DS-GA 1001 INTRODUCTION TO DATA SCIENCE

# Data Analysis of Restaurant Ratings

Team S.H.A.R.P

| Siri Desiraju | Hoa Duong | Anna Dominic | Ridhika Agrawal | Palak Bansal |
|---|---|---|---|---|
| N14355365 | N11455685 | N18538396 | N13428967 | N11465209 |

## 1. Introduction

As crowd-sourced review and recommendation platforms continue to expand, online reviews have become a digital version of word-of-mouth marketing and evolved into a critical form of information. Many individuals are increasingly basing their decisions on consumer-generated reviews and ratings. In fact, for instant, the popularity of a restaurant is positively associated with ratings about its quality, its environment and services provided, and its number of online reviews[1]. On the other hand, online reviews help business owners identify their strengths and weaknesses, understand customers' needs and potential areas of improvement, and create a healthy competition in the industry[2]. As such, effective analysis of reviews data becomes of paramount importance.

Our project utilizes the datasets on businesses, reviews, and users data provided by Yelp[3]. The business dataset contains information on location, business rating, number of reviews, category, whether the business is operating, operating hours, and additional attributes specific to each of the 150,346 businesses. The reviews dataset contains the rating, review text, date, and the number of other users who find the review useful, funny, or cool of close to 7 million reviews. These datasets will allow us to dive deeper into the relationship between reviews and business popularity as well as their operating status from one of the largest online review and recommendation venues. Motivated by the fact that many small businesses had to close down during COVID, we seek to use customers' reviews and various business features to understand and predict businesses' popularity, quality, and status of operation. This information will provide insights to both the customers and the business owners as they navigate through the pandemic recovery.

[1] https://doi.org/10.1016/j.ijhm.2010.02.002.

[2] https://www.e-satisfaction.com/7-reasons-why-customer-reviews-are-important/

[3] https://www.yelp.com/dataset

## 2. Data Preprocessing

Our datasets do not have missing data, with all columns having data for all observations. Our ratings data in both the business dataset and the reviews dataset are in the range of 1-5. Through exploratory data analysis, we found that most businesses have an average rating of 3.5 to 4.5 stars, as shown in Figure 1. The ratings distribution is slightly skewed, which makes sense intuitively, as people are more prone to write positive reviews and leave higher ratings[4].

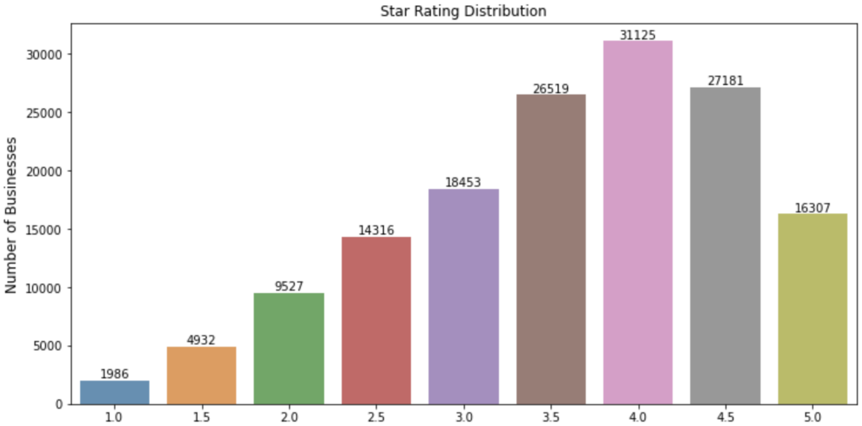


Figure 1. Distribution of Business Rating

Taking a closer look into the locations of the businesses in our dataset, we found that they are located around the world. However, a majority of the businesses are located in the United States (144,773 businesses out of 150,346 businesses). The map in Figure 2 shows the heatmap of the businesses within the United States.

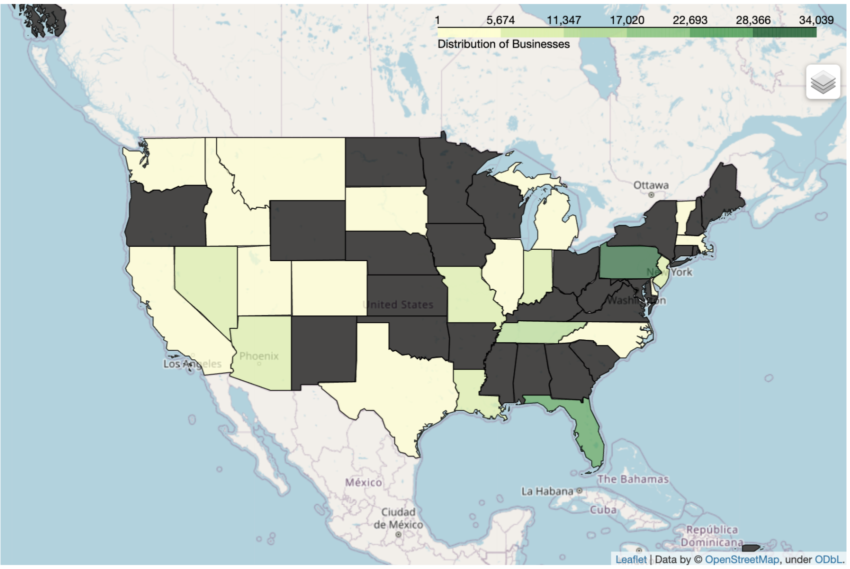


Figure 2. Heatmap of Businesses in the United States

[4] https://doi.org/10.1016/j.bushor.2011.11.001.

In order to extract meaningful insights from reviews and business features, we first need to identify the important features. As aforementioned, our business dataset includes information about various features specific to each business. Thus, we want to focus on one business category and find the common features among businesses in the same industry. As shown in Figure 3, the two most common categories in the business dataset are restaurants and food. Thus, we decided to focus on these two categories, so as to extract meaningful features. As some businesses are in both the restaurants category and food category, restricting the business dataset to these two categories leaves us with 64,616 rows.

