Sucheta Gangullapattu

**NYC data Property Analysis - Capstone final project report**

**Introduction / Business Problem**

New York City (NYC) is one of the most populated cities in USA and second most populous city in North America. Its is comprised of 5 boroughs namely Brooklyn, Manhattan, Staten Island, Bronx and Queens. The 5 Boroughs cover overall land area of about 784 Km square with a population of about roughly 8,398,748 in year 2018. As a Neighbor of New York City, I have chosen this location for my capstone project.

New York City is culturally very diverse with population density of 159 people per square Km and described as financial, and media capital of the world and it’s a center for commerce entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

NYC is one of the largest metropolitan cities with over 20 million people and home to headquarters of United Nations. There are more than 100 neighborhoods divided among 5 boroughs with Manhattan titled the most expensive real estate Markets. New York city is most powerful city economically and financially, it is also home to largest stock exchanges the NASDAQ and New York Stock Exchange.

As we can see from the statistics NYC is a very diverse and financial capital, we can derive many ideas and problems like: if I am looking to open a restaurant or business, I would like to explore neighborhoods /areas with low real estate property values? If someone is looking for office / house to rent which area should they prefer and why?

With help of foursquare location data and raw data (NYC property data) and other tools I explore further to cluster based on borough information and venue data obtained using foursquare and come up with a solution to some of the problems mentioned above.

**Reference:**

<https://en.wikipedia.org/wiki/New_York_City>

Note: some of the statistical information (population info) has been taken from the link above.

**Data Section:**

1. <https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>

2. Foursquare location data will also be used.

Data consists of rolling property sales data for all 5 boroughs and information about taxes, type of property, neighborhood, date of sale, square footage info etc. The data corresponds to 12-month period (year 2018). The source had sales data recorded per each Borough. I have consolidated the data in one data source.

**Description:**

The data contain property sales data across all 5 boroughs

Manhattan (1), Brooklyn (3), Staten Island (5), Bronx (2), Queens (4)

Neighborhood info: Name of the Neighborhood where the property dwells

Building Class Category: 01 ONE FAMILY DWELLINGS, 21 OFFICE BUILDINGs, COMMERICAL CONDOS etc. (There are about 44 Categories)

Tax class at Present: There 3 to 4 diff tax classes applied based on building class category

Property Details: BLOCK, LOT, EASE-MENT, BUILDING CLASS AT PRESENT, ADDRESS,

APARTMENT NUMBER, ZIP CODE, RESIDENTIAL UNITS, COMMERCIAL UNITS, TOTAL UNITS,

LAND SQUARE FEET, GROSS SQUARE FEET, YEAR BUILT, TAX CLASS AT TIME OF SALE, BUILDING CLASS, TAX TIME OF SALE, SALE PRICE, SALE DATE

**Date Understanding and Preparation:**

The data is obtained from the following source:

<https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>

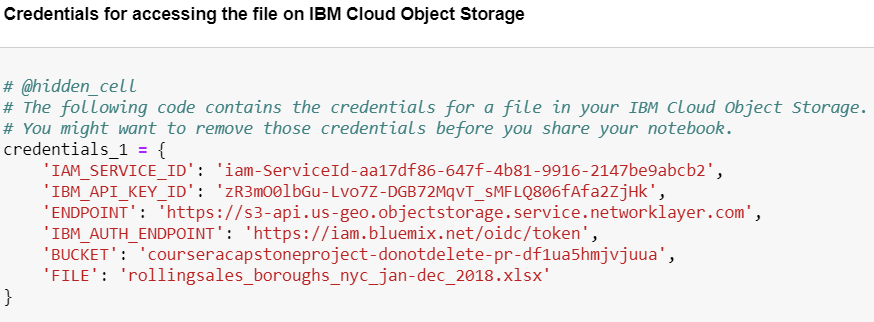
Sales data of NYC for 12-months sorted by Borough (Manhattan, Bronx, Brooklyn, Queens, Staten Island)

The source page contains a downloadable Excel file for each borough.

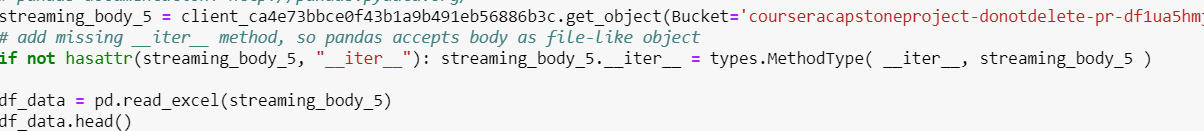


I combined the data from 5 boroughs into one Master Excel File.

I used the Master file saved on the IBM Cloud and retrieved using Credential’s.

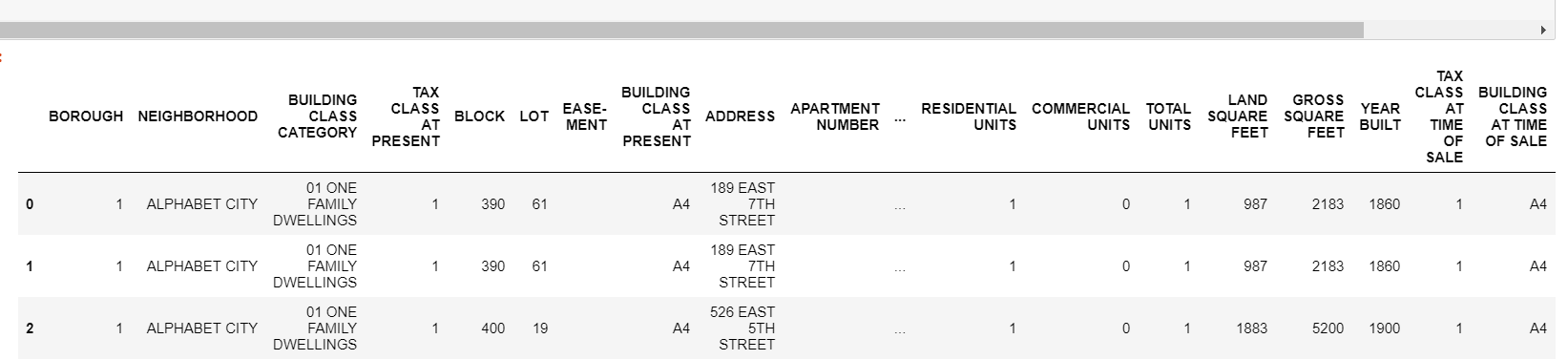


**Retrieving the file from cloud storage:**



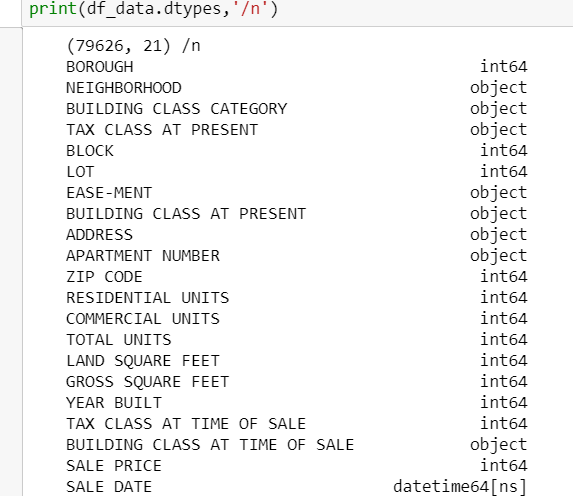
As we can see the dataset contains the following info:

The following snapshot shows the consolidated data from all five boroughs and their corresponding neighborhood’s.

