Finding the Ideal Location for a Mexican Restaurant in Boston, using k-Means Clustering

1. Introduction/Business Problem

A client wants to open a brand-new Mexican restaurant in Boston. The location of a restaurant can be crucial for its success. Thus, finding the ideal place to establish it is a very important step that has to be planned meticulously. The location of a restaurant can be determined by many factors, such as per capita income of a neighborhood, accessibility (e.g., metro stations) and how many similar restaurants exist in the area. Since it is a Mexican restaurant it is thoughtful to consider the number of Mexican people that live in the area. Acquiring the above data will help us to suggest potential locations to the stakeholders which will lead to their business success.

2. Data Acquisition and Cleaning

As mentioned in the introduction, when considering opening a restaurant it is useful to find data about the area, in this case Boston, and potential customers with specific demographics (e.g., high income, ethnicity). Consequently, finding data about Boston and its neighborhoods is an essential part of the project. For each neighborhood information is needed about the total population, Mexican population, and per capita income. Such data can be found here. Furthermore, using Foursquare's API we will obtain the number of Mexican restaurants and other venues in each neighborhood. The number of Mexican restaurants is an indicator of the demand and the competition that exists in the area. Existing demand suggests valuable information and having quality data is key to an accurate prediction.

In order to use the data correctly they should be organized in a single table and decide to erase or fill any missing data. To begin with, we formed a .csv file which contained a table with all the information that was necessary for the project (e.g., neighborhoods, total population, Mexican population, per capita income). The data file was loaded to Jupyter Notebook and a pandas data frame was created.

	Neighbourhood	Total Population	Mexicans	Per capita income
0	Allston	19,363	73	$28,986
1	Back Bay	18,176	119	$98,495
2	Beacon Hill	9,751	0	$90,227
3	Brighton	51,785	109	$35,876
4	Charlestown	18,901	12	$69,219

Figure 1: The created pandas data frame

The next step was to clean the data. For instance, we erased commas and the sign dollar from each value of any column. That way the values were more readable. After renaming the columns we ended up with the table below.

	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)
0	Allston	19363	73	28986
1	Back Bay	18176	119	98495
2	Beacon Hill	9751	0	90227
3	Brighton	51785	109	35876
4	Charlestown	18901	12	69219

Figure 2: Updated data frame with changed column names

3. Methodology

3.1. Exploratory data analysis

The next phase of the project was to visualize the neighborhood on a map. For that reason, we needed to obtain the geospatial data for each neighborhood. We used the "geopy" library to do that.

	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude
0	Allston	19363	73	28986	42.355434	-71.132127
1	Back Bay	18176	119	98495	42.350549	-71.080311
2	Beacon Hill	9751	0	90227	42.358708	-71.067829
3	Brighton	51785	109	35876	42.350097	-71.156442
4	Charlestown	18901	12	69219	42.377875	-71.061996

Figure 3: Data frame with geospatial data

Moreover, it was decided to drop the rows with null values because they did not offer quality information to our research.

	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude
0	Allston	19363	73	28986	42.355434	-71.132127
1	Back Bay	18176	119	98495	42.350549	-71.080311
2	Brighton	51785	109	35876	42.350097	-71.156442
3	Charlestown	18901	12	69219	42.377875	-71.061996
4	Dorchester	125947	229	26292	42.297320	-71.074495

Figure 4: Updated data frame after erasing null values

Using the folium library, a map of Boston was created, and each neighborhood was visualized with a blue marker.