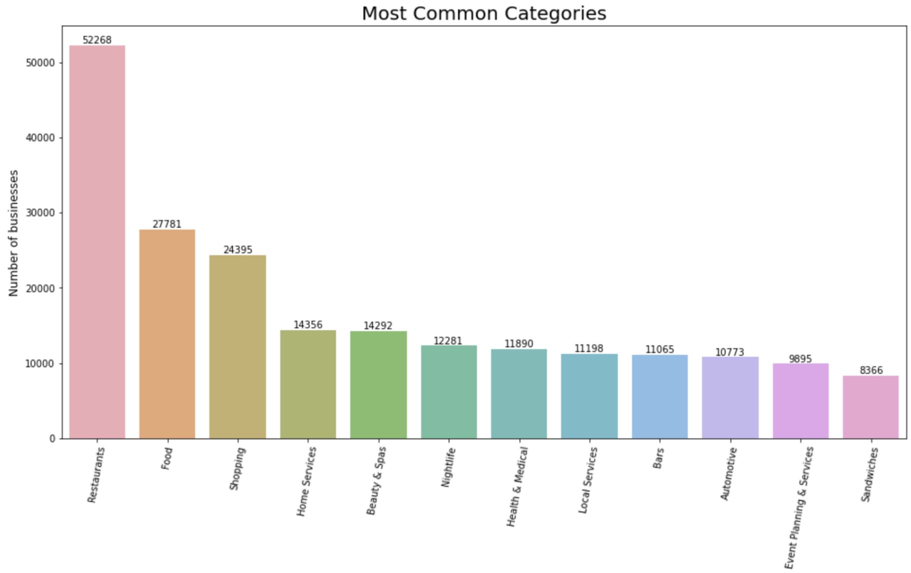


Figure 3. Number of Businesses for Most Common Categories

We also wanted to capture the restaurant opening and closing timings as features in the model since a business' hours might affect business popularity. To that end, we added a dummy for whether a restaurant is open on weekends or not and another variable which calculates the total hours a restaurant is open in a week using open and close times.

We then proceeded to extract features from the additional features column. We found that some of the most common features for these restaurants are take-out, parking, credit-card acceptance, price range, and delivery, as shown in Figure 4, which all intuitively will affect the restaurant's popularity and operating status. Each feature was converted into a column with a binary option, 1 if feature present, 0 elsewise, thus eliminating the presence of missing values.

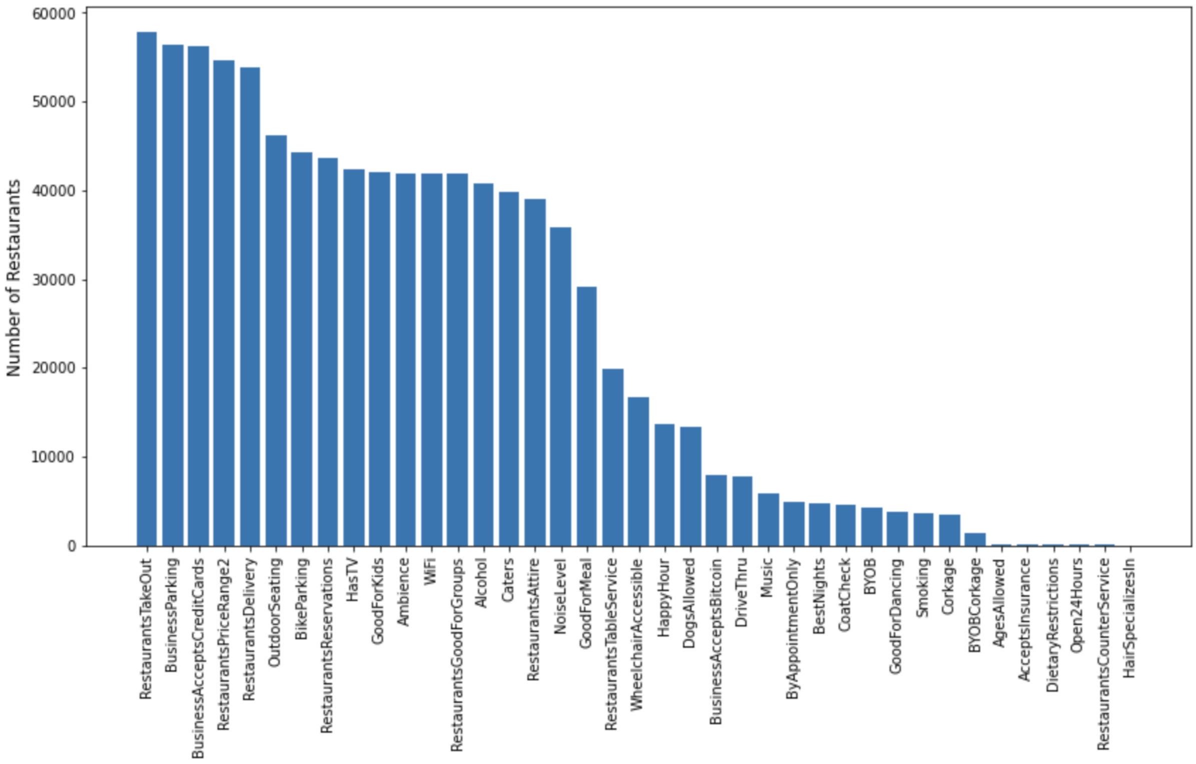


Figure 4. Count of Additional Features for Restaurants

In addition to using business-specific features, we considered the information from customers reviews for each restaurant. Firstly, since ratings are slightly skewed, using average ratings might be misleading. As such, we included median rating obtained from the reviews dataset as an additional column to the business dataset. Furthermore, we gathered the total number of users noting a specific review for a restaurant as useful, funny, or cool and included these counts as additional features for the restaurants. This is because intuitively, a restaurant with more reviews that are rated as useful, funny, or cool will more likely be popular and continue to stay open. Lastly, we noticed some users who posted explicitly negative review texts yet provided a higher-than-average rating. Thus, we performed sentiment analysis on the Reviews Dataset using the Vader algorithm. Doing so, enabled us to classify each review as either positive or negative and we proceeded to add this information as a feature in our analysis.

We hypothesized that demographic information about the neighborhood the restaurant is in might affect its ratings as well as its closure. To capture this, we used US Census Tract[5] data and added zip-code level information on total population in neighborhood, median age, percentage of white population, and sex ratio (number of women per 100 men).

[5] https://data.census.gov/

# 3. Inference Questions

We aim to explore if there is a meaningful relationship between the restaurants' features in predicting the restaurants' popularity (proxied as their ratings) and their operating status (whether they are closed or open). The dataset contains over 25 various features for each restaurant. However, since details of features are not provided for every restaurant, for the purposes of this analysis, we decided to take the 11 most occurring features (refer to Figure 4) and implement the Mann-Whitney U Test hypothesis test.