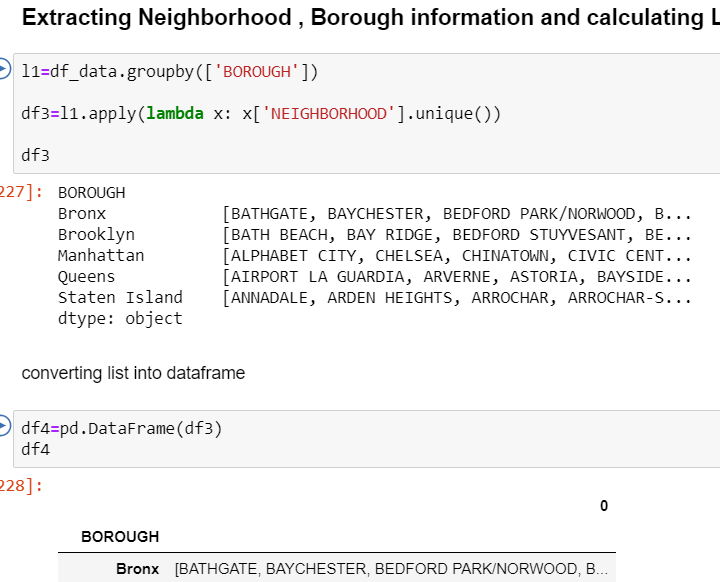


**Obtaining Longitude and Latitude information:**

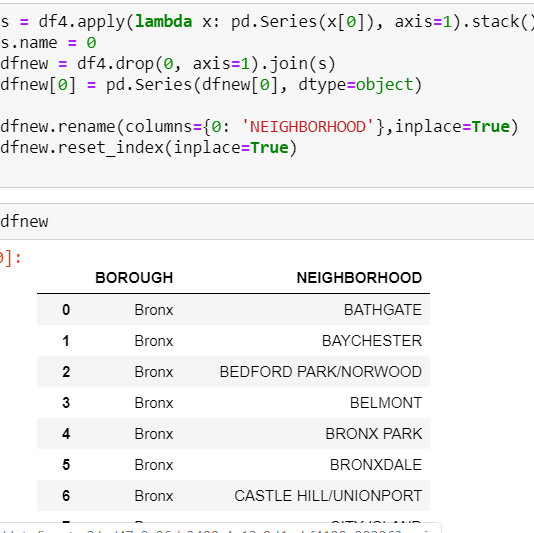
As we can see from the snapshot below that we need Longitude and latitude information in the raw Data Source (dataset), geopy Geocoding library for python has been used to generate Coordinates for Boroughs and its corresponding Neighborhood’s.



