IBM DATA-SCIENCE CERTIFICATION

Coursera Capstone

The Battle of Neighborhoods

22.05.2020

The Battle of Neighborhoods

Abstract

This study provides an decision framework for young families willing to start a new life and New York city. By k-Means clustering approaches New York City's boroughs and neighborhoods will be examined in detail focusing on which borough and which neighborhood will fit young family's preferences best. The study comes to the end, that Staten Island will be a good location for young families settling over to New York and provides clusters of interesting neighborhoods based on most common venues.

Table of Contents

1 Inhalt

1 Introduction

New York City (NYC) is an extremely popular city and, of course, one of the most attractive cities in the world. With an estimated population of 8,336,817 (2019) distributed over about 302.6 square miles, NYC is also one of the most densely populated cities in the world. Many districts, landmarks and organizations of NYC are well known and attracting a huge number of immigrants yearly. So many people are coming to NYC for living and working, but before they migrate, they will have to check the opportunities of its boroughs namely Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

1.1 Scope of Interest

The aim of this project is to help people explore different living opportunities and to help make better decisions before migrating to NYC. The projects aim to create an analysis of features for a comparative analysis. The features include statistics like population, land area, population per square mile, household income, per capita income, housing units, mean travel time to work, housing owner costs (with a mortgage as well as without a mortgage), and median gross rent. These features will be potential factors to help people to get awareness before starting a new life.

This study focuses on the moving decision of young families, because their decision usually needs to be better prepared and is usually irreversible. By doing so, parents of a young families are the main audience of this study. Young Families will be defined as young couples with at least one child in school age.

1.2 Analytical Approach

The project aims to do a comparative analysis between NYC's boroughs in a first step. In a second step, the individual neighborhoods of the selected borough are analysed among themselves. A k-Means clustering method is applied in both steps of the investigation. For a better exploration, the cluster analysis is supplemented by some descriptive statistics.

2 Data

According to the prior Labs in the IBM data-science course, this study will based on location data provided by Foursquare. To make the location data alive, some socio-economic statistics will be taken in account also. Once again NYC acts as the central location of interest.

2.1 Data requirements

For this study geo-locational information about the boroughs and the particular neighborhoods of NYC is needed. Also the socio-economic statistics are needed on borough level. For quality reasons the needed data should be actual and accurate. So, the last US Census data will be gathered for this project.

2.2 Data collection

The project uses Foursquare API as primarily data source. By doing so, features of nearby places of the neighborhoods will be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100. As mentioned, data from US Census Bureau (https://www.census.gov/) is used to supplement the Foursquare data. To explore socio-economic statistics, this study focus on "QuickFacts" provided by US Census Bureau.

QuickFacts data are derived from: Population Estimates, American Community Survey, Census of Population and Housing, Current Population Survey, Small Area Health Insurance Estimates, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits.

2.3 Data preparation

Data will be prepared directly in the Jupyter notebook with the help of some essential python libraries. These libraries are numpy, pandas, geopy, matploblib, sklearn and folium. For handling JSON files, XLSX files and requests adequate modules will be imported also.

For data preparation some functions are defined to handle data more convenient. For example there is a function that extracts the category of the venue, a function that extracts nearby venues, and one function for return most common venues. These functions are adapted from prior labs in this course.

3 Methodology

Based on the New York dataset from on a prior lab within this course program, I transform the geospatial data in a proper data frame. The data frame consists of borough, neighborhood, latitude, and longitude. For better exploration, I slice this data frame on borough level to gain deeper understanding of the particular boroughs. In a first step, some maps were created to gain geographical insights.

After that, the borough-level data were aggregated by the socio-economic statistics from US Census Bureau in order to make an exploratory data analysis. For this, I create some bar charts about potentially relevant factors concerning to the aim of the study. In order to choose a borough for deeper analysis, a k-Means clustering approach was applied for comparative analysis.

With the chosen borough in mind, the Foursquare API was utilized to get venue data on neighborhood level for all neighborhoods within the chosen borough. The JSON-file was normalized and filtered, to gather venue name, venue categories, and venue location data (zip code, latitude, and longitude).

After that, the data frame was restructured for a comparative analysis of the most common venues within the particular neighborhoods. Based on the restructured data frame a further k-means clustering approach was applied in order to analyze the (dis-)similarities of the neighborhoods of the chosen NYC borough.

The k-means clustering algorithm was approached because this form of unsupervised machine learning is most suitable to cluster the given data. It helps to cluster data points based on feature similarity which supports the aim of the study totally.

