NYPD Shooting

NYPD Shooting Incident History

The goal of this study is to find what factors could be contributing to shootings in New York. We will look at factors such as victim traits, location, day of the week, and time of day. We will then discuss other factors not seen in this data set and how those could be impacting our data as an outside influence. Finally, we will discuss the impact that those outside factors and biases have on studies.

Import Packages

```
library(tidyverse)
library(lubridate) #change data to date object
library(ggplot2)   # to plot data
library(mice)      # to visualize missing data
library(dplyr)
set.seed(550)      #set seed for consistency
```

Import CSV files and view

First we need to import the dataset as a csv file.

```
raw_data <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

Now let's look at the summary of the data so we get an idea of what we have to work with.

```
summary(raw_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class1:hms         Class :character
##  Median : 83365370   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :102218616                      Mode  :numeric
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  Mean   : 66.21   Mean   :0.3323
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
```

```
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:23568       Length:23568       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

We can see many different variables with different types. We will improve that later.

Now let's look at the first few rows of data. This is another way to visualize the data before starting our cleaning and analysis.

```
head(raw_data)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>            <dbl>              <dbl>
## 1    201575314 08/23/2019 22:10      QUEENS             103                  0
## 2    205748546 11/27/2019 15:54      BRONX               40                  0
## 3    193118596 02/02/2019 19:40      MANHATTAN           23                  0
## 4    204192600 10/24/2019 00:52      STATEN ISLAND      121                  0
## 5    201483468 08/22/2019 18:03      BRONX               46                  0
## 6    198255460 06/07/2019 17:50      BROOKLYN            73                  0
## # ... with 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

Cleaning and Transforming Data

Now that we have an idea of what we are working with, first we want to find missing values and deal with them.

We will use the 'mice' package to find the missing data quickly. After running each column in the console, we find that 5 columns have missing data. Now we must figure out how to deal with the missing data.

1. JURISDICITION is only missing two rows so we will replace these values with the most common jurisdiction (either 0, 1, or 2).
2. LOCATION_DESC is missing the most values which is also almost 10% of the data. Therefore we will just delete this column as to not skew the data unnecessarily.
3. PERP_AGE_GROUP, PERP_SEX and PERP_RACE are each missing more than 35% of the data. Therefore, we will also delete these columns too.
4. The INCIDENT_KEY is not important for our analysis because it is just a way to number the incidents. This is important for police files but not for data analysis. So we delete that column too.

```
shooting_data$JURISDICTION_CODE[is.na(shooting_data$JURISDICTION_CODE)] <- names(which.max(table(shooti
shooting_data$LOCATION_DESC <- NULL
shooting_data$PERP_AGE_GROUP <- NULL
shooting_data$PERP_RACE <- NULL
shooting_data$PERP_SEX <- NULL
shooting_data$INCIDENT_KEY <- NULL
```

Now we see that we have two different ways of identifying where the incident took place, by using either XY coordinates or Latitude-Longitude coordinates. Since we are most familiar with Lat-Long globally (most people do not know the XY coordinates for NYC), we will delete the X_COORD_CD and Y_COORD_CD.

