

# **ANLY\_501\_Project\_Report Part Two**

# **What Makes a Great Restaurant?**

**Group 6**

Shaoyu Feng, Jiaxuan Sun, Jen Wang, Chelsea Wang

Contents

[ANLY\_501\_Project\_Report Part Two 1](#_Toc529914822)

[What Makes a Great Restaurant? 1](#_Toc529914823)

[Overview 2](#_Toc529914824)

[Exploratory Analysis 3](#_Toc529914825)

[Basic Statistical Analysis and Data Cleaning Insights: 3](#_Toc529914826)

[Basic Statistics Analysis: 3](#_Toc529914827)

[Handling with Missing Values 5](#_Toc529914828)

[Outlier Detection 6](#_Toc529914829)

[Binning Features 9](#_Toc529914830)

[Additional Feature Engineering 10](#_Toc529914831)

[Histograms and Correlations: 10](#_Toc529914832)

[Data cleaning before data visualization 10](#_Toc529914833)

[Assess relationships between explanatory and outcome variables 11](#_Toc529914834)

[Check correlation 11](#_Toc529914835)

[Determine relationships among the explanatory variables 14](#_Toc529914836)

[Cluster Analysis 16](#_Toc529914837)

[Data Binning 16](#_Toc529914838)

[Selecting Homogenous Properties 18](#_Toc529914839)

[Applying Clustering Algorithms for Neighborhood Subset 22](#_Toc529914840)

[Association Rules / Frequent Itemset Mining Analysis 26](#_Toc529914841)

[Data Binning 26](#_Toc529914842)

[Applying Association Rules Analysis 26](#_Toc529914843)

[Predictive Analysis 28](#_Toc529914844)

[Hypothesis Generation and Proposed Methods 29](#_Toc529914845)

[Class Label Generation 29](#_Toc529914846)

[Parametric Statistical Methods 30](#_Toc529914847)

[Logistics Regression Classifier and other Data Driven Predictive Models 32](#_Toc529914848)

[Results Comparison and Discussion 39](#_Toc529914849)

# 

# Overview

The formula for maintaining a great restaurant has changed quite a bit as of late. Customers have new expectations and tend to gravitate toward restaurants that cater to their various needs. The discussion has raised to dissect the qualities of a successful restaurant in business. It encouraged data scientists to find new standards in defining the characters and qualities of a restaurant. The purpose of the study is to utilize data-driven methods in analyzing the critical factors contributing towards the success of restaurants and thus provide data-science solutions to restaurant owners to help them to improve their business.

Social media nowadays allow consumers to share their experiences with the public on designated platforms to help express their opinions, online reviews have become one of the most influential factors in restaurant selection. Yelp, a popular diner review site, is the primary source to collect data about restaurants in large cities. Moreover, the associated information based on the location of restaurants are web scrapped from city-data and extracted from Google API.

Picking up from project part one, the data set collected from yelp fusion API and some external resources is formatted and joint in one piece. Data manipulation and exploratory data analysis is conducted at this stage to maximize insight and uncover the underlying structure in the dataset. At the same time, several techniques such as cluster analysis and Association Rules mining is employed to help diagnose the dataset. Finally, statistical hypothesis testing and machine learning predictive analytics are conduct to make an insightful prediction from the dataset.

# Exploratory Analysis

## Basic Statistical Analysis and Data Cleaning Insights:

Data Cleaning and feature engineering are essential steps prior to any statistical testing or data analytics in a data science project. This section illustrates the procedures and steps taken for data cleaning and feature transformation including basis statistics analysis, dealing with missing data, outlier detection and feature categorization.

### Basic Statistics Analysis:

There are overall four aspects in the restaurant dataset. The first dimension is the basic information of restaurants, including the number of reviews, ratings, and category of the results and so on. The second dimension is the internal attribute of the restaurants, like WIFI option, Ambience, and Alcohol Availability. The third dimension is information about surrounding facilities near the restaurants, like the number of schools, shopping malls or bus stops. The last dimension consists of the demographical information of the restaurants nearby, including the proportion of white people, unemployment rate, education levels and so on. In summary, there are altogether 5133 rows of data, extracting from 7 major cities and metropolitan areas, and 59 dimensions representing different aspects of the restaurants.

Among the 79 features, there are 56 numerical features and 23 categorical features. Table1-1 shows the physical meaning and types of each column in the data.

Table -1: Dataset Feature List



Table 1-2 shows a snapshot of basic statistics of numerical columns, including distinctive count, mean, min, max and quantiles. Table 1-3 shows the mode of some categorical features in the dataset. As shown in the following tables, there are missing values in some columns like ‘bank’, ’bar’. Therefore, dealing with the missing values will be the first step for data cleaning.

Table 1-: Basic Statistics of Numerical Columns

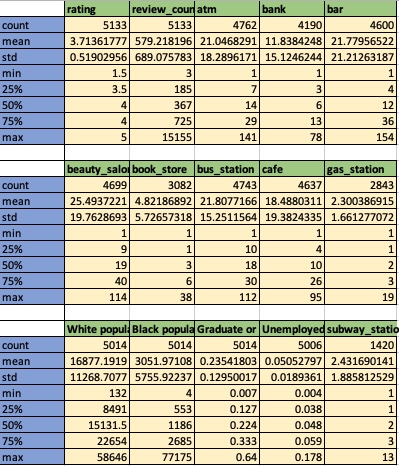
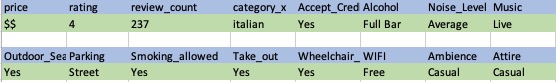
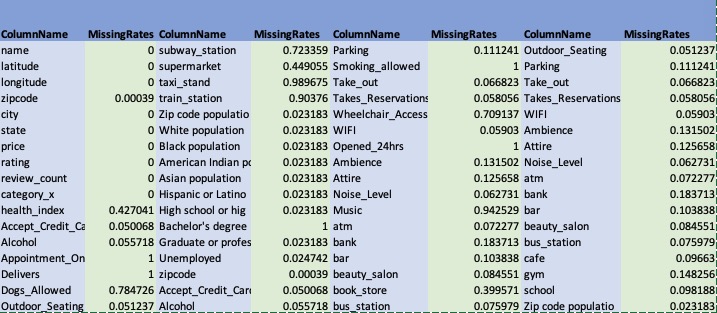


Table 1-3: Mode of Categorical Features



### Handling with Missing Values

Table 1-: Missing Rate for Columns



As shown in table 1-4, the missing rate varies significantly over different features. Therefore, different missing value handling techniques need to be applied. The missing values are mainly due to information missing in the Yelp website and Google Map, as not all restaurant has all the features above. In general, three different techniques have been applied to deal with NA values in this dataset.

First, for all features with more than 40% of missing rate are dropped off, as there will never be a fair imputation method to fill in the blanks. Second, for the majority of the remaining features, imputation by median/mode to fill in the blanks. The rationale behind is that variance in some of the numerical columns is large, imputation by median will introduce less variance towards the dataset. Thirdly, for missing data in columns like ‘LowPrice’, the values can be inferred by the price range column in the dataset.

### Outlier Detection

Although the dimensions in this dataset all have physical meanings and extreme values happen occasionally, outlier detection and handling are still needed as outliers will create barriers for obtaining high accuracy machine learning models. Three outlier detection methods are applied to detect extreme values in this dataset.

#### Z-Score Based Methods

The most common outlier detection method is to make use of box plot and z-score to flag out any value that is far away from the population mean in a univariate manner. Figure 1-1 illustrates the boxplot for 5 numerical features in the dataset, where there are some high spikes exist in those columns. A z-score threshold of 3 is applied to flag out any point far away from its mean.

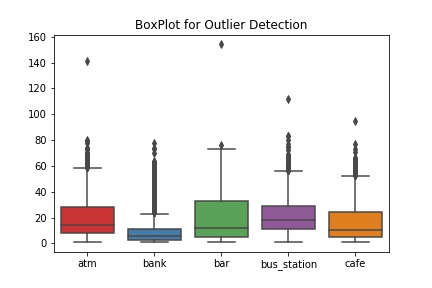


Figure 1-: Outlier Detection using Box Plot

#### Local Outlier Factor (LOF):

Local outlier factor is an outlier detector for finding anomalous data point by measuring the local deviation of a given data point with respect to its neighborhood. Figure 1-2 illustrates the results for applying LOF methods on the 5 features shown above in a multi-dimensional way. The following plot is plotted using PCA decomposition, and the radius of the cycle denotes the outlier score. The larger the radius of the cycle, the higher the chance the data point is an outlier.