We separated the restaurants into 2 groups, open and closed. Using a bar chart, we visualized the count of features within the two groups, with some examples shown in Figure 5. We can see that in most cases, among open restaurants, there are more restaurants with the feature, while among closed restaurants, there are more restaurants without the feature. This suggests that the existence of the feature has an implication on the operating status of the restaurant.

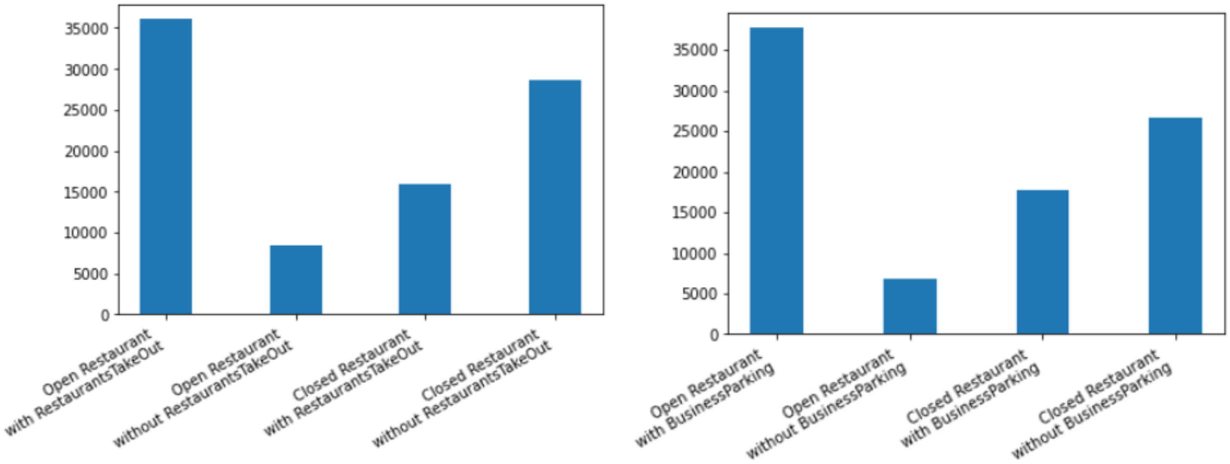


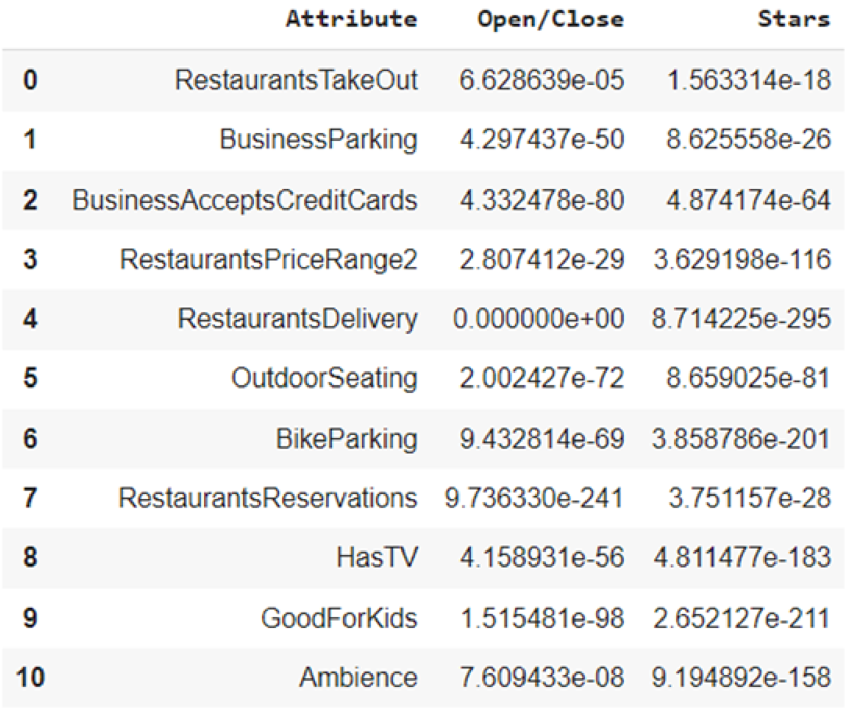
Figure 5: Number of Open and Close Restaurants

We test the following hypotheses:

$H_0$: There is no relationship between an attribute and a restaurant's closure or its rating

$H_a$: There is a relationship between the attribute and the restaurant's closure or its rating

The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed. We found that all of the 11 features selected were statistically significant, with p-values less than 0.005, as shown in Figure 6. This indicates that these features are important in predicting the operational status of the restaurants and the popularity of the restaurants. In summary, these 11 features should be prioritized.



| | Attribute | Open/Close | Stars |
|---|---|---|---|
| 0 | RestaurantsTakeOut | 6.628639e-05 | 1.563314e-18 |
| 1 | BusinessParking | 4.297437e-50 | 8.625558e-26 |
| 2 | BusinessAcceptsCreditCards | 4.332478e-80 | 4.874174e-64 |
| 3 | RestaurantsPriceRange2 | 2.807412e-29 | 3.629198e-116 |
| 4 | RestaurantsDelivery | 0.000000e+00 | 8.714225e-295 |
| 5 | OutdoorSeating | 2.002427e-72 | 8.659025e-81 |
| 6 | BikeParking | 9.432814e-69 | 3.858786e-201 |
| 7 | RestaurantsReservations | 9.736330e-241 | 3.751157e-28 |
| 8 | HasTV | 4.158931e-56 | 4.811477e-183 |
| 9 | GoodForKids | 1.515481e-98 | 2.652127e-211 |
| 10 | Ambience | 7.609433e-08 | 9.194892e-158 |

Figure 6: Hypothesis Test Results

# 4. Prediction Questions

As shown above, all of our additional features provide meaningful information about the restaurants' popularity. Here, we predict each restaurant's rating using its location (as given by its longitude/latitude pair), zip-code level information on race, age, sex and total population, number of reviews, some of the common restaurant features, open hours, open on weekends, and the Vader results.

