Are Yelp Reviews Providing Insight to a Restaurant Closing Down?

|  |  |  |
| --- | --- | --- |
| ***Jessica Cruz-Nagoski***  Bloustein School of Planning and Public Policy  Rutgers University | ***Dayvonn Jones***  Bloustein School of Planning and Public Policy  Rutgers University |  |

**ABSTRACT**

**Keywords**Yelp, Review, Text, Textual, Analysis, Sentiment, Score, Regression

**1 INTRODUCTION**

Restaurant closure is an issue that has been on many restaurant owner’s minds after the pandemic. Even in normal economic conditions, as many as 61% of independently operated restaurants fail within three years of opening, according to a Times article (Vesoulis, 2021). The ongoing Covid-19 pandemic has made those odds infinitely worse. 60% of business closures due to the coronavirus pandemic are now permanently closed (Sundaram, 2020), and those that are still open are barely surviving. Unfortunately, aid from the government is not enough to support all the restaurants that need help. With all the applicants applying to the Restaurant Relief Fund, there was a need of nearly triple the amount of funding available (Beckett, 2021).

The restaurant industry is lacking research into what factors determine the success of the restaurant. We want to use factors found in Yelp to determine if they will predict if a restaurant will stay open or closed. Businesses have indicated to their customers via Yelp that they will not reopen their restaurant and the closure is permanent. The characteristics we will be analyzing are the customer satisfaction of a restaurant, a restaurant's price range, ambience, location, yelp star rating, and if the restaurant provides takeout, delivery, free parking, and WiFi. The customer satisfaction will be measured by extracting sentiment scores from the Yelp reviews of restaurants. The sentiment score and other characteristics will be extracted from the Yelp dataset to compare user experience and analyze the restaurant quality. Using yelp measures business activity with their users better than government collected data because Yelp data consists of real time data. The data is constantly being updated due to user experience and it provides sentiment analysis based on geographical location. The data is finer and can be used to analyze restaurant success.

With this project, we will be able to help policymakers and business owners understand the restaurant market better. Both big restaurant corporations and small business owners can understand how market demands influence the success of their restaurant, and put time and effort into restaurant upgrades that matter the most. Policymakers can also understand what factors they should focus on when trying to help restaurants stay open during hard times like COVID-19.

**2 LITERATURE REVIEW**

We divided our literature review into 3 sections: Literature Review of Textual Analysis, Literature Review of Using Yelp Reviews for Restaurant Dynamics, and Literature Review of Restaurant Success. These were the three main key words we were using when doing our literature review. For each section, we provide a table summarizing all the articles.

**Literature Review of Textual Analysis**

To have a further understanding of textual analysis and what can be done, we read through the required articles for our literature review. Below in **Table 1**, we have summarized the 12 articles by research question, methodology, and takeaway.

**Table 1. Text Analysis Literature Review**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Article** | **Research Question** | **Methodology** | **Takeaway** |
| A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community | Introducing the concept of LDA/LSA in text analysis (specifically for behavioral scientist) and showing how to do it in R | Pre-processing:   * Stemming * Remove stop words * Orthographic transformations * Stripping punctuation * Identifying named entities * Substitution   Modeling:   * Build tdm * Run LSA | LDA/LSA is great for topic modeling |
| Implicit Gender Bias in Linguistic  Descriptions for Expected Events:  The Cases of the 2016 United States  and 2017 United Kingdom Elections | Does gender biases in language reflect subjective beliefs or is translating our thoughts to language the thing that is biased? | Data:   * Between June ‘16 to Jan ‘17, they collected data relating to US presidential campaign and UK General Election * Recorded who they thought was going to win and then had them do text-completion tasks to see what pronoun they used?   Methodology:   * Used Bayesian Hierarchical generalized linear models | Use of pronouns changed as events changed  Severe bias against she |
| Public Perceptions of COVID-19 Vaccines: Policy Implications  from US Spatiotemporal Sentiment Analytics | What is the temporal dynamics of public sentiment towards the COVID-19 vaccine | * Sentiment analysis on vaccine tweets * monitor changes in public sentiment over time * contrast vaccine sentiment scores with actual vaccination data from the US CDC and HPS * explore the influence of maturity of Twitter user-accounts * generate geographic mapping of tweet sentiments * Use the Public Sentiment Scenarios (PSS) framework for implications for vaccines | Despite overall strength of positive sentiment and despite increasing numbers of Americans being fully vaccinated, negative sentiment towards COVID-19 vaccines still persist ppl who are hesitant towards it |
| COVID-19 Public Sentiment Insights and Machine  Learning for Tweets Classification | Address and better understand COVID-19s informational crisis and gauge public sentiment | Data:   * Coronavirus specific Tweets and R statistical software   Method:   * Sentiment analysis of fear-sentiment over time * Naive Bayes vs logistic regression to classify Tweets |  |
| Automating Discovery of Dominance in Synchronous Computer-Mediated  Communication | Investigates dynamics and characteristics of dominance in virtual interaction by analyzing electronic chat transcripts of groups solving a hidden profile task | Data:   * Transcripts of electronic communications of groups solving a decision-making task   Methodology:   * extracted relevant information both manually and automatically and generated descriptive summaries * Automatic program used custom-developed parsing program that identified variables | Automatic content search is easier than manual and still gives similar results |
| The Psychological Meaning  of Words: LIWC and  Computerized Text  Analysis Methods | Linguistic Inquiry and Word Count (LIWC) was created and validated. LIWC is a transparent text analysis that counts words in psychologically meaningful categories | Processing Component   * goes through each words in a text file and compares it to the dictionary   Dictionaries   * Attentional focus, emotional state, social relationships, thinking styles, and individual differences | A new way of contextualizing text analysis using social and psychological meanings of words rather than just the content of words |
| Feeling Positive About Reopening?  New Normal Scenarios From  COVID-19 US Reopen  Sentiment Analytics | The public have different feelings about the covid reopening. People were feeling and thinking more positively about the covid reopening. In addition people were unhappy with the fake news on COVID-19. | Data:   * Twitter tweets during the time between January 4 and May 4   Methodology   * Sentiment analysis of reopening using public sentiment scenarios. * Descriptive analysis of tweets using R. Word frequency and N-grams. | Research conceptualized useful sentiment analysis which can be used for future crises. |
| Natural Language Based Financial  Forecasting: A Survey. | Investigates if Natural Language Based Financial Forecasting can yield accurate results in stock prediction. | Data   * Corporate disclosures, financial reports, professional periodicals, aggregated news, message boards, social media.   Methodology   * Various algorithms with linear regression models. * Natural Language PRocessing (NLP) | Semantic, sentiment, and event representation is extracted. There is a direct correlation with public sentiment and market trends, illusion of growth, positive effect with predictions and poor market trends. |
| Text Analytics: the convergence of Big Data and  Artificial Intelligence | Describes the applications of textual analytics and the emergence of computational linguistics. | Techniques   * Information extraction through named entity recognition. * Topic tracking through keyword analytics. * Text summarization with deep and shallow analysis. * Clustering with defined topics. * Categorization and classification by theming each document source. * Concept linking by combining attributes in each document. * Information visualization by mapping to produce a browsing capable source. * Question answering with predictive answers to fit the themes and typologies. * Deep learning involving recurrent neural networks, and convolutional neural networks. | Information that is expressed as text contains data that can be used with cognitive applications to filter information and provide pathways and clarity to important and relevant information. |
| Advancing the Field of Writing  Analytics: Lessons from “Text-as-  Data” in the Social Sciences | Review of textual data as political science, burgeoning area of research, and quantitative tool for the analysis of political text. | Key Points   * Behavioral revolution changed the way political scientists review text. * Political Text Transitioned into Data * Exploring lexical data and tone from judicial records because of the source of descriptive detail. * Elitist communication effort through speeches, press releases, political advertisement, and lobbying data. * Public opinion through polls and survey data. | Text as data is more that writing analytics. Writing analytics is more interdisciplinary and multidisciplinary than political text as data. Learning new techniques would allow more substantive research results. |
| Conceptual Frameworks for Big-Data Visualization: Discussion on Models,  Methods and Artificial Intelligence for Graphical Representations of Data. | How can we meaningfully and effectively exploit big data so as to extract insights in ways that are friendly to human sensory abilities and cognitive capabilities? | Methodology   * Visual analytics by analytical problems and general application areas. * Principal Component Analysis to represent data in a condensed dimensional structure. * Multi-Dimensional Scaling to map higher dimensional data to lower dimensional data.   Model   * Line and bar graphs, word frequency visualizations, | The growing availability and notoriety of artificial intelligence will help people perform better data visualization. |
| Semantic Network Analysis as a Method for Visual Text Analytics | If semantic models can be used to support knowledge building, analytical reasoning, and explorative analysis. | Model   * Network structure by statistical methods. * Edges that represent the relationship between two models using statistical quantities of adjacent nodes. * Path is a set of connected edges which describes the semantic relations. * Nodes, hubs, subgraphs, and clusters. | General methods for semantics that reviews quantitative and qualitative perspectives from the analysis and the interpretation of network structures. |

**Literature Review of Using Yelp Reviews for Restaurant Dynamics**

We tried looking for studies as similar as possible to our research questions: Yelp Reviews for Restaurant Success. We couldn’t find studies exactly on what we were looking for, but we found various studies using yelp reviews for other types of restaurant dynamics like exploring gentrification, sentiment analysis for new restaurant features, and neighborhood similarities. These were all very cool projects that are outlined in **Table 2** that gave us confidence to use Yelp as a valid economic indicator.

