# BATTLE OF NEIGHBORHOODS

**NEW YORK CITY - PROPERTY SALES ANALYSIS TO FIND THE BEST HOME FOR YOU**

***KARTIK NAIR***

***JULY 2020***

1. INTRODUCTION
   1. BACKGROUND

For people trying to find the best places to live, it's always a good idea to compare cities and if possible, to compare neighborhoods to see if its suites your taste and fits your budget. The cost of living in the neighbour hood and the amenities with the property and nearby it is a top concern when moving to a new area. For some its a restaurant nearby for some a coffee shop , in all everyone has their requirements which are needeed to be looked out for.

1.2 PROBLEM

The Property Sales dataset of New York has sales details with price of different types of houses sold in each borough of New York over the years 2010-2019. The market price of each place changes with time. This project aims to select the boroughs in NYC based on the highest number of sales, explore the neighborhoods of that borough to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

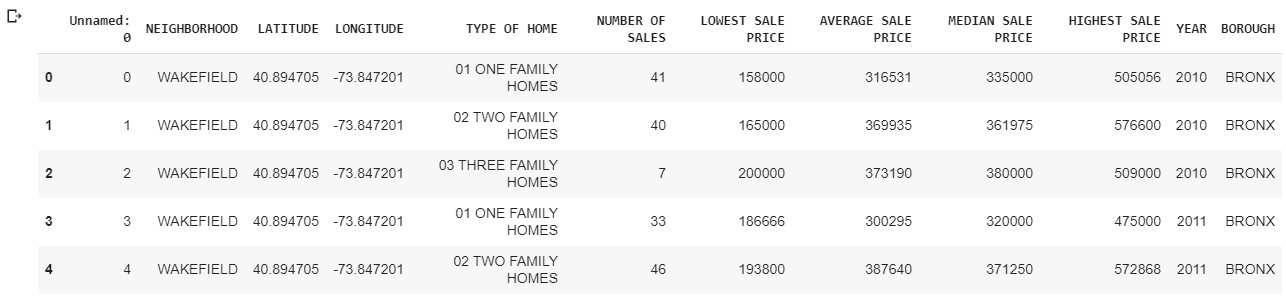
**2. Data**

There were two main datasets used in this project, one being the New York City Property Sales dataset 2010-2019 and the other dataset contained all the neighbourhoods of NYC with their geographical coordinates. But for Comfort The Geographical coordinates was merged with the Property sales dataset

**2.1 Dataset**

2.1.1 NYC Property Sales Dataset

Let’s have a quick look at our dataset:-



* Some Important Columns :

1.NEIGHBORHOOD:- This column has the neighborhood name of each sale

2. NUMBER OF SALES :- This numeric column indicates the total sales done in each neighborhood of a particular type of house.

3. BOROUGH :- This column is very important for our project this column indicates the borough of the sale

And 11 more columns for now , some will be cleaned later and these columns are not relevant to our analysis either.

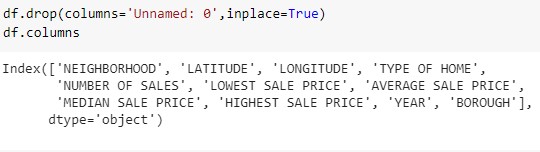
**2.2 Data Cleaning**

The New York Sales Dataset needs a few minor cleaning,

1. Duplicate rows needed to be removed
2. Uneccessary columns to be dropped
3. And Column Names to be renamed for further processes
4. Total number of sales per Borough needed