4 Results

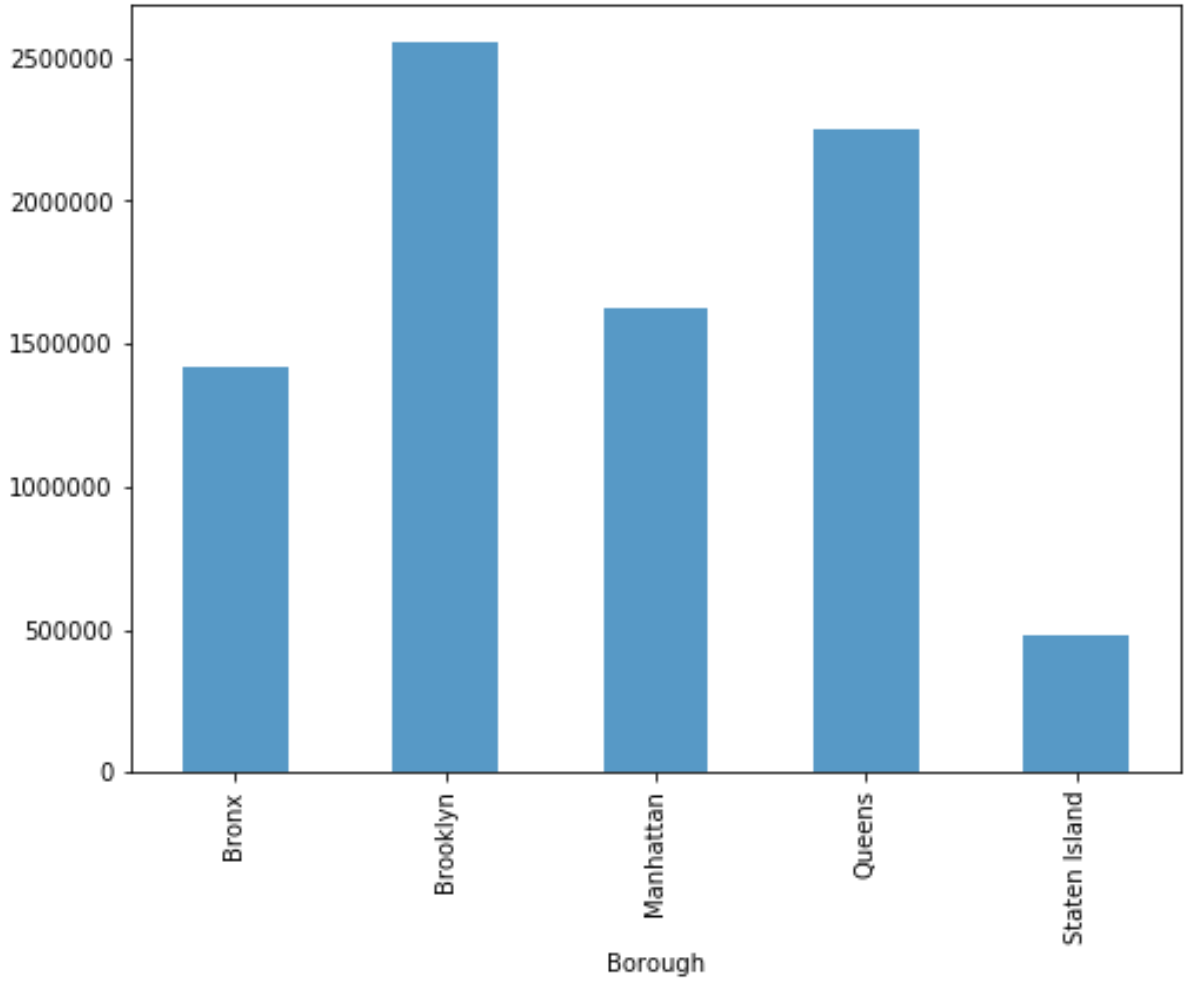

Figure 1: Population estimates 2019

Figure 1 shows the distribution of the population within the particular boroughs. Brooklyn is the biggest borough of NYC with 2,559,903 people while Staten Island is the smallest with 476,143 people.

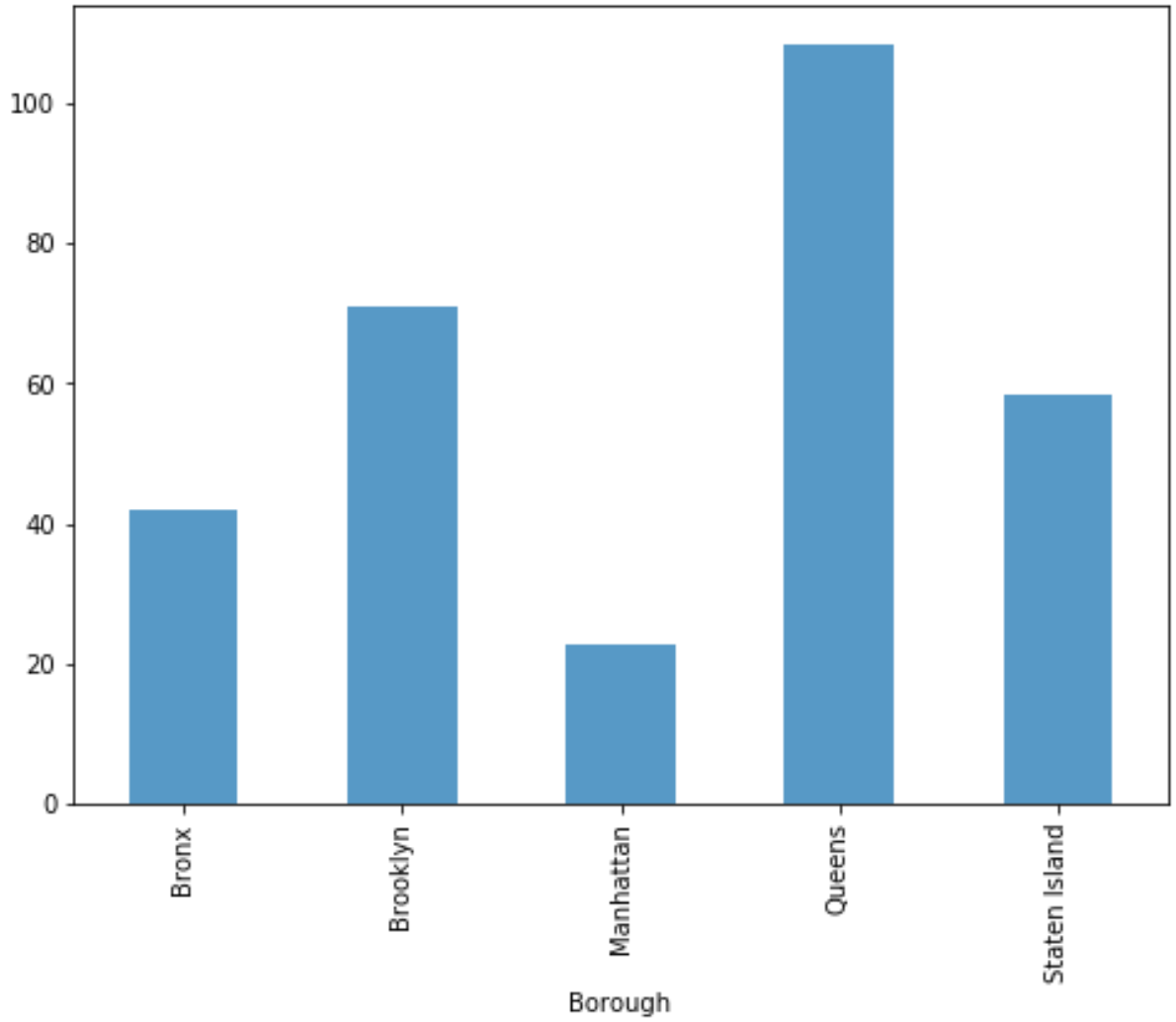

Figure 2: Land area in square miles

Figure 2 shows the distribution of the land area within the particular boroughs. Queens is the biggest borough of NYC with 2,559,903 people while Staten Island is the smallest with 476,143 people.

