Capstone Project - The Battle of Neighbourhoods' to get optimal Real-Estate properties

1. Introduction

Business Problem section

Background

New York City's housing market has largely recovered from the financial crisis of 2008, but that doesn't necessarily mean that buying a home here is, in the long run, a good investment. That's the conclusion from a new report by Street Easy, which looks at how home values in the city have changed in the 10 years since the Great Recession. Additionally, home values have overall gone up since the post-crisis low of November 2011 [1] . Street Easy found that those have risen by a whopping 30 percent in the past seven years, at an average of nearly four percent per year.

Business Problem

The problem scenario is to suggest the homebuyers clientele to purchase a suitable real estate in New York using Machine Learning Algorithms.

As a result, the business problem we are currently posing is:

How could we provide suggestions to homebuyers clients to purchase a suitable real estate in New York street in this depreciating economy?

To solve this business problem, we are going to cluster New York neighbourhoods in order to recommend venues and the current average price of real estate where homebuyers can make a real estate investment. Also we will recommend profitable venues venues i.e. pharmacy, restaurants, hospitals & grocery stores.

Full Implementation of Project:

https://github.com/chandrashek1007/Capstone-Project---The-Battle-of-Neighborhoods-to-get-optimal-Real-Estate-properties/blob/master/Capstone_Project%20.ipynb

2. Data Description

The Department of Finance (DOF) maintains records for all property sales in New York City, including sales of family homes in each borough(https://data.cityofnewyork.us/api/views/948r-3ads/rows.csv?accessType=DOWNLOAD) [2].

This list includes all sales of 1-, 2-, and 3-Family Homes' from January 1st, 2009 to December 31, 2009, whose sale price is equal to or more than $150,000. The Building Class Category for Sales is based on the Building Class at the time of the sale.

To explore and target recommended locations across different venues according to the presence of amenities and essential facilities, we will access data through FourSquare API [2] interface and arrange them as a data frame for visualization. By merging data on New York properties and the relative price paid data from the HM Land Registry and data on amenities and essential facilities surrounding such properties from FourSquare API interface, we will be able to recommend profitable real estate investments.

Methodology

To consider the problem we can list the data as below

1. Collecting Data

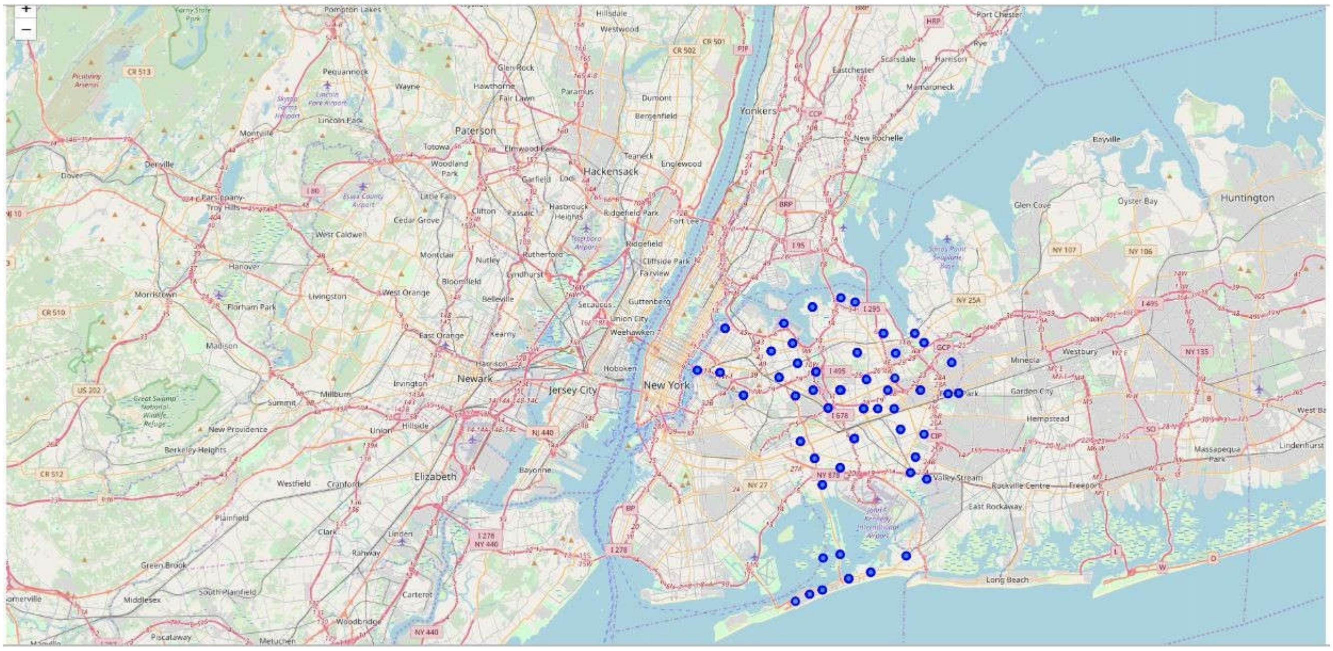
We take the data set to a pandas data frame and analyse that data. the data set columns are

