Restaurant Location Optimization in Washington D.C.

Richard Boulet

June 30, 2020

1. Introduction

Washington D.C. is home to more than 2,200 restaurants, all packed into an area of only 68 square miles. With numerous options for the consumer at every street corner, picking the best location to open a new restaurant is critical. By observing where existing similar restaurants are located in the city, and how accessible those areas of the city are, our goal is to assist our client in finding the optimal location to open a new high-end French restaurant in Washington D.C.

2. Data

In order to assess the best location for our new restaurant, we will use two sources for data. The first is the Foursquare API, and the second is the DC Open Data API.

2.1. Foursquare API

The Foursquare API provides an easy and free way to find the locations of other restaurants in the D.C. area, and we can use a category identifier in the API call to only return restaurants that meet a criteria, such as French. Additionally, the Foursquare API can give us additional metrics about the French restaurants such as amount of likes, the overall rating of the restaurant, and the price tier of the restaurant. These features in our final data set will aid in clustering our Census Tracts in unique ways. You can find the Foursquare API here.

2.2. DC Open Data API

In addition to our Foursquare data, DC Open Data is a great source for finding data related to DC, and many of it is even already nicely

formatted into shapefiles or GeoJSON files for import into mapping packages. The DC Open Data portal can be found here. From the data portal we will be obtaining the geographic boundaries of US Census Bureau Tracts, which are Census-defined geographic areas, and their respective median income values of the people living within these tracts. Additionally, we will also obtain a data set of D.C. Metro station entrances, as well as the roadway network in D.C. with average daily traffic values. These two transportation data sets will help us identify clusters that are easily accessible, which is ideal when opening a new restaurant.

3. Methodology

To obtain our final feature data set, which we will use to perform k-means clustering analysis on, we start by cleaning our data, exploring our data, accounting for any anomalies, and then merging our separate data sets together.

3.1. Data Cleaning

In order to speed up cleaning the data from the APIs, and ultimately JSON files, a new class was defined that allows for methods to be used on that class. This means that for each JSON object that was passed to the class, a series of operations were performed on the file, and an output was created. In this instance, a temporary normalized dataframe was used with json_normalize function in Pandas, and the columns that were renamed by the user were the only columns kept. This allowed far less code to be required for dropping and renaming the numerous unnecessary columns that are imported from json_normalize function.

After data was cleaned, each independent data set was loaded into a GeoPandas dataframe, which allowed the user to set the geometry used when mapping geographically. This is especially important when mapping polygons, like our Census Tracts, since they have a unique geometry versus single latitude and longitude pairings. Each of these GeoPandas dataframes was then mapped. Additionally, each data set was checked to identify any odd outliers in the data, as well as any null values.

When dealing with null values, each null was dealt with independently in each data set, prior to merging the data sets, to aid in keeping the use of the .fillna function simpler. For median income by census tract, the median value of the entire data set was used to fill missing values. For ratings and price tiers of other French restaurants, a median value was also used in order to not bias the specific restaurant as appearing to have zero rating, when it may be possible that in reality it may be a good place to eat but had no current reviews, and would be clustering into low-tier rated restaurants.

3.2. Exploratory Analysis

Once our data sets were clean and modified to be mapped geographically, each data set was combined into a map, a choropleth of the Census Tract median income bins used as a background. This allowed for easy exploration of obvious clusters of restaurants located in high income areas of D.C.