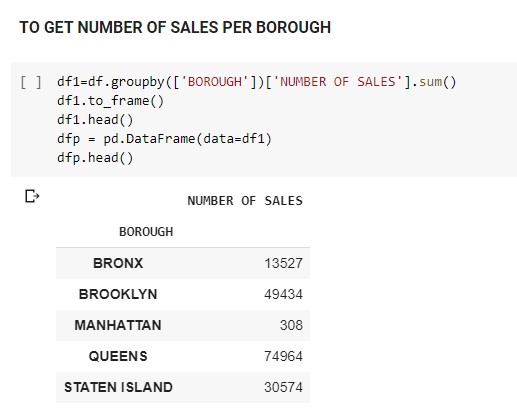
These all processes were done using the Pandas Library as follows :

duplicate



rename

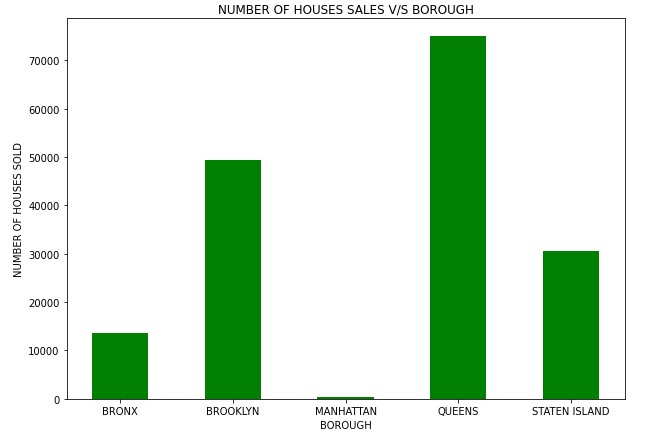
3.

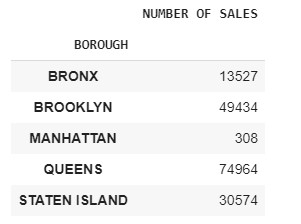


4.

1. **METHODOLOGY**
   1. **Exploratory Data Analysis**

Let’s Visualize The Sales in each borough





Comparing five boroughs it is evident that Queens has the highest Property Sales recorded followed by Brooklyn, Staten Island, Bronx and Manhattan

It’s clearly seen that people prefer QUEENS than any other borough so that’s our first preference so we’ll explore this.

**3.1.2 Neighborhoods in QUEENS**

Exploring the Queens Borough showed that only a single borough won’t have enough neighbourhoods for a person to choose from . Hence under this case we will take 3 Borough with most property sales which are

* QUEENS
* BROOKLYN
* STATEN ISLAND

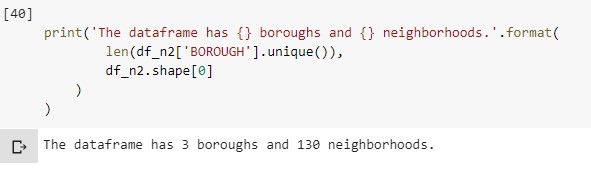
**3.1.3 Visualize the neighbourhoods**

Firstly, we select the neighbourhoods only from these 3 Boroughs,

that is Bronx, Queens and Staten Island, and save them in a new

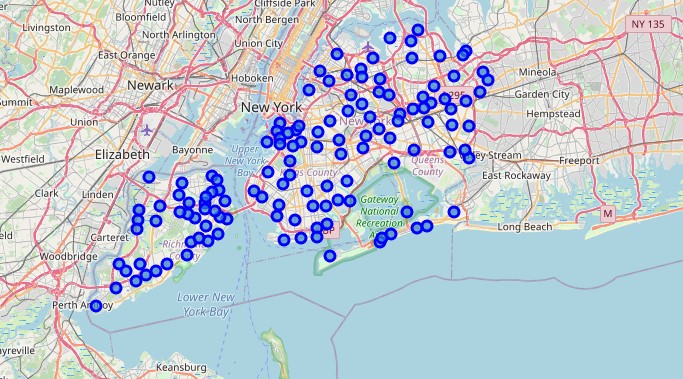
dataframe.





