

Location Selection by K-Means Clustering

Leo Charles, Aug 2019

1 Introduction

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 millennials. Now they would like to do a beta launch to attract early adopters for their solution. The obvious choice for them is to launch it in New York City because it's the financial capital of United States, and also the average income of the workforce is much higher. Their plan is to set up a self serviced kiosks in prominent locations, so that they can attract professionals during their break time. Along with the kiosks they also want to have their employees to answer questions and to help the customers sign up.

Having said that, as a start up, the company is constrained on resources, so they have to be careful in spending, and at the same time get a maximum return on this initiative. Since New York City is a big city in terms of its size, population, and business activities it becomes difficult to identify the appropriate locations that are most frequented by the population. Also, the city's boroughs compounds the selection because of the cultural diversity each one brings.

The key objective of this project is to explore and analyze the neighborhoods of the city, and come up with a suggested list of venues on their respective neighborhoods that are suitable to set up kiosks.

2 Datasets

To accomplish the stated objective, the following data requisites, and their sources are identified.

2.1 Active Businesses in New York City

The target customers for the yet to be launched Retirement Portfolio solution are the working professionals. For that data we need to get the database of the active businesses in the city, and it's available in NYC Open data portal,

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

NYC OpenData Home Data About ▾ Learn ▾ Alerts Contact Us Blog | 🔍 Sign In

Legally Operating Businesses

This data set features businesses/individuals holding a DCA license so that they may legally operate in New York City. Temporary street fair vendors are not included in this data set.

View Data Visualize ▾ Export API ...

Download Legally Operating Businesses ✕

Download Legally Operating Businesses for offline use in other applications.

CSV CSV for Excel

Additional Formats

[CSV for Excel \(Europe\)](#) [TSV for Excel](#)

[RDF](#) [XML](#)

[RSS](#)

About this Dataset

Updated **August 23, 2019**

Data Last Updated	August 23, 2019	Metadata Last Updated	November 23, 2018
Date Created	March 6, 2015		

Update

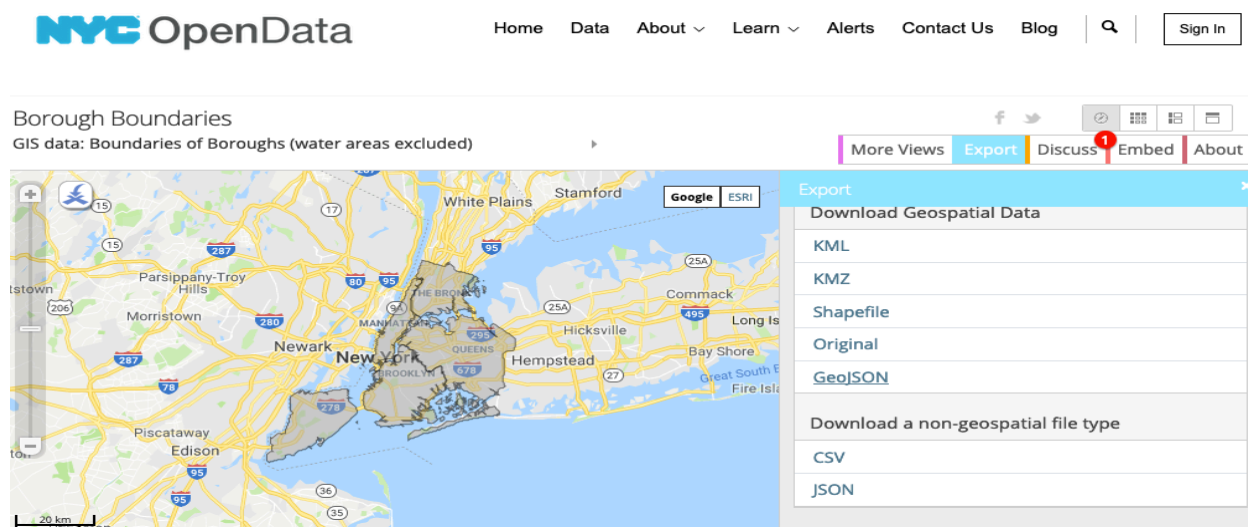
Update Frequency	Weekly
Automation	Yes
Date Made Public	2/22/2016

