# Predicting the best neighborhood for a new restaurant

Dharni Shah

April 30, 2020

## 1. Introduction

### 1.1 Background

The quick-service restaurant industry is a crowded market with all the players in the space competing closely for a piece of the revenue pie. Each quick-serve is constantly trying to get increasingly creative with marketing and advertising campaigns in order to win customers over to their restaurants. One of the ways companies are currently gaining a competitive edge is via technology and data-driven methods like real-time geofencing and location intelligence to achieve this. More and more quick-service restaurants and franchise owners are discovering the power of GIS to find the best sites for their restaurants. Owners of successful franchises have relied on GIS technology to discern markets for many years. The technology provides tools that help organize information by using location as the common identifier for data. By understanding where franchises, the competition, and customers are located, franchisors can make informed decisions, improve communication, and share their knowledge with others.

### 1.2 Problem

This project aims to predict the best location options to open a new restaurant in New York city. The neighborhoods are analyzed to find out the best ones for a new restaurant depending on the type, budget and geographical limitations of the restaurant. Exploring the neighborhoods further is to determine which type of cuisine would be favorable for the selected neighborhood. The factors that would be taken into considerations while selecting the suitable neighborhoods location demographics, population of the neighborhood, competitors, restaurant category (American, Italian, etc.) and their locations.

### 1.3 Interest

Quick-serve restaurants and franchise owners would be able to use this technique to make critical investment decisions for new outlets considering all the location demographics, competitors, etc. Site selection and predictive modelling not will assist them in optimizing their process but also will keep them ahead and well-informed of the competitors.

## 2. Data acquisition, cleaning and pre-processing

### 2.1 Data sources

The neighborhood demographics data was obtained from census data https://www.census.gov/ and https://popfactfinder.planning.nyc.gov/ . The data consisting of the venues for each neighborhood was obtained using the Foursquare API for the city of New York.

### 2.2 Data cleaning

The data downloaded by the census website consisted of many demographics out of which only the relevant information was used. Also, the geographical co-ordinates of the neighborhoods are found using geocoder. The final dataframe consists of population size, median income and median age along with the neighborhoods geographical data. The dataframe has 5 boroughs and 172 neighborhoods.

Table 1. Demographics data.

| | Borough | Ncode | Neighborhood | Latitude | Longitude | population total | Median income | Median age |
|---|---|---|---|---|---|---|---|---|
| 0 | Brooklyn | BK45 | Georgetown-Marine Park-Bergen Beach-Mill Basin | 40.623845 | -73.916075 | 48351 | 1520979 | 36.8 |
| 1 | Brooklyn | BK17 | Sheepshead Bay-Gerritsen Beach-Manhattan Beach | 40.577914 | -73.943537 | 61584 | 1054259 | 40.3 |
| 2 | Brooklyn | BK61 | Crown Heights North | 40.670829 | -73.943291 | 100130 | 980637 | 34.6 |
| 3 | Brooklyn | BK90 | East Williamsburg | 40.708492 | -73.938858 | 33155 | 519058 | 34.1 |
| 4 | Queens | QN23 | College Point | 40.784903 | -73.843045 | 24199 | 354073 | 38.7 |

Foursquare API is used to explore the neighborhoods in New York City. The venue category is filtered for “Restaurants” and the number of restaurants in each neighborhood and both the dataframes are merged. Also, for further analysis each neighborhood’s venue data is converted to one hot encoded frame to find which type of cuisine would be a best choice for the new restaurant. In Table 3. column name “Venue” represents the number of restaurants each neighborhood consists.

Table 2. Foursquare API data.

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Georgetown-Marine Park-Bergen Beach-Mill Basin | 40.623845 | -73.916075 | Landi's Pork Store | 40.619633 | -73.917918 | Italian Restaurant |
| Georgetown-Marine Park-Bergen Beach-Mill Basin | 40.623845 | -73.916075 | Chipotle Mexican Grill | 40.626405 | -73.916976 | Mexican Restaurant |
| Georgetown-Marine Park-Bergen Beach-Mill Basin | 40.623845 | -73.916075 | Gourmet Grill | 40.619543 | -73.916111 | American Restaurant |

Table 3. Merged data.

