capstone-w5-report

June 22, 2020

1 New York - Population, Venues, Housing Price Analysis

1.0.1 Arun Leo Prakash

18-Jun-2020

1.1 Table of Contents

Introduction New York City (NYC), often called New York (NY), is the most populous city in the United States. With an estimated 2019 population of **8,336,817** distributed over about **302.6** square miles (784 km2), New York is also the most densely populated major city in the United States.

Situated on one of the world's largest natural harbors, New York City is composed of five boroughs, each of which is a county of the State of New York. The five boroughs—Brooklyn, Queens, Manhattan, the Bronx, and Staten Island—were consolidated into a single city in 1898. The city and its metropolitan area constitute the premier gateway for legal immigration to the United States. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. New York is home to more than 3.2 million residents born outside the United States, the largest foreign-born population of any city in the world as of 2016. As of 2019, the New York metropolitan area is estimated to produce a gross metropolitan product (GMP) of $2.0 trillion. If the New York metropolitan area were a sovereign state, it would have the eighth-largest economy in the world. New York is home to the highest number of billionaires of any city in the world.

Business Problem With a population of 8 million, New York is a city with a high population and produces high GDP. Being such a crowded city leads the owners of shops and social sharing places in the city where the population is dense. Business investors expects lower real estate cost,

with high density of population and the type of business they want to install is less intense. It is difficult to obtain information that will guide investors in this direction, nowadays.

When we consider all these problems, we can create a map and information chart where the real estate index is placed on New York and each district is clustered according to the venue density. This would help the investor to decide the ideal location to run the business based on the factors mentioned above

Data Data requirements includes a. spatial data of new york to build maps with boundaries, b. average sales price per sqm for every borough, c. venue data of the neighborhoods. Venue like coffee shops, entertainment related venues depending on the business requirements would be considered as a baseline parameter.

Data Source 1. Spatial data for NewYork will be downloaded from https://cocl.us/new_york_dataset/newyork_data.json 2. Population Data per borough from https://en.wikipedia.org/wiki/New_York_City 3. Property sales data to collect average sales price from https://www.kaggle.com/new-york-city/nyc-property-sales

First step, load the new york Geospatial data into dataframe for map generation.

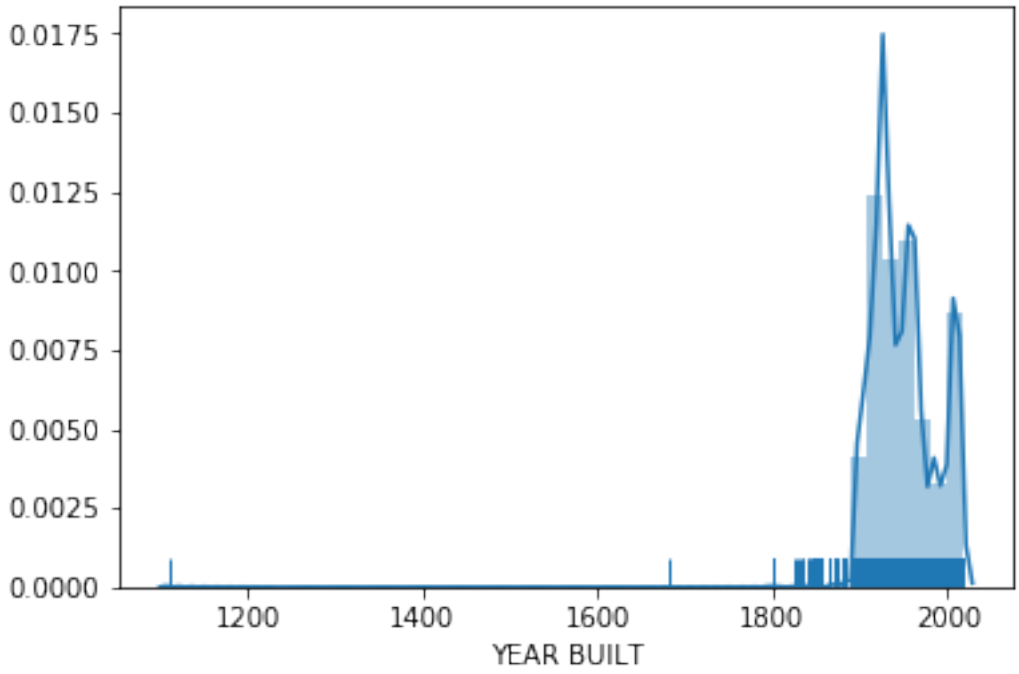
```
The dataframe has 5 boroughs and 306 neighborhoods.
```

House Sales Data In the below section New York Housing Sales Data is loaded, cleansed for using it in our comparison

```
0
```

There are too many columns we need only YEAR BUILT, BOROUGH, SALE PRICE. First step let us remove the columns and load it into a new data frame.