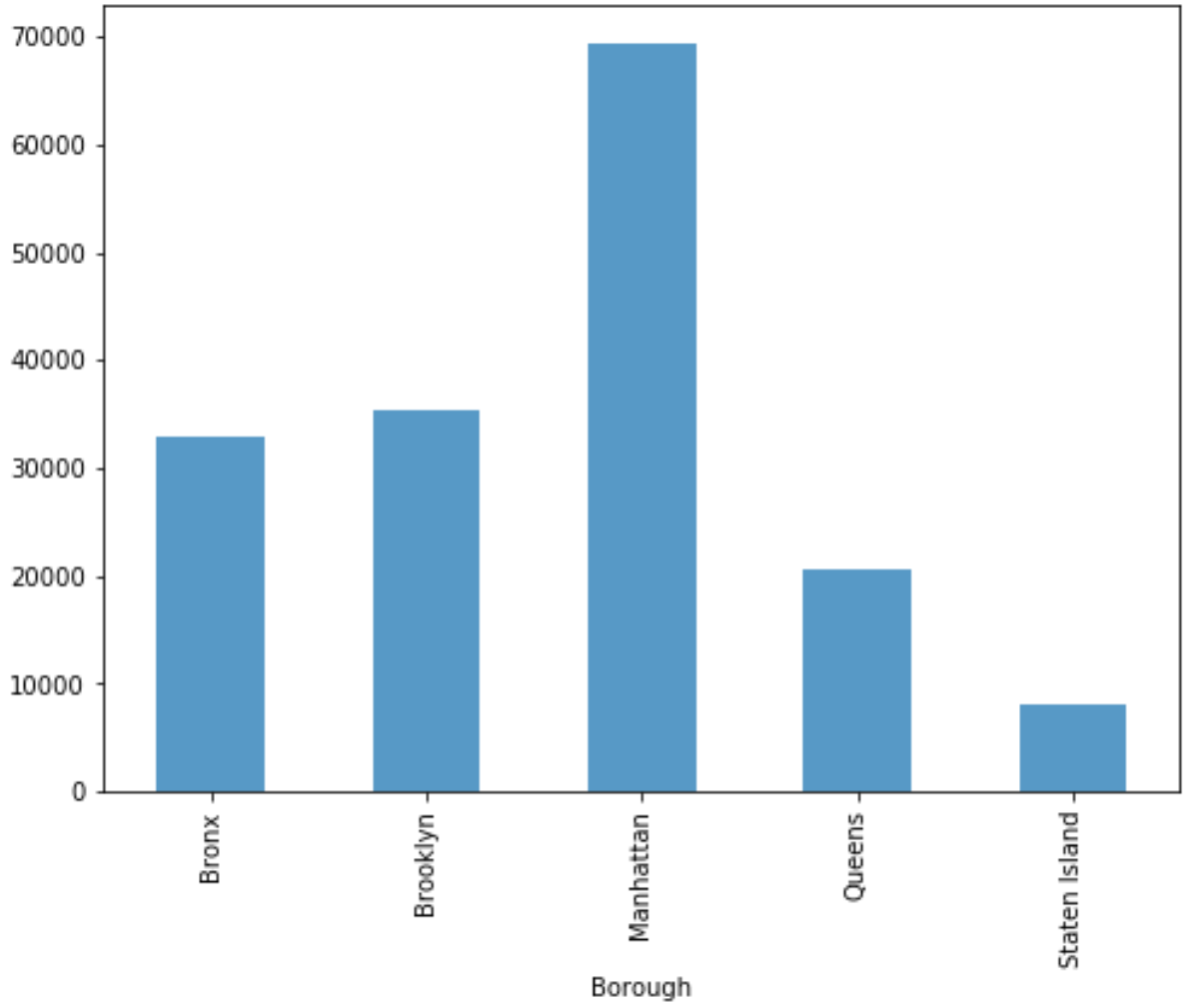

Figure 3: Population per square mile

Figure 3 shows the distribution of the population per square mile within the particular boroughs. Manhattan has the biggest population density of NYC's boroughs with 69,467.5 people per square mile while Staten Island has the smallest with 8,030.3 people per square mile.

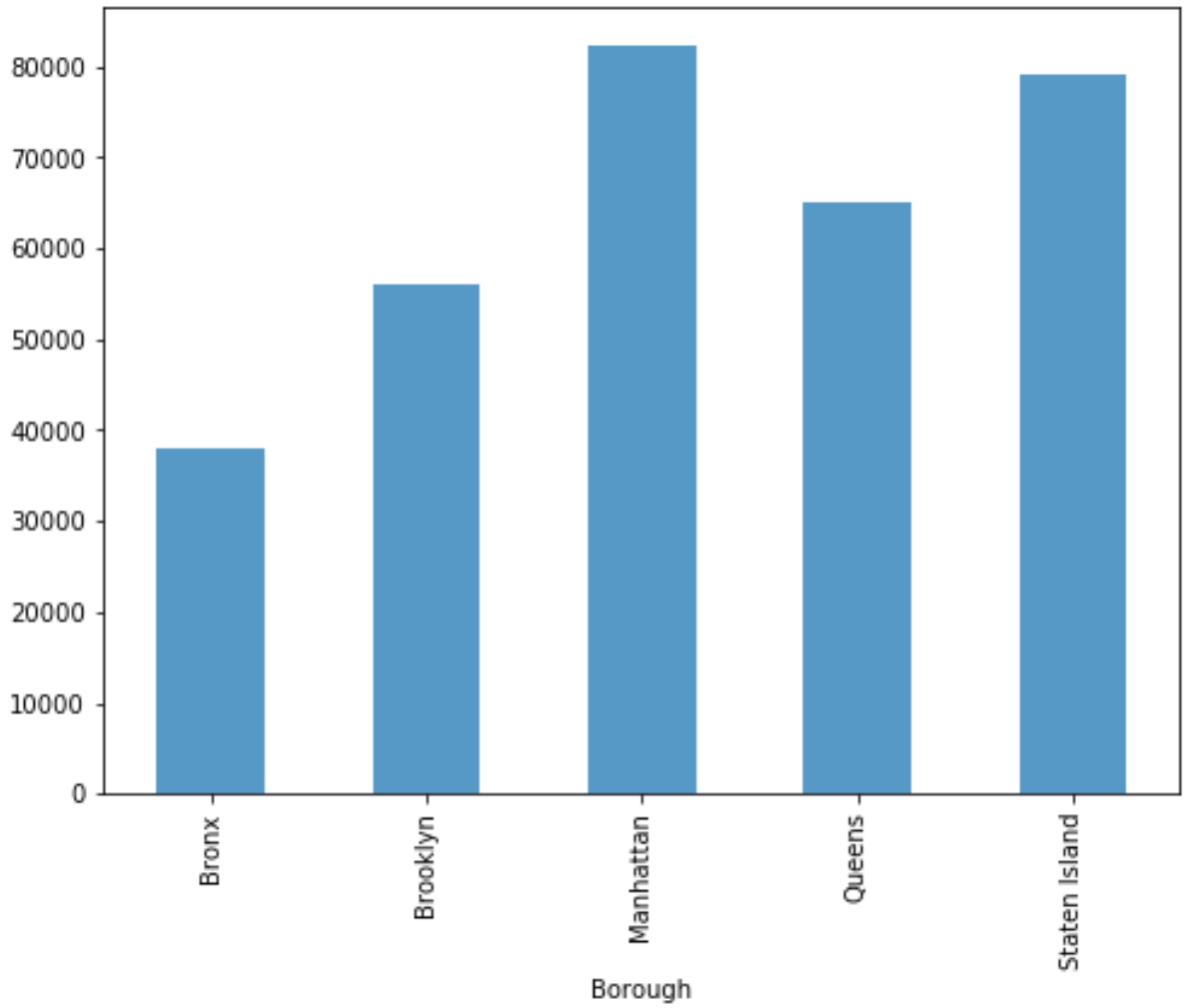

Figure 4: Median household income

Figure 4 shows the distribution of the median household income within the particular boroughs. Manhattan has the highest household income with 82,459$ people while Bronx has the smallest with 38,085$.

