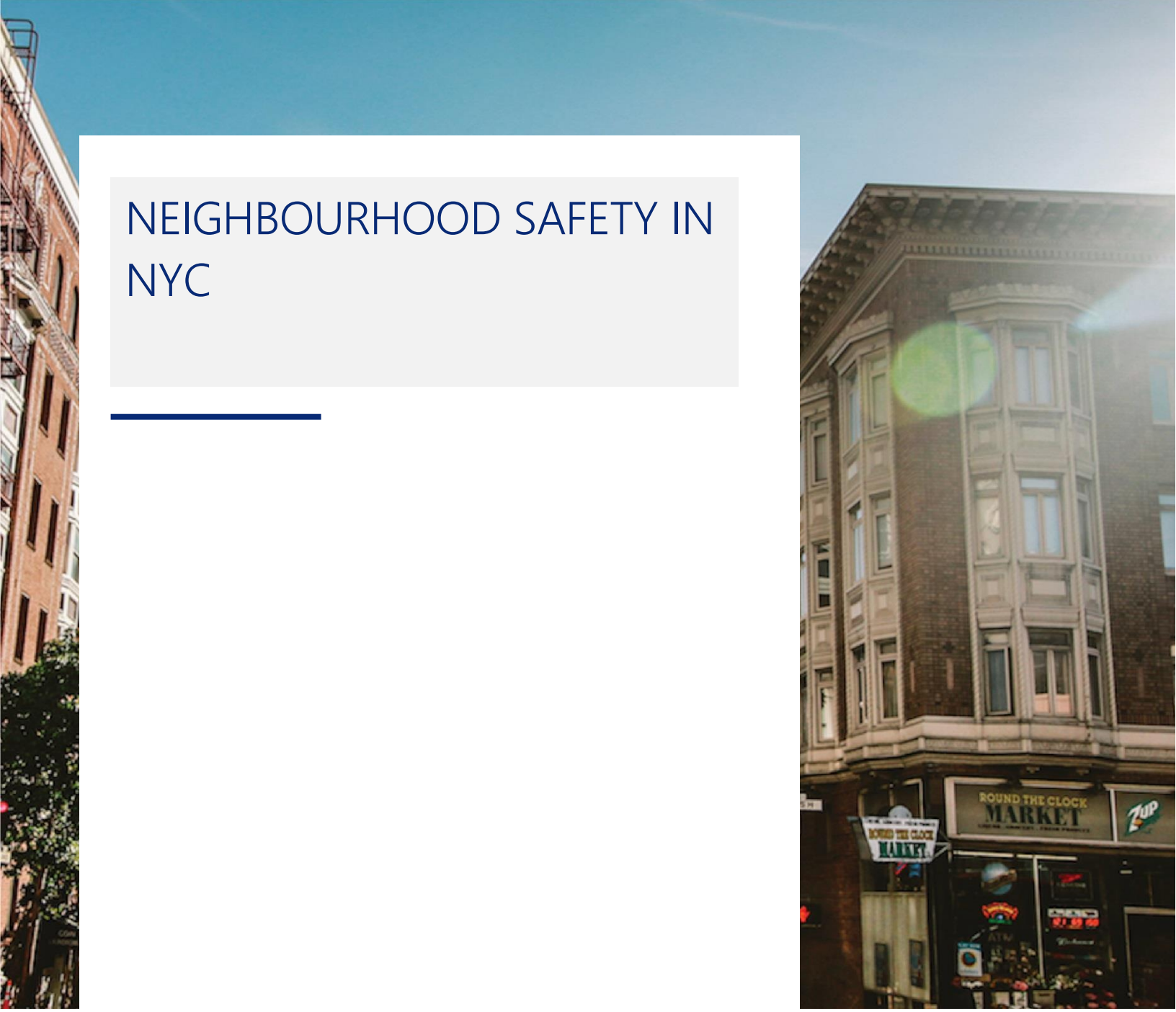
NEIGHBOURHOOD SAFETY IN NYC



AUGUST 3

IBM DATA SCIENCE CAPSTONE PROJECT

Authored by: Ifeoluwa Anjorin

INTRODUCTION

When you choose to live in New York, NY, you are choosing a lot more than a place to call home. You are choosing a lifestyle, defined by a city teeming with possibilities. Encompassing 8.4 million residents spread throughout five boroughs, New York's intersecting cultures and communities have established the city as a global destination for arts and culture, commerce and cuisine, and everything in between.

