NYPD Shooting Incident Data

VM

2024-06-16

NYPD Shooting Incident Data Analysis and Model

```
library(tidySEM)
library(lavaan)
library(ggplot2)
library(dplyr)
library(tidyr)
library(PerformanceAnalytics)
library(prophet)
```

Loading the Data

The data is loaded from the link data.gov

```
data <- read.csv("./NYPD_Shooting_Incident_Data__Historic_.csv")
glimpse(data)
```

```
## Rows: 27,312
## Columns: 21
## $ INCIDENT_KEY            <int> 228798151, 137471050, 147998800, 146837977, 58~
## $ OCCUR_DATE              <chr> "05/27/2021", "06/27/2014", "11/21/2015", "10/~
## $ OCCUR_TIME              <chr> "21:30:00", "17:40:00", "03:56:00", "18:30:00"~
## $ BORO                    <chr> "QUEENS", "BRONX", "QUEENS", "BRONX", "BRONX",~
## $ LOC_OF_OCCUR_DESC       <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ PRECINCT                <int> 105, 40, 108, 44, 47, 81, 114, 81, 105, 101, 2~
## $ JURISDICTION_CODE       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2~
## $ LOC_CLASSFCTN_DESC      <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ LOCATION_DESC           <chr> "", "", "", "", "", "", "", "", "", "MULTI DWE~
## $ STATISTICAL_MURDER_FLAG <chr> "false", "false", "true", "false", "true", "tr~
## $ PERP_AGE_GROUP          <chr> "", "", "", "", "25-44", "", "", "", "", "25-4~
## $ PERP_SEX                <chr> "", "", "", "", "M", "", "", "", "", "M", "", ~
## $ PERP_RACE               <chr> "", "", "", "", "BLACK", "", "", "", "", "BLAC~
## $ VIC_AGE_GROUP           <chr> "18-24", "18-24", "25-44", "<18", "45-64", "25~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "WHITE", "WHITE HISPANIC", "~
## $ X_COORD_CD              <dbl> 1058925.0, 1005028.0, 1007667.9, 1006537.4, 10~
## $ Y_COORD_CD              <dbl> 180924.0, 234516.0, 209836.5, 244511.1, 262189~
## $ Latitude                <dbl> 40.66296, 40.81035, 40.74261, 40.83778, 40.886~
## $ Longitude               <dbl> -73.73084, -73.92494, -73.91549, -73.91946, -7~
## $ Lon_Lat                 <chr> "POINT (-73.73083868899994 40.662964620000025)~
```

Exploratory Data Analysis

EDA - Exploratory data analysis is the first step in data science process to understand the data in hand. The basic step is identify the integrity of the data, whether it has all necessary information, data types, missing values and number of columns and records.

`glimpse` function from `dplyr` package gives a short summary view of high level data loaded into data frame.

This indicates that we have 21 different dimensions (columns) and 27, 312 observations (rows).

Filter required columns

Filtering key columns of interest to get more insite about the data

Column	Description	Data Type
OCCUR_DATE	Incident Date	Date format mm/dd/yyyy
OCCUR_TIME	Incident Time	Time format HH:mm:ss
BORO	Borough in which a crime occurred	string
STATISTICAL_MURDER_FLAG	field that indicates whether a homicide is considered to be a statistical murder	boolean
PERP_AGE_GROUP	perpetrator age group	range string
PERP_SEX	perpetrator sex	string
PERP_RACE	perpetrator race	string
VIC_AGE_GROUP	victim age group	string
VIC_SEX	victim age group	string
VIC_RACE	victim age group	string
Latitude	Latitude	number
Longitude	Longitude	number

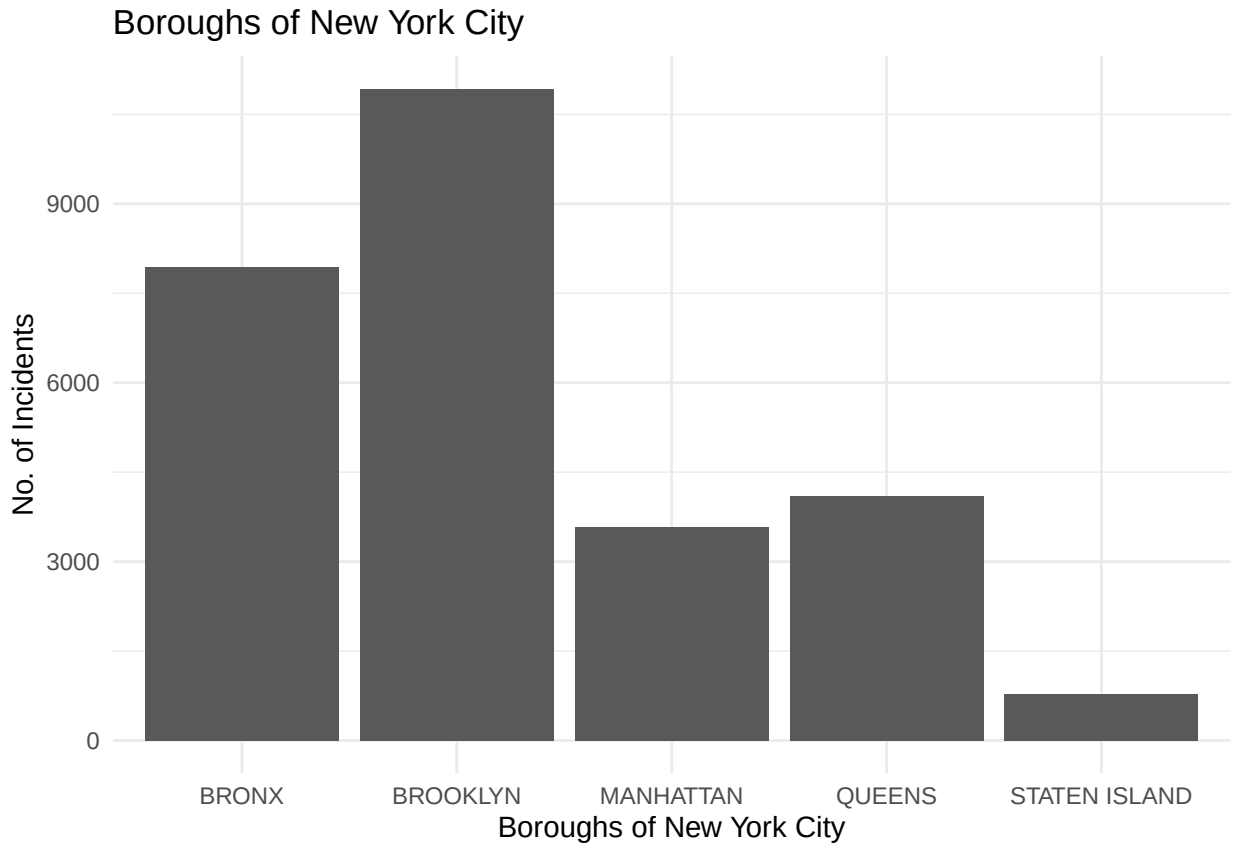
```
data[data==""]<-NA
data[data=="(null)"]<-NA
data_filtered <- data %>% select(OCCUR_DATE, OCCUR_TIME, BORO,STATISTICAL_MURDER_FLAG,
                                 PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                                 VIC_AGE_GROUP, VIC_SEX, VIC_RACE, Latitude, Longitude) %>%
                      mutate(PERP_AGE_GROUP = ifelse(is.na(PERP_AGE_GROUP), 'UNKNOWN', PERP_AGE_GROUP))
                      mutate(PERP_SEX = ifelse(is.na(PERP_SEX), 'UNKNOWN', PERP_SEX)) %>%
                      mutate(PERP_RACE = ifelse(is.na(PERP_RACE), 'UNKNOWN', PERP_RACE))
data_filtered$STATISTICAL_MURDER_FLAG_INT <- as.integer(as.logical(data_filtered$STATISTICAL_MURDER_FLA
head(data_filtered)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1 05/27/2021   21:30:00   QUEENS                  false        UNKNOWN
## 2 06/27/2014   17:40:00    BRONX                  false        UNKNOWN
## 3 11/21/2015   03:56:00   QUEENS                   true        UNKNOWN
## 4 10/09/2015   18:30:00    BRONX                  false        UNKNOWN
## 5 02/19/2009   22:58:00    BRONX                   true          25-44
## 6 10/21/2020   21:36:00 BROOKLYN                   true        UNKNOWN
##   PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE Latitude Longitude
## 1  UNKNOWN   UNKNOWN         18-24       M          BLACK 40.66296 -73.73084
## 2  UNKNOWN   UNKNOWN         18-24       M          BLACK 40.81035 -73.92494
## 3  UNKNOWN   UNKNOWN         25-44       M          WHITE 40.74261 -73.91549
## 4  UNKNOWN   UNKNOWN           <18       M WHITE HISPANIC 40.83778 -73.91946
## 5        M     BLACK         45-64       M          BLACK 40.88624 -73.85291
```

```
## 6  UNKNOWN   UNKNOWN         25-44       M          BLACK 40.67846 -73.92795
##   STATISTICAL_MURDER_FLAG_INT
## 1                           0
## 2                           0
## 3                           1
## 4                           0
## 5                           1
## 6                           1
```