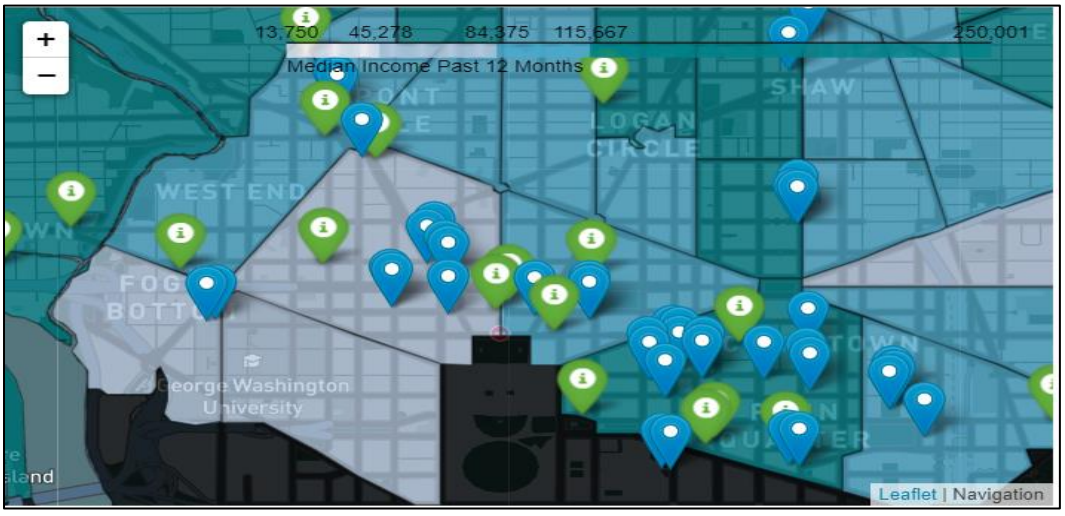


Figure 1 - D.C. Restaurants and Metro Stops in Relation to Median Income

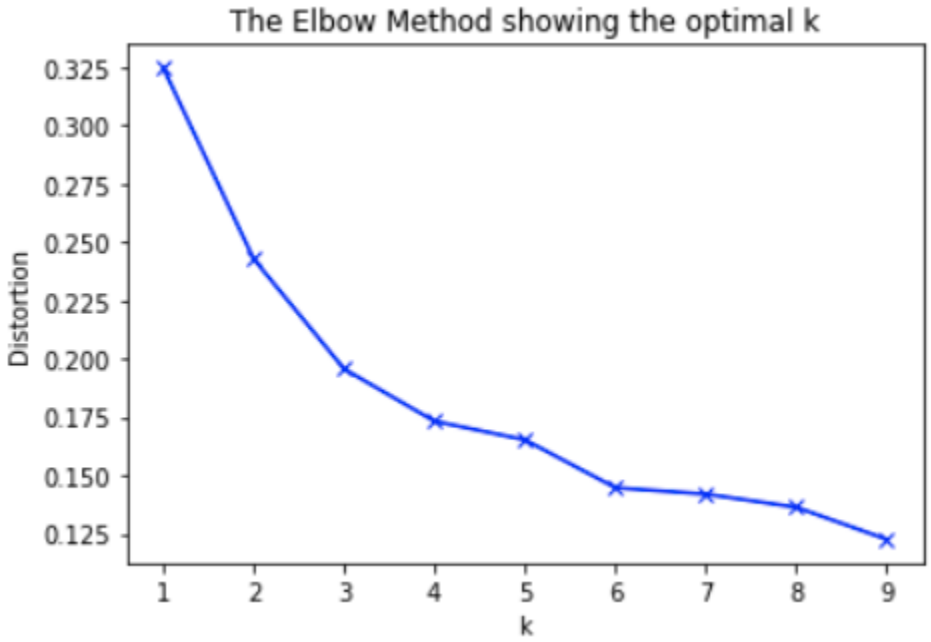
After plotting all of the data, the decision was made to choose the Census Tracts to act as the location targets for our new restaurant. Once this was decided, the data sets were all merged to create our feature data set.

3.3. K-Means Clustering

Based on the mapping techniques that we learned, I settled on using k-means clustering in order to assess the best Census Tracts that shared similar ideal characteristics. In the end I had 7 features that were used to cluster our target Census Tract locations by:

Feature	Source
Median Income	DC Open Data
Count of Metro Entrances	DC Open Data
Count of French Restaurants	Foursquare
Average Daily Traffic	DC Open Data
French Restaurant Likes	Foursquare
French Restaurant Ratings	Foursquare
French Restaurants Price Tier	Foursquare

Using these features, I was able to perform k-means clustering with various starting amounts of clusters and observed the point in which distinctions were not being made as a higher cluster value was used and shown below in the Elbow Method chart.



From the chart above I chose to use a starting cluster amount of 4 when performing k-means clustering on the feature data set.