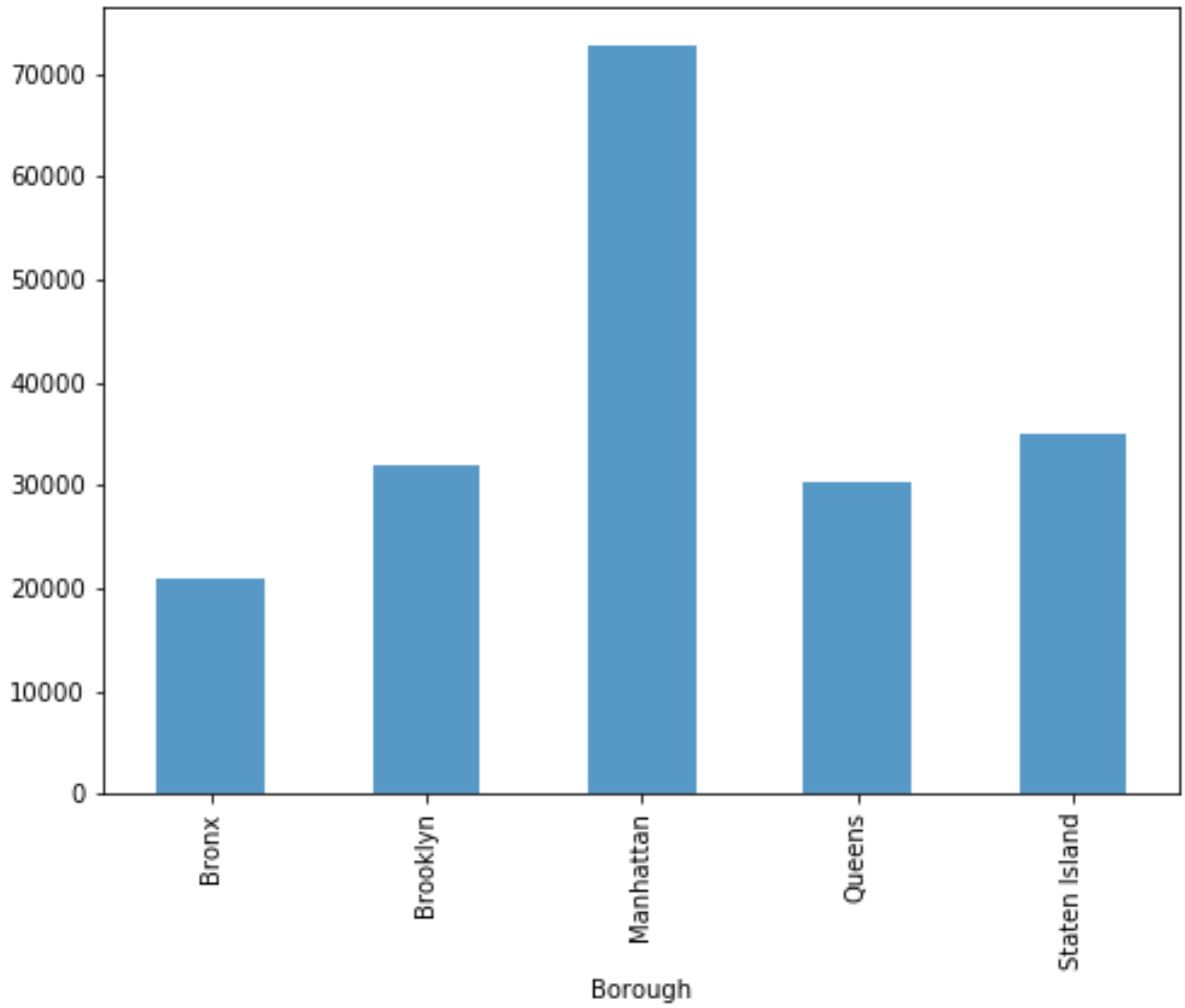

Figure 5: Per capita income in past 12 months

Figure 5 shows the distribution of the per capita income within the particular boroughs. Manhattan has the per capita income with 72,832$ while Bronx has the smallest with 20,850$.

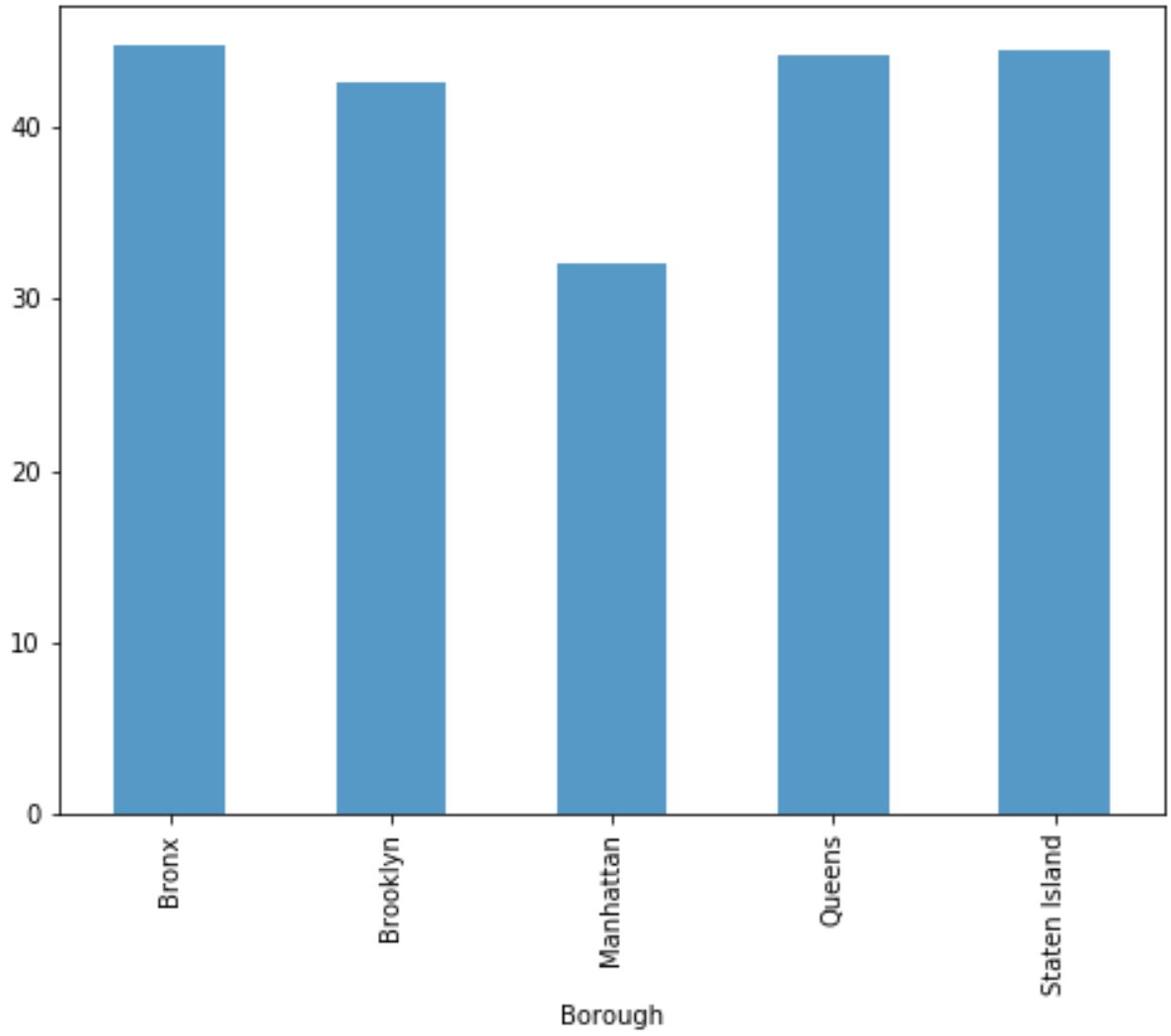

Figure 6: Mean travel time to work (minutes) workers age 16 years+

Figure 6 shows the distribution of the mean travel time to work within the particular boroughs. Manhattan has the shortest mean travel time with 32.1 minutes while Bronx has the longest with 44.8 minutes.