```
shooting_data$X_COORD_CD <- NULL
shooting_data$Y_COORD_CD <- NULL
```

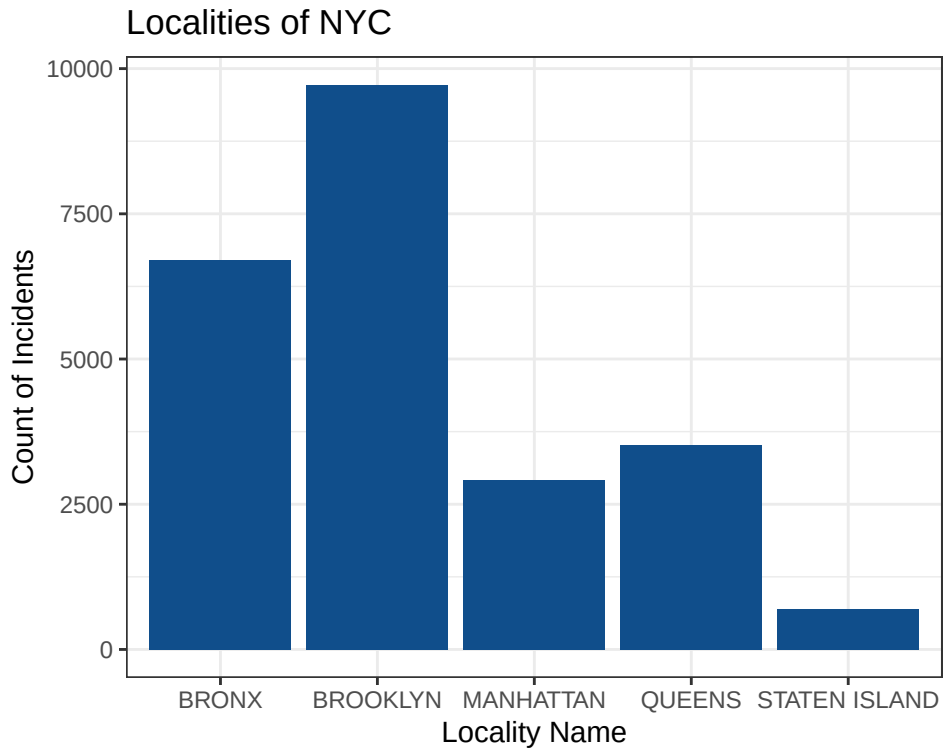
We want to change the column names to make it easier for us to read and use. Then, let's have one more look at the data before we start to visualize and analyze it.

```
shooting_data <- rename(shooting_data, Dates=OCCUR_DATE, Time=OCCUR_TIME, Locality=BORO, Precinct=PRECI
summary(shooting_data)
```