Dataset Information

2.2 GeoJSON of New York City

To visualize the distribution of the businesses across the NYC boroughs, Choropleth maps could be utilized, for that GeoJSON file is required, and this available in NYC Open data portal in the below link,

url: <https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-i8zm>



NYC OpenData

Home Data About Learn Alerts Contact Us Blog Sign in

Borough Boundaries

GIS data: Boundaries of Boroughs (water areas excluded)

More Views Export Discuss Embed About

Export

Download Geospatial Data

- KML
- KMZ
- Shapefile
- Original
- GeoJSON

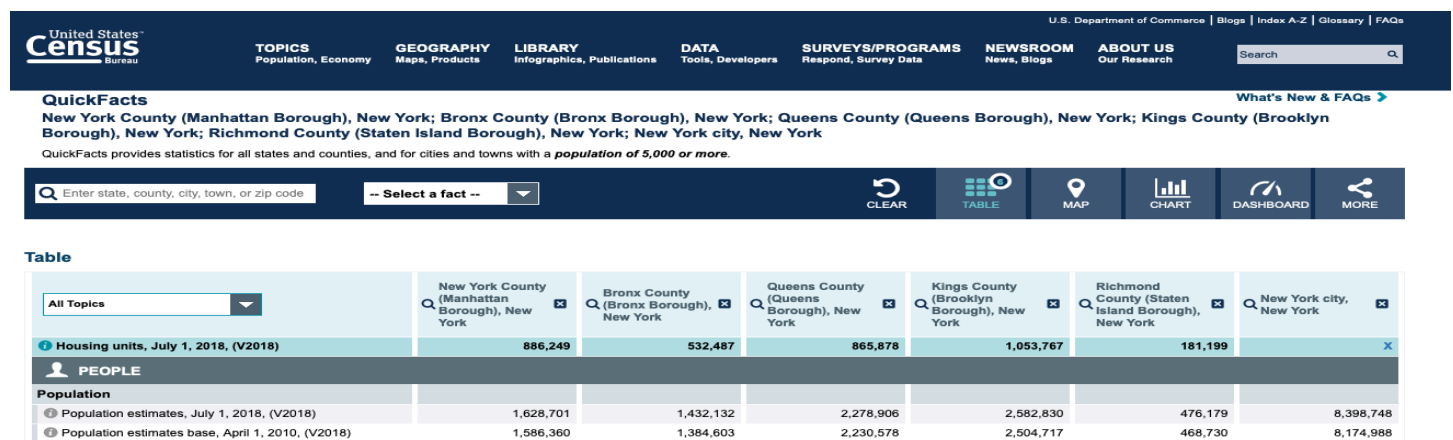
Download a non-geospatial file type

- CSV
- JSON

2.3 Demography info of New York City

Apart from the businesses, to analyze the boroughs and its population, demography data like borough's business volume, residents age, education, and income are required, which are available in the US Census portal.

url: <https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork,bronxcountybronxboroughnewyork,queenscountyqueensboroughnewyork,kingscountybrooklynboroughnewyork,richmondcountystatenislandboroughnewyork,newyorkcitynewyork/HSG010218>



United States Census Bureau

TOPICS: Population, Economy | GEOGRAPHY: Maps, Products | LIBRARY: Infographics, Publications | DATA: Tools, Developers | SURVEYS/PROGRAMS: Respond, Survey Data | NEWSROOM: News, Blogs | ABOUT US: Our Research

QuickFacts

New York County (Manhattan Borough), New York; Bronx County (Bronx Borough), New York; Queens County (Queens Borough), New York; Kings County (Brooklyn Borough), New York; Richmond County (Staten Island Borough), New York; New York city, New York

QuickFacts provides statistics for all states and counties, and for cities and towns with a population of 5,000 or more.

Enter state, county, city, town, or zip code

Table

All Topics	New York County (Manhattan Borough), New York	Bronx County (Bronx Borough), New York	Queens County (Queens Borough), New York	Kings County (Brooklyn Borough), New York	Richmond County (Staten Island Borough), New York	New York city, New York
Housing units, July 1, 2018, (V2018)	886,249	532,487	865,878	1,053,767	181,199	
PEOPLE						
Population						
Population estimates, July 1, 2018, (V2018)	1,628,701	1,432,132	2,278,906	2,582,830	476,179	8,398,748
Population estimates base, April 1, 2010, (V2010)	1,586,360	1,384,603	2,230,578	2,504,717	468,730	8,174,988

2.4 Borough Neighborhoods

To perform k-means clustering, we need the details of the boroughs' neighborhoods, their geo coordinates. Those details can be accessed from New York university's open database.

