 9 displays the results from our minimization and maximization techniques, ranging from 2 to 15 topics. We aim at minimizing the CaoJuan2009 (Cao et al., 2009) and the Arun2010 (Arun et al., 2010) and we aim at maximizing the Deveaud2014 (Deveaud et al., 2014) and Griffith2004 (Griffiths & Steyvers, 2004). Based on these figures and considering that fewer topics increase the interpretability of our model, we decided to use ten topics ($k = 10$). In addition, we need to tune our second parameter, the alpha (α). For this study, we decide that a lower α is more suitable because the model will then identify the dominant or distinctive topics within each restaurant review. Therefore, we choose $\alpha = 0.1$.

Table 8: Topics and their assigned labels

Topic	Label	15 most-likely terms	Counts
1	Positive (Vegan) Dining Experience	food great place vegan love staff best will servic back recommend friendli time alwai good	15140
2	Ordering and Waiting	number order food time wait get ask just place back minut said came even take	8534
3	Delicious Burgers and Sandwiches	good chicken order fri burger sandwich got number chees taco try food place hot mac	8035
4	Tasty Thai and Indian Dishes	food chicken order good thai indian dish rice pho place flavor number try great spici	7453
5	Great Fresh Salads and Bowls	bowl salad fresh good place order food chicken love also great like get option got	7025
6	Cozy Coffee and Tea Place	place great coffe love nice tea also good drink shop can staff like park get	6759
7	Great Food and Service for Dinner	great food menu good servic dinner number dish back night restaur order server wine meal	6302
8	Good Pizza and Vegan Options	pizza good vegan donut place like love cream try number order crust flavor also great	5947
9	Prices and Positive Experience	like number food just good place dollarvalue price get tast realli much better can think	5312
10	Good Breakfast and Brunch	breakfast good great place back egg food got order try time sandwich brunch biscuit also	5114

6.1.2 Extracting the Topics

After choosing the number of topics, we compute the LDA model with an inference via 500 iterations of Gibbs sampling. The results from re-ranking the topics based on the 'Rank-1' method described in Section 4.3 can be found in Table 8. We have included the 15 most likely terms per topic and assigned a label that we believe is suitable based on the most likely terms.

We have included three randomly chosen reviews from our data set in Table 9. In Figure 10, these three examples are illustrated based on the theta (θ) distribution, which represents the probability distribution of each topic within each document, as explained in Section 4.3. This shows how the topics are distributed among the documents (reviews). To start, the reviewer of the first review states how the reviewer and their daughter had to wait for at least 10 minutes before ordering and, therefore, had a bad experience with the service. It makes sense that this review is mainly distributed in Topic 2: 'Ordering and Waiting'. However, as the review starts with a comment that the food is great, it is also distributed a little bit in topic 10: 'Good Breakfast and Brunch'.

Furthermore, the second review is very positive about their dinner experience at the restaurant. For example, they state that it is their 'FAVORITE restaurant in the city'. The review is mainly distributed in Topic 7: 'Great Food and Service for Dinner', which we can argue is accurate as the reviewer talks about a date night. Moreover, it is slightly distributed in Topic 6: 'Cozy Coffee and Tea Place', probably because the review mentions drinks, which is one of the most likely terms of this topic. Finally, the third review is a review regarding the sandwiches of a restaurant. The review is positive, although the reviewer warns customers that the prices are slightly higher. Therefore, the review is distributed mainly in Topic 9: 'Prices and Positive Experience'. It is also distributed slightly in Topic 8 ('Good Pizza and Vegan Options') and Topic 1 ('Positive (Vegan) Dining Experience'), which is probably caused by the reviewer talking about the 'healthy vegan type of sandwiches'.

In addition, Figure 11 displays the average probability of each topic occurring in a review at the restaurant segment level. Based on this figure, we would argue that the reviews from vegetarian-friendly restaurants are more likely to be regarding the topics 'Positive (Vegan) Dining Experience' and 'Great Fresh Salads and Bowls'. In contrast, we argue that the reviews from non-vegetarian restaurants have a higher likelihood to be regarding the topics 'Delicious Burgers and Sandwiches', 'Cozy Coffee and Tea Place', 'Great Food and Service for Dinner', and 'Good Breakfast and Brunch'. In line with this, Figure 12 is included, which shows the number of restaurants that, on average, have the highest probability of that topic occurring compared to other topics. This figure shows a similar pattern as Figure 11. However, Figure 12

Table 9: Example reviews.

Example	Review
1	Food is great. Service kinda sucks. Came in with my daughter today to kinda be treated like nothing. Sat & waited to order for at least 10 minutes with out as much as a song left acknowledgement that were standing there. Then after we order our food we literally wait 30 minutes for a sandwich. Then watch people who came in after us get their food before us.
2	Love the cozy, intimate atmosphere in here. My boyfriend and I have eaten twice at the bar for a date night, and the bartenders are lovely, attentive, and a great source of recommendations (which comes in handy in light of the lengthy menu). My two fav food items are the seafood stuffed mushrooms - truly a cannot miss - and the seafood paella (delicious, albeit slightly over salted). They also make a mean Paloma. While not my FAVORITE restaurant in the city, the food IS above average, and the drinks and atmosphere make up for whatever shortcomings in the food-department there may be.
3	[Restaurant] has very yummy sandwiches. They are fast and are always cooked perfect. You can get cold sandwiches hot sandwiches or even healthy vegan type of sandwiches. The only problem for some people may be the price of there sandwiches. They are a little on the higher end and I wouldn't say you get a lot for it. However in this case it is quality for quantity. I would recommend this to people who haven't had it before. What I mean when I say price is high it is roughly around 8−14 for about 8 inches. However it is very worth it and you should try it out if you haven't already. Plus they give you a 3 dollar off couple your next sandwich when you go and purchases a sandwich so keep that and you get get 3 dollars off the next time you go.

Figure 10: Example reviews per topic.

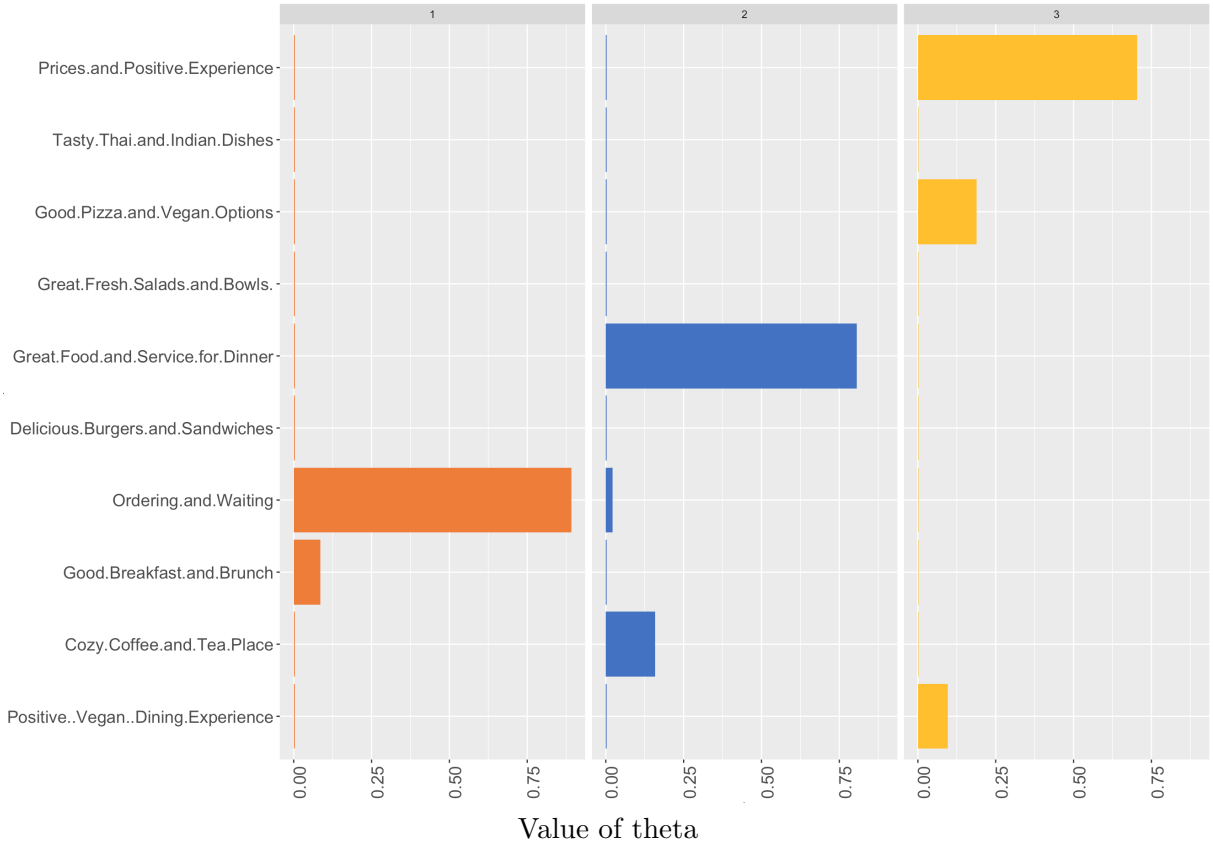
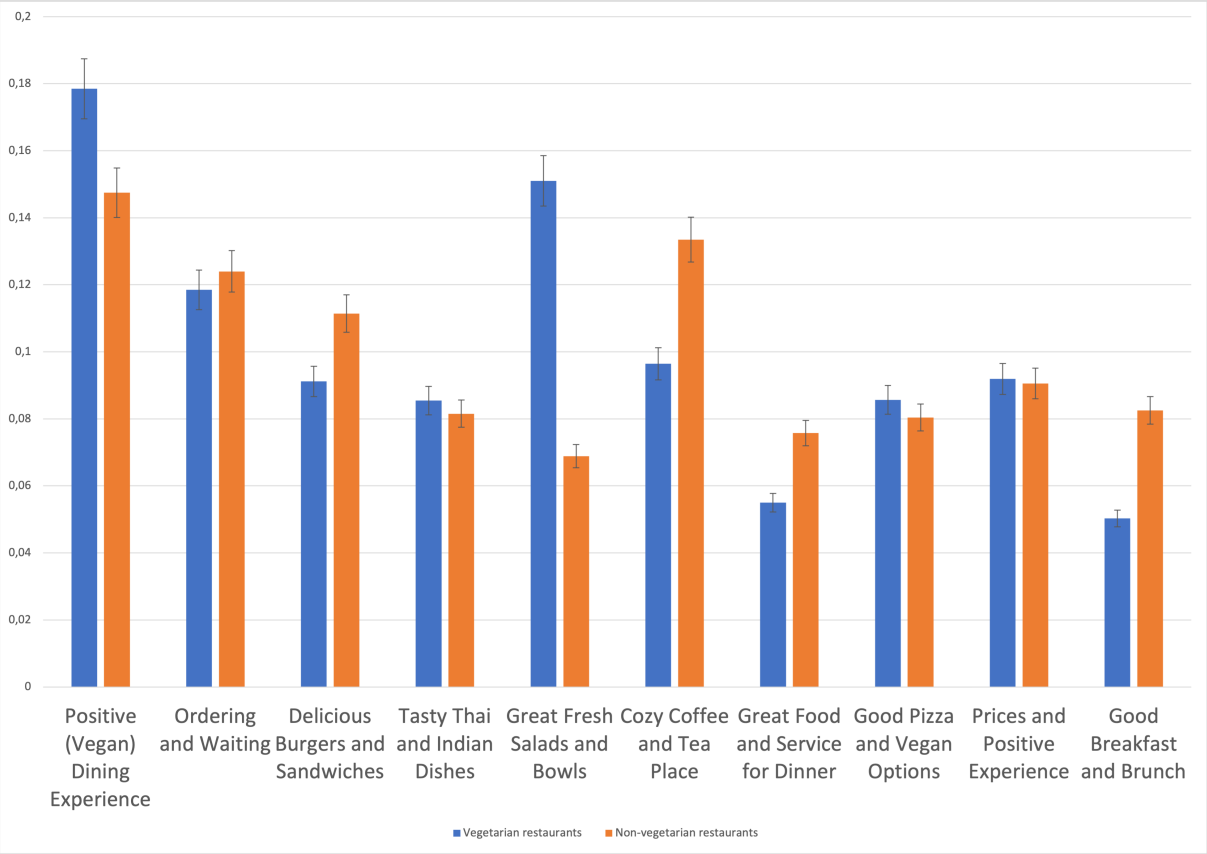