We find that there are 130 neighbourhoods recorded in the 3 Boroughs

To visualize this we use the *Folium* Library



* 1. **Modelling**

Using the final dataset containing the neighbourhoods with the

latitude and longitude, we can find all the venues within a 500 meter

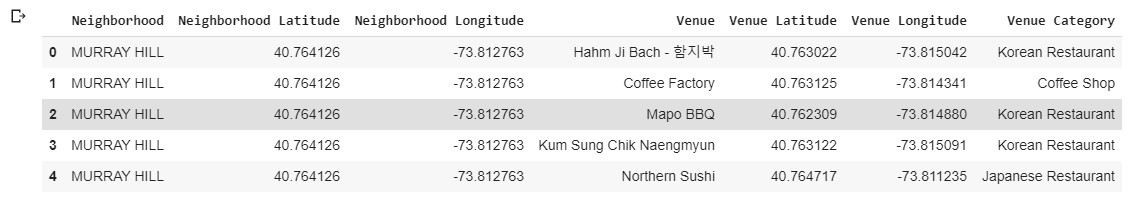
radius of each neighbourhood by connecting to the Foursquare API.

This returns a json file containing all the venues in each

neighbourhood which is converted to a pandas dataframe. This data

frame contains all the venues along with their coordinates and

category.



One hot encoding is done on the venues data. (One hot encoding is a

process by which categorical variables are converted into a form that

could be provided to ML algorithms to do a better job in prediction).

The Venues data is then grouped by the Neighbourhood and the

mean of the venues are calculated, finally the 10 common venues

are calculated for each of the neighbourhoods.

To help people find similar neighbourhoods in the safest borough we

will be clustering similar neighbourhoods using K - means clustering

which is a form of unsupervised machine learning algorithm that

clusters data based on predefined cluster size. We used the elbow

method to find the best cluster size and found 5 clusters to be ideal.

The reason to conduct a K- means clustering is to cluster

neighbourhoods with similar venues together so that people can

shortlist the area of their interests based on the venues/amenities

around each neighbourhood.

1. **Results**

After running the K-means clustering we can access each cluster

created to see which neighborhoods were assigned to each of the

five clusters. Looking into the neighborhoods in the first cluster



Upon closely examining these neighborhoods we can see that the

most common venues in these neighborhoods are Bus Stop, Coffee

shops and restaurants.

Similarly looking at the 2nd cluster we see that it mainly consists of

the neighbourhoods with venues such as be Bar, Pizza Place, Pharmacy.

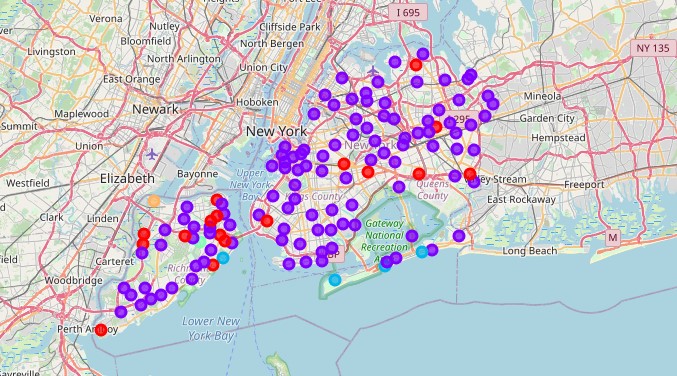


Similarly, we can examine each cluster to find out which

neighbourhoods suits our best interest by looking at the most

common venues.

Finally lets visualize the clustered neighbourhoods using Folium Library.



**5. Discussion**

The aim of this project is to help people who want to relocate to the

Best borough in New York city, expats can choose the

neighbourhoods to which they want to relocate based on the most

common venues in it. For example, if a person is looking for a

neighbourhood with good connectivity and public transportation we

can see that Cluster 1 has and Bus stops as the most common

venues. If a person is looking for a neighbourhood with stores and

restaurants in a close proximity, then the neighbourhoods in the

second cluster is suitable. The choices of neighbourhoods may vary

from person to person.

**6.Conclusion**

This project helps a person get a better understanding of the

neighbourhoods with respect to the most common venues in that

neighbourhood. It is always helpful to find out more about places

before moving into a neighbourhood.We have just taken Price(Budget) as a

primary concern to shortlist the best boroughs in New York city.

The future of this project includes taking other factors such as saefty in the areas into consideration to shortlist the borough, such as

filtering areas based on a Number of Crimes Recorded in the Borough.