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

2.5 Four Square API

The key datapoint of this project is to identify the locations where the people frequently visits during anytime of the day. To find the places we need the details about the venues of the boroughs, and its available in Foursquare's developers portal. With its API, the below link could be accessed.

url:https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

3 Methodology

3.1 Data Cleansing & Pre-Processing

The main focus in the data cleansing process was to refine the raw master data into a proper form suitable to do exploratory analysis.

In the of the active businesses database of New York City, which is sourced from New York open data portal, the initial data frame of the raw data had many NaN values and other irrelevant columns for our analysis.

	DCA License Number	License Type	License Expiration Date	License Status	License Creation Date	Industry	Business Name	Business Name 2	Address Building	Address Street Name	...	Community Board	Council District	BIN	BBL	NTA
0	1232665-DCA	Individual	02/28/2021	Active	07/10/2006	Home Improvement Salesperson	CATALFUMO, DANIEL J	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
1	1217192-DCA	Individual	02/28/2021	Active	01/09/2006	Home Improvement Salesperson	MICHILLI, ANGELO	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
2	0868067-DCA	Individual	02/28/2021	Active	10/04/1994	Home Improvement Salesperson	BURKE, EDWARD	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
3	2012563-DCA	Individual	02/28/2021	Active	08/27/2014	Home Improvement Salesperson	CHEN, YI FA	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
4	1351461-DCA	Individual	04/30/2020	Active	04/23/2010	Pedicab Driver	HASANOV, JEYHUN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN

5 rows x 27 columns

Also, in the dataframe there were many businesses listed under Individual license type. Since the objective is to locate appropriate locations which are frequented by people, it was assumed that only businesses establishments will have traffic of people, so only those type was considered. Finally,

after removing the rows with null values on key columns like Business Name, Address, Latitude, and Longitude the cleansed data frame appear as below.

	Business Name	zip	Borough	Longitude	Latitude
82	Rosalie Tuzzino	10013	Manhattan	-73.997619	40.718663
84	NEPTUNE AIR CONDITIONING INC	10022	Manhattan	-73.971827	40.757216
85	TITAN CONTRACTING CORP	10004	Manhattan	-73.980236	40.641462
86	SPRINT SPECTRUM L.P.	10475	Bronx	-73.830453	40.865956
88	RUN FENG TRADING INC.	10013	Manhattan	-73.996970	40.718101

The next cleansing was done on the demography dataset, which was sourced from US Census portal. The original data appeared as below.