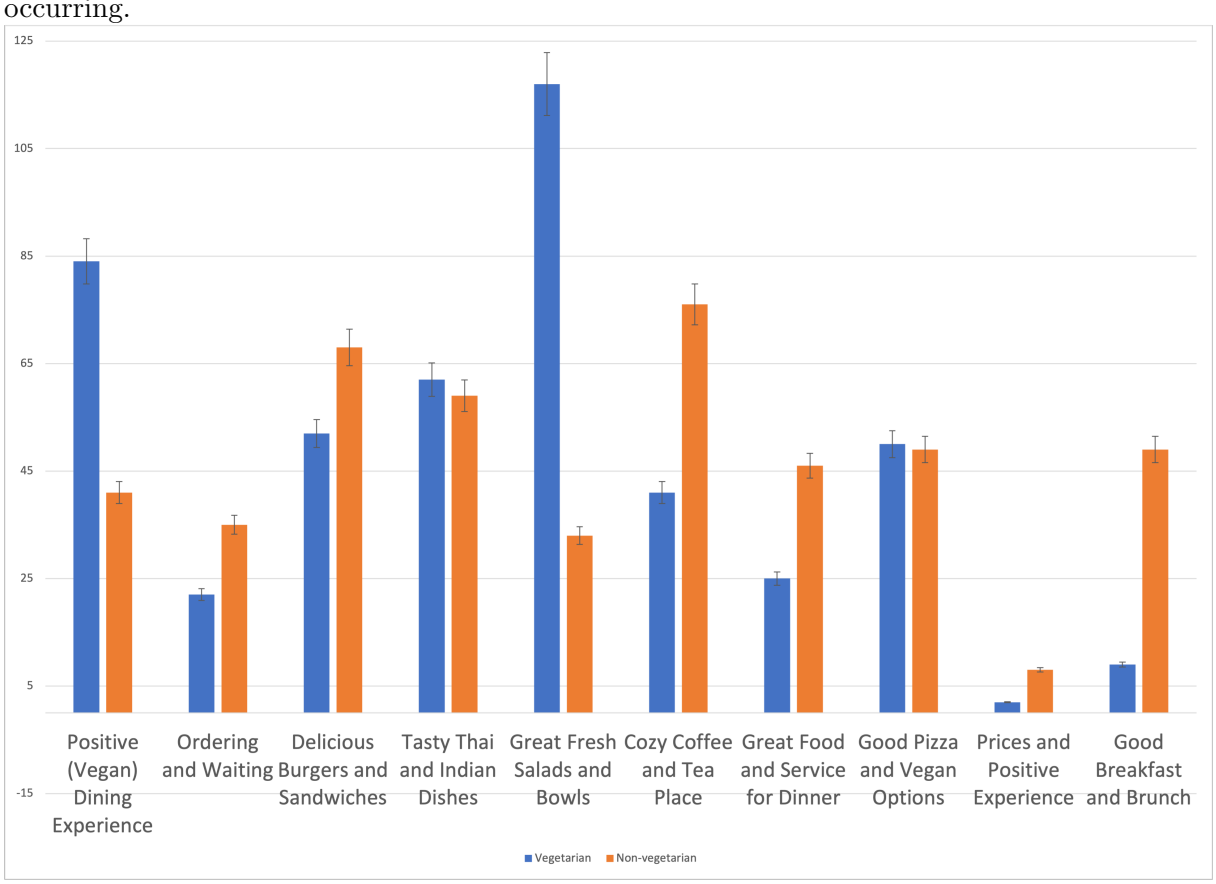
Figure 11: Average theta per topic.



Note: The error term corresponds to a range of 5% around the average value.

also suggests that non-vegetarian-friendly restaurants are more likely to be regarding the topics 'Ordering and Waiting' and 'Prices and Positive Experience'. Following from these topics, the average probability of each topic is derived per restaurant for our main analysis. The final data set contains 928 restaurants, of which 464 are vegetarian and 464 are non-vegetarian. We have included the descriptive statistics of the average theta per restaurant in Table 10.

Figure 12: Number of restaurants that, on average, have the highest probability of that topic occurring.



Note: The error term corresponds to a range of 5% around the average value.

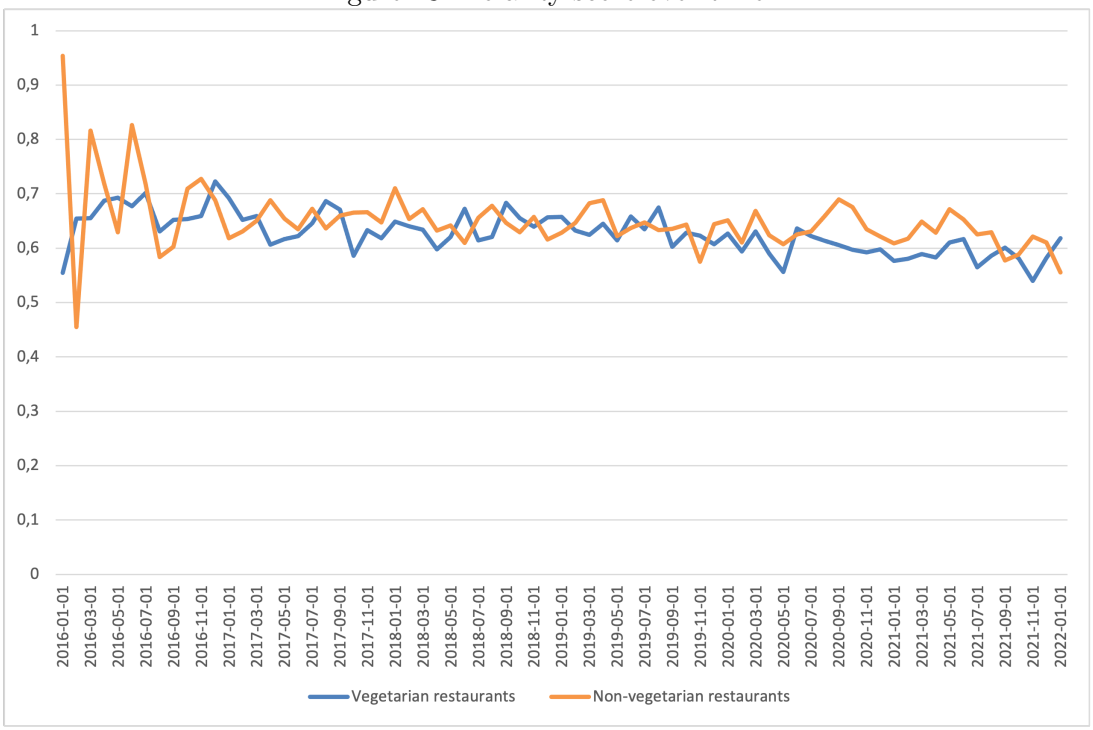
6.2 Sentiment Analysis Results

As explained in Section 4.4, we also used polarity scores to find the sentiment of reviewers towards a restaurant. The descriptive statistics of these sentiment scores are displayed in Table 10. In addition, Figure 13 displays the average polarity scores of vegetarian and non-vegetarian-friendly restaurants over time. From this, we find that these scores are relatively stable for both restaurant segments between 0.6 and 0.7, indicating an average positive sentiment towards the restaurants.

Table 10: Descriptive Statistics

Variables	**N**	**Mean**	**Min.**	**1st Quartile**	**3rd Quartile**	**Max.**
Topic 1: Positive (Vegan) Dining Experience	21,770	0.17	0.00	0.01	0.26	0.98
Topic 2: Ordering and Waiting	21,770	0.12	0.00	0.01	0.17	0.99
Topic 3: Delicious Burgers and Sandwiches	21,770	0.10	0.00	0.00	0.11	0.98
Topic 4: Tasty Thai and Indian Dishes	21,770	0.09	0.00	0.00	0.07	0.98
Topic 5: Great Fresh Salads and Bowls	21,770	0.10	0.00	0.00	0.11	0.98
Topic 6: Cozy Coffee and Tea Place	21,770	0.10	0.00	0.00	0.11	0.99
Topic 7: Great Food and Service for Dinner	21,770	0.07	0.00	0.00	0.07	0.99
Topic 8: Good Pizza and Vegan Options	21,770	0.08	0.00	0.00	0.06	0.98
Topic 9: Prices and Positive Experience	21,770	0.09	0.00	0.01	0.13	0.96
Topic 10: Good Breakfast and Brunch	21,770	0.07	0.00	0.00	0.06	0.98
Polarity Score	21,770	0.29	-0.64	0.19	0.39	0.97

Figure 13: Polarity score over time.



6.3 Results from Models

In this section, we present the main results from our analysis. We divided the data set into an 80% training data set and a 20% testing data set. In Subsection 6.3.1, we will compare the performance of several models: an OLS regression with time-fixed effects and interaction terms, a LASSO regression with time-fixed effects and interaction terms, a Random Forest model, and a Support Vector Regression. We evaluate their performance and identify the best-performing model. Based on this, Subsection 6.3.2 focused on the Random Forest model, analyzing its results and uncovering the key drivers behind the success of a restaurant. Finally, we interpret the results from a conclusive regression analysis with time and restaurant fixed effects, encompassing interaction terms between the key drivers and a binary variable indicating whether a restaurant is vegetarian-friendly.

6.3.1 Model Performance

In this subsection, we compare the performance of each model using the following metrics: Residual Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the R-squared metrics for each model. A detailed description of these metrics can be found in Section 4.5.5.

Table 11 displays the performance metrics of our models for predicting the log of the number of reviews per restaurant. The results indicate that the Random Forest model achieved the lowest RMSE, closely followed by the Support Vector Regression. To provide a more comprehensive evaluation of each model's performance, we also consider the MAE. Unlike the RMSE, which amplifies larger errors due to its squaring effect, the MAE treats all errors equally. Therefore, the RMSE is more sensitive to outliers and better assesses the model's ability to handle such cases. Consistent with the RMSE, the Random Forest model exhibits the lowest MAE compared

Table 11: Model performance comparison.

Metric	OLS regression with FE	LASSO regression with FE	Random Forest	Support Vector Regression
RMSE	1.03	0.49	0.37	0.41
MAE	0.83	0.39	0.29	0.32
R-squared	0.63	0.63	0.79	0.74

to the other models. Again, it is closely followed by the SVR.