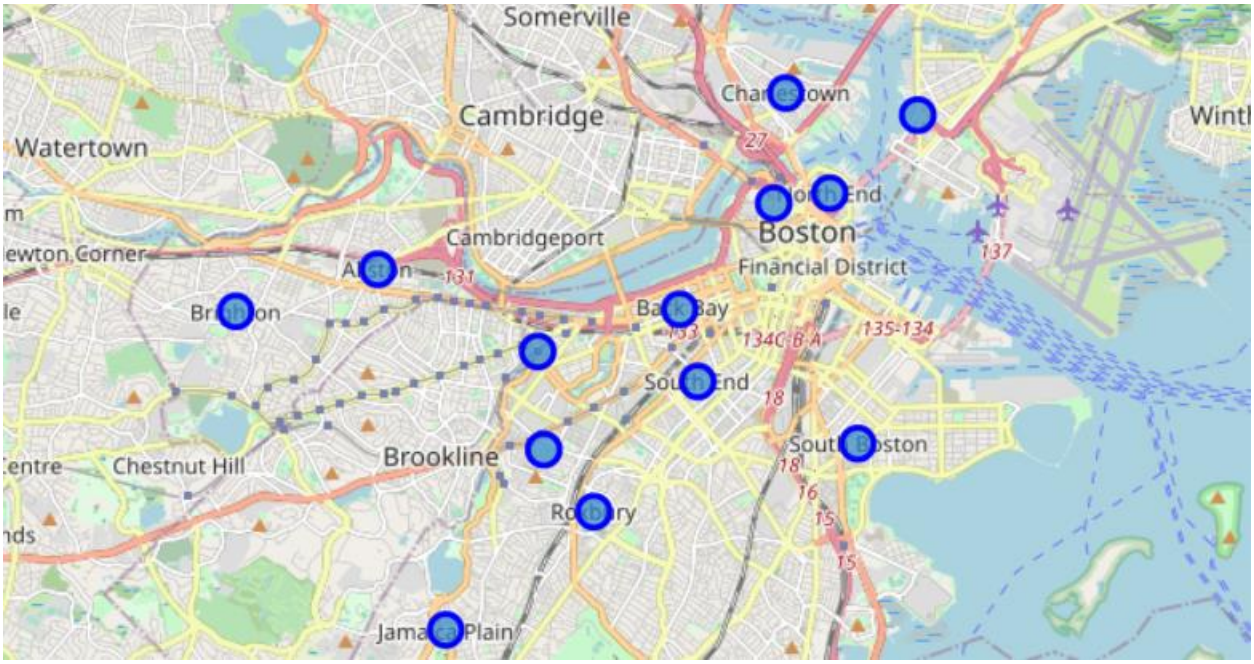

Figure 5: Map of Boston with neighborhoods using Folium library

As already mentioned, the number of Mexican restaurants that already exist in the area is a key indicator of the existing but also future demand. Thus, it is of great importance to locate and count similar restaurants around each neighborhood. Using the Foursquare API's explore function we were able to find Mexican restaurants in each neighborhood and plot them.

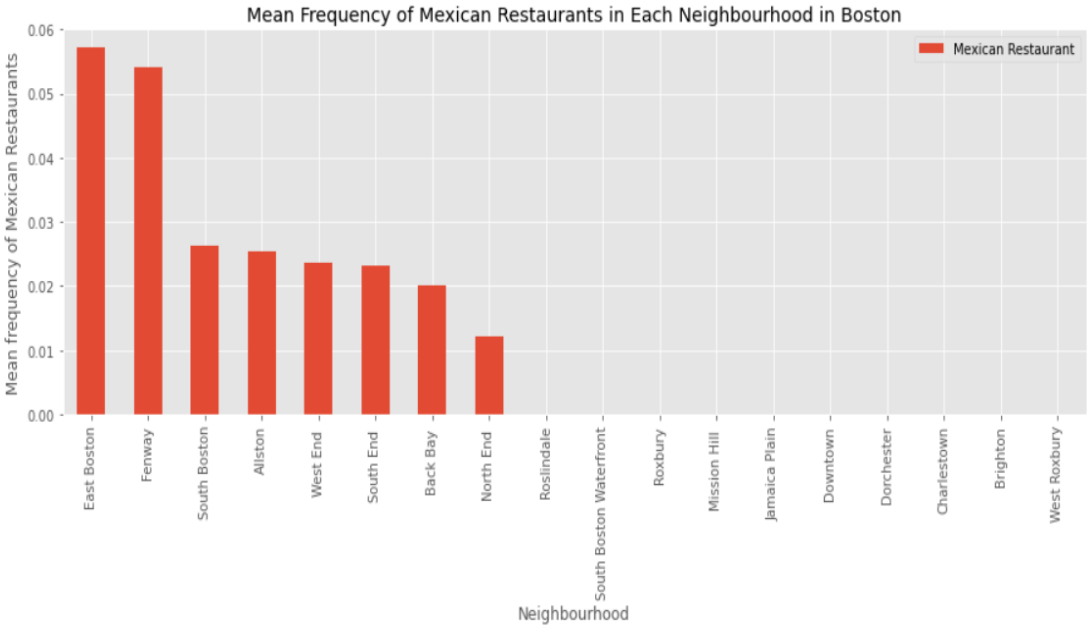

Figure 6: Frequency of Mexican Restaurants in Each Neighborhood in Boston

Apart from Mexican restaurants, we assume that neighborhoods with high Mexican population means more potential customers for a new business. In other words, we hypothesize that there is positive correlation between Mexican population and the demand for a Mexican restaurant. As above, we plot the neighborhoods according to their local Mexican population.

