Christopher Kang
https://github.com/chriskang97/Coursera_Capstone

Classifying a good BBQ Restaurant

1 Introduction

1.1 Background

Starting up a restaurant from scratch is a risky action to take as the likelihood of the restaurant failing is high due to many of the existing dominant restaurants that are in the industry already. Typically, people define a good or successful restaurant as having good food, good services, low prices, etc. The scope of this analysis will attempt to show how much this claim holds up and whether there are specific restaurant properties that hold more significance. Results of this analysis will be able to help owners who want to enter into the restaurant scene along with helping potential investors determine whether the restaurant will be a good investment to make.

1.2 Focused Area and Constraints

The specific restaurant type was chosen due to being a common restaurant type seen in most states along with the author of this project having a strong interest in this cuisine. This limited field will also help showcase common attributes and patterns easier as these restaurants will all focus on a certain type of cuisine. In addition, all restaurants acquired will be located in the USA as entry level market properties can vary between countries.

1.3 High Level Goals

1. Acquire many different BBQ restaurants located all around the USA and clean the data to include specific restaurant properties.

2. Visualize the data by showcasing each of the restaurant properties and determine if a common trend can be shown.

3. Utilize different classifying algorithms such as KNN, SVM, Random Forest, etc and measure how well our model does with existing restaurants and randomly generated ones.

4. Analyze results and determine how much of the result can be used for other types of restaurants.

2. Data Explanation and Acquisitions

2.1 Data Usage

In order to address the problem stated above, individual restaurant's data will be needed in constructing a model. Restaurant's service and quality can be represented by the restaurant's overall rating, good food can be represented on the items displayed on the restaurant's menu, and price can be represented by the overall price of the restaurant's food. Location will also be used as another factor to address the problem stated above as well.

2.2 Data Sources

Data that will be used in this analysis will come from the Foursquare database and be obtained by making multiple POST/REQUEST calls. Since the analysis focuses on good restaurant properties, each data can be categorized in these following categories: types of food being served, location of the restaurant, price range, general comments, number of people checked in, number of likes in the restaurant, etc. To obtain these restaurants, a venue search will be done and we will be searching for venues that fall under the category of BBQ Joints. In addition, locations will be limited to the state and capital.

To choose the representative states for our model, three states from each region across the United States. There are roughly nine representative regions/divisions in the USA so approximately 27 states will be used in constructing the restaurant dataset [1]. States chosen were done manually as there are certain states that can be classified under a BBQ state [2]. If no state was classified as a BBQ state for that division, the states that have the highest population count would then be chosen in order to increase the overall quantitative values for each of the restaurant's data attributes. Once this was done, a State and Capital Dataframe table was created using a csv file containing each state's respective state and capital.

The biggest issue in obtaining the data came from the actual Foursquare API retrieval. Firstly, a function was made to find all restaurants located in each of the state's capital states and stored into a DF table. The main information that was needed from each restaurant were the ID, name, latitude/longitude, rating value, price tier, category of restaurant, number of ratings, menu items and price of each item. However, multiple premium calls would be required to obtain each restaurant's information. In order to approach this limitation, a physical web browser html retrieval was done to the foursquare website. After a few restaurants retrieval, this method ended up failing due to the website blocking the web scraper in obtaining information. Finally, a macro procedure was programmed to physically copy the text displayed on the foursquare developer console page [3]. Each text was copied onto a text file and was then turned into a pickle to

reduce the amount of storage data. Afterwards, each pickle file was then read and each subsequent information was parsed and added onto the DF table.