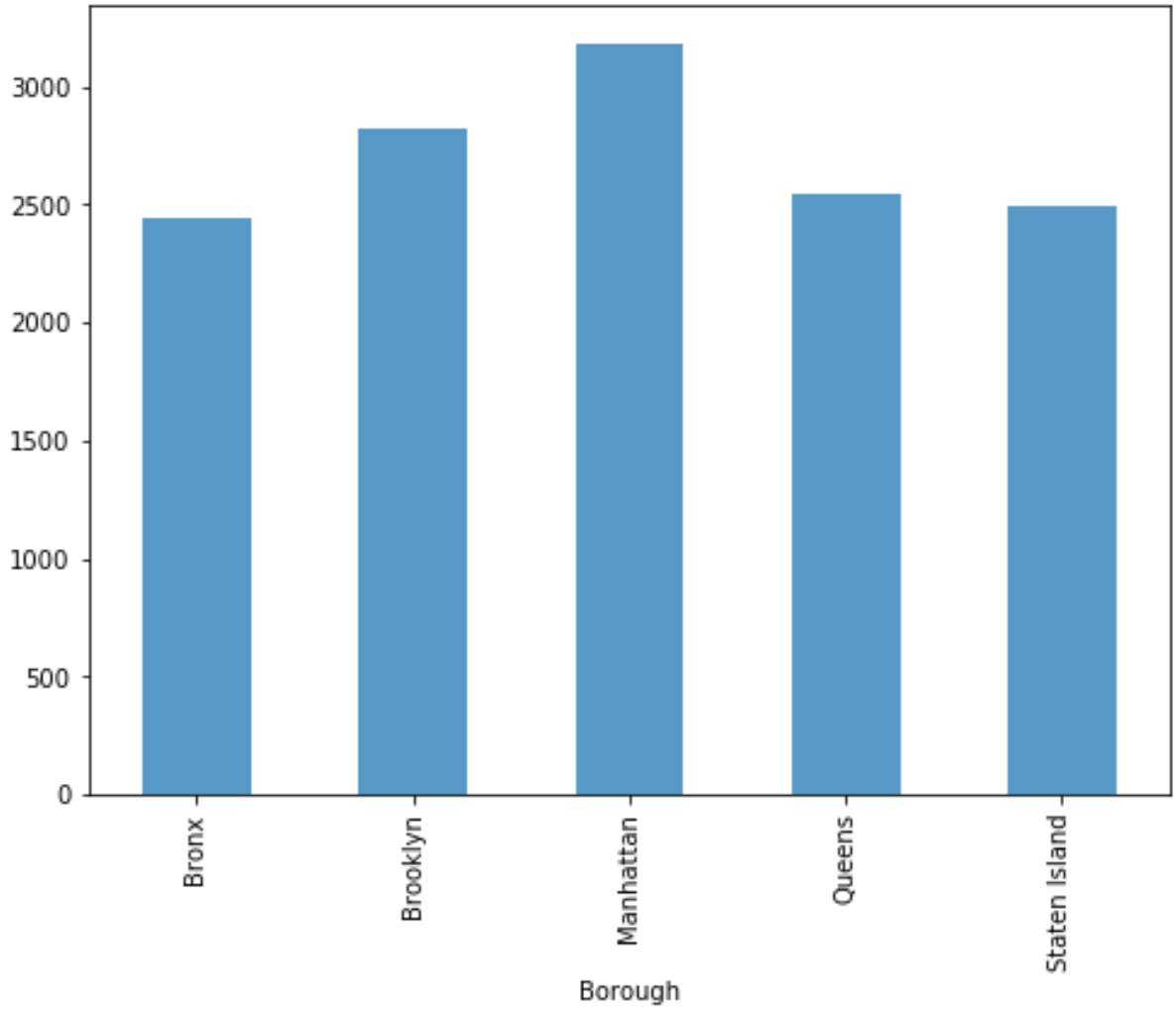

Figure 7: Median selected monthly owner costs -with a mortgage

Figure 7 shows the distribution of the median selected monthly owner costs (with a mortgage) within the particular boroughs. Manhattan has the highest monthly owner costs of NYC with 3,186$ while Bronx has the smallest with 2,440$.