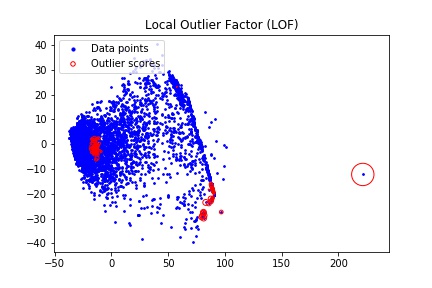


Figure 1-: Outlier Detection using LOF

#### Isolation Forest:

Employing decision tree-based technique, isolation forest detects the outliers by assuming that outliers are rare and different from the main population, and therefore it is easier to be split out using shallow decision trees. Figure 1-3 illustrates the results of isolation forest using the same feature stated above, the dot in red are treated as outliers.



Figure 1-: Isolation Forest Outlier Detection

Data point will be treated as outlier if all three of the methods repost certain data point as outlier. In this case, there is one common data point being flagged by all three methods, and it will be removed from the analysis.

### Binning Features

As shown in table 1-2, some of the features will have a large standard deviation, and the distribution is highly skewed. Data binning or categorization is a useful method to deal with such a situation. In this dataset, one crucial data issue is that the review counts are highly variant, and it adds difficulty in the classification task. Figure 1-4 shows an effective binning method dealing with high variance review count column. The binning strategy here is to make sure the frequency in each category is comparable, in order to facilitate later data analytics.

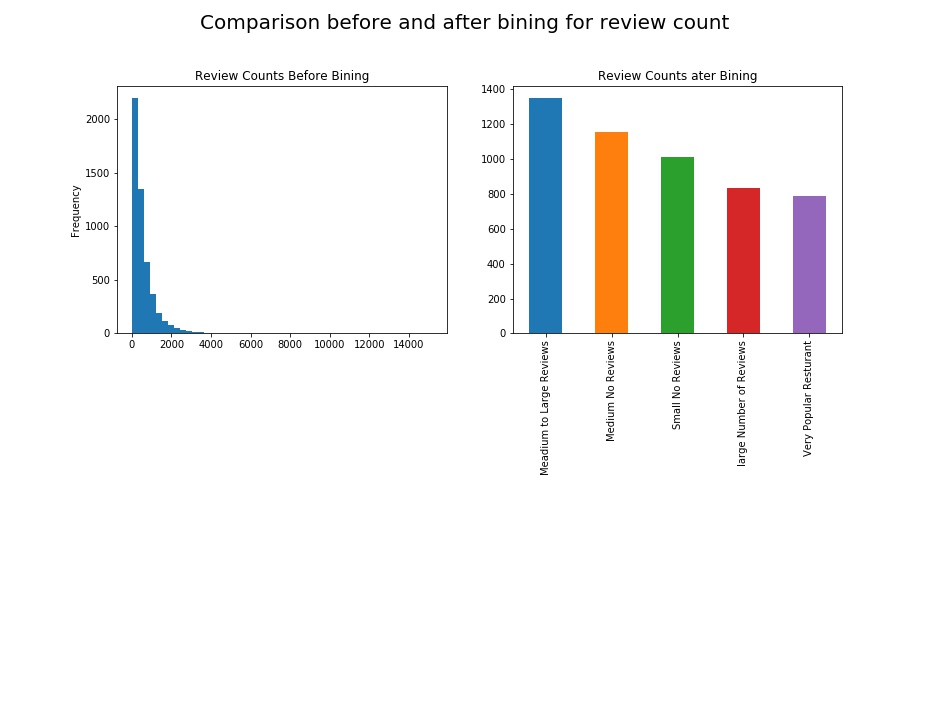


Figure 1-: Data Binning Example

### Additional Feature Engineering

In addition to the stated data cleaning and feature generation method, there several other steps being taken to transform the data features to a more user-friendly format. For example, any Boolean column with value ‘Yes’ or ‘No’ will be filled by ‘1’ and ‘0’ so that it can be used directly in any numerical analysis. Also, instead of using absolute population number, the ratio of each race has been calculated against the total population within a neighborhood.

## Histograms and Correlations:

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques to maximize insight into a data set, uncover the underlying structure and determine optimal factor settings. This section performs graphical procedures to roughly assess relationships between explanatory and outcome variables, check assumptions and identify relationships among the explanatory variables.

### Data cleaning before data visualization

Upon the nature of the data, review rating of restaurant would be inflated when there are few reviews posted. Also, some restaurants with ridiculously high review counts are not helpful to explore the general trend and provide insights to the business. For interpretation purpose, the restaurants with fewer than 10 reviews or more than 3000 reviews are removed from the data frame in this section.

### Assess relationships between explanatory and outcome variables

Review counts and rating are two primary outcome variables to measure the popularity and quality of a restaurant. As shown in figure 2-1, costumers preferably give a rating of 4 and 4.5 at every price range. On the other side, expensive restaurants are likely to gain more popularity than restaurants at a lower price level.

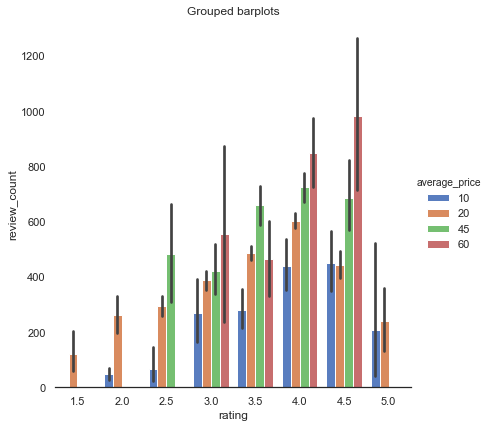


Figure 2-1: Grouped bar plots for rating vs review counts grouped by average price

### Check correlation

Correlation is the first step to understanding the relationships between quantities and subsequently building better models. Correlation can help in predicting one quantity from another and also can be used as a basic quantity and foundation for many other modeling techniques.

Pearson's Correlation test on the all the continuous variables gives result shown in this diagonal correlation matrix. The gradients in the heatmap vary based on the strength of the correlation. Warm means a positive correlation and cool means a negative correlation. The stronger the correlation is represented by more saturated color.

As shown in figure 2-2, features about the availability of public facilities (atm, gym, school, and so on) have a strong positive relationship with each other. Also, counts of racial categories are likely to have a negative relationship with education level. Moreover, education level and unemployment are moderately correlated. These findings motivate further exploration of three variables (café, bar, unemployment) and their relationships.

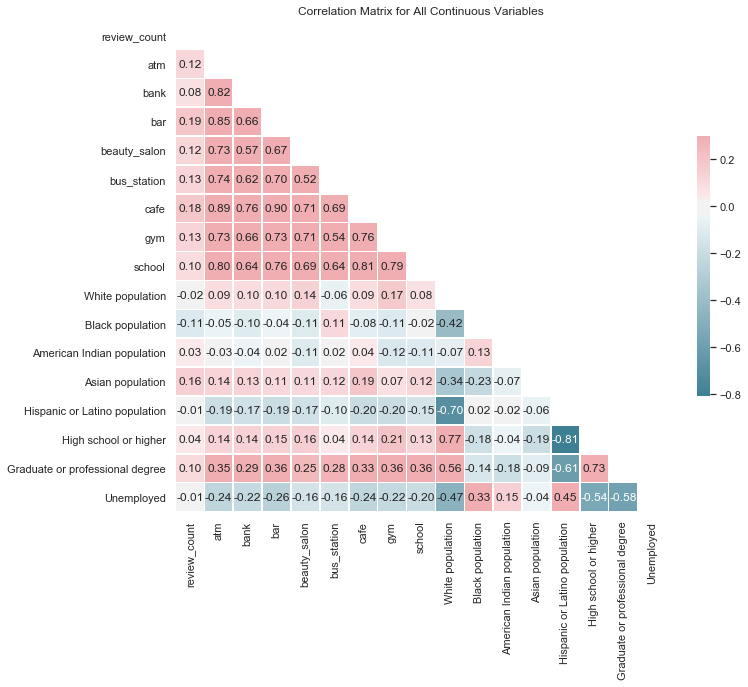


Figure 2-2: Correlation Matrix for Continuous Variables

In Figure 2-3, pair-wise scatter plots show that the distributions of three attributes of interest are right-skewed. Also, both histogram of café and bar shows an abnormally large bin at the tail. It indicates that majority of the restaurants have few bar and café nearby, but there exist some restaurants with a relatively large number of bars and cafés around.