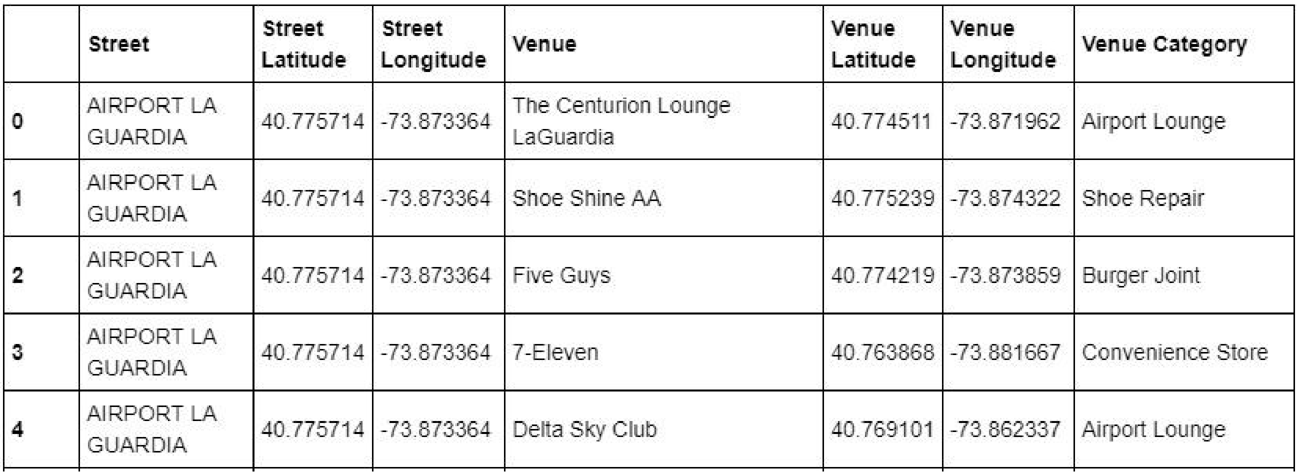
- NEIGHBOURHOOD : contains the Neighbourhood of New York , USA.
- TYPE OF HOME : It has three categories 01 ONE FAMILY HOMES,02 ONE FAMILY HOMES and 03 ONE FAMILY HOMES.
- TOTAL NO. OF PROPERTIES : No. of Properties Available.
- NUMBER OF SALES : No. of Properties already sold out.
- LOWEST SALE PRICE : Minimum Price of property in that neighbourhood.
- AVERAGE SALE PRICE : Average Price of property in that neighbourhood.
- MEDIAN SALE PRICE: Median Price of property in that neighbourhood.
- HIGHEST SALE PRICE : Maximum Price of property in that neighbourhood.

2. Exploring Data

We get the latitude and longitude of the Neighbourhood using geopy.geocoders package of python and we append the longitude and latitude to the pandas data frame. Plotting the Neighbourhoods of New York present in Dataset using Folium package. We added the marker in the World Map for better visualisation.

I have utilised the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venue and the radius 2500 meter for each borough from their given latitude and longitude informations. Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API.