Visualizing the Incidents

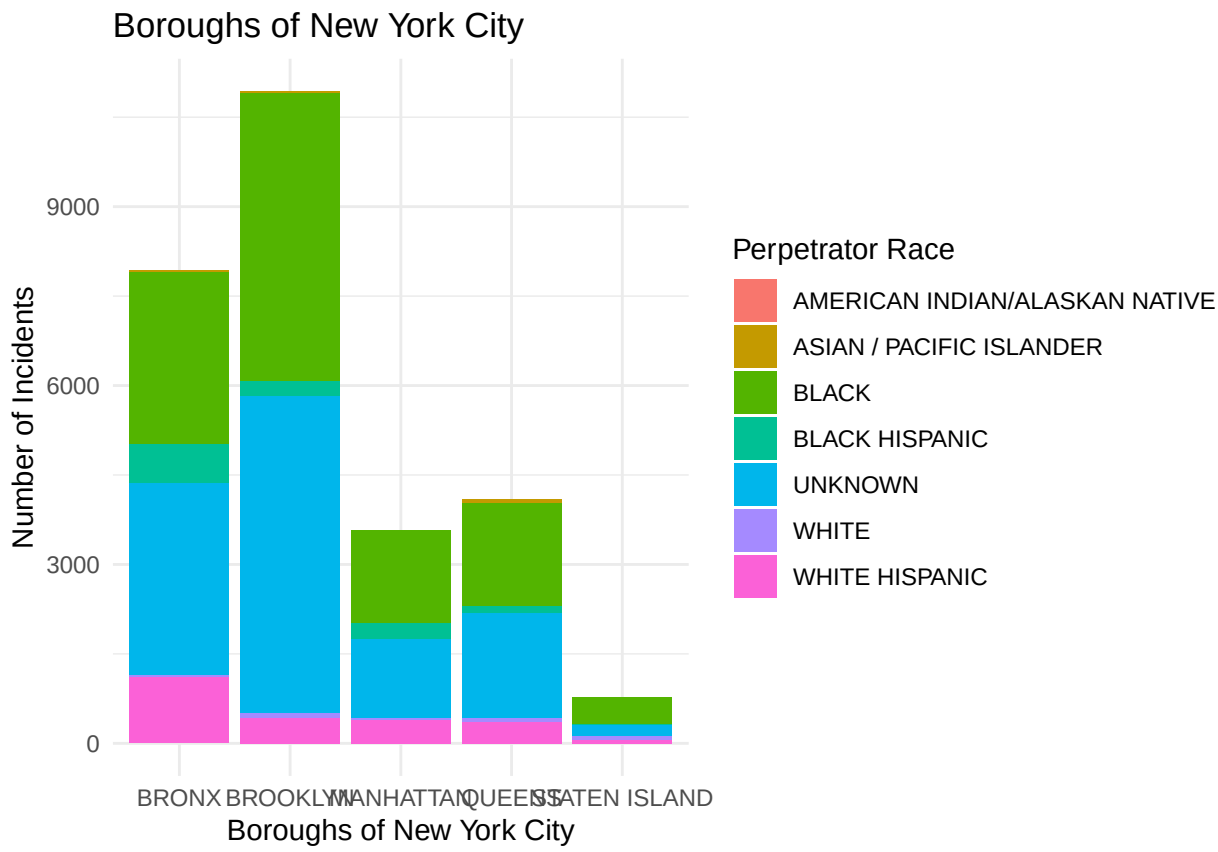
```
g <- ggplot(data_filtered, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Boroughs of New York City",
       x = "Boroughs of New York City",
       y = "No. of Incidents") +
  theme_minimal()
g
```



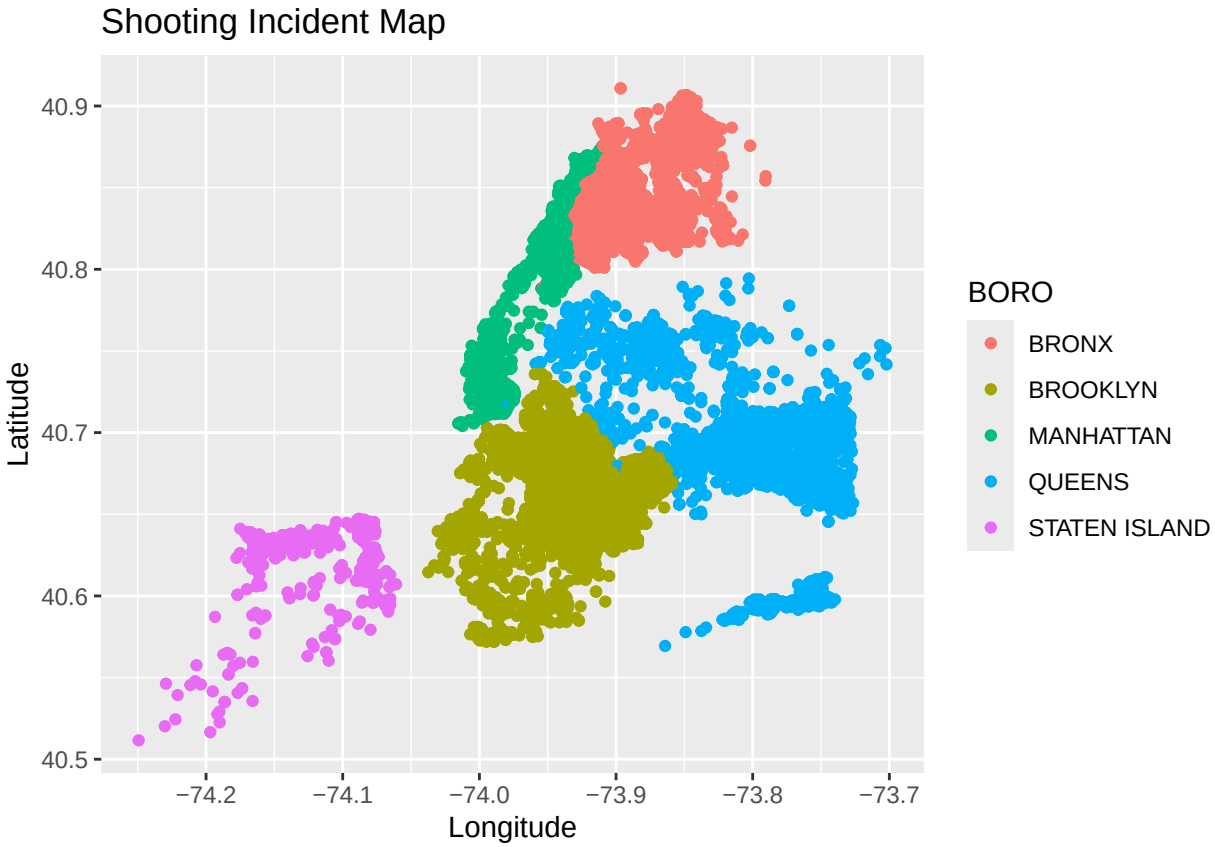
Visualizing the Incidents including Race

```
g <- ggplot(data_filtered, aes(x = BORO, fill = PERP_RACE)) +
  geom_bar(position = "stack") +
  labs(title = "Boroughs of New York City",
       x = "Boroughs of New York City",
       y = "Number of Incidents",
```

```
       fill = "Perpetrator Race") +
  theme_minimal()

# Display the plot
print(g)
```




```
map <- data_filtered %>%
          ggplot(aes(x=Longitude, y=Latitude, col=BORO)) +
          geom_point() +
          labs(title = "Shooting Incident Map",
                  x = "Longitude",
                  y = "Latitude")
map
```



