Applied Data Science Capstone

Predicting customer ratings of restaurants in Philadelphia, USA

By: Amos Johnson

22 Jan 2021

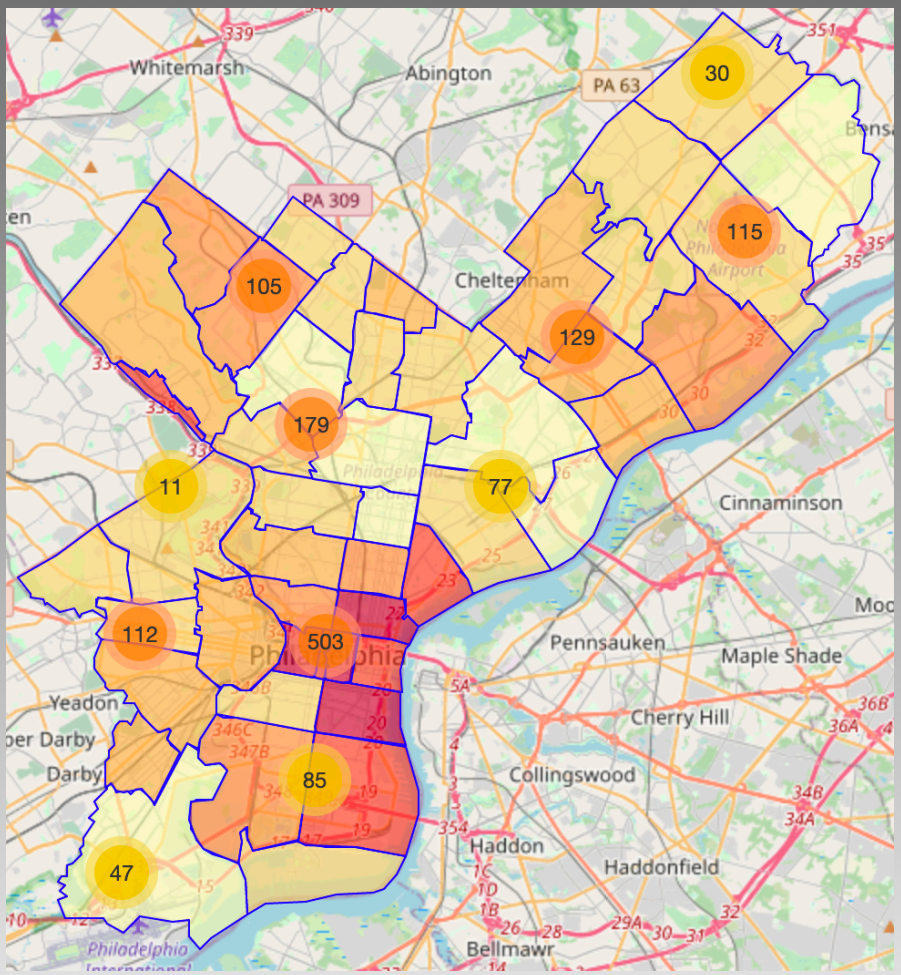


Table of Contents

1. Introduction

The US restaurant industry serves an enormous market size of 800 billion USD in 2017 and is one of the countries top employers. The abundance of restaurants make the market a particularly competitive market with some sources claiming that up to 90% of restaurants fail within their first five years. While this statistic has been disputed, the restaurant business is generally agreed to be a tough business environment riddled with risk. Intuitively, restaurants are among the businesses most affected by the COVID-19 pandemic, with restaurants closures skyrocketing. A model that is able to leverage readily available data to generate a prediction of customer satisfaction rating may offer new insights as to what types of restaurants tend to do well in the eyes of the customer. Such a model is assumed to be of interest to both aspiring restauranteurs seeking to increase their prospects of success by studying what parameters tend to be appreciated by customers, as well as to existing restaurants aiming to revitalize or improve their business through data driven business venture recommendations.

Thus, the purpose of this project is to create a machine learning model able to predict the rating of a restaurant based on data gathered through a publicly usable API.

2. Data

The subject of my study was the restaurant scene in Philadelphia PA, USA where I am currently residing. Data was primarily obtained using the foursquare API endpoints *explore venues* and *venue details*. Data was then wrangled, which in this project involved completing the dataset by merging data with prior datasets, dropping / imputing missing values and transforming categorical data into numeric variables.

2.1 Zip Code Data

A geodataset was used to obtain the zip codes and their geographical borders, the dataset *philly_df* was constructed.

TABLE 1: PHILLY_DF

Zip Code	Latitude	Longitude
19102	39.952719	-75.165150
19103	39.954636	-75.175518
19104	39.958853	-75.196545
19106	39.952621	-75.147081
19107	39.952426	-75.15922

The latitude and longitude of each zip code was inferred as the mean of the geographical border points given through the geodataset.

2.2 General Venue Data

The coordinates were used as arguments for the foursquare endpoint *explore* to obtain the venue IDs and other general information for popular restaurants in each zip code. This information was used to generate the second data frame, *philly_venues_df*.

TABLE 2: PHILLY_VENUES_DF

Zip Code	ID	Name	Category	Lat	Long
19102	4ab2ac0bf964a520d66b20e3	Del Frisco's Double Eagle Steak House	Steakhouse	39.9510	-75.1655
19102	4a97fea2f964a5200a2a20e3	Chick-fil-A	Fast Food Restaurant	39.9524	-75.1682
19102	4b09d9def964a520b41e23e3	Su Xing House	Chinese Restaurant	39.9505	-75.1663
19102	565f615b498eeef346958da1	Tredici Enoteca	Mediterranean Restaurant	39.9498	-75.1620
19102	40b28c80f964a520daf81ee3	Alma De Cuba	Cuban Restaurant	39.9500	-75.1685

2.3 Detailed Venue Data

The restaurant IDs were used to call the endpoint *venue details* which yielded all the data used in the following sections of the project (note that new variables were added, however they were generated from the data in this dataset). The dataset, *venues_full_df,* contained the general information along with additional attributes on each dataset. The primary difficulty of this stage was that *venue details* is offered as a premium endpoint, subsequently I was limited to 500 daily calls to this endpoint. Since I deemed it necessary to obtain over 1000 observations, the final dataset of venue details was compiled in increments over three days, each day making the allotted 500 daily calls.

After significant wrangling of the data, the dataset took its final form, the first columns of which can be seen below.