Hence, if anyone is considering moving to a new area of the city, it does not matter whether they are making the switch from Manhattan or the Midwest — people are likely to be following the same basic checklist. After the standard affordability factor, people also ask; where are the safest areas in NYC? Does this potential new neighborhood have relatively easy access to the subway? Access to good restaurants and bars? Do I feel comfortable walking there alone at night?

Cost, convenience, neighborhood amenities and security are only some of the factors that renters and buyers consider in selecting a home, and it is not likely that any one factor would be decisive. While safety statistics are probably not the only factor a prospective renter or buyer should consider in choosing a home, safety data is one type of information that may help to evaluate housing choices.

It is important to be proactive about one's safety, especially when it comes to choosing a place to live. After all, the safety of your neighborhood affects everything from one's happiness and sense of security to one's home resale value.

PROBLEM STATEMENT

The objective of this capstone project is to analyze and profile neighborhoods across all boroughs in New York City based on safety. Using data science methodologies and machine learning techniques like clustering, this project aims to provide solutions to answer the question: If a homebuyer or renter were looking for the safest place to stay in NYC, where would you recommend?

DATA

To solve the problem, we will need the following data:

- Neighborhoods across New York. This defines the scope of the project that is confined to the city of New York and its five boroughs.
- Latitude and longitude coordinates of those neighborhoods. This is required in other to plot the map and give venue data
- Crime data related to arrests for some specific offences and venue data related to police departments. We will use this data to perform clustering on the Neighborhoods.

DATA SOURCES

The Crime Data is a List of every arrest in NYC in 2017. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents an arrest effected in NYC by the NYPD and includes information about the type of crime, the location and time of enforcement. The data can be found here :

https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

In addition, Foursquare API will be used to get the venue data for those neighborhoods. Foursquare has one of the largest database for millions of locations and is used by thousands of developers. Foursquare provides many categories of venue data but for this project, the interest lies in Police department category.

METHODOLOGY

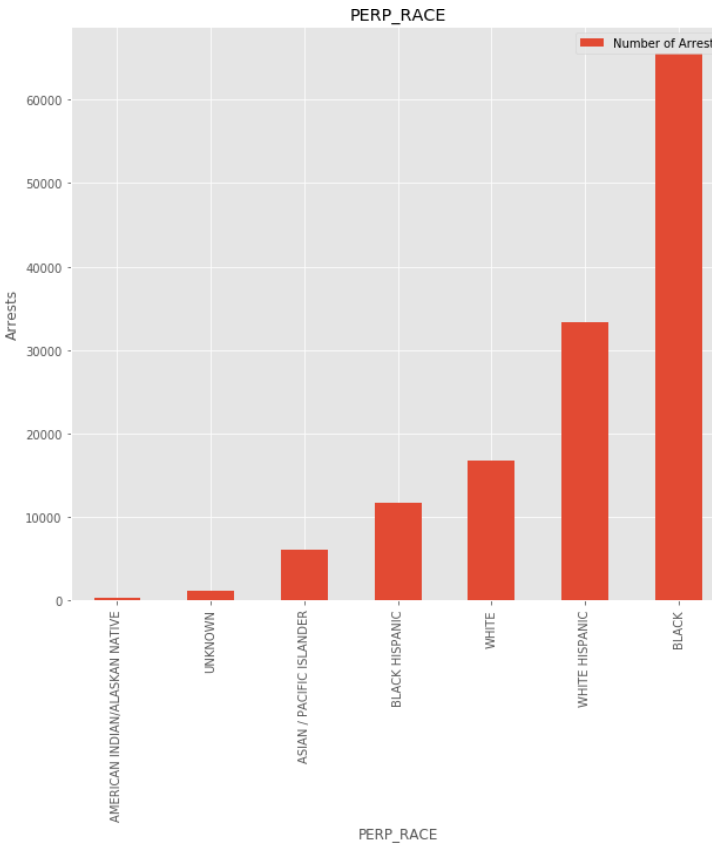
The data contains 18 attributes or columns that describe every arrest in New York documented in the year 2017. The data has exactly 283,853 instances or rows that represent each unique arrest. There was no need to use Foursquare to generate coordinates as it was already supplied with the data