Table

All Topics	New York County (Manhattan Borough), New York	Bronx County (Bronx Borough), New York	Queens County (Queens Borough), New York	Kings County (Brooklyn Borough), New York	Richmond County (Staten Island Borough), New York	New York city, New York
Housing units, July 1, 2018, (V2018)	886,249	532,487	865,878	1,053,767	181,199	X
PEOPLE						
Population						
Population estimates, July 1, 2018, (V2018)	1,628,701	1,432,132	2,278,906	2,582,830	476,179	8,398,748
Population estimates base, April 1, 2010, (V2018)	1,586,360	1,384,603	2,230,578	2,504,717	468,730	8,174,988
Population, percent change - April 1, 2010 (estimates base) to July 1, 2018, (V2018)	2.7%	3.4%	2.2%	3.1%	1.6%	2.7%
Population, Census, April 1, 2010	1,585,873	1,385,108	2,230,722	2,504,700	468,730	8,175,133
Age and Sex						
Persons under 5 years, percent	▲ 4.7%	▲ 7.2%	▲ 6.2%	▲ 7.2%	▲ 5.8%	▲ 6.5%
Persons under 18 years, percent	▲ 14.3%	▲ 24.8%	▲ 20.1%	▲ 22.8%	▲ 21.8%	▲ 21.0%
Persons 65 years and over, percent	▲ 16.5%	▲ 12.8%	▲ 15.7%	▲ 13.9%	▲ 16.2%	▲ 13.6%
Female persons, percent	▲ 52.7%	▲ 52.9%	▲ 51.5%	▲ 52.6%	▲ 51.5%	▲ 52.3%
Race and Hispanic Origin						
White alone, percent	▲ 64.5%	▲ 44.9%	▲ 47.9%	▲ 49.5%	▲ 75.2%	▲ 42.8%
Black or African American alone, percent (a)	▲ 17.9%	▲ 43.6%	▲ 20.7%	▲ 34.1%	▲ 11.7%	▲ 24.3%
American Indian and Alaska Native alone, percent (a)	▲ 1.2%	▲ 2.9%	▲ 1.3%	▲ 0.9%	▲ 0.6%	▲ 0.4%
Asian alone, percent (a)	▲ 12.8%	▲ 4.5%	▲ 26.8%	▲ 12.7%	▲ 10.2%	▲ 14.0%
Native Hawaiian and Other Pacific Islander alone, percent (a)	▲ 0.2%	▲ 0.4%	▲ 0.2%	▲ 0.1%	▲ 0.1%	▲ 0.1%
Two or More Races, percent	▲ 3.4%	▲ 3.7%	▲ 3.0%	▲ 2.7%	▲ 2.1%	▲ 3.3%
Hispanic or Latino, percent (b)	▲ 25.9%	▲ 56.4%	▲ 28.1%	▲ 19.1%	▲ 18.7%	▲ 29.1%
White alone, not Hispanic or Latino, percent	▲ 47.0%	▲ 9.1%	▲ 25.0%	▲ 36.4%	▲ 60.3%	▲ 32.1%
Population Characteristics						
Veterans, 2013-2017	32,192	27,604	45,662	40,990	17,017	163,465
Foreign born persons, percent, 2013-2017	28.9%	35.3%	47.5%	36.9%	22.2%	37.2%
Housing						
Housing units, July 1, 2018, (V2018)	886,249	532,487	865,878	1,053,767	181,199	X
Owner-occupied housing unit rate, 2013-2017	24.1%	19.7%	44.5%	30.0%	69.5%	32.6%
Median value of owner-occupied housing units, 2013-2017	\$915,300	\$371,800	\$481,300	\$623,900	\$460,200	\$538,700
Median selected monthly owner costs -with a mortgage, 2013-2017	\$3,112	\$2,370	\$2,482	\$2,723	\$2,433	\$2,588

In this case, data cleansing happened in the source csv file itself. All the irrelevant data was deleted, and a new dataframe was created with the desired information.

	Manhattan	Bronx	Queens	Brooklyn	Staten Island	NYC Total
Fact						
Population	1628.70	1432.13	2278.91	2582.83	476.18	8398.75
population between 18 - 65 years in numbers	1127.06	893.65	1463.06	1634.93	295.23	5492.78
Bachelors degree in numbers	988.62	277.83	701.90	909.16	152.85	3082.34
Labor Force in Numbers	1094.49	852.12	1460.78	1637.51	279.04	5333.20
Hotel and Food Service Sales	2038.26	100.51	313.91	245.34	47.26	2745.29
Total Retail Sales	4404.00	687.28	1700.32	2053.31	381.60	9226.50
Per Capita Income	695.29	197.21	288.14	299.28	339.22	357.61
Land area in square miles, 2010	22.83	42.10	108.53	70.82	58.37	302.64

The next dataset was New York City's neighborhoods, since their geo coordinates was an essential data, it was sourced from New York University's open database portal. The raw data from the json file appeared as below,

```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'type': 'Feature',
  'id': 'nyu_2451_34572.1',
  'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]}},
  'geometry_name': 'geom',
  'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
  40.89470517661,
  -73.84720052054902,
  40.89470517661]}},
  {'type': 'Feature',
  'id': 'nyu_2451_34572.2',
  'geometry': {'type': 'Point',
  'coordinates': [-73.82993910812398, 40.87429419303012]}},
  'geometry_name': 'geom',
  'properties': {'name': 'Co-op City',
  'stacked': 2,
  'annoline1': 'Co-op'.
```

With a careful analysis, it was identified that features key in the dictionary list has all the required details. Hence a new dataframe with details in the features key was created as below,

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

Similary, to get the venues of each neighborhood, data has to be sourced from Foursquare.com, with their API provision, by passing the relevant geo coordinates as input, a dataframe with list of venues of each neighborhood was created as below,

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	Ho' Brah Taco Joint	40.622960	-74.031371	Taco Place

3.2 Exploratory Analysis

With the cleansed data, the explorary 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.

