GEORGETOWN UNIVERSITY
The Graduate School of Arts & Sciences

ANLY_501_Project_Report Part Two

What Makes a Great Restaurant?

Group 6

Shaoyu Feng, Jiaxuan Sun, Jen Wang, Chelsea Wang

Contents

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 1-1: Dataset Feature List

Dimension Name	DataType	Description	Dimension Name	DataType	Description
name	String	Resturant Name	*Delivers*	Boolean	Can Deliver
latitude	Float	Latitude	*Dogs_Allowed*	Boolean	Whether pets allowed
longitude	Float	Longitude	*Outdoor_Seating*	Boolean	Whether has outdoor seating
is_closed	Boolean	Resturant Still Open	*Parking*	String	Parking Options
zipcode	String	Zipcode	*Smoking_allowed*	Boolean	Whether Smoking allowed
city	String	City	*Take_out*	Boolean	Take out option
state	String	State Name	*Takes_Reservations*	Boolean	Whether accepts booking
price	String	Price Level, $ or $$	*Wheelchair_Accessible*	Boolean	Disabled frendly
rating	Float	Resturant Rating	*WIFI*	String	WIFI Option
url	String	Web Link	*Opened_24hrs*	Boolean	Is 24 hours opening
review_count	Int	Number of Reviews	*Ambience*	String	Ambience
transactions	String	Types of Business	*Attire*	String	Dress code
category_x	String	Resturant Type	*Noise_Level*	String	Noise level
id	String	Unique ID	*Music*	String	Background music type
Name	String	Resturant Name	*atm*	Int	No of ATMS nearby
category_y	String	Catgorical Columns	*bank*	Int	No of banks nearby
lowprice	String	Price Range Min	*bar*	Int	No of bars nearby
highprice	String	Price Range High	*beauty_salon*	Int	No of Beauty Salon nearby
health_index	String	Health Score	*book_store*	Int	No of Book Stores nearby
star1	Int	Number of 1 star reviews	*bus_station*	Int	No of Bus Station nearby
star2	Int	Number of 2 star reviews	*cafe*	Int	No of cafe nearby
star3	Int	Number of 3 star reviews	*gas_station*	Int	No of Gas Station nearby
star4	Int	Number of 4 star reviews	*gym*	Int	No of Gyms nearby
star5	Int	Number of 5 star reviews	*movie_theater*	Int	No of Theather nearby
Accept_Credit_Card	Boolean	Payment methods	*museum*	Int	No of Museums nearby
Alcohol	String	Alcolhol Availibility	*school*	Int	No of Bus Schools nearby
Appointment_Only	Boolean	Walk in Possible	*shopping_mall*	Int	No of Bus schopping malls nearby
Caters	Boolean	Holding Catering	*subway_station*	Int	No of subway Station nearby
taxi_stand	Int	Number of taxi stands	*supermarket*	Int	No of supermarket nearby
train_station	Int	Number of train stations	*Bachelor's degree or higher*	Float	Percentage of pelple with bachelor degree
White population	Int	Number of White Population	*Graduate or professional degree*	Float	Percentage of pelple with graduate education
Black population	Int	Number of Black Population	*Unemployed*	Float	Unemployment ratio
American Indian population	Int	Number of American Indian Population	*Mean travel time to work*	Float	Travel Time need to work
Asian population	Int	Number of Asian Population	*Now married*	Float	Mariage Ratio
Native Hawaiian	Int	Number of Native Population	*Divorced*	Float	Devorced Ratio
Hispanic or Latino population	Int	Number of Hispanic/Latino Population	*High school or higher*	Float	Percentage of pelple with High school education

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 1-2: Basic Statistics of Numerical Columns

	rating	review_coun	atm	bank	bar
count	5133	5133	4762	4190	4600
mean	3.71361777	579.218196	21.0468291	11.8384248	21.77956522
std	0.51902956	689.075783	18.2896171	15.1246244	21.21263187
min	1.5	3	1	1	1
25%	3.5	185	7	3	4
50%	4	367	14	6	12
75%	4	725	29	13	36
max	5	15155	141	78	154

	beauty_salo	book_store	bus_station	cafe	gas_station
count	4699	3082	4743	4637	2843
mean	25.4937221	4.82186892	21.8077166	18.4880311	2.300386915
std	19.7628693	5.72657318	15.2511564	19.3824335	1.661277072
min	1	1	1	1	1
25%	9	1	10	4	1
50%	19	3	18	10	2
75%	40	6	30	26	3
max	114	38	112	95	19

	White popul	Black popula	Graduate or	Unemployed	subway_statio
count	5014	5014	5014	5006	1420
mean	16877.1919	3051.97108	0.23541803	0.05052797	2.431690141
std	11268.7077	5755.92237	0.12950017	0.0189361	1.885812529
min	132	4	0.007	0.004	1
25%	8491	553	0.127	0.038	1
50%	15131.5	1186	0.224	0.048	2
75%	22654	2685	0.333	0.059	3
max	58646	77175	0.64	0.178	13

Table 1-3: Mode of Categorical Features

price	rating	review_count	category_x	Accept_Cred	Alcohol	Noise_Level	Music
$$	4	237	italian	Yes	Full Bar	Average	Live
Outdoor_Sea	Parking	Smoking_allowed	Take_out	Wheelchair_	WIFI	Ambience	Attire
Yes	Street	Yes	Yes	Yes	Free	Casual	Casual

Handling with Missing Values

Table 1-4: Missing Rate for Columns

ColumnName	MissingRates	ColumnName	MissingRates	ColumnName	MissingRates	ColumnName	MissingRates
name	0	subway_station	0.723359	Parking	0.111241	Outdoor_Seating	0.051237
latitude	0	supermarket	0.449055	Smoking_allowed	1	Parking	0.111241
longitude	0	taxi_stand	0.989675	Take_out	0.066823	Take_out	0.066823
zipcode	0.00039	train_station	0.90376	Takes_Reservations	0.058056	Takes_Reservations	0.058056
city	0	Zip code populatio	0.023183	Wheelchair_Access	0.709137	WIFI	0.05903
state	0	White population	0.023183	WIFI	0.05903	Ambience	0.131502
price	0	Black population	0.023183	Opened_24hrs	1	Attire	0.125658
rating	0	American Indian pc	0.023183	Ambience	0.131502	Noise_Level	0.062731
review_count	0	Asian population	0.023183	Attire	0.125658	atm	0.072277
category_x	0	Hispanic or Latino	0.023183	Noise_Level	0.062731	bank	0.183713
health_index	0.427041	High school or hig	0.023183	Music	0.942529	bar	0.103838
Accept_Credit_Ca	0.050068	Bachelor's degree	1	atm	0.072277	beauty_salon	0.084551
Alcohol	0.055718	Graduate or profes	0.023183	bank	0.183713	bus_station	0.075979
Appointment_On	1	Unemployed	0.024742	bar	0.103838	cafe	0.09663
Delivers	1	zipcode	0.00039	beauty_salon	0.084551	gym	0.148256
Dogs_Allowed	0.784726	Accept_Credit_Car	0.050068	book_store	0.399571	school	0.098188
Outdoor_Seating	0.051237	Alcohol	0.055718	bus_station	0.075979	Zip code populatio	0.023183

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.

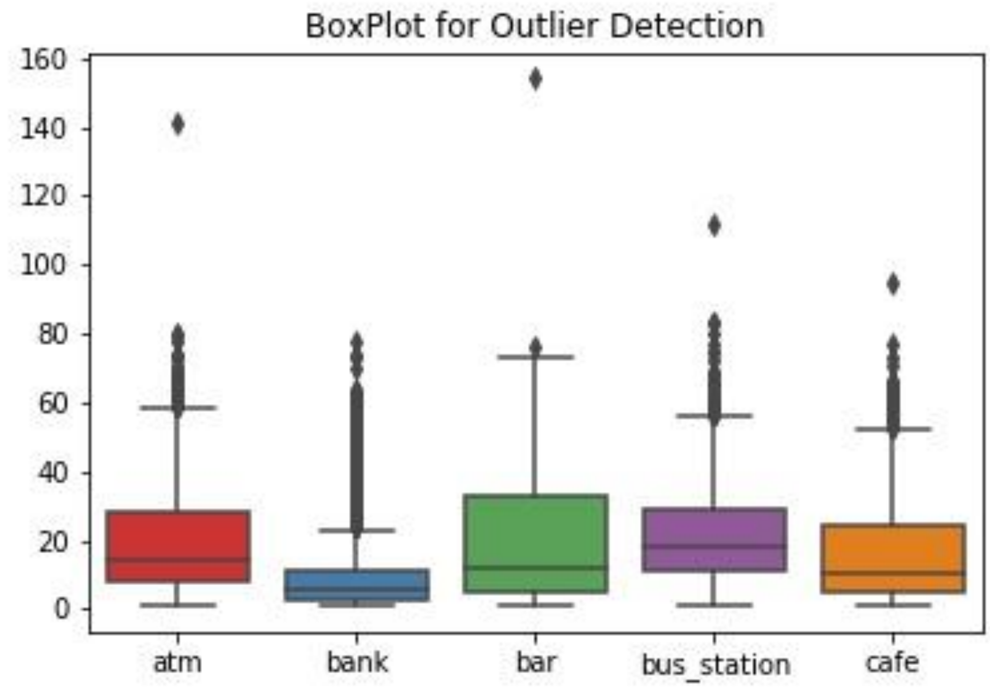

Figure 1-1: Outlier Detection using Box Plot

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.

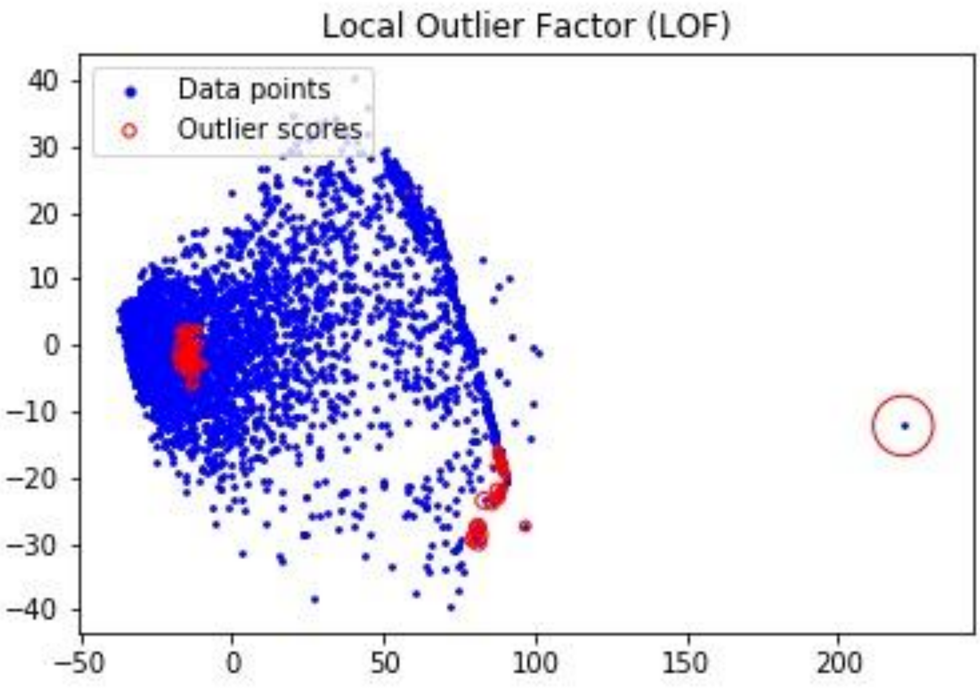

Figure 1-2: Outlier Detection using LOF

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.

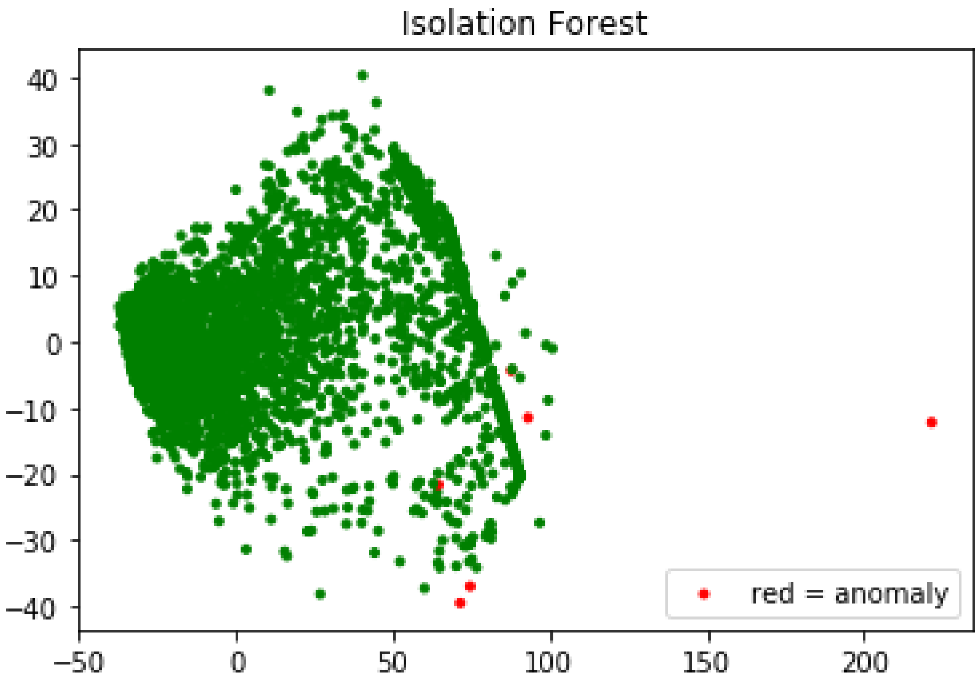

Figure 1-3: Isolation Forest Outlier Detection

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.