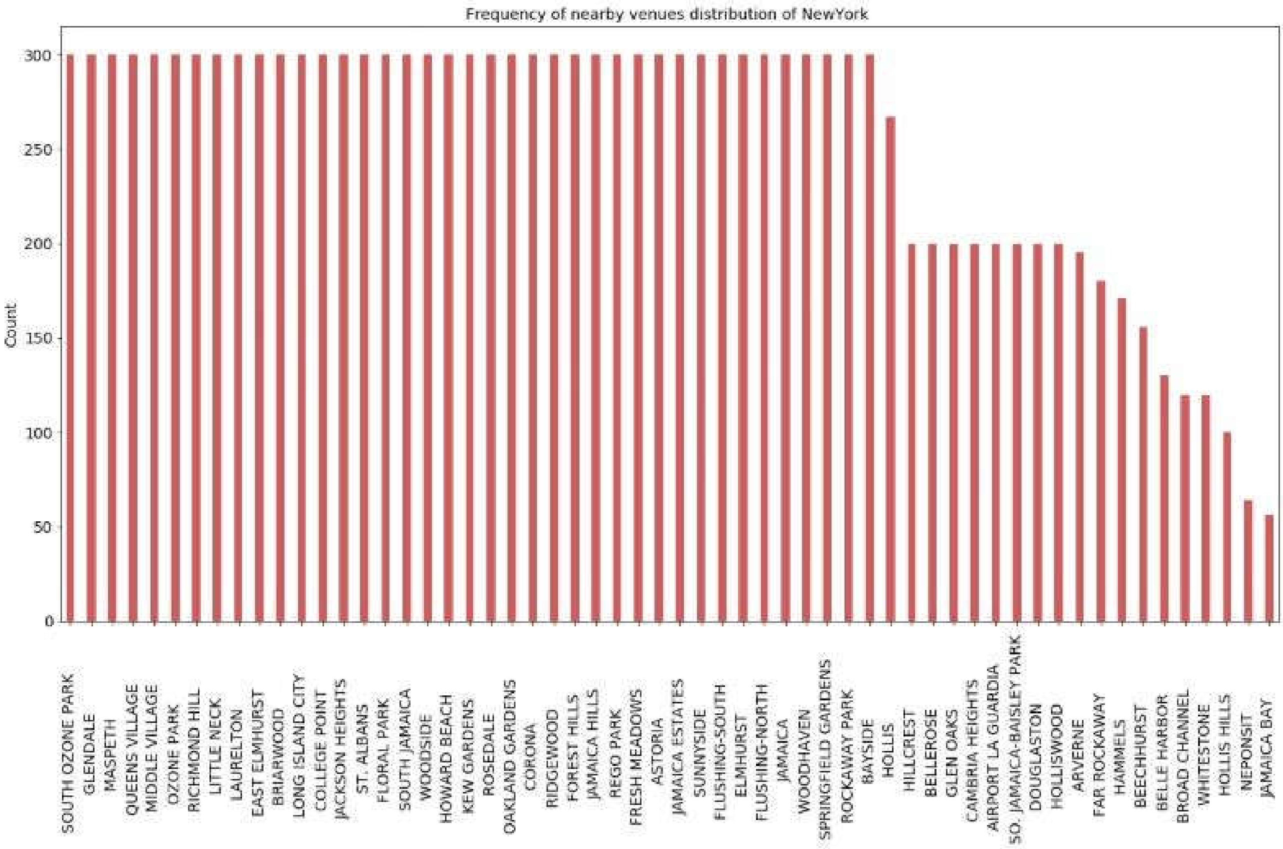




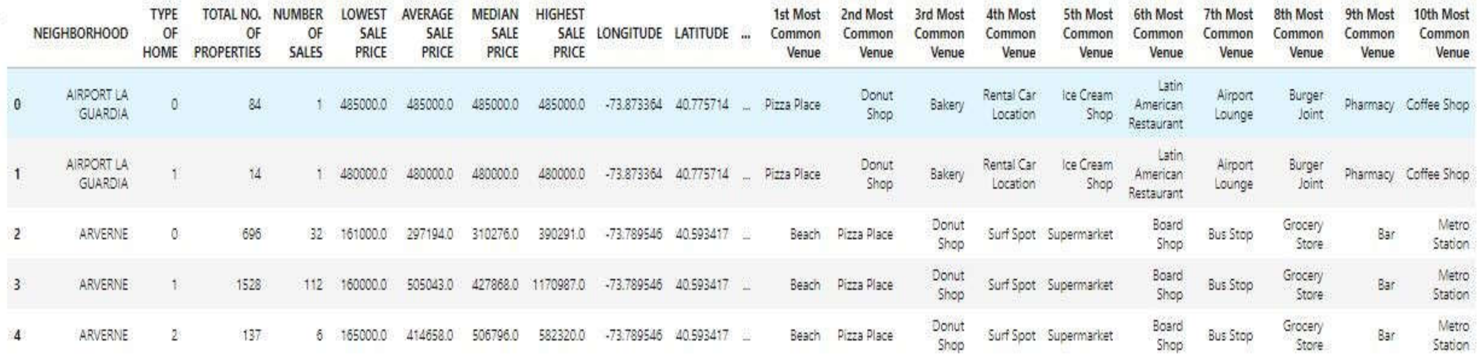
	Street	Street Latitude	Street Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	AIRPORT LA GUARDIA	40.775714	-73.873364	The Centurion Lounge LaGuardia	40.774511	-73.871962	Airport Lounge
1	AIRPORT LA GUARDIA	40.775714	-73.873364	Shoe Shine AA	40.775239	-73.874322	Shoe Repair
2	AIRPORT LA GUARDIA	40.775714	-73.873364	Five Guys	40.774219	-73.873859	Burger Joint
3	AIRPORT LA GUARDIA	40.775714	-73.873364	7-Eleven	40.763868	-73.881667	Convenience Store
4	AIRPORT LA GUARDIA	40.775714	-73.873364	Delta Sky Club	40.769101	-73.862337	Airport Lounge

We got **14859** rows returned by Foursquare API. We merged the street with latitude and longitude with the Venues data. There are **293** unique categories.

We plot the total number of venues in Neighbourhoods of the New York.



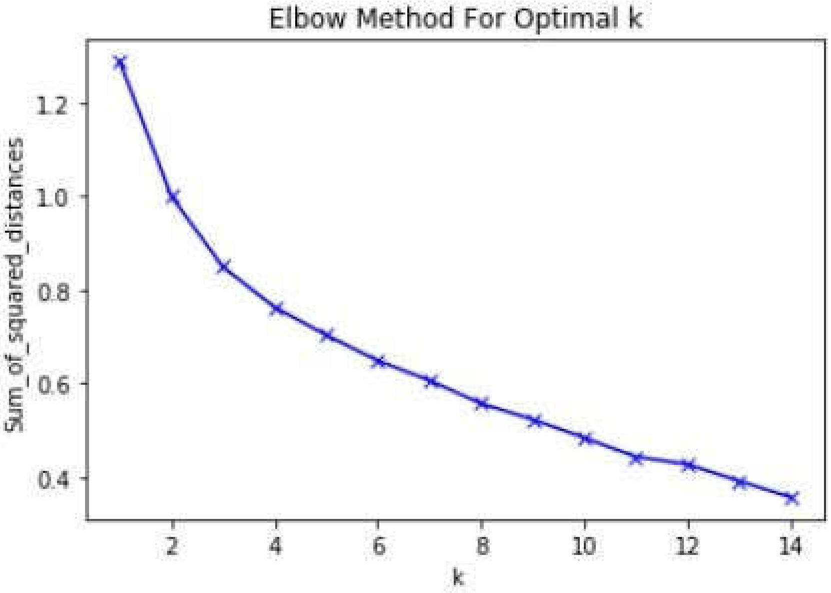
In summary of this graph 293 unique categories were returned by Foursquare, then I created a table which shows list of top 10 venue category for each borough in below table.