Exploratory analysis - the reports below helps in understanding outliers

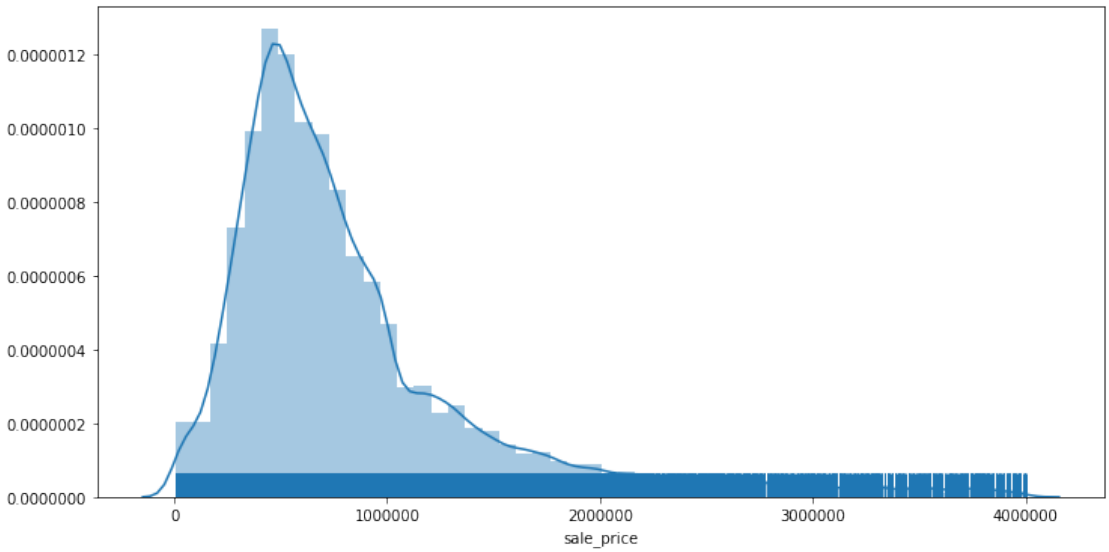


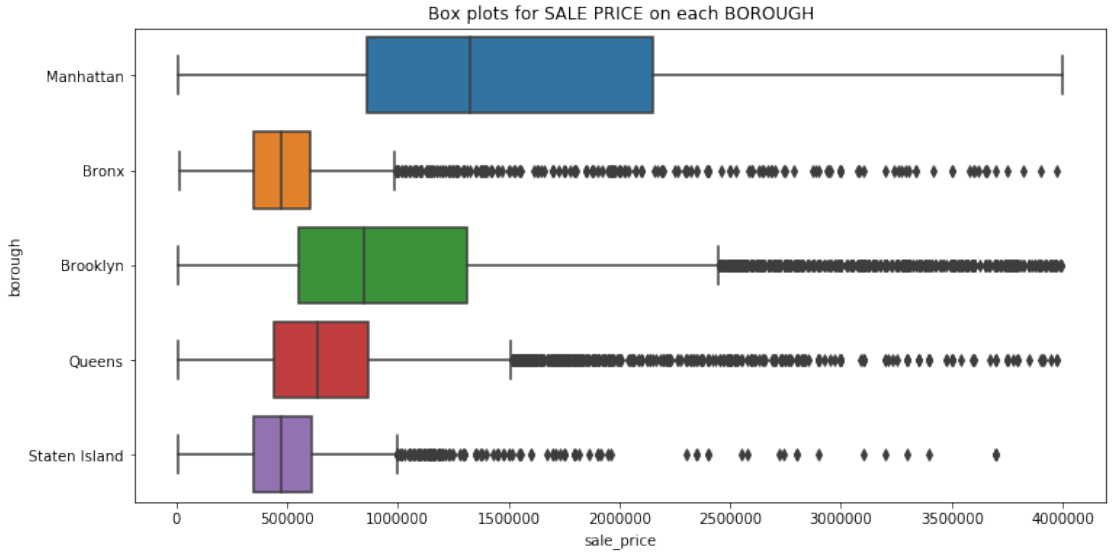
next step, convert the sale_price to numeric

```
Out[26]: 0.2159664549747598
```