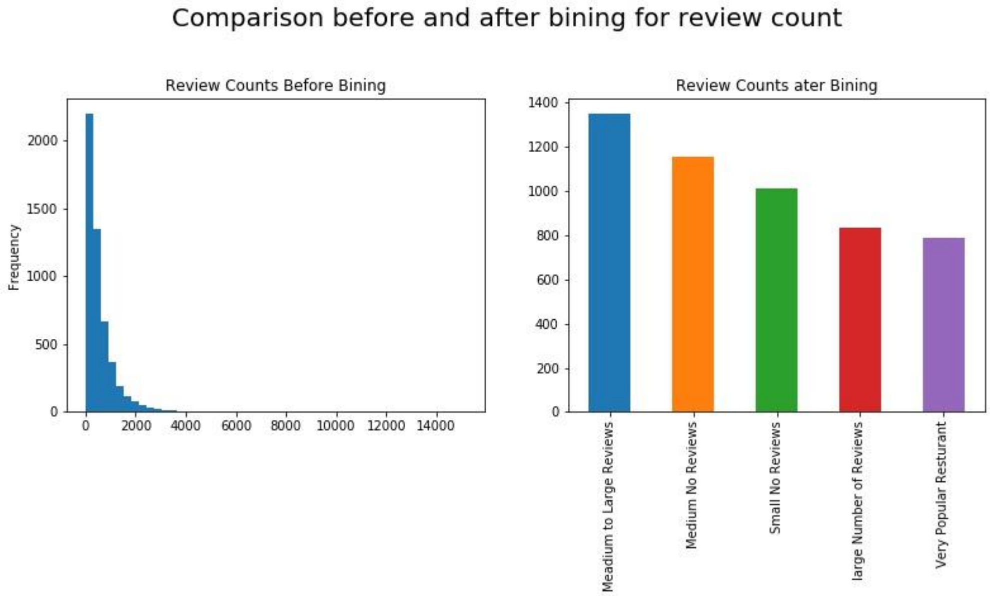

Figure 1-4: Data Binning Example

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.

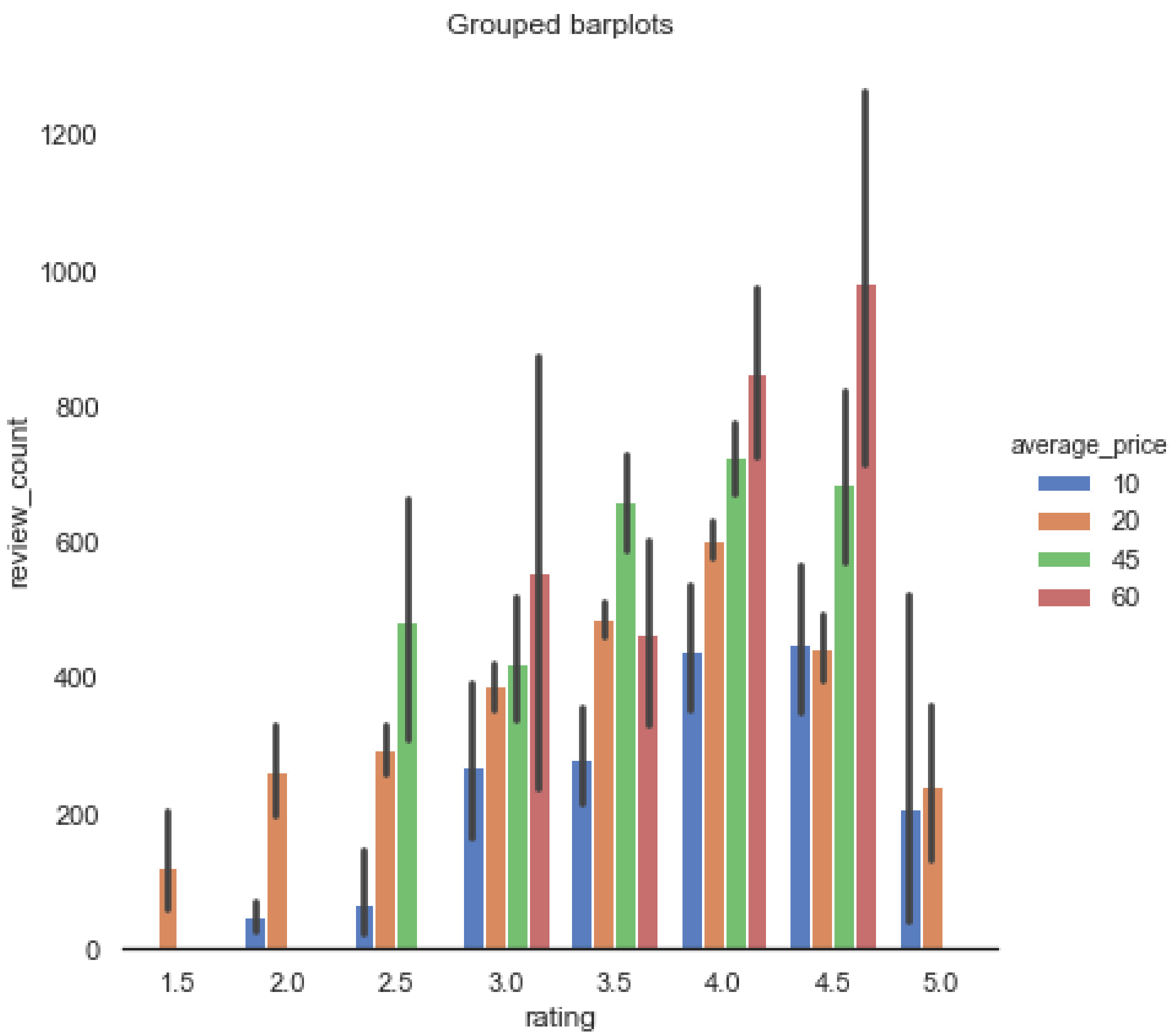

Figure 2-1: Grouped bar plots for rating vs review counts grouped by average price

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.

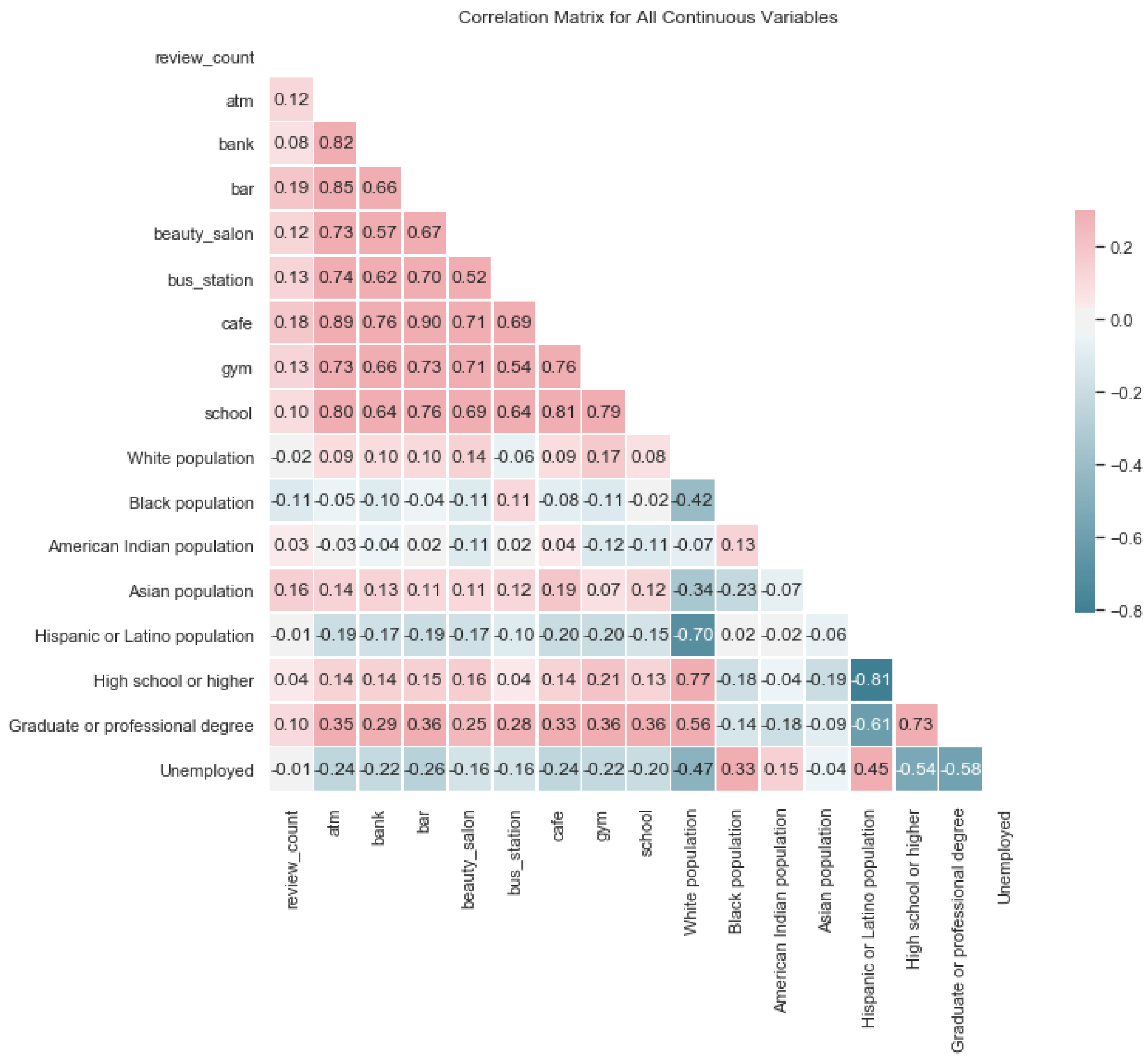

Figure 2-2: Correlation Matrix for Continuous Variables

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.

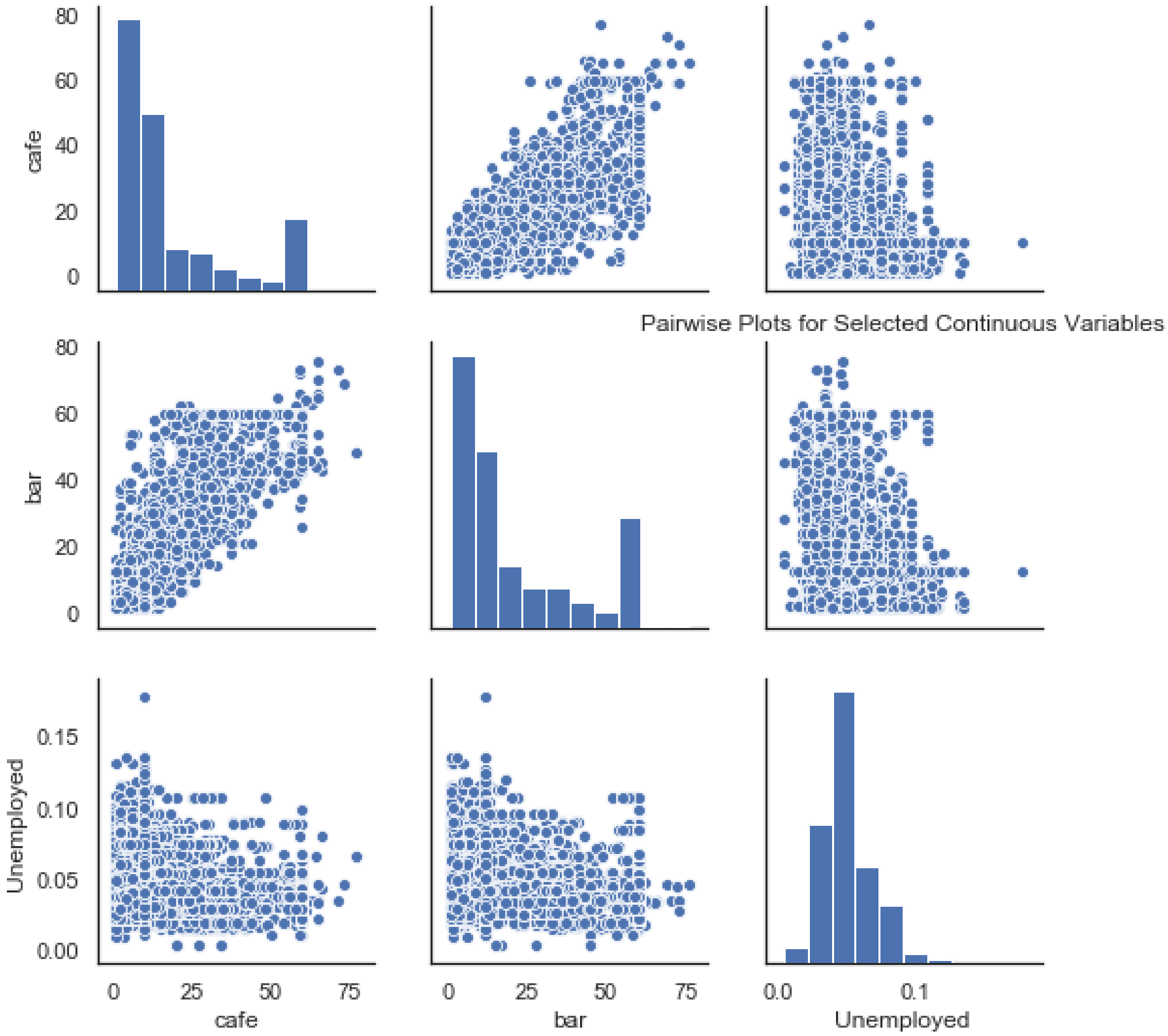

Figure 2-3: Pairwise Plot for three selected Continuous Variables

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.