	NEIGHBORHOOD	TYPE OF HOME	TOTAL NO. OF PROPERTIES	NUMBER OF SALES	LOWEST SALE PRICE	AVERAGE SALE PRICE	MEDIAN SALE PRICE	HIGHEST SALE PRICE	LONGITUDE	LATITUDE	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	AIRPORT LA GUARDIA	0	84	1	485000.0	485000.0	485000.0	485000.0	-73.873364	40.775714	...	Pizza Place	Donut Shop	Bakery	Rental Car Location	Ice Cream Shop	Latin American Restaurant	Airport Lounge	Burger Joint	Pharmacy	Coffee Shop
1	AIRPORT LA GUARDIA	1	14	1	480000.0	480000.0	480000.0	480000.0	-73.873364	40.775714	...	Pizza Place	Donut Shop	Bakery	Rental Car Location	Ice Cream Shop	Latin American Restaurant	Airport Lounge	Burger Joint	Pharmacy	Coffee Shop
2	ARVERNE	0	696	32	161000.0	297194.0	310276.0	390291.0	-73.789546	40.593417	...	Beach	Pizza Place	Donut Shop	Surf Spot	Supermarket	Board Shop	Bus Stop	Grocery Store	Bar	Metro Station
3	ARVERNE	1	1528	112	160000.0	505043.0	427868.0	1170987.0	-73.789546	40.593417	...	Beach	Pizza Place	Donut Shop	Surf Spot	Supermarket	Board Shop	Bus Stop	Grocery Store	Bar	Metro Station
4	ARVERNE	2	137	6	165000.0	414658.0	506796.0	582320.0	-73.789546	40.593417	...	Beach	Pizza Place	Donut Shop	Surf Spot	Supermarket	Board Shop	Bus Stop	Grocery Store	Bar	Metro Station