TABLE 3: VENUES_FULL_DF

ID	Name	Zip	Lat	Long	Category	Distance	Rating	Price Tier
4ab2ac0bf964a520d66b20e3	Del Frisco's Double Eagle Steak House	19102	39.9510	-75.1655	Steakhouse	0.00234	8.7	4.0
4a281e64f964a520f4941fe3	Oyster House	19102	39.9504	-75.1665	Seafood Restaurant	0.00353	9.3	3.0
56cc831ccd10c5927d30dc1d	Snap Custom Pizza	19102	39.9504	-75.1662	Pizza Place	0.00322	8.8	1.0
4af2d4cef964a520a9e821e3	The Capital Grille	19107	39.9507	-75.1639	American Restaurant	0.00170	8.5	4.0
4a4268fdf964a520d4a51fe3	Fogo De Chão	19107	39.9509	-75.1630	Churrascaria	0.00160	8.8	4.0

The remaining 108 columns were omitted from this report (it obviously would not fit on a page). They contained categorical attribute (Wi-Fi, Alcohol, Private Parking Lot etc.) and category dummy variables (Fast Food, American Restaurant, Sushi Restaurant etc).

2.4 Issues with data

It should be noted that many of the columns were sparsely populated (as is natural with dummy variables), but in many cases it was impossible to determine whether the restaurant truly did not have an attribute (e.g., Wi-Fi) or if they had simply neglected to indicate that information. In some cases, such as with the *reservations* attribute, a manual check revealed that some of the restaurants which had not indicated that they accept reservation did in fact accept reservations.

Another issue was the lack of a universal system for categorizing the restaurants. Some restaurants categorized by country or continent (e.g., American, Chinese or Asian, European.), some by special food items (Sushi, Burger, Noodles). This led to a wide range of

categories as well as overlapping information (all Chinese restaurants are technically also Asian restaurants). Attempts to re-categorize were futile and deemed beyond the scope of this assignment.

3. Methodology

The overall progression of the project was as follows:

1. Data Acquisition
2. Data Wrangling
   2.1.Completion of variables (columns)
   2.2.Missing values
   2.3.Transforming variables
3. Data visualization (exploration)
   3.1. By venue
   3.2. By category
   3.3. By zip code
4. Modelling (inkl. evaluation)

This section of the report pertains to explanation of step *3. Data Visualization* and *4. Modeling*.

3.1 Data Visualization

Data was visualized in two parts. Firstly, the dataset *venues_full_df* was explored to examine the venues and their behaviors individually. Then, grouping of the categories allowed for a quick overview of the pertinent categories and their relationship to some important variables. Finally, in order to probe the variables' behavior from a geographical perspective, *venues_full_df* was grouped by zip code showing mean values for each zip.

3.1.1. By Venue

First, the venues were visualized as marker clusters on a map of Philly with boarders of the zip codes, see *figure 1*. This provided a broad understanding of the density of venues as across Philly.

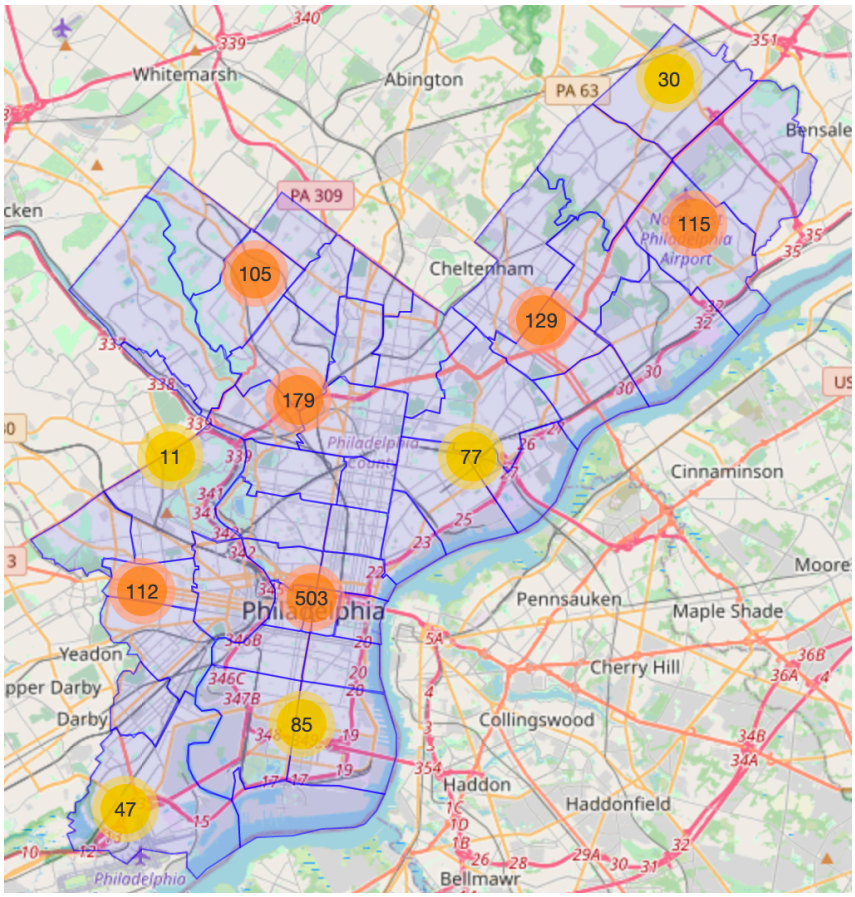


FIGURE 1: CLUSTERED VENUES IN PHILADELPHIA

Because it was an aspiration to obtain a balanced number of venues across zip codes, we cannot observe a perhaps expected dissipation of restaurants as we move further away from the city center. At default zoom level, the center appears to have a higher concentration of venues, but this merely owes to the fact that the zip codes are smaller toward the center of the city.

To gain an overview of the relationships between the variables, a correlation matrix was plotted with a heat map colormap, see *figure 2*.

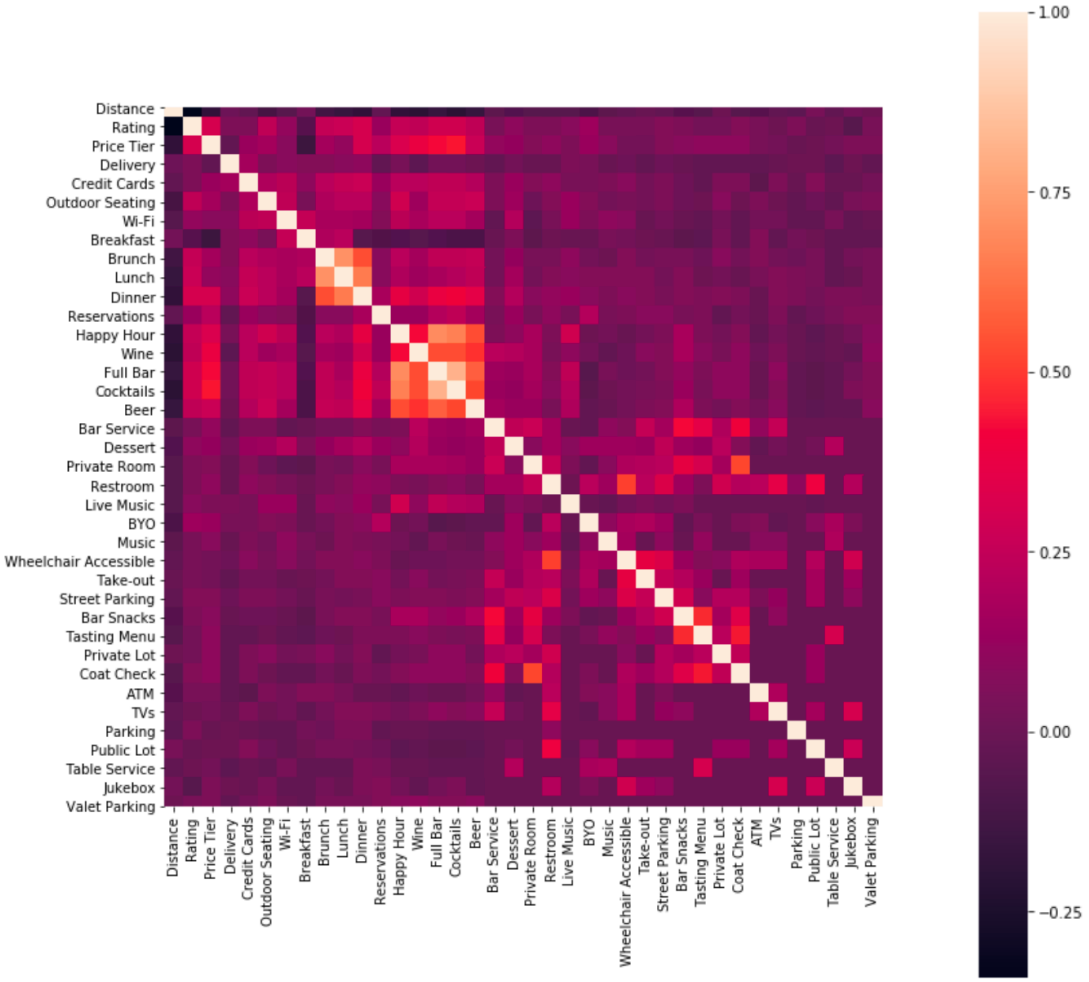


FIGURE 1: CLUSTERED VENUES IN PHILADELPHIA

Nothing appeared particularly noteworthy, save for the correlation between the various alcohol types and the variables brunch and lunch. Because of the strong correlation between the forms of alcohol, a new ordinal variable, *Alcohol*, was created whose value is *1* if the venue serves any of the alcohol forms. It can also be observed that *Rating* is somewhat correlated to the alcohol types as well as *Price Tier*.

Next, the dependent variable *Rating* was studied specifically. To begin, a bar plot was created to visualize the distribution of ratings, see *figure 3*.

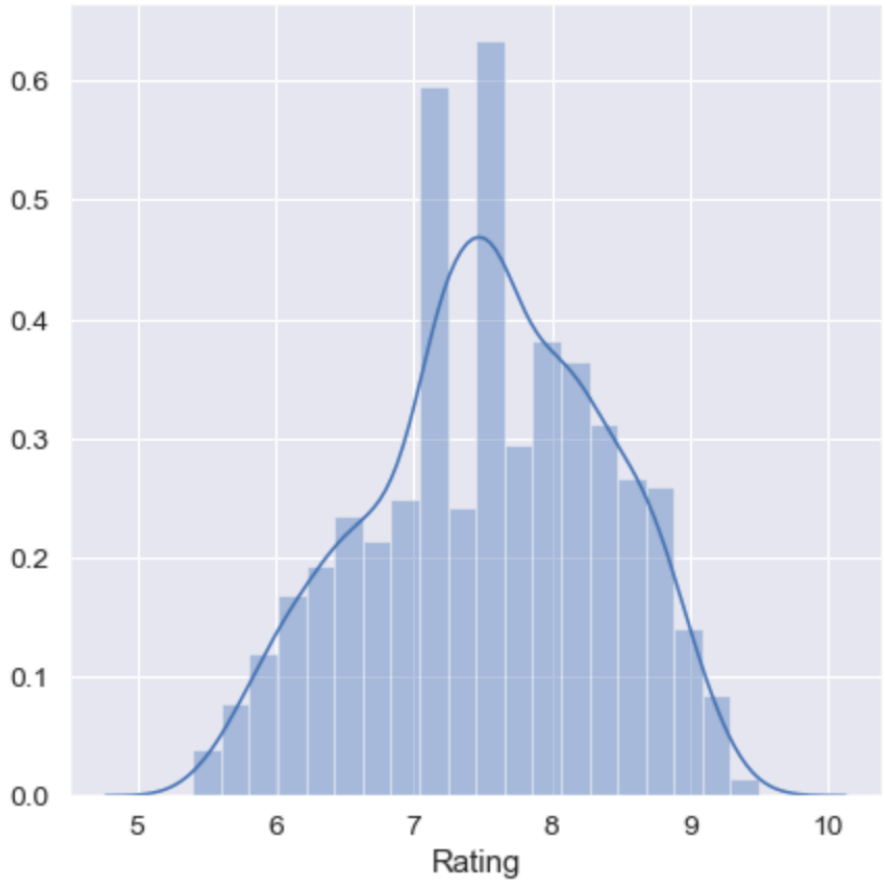


FIGURE 3: BAR PLOT OF RATING

It would be far fetched to call the above figure normal, however, it is close enough for our purposes. I was primarily vigilant of extreme deviations from normality. The two peaks at 7.3 and 7.6 do look unnatural, I suspect some UI issues for inputting of ratings that leads these two very specific numbers to have higher frequencies. I also began to speculate at the inaccuracy of humans rating on a 0-10 scale with decimals (de facto a 1-100 scale). I believe that the difference between a 7.1 and a 7.2 owes mostly to noise, thus I do not find it valuable to have a predictive model to predict at to this level of accuracy. Therefor, the *Rating* variable was later transformed into an ordinal (1-3) variable.

Next, the relationship between *Rating, Price Tier* and *Distance* was explored.