	ID	Name	Latitude	Longitude	State	City	Category Name	Rating	Number of Rating	Price Tier	Menu Item	Menu Item Price
0	534f07e9498e5cc70137182b	The Causeway Restaurant and Pub	42.364659	-71.062912	MA	Boston	BBQ Joint	7.0	55	Moderate	[]	[]
1	593d9fb5c876c8327eef128d	Rusty Can	42.755437	-70.938839	MA	Byfield	BBQ Joint	8.0	14	Moderate	[]	[]
2	4ba2b7caf964a520211338e3	Joff's Backyard Grill	42.084574	-71.471883	MA	Bellingham	BBQ Joint	7.9	18	Moderate	[]	[]
3	5c5efed8419a9e002ce8ea9c	The Smoke Shop BBQ - Assembly Row	42.392249	-71.078180	MA	Somerville	Restaurant	-1.0	0	Moderate	[17Th Street Soul Rolls, Hot Links & Pimento C...	[9.00, 7.50, 9.00, 10.00, 12.00, 9.00, 9.00, 8...
4	5ca25d74dd12f8002c74364b	Flip The Bird	42.559270	-70.881620	MA	Beverly	BBQ Joint	-1.0	0	Moderate	[]	[]

Figure 1: Dataframe Table showcasing Restaurants Information

3. Data Methodology

3.1 Data Visualization

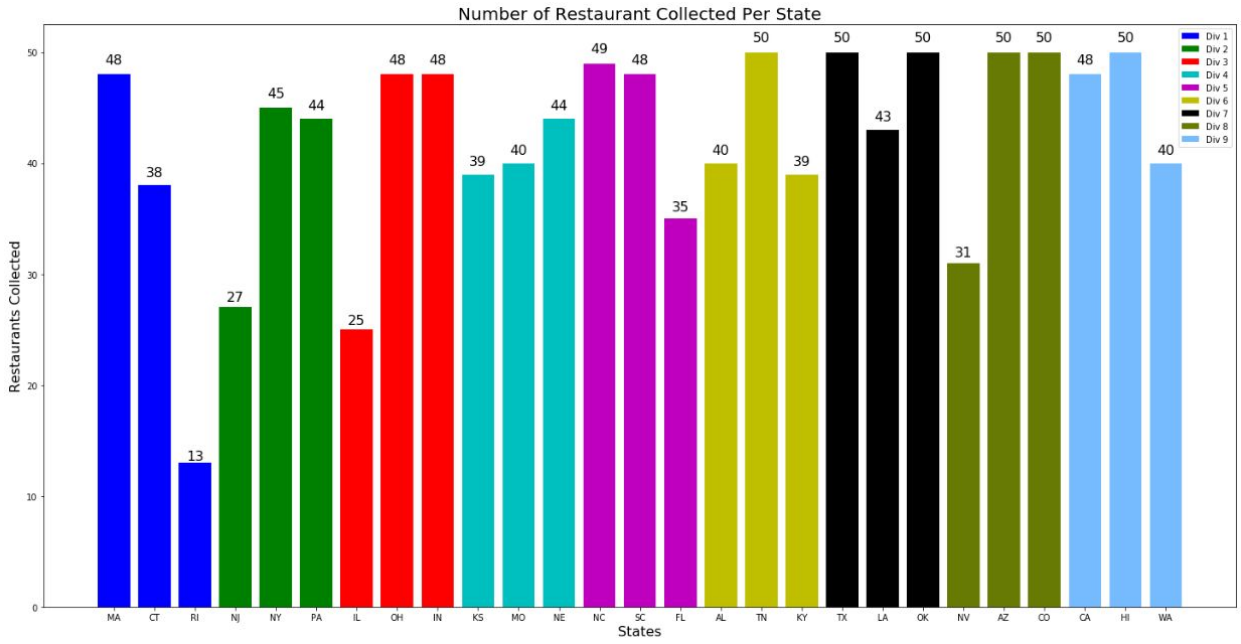
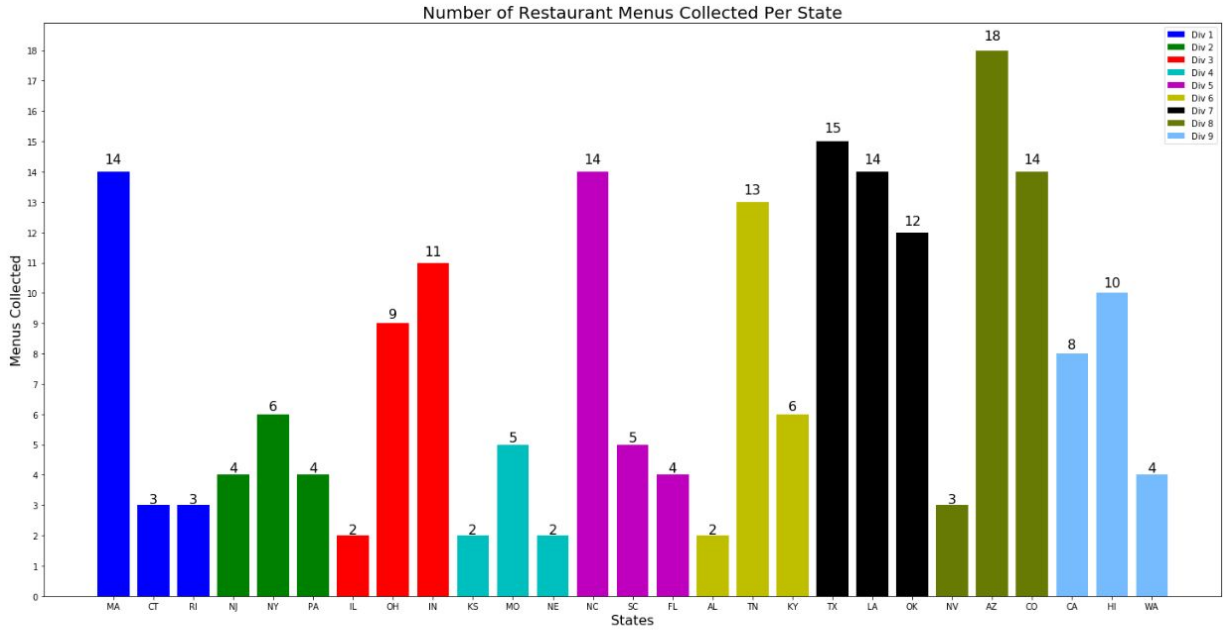