Additionally, we examine the R-squared of our models. This measure indicates the proportion of variance explained by the model, with a higher value indicating a better fit. Again, the Random Forest model outperforms the other models. Moreover, we find that the R-squared of the OLS regression and the LASSO regression are very similar. In contrast, the error metrics of the OLS regression are much higher than the error metrics of the LASSO regression. This suggests that the OLS regression model may be overfitting the data, resulting in larger prediction errors. By shrinking the coefficients of less important variables towards zero, LASSO seems to be successful at preventing overfitting the data. This observation also highlights the importance of considering multiple evaluation metrics to assess different models' performance and predictive accuracy.

In conclusion, the Random Forest model is the best-performing model. The metrics demonstrate superior predictive accuracy and its ability to explain a significant amount of variance in the data. As a result, we will focus our analysis on the Random Forest model to investigate the impact of different variables on the success of a restaurant. By leveraging the insights provided by this model, we gain a deeper understanding of the factors that contribute to the success of the restaurants in our data set.

6.3.2 Key Drivers Behind the Success of a Restaurant

In this subsection, we will analyze the outcomes of our model, aiming to uncover the key factors driving a restaurant's success. Our analysis first examines the results derived from running the Random Forest model on the complete restaurant data set, including a binary variable indicating whether a restaurant is vegetarian-friendly or not. In addition, we extend our Random Forest analysis by exploring whether these significant drivers vary across different restaurant segments. To accomplish this, we will also analyze two separate Random Forest models: one on the subset of vegetarian-friendly restaurants and one on the subset of non-vegetarian-friendly restaurants. We employ partial dependence plots to interpret our initial observations concerning variations across distinct restaurant segments.

Subsequently, we expand our analysis to explore whether the most significant drivers exhibit variation across different restaurant segments and their impact on a restaurant's success. To this end, we employ a final regression model with time and restaurant-fixed effects. This regression includes the variables identified as highly predictive by the Random Forest model, interacted with a binary variable distinguishing vegetarian-friendly restaurants from others. This allows us to discern whether the prominence of specific topics within restaurant reviews during a particular

month and within a specific restaurant leads to more reviews. Ultimately, this approach will move us closer to achieving 'causal' identification, providing deeper insights into restaurant success determinants.

Figure 14 presents the variable importance plot derived from the Random Forest model applied to the entire set of restaurants. The plot reveals that several topics emerge as significant drivers of restaurant success, which include: 'Prices and Positive Experience', 'Positive (Vegan) Dining Experience', 'Ordering and Waiting', 'Number of Friends', 'Great Food and Service for Dinner', 'Cozy Coffee and Tea Place', 'Good Breakfast and Brunch' and 'Good Pizza and Vegan Options'.

The binary variable indicating whether a restaurant is vegetarian-friendly appears to be the least important in our model. This finding suggests that a vegetarian label does not significantly impact a restaurant's success. Furthermore, we note that the environmental variables incorporated in our model (the Google Trends search terms and ratios) exhibit low importance. This implies that these variables have limited explanatory power in predicting restaurant success, compared to other factors identified in the analysis.

Additional insights are gained from Figure 16, which presents the partial dependence plots revealing the relationships between the dependent variable and specific independent variables. Firstly, for the topic 'Prices and Positive Experience', we observe an initial steep increase relative to the logarithm of the number of reviews, followed by a subsequent negative relationship that stabilizes. This suggests that when restaurant reviews are more likely to focus on this topic, it will influence the restaurant's success negatively until it reaches a point of stability. Secondly, the relationship between the topics 'Positive (Vegan) Dining Experience' and 'Ordering and Waiting' and the logarithm of the number of reviews show similar patterns. They exhibit an initial steep increase, followed by a negative relationship. This implies that when reviews are more likely to be regarding these topics, this negatively influences restaurant success. Thirdly, the 'Number of Friends' variable demonstrates an initial sharp increase, followed by a slight decline and eventually stabilization. This suggests that zero to a few friends may negatively influence the logarithm of the number of reviews, but the influence remains constant beyond a certain threshold. Fourthly, the topic 'Great Food and Service for Dinner' displays an initial strong positive relationship, followed by a brief decline and a gradual increase as the topic's theta value increases. This indicates that when reviews are more likely to highlight this topic, this positively impacts restaurant success. Fifthly, the topic 'Cozy Coffee and Tea Place' exhibits an initial steep positive relationship, which then transitions to a negative relationship before experiencing a slight increase beyond a certain threshold. Finally, the relationship between the polarity score and the logarithm of the number of reviews follows an inverted U-shaped relationship. The shape starts at the lowest level of the logarithm of the number of reviews, peaks, and then decreases until reaching a stable medium level.

In addition, Figure 32(a) presents the variable importance plot of the Random Forest model applied to the vegetarian-friendly restaurants subset. The graph reveals that the topics 'Prices and Positive Experience', 'Positive (Vegan) Dining Experience', 'Ordering and Waiting', 'Cozy Coffee and Tea Place', 'Number of Friends' and 'Great Food and Service for Dinner' are significant drivers of success for vegetarian-friendly restaurants.

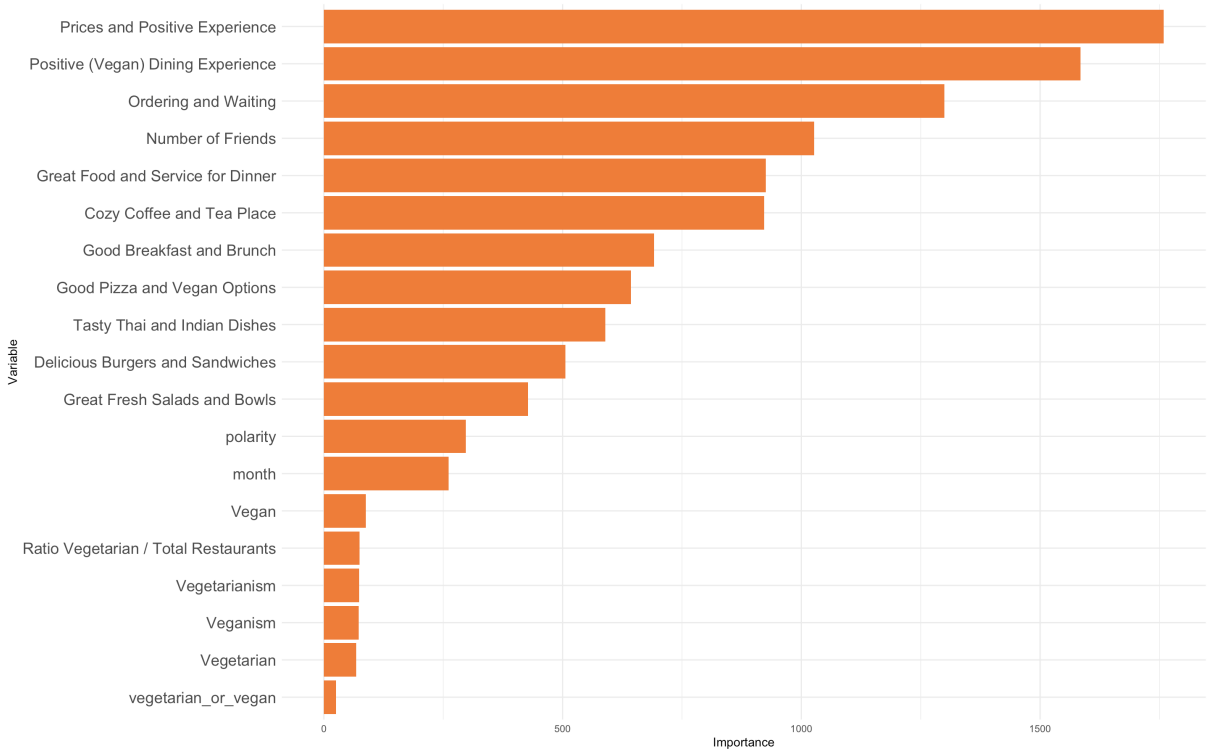
In contrast, Figure 32(b) displays the variable importance plot of the Random Forest model using the non-vegetarian friendly restaurants subset. This graph shows that the topics 'Prices and Positive Experience', 'Ordering and Waiting', 'Positive (Vegan) Dining Experience', 'Number of Friends', 'Great Food and Service for Dinner' and 'Cozy Coffee and Tea Place' are the most significant contributors to non-vegetarian restaurants' success.

Based on these variable important plots, we can argue that the significant drivers differ slightly between the two restaurant segments. For non-vegetarian friendly restaurants, the topics 'Ordering and Waiting', 'Great Food and Service for Dinner', 'Good Breakfast and Brunch', 'Great Fresh Salads and Bowls', and the number of friends appear to be more important. On the other hand, the topics 'Positive (Vegan) Dining Experience', 'Cozy Coffee and Tea Place', and 'Delicious Burgers and Sandwiches' appear to be more important for vegetarian-friendly restaurants.

These insights suggest that while there are some shared important factors for both restaurant segments, certain variables hold different degrees of importance depending on whether the restaurant is vegetarian-friendly or not. We present the partial dependence plots revealing the relationships between the dependent and independent variables to gain additional insights. These plots allow us to identify potential differences in the relationships between the two restaurant segments. We observe that most relationships are very similar for both restaurant segments (See Appendix in Appendix A). However, we argue that there appear to be differences in the relationships of diverse independent variables with the dependent variable between the two restaurant segments.

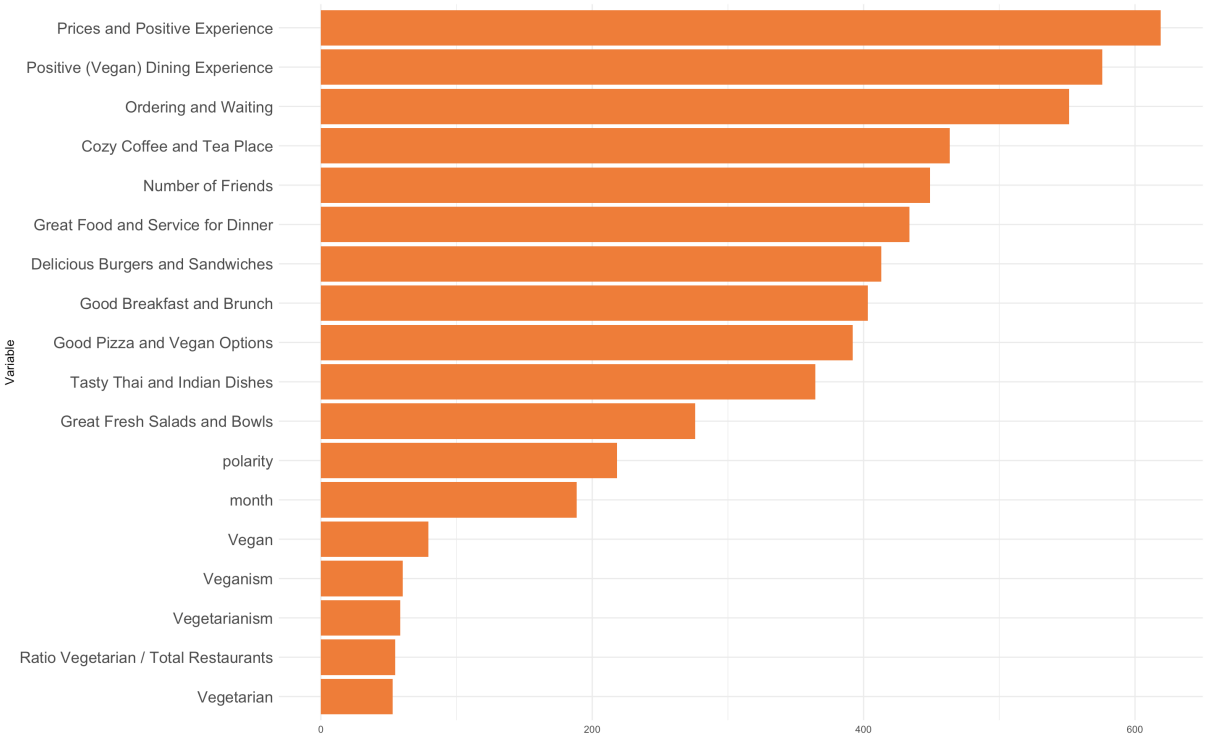
Firstly, for the topic 'Positive (Vegan) Dining Experience', we observe that the relationship between the topic probability and the logarithm of the number of reviews declines slower for vegetarian-friendly restaurants compared to non-vegetarian restaurants, which shows a steeper decline (See Figure 17). Secondly, for the topic 'Great Food and Service for Dinner', we observe that after the initial steep increase, the relationship for vegetarian-friendly restaurants is slightly negative. In contrast, for non-vegetarian restaurants, it first declines and then becomes positive (See Figure 18). Thirdly, for the topic 'Great Fresh Salads and Bowls', the relationship is initially very positive for both restaurant segments. After that, the relationship for vegetarian-friendly restaurants becomes slightly negative and remains constant. At the same time, for non-vegetarian friendly restaurants, it becomes more negative but also remains constant after that (See Figure 19). Fourthly, for the topic 'Delicious Burgers and Sandwiches', vegetarian-friendly restaurants show a small negative relationship, while non-vegetarian restaurants show a negative relationship, followed by a positive relationship, and then another negative relationship (See Figure 20). Fifthly, for the search term 'Vegetarianism', the relationship declines slower for vegetarian-friendly restaurants than for non-vegetarian-friendly restaurants (See Figure 21). Sixthly, for the search term 'Vegetarian', we observe that the relationship is initially stable and then takes on a U-shaped pattern for vegetarian-friendly restaurants. In contrast, for non-vegetarian restaurants, the relationship immediately starts to decline and only slightly increases towards the end (See Figure 22). Finally, for the search term 'Vegan', we find that the relationship of non-vegetarian-friendly restaurants initially declines but then starts to increase. In contrast, vegetarian-friendly restaurants immediately show a positive relationship (See Figure