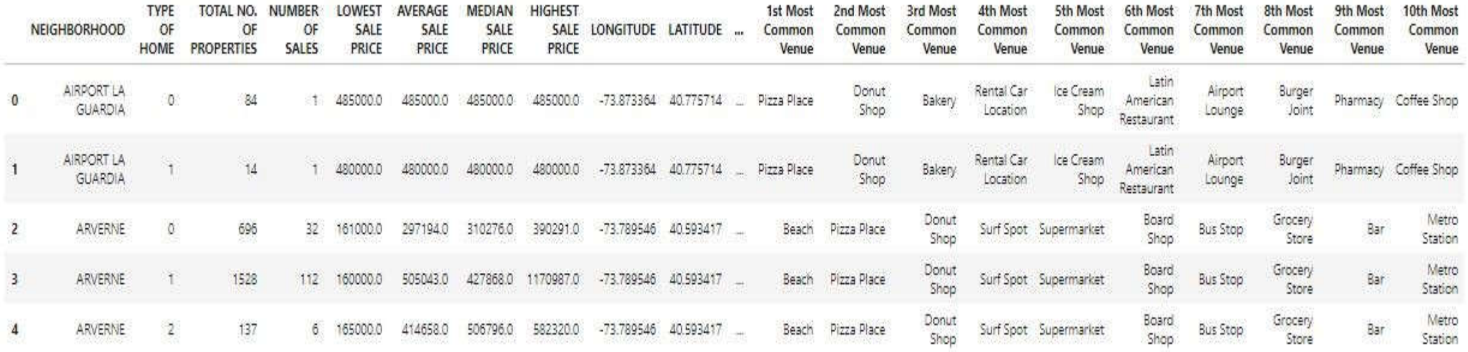
3. Modelling

We have some common venue categories of the Neighbourhoods of New York. We use K-Means Algorithm to cluster the Neighbourhoods of New York. K-Means algorithm is one of the most common cluster method of unsupervised learning.



From above graph the Elbow point is in **k=4** which is the optimal value of K-Means clustering

Here is my merged table with cluster labels for each neighbourhoods of New York.

