NYPD Shooting Incident Project

Project: NYPD Shooting Incident Project
Author: Larry Aruna
Date: 2024-06-05

1) Introduction:

As a Data scientist, am interested in analyzing list of every shooting that occurred in NYC going back 2006 through end of the previous calendar year. Also interested in visualization reports and determining which factors are statistical significant associated and explaining shooting resulted in the victim's death which would be counted as a murder (Statistical Murder Flag).

This data was manually extracted every quarter and review by the office of management and planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location, victim demographics and time of occurrence.

2) Method:

The NYPD Data contains.

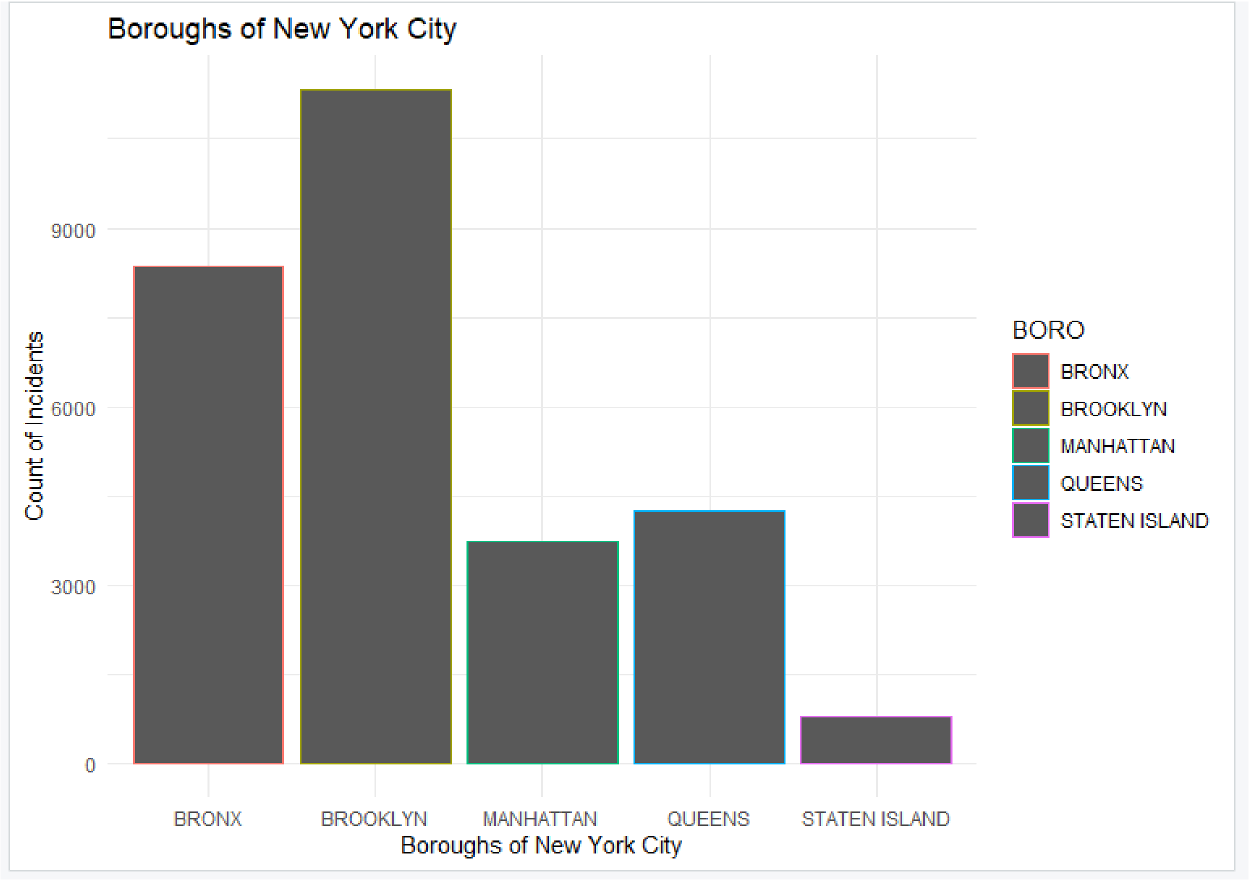
- Incident Key, Occurrence Date, Occurrence Time, Borough where the shooting incident occurred.
- Precinct where the shooting incident occurred, Jurisdiction code, location description, Statistical Murder Flag, Victim's age within a category, Victim's sex description, Victim's race description, Latitude and Longitude.