3.2.1 Choropleth map to visualize NYC's business distribution

Based on the active business data, Choropleth map was plotted to visualize the distribution of the NYC's businesses. Here the segmentation was done based on the NYC boroughs, and appropriate GeoJSON file was used to plot the below figure,



Fig 1. NYC's Borough Boundaries (GeoJSON)

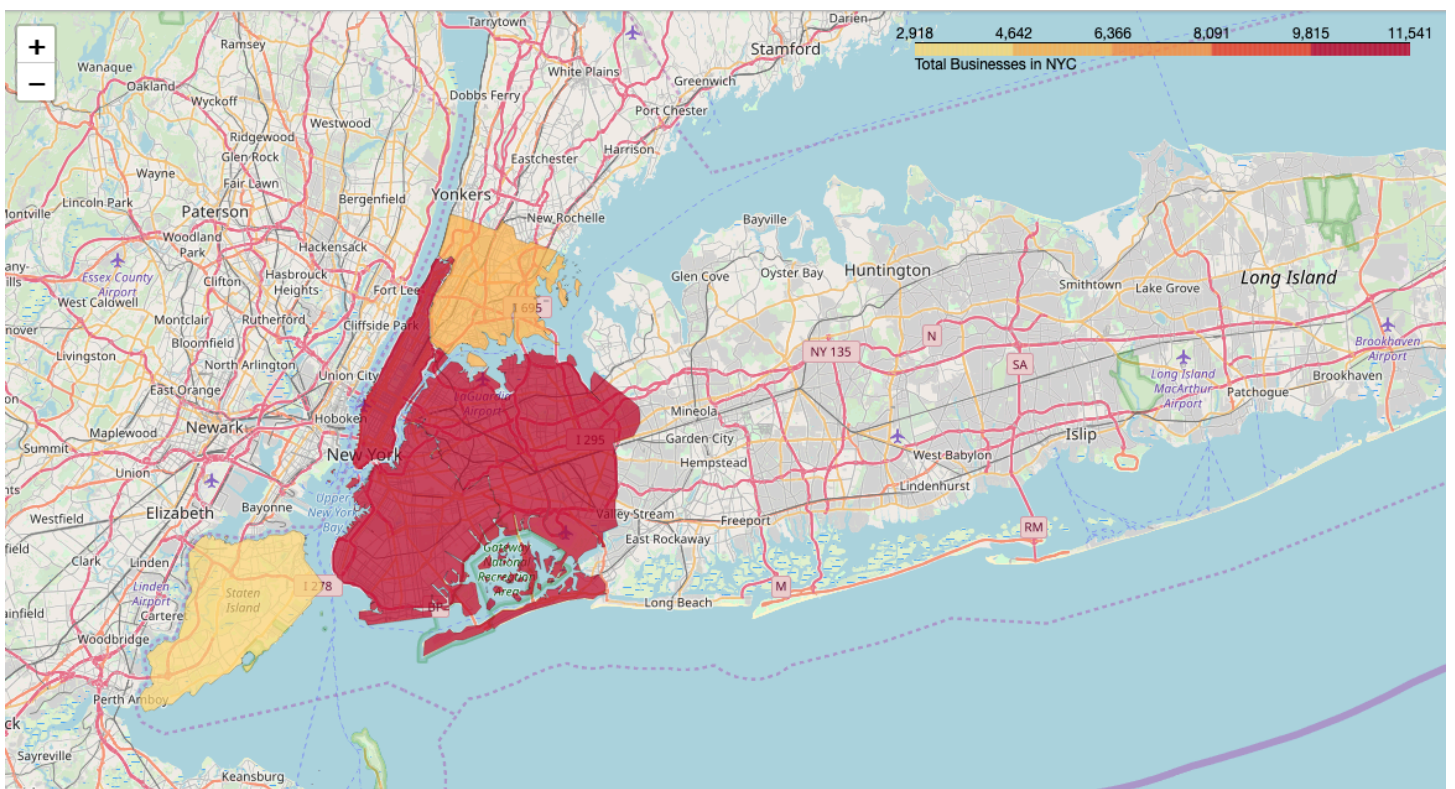


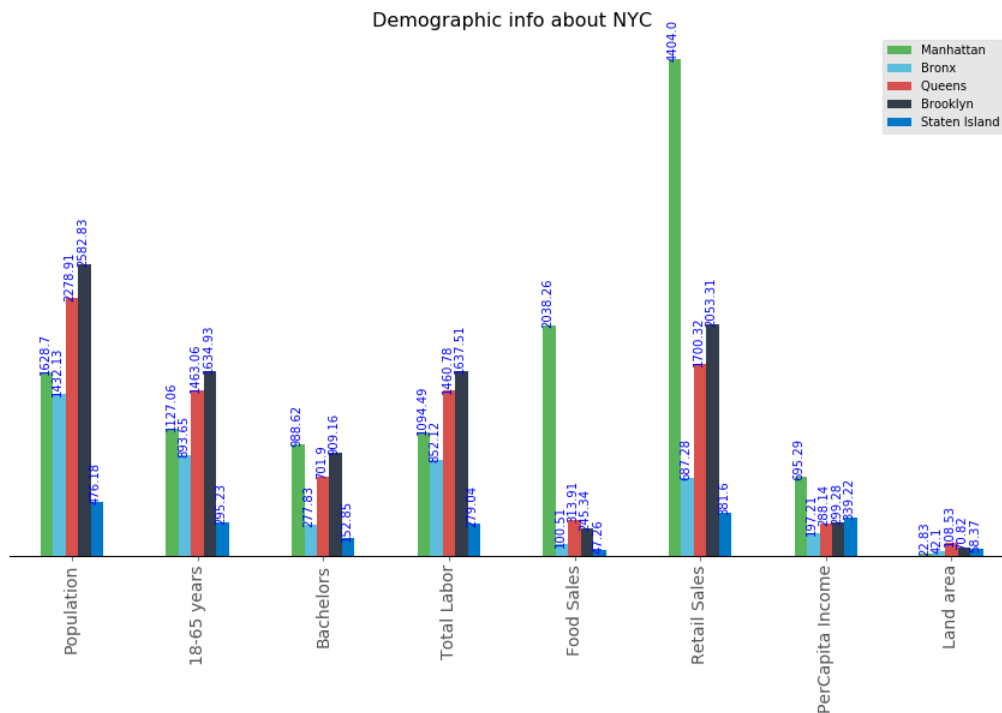
Fig 2. Choropleth Map showing the distribution of active businesses in NYC's boroughs.

It was clearly evident that Manhattan, Queens and Brooklyn boroughs has the same range of 9800 to 11600 business establishments. With this multiple choices, a clear selection of location was not

possible based on only the business data. Hence further analysis was required to narrow down the selection.

3.2.2 Analyzing NYC's population characteristics.

US census data on NYC boroughs has the required demography details like population, education level, age, sales volume.etc. Based on the details a bar chart was plotted, which enables to compare the respective boroughs. The plotted bar chart appears as below,



As seen in the bar chart, Manhattan borough (Green) and Brooklyn borough (Black) are almost comparable in most of the parameters. Also, considering the fact that the start-up company has resource constraints, it has to stay low on the budget. So the choice was Brooklyn borough.

3.2.3 Finding Top 10 common venues in Brooklyn Neighborhoods.

Having the borough identified, the next step was to drill down to identify the venues in the borough's neighborhoods, and in those venues have to identify the top common venues. Since the objective is to set up kiosks explaining the new solution on retirement portfolio management, its imperative to have those set up in locations that has potential to attract major traffic. This exercise involves multiple stages, the summary of those stages are as below,

- In the Brooklyn venues data frame, performed one hot encoding (converting categorical variables as binary vectors). In this case those binary vectors represents the frequency of occurrence in each of the venues.
- Create a dataframe grouped by the means of frequency of occurrence.
- Create a list with top 5 venues in each neighborhood like below based on the new dataframe

----Bath Beach----

	venue	freq
0	Pizza Place	0.10
1	Chinese Restaurant	0.06
2	Bank	0.06
3	Bakery	0.05
4	Cantonese Restaurant	0.04