**Table 2. Yelp Literature Review**

|  |  |
| --- | --- |
| **Name** | **Findings** |
| Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change | * Combined data on businesses from Yelp with data on gentrification from the Census, Federal Housing Finance Agency, and Streetscore (an algorithm using Google Streetview)   + Didn’t use text from yelp reviews; instead they used changes in restaurant activity as their data * Found that gentrifying neighborhoods tend to have growing numbers of local groceries, cafes, restaurants, and bars. * For example, the entry of a new coffee shop into a zip code in a given year is associated with a 0.5 percent increase in housing prices. * Yelp measures of local business activity provide leading indicators for housing price changes and help to forecast which neighborhoods are gentrifying |
| Are Yelp's tips helpful in building influential consumers? | * The recommendations are built on value co-creation of content and history that examines the credibility and trust of the author. * Affective trust, cognitive trust, and perceived credibility are importance factors in a user’s comments. * Materials were yelp reviews and the methods were a support vector machine for data mining algorithms. * A group of specific keywords were determined to have reproducibility and a strong effect on a reader’s trust for the review. |
| Why do you use Yelp? Analysis of factors influencing customers’ website adoption and dining behavior | * Analysis of the adoption of behavior of a reader to a review. Restaurant review websites have a directly relationship with a customer’s dining decision. * A literature review was performed to create a Technology Acceptance Model that has explanatory power. * Several proposed relationships were used a hypothesis to create hypothetical decision making processes. * The study examined the reason why customers preferred different websites to read reviews that guide their decision making. |
| The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews | Research Question:   * This study examined Japanese restaurant reviews in English at Yelp.com and those in Japanese at Yelp.co.jp from a cross-cultural perspective   Methodology:   * Using bilingual text mining software, these findings shed insights on how review contents and ratings may vary between local and foreign customers |
| Reading the city through its neighbourhoods: Deep text embeddings of Yelp reviews as a basis for determining similarity and change | Research Question:   * This paper develops novel methods for using Yelp reviews as a window into the collective representations of a city and its neighbourhoods   Methodology:   * Clean and vectorize the reviews to 1024 words * Put the reviews through an autoencoder model (a neural net)   Data:   * The dataset includes approximately 24,000 businesses and 500,000 reviews. * our dataset also includes information about each POI's category (e.g. ‘sushi restaurant’ or ‘rock climbing’). * Each review is also indexed to the reviewer who wrote it, Business ID, a User ID, and a location (lat/lon) associated with each review.   Takeaway:   * They do reviews over time and map cosine similarities between each neighborhood |
| Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models | Research Question:   * We use over 350,000 Yelp reviews on 5,000 restaurants to compare the effectiveness of several machine learning and deep learning models on predicting user sentiment (negative, neutral, or positive).   Methodology:   * we perform an ablation study on several text preprocessing techniques (e.g., stop word removal, normalization) using a simple multinomial Naive Bayes model * Afterwards, using the preprocessed data, we train several machine learning (e.g., Logistic Regression, Support Vector Machine) and deep learning (e.g., LSTM, BERT) models and compare their effectiveness at predicting sentiments   Takeaway:   * For machine learning models, we find that using binary bagof-word representation, adding bi-grams, imposing minimum frequency constraints and normalizing texts have positive effects on model performance. For deep learning models, we find that using pre-trained word embeddings and capping maximum length often boost model performance |
| Identifying Restaurant Features via Sentiment  Analysis on Yelp Reviews | Research Question:   * Yelp offers not enough information for independently judging its various aspects such as environment, service or flavor, therefore they use sentiment analysis to measure certain features   Methodology:   * The attributes we used were business id, business categories, review content and rating * Use bag of words (did a dictionary instead of vectors) and preprocessing on each tweet * In order to find specific words that were used to indicate customers’ concerns for the restaurant, or by moving forward exploring the unique characteristic of each restaurant category, adjectives that simply describing the polarity of sentiment (i.e. “good”, “amazing”, “terrible” and etc.) were neglected * SVM model actually calculate a total score for each review, and this score to some extent indicates how satisfied or discontented the customer is. The polarity score we calculated shows how much a word contributes to the score of all restaurants of a certain type * Then for each category of restaurants, the top positive and negative words are extracted. We may discover what are the special features for each type and the discrepancy of those restaurants providing great food around the world |

**Literature Review on Restaurant Success**

Finally, we wanted to see what indicators were important in understanding a restaurant’s success, because we wanted to test for other potential factors besides Yelp reviews. From the studies that are outlined in **Table 3**, we found out that sociodemographic factors have a significant impact on restaurant success due to demand characteristics.

**Table 3. Restaurant Success Literature Review**

|  |  |
| --- | --- |
| **Article Name** | **Findings** |
| Exploring advertising strategy for restaurants sourcing locally: The interplay of benefit appeal and regulatory focus by Sun Hwa Kim, Ran Huang, Seeun Kim | * Restaurants have not found a proper way to advertise that the food they purchased is sourced locally. * Content characteristics in local food advertising is not a widely explored topic in restaurant advertising. * Sourcing local food to be served as restaurants benefit the producers and make the consumers feel more self benefit. * Authenticity, positive affects on customers, and patronage intention are all displayed when a restaurant effectively advertises their local food source. * Four (4) hypotheses are explored. * Methodology includes research design with participants, and manipulation checks. * It is unclear how the restaurants can deliver this information to the consumers. But an intention behavior gap exists and for that the context of the advertisement must be specifically focused. |
| Understanding and projecting the restaurantscape: The influence of neighborhood sociodemographic characteristics on restaurant location | * Analysis of the relationship of restaurant location and socio-demographic status. * Nationwide data base review of restaurants with different cuisine types based on frequency and location. Some major chain locations were included in the location patterns. * Initially certain types of restaurants showed a pattern for location choosing by specific demographic characteristics of a neighborhood. * Negative binomial regressions were used to explore chain restaurant affiliations with demographic characteristics. * Used Esri to predict future restaurant growth with projected demographics. * Restaurant locations are a strategic decision based on the market demand characteristics. |
| How restaurant owners manage strategic risk | * Case study of restaurants as they experienced operational trades and the responses to these operations changes. * The restaurateur’s perception of risk was examined and categorized into financial, marketing, or operational. * The five case studies reviewed how the restaurant analyzed their risk and their intended response. |
| Language representation of restaurants: Implications for developing online recommender systems | Research Question   * Do restaurants understand the implications for online recommendations?   Methods   * Collected customer’s verbal reviews of the restaurants. * Examined descriptive text in the restaurant’s online publications. * Multivariate text analysis.   Results   * Restaurant operators and owners often ignore the opinions of the consumers when they create the opinions of the restaurant qualities. * The consumer's opinion creates a more accurate description of the quality and service for a restaurant. |
| Yelp data shows 60% of business closures due to the coronavirus pandemic are now permanent | Article Summary |
| Most restaurants fail. covid-19 made the odds even worse. | Article Summary |

**3 LITERATURE SYNTHESIS**

When doing our literature review, we were trying to find ways people had used data to better understand the restaurant industry. Upon doing so, we found two studies. These are outlined in the section below called “Comparing studies on restaurant dynamics”. From these studies, we understood that we would be the first to use data on restaurant characteristics to understand how restaurants were doing. The studies we found used environmental factors like real estate and the commercial layout of an area as factors of restaurant dynamics. One of the studies found significant results using Yelp, which gave us confidence that Yelp was a valid indicator of restaurant success. After that, we compared studies on sentiment analysis methodologies using Yelp Reviews in the section below called “Comparing Yelp review textual analysis methods”, since we were interested in doing some sort of text analysis on the reviews.

**Comparing studies on restaurant dynamics**

Studies on restaurant success have all done so by including external factors such as rent prices and demographics. Since we are only using Yelp Reviews, it is important to understand how external, environmental factors can influence the success of a restaurant. To explore this further, we focused on two studies.

The first study we saw was a Kaggle competition that called all data scientists to create a model that used demographic, real estate, and commercial data to determine whether the investment of a restaurant was worth it. They determined the worth of a restaurant by using restaurant revenue. We don’t have restaurant revenue, but we do have whether a restaurant has stayed open or not from the Yelp dataset. Therefore, we plan on doing a similar experiment to the Kaggle competition, but instead we would use Yelp reviews as a metric in combination to the other factors, and have whether the restaurant closed down or not as our dependent variable.

The second study used Yelp data to gain insight into gentrification of a neighborhood. By combining data on businesses from Yelp with data on gentrification from the Census, Federal Housing Finance Agency, and Streetscore (an algorithm using Google Streetview), they were able to find that gentrifying neighborhoods tend to have growing numbers of local groceries, cafes, restaurants, and bars. For example, the entry of a new coffee shop into a zip code in a given year is associated with a 0.5 percent increase in housing prices. From this paper, we can assume that Yelp can provide leading indicators for things like housing price changes which can be used to determine the gentrification of a neighborhood. This made us believe that we can use Yelp as a valid indicator to determine how the restaurant market is doing.