Figure 14: Variable Importance Plot (All restaurants)

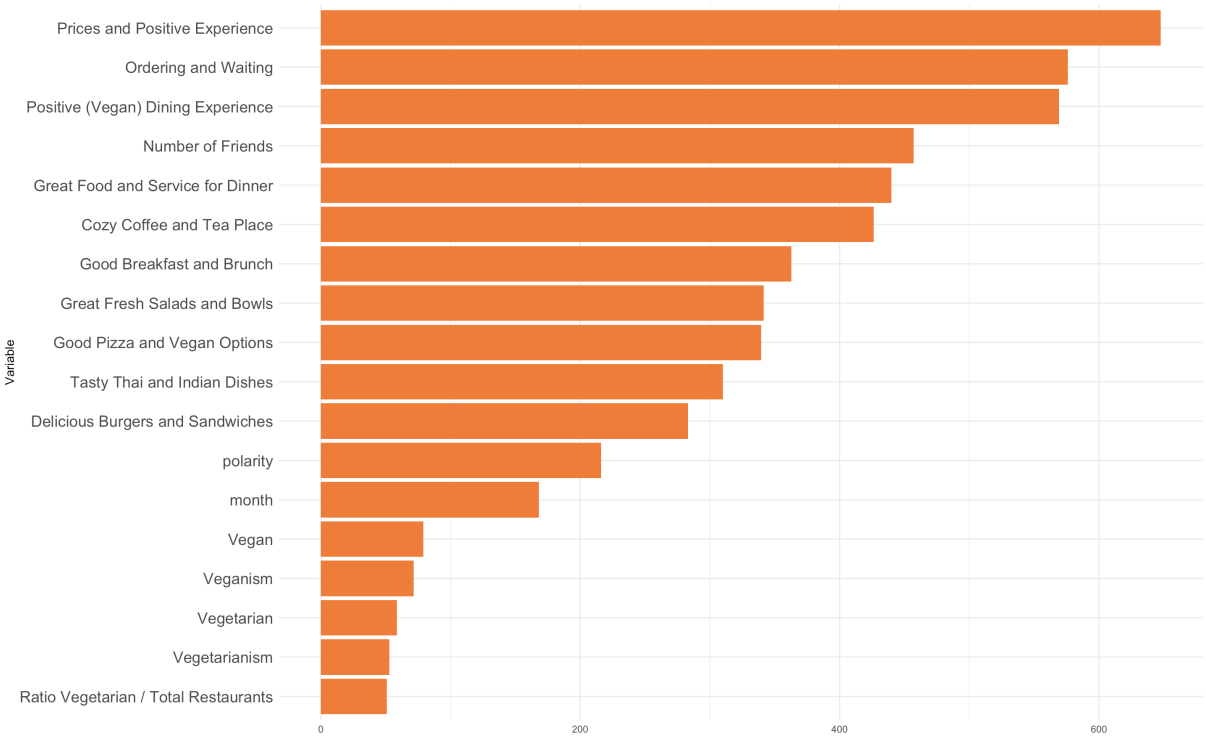


23).

Figure 15: Variable Importance Plot

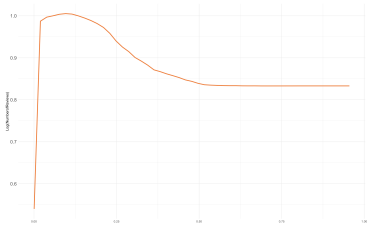


(a) Vegetarian friendly restaurants

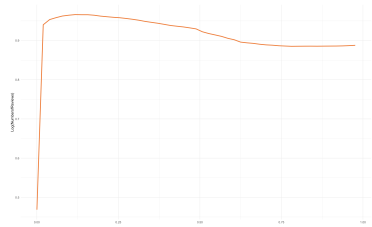


(b) Non-vegetarian friendly restaurants

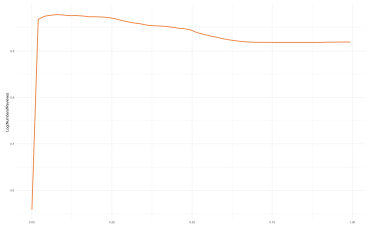
Figure 16: Partial dependence plots (all restaurants).



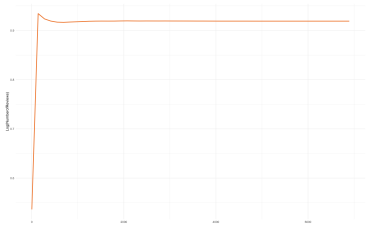
(a) Prices and Positive Experience



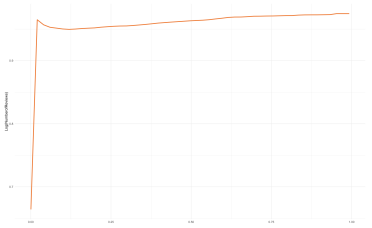
(b) Positive (Vegan) Dining Experience



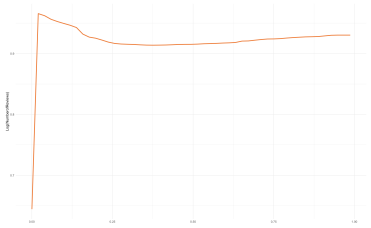
(c) Ordering and Waiting



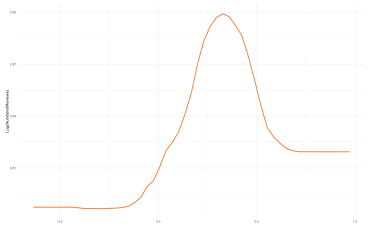
(d) Number of Friends



(e) Great Food and Service for Dinner

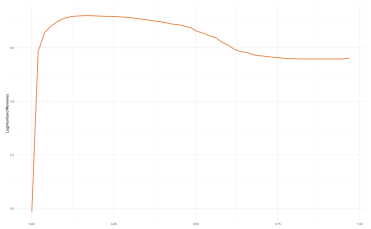


(f) Cozy Coffee and Tea Place

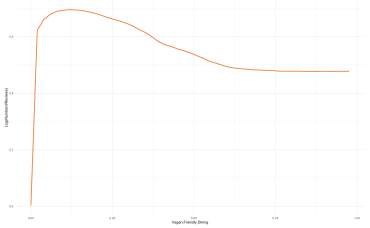


(g) Polarity score

Figure 17: Partial Dependence Plot: Positive (Vegan) Dining Experience



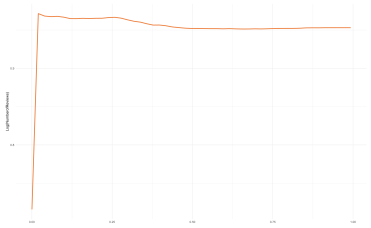
(a) Vegetarian friendly restaurants



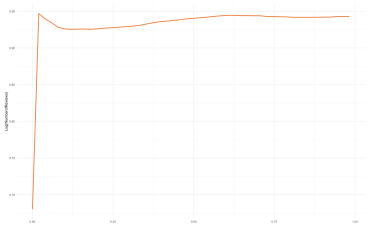
(b) Non-vegetarian friendly restaurants

To gain a deeper understanding of the key variables influencing restaurant success, we conducted a final regression analysis with time and restaurant fixed effects. This analysis also includes interaction terms between the variables identified as most important in our previous Random Forest model (See Figure 14) and a binary variable indicating whether a restaurant is vegetarian-friendly. The results of this analysis are presented in Table 12. From these results, several insights emerge. We identify several variables that significantly and positively affect the number of restaurant reviews, all at a significance level of $\alpha = 0.001$. These variables include 'Prices and Positive Experience', 'Positive (Vegan) Dining Experience', 'Ordering and Waiting',

Figure 18: Partial Dependence Plot: Great Food and Service for Dinner

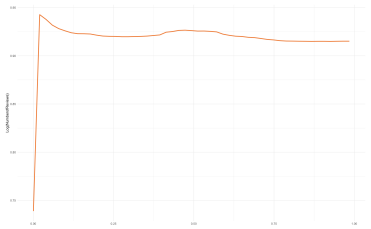


(a) Vegetarian friendly restaurants

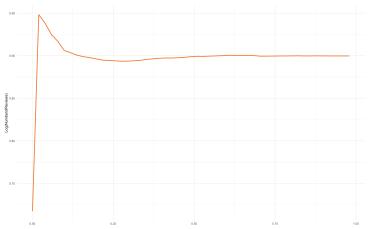


(b) Non-vegetarian friendly restaurants

Figure 19: Partial Dependence Plot: Great Fresh Salads and Bowls

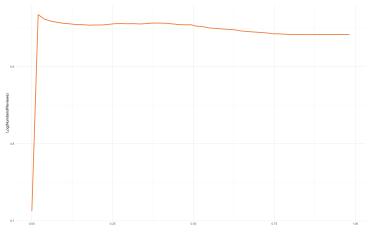


(a) Vegetarian friendly restaurants

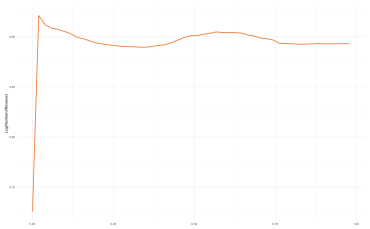


(b) Non-vegetarian friendly restaurants

Figure 20: Partial Dependence Plot: Delicious Burgers and Sandwiches

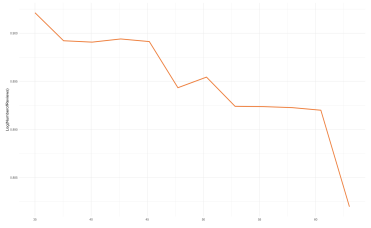


(a) Vegetarian friendly restaurants

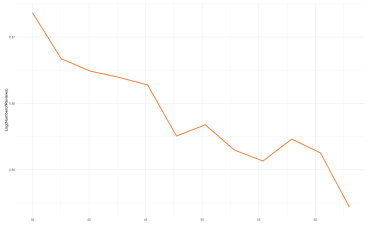


(b) Non-vegetarian friendly restaurants

Figure 21: Partial Dependence Plot: Search Term (Vegetarianism)



