**Predicting the Best Italian Restaurant to Eat at While Visiting Washington, D.C.**

Logan Walker

December 23rd, 2019

**1. Introduction**

**1.1 Background**

Washington, District of Columbia, the capital of the United States of America, is memorably known as not only a hub for a majority of the nation’s federal powerhouses and institutions, but also can be quite enjoyable due to its wide variety of historical attractions. Tourists from both inside and outside of the United States often choose Washington, D.C. as a top destination to visit on their list of must-see places across the country. With this influx of visitors to the nation’s capital only growing, a handful of notable and curious business owners have chosen the District of Columbia as an excellent candidate when deciding where to open a new restaurant in order to allow for tremendous success. As this development of food options continues to expand, the District of Columbia is attracting even more tourists from all over the country who hope to find new and exciting restaurants to try in an unfamiliar environment. One particular cuisine category that is one of the most popular is Italian.

**1.2 Problem**

With limited knowledge of the city’s layout and so many neighborhoods to choose from, these groups of people might wonder where exactly the best location to try an Italian restaurant might be. As there are numerous Italian restaurants in the Washington, D.C. area, it would be helpful to focus on locations that contain numerous options for Italian restaurants. It would also be especially valuable to target locations that are as close to the D.C. city center as possible, in order to make it easier for tourists who are already visiting major attractions around the city to walk to these locations.

**1.3 Interest**

Obviously, groups of tourists visiting Washington, D.C. would very interested in this idea of exploring top rated Italian restaurants that are reasonably close to attractions that they are already close to. Additionally, potential entrepreneurs or restaurant owners might also be interested in this topic and maybe even further exploration to determine why some of these restaurants have such high reviews or a large volume of customers. They could use such information to help decide which variables play a larger role to customers when deciding which restaurant to eat at and factor this in to determine where the best location to open a restaurant would be.

**2. Data acquisition and cleaning**

**2.1 Data sources**

Fortunately, for those whom this might appeal to, two applications in particular make this topic very easy to explore further. Through an open-source geocoding Python library package, Geocoder, one can easily enter an address into his or her Jupyter Notebook and immediately get this address’s geographical coordinates with just a few simple lines of code. Also, through Foursquare Places, a location data provider application, a few API calls can be made to trace popular restaurants by cuisine nearby a particular location in Washington, D.C. This information can then be used by tourists hoping to eat at a restaurant in order to carefully target food options that are closer to the tourist attractions they might experience and possess the most positive feedback.

First, utilizing Geocoder’s Nominatim feature, the coordinates of the city center of Washington, District of Columbia needed to be retrieved in order to serve as a starting point. An address object was defined and then plugged into the geolocator function to retrieve latitude and longitude values. The geographical coordinates were found to be 38.8948932 and -77.0365529, respectively.

Next, in order to focus specifically on Italian restaurants within a defined distance from the Washington, D.C. city center, a search query object was defined with a value of “Italian”. The radius was set to 5000 meters from the starting point and limited to 50 venues. Foursquare credentials including API Client ID, Client Secret, and Version were defined and then utilized within a created Foursquare URL. The URL was then used to conduct a search of venues with the defined criteria and the requests package converted these results into a JSON file to be filtered.

**2.2 Data cleaning and feature selection**

Data retrieved from the JSON file was combined into one Pandas data frame, with each row containing information pertaining to a specific restaurant from the search. There was a total of 47 venues returned and 21 features for each venue. Several of these features, however, contained missing values and some features were determined to be inapplicable. In order to filter out irrelevant features for each venue, a function was defined that preserved only the columns of the data frame including venue name and any other category associated with the restaurant location. This was sufficient enough to determine distance of each venue from the city center of Washington, D.C.

Additionally, popularity of each venue needed to be incorporated into this data frame to help assess the role of another independent variable in determining factors that influence decision-making of Italian restaurants to visit in the nation’s capital. In order to explore this, venue ID’s were extracted from the dataset and a loop was performed on each ID to create a list of the number of Foursquare users likes from each venue. This list was then transformed into a data frame and merged to the original data frame so that a new column was added and the number of likes for each Italian restaurant was incorporated for each row. The data frame was sorted into descending order based on the number of likes to provide the clearest depiction of venue popularity.