The Vader results were used to create an additional feature that gave the percentage of positive reviews that each business received. This additional column significantly improved our results. We used a Neural Network here to perform the Regression and predict the median rating that each business received. The Network has three layers with 9, 5 and 1 neurons respectively. We achieved a score of 54% on the training data, and 52% on the testing data. If we were predicting the median rating for each business, we would have a 20% chance of guessing the right rating, so we posit that the model still performs better than random guessing.

# 5. Classification Questions

We are also interested in the likelihood that the restaurants might be closed down. We hypothesize that restaurants' popularity might be related to its closure. We used a few of the same features that were used in the Regression question, in addition to the feature - number of users who thought reviews were funny, useful or cool.

We first performed k-means clustering on the restaurants from the dataset. First we scaled the data by applying min-max scaling to bring all features to the same scale for efficient clustering. We applied the elbow method to identify the number of clusters for good clustering as 30 from Figure 7 (as 3 is an elbow point).

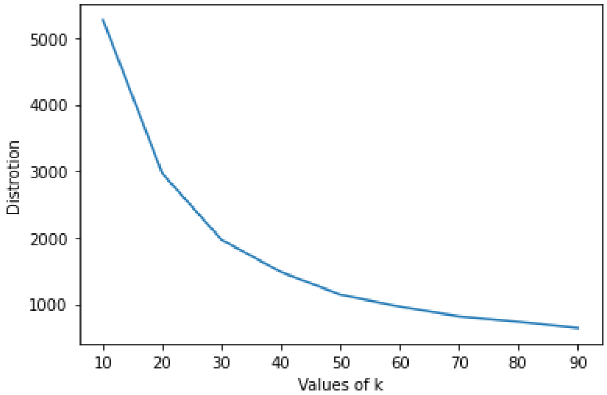


Figure 7: Elbow Method Showing Elbow Point

The three features that give us the best scores for k-means are review count, median rating, and stars. The visualizations can be found in Figure 8. The inertia of the k-means model, i.e., 'Sum of squared distances of samples to their closest cluster center', is 0.58. Clustering on median rating, average rating, and stars gives a similar score.

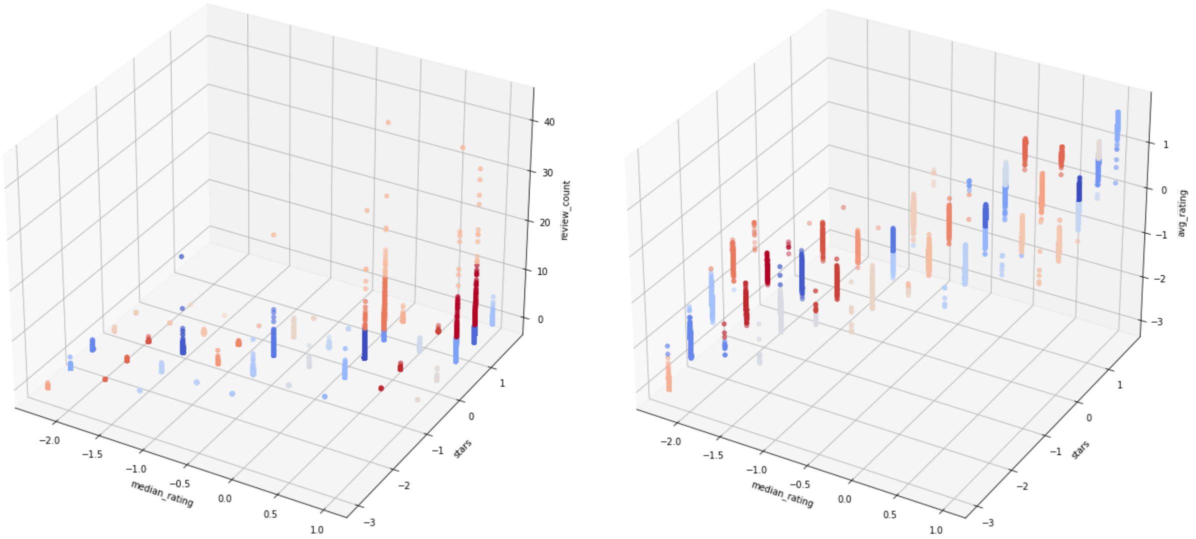


Figure 8: k-means Clustering using Median Rating, Stars, and Review Count (left) and k-means Clustering using Median Rating, Stars, and Average Rating (right)

The classification itself, to predict whether or not a restaurant would get closed down or not, was performed using a Neural Network, with a total of 3 layers. The input layer has 10 neurons, the hidden layer has 10 neurons as well, and the output layer has a single neuron. The activation functions used were the tanh, relu and sigmoid functions respectively. We achieved an accuracy score of 74% on the training data, and 73% on the testing data.

We also used a Logistic Regression Model and achieved similar accuracy scores. Plotting the ROC-AUC scores gave us the following results as seen in Figure 9.