21% of the sale prices are either less than 10,000 or greater than $10,000,000. We have to drop all these observations from the data

```
Out[27]:          sale_price    YEAR BUILT   TOTAL UNITS
         count  3.845700e+04  38457.000000  38457.000000
         mean   1.077431e+06   1953.056115      2.191123
         std    1.270522e+06     37.838285     17.841976
         min    1.000100e+04   1800.000000      1.000000
         25%    4.550000e+05   1923.000000      1.000000
         50%    6.880000e+05   1941.000000      1.000000
         75%    1.110000e+06   1989.000000      2.000000
         max    9.999999e+06   2017.000000   2261.000000
```





Out[32]:

	borough	sale_price
0	Staten Island	5.067751e+05
1	Bronx	5.579480e+05
2	Queens	7.098848e+05
3	Brooklyn	1.038700e+06
4	Manhattan	1.571737e+06

New York population data below Population, Density & GDP data is extracted manually from wikipedia.

Out[36]:

	borough	county	population	gdp_billions	gdp_capita	sqms	\
4	Staten Island	Richmond	476143	14.514	30500	58.37	
0	Bronx	Bronx	1418207	42.695	30100	42.10	
2	Manhattan	New York	1628706	600.244	368500	22.83	
3	Queens	Queens	2253858	93.310	41400	108.53	
1	Brooklyn	Kings	2559903	91.559	35800	70.82	

	sqkm	pers_sqms	pers_sqkm
4	151.18	8157	3150
0	109.04	33867	13006
2	59.13	71341	27544
3	281.09	20767	8018
1	183.42	36147	13957

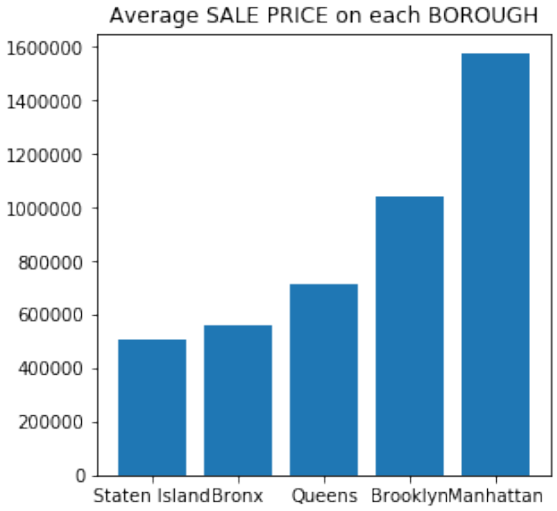
Out[40]:

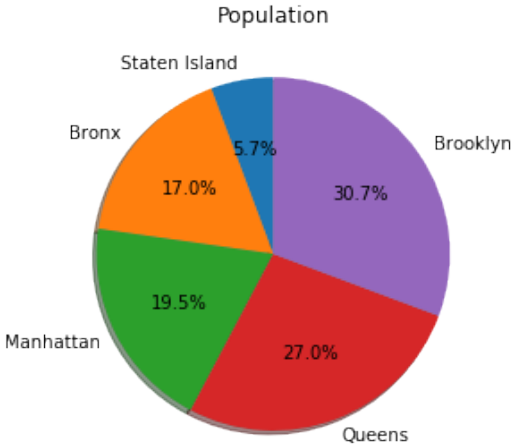
	borough	county	population	gdp_billions	gdp_capita	sqms	\
0	Staten Island	Richmond	476143	14.514	30500	58.37	
1	Bronx	Bronx	1418207	42.695	30100	42.10	
2	Manhattan	New York	1628706	600.244	368500	22.83	
3	Queens	Queens	2253858	93.310	41400	108.53	

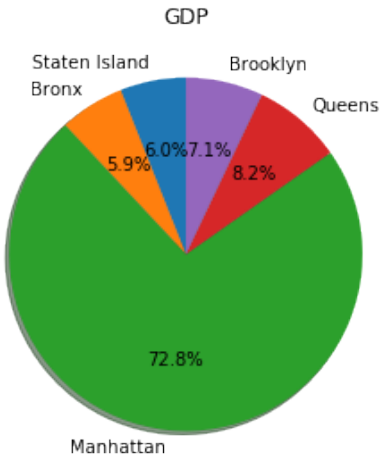
```
4       Brooklyn     Kings     2559903        91.559       35800    70.82

     sqkm  pers_sqms  pers_sqkm    sale_price
0  151.18       8157       3150  5.067751e+05
1  109.04      33867      13006  5.579480e+05
2   59.13      71341      27544  1.571737e+06
3  281.09      20767       8018  7.098848e+05
4  183.42      36147      13957  1.038700e+06
```