**Extracting Borough and Neighborhood Information:**

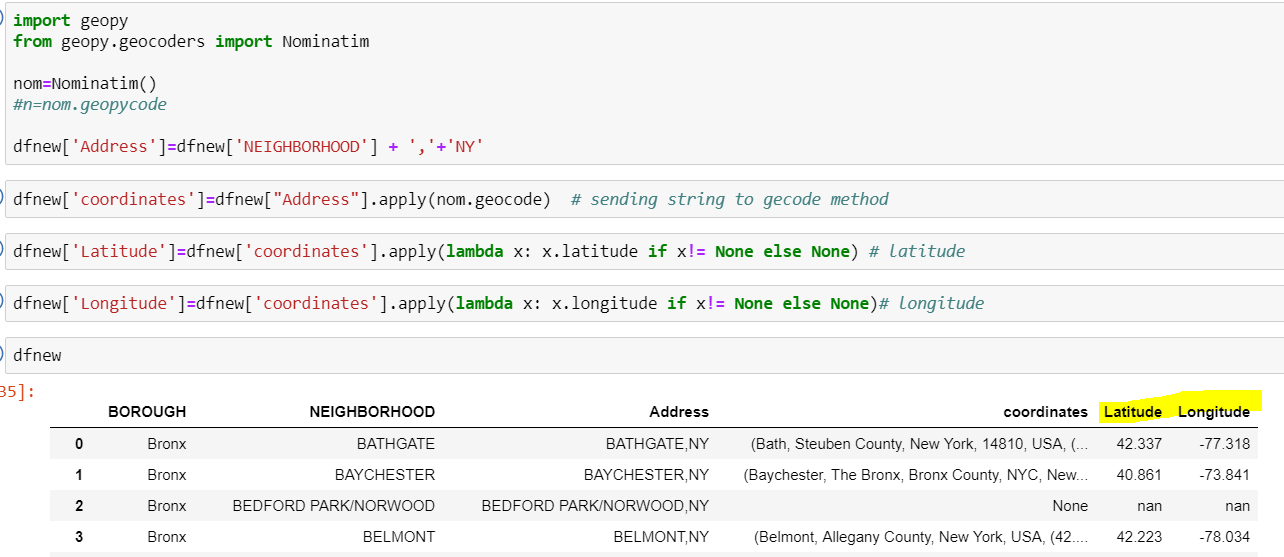


**Splitting the Neighborhood data:**

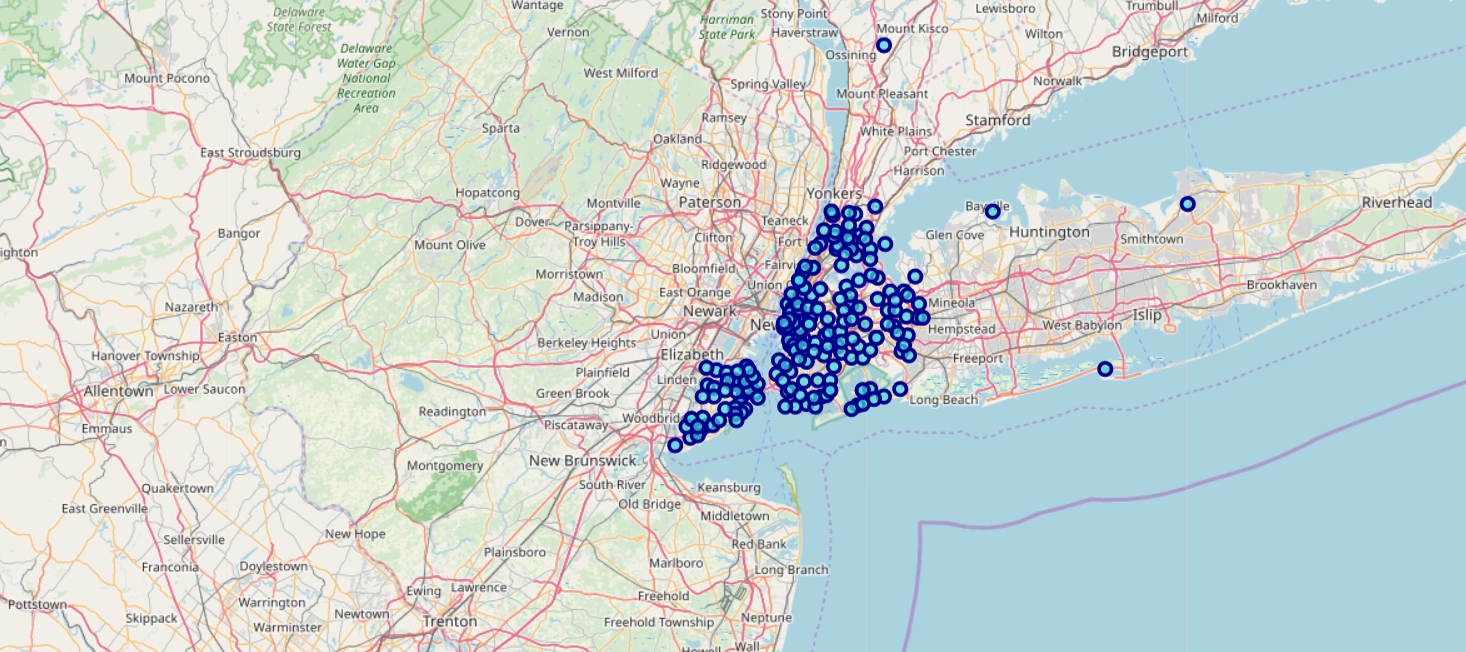


Now that the new data set has Borough/ Neighborhood data

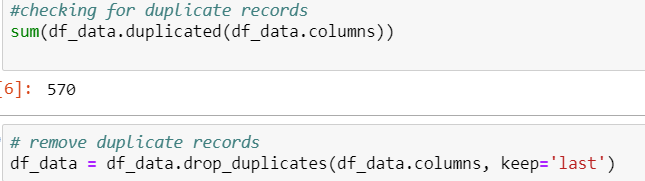
Library gepoy is used to generate Latitude and longitude information



Next, Creating a NYC map using folium with Boroughs and Neighborhood superimposed on the map.

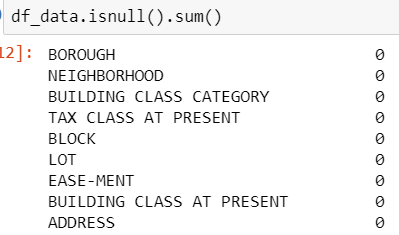


**DATA CLEANSING:**

1. Does the data contain duplicates?

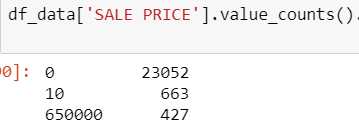
Checking and removing duplicate values

1. Checking for Null (NaN) values?



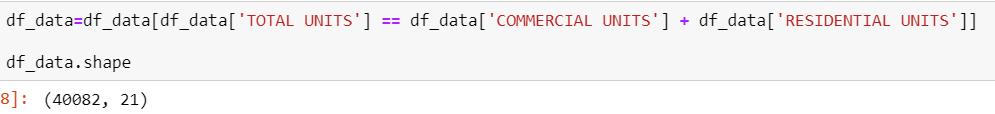
No Null Values found

1. Checking for any invalid entries?

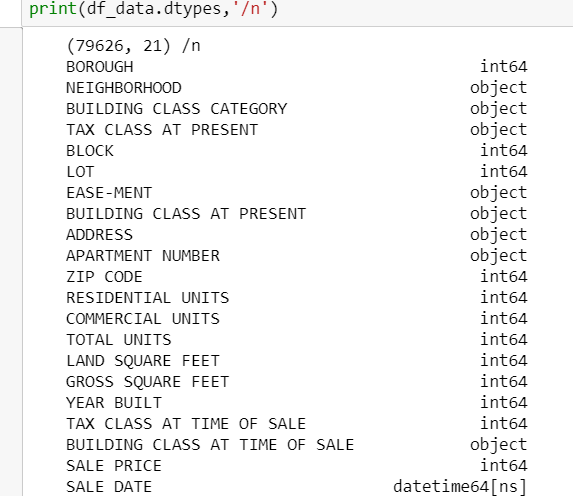


As we can see there are some entries where sales price is equal to 10 / 0 getting rid of invalid entries.

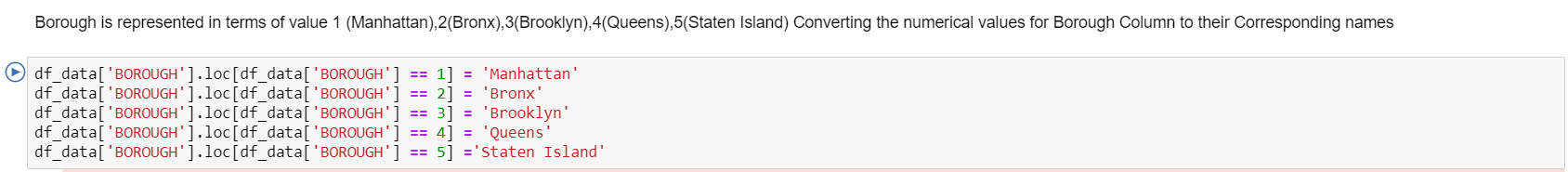
1. Checking if Total units ==0?



Rows where Total units = sum (commercial units, Residential units) are taken into account.



1. Changing the Numerical representation of Boroughs to their actual Names.



**Methodology**

Introduction: Foursquare API

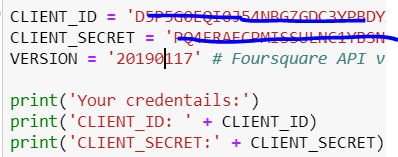
FourSquare API is used to for exploring / obtaining venue information, Foursquare user information, explore geographical information and to get trending venues around a location.

Foursquare API can be used by Constructing an URL with credentials obtained by signing up into Foursquare and sending a request to the API for search of a specific venue, to explore the geographical locations around a venue etc.

Note: The snap shots are taken to show as results, are from Borough Manhattan.

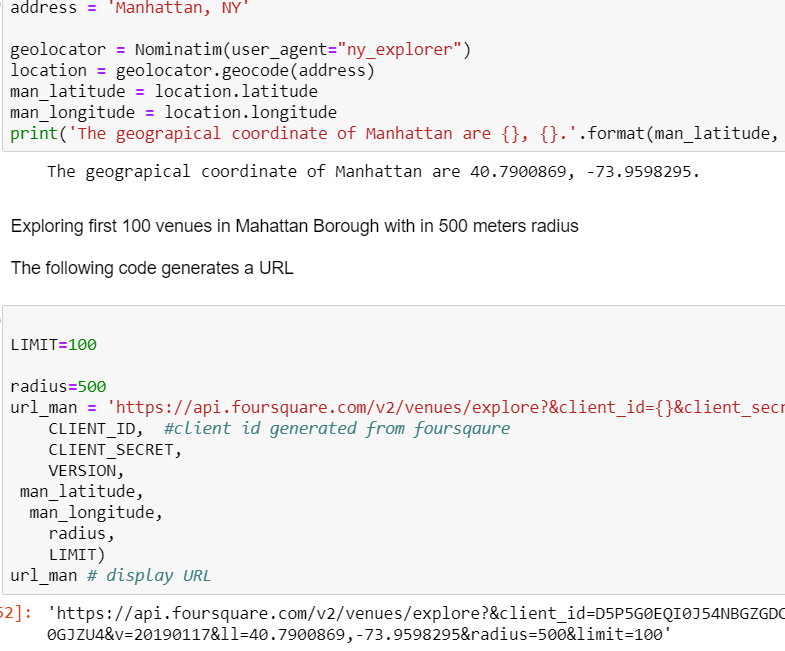
The above process is carried out for Each Borough and their neighborhoods (Brooklyn, Bronx, Staten Island, Queens, Manhattan).

1. Credentials for generating URL request to



1. Exploring the Manhattan Borough

A URL request is the outcome here



1. Processing the URL obtained



The result is a Json file.