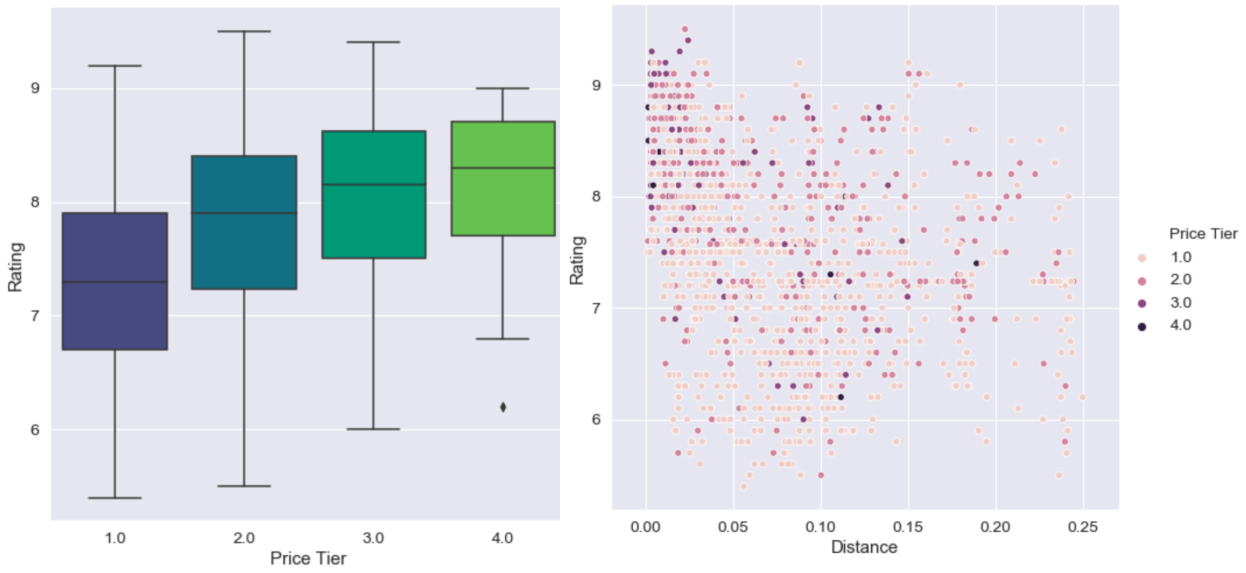


FIGURE 4: RATINGS VS PRICE TIER & DISTANCE

A small positive correlation can be observed between the two variables in figure 4a, as ratings tend to increase with price tier. Again in 4b, a slight tendancy toward higher price tiers are observed for higher ratings. More significant is the higher concentration of highly rated restaurants closer to the city.

3.1.2 By Category

It quickly became apparent that the *Category* column contained numerous unique categories. In order to investigate the variable further, a data frame where *venues_full_df* was grouped by category was created.As previously mentioned, several problems pertained to the *Category* of the restaurant, including the lack a universal categorization system as well overwhelming prominence of a few categories. This was visualized and discovered with a count plot, see figure 5:

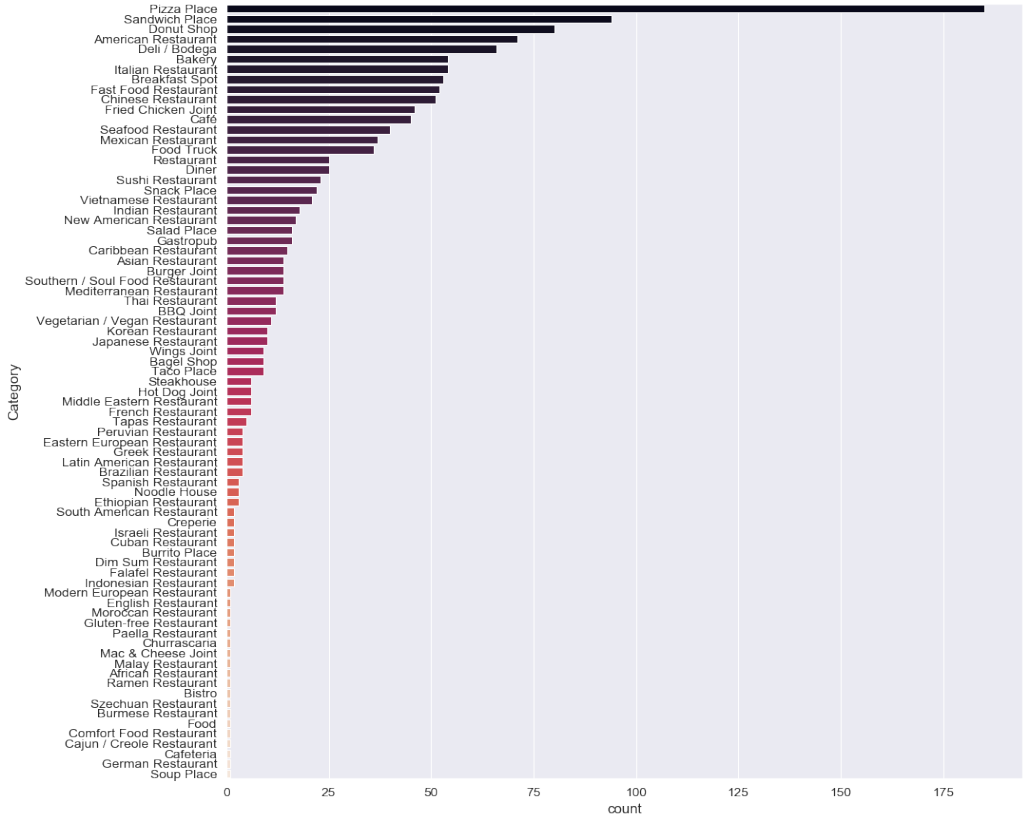


FIGURE 5: COUNTPLOT OF CATEGORIES

There is heavy skewness toward the top of the plot; the majority of venues belonging to a small set of categories. Subsequntly, extracting the top five category types for the zip codes (as we have done in the labs) would not yield any insight as the probability is so highly in the top categories' favor.

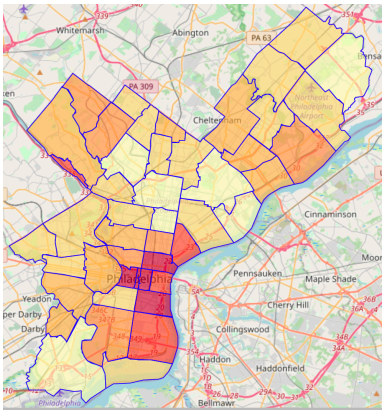
The following plots, figure 6, show how the variables rating and price tier (respectively) depend on category.