Figure 2: Number of Restaurants Collected Per State

From this graphical representation, the number of restaurants obtained per each capital and state seems to cap off at 50. It seems most division has two states that contained a sizable count of restaurants while the other state seems to have a lower count of restaurants.




```
Total Menus Collected: 207
Menus Collected Per Division: [20. 14. 22.  9. 23. 21. 41. 35. 22.]
```

Figure 3: Graphical Visualization on the Number of Menus Collected

With the total menus collected being only 207, this is approximately 17-18% of the total restaurants collected. Most divisions seem to have an equal amount of menus collected with the exception of Division 7.

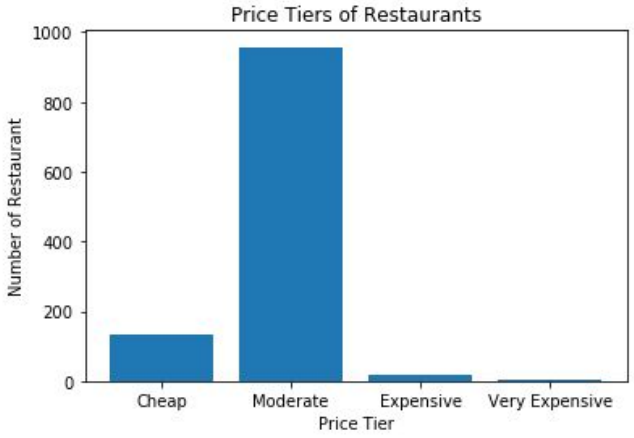


Figure 4: Graphical Figure of Restaurant Price Tier

Most of the restaurants that were collected fell under the moderate price tier category. The following runner up price tier was the cheap price tier category.