```
##     Dates              Time            Locality           Precinct
##  Length:23568       Length:23568      Length:23568       Min.   :  1.00
##  Class :character   Class1:hms        Class :character   1st Qu.: 44.00
##  Mode  :character   Class2:difftime   Mode  :character   Median : 69.00
##                     Mode  :numeric                      Mean   : 66.21
##                                                         3rd Qu.: 81.00
##                                                         Max.   :123.00
##    Jur.Code         Stat.Flag        Victim.Age         Victim.Sex
##  Length:23568       Mode :logical   Length:23568       Length:23568
##  Class :character   FALSE:19080     Class :character   Class :character
##  Mode  :character   TRUE :4488      Mode  :character   Mode  :character
##
##
##
##  Victim.Race            Lat             Long          Lon_Lat
##  Length:23568       Min.   :40.51   Min.   :-74.25   Length:23568
##  Class :character   1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Mode  :character   Median :40.70   Median :-73.92   Mode  :character
##                     Mean   :40.74   Mean   :-73.91
##                     3rd Qu.:40.82   3rd Qu.:-73.88
##                     Max.   :40.91   Max.   :-73.70
```

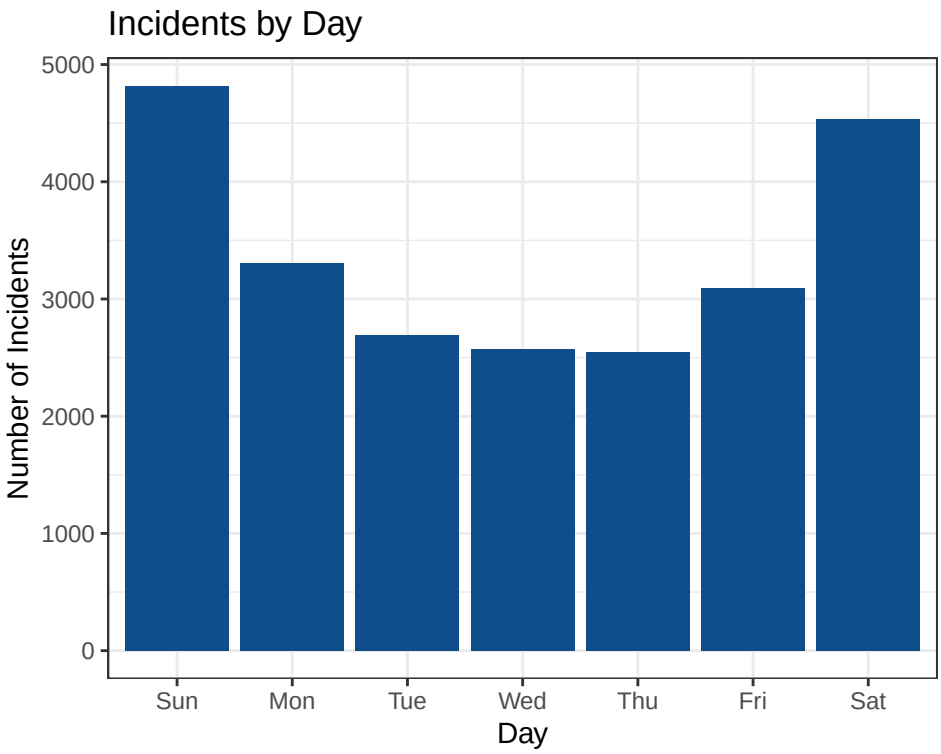
Visualization and Analysis

First we will look at the most active localities (boroughs/locations) in NYC.

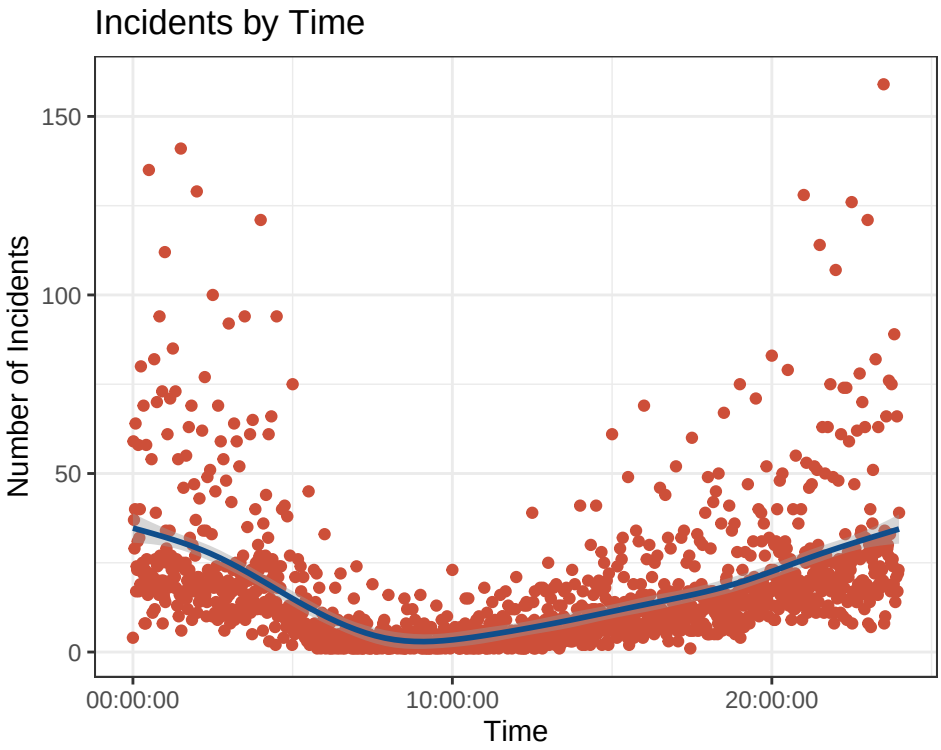


Here we can see that Brooklynn is has the most incidents, then the Bronx. Manhatten, Queens and Staten Island have significantly fewer incidents.

Now let's look at the incidents by day and time.



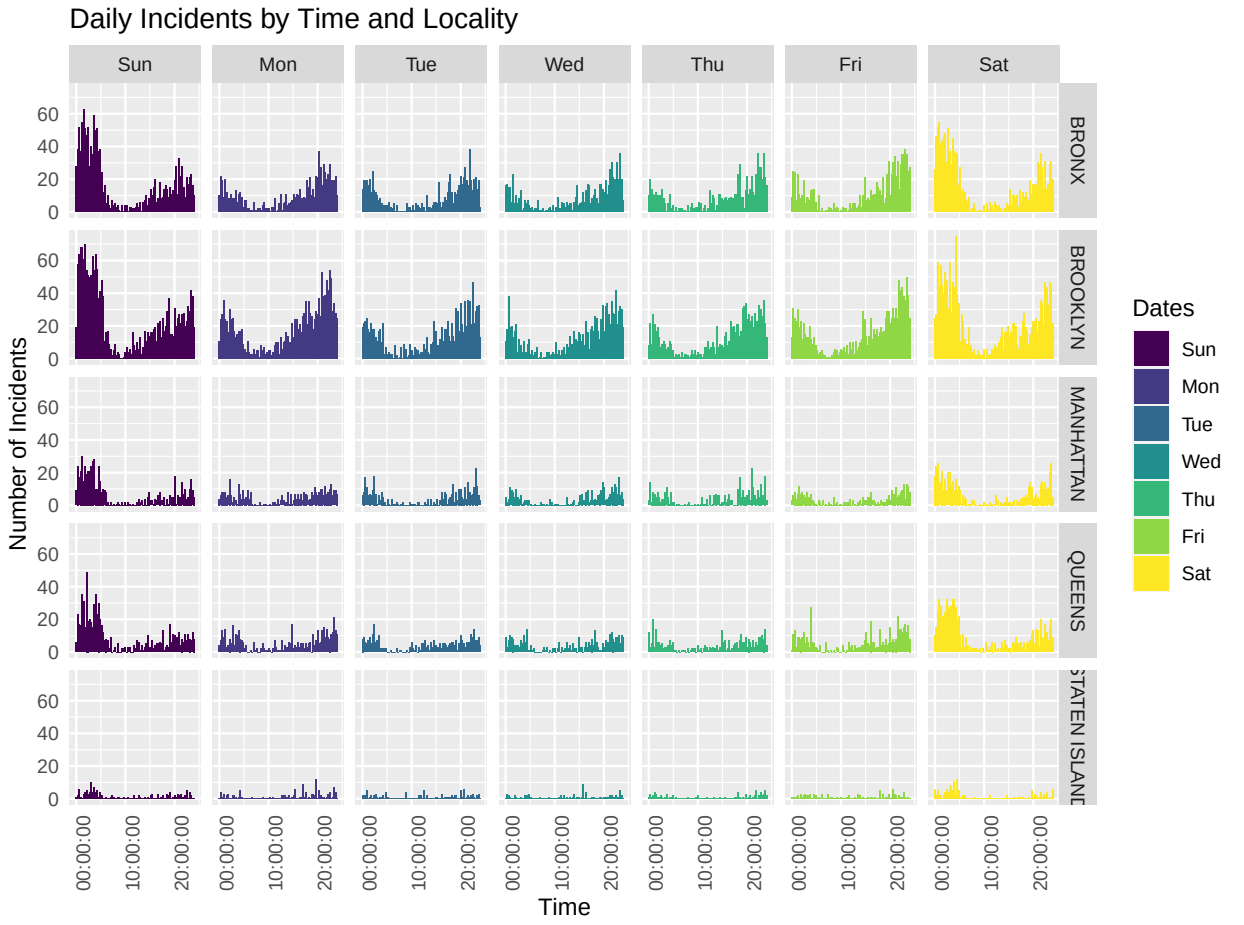

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



We can see that the middle of the week (Tuesday, Wednesday, Thursday) are safer days because of fewer incidents of gun violence. Saturday and Sunday are the most dangerous.

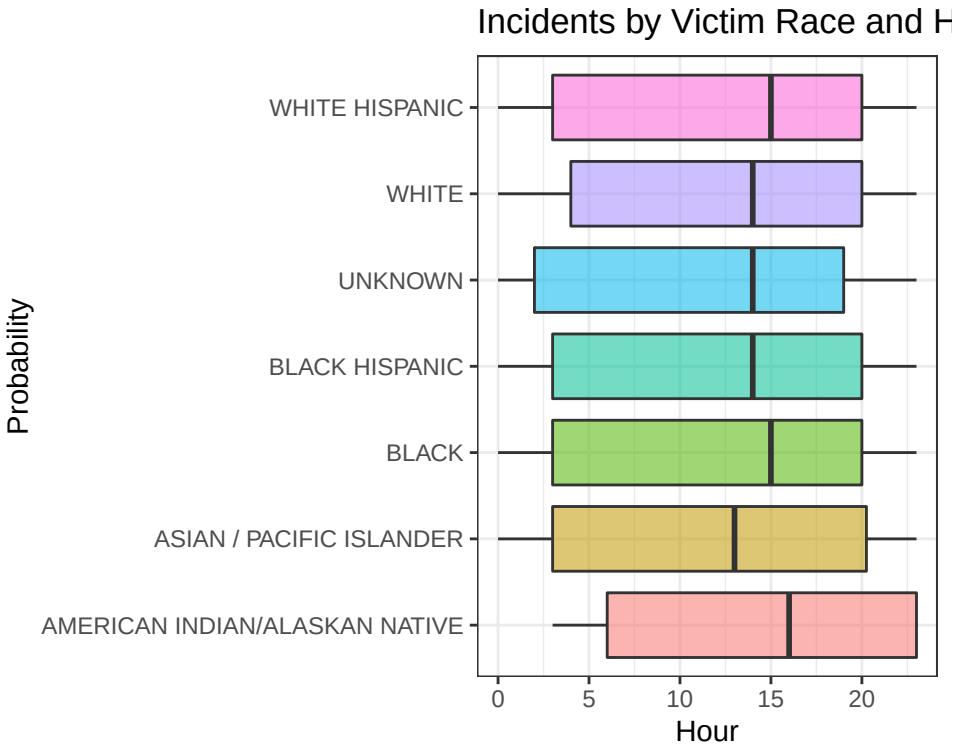
By looking at the times of incidents, we can see that mornings (around 9:00am) are the safest. We notice that there are more extreme values during at night but we can contrast that with the best-fit curve line which shows that the incidents on average don't vary that much. Therefore, we can say that mid-morning is safest and night is more dangerous.

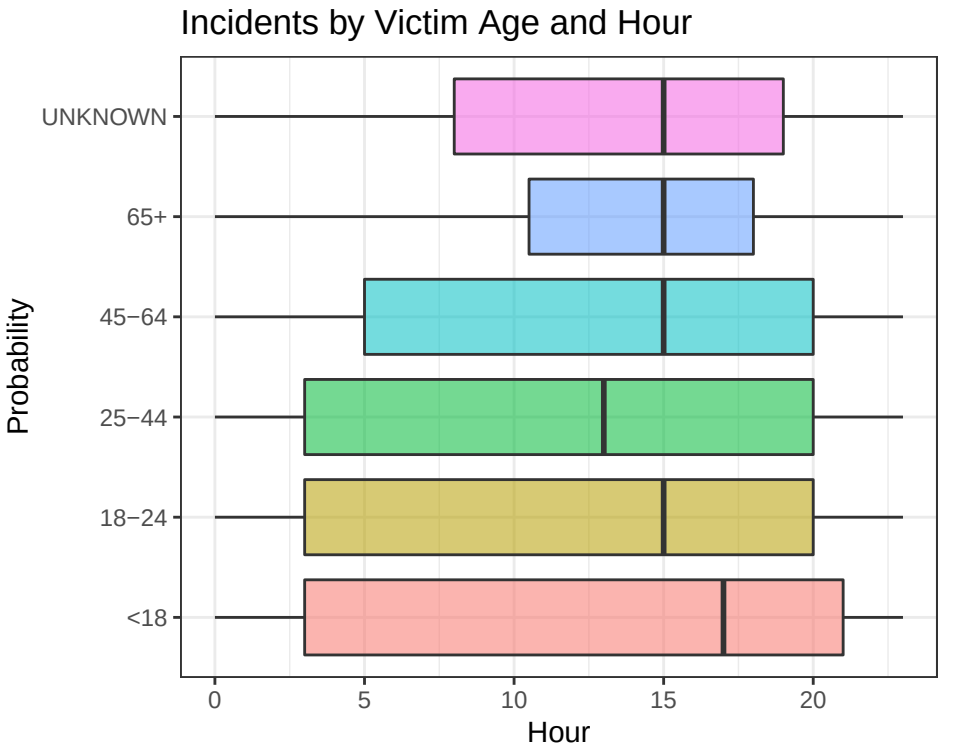
By looking at the prior plots, we start to question if there are patterns between these three variables. Therefore, we can plot all three on a facet grid to get a better idea of their relationship.



In the plot, we can see that the Bronx and Brooklynn have the highest number of incidents early on Saturday and Sunday mornings/nighttime. We can begin to question as to why these could be occurring and how connected these variables are. For this, we will run some models to find correlations.

We also want to see if race, age or gender of the victim affects the incident level.


Incidents by Victim Race and H
WHITE HISPANIC
WHITE
UNKNOWN
BLACK HISPANIC
BLACK
ASIAN / PACIFIC ISLANDER
AMERICAN INDIAN/ALASKAN NATIVE
Probability
0
5
10
15
20
Hour


Incidents by Victim Age and Hour
UNKNOWN
65+
45−64
25−44
18−24
<18
Probability
0
5
10
15
20
Hour

By looking at the plots, we can see that generally race and gender do not affect the number of incidents (note "U" is unknown gender). We also can see that 18- 64-year-olds are most targeted while 65+ are least targeted.

But lets also check the numbers separately for victim race, victim age, and victim sex by using "table()".