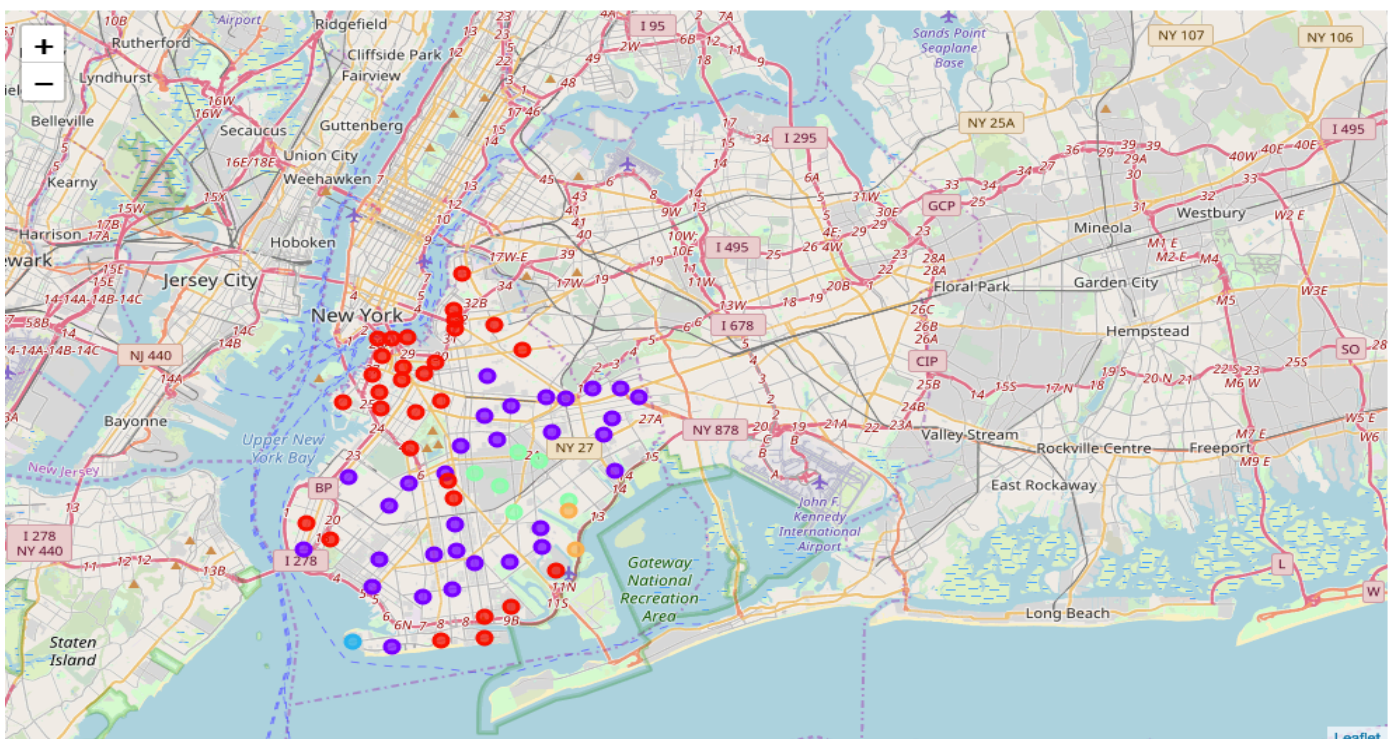
----Bay Ridge----

	venue	freq
0	Spa	0.07
1	Italian Restaurant	0.06
2	Pizza Place	0.05
3	Bar	0.05
4	Cosmetics Shop	0.04

-
- Next is conversion of the above list as a dataframe, and to generate cluster labels by running k-means clustering, and finally to create a new data frame which is a combination of cluster data and the venues data.