Finally, now that the data frame contained both distance and number of likes for each restaurant, the additional columns containing irrelevant location information could be removed. A data frame object that was simply a list of all venue ID’s was created, and this filter object was used to create an even more simplified data frame containing columns: “Name”, “Distance”, and “Likes”. This would allow easy retrieval of the independent variables being assessed in order to move forward with careful statistical analysis.

**3. Exploratory Data Analysis**

**3.1 K-Means Clustering**

One of the first pieces of analysis used to explore the data involved K-Means Clustering, partitioning Italian restaurant venues into groups sharing similar characteristics. This unsupervised machine learning approach allowed careful dividing of the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. The first step of this process required normalizing the feature set, focusing on distance and number of likes from the cleaned data frame in this case, to retrieve an accurate dissimilarity measure. Once values were generated for both of these features for each venue of the dataset, a new cluster data frame was created for all 47 venues depicting the dissimilarity values. The number of clusters, or k, then needed to be initialized to assist with grouping venues based on these dissimilarity values. Given such a small sample size of 47 and a two-dimensional feature set, the appropriate value of k was determined to be 3. At this point, the initial centroid for each cluster was chosen at random and each datapoint was assigned to its closest centroid. Through this algorithm, after 12 simulations of re-calculating the closest centroid for each datapoint, labels of 0, 1, and 2 were assigned to each venue, representing which cluster that particular venue was assigned to. These cluster labels generated through the k-means algorithm were then added to the simplified data frame for easy retrieval.

A screenshot of a cell phone

Description automatically generated

Figure 1. Histogram visualizing the number of Italian restaurants that were delegated to each cluster based on k-means algorithm running 12 times.

To check the centroid values for the data frame, the features of each cluster were averaged together. Then, a new data frame was created, describing the k-means cluster centers for each k-means cluster label generated through the algorithm.

While the histogram was helpful in confirming that the unsupervised machine learning algorithm was executing properly and the appropriate number of clusters was used, a k-means cluster map would be needed to illustrate actual plotting of datapoints. Using the Matplotlib plotting library, a cluster map was created. Using color mapping, an array of colors was used to associate each datapoint to one of the 3 clusters. The cluster labels for each venue that were added to the data frame were looped through and associated with calculated centroids.

A picture containing text

Description automatically generated

Figure 2. Color map depicting each Italian restaurant venue and its associated cluster label. Number of cluster centroids are depicted in the lower left-hand corner and individual datapoints were mapped based on values of features (distance and number of likes). Distance in meters is assigned to the X-axis and Number of Likes is assigned to the Y-axis.

Labels: Red = Cluster 0, Yellow = Cluster 1, Blue = Cluster 2

**3.2 Agglomerative Divisive Clustering**

As an alternative form of clustering to analyze the datapoints for Italian restaurant distance and number of likes from the Washington, D.C. city center, agglomerative divisive clustering was utilized. This form of clustering builds a hierarchy of clusters where each node is a cluster consisting of the clusters of its daughter nodes. Since divisive is indicative of top-down, all of the observations are initially considered as one large cluster and broken down into smaller pieces. This machine learning model functioned in a manner similar to the k-means analysis performed above. The number of clusters was set to 3 and new labels were assigned to each venue within the dataset. The new labels assigned to each venue were then plotted as a scatter plot and the image was nearly identical to that shown above, but without cluster centroids.

A close up of a piece of paper

Description automatically generated

Figure 4. Scatter plot of number of likes (Y-axis) vs. distance from city center (X-axis). New cluster label colors were used for agglomerative divisive clustering.

Labels: Red = Cluster 0, Blue = Cluster 1, Green = Cluster 2

**3.3 Density-Based Spatial Clustering of Applications with Noise**

The final method of analysis performed on the filtered data frame involved a third clustering method known as Density-Based Spatial Clustering of Applications with Noise, or DBSCAN. This form of clustering seemed appropriate as spatial data was examined in the form of a scatter plot comparing the two variables: number of likes and distance from the city center. Since definitive differences in the previous clusters discussed may not have accounted for completely accurate grouping by similarity, it seemed appropriate to test all of the venues using an additional method. Also, DBSCAN is especially useful in detecting outliers within the data, in this case the 3 restaurants that experienced a much higher volume of positive feedback than the others. In this sense, DBSCAN identifies arbitrarily shaped clusters without being significantly affected by these 3 outlier datapoints. It works on the notion that if a particular point belongs to a cluster it should be to a handful of other points in that cluster.