```
## [1] "***Grouped by Age***"

##
##     <18   18-24   25-44   45-64     65+ UNKNOWN
##    2525    9000   10287    1536     155      65

## [1] "***Grouped by Sex***"

##
##     F     M     U
##  2195 21353    20

## [1] "***Grouped by Race***"

##
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                              9                            320
##                          BLACK                  BLACK HISPANIC
##                          16846                           2244
##                        UNKNOWN                          WHITE
##                            102                            615
##                WHITE HISPANIC
##                           3432
```

Just by looking at the numbers, we can see that the ages 18-44 are most targeted, that males are much more targeted than females, and that blacks out numbers the other races in this data set.

Modeling Data

First let's look at any relationships between the variables by running multiple linear regressions. We chose to use the following criteria in our regression: Dates (day of the week), Time (of day), Locality (location), victim age, victim sex, victim race. To run these variables as a multiple regression, we must set them to numerical data so that the algorithm can compute relationships. Then let's have a look at the summary and regression data head to make sure it looks okay.

```
##      dates        time_of_day       location        vic_age        vic_race
##  Min.   :1.000   Min.   :    0   Min.   :1.000   Min.   :1.00   Min.   :1.00
##  1st Qu.:2.000   1st Qu.:12000   1st Qu.:1.000   1st Qu.:2.00   1st Qu.:3.00
##  Median :4.000   Median :54000   Median :2.000   Median :3.00   Median :3.00
##  Mean   :3.939   Mean   :45179   Mean   :2.228   Mean   :2.49   Mean   :3.75
##  3rd Qu.:6.000   3rd Qu.:74655   3rd Qu.:3.000   3rd Qu.:3.00   3rd Qu.:4.00
##  Max.   :7.000   Max.   :86340   Max.   :5.000   Max.   :6.00   Max.   :7.00
##     vic_sex
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :1.908
##  3rd Qu.:2.000
##  Max.   :3.000
```

```
##   dates time_of_day location vic_age vic_race vic_sex
## 1     6       79800        4       3        3       2
## 2     4       57240        1       3        3       1
## 3     7       70800        3       2        4       2
## 4     5        3120        5       3        3       1
## 5     5       64980        1       2        3       2
## 6     6       64200        2       3        3       2
```

Now we will run the multiple regression model. We set the target value to victim race because we want to see if is correlated to any other variables.