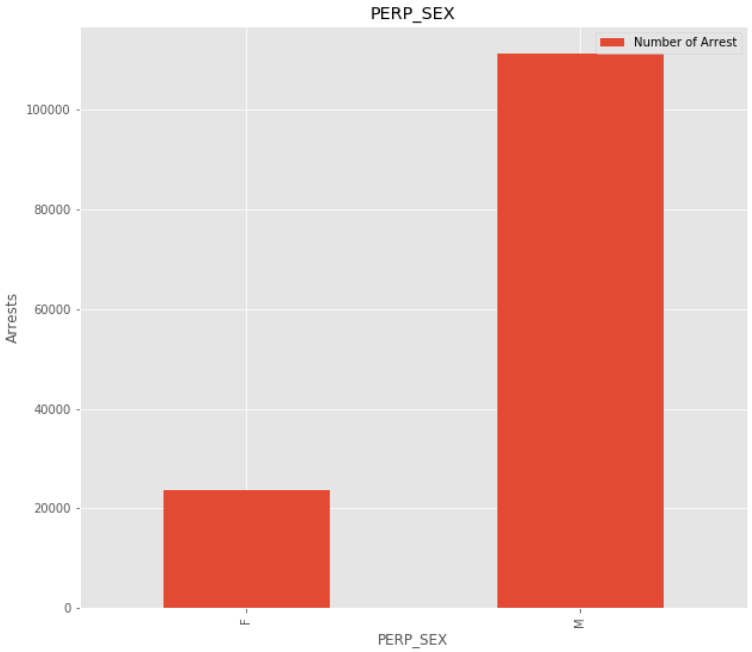
However there isn't any attribute indicating which neighborhood in New York city each arrest coordinate belongs to. To extract the coordinates, the geo-names reverse geo-coding API was used to import neighborhoods for each arrest coordinate. The API provided neighborhood attribute for 134,832 arrest instances while other arrest instances with unknown neighborhoods were removed from the final data. Meta-attributes such as Weekday, Day type and Month name were also added to the data to allow for more in-depth time intelligence analysis. The data was then analyzed to show arrest trends for the year, the top most crimes with the highest number of arrests in New York. Also the data was analyzed to extract the top 5 crimes in each of the 5 boroughs along with the top 20 and least 20 neighborhoods with the highest and lowest frequencies of crimes respectively.

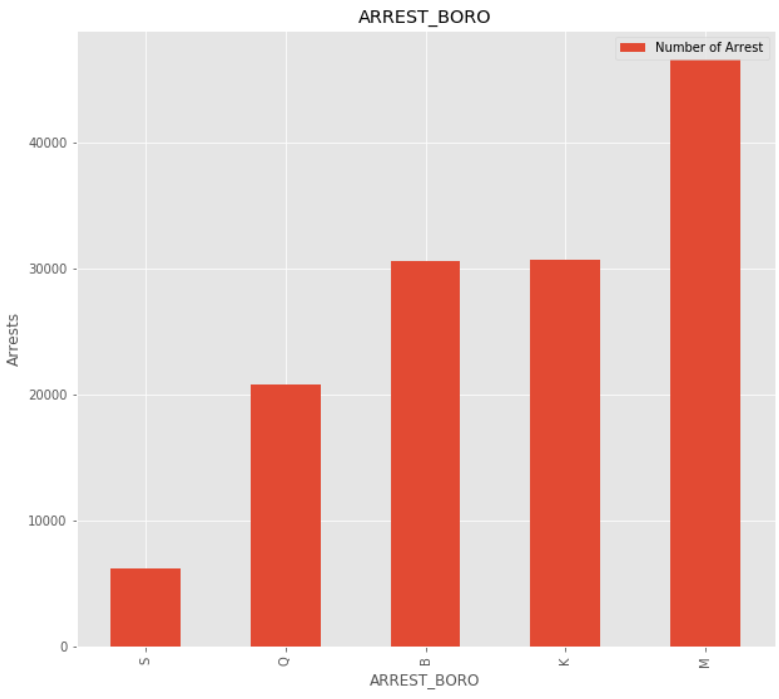
The data consists of 265 unique crime categories; therefore, one cannot simply choose a neighborhood to relocate to just because it is in a borough with lower overall crimes committed. Hence, the top 5 crime categories in each neighborhood was also highlighted. Using crime average frequencies for each neighborhood and crime categories, the neighborhoods were clustered into 10 using the K-means clustering algorithm and plotted on a map in other to visually identify which cluster each neighborhood belongs to.