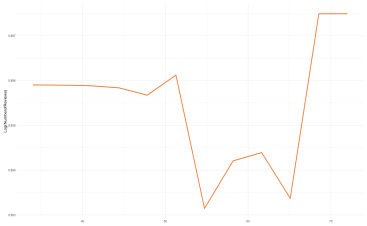
(a) Vegetarian friendly restaurants



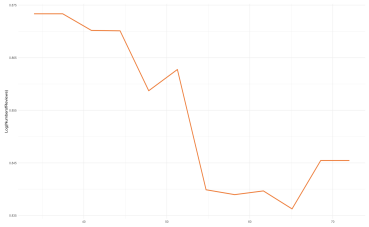
(b) Non-vegetarian friendly restaurants

'Number of Friends', 'Great Food and Service for Dinner', 'Cozy Coffee and Tea Place', 'Good Breakfast and Brunch' and 'Good Pizza and Vegan Options'. For example, an increase of 1% in the number of friends is associated with a 0.067% increase in the number of reviews, with all

Figure 22: Partial Dependence Plot: Search Term (Vegetarian)

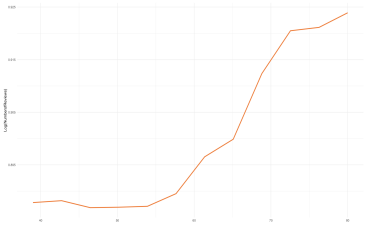


(a) Vegetarian friendly restaurants

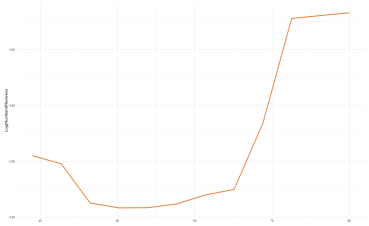


(b) Non-vegetarian friendly restaurants

Figure 23: Partial Dependence Plot: Search Term (Vegan)



(a) Vegetarian friendly restaurants



(b) Non-vegetarian friendly restaurants

other variables in the model held constant.

Moreover, the interaction terms between particular variables and the binary variable indicating whether a restaurant is vegetarian-friendly also yielded notable findings. The interaction term between the topic 'Prices and Tasty Options' and vegetarian-friendliness is positively significant at $\alpha = 0.05$. Therefore, this variable significantly, positively influences the number of reviews of a vegetarian-friendly restaurant. Besides that, we find that the interaction term with the variables 'Positive (Vegan) Dining Experience', 'Number of Friends', 'Cozy Coffee and Tea Place', 'Good Breakfast and Brunch' and 'Good Pizza and Vegan Options' demonstrate a positive relationship with the number of reviews, although the significance level $\alpha = 0.05$ was not met. This suggests a potential positive influence on the success of vegetarian-friendly restaurants compared to non-vegetarian-friendly restaurants. Conversely, we find that the interaction terms with the variables 'Ordering and Waiting' and 'Great Food and Service for Dinner' negatively affect the number of reviews, although significance was not achieved. This suggests a potential negative influence on the success of vegetarian-friendly restaurants compared to non-vegetarian-friendly restaurants. However, these interpretations should be considered unreliable due to their lack of statistical significance, and no conclusive evidence of their effect can be established. These findings provide valuable insights into the factors driving the success of vegetarian-friendly and non-vegetarian-friendly restaurants. Further conclusions and implications based on these results are presented in Section 7.

Table 12: Results from OLS model with Time and Restaurant Fixed Effects including the most important variables from RF

Variables	**Coefficient**	**Std. Error**
log(Prices.and.Tasty.Options)	0.096***	0.003
log(Vegan.Friendly.Dining)	0.081***	0.003
log(Ordering.and.Waiting)	0.096***	0.003
log(friends_count)	0.067***	0.003
log(Great.Food.and.Service.for.Dinner)	0.087***	0.004
log(Cozy.Coffee.and.Tea.Place)	0.087***	0.003
log(Good.Breakfast.and.Brunch)	0.074***	0.004
log(Good.Pizza.and.Vegan.Options)	0.070***	0.004
vegetarian_or_vegan:log(Prices.and.Tasty.Options)	0.009*	0.004
vegetarian_or_vegan:log(Vegan.Friendly.Dining)	0.003	0.004
vegetarian_or_vegan:log(Ordering.and.Waiting)	-0.003	0.004
vegetarian_or_vegan:log(friends_count)	0.002	0.004
vegetarian_or_vegan:log(Great.Food.and.Service.for.Dinner)	-0.002	0.005
vegetarian_or_vegan:log(Cozy.Coffee.and.Tea.Place)	0.003	0.005
vegetarian_or_vegan:log(Good.Breakfast.and.Brunch)	0.003	0.005
vegetarian_or_vegan:log(Good.Pizza.and.Vegan.Options)	0.006	0.006
Restaurant Fixed Effect	Yes	
Time Fixed Effect	Yes	

Significance levels:. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

7 Conclusion and Implications

Based on the hypotheses established in Section 3, this section provides an overview of the conclusions and implications of this study.

7.1 Topics Discussed in Reviews

The results from our analysis suggest that the most important topics derived from our Random Forest model are related to price, food quality, service quality, and ambience. These variables include 'Prices and Positive Experience', 'Positive (Vegan) Dining Experience', 'Ordering and Waiting', 'Great Food and Service for Dinner', 'Cozy Coffee and Tea Place', 'Good Breakfast and Brunch' and 'Good Pizza and Vegan Options' and they are found to have a significantly, positive effect on a restaurant's success.

To investigate if these topics vary between vegetarian and non-vegetarian friendly restaurant segments (H_1), our final regression with time and restaurant fixed effects reveals that only the topic 'Prices and Positive Experiences' significantly and positively influences the success of vegetarian-friendly restaurants compared to their non-vegetarian counterparts. In contrast, the other crucial features exhibit no significant interaction with the vegetarian-friendly binary variable when utilizing restaurant and time-fixed effects, consistent with the Random Forest model on the complete restaurant data. This Variable Importance Plot unveils the limited importance of the vegetarian-friendly binary variable, implying that the presence of this label has a minor impact on restaurant success.

Further insights from the Random Forest models per restaurant segment reveal differences

in the importance rankings of specific topics. In non-vegetarian restaurants, 'Great Fresh Salads and Bowls' take precedence, emphasizing healthfulness, while in vegetarian-friendly establishments, 'Delicious Burgers and Sandwiches' gain prominence, marking the relevance of vegetarian menu options for such items. In addition, analyzing the relationships between these topics and the number of reviews exposes distinct dynamics between restaurant segments.

Firstly, the topic 'Positive (Vegan) Dining Experience' has a lower negative effect on vegetarian-friendly restaurants than non-vegetarian-friendly ones. This suggests that an increased probability of reviews discussing vegan dining experiences has less impact on vegetarian-friendly restaurants, which is supported by the results from the OLS and LASSO regression (See Table 13 and 14 in Appendix B).

Secondly, for the topic 'Great Food and Service for Dinner', an increased probability of reviews discussing this topic has a positive effect on non-vegetarian restaurants but a negative effect on vegetarian-friendly restaurants. This implies that reviews emphasizing great food and service during dinner are beneficial for non-vegetarian restaurants but may not have the same positive impact on vegetarian-friendly establishments. This is also supported by the results from the OLS and LASSO regression (See Table 13 and 14 in Appendix B).

Lastly, regarding the topic 'Great Fresh Salads and Bowls', previously identified as more important for non-vegetarian-friendly restaurants, we observe a smaller decrease in its influence on vegetarian-friendly restaurants. This suggests that an increased probability of reviews discussing great fresh salads and bowls has a more negative effect on non-vegetarian restaurants compared to vegetarian-friendly restaurants, which is in line with the results from the OLS and LASSO regression (See Table 13 and 14 in Appendix B).

In conclusion, differences in topic importance are evident among restaurant segments, with 'Price and Positive Experiences' standing out as a topic that significantly and positively affects vegetarian-friendly restaurants compared to non-vegetarian-friendly restaurants. Therefore, we reject the null hypothesis that the topics influencing the number of reviews received by restaurants do not differ between the two segments.

7.2 Sentiment Captured in Reviews

The relationship between sentiment and the number of reviews is similar for both segments. As the polarity score increases, the number of reviews initially also increases. However, after reaching a certain threshold, the number of reviews declines until it reaches another threshold level. This inverted U-shaped relationship suggests that extremely positive or extremely negative sentiment may not necessarily lead to a higher number of reviews. Instead, there is an optimal range of sentiment that drives more reviews. This conclusion is in line with the results from the OLS regression (See Table 13 in Appendix B), which indicates that the interaction term between vegetarian or not and the polarity score is not significant. However, the LASSO regression model (See Table 14 in Appendix B) does find a significant, negative coefficient for this interaction term. This would suggest a more negative sentiment effect for vegetarian-friendly restaurants compared to non-vegetarian-friendly restaurants.

Based on the analysis of sentiment captured in the restaurant reviews, it is concluded that the effect of sentiment on restaurant success does not significantly differ between the two segments

(H_2). Therefore, we fail to reject the null hypothesis that the effect of sentiment on the number of reviews received by restaurants does not differ between the two segments.

7.3 Social Networks of Reviewers

Several findings are observed based on the analysis of social networks and their impact on the number of reviews. The third hypothesis (H_3) suggests a positive relationship between the number of friends a reviewer has on Yelp and the number of reviews the restaurant receives. The partial dependence plot derived from the Random Forest model partly supports this hypothesis (See Figure 16). Initially, the number of reviews sharply increases when a reviewer has a few friends on Yelp, indicating that having reviewers with some friends review the restaurant is beneficial for restaurant success. However, after that, the relationship is negative shortly and then stabilizes.

Moreover, our final regression analysis supports a significant, positive relationship between the number of friends and the number of restaurant reviews. This is supported by the results from the OLS and the LASSO model, which also indicate a positive relationship between the number of friends and the number of restaurant reviews (See Table 13 and 14 in Appendix B). Therefore, we reject the null hypothesis that the number of friends negatively or not influences the number of restaurant reviews.

The fourth hypothesis (H_4) proposes that the effect of the number of friends on the number of reviews differs per restaurant segment. The variable importance plot shows that the number of friends is slightly more important for non-vegetarian-friendly restaurants. However, this difference is minimal. From the results of our final regression with time and restaurant fixed effects, we conclude that there is no significant effect between the interaction term and the number of restaurant reviews (See Table 12). Therefore, we fail to reject the null hypothesis that the effect of the number of friends on the number of reviews does not differ per restaurant segment.

In summary, we reject the null hypothesis that the number of friends negatively or not influences the number of restaurant reviews. We fail to reject the null hypothesis that the effect of the number of friends on the number of reviews does not differ per restaurant segment.