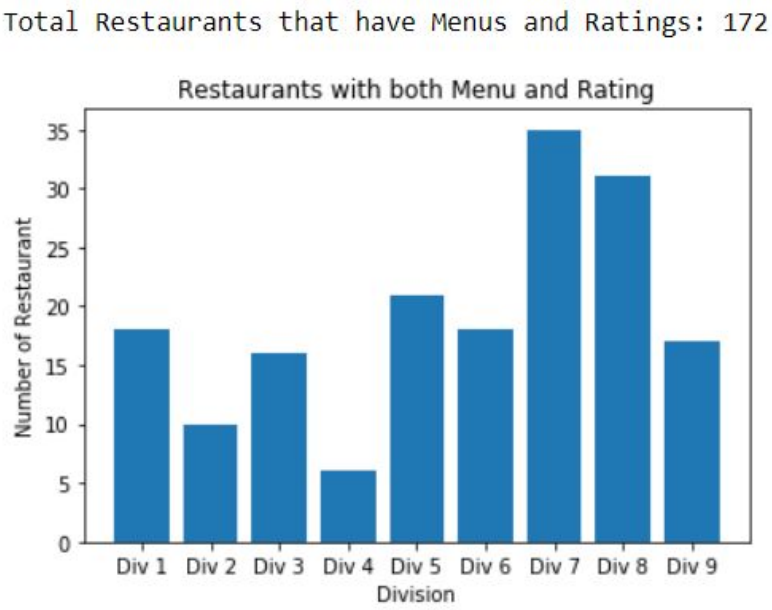


Figure 5: Restaurant Rows containing both a Rating and Menu

With the number of restaurants having a menu, the number of restaurants containing both a menu and rating isn't too surprising to see.

```
restaurant_df.groupby("Category Name").size()
```

```
Category Name
American Restaurant                21
Asian Restaurant                    4
BBQ Joint                         959
Bakery                              1
Bar                                10
Beer Bar                            2
Brazilian Restaurant                2
Breakfast Spot                      2
Brewery                             5
Building                            1
Burger Joint                        3
Cajun / Creole Restaurant           3
Chinese Restaurant                  1
Churrascaria                        1
Clothing Store                      1
College Classroom                   1
Comfort Food Restaurant             1
Cuban Restaurant                    2
Diner                               5
Dive Bar                            3
Fast Food Restaurant                4
Filipino Restaurant                 1
Food Truck                          9
General Entertainment               1
Grocery Store                       1
Hawaiian Restaurant                 7
Italian Restaurant                  1
Japanese Restaurant                 4
Karaoke Bar                         1
Kebab Restaurant                    1
Korean Restaurant                  12
Mexican Restaurant                  3
Mongolian Restaurant                1
Pizza Place                         3
Ramen Restaurant                    1
Restaurant                         10
Sandwich Place                      1
Seafood Restaurant                  2
Southern / Soul Food Restaurant     5
Sports Bar                          4
Steakhouse                         26
Wings Joint                         6
dtype: int64
```

Figure 6: Restaurant Category Collected

Finally, the highest number of restaurant categories picked up was BBQ Joint. Although other international BBQ restaurants were considered as part of this analysis, the data sample that was collected lies mostly in this category and thus narrows our analysis further to contain roughly typical American BBQ joint restaurants.

3.2 Data Cleaning

In order to use menu item and menu item price, we need to be able to turn them into numerical values that we can use in our model making. The original approach was to find all of the BBQ food items listed in the menu and count how many of those occur in the menu. The issue that this comes with is assuming that each BBQ item that the restaurant has can be considered as a food item that the customer would rate positively. Instead of relying on potential outside factors, the entire menu item length was considered and this attribute can describe a restaurant's specialty in a wide or narrow variety of BBQ foods. Menu item price can then be turned into average menu item price by taking the average of all of the food price's in the restaurant.

As seen in Figure 1, some of the restaurant's attributes are empty. This can occur if no one has inputted the formatted data onto the Foursquare website or if no one has submitted any reviews/rating for that restaurant. The first approach in solving this issue was to replace all the empty values with the mean for that column. However, due to many of the restaurants having missing attributes, the correlation scatterplot between the variables ended up being extremely skewed towards the mean. This also happens if each of the empty values were replaced with a linear fit of the column's data as well.