4. Results

Based on using a starting cluster amount of 4 clusters, we ended up with some interesting results. In the tables below, we can see the differences between the clusters, and why Cluster 2 and Cluster 3 are ideal locations to open our high-end French restaurant in. In Figure 2, we can see that on average Cluster 2 has a very high median income, lots of metro entrances, large amounts of traffic, as well as other competitors that are well-liked. It also appears in Figure 3 that much of Cluster 2 is located in the Downtown area of D.C. where many restaurants currently already exist. This shows us that it may be an excellent place to open our new restaurant, given that the client is open to accepting lots of local competition.

	median_income	metro_sum	price_tier	likes	rating	french_sum	avg_daily_traffic
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	125805.071429	2.785714	2.814286	125.533333	7.942619	2.000000	10914.399124
std	42272.698222	4.209487	0.546115	306.796917	0.844921	1.300887	2445.197123
min	73688.000000	0.000000	2.000000	0.000000	5.700000	1.000000	8206.244898
25%	98047.250000	0.000000	2.350000	17.500000	7.835000	1.000000	9649.196139
50%	106161.500000	0.500000	3.000000	30.500000	8.075000	1.000000	10267.346734
75%	144723.000000	3.750000	3.000000	71.166667	8.391667	3.000000	11927.234037
max	217708.000000	14.000000	4.000000	1181.000000	9.200000	5.000000	15864.302083

Figure 2 - Cluster 2 Descriptive Statistics

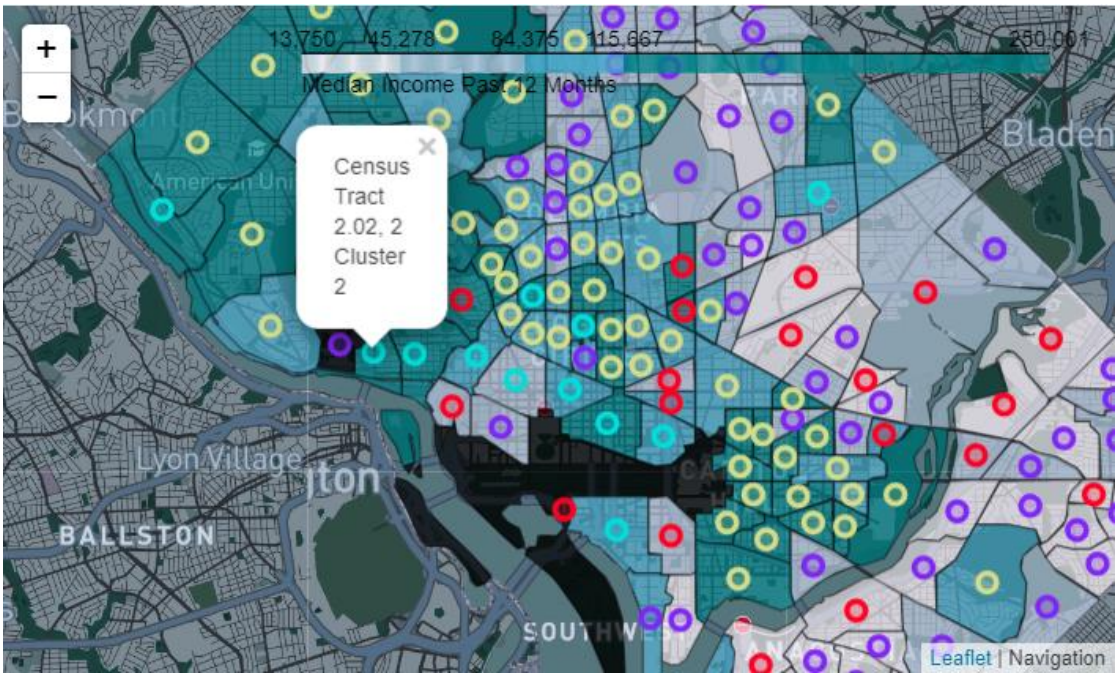


Figure 3 - Cluster 2 Mapping

On the other hand, Cluster 3 was also shown to have some promise, and far more locations, to open our restaurant in. In Figure 4 we can see that on average in Cluster 3 areas, the median income was even higher than Cluster 2, but there is little to no competition in the form of other French restaurants. Additionally, there appears to be very little metro access on average in the Cluster 3 areas. However, there is plenty of roadway access and nearly as much daily traffic on average when compared to the Cluster 2 areas.