While this paper provided us a lot of good insight, there are changes we need to make in order to better answer our research question. We want to take this a step further by using text from the yelp reviews rather than collecting the changes in restaurant activity. To do so, we need to do a literature review of different text analysis methods people have applied to yelp reviews in order to extract information from them.

**Comparing Yelp review textual analysis methods**

After doing the literature review on textual analysis methods on yelp reviews, we realized that sentiment analysis is the most popular method applied from Yelp Reviews. While Yelp’s five-star rating system is good, most people use Yelp for the reviews rather than the star rating. Therefore, sentiment analysis has been an effective way to turn the text of the reviews into a numerical number that people can use rather than look through all the reviews. Sentiment Analysis would then be a good metric for us to extract out of the Yelp reviews that we can use to determine restaurant quality. To learn how to do that, we focused on two studies that will be explained more below.

In a study titled “Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews”, the authors wanted to use sentiment analysis to provide more insight to a restaurant’s environment, service, or flavor. Therefore, they preprocessed each review by removing punctuation and used the bag-of-words method for a SVM model to get a polarity score. A polarity score seems like a good metric from text to use for analyzing customer satisfaction, which can be a good indicator of restaurant quality.

In another paper titled “Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models”, once again sentiment analysis was applied to yelp reviews. They tested out many different sentiment analysis methods and came to the conclusion that bag-of-words improved the sentiment analysis model performance the best in addition to adding bi-grams, imposing minimum frequency constraints and normalizing texts.

These papers are useful for us to understand the processing steps other researchers took when using Yelp Reviews. After reading about the most effective methods, we will apply those steps to our own preprocessing and methodology of sentiment analysis for Yelp Reviews.

**4. METHOD**

To do our project, we first cleaned the original data. On our cleaned data, we added a sentiment score of the reviews. After our data includes the sentiment scores, we performed an Exploratory Data Analysis which helped us understand the weaknesses of our dataset. Finally, we performed a logistic regression to better understand which restaurant characteristic was the most significant in restaurant closure. Below we have described in detail the steps we took to clean the data, do the sentiment analysis, logistic regression, and visualizations. All of our code can be found at our github[[1]](#footnote-1).

**Cleaning Original Dataset**

We created a shortened dataset from the original dataset, because the original dataset had many different types of businesses and variables that we weren’t interested in for our research question. To get this shortened dataset, we performed an inner join by the business\_id on the business and review json files. Then, we filtered out any reviews that didn’t contain the words “Food” or “Restaurant” in the categories column. We ended up with 4,696 reviews to do our analysis on. We then needed to clean the independent variables in our dataset so that we could do an exploratory data analysis and a regression analysis. Below we will discuss which variables we cleaned up and how we did it. You can find our code on how we cleaned the dataset under the “Data Cleaning” folder in our github.

*Attributes Variable*

This was definitely the variable that took the most work to clean. Our attributes variable was in the form of a dictionary with 35 distinct attributes. Out of the 35 different attributes, we were interested in only 6. These attributes included the restaurant’s price range, ambience, take out status, delivery status, parking status, and WiFi status. Therefore, we extracted these 6 attributes from the ambience column and put them into their own columns in our dataset. The parking status and ambience of a restaurant was further split up into more columns because they had their own dictionaries. From the ambience column, we extracted whether the restaurant was trendy, hipster, upscale, and touristy as boolean variables. From the parking column, we extracted whether the restaurant had free parking or not as a boolean variable. In the end, we were left with 9 new columns: ‘RestaurantPriceRange’, ‘RestaurantTakeOut’, ‘RestaurantDelivery’, ‘Wifi’, ‘FreeParking, ‘Trendy’, ‘Hipster’, ‘Upscale’, ‘Touristy’.

*Categorizing and encoding the variables*

It is necessary to encode any variables that are not numerical for the regression model to work. **Table 4** shows how we encoded many of our variables into numerical values. Many of our newly extracted variables from the attributes column had n/a values, so we encoded n/a values to 0. We assumed that if a restaurant did not have that variable recorded, it was because it didn’t exist.

**Table 4. The Variables In Our Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Original Values** | **Encoded Values** |
| business\_id | ID for the restaurant given by Yelp | strings | Didn’t have to encode |
| text | Reviews of the restaurant | strings | Didn’t have to encode |
| latitude | Latitude of where the restaurant is | Numerical points | Didn’t have to encode |
| longitude | Longitude of where the restaurant is | Numerical points | Didn’t have to encode |
| star\_y | Star rating of the business | [1.5 - 5] | Didn’t have to encode |
| RestaurantsPriceRange | Price range of restaurant ;  Extracted from ‘Attributes’ column | [0,1,2,3] | Didn’t have to encode |
| RestaurantsTakeOut | If the restaurant has take out ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| RestaurantsDelivery | If the restaurant has delivery ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| WiFi | If the restaurant has WiFi ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| Trendy | If the restaurant is “trendy” ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| Hipster | If the restaurant is “hipster” ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| Upscale | If the restaurant is upscale ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| Touristy | If the restaurant is popular with tourists ;  Extracted from ‘Attributes’ column | [True, False, n/a] | n/a = 0,  False = 0,  True = 1 |
| Regions | What region the restaurant is according to the regions set by the U.S. Census Bureau ;  Extracted from ‘state’ column | ['AZ', 'CA', 'NV', 'AB', 'DE', 'FL', 'LA', 'TN', 'ID', 'IL', 'IN', 'MO', 'NJ', 'PA'] | {'AZ', 'CA', 'NV'} = 0,  {'AL', 'DE', 'FL', 'LA', 'TN'} = 1,  {'ID', 'IL', 'IN', 'MO'} = 2,  {'NJ', 'PA'} = 3  {‘AB’} = 4 (Canada) |
| Compound | The sentiment score of a restaurant from the reviews;  Extracted from the ‘text’ column | numerical | Did not need to encode |

**Sentiment Analysis on the Reviews**

We wanted to extract a sentiment score for each restaurant, therefore we decided to use the VADER model. The cool thing about VADER is that you can apply it directly to unlabeled text data. VADER actually prefers untouched text because every capitalization and punctuation is used in VADER to analyze emotion intensities. For example, if a word was in all capitals and had lots of exclamation points, then VADER would recognize the word as very intense and take it into account when calculating the sentiment score. This is great for text from social media since people tend to talk on social media in an expressive way. After we applied VADER to our reviews, we extracted the compounded score (computed by normalizing the negative, positive, and neutral scores) as our sentiment score, and then found the median sentiment score per restaurant. We picked median over mean, because outliers tend to skew the results when using mean. You can find the code for the sentiment analysis at Sentiment Analysis.ipynb in the “Results” folder in our github.

**Logistic Regression Analysis**

To better understand what factors influence a restaurant’s ability to stay open, we ran a logistic regression model with the binary variable ‘is\_open’ as our dependent variable and 12 variables from our dataset shown in **Table 5** as our independent variables. We chose logistic regression, because our dependent variable is binary. To do this, we used glm() in R with family = ‘binomial’. The results of the logistic regression will be discussed in the discussion section. You can find the code for the logistic regression analysis at LogisticRegression.R in the “Results” folder in our github.

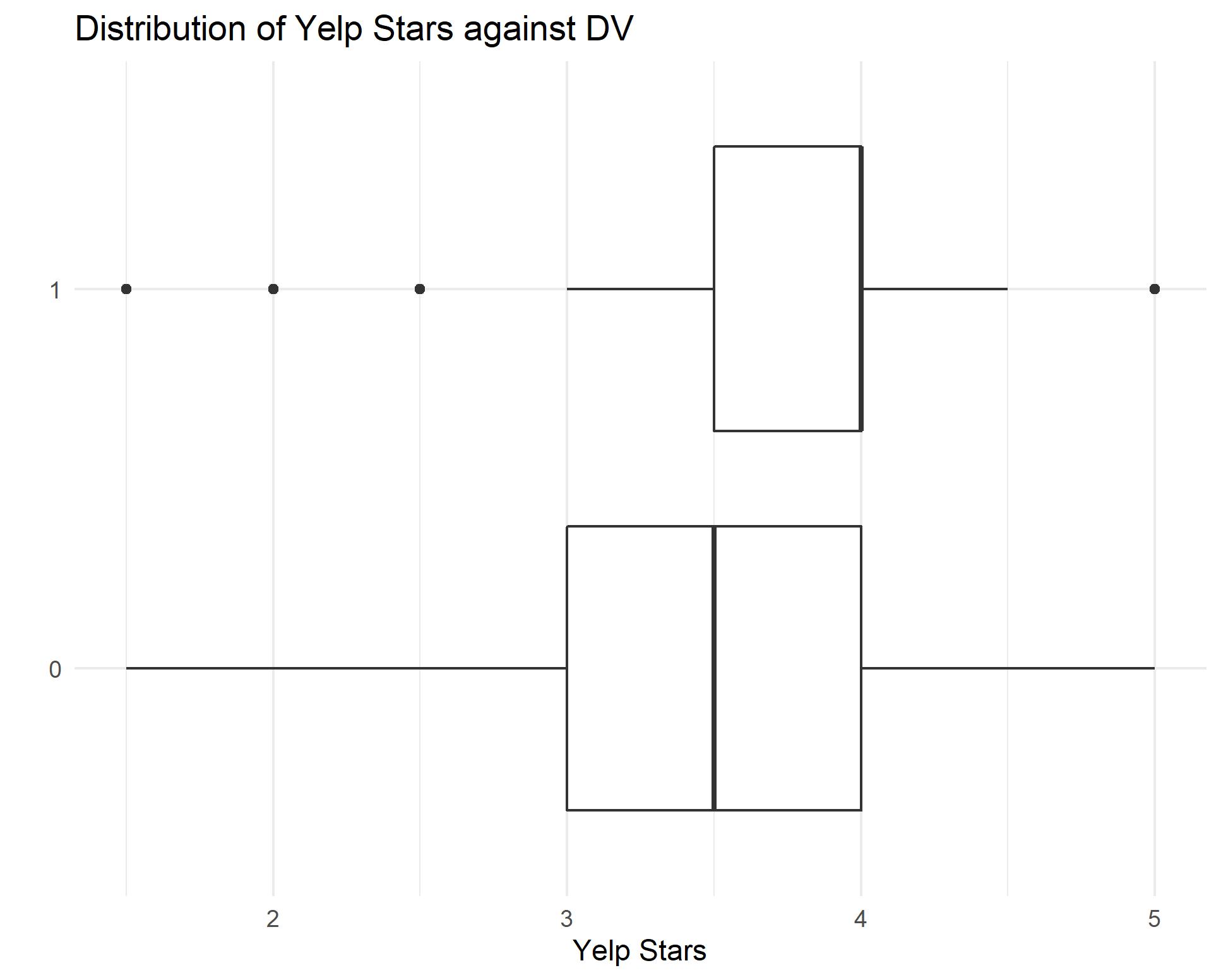
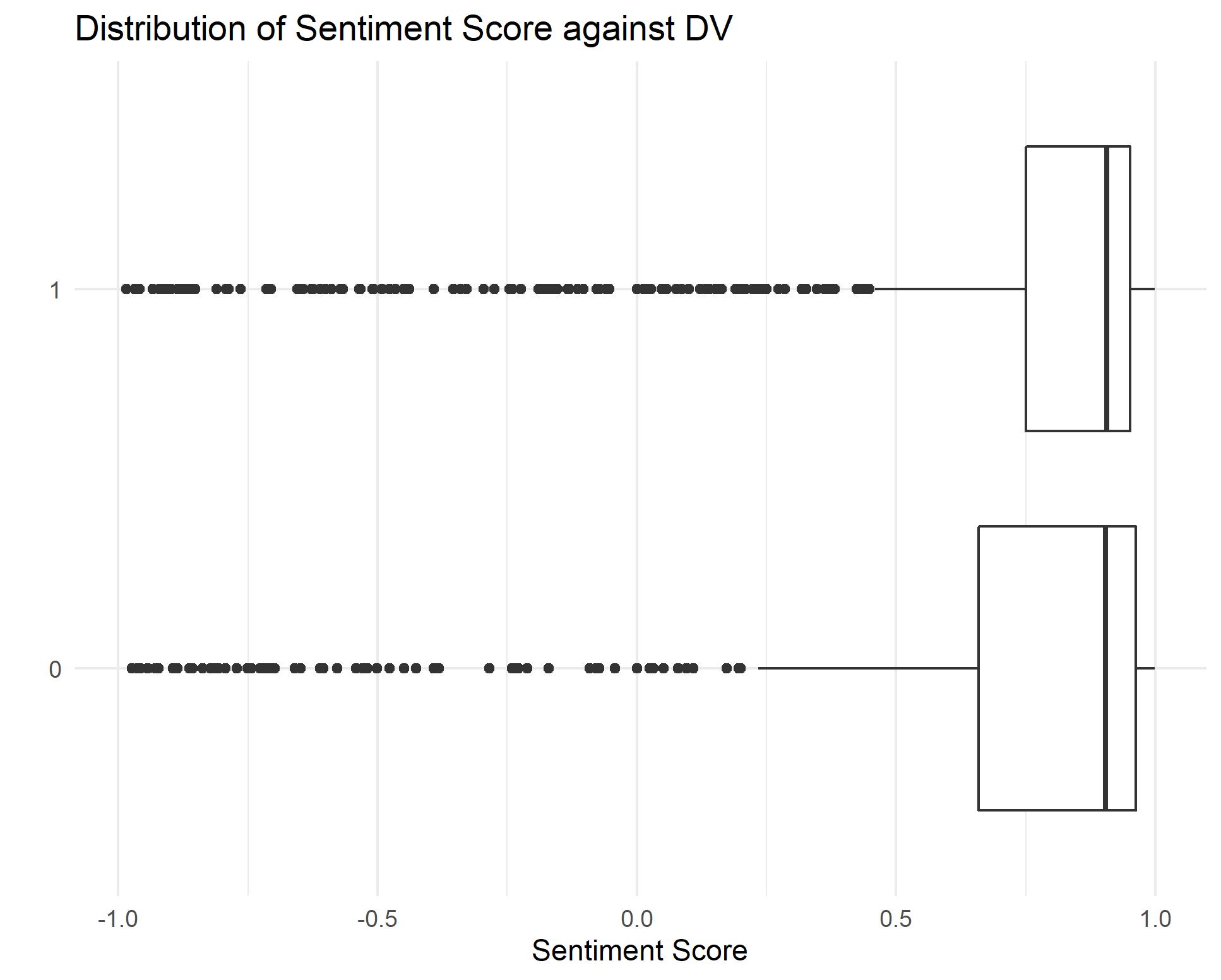
**Visualizations**

We performed all of our visualizations for our project in R using ggplot and wordcloud. We mainly used visualizations for our Exploratory Data Analysis where you can find the code for in the “Exploratory Data Analysis” folder on github. The few visualizations we used in our regression can be found at LogisticRegression.R in the “Results” folder in our github.

**5. DISCUSSION**

**Sentiment Analysis**

After we computed the median sentiment scores as outlined in the Methods section, we did a box plot to understand how the sentiment score differed between open and closed restaurants as seen in **Figure 1**.

****

**Figure 1. Sentiment Score (left) vs. Yelp Star Rating (right) Between Open and Closed Restaurants**

Our sentiment score box plot showed us that restaurants that were open tended to have a more positive distribution of sentiment scores than restaurants that were closed. This means that consumer satisfaction of a restaurant does have an influence on whether a restaurant stays open or closed. But since our yelp reviews were heavily biased to positive reviews, we needed to compare it to yelp star rating. In **Figure 1**, you can see that Yelp Stars box plot showed similar results to the sentiment score. This means a restaurant’s Yelp star rating is reflective of the sentiment of the review, which confirms that sentiment of a review is a good indicator of customer satisfaction. This box plot showed us that there was a correlation between positive sentiment and a restaurant staying open, but now we need to see if we could state that this variable was significant in comparison to other restaurant characteristics, therefore we performed a logistic regression.