In order to successfully carry out analysis with this method, two parameters, eps and minimum points, are used within the algorithm. Eps determines the maximum distance between two samples for one to be considered as in the neighborhood of the other. The minimum points parameter determines the minimum number of datapoints within a neighborhood in order to define a cluster. Additionally, through setting a colormap parameter to a color sequence that differentiates datapoints based on density, relevance of each point can be assessed in reference to the problem at stake.

A screenshot of a cell phone

Description automatically generated

Figure 5. Scatter plot of number of likes (Y-axis) vs. distance from city center (X-axis) utilizing DBSCAN. Datapoints with darker colors are more indicative of restaurants that have more favorable popularity and distance to the city center. Lighter colors have a much larger margin for these two variables and are more difficult to interpret.

**4. Results/Discussion**

**4.1 K-Means Clustering Results/Discussion**

Based on the k-means machine learning model generated, there were several insights gathered. One key finding was that cluster 1 had significantly fewer venues with this label than both clusters 0 and 2. While this could be indicative of outliers within the dataset, it also could show that these points might be of more interest to tourists deciding which Italian restaurant to choose when accessing numerous options. Additionally, the fact that only 3 restaurants were part of this cluster helps show the limitation of a smaller sample size (47 venues) for this particular study. If the radius and limit were greater, especially in city like Washington, D.C., there could likely be more restaurants associated with this cluster and thus, a more accurate depiction that this cluster might be something for tourists to focus more on when deciding restaurants to choose from.

Also, another major finding from this model was that in general, the 3 clusters generated could be categorized as such:

* Cluster 0 = Far from city center, low number of likes
* Cluster 1 = Close to OR far from city center, high number of likes
* Cluster 2 = Close to city center, low number of likes

While these are simply generalizations based on a small sample size, such results could be especially relevant to the specific preferences of tourists making the decision of which Italian restaurant to go to. However, it should be noted that cluster 1 experienced a significantly higher number of likes compared to the other two, so it may be hard to gauge exactly how accurate these popularity ratings truly are. Additionally, because the number of likes was limited to only Foursquare account holders who choose to review a venue out of personal interest, it is likely that most of restaurants from this small sample are underrated in popularity and aren’t given a fair evaluation. If another API endpoint system were utilized, such as Google Maps, a more accurate representation of popularity and positive ratings could be used. This would provide those visiting the D.C. area with a more realistic approach to their decision-making process.

**4.2 Agglomerative Divisive Clustering Results/Discussion**

Based on the agglomerative divisive clustering machine learning model, it is clear that these new cluster labels are more oriented towards distance as a feature opposed to the number of likes. Thus, the 3 clusters generated could be categorized as follows:

* Cluster 0 = Closest to city center
* Cluster 1 = Moderately close to city center
* Cluster 2 = Furthest from city center

This division of clusters could be representative of the different options for restaurants when a group of tourists was more interested on how close a restaurant was to the Washington, D.C. center rather than its popularity. Although this may hold true for some samples of the relevant population, it would be more accurate to assume that most people who focus more closely on both distance to the city center and number of likes rather than one or the other. Perhaps, if there was a more uniform margin assessing the likes each venue received from Foursquare users, the agglomerative divisive clustering algorithm would have likely factored popularity into its clustering to a larger degree. This would have provided more representative clusters considering most of the data points depicted have about 50 or fewer user likes provided through the Foursquare app. Additionally, if the number of clusters value was increased to 4, it is likely that the 3 outlier datapoints with much higher popularity would have been formed into a cluster of their own. If number of likes was a priority, then relevant decision-makers would likely seek out Italian restaurants of this group.