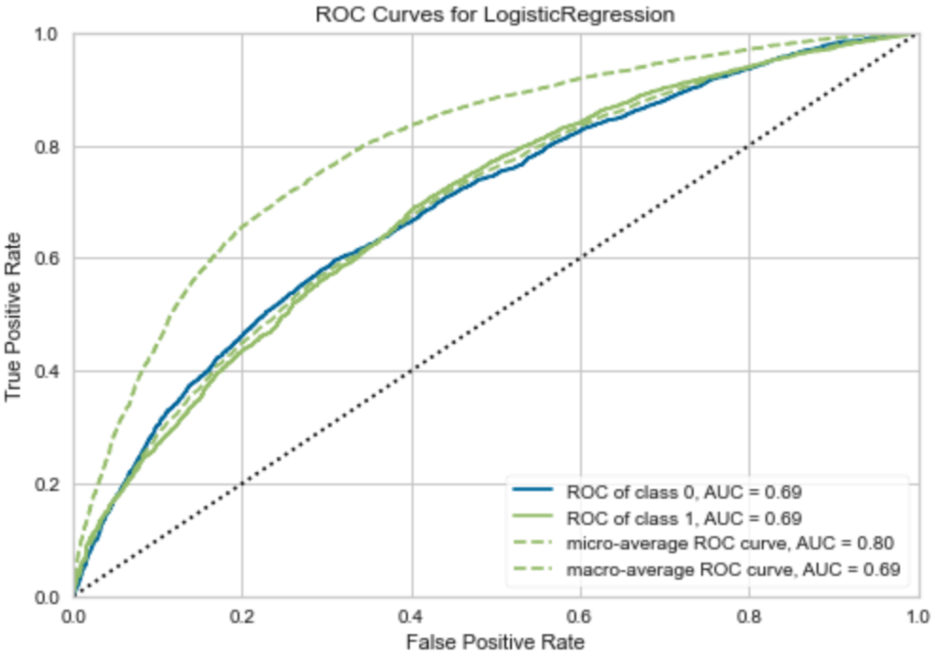


Figure 9: Logistic Regression Results

## 6. Summary and Conclusions

From our analysis, we can see that varied features impact a restaurant's rating and its closure. The hypothesis tests revealed that there is a significant difference between attributes (e.g. takeout, price range) for closed vs. open restaurants and for restaurants with different ratings. At the same time, there were still several features that surprisingly did not have a huge impact as revealed by our predictive analysis. Restaurant Delivery and Restaurant Takeout were two features that seemed to be good predictor variables, but including the Price Range feature did not change or improve our results in the Prediction question. Features like the number of reviews received by each business and the percentage of positive reviews also had an impact on its popularity and closure.

Further, we found that it was easier to predict the restaurant closure than to predict the ratings. This could be due to the fact that ratings are more subjective and not consistent among different users, i.e., different individuals could rate the same restaurant differently even with similar experiences since they have different standards for a 5 star vs. 4 star. As opposed to closure, where it might be more consistently determined by factors that influence e.g. revenue.

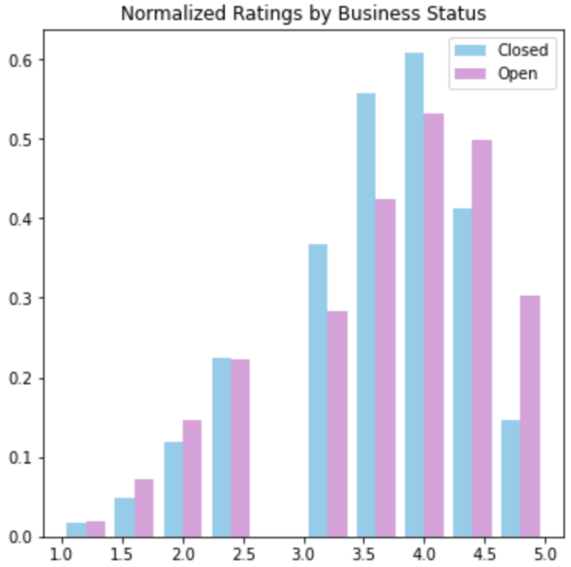
Thus, we think our model would be most useful for business owners who want to understand what they could do differently to prevent closure. This insight is especially valuable for small business owners. They can learn about what factors influence closure of their restaurant, and make resource allocation decisions based on that.

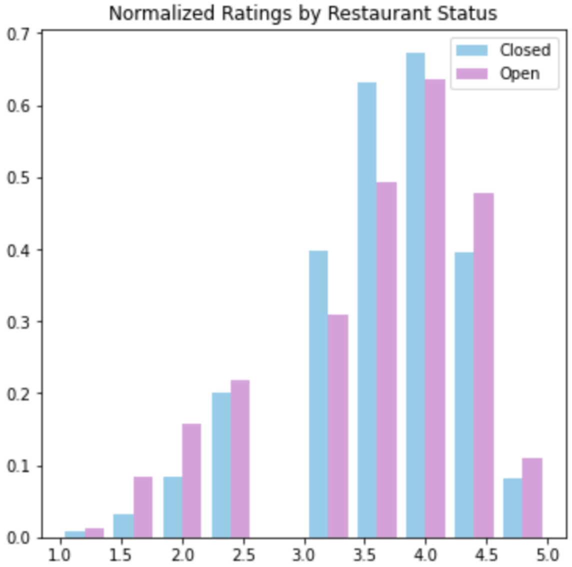
We have several limitations in our models. Our dataset was imbalanced with 31% of the restaurants classified as closed. Thus, for the classification question, we could have oversampled the minority class so that we get a more balanced dataset. Additionally, the restaurant attributes were not consistent across all different restaurants because of which we were not able to use potentially impactful features in our model. For the features that were most common, we imposed a binary input with 1 being present and 0 otherwise. However, there is a difference between knowing that the feature was not present versus not having any information about the presence of the feature.

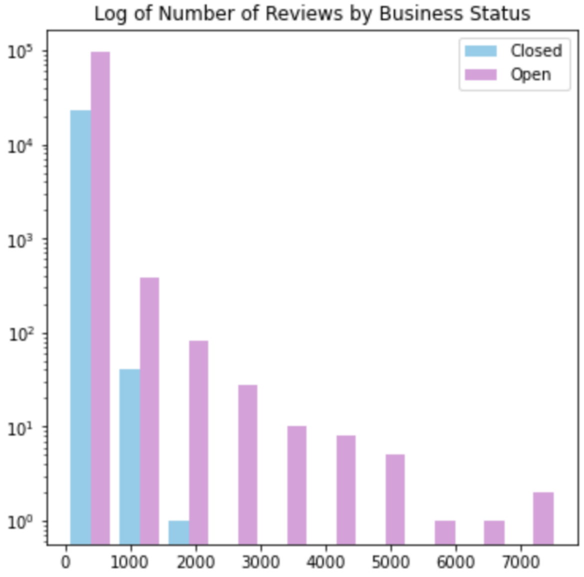
We would like to further extend our analysis to differentiate between different types of restaurants, e.g. bars, cafes, breakfast places. If we had the aforementioned information, we might be able to predict ratings better since there might be more consistency within restaurant types. Our predictive model could then be used by specific restaurants to know how they are performing compared to other restaurants of the same type to increase revenue and footsteps.