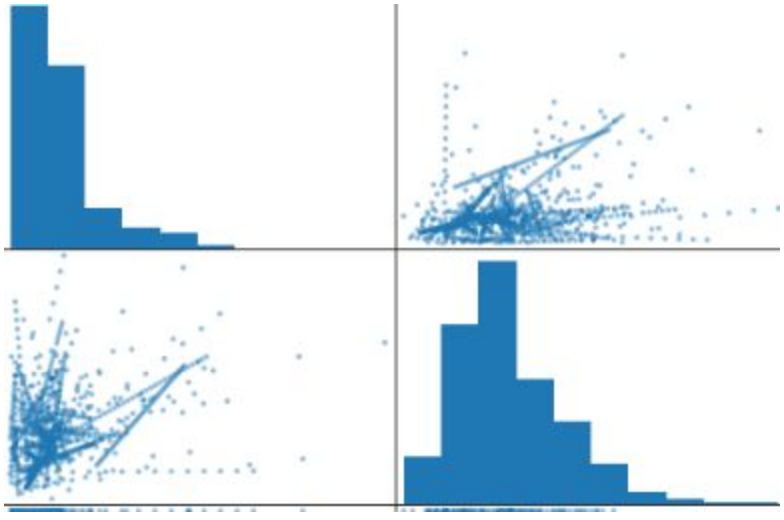


Figure 7: Correlation Scatterplot between Average Menu Length and Average Item Price

Therefore, each of the missing values in each column was going to be a value randomly picked from other restaurant values that were inputted. Missing restaurant's price tier would be labeled as moderate due to a majority of restaurants gathered falling under the moderate price tier category. When finding that a huge number of the restaurants contained missing values, several assumptions can then be made with analysis going forward. The model might not contain the entire sample size data that can represent the BBQ restaurant industry in the USA due to lack of sample size, a huge chunk of missing values, and a majority of restaurants falling under the moderate price tier category.

3.3 Data Modeling

The original intent of the model was to be able to predict whether an inputted restaurant would be a good restaurant. However, as stated in the previous section, due to many restaurants having missing attribute values, designing a predictive model would be hard to do and is more likely to be under fit or biased predictions. Thus, the model that would be made will be focused on unsupervised learning algorithms and specifically, clustering algorithms.

The clustering algorithms used were K-means, spectral clustering, and DBSCAN. K-means focuses on finding all the points that lie in the given inputted number of clusters through a distance based metric. Spectral clustering also attempts to find a given set input of clusters in a given set of features and accomplishes this through a set of eigenvectors and can find clusters that aren't necessarily a fine tuned shape. DBSCAN on the other hand attempts to find the number of clusters that are presented and does this based on the hyperparameters of maximum distance and minimum number of points to be considered as a cluster. Due to the small set of data, all of the data will be used in constructing the model; meaning no train/test/dev set will be made in the process.

4. Results and Conclusion

4.1 Results

In order to see the results for K-Means and spectral clustering, a silhouette analysis was done between each of the features. Silhouette analysis allows us to see the clusters that each feature belonged to by plotting them on a 2D graph and the silhouette scores tells us how well each of the clusters are distributed by comparing each of the data points inside the cluster. Before looking at each of the model results, many simulations and model creations were done ranging from hyperparameter tuning to determining which features would be used. The final models were created by using only rating, number of rating, and average item length.