| Borough | Ncode | Neighborhood | Latitude | Longitude | population total | Median income | Median age | Venue |
|---|---|---|---|---|---|---|---|---|
| Brooklyn | BK45 | Georgetown-Marine Park-Bergen Beach-Mill Basin | 40.623845 | -73.916075 | 48351 | 1520979 | 36.8 | 4.0 |
| Brooklyn | BK61 | Crown Heights North | 40.670829 | -73.943291 | 100130 | 980637 | 34.6 | 1.0 |
| Brooklyn | BK90 | East Williamsburg | 40.708492 | -73.938858 | 33155 | 519058 | 34.1 | 2.0 |
| Queens | QN23 | College Point | 40.784903 | -73.843045 | 24199 | 354073 | 38.7 | 8.0 |
| Staten Island | SI11 | Charleston-Richmond Valley-Tottenville | 40.530531 | -74.232158 | 24083 | 342708 | 39.5 | 2.0 |

### 2.3 Data Pre-processing

The merged dataframe is normalized to bring the values between 0-1. The goal of **normalization** is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. The final dataframe consists of 153 neighborhoods with 4 features to be used for clustering. OneHotEncoding was used to create dummy variables for the "Venue Category" as it is categorical variable in the Foursquare API data.

## 3. Exploratory Data Analysis

Folium map is created with neighborhoods are superimposed on top.

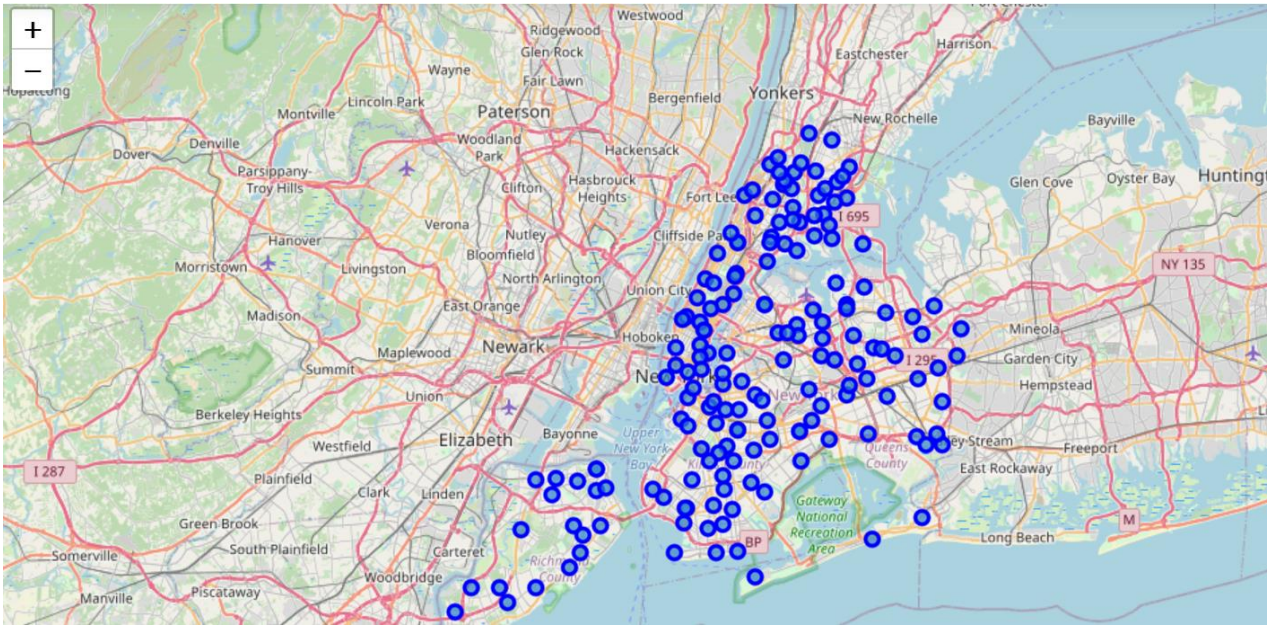


Figure 1. Neighborhoods in New York City.

The frequency of demographics parameters is determined by visualizing the population size, median age and median income of all the neighborhoods.