3.2.4 Visualize Brooklyn Clusters

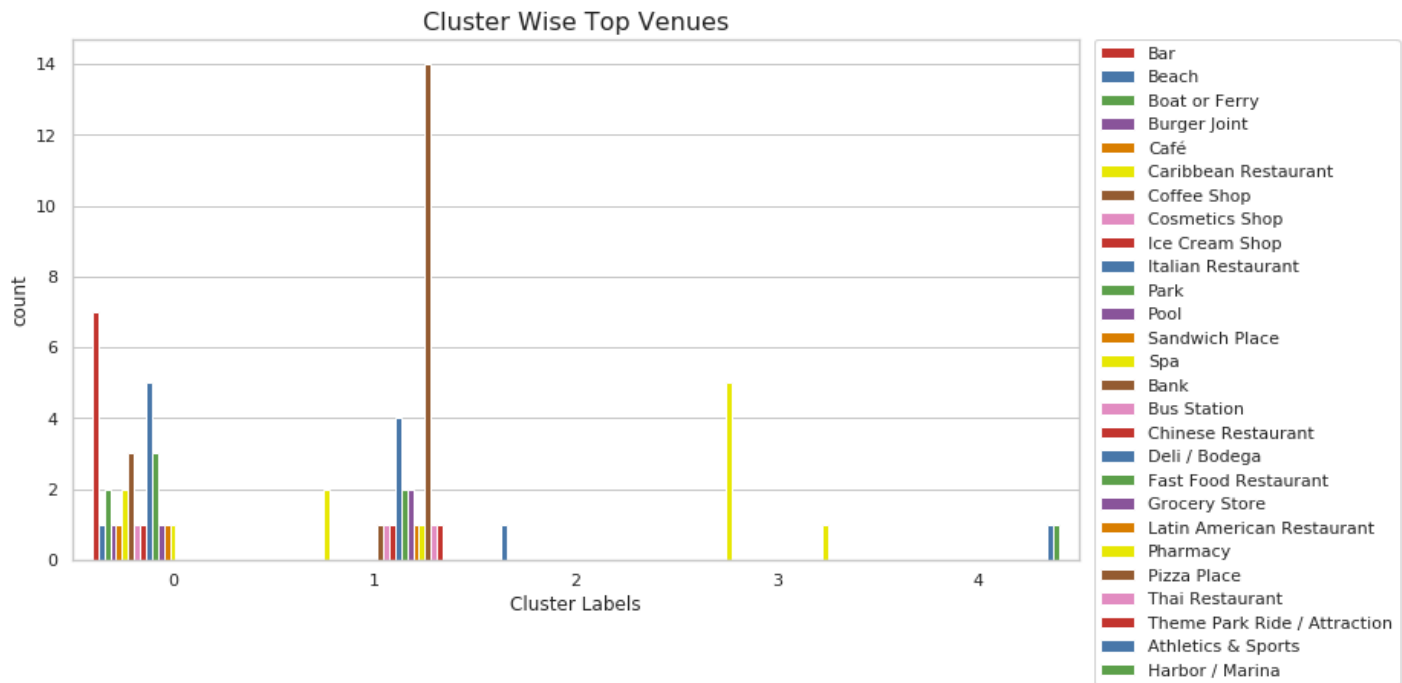
With the help of cluster data and the venues data in the merged dataframe, the below folium map was plotted.



The map shows that the red clusters are concentrated near Manhattan and others are wide spread across the neighborhood. We can assume that the venues in the red clusters are the optimum choice to set up the kiosks. However, this assumption could be validated by analyzing the clusters.

3.2.5 Analyze Brooklyn Clusters

To perform further analysis on Brooklyn clusters, the master dataframe with cluster data and venues data was sliced to create new dataframe with 1st common venues across all clusters. From the new data frame a seaborn bar chart was plotted to visualize the venues distribution cluster wise, and the bar chart appear as below



The above bar chart provides more insights about the venues. With high level 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.

4 Conclusion & Recommendations

In this analysis, data on New York City's boroughs were analyzed based on their neighborhood's geo coordinates, and with its demography details of population. It was also observed through exploratory analysis that the data was sufficient to identify the borough which fits the selection criteria. There was assumption made that the areas having more business establishments will have more working force, and there will significant number of them who save for their retirements. In the venue selection also, it was assumed that the target customers frequent neighborhood venues during the day and having kiosks near them will give the maximum returns in the initiative. So, with the data, assumptions and analysis we can conclude that Coffee Shops & Pizza Restaurants located in Brooklyn cluster '0' is the ideal choice to set up the kiosks.

Though we have suggested 1st common venues in cluster '0' as part of this analysis, we also recommend to analyze 2nd common venues across the clusters or have an hybrid approach of having a mix of 1st common venues and 2nd common venues of cluster '0' to set up the kiosks.