IBM CAPSTONE PROJECT

The Battle of Neighborhoods

Cluster Analysis of Manhattan Real Estate Market

Godkowicz Paweł, 9.05.2019

1. Introduction

Background

New York is also the most densely populated major city in the United States. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. Manhattan is the central island in New York City and is home to Times Square, the financial district with Wall Street and New York Stock Exchange, United Nations, Central Park, Metropolitan Museum of Art and many other famous landmarks and tourist attractions.

Problem

In this problem we use machine learning tools in order to help people, investors to make wise and effective decision. The business problem we put is: How can we offer support to investors, businessmen, to make a decision to buy property in New York's Manhattan district? In order to solve the problem, I intend to cluster neighborhoods in Manhattan to recommend venues and the current average price of real estate where homebuyers can make a real estate investments. We will recommend profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, restaurants, high schools, hospitals & grocery stores etc.

2. Data

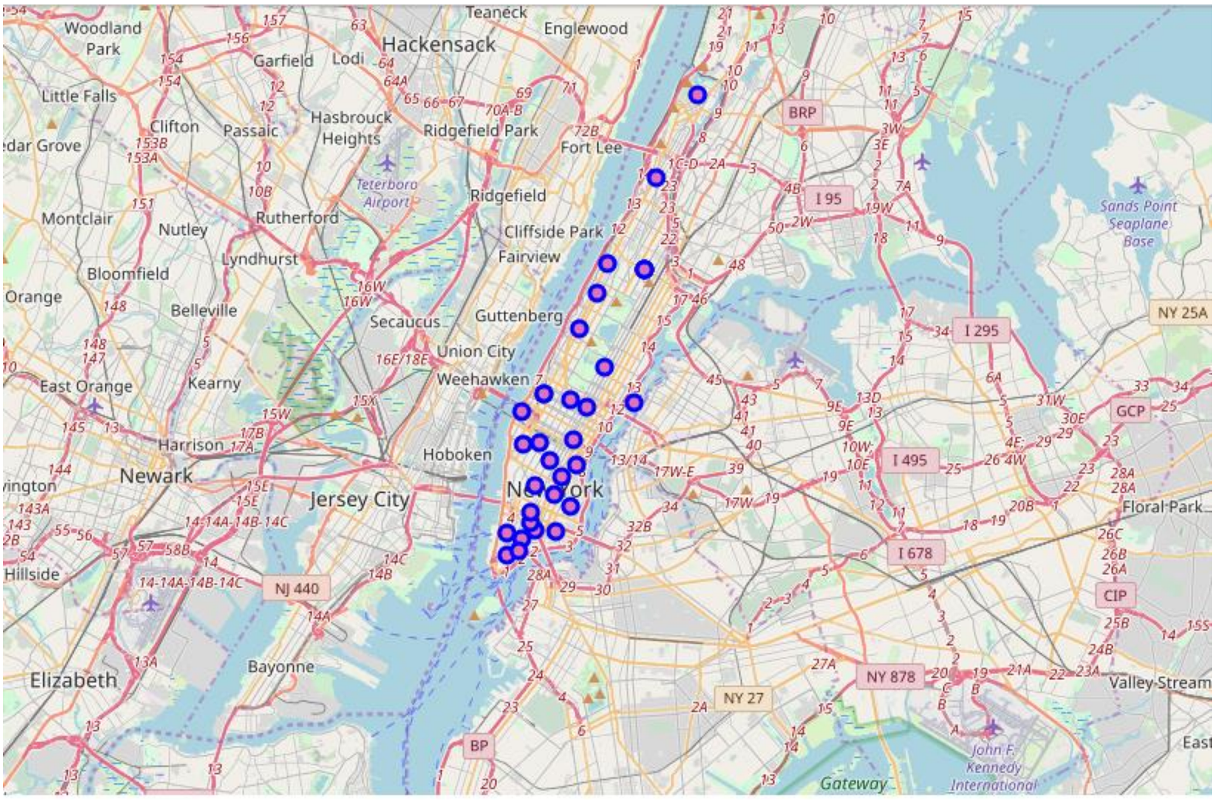
The Rolling Sales list has detailed sales information for the current twelve-month period in all five boroughs. Data on Manhattan properties and the relative price paid data were extracted from the NYC Department of Finance (https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page). The Rolling Sales list has detailed sales information for the current twelve-month period in all five boroughs.The The following data contains following columns: Borough, Neighborhood, Building Class Category, Tax Class At Present, 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 At Time Of Sale, Sale Price, Sale Date. We reduce the data in excel to 4 columns just for simplicity: Neighborhood, Address, Price, Date.

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 Manhattan properties and the relative price paid data from the NYC Department of Finance and data on amenities and essential facilities surrounding such properties from FourSquare API interface, we will be able to recommend profitable real estate investments.

