**1. Introduction**

Nashville, Tennessee is known for country music, tourism, and its nightlife scene. This project focuses on analyzing businesses in Nashville that have a category label of either “bar” or “nightlife”. We define a successful business based on the star rating. We aim to focus on answering the following questions: what business attributes affect success, what words found in a review are associated with success, and what demographic features in an area affect success. We discovered that happy hours are not important, location is important, what matters to the customer, population size is negatively related to average stars of businesses, and cocktails are the best drink to serve.

**2. Data Sources and Data Cleaning**

We used three main sources of data for our analysis. The first being from Yelp which contained information on businesses as well as reviews given to those businesses. First we filtered the business data to only include businesses located in Nashville and that had “bar” or “nightlife” listed in their category. To merge the business data file with the review data file, we connected them by using the unique identifier of business\_id. We used data from the US Department of Transportation to study how trips affect the reviews. Again we needed to filter this data to only contain trips that were in Nashville. Lastly we used data from the US Census Bureau to answer demographic questions. For this data we only selected data where the postal code matched with the postal codes in Nashville. We chose elements including population, median income, educational level, employment rate, number of households and number of businesses in every postal code area.

We completed several tasks to clean our dataset. One of the data cleaning aspects involved converting certain fields to factors for analysis. For example converting postal codes to factors to do analysis based on location. For the US Census Bureau data, we averaged stars of businesses sharing the same postal code and added average stars to census data to find out how the demographic factors influence stars rating. Objects with missing values were then removed. For the sentiment analysis we first removed stopwords (e.g. “the”), special symbols, and punctuation. For part of the sentiment analysis we filtered reviews who contained values for the four aspects of: food, drink, service, and ambience.

**3. Exploratory Data Analysis**

Our exploratory data analysis focused on five main areas: text analysis on reviews, business attributes, demographic factors, location, and trips.

**3.1 Text Analysis on Reviews**

For the text analysis on reviews we focused on finding common words used to see what effect they might have on the star rating of a business. We discovered that if a review contained the word “bouncer” it was likely to be a one star review whereas if a review contained the word “cocktails'' it was likely to be a 4 or 5 star review. This led us to wanting to explore sentiment analysis of the reviews as our next steps.

**3.2 Business Attributes**

Next we looked to see if some of the business attributes have an effect on the star rating. We discovered that on average if a business had a no outdoor smoking policy that it tended to have a higher review than the businesses that did have a outdoor smoking policy. We also noticed that the price range did not significantly have an impact on the average ratings of stars. For example the cheapest price range, indicated by a 1 on Yelp, had an average star rating of about 3.8 whereas the highest price range, indicated by a 4 on Yelp, had an average star rating of about 3.7. We also found free Wifi has no contribution to positive reviews.

**3.3 Demographic Factors**

In our initial exploratory data analysis we created plots using various variables such as median household income from the US Census Bureau. From the plots we noticed that the average stars are affected by some factors. To analyze this further a regression model was built which discovered that the most significant factor from the US Census Bureau data is population. Population size is negatively related to average stars of businesses in one area. From the regression model, the p-value of the coefficient is 0.002, which is assumed significant. For every 6,250 additional people in the population, a region's average stars will decrease by 0.1 points. Thus we suggest that business owners establish their businesses in lower populated regions in order to be more successful.

**3.4 Location**

We wanted to see if the postal code of businesses influenced the star rating of a business. It was found that the postal codes with the highest average star ratings were 37015, 37243, 37238, 37220, 37135, and 37115.

**3.5 Trips**

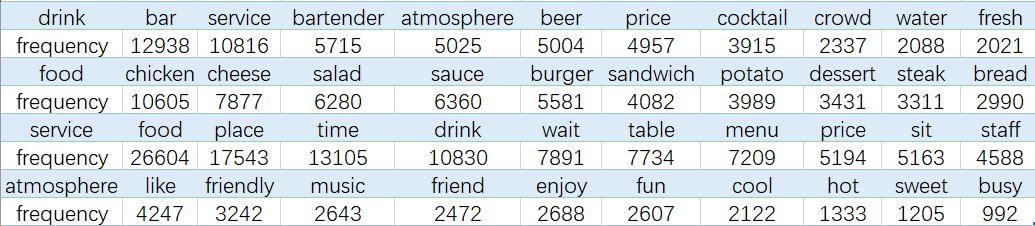
One of our most interesting discoveries was that it seemed like the number of reviews and number of trips taken in 2019 had a relationship as shown below. We can see a decrease from January to February and then an increase after February in both. We also see a sharp decline after September in both.