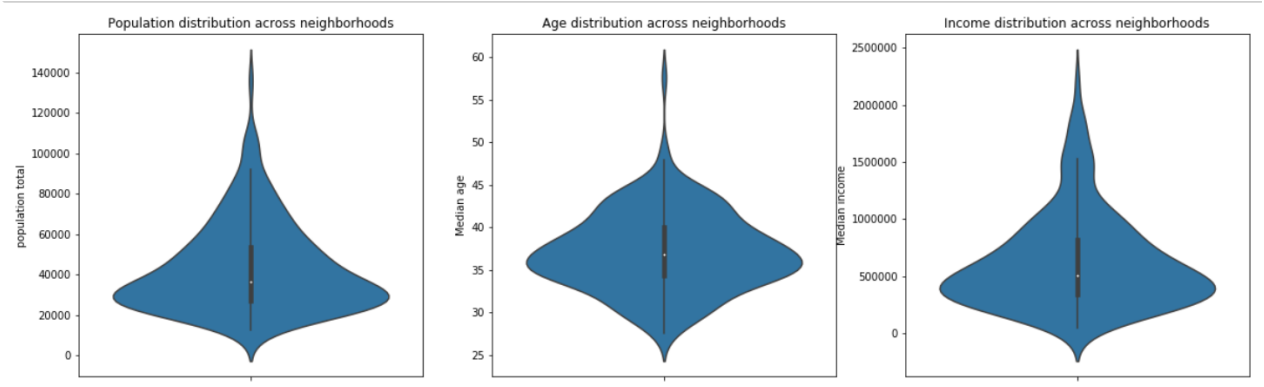


Figure 2. Frequency of Median age, median income and population within the 153 neighborhoods.

Since the median age for majority of the neighborhoods is between 30-45 years, hence it is not a useful feature for clustering. Box plots to better visualize the distribution range.

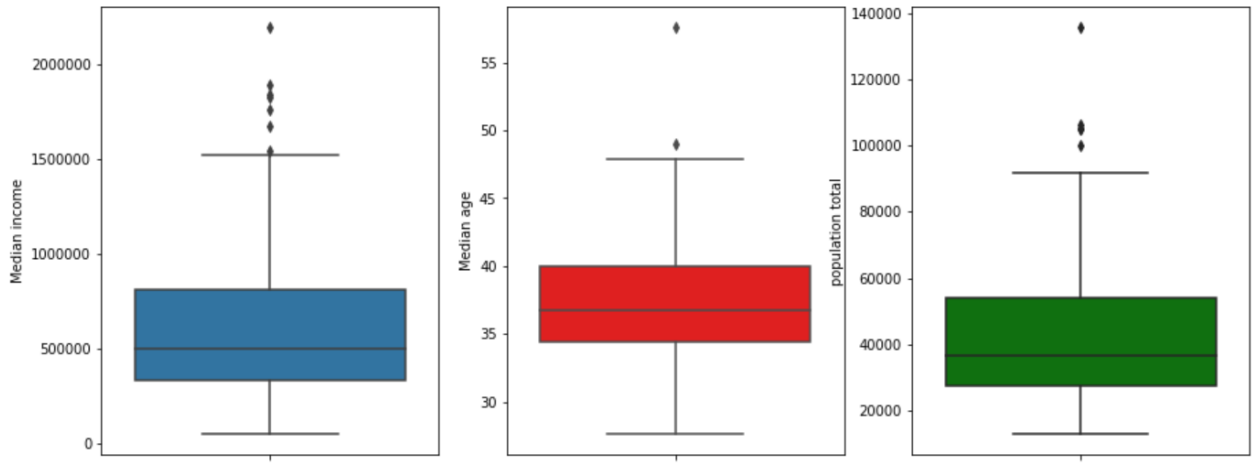


Figure 3. Box plot of Median age, median income and population within the 153 neighborhoods.

## 4. K-means clustering

### 4.1 Within Cluster Sum Of Squares (WCSS)

Within Cluster Sum Of Squares (WCSS) against the the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where *Yi* is centroid for observation *Xi*. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

### 4.2 The Elbow Method

Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow. The optimal K value is found to be 4 using the elbow method.