FIGURE 6: CATEGORY VS RATING & PRICE TIER

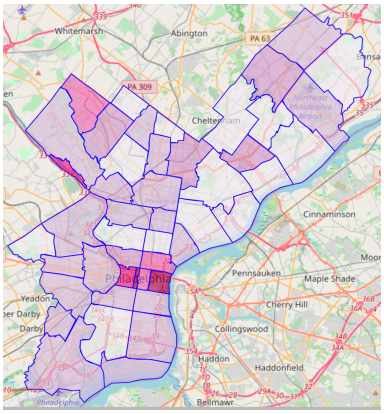
Because of the sparsity of data in the categories column (many categories, including the top and least rated categories, had only one observation), one should exercise caution in drawing too many conclusions from figure 6. However, what can be said is that *Friend Chicken Joint, Fast Food Restaurant* and *Donut Shop* generally have low ratings in Philadelphia (note that these were among the most popular categories which is why we can safely make this statement). The price tier does not appear to increase with rating.

3.1.3 By Zip

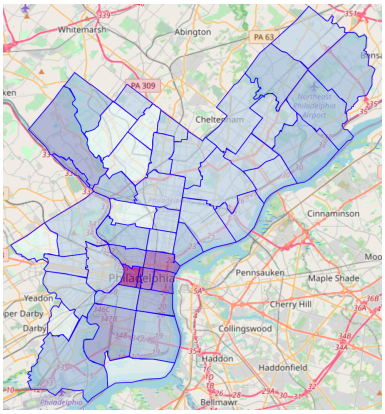
This part of the explorative analysis was conducted on a version of venues_full_df that was grouped by zip code showing mean values per zip. Extra columns were added that would be interesting to this geospatial analyses: {*Count:* num venues in zip, *Density:* Count / Area of zip*}*. Because the actual value of *Density* was uninteresting, it was substituted for *Density Rank,* a natural number ranking of Density (from 1 to len(Zip.unique()). Then, a map with multiple layers of choropleths was generated. Using the geo-data (previously described) and by binding to variables in the grouped dataset, the choropleths visualize in a gradual color map style how the variable changes for the zip codes. See figure 7.



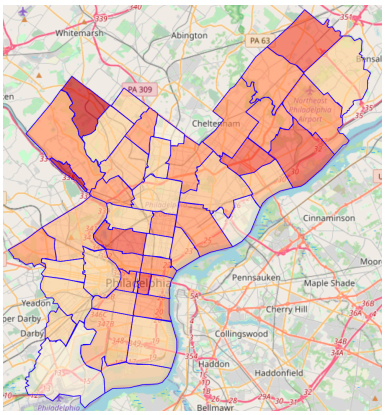
RATING



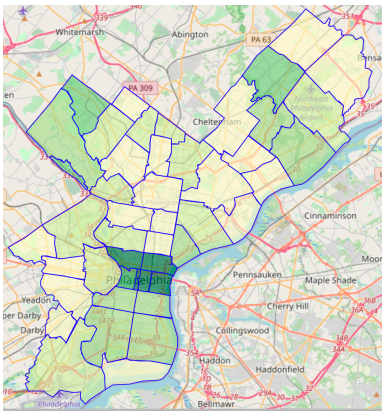
PRICE TIER



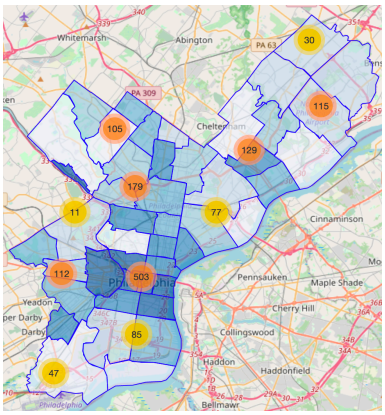
DINNER



DELIVERY



ALCOHOL



DENSITY

From figure 7, the following assumptions were made:

1. Zip codes around center city and west Philly tend to have higher concentrations of popular restaurants
2. Zip codes around center city and South Philly tend to have more highly rated restaurants
3. Zip codes around center city tend to be more expensive

Conclusion 1-3 were basically summarized in the earlier scatterplot of distance vs rating with price tier reflected in the hue.

4. Delivery appears more randomly distributed, although a slight tendency can be seen for venues further away from the city center to offer delivery.
5. Zip codes around center city tend to have more restaurants offering alcohol and dinner

Finally, a pair plot of the above variables with regression lines gives further substance to the assumptions. See figure 8.