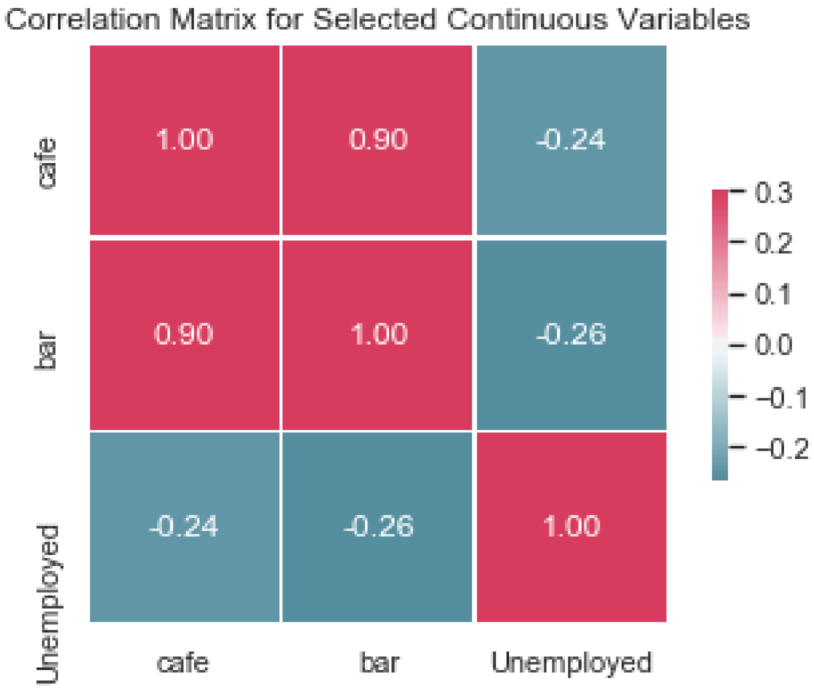

Figure 2-4: Correlation Matrix for three selected Continuous Variables

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.

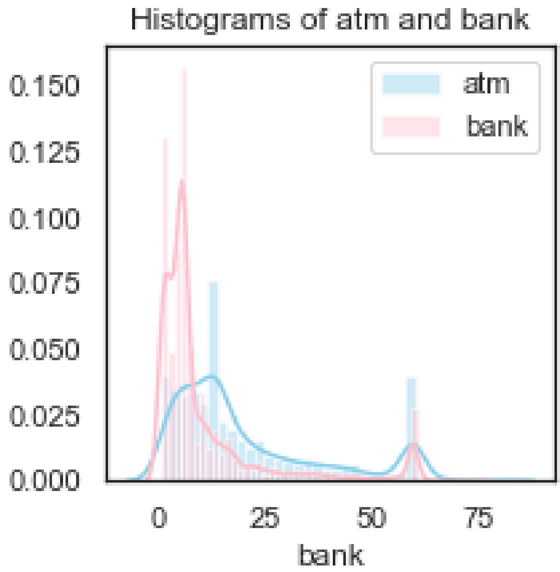

Figure 2-5: Histogram of atm and bank

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.

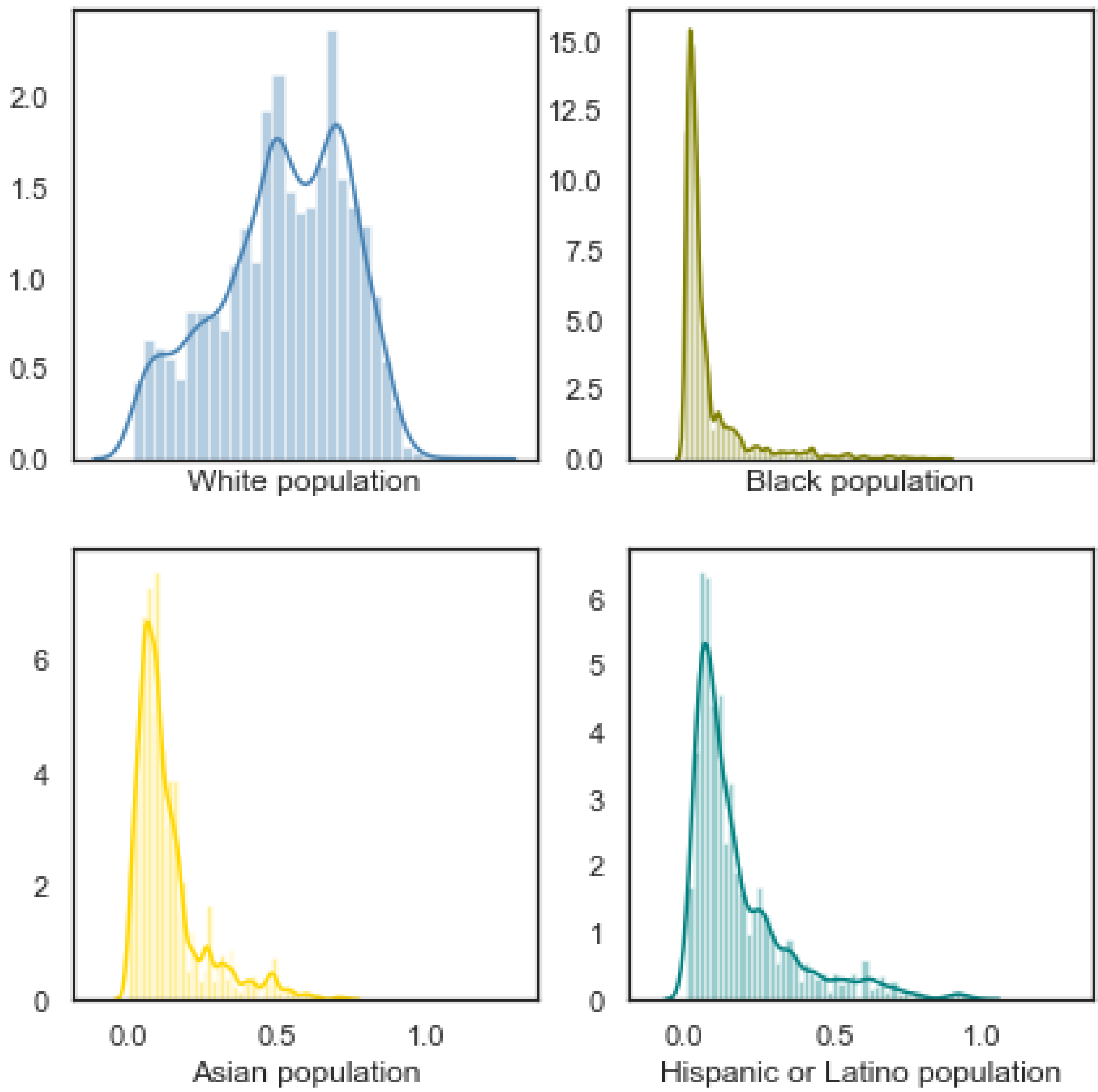

Figure 2-6: Histograms of four races

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 3-5: Full Dataset Part 1

name	latitude	longitude	zipcode	city	state	price	rating	review_count	category_x	Accept_Cred	Alcohol	Outdoor_Seating	Parking	Take_out	Takes_Reser
Middle East Restaurant and Nightclub	42.3637902	-71.1012764	2139	Cambridge	MA	$$	3.5	531	mideastern,musicvenues,bars	1	Full Bar	0	Street	1	
Boston Beer Works	42.3472099	-71.09913	2215	Boston	MA	$$	3.5	691	breweries,tradamerican,beerbar	1	Full Bar	0	Street	1	
Craigie On Main	42.3634818	-71.09861289	2139	Cambridge	MA	$$$$	4	1216	newamerican,french,bars	1	Full Bar	0	Valet, Garage, Street	0	
Gyro City	42.34316228	-71.09899293	2215	Boston	MA	$	4	283	greek,mediterranean	1	No	1	Street	1	
Porcinis Italian Restaurant	42.36483	-71.1667	2472	Watertown	MA	$$	4	167	italian,seafood,mediterranean	1	Full Bar	1	Street	1	
Eastern Standard Kitchen & Drinks	42.34874032	-71.09601684	2215	Boston	MA	$$$	4	1749	newamerican,lounges,breakfast_brunch	1	Full Bar	1	Valet, Street	1	
Cuchi Cuchi	42.36334	-71.09713	2139	Cambridge	MA	$$$	4	1066	tapasmallplates,wine_bars,cocktailbars	1	Full Bar	0	Street	0	
Island Creek Oyster Bar	42.34868224	-71.09511845	2215	Boston	MA	$$$	4.5	2486	seafood,bars	1	Full Bar	0	Valet, Street	0	
Trattoria Pulcinella	42.3827371	-71.13073979	2138	Cambridge	MA	$$$	3	47	italian	1	Beer & Wine Only	1	Street	1	
Bondir Cambridge	42.3682904	-71.0977689	2139	Cambridge	MA	$$$$	4	313	newamerican	1	Beer & Wine Only	0	Street	1	
Giulia	42.38251	-71.1200999	2138	Cambridge	MA	$$$	4.5	393	italian	1	Full Bar	0	Street	0	
Temple Bar	42.38273	-71.12006	2138	Cambridge	MA	$$	3.5	487	newamerican,bars	1	Full Bar	1	Private Lot	1	
The Kirkland Tap & Trotter	42.378255	-71.106949	2143	Somerville	MA	$$$	3.5	422	newamerican	1	Full Bar	0	Street	1	
Oleana Restaurant	42.37055	-71.09713	2139	Cambridge	MA	$$$	4.5	1366	mediterranean	1	Beer & Wine Only	1	Street	0	
Deuxave	42.349521	-71.089166	2115	Boston	MA	$$$	4	479	newamerican,bars	1	Full Bar	1	Valet	0	
Midwest Grill Brazilian BBQ	42.37285302	-71.09603	2139	Cambridge	MA	$$	3	360	steak,brazilian,bbq	1	Full Bar	0	Street	1	
Woody's Grill and Tap	42.344859	-71.089571	2115	Boston	MA	$$	4	257	bars,tradamerican,pizza	1	Beer & Wine Only	0	Street	1	
Shaking Crab - Cambridge	42.38729	-71.11839	2140	Cambridge	MA	$$	3.5	85	seafood,cajun	1	Full Bar	1	Street, Private Lot	1	
Pho Basil	42.34576797	-71.08737183	2115	Boston	MA	$$	4	1017	vietnamese,thai,seafood	1	Beer & Wine Only	0	Street	1	
Cafe Mami	42.38743289	-71.11869726	2140	Cambridge	MA	$	4	469	japanese,salad,steak	0	No	0	Street, Private Lot	1	
Summer Shack-Boston	42.347067	-71.085658	2115	Boston	MA	$$	3.5	776	seafood,bars,tradamerican	1	Full Bar	0	Garage, Street, Validated	1	
Boston Shawarma	42.34111	-71.08778	2115	Boston	MA	$	4	357	mideastern,halal,mediterranean	1	No	0	Street	1	
The Capital Grille	42.347811	-71.08511109	2115	Boston	MA	$$$$	4.5	830	steak,seafood,wine_bars	1	Full Bar	1	Valet, Garage, Street	0	
Select Oyster Bar	42.34861	-71.08411	2115	Boston	MA	$$$	3.5	293	seafood,bars	1	Full Bar	1	Street	0	
Towne Stove and Spirits	42.3482841	-71.0834741	2115	Boston	MA	$$$	3	579	newamerican,breakfast_brunch	1	Full Bar	1	Validated	1	
Buk Kyung	42.3794757	-71.0955337	2143	Somerville	MA	$$	3.5	239	korean,seafood	1	No	0	Street, Private Lot	1	
The Independent	42.379665	-71.094874	2143	Somerville	MA	$$	3.5	399	bars,newamerican	1	Full Bar	1	Street	1	
Top of the Hub	42.34856613	-71.08228754	2199	Boston	MA	$$$	3.5	2019	newamerican,wine_bars	1	Full Bar	0	Garage	0	
Canary Square	42.31977	-71.11197	2130	Jamaica Plain	MA	$$	3	527	newamerican,breakfast_brunch,burgers	1	Full Bar	1	Street	1	
Casa B Tapas and Cocktail & Rum Bar	42.3797093	-71.0942732	2143	Somerville	MA	$$$	4	258	caribbean,tapasmallplates,cocktailbars	1	Full Bar	0	Street	0	
Piattini	42.34969	-71.08122	2116	Boston	MA	$$	4	740	italian,bars,salad	1	Full Bar	1	Street	1	
The Cheesecake Factory	42.345868	-71.081993	2199	Boston	MA	$$	3	602	newamerican,desserts	1	Full Bar	1	Garage	1	
Atlantic Fish	42.34922179	-71.08112729	2116	Boston	MA	$$$	4	2251	seafood,raw_food,cocktailbars	1	Full Bar	1	Valet, Street	1	
Demos	42.3663721	-71.1821673	2472	Watertown	MA	$	3.5	94	mediterranean,greek	1	Beer & Wine Only	0	Street	1	
Lucca Back Bay	42.3465561	-71.0799281	2116	Boston	MA	$$$	4	300	italian,wine_bars	1	Full Bar	1	Valet, Garage	1	
SRV Boston	42.340918	-71.08161505	2118	Boston	MA	$$$	4.5	306	italian,wine_bars	1	Full Bar	1	Street	0	

