NYPD - Data Analysis

David Forero Botia

2024-02-03

Importing the Data

To import the data we will use the following code that download the data directly from the official page, this allowing that anyone with the Rmarkdown can download the data.

```
nypd_data<-read.csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')
```

The data look like this:

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class
:character
##  Median : 90372218   Mode  :character   Mode  :character   Mode
:character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Cleaning the Data

The data will be clean with these steps:

1. Changing the corresponding columns to its correct data type.

```
nypd_data$OCCUR_DATE<-as.Date(nypd_data$OCCUR_DATE, , format = "%m/%d/%Y")
nypd_data$OCCUR_TIME <- as_hms(nypd_data$OCCUR_TIME)
nypd_data$STATISTICAL_MURDER_FLAG<-
as.logical(nypd_data$STATISTICAL_MURDER_FLAG)
```

2. Deleting the columns that I will not use for the analysis

```
nypd_data<- subset(nypd_data,
select=c('OCCUR_DATE','OCCUR_TIME','LOCATION_DESC','STATISTICAL_MURDER_FLAG',
'PERP_AGE_GROUP','PERP_SEX','PERP_RACE','VIC_AGE_GROUP','VIC_SEX','VIC_RACE')
)
```

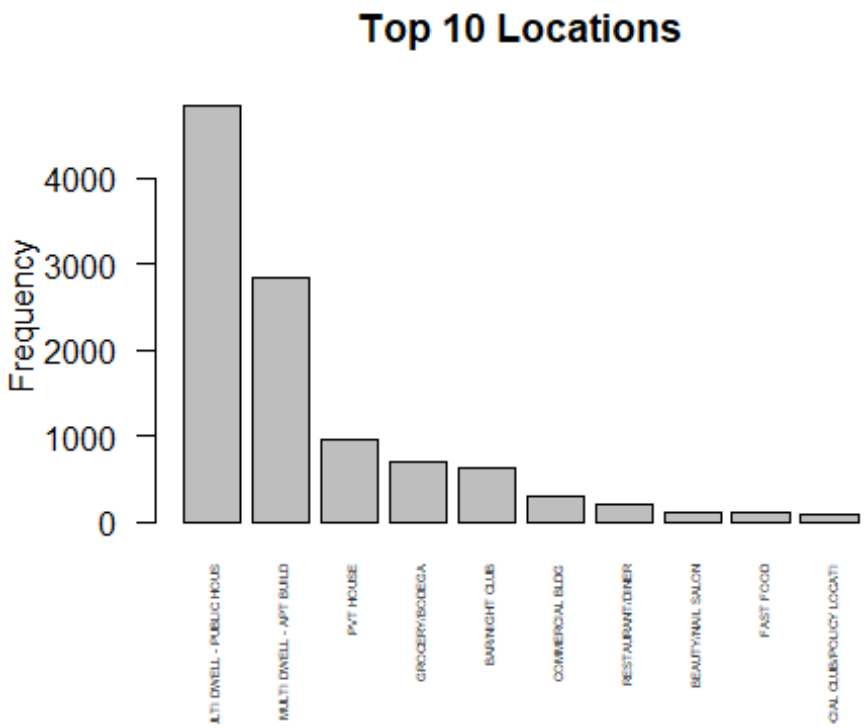
The data at the end will look like this:

```
##    OCCUR_DATE          OCCUR_TIME       LOCATION_DESC
##  Min.   :2006-01-01   Length:27312      Length:27312
##  1st Qu.:2009-07-18   Class1:hms        Class :character
##  Median :2013-04-29   Class2:difftime   Mode  :character
##  Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:2018-10-15
##  Max.   :2022-12-31
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:27312       Length:27312
##  FALSE:22046             Class :character   Class :character
##  TRUE :5266              Mode  :character   Mode  :character
##
##
```

```
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

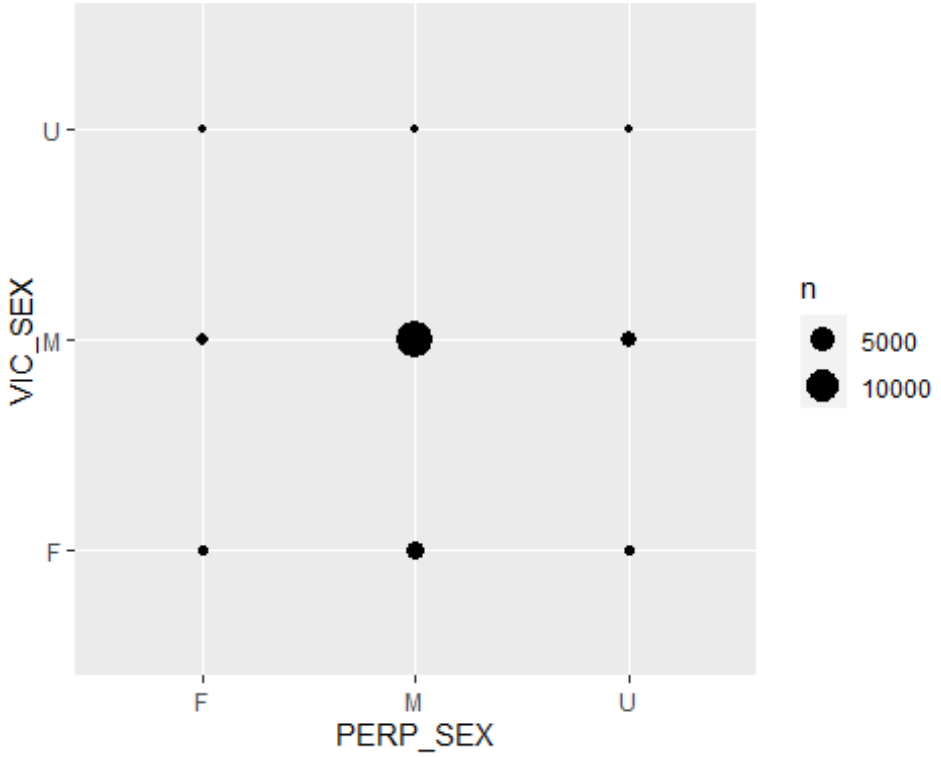
Analyzing the Data

We can see the top 10 locations where have been any incident:



Here we can see that the most frequent places for these incidents are in houses or apartments, and less than half are elsewhere.

Also, below we show the sex of the perpetrator against the sex of the victim:



Here we can see that the most common perpetrators are Men and these attacks most of the time are to other men, also we can see that men have more attacks on women than women against both sexes.

Conclusion

This data has been really useful in showing some historical data related to incidents, where we dive into different patterns that have been observed, and that we can quantify them, for example, the most usual locations for an incident to occur that we notice are the Homes, or the relationship between the perpetrator and victim genders. Although this was an exercise where we can make conclusions of the past we can not predict the future based on this historical data.

Bias

We have to look very carefully for bias in this dataset, specifically because it shows data with so much weight in our day-to-day. For example, we can not conclude that a man has a higher profile as a perpetrator, instead, we should think that this data can not predict the future type of perpetrator and is just a form to show historical data.

Another example is the type of race with more incidents related, this doesn't mean that this race has a higher probability to do a murder, instead, we will need to have more sociopolitical data that allows us to get to the bottom of the issue, and find the real root of this problematic, something that we can not display with the amount of data we have available right now.