1. Function to retrieve venues across each Neighborhood in Manhattan and using this function and making calls to Foursquare API lopping through each Neighborhood in NYC Master Data set.



The result of this Function is collection of all venues corresponding to each Neighborhood resulting into a data frame containing Latitude, longitude, Venue, Venue category, Neighborhood information.

Now we use this Function to write the code to run the above function on each neighborhood and create a new data frame called

manhattan\_venues (For Borough Manhattan)

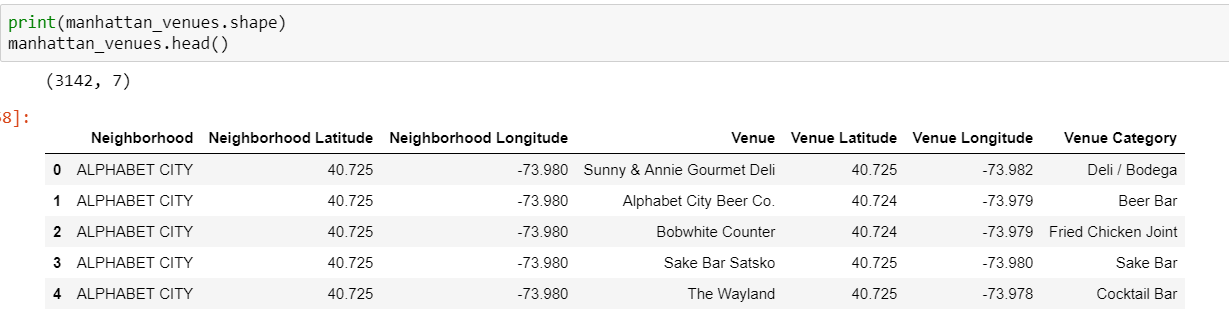
Brooklyn\_venues(For Borough Brooklyn )

Queens\_venues(For Borough Queens)

SI\_\_venues(For Borough Staten Island)

BR\_data.shape(For Borough Bronx)

For Example: manhattan\_venues



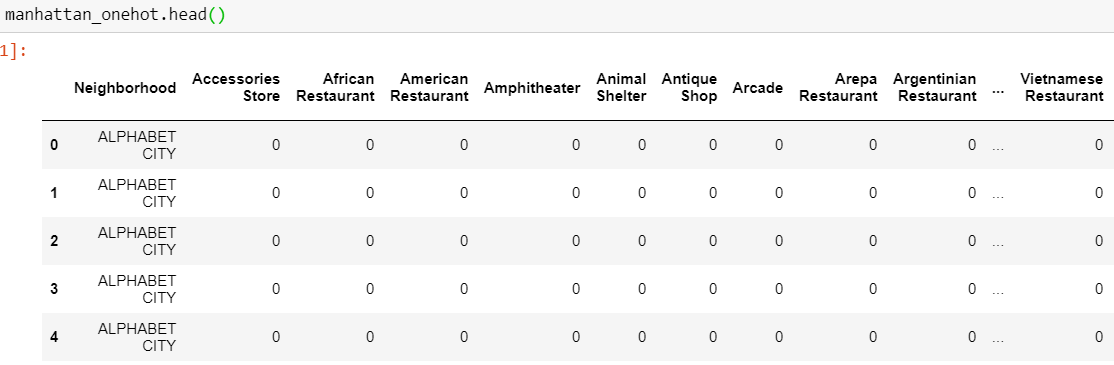
And we can also see number of unique venue categories returned by Manhattan Venues.



1. One -hot coding -Analyzing each Neighborhood

One-Hot encoding helps analyze frequency of each category (Venues) in a Neighborhood.

For Example: Manhattan



Group by rows by each neighborhood and by taking the mean of the frequency of occurrence of each category.



1. Displaying top 10 most common venue categories for Each Neighborhood.

For Example: in Borough Manhattan



Note: The snap shots are taken to show as results are from Borough Manhattan.

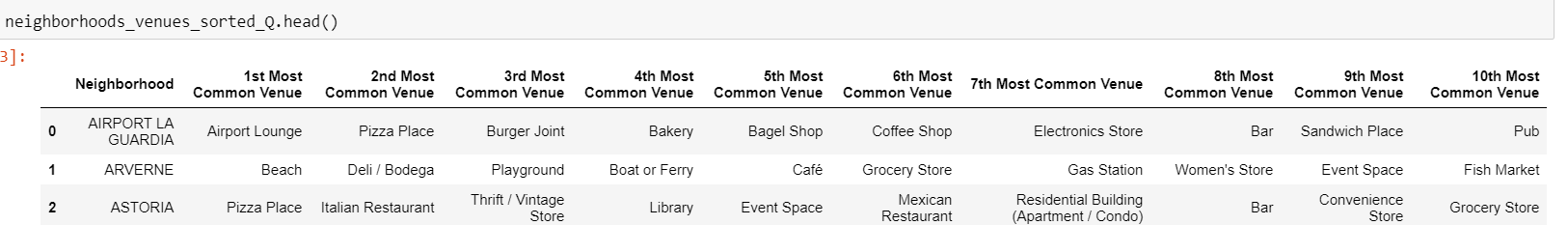
The above process is carried out for Each Borough and their neighborhoods (Brooklyn, Bronx, Staten Island, Queens, Manhattan).

The result sets for Boroughs Brooklyn, Bronx, Staten Island, Queens:

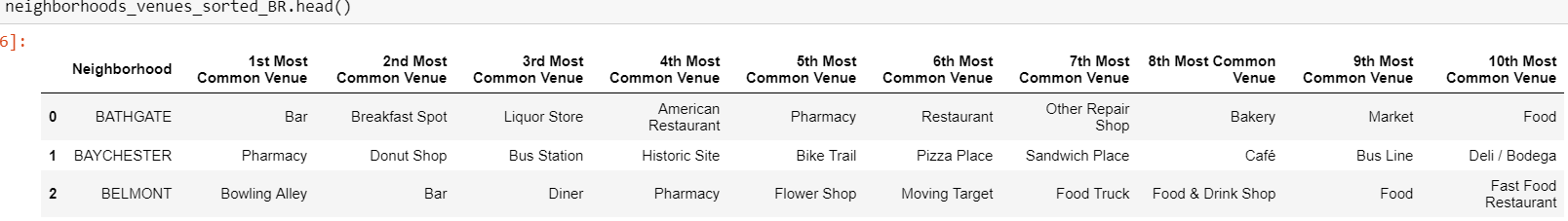
Brooklyn: Top 10 most common Venues



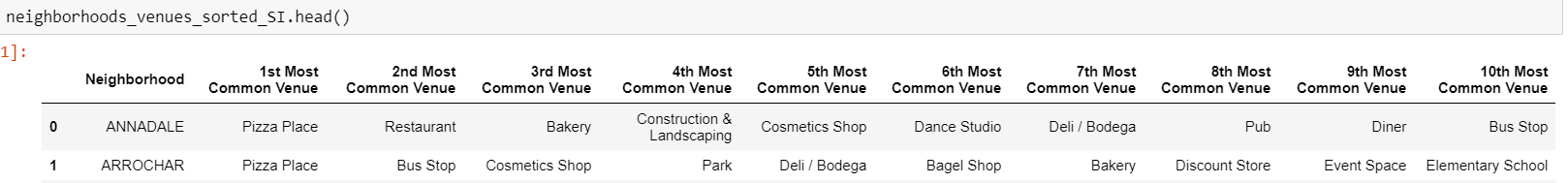
Queens: Top 10 most common Venues



Bronx: Top 10 most common Venues



Staten Island: Top 10 most common Venues



**K-means Clustering and Elbow Method**

K-means is an unsupervised learning methods of clustering unlabeled data into k clusters.

