Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

New York City(NYC) is the most populous city in the United States. NYC has been described as the cultural, financial, and media capital of the world. Also, it is an important center for commerce, entertainment, research, technology, art, and fashion. NYC is composed of five boroughs. which are Brooklyn, Queens, Manhattan, the Bronx, and Staten Island. It is estimated that population of about 8.4 million distributed over the city. NYC is considered as the busiest and largest financial centers in the world.

According to NYC & Company, 62.8 million tourists have visited New York City, and the number of visitors is increasing. However, not like the beautiful appearance of the city, about five hundred thousand of crimes is occurring, and it is a threat not only to New Yorkers but also to visitors. So, we are going to explore NYC to find which borough need to increase public security by analyzing the number of crime occurred and distribution of police stations in NYC.

Obviously, city officers in New York will be very interested in the result of this analysis. Also, others who are planning to visit New York City will be interested.

Methodology

In this project we will take a closer look at the distribution of police stations and the number and location of crimes occured by boroughs. We will limit our crime data set to 100000 rows.

In the first step, we will collect and clean information on distribution of police stations and crimes occured in NYC. This will give us a rough sense of each borough if they are likely to be safe or not. This step includes getting additional data such as latitude and longitude using Google Maps API.

Second step will be exploring the processed data. We will use basic visualization tools like histogram and bar charts to identify relatively safe and dangerous boroughs.

Last step will be visualizing the results from the second step using Folium library. Through this, we will be able to easily understand which boroughs need to increase public security, and the relation between distribution of police station and crime situation.

Step1_Data collection and cleaning

Based on definition of our problem, factors that will influence our decision are:

- the number and location of crimes and occurred in NYC
- address of police stations in NYC

In [94]:

```
import pandas as pd
import numpy as np
import googlemaps
gmaps=googlemaps.Client(key='KEY')

!conda install -c conda-forge folium=0.5.0 --yes
import folium

import string
import requests
import geopy
from geopy.geocoders import Nominatim

%matplotlib inline
import matplotlib.pyplot as plt
```

```
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

# All requested packages already installed.
```

In [10]:

```
#crime data
crime = pd.read_excel('c://data//crime_nyc.xlsx')
crime.drop('Unnamed: 0',axis=1, inplace=True)
crime.dropna(inplace=True)
crime.dropna(inplace=True)
crime.head()
```

Out[10]:

	Level	Borough	Latitude	Longitude
0	FELONY	BRONX	40.828848	-73.916661
1	FELONY	QUEENS	40.697338	-73.784557
2	FELONY	MANHATTAN	40.802607	-73.945052
3	MISDEMEANOR	QUEENS	40.654549	-73.726339
4	MISDEMEANOR	MANHATTAN	40.738002	-73.987891

In [13]:

```
#address
url = 'https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page'
df = pd.read_html(url)[0]

lst = []
for i in range(0,df.shape[0]):
    if df.iloc[i,0] in ['Manhattan','Brooklyn','Queens','Bronx','Staten Island']:
        lst.append(i)

police_df = df.drop(lst)
police_df['Precinct'] = police_df['Precinct'] + ', NY'
print('police_df consists of ', police_df.shape[0],'rows and ',police_df.shape[1],' columns')
```

```
police_df consists of  77 rows and  3  columns
```

Following data source will be needed to generate the required information

- Latitude and Longitude of each police station will be obtained by Google Maps API geocoding.

In [16]:

```
search_keyword = police_df['Precinct'].tolist()
station_lat = []
station_lng = []
station_address = []

for name in search_keyword:
    tmp = gmaps.geocode(name)

    station_address.append(tmp[0].get('formatted_address'))

    tmp_loc =tmp[0].get('geometry')
    station_lat.append(tmp_loc['location']['lat'])
    station_lng.append(tmp_loc['location']['lng'])

police = pd.DataFrame({'Address':station_address,
                       'Latitude':station_lat,
                       'Longitude':station_lng})
police.head()
```

Out[16]:

	Address	Latitude	Longitude
0	16 Ericsson Pl, New York, NY 10013, USA	40.720369	-74.006969
1	19 Elizabeth St, New York, NY 10013, USA	40.716201	-73.997477
2	233 W 10th St, New York, NY 10014, USA	40.734000	-74.005433
3	19 1/2 Pitt St, New York, NY 10002, USA	40.716497	-73.983989
4	321 E 5th St, New York, NY 10003, USA	40.726538	-73.987820

Step2_Data exploration

Let's perform baic explanatory analysis and derive some additional info from our raw data.

First we need to get borough information of police station to see rough distribution.