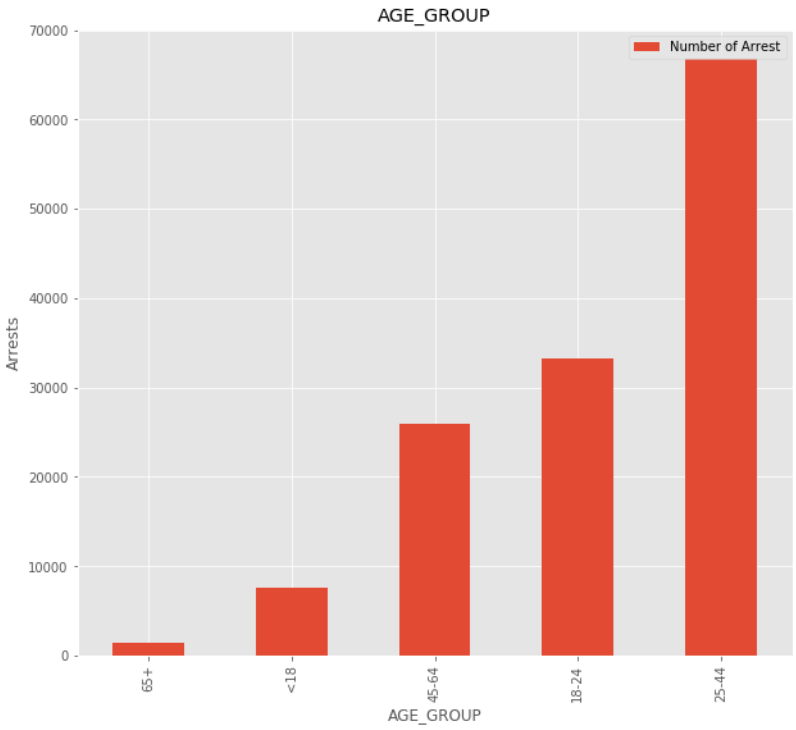
ANALYSIS

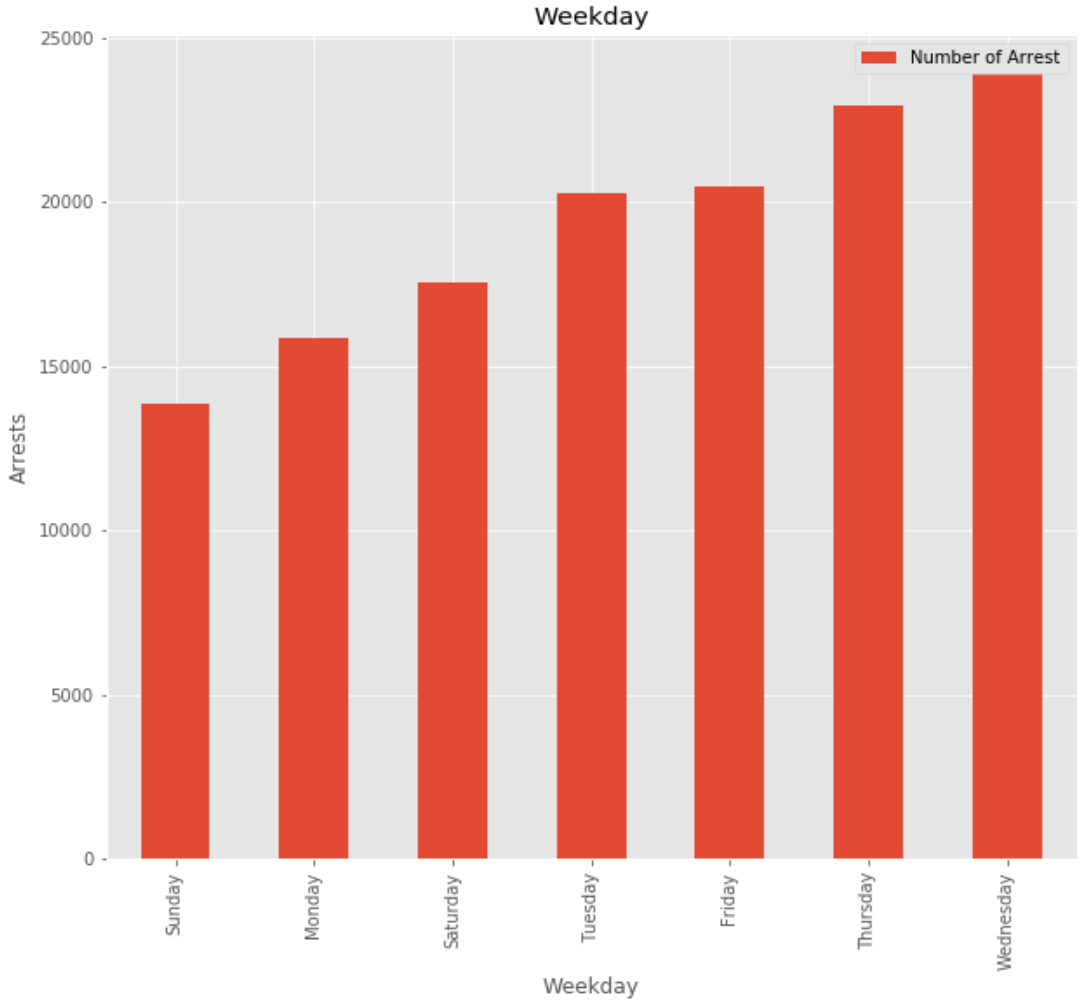
Visual exploratory analysis of categorical attributes in the data is has shown in the series of plots highlighted below:

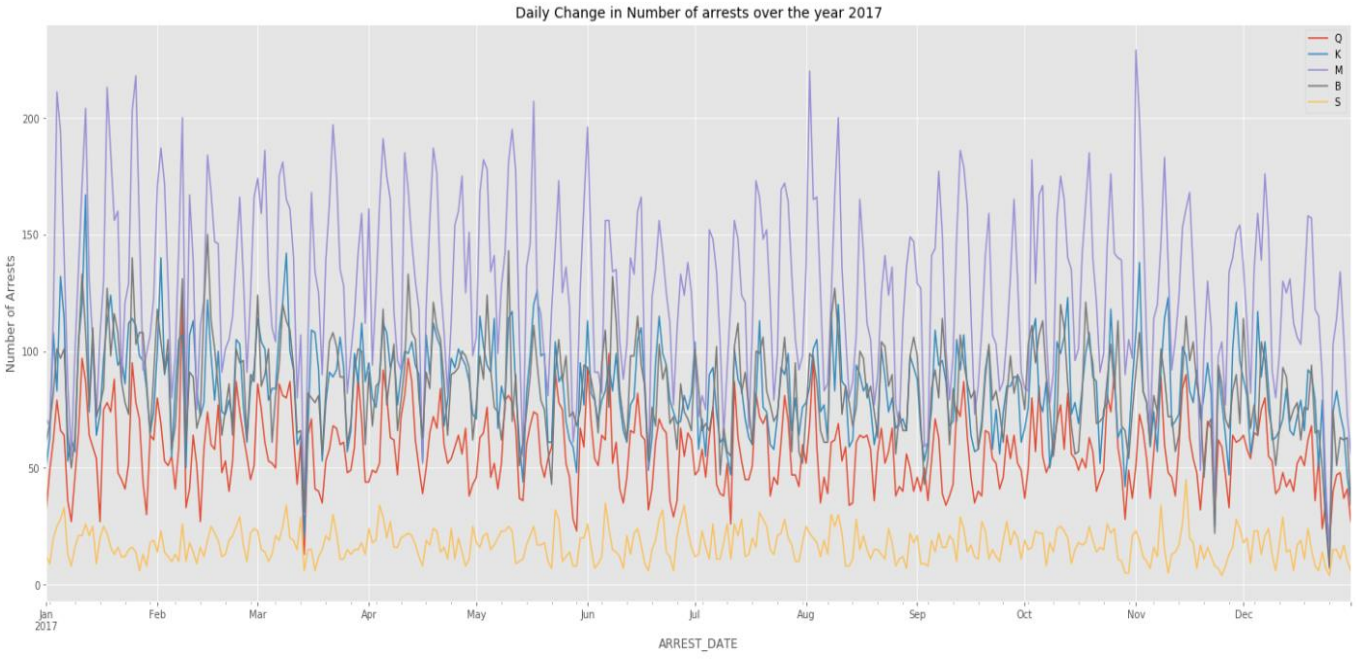


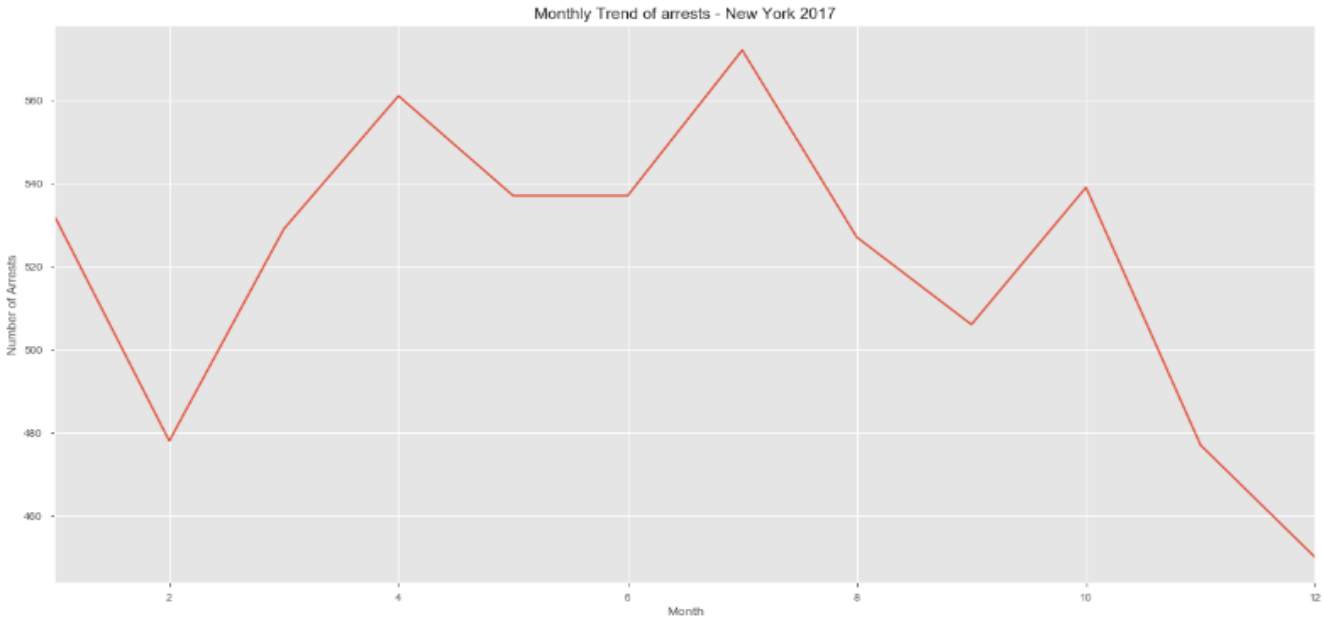


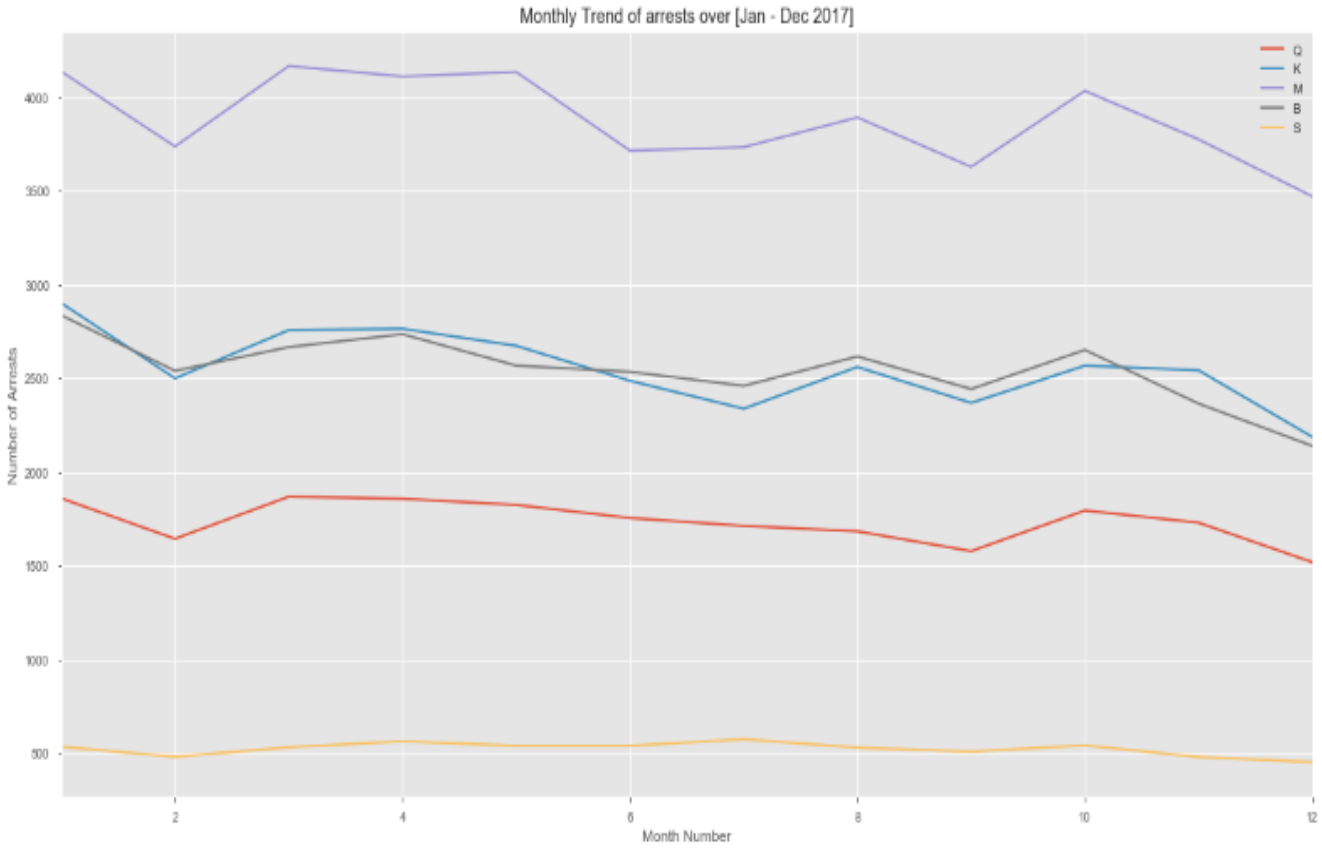


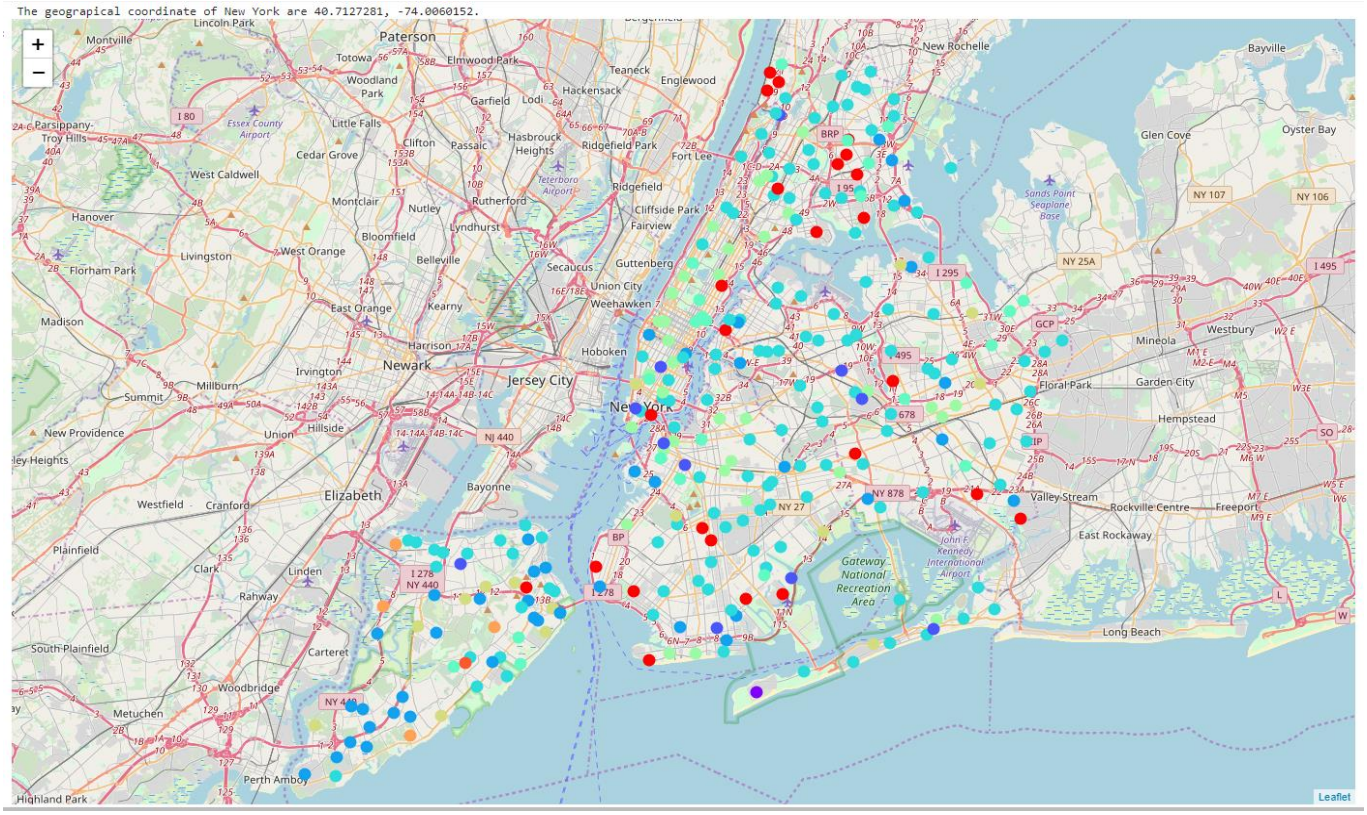



Weekday
Number of Arrest
25000
20000
15000
10000
5000
0
Arrests
Sunday
Monday
Saturday
Tuesday
Friday
Thursday
Wednesday
Weekday


Daily Change in Number of arrests over the year 2017
Q
K
M
B
S
200
150
100
50
0
Number of Arrests
Jan
2017
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
ARREST_DATE


Monthly Trend of arrests - New York 2017
Number of Arrests
560
540
520
500
480
460
2
4
6
8
10
12
Month


Monthly Trend of arrests over [Jan - Dec 2017]
Q
K
M
B
S
Number of Arrests
4000