K-means is used in this project to cluster Neighborhoods of NYC and their Boroughs.



* Feature Selection

I have used Longitude and latitude of 5 Borough to calculate clusters

The dataset of neighborhood venues for all 5 boroughs in New York City is consolidated to one dataset. For each venue category, the mean of frequency of venues across each neighborhood was calculated. This information would then be used to fit a K-Means clustering algorithm to the data to determine neighborhoods of similar venue profile.

First, the total number of venues for each category was determined:

For Example:

The result of One -hot Encoding was taken and Group by was applied to rows by each neighborhood and by taking the mean of the frequency of occurrence of each category.



Note: I m just showing the results of Borough Manhattan

* I have applied One – Hot encoding to all 5 Boroughs.

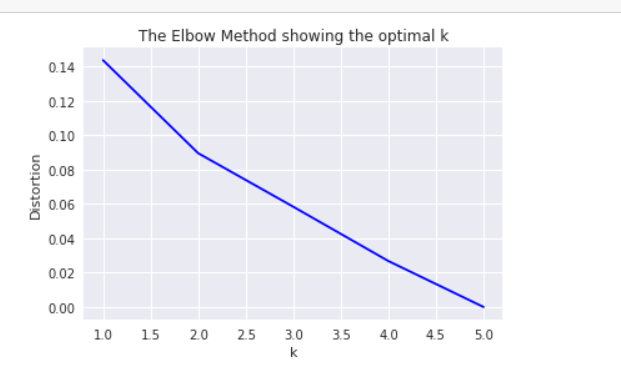
Elbow Method

Determining optimal k

The technique to determine K, the number of clusters, is called the elbow method.

Values for k on horizontal axis and Distortion (% of variance) Vertical axis

1.When K increases, the centroids are closer to the cluster’s centroids. The improvements will decline, creating the elbow shape.



From the figure we can say that optimal Value of K=2

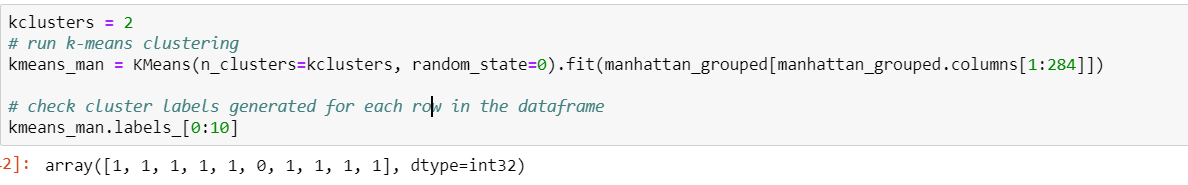
* Creating cluster labels for all 5 Boroughs

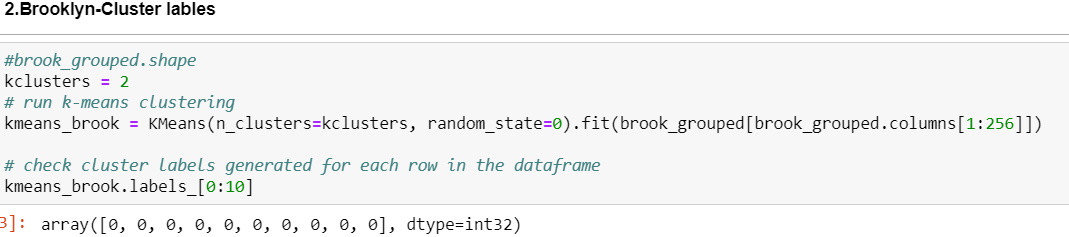
Since I have used Foursquare API to Analyze each Borough and its neighborhoods,

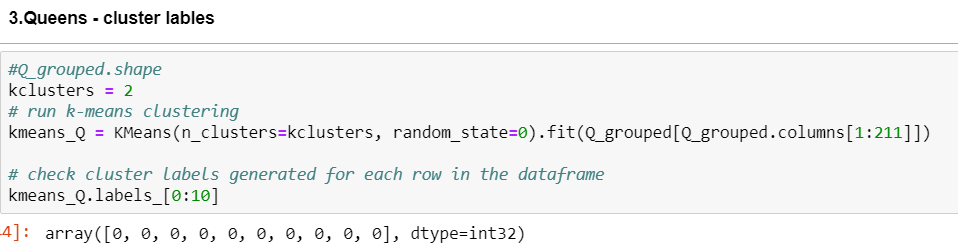
Cluster labels for all 5 Boroughs have been created and the added to their corresponding Datasets.

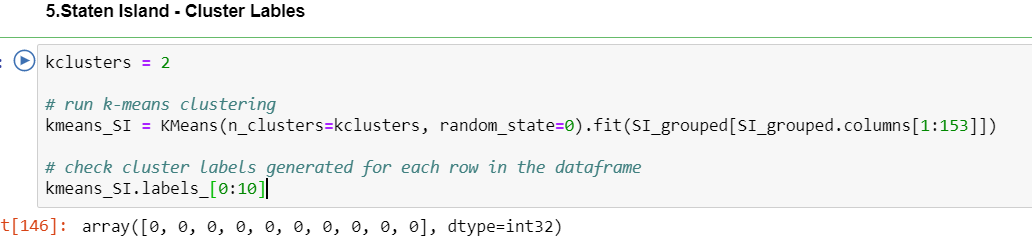
The snap shots should Explain the process more clearly:

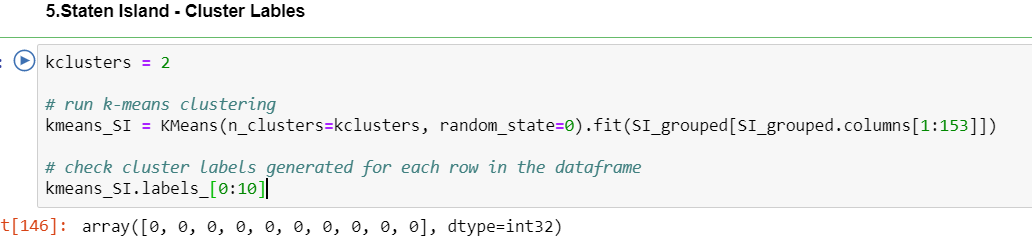
1. Manhattan Clusters









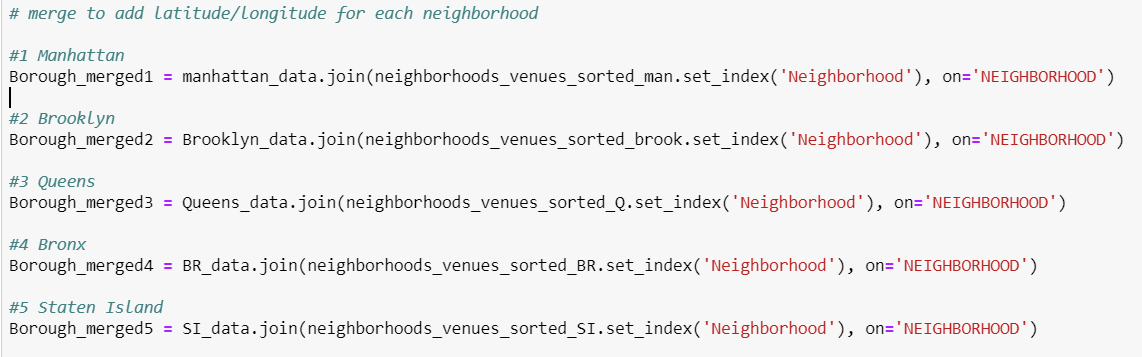


Adding the labels created above to Corresponding Boroughs:



Let's create a new data frame that includes the cluster as well as the top 10 venues for each neighborhood for Each Borough.