Table 3-2: Full Dataset Part 2

WIFI	Ambience	Attire	Noise_Le	atm	bank	bar	beauty_s	bus_stati	cafe	gym	school	White population	Black population	American India	Asian population	Hispanic or Lat	High scho	Graduate o	Unempl	average	review_count_binned
0	Hipster	Casual	3	18	10	19	9	24	16	9	22	0.568382213	0.133807211	0.001627825	0.155647196	0.082232291	0.957	0.449	0.035	20	Meadium to Large Reviews
0	Casual	Casual	2	24	6	34	14	21	21	6	23	0.674716589	0.039477886	0.001157592	0.206929586	0.084863484	0.982	0.471	0.074	20	large Number of Reviews
0	Casual	Casual	2	16	8	17	8	31	15	9	17	0.568382213	0.133807211	0.001627825	0.155647196	0.082232291	0.957	0.449	0.035	60	Very Popular Resturant
0	Casual	Casual	2	15	3	20	3	17	14	3	11	0.674716589	0.039477886	0.001157592	0.206929586	0.084863484	0.982	0.471	0.074	10	Medium No Reviews
0	Casual	Casual	2	4	1	3	1	7	1	3	7	0.76009197	0.026106697	0.000727611	0.067056666	0.049302948	0.961	0.311	0.044	20	Medium No Reviews
0	Classy	Casual	2	18	4	26	11	16	8	8	20	0.674716589	0.039477886	0.001157592	0.206929586	0.084863484	0.982	0.471	0.074	45	Very Popular Resturant
0	Trendy	Casual	3	12	5	17	5	23	13	10	20	0.568382213	0.133807211	0.001627825	0.155647196	0.082232291	0.957	0.449	0.035	45	Very Popular Resturant
0	Trendy	Casual	2	16	4	27	6	14	8	8	24	0.674716589	0.039477886	0.001157592	0.206929586	0.084863484	0.982	0.471	0.074	45	Very Popular Resturant
0	Romantic	Casual	4	3	1	1	3	10	2	1	7	0.663173573	0.060056386	0.001489441	0.143225703	0.059710623	1.001	0.561	0.039	45	Small No Reviews
0	Intimate	Casual	4	2	1	2	8	24	5	5	19	0.568382213	0.133807211	0.001627825	0.155647196	0.082232291	0.957	0.449	0.035	60	Meadium to Large Reviews
0	Intimate	Casual	2	3	1	4	6	4	3	2	21	0.663173573	0.060056386	0.001489441	0.143225703	0.059710623	1.001	0.561	0.039	45	Meadium to Large Reviews
0	Casual	Casual	2	4	1	4	6	4	4	2	20	0.663173573	0.060056386	0.001489441	0.143225703	0.059710623	1.001	0.561	0.039	20	Meadium to Large Reviews
0	Hipster	Casual	2	3	6	4	19	24	3	2	12	0.657335197	0.043996469	0.00110359	0.090494408	0.069121542	0.931	0.339	0.033	45	Meadium to Large Reviews
0	Romantic	Casual	2	8	2	9	8	33	6	7	19	0.568382213	0.133807211	0.001627825	0.155647196	0.082232291	0.957	0.449	0.035	45	Very Popular Resturant
0	Upscale	Dressy	2	13	2	23	28	14	20	12	27	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	45	Meadium to Large Reviews
0	Casual	Casual	2	10	2	13	7	28	6	4	13	0.568382213	0.133807211	0.001627825	0.155647196	0.082232291	0.957	0.449	0.035	20	Meadium to Large Reviews
0	Casual	Casual	2	19	4	12	8	18	24	10	30	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	20	Medium No Reviews
0	Hipster, Casual, Trend	Casual	2	13	4	5	12	18	8	7	16	0.51772693	0.156399296	0.001257229	0.118783002	0.056675886	0.952	0.433	0.049	20	Small No Reviews
0	Casual	Casual	2	28	10	22	28	18	38	14	31	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	20	Very Popular Resturant
0	Casual	Casual	2	13	4	6	12	19	8	6	16	0.51772693	0.156399296	0.001257229	0.118783002	0.056675886	0.952	0.433	0.049	10	Meadium to Large Reviews
0	Casual	Casual	2	38	15	31	48	16	43	19	31	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	20	large Number of Reviews
0	Casual	Casual	2	18	3	7	7	11	20	2	22	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	10	Meadium to Large Reviews
0	Classy, Upscale	Dressy	2	38	16	36	57	16	42	19	29	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	60	large Number of Reviews
0	Trendy	Casual	2	32	16	41	60	15	44	19	30	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	45	Medium No Reviews
0	Trendy	Dressy	2	33	16	40	60	14	46	19	29	0.626382226	0.074410534	0.001394985	0.143649416	0.095131163	0.915	0.359	0.09	45	Meadium to Large Reviews
0	Casual	Casual	2	7	2	12	10	34	8	6	14	0.657335197	0.043996469	0.00110359	0.090494408	0.069121542	0.931	0.339	0.033	20	Medium No Reviews
0	Casual	Casual	2	7	2	12	10	34	9	6	15	0.657335197	0.043996469	0.00110359	0.090494408	0.069121542	0.931	0.339	0.033	20	Meadium to Large Reviews
0	Classy, Upscale	Dressy	2	38	17	39	60	14	43	20	30	0.731611894	0.011737089	0.000782473	0.111893584	0.027386541	1.02	0.452	0.035	45	Very Popular Resturant
0	Casual	Casual	2	7	2	5	9	13	1	1	8	0.506757943	0.106647609	0.001479475	0.045688207	0.200180546	0.941	0.358	0.045	20	Meadium to Large Reviews
0	Casual	Casual	2	7	2	11	8	34	9	6	15	0.657335197	0.043996469	0.00110359	0.090494408	0.069121542	0.931	0.339	0.033	45	Medium No Reviews
0	Intimate	Casual	2	40	21	39	60	14	39	24	33	0.628208974	0.052929328	0.001028586	0.132601894	0.048643552	0.949	0.441	0.041	20	large Number of Reviews
0	Casual	Casual	2	38	16	38	60	13	50	17	22	0.731611894	0.011737089	0.000782473	0.111893584	0.027386541	1.02	0.452	0.035	20	large Number of Reviews
0	Classy	Casual	2	43	23	39	60	16	43	23	34	0.628208974	0.052929328	0.001028586	0.132601894	0.048643552	0.949	0.441	0.041	45	Very Popular Resturant
0	Casual	Casual	2	13	9	3	18	16	4	6	14	0.76009197	0.026106697	0.000727611	0.067056666	0.049302948	0.961	0.311	0.044	10	Small No Reviews
0	Romantic	Dressy	2	42	18	39	60	17	40	17	25	0.628208974	0.052929328	0.001028586	0.132601894	0.048643552	0.949	0.441	0.041	45	Medium No Reviews
0	Trendy	Dressy	2	10	2	11	24	25	15	3	16	0.432965075	0.181004989	0.001781896	0.128724163	0.176478974	0.849	0.32	0.047	45	Meadium to Large Reviews

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.