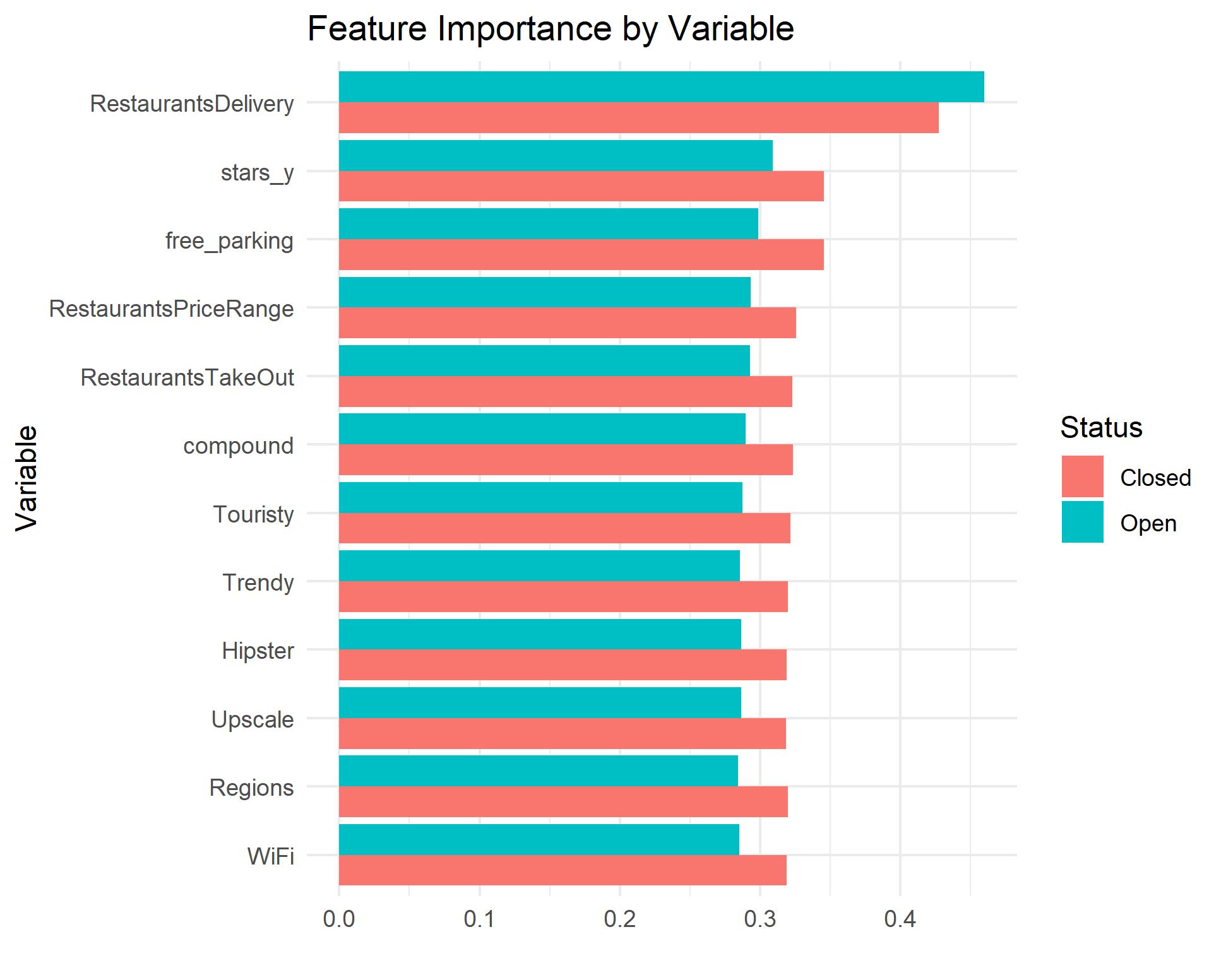
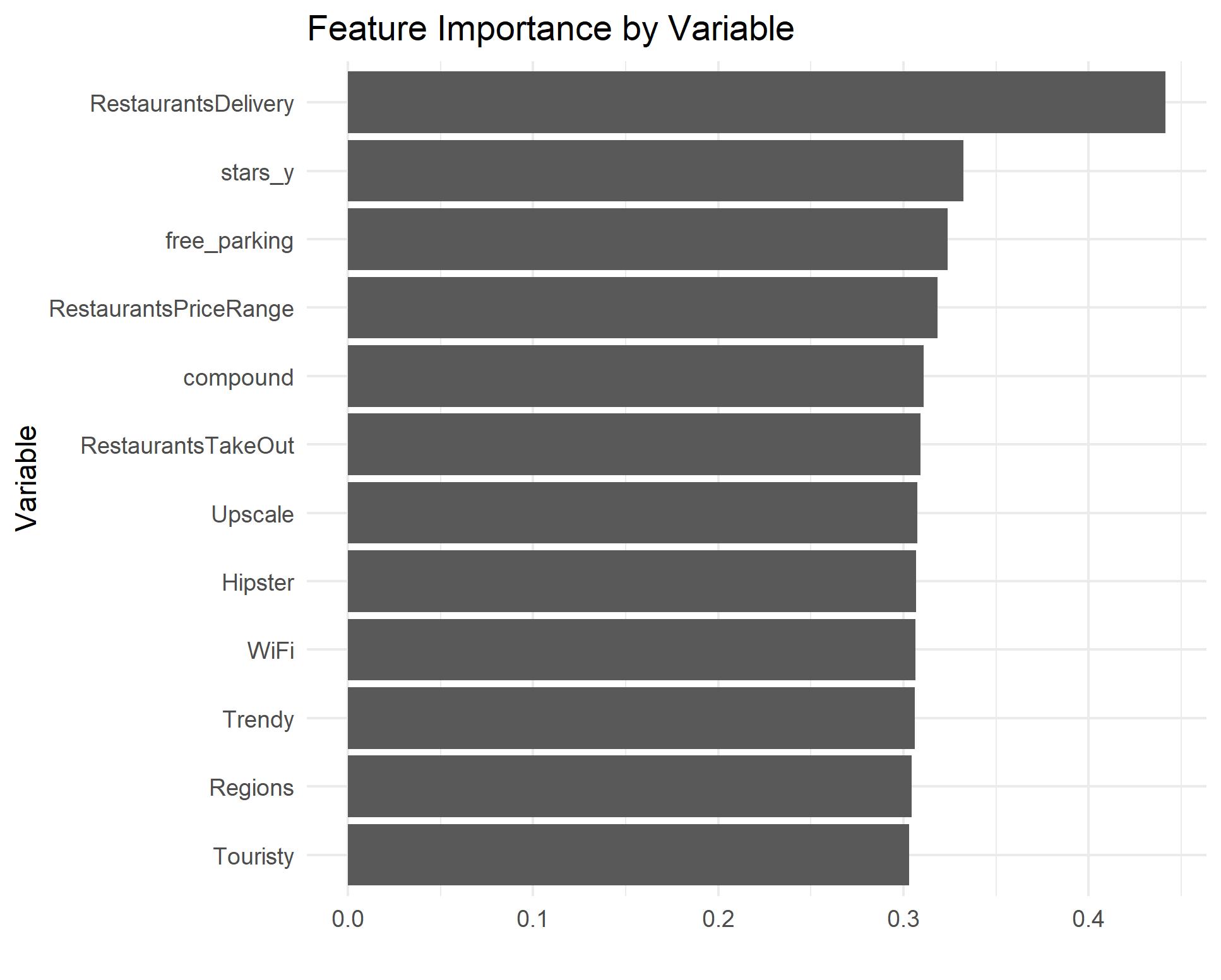
**Logistic Regression**

Performing a logistic regression showed us that the sentiment score, which we called “Customer Satisfaction” was not a significant predictor in restaurant closure as seen in Table 5. Only the coefficients with highlighted P Values are considered significant. This means that Yelp Stars, Delivery Status, Free Parking, and Price Range were all significantly variables of restaurant closure. The higher the rating, lower the price, and if the restaurant had delivery and free parking meant there was a higher chance the restaurant would stay open as seen by the estimates.

**Table 5. Summary of Logistic Regression Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Std. Error** | **Z Value** | **P Value** |
| **Yelp Stars** | 0.39152 | 0.09944 | 3.937 | 8.24e-05 \*\*\* |
| **Delivery** | 1.35208 | 0.12850 | 10.522 | < 2e-16 \*\*\* |
| **Free Parking** | 0.41544 | 0.12501 | 3.323 | 0.00089 \*\*\* |
| **Price Range** | -0.28659 | 0.11573 | -2.476 | 0.01327 \* |
| **Customer Sat.** | 0.14931 | 0.14104 | 1.059 | 0.28976 |
| **Takeout** | -0.36034 | 0.25051 | -1.438 | 0.15031 |
| **Touristy** | 0.88322 | 0.51519 | 1.714 | 0.08646 |
| **WiFi** | 0.10152 | 0.12456 | 0.815 | 0.41504 |
| **Trendy** | -0.04981 | 0.24180 | -0.206 | 0.83680 |
| **Hipster** | -0.20376 | 0.33205 | -0.614 | 0.53945 |
| **Upscale** | 0.37376 | 0.51851 | 0.721 | 0.47101 |
| **Regions** | -0.06160 | 0.05985 | -1.029 | 0.30333 |
| **(Intercept)** | -1.16314 | 0.43317 | -2.685 | 0.00725 \*\* |