For quantitative variables, it is worthwhile looking at the central tendency, spread, and skewness of the data for a particular variable from an experiment. The most common measure of central tendency is the mean. For skewed distribution or when there is concern about outliers, the median is much more robust.

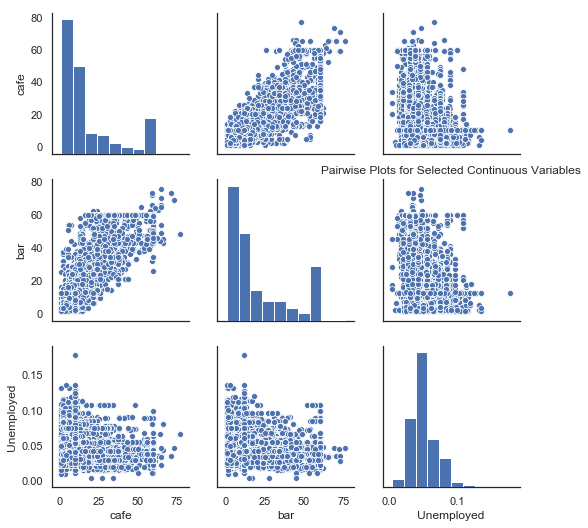


Figure 2-3: Pairwise Plot for three selected Continuous Variables

Figure 2-4 illustrates that bar and café are highly correlated since Pearson correlation coefficient equals to 0.90. This result is consistent with the scatterplot trend between these two variables in Figure 2-3. Besides, both bar and café have a weak negative correlation with unemployment.

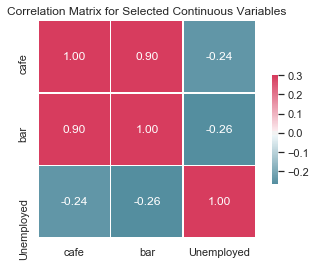


Figure 2-4: Correlation Matrix for three selected Continuous Variables

### Determine relationships among the explanatory variables

As shown in figure 2-5, the distribution of atm and bank have similar peak and trend. However, bank has a larger variance than atm. Since these two explanatory variables are highly correlated shown in Figure 2-2, multicollinearity occurs in the regression model. The independence assumption is not met among these two variables.

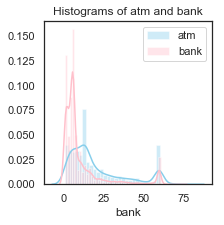


Figure 2-5: Histogram of atm and bank

Figure 2-6 shows the distribution of four races. Black population, Asian population and Hispanic/Latino population have similar distribution and are highly right-skewed, which fit the overall distribution of data set. However, white population is slightly left distributed and has two peaks. This finding is noticeable and should be aware in the future analysis.

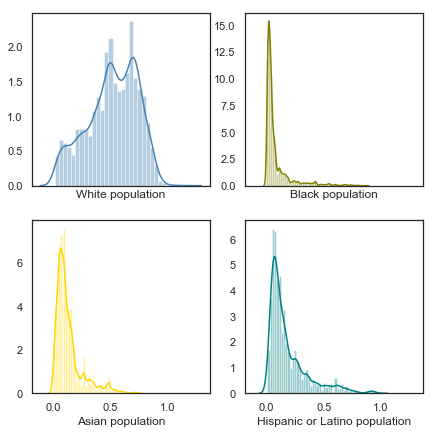


Figure 2-6: Histograms of four races

## Cluster Analysis

Clustering is an essential process of discovering data distribution. Machine learning techniques identify how data are related or unrelated. This section illustrates the procedures of conducting three cluster analyses on our data including a hierarchical clustering method, K-means method, and the dbscan clustering analysis.

### Data Binning

One of essential limitation of the clustering algorithm is that it can be only applied to the numeric data; thus columns with categoric values should be binned into numeric in our dataset before applying the clustering algorithm. Table 3-1 and Table 3-2 shows a snapshot of the cleaned dataset. Features like ‘price’, ‘category\_x’, ‘Alcohol’, ‘Parking’ are categorical, so that these columns need to be binned.

For ‘price’ columns, different numbers of ‘$’ represent the average price of restaurants. Values in this column are binned into 1 to 4 corresponding to the ‘$’ to ‘$$$$’. For ‘category\_x’, ‘Alcohol’, ‘Parking’ columns, values represent the richness of restaurants. For example, if a restaurant has ‘breweries, trad American, beer bar’ category, it can provide three kinds of food. For those three columns, it should be binned according to the number of values each cell has. If a restaurant’s ‘Parking’ feature has three values like ‘Garage, Street, Validated’, the binning value for this feature should be 3 to represent the richness of parking ways it can provide.

After applying binning method on categorical columns, all the values in the dataset are numeric and can be applied clustering algorithms now.

