IBM DATA-SCIENCE CERTIFICATION

Coursera Capstone

The Battle of Neighborhoods

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.

Introduction

- aim of this project: help people to explore different living opportunities and to help make better decisions before migrating to NYC
- create an analysis of features for a comparative analysis
  - 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
  - features will be potential factors to help people to get awareness before starting a new life
- focuses on the moving decision of young families, because their decision usually needs to be better prepared and is usually irreversible
- young families as main audience
  - Young Families will be defined as young couples with at least one child in school age

Data

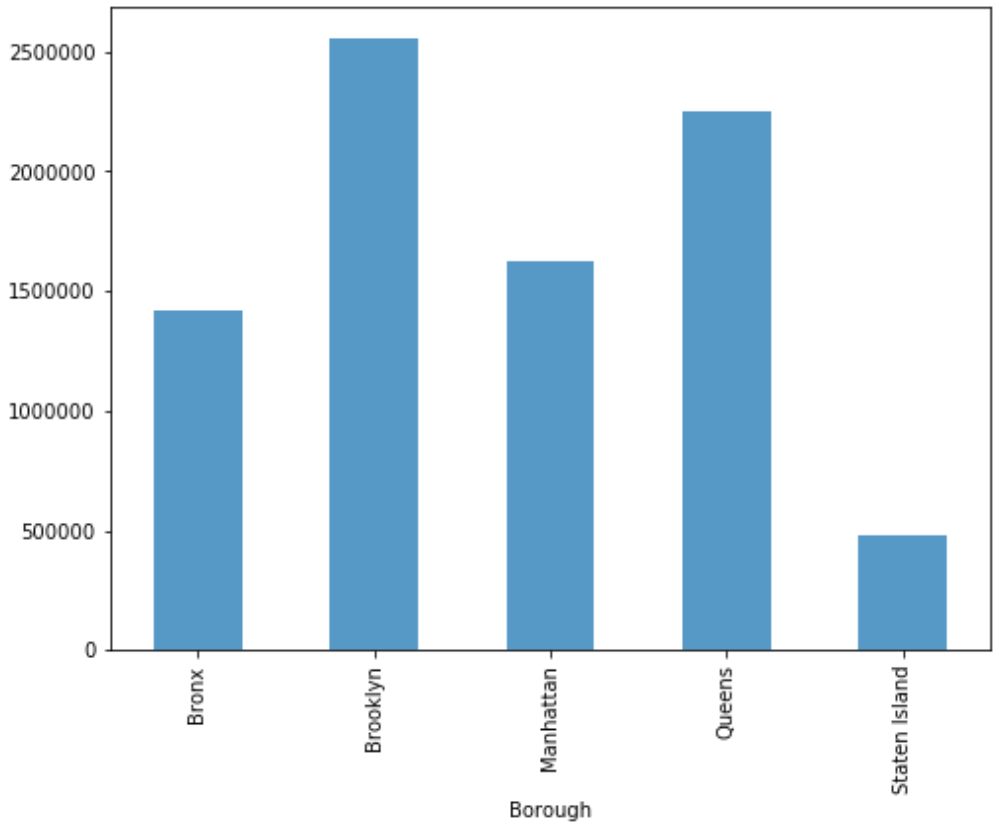
- study will based on location data provided by Foursquare, some socio-economic statistics will be taken in account also
- Foursquare API as primarily data source; due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100
- QuickFacts Data from US Census Bureau (https://www.census.gov/) is used to supplement the Foursquare data
- Data will be prepared directly in Jupyter notebook with the help of some essential python libraries (numpy, pandas, geopy, matploblib, sklearn and folium)
- For data preparation some functions are defined to handle data more convenient