7.4 Social Environment

To investigate the influence of the social environment on the number of reviews for vegetarian-friendly restaurants, we included five variables in our analysis. These variables aimed to address our fifth hypothesis (H_5), which suggests that a stronger focus on the vegetarian and vegan diet within the social environment would increase the number of reviews for vegetarian-friendly restaurants. The variables included were four Google Trends search terms and the ratio of vegetarian-friendly restaurants to total restaurants in a state.

Based on the Partial Dependence Plots derived from the Random Forest model on the data set including vegetarian-friendly restaurants (See Appendix A), we have drawn several conclusions. Firstly ($H_{5.1}$), the relationship between the ratio of vegetarian-friendly restaurants to total restaurants in a state and the number of reviews was turbulent, but eventually turned out to be positive, partially supporting our initial hypothesis for this variable. Secondly ($H_{5.2}$),

for the search term 'Vegetarian', we found that the relationship was initially stable but started to decline after reaching a certain threshold, followed by an increase. This finding contradicts our initial hypothesis regarding this search term. Thirdly ($H_{5.3}$), the relationship between the search term 'Vegetarianism' and the number of reviews was negative, thereby rejecting our initial hypothesis for this search term. Fourthly ($H_{5.4}$), for the search term 'Vegan', a positive relationship between the number of searches for 'Vegan' and the number of reviews was identified, which results in accepting our initial hypothesis for this search term. Finally ($H_{5.5}$), the relationship between the search term 'Veganism' and the number of reviews was positive, supporting our initial hypothesis for this search term.

These results suggest that the social environment may influence the success of a restaurant. However, as these variables do not exhibit significant importance in the Variable Importance Plots, we fail to reject our fifth null hypothesis that when the social environment is more focused on the vegetarian and vegan diet, the number of reviews of vegetarian-friendly restaurants decreases or remains the same.

7.5 Implications

7.5.1 Academic Implications

The academic implications of this study are significant and contribute to the existing literature on electronic Word-Of-Mouth (eWOM). Specifically, the study expands knowledge in four different areas.

Firstly, by applying a Latent Dirichlet Allocation topic model, this study uncovers hidden patterns and insights related to the topics discussed in eWOM. This contributes to the understanding of the key factors that influence restaurant success in both the vegetarian-friendly and non-vegetarian-friendly segments. The methodology employed in this study can serve as a framework for future research exploring the topics discussed in eWOM across various industries.

Secondly, the study provides insights into the relationship between sentiment and eWOM volume in the context of vegetarian and non-vegetarian-friendly restaurants. By examining the polarity score of reviews and its impact on the number of reviews, this research contributes to the understanding of how sentiment influences the online behavior of customers in terms of reviewing diverse restaurant segments.

Thirdly, the study investigates the impact of reviewers' online social networks on eWOM volume. By analyzing the number of friends a reviewer has on Yelp and its influence on the number of reviews, the research provides insights into the social dynamics of eWOM. This contributes to understanding social networks' role in shaping restaurant success.

Lastly, the study explores the influence of the social environment on eWOM volume by considering Google Trends search terms and the ratio of vegetarian-friendly restaurants to total restaurants in a state. This expands the understanding of how external factors, such as online search trends and the local restaurant landscape, impact eWOM volume. The research provides context-specific insights that contribute to existing literature by focusing on specific restaurant segments.

To summarize, the academic implications of this study lie in expanding the knowledge base on eWOM by exploring topics, sentiment, social networks, and the social environment in the

context of vegetarian and non-vegetarian-friendly restaurants. The findings contribute to a deeper understanding of customer behavior and provide a foundation for future research in eWOM.

7.5.2 Managerial Implications

The conclusions of this study also have several managerial implications and provide valuable recommendations for restaurant managers who rely on eWOM to promote their restaurants. By understanding the factors that influence eWOM volume and the nuances within different restaurant segments, managers can make informed decisions to enhance customer experience, improve their offerings, and ultimately drive the success of their restaurants.

One of the critical managerial implications is that this study uncovers hidden topics that influence eWOM volume, providing restaurant managers with valuable insights into their restaurant's quality and performance. By analyzing the topics discussed in reviews, managers can understand which aspects of their restaurant are driving more reviews. This knowledge helps them identify areas of strength and areas that may require improvement, allowing for more targeted efforts to enhance the customer experience. To leverage this implication, restaurant managers should regularly analyze and monitor the topics expressed in reviews to gain insights into what aspects of the restaurant drive reviews. Addressing negative sentiments and customer concerns can improve customer satisfaction and a higher eWOM volume.

Another important implication is the potential variation in the influence across different restaurant segments. Managers can leverage these findings by developing tailored strategies that align with the preferences and expectations within their specific segment. The study highlights that certain topics hold greater importance for the vegetarian-friendly restaurant segment, such as the topics regarding price and sandwiches and burgers. An implication for restaurant managers of vegetarian-friendly restaurants is to ensure that their vegetarian alternatives are price competitive and that they serve good vegetarian alternatives. By offering delicious burgers and sandwiches tailored to vegetarian preferences, managers can attract a broad customer base and increase their chances of receiving positive reviews.

Moreover, the study highlights that the topic of fresh salads and bowls holds greater importance for non-vegetarian-friendly restaurants. This implies that managers of non-vegetarian-friendly restaurants ensure having enough healthy alternatives on the menu. Managers can emphasize these healthy options in their menus, marketing materials, and customer interactions to attract more (positive) reviews and enhance their restaurant's reputation.

Besides the implications for managers, platforms can also benefit from the results of this paper. By recognizing the importance of the number of friends of a reviewer on eWOM volume, the platform can encourage reviewers to connect on the platform. They could do this by increasing promotional efforts to facilitate reviewer connections. This can help increase eWOM volume and foster a sense of community among reviewers.

To conclude, the findings of this study emphasize the importance of improving a restaurant's reputation and customer satisfaction. By addressing the factors that significantly affect eWOM volume, restaurant managers can actively work towards enhancing their restaurant's reputation and increasing customer satisfaction. Adaption and responsiveness to customer reviews will

contribute to building a positive online presence and drive success in the competitive restaurant industry. In addition, platforms can benefit from the results by increasing their promotional efforts to facilitate reviewer connections.

8 Discussion

This section will summarise the limitations of this research and provide recommendations for future research.

8.1 Limitations

The first limitation of this research is that this study utilized a 1:1 propensity score matching approach to match restaurants from two different segments. However, it is essential to recognize that this matching approach does not guarantee a perfect comparison between the two segments. There may still exist differences that could impact the results and conclusions drawn from the analysis.

The second limitation of this research is that findings are based on a specific data set and may not be fully generalizable to all restaurants or different geographic regions. The analysis focused on a particular set of restaurants and may not capture the full diversity of the restaurant industry.

The third limitation of this research is that the analysis of social networks relied on the number of friends a reviewer has on Yelp. However, this measure may not capture the true extent and influence of a reviewer's social network.

The fourth limitation of this research is that the study primarily focused on identifying associations and correlations between variables rather than establishing causal relationships. While the analysis provides valuable insights into the factors influencing eWOM volume, further research using experimental or longitudinal designs is needed to establish causality and understand the direction of these relationships.

The last limitation of this research is the limited explanatory power of the environmental variables. Despite including Google Trends search terms and ratios as environmental variables in the model, their relatively low importance suggests that they may have limited influence on predicting restaurant success compared to other factors examined in the analysis. This indicates that there may be additional environmental factors beyond those considered in this study that have a more substantial impact on eWOM volume.

8.2 Future Research

Based on the limitations that were identified, we propose four recommendations for future research that address these limitations to further advance our understanding of eWOM in the context of the restaurant industry.

First, future research should employ more rigorous methods to compare restaurant segments, such as quasi-experimental designs or natural experiments. This can help overcome the limitations of the propensity score matching approach and provide a more robust comparison between segments.

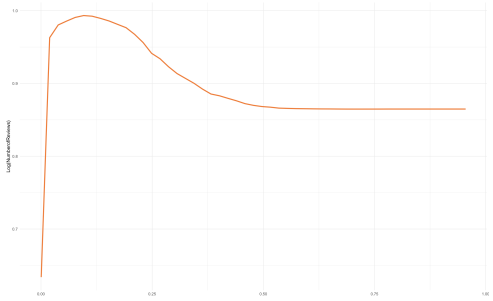
Second, future research should collect data from a broader range of restaurants and geographic locations, but also from multiple review platforms or social media channels. This will enhance the generalizability of the findings. Including restaurants from various cuisines, price ranges, geographical regions, and review platforms can provide a more comprehensive understanding of eWOM dynamics in the restaurant industry.

Third, future research could explore more comprehensive measures of reviewer networks. This could include considering the quality and strength of connections, influencers within the network, and the spread of eWOM beyond immediate social circles. This approach can help better capture the influence of social networks on eWOM volume.

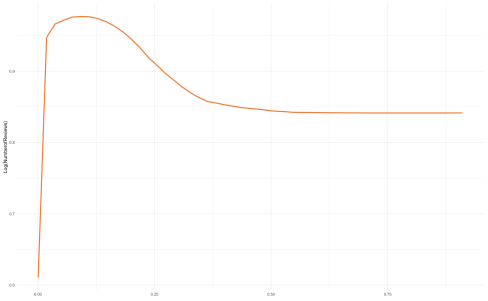
Last, future research could explore and identify relevant environmental variables that could provide a more comprehensive understanding of the contextual factors influencing eWOM dynamics in the (vegetarian) restaurant industry.