```
regressor1 = lm(formula=vic_race ~ . , data=regression_df)
print("Summary with all variables")
```

```
## [1] "Summary with all variables"
```

```
summary(regressor1)
```

```
##
## Call:
## lm(formula = vic_race ~ ., data = regression_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8013 -0.8026 -0.7120  0.1166  3.6343
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.979e+00  7.507e-02  53.003  < 2e-16 ***
## dates        2.318e-03  4.295e-03   0.540    0.589
## time_of_day  9.395e-08  3.084e-07   0.305    0.761
## location    -8.863e-02  8.587e-03 -10.321  < 2e-16 ***
## vic_age      8.202e-02  1.161e-02   7.062 1.68e-12 ***
## vic_sex     -1.304e-01  3.242e-02  -4.022 5.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.454 on 23562 degrees of freedom
## Multiple R-squared:  0.007089,   Adjusted R-squared:  0.006878
## F-statistic: 33.64 on 5 and 23562 DF,  p-value: < 2.2e-16
```

```
regressor2 = lm(formula=vic_race ~ location + vic_age + vic_sex, data=regression_df)
print("Summary without data nor time")
```

```
## [1] "Summary without data nor time"
```

```
summary(regressor2)
```

```
##
## Call:
## lm(formula = vic_race ~ location + vic_age + vic_sex, data = regression_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8070 -0.8003 -0.7183  0.1110  3.6300
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.991793   0.072029  55.419  < 2e-16 ***
## location    -0.088739   0.008578 -10.346  < 2e-16 ***
## vic_age      0.082026   0.011607   7.067 1.63e-12 ***
## vic_sex     -0.130042   0.032414  -4.012 6.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.454 on 23564 degrees of freedom
## Multiple R-squared:  0.007072,   Adjusted R-squared:  0.006946
## F-statistic: 55.94 on 3 and 23564 DF,  p-value: < 2.2e-16
```

We can see that date and time are not significant in this model which suggests they do not play an important role in affecting which race is targeted. Let's run again without date and time in this model. We can see that these variables are all highly statistically significant (p is close to 0).

Let's run another regression focusing on time of day.

```
regressor3 = lm(formula=time_of_day ~ . , data=regression_df)
summary(regressor3)
```

```
##
## Call:
## lm(formula = time_of_day ~ ., data = regression_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49946 -32972   8779  29073  46431
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46393.29    1649.96  28.118  < 2e-16 ***
## dates         563.36      90.63   6.216 5.19e-10 ***
## location    -1304.88     181.59  -7.186 6.88e-13 ***
## vic_age     -1008.77     245.46  -4.110 3.98e-05 ***
## vic_race       41.91     137.60   0.305    0.761
## vic_sex       958.57     684.95   1.399    0.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30720 on 23562 degrees of freedom
## Multiple R-squared:  0.004711,   Adjusted R-squared:  0.0045
## F-statistic: 22.31 on 5 and 23562 DF,  p-value: < 2.2e-16
```

Running a regression on the time has high statistical significance with date, location and victim age, suggesting that these variables are closely related.

Lastly, let's run a regression with dates as the target.

```
regressor4 = lm(formula=dates ~ . , data=regression_df)
summary(regressor4)
```