Table 3-1: Full Dataset Part 1

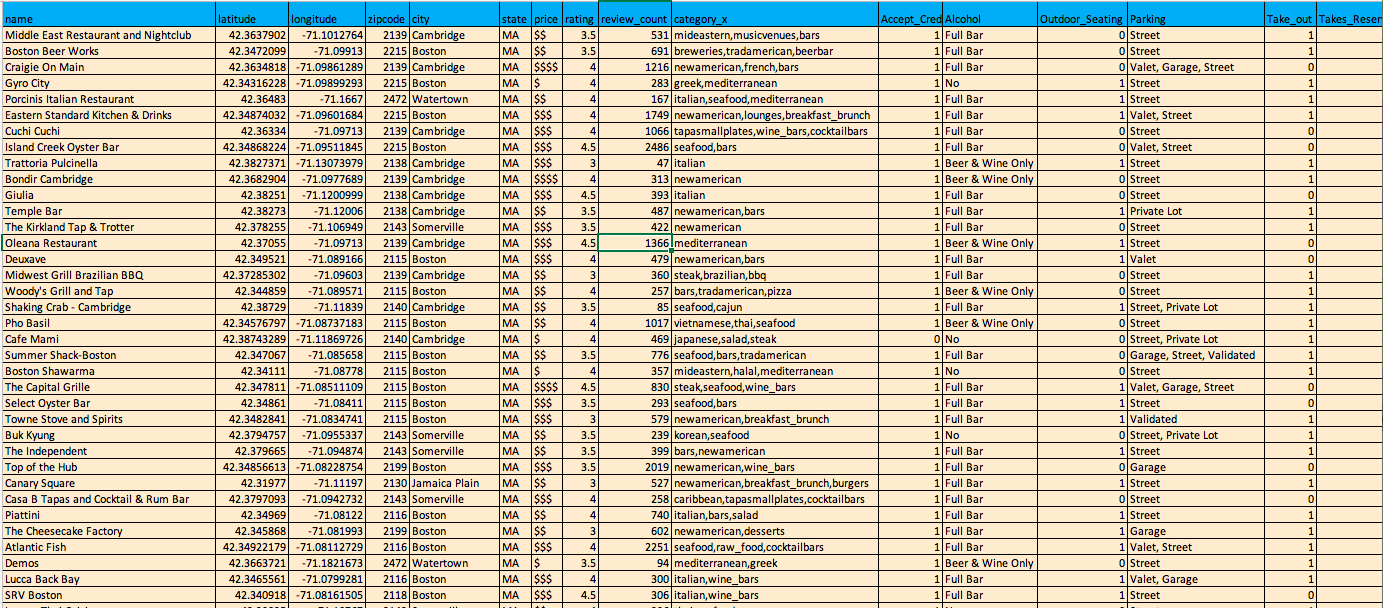
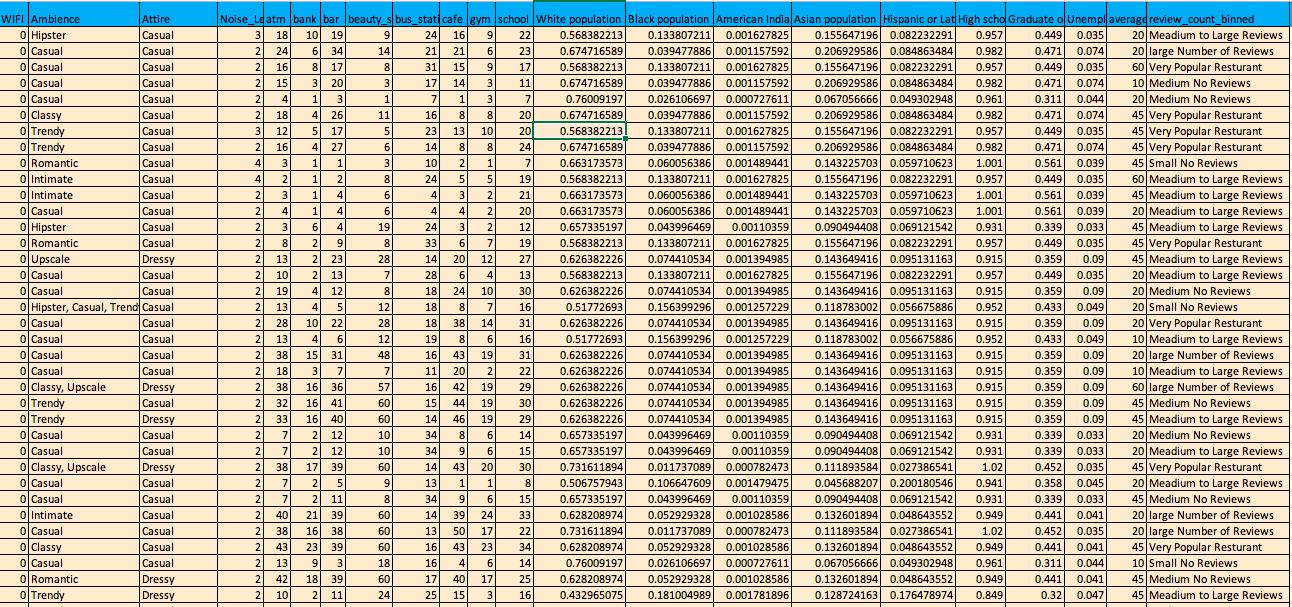


Table 3-2: Full Dataset Part 2



### Selecting Homogenous Properties

At first, three different clustering algorithms are applied to the whole dataset. However, the silhouette score for each algorithm is very low which indicates the poor performance of clustering methods. Figure 3-1 to 3-6 show the clustering results for three algorithms applied to all valuable columns. It is also hard to interpret the meaning of clustering result based on these graphs.

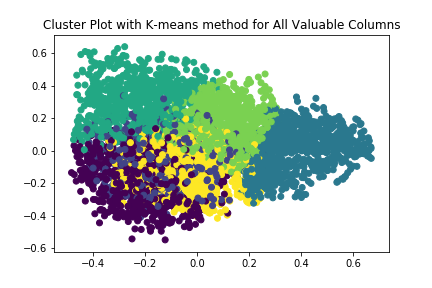


Figure 3-1: 2D graph for k-means method for All Valuable Columns with n=6

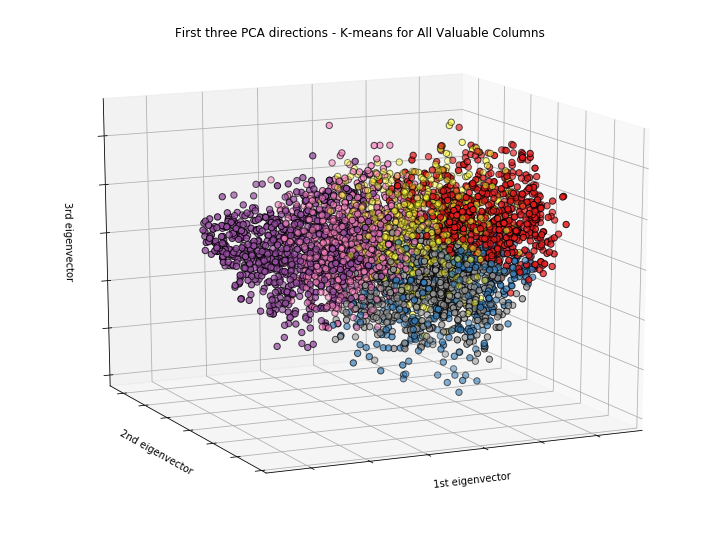


Figure 3-2: 3D graph for k-means method for All Valuable Columns with n=6

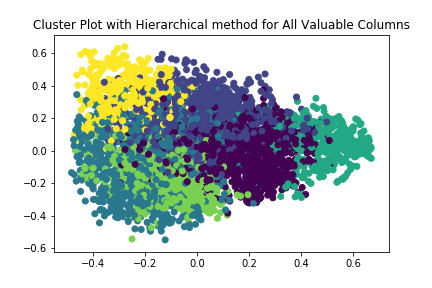


Figure 3-3: 2D graph for hierarchical method for All Valuable Columns with n=6

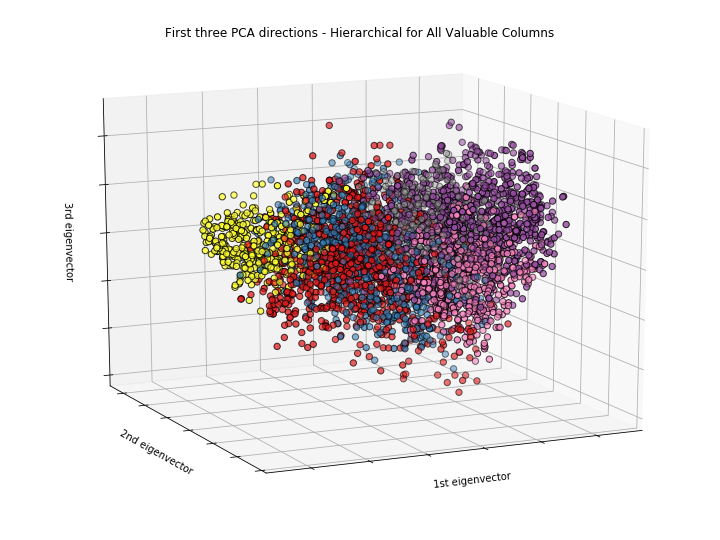


Figure 3-4: 3D graph for hierarchical method for All Valuable Columns with n=6

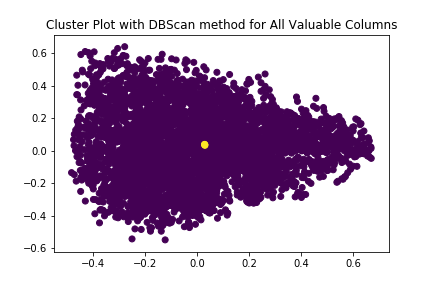


Figure 3-5: 2D graph for dbscan method for All Valuable Columns

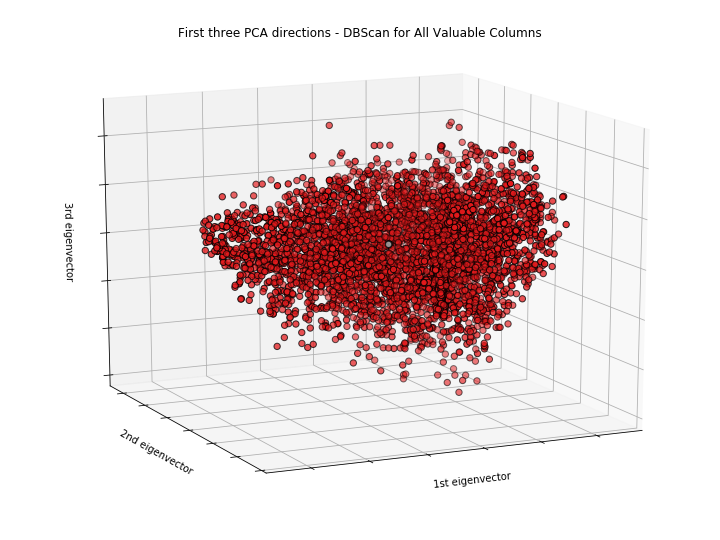


Figure 3-6: 3D graph for dbscan method for All Valuable Columns

This result illustrates that it is meaningless to apply clustering algorithms on the whole dataset since different features have a diverse range of values. For neighborhood features like ‘bank’, ‘school’, the range of these values are about 10 to 30. For features like ‘review count’, most of its values are larger than 100. Therefore, mistakes would occur if applying clustering technique on the above features since the values of these features represent different meaning.