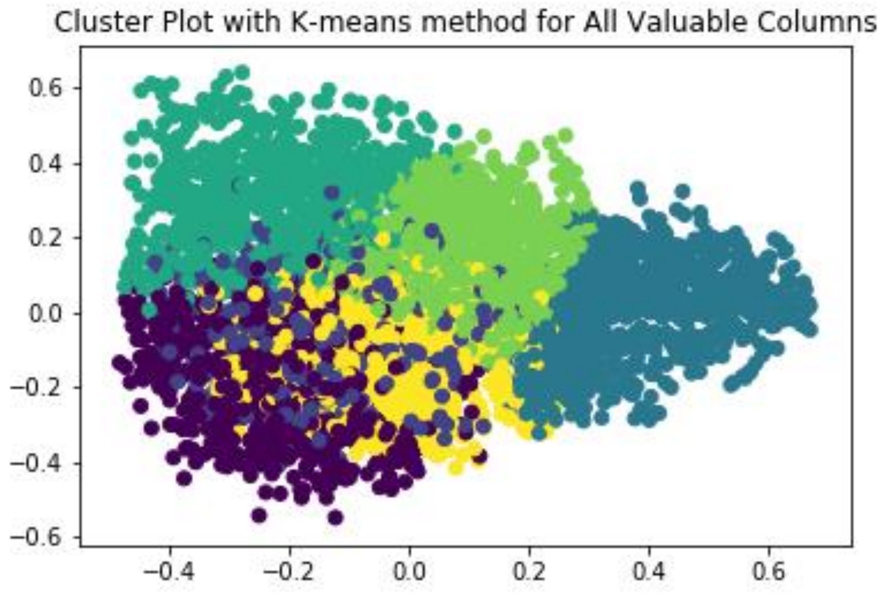

Figure 3-1: 2D graph for k-means method for All Valuable Columns with n=6

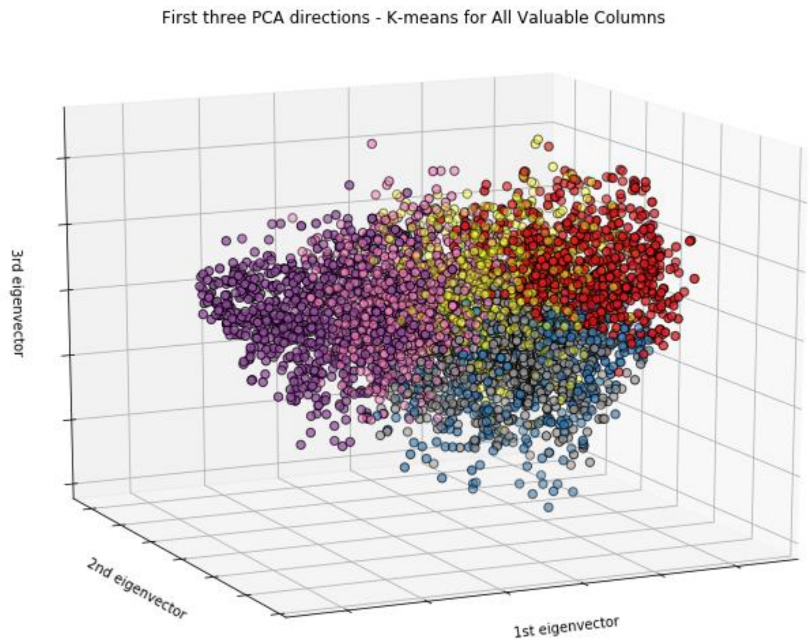

Figure 3-2: 3D graph for k-means method for All Valuable Columns with n=6

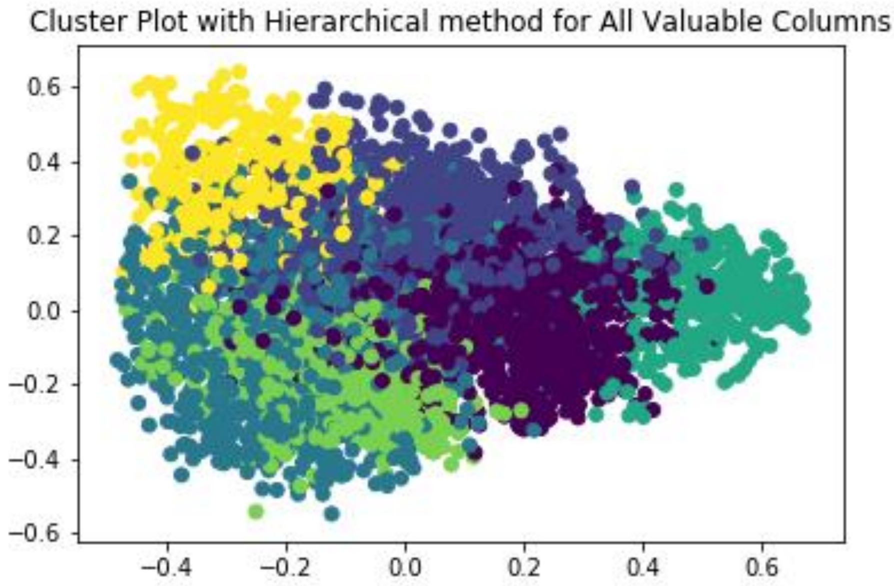

Figure 3-3: 2D graph for hierarchical method for All Valuable Columns with n=6

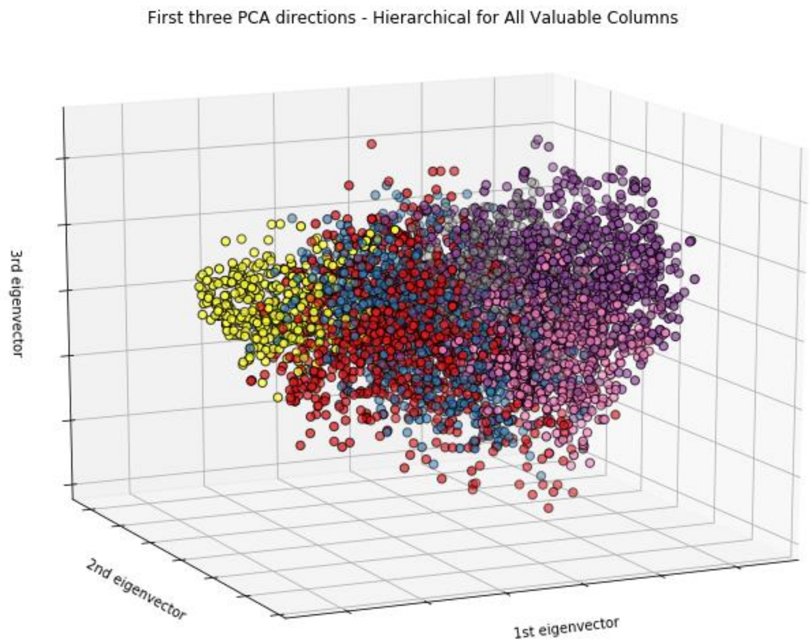

Figure 3-4: 3D graph for hierarchical method for All Valuable Columns with n=6

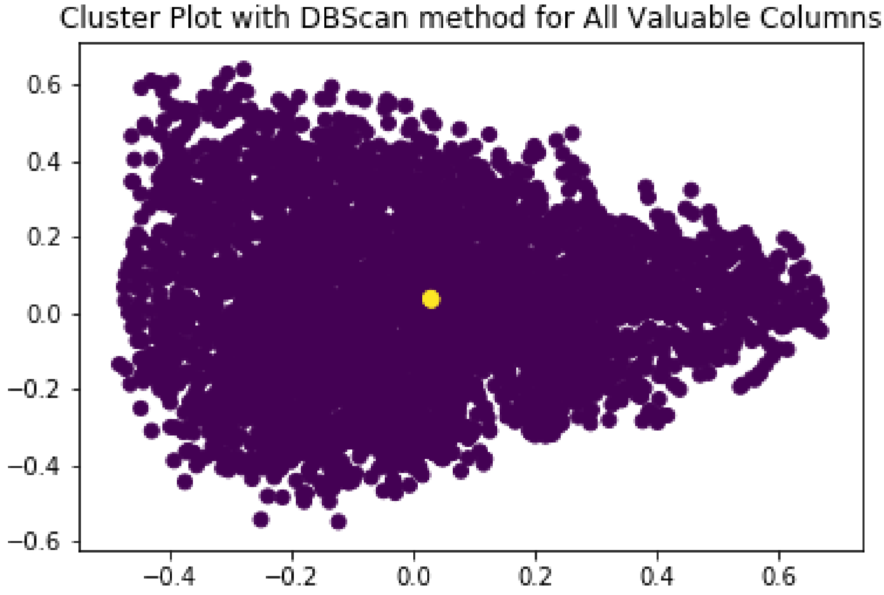

Figure 3-5: 2D graph for dbscan method for All Valuable Columns

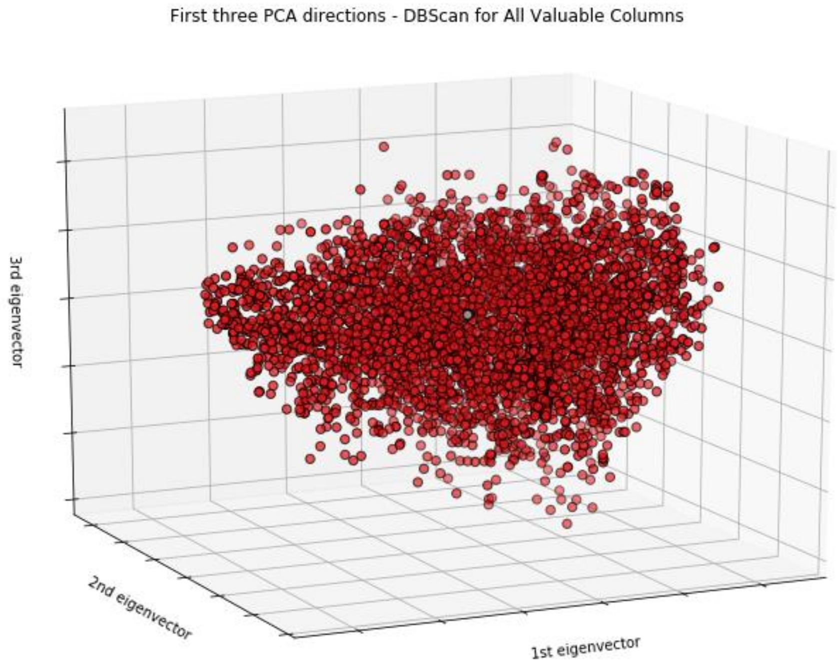

Figure 3-6: 3D graph for dbscan method for All Valuable Columns

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.

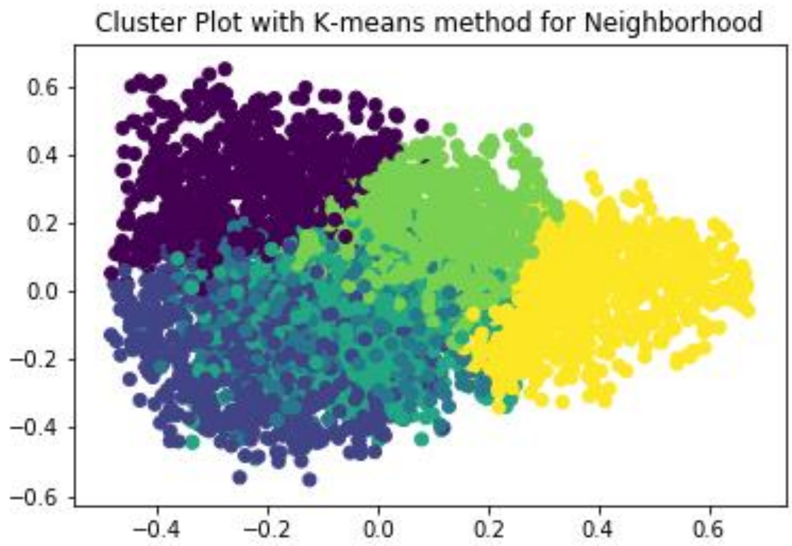

Figure 3-7: 2D graph for k-means method for Neighbourhood with n=6

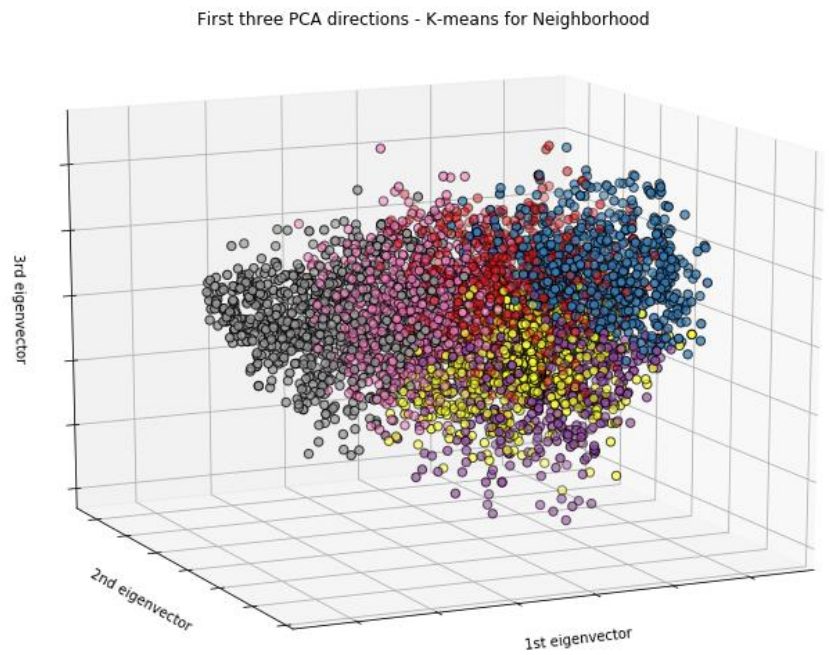

Figure 3-8: 3D graph for k-means method for Neighbourhood with n=6

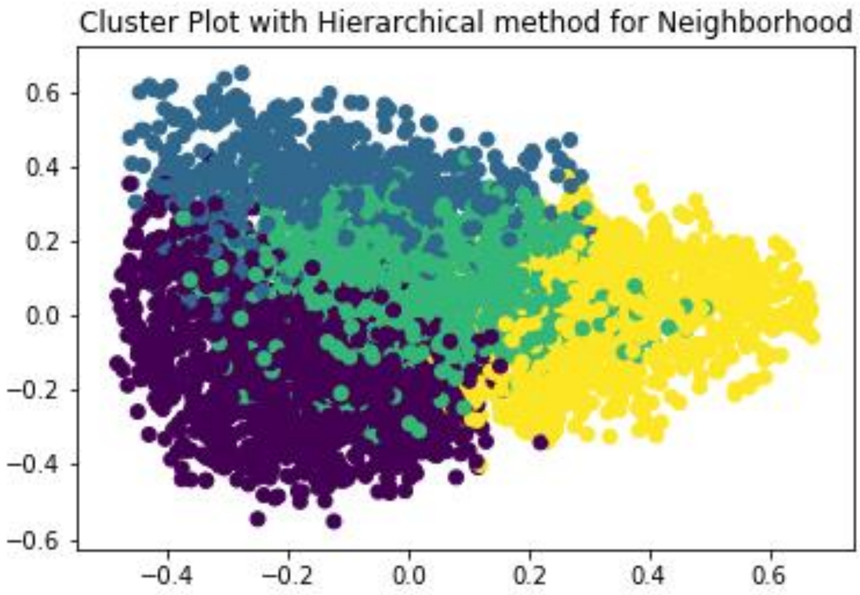

Figure 3-9: 2D graph for hierarchical method for Neighbourhood with n=4

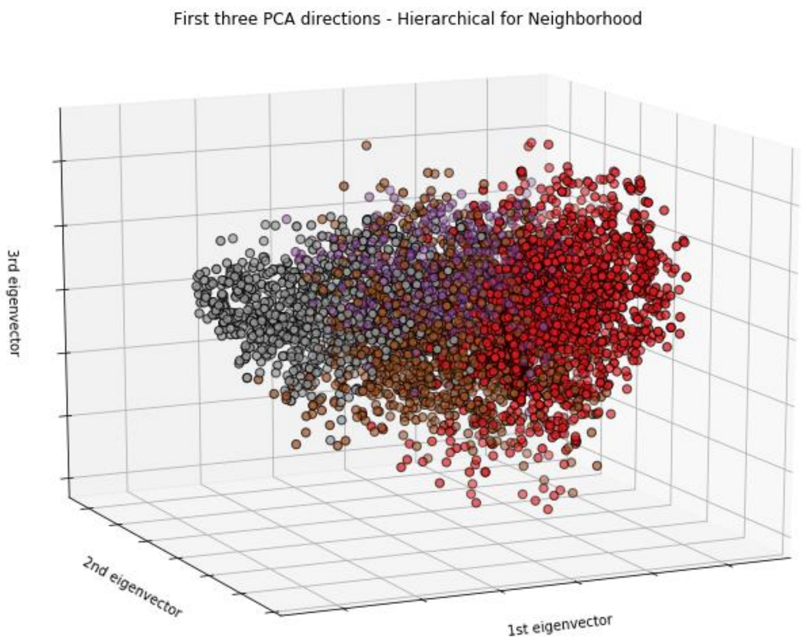

Figure 3-10: 3D graph for hierarchical method for Neighbourhood with n=4

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.