	median_income	metro_sum	price_tier	likes	rating	french_sum	avg_daily_traffic
count	65.000000	65.000000	65.0	65.0	65.0	65.0	65.000000
mean	127458.261538	0.615385	0.0	0.0	0.0	0.0	9183.159392
std	31424.980699	1.155047	0.0	0.0	0.0	0.0	1208.422356
min	90000.000000	0.000000	0.0	0.0	0.0	0.0	6196.321429
25%	102424.000000	0.000000	0.0	0.0	0.0	0.0	8539.781609
50%	117045.000000	0.000000	0.0	0.0	0.0	0.0	9073.440476
75%	144922.000000	1.000000	0.0	0.0	0.0	0.0	9862.611111
max	250001.000000	4.000000	0.0	0.0	0.0	0.0	12048.457143

Figure 4 - Cluster 3 Descriptive Statistics

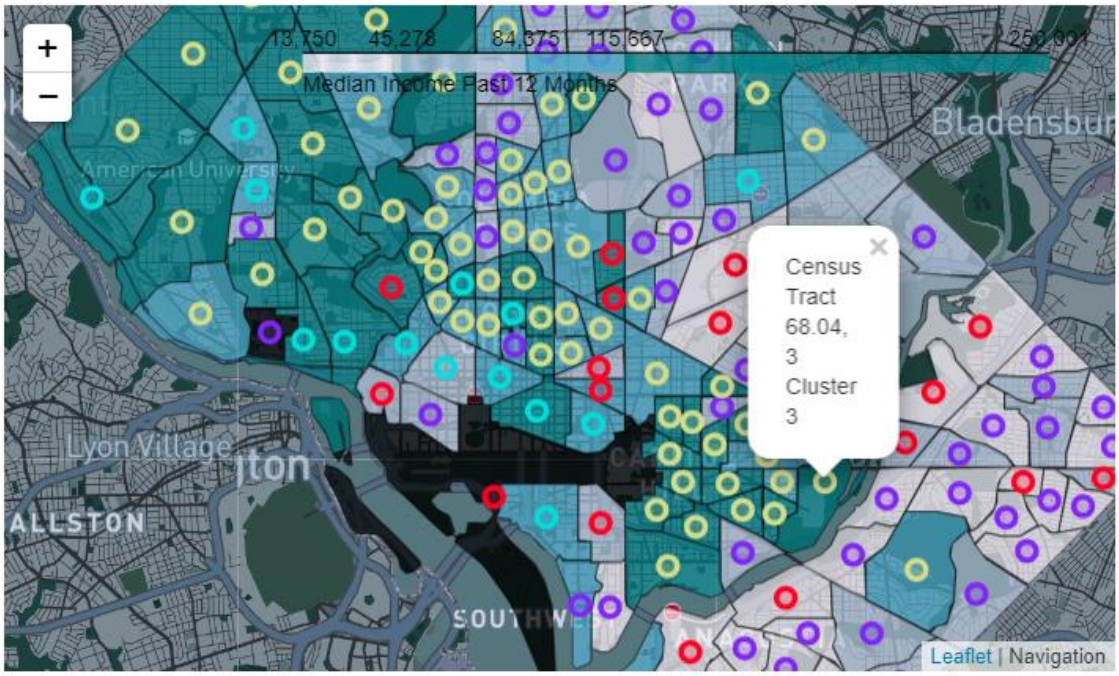


Figure 5 - Cluster 3 Mapping

5. Conclusions

In this case, when our client asks us what the best location is for their new high-end French restaurant, I would recommend that they open their new restaurant in one of the Census Tracts in Cluster 2, as long as they are comfortable with competition. On the other hand, if the client wants a quieter location, but still a large base of potential customers with low amounts of local competition, then I would recommend that they choose a Census Tract in one of the Cluster 3 Census tracts.