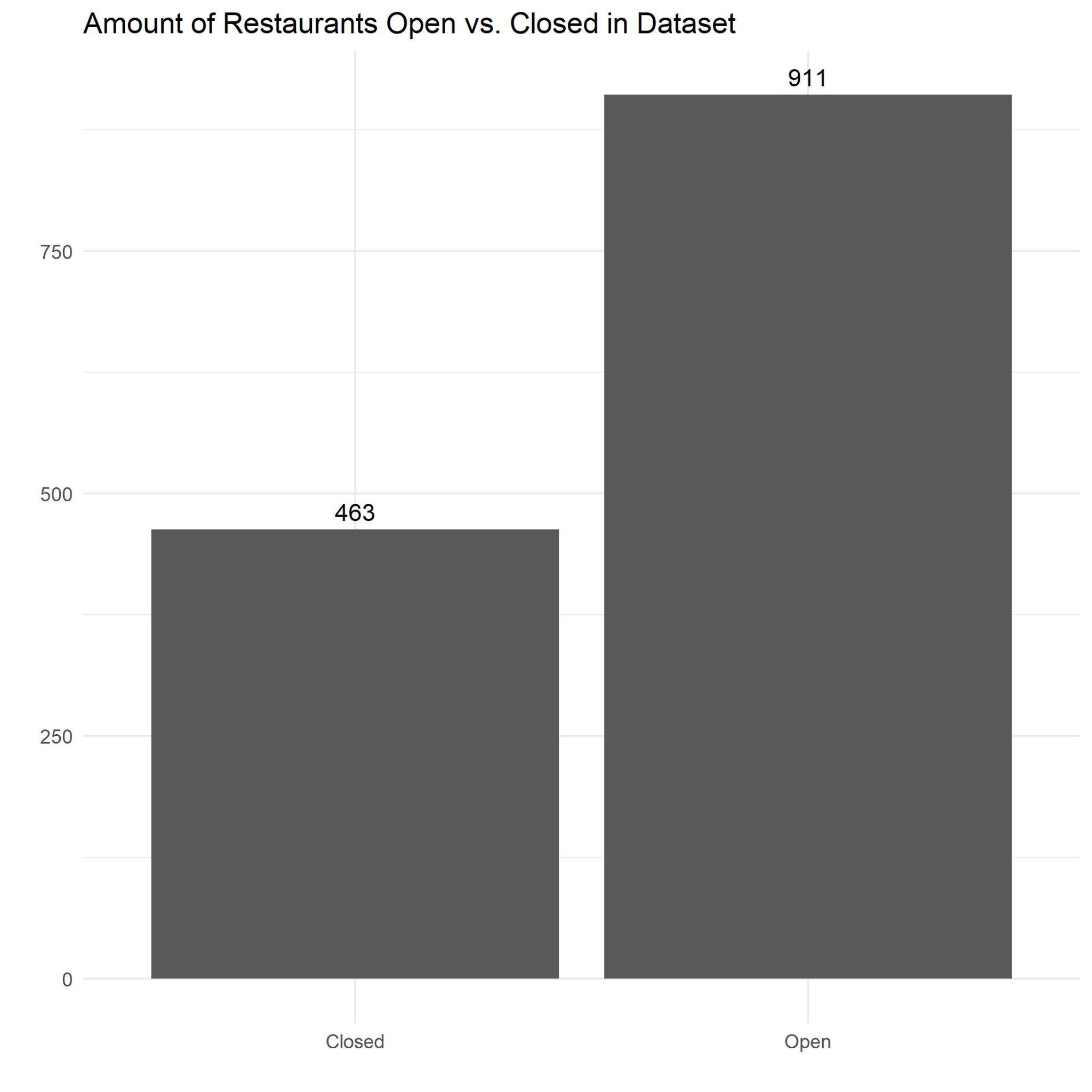
Out of the significant variables, Restaurant Delivery was the most important feature in predicting restaurant closure as seen on the left of **Figure 2**. When looking at the right of **Figure 2**, you can see that restaurant delivery was more important in predicting an open restaurant rather than restaurant closure. All the other characteristics were more important in predicting a closed restaurant.

****

**Figure 2. Feature Importance (left) and Feature Importance for Open vs Closed (right)**

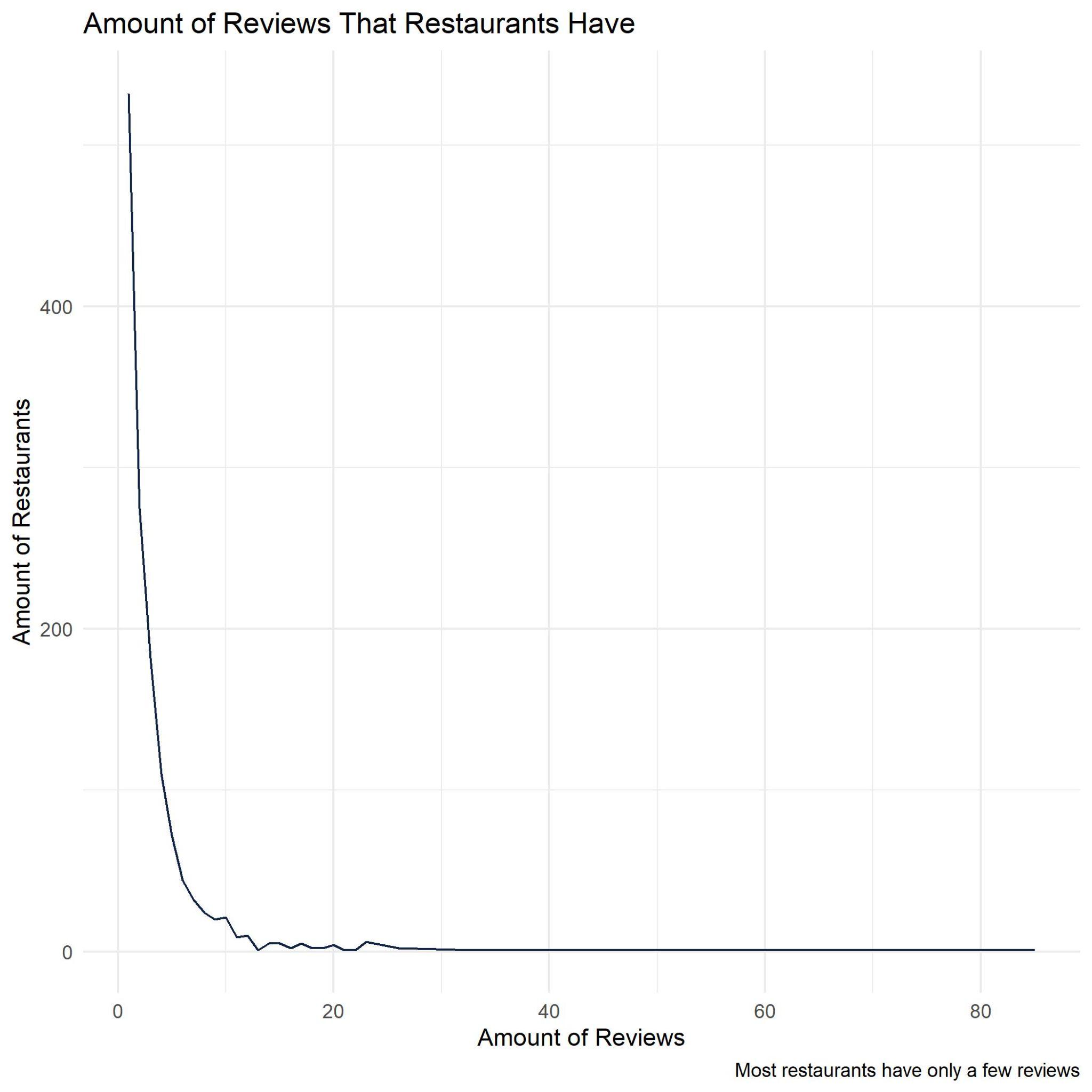
**6. WEAKNESSES, FUTURE RESEARCH**

Performing an exploratory analysis on our shortened dataset allowed us to spot any limitations in our dataset. Our first limitation was that we only had 4,696 reviews of a total of 1,371 restaurant. This is not nearly enough to understand the current restaurant industry as there are about 700,000 restaurants in the United States alone (Statista, 2018). This analysis provides just a small snapshot of how the restaurant industry is performing. Out of the 1,374 restaurants in the dataset, around 75% of the restaurants were still open (as seen by **Figure 3**) during the date the dataset was released in 2018. Ideally we would like a 50/50 distribution, but we felt that this was a good enough distribution that we didn’t need to oversample our data.



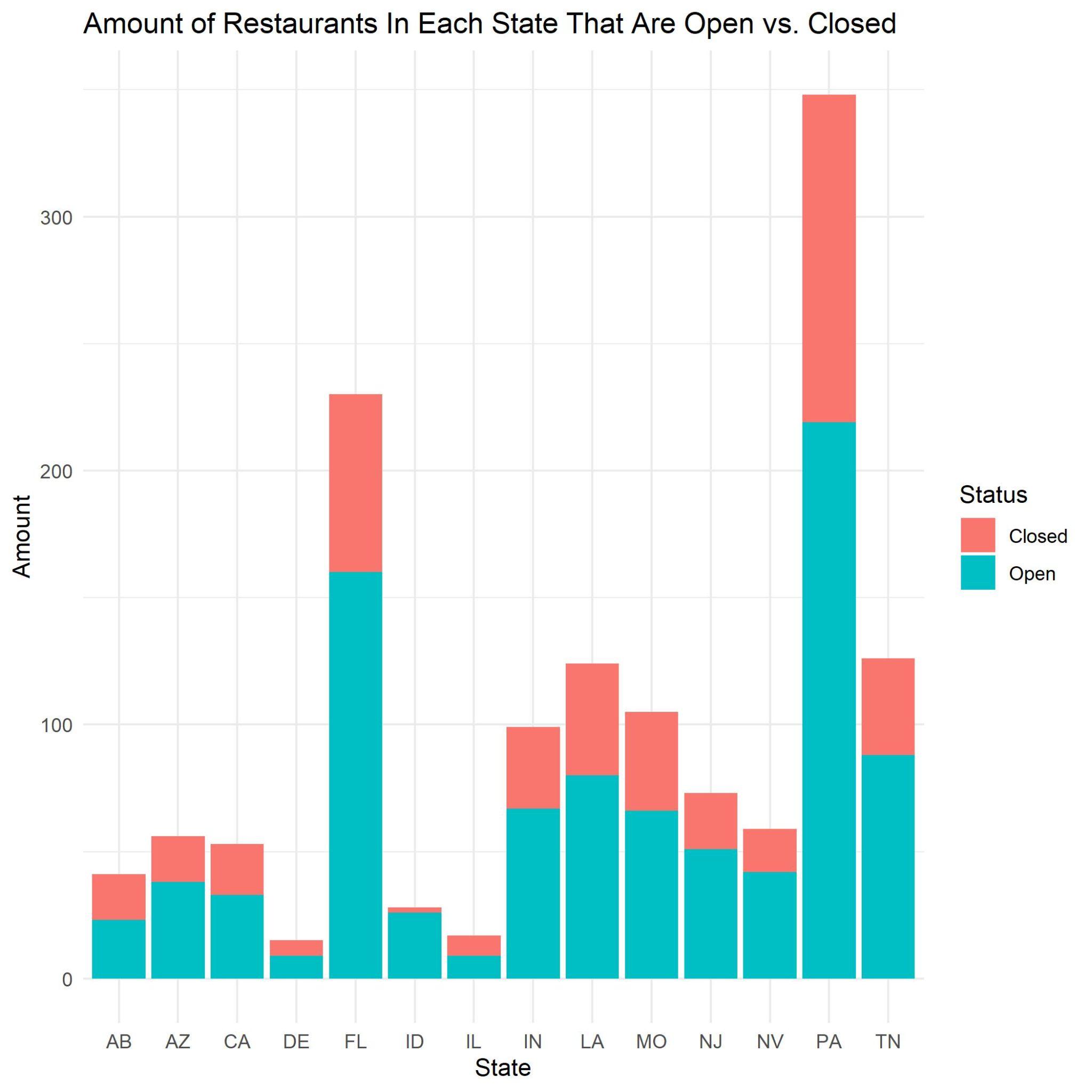
**Figure 3. Open vs. Closed Restaurants**