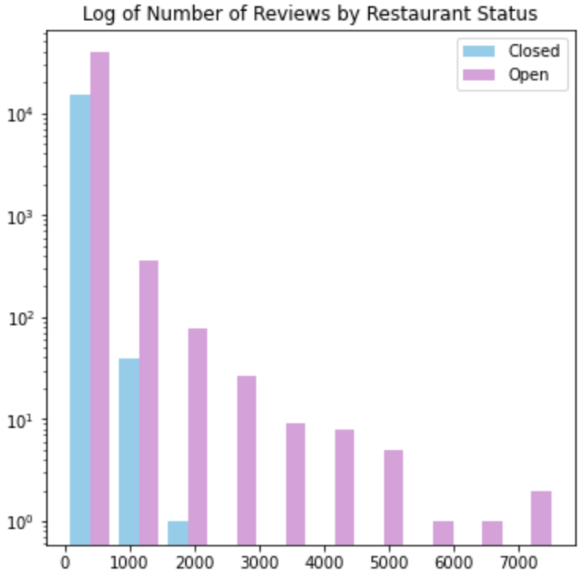
# 7. Appendix

### Exploratory Data Analysis:



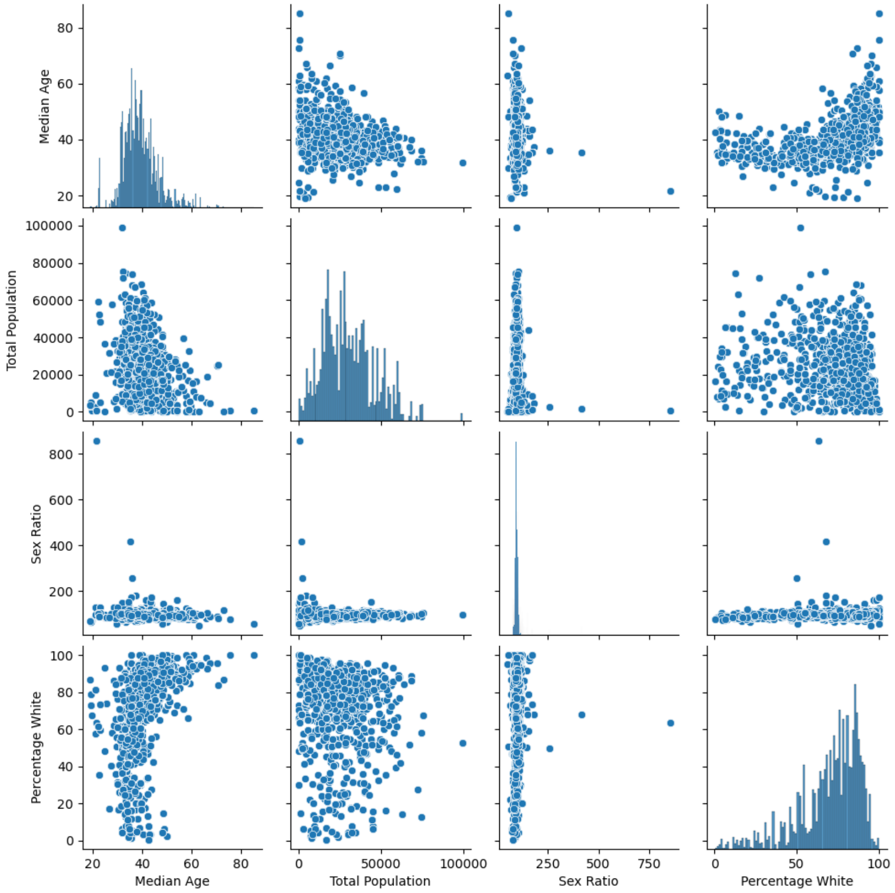






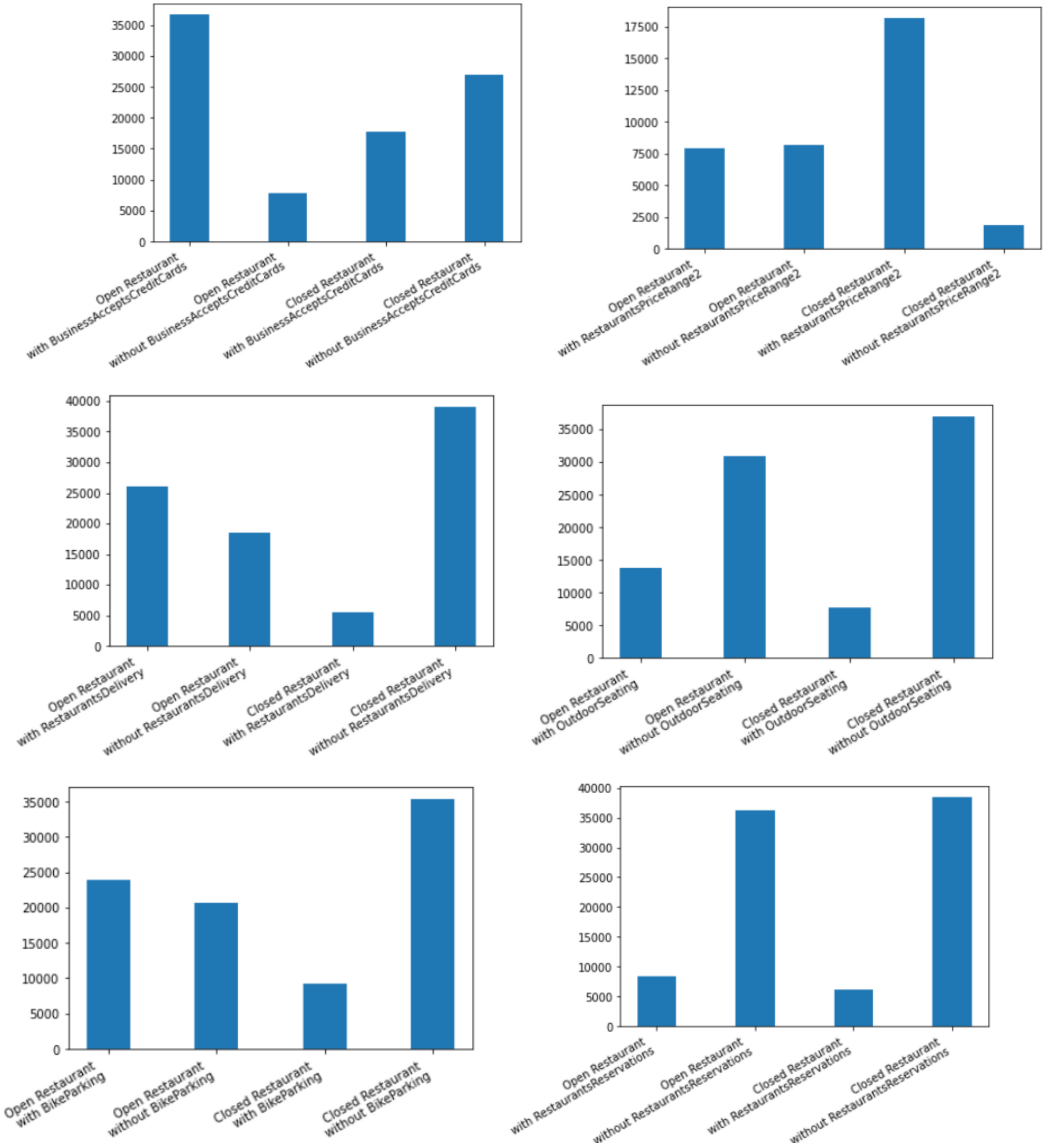
## Data Preprocessing:

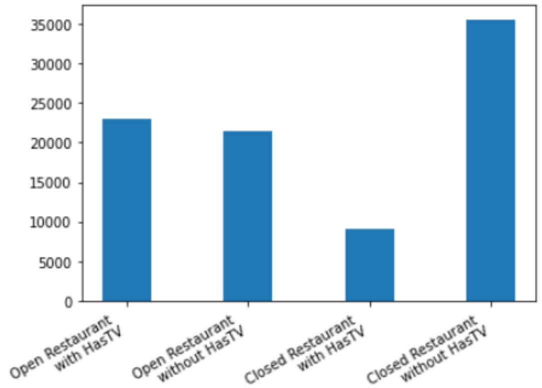
Distribution of the neighborhood demographic information:

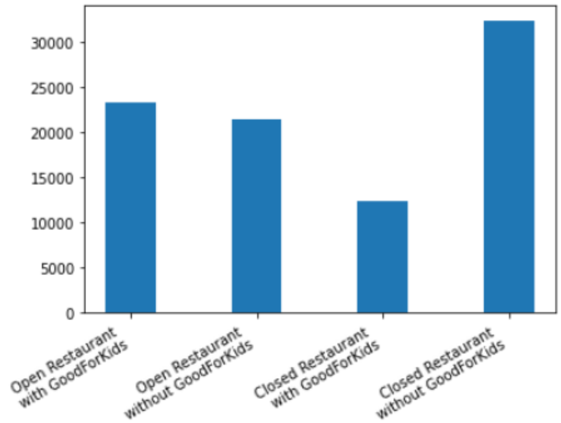


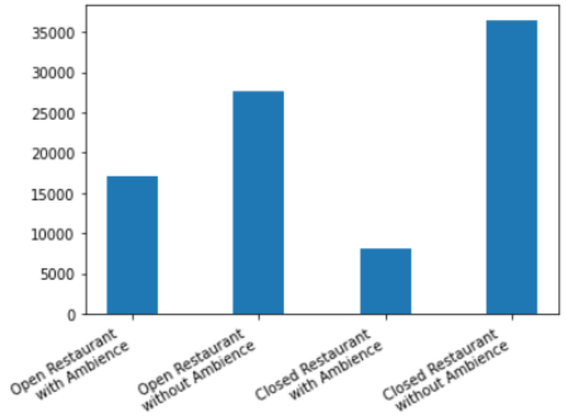
**Inference Question**:

Graphs counting the opening and closing of the restaurant as compared to the features :

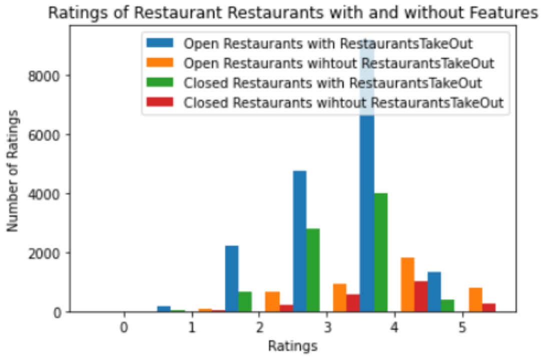


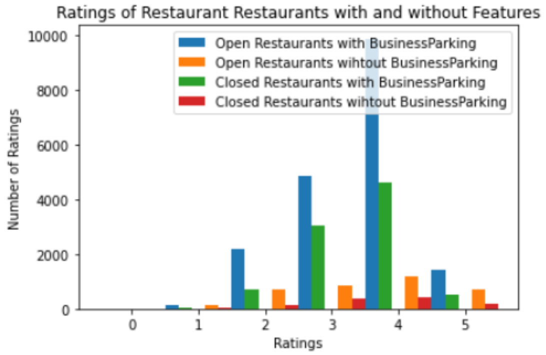


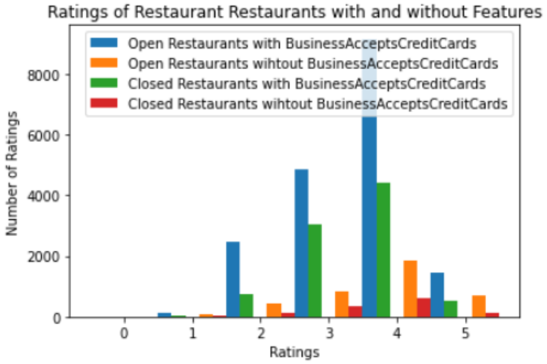


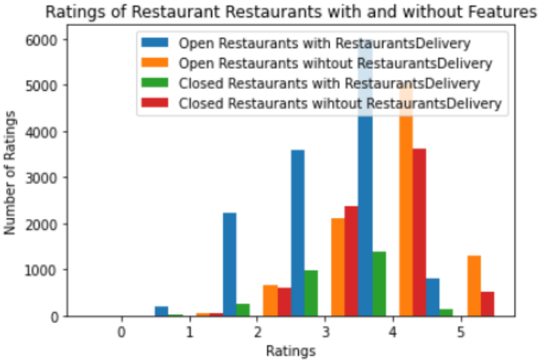


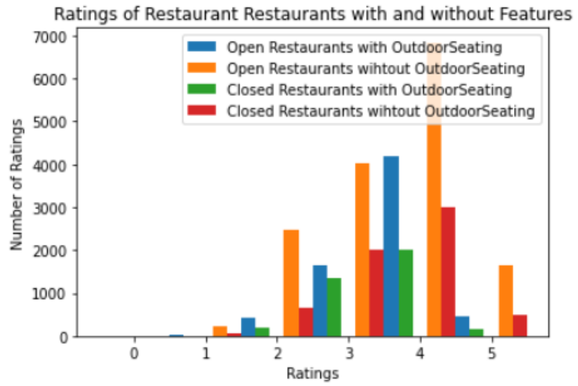
Graphs plotting ratings against the open and closed restaurants with and without features :