The next step needed to be done is to select features have the same meaning. By looking at the dataset, some features with homogeneous properties can be easily found. There are five features represent the population composition near the restaurants, they are ‘White population’, ‘Black population’, ‘American Indian population’, ‘Asian population’, ‘Hispanic or Latino population’. For the surrounding facilities near the restaurants, there are eight features measure it: ‘atm’, ‘bank’, ‘bar’, ‘beauty\_salon’, ‘bus\_station’, ‘cafe’, ‘gym’, ‘school’. Moreover, feathers like ‘Accept\_Credit\_Card’, ‘Outdoor\_Seating’, ‘Take\_out’, ‘Takes\_Reservations’, ‘WIFI’ represent the internal factors of restaurants. ‘High school or higher’, ‘Graduate or professional degree’ and ‘Unemployed’ features represent the people’s education level in the neighborhood of restaurants. Category\_x’, ‘Alcohol’, ‘Parking’ can be used to represent the internal richness of restaurants.

According to above analysis, clustering algorithms will be applied on five subsets of the full dataset including ‘Population composition’, ‘Neighborhood’, ‘Internal Factors’, ‘Internal Richness’ and ‘Education Level’.

### Applying Clustering Algorithms for Neighborhood Subset

In order to get the most accurate clustering result, three different n values are used for k-means and the hierarchical algorithms. For dbscan method, three different eps values and minimum sample data values are used. The most accurate method for each subset of the dataset is found according to the silhouette score.

Take ‘Neighbourhood’ subset as an example, both k-means and hierarchical algorithms with n=4, 6 and 8 are implemented respectively. From the k-means algorithm, the average silhouette score raises from 0.1759 to 0.1775, and then decreases to 0.1708, so it’s apparent that n=6 is the most suitable value. For the hierarchical algorithm, the average silhouette score drops from 0.1374 to 0.1370 and finally 0.1271, so n=4 should fit this algorithm best.

Figure 3-7 and 3-8 show the clustering results for the k-means method with n=6. Figure 3-9 and 3-10 show the clustering results for the hierarchical method with n=4.

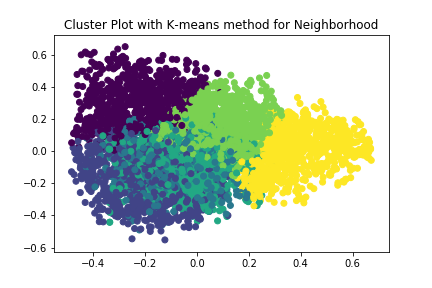


Figure 3-7: 2D graph for k-means method for Neighbourhood with n=6

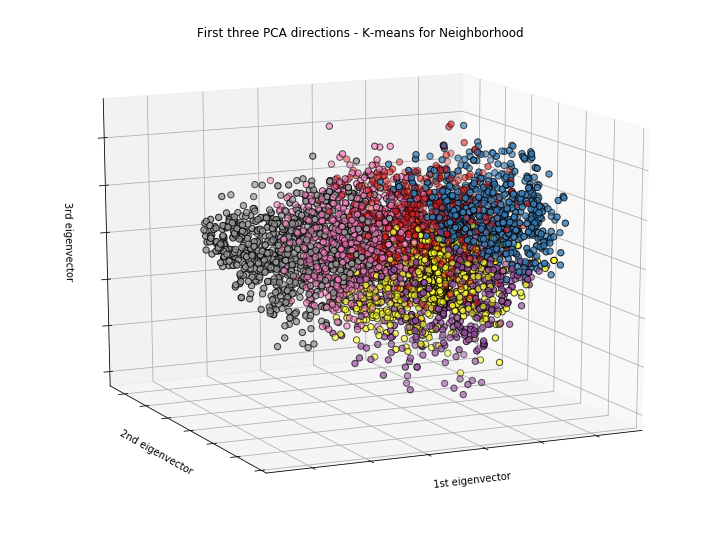


Figure 3-8: 3D graph for k-means method for Neighbourhood with n=6

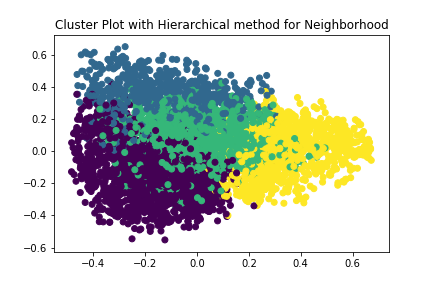


Figure 3-9: 2D graph for hierarchical method for Neighbourhood with n=4

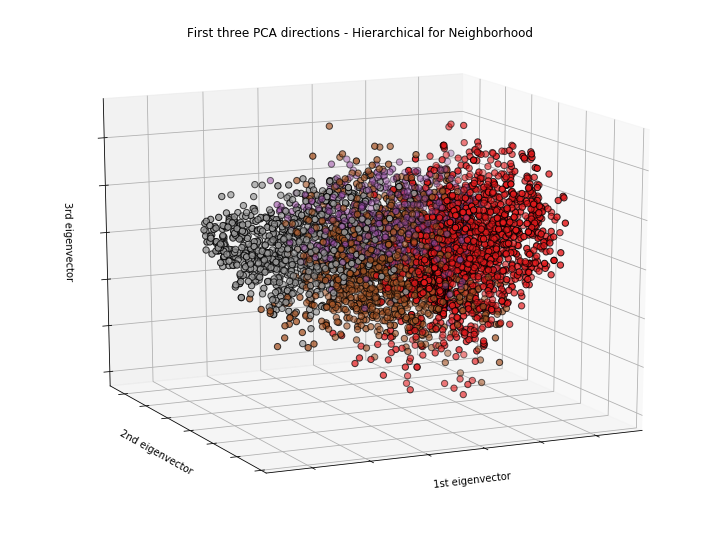


Figure 3-10: 3D graph for hierarchical method for Neighbourhood with n=4

As for dbscan method, average silhouette score with the different pair of eps values and msdv (minimum sample data values) are calculated. When eps is 0.2 and msdv is 100, average silhouette score equals to 0.0281; when eps is 0.25 and msdv is 100, average silhouette score raises to 0.1880; when eps is 0.3 and msdv is 100, average silhouette score increases to 0.2582. As a result, the third pair indicates the best clustering result. Figure 3-11 and 3-12 show the clustering result for dbscan methods with eps=0.3, msdv=100.