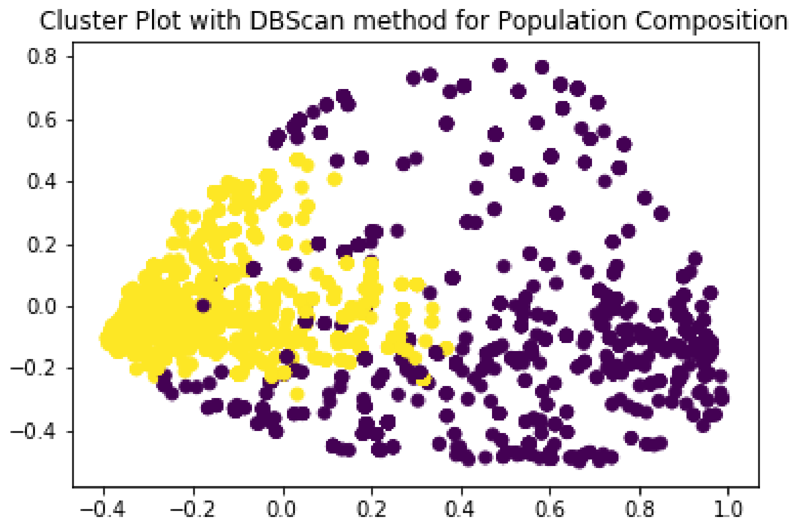

Figure 3-11: 2D graph for dbscan method for Neighbourhood

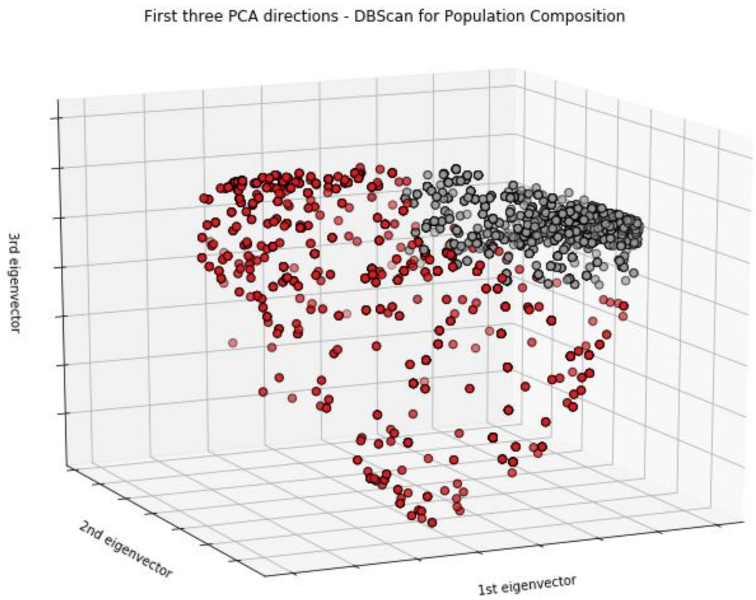

Figure 3-12: 3D graph for dbscan method for Neighbourhood

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

	Rule	Transaction	Antecedent	Consequent	Support	Confidence	Lift
0	{10 -> Low_Noise}	['10', 'Low_Noise']	10	Low_Noise	0.04928891	0.85185185	0.98303857
1	{3.0 -> 20}	['3.0', '20']	3	20	0.09877265	0.83663366	1.1339954
2	{3.5 -> 20}	['3.5', '20']	3.5	20	0.23241769	0.77166882	1.04594034
3	{4.0 -> 20}	['4.0', '20']	4	20	0.29417495	0.69393382	0.94057627
4	{4.5 -> 20}	['4.5', '20']	4.5	20	0.06993961	0.64336918	0.8720396
5	{CA -> 20}	['CA', '20']	CA	20	0.31131892	0.73878872	1.00137378
6	{Few_atm -> 20}	['Few_atm', '20']	Few_atm	20	0.21274109	0.77667141	1.05272098
7	{Few_bank -> 20}	['Few_bank', '20']	Few_bank	20	0.20553283	0.78555473	1.06476166
8	{Few_bar -> 20}	['Few_bar', '20']	Few_bar	20	0.20670173	0.76111908	1.03164094
9	{Few_beauty_salon -> 20}	['Few_beauty_salon', '20	Few_beauty_s	20	0.20358465	0.75724638	1.02639178
10	{Few_bus_station -> 20}	['Few_bus_station', '20']	Few_bus_stati	20	0.2183908	0.78556412	1.06477439
11	{Few_cafe -> 20}	['Few_cafe', '20']	Few_cafe	20	0.22832651	0.79296346	1.07480366
12	{Few_gym -> 20}	['Few_gym', '20']	Few_gym	20	0.22969024	0.79340511	1.07540228
13	{Few_school -> 20}	['Few_school', '20']	Few_school	20	0.20066238	0.77912254	1.05604331
14	{High_Noise -> 20}	['High_Noise', '20']	High_Noise	20	0.04110657	0.63746224	0.86403318
15	{20 -> Low_Noise}	['20', 'Low_Noise']	20	Low_Noise	0.64114553	0.86902561	1.00285712
16	{Low_Noise -> 20}	['Low_Noise', '20']	Low_Noise	20	0.64114553	0.73988309	1.00285712
17	{Many_atm -> 20}	['Many_atm', '20']	Many_atm	20	0.16754335	0.77060932	1.04450426
18	{Many_bank -> 20}	['Many_bank', '20']	Many_bank	20	0.10617573	0.74759945	1.01331608
19	{Many_bar -> 20}	['Many_bar', '20']	Many_bar	20	0.15098383	0.76656775	1.03902622
20	{Many_beauty_salon -> 20}	['Many_beauty_salon', '2	Many_beauty_	20	0.16072472	0.76388889	1.03539521
21	{Many_bus_station -> 20}	['Many_bus_station', '20	Many_bus_sta	20	0.14806156	0.72796935	0.98670892
22	{Many_cafe -> 20}	['Many_cafe', '20']	Many_cafe	20	0.14591857	0.75733064	1.02650598
23	{Many_gym -> 20}	['Many_gym', '20']	Many_gym	20	0.13578804	0.72153209	0.97798369
24	{Many_school -> 20}	['Many_school', '20']	Many_school	20	0.15020456	0.74781765	1.01361183
25	{Meadium to Large Reviews -> 20}	['Meadium to Large Revi	Meadium to La	20	0.20494837	0.78041543	1.05779572
26	{Median_Noise -> 20}	['Median_Noise', '20']	Median_Noise	20	0.05552309	0.80508475	1.09123317
27	{Medium No Reviews -> 20}	['Medium No Reviews', '	Medium No Re	20	0.16773816	0.74674761	1.01216148
28	{Moderate_atm -> 20}	['Moderate_atm', '20']	Moderate_atm	20	0.20358465	0.76	1.03012411
29	{Moderate_bank -> 20}	['Moderate_bank', '20']	Moderate_ban	20	0.27683616	0.7710255	1.04506837
30	{Moderate_bar -> 20}	['Moderate_bar', '20']	Moderate_bar	20	0.22423534	0.78781656	1.06782742
31	{Moderate_beauty_salon -> 20}	['Moderate_beauty_salo	Moderate_bea	20	0.20981882	0.76437189	1.03604989
32	{Moderate_bus_station -> 20}	['Moderate_bus_station'	Moderate_bus	20	0.21274109	0.75466482	1.02289267
33	{Moderate_cafe -> 20}	['Moderate_cafe', '20']	Moderate_cafe	20	0.21235145	0.76011158	1.03027534
34	{Moderate_gym -> 20}	['Moderate_gym', '20']	Moderate_gym	20	0.22248198	0.76644295	1.03885706
35	{Moderate_school -> 20}	['Moderate_school', '20']	Moderate_sch	20	0.2238457	0.76143141	1.03206428
36	{NY -> 20}	['NY', '20']	NY	20	0.1768946	0.7189232	0.97444753

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

Rule	Transaction	Antecedent	Consequent	Support	Confidence	Lift
{bars,45 -> 4.0}	['bars,45', '4.0']	bars,45	4	0.03097604	0.51456311	1.21381086
{CA,Many_gym -> 4.0}	['CA,Many_gym', '4.0']	CA,Many_gym	4	0.04948373	0.5059761	1.19355482
{CA,Super many_beauty_salon -> 4.0}	['CA,Super many_beauty_salon', '4.0']	CA,Super many_beauty_salon	4	0.04169102	0.50234742	1.18499508
{CA,Very Popular Resturant -> 4.0}	['CA,Very Popular Resturant', '4.0']	CA,Very Popular Resturant	4	0.05727645	0.5505618	1.29872873
{CA,italian -> 4.0}	['CA,italian', '4.0']	CA,italian	4	0.04324956	0.52112676	1.22929396
{Low_Noise,Very Popular Resturant -> 4.0}	['Low_Noise,Very Popular Resturant', '4.0']	Low_Noise,Very Popular Resturant	4	0.07519969	0.58751903	1.3859077
{Low_Noise,large Number of Reviews -> 4.	['Low_Noise,large Number of Reviews', '4.0']	Low_Noise,large Number of Reviews	4	0.07714787	0.53297443	1.25724161
{Low_Noise,mediterranean -> 4.0}	['Low_Noise,mediterranean', '4.0']	Low_Noise,mediterranean	4	0.03915839	0.51015228	1.2034061
{Many_bus_station,Super many_gym -> 4.0	['Many_bus_station,Super many_gym', '4.0']	Many_bus_station,Super many_gym	4	0.03039158	0.5	1.17945772
{Many_gym,Many_cafe -> 4.0}	['Many_gym,Many_cafe', '4.0']	Many_gym,Many_cafe	4	0.03643094	0.50134048	1.18261981
{Moderate_bar,italian -> 4.0}	['Moderate_bar,italian', '4.0']	Moderate_bar,italian	4	0.03233976	0.5030303	1.18660595
{Moderate_beauty_salon,italian -> 4.0}	['Moderate_beauty_salon,italian', '4.0']	Moderate_beauty_salon,italian	4	0.03156049	0.52941176	1.24883759
{Super many_bar,Very Popular Resturant ->	['Super many_bar,Very Popular Resturant', '4.0']	Super many_bar,Very Popular Restura	4	0.03350867	0.56953642	1.34348826
{Super many_cafe,Very Popular Resturant -	['Super many_cafe,Very Popular Resturant', '4.0']	Super many_cafe,Very Popular Restur	4	0.03311903	0.56291391	1.32786631
{newmexican,Very Popular Resturant -> 4.0	['newmexican,Very Popular Resturant', '4.0']	newmexican,Very Popular Resturant	4	0.03915839	0.54324324	1.28146487
{newmexican,breakfast_brunch -> 4.0}	['newmexican,breakfast_brunch', '4.0']	newmexican,breakfast_brunch	4	0.03039158	0.50322581	1.18706713
{CA,Very Popular Resturant,20 -> 4.0}	['CA,Very Popular Resturant,20', '4.0']	CA,Very Popular Resturant,20	4	0.0360413	0.5393586	1.27230133
{CA,italian,20 -> 4.0}	['CA,italian,20', '4.0']	CA,italian,20	4	0.03195013	0.51572327	1.21654759
{Low_Noise,Very Popular Resturant,20 -> 4.	['Low_Noise,Very Popular Resturant,20', '4.0']	Low_Noise,Very Popular Resturant,20	4	0.04714592	0.59605911	1.40605305
{Low_Noise,20,breakfast_brunch -> 4.0}	['Low_Noise,20,breakfast_brunch', '4.0']	Low_Noise,20,breakfast_brunch	4	0.03701539	0.50131926	1.18256975
{Low_Noise,large Number of Reviews,20 ->	['Low_Noise,large Number of Reviews,20', '4.0']	Low_Noise,large Number of Reviews,2	4	0.05318527	0.51412429	1.21277574
{Low_Noise,CA,45 -> 4.0}	['Low_Noise,CA,45', '4.0']	Low_Noise,CA,45	4	0.03097604	0.55017301	1.29781161
{Low_Noise,Super many_bar,45 -> 4.0}	['Low_Noise,Super many_bar,45', '4.0']	Low_Noise,Super many_bar,45	4	0.03078122	0.52491694	1.23823468
{Low_Noise,CA,Many_beauty_salon -> 4.0}	['Low_Noise,CA,Many_beauty_salon', '4.0']	Low_Noise,CA,Many_beauty_salon	4	0.04324956	0.5248227	1.23801236
{Low_Noise,CA,Many_bus_station -> 4.0}	['Low_Noise,CA,Many_bus_station', '4.0']	Low_Noise,CA,Many_bus_station	4	0.03039158	0.51655629	1.21851261
{Low_Noise,CA,Many_cafe -> 4.0}	['Low_Noise,CA,Many_cafe', '4.0']	Low_Noise,CA,Many_cafe	4	0.03759984	0.51742627	1.22056483
{Low_Noise,CA,Many_gym -> 4.0}	['Low_Noise,CA,Many_gym', '4.0']	Low_Noise,CA,Many_gym	4	0.04441847	0.52413793	1.23639706
{Low_Noise,CA,Super many_beauty_salon -	['Low_Noise,CA,Super many_beauty_salon', '4.0']	Low_Noise,CA,Super many_beauty_sal	4	0.0360413	0.50824176	1.19889933
{Low_Noise,CA,Very Popular Resturant -> 4.	['Low_Noise,CA,Very Popular Resturant', '4.0']	Low_Noise,CA,Very Popular Resturant	4	0.05143191	0.57768053	1.36269951
{Low_Noise,CA,italian -> 4.0}	['Low_Noise,CA,italian', '4.0']	Low_Noise,CA,italian	4	0.03896357	0.52770449	1.24481026
{Low_Noise,CA,large Number of Reviews ->	['Low_Noise,CA,large Number of Reviews', '4.0']	Low_Noise,CA,large Number of Review	4	0.04169102	0.50831354	1.19906866
{Low_Noise,Many_gym,Many_beauty_salor	['Low_Noise,Many_gym,Many_beauty_salon', '4.0']	Low_Noise,Many_gym,Many_beauty_s	4	0.03039158	0.50649351	1.19477535
{Low_Noise,Many_bus_station,Super many	['Low_Noise,Many_bus_station,Super many_school', '4.0']	Low_Noise,Many_bus_station,Super m	4	0.03000195	0.50657895	1.1949769
{Low_Noise,Many_gym,Many_cafe -> 4.0}	['Low_Noise,Many_gym,Many_cafe', '4.0']	Low_Noise,Many_gym,Many_cafe	4	0.0327294	0.51851852	1.22314134
{Low_Noise,newmexican,Very Popular Rest	['Low_Noise,newmexican,Very Popular Resturant', '4.0']	Low_Noise,newmexican,Very Popular	4	0.03409312	0.56451613	1.33164581
{Low_Noise,CA,Many_beauty_salon,20 -> 4.	['Low_Noise,CA,Many_beauty_salon,20', '4.0']	Low_Noise,CA,Many_beauty_salon,20	4	0.03156049	0.51104101	1.20550253
{Low_Noise,CA,20,Many_gym -> 4.0}	['Low_Noise,CA,20,Many_gym', '4.0']	Low_Noise,CA,20,Many_gym	4	0.03156049	0.51757188	1.22090831