Most restaurants also only had a few reviews as seen in **Figure 4**. We understand that only one or two reviews isn’t really enough to understand a consumer’s sentiment on a restaurant, and one person’s idea of a restaurant will determine the sentiment score of many of our restaurants. This is why we included the restaurant’s yelp star rating in our independent variables.



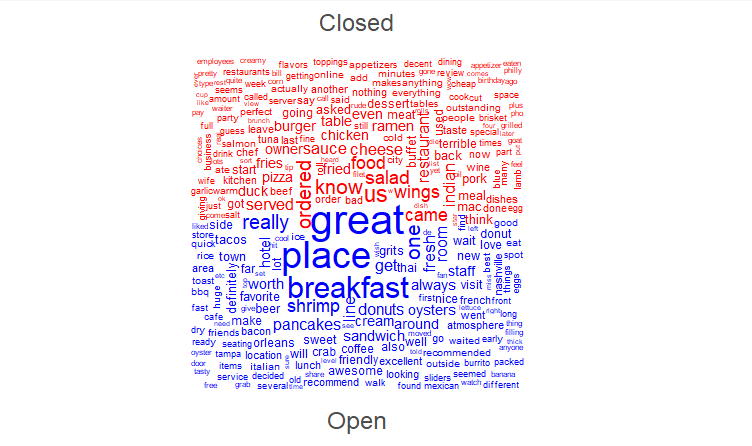
**Figure 4. Amount of Reviews Each Restaurant had in the Dataset**

We were curious about the location of these restaurants since regions can affect business success. It can be harder to keep a restaurant open in areas where rent is expensive, the minimum wage is higher, or there’s a lot of competition. That doesn’t necessarily seem to be reflected in our data as seen in **Figure 5**, so we just have to keep in mind that our location variable might not be the strongest indicator of what’s going on in the real world. If we had a more balanced dataset based on location, we might see a stronger influence from the location variable.



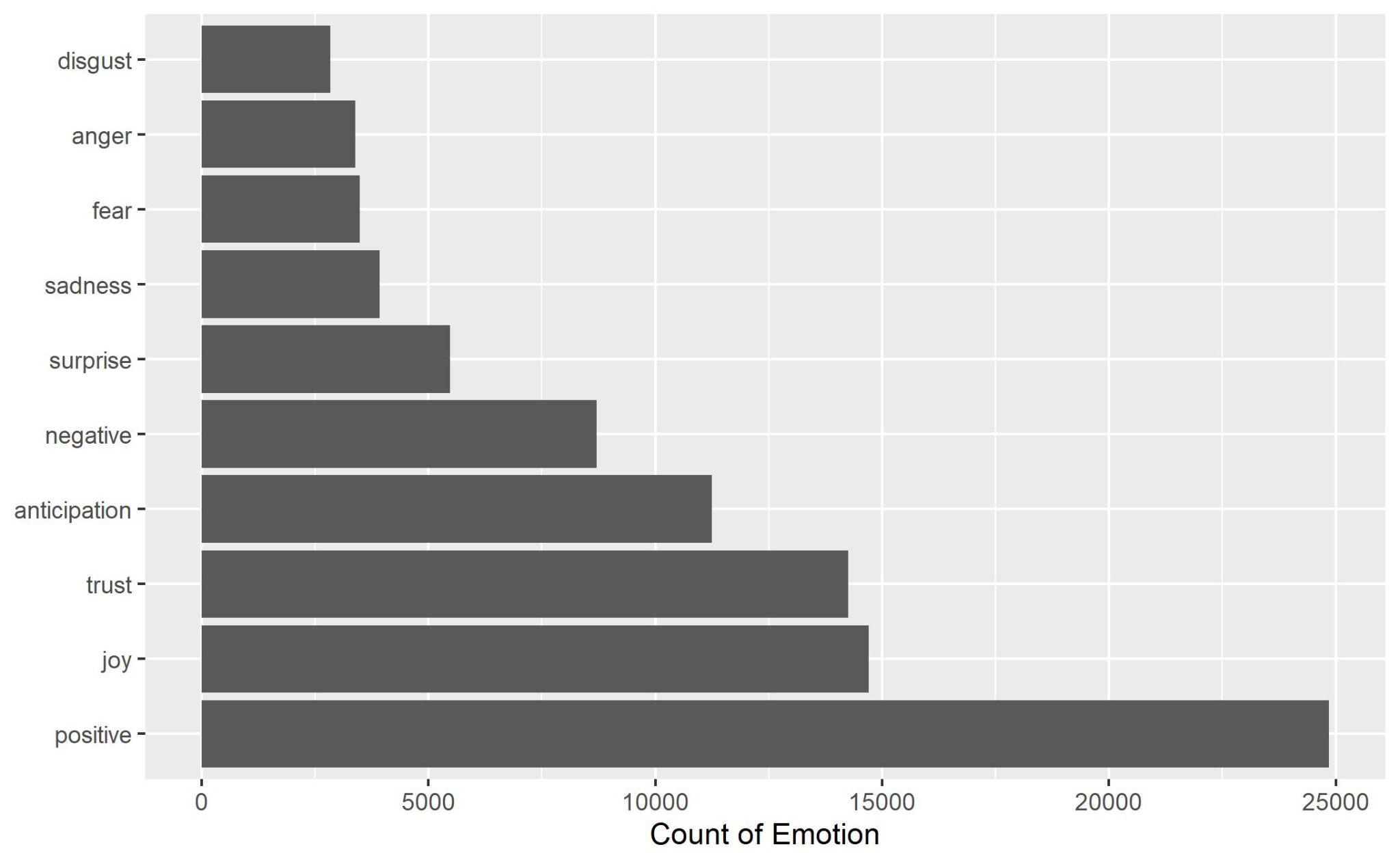
**Figure 5. Restaurants Per State**

To understand our reviews better, we did a word cloud comparing open vs. closed restaurants as seen in Figure 6. The word that stands out is ‘great’ for open restaurants. Other than the word great, there doesn’t seem to be a difference between Open vs. Closed restaurants. It’s just a bunch of food words. This showed us that the closed reviews might not have the strongest vocabulary.