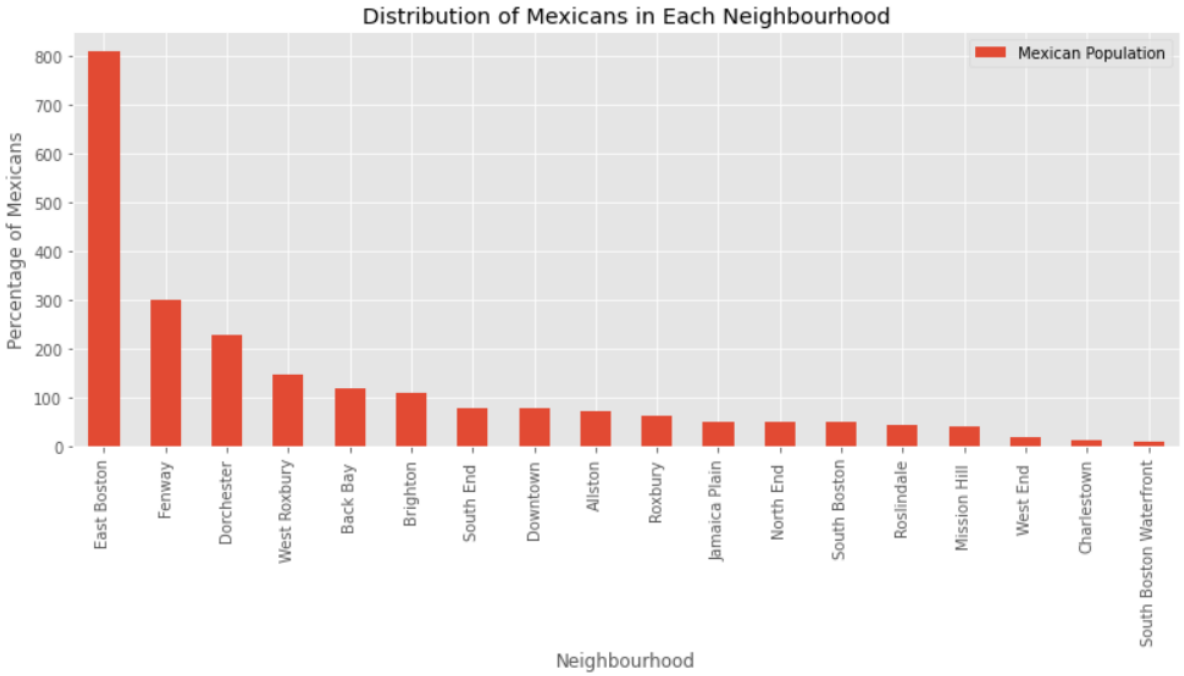

Figure 7: Distribution of Mexicans in Each Neighborhod

Undoubtedly, the distirbution of income plays a vital role in the future demand and consequently in the location of the restaurant. Again we plot the distribution of per capita income in each neighborhood.

