NYPD shooting analysis, modelization and bias study

2025-02-25

NYPD shooting analysis

This document describes the work done on NYPD shooting cases dataset for CU Boulder MDS 2025

This work tries to answer the following research questions:

- When the shootings happens, which may lead to some correlation with the time.
- Is there any potential bias in the data
- It is safe using machine learning on this dataset
- Its safe to use machine learning to estimate when a possible perpetrator is guilty

Datasources

Data comes from following repository: "https://data.cityofnewyork.us>"

File is:

- **New York shooting cases:** "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

  From the NYPD description:
  This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis 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 and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.
  https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

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

nypd_shootings <- read_csv(url, show_col_types = FALSE)

# Check structure of dataset
glimpse(nypd_shootings)
```

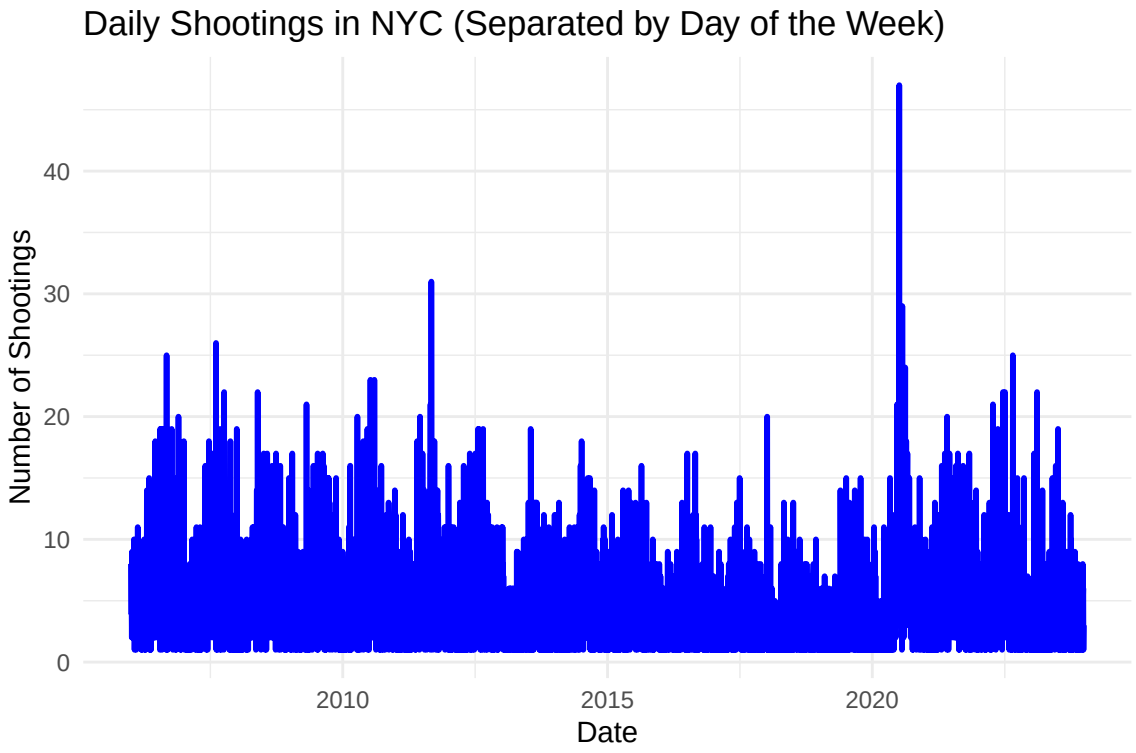
```
## Rows: 28,562
## Columns: 21
## $ INCIDENT_KEY           <dbl> 231974218, 177934247, 255028563, 25384540, 726~
## $ OCCUR_DATE             <chr> "08/09/2021", "04/07/2018", "12/02/2022", "11/~
## $ OCCUR_TIME             <time> 01:06:00, 19:48:00, 22:57:00, 01:50:00, 01:58~
## $ BORO                   <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN", "BRO~
## $ LOC_OF_OCCUR_DESC      <chr> NA, NA, "OUTSIDE", NA, NA, NA, NA, NA, NA, NA,~
## $ PRECINCT               <dbl> 40, 79, 47, 66, 46, 42, 71, 69, 75, 69, 40, 42~
## $ JURISDICTION_CODE      <dbl> 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0, 2, 0~
## $ LOC_CLASSFCTN_DESC     <chr> NA, NA, "STREET", NA, NA, NA, NA, NA, NA, NA, ~
## $ LOCATION_DESC          <chr> NA, NA, "GROCERY/BODEGA", "PVT HOUSE", "MULTI ~
```