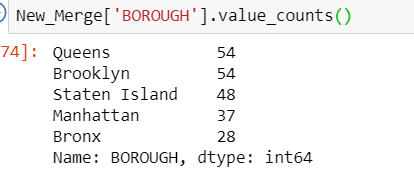
The Data Frames contain top 10 common venues for all 5 boroughs and Neighborhoods with Coordinates information and cluster labels.



Combining the data showed in the above figure to New\_Merge data set with clusters labels and top 10 most common venues for all 5 Boroughs and its Neighborhoods.

The snapshots of the data set New Merge:





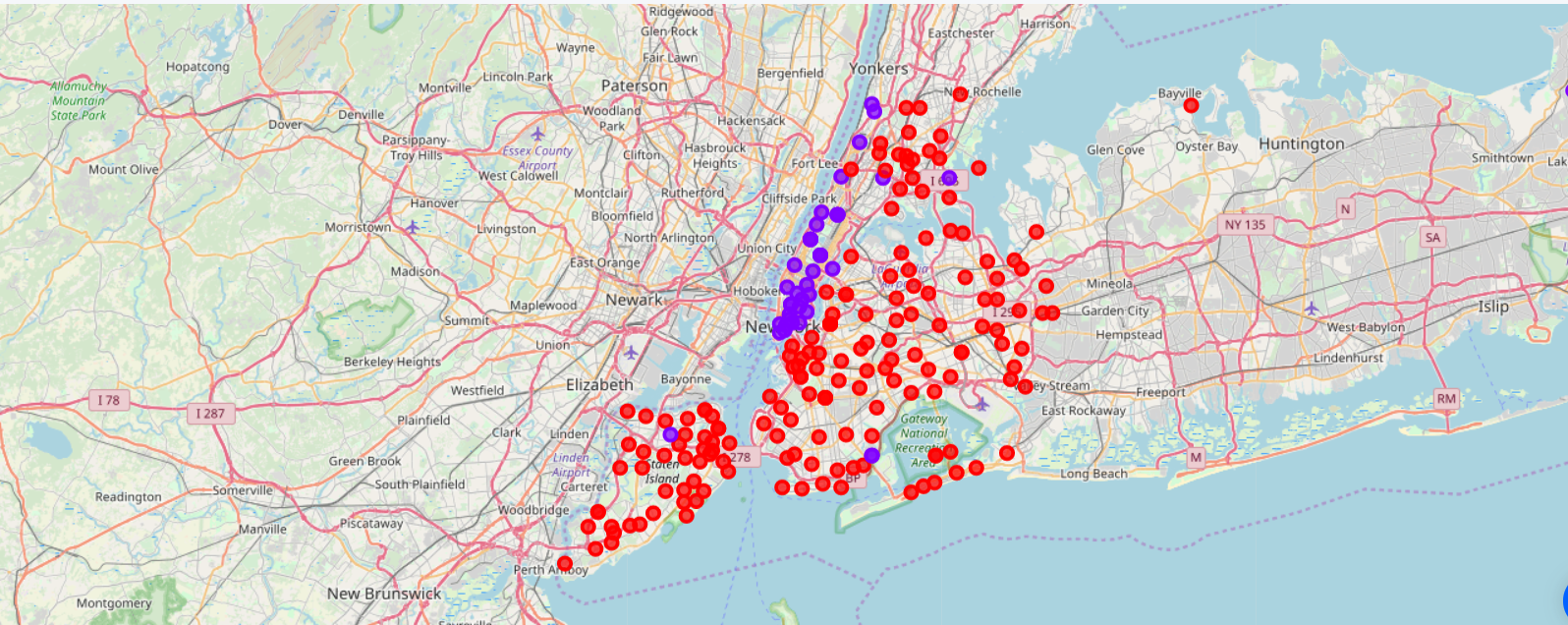
**Clusters:**





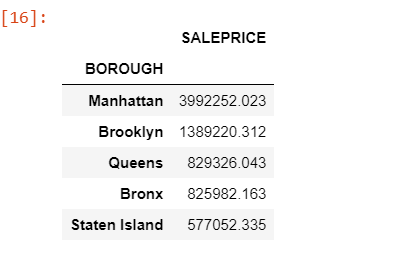
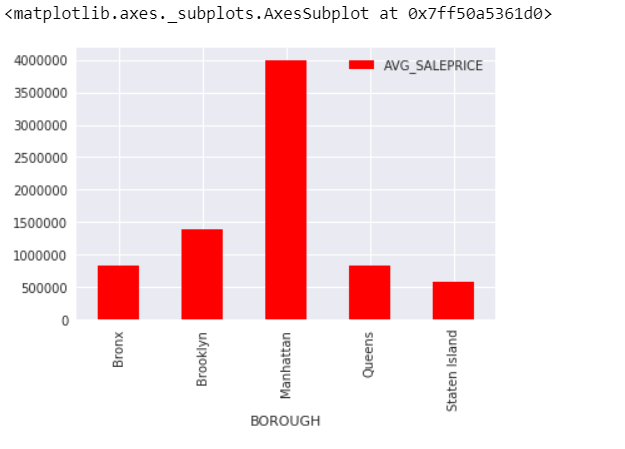
**Data Visualization**

Map with Cluster Label’s:



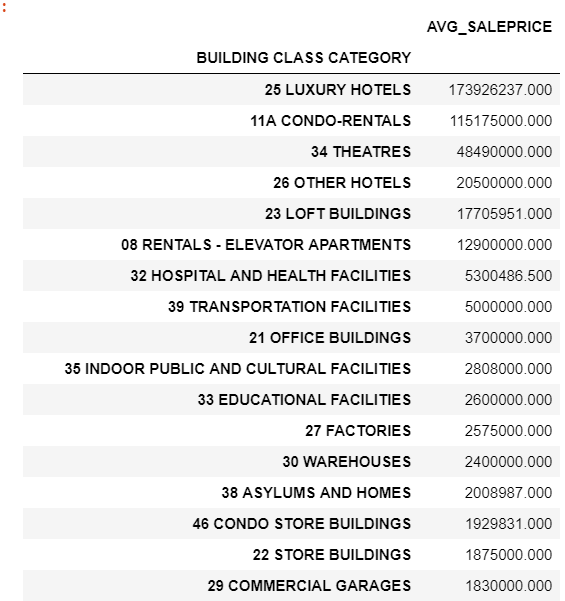
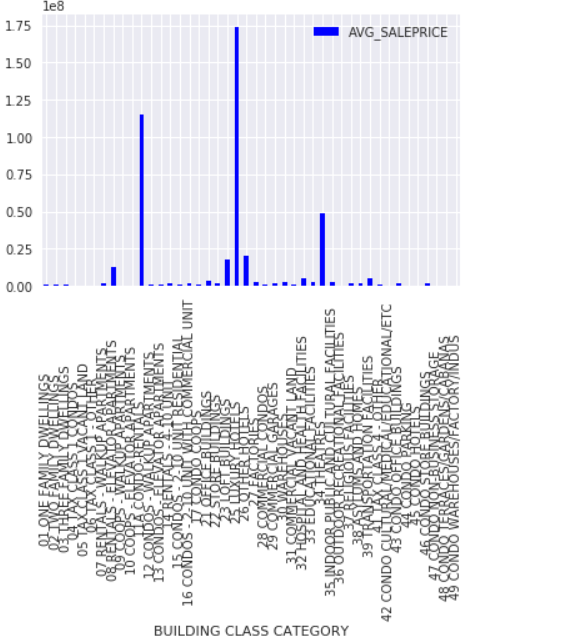
Analyzing Sale Price and other Features