**4.3 Density-Based Spatial Clustering of Applications with Noise Results/Discussion**

In this experiment, since the majority of datapoints were very close together, a small eps value of 0.123 was used as to ensure the local radius for expanding clusters was very low. Additionally, in order to account for the outliers that were spaced much further apart than the rest of the datapoints, a minimum points value of 1 was used in the DBSCAN algorithm. Through utilizing a “plasma” colormap to differentiate between the various points in the scatter plot, a reasonable assumption was made that darker datapoints would be better options for Italian restaurants to explore while lighter datapoints would be of less interest. With this generalization, navy blue points were venues that received greater than 50 likes on Foursquare and/or were within 1000 to 5500 meters of the city center coordinates. Magenta datapoints fell within a similar distance (closer to city center for some) but received about 10 to 50 likes. It was more difficult to interpret orange and yellow points since they were spaced much closer together, but it was assumed that such datapoints received the fewest number of likes from Foursquare users and spanned from the closest Italian restaurants to the city center to the furthest Italian restaurants to the city center (orange being closer to city center and yellow being further).

While this model seemed to represent the most accurate results in terms of restaurant recommendations, there were a few setbacks and areas for improvement. For example, since there were only 47 datapoints being plotted here, this easily could have slightly altered density values for each point. More Italian restaurants to choose from could have allowed better precision for these results. Additionally, perhaps altering the eps value to a slightly higher number may have also altered the results and experimenting with more values would be necessary to differentiate between the most accurate algorithm. Despite these few setbacks, this model still proved to be better overall than other two used to test the data.

**5. Further Limitations and Recommendations**

After moderate success in identifying Italian restaurants of interest, it was very apparent that several limitations were experienced during data exploration. For example, one aspect that could be improved was the address used to find the city center coordinates. “Washington, District of Columbia” was used as the pertinent location to reference the 47 venues, but in order to allow for more accurate restaurant results, it would be useful to use a more specific location address. Since the actual city center (where all 4 quadrants meet) is located within the United States Capitol building, perhaps using the coordinates of this location could have eliminated Italian restaurants returned that were unreasonably distant and thus, not beneficial for tourists to consider.

Another limitation to consider was the fact that not all venues returned by the Foursquare API Search were in fact Italian restaurants. For example, the third result in the data frame was “Rita’s Italian Ice & Frozen Custard”. While this result has the phrase “Italian” in the title, it is actually an ice cream shop and could be misleading to groups hoping to dine at a restaurant serving actual Italian food. This limitation is somewhat difficult to fix through the Foursquare search itself, but if users are able to identify cases like this, then they can remove these rows from the data table so they aren’t accounted for during analysis.

As several of the venues were also scattered around the central city coordinates with no consideration for neighborhood, it may have also been useful to factor in neighborhoods to the search query. Through creating an algorithm that generates neighborhoods of equal size around the city center, it would have been much easier to group Italian restaurants within the same neighborhood to simplify decision-making. Also, if some neighborhoods contained other tourist attractions that made it more convenient to choose from since groups of people might have already been traveling through this area, these neighborhoods would have likely been more of interest. Generating a map of Washington, D.C. containing neighborhood borders could have also been helpful to simplify this process and even filter out some of the venues that were returned in areas of Arlington, Virginia.

Lastly, since analysis was limited to a simple search query and a few clustering methods, this study experienced the limitation of statistical analysis. Results depicted through color mapping were somewhat subjective to interpretation as no direct calculations were displayed in doing so. Through additional analysis methods such as linear regression, descriptive statistics, and correlation values, the data retrieved could have easily been made much more translatable and less subjective. Actual mathematical values could have been used to help provide stakeholders with a more accurate idea of the weight each independent variable possessed when utilizing each clustering method. Also, if more independent variables were factored into the models, this study would likely have been more realistic in considering additional factors that stakeholders, especially tourists, use when evaluating the best Italian restaurant to visit while in the D.C. metropolitan area.

**6. Conclusion**

In this study, the relationship between distance from the Washington, D.C. city center and the number of Foursquare user likes was analyzed to determine the most suitable Italian restaurants for tourists to dine at while visiting. Foursquare API Search was utilized to retrieve information pertaining to relevant venues within a defined radius of the D.C. city center address. Then, clustering models such as K-Means, Agglomerative Divisive Clustering, and DBSCAN were built to further analyze these restaurants and explore whether distance to city center or number of likes possessed more value in helping decide the best venue options. All three of these models were unique, yet supportive in Italian restaurant choice assessment for the relevant stakeholders. However, due to number of limitations experienced during this study, there is no doubt that there are several areas to improve if a similar study were conducted in the future.