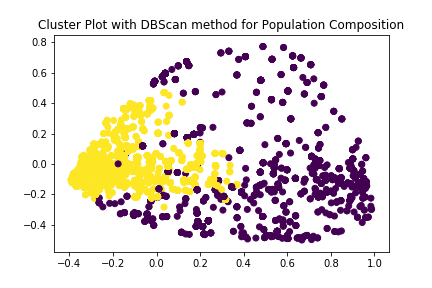


Figure 3-11: 2D graph for dbscan method for Neighbourhood

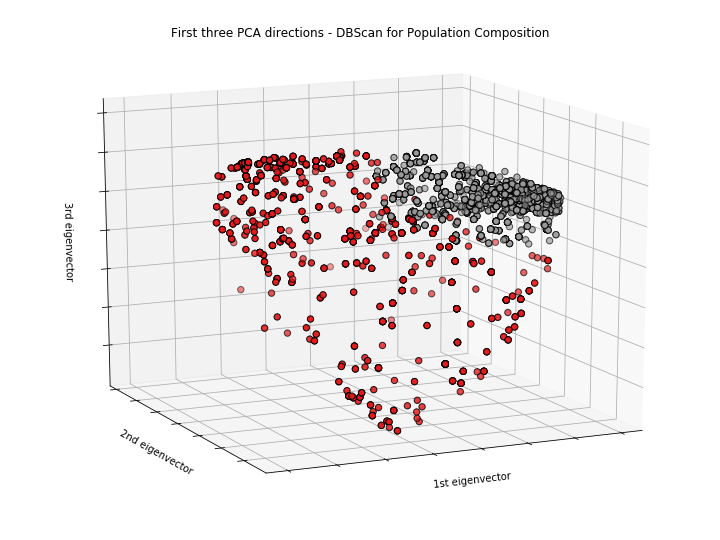


Figure 3-12: 3D graph for dbscan method for Neighbourhood

From all above analysis of the clustering results with three different result, the most reasonable and effective result is from the k-means algorithm. Since one sample contains eight columns, the clustering group should not be as small as two or four. Meanwhile, with six groups, those restaurants are divided into different levels easily. Hence it is an effective way to evaluate the neighborhood’s environment.

Moreover, implementation of the same algorithms to different subsets are achieved, including ‘population composition’, ‘internal factors’, ‘internal richness’ and ‘education level’. A detailed version of those analyses is attaching in a document named ‘clustering results.doc’.

Applying three different clustering algorithms on five subsets of the dataset can comprehensively interpret the distribution of data from multiple angles. To get the best clustering result, three different n values are applied for k-means and hierarchical algorithm respectively and three eps values are applied for DBscan algorithm. 2D and 3D PCA graph are plotted to visualization the clustering result for each algorithm.

## Association Rules / Frequent Itemset Mining Analysis

Association rules mining is to find items that frequently occur together in the same transaction, so it’s useful to find a co-occurrence relationship in our dataset. In order to explore the relationships between other factors and the restaurants’ popularity/rating, utilization of this frequent itemset mining is necessary to find meaningful elements among them. Also, Apriori algorithm improve accuracy and efficiency of the analysis process by reducing itemset from bottom to top.

### Data Binning

In the data binning process, categorization is performed on some numeric factors, including the number of ATMs, number of banks, percentage of the white population and so on, to make sure data within some range could be recognized as the same level.

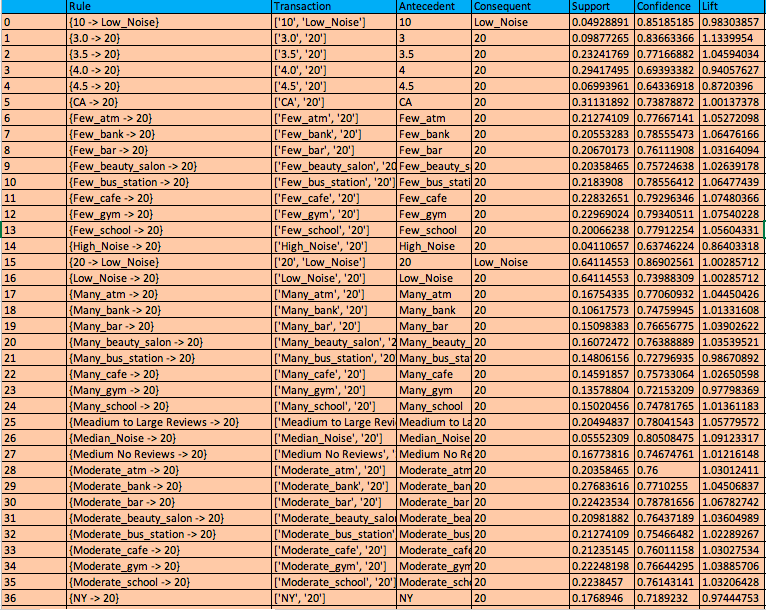
More specifically, data are divided into four parts according to its quantiles – values lower than Q1 would be replaced as ‘small values’; values between Q1 and Q2 are replaced by ‘median values’; values between Q2 and Q3 are replaced by ‘large values’, and values larger than Q3 are replaced by ‘super large values’. Therefore, some columns of table 3-1 and 3-2 are binned into four levels, which could be useful in the later association rules mining process.

Furthermore, categorization is also applied to ‘category\_x’ factor since it contains too many different values that could be defined as the same meaning. To ensure this feature wouldn’t be filtered out because of this problem, values with same or similar meanings and significantly lower the number of unique values are combined.

### Applying Association Rules Analysis

When applying the association rules mining method, msvs (minimum support values) and mcvs (minimum confidence values) for the Apriori method are selected differently. Starting with msvs=0.02 and mcvs=0.5, then increase minimum support threshold to 0.03, 0.04 and 0.05. As a result, four association rules datasets are achieved. Table 4-1 is a snapshot of association rules that we get when msvs=0.04 and mcvs=0.5.

Table 4-1: Association Rules Dataset with support = 0.04



In the dataset, rating and counts of reviews represent the quality and popularity of restaurants. Since the goal is to find factors leading to a good restaurant, only rules which related to the association between factors and the rating or reviews counts are valuable to keep, so that all irrelevant rules are filtered out.

Table 4-2: Filtered Association Rules Dataset with support = 0.03

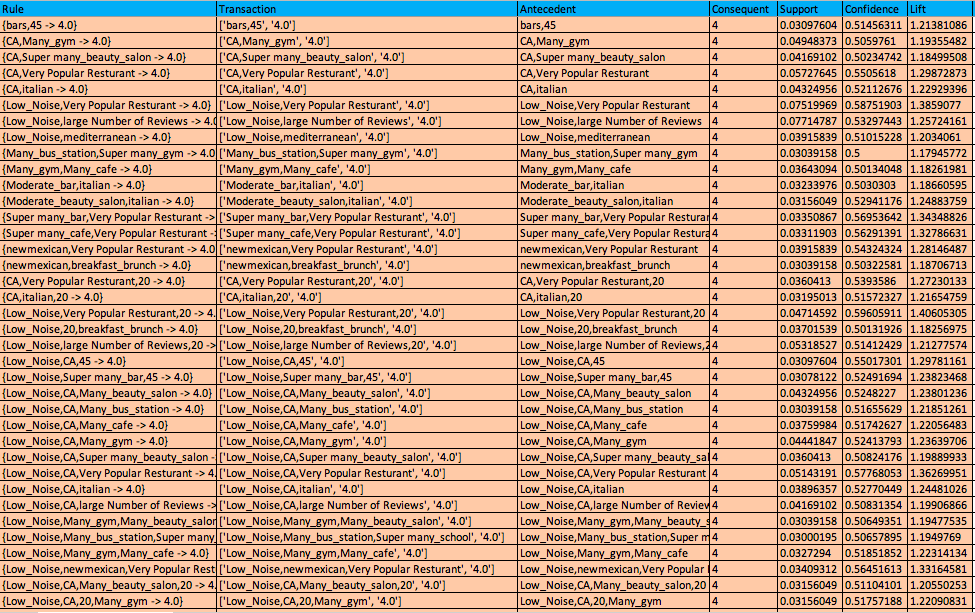


Table 4-2 is a snapshot of the filtered association rules dataset with msvs=0.03 and mcvs=0.5. The dataset gives insights about useful relationships.

The second rule illustrates that when a restaurant has a larger number of reviews and the average price of this restaurant is 20, then it is very likely to be rated as 4. Therefore it is reasonable to state that people prefer to write reviews if restaurants are good. Also, a lower price could gain more popularity.

From rule 5,6,7, the environment has a significant impact on the restaurants’ score since restaurants with super many bars, beauty salons and gyms nearby are more likely to have rating 4, and with a higher consumption level of average 45 dollars for those people have higher life quality. In rule 13 to 15 and 26 to 37, it shows that low noise is an essential factor to be taken into account, and those rules with this same factor can normally lead to a score 4.