|  |  |
| --- | --- |

**4. Key Findings**

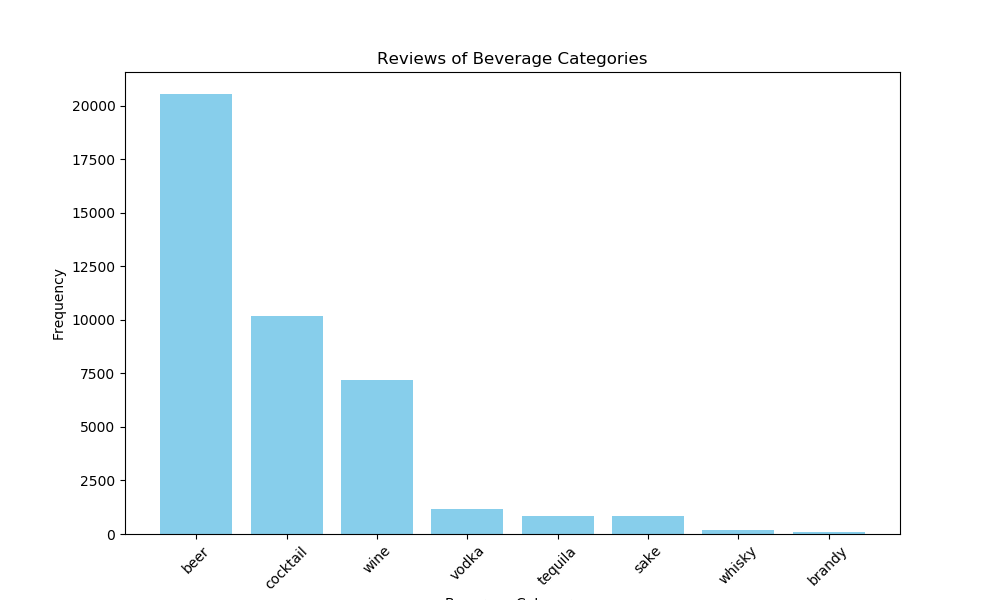
**4.1 Sentiment Analysis on Reviews**

One of our main focuses of analysis was sentiment analysis on the reviews. Firstly we got a sentiment score(compound score) for every review and noticed the sentiment score increases as the star rating increases. We assumed reviews having stars 1 and 2 are negative and reviews having stars 4 and 5 are positive. We used the TF-IDF model to extract keywords from processed reviews. Words with high TF-IDF value can be divided into 4 aspects, that is service, food, drink and atmosphere. Then we separately calculated the most frequent words occurring together with service, food, drink and atmosphere. The frequencies of words are shown in the table below.



We can see when talking about drink, the most frequent word is bar, which is consistent with our common sense. The most popular category of food is chicken. When talking about service, food, place, time and drink are mostly mentioned. Customers tend to write reviews for a business when the atmosphere is friendly.

As for categories of beverage, the most popular type of alcohol is beer, followed by cocktails and wine as shown below. Their frequency of mention accounts for 92.4% of the frequency of mentions of all alcohol names.



The average sentiment scores of alcohol shows customers prefer cocktails and wine than other categories as shown below. We would recommend to business owners to serve cocktails and wine. Sake and vodka received the lowest sentiment scores.



As to the service and atmosphere of businesses, we picked up some keywords and calculated their contained-reviews’ frequencies. We found people value friends when entering the bar or nightlife business and are sensitive to wait. More than one in five of reviews mentioned “friend/friends” (23.7%) and mentioned “wait” (22.6%). The next mentioned words in regards to the atmosphere of a business were “menu”, “music”, “bartender”, and “dance”. Dance was the least popular of these indicating less customers care about dancing.

