

Find the neighborhood that better suits your desires

1. Introduction

1.1 Background

I recently moved to New York and, while struggling to decide the appropriate neighborhood where to move to. Several days spent on Google to look for the best suited neighborhood for me and my girlfriend, in such a huge unknown city, was only causing me headache; but the problem with all these lists were they just looked like opinions and not based on data. Coming from a science background, I understood that what I needed were data to support my choice. I want to avoid my choice to be a coin toss, therefore my idea is to use available data to support my decision and possibly avoid surprises...

1.2 Problem

In this project we will try to identify neighborhood in New York that are valuable for the life of a young couple. This project is targeted towards young couples looking for new homes or stakeholders who offer housing to young couples.

Finding a suitable neighborhood to base your future life is a challenge. Especially, in big cities like London. There are many variables to consider such as crime rates and easy access to venues such as Restaurants, Gyms, Pubs, and Libraries.

To determine which district to select, we will focus on neighborhoods with low crime rates and close to venues that meet young couple expectation, considering the following data

- Crime Rates for each neighborhood
- Types of venues
- The amount of most liked venues in each neighborhood

2. Data

2.1 Data sources

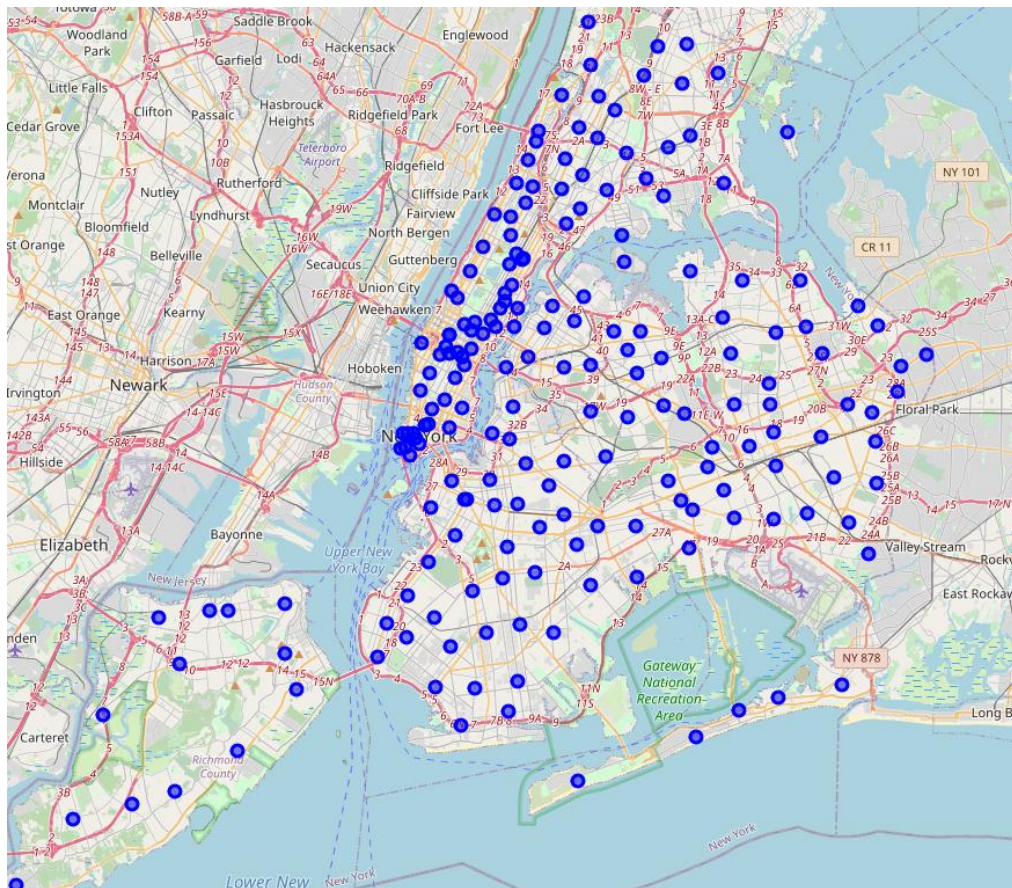
Crime Rate data are collected from NYPD data where all registered crimes are stored. Postal code of each location is obtained by Nominatim function. Venues data are obtained by using Foursquare API. The total number of venues within each category are calculated for each postal code.

