LOCATION SELECTION BY K-MEANS CLUSTERING

LEO CHARLES, AUG 2019

CONTENTS

- Problem Statement
- Techniques
- Data Acquisition & Cleansing
- Exploratory Analysis
- Findings
- Assumptions
- Conclusion

PROBLEM STATEMENT

Beyond401K Inc (Fictitious Company) is a Houston, TX based Financial Services start up company which has developed an AI/ML driven Retirement Portfolio Management Solution targeting professionals who save for retirement, especially millennials. To do beta launch the company with its resource constraints has to identify appropriate locations to set up kiosks in New York City. With the kind of population diversity and geographical spread, it becomes difficult to find the optimum location.

TECHNIQUES

- To identify the best location, Data Science techniques learnt in the IBM Data Science Professional Certificate program. The techniques encompass Data manipulation, Data analysis and Data visualization in Python.

DATA ACQUISITION & CLEANSING

- New York City related data was acquired by following data sources
  - New York Open data portals

  url : (https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh)

  - New York University's open data base

  url: https://geo.nyu.edu/catalog/nyu_2451_34572

  - US gov census portal
  - Foursquare.com
- All the raw data acquired from the above sources was cleansed and got converted as Pandas dataframe.

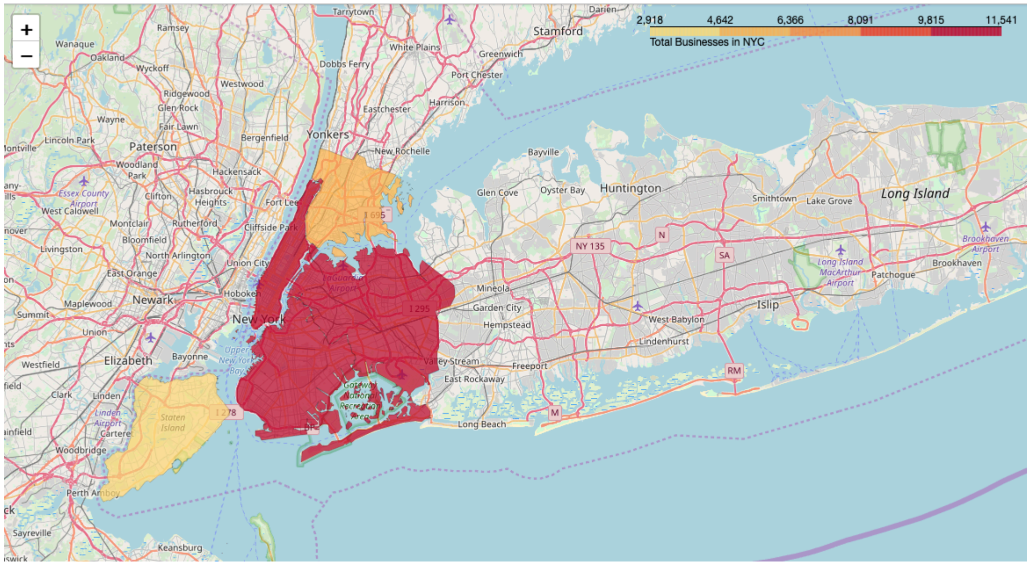
EXPLORATORY ANALYSIS

- With the dataframe, the exploratory analysis was carried out in the following stages ,
  - Analyzing the New York City's business data to identify the areas of high business concentration (Choropleth map)
  - Selecting appropriate borough based on demography data (Bar Chart)
  - Finding out the top 10 most common venues in each neighborhood.
  - Plotting K-means clustering to identify the appropriate neighborhood cluster

FINDINGS

- Choropleth Maps on the density of the business establishments
  - Manhattan, Queens and Brooklyn has higher concentration of business establishment
- Brooklyn neighborhood Clustering.
- Analyzing the Clusters

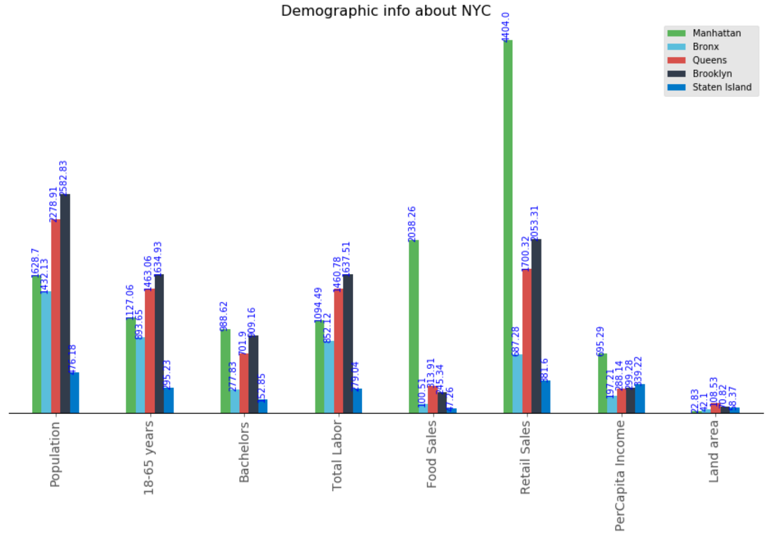
CHOROPLETH MAPS OF NEW YORK CITY



FINDINGS

- Bar Chart comparing the demography info of the boroughs.
  - Brooklyn and Manhattan are comparable boroughs.
  - Based on the criteria Brooklyn borough is the good choice.