Association rules mining analysis could interpret important relationships among different features of data. Binning method is used as a pre-processing step on the dataset to transfer numeric columns into categoric columns to guarantee the correctness of the algorithm. By changing different minimum support values and confidence values, rules of different credibility are generated. Applying filter method on rules dataset, rules with ‘rating’ and ‘review count’ will be preserved and rules with irrelevant information will be filtered out.

# Predictive Analysis

## Hypothesis Generation and Proposed Methods

Based on the exploratory data analysis and heuristics, several hypotheses have been brought out to extract more insights from the dataset:

First, two of the most significant attributes in the restaurant data set are review counts and price. The former is a great indicator of the popularity of the restaurant, and the latter is an important metric for the customer to rate the restaurant. Therefore, the first hypothesis being brought up is whether the review count distributions are the sample among different price groups. The null hypothesis is that the distribution of review counts are the sample among all 4 price groups, namely ‘$’,’$$’,’$$$’ and ‘$$$$’ To verify the hypothesis, Analysis of Variance(ANOVA) method will be performed. Since there are multiple levels in the price features. T-test will also be applied to support the hypothesis further.

Second, intuitively, popularity and ratings of the restaurants should have a strong association. Popularity can be measured by review counts. The second hypothesis is that the review count and ratings of the restaurant have a linear relationship. This can be tested using a linear regression model.

The third hypothesis states as the good rating restaurants, moderate rating restaurants, and good restaurants can be well-separated using the features listed in the dataset. In other words, the hypothesis states that there is a clear decision boundary between different classes of the restaurants. Testing this hypothesis is the processing of building a multi-class classification model. To verify the hypothesis, logistics regression and other data-driven machine learning models will be applied.

### Class Label Generation

One of the most important tasks prior to any supervised classification task is to make sure the data is properly labeled. The objective in this dataset is to predict whether a restaurant is good, and the most direct metric is the rating provided by Yelp. However, one issue with this label is that the number of reviews will potentially influence this rating. To compensate for the bias introduced by review count, a new fusion metric has been introduced. The new rating is calculated by treating the original rating minus 0.5 as the base score and the review count z-score will penalize within its original base score group. For example, if a restaurant has a rating of 4, its base score will be 3.5 and the final score is calculated by 0.5 times the ratio of the review count of this restaurant to the maximum review count of restaurants whose rating are also 4.

After the final score has been calculated, the class label is generated by binning the final score. Three classes have been created, namely 0, 1 and 2, they represent poor rating restaurant, moderate rating restaurant and good rating restaurant respectively. The objective of the classification task is to correctly classify each restaurant to the appropriate class using supervised classification techniques. Table 5-1 shows the class distribution in the class label.

Table 5-1 Class Label Distribution

|  |  |  |
| --- | --- | --- |
| Class Label | Count | Physical Meaning |
| 0 | 2482 | Poor Rating |
| 1 | 2151 | Moderate Rating |
| 2 | 499 | Good Rating |

### Parametric Statistical Methods

#### T-Test and ANOVA

ANOVA test is used to test the first hypothesis, which states that there is no significant difference in the number of reviews(review\_counts) and price range. Figure 5-1 illustrated a two way ANOVA test results. Moreover, Figure 5-2 shows the Quantile-Quantile Plot of theoretical quantiles and sample quantiles in the test.

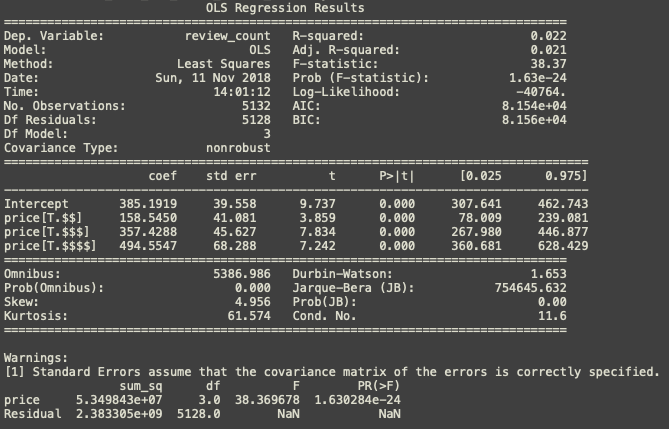


Figure -1: ANOVA Test Result



Figure 5-2: ANOVA Q-Q Plot

As illustrated above, there is a significant sum of square error, an F score of 38.37, and a near 0 p-value. All those statistics show strong evidence against the null hypothesis; therefore, the hypothesis is not valid. There are significant differences in review counts among different price groups.

To further verify the results, the t-test is conducted among pair-wisely among different price groups. Figure 5-3 shows an example of price group ‘$’ against other price groups. Again, this is evidence against the proposed hypothesis.

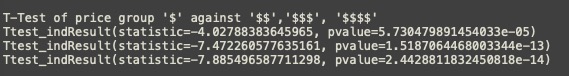


Figure 5-3: Pair-wise T-test Example

#### Linear Regression

Linear regression is one of the most effective ways in examining the linear relationship between two variables. To verify the second hypothesis, a linear model has been fitted into the data and figure 5-4 shows the results of the linear regression model. The R-score obtained by this model is 0.16, which suggest that the linear relationship between these two variables is weak.



Figure 5-4: Linear regression Model Result

### Logistics Regression Classifier and other Data Driven Predictive Models

In next portion, six data-driven predictive models are applied to test the third hypothesis. Also, all six methods utilize the same set of training and testing data. K-fold cross-validation method has been used to test the robustness of the model, and ROC-AUC plot and confusion matrix for each model will be generated separately.

#### Logistic Regression

Logistics Regression is a widely used statistic model which utilize logistic function to model binary/multiclass dependent variables, in this case, the restaurant classes. Figure 5-5 and 5-6 illustrated the Receiver Operating Characteristic (ROC) curve for all class label and the confusion matrix heat map.

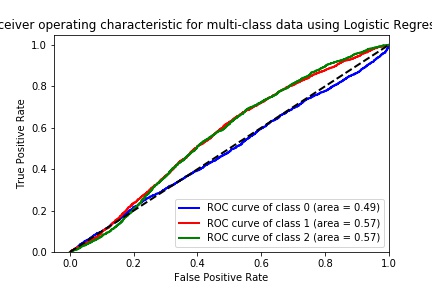


Figure 5-5: ROC for Logistic Regression

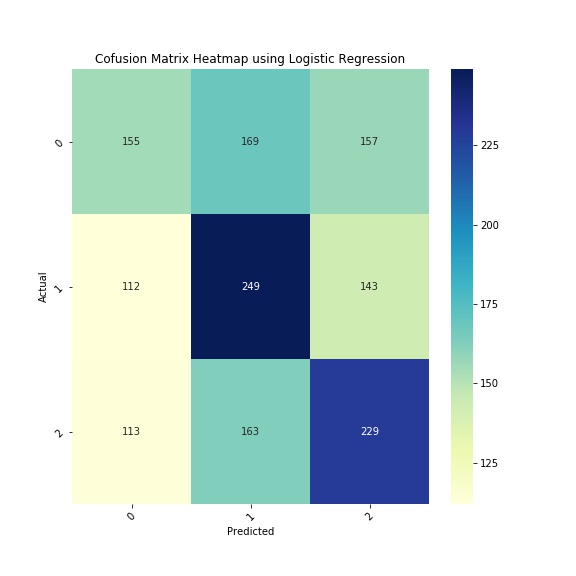


Figure 5-6: Confusion Matrix Heatmap for Logistic Regression

#### Decision Tree Classifier

Decision tree-based classifier is one of the most-easy-to-understand classification techniques. The properties that the results of a decision tree classifier can be easily interpreted has make it popular. Figure 5-7 and 5-8 illustrated the Receiver Operating Characteristic (ROC) curve for all class label and the confusion matrix heat map.