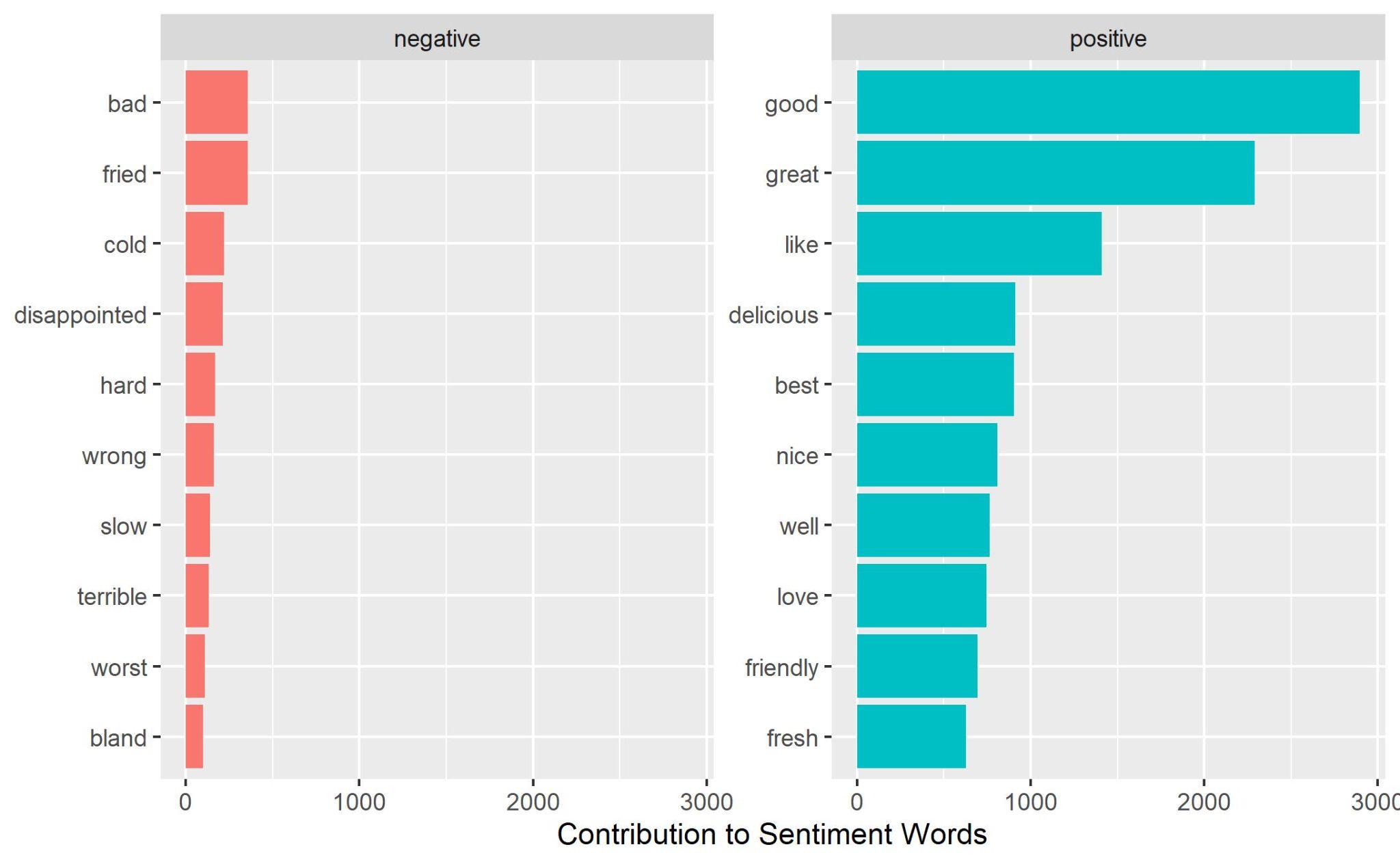
**Figure 6. Word Cloud of Closed vs Open Restaurants**

In Figure 7 we have a significantly higher amount of positive emotions in our dataset than negative. This tells us that our sentiment score will be biased towards positive words.

****

**Figure 7. Count of Emotion**

Out of the negative and positive sentiments, Figure 8 shows the top 10 words for each sentiment. These words align to what we might think negative or positive reviews might contain. Negative reviews will have words like “cold, slow, terrible, and bland” while positive reviews will have words like “delicious, friendly, and fresh”.

****

**Figure 8. Contribution to Sentiment Words**

In the future, we want to gain a larger dataset that has an even distribution of positive and negative words as well as more reviews per restaurant. This means we may have to scrape from Yelp.

**7. CONCLUSION**

**What and where did we learn from the data?**

The Yelp Review data described negative and positive sentiment that was then calculated to show how the score of the sentiment in regards to each categorized word (Samuel J. *et al*., 2020). There was a greater frequency of positive reviews and it that aided open restaurants and distributed more scores in different sentiment categories (Samuel J. *et al*., 2020). The data provided by Yelp indicates that a restaurant with a greater number of reviews, has a higher sentiment score (Redondo, T., & Sandoval, A. M., 2016).

**How did it show it?**

It appeared and was obvious that restaurants with positive reviews stayed open. When the distribution of sentiment scores were created it showed open restaurants had higher scores on average than closed stores. But that does not insist that better reviews corroborate an open restaurant, it is only an observation that says this is possible. The *Feature Importance by Variable* highlights RestaurantsDelivery as the highest rated score in the linear regression analysis. If a restaurant has good food and or if a restaurant was reliable to the consumers then, the restaurant would have good ratings (Kim, S.-H., Huang, R., & Kim, S., 2022). The highest rated feature was restaurant delivery, and the sentiment in that is determined by the customer satisfaction of the restaurant. Service interaction with the restaurant and the patronage is very important because the patronage intention when visiting a restaurant either by delivery or dining in is crucial to the success of the service organization (Kim, S.-H., Huang, R., & Kim, S., 2022). Restaurant deliveries were vital, and if not there was pick up to receive food during the early months of the COVID-19 pandemic, because of restrictions (Ali G. *et al.,* 2021).

**What conclusions can we draw from these articles?**

When restrictions due to COVID-19 began to roll back, people had a strong desire to return to normal activities following health guidelines (Samuel J. *et al*., 2020). The restaurants that performed well with patrons and were able to remain open received a greater number of positive reviews than negative on their Yelp profiles. Restaurants with the highest sentiment scores were three times as likely to be open as seen in **Figure 3.** The sentiment scores showed a preference for service which could imply the customers select food that they like in agreement with good service (Yu B. *et al.,* 2017).

**How does the relate to the culture**

In addition to having great service, restaurant culture was another important factor in determining the performance for customer satisfaction (Nakayama & Wan, 2019). Word-of-mouth is considered credible and trustworthy because it was from an acquaintance, and now this practice has been included in online spaces such as Yelp (Nakayama & Wan, 2019). Word-of-mouth recommendations whether in person or online create product or service oriented social commerce. Customers are more likely to write a good review if the food quality is good, and provide a review on the service to indicate how the restaurant culture is (Nakayama & Wan, 2019).

**8. REFERENCES *–***

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. (2017). A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code Studying Online Reviews and the Stack Exchange Community. Communications of the *Association for Information Systems*, *41*, 450–496. <https://doi.org/10.17705/1cais.04121>

von der Malsburg, T., Poppels, T., & Levy, R. P. (2020). Implicit Gender Bias in Linguistic Descriptions for Expected Events: The Cases of the 2016 United States and 2017 United Kingdom Elections. *Psychological Science*, *31*(2), 115–128. <https://doi.org/10.1177/0956797619890619>

Ali, G. G. Md. N., Rahman, Md. M., Hossain, Md. A., Rahman, Md. S., Paul, K. C., Thill, J.-C., & Samuel, J. (2021). Public Perceptions of COVID-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics. *Healthcare*, *9*(9), 1110. <https://doi.org/10.3390/healthcare9091110>

Samuel, J., Ali, G. G. Md. N., Rahman, Md. M., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, *11*(6), 314. <https://doi.org/10.3390/info11060314>

Samuel, J., Holowczak, R., Benbunan-Fich, R., & Levine, I. (2014, January). Automating

discovery of dominance in synchronous computer-mediated communication. In 2014 47th Hawaii

International Conference on System Sciences (pp. 1804-1812). *IEEE Access*.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. <https://doi.org/10.1177/0261927x09351676>

Samuel, J., Rahman, Md. M., Ali, G. G. Md. N., Samuel, Y., Pelaez, A., Chong, P. H. J., & Yakubov, M. (2020). Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics. *IEEE Access*, 8, 142173–142190. <https://doi.org/10.1109/access.2020.3013933>

Xing, F. Z., Cambria, E., & Welsch, R. E. (2017). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49–73. <https://doi.org/10.1007/s10462-017-9588-9>

Redondo, T., & Sandoval, A. M. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*(Special Issue on Big Data and AI). <https://www.ijimai.org/journal/bibcite/reference/2540>

Lang, S. (2020). Writing Analytics: Expanding the Body of Knowledge. The *Journal of Writing Analytics*, *4*(1), i–iv. <https://doi.org/10.37514/jwa-j.2020.4.1.01>

Samuel, J. (2020, July 1). *Conceptual Frameworks for Big-Data Visualization: Discussion on Models, Methods and Artificial Intelligence for Graphical Representations of Data - Call for Extension*. Papers.ssrn.com. <https://ssrn.com/abstract=3640058>

Drieger, P. (2013). Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences*, *79*, 4–17. <https://doi.org/10.1016/j.sbspro.2013.05.053>

Sundaram, A. (2020, December 11). *Yelp data shows 60% of business closures due to the coronavirus pandemic are now permanent*. CNBC. Retrieved May 2, 2022, from <https://www.cnbc.com/2020/09/16/yelp-data-shows-60percent-of-business-closures-due-to-the-coronavirus-pandemic-are-now-permanent.html>

Vesoulis, A. (2021, December 28). *Most restaurants fail. covid-19 made the odds even worse*. Time. Retrieved May 2, 2022, from <https://time.com/6129713/restaurants-closing-covid-19/>

Published by S. Lock, S., & 7, J. (2022, January 7). *Number of restaurants in the U.S.* Statista. Retrieved May 6, 2022, from https://www.statista.com/statistics/244616/number-of-qsr-fsr-chain-independent-restaurants-in-the-us/#:~:text=How%20many%20restaurants%20are%20in,a%20little%20over%20two%20percent.

1. <https://github.com/Jessicacruzn/PI-Studio-Final-Project> [↑](#footnote-ref-1)