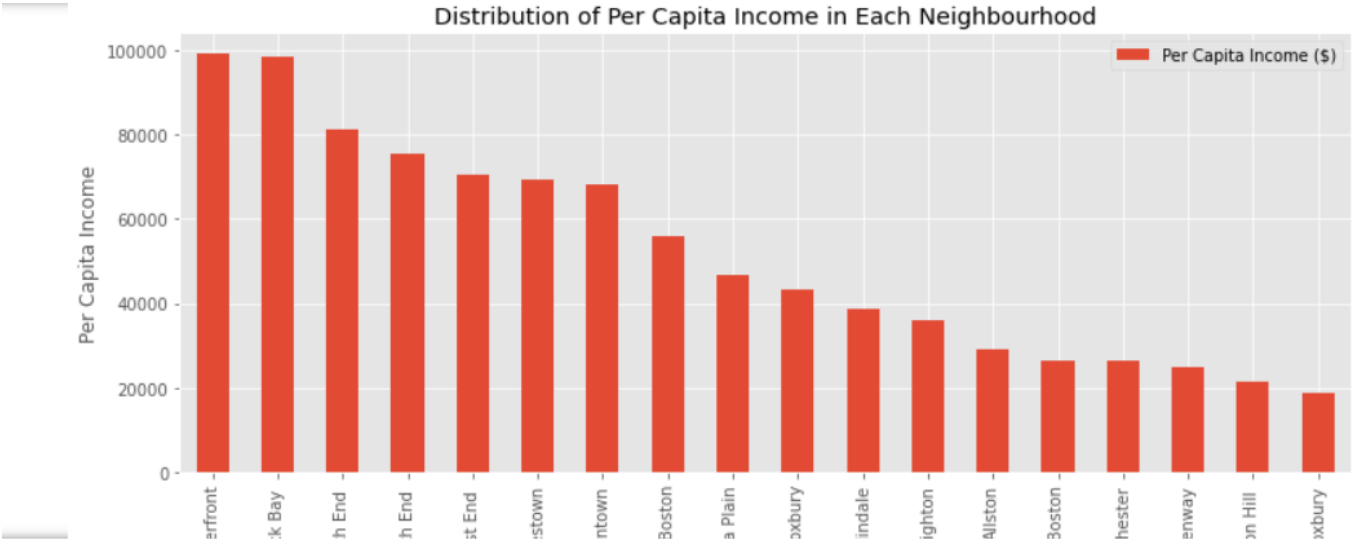

Figure 8: Distribution of Per Capital Income in Each Neighborhood

3.2 K-means Clustering

Our final data frame consists of 6 columns as shown below.

	Neighbourhood	Latitude	Longitude	Mexican Population	Per Capita Income ($)	Mexican Restaurant
0	Allston	42.355434	-71.132127	73	28986	0.025316
1	Back Bay	42.350549	-71.080311	119	98495	0.020000
2	Brighton	42.350097	-71.156442	109	35876	0.000000
3	Charlestown	42.377875	-71.061996	12	69219	0.000000
4	Dorchester	42.297320	-71.074495	229	26292	0.000000
5	Downtown	52.971149	-0.059809	77	67994	0.000000
6	East Boston	42.375097	-71.039217	810	26569	0.057143
7	Fenway	42.345187	-71.104599	302	24997	0.054054
8	Jamaica Plain	42.309820	-71.120330	51	46721	0.000000
9	Mission Hill	42.332560	-71.103608	40	21386	0.000000
10	North End	42.365097	-71.054495	49	81149	0.012048
11	Roslindale	42.291209	-71.124497	45	38760	0.000000
12	Roxbury	42.324843	-71.095016	63	18932	0.000000
13	South Boston	42.333431	-71.049495	49	55709	0.026316
14	South Boston Waterfront	40.318570	-79.817939	10	99073	0.000000
15	South End	42.341310	-71.077230	79	75387	0.023256
16	West End	42.363919	-71.063899	18	70389	0.023529
17	West Roxbury	42.279265	-71.149497	148	43415	0.000000

Figure 9: Final data frame

Because each column is measured in different units it is essential to normalize them in order for K-means clustering method to be effective.

	Mexican Population	Per Capita Income ($)	Mexican Restaurant
0	-0.297045	-0.884043	0.655906
1	-0.043223	1.829594	0.362645
2	-0.098402	-0.615057	-0.740575
3	-0.633635	0.686657	-0.740575
4	0.563742	-0.989217	-0.740575

Figure 10: Normalized values

The machine learning method that we will use is K-means clustering. In order to determine the optimal number of clusters we will use the elbow method as shown below. The optimal number of clusters is where an angle (like elbow) begins to appear.