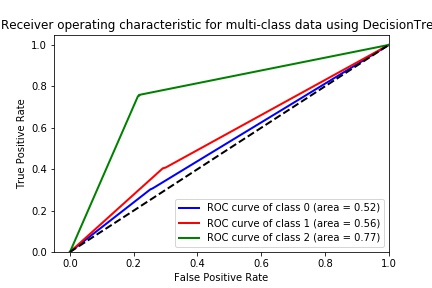


Figure 5-7: ROC Curve for Decision Tree



Figure 5-8: Confusion Matrix Heatmap for Decision Tree

#### Naïve Bayes Classifier

Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem by assuming the data feature are independent. Gaussian Naïve Bayes is used here to deal with the continuous values in this data set. Figure 5-9 and 5-10 illustrated the ROC curve for all class label and the confusion matrix heat map.

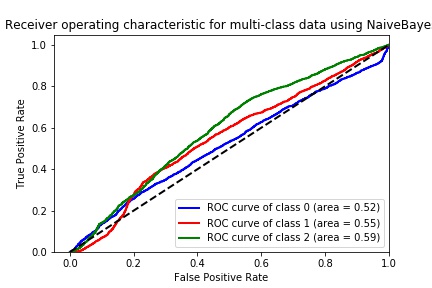


Figure 5-9: ROC curve for Naive Bayes Classifier



Figure 5-10: Confusion Matrix Heatmap of Naive Bayes

#### K-Nearest-Neighborhood (KNN) Classifier

As a lazy learner, KNN classifier is an instance-based learning technique by applying majority vote principle in its neighborhood data point, while the neighborhood is obtained the pairwise distance measure. Feature 5-11 and 5-12 illustrate the ROC curve for all class label and the confusion matrix heat map.

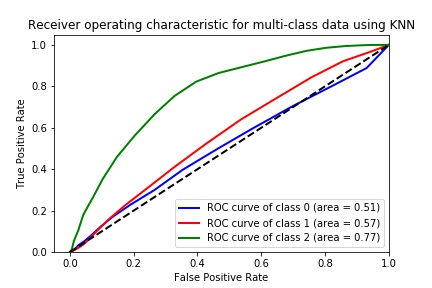


Figure 5-11: ROC for KNN

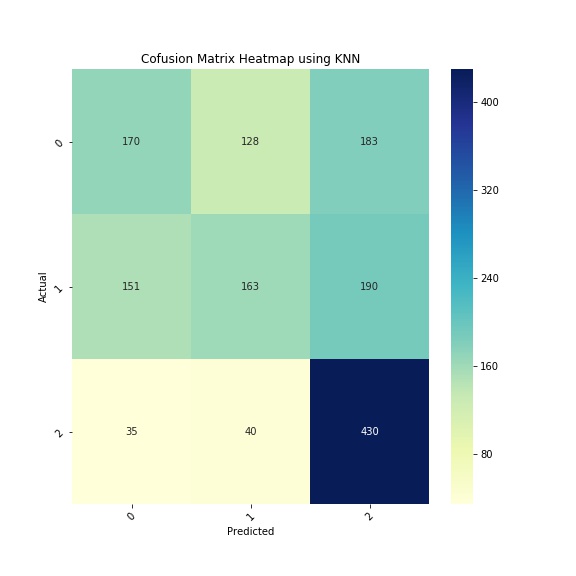


Figure 5-12: Confusion Matrix Heatmap for KNN

#### Support Vector Machine (SVM)

Support Vector Machine is a popular supervised technique to deal with non-linear classification or regression problems. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by an apparent gap that is as wide as possible. In this case, a radial basis function kernel is applied. Figure 5-13 and 5-14 illustrate the ROC curve for all class label and the confusion matrix heat map.

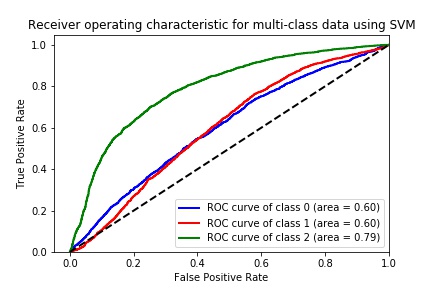


Figure 5-13: ROC for SVM

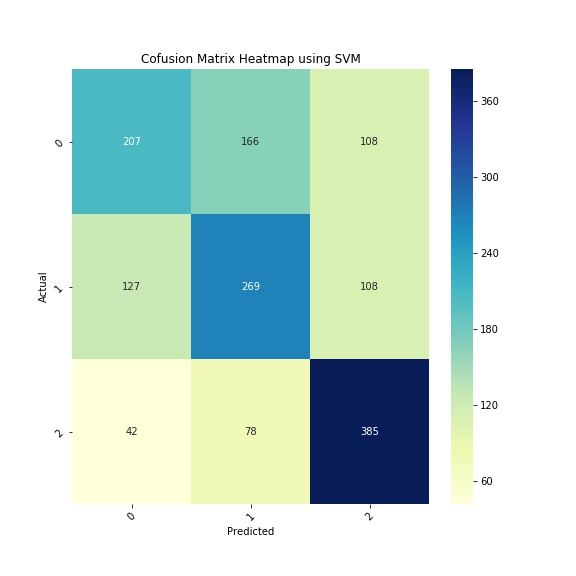


Figure 5-14: Confusion Matrix Heatmap for SVM

#### Random Forest

Employing bagging principle, the random forest is an ensemble-tree based machine learning technique in classification and regression problems. Figure 5-15 and 5-16 illustrate the ROC curve for all class label and the confusion matrix heat map.

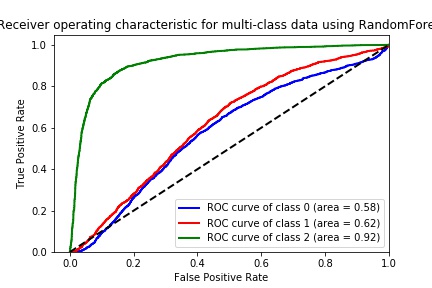


Figure 5-15: ROC for random Forest

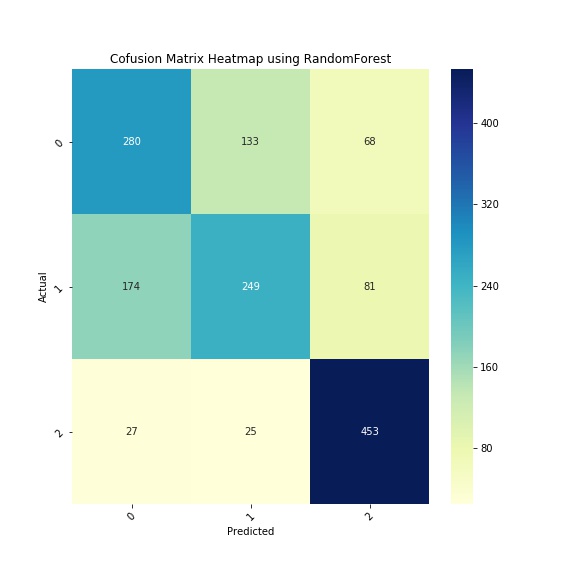


Figure 5-16: Confusion Matrix Heatmap for Random Forest

### Results Comparison and Discussion

Based on the model results above, there are several observations from the ROC curves and confusion matrix:

• The data features are not independent; this is the primary reason why Naïve Bayes classier generated near random results.

• The problem is not a linear-separable problem. Therefore, SVM or decision tree-based methods achieved relatively better classification accuracy.

• Class 0 and Class 1 are not as distinctive as Class 2, none of the classier produce an excellent separation between Class 0 and Class 1. This observation suggests that the class label generation may be biased.

• Out of the 5 classifiers, Random Forest achieved the best AUC score for Class 2. An AUC of 0.92 suggests that the dataset does have predictive power for the rating of a restaurant based on the information given.

**Conclusion and Future Work:**

To summarize, the restaurant dataset has been analyzed from multiple accepts and dimensions including histogram and correlation study, association rule mining, clustering analysis, hypothesis testing, and data-driven machine learning techniques. Numerous insights extracted from the dataset suggest that the dataset does have predictive power in differentiating restaurants with different classes.

However, to complete a data science project, transforming data analytics results to business recommendations is an essential step to make sure stakeholders entirely make use of the data insights. To complete this step, future work needed will be further understanding the decision rules in splitting the class labels, integrating visualization solutions to more insightful reports as well as fine-tuning the current predictive model to achieve better results.