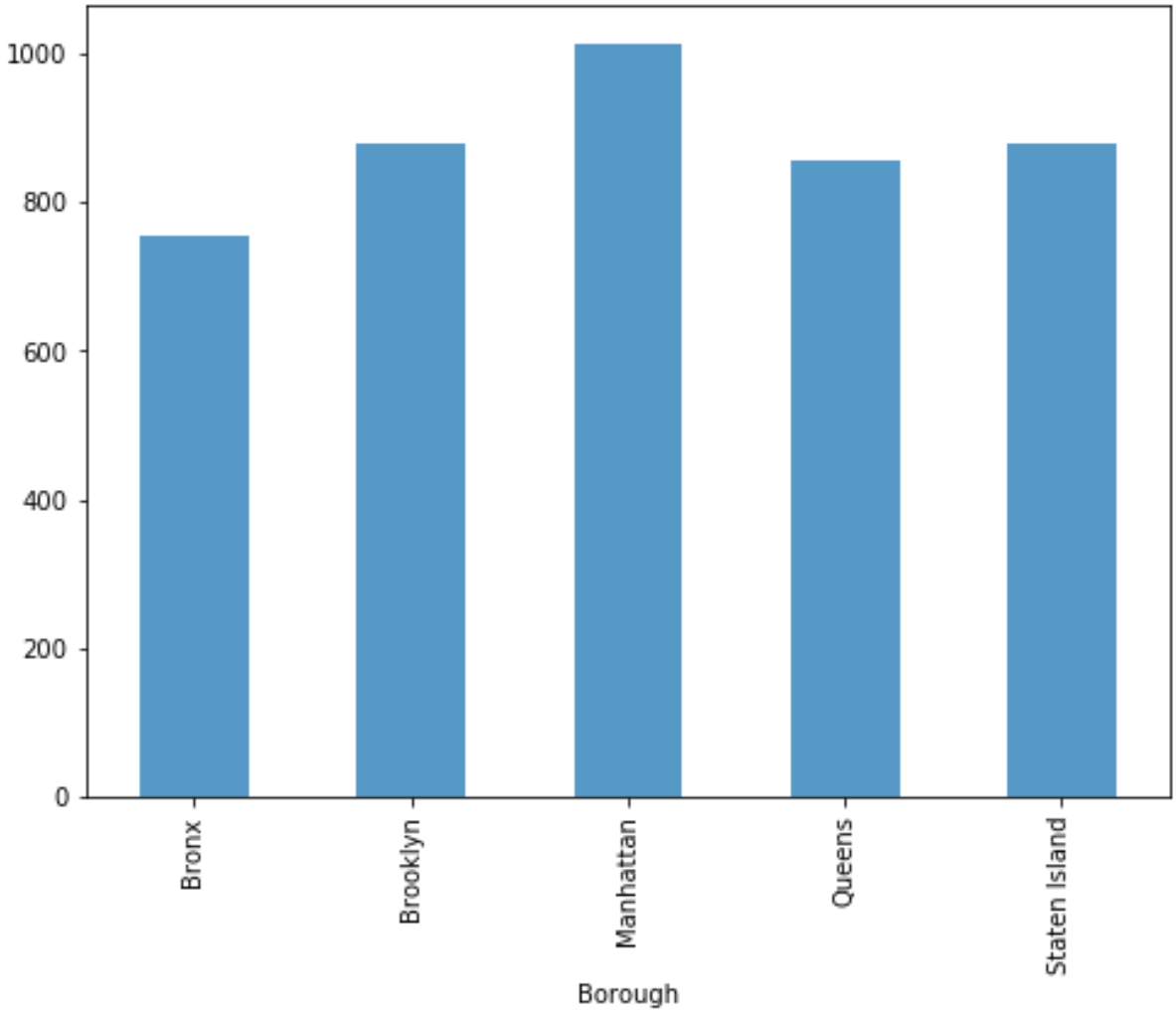

Figure 8: Median selected monthly owner costs -without a mortgage

Figure 8 shows the distribution of the median selected monthly owner costs (without a mortgage) within the particular boroughs. Manhattan has the highest with 1,014$ while Bronx has the lowest with 755$.

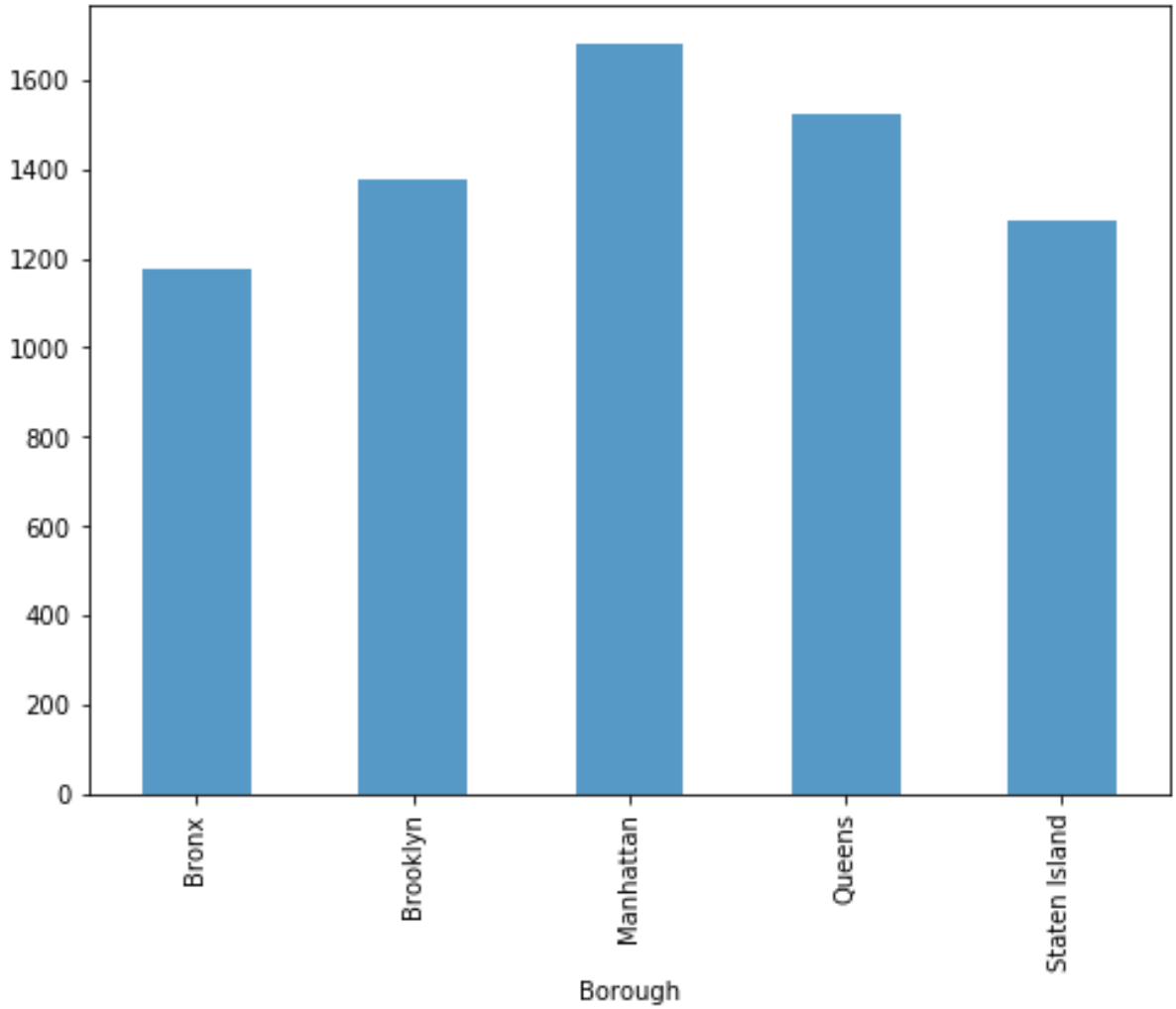

Figure 9: Median gross rent

Figure 9 shows the distribution of the median gross rent within the particular boroughs. Manhattan has the highest with 1,682$ while Bronx has the smallest with 1,176$.