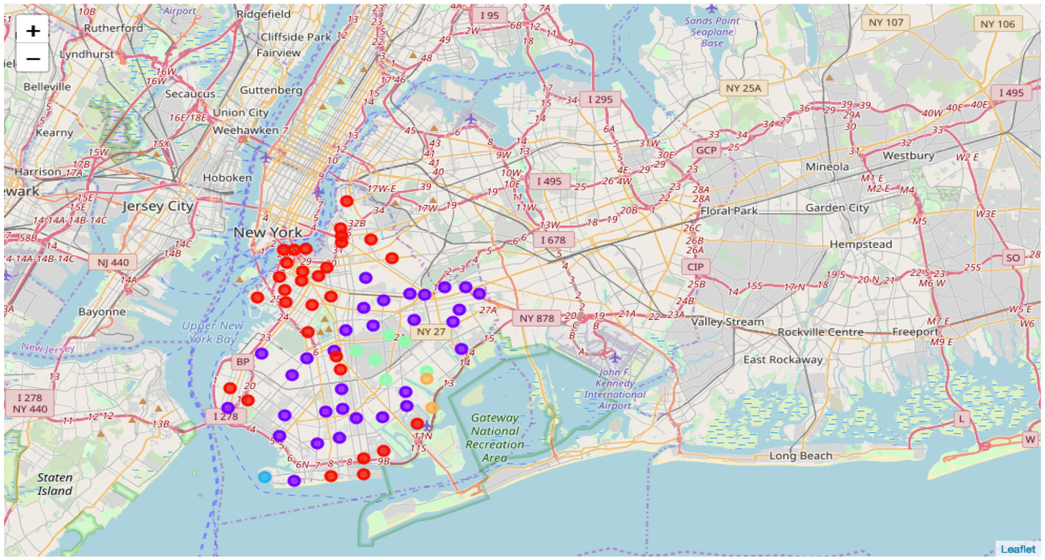
DEMOGRAPHY INFO - NEW YORK CITY



FINDINGS

- Brooklyn neighborhood Clustering.
  - Red Clusters (cluster '0') are more densely located.

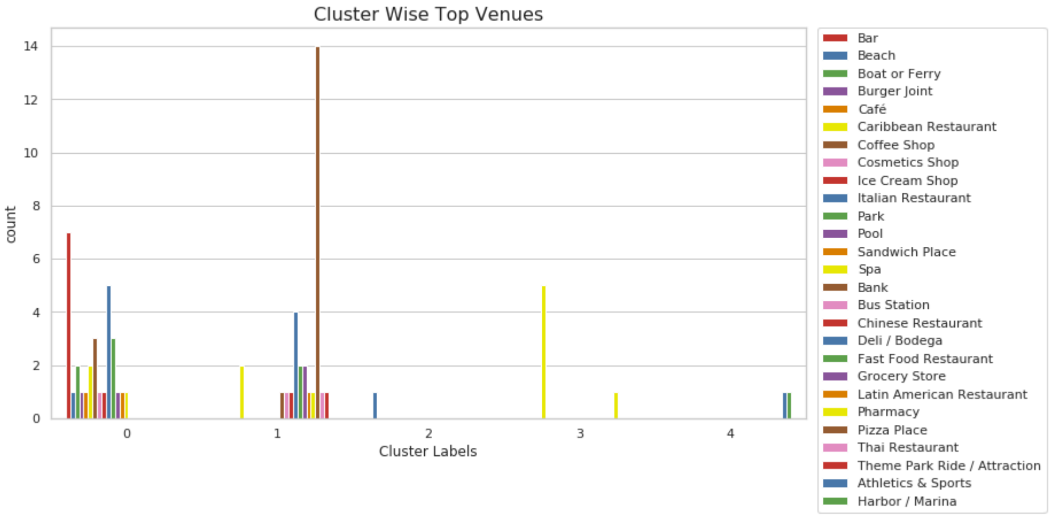
CLUSTERING – BROOKLYN NEIGHBORHOODS



FINDINGS

- Analyzing the Clusters.
  - Pizza Restaurants and Coffee shops in Cluster ‘0’ are the good venue choices

ANALYZING BROOKLYN CLUSTERS



ASSUMPTIONS

- There was assumption made that the areas having more business establishments will have more working force.
- The significant number of them in work force save for their retirements.
- In the venue selection also, it was assumed that the target customers frequent the neighborhood venues during the day and having kiosks near them will give the maximum returns in the initiative

CONCLUSION

- Based on the analysis, it's quite reasonable to suggest to the company to look for locations in (or) around cluster '0' - Italian Restaurants and Coffee Shops, in that way the campaign can generate more visibility among the target group of customers.

THE END

WORKS CITED