

Capstone Project - The Battle of Neighbourhoods

1. Introduction

Business Problem section

Background

New York city is world's major financial and cultural centre. As a result of this, real estate properties see lot of activity in terms of buying and selling. Due to financial crisis of 2008, NYC has seen a bit of downturn in real estate prices but recently it has seen some recovery. Investors looking for real estate investment are very curious to know the best way to invest in the current scenario.

Business Problem

Homebuyer client want optimal recommendation based on his requirements to buy a house in NYC. The house should meet his requirements, that is to be able to take a decision of buying in NYC.

Full Implementation of Project:

<https://github.com/curiouspushkar/Segmenting-and-Clustering-/blob/master/The%20Battle%20of%20Neighbourhoods%20WK1.ipynb>

2. Data Description

Using Data Science techniques learnt in this course and using FourSquare location data we will provide recommendations to client. We are going to cluster New York neighborhoods 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 i.e. pharmacy, restaurants, hospitals & grocery stores.

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.

Methodology


1. Collect Inspection Data
2. Explore and Understand Data
3. Data preparation and preprocessing
4. Modeling

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

| | 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 |



informations from Forsquare API.

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.

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.

Result

First of all, even though the London Housing Market may be in a rut, it is still an "ever-green" 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

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

FourSquare API