```
##
## Call:
## lm(formula = dates ~ ., data = regression_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2377 -2.0156 -0.0089  2.0380  3.3708
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.451e+00  1.184e-01  29.158  < 2e-16 ***
## time_of_day 2.906e-06  4.675e-07   6.216 5.19e-10 ***
## location    1.159e-02  1.306e-02   0.888  0.37452
## vic_age     4.636e-02  1.763e-02   2.629  0.00856 **
## vic_race    5.334e-03  9.882e-03   0.540  0.58938
## vic_sex     1.027e-01  4.919e-02   2.087  0.03692 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.206 on 23562 degrees of freedom
## Multiple R-squared:  0.002113,   Adjusted R-squared:  0.001901
## F-statistic: 9.978 on 5 and 23562 DF,  p-value: 1.493e-09
```

We see that only time of day is very closed related to the date. Knowing this, let's remove location and victim race from the regression and see how the relationships change. We can see that time of day is closely relate. However, victim age and victim sex are connected, but to a less extent.

From these analysis, we are given the following information to decipher: - Victim race is highly connected to location, victim age, and victim sex. - Time of day is highly connected to date, location, and victim age - Date is highly connected to time of day, and less so (but still significantly) with victim age and victim sex.

##Conclusions

Based on the information we are given, we can conclude that location (borough/locality), day of the week, and time of day are very influential to the number of gun shooting incidents. We further conclude that the victim age, race, and sex may be targeted as well.

By looking at the relationships through the regression models and by looking at the numbers for targeted victims, we can confidently say that the following would be the most dangerous situation would be to be a black male, aged 18 to 44, in the Bronx or Brooklynn at nighttime hours Friday-Sunday. On the contrary, the safest situation would be an 65+ year old female American Indian/Alaskan Native on Staten Island between Tuesday and Wednesday in the daylight morning hours. However, more analysis is needed to make concrete conclusions.

We must also be aware that other, outside variables could influence our dataset. For example: (1) Are there known gangs in the more active areas? (2) Do poverty levels affect the the number of incidents? (3) What is the rate of school graduation or attendance? (4) Are gun sales connected to neighborhood, incidents, gender, race, etc? (5) many other questions to answer as well. To find these answers, we would need to find relevant data online, import that data and run more analysis to find any connections.

Bias and Ethics

When doing any analysis, we will have biases with ourselves, the data and the algorithms. We must try our hardest not to assume anything and work through the process without our personal opinions.

In my personal case, I have mixed feeling about the gun laws in the United States which may have influenced how I handled the data. To help prevent my personal biases, I compared much of the data with different parts to get the whole picture. In addition, from watching TV shows and movies, we always hear about gangs in LA and NYC. Compton is famous for gang violence, while in NYC the Bronx is famous for gang violence. To deter any prior biases we may have while working with new data sets, it is imperative to ensure a full analysis without making any assumptions. In addition, once the data is analyzed, we can return and rerun analysis to ensure few biases have appeared. Also, by adding more outside data, such as the suggestions above, can help as well.

In this dataset, we chose to remove the missing variables because there were so many (perp race, perp age, perp sex). However, a different analysis could have, instead, replaced these missing variables with the most common value or mean value. Doing so would have resulted in much different algorithms and analysis, potentially skewing the data as well.

Another important part of biases is outliers. Outliers are sometimes seen as just that, an outlier, something that is not normal and should usually be disregarded. However, a good analysis would try to understand the underlying cause of those outliers. For example, in our analysis, we have many outliers (high incident numbers away from the average) if we just look at time vs. incidents. Upon further investigation, we can suggest that those values are from the Bronx or Brooklynn areas. Therefore, we can begin to question is it the time, the location, or another outside factor that is increasing shootings at those times? Further data and analysis would be needed to determine this.

Lastly, machine learning algorithms themselves cause biases when running. When we do linear regressions or train-test models, we are allowing the computer to make decisions which will always have some bias. Every machine learning algorithm has a diffferent algorithm to learn, different rules to learn by. By insuring our data is clean, has sufficient information and is handled correctly before testing, we can assure ourselves that

it is as least biased as possible. In addition, after running a regression and model, we should look at the results to make sure they make sense and do further tests if needed.

Appendix

- Data was collected from opendata.cityofnewyork.us with the specific csv file downloadable from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD
- Further data should be found online and downloaded. A good way to find more data is to simply Google and use reputatal sites. Sites to begin your search are: Google Dataset Search, Kaggle, Data.gov, Datahub.io, Earth Data, CERN Open Data Portal, and even github.