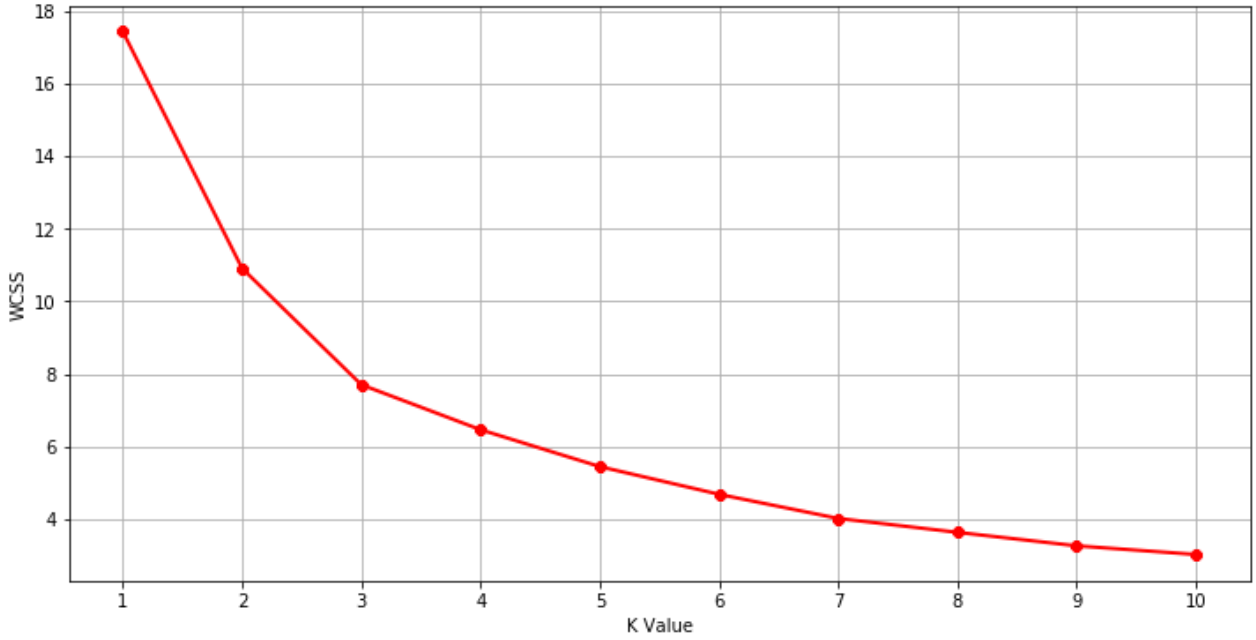


Figure 4. WCSS curve plot.

### 4.3 Results

After examining the centers of 4 clusters. The following description is added across the corresponding labels.

Table 4. Clusters formed using K-means.

| Cluster label | Cluster description |
|---|---|
| 0 | Very small population, low income, very less number of restaurants |
| 1 | Very large population, high income, very high number of restaurants |
| 2 | Small population, very low income, high number of restaurants |
| 3 | Large population, very high income, less number of restaurants |

The clusters formed are visualized using 3D scatter plot. The parameters used for clustering are Median income, population and number of restaurants. There 4 clusters formed which separates the neighborhoods w.r.t these features.

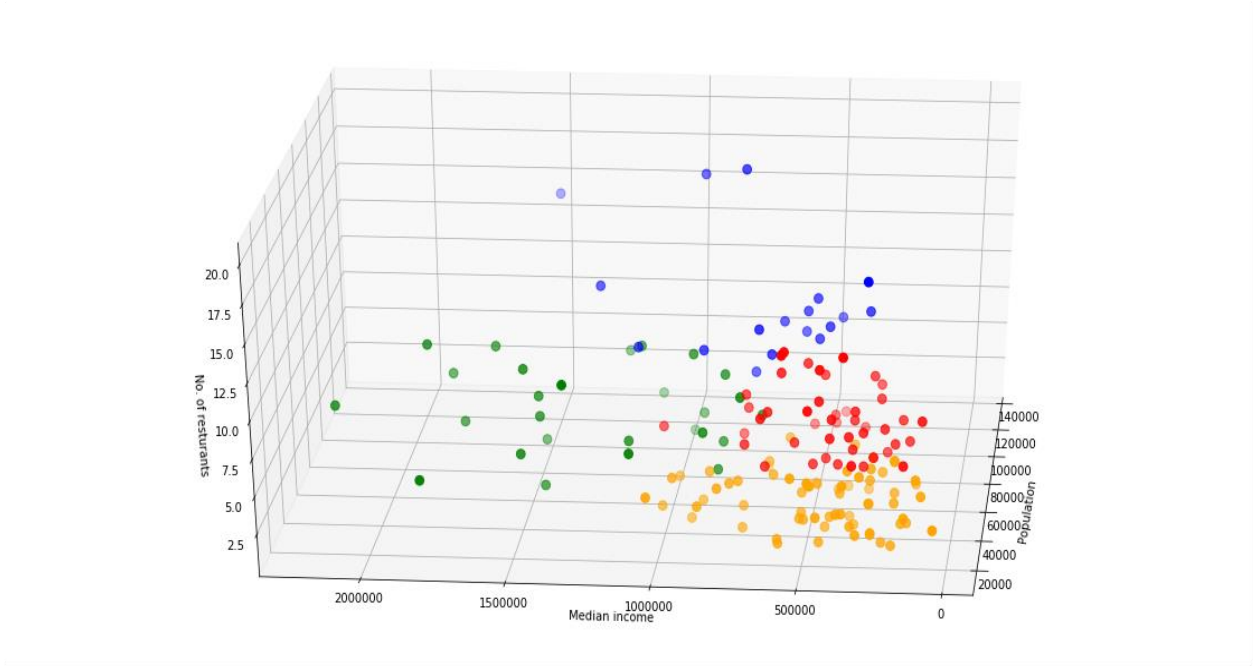


Figure 5. Clusters formed.