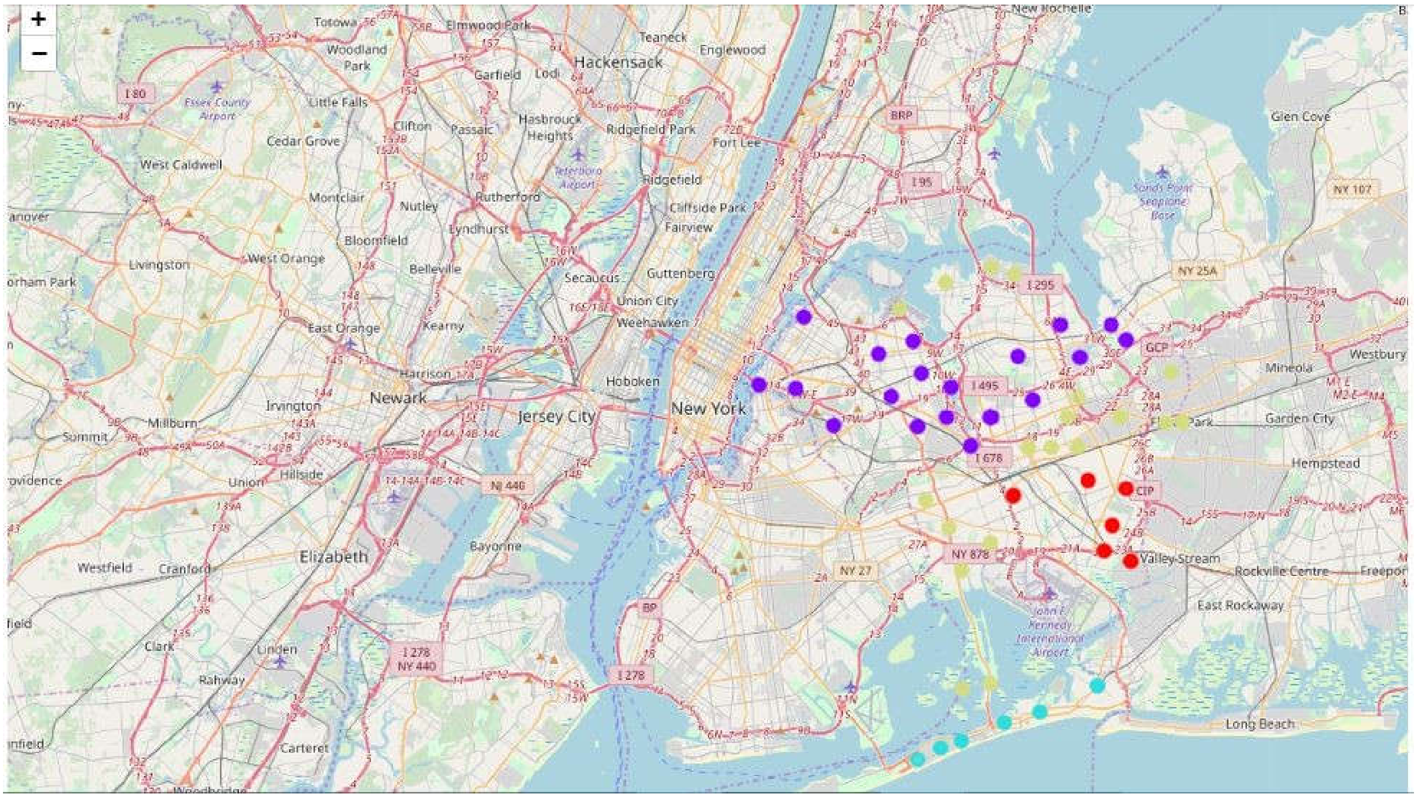


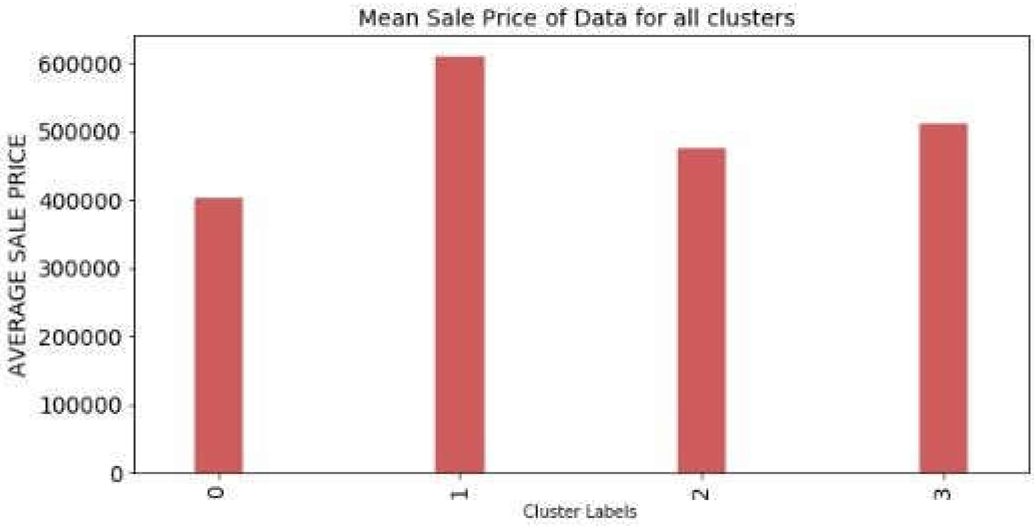
	NEIGHBORHOOD	TYPE OF HOME	TOTAL NO. OF PROPERTIES	NUMBER OF SALES	LOWEST SALE PRICE	AVERAGE SALE PRICE	MEDIAN SALE PRICE	HIGHEST SALE PRICE	LONGITUDE	LATITUDE	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	AIRPORT LA GUARDIA	0	84	1	485000.0	485000.0	485000.0	485000.0	-73.873364	40.775714	...	Pizza Place	Donut Shop	Bakery	Rental Car Location	Ice Cream Shop	Latin American Restaurant	Airport Lounge	Burger Joint	Pharmacy	Coffee Shop
1	AIRPORT LA GUARDIA	1	14	1	480000.0	480000.0	480000.0	480000.0	-73.873364	40.775714	...	Pizza Place	Donut Shop	Bakery	Rental Car Location	Ice Cream Shop	Latin American Restaurant	Airport Lounge	Burger Joint	Pharmacy	Coffee Shop
2	ARVERNE	0	696	32	161000.0	297194.0	310276.0	390291.0	-73.789546	40.593417	...	Beach	Pizza Place	Donut Shop	Surf Spot	Supermarket	Board Shop	Bus Stop	Grocery Store	Bar	Metro Station
3	ARVERNE	1	1528	112	160000.0	505043.0	427868.0	1170987.0	-73.789546	40.593417	...	Beach	Pizza Place	Donut Shop	Surf Spot	Supermarket	Board Shop	Bus Stop	Grocery Store	Bar	Metro Station
4	ARVERNE	2	137	6	165000.0	414658.0	506796.0	582320.0	-73.789546	40.593417	...	Beach	Pizza Place	Donut Shop	Surf Spot	Supermarket	Board Shop	Bus Stop	Grocery Store	Bar	Metro Station