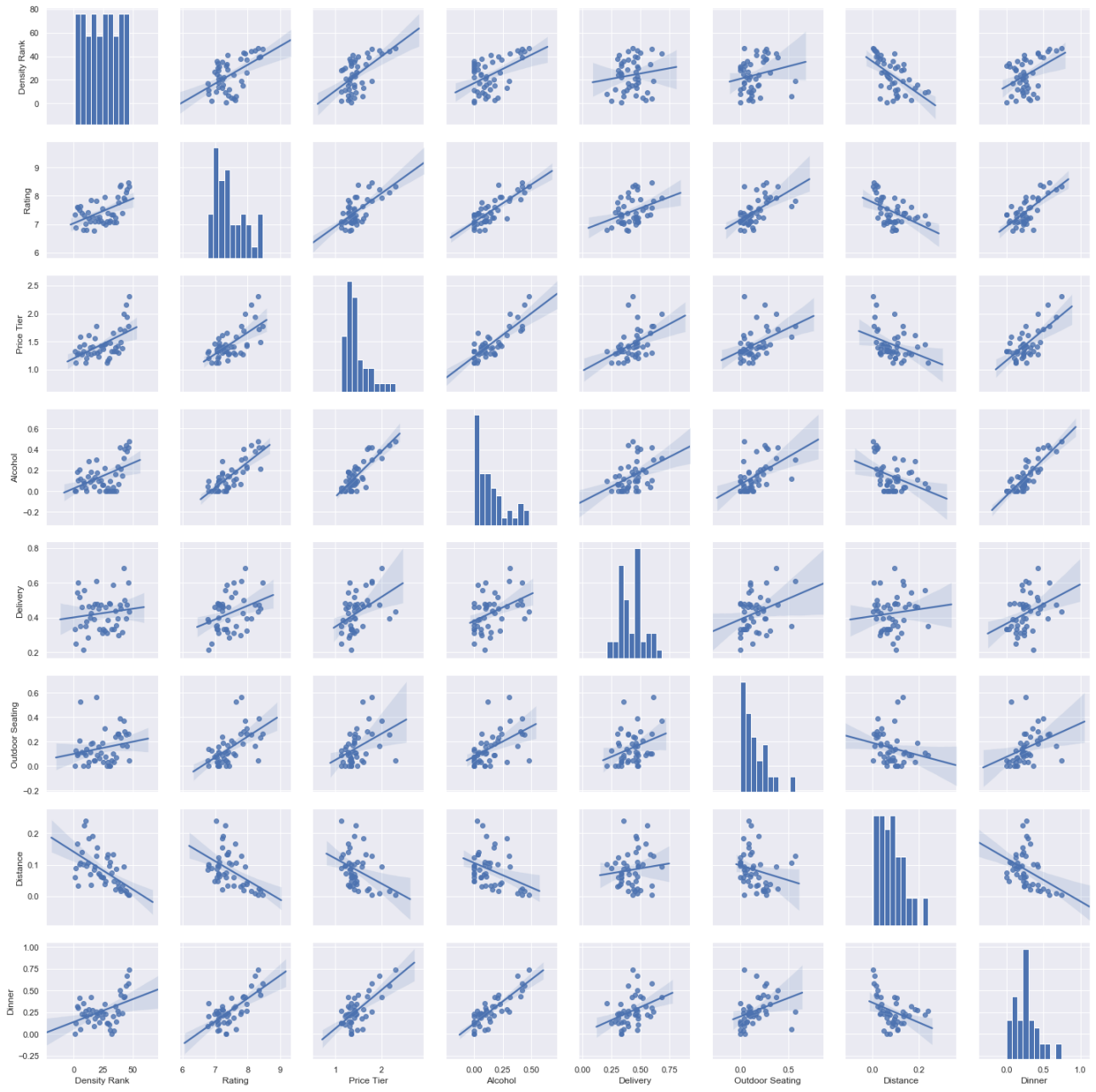


FIGURE 8: PAIRPLOT OF INTERESTING VARIABLES

Indeed, the pairplot offers no contradiction to our assumptions. They also reveal that price and rating are positively correlated with all other variables except distance, which is negatively correlated with all other variables (except delivery).

3.2 Predictive models

Because of the sparsity of data and because it was mentioned in an earlier class, I wanted to try combining clustering and classification algorithms. I chose K-means clustering because of its aptitude for and our familiarity in using it in geospatial analyses.

As mentioned, I did not find it valuable to predict a restaurants ratings to the decimal point, so the *Ratings* variable was transformed to a 1-3 ordinal variable. This was achieved by creating an array of equidistant points with start and stop at the minimum and maximum values of *Rating* (np.linspace()) and then transforming the rating of a restaurant according the interval they belonged to within the array. Thus, the classification problem was of multi-class nature instead of continuous.

The training and test data used for these algorithms consisted of all variables except *location data (lat/long/zip*), *name of restaurant* and *restaurant ID*. The data was standardized prior to splitting into a training set (80%) and test set (20%).

A range of classifiers were trained. Thanks to the versatility of sklearn, classifiers that are not inherently multi-class could be easily implemented as well. The following classifiers were implemented: Support Vector Machine, Decision Tree, Logistic Regression and Ridge Regression. Because there was significant class imbalance, the hyperparamer class_weight was set to ‘balanced’ for all models. For the support vector machine and decision tree models, several models were evaluated each with different kernel / max_depth.

4. Results

4.1 K-Means Clustering

The elbow method was used to determine the optimal k. See figure 9:

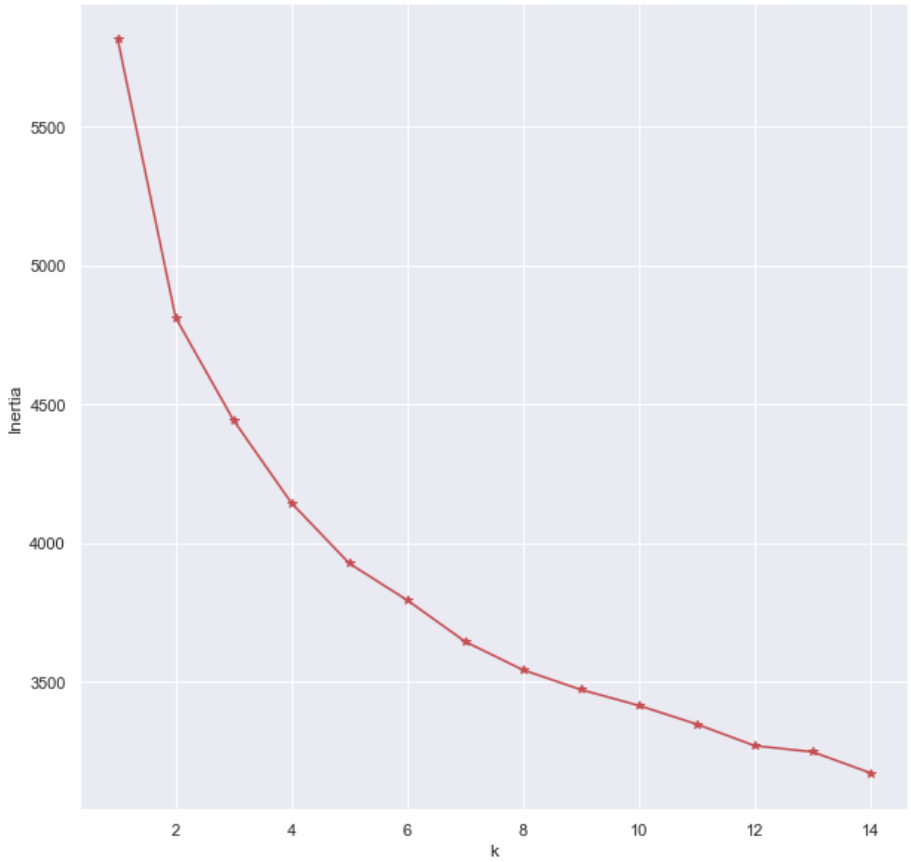


FIGURE 9: ELBOW METHOD FOR OPTIMAL K

While not completely apparent, the point most resembling an "elbow" is at k=2. After testing with a few variants of k, this was indeed found to yield the best results. So, the analysis proceeded with k=2.

A dataframe grouped by cluster label showing means was created to get an overview of what the clusters entailed. See table 4.

TABLE 4: VENUE DATA GROUPED BY CLUSTER*

Clst	Distance	Rating	Price Tier	Delivery	Alcohol	Credit Cards	Outdoor Seating	Wi-Fi	Breakfast	Brunch
0	0.0894	7.0926	1.2323	0.3732	0.0293	0.3720	0.0633	0.0633	0.1725	0.0892
1	0.0541	8.2338	1.8687	0.5027	0.3900	0.7264	0.3216	0.2033	0.1959	0.5693

*All columns not included

Fortunately, the clustering did indeed create separation for rating, with cluster 1 having significantly higher rated restaurants. As speculated, the values for *Price Tier, Alcohol* and

Dinner, exhibits positive correlation with *Rating* while *Distance* shows small negative correlation with *Rating*. Similar to the labs, the most common categories were also extracted; however, this time for the cluster labels and not the neighborhoods. The 20 most popular venues were also extracted for each cluster, see table 5.