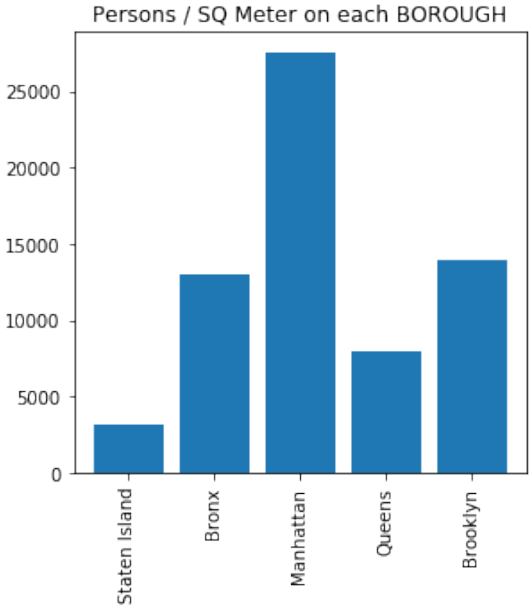
Data Exploration

#Comparison of GDP, Population, Sale Price









GDP The Major contributor of GDP is Manhattan, followed by Queens, Brooklyn, Stalen Island & Bronx. **Avg sale price** Prices were Manhattan, Brooklyn, Queens, Staten Island, Bronx. **Person/SQM** Person per SQM is high in the order Manhattan, Brooklyn, Bronx, Queens. **Population** Brooklyn, Queens, Manhattan, Bronx & Staten Island

I wish to perform neighborhoods exploration for **Queens** considering the facts mentioned in our problem description - lower real estate cost, population and better GDP. This will help in lowering the investment as well as consider better returns. In the next section lets explore the neighborhood data followed by segmentation to find top venues.

Methodology **Google Map API**, 'Search Nearby' option to get the center coordinates of the each Borough. **Population data** is captured from wiki page. We will be using **Foursquare API** for Neighborhoods data exploration. The information we want to focus on are shopping venues, coffee shops, and entertainment venues. We will choose top 2 boroughs based on population: **Manhattan & Brooklyn**. We need to apply Neighborhood Segmentation and Clustering to analyzing the neighborhood data and prioritize the best shopping location in both boroughs based on foot traffic and type of venues available. This helps the investor to choose the best place for business investment.

I am using Google geolocator API for finding latitude / longitude details. Folium library to load the New York map

```
The geograpical coordinate of New York City are 40.7127281, -74.0060152.
```

```
Out[54]: <folium.folium.Map at 0x2ecb3fa3eb8>
```

1.1.1 Neighborhoods exploration

we are going to start utilizing the Foursquare API to explore the neighborhoods for segment them.

Define Foursquare Credentials and Version

Get the neighborhood's name.

I have Limit to 100 venues within a radius of 500 meters.

Method is created to reuse the code for both the borough

```
There are 315 uniques categories.
```

```
Out[60]: (3785, 315)
```

```
(81, 315)
```

```
Out[63]:       Neighborhood 1st Most Common Venue 2nd Most Common Venue  \
         0          Arverne             Surf Spot                 Beach
         1          Astoria                   Bar      Greek Restaurant
         2  Astoria Heights   Rental Car Location                Bakery
         3       Auburndale     Korean Restaurant        Cosmetics Shop
         4      Bay Terrace        Clothing Store        Cosmetics Shop
```

```
       3rd Most Common Venue 4th Most Common Venue 5th Most Common Venue  \
0              Deli / Bodega        Sandwich Place              Bus Stop
1  Middle Eastern Restaurant           Pizza Place    Seafood Restaurant
2               Bus Station           Pizza Place                 Café
3               Pizza Place      Sushi Restaurant      Greek Restaurant
4         Mobile Phone Shop         Women's Store            Kids Store

  6th Most Common Venue   7th Most Common Venue 8th Most Common Venue  \
0           Donut Shop  Furniture / Home Store                 Café
1          Coffee Shop              Hookah Bar         Grocery Store
2   Italian Restaurant                   Hotel            Laundromat
3       Sandwich Place                    Bank             Pet Store
4       Lingerie Store     American Restaurant            Donut Shop

  9th Most Common Venue 10th Most Common Venue
0       Bed & Breakfast            Gas Station
1        Sandwich Place      Indian Restaurant
2        Baseball Field     Chinese Restaurant
3               Pharmacy   Gym / Fitness Center
4            Shoe Store            Men's Store
```

1.1.2 Segmentation (Clustering)

Out[68]:
```
array([1, 0, 0, 0, 0, 0, 2, 0, 3, 1])
```

Out[69]:
```
   Borough     Neighborhood   Latitude  Longitude  Cluster Labels  \
0  Queens          Astoria  40.768509 -73.915654               0
1  Queens         Woodside  40.746349 -73.901842               0
2  Queens  Jackson Heights  40.751981 -73.882821               0
3  Queens         Elmhurst  40.744049 -73.881656               0
4  Queens     Howard Beach  40.654225 -73.838138               0

       1st Most Common Venue      2nd Most Common Venue  \
0                        Bar           Greek Restaurant
1           Thai Restaurant                Pizza Place
2  Latin American Restaurant  South American Restaurant
3           Thai Restaurant         Mexican Restaurant
4        Italian Restaurant                 Bagel Shop

       3rd Most Common Venue 4th Most Common Venue      5th Most Common Venue  \
0  Middle Eastern Restaurant           Pizza Place         Seafood Restaurant
1                        Bar         Grocery Store                    Bakery
2                    Bakery    Mexican Restaurant        Peruvian Restaurant
3        Chinese Restaurant                Bakery  South American Restaurant
4                   Pharmacy  Fast Food Restaurant             Sandwich Place

       6th Most Common Venue 7th Most Common Venue 8th Most Common Venue  \
```

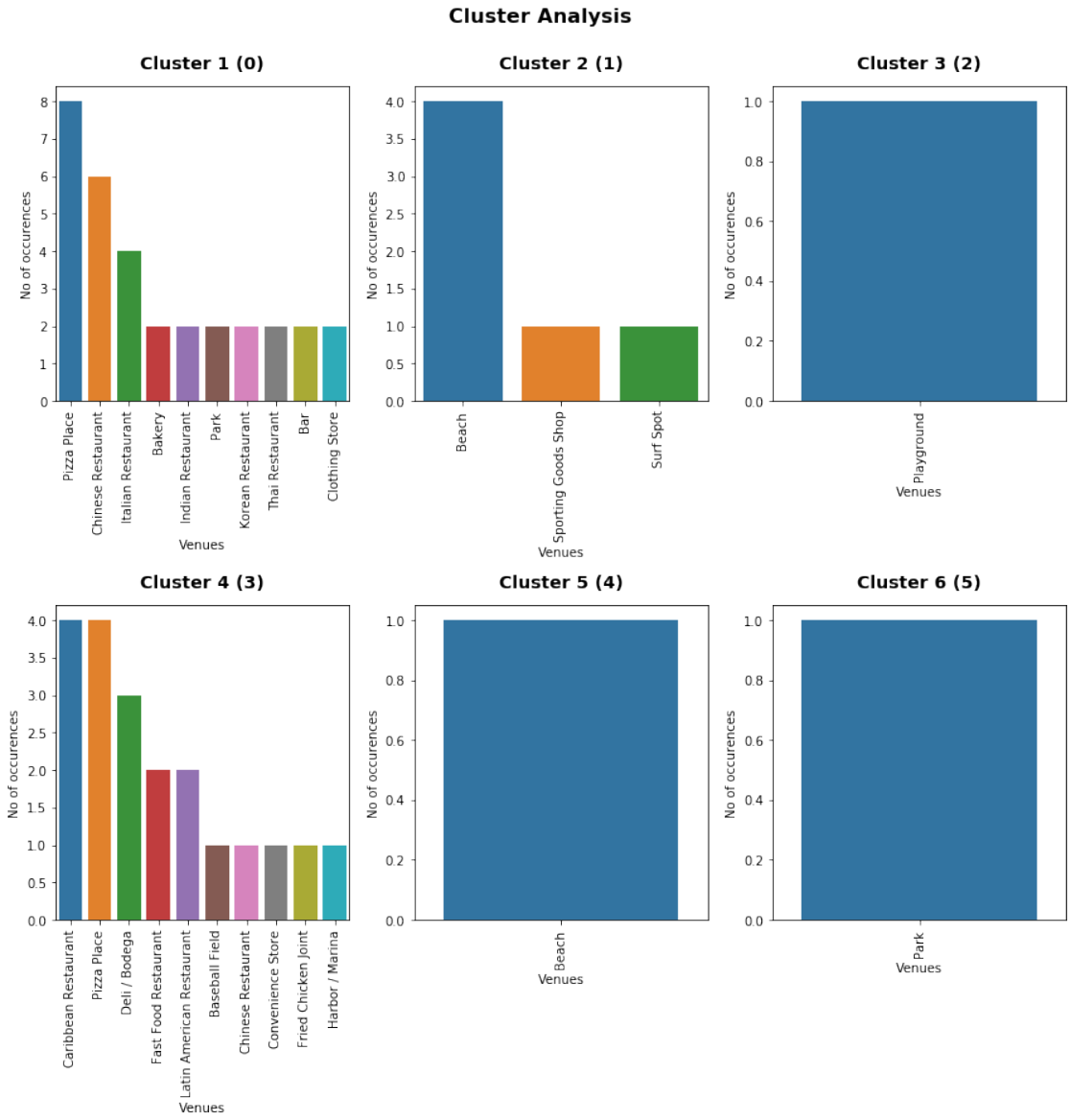
```
0                Coffee Shop            Hookah Bar         Grocery Store
1  Latin American Restaurant   American Restaurant                  Pub
2                Pizza Place       Thai Restaurant         Grocery Store
3  Latin American Restaurant         Grocery Store           Supermarket
4              Ice Cream Shop                  Park      Sushi Restaurant

  9th Most Common Venue 10th Most Common Venue  county  population  \
0        Sandwich Place      Indian Restaurant  Queens     2253858
1   Filipino Restaurant             Donut Shop  Queens     2253858
2           Coffee Shop             Donut Shop  Queens     2253858
3       Bubble Tea Shop            Snack Place  Queens     2253858
4       Other Nightlife            Bus Station  Queens     2253858

   gdp_billions  gdp_capita    sqms    sqkm  pers_sqms  pers_sqkm  \
0         93.31       41400  108.53  281.09      20767       8018
1         93.31       41400  108.53  281.09      20767       8018
2         93.31       41400  108.53  281.09      20767       8018
3         93.31       41400  108.53  281.09      20767       8018
4         93.31       41400  108.53  281.09      20767       8018

      sale_price
0  709884.787686
1  709884.787686
2  709884.787686
3  709884.787686
4  709884.787686
```

1.1.3 Results

A map plotted to visualize the neighborhoods by its clusters - chloropleth map is created to highlight the boroughs by its house average sales.

```
Out[81]: <folium.folium.Map at 0x2ecb4210710>
```

Cluster visualization, bar graphs developed to visualize what venue categories are available for each clusters.



Clusters are categorized as follows

Cluster 1 - Food/Clothing.
Cluster 2 - Beaches/sports
Cluster 3 - Sports
Cluster 4 - Food/Harbor
Cluster 5 - Beach
Cluster 6 - Park

Data exploration results helps to understand the population denisty , GDP and housing pricing to compare and decide which borough to choose for clustering analysis. From the results on clustering, based on the type of business one is investing can choose the cluster and the map helps in selecting the area in which the business can be built.

1.1.4 Discussion

Newyork is a popular city in US. The city has 5 boroughs and for this project I have considered **Queens** based on the real estate pricing (low) with good population density and good GDP rate. The interactive map could help the investors in assessment of finding a suitable neighborhood based on the business the investor is interested in.

1.1.5 Conclusion

The popular city new york in United State has Manhattan as the most expensive borough with median population comparing with its peer boroughs. Manhattan also provides 72% of the GDP. Business investors with high budget can opt for Manhattan which has better GDP. Queens with its high population and low average housing cost could be a better investment option. Cluster 1/4 with high number of restaurant venues should be a good place for investing on related business that supports restaurants and bars. eg. Food supplies etc.

Also further analysis can be done for different boroughs, cities depending on the investors needs. I could imagine this can be extended further by having webbased UI that could ask user to choose city ie. New York and then the map with choropleth can be diplayed and based on selection of the borough the report could be dynamically generated. Not just for retailers, this can be usable by home investors, government etc.

1.1.6 References

https://en.wikipedia.org/wiki/New_York_City
https://www.kaggle.com/new-york-city/nyc-property-sales
https://services5.arcgis.com/GfwWNkhOj9bNBqoJ/arcgis/rest/services/NYC_Borough_Boundary/Featu
https://geo.nyu.edu/catalog/nyu_2451_34572
https://developer.foursquare.com/
https://www.google.com/maps/
https://cdn.wallpapersafari.com/34/75/CwGD1o.jpg