2.2 Data cleaning

NYPD crime data have been cleaned from the incomplete rows. Based on the severity of the crime, a factor has been assigned to each crime in order to have a weighted value of the crime. From geographical coordinates a Postal Code is assigned to each crime from latitude and longitude of each crime.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes.

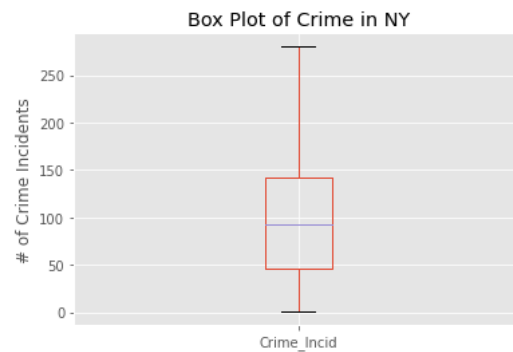
Map of New York



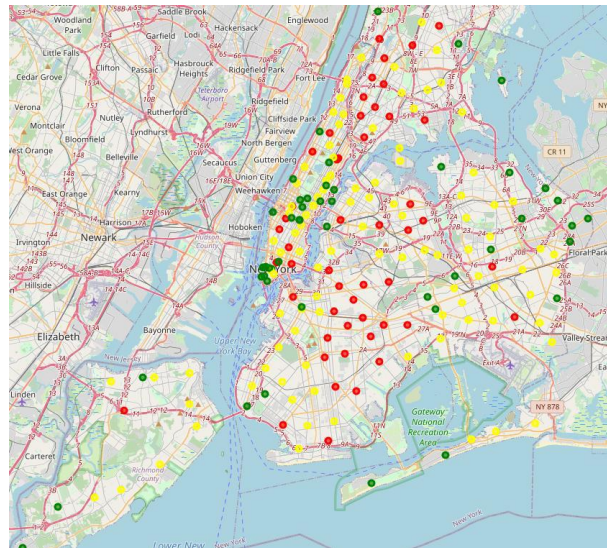
3. Methodology

3.1 Exploratory data analysis

Considering the crime incidents population in NY neighborhoods, it was decided to discard the unsafe areas.



The neighborhoods in the first quartile are considered safe, marked in green and considered for deeper analysis on the neighborhood selection.

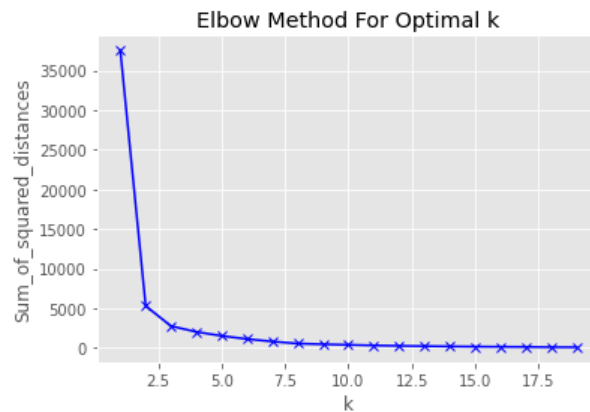


3.2 Clustering

In order to understand the neighborhood that would better suit us, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big city like New York. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

The quantity of important venues for us, like Gyms, Italian Restaurants and Nightlife spots are associated to each safe neighborhood of New York, using Foursquare API features.

Using **K-Means Clustering** approach to classify the neighborhoods based on the distribution of the liked venues, we are able to narrow down more the neighborhood in New York where we should look to start our new life. Elbow method is used to find the optimal number of clusters.



4. Results

Analyzing the average quantity of venues for each cluster defined, it is find out that cluster 1 is the one that more reflects our desires.

Therefore we can update the NY map look at the optimal location where this analysis has led us, and we find some neighborhoods in Financial District, Central Park area and Midtown.



5. Discussion

The purpose of this project was to find a good solution to analyze a wide area like New York and narrow down the possible neighborhoods where to buy a house, considering safety and personal information.

The remaining analysis can be manually done from considering the square-meter costs seen in the map.



And finally... let's start to look for some ads in Lower Manhattan!



6. Conclusion

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation.

This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools.

The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

Further steps can be added to this application tool, like deeper economical analysis for house costs and schooling, population distribution, and then this can become a really powerful tool that can help many people to at least narrow down the areas to look for a new house, based on a scientific method.