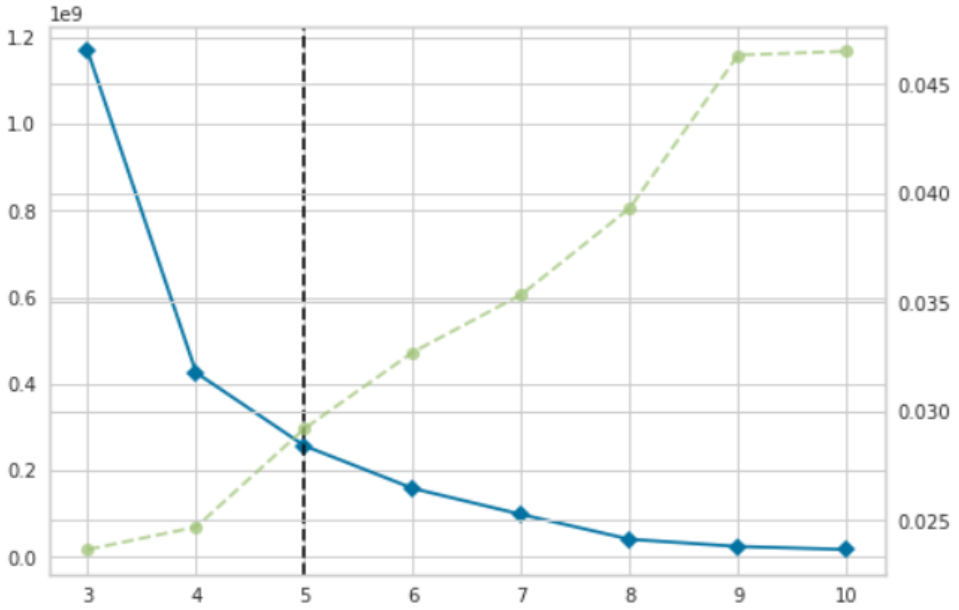

Figure 11: Choosing optimal number of clusters with'' elbow'' method

As shown in the figure we will make five clusters that contain the total data that we have. Then we will find the characteristics for each cluster and decide where the ideal location for a Mexican restaurant in Boston.

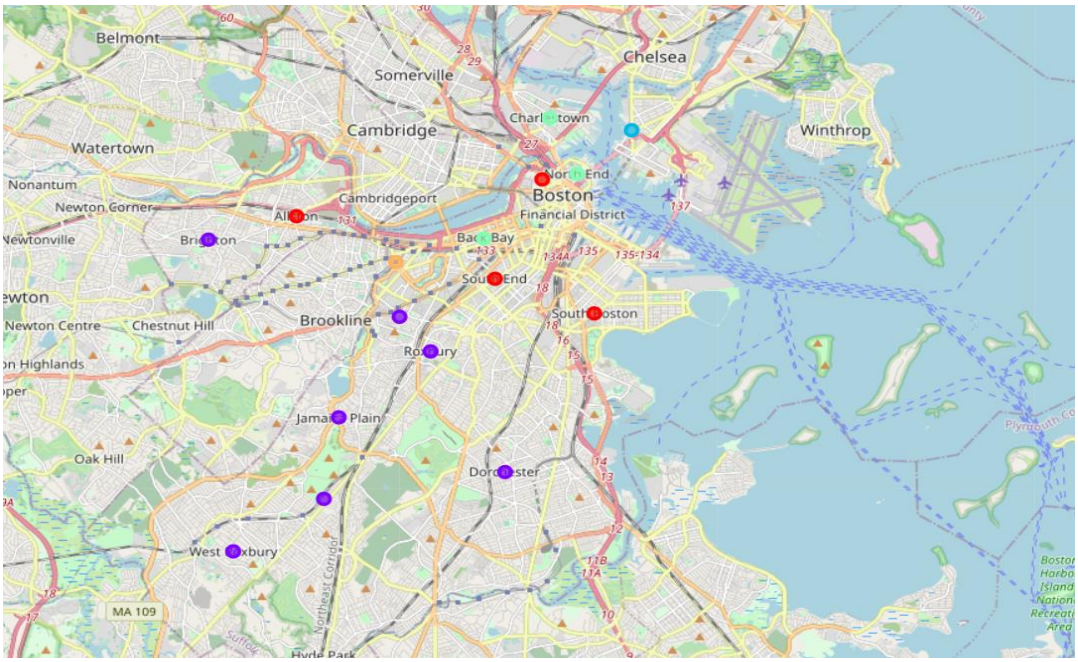

Figure 12: Mapping the clusters

4. Results

Cluster Label 0

As we can see this cluster contains neighborhoods that have many Mexican restaurants which is a good indicator for the current and the future demand. The household income is medium as well as the Mexican population that resides in the area.

	Cluster Label	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude	Mexican Restaurant
0	0	Allston	19363	73	28986	42.355434	-71.132127	0.737351
13	0	South Boston	36212	49	55709	42.333431	-71.049495	0.775507
15	0	South End	32040	79	75387	42.341310	-71.077230	0.544028
16	0	West End	6173	18	70389	42.363919	-71.063899	0.572587

Figure 13: Cluster 0

Cluster Label 1

Cluster number 1 has the most neighborhoods. Specifically, it has a high population of Mexicans, but both the income and the number of competitors is low.

	Cluster Label	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude	Mexican Restaurant
2	1	Brighton	51785	109	35876	42.350097	-71.156442	-0.712572
4	1	Dorchester	125947	229	26292	42.297320	-71.074495	-0.712572
8	1	Jamaica Plain	39314	51	46721	42.309820	-71.120330	-0.712572
9	1	Mission Hill	17406	40	21386	42.332560	-71.103608	-0.712572
11	1	Roslindale	29206	45	38760	42.291209	-71.124497	-0.712572
12	1	Roxbury	52944	63	18932	42.324843	-71.095016	-0.712572
17	1	West Roxbury	33930	148	43415	42.279265	-71.149497	-0.712572

Figure 14: Cluster 1

Cluster Label 2

This cluster has only one neighborhood. It has the highest number of Mexican citizens, but their income is very low. On the other hand, many Mexican restaurants already exist in the area.

	Cluster Label	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude	Mexican Restaurant
6	2	East Boston	46655	810	26569	42.375097	-71.039217	2.518684

Figure 15: Cluster 2

Cluster Label 3

This cluster contains neighborhoods with very high income but not so many Mexican restaurants. Moreover, Mexicans residing in those areas are not so many.

	Cluster Label	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude	Mexican Restaurant
1	3	Back Bay	18176	119	98495	42.350549	-71.080311	-0.147102
3	3	Charlestown	18901	12	69219	42.377875	-71.061996	-0.712572
5	3	Downtown	17581	77	67994	52.971149	-0.059809	-0.712572
10	3	North End	9271	49	81149	42.365097	-71.054495	-0.062607
14	3	South Boston Waterfront	3443	10	99073	40.318570	-79.817939	-0.712572

Figure 16: Cluster 3

Cluster Label 4

Another cluster that contains only one neighborhood. It has a medium amount of Mexican population and very low income. However, Mexican restaurants appear to be in large numbers in the area.

	Cluster Label	Neighbourhood	Total Population	Mexican Population	Per Capita Income ($)	Latitude	Longitude	Mexican Restaurant
7	4	Fenway	32598	302	24997	42.345187	-71.104599	2.187273

Figure 17: Cluster 4

5. Discussion

In this project, we clustered the neighborhoods of Boston according to some characteristics. These characteristics were the per capita income, the number of Mexican restaurants that exist in the area and the Mexican population that resides in each neigborhood.

As we can see from the clusters the ideal location for a Mexican restaurant depends a lot on the target group that the stakeholders have. If the restaurant offers food for a low-normal price then Cluster 2 and Cluster 4 would be very good alternatives. The reason behind this is because these neighborhoods (East Boston, Fenway) have many Mexican residents and a lot of competition as far as Mexican restaurants are concerned. As we have already mentioned, copmetition is a sign of increased demand and that is always good for a new business. However, with all the competitors around it means that there is already restaurants with a brand name and might be a little tough to penetrate the market.

If stakeholders plan to open a restaurant that offers pricy Mexican food then probably cluster 0 would be the first option as it consists of the more wealthy neighborhoods and there is an average competition as far as Mexican cuisine is concerned.

6. Conclusion

Concluding this report, we could presume that this model or this methodology could be used for similar projects that concern new businesses. It is essential to mention that the model could be improved with more accurate data.

6.1. Future Dierections

The model definitely has some room for improvement as it could take into account the style of the restaurants that already existed in each neighborhood. By style, we mean if it is a fine dining restaurant or a take away restaurant. In that way we could measure the competition more precisely and have a more clear option for the final location of the Mexican restaurant.