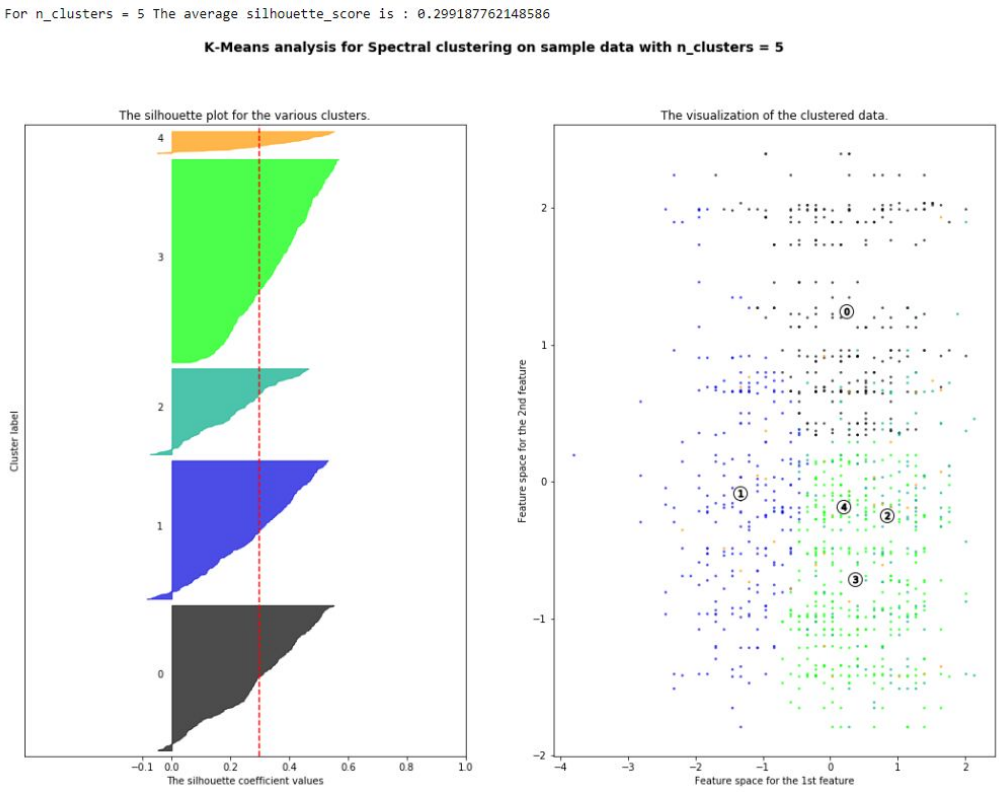


Figure 8: Silhouette Analysis with K-Means with Number of Clusters = 5

The current features that are being compared right now are ratings and average item price. Though each of clusters are roughly distributed graphically, there seems to be a lot of overlaps and the silhouette score being roughly 0.3 means that each of the points in the cluster aren't similar to one another.