* **Borough with Highest Average Sales Price**

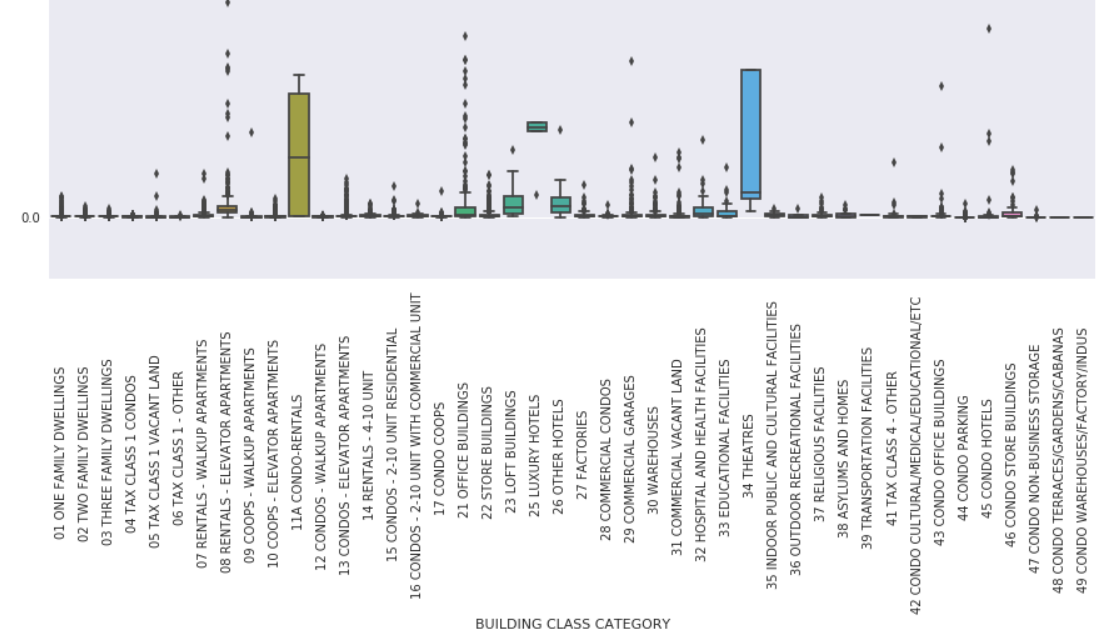
 

The figure clearly shows that Borough Manhattan has the highest Averaged sale Priced Properties.

* **Building Class Category**

* **Sale Price Distribution over Building class category**



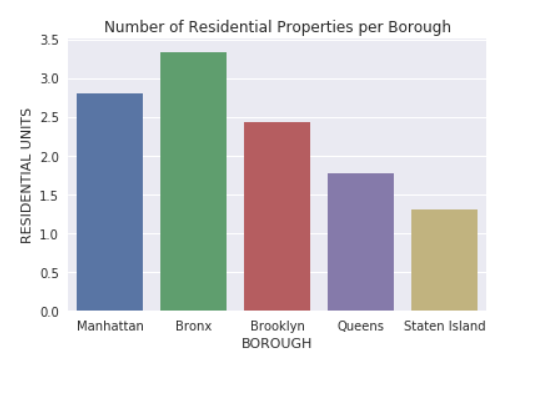
From the figure

Observation: From the above plot we can state that

25 LUXURY HOTELS, 11A CONDO-RENTALS, 34 THEATRES

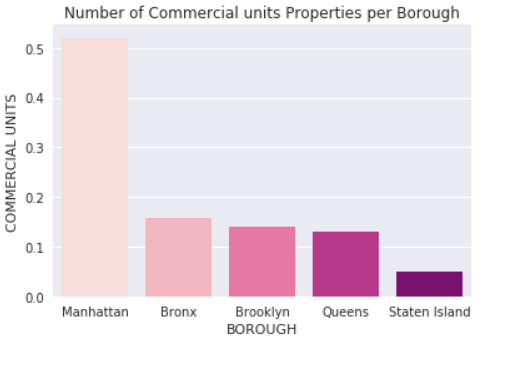
are highly Priced Building Class Categories

* **Residential Properties Per Borough**



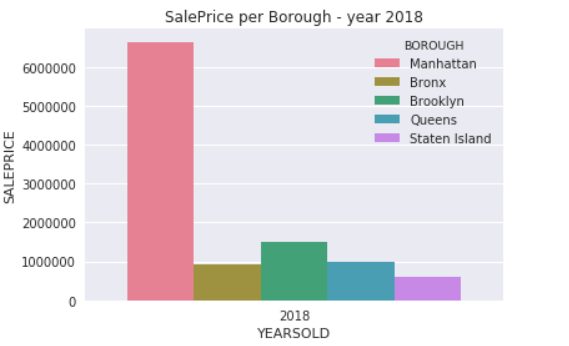
This plot shows that Borough Bronx has more Residential properties than the other Boroughs.

* **Commercial Properties per Borough**



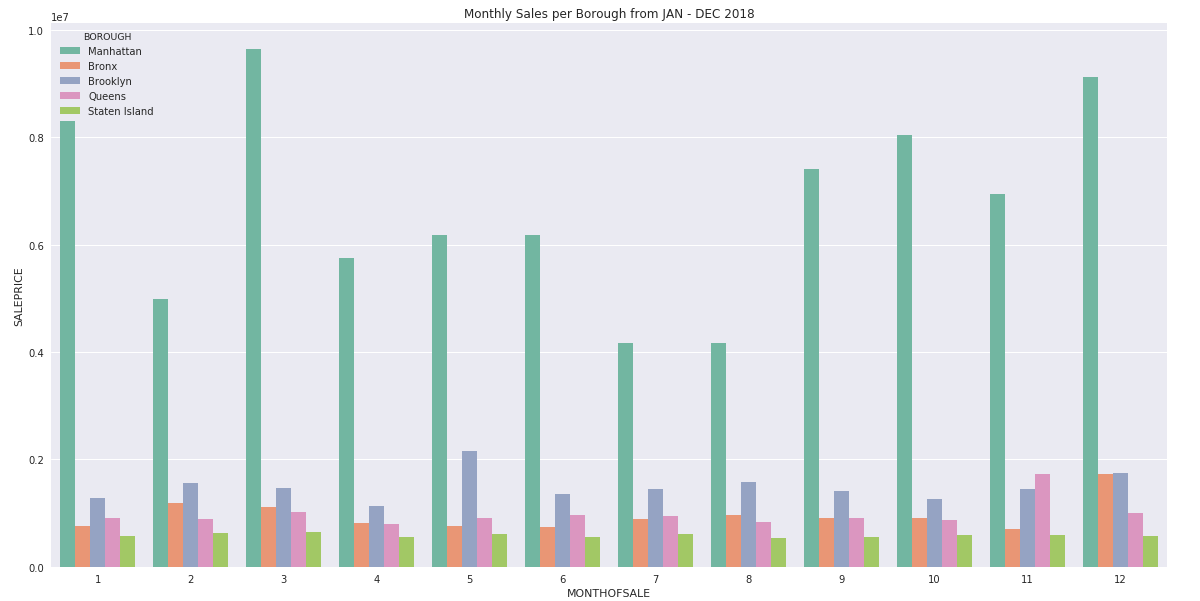
This plot shows That Manhattan houses the highest number of commercials properties.