3) Statistical Approach: This project involved Data mining Pipeline and the statistical approach on the studies:
   - Data mining Pipeline: Data Knowledge Application and Technique
   - Data understanding, Data Preprocessing, Data Warehousing, Data Modeling and Pattern Evaluation

Data Understanding: Using tidy and transformation and then Summary of the NYPD Data.

```
INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                   BORO
 Length:28503       Length:28503       Length:28503       BRONX        : 8363
 Class :character   Class :character   Class :character   BROOKLYN     :11331
 Mode  :character   Mode  :character   Mode  :character   MANHATTAN    : 3746
                                                          QUEENS       : 4263
                                                          STATEN ISLAND:  800

STATISTICAL_MURDER_FLAG
false:22981
true : 5522


 VIC_AGE_GROUP     X_COORD_CD        Y_COORD_CD     VIC_SEX
 <18    : 2946   Min.   : 914928   Min.   :125757   F: 2753
 1022   :    1   1st Qu.:1000068   1st Qu.:182905   M:25738
 18-24  :10363   Median :1007776   Median :194872   U:   12
 25-44  :12946   Mean   :1009437   Mean   :208375
 45-64  : 1978   3rd Qu.:1016807   3rd Qu.:239814
 65+    :  205   Max.   :1066815   Max.   :271128
 UNKNOWN:   64

VIC_RACE
AMERICAN INDIAN/ALASKAN NATIVE:   11
ASIAN / PACIFIC ISLANDER      :  440
BLACK                         :20202
BLACK HISPANIC                : 2787
UNKNOWN                       :   70
WHITE                         :  728
WHITE HISPANIC                : 4265


    Latitude       Longitude
 Min.   :40.51   Min.   :-74.25
 1st Qu.:40.67   1st Qu.:-73.94
 Median :40.70   Median :-73.92
 Mean   :40.74   Mean   :-73.91
 3rd Qu.:40.82   3rd Qu.:-73.88
 Max.   :40.91   Max.   :-73.70
```

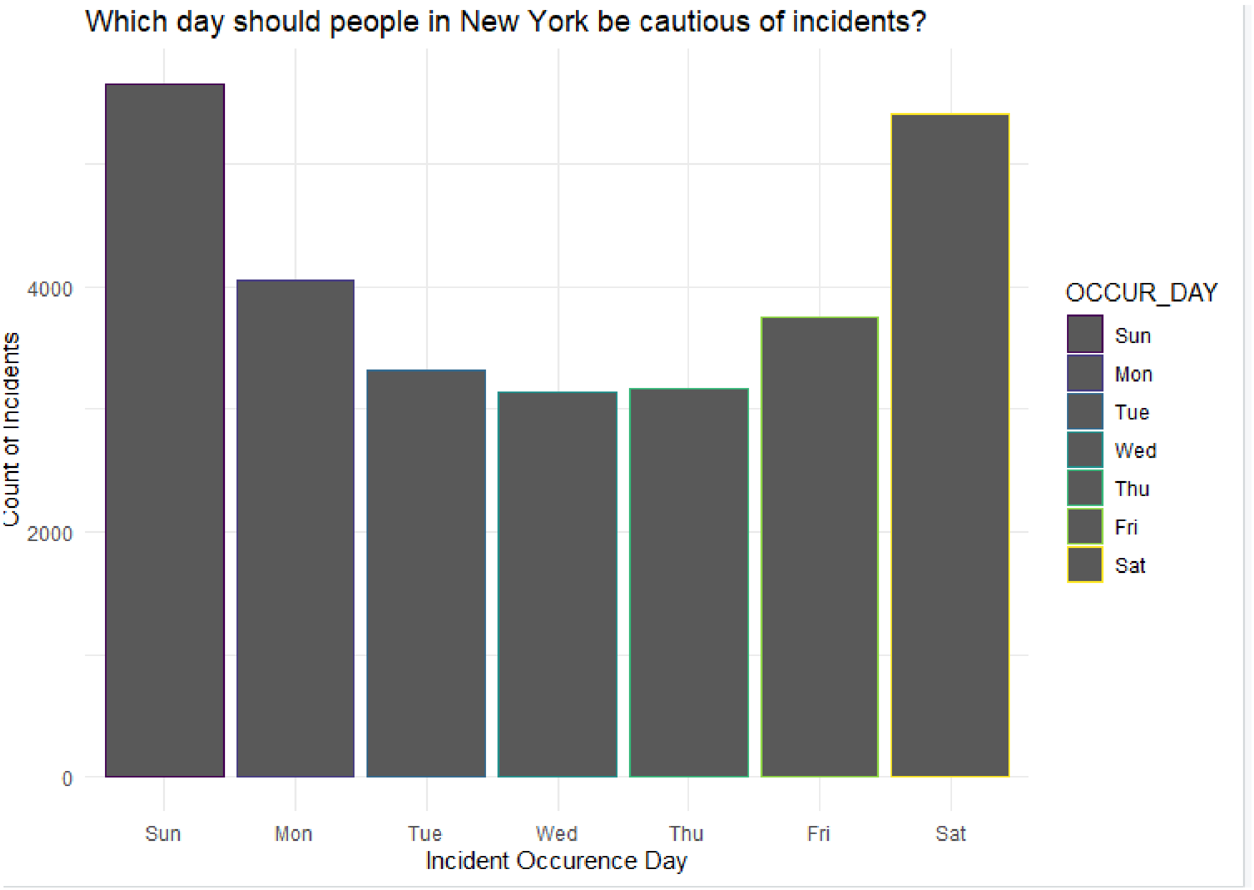
Visualization: Question1. Visualize which part of New York has the most number of incident



Question2 : Group the data based on shooting resulted in the victim's death which would be counted as a murder

```
               False   True
BRONX            6729   1634
BROOKLYN         9124   2207
MANHATTAN        3074   672
QUEENS           3423   840
STATEN ISLAND     631   169
```

Question3 : Which day should people in New York be Cautious of incidents ?
Sunday and Saturday turn out to be the highest crime days.