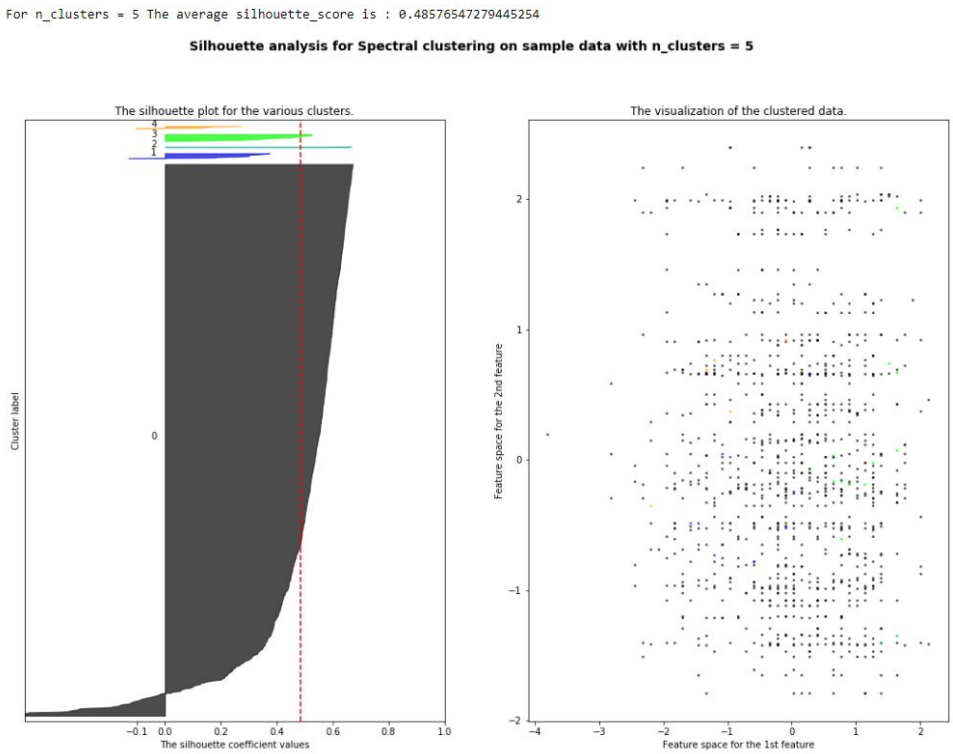


Figure 9: Silhouette Analysis with Spectral Clustering with Number of Clusters = 5

When attempting to do a spectral clustering, one cluster ends up clustering the entire dataset and thus ends up skewing the amount of data points per cluster set. Therefore the silhouette can't be depicted as a meaningful result here.

```
Estimated number of clusters: 2
Estimated number of noise points: 21
Silhouette Coefficient: 0.494
```

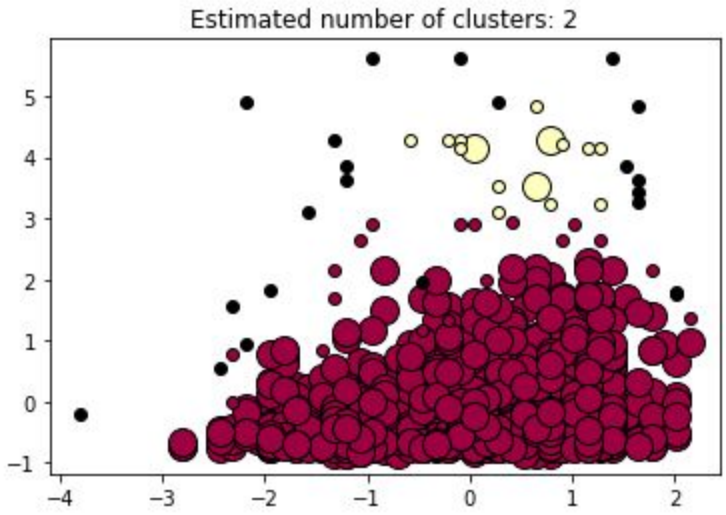


Figure 10: Graphical Result of DBSCAN and Silhouette Score

Using DBSCAN resulted in roughly two clusters forming along with several black colored data points considered as noise. With the silhouette coefficient being reasonable and a clear pattern shown visually, this was the algorithm that was chosen on this dataset. When analyzing the restaurants in the yellow cluster, the following results were shown:

	ID	Name	Latitude	Longitude	State	City	Category Name	Rating	Number of Rating	Price Tier	Average Menu Length	Average Item Price	Division
119	4cd573d7122ba143c4192ca1	Dinosaur Bar-B-Que	42.734600	-73.689244	NY	Troy	BBQ Joint	8.4	385.0	1.0	25.0	7.318182	2
127	5b8ad0a631fd14002c758e7c	Unihog	42.898373	-73.352850	NY	Hoosick Falls	BBQ Joint	7.6	388.0	1.0	25.0	6.476000	2
158	51b2306b498e43d578d7bd29	Bullbq's Burger Cafe	40.269864	-76.888302	PA	Harrisburg	BBQ Joint	8.3	311.0	1.0	10.0	6.272800	2
169	503923eee4b06348eb4d6260	Little Everett's BBQ	39.808870	-77.001290	PA	Hanover	BBQ Joint	7.7	379.0	1.0	10.0	7.168000	2
259	520e9ff911d2b12d923e69e1	Tamarack chicken and fish	40.076410	-82.966952	OH	Columbus	BBQ Joint	7.5	388.0	1.0	18.0	6.272800	3
271	4b64f537f964a52045dc2ae3	Weber Grill Restaurant	39.767612	-86.159902	IN	Indianapolis	American Restaurant	7.9	333.0	1.0	29.0	7.600000	3
442	5291360e11d20de90e42c30e	The Pit	36.003458	-78.899893	NC	Durham	BBQ Joint	7.9	301.0	1.0	46.0	5.610580	5
596	4c2b7217b34ad13ab4c3e9ce	Five Brothers BBQ	32.343436	-86.222641	AL	Montgomery	BBQ Joint	7.6	379.0	1.0	54.0	8.137143	6
612	515484d6e4b0bdae55d49c6d	Edley's East	36.175910	-86.756398	TN	Nashville	BBQ Joint	8.7	311.0	0.0	40.0	7.303571	6
667	4b747dfef964a520dddf2de3	Attus Apparel	38.000664	-84.524555	KY	Lexington	Clothing Store	7.2	388.0	1.0	5.0	5.843750	6
678	5676037b498e1b7d98c935c7	Cooper's Old Time Pit Bar-B-Que	30.264966	-97.743747	TX	Austin	BBQ Joint	8.6	379.0	1.0	10.0	7.303571	7
688	4a69f8d0f964a5204ecc1fe3	County Line on the Lake	30.357139	-97.785685	TX	Austin	BBQ Joint	8.3	388.0	1.0	25.0	7.360000	7
833	58fe9edf340a5840405cc618	Naked BBQ	33.579469	-111.887248	AZ	Scottsdale	BBQ Joint	8.2	430.0	1.0	49.0	7.855000	8
885	5dd705446292a10008685d70	AJ's Pit Bar-B-Q	39.677094	-104.992101	CO	Denver	BBQ Joint	8.7	379.0	1.0	48.0	7.716400	8
1027	4b36a724f964a520fe3925e3	The Butcher Boys Beef Outlet	47.119359	-122.293721	WA	Puyallup	Steakhouse	8.2	333.0	3.0	65.0	7.360000	9

Figure 11: Restaurants Falling under the Yellow Cluster

Majority of restaurants seem to fall under the east and south division of the USA. Additionally, a majority of them have above average rating and have a huge number of ratings given. Finally, the price tier lies within the moderate category with average menu length lying from small to large menu sizes.

4.2 Discussion

Ideally, the cluster shown above would represent the restaurants that can be considered good to the consumers. Common trends such as low price along with relatively positive ratings seems to be the main pattern seen in this cluster. As stated above, features had to be removed due to them producing results that either added more outliers on the graph or drastically reducing the silhouette score.

In addition, the biggest assumption this analysis makes is that each of the restaurant's randomized value inputs were accurate enough to represent that restaurant. These may or may not correctly represent the restaurant's values and thus not enough confidence can be placed in the accuracy of this analysis overall. As stated before, the main issue seems to be from a lack of

total sample representation of the BBQ restaurant industry as a whole and having more restaurants with accurate data will help the analysis be more concrete.

4.3 Conclusion

In terms of the overall analysis, the result showing that the number of people rating the restaurant and the price being fair is still a reasonable conclusion to come to. People who frequent the restaurants are more likely to leave a positive rating and each of those visits can attest to the restaurant's quality and standard. In addition, people value the restaurant's item being fairly reasonable and this goes in line for people who are able to eat at the restaurants multiple times rather than a one time occasion.

Once again, further data collection will be needed in order to prove the previous statements assertion. In terms of what I learned while doing this project, I've realized that data collection and cleaning are incredibly vital in making a concrete analysis. Where the data is sourced from and how well they represent the sample as a whole are important to consider along with making sure that the data is cleaned and formatted properly. In addition, outliers can come from wrong inputs from the data source or scraping the data incorrectly.

In order to continue this analysis and project, a proper documentation and classes/object would be needed in order to help reduce the amount of time spent on creating visual graphics and performing the analysis.

5. Reference

[1] https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States
[2] https://en.wikipedia.org/wiki/Barbecue_in_the_United_States
[3] https://foursquare.com/developers/explore#req=users%2Fself