TABLE 5: MOST COMMON CATEGORIES FOR CLUSTERS*

Clst	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
0	Pizza Place	Donut Shop	Sandwich Place	Fast Food	Deli / Bodega	Breakfast Spot	American	Chinese	Fried Chicken Joint	Bakery
1	Pizza Place	Italian	American	Sandwich Place	Mexican	Café	Deli / Bodega	Bakery	Seafood	New American

* 11-20 most common not included

As mentioned, we must take care to not be too naive about making assumptions based on the category data due to its sparsity; however Donut Shop, Fast Food and Friend Chicken occurred frequently enough to draw insight from. Recall that these three categories were among the lowest rated categories on average, and they all appear in the top 10 of cluster 0; whereas, cluster 1 does not feature them once in its top 10.

Finally, the clusters were visualized geographically, see figure 10:

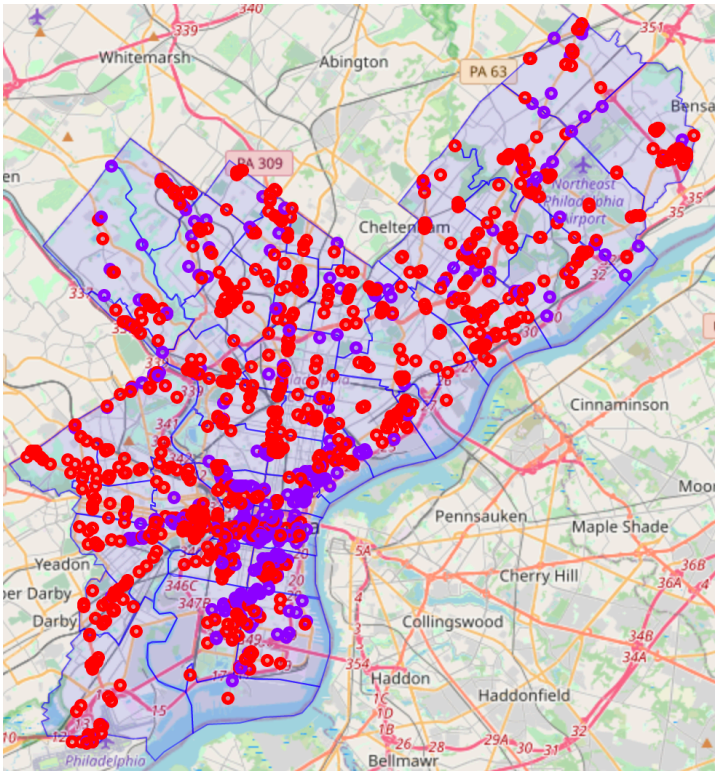


FIGURE 10: CLUSTER MAP (PURPLE = CLST 1)

In figure 10, we see a higher concentration of purple markers (cluster 1) around Center City and South Philly. These regions have higher concentrations of highly rated restaurants, so this too is unsurprising.

4.2 Classifiers

The classifiers were evaluated based on accuracy, f1 score and jaccard score. Precision and recall scores for each classifier was also extracted, see table 6.

TABLE 6: PERFORMANCE OF CLASSIFIERS

	Accuracy	F1 Score	Jaccard Score	Precision	Recall
SVM	0.7025	0.7155	0.5674	0.7512	0.7025
DTree	0.7491	0.7492	0.6104	0.7503	0.7491
Log Reg	**0.7706**	**0.7449**	**0.6196**	**0.7512**	**0.7706**
Ridge Reg	0.6667	0.6856	0.5339	0.7376	0.6667

The logistic regression model had the best scores in all metrics with the decision tree coming in a close second. Figure 11 shows a count plot of the logistic regression model's prediction on the test set (blue) along with the actual test labels (orange).

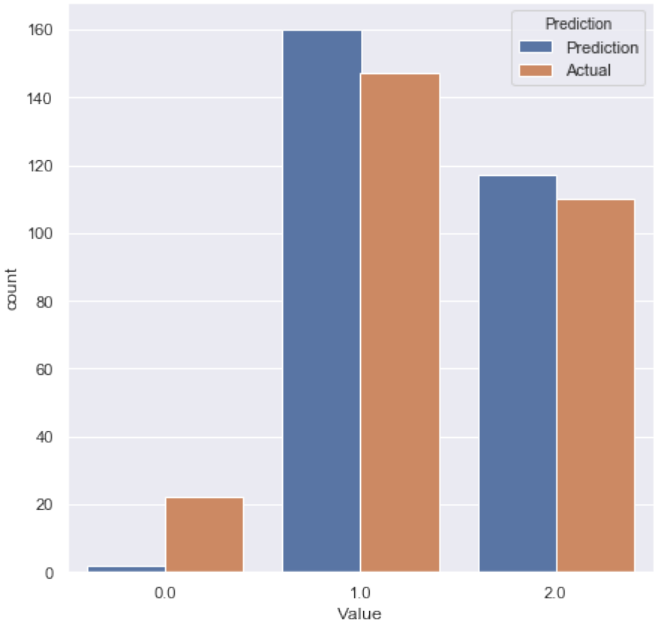


FIGURE 11: LOGREG COUNT PLOT OF PREDICTIONS AND ACTUAL VALUES

Unfortunately, the class imbalance has clearly resulted in the model having a bias toward predicting class 1 and 2. For this reason, a similar count plot for the decision tree model was created, see figure 12:

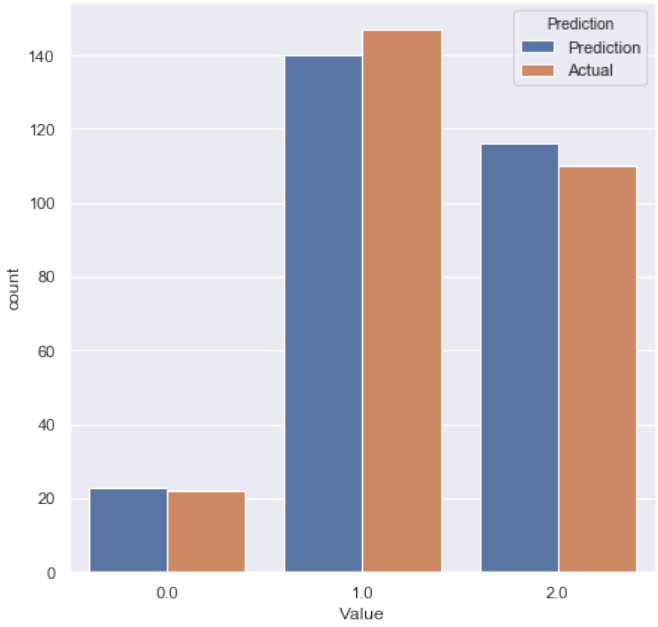


FIGURE 12: DECISION TREE COUNT PLOT OF PREDICTIONS AND ACTUAL VALUES

Despite the heavy class imbalance, the decision tree model performs well as the distribution of predictions rather closely mimics the distribution of the actual labels. Perhaps this model would prove more robust given more data. Finally, figure 13 shows visualizes the decision tree's path to predictions.