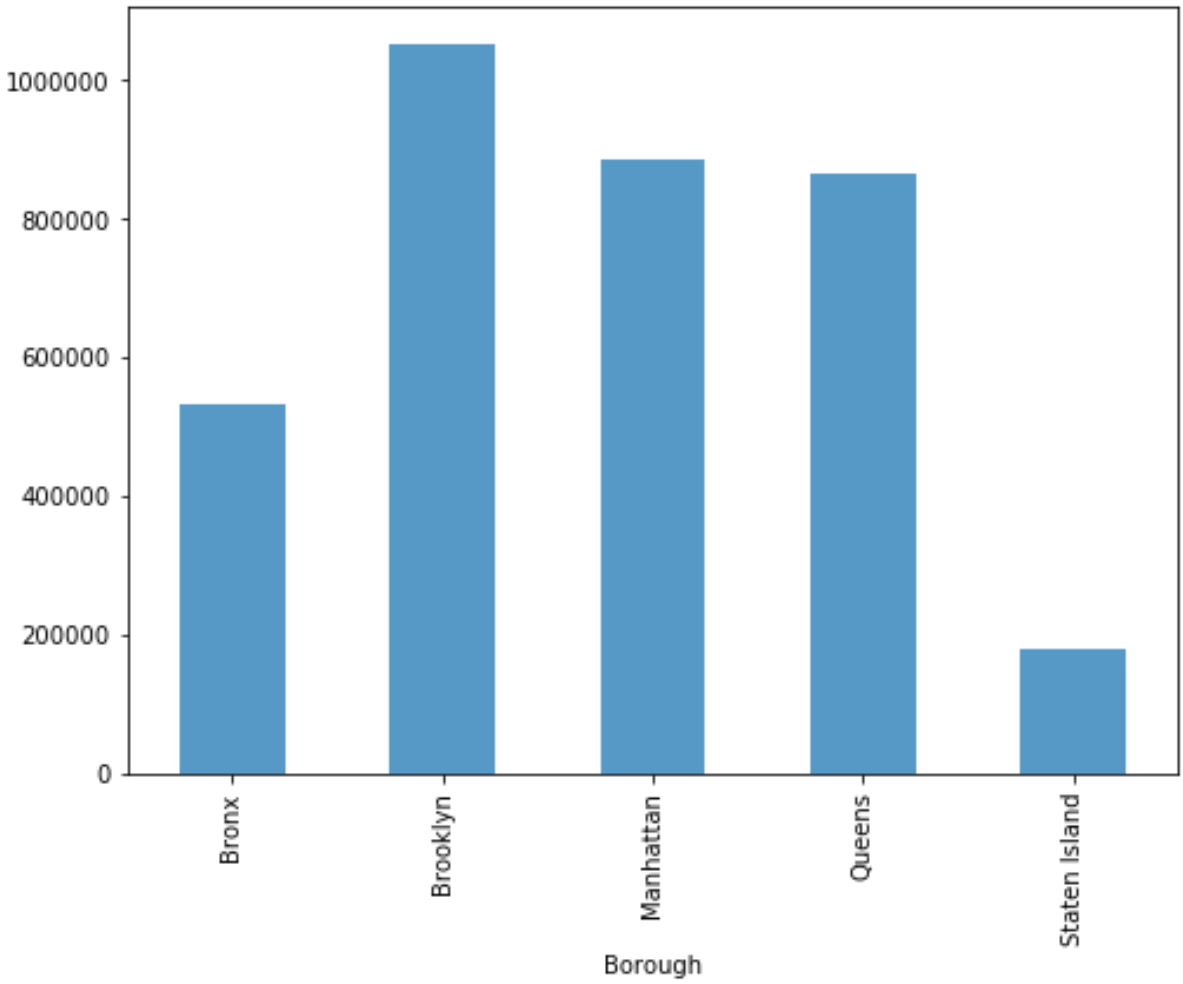

Figure 10: Housing units

Figure 10 shows the distribution of the housing units within the particular boroughs. Brooklyn has the highest amount with 1,053,767 housing units while Staten Island has the lowest amount with 181,199 housing units.

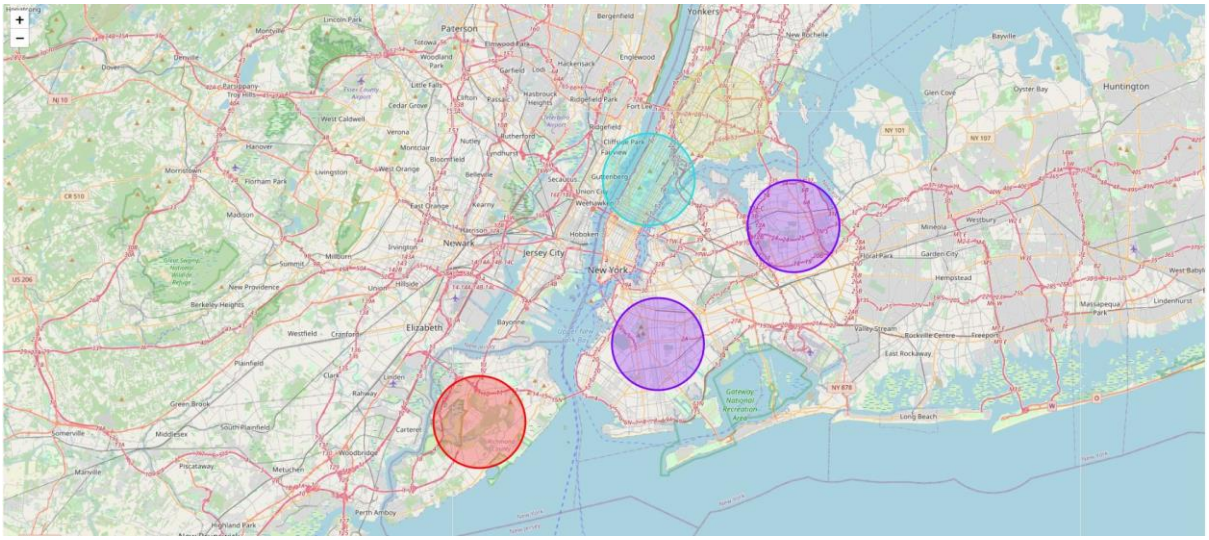

Figure 11: Borough Clustering

Figure 11 shows the results of k-Means Cluster analysis of the particular boroughs based on the socio-economic statistics. 4 clusters were built, whereas Queens and Brooklyn are united to one cluster because of similarity. Bronx, Manhattan and Staten Island stay as unique as stand-alone clusters because of dissimilarity among each others. By using the derived clusters, young families can choose a borough based on the cluster features.

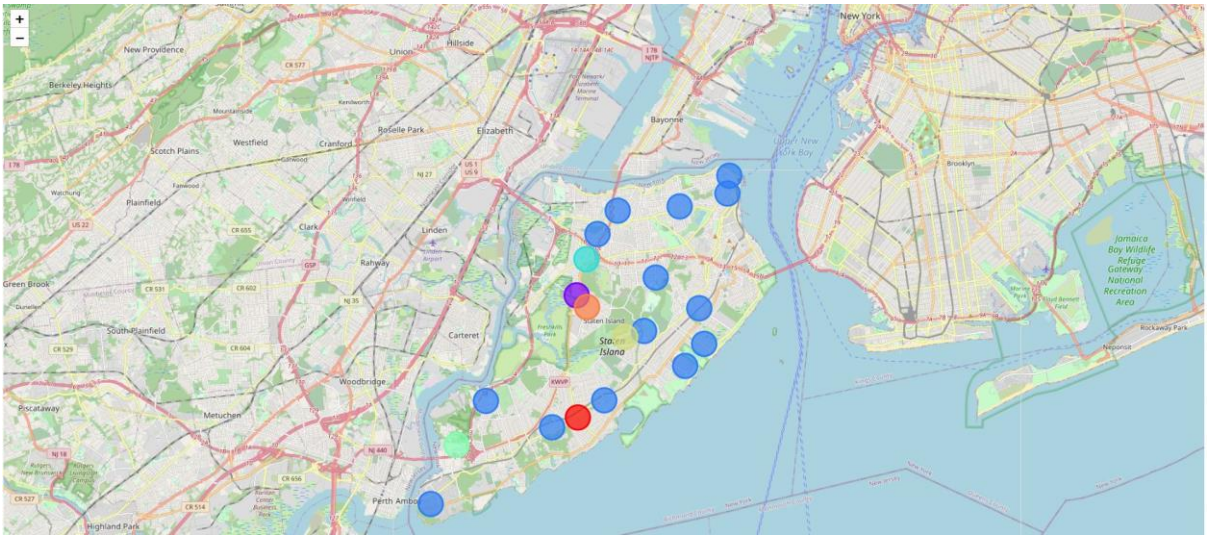

Figure 12: Neighborhood Clustering

Figure 12 shows the results of k-Means Cluster analysis of Staten Island's neighborhoods based on the Foursquare venue data. 7 clusters were derived, whereas Bulls Head, Heartland Village, New Springville, Lighthouse Hill and Eltingville stay unique as stand-alone clusters because of dissimilarity among each others. By using the derived clusters, young families can choose a neighborhood based on the cluster features.

5 Discussion

The exploratory analysis of NYC's socio-economic statistics (shown in Figure 1-10) shows different socio-economic situations within the particular boroughs. This study runs a k-Means Clustering approach for a comparative analysis. Its results are shown in Figure 11.

By focusing on the borough cluster features we can make a decision on which borough we will choose for our main audience defined in chapter 1 of this study. This study assume that young families (= main audience) will focus on relatively low housing costs (rent as well as ownership), enough area for play and relax, relatively short time to travel to work, as well as relatively good income level and relatively good local supply by retail and gastronomy.

Borough	Population per square mile	Median household income	Per capita income	Mean travel time to work	owner costs with a mortgage	owner costs without a mortgage	Median gross rent
Bronx	0,47	0,46	0,29	1,00	0,77	0,74	0,70
Brooklyn	0,51	0,68	0,44	0,95	0,89	0,87	0,82
Manhattan	1,00	1,00	1,00	0,72	1,00	1,00	1,00
Queens	0,30	0,79	0,42	0,98	0,80	0,84	0,90
Staten Island	0,12	0,96	0,48	0,99	0,78	0,87	0,76

Figure 13: Normalized socio-economic statistics for clustering

Figure 13 shows the normalized socio-economic statistics for the clustering approach. According to the cluster analysis results Brooklyn and Queens are quite similar to another. Each others are more specific by itself. The Brooklyn-Queens-Cluster suits to the young family decision assumptions defined earlier, but Staten Island suits much better. By this, Staten Island is chosen for deeper analysis.

By focusing on the neighborhood cluster features we can make a decision on which neighborhood our main audience my will choose for starting a new life in NYC. This decision highly depends on the family's preferences. This study will not anticipate the individual preferences, instead provide a decision framework based on the clusters.

Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
St. George	2	Ice Cream Shop	Clothing Store	Coffee Shop	Women's Store	Discount Store
West Brighton	2	Donut Shop	Women's Store	Department Store	Gym / Fitness Center	Grocery Store
New Springville	1	Clothing Store	Toy / Game Store	Furniture / Home Store	Cosmetics Shop	Department Store
Great Kills	2	Bakery	Women's Store	Department Store	Gym / Fitness Center	Grocery Store
Eltingville	0	Pharmacy	Video Game Store	Cosmetics Shop	Grocery Store	Golf Course
Annadale	2	Bagel Shop	Women's Store	Department Store	Gym / Fitness Center	Grocery Store
Tottenville	2	Bank	Women's Store	Department Store	Gym / Fitness Center	Grocery Store
Tompkinsville	2	Gym / Fitness Center	Fast Food Restaurant	Women's Store	Department Store	Grocery Store
Graniteville	2	Coffee Shop	Women's Store	Department Store	Gym / Fitness Center	Grocery Store
Dongan Hills	2	Bank	Women's Store	Department Store	Gym / Fitness Center	Grocery Store
Midland Beach	2	Paper / Office Supplies Store	Discount Store	Women's Store	History Museum	Gym / Fitness Center
New Dorp Beach	2	Supplement Shop	Furniture / Home Store	Burger Joint	Sandwich Place	Women's Store
Charleston	4	Cosmetics Shop	Bakery	Health & Beauty Service	Supplement Shop	Pet Store
Rossville	2	Bagel Shop	Mexican Restaurant	Women's Store	Discount Store	Gym / Fitness Center

Heartland Village	6	Hookah Bar	Shoe Store	Furniture / Home Store	Food & Drink Shop	Liquor Store
Bulls Head	3	Baseball Field	Chinese Restaurant	Playground	Women's Store	Discount Store
Elm Park	2	Department Store	Mobile Phone Shop	Sandwich Place	Women's Store	Grocery Store
Manor Heights	2	Trail	Campground	Women's Store	Department Store	Gym / Fitness Center
Egbertville	2	Park	Women's Store	History Museum	Gym / Fitness Center	Grocery Store
Lighthouse Hill	5	Italian Restaurant	Art Museum	Golf Course	Spa	Women's Store

Figure 14: Most common venues of clustered neighborhoods

Figure 14 shows the most common venues of the clustered neighborhoods. The largest cluster is cluster 2 with 14 neighborhoods. Their venues are quite similar to another but also dissimilar among each others.

5.1 Recommendations

Young families can decide on which neighborhood to choose based on the derived clusters. For example, if the family is highly interested in arts it may will tend to choose cluster 5 (= Lighthouse Hill) more than the others. This study will not make further recommendations for decision-makers because there are no information about possible preferences.

For future studies I recommend to take a bigger data set into account and to survey young family's preferences. Without these information, such a study like this here, will be not really representative.

6 Conclusion

This study provides a decision framework for young families which will start a new life at NYC. By doing so, the study clustered the particular boroughs of NYC according to socio-economic statistics first, and clustered the particular neighborhoods of the chosen borough according to their most common venues second.

The first step is highly based on the assumptions of the main audience's preferences. By this, it is highly sensitive and susceptible to criticism. Even smallest changes in the preferences can significantly alter the results and turn the framework into nonsense. So, more information on possible preferences will be desirable, but go beyond the frame of this course.

The second step is not based on assumptions of unknown preferences but depends highly on the Foursquare venue data. Because of API limitations the dataset is only a small snapshot of Staten Island neighborhood's features. By this, the derived clusters will be not representative and much more data will be desirable, but also go beyond the frame of this course.