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 6-1 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.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_count   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.021
Method:                 Least Squares   F-statistic:                     38.37
Date:                Sun, 11 Nov 2018   Prob (F-statistic):           1.63e-24
Time:                        14:01:12   Log-Likelihood:                -40764.
No. Observations:                5132   AIC:                         8.154e+04
Df Residuals:                    5128   BIC:                         8.156e+04
Df Model:                           3
Covariance Type:            nonrobust
=================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept     385.1919     39.558      9.737      0.000     307.641     462.743
price[T.$$]   158.5450     41.081      3.859      0.000      78.009     239.081
price[T.$$$]  357.4288     45.627      7.834      0.000     267.980     446.877
price[T.$$$$] 494.5547     68.288      7.242      0.000     360.681     628.429
==============================================================================
Omnibus:                     5386.986   Durbin-Watson:                   1.653
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           754645.632
Skew:                           4.956   Prob(JB):                         0.00
Kurtosis:                      61.574   Cond. No.                         11.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                sum_sq      df          F        PR(>F)
price     5.349843e+07     3.0  38.369678  1.630284e-24
Residual  2.383305e+09  5128.0        NaN           NaN
```

Figure 5-1: ANOVA Test Result

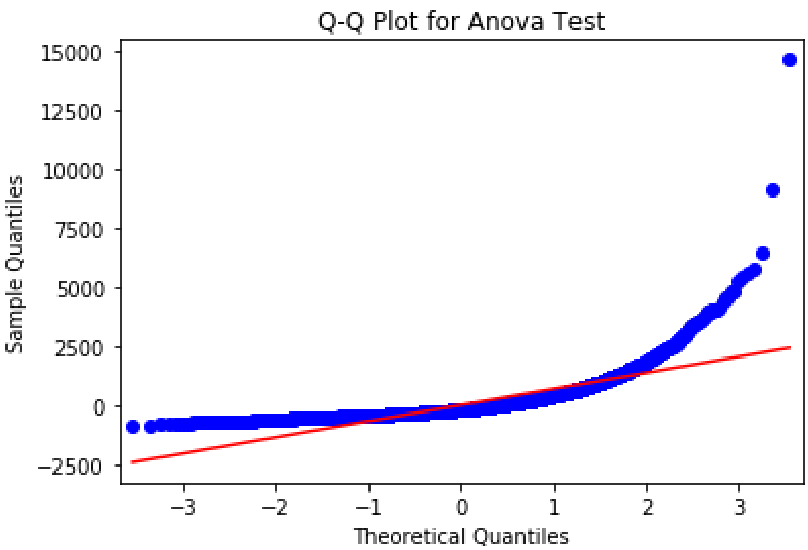

Figure 5-2: ANOVA Q-Q Plot

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.