* **Borough with Highest SalePrice for sale year -2o18**



Borough Manhattan = Highest Sale Priced Properties in Year -2018

* **Monthly Sales Per Borough from Jan -Dec 2018**



From the plots above we can say that Borough Manhattan has highest priced properties.

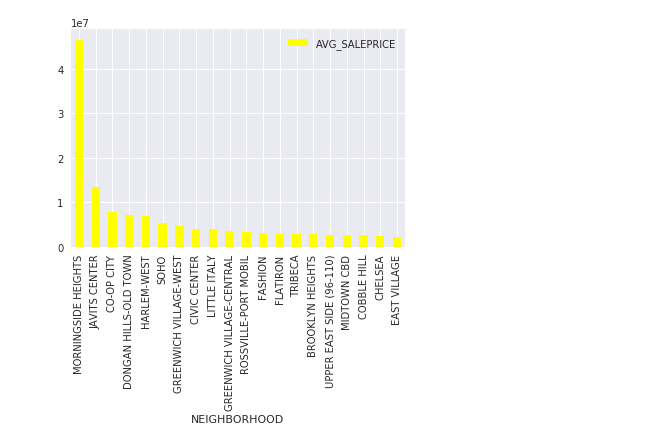
In year 2018 Manhattan has sold highest valued Properties

Borough Manhattan experienced high Property sale Prices in Jan and Dec months of year -2018

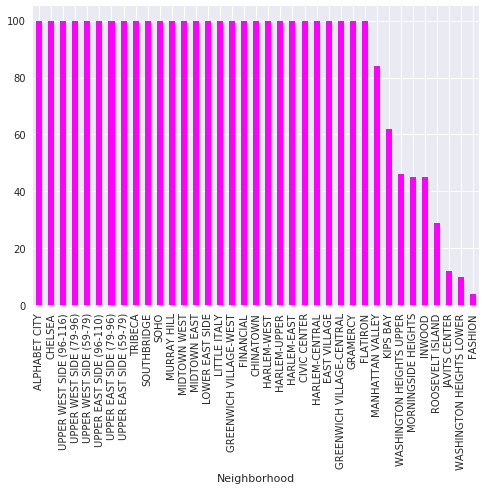
* **Average Sale Price of top 20 Neighborhoods**

As we can see from the figure below Neighborhoods Morning side heights, Javits center

Co-op city are top three Neighborhoods with highest property sale price, and cobble hill, East Village, Chelsea being the bottom three neighborhoods with least priced Properties.

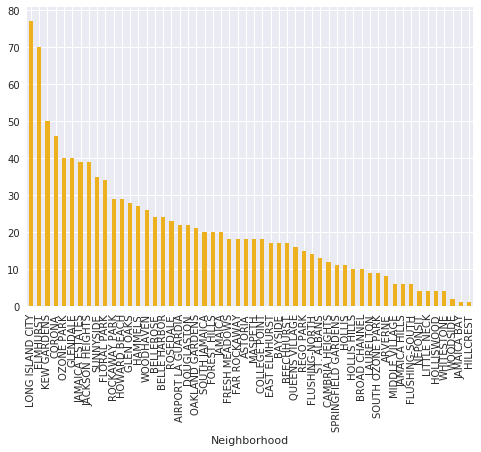


* **Number of venues per neighborhood in Manhattan Borough?**



As the figure shows majority of the Neighborhoods has many venues.

* **Number of venues per neighborhood in Queens Borough?**



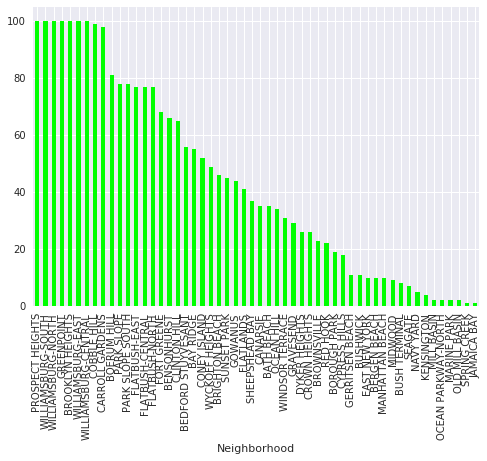
Neighborhoods Long Island City, Elmhurst, Kew Gardens have a greater number of venues then rest of the Neighborhoods in Queens Borough.

* **Number of venues per neighborhood in Brooklyn Borough?**

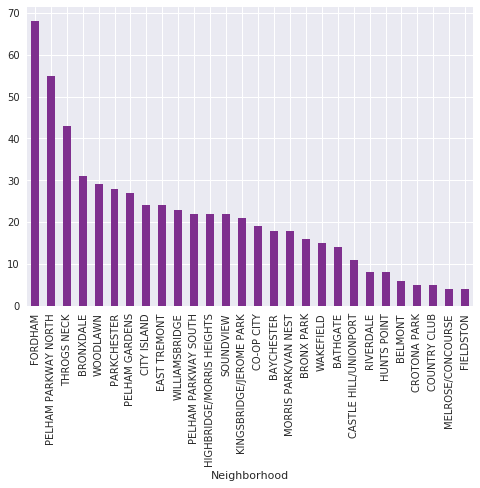
As we can see from the figure Prospect Heights, Williamsburg-south, Williamsburg-North,

Green points and Brooklyn Heights have larger group of venues in Brooklyn Borough.

Whereas Jamaica Bay, Spring creek and old mill basin have least number of venues.

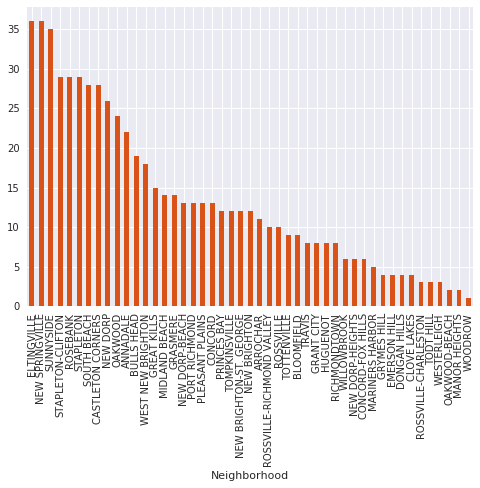


* **Number of venues per neighborhood in Bronx Borough?**



Fordham, Pelham parkway station and Throgs Neck Neighborhood have a greater number of venues than the rest of the neighborhoods in Bronx.

* **Number of venues per neighborhood in Staten Island Borough?**



From the figure above Eltingville , New Springville , Sunside hold maximum number of venues in Staten Island Borough.

* **Top 2 Neighborhoods with Max sale price in Each borough.**

Manhattan: CHELSEA, UPPER WEST SIDE (59-79)

Brooklyn: SPRING CREEK, RED HOOK

Queens: LONG ISLAND CITY, REGO PARK

Bronx: WESTCHESTER, PELHAM GARDENS

Staten Island: ROSEBANK, ROSSVILLE-CHARLESTON