Methodology

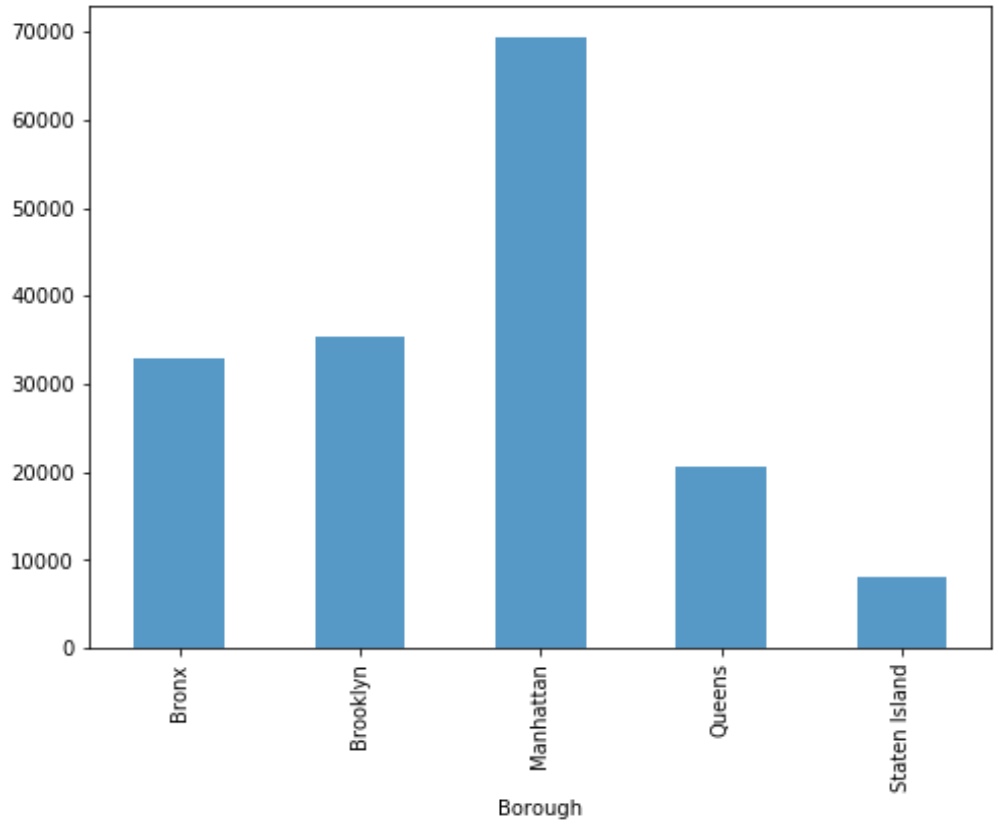
- Transformation of geospatial data into a proper data frame consists of borough, neighborhood, latitude, and longitude
- Slicing the data frame for exploratory analysis and data Aggregation with socio-economic statistics from US Census Bureau
- Exploratory data analysis with maps and bar charts, then comparative analysis by k-Means clustering approach
- Foursquare API was utilized to get venue data on neighborhood level; resulting JSON-file was normalized and filtered, to gather venue name, venue categories, and venue location data (zip code, latitude, and longitude).
- Restructuring data frame for a comparative analysis of the most common venues by a further k-means clustering approach

Results (1/7)