For further sentiment analysis, we performed an Aspect-based Sentiment Analysis. A function is built to filter the data based on the following aspects: service, ambiance, food, drinks, bar, vibe, music, crowd, employees, and price. These aspects are chosen based on the most frequently occurring words in the reviews. The text is tokenized and we compute the sentiment scores for each aspect using the AFINN lexicon. For each business, we aggregate all the scores and make them consistent using min-max normalization. We also incorporate postal codes to compare the performance of various locations based on the chosen aspects. To analyze the significance of these aspects on the star ratings, we apply a multiple linear regression model. The factors that significantly impact the star ratings are:

1. **Service:** A one-unit increase in the sentiment score of 'service' is associated with an estimated **increase** of approximately **0.81** units in star ratings (p-value < 0.001).
2. **Time:** A one-unit increase in the sentiment score of 'time' is associated with an estimated **increase** of approximately **1.61** units in star ratings (p-value < 0.001).
3. **Food:** A one-unit increase in the sentiment score of 'food' is associated with an estimated **increase** of approximately **0.51** units in star ratings (p-value < 0.001).

The other factors do not demonstrate a significant impact on the star rating based on the p-values. This model explains about 47.31% variability in the star ratings. This model could be improved in the future but this agrees with our analysis done earlier that we suggest business owners to focus more on their service, wait times and food to improve their star ratings.

**4.2 Business Attributes Analysis**

We suggest that it is not worth it to have a happy hour. The star difference between those that have a happy hour vs those that do not have a happy hour is less than 0.5 stars. To come to this conclusion we performed a Welch two sample t-test. The results were a p-value of 0.007 which indicates that there is a significant difference between the star rating of the business who have a happy hour and those who do not have a happy hour. However the star difference is not that large as we are 95% confident that the actual difference in average star rating is between 0.093 and 0.342. These results are interesting because one would expect that having a happy hour would increase star ratings significantly.

To expand on our exploratory data analysis of location, we conducted an ANOVA test using the postal codes. The p-value was very small, less than 0.001, indicating that there are significant differences in average star ratings across the different postal codes. Next we conducted a Tukey HSD test to see which postal codes pairs had the largest difference. We found that on average the star ratings for postal code 37221 are lower by approximately 0.776 stars than those in postal code 37216. We consider this difference statistically significant as the adjusted p-value is less than 0.05. Other postal code pairings did not have a low enough adjusted p-value so we are unable to comment on how other locations impact star rating. Thus if business owners are deciding to open a business in either 37221 or 37216, we recommend that they open it in 37216.

**5. Conclusion**

We advise the business owners to consider location, types of drinks, food, and service times to increase their rating on Yelp. This project does have some limitations. One is that some reviews could be sarcastic which could affect our sentiment analysis. Another is that for our ANOVA test we had to make certain assumptions about the data. We also made assumptions on what a positive or negative review, which can be subjective. We could also do a more in-depth analysis to see how other aspects would impact the star rating of a business. Overall this project provides Nashville bar and nightlife business owners some information to improve their Yelp rating.

**6. Contributions**

| **Contributions** | **Vaishnavi Borwankar** | **Abigail Sikora** | **Shuangyu Wang** |
| --- | --- | --- | --- |
| **Presentation 1** | Responsible for slides 8-10.  Reviewed all slides. | Responsible for slides 1-4. Reviewed all slides. | Responsible for slides 5-7.  Reviewed all slides. |
| **Presentation 2** | Responsible for slides 9-10.  Reviewed all slides. | Responsible for slides 1-4, 11. Reviewed all slides. | Responsible for slides 5-8. Reviewed all slides. |
| **Executive Summary** | Responsible for the second half of sentiment analysis, Trips, and data cleaning. Reviewed everything. | Responsible for intro, data sources, business attributes, business attributes analysis, and conclusion. Reviewed everything. | Responsible for the first half of sentiment analysis, Demographic factors, and data cleaning. Reviewed everything. |
| **Shiny App** | Co-Responsible for Shiny App | Co-Responsible for Shiny App | Reviewed/provided feedback on Shiny App. |