3. Methotology

We download data from https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page and after cleaning and after removing the columns that are not interesting for us, we get data frame with 2 columns and 22924 rows. There are some missing value in column „Price”, so we remove them. Next we combine some neighborhoods for simplicity. After that we count our neighborhoods. Next we group the Neighborhoods by average price. The highest average price have Midtown Central neighborhood: $19 053 401. The lowest price 860 169 $, the Inwood neighborhood. In the next step we use „geolocator” library to get the each neighborhood coordinate (latitude and longitude).

Using „Folium” library we can see each neighborhood on the map:



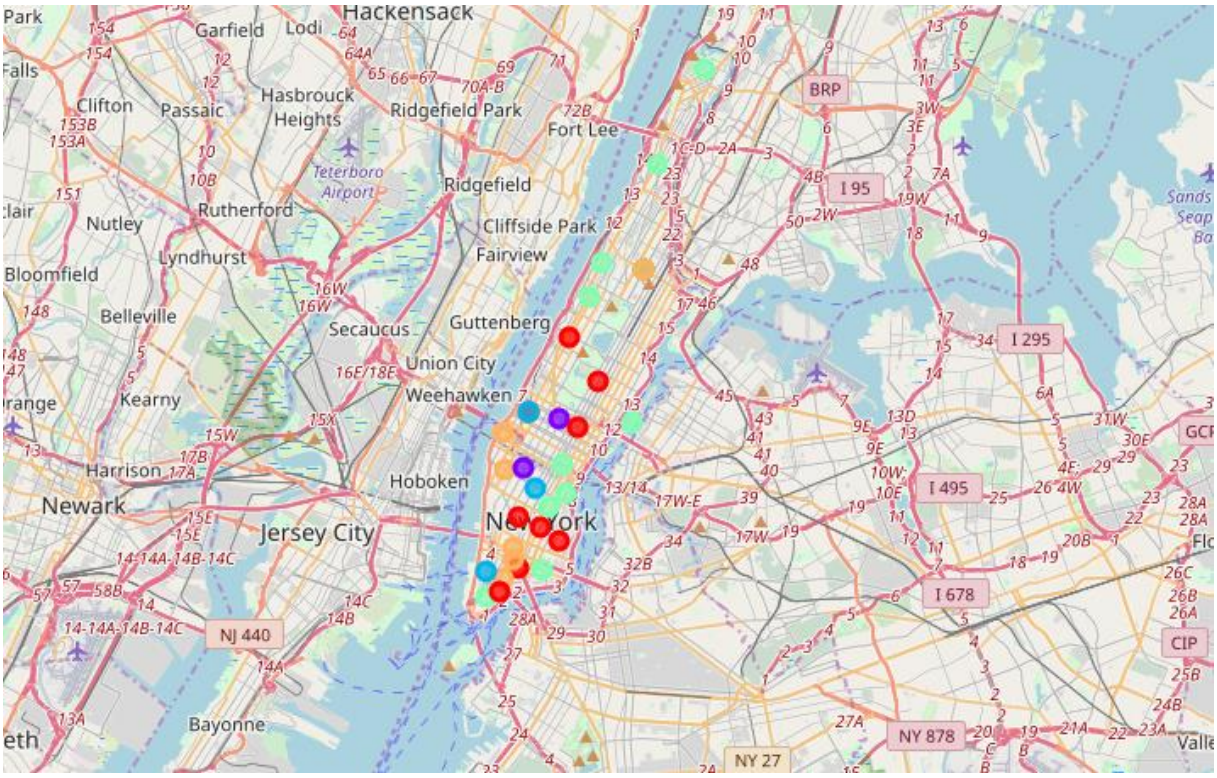
Modeling

Now we can deal with modeling. We will analyze neighborhoods to recommend real estates where home buyers can make a real estate investment. We will then recommend profitable venues according to amenities and essential facilities surrounding such venues.

Now we are going to use the cluster methotology to analyze our data. We will use the k-means clustering technique as it is fast and efficinet in terms of computational cost, is highly flexible to account for mutations in real estate market in Manhattan and is accurate.

4. Conclusion

"As New York’s population and economy continue to grow, every sector of the building industry—commercial, residential, healthcare, education, cultural and infrastructure-remains robust, offering opportunity for both the labor force and contractors,” says Carlo A. Scissura, the organization’s president and CEO. He points to the high costs of land and materials and regulations as the primary drivers of cost increases in 2018. He nonetheless adds, “While the cost of construction is high, the rewards for doing business in New York have never been greater.”.



By analyzing the results according to our five clusters, we can see that all clusters could praise an optimal range of facilities and amenities. The first pattern we are referring to, i.e. Cluster 1 (purple) which have the bigest average price may target investor, who value the neighborhood of hotels, Italian restaurants, gym, concert hall, maybe for potential Italian investors. Cluster 2 (blue) also have high average price, but with a larger neighborhood of the restaurant. The CLusters 0, 3, 4 (red, green, orange) may target potential buyers who are more interest to live in area full of restaurants, shops and parks.