Visualize the Data based on Victim’s Age group

```
Age :    <18     1022    18-24    25-44    45-64      65+      UNKNOWN
Total:    2946         1    10363    12946     1978      205       64
```

Data Warehousing and Modeling / Pattern Evaluation

Building the Logistics Regression. The process of this model is to determine which factors are contributing to shooting resulted in the victim's death which would be counted as a murder. With below model summary report. Only variable Victim's Age group has a small p-value less than alpha 0.05. That means the any changes in Age group value will change the value in shooting resulted in the victim's death which would be counted as a murder (Statistical Murder Flag)

```
Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        2.407503  99.359858   0.024  0.98067    
VIC_RACEASIAN / PACIFIC ISLANDER  11.316996  97.405627   0.116  0.90751    
VIC_RACEBLACK                     11.061930  97.405565   0.114  0.90958    
VIC_RACEBLACK HISPANIC            10.900010  97.405577   0.112  0.91090    
VIC_RACEUNKNOWN                   10.231722  97.406476   0.105  0.91634    
VIC_RACEWHITE                     11.378755  97.405602   0.117  0.90700    
VIC_RACEWHITE HISPANIC            11.178394  97.405571   0.115  0.90863    
VIC_SEXM                          -0.034850   0.050866  -0.685  0.49326    
VIC_SEXU                          -0.576213   1.080844  -0.533  0.59395    
VIC_AGE_GROUP1022                -10.581589 324.743703  -0.033  0.97401    
VIC_AGE_GROUP18-24                 0.291893   0.061052   4.781 1.74e-06 ***
VIC_AGE_GROUP25-44                 0.624751   0.059079  10.575  < 2e-16 ***
VIC_AGE_GROUP45-64                 0.759025   0.076093   9.975  < 2e-16 ***
VIC_AGE_GROUP65+                   1.072494   0.160455   6.684 2.32e-11 ***
VIC_AGE_GROUPUNKNOWN               0.861718   0.316691   2.721  0.00651 ** 
OCCUR_HOUR                         0.001301   0.001846   0.705  0.48077    
OCCUR_DAY.L                       -0.019099   0.037233  -0.513  0.60798    
OCCUR_DAY.Q                       -0.068395   0.039874  -1.715  0.08630 .  
OCCUR_DAY.C                       -0.050970   0.040213  -1.267  0.20498    
OCCUR_DAY^4                        0.002532   0.040888   0.062  0.95062    
OCCUR_DAY^5                        0.024403   0.042997   0.568  0.57034    
OCCUR_DAY^6                       -0.098777   0.044135  -2.238  0.02522 *  
Latitude                          -0.051490   0.181480  -0.284  0.77662    
Longitude                          0.179415   0.230919   0.777  0.43718    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28024  on 28502  degrees of freedom
Residual deviance: 27721  on 28479  degrees of freedom
AIC: 27769

Number of Fisher Scoring iterations: 11
```

Evaluate the mode and fit it based on Age group predictor column.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -1.90090    0.05478 -34.699  < 2e-16 ***
VIC_AGE_GROUP1022   -7.66506   72.46288  -0.106   0.9158    
VIC_AGE_GROUP18-24   0.29273    0.06079   4.816 1.47e-06 ***
VIC_AGE_GROUP25-44   0.63022    0.05876  10.726  < 2e-16 ***
```

```
VIC_AGE_GROUP45-64    0.79553    0.07554  10.531  < 2e-16 ***
VIC_AGE_GROUP65+      1.15608    0.15921   7.261 3.83e-13 ***
VIC_AGE_GROUPUNKNOWN  0.71713    0.30013   2.389   0.0169 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28024  on 28502  degrees of freedom
Residual deviance: 27782  on 28496  degrees of freedom
AIC: 27796
```

Identify Bias

The first thinking about this topic project without review and analyze the data was based on discrimination and hate crime and social media contribute to bias as well. By looking at some specific interested columns in my analyzing with column city where crime occurred, city like Bronx, Brooklyn, Manhattan, Queens and it turns out Brooklyn has a highest hate crime. Also looking at the specific days column (Sun, Mon, Tue, Wed, Thursday, Friday and Saturday) and it turns out Sunday and Saturday are the highest days for hate crimes. All these contribute to bias in NYPD Shooting incident. How do I handle a missing value and fit a model on Statistical murder flag also contribute to a bias.

Conclusion

Based on the Analyzing of the NYPD data and the visualization reports. The model suggested that Age Group variable which has P-vale 0,005 less than alpha is significantly associated with dependence Statistical Murder flag variable response. That suggests, if any changes in Age group value will change the Statistical Murder flag. Based on the summary report, is good to evaluate the model with either Backward or forward selection on variables predictors. Do more analyses on model interactions.