3500
3000
2500
2000
1500
1000
500
2
4
6
8
10
12
Month Number



RESULTS AND DISCUSSION

Exploratory analysis shows Manhattan has the highest number of crimes (46,558) which is about 35% of the entire crimes dataset. These crimes seem to be mostly committed by male black individuals in the 25-44 age group with majority of the crime related arrests occurring during the summer and lowest in the winter period. Arrests seem to happen a lot on Wednesdays and Thursdays. Based on these inference, one might say it's best to relocate to Staten highland borough, which has the lowest crime rates, however, the distinction between neighborhoods is not clearly delineated by the borough they belong to.

A neighborhood in Staten highland might have crime characteristics that is similar to neighborhoods in boroughs with higher crime rates and vice versa. Therefore, to properly identify neighborhoods that share similar properties, it is important to reference the plot of clustered neighborhoods on the Map plot. This is helpful especially when trying to find a neighborhood in another borough that shares similar crime properties but has lower real-estate cost compared to another neighborhood of interest in a similar or different borough.

CONCLUSION

This analysis will be very important for individuals who intend to relocate to New York City but would like to identify the safer neighborhoods in any borough of choice. In addition, tourists and real estate agents can benefit from this analysis. The results can be further improved on if neighborhood attribute information is available for all the data instance of arrests, data from other years is also analyzed and average cost of houses in each of the neighborhood is put into consideration as part of the analysis.

References

- https://www.forbes.com/sites/trulia/2016/10/04/the-top-10-new-york-city-neighborhoods-to-live-in-according-to-the-locals/#8ee653314940

- https://www.safewise.com/blog/confirm-safety-neighboorhood-online-tools/