In [21]:

```
police_ad = []
police_address = []
for i in range(0,len(police['Address'])):
    police_ad.append(police['Address'].tolist()[i].split(','))
    police_address.append(police_ad[i][1])

for i in range(0,len(police_address)):
    if police_address[i] == ' New York':
        police_address[i] = 'Manhattan'
    elif police_address[i] == ' Brooklyn':
        police_address[i] = 'Brooklyn'
    elif police_address[i] ==' Staten Island':
        police_address[i] = 'Staten Island'
    elif police_address[i] ==' The Bronx':
        police_address[i] = 'Bronx'
    else:
        police_address[i] = 'Queens'

police['Borough'] = police_address
```

In [90]:

```
police_group = police.groupby(police['Borough']).count().iloc[:,0]
police_group=pd.DataFrame(police_group).rename(columns={'Address':'Number'})
police_group
```

Out[90]:

Borough	Number
Bronx	12
Brooklyn	23
Manhattan	22
Queens	16
Staten Island	4

This table shows that more than half of police station is located in Manhattan and Brooklyn. But, there are only 4 in Staten Island. Let's get grouped crime data.

In [28]:

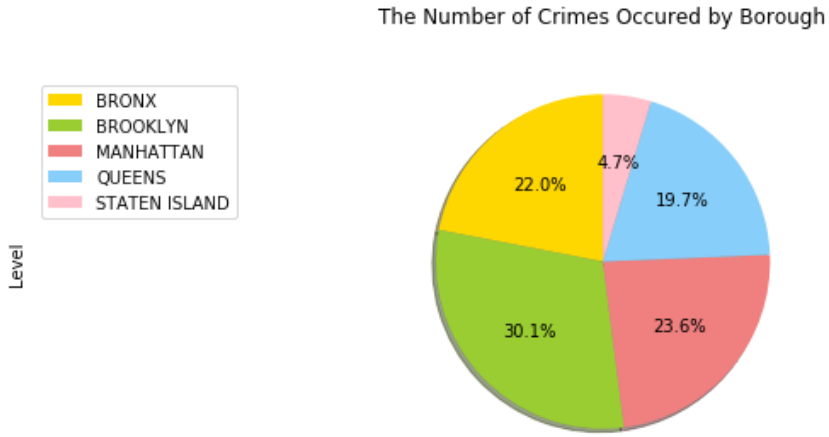
```
crime_sum=crime.groupby(crime['Borough']).count().iloc[:,0]
pd.DataFrame(crime_sum)
```

Out[28]:

Borough	Level
BRONX	102258
BROOKLYN	139999
MANHATTAN	110033
QUEENS	91723
STATEN ISLAND	21656

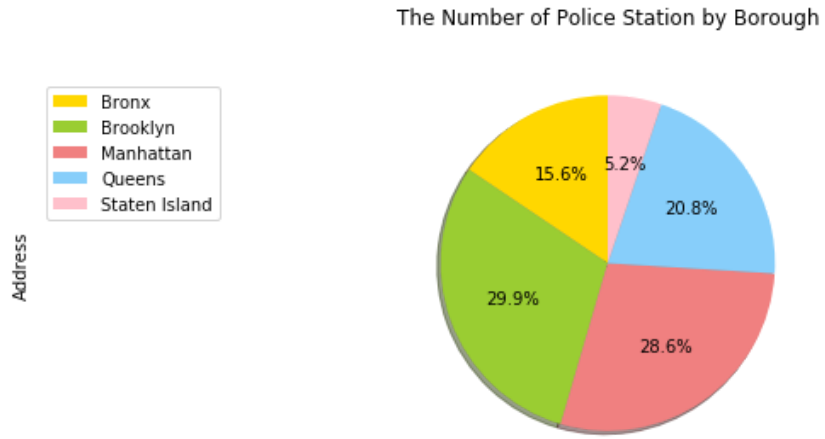
In [49]:

```
colors_list = ['gold', 'yellowgreen', 'lightcoral', 'lightskyblue','pink']
crime_sum.plot(kind='pie',
               figsize=(12, 4),
               autopct='%1.1f%%',
               startangle=90,
               shadow=True,
               labels=None,
               colors=colors_list)

plt.title('The Number of Crimes Occured by Borough', y=1.12)
plt.axis('equal')
plt.legend(labels=crime_sum.index, loc='upper left')
plt.show()
```



In [79]:

```
colors_list = ['gold', 'yellowgreen', 'lightcoral', 'lightskyblue','pink']
police_group.plot(kind='pie',
                  figsize=(12, 4),
                  autopct='%1.1f%%',
                  startangle=90,
                  shadow=True,
                  labels=None,
                  colors=colors_list)

plt.title('The Number of Police Station by Borough', y=1.12)
plt.axis('equal')
plt.legend(labels=police_group.index, loc='upper left')
plt.show()
```



It seems like relatively bronx is not that safe compared to its proportion of police stations.

Step3_Visualization with Map

To easily understand, let's visualize this result. We will plot the location of police stations and count each crime case.