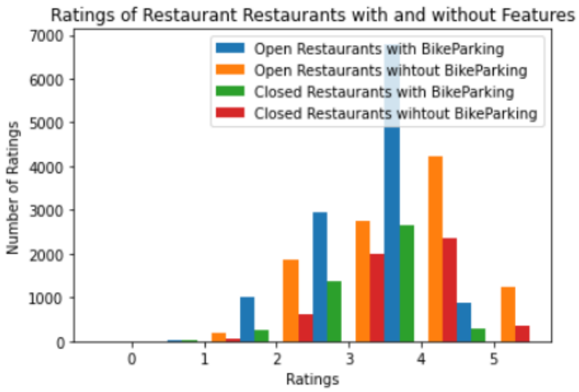


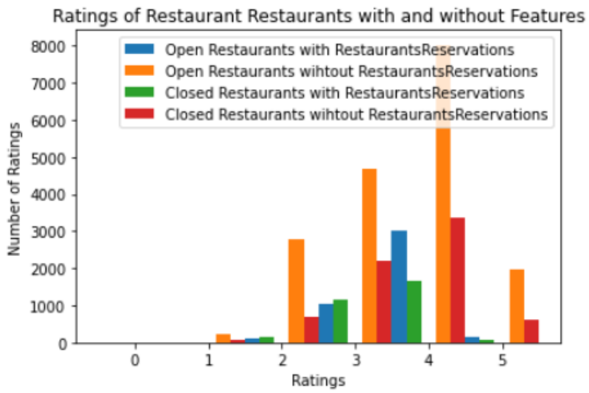


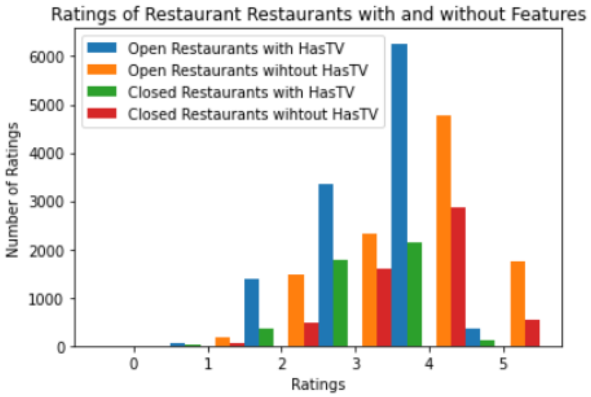


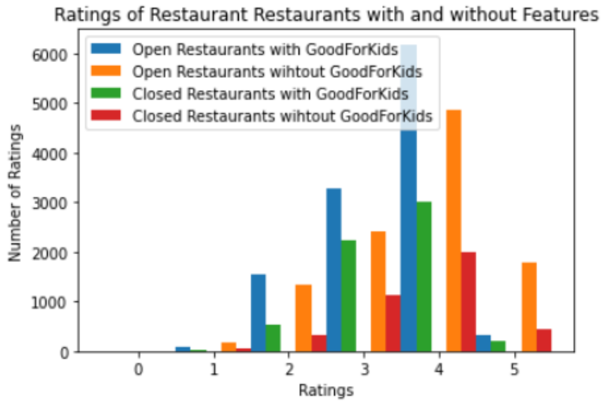


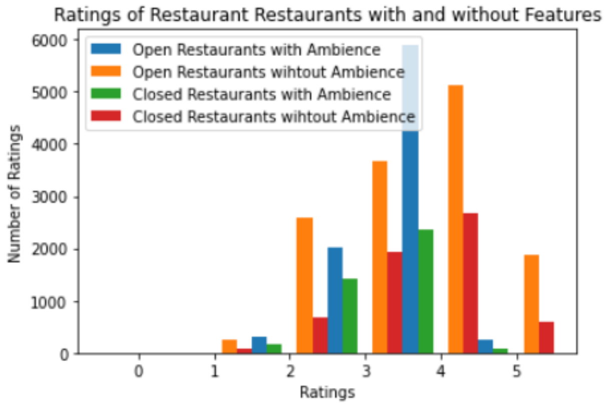

Ratings of Restaurant Restaurants with and without Features
Open Restaurants with OutdoorSeating
Open Restaurants wihtout OutdoorSeating
Closed Restaurants with OutdoorSeating
Closed Restaurants wihtout OutdoorSeating
Number of Ratings
7000
6000
5000
4000
3000
2000
1000
0
Ratings
0
1
2
3
4
5


Ratings of Restaurant Restaurants with and without Features
Open Restaurants with BikeParking
Open Restaurants wihtout BikeParking
Closed Restaurants with BikeParking
Closed Restaurants wihtout BikeParking
Number of Ratings
7000
6000
5000
4000
3000
2000
1000
0
Ratings
0
1
2
3
4
5


Ratings of Restaurant Restaurants with and without Features
Open Restaurants with RestaurantsReservations
Open Restaurants wihtout RestaurantsReservations
Closed Restaurants with RestaurantsReservations
Closed Restaurants wihtout RestaurantsReservations
Number of Ratings
8000
7000
6000
5000
4000
3000
2000
1000
0
Ratings
0
1
2
3
4
5


Ratings of Restaurant Restaurants with and without Features
Open Restaurants with HasTV
Open Restaurants wihtout HasTV
Closed Restaurants with HasTV
Closed Restaurants wihtout HasTV
Number of Ratings
6000
5000
4000
3000
2000
1000
0
Ratings
0
1
2
3
4
5


Ratings of Restaurant Restaurants with and without Features
Open Restaurants with GoodForKids
Open Restaurants wihtout GoodForKids
Closed Restaurants with GoodForKids
Closed Restaurants wihtout GoodForKids
Number of Ratings
6000
5000
4000
3000
2000
1000
0
Ratings
0
1
2
3
4
5


Ratings of Restaurant Restaurants with and without Features
Open Restaurants with Ambience
Open Restaurants wihtout Ambience
Closed Restaurants with Ambience
Closed Restaurants wihtout Ambience
Number of Ratings
6000
5000
4000
3000
2000
1000
0
Ratings
0
1
2
3
4
5