A Appendix

Figure 24: Partial Dependence Plot: Prices and Positive Experience

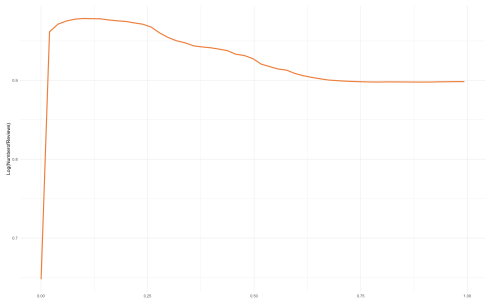


(a) Vegetarian friendly restaurants

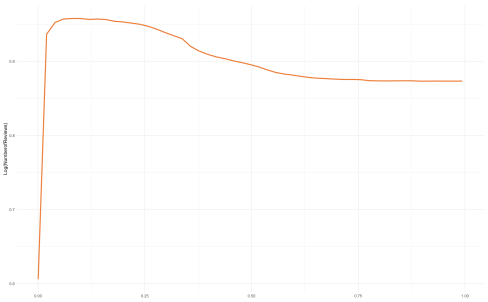


(b) Non-vegetarian friendly restaurants

Figure 25: Partial Dependence Plot: Ordering and Waiting

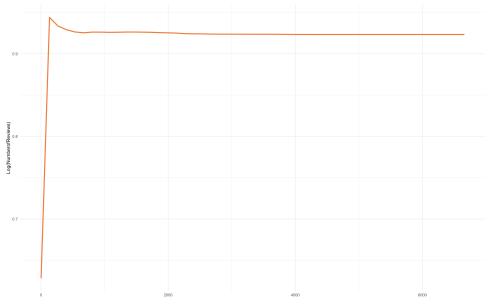


(a) Vegetarian friendly restaurants

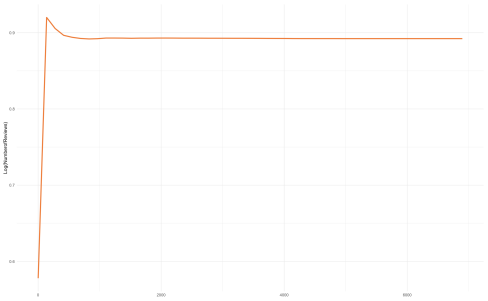


(b) Non-vegetarian friendly restaurants

Figure 26: Partial Dependence Plot: Number of Friends

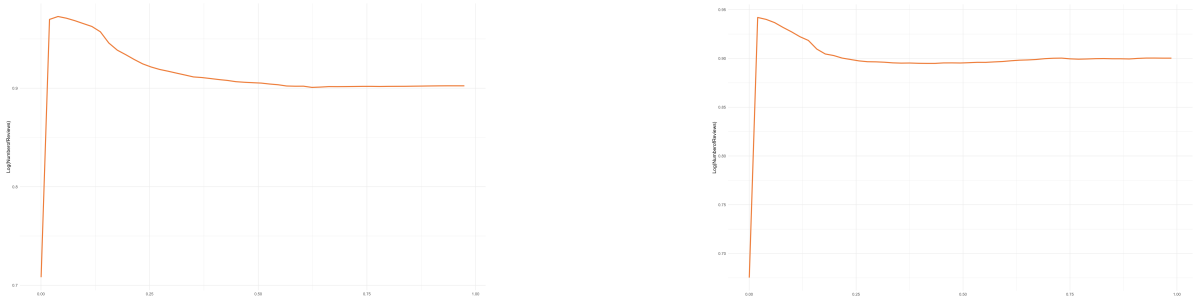


(a) Vegetarian friendly restaurants



(b) Non-vegetarian friendly restaurants

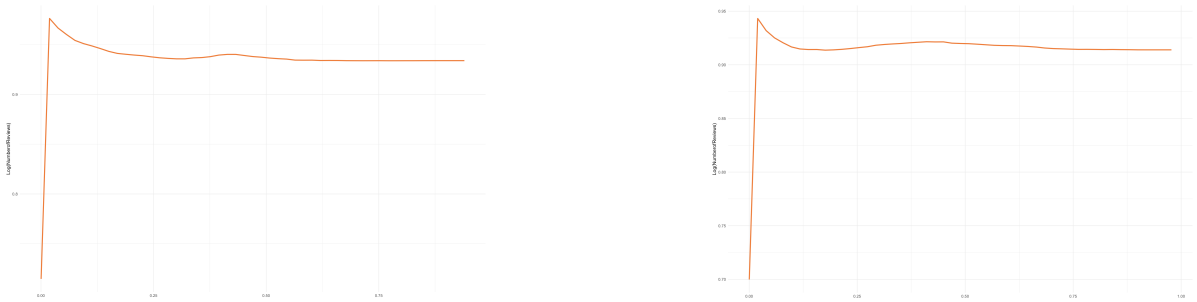
Figure 27: Partial Dependence Plot: Cozy Coffee and Tea Place



(a) Vegetarian friendly restaurants

(b) Non-vegetarian friendly restaurants

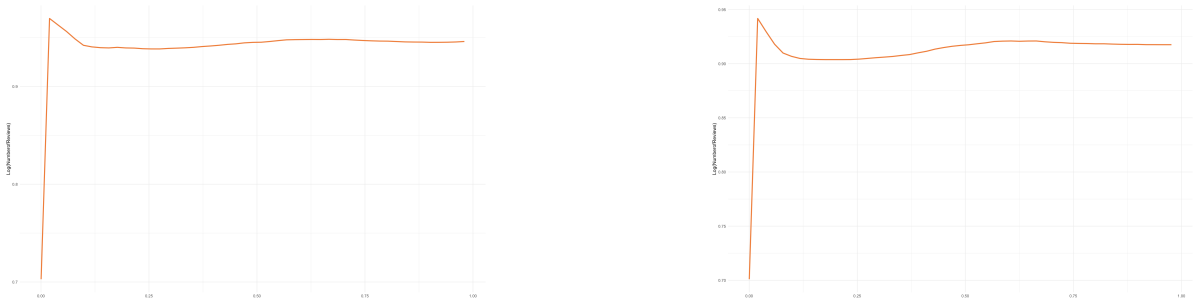
Figure 28: Partial Dependence Plot: Good Breakfast and Brunch



(a) Vegetarian friendly restaurants

(b) Non-vegetarian friendly restaurants

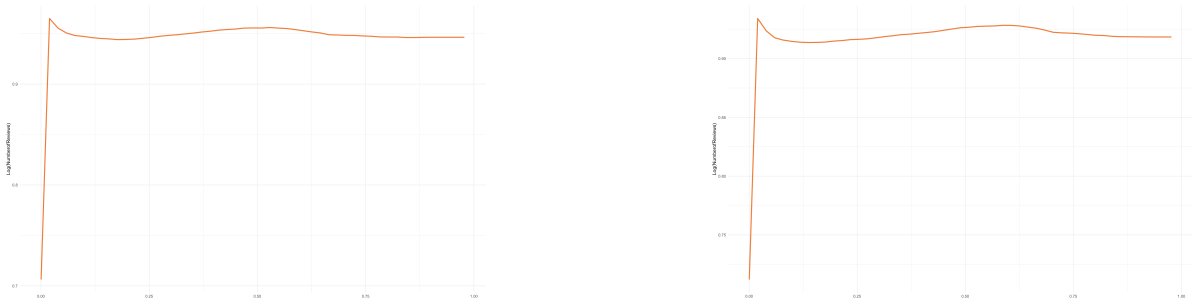
Figure 29: Partial Dependence Plot: Good Pizza and Vegan Options



(a) Vegetarian friendly restaurants

(b) Non-vegetarian friendly restaurants

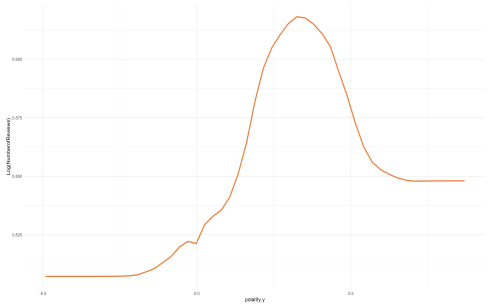
Figure 30: Partial Dependence Plot: Tasty Thai and Indian Dishes



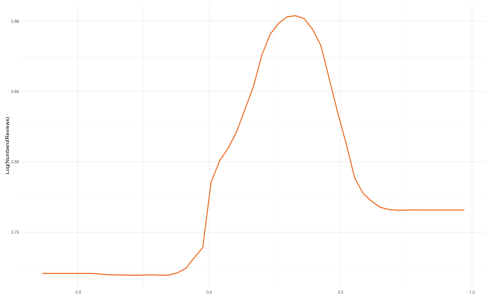
(a) Vegetarian friendly restaurants

(b) Non-vegetarian friendly restaurants

Figure 31: Partial Dependence Plot: Polarity Score

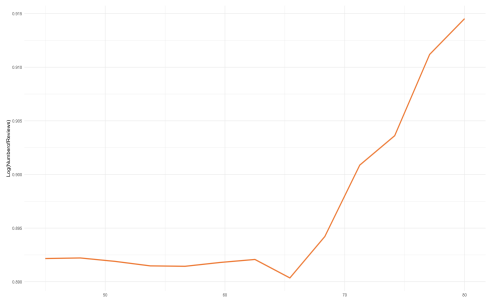


(a) Vegetarian friendly restaurants

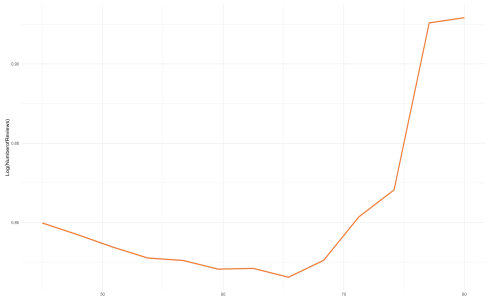


(b) Non-vegetarian friendly restaurants

Figure 32: Partial Dependence Plot: Search Term (Veganism)

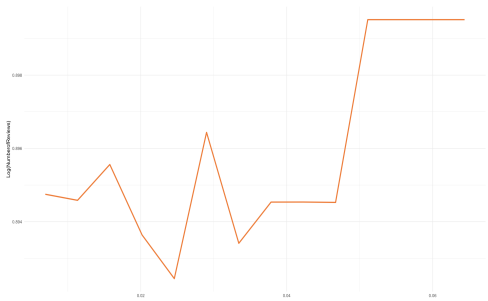


(a) Vegetarian friendly restaurants

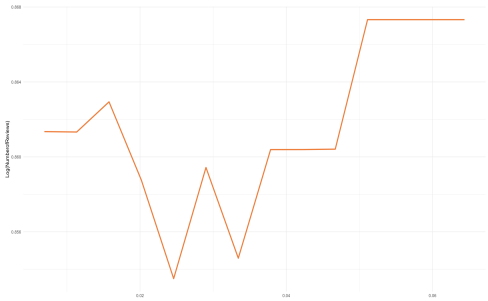


(b) Non-vegetarian friendly restaurants

Figure 33: Partial Dependence Plot: Ratio Vegetarian Restaurants to Non-vegetarian Restaurants



(a) Vegetarian friendly restaurants



(b) Non-vegetarian friendly restaurants

B Appendix

Table 13: Results from OLS regression with Fixed Effects

Variables	**Coefficient**	**Std. Error**
log(Prices.and.Tasty.Options)	0.092***	0.005
log(Tasty.Thai.and.Indian.Dishes)	0.071***	0.005
log(Good.Pizza.and.Vegan.Options)	0.069***	0.005
log(Great.Fresh.Salads.and.Bowls)	0.051***	0.005
log(Great.Food.and.Service.for.Dinner)	0.086***	0.005
log(Delicious.Burgers.and.Sandwiches)	0.060***	0.004
log(Ordering.and.Waiting)	0.127***	0.004
log(Good.Breakfast.and.Brunch)	0.062***	0.005
log(Cozy.Coffee.and.Tea.Place)	0.082***	0.005
log(Vegan.Friendly.Dining)	0.096***	0.004
log(ratio)	-0.081***	0.023
log(Vegetarianism)	-0.321***	0.049
log(Vegetarian)	-0.370***	0.055
log(Vegan)	0.251***	0.049
log(Veganism)	0.261**	0.080
log(friends_count)	0.080***	0.003
vegetarian_or_vegan	-1.147***	0.268
polarity	2.341***	0.168
log(Prices.and.Tasty.Options):vegetarian_or_vegan	-0.003	0.005
log(Tasty.Thai.and.Indian.Dishes):vegetarian_or_vegan	-0.001	0.004
log(Good.Pizza.and.Vegan.Options):vegetarian_or_vegan	0.003	0.004
log(Great.Fresh.Salads.and.Bowls):vegetarian_or_vegan	0.010*	0.004
log(Great.Food.and.Service.for.Dinner):vegetarian_or_vegan	-0.010*	0.004
log(Delicious.Burgers.and.Sandwiches):vegetarian_or_vegan	0.008*	0.004
log(Ordering.and.Waiting):vegetarian_or_vegan	-0.010*	0.004
log(Good.Breakfast.and.Brunch):vegetarian_or_vegan	-0.010*	0.004
log(Cozy.Coffee.and.Tea.Place):vegetarian_or_vegan	-0.013**	0.004
log(Vegan.Friendly.Dining):vegetarian_or_vegan	0.008 .	0.004
log(ratio):vegetarian_or_vegan	0.085**	0.032
log(Vegetarianism):vegetarian_or_vegan	0.102	0.067
log(Vegetarian):vegetarian_or_vegan	0.272***	0.076
log(Vegan):vegetarian_or_vegan	-0.093	0.068
log(Veganism):vegetarian_or_vegan	0.076	0.112
polarity:vegetarian_or_vegan	-0.011	0.052
log(friends_count):vegetarian_or_vegan	-0.010*	0.004
log(Prices.and.Tasty.Options):polarity	0.099***	0.012
log(Tasty.Thai.and.Indian.Dishes):polarity	0.077***	0.012
log(Good.Pizza.and.Vegan.Options):polarity	0.049***	0.013
log(Great.Fresh.Salads.and.Bowls):polarity	0.060***	0.012
log(Great.Food.and.Service.for.Dinner):polarity	0.046***	0.013
log(Delicious.Burgers.and.Sandwiches):polarity	0.056***	0.012
log(Ordering.and.Waiting):polarity	0.021 .	0.012
log(Good.Breakfast.and.Brunch):polarity	0.060***	0.013
log(Cozy.Coffee.and.Tea.Place):polarity	0.050***	0.012
log(Vegan.Friendly.Dining):polarity	-0.025*	0.011
Time Fixed Effect	Yes	