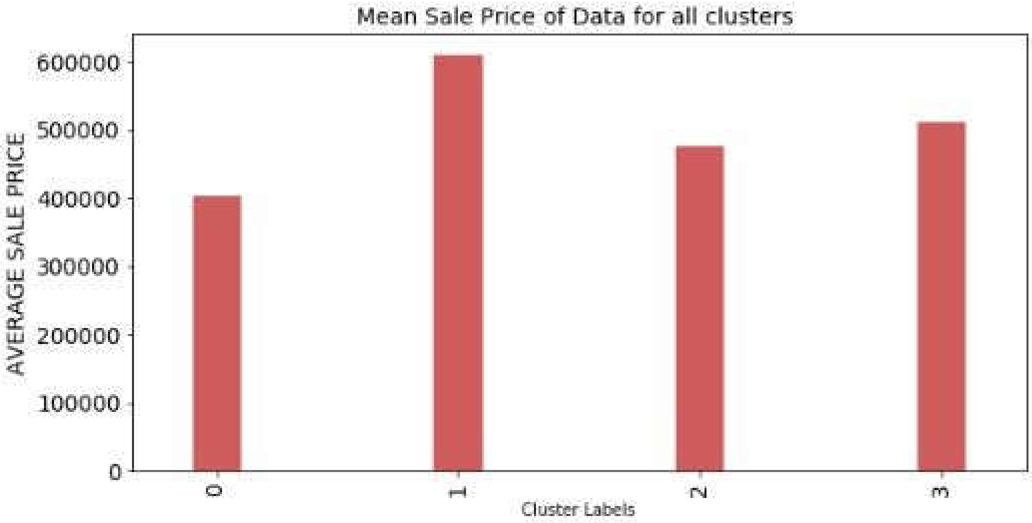
After merging the dataframes we plot each clusters in the World Map Using Folium Library.

So we can see that 4 clusters has been created based on the respective attributes.

Now we will analyse each cluster data and finds its mean and median price.







Result

First of all, even though the London Housing Market may be in a rut, it is still an "evergreen" for business affairs.

Key Observations under the Results:

First, we may examine them according to neighborhoods of New York Areas.

Cluster 0:

1. The average and Median price of Cluster one Neighborhoods are 403649.400000 and 406267.950000 respectively.

2. The cluster contains following places -

CAMBRIA HEIGHTS, JAMAICA, LAURELTON, ROSEDALE , SOUTH,JAMAICA ,SPRINGFIELD GARDENS and ST. ALBANS

3. The most common venues nearby are Food Corner , Restaurants, Bank , Park. The no of Sales is less with respect to available properties.

4. The properties are best to buy as it has very reasonable average and median rates and in addition to that it has elementary stuffs for daily needs .

5. The place is best for food and restaurants but frequency of other amenities like hospital, schools is less.

Cluster 1:

1. The average and Median price of Cluster one Neighborhoods are 610196.027397 and 607131.506849 respectively.

2. The cluster contains following places ASTORIA , BAYSIDE , BRIARWOOD , CORONA , DOUGLASTON , EAST ELMHURST , ELMHURST , FLUSHING-NORTH , FLUSHING-SOUTH , FOREST HILLS ,FRESH

MEADOWS ,GLENDALE , HILLCREST , JACKSON HEIGHTS , KEW GARDENS , LITTLE NECK , LONG ISLAND CITY , MASPETH , MIDDLE VILLAGE , OAKLAND GARDENS , REGO PARK , RICHMOND HILL , RIDGEWOOD , SUNNYSIDE , WOODSIDE

3. The average and median price is more compare to all other clusters .The most common venues nearby are Supermarkets , Restaurants, Bar , Park and Bagel Shop.

Cluster 2:

1. The average and Median price of Cluster one Neighborhoods are 474991.333333 and 458104.6 respectively.

2. The cluster contains following places ARVERNE , BELLE HARBOR , FAR ROCKAWAY , HAMMELS , NEPONSIT and ROCKAWAY PARK

3. The most common venues nearby are Beach, Pizza place,Bank,Bus stop and all kinds of Food Corners.

4. This should be second most preferred properties after Cluster 0 properties due to its average and median rates.

Cluster 3:

1. The average and Median price of Cluster one Neighborhoods are 511496.795918 and 458104.600000 respectively.

2. The cluster contains following places-

AIRPORT LA GUARDIA ,BEECHHURST ,BELLEROSE ,BROAD CHANNEL , COLLEGE POINT ,FLORAL PARK ,GLEN OAKS ,HOLLIS ,HOLLIS HILLS ,HOLLISWOOD ,HOWARD BEACH ,JAMAICA BAY , JAMAICA ESTATES, JAMAICA HILLS,OZONE PARK,QUEENS VILLAGE,SO. JAMAICA-BAISLEY PARK ,SOUTH OZONE PARK , WHITESTONE and WOODHAVEN

3. The most common venues nearby are Airport Lounge,Burger Joint,Pharmacy,Coffee Shop ,Parks etc.

4. The real estate properties are more expensive after cluster 1 properties.

Conclusion

At Last we state the problem scenario.

The problem scenario is to suggest the home buyers clients to purchase a suitable real estate in New York using Machine Learning Algorithms.

As a result, the business problem we are currently posing is:

How could we provide suggestions to home buyers clients to purchase a suitable real estate in New York street in this depreciating economy?

To solve this business problem, we are going to cluster New York neighborhoods in order to recommend venues and the current average price of real estate where home buyers can make a real estate investment.Also we will recommend profitable venues venues i.e. pharmacy , restaurants, hospitals & grocery stores.

First, we gathered data from The Department of Finance (DOF) maintains records for all property sales in New York City, including sales of family homes in each borough(https://data.cityofnewyork.us/api/views/948r-3ads/rows.csv?accessType=DOWNLOAD).

This list includes all sales of 1-, 2-, and 3-Family Homes' from January 1st, 2009 to December 31, 2009, whose sale price is equal to or more than $150,000. The Building Class Category for Sales is based on the Building Class at the time of the sale.

To explore and target recommended locations across different venues according to the presence of amenities and essential facilities, we will access data through FourSquare API interface and arrange them as a dataframe for visualization. By merging data on New York properties and the relative price paid data from the HM Land Registry and data on amenities and essential facilities surrounding such properties from FourSquare API interface, we will be able to recommend profitable real estate investments.

At last , We may analyze our results according to the five clusters we have produced. Even though, all clusters could praise an optimal range of facilities and amenities.

Cluster 3 - It have properties with almost average and median nearly close to each other and also the common venues also matching to each other but properties has more expensive than Cluster 1.

Cluster 0 and 2 - The average and median price is less compare to other clusters.

Cluster 1 - The average and median price is more compare to other clusters.

Reference

[1] https://ny.curbed.com/2018/9/18/17873488/new-york-home-prices-financial-crisisrecovery

[2] https://data.cityofnewyork.us/api/views/948r-3ads/rows.csv?accessType=DOWNLOAD

[3] FourSquare API

[4] Full Implementation of Project:

https://github.com/chandrashek1007/Capstone-Project---The-Battle-of-Neighborhoods-to-get-optimal-Real-Estate-properties/blob/master/Capstone_Project%20.ipynb