```
T-Test of price group '$' against '$$','$$$', '$$$$'
Ttest_indResult(statistic=-4.02788383645965, pvalue=5.730479891454033e-05)
Ttest_indResult(statistic=-7.472260577635161, pvalue=1.5187064468003344e-13)
Ttest_indResult(statistic=-7.885496587711298, pvalue=2.4428811832450818e-14)
```

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.

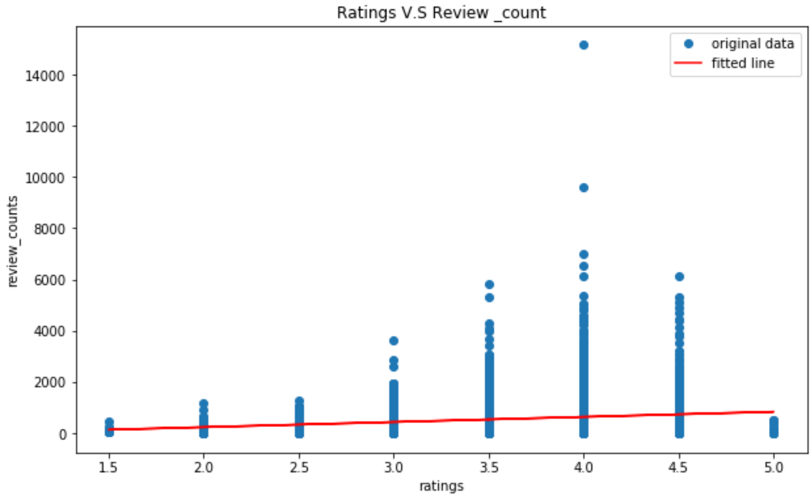

Figure 5-4: Linear regression Model Result

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.

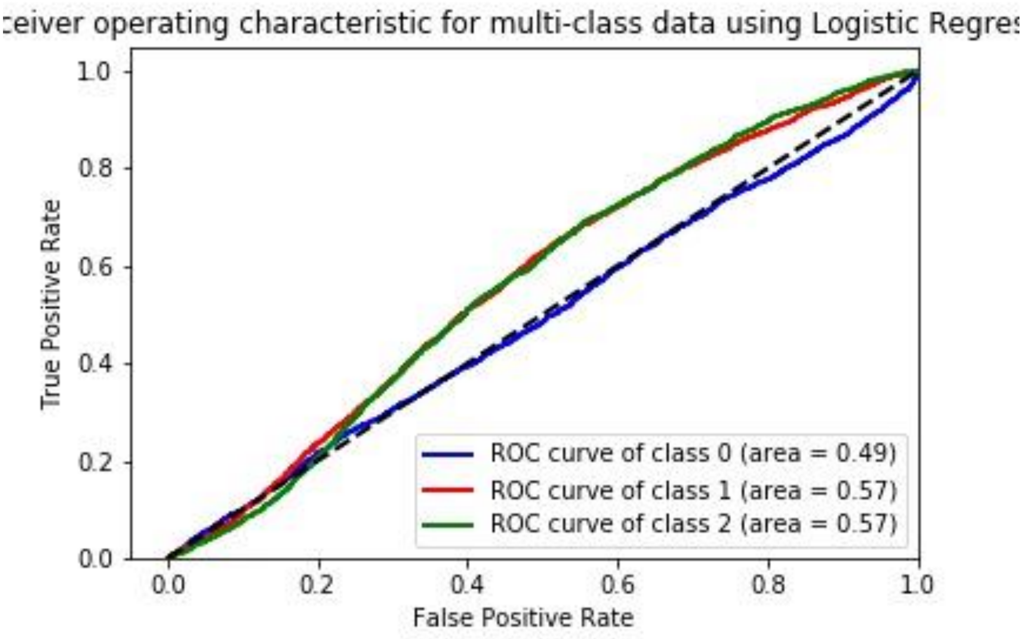

Figure 5-5: ROC for Logistic Regression

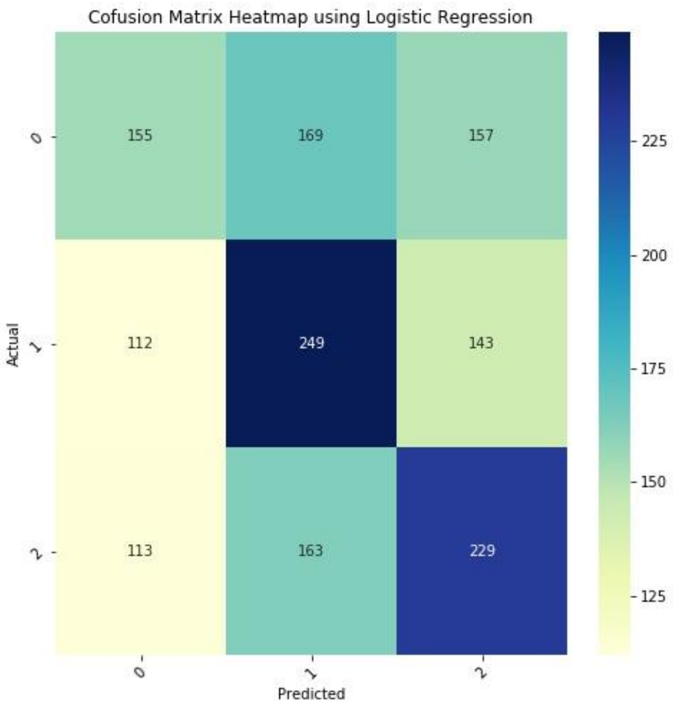

Figure 5-6: Confusion Matrix Heatmap for Logistic Regression

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.

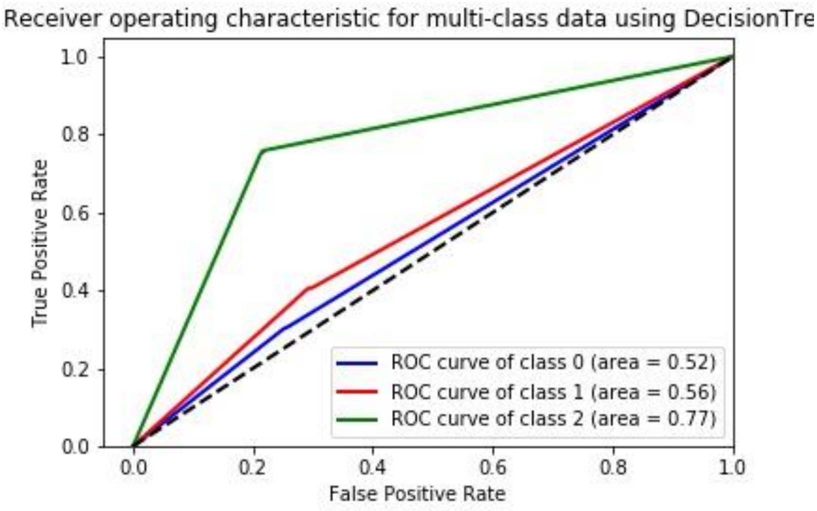

Figure 5-7: ROC Curve for Decision Tree

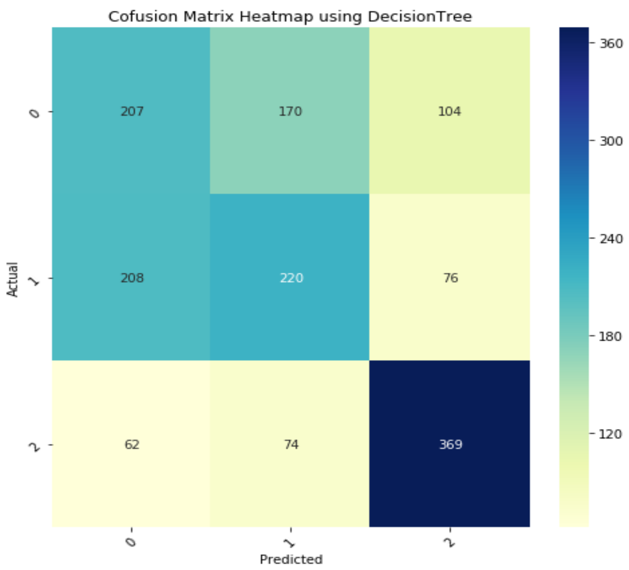

Figure 5-8: Confusion Matrix Heatmap for Decision Tree

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.

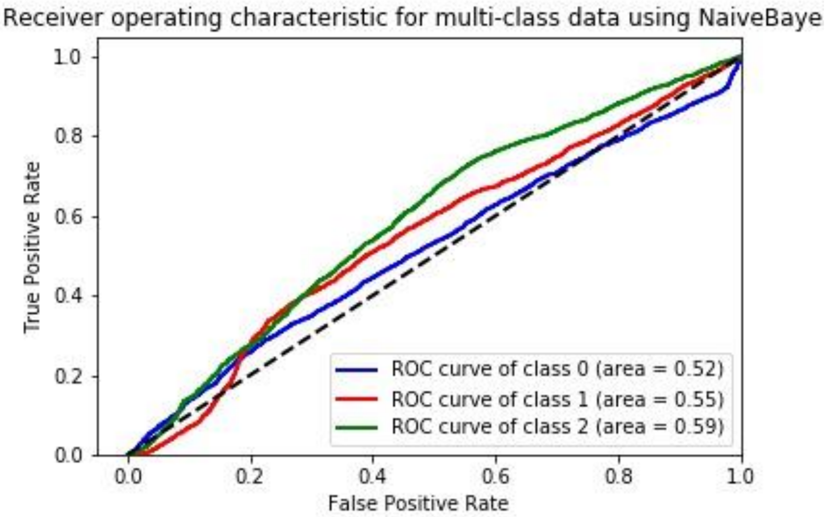

Figure 5-9: ROC curve for Naive Bayes Classifier

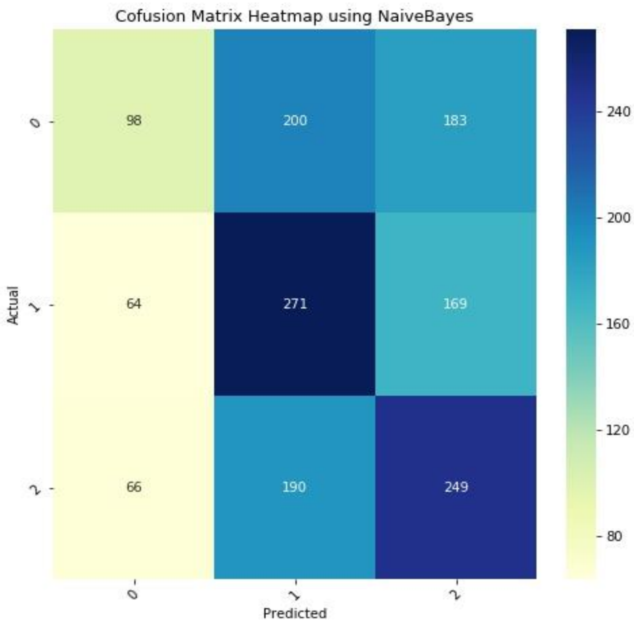

Figure 5-10: Confusion Matrix Heatmap of Naive Bayes

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.

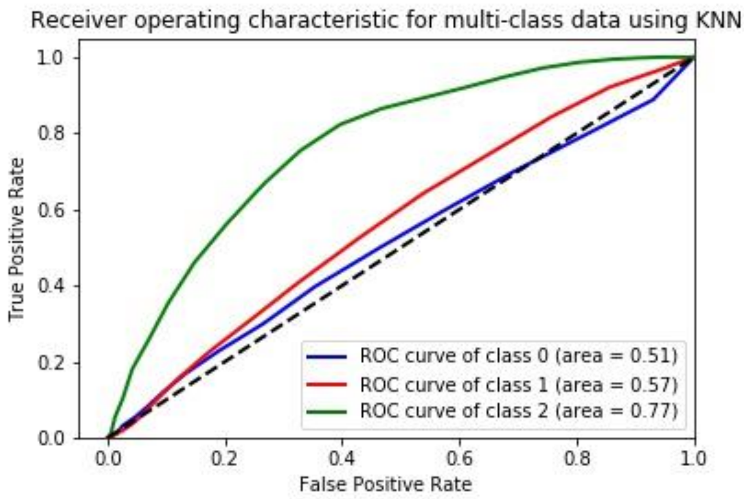

Figure 5-11: ROC for KNN

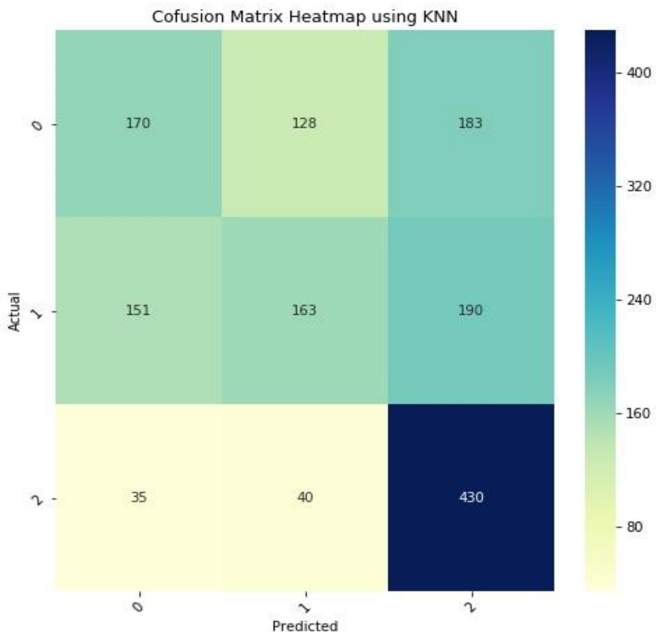

Figure 5-12: Confusion Matrix Heatmap for KNN

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.

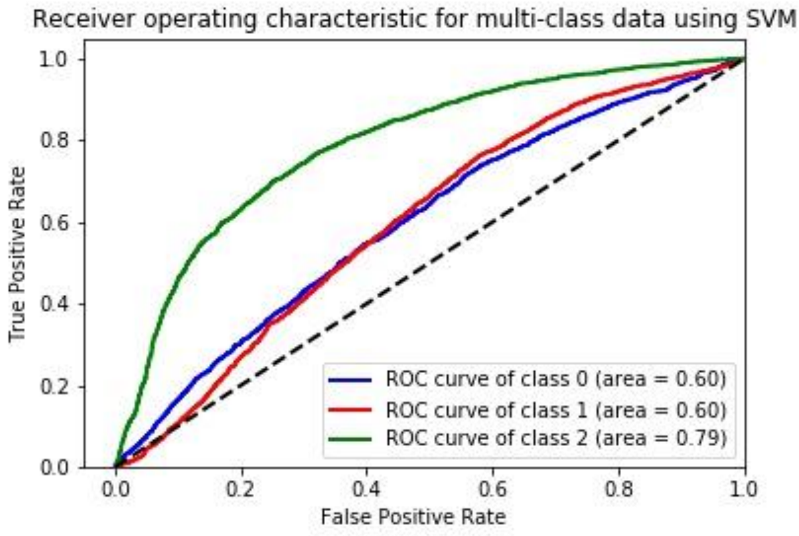

Figure 5-13: ROC for SVM

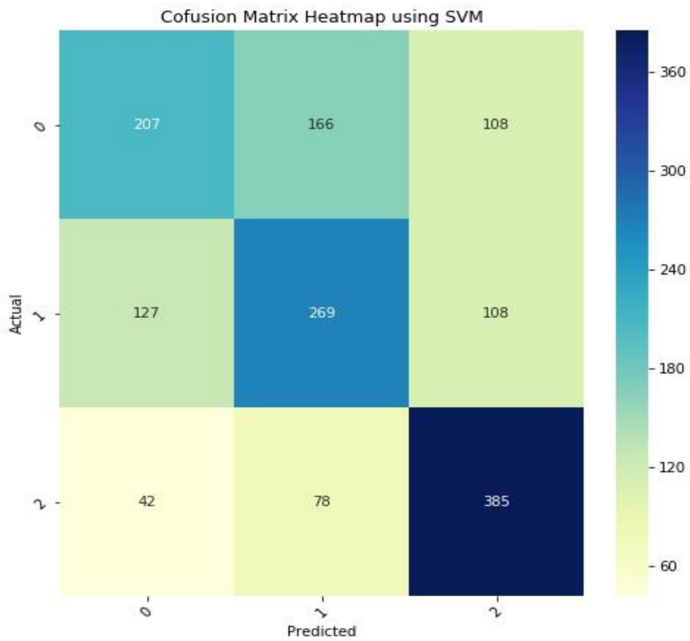

Figure 5-14: Confusion Matrix Heatmap for SVM

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.

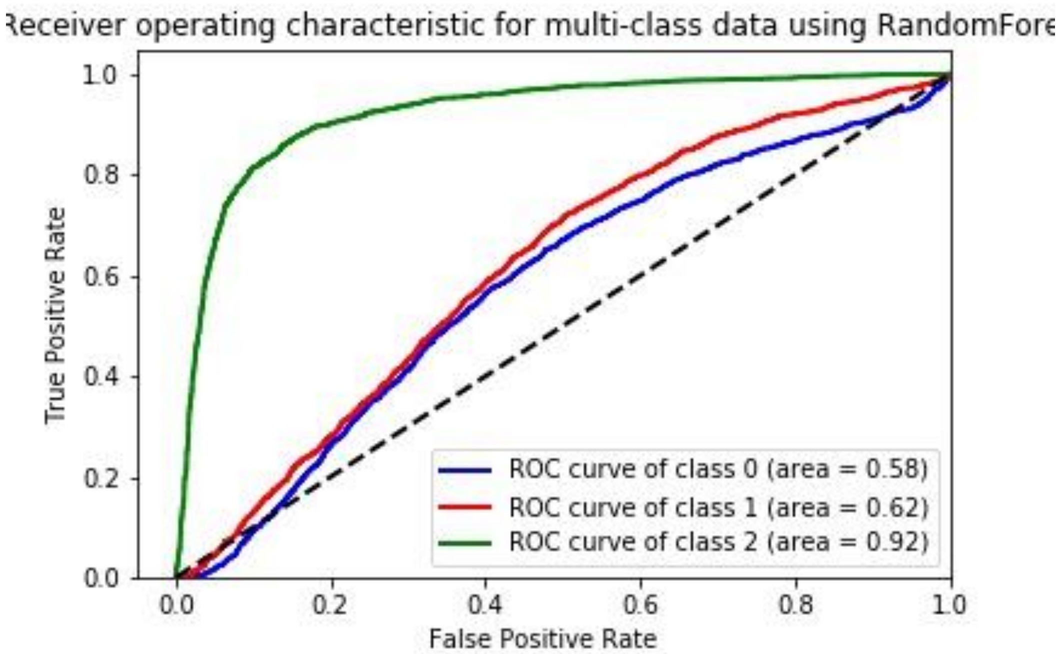

Figure 5-15: ROC for random Forest

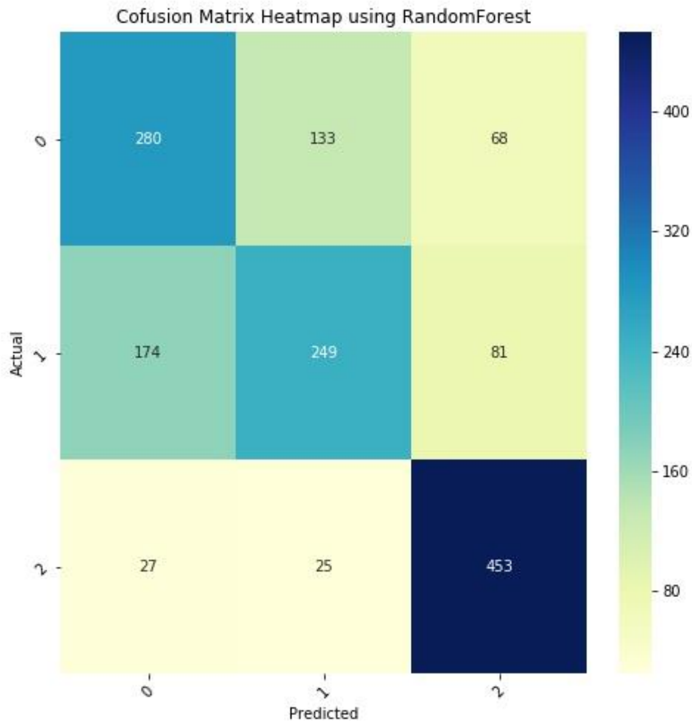

Figure 5-16: Confusion Matrix Heatmap for Random Forest

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.