Significance levels:. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 14: Results from LASSO model with Fixed Effects

Variables	**Coefficient**
(Intercept)	3.449
log(Prices.and.Tasty.Options)	0.088
log(Tasty.Thai.and.Indian.Dishes)	0.072
log(Good.Pizza.and.Vegan.Options)	0.069
log(Great.Fresh.Salads.and.Bowls)	0.048
log(Great.Food.and.Service.for.Dinner)	0.085
log(Delicious.Burgers.and.Sandwiches)	0.061
log(Ordering.and.Waiting)	0.126
log(Good.Breakfast.and.Brunch)	0.061
log(Cozy.Coffee.and.Tea.Place)	0.079
log(Vegan.Friendly.Dining)	0.094
log(ratio)	-0.069
log(Vegetarianism)	-0.267
log(Vegetarian)	-0.321
log(Vegan)	0.24
log(Veganism)	0.265
log(friends_count)	0.079
vegetarian_or_vegan	-0.395
polarity	2.199
log(Prices.and.Tasty.Options):vegetarian_or_vegan	-0.004
log(Tasty.Thai.and.Indian.Dishes):vegetarian_or_vegan	-0.002
log(Good.Pizza.and.Vegan.Options):vegetarian_or_vegan	0.003
log(Great.Fresh.Salads.and.Bowls):vegetarian_or_vegan	0.009
log(Great.Food.and.Service.for.Dinner):vegetarian_or_vegan	-0.01
log(Delicious.Burgers.and.Sandwiches):vegetarian_or_vegan	0.009
log(Ordering.and.Waiting):vegetarian_or_vegan	-0.008
log(Good.Breakfast.and.Brunch):vegetarian_or_vegan	-0.012
log(Cozy.Coffee.and.Tea.Place):vegetarian_or_vegan	-0.013
log(Vegan.Friendly.Dining):vegetarian_or_vegan	0.008
log(ratio):vegetarian_or_vegan	0.081
log(Vegetarianism):vegetarian_or_vegan	0.006
log(Vegetarian):vegetarian_or_vegan	0.157
log(Vegan):vegetarian_or_vegan	.
log(Veganism):vegetarian_or_vegan	.
polarity:vegetarian_or_vegan	-0.025
log(friends_count):vegetarian_or_vegan	-0.013
log(Prices.and.Tasty.Options):polarity	0.093
log(Tasty.Thai.and.Indian.Dishes):polarity	0.069
log(Good.Pizza.and.Vegan.Options):polarity	0.045
log(Great.Fresh.Salads.and.Bowls):polarity	0.061
log(Great.Food.and.Service.for.Dinner):polarity	0.04
log(Delicious.Burgers.and.Sandwiches):polarity	0.051
log(Ordering.and.Waiting):polarity	0.018
log(Good.Breakfast.and.Brunch):polarity	0.058
log(Cozy.Coffee.and.Tea.Place):polarity	0.051
log(Vegan.Friendly.Dining):polarity	-0.024
Time Fixed Effect	Yes

References

Andrzejewski, D. & Zhu, X. (2009, 01). Latent dirichlet allocation with topic-in-set knowledge. *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. doi: 10.3115/1621829.1621835

Arun, R., Suresh, V., Madhavan, C. & Murty, M. (2010, 06). On finding the natural number of topics with latent dirichlet allocation: Some observations. In (p. 391-402). doi: 10.1007/978-3-642-13657-3$_4$3

Babić Rosario, A., Sotgiu, F., De Valck, K. & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, *53*(3), 297–318.

Bellman, R. & Kalaba, R. (1959). On adaptive control processes. *IRE Transactions on Automatic Control*, *4*(2), 1-9. doi: 10.1109/TAC.1959.1104847

Blei, D., Ng, A. & Jordan, M. (2001, 01). Latent dirichlet allocation. In (Vol. 3, p. 601-608).

Bufquin, D., DiPietro, R. & Partlow, C. (2017). The influence of the dinex service quality dimensions on casual-dining restaurant customers' satisfaction and behavioral intentions. *Journal of Foodservice Business Research*, *20*(5), 542–556. Retrieved from `https://doi.org/10.1080/15378020.2016.1222744` doi: 10.1080/15378020.2016.1222744

Buntine, W. & Jakulin, A. (2006, 01). Discrete component analysis. In (p. 1-33).

Buttle, F. A. (1998). Word of mouth: Understanding and managing referral marketing. *Journal of Strategic Marketing*, *6*(3), 241–254.

Cao, J., Xia, T., Li, J., Zhang, Y. & Tang, S. (2009, 03). A density-based method for adaptive lda model selection. *Neurocomputing*, *72*, 1775-1781. doi: 10.1016/j.neucom.2008.06.011

Chevalier, J. A. & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Choi, H., Joung, H.-W., Choi, E.-K. & Kim, H.-S. (2022). Understanding vegetarian customers: the effects of restaurant attributes on customer satisfaction and behavioral intentions. *Journal of Foodservice Business Research*, *25*(3), 353–376. Retrieved from `https://doi.org/10.1080/15378020.2021.1948296` doi: 10.1080/15378020.2021.1948296

Dearing, J. W. (2009, 1st Sep). Applying diffusion of innovation theory to intervention development. *Research on Social Work Practice*, *19*(5), 503–518. doi: 10.1177/1049731509335569

Dellarocas, C., Zhang, X. M. & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45.

Deveaud, R., Sanjuan, E. & Bellot, P. (2014, 06). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*. doi: 10.3166/dn.17.1.61-84

Díaz, E. (2017). The second-curve model: A promising framework for ethical consumption? veganism as a case study. In C. Bala & W. Schuldzinski (Eds.), *The 21st century consumer: Vulnerable, responsible, transparent? ; proceedings of the international conference on consumer research (iccr) 2016* (pp. 235–244). Düsseldorf: Kompetenzzentrum Verbraucher-

forschung NRW. Retrieved from https://doi.org/10.15501/978-3-86336-918-7_20 doi: 10.15501/978-3-86336-918-7_20

Gan, Q., Ferns, B. H., Yu, Y. & Jin, L. (2017). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality & Tourism*, *18*(4), 465-492. Retrieved from https://doi.org/10.1080/1528008X.2016.1250243 doi: 10.1080/1528008X.2016.1250243

Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A. & Marteau, T. M. (2019, 15th Oct). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(42), 20923–20929. doi: 10.1073/pnas.1907207116

Godfray, H. C. J., Aveyard, P., Garnett, T., Hall, J. W., Key, T. J., Lorimer, J., ... Jebb, S. A. (2018). Meat consumption, health, and the environment. *Science (New York, N.y.)*, *361*(6399). Retrieved from https://doi.org/10.1126/science.aam5324 doi: 10.1126/science.aam5324

Griffiths, T. & Steyvers, M. (2004, 04). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101 Suppl 1*, 5228-35. doi: 10.1073/pnas.0307752101

Hargreaves, S. M., Raposo, A., Saraiva, A. & Zandonadi, R. P. (2021, April). Vegetarian Diet: An Overview through the Perspective of Quality of Life Domains. *IJERPH*, *18*(8), 1-23.

Harrington, R. J., Staggs, A., Powell, F. A. & Ottenbacher, M. C. (2012). Generation y consumers: key restaurant attributes affecting positive and negative experiences. *Journal of Hospitality and Tourism Research*, *36*(4), 431–449. Retrieved from https://doi.org/10.1177/1096348011400744 doi: 10.1177/1096348011400744

Hu, N., Koh, N. S. & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, *57*, 42-53. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923613001942 doi: https://doi.org/10.1016/j.dss.2013.07.009

Janssen, M., Busch, C., Rödiger, M. & Hamm, U. (2016). Motives of consumers following a vegan diet and their attitudes towards animal agriculture. *Appetite*, *105*, 643–651. doi: 10.1016/j.appet.2016.06.039

Jeong, E. & Jang, S. (2011). Restaurant experiences triggering positive electronic word-of-mouth (ewom) motivations. *International Journal of Hospitality Management*, *30*(2), 356–366.

Knutson, B. J. (2000). College students and fast food: How students perceive restaurant brands. *Cornell Hotel and Restaurant Administration Quarterly*, *41*, 68–74. Retrieved from https://doi.org/10.1177/001088040004100316 doi: 10.1177/001088040004100316

Lia, X., Wu, C. & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information Management*, *56*, 172–184.

Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. (2016, 09). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*. doi: 10.1186/s40064-016-3252-8

Nilashi, M., Ahmadi, H., Arji, G., Alsalem, K. O., Samad, S., Ghabban, F., ... Alarood, A. A. (2021). Big social data and customer decision making in vegetarian restaurants: a combined machine learning method. *Journal of Retailing and Consumer Services*, *62*. Retrieved from https://doi.org/10.1016/j.jretconser.2021.102630 doi: 10.1016/j.jretconser.2021.102630

Phills, J. A., Deiglmeier, K. & Miller, D. T. (2008). Rediscovering social innovation. *Stanford Social Innovation Review*, *6*(4).

Ploll, U., Petritz, H. & Stern, T. (2020). A social innovation perspective on dietary transitions: diffusion of vegetarianism and veganism in austria. *Environmental Innovation and Societal Transitions*, *36*, 164–176. Retrieved from https://doi.org/10.1016/j.eist.2020.07.001 doi: 10.1016/j.eist.2020.07.001

Rogers, E. M. (1962). *Diffusion of innovations*. New York: Free Press.

Rosenbaum, P. & Rubin, D. (1983, 04). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55. doi: 10.1093/biomet/70.1.41

Tirunillai, S. & Tellis, G. (2014, 08). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*, 463-479. doi: 10.1509/jmr.12.0106

Waite, K. & Perez-Vega, R. (2017). Digital media and marketing interactivity. In G. Bell & B. Taheri (Eds.), *Marketing communications: An advertising, promotion and branding perspective* (pp. 151–166). Oxford: Goodfellow Publishers.

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

You, Y., Vadakkepatt, G. G. & Joshi, A. M. (2015). A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, *79*(2), 19–39.

Yuan, H., Xu, W., Li, Q. & Lau, R. (2018). Topic sentiment mining for sales performance prediction in e-commerce. *Annals of Operations Research*, *270*, 553–576.

Zhang, J. & Liu, R. R. (2019). The more the better? exploring the effects of reviewer social networks on online reviews. *Journal of Marketing Management*, *35*(17-18), 1667–1688. Retrieved from https://doi.org/10.1080/0267257X.2019.1666157 doi: 10.1080/0267257X.2019.1666157