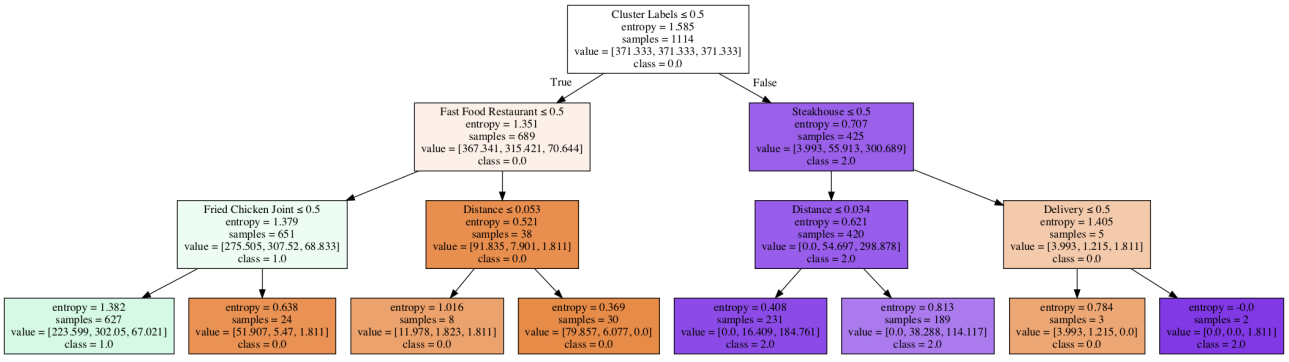


FIGURE 13: DECISION TREE

Figure 13 and the relative importance attribute of the decision tree object showed that cluster label was indeed the most important variable leading to the highest information gain. After that, the algorithm appears to check for highly / lowly rated categories (i.e., steakhouse, fast food & fried chicken). At this point I was concerned that overfitting may have occurred. Steakhouse was a fairly uncommon category in Philly, however all occurrences of

steakhouses were highly rated. Surely there exists lowly rated steakhouses as well. Conversely, there must exist highly rated fried chicken joint.

To avoid this behavior, a decision tree was trained without the categories. The resulting trained model performed to a satisfactory level, see table 7. In figure 14, the decision path is visualized.

TABLE 7: PERFORMANCE OF DECISION TREE 2

	Accuracy	F1 Score	Jaccard Score	Precision	Recall
DTree	0.6953	0.7082	0.5638	0.7311	0.6953

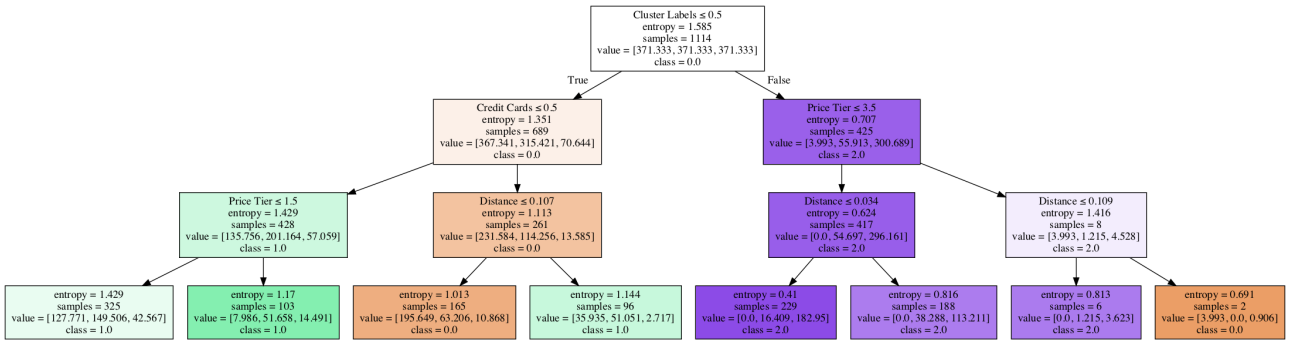


FIGURE 14: DECISION TREE 2

This time around, higher importance is placed on the attributes *Price Tier* and *Credit Cards*. In both decision trees, distance is an important variable with lower values leading to higher rating tiers.

5. Discussion

Through the data exploration, it was found that it general, the existence of attributes (value 1) tended to have a positive affect on both rating and price tier and that higher price tiers tended to drive up ratings. The variables that had the highest pairwise correlation with rating was price tier, alcohol and dinner. It was also concluded that distance from center city negatively affected rating, i.e., higher rated restaurants tended to be located closer to the city center.

This was reflected in the results of the K-Means clustering which yielded two clusters. Cluster 0 represented lower rater restaurants and thus had lower average vales for alcohol, dinner and price tier as well as outdoor seating and Wi-Fi. Restaurants that had been clustered into cluster 1, the cluster with higher rated restaurants, tended to be closer to the city center as well.

The classifiers, that were shown the cluster labels when training, placed heavy importance on the cluster labels. The decision tree model, which performed the best, had cluster label as its most important attribute for information gain. The cluster label represented a number of key pieces of information that all correlate to higher ratings, showing the fortuitous synergy between the algorithms in this case. Indeed, when the cluster label was omitted from the training set, the same highly discriminative variables (alcohol, dinner etc) became important for the decision tree algorithm.

The decision tree model achieved accuracy levels of up to 75%, however I suspect overfitting has occurred. Thanks to the intuitive nature of decision trees, they can be analyzed visually, and their paths to a prediction followed. In the case of this project, the branching that occurred suggests that the model has learned some patterns that are very specific to our training set, e.g., that steakhouses are highly rated and that a restaurants with high price tier but distances just over 0.109 are likely lowly rated, had the restaurants however been just a little bit closer tot the center city it would have had the maximum rating tier.

6. Conclusions

In this project, I have explored the restaurant scene in Philadelphia in order to create a machine learning model able to predict the customer rating of a given restaurant. This was done by creating and analyzing a dataset comprised of restaurant data obtained from the foursquare API. Exploratory data analysis revealed that the majority of highly rated restaurants are located in and around center city. The preliminary analysis also suggested that highly rated restaurants tend to be more expensive and offer more amenities/services such as alcohol, outdoor seating, credit card support etc. This was supported by subsequent machine learning models, including a decision tree that was attained 75% accuracy when predicting ratings.