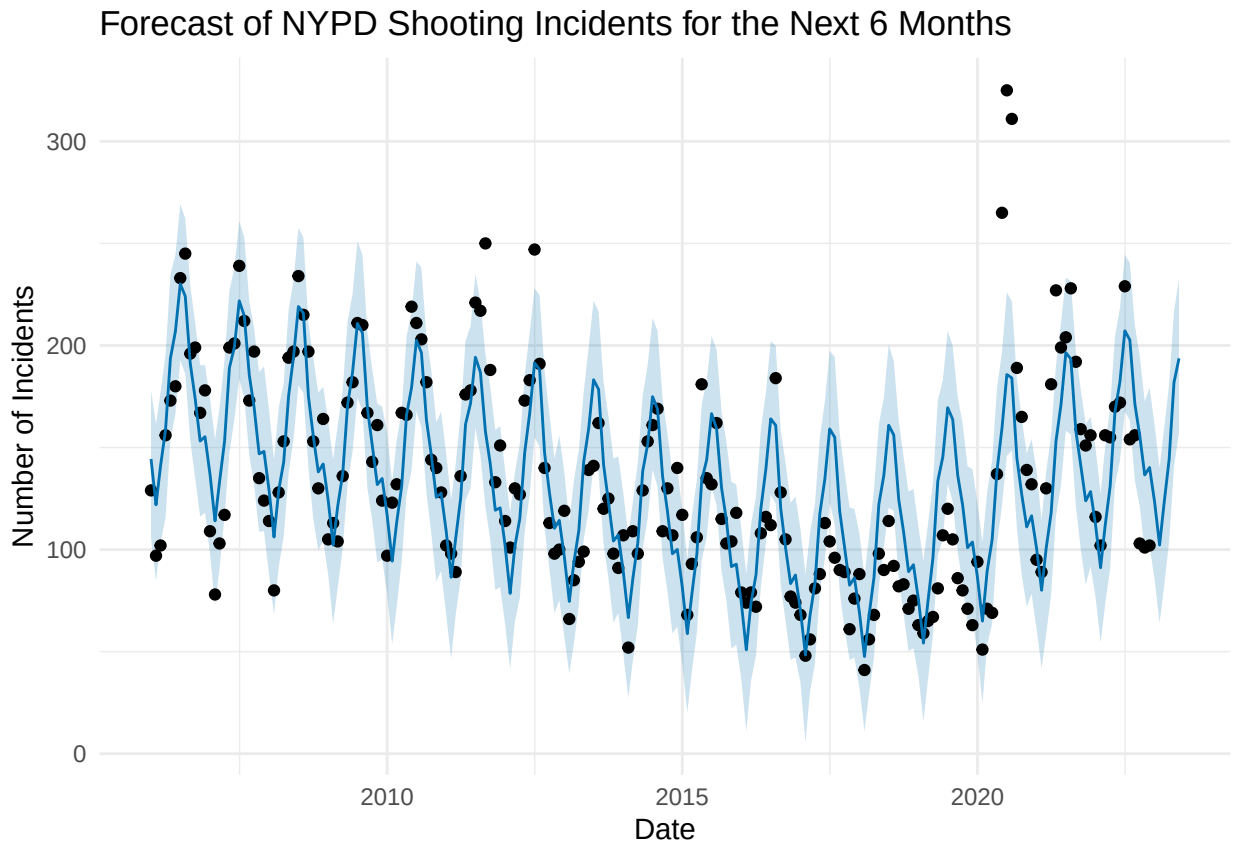
Feature Selection before modeling

This is a open end problem we are not expected to produce a particular model. So try to make it simple but still cover the aspects of DS project. I wanted to make a simple assumption about the goal.

The idea is to find whether we can create a simple forecast of incidents for next six months, and plot the results with confidence interval to understand the pattern of how these incidence happen and possible future values. This will have two model, one without including race filter another including race filter to find how particular race influence the total incidence. ## Basic Model

```
data_filtered <- data %>%
  select(OCCUR_DATE, BORO) %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))

# Aggregate data by month
monthly_data <- data_filtered %>%
  group_by(month = as.Date(cut(OCCUR_DATE, breaks = "month"))) %>%
  summarise(incidents = n()) %>%
  ungroup()

# Prepare the data for Prophet
prophet_data <- monthly_data %>%
  rename(ds = month, y = incidents)

# Fit the Prophet model
m <- prophet(prophet_data)
```

```r
# Make a future dataframe for 6 months
future <- make_future_dataframe(m, periods = 6, freq = "month")

# Forecast
forecast <- predict(m, future)

# Plot the forecast
plot(m, forecast) +
  ggtitle("Forecast of NYPD Shooting Incidents for the Next 6 Months") +
  xlab("Date") +
  ylab("Number of Incidents") +
  theme_minimal()
```