In [71]:

```
address = "New york City, NY"

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude

print(address,'Latitude & Longitude ',latitude,longitude)
```

```
New york City, NY Latitude & Longitude  40.7127281 -74.0060152
```

In [72]:

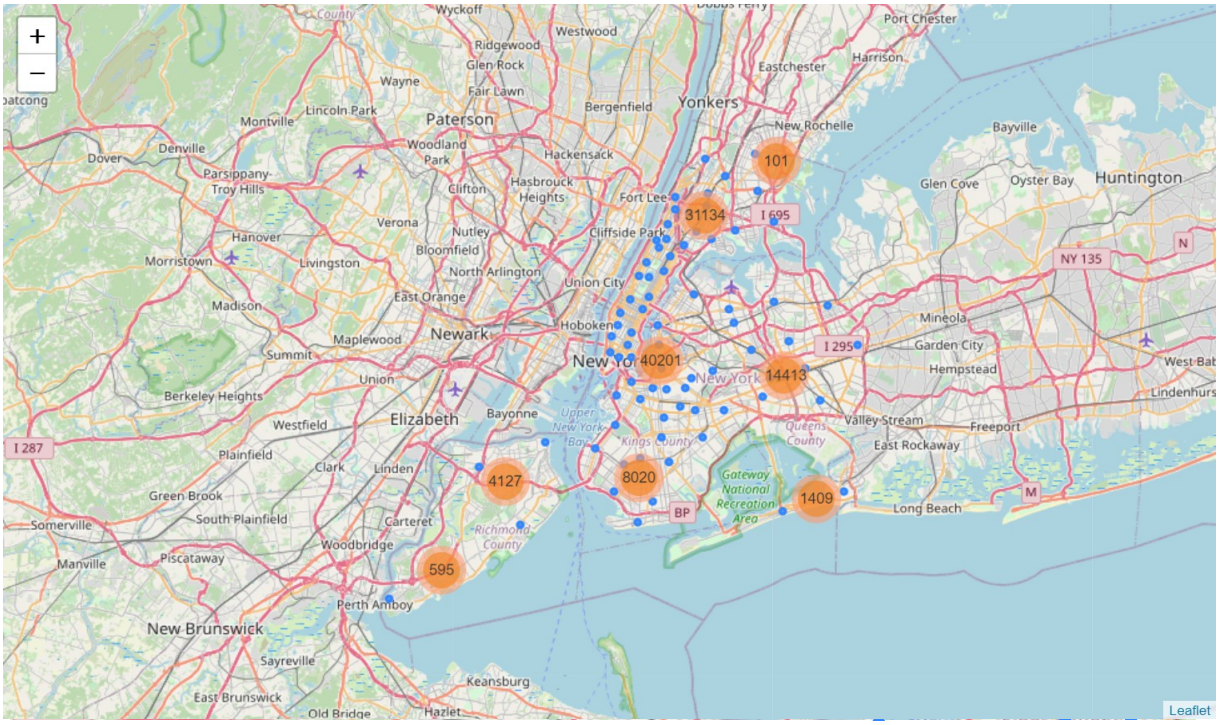
```
map_newyork2 = folium.Map(location=[latitude, longitude], zoom_start = 10)
```

using for statement, we will plot points indicating location of police stations and aggregated crime cases

In [75]:

```
for lat, lng in zip(police.loc[:,'Latitude'].tolist(),police.loc[:,'Longitude'].tolist()):
    folium.CircleMarker(
        [lat,lng],
        radius=2,
        fill=True,
        fill_color='#3186cc',
        fiil_opacity=0.5).add_to(map_newyork2)

from folium import plugins
incidents = plugins.MarkerCluster().add_to(map_newyork2)

for lat, lng in zip(crime['Latitude'].head(100000),crime['Longitude'].head(100000)):
    folium.Marker(
        location=[lat, lng],
        icon=None,
    ).add_to(incidents)


map_newyork2
```

Out[75]:



The map above shows the distribution of police station and the number of crime occured by areas. We can see that the majority of crime happens in manhattan and brooklyn, where more than half of police stations are. This seems resonable, because most of police stations are located where they need.

Similar pattern can be found in Staten Island. Like we saw at step2, there were only 4 stations in Staten Island. Here, 4722 crime cases out of 100000, which is less then 5 percent, happened in Staten Island.

However, Bronx borough seems like it needs more police station.

Results and Conclusion

Our analysis shows that Bronx has less police stations than other boroughs in comparison with crime cases

happened in the borough.

As we found from the second step, although most crimes happen in Manhattan and Brooklyn, they have enough facilities for public security. But, Bronx has about a half of them.

Purpose of this project was to figure out boroughs which need more security facilities. Though this analysis, we can provide an idea that Bronx needs more police station with city officials and administrators in NY to choose where to build new facilities for public security. Plus, tourists planning to visit new york city can figure out which spot is relatively safer.

In [ ]: