



# New York City Crime analysis

**Team - The Dorm Room Developers**

- Abhishek Kuntal
- Nanda Kishore Kalidindi
- Saurav Mawandia

# Contents

<b>1. Project Overview</b>	3
<b>2. Crime Dataset</b>	4
<b>3. Visualizing crime data</b>	6
3.1 Crimes week visualization for the Week of June 12, 2016	6
3.2 Crimes in NYC	6
3.3 Per Hour Average Crime count	7
3.4 Month Vs Crime	7
3.4 Crime at hour of day, season of the year	8
<b>4. Correlation of Crime with weather</b>	9
4.1 Crime Count vs Temperature in Degree Celsius	10
4.2 Number of Crimes Vs Temperature vs Hour of the Day	12
<b>5. Correlation of Crime with population and Median Income</b>	13
5.1 Crime in different areas of NYC	13
5.2 Median Income in different areas of NYC	13
5.3 Population in different areas of NYC	14
5.4 Mean Income Vs Crime	14
<b>6. Correlation of Crime with Markets</b>	15
6.1 Decrease in Crime with Increase in Dow Jones Index	15
6.2 Impact of 2008 recession on crime	15
<b>7. Recommendation based on Crime</b>	16
7.1 Hotels in NYC	16
7.2 Bars in NYC	17
<b>8. Technology Used</b>	19
<b>9. Conclusion</b>	19

# 1. Project Overview

With the number of crime incidents increasing daily, we thought of analyzing NYC crime data and find out whether the below factors have any influence on perpetrator

- Weather (Gathered last 10 years data from Open data)

Data Source: <http://w2.weather.gov/climate/index.php?wfo=okx>

- Stock Market

Data Source: <http://www.macrotrends.net/1358/dow-jones-industrial-average-last-10-years>

- Mean Household Income (Collected last 10 years of population in NYC on various zip code and there median and mean Income)

Data Source: <https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/>

We also took this opportunity to find number of crime incidents near frequently visited places

- Compiled the list of safe and unsafe hotels in New York city

Data Source: Trip Advisor <http://times.cs.uiuc.edu/~wang296/Data/LARA/TripAdvisor/TripAdvisorJson.tar.bz2>

- Compiled the list of safe and unsafe Bars in New York city

Data Source: <https://data.ny.gov/Economic-Development/Liquor-Authority-Quarterly-List-of-Active-Licenses/hrvs-fxs2/data>

## 2. Crime Dataset

The Crime [dataset](#) has 24 columns as shown below.

---

<i>CMPLNT_NUM</i>	<i>CMPLNT_FR_DT</i>	<i>CMPLNT_FR_TM</i>	<i>CMPLNT_TO_DT</i>
<i>CMPLNT_TO_TM</i>	<i>RPT_DTKY_CD</i>	<i>OFNS_DESC</i>	<i>PD_CD PD_DESC</i>
<i>CRM_ATPT_CPTD_CD</i>	<i>LAW_CAT_CD</i>	<i>JURIS_DESC</i>	<i>BORO_NM ADDR_PCT_CD</i>
<i>LOC_OF_OCCUR_DESC</i>	<i>PREM_TYP_DESC</i>	<i>PARKS_NM</i>	<i>HADEVELOPT X_COORD_CD</i>
<i>Y_COORD_CD</i>	<i>Latitude</i>	<i>Longitude</i>	<i>Lat_Lon</i>

---

We started by analyzing importance of each field. As, we did not find good metadata for the data.

- After analyzing the data, we found out there are three major types of crime

Code: `crime_df.select("LAW_CAT_CD").distinct().show()`

Results: *Felony, Misdemeanor, Violation*

- We also saw how many unique OFNS\_Description are there in dataset

Code: `crime_df.select("OFNS_DESC").distinct().count()`

Result: We found out 71 unique description.

- We wanted to plot data in map based on zip code, but zip code was not present in dataset. So, we extracted zip code from latitude and longitude using google reverse geocoding API. This was the most challenging part of the project. As we ran out of limitation on calling the API and we had to gather the data for almost a week to get complete data (P.S we upgraded our google account to premium account to get the data)

```

import requests
from pyspark.sql.types import StringType

def get_pincode(lat, lng):
    try:
        sensor = 'true'
        base = "https://maps.googleapis.com/maps/api/geocode/json?key= &"
        params = "latlng={lat},{lon}&sensor={sen}".format(lat=lat,lon=lng,sen=sensor)
        url = "{base}{params}".format(base=base, params=params)
        response = requests.get(url).json()
        results = response["results"]
        address_components = list(map(lambda x: x["address_components"], results))
        flatten = [item for sublist in address_components for item in sublist]
        postal_objects = list(filter(lambda x: x["types"] == ["postal_code"], flatten))
        postal_codes = list(map(lambda x: x["long_name"], postal_objects))
        return str(postal_codes[0]);
    except Exception as e:
        return None

udf_getZip = udf(get_pincode, StringType())

```

*Figure 1 Code Snippet to get zipcode form Lat,Lon*

We then visualized the crime data we had and gathered some meaningful insights from it. We also normalized the date and time to visualize data as shown below.

```

# Date conversion
from pyspark.sql.functions import udf

crime_df_date = crime_df.select("CMPLNT_FR_DT", "CMPLNT_FR_TM", "OFNS_DESC", "LAW_CAT_CD",
"Latitude", "Longitude", "PD_DESC")

def crime_date_convert(date, time):
    try:
        k = date.split("/")
        full_date = "-".join(k[2:] + k[1:2] + k[0:1])
        full_date = full_date + " " + time.split(":")[0]
        return full_date
    except:
        return "2021-01-01 00"

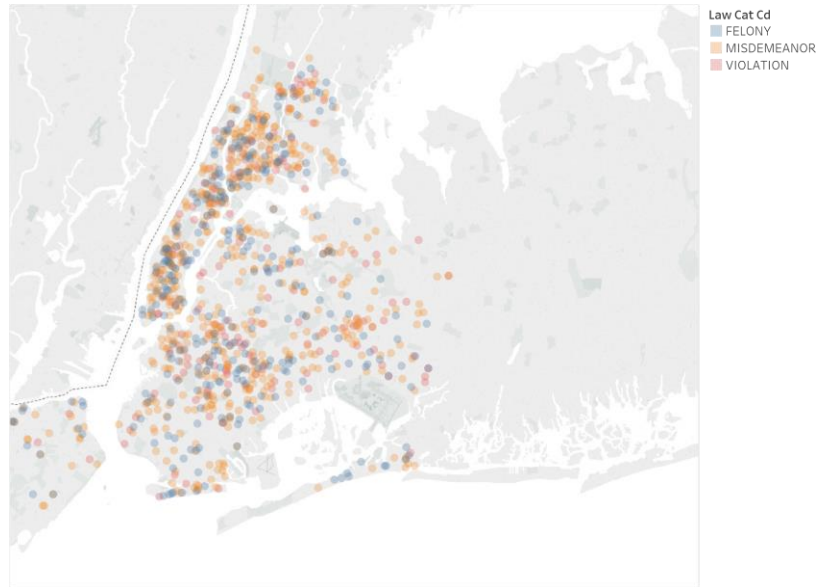
udf_crime_date_convert = udf(crime_date_convert, StringType())

with_crime_date_conversion = crime_df_date.withColumn("normalized_date",
udf_crime_date_convert("CMPLNT_FR_DT", "CMPLNT_FR_TM")).select("normalized_date",
"OFNS_DESC", "LAW_CAT_CD", "Latitude", "Longitude", "PD_DESC")

```

### 3. Visualizing crime data

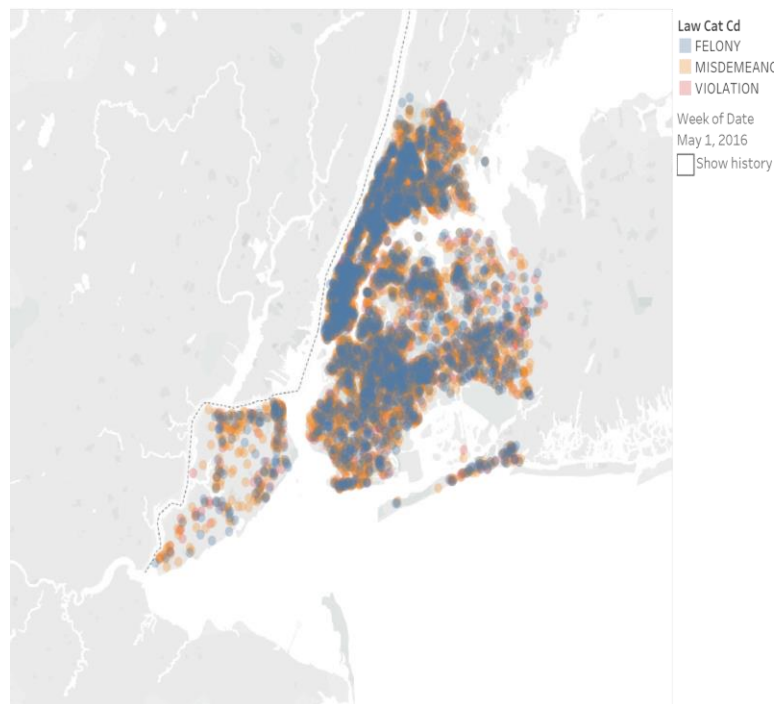
#### 3.1 Crimes week visualization for the Week of June 12, 2016



**Map based on Longitude and Latitude. Color shows details about Law Cat Cd.**

We visualized the three-different type of crime on map for the week of June 12, 2016 and tried understanding what are the locations which has most number of Incidents

#### 3.2 Crimes in NYC



*Figure 2 Crime in NYC for Last 10 years*

We also tried to see that over a period of time which type of crime occurs at a particular location

### 3.3 Per Hour Average Crime count

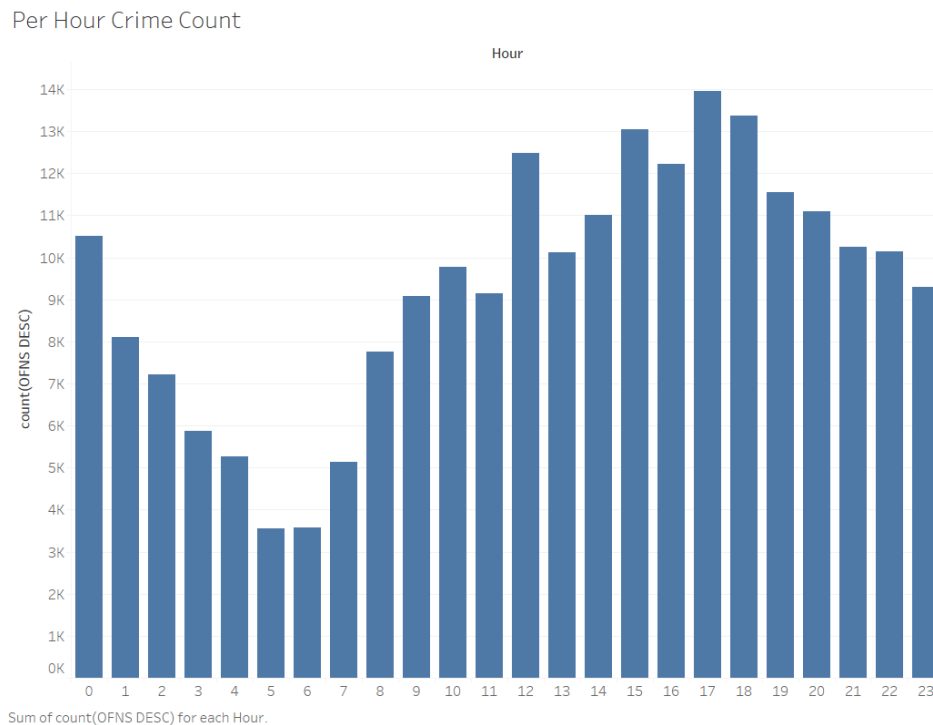


Figure 3 Average Crime Count every hour

We then tried to see that hour in the day has any co-relation with crime. The finding was interesting, we saw that crime was maximum around 06:00' clock in the evening when people return to their home.

### 3.4 Month Vs Crime

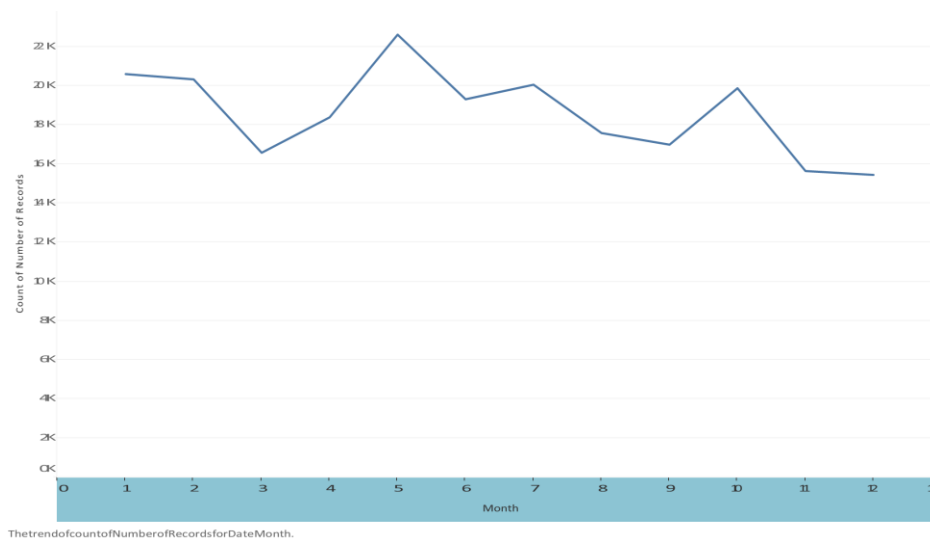
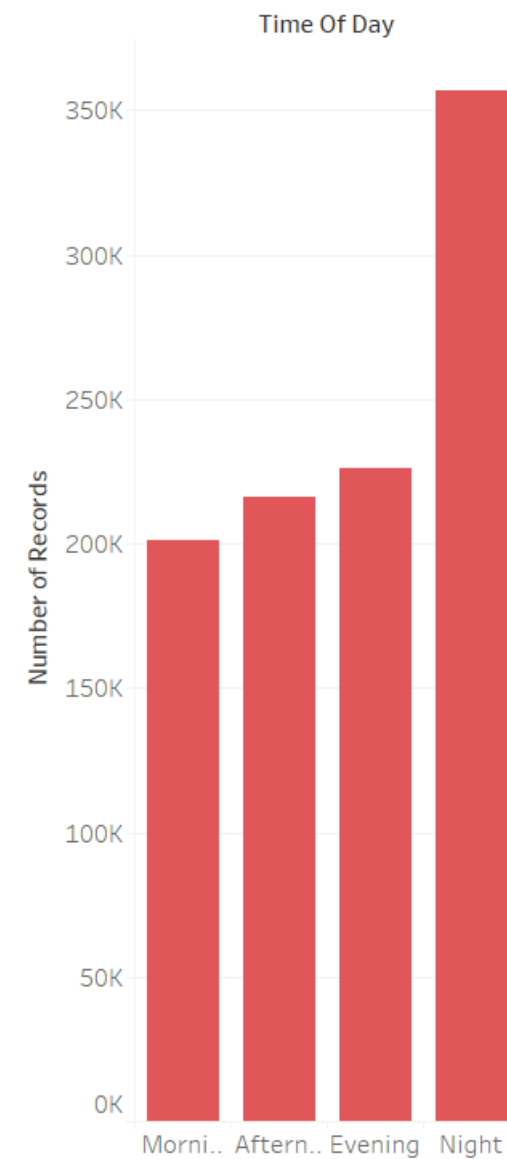


Figure 4 Month and Crime

We then tried to see if crime depends on the month of the year. It was interesting to see that the crime drastically dropped in month of December. This correlation was difficult to digest as the tourist come in this part of the year. May be the chilly weather stops the perpetrator.

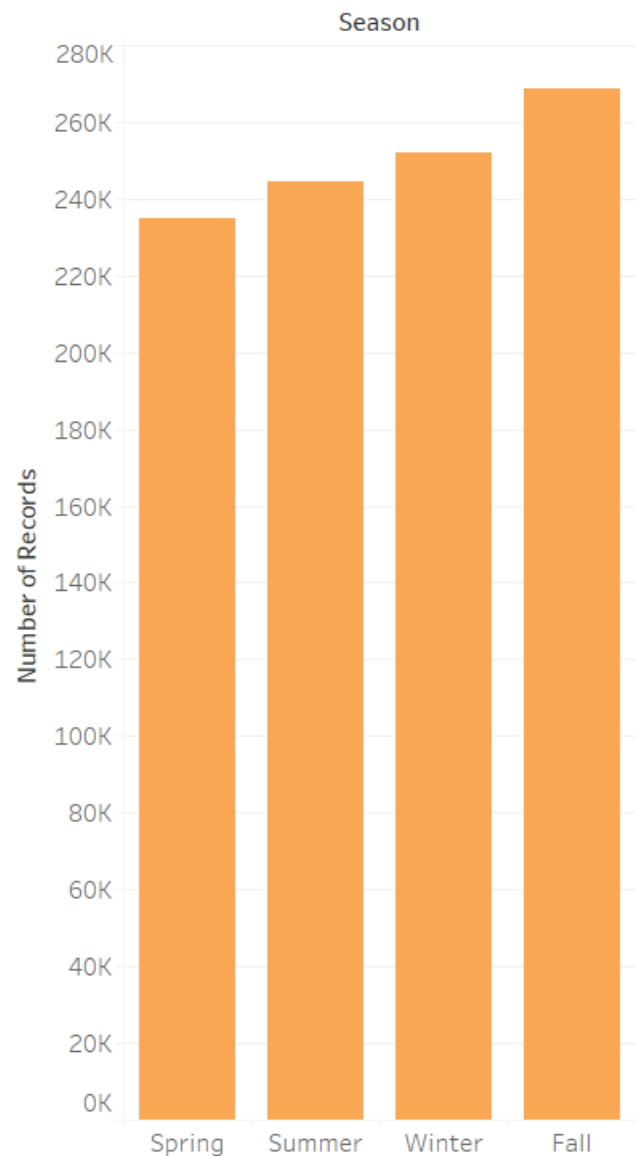
### 3.4 Crime at hour of day, season of the year

Time of the Day vs Count



Sum of Number of Records for each Time Of Day.

Seasonal Variation



Sum of Number of Records for each Season.

We then plotted the graph to see if time of day or season has impact on crime. The results were as expected, the crime was more at Night time and was lowest in morning.

The crime was also highest in fall, which was an interesting correlation.



## 4. Correlation of Crime with weather

We then gathered weathered data for NYC and imported it in spark. We got hourly weather data for last 10 years and it had 90 columns as shown below.

---

'STATION', 'STATION\_NAME', 'ELEVATION', 'LATITUDE', 'LONGITUDE', 'DATE', 'REPORTTYPE',  
'HOURLYSKYCONDITIONS', 'HOURLYVISIBILITY', 'HOURLYPRESENTWEATHERTYPE',  
'HOURLYDRYBULBTEMPF', 'HOURLYDRYBULBTEMPC', 'HOURLYWETBULBTEMPF',  
'HOURLYWETBULBTEMPC', 'HOURLYDewPointTempF', 'HOURLYDewPointTempC',  
'HOURLYRelativeHumidity', 'HOURLYWindSpeed', 'HOURLYWindDirection', 'HOURLYWindGustSpeed',  
'HOURLYStationPressure', 'HOURLYPressureTendency', 'HOURLYPressureChange',  
'HOURLYSeaLevelPressure', 'HOURLYPrecip', 'HOURLYAltimeterSetting',  
'DAILYMaximumDryBulbTemp', 'DAILYMinimumDryBulbTemp', 'DAILYAverageDryBulbTemp',  
'DAILYDeptFromNormalAverageTemp', 'DAILYAverageRelativeHumidity',  
'DAILYAverageDewPointTemp', 'DAILYAverageWetBulbTemp', 'DAILYHeatingDegreeDays',  
'DAILYCoolingDegreeDays', 'DAILYSunrise', 'DAILYSunset', 'DAILYWeather', 'DAILYPrecip',  
'DAILYSnowfall', 'DAILYSnowDepth', 'DAILYAverageStationPressure', 'DAILYAverageSeaLevelPressure',  
'DAILYAverageWindSpeed', 'DAILYPeakWindSpeed', 'PeakWindDirection',  
'DAILYSustainedWindSpeed', 'DAILYSustainedWindDirection', 'MonthlyMaximumTemp',  
'MonthlyMinimumTemp', 'MonthlyMeanTemp', 'MonthlyAverageRH', 'MonthlyDewpointTemp',  
'MonthlyWetBulbTemp', 'MonthlyAvgHeatingDegreeDays', 'MonthlyAvgCoolingDegreeDays',  
'MonthlyStationPressure', 'MonthlySeaLevelPressure', 'MonthlyAverageWindSpeed',  
'MonthlyTotalSnowfall', 'MonthlyDeptFromNormalMaximumTemp',  
'MonthlyDeptFromNormalMinimumTemp', 'MonthlyDeptFromNormalAverageTemp',  
'MonthlyDeptFromNormalPrecip', 'MonthlyTotalLiquidPrecip', 'MonthlyGreatestPrecip',  
'MonthlyGreatestPrecipDate', 'MonthlyGreatestSnowfall', 'MonthlyGreatestSnowfallDate',  
'MonthlyGreatestSnowDepth', 'MonthlyGreatestSnowDepthDate', 'MonthlyDaysWithGT90Temp',  
'MonthlyDaysWithLT32Temp', 'MonthlyDaysWithGT32Temp', 'MonthlyDaysWithLT0Temp',  
'MonthlyDaysWithGT001Precip', 'MonthlyDaysWithGT010Precip', 'MonthlyDaysWithGT1Snow',  
'MonthlyMaxSeaLevelPressureValue', 'MonthlyMaxSeaLevelPressureDate',  
'MonthlyMaxSeaLevelPressureTime', 'MonthlyMinSeaLevelPressureValue',  
'MonthlyMinSeaLevelPressureDate', 'MonthlyMinSeaLevelPressureTime',  
'MonthlyTotalHeatingDegreeDays', 'MonthlyTotalCoolingDegreeDays',  
'MonthlyDeptFromNormalHeatingDD', 'MonthlyDeptFromNormalCoolingDD',  
'MonthlyTotalSeasonToDateHeatingDD', 'MonthlyTotalSeasonToDateCoolingDD'

---

We were perplexed by the temperature value in this column and then finally decided to go with HourlyBulbTemperature to co-relate with crime. We normalized the dates and then cleaned the dataset,

Finally, we merged the crime and weather data set based on hour and day and gathered some insights as discussed below.

### 4.1 Crime Count vs Temperature in Degree Celsius

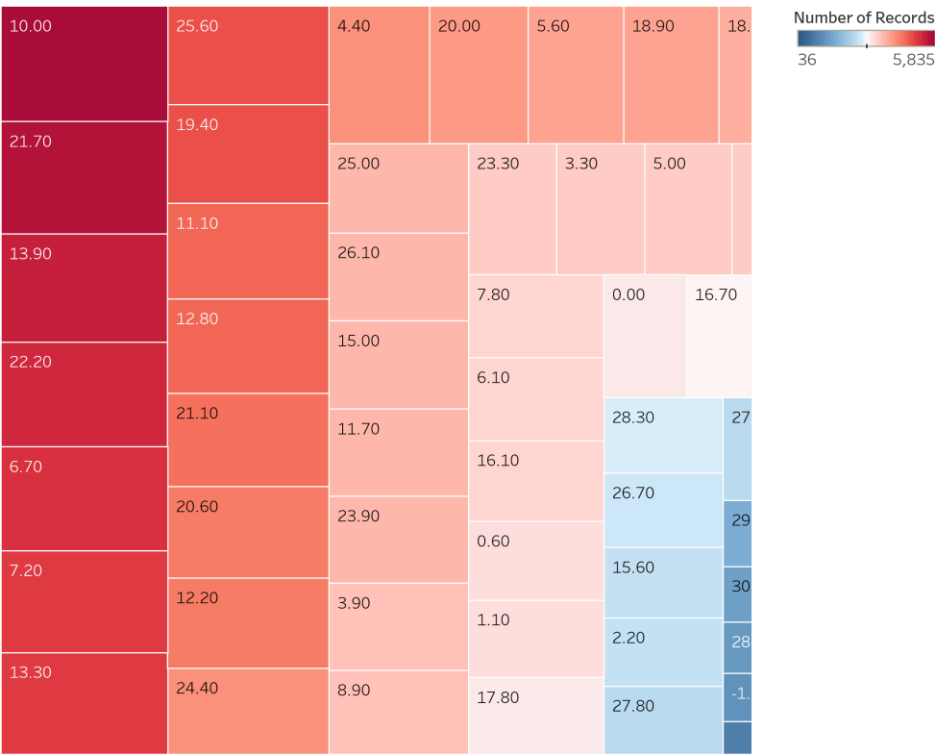


Figure 5 Temperature and Crime count

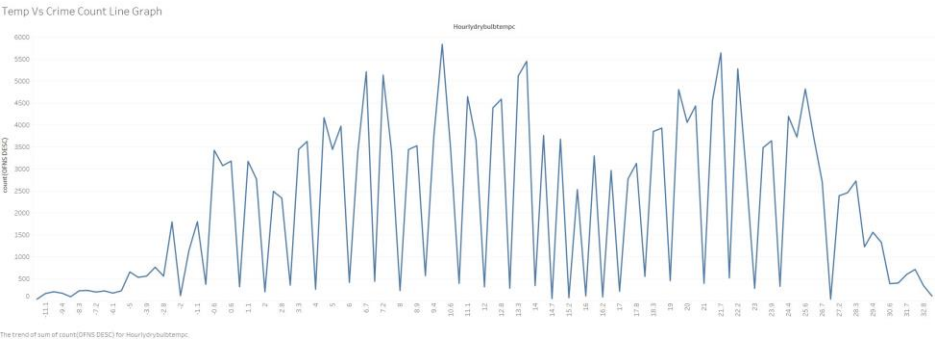


Figure 6 Temperature and Crime count for 2016

We started plotting crime count based on temperature and saw the crime was highest at 10 degrees and lowest at -1. This correlation was kind of expected, as we expect perpetrator to commit less crime on a very chilly day.

Temp Range vs Count

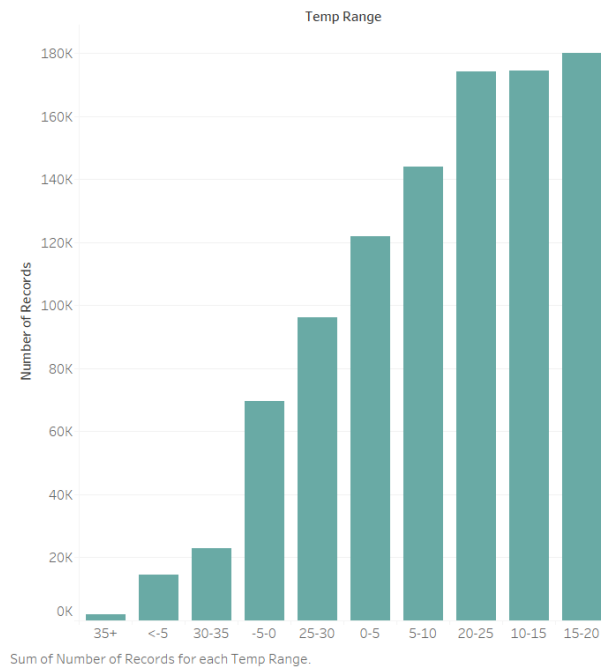


Figure 7 Range of Temperature and Crime

Temp Type vs Count

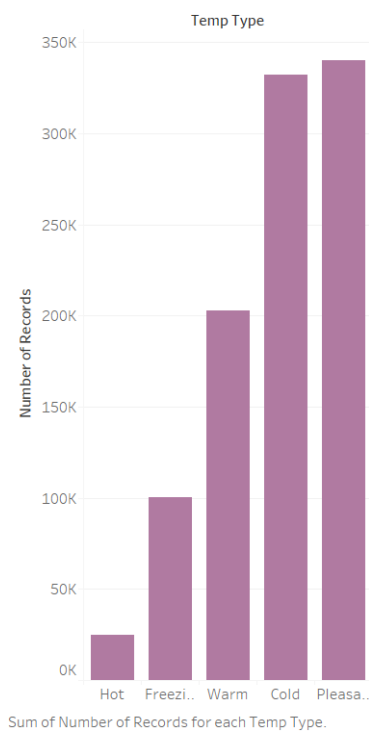
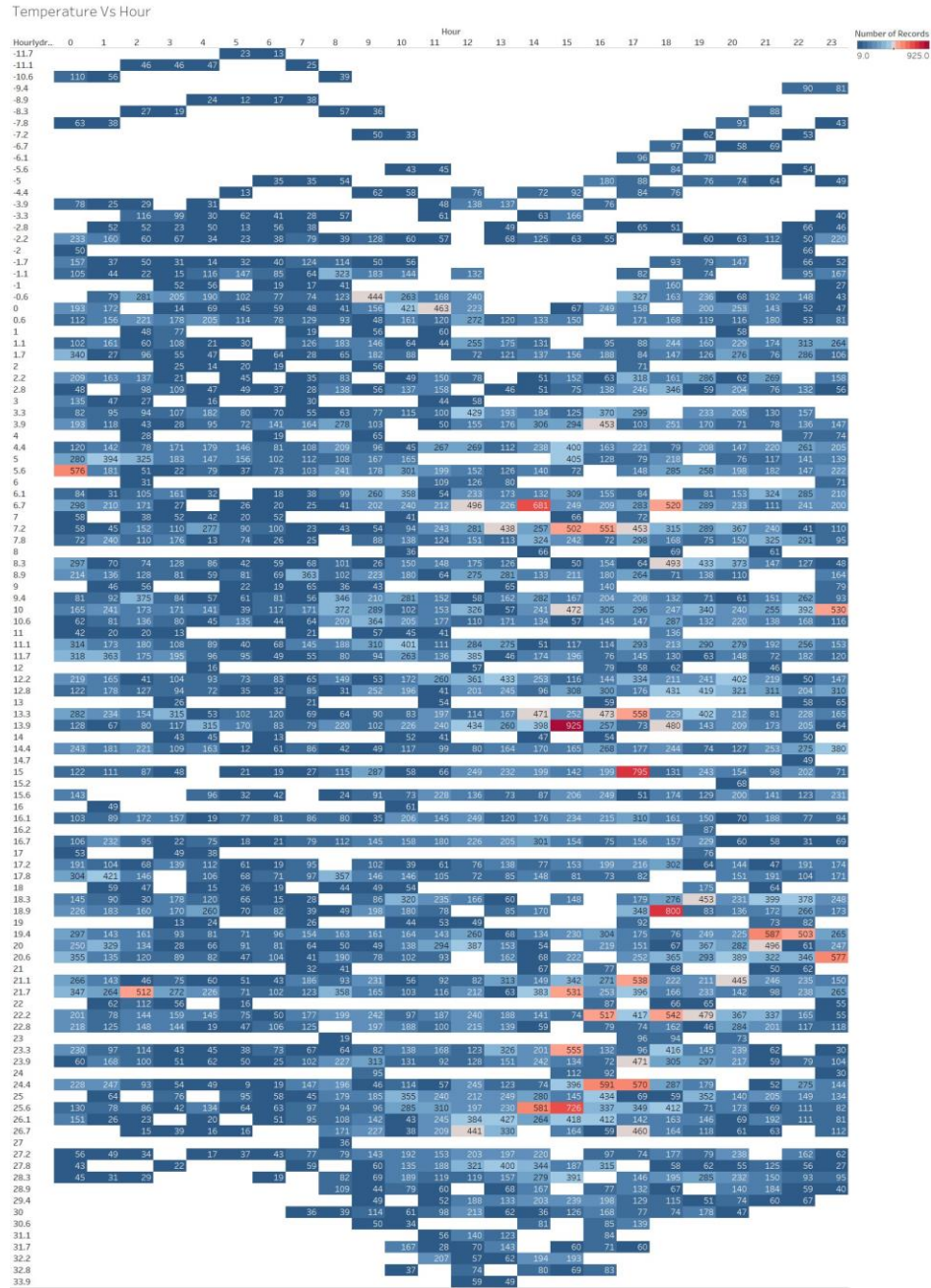


Figure 8 Temperature type and Crime

We visualized the crime was maximum in pleasant temperature and was minimum In hot temperature.

## 4.2 Number of Crimes Vs Temperature vs Hour of the Day



We then plotted a graph from temperature and time of the day to see what is the most favorable condition and find out that if it is 03:00pm and it is 20 degrees Celsius the chances of you being a victim are very high. Which was an interesting finding.

## 5. Correlation of Crime with population and Median Income

We also download the population and median income dataset and cleaned it. We then merged the census data based on zip code and started analyzing crime based on population and wealth.

### 5.1 Crime in different areas of NYC

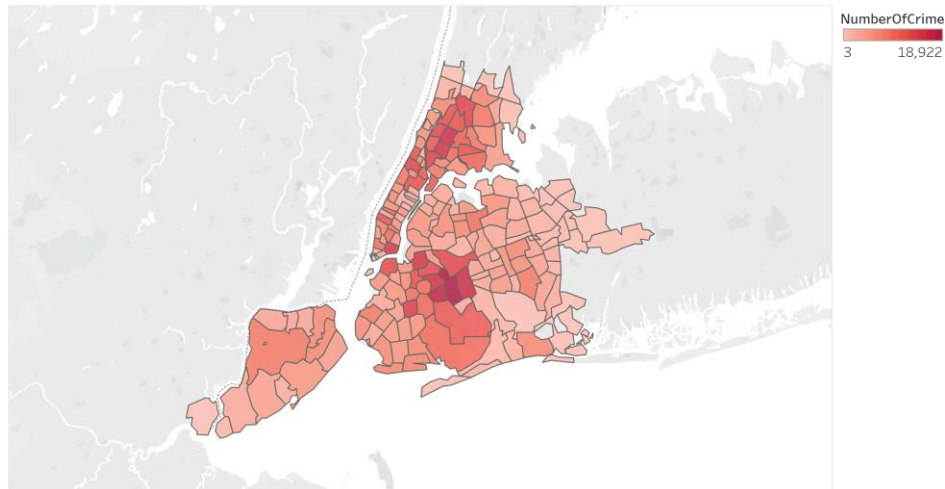
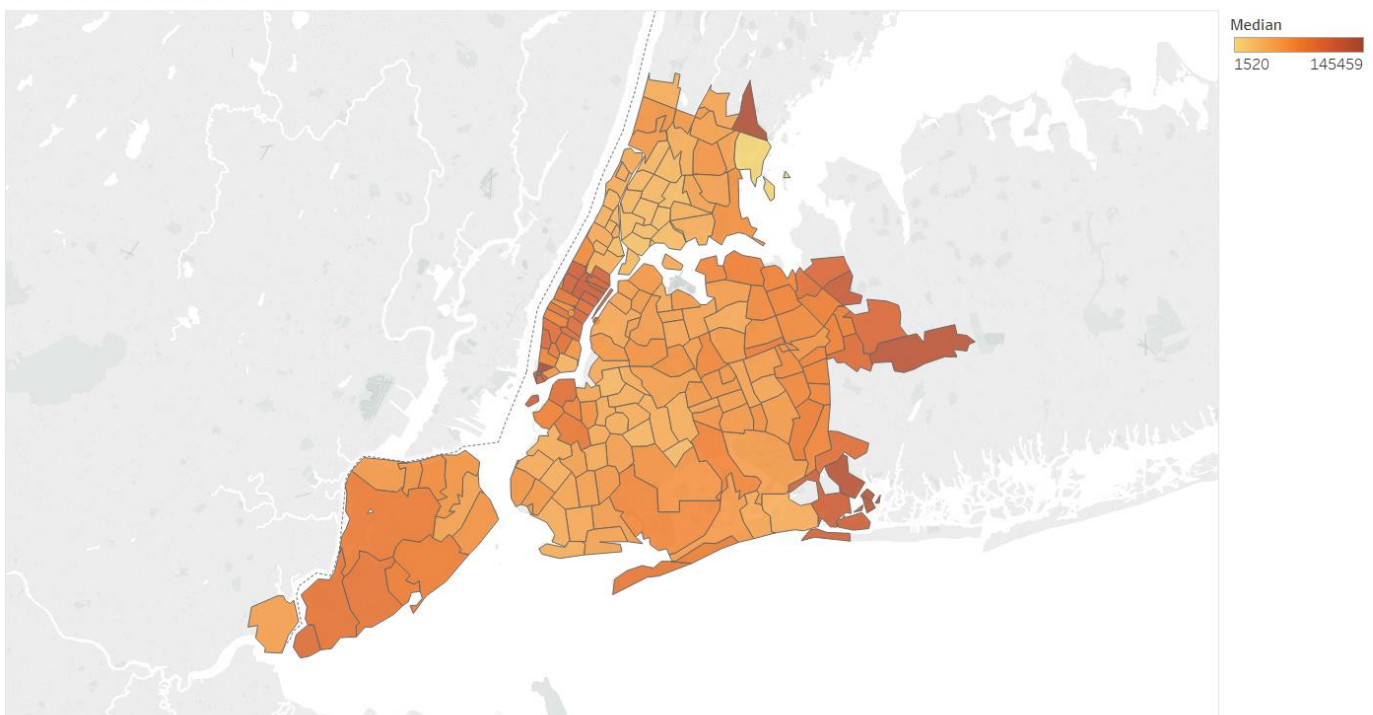


Figure 9 Map based on Longitude (generated) and Latitude (generated). Color shows sum of NumberOfCrime. Details are shown for Zipcode.

We visualized that crime was more in some parts of Queens and Upper Manhattan. We were interested to see what is the population in that part and what is the median income there.

### 5.2 Median Income in different areas of NYC

Wealth Distribution Map



Map based on Longitude (generated) and Latitude (generated). Color shows details about Median. Details are shown for Zip.

The results showed that where the crime was more the income was less. This was kind of expected.

### 5.3 Population in different areas of NYC

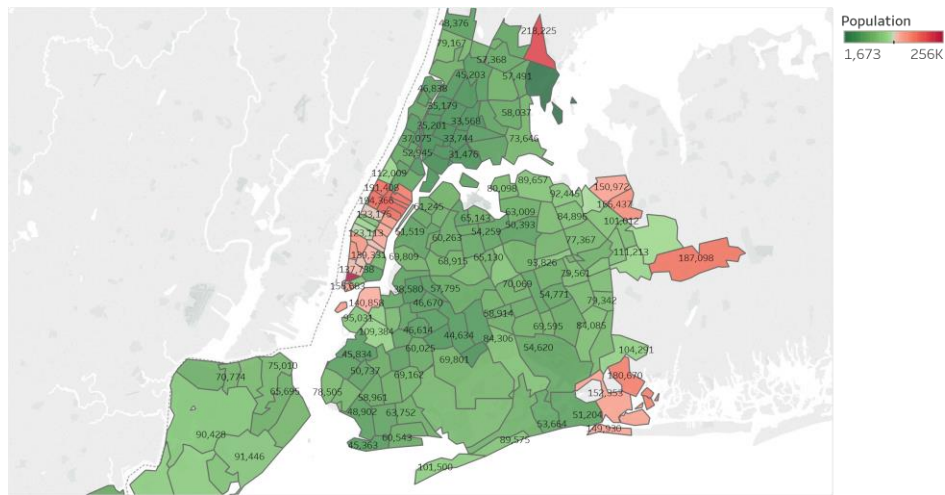


Figure 10 Map based on Longitude (generated) and Latitude (generated). Color shows sum of Population. The marks are labeled by sum of Population. Details are shown for Zipcode.

The population had a minor impact on crime. The conclusion we derived was that with low population the crime gradually increases which can be seen from the above map.

### 5.4 Mean Income Vs Crime



We then plotted graph to visualize the crime count and median income. Crime is clearly inversely proportional to income



## 6. Correlation of Crime with Markets

We then downloaded the stock market data for last 10 years. We then cleaned the dataset and normalized the date of the data and merged it with crime data.

### 6.1 Decrease in Crime with Increase in Dow Jones Index

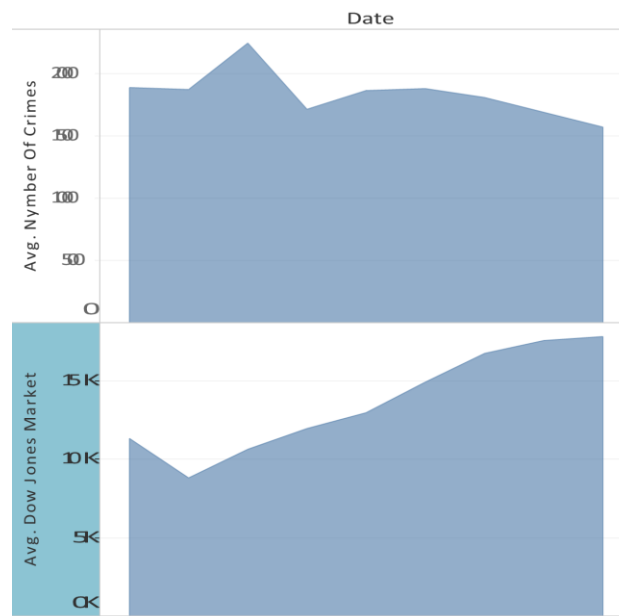


Figure 11 Average of Count and average of Value for each Date Year.

We saw that as the Dow Jones index increased the crime has decreased. But this can be a coincidence as well, It is a possibility that crime has decreased over year due to strict law enforcement.

### 6.2 Impact of 2008 recession on crime

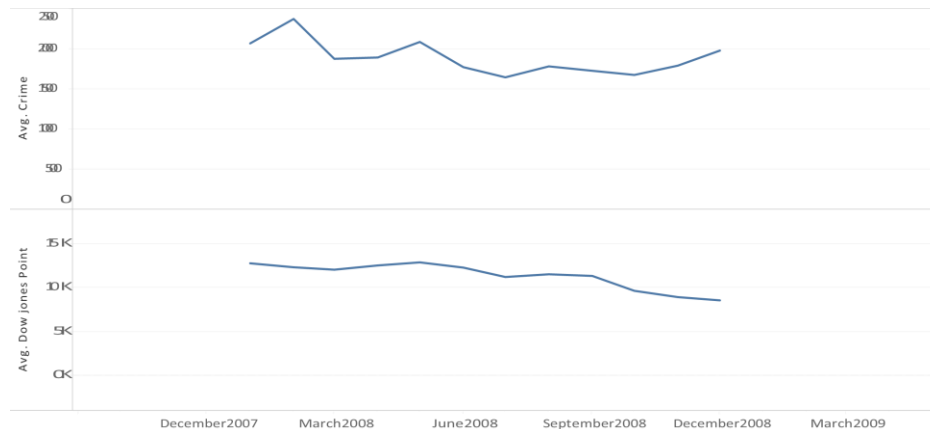


Figure 12 The trends of average of Count and average of Value for Date Month. The data is filtered on Date Year, which keeps 2008.

We then wondered that when the market crashed in 2008 did it impact crime? The answer to this is visible in the graph above. We can clearly see that there was an increase in crime when the market crashed. So, there is for sure some correlation between stock market and crime.

## 7. Recommendation based on Crime

When a person travels to a new city his biggest concern is wherever he is booking the hotel, the area should be safe. We downloaded the trip advisor dataset and filtered hotels of New York and then extracted zip code of each hotel and merged the hotel dataset with crime dataset based on zip code. We then visualized few hotels and crime near them.

### 7.1 Hotels in NYC

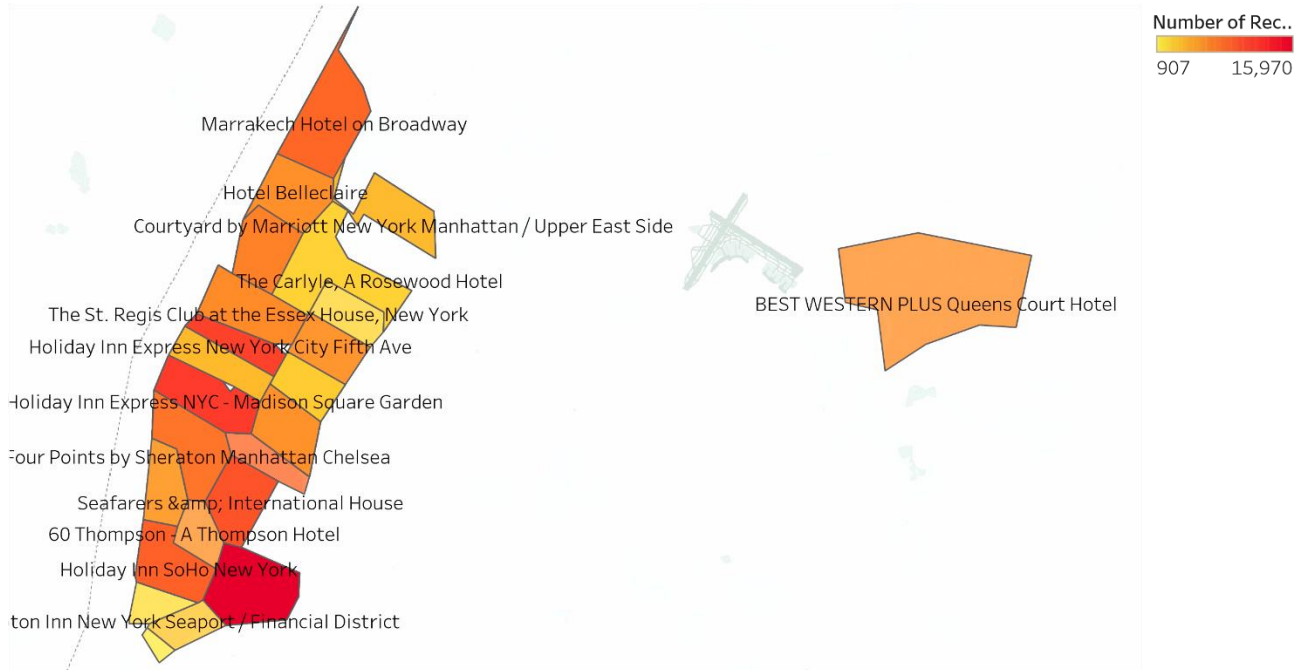


Figure 13 Map based on Longitude (generated) and Latitude (generated).

The above map gives some valuable insights about the hotels one should book based on location.

We also came up with top 10 and bottom 10 hotels based on location

Hotel Name	Number of reported Incidents near them
Club Quarters Downtown	1981
Millenium Hilton	4004
254 East Vacation	4636
Loews Regency Hotel	4636
The Carlyle, A Rosewood Hotel	4636
The Helmsley Carlton House	4636
The Bentley Hotel	5018
Dylan Hotel	5372
Helmsley Middletowne Hotel	5372
Hilton Manhattan East	5372
New York Marriott East Side	5372
Roger Smith Hotel	5372
The Alex Hotel	5372
The New York Helmsley	5372
The Roosevelt Hotel	5372



### Safe hotels in NYC

Hotels	Number of Incidents Near them
Chelsea Star Hotel	24539
Comfort Inn Manhattan	24539
Herald Square Hotel	24539
Holiday Inn Express NYC - Madison Square Garden	24539
Hotel Pennsylvania New York	24539
Hotel Stanford	24539
La Quinta Inn Manhattan	24539
Radisson Martinique on Broadway	24539
Wingate by Wyndham Manhattan Midtown	24539
nyma, the New York Manhattan Hotel	24539
Hotel on Rivington	33605
Off Soho Suites	33605
Windsor Hotel	33605

### Not so Safe hotels in NYC

## 7.2 Bars in NYC

We then downloaded the bar license dataset, cleaned it and merged it with crime data set to come out with list of safe and not so safe bars.

Bars	Crime Count
111 AGAM CORPORATION, 10704	1
1132 CONVENIENCE CORP, 10704	1
1160 YONKERS AVE FOOD CORP, 10704	1
115 WOLFS LANE RESTAURANT CORP, 10803	6
2016 NEW HOPE GROCERY CORP, 10705	19
2020 YONKERS DELI GROCERY CORP, 10705	19
208 SARATOGA FOOD CORP, 10705	19
108 COVERT AVENUE PUB INC, 11530	93
2048 HILLSIDE RESTAURANT CORP, 11040	188
205 SERVICE STATION INC, 11040	188
2201 GAS CORP, 11040	188

### Safe bars in NYC

Bars	Crime Count
230 JC DELI GROCERY INC, 10457	36327
1155 NEW WAY DELI CORP, 11208	36558
1176 LIBERTY FOOD CORP, 11208	36558
1290 SUTTER GROCERY INC, 11208	36558
160 CRESCENT DELI & GROCERY CORP, 11208	36558
208 CRESCENT MEAT & PRODUCE INC, 11208	36558
2552 DELI CORP, 11208	36558
255 E 95 ST FOOD CORP, 11212	38172
1115 PENNSYLVANIA MEAT CORP, 11207	42442
200 JAMAICA FOOD CORP, 11207	42442
2053 PITKIN MINIMARKET INC, 11207	42442
106 AMSTERDAM REST CORP, 10029	55808

Not so safe bars in NYC

## 8. Technology Used

- Apache Zeppelin
- Apache Spark
- Scala
- Python
- Tableau
- Databricks cloud platform
- Google Reverse geocoding API

## 9. Conclusion

As we write this report there is a crime happening somewhere in NYC. The crimes in New York is mundane. This project gave us a good insight on how crime occurs in different areas of NYC. These insights can be used to control the crime. We have good insights how crime varies in month and the location which are impacted the most. If these insights are considered by NYPD we can work together for a better New York. Crime rate in a locality can become one factor for people to decide which hotel to stay or which bars to visit.