Population estimates

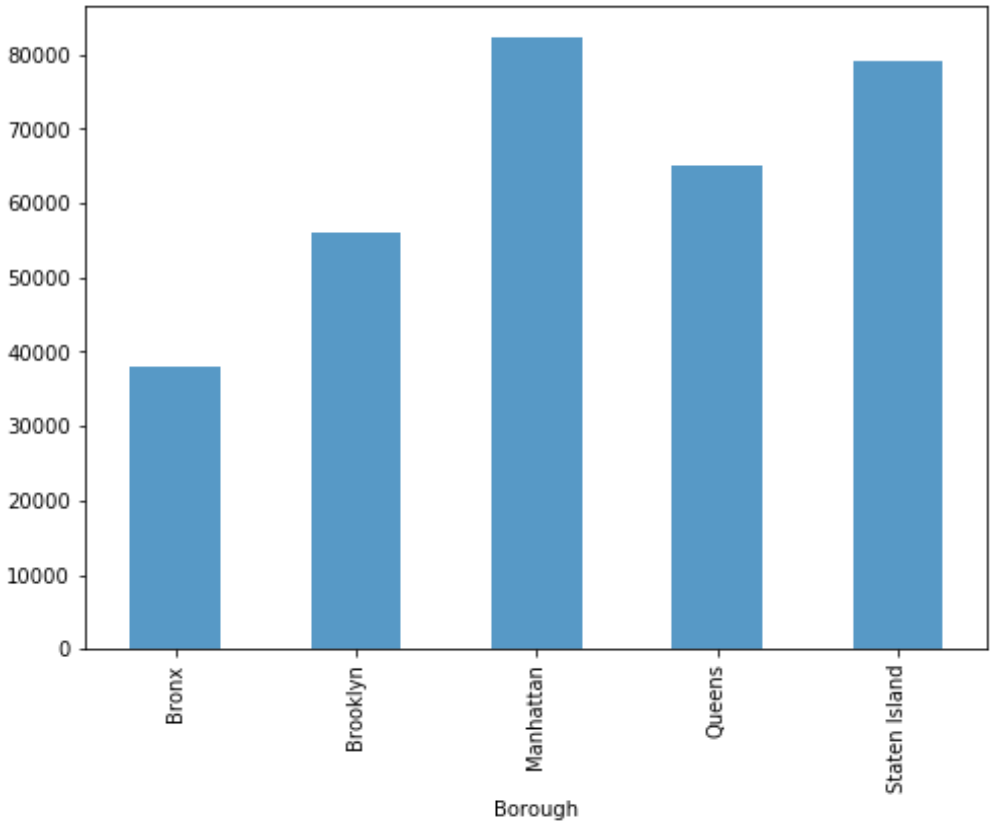


Population density

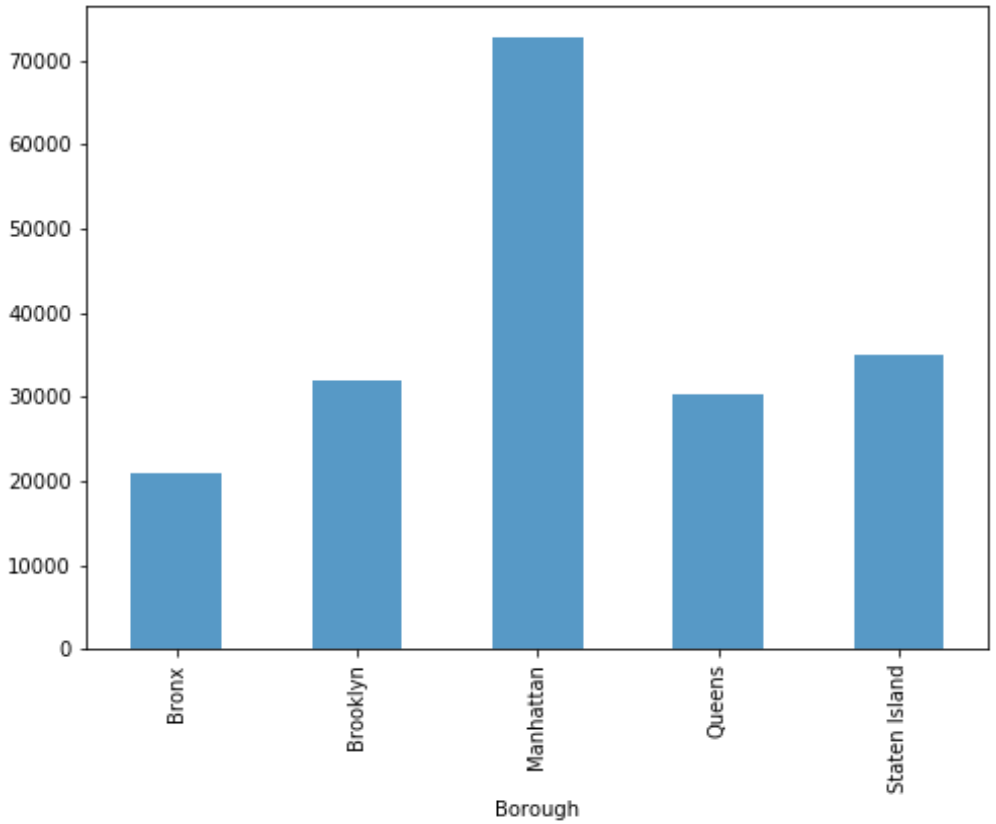


Results (2/7)

Median Household Income

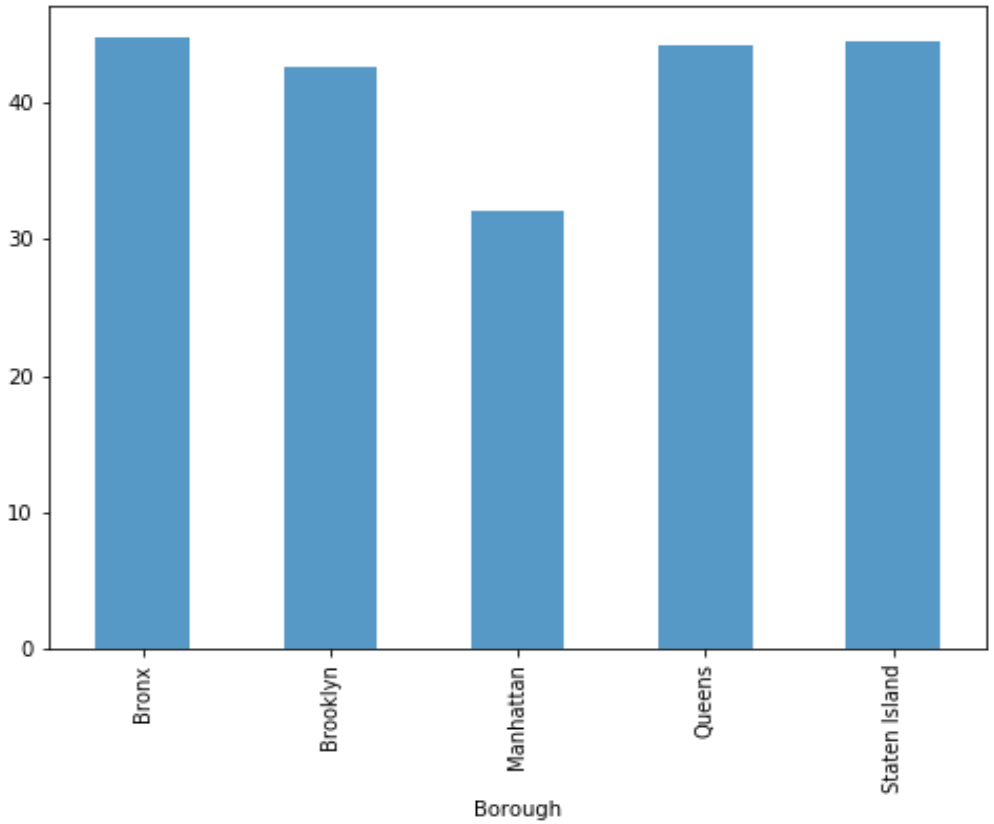


Per Capita Income

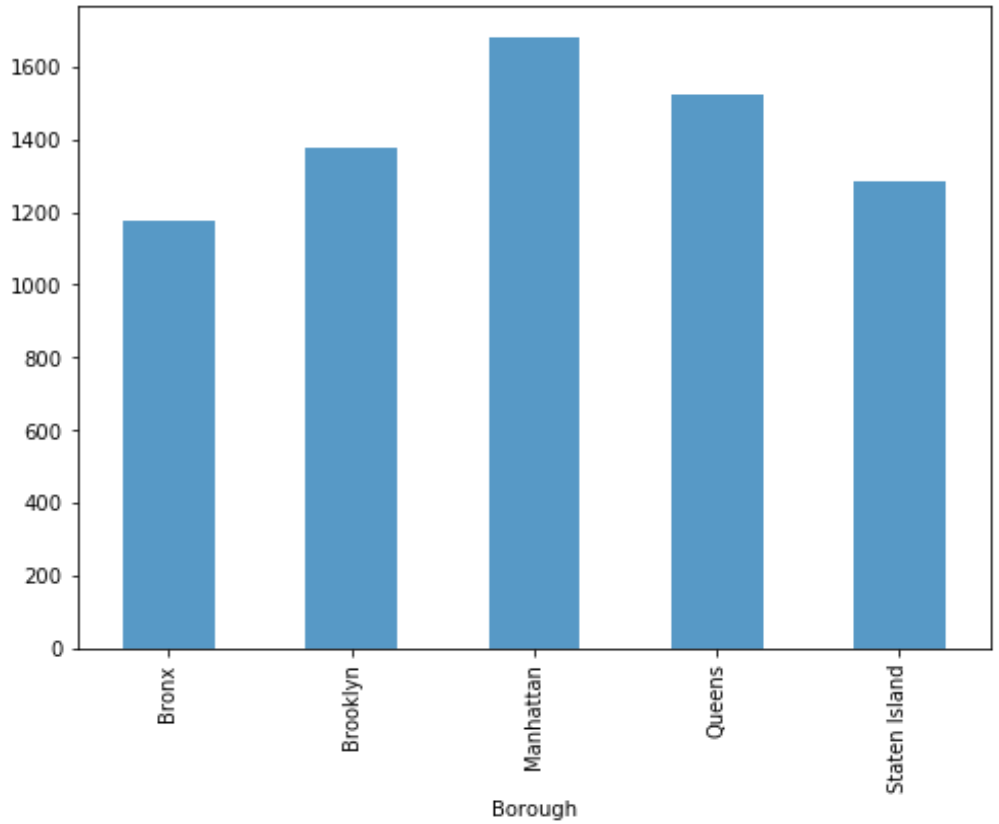


Results (3/7)

Mean Travel Time to Work

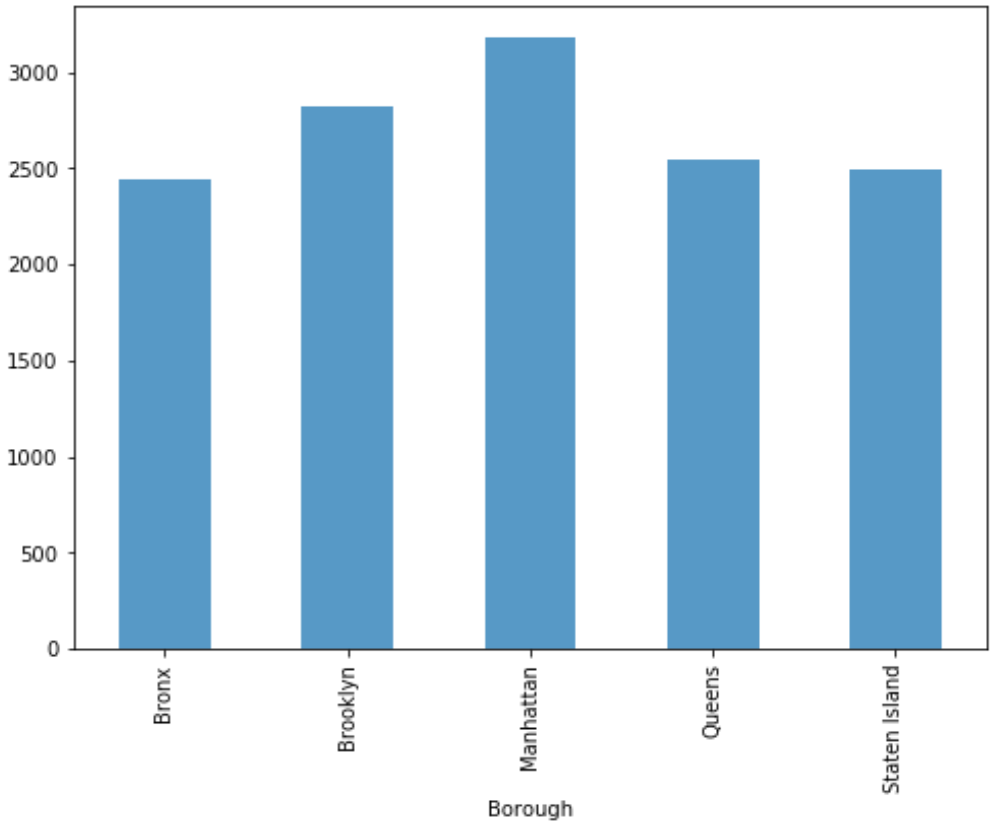


Median Gross Rent

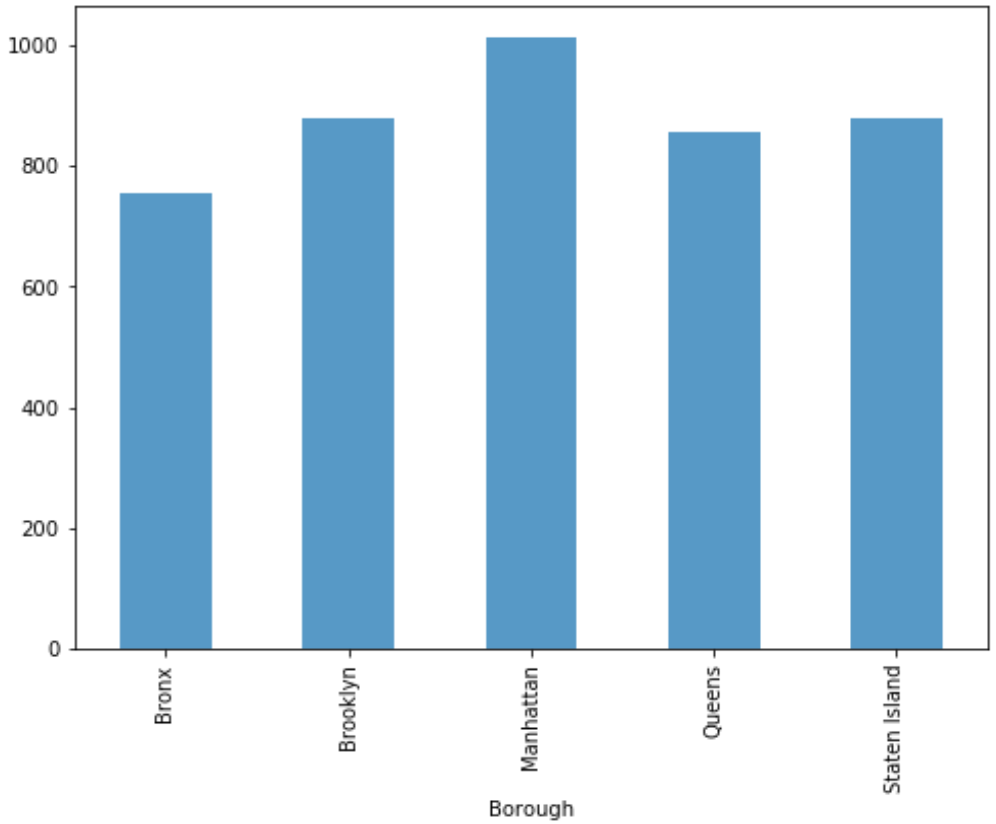


Results (4/7)

Ownership Costs
(monthly, with mortgage)



Ownership Costs
(monthly, without mortgage)



Results (5/7)

Borough Clustering

- 4 clusters:

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

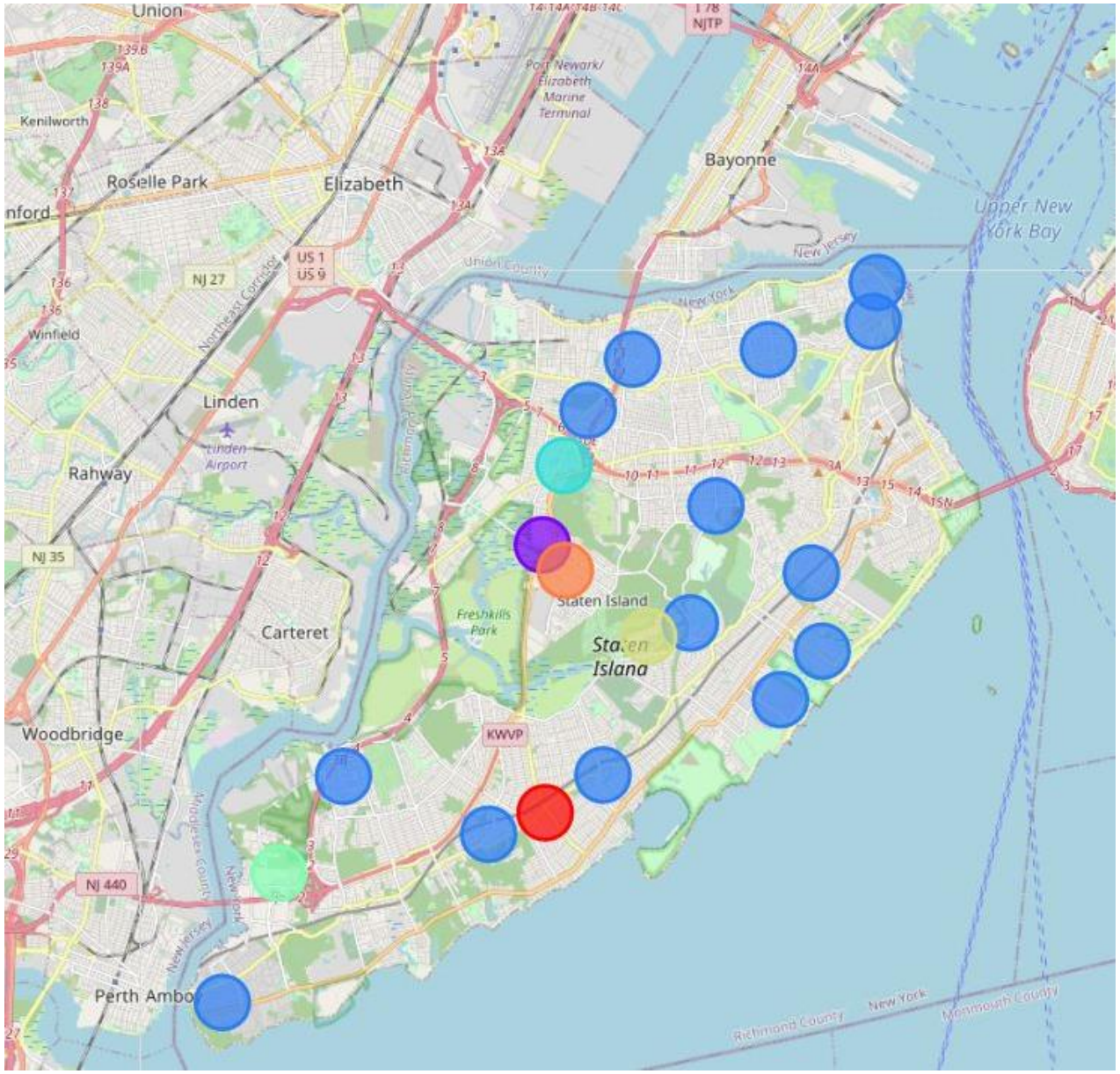
Staten Island seems to be a good choice for young families because of low population density, proper income levels, and relatively low housing costs

Results (6/7)

Neighborhood Clustering

- 7 clusters:

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.



Results (7/7)

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

Discussion

- Exploratory analysis of NYC's socio-economic statistics is showing different socio-economic situations within the particular boroughs
- Assumption: 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
  - The Brooklyn-Queens-Cluster suits to the young family decision assumptions, but Staten Island suits much better. By this, Staten Island was chosen for deeper analysis.
- Young families can decide on which neighborhood to choose based on the derived clusters.
  - This study will not make further recommendations for decision-makers because there are no information about possible preferences

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.

Thanks ☺

For reading and grading as well as for this beautiful course!