Model interpretation

Based on the forecast plot generated using the `prophet` model, here are some key interpretations:

1. **Seasonal Patterns**: The plot shows a clear seasonal pattern in the number of incidents. There are recurring peaks and troughs within each year, indicating that incidents are likely influenced by seasonal factors. This is typical in many crime datasets, where certain times of the year see higher crime rates.
2. **Trend Over Time**: The overall trend shows some variability, with periods of increase and decrease in the number of incidents over the years. There is an observable decline in incidents around 2014-2017, followed by a recent increase.
3. **Recent Increase**: The most recent data points (2020 and onward) show an increase in the number of incidents. This uptick might be due to various socio-economic factors, changes in law enforcement practices, or other external events such as the COVID-19 pandemic.

4. **Forecast for the Next 6 Months*** The forecast for the next 6 months shows a continuation of the seasonal pattern with an expected increase in incidents. The confidence intervals (shaded area) indicate the range within which the actual number of incidents is likely to fall. The intervals widen slightly over time, reflecting increasing uncertainty in the forecast further into the future.
5. **Anomalies and Outliers**: Some data points appear as outliers (e.g., very high or low compared to the overall pattern). These could be due to specific events or reporting anomalies. The model seems to capture these outliers but does not let them heavily influence the overall trend and seasonality.
6. **Model Fit**: The model appears to fit the historical data well, capturing the underlying seasonal patterns and long-term trends. However, it's important to validate the model using out-of-sample data to ensure its robustness.

Additional Considerations:

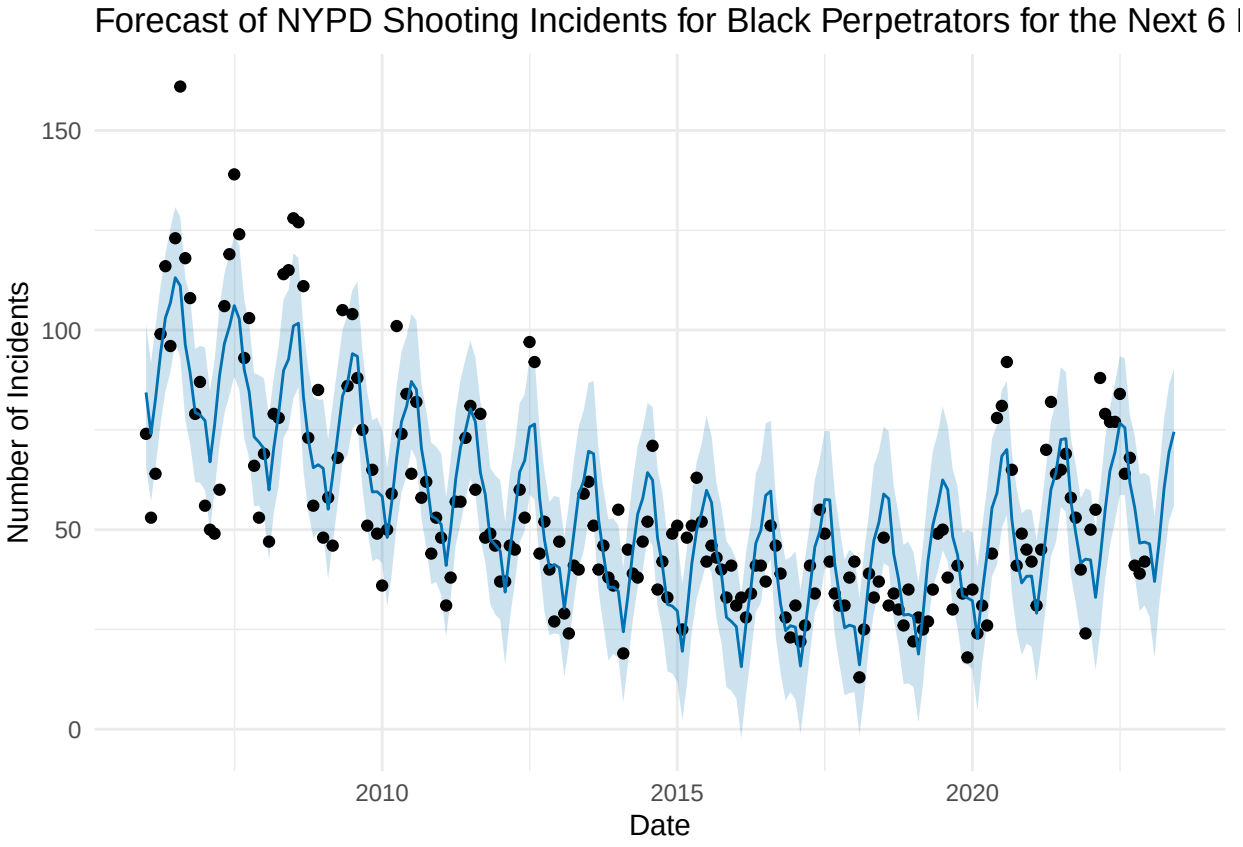
1. **External Factors:** It's crucial to consider external factors that might influence the number of incidents, such as policy changes, economic conditions, or public health crises.

Overall, the `prophet` model provides a good starting point for understanding and forecasting the number of NYPD shooting incidents, with clear insights into seasonal patterns and recent trends.

Second model without perpetrator variables

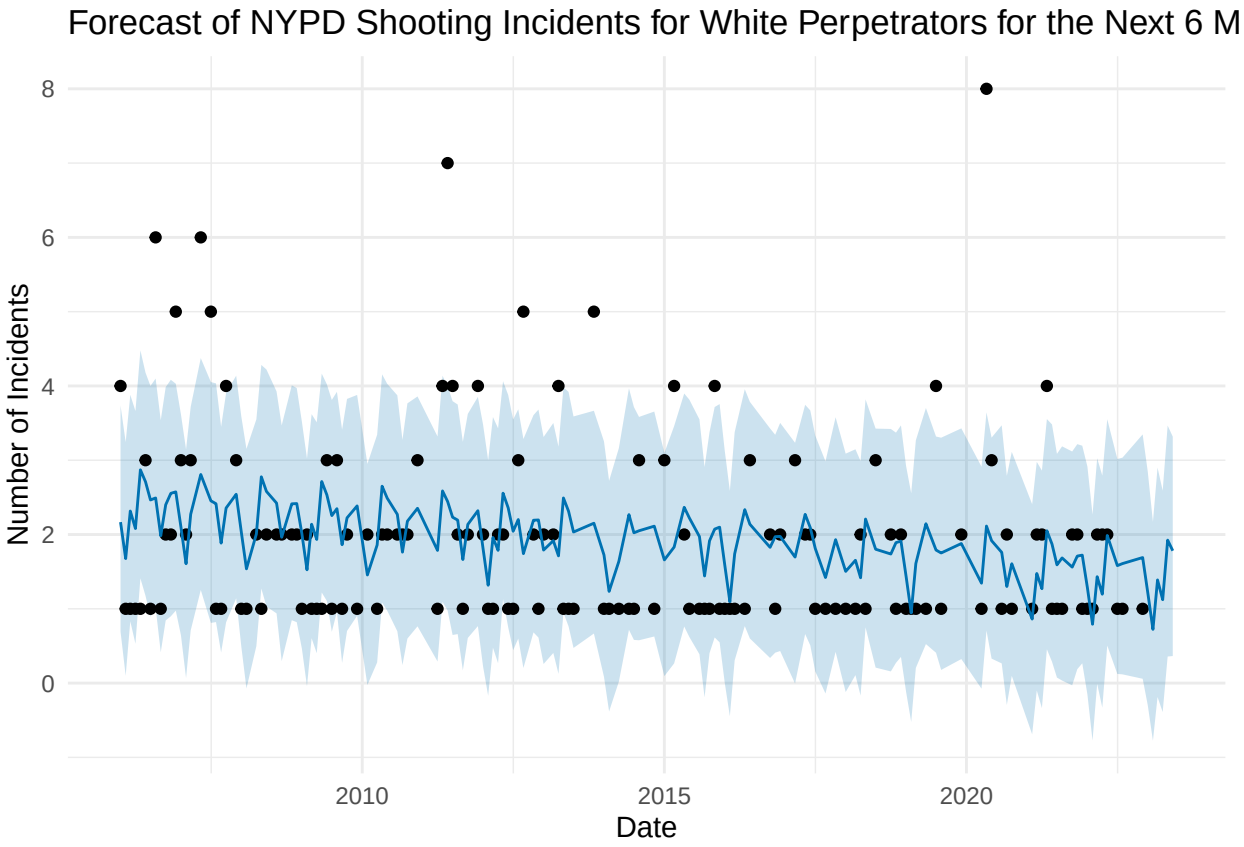
```
data_filtered <- data %>%
  select(OCCUR_DATE, PERP_RACE) %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))

# Filter data for Black perpetrators
data_black <- data_filtered %>%
  filter(PERP_RACE == "BLACK")

# Filter data for White perpetrators
data_white <- data_filtered %>%
  filter(PERP_RACE == "WHITE")

# Aggregate data by month for Black perpetrators
monthly_data_black <- data_black %>%
  group_by(month = as.Date(cut(OCCUR_DATE, breaks = "month"))) %>%
  summarise(incidents = n()) %>%
  ungroup()

# Aggregate data by month for White perpetrators
monthly_data_white <- data_white %>%
  group_by(month = as.Date(cut(OCCUR_DATE, breaks = "month"))) %>%
  summarise(incidents = n()) %>%
  ungroup()

# Prepare the data for Prophet for Black perpetrators
prophet_data_black <- monthly_data_black %>%
  rename(ds = month, y = incidents)

# Prepare the data for Prophet for White perpetrators
prophet_data_white <- monthly_data_white %>%
  rename(ds = month, y = incidents)

# Fit the Prophet model for Black perpetrators
```

```r
m_black <- prophet(prophet_data_black)

# Fit the Prophet model for White perpetrators
m_white <- prophet(prophet_data_white)

# Make a future dataframe for 6 months for Black perpetrators
future_black <- make_future_dataframe(m_black, periods = 6, freq = "month")

# Make a future dataframe for 6 months for White perpetrators
future_white <- make_future_dataframe(m_white, periods = 6, freq = "month")

# Forecast for Black perpetrators
forecast_black <- predict(m_black, future_black)

# Forecast for White perpetrators
forecast_white <- predict(m_white, future_white)

# Plot the forecast for Black perpetrators
plot(m_black, forecast_black) +
  ggtitle("Forecast of NYPD Shooting Incidents for Black Perpetrators for the Next 6 Months") +
  xlab("Date") +
  ylab("Number of Incidents") +
  theme_minimal()
```




```r
# Plot the forecast for White perpetrators
plot(m_white, forecast_white) +
```

```
  ggtitle("Forecast of NYPD Shooting Incidents for White Perpetrators for the Next 6 Months") +
  xlab("Date") +
  ylab("Number of Incidents") +
  theme_minimal()
```



Model interpretation

From the model outputs for the forecast of NYPD shooting incidents involving Black and White perpetrators, several interpretations can be drawn:

Forecast for Black Perpetrators:

1. **Seasonal Pattern**:

- The data for incidents involving Black perpetrators shows a clear seasonal pattern. There are regular peaks and troughs, indicating that the number of incidents varies significantly throughout the year.
- This seasonality has been consistently captured over the years, reflecting periodic fluctuations.

2. **Trend Over Time**:
   - There was a noticeable decline in incidents from 2006 to around 2015. After that, the number of incidents appears to stabilize with less pronounced fluctuations.
   - The recent data (2020 onwards) shows a slight uptick, indicating a potential rise in incidents.
3. **Forecast for Next 6 Months**:
   - The forecast suggests an increase in incidents, with the model predicting periodic peaks similar to historical patterns.
   - The confidence intervals are relatively tight, indicating a high level of certainty in the forecasted values.

Forecast for White Perpetrators:

1. **Seasonal Pattern**:
   - The incidents involving White perpetrators do not show a clear seasonal pattern. The data is much more sporadic with fewer discernible peaks and troughs.
   - This lack of clear seasonality makes it harder to predict specific times of increased or decreased incidents.
2. **Trend Over Time**:
   - The overall number of incidents involving White perpetrators is significantly lower compared to Black perpetrators.
   - There is a consistent level of low incidents throughout the years, with occasional spikes.
3. **Forecast for Next 6 Months**:
   - The forecast for incidents involving White perpetrators suggests a stable trend with low incident counts.
   - The confidence intervals are wider, reflecting higher uncertainty in the predictions. This is likely due to the lower number of incidents and the lack of clear seasonal patterns.

Comparative Insights:

1. **Incident Volume**:
   - Incidents involving Black perpetrators are substantially higher than those involving White perpetrators.
   - This difference in volume is consistent over the years and is reflected in the forecasts.
2. **Seasonality and Trends**:
   - The seasonal pattern is strong and clear in the data for Black perpetrators, while it is almost non-existent for White perpetrators.
   - The trend for Black perpetrators shows periodic fluctuations and recent increases, while for White perpetrators, the trend remains relatively flat with occasional spikes.
3. **Forecast Uncertainty**:
   - The model for Black perpetrators has tighter confidence intervals, indicating more confidence in the predictions.
   - The model for White perpetrators has wider confidence intervals, suggesting less confidence and higher variability in the forecasts.

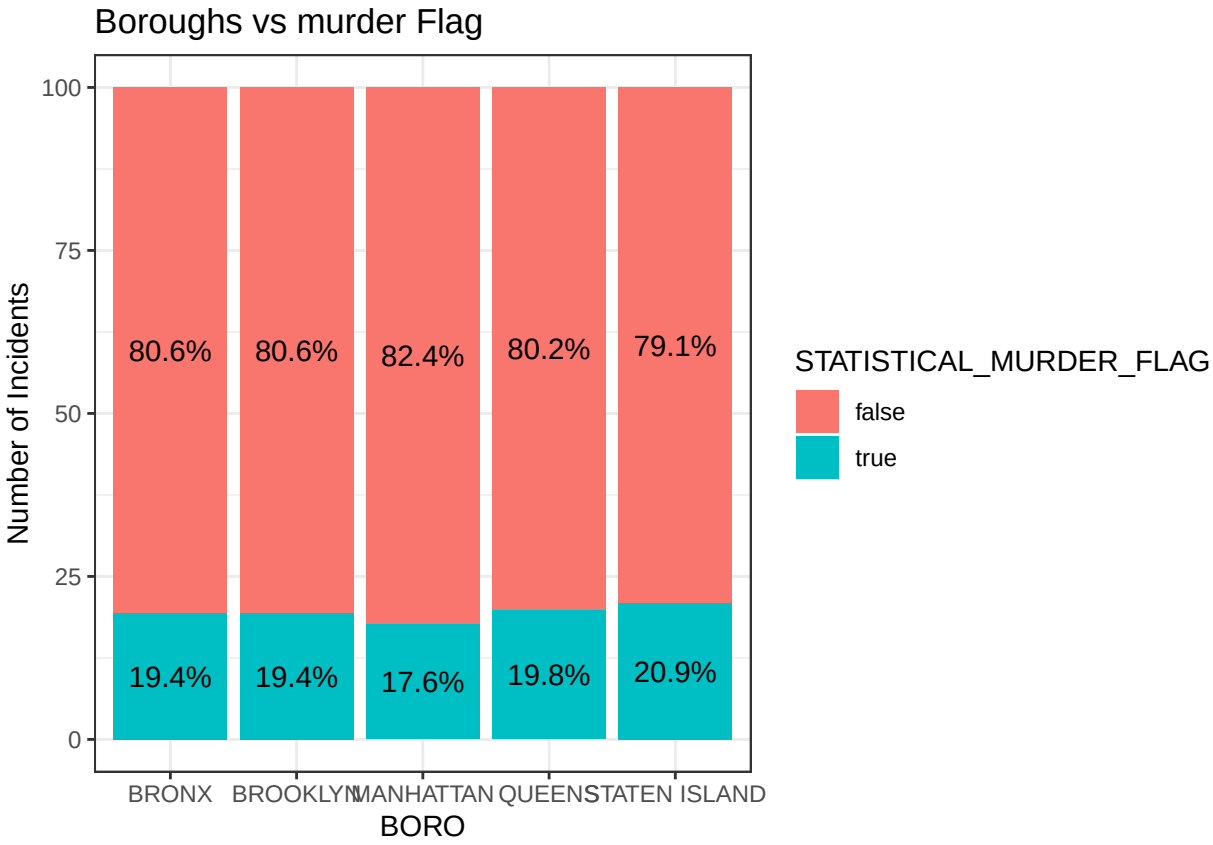
Implications:

1. **Policy and Resource Allocation**: The distinct patterns and higher volume of incidents involving Black perpetrators may warrant targeted interventions and resource allocation.
2. **Seasonal Interventions**: For incidents involving Black perpetrators, seasonal patterns could inform the timing of increased police presence or community outreach programs.
3. **Further Analysis:** The sporadic nature of incidents involving White perpetrators suggests the need for different analytical approaches or more granular data to understand underlying causes.

These interpretations provide a comprehensive understanding of the forecasted incidents for both racial groups and highlight the differences in patterns, trends, and forecast uncertainties.

```
data_filtered <- data %>% select(OCCUR_DATE, OCCUR_TIME, BORO,STATISTICAL_MURDER_FLAG,
                                 PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                                 VIC_AGE_GROUP, VIC_SEX, VIC_RACE, Latitude, Longitude) %>%
                      mutate(PERP_AGE_GROUP = ifelse(is.na(PERP_AGE_GROUP), 'UNKNOWN', PERP_AGE_GROUP))
                      mutate(PERP_SEX = ifelse(is.na(PERP_SEX), 'UNKNOWN', PERP_SEX)) %>%
                      mutate(PERP_RACE = ifelse(is.na(PERP_RACE), 'UNKNOWN', PERP_RACE))
data_filtered %>%
  count(BORO, STATISTICAL_MURDER_FLAG) %>%
  group_by(BORO) %>%
  mutate(pct= prop.table(n) * 100) %>%
```

```
ggplot() + aes(BORO, pct, fill=STATISTICAL_MURDER_FLAG) +
geom_bar(stat="identity") +
ylab("Number of Incidents") +
geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
          position=position_stack(vjust=0.5)) +
ggtitle("Boroughs vs murder Flag") +
theme_bw()
```



conclusion

It appears that there is a clear bias by race, but considering above chart we can do further analysis to understand how many incidents are classified as statistical murders.

The bar chart also provides a clear visualization of the distribution of shooting incidents across New York City's boroughs, highlighting the proportion of incidents classified as statistical murders. Staten Island and Queens have the highest percentages of statistical murders, while Manhattan has the lowest. This information is crucial for understanding the dynamics of violent incidents across the city and for planning targeted interventions.

We can keep exploring various aspects by the objective here is to demonstract the various steps. So I am limiting to only this. This also cover more than one visualization as required for the assignment.

We can extend this analysis by sex, different race categories and much more to infer and understand the pattern.

Concluding bias in the data solely from this chart requires careful consideration and context. However, the distribution of statistical murder flags across boroughs could suggest potential biases or differences in classification practices. Here are some points to consider:

Possible Indicators of Bias:

1. **Uniformity in Non-Statistical Murders**: The high and relatively uniform percentages of non-statistical murders across boroughs (around 79-82%) might suggest that there is a consistent criterion being applied. However, if certain boroughs systematically report fewer incidents as statistical murders, this could indicate a reporting or classification bias.
2. **Variation in Statistical Murders:** The variation in the percentage of statistical murders (17.6% in Manhattan to 20.9% in Staten Island) could indicate differences in how incidents are classified. This might be due to:
3. **Reporting Practices**: Different standards or thresholds for classifying an incident as a statistical murder.
    - **Law Enforcement Policies**: Varied approaches to handling and reporting violent incidents.
    - **Socio-economic Factors**: Differences in socio-economic conditions and community-police relations might influence reporting and classification.
4. **Demographic and Socio-economic Differences**: Bias can also stem from demographic and socio-economic disparities between boroughs. For example, boroughs with higher minority populations might experience different policing practices, which can influence the classification of incidents.

Steps to Investigate Bias:

1. **Compare Incident Details**: Compare detailed characteristics of incidents classified as statistical murders across boroughs to identify any systematic differences.
2. **Demographic Analysis**: Analyze the demographics of perpetrators and victims for statistical murders and non-statistical murders to identify any potential bias in classification.
3. **Historical Context**: Consider historical crime data and socio-economic context to understand if certain boroughs have historically been treated differently in terms of crime reporting and classification.
4. **Policy and Training Review**: Investigate the policies and training provided to law enforcement in different boroughs to understand if there are discrepancies in how incidents are reported and classified.

Conclusion:

While the chart itself does not provide definitive evidence of bias, the variation in the percentage of statistical murders across boroughs suggests the need for further investigation. Differences in reporting practices, law enforcement policies, and socio-economic factors could contribute to potential biases. To draw more robust conclusions, a comprehensive analysis considering these factors is essential.