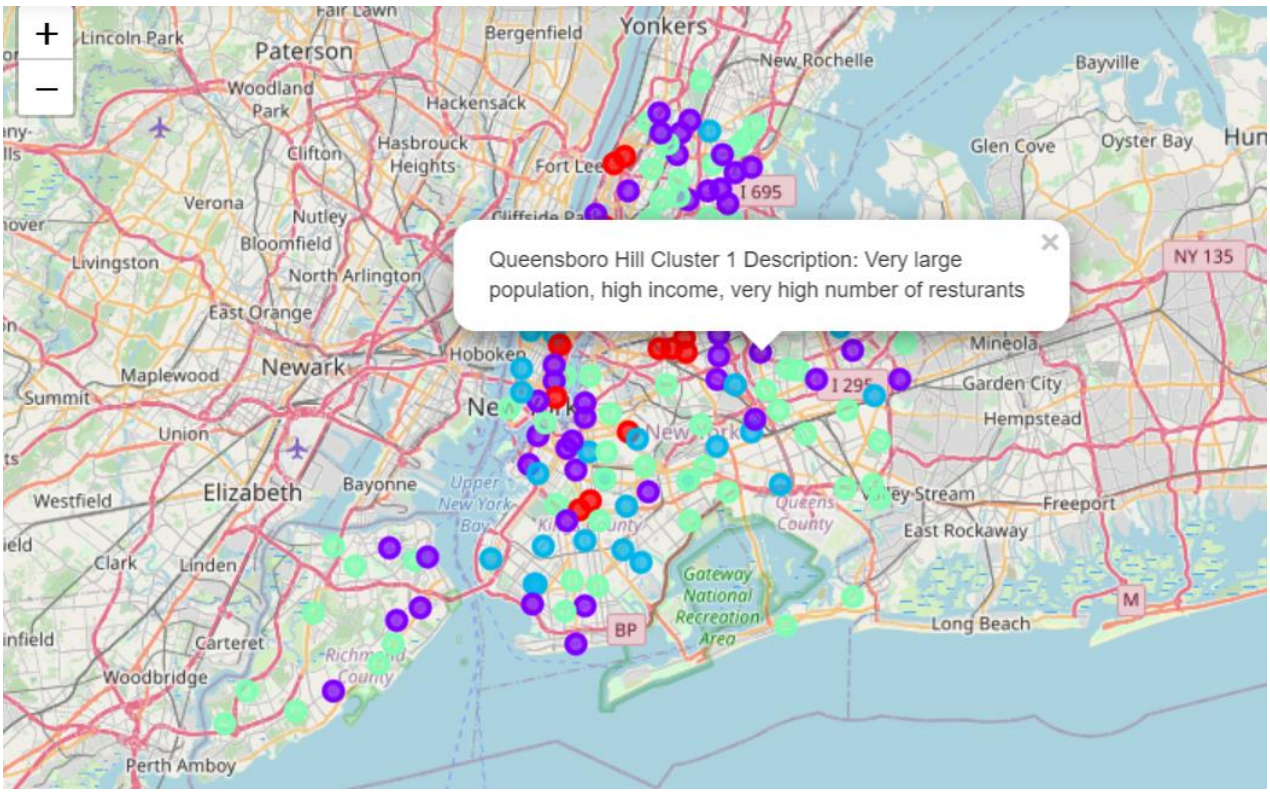


Figure 6. Clusters of neighborhoods.

To visualize which neighborhoods belongs to which cluster, Folium map with neighborhoods superimposed on it was created. The label for each neighborhood shows the cluster number and its description. Further, on using K-means clustering on all the “Venue Category” to predict the best type of restaurant for the new outlet.

## 5. Conclusions

In this study, I analyzed cluster 3 would be the best option from which the neighborhoods for a new restaurant should be selected. Depending on the budget and size of population the owner plans to serve, other clusters can be investigated.

## 6. Future directions

The clusters can be made non-overlapping with more demographics, geographically accessibility of the location and competitors’ data.