```
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, TRUE, F~
## $ PERP_AGE_GROUP          <chr> NA, "25-44", "(null)", "UNKNOWN", "25-44", "18~
## $ PERP_SEX                <chr> NA, "M", "(null)", "U", "M", "M", NA, NA, "M",~
## $ PERP_RACE               <chr> NA, "WHITE HISPANIC", "(null)", "UNKNOWN", "BL~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "18-24", "<18", "18~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "F", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1006343.0, 1000082.9, 1020691.0, 985107.3, 100~
## $ Y_COORD_CD              <dbl> 234270.0, 189064.7, 257125.0, 173349.8, 247502~
## $ Latitude                <dbl> 40.80967, 40.68561, 40.87235, 40.64249, 40.845~
## $ Longitude               <dbl> -73.92019, -73.94291, -73.86823, -73.99691, -7~
## $ Lon_Lat                 <chr> "POINT (-73.92019278899994 40.80967347200004)"~
```

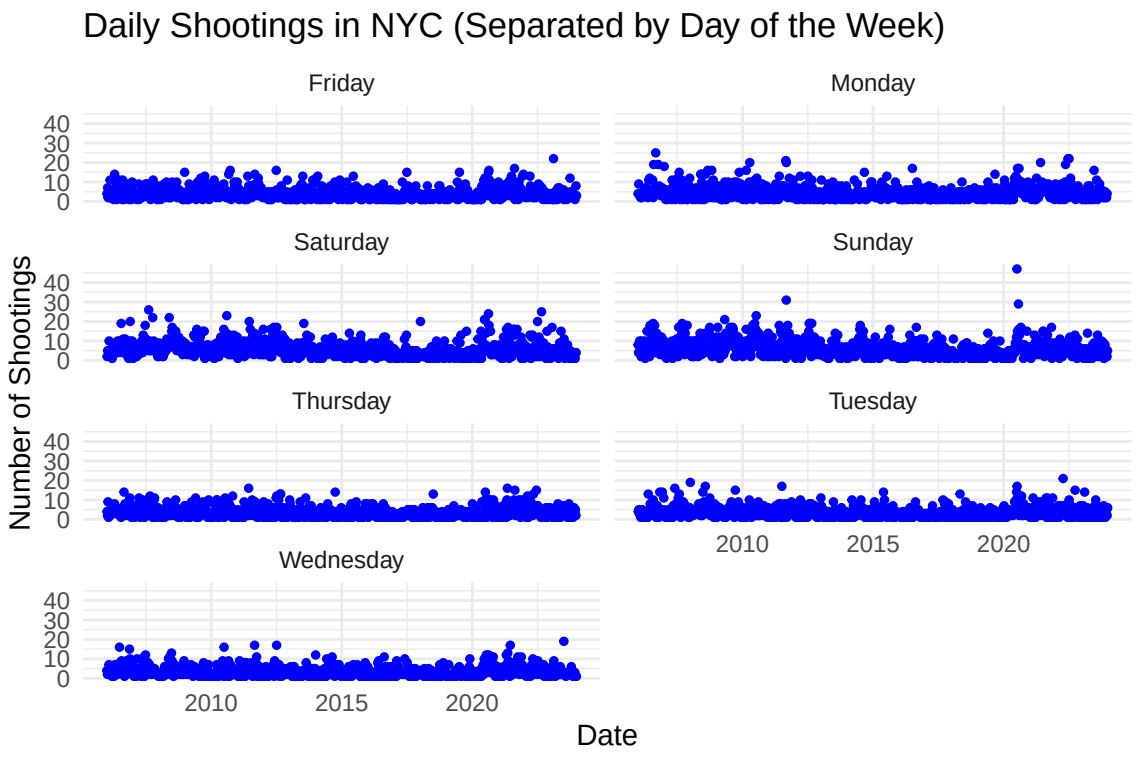
Data pre-process

DATA AGGREGATION

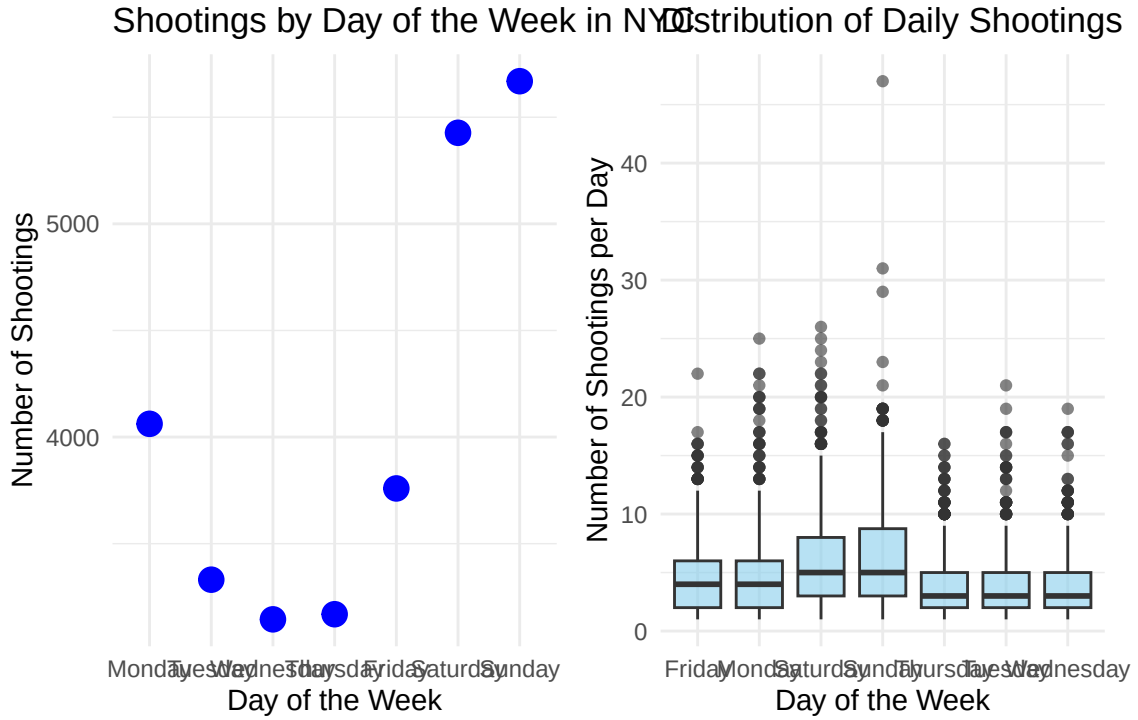
PLOT THE TRAJECTORY OF THE SHOOTINGS ALONG TIME



PLOT THE TRAJECTORY OF THE SHOOTINGS BY WEEKDAY BECAUSE THE TREND IS NOT CLEAR



PLOT THE DISTRUBUTION PER WEEK DAY TO UNDERSTAND IF ITS SOME PATTERN THERE



Its a clear pattern on the weekends, more shoots.
There are more shootings during the weekends.

Also, the boxplot confirms the scatter and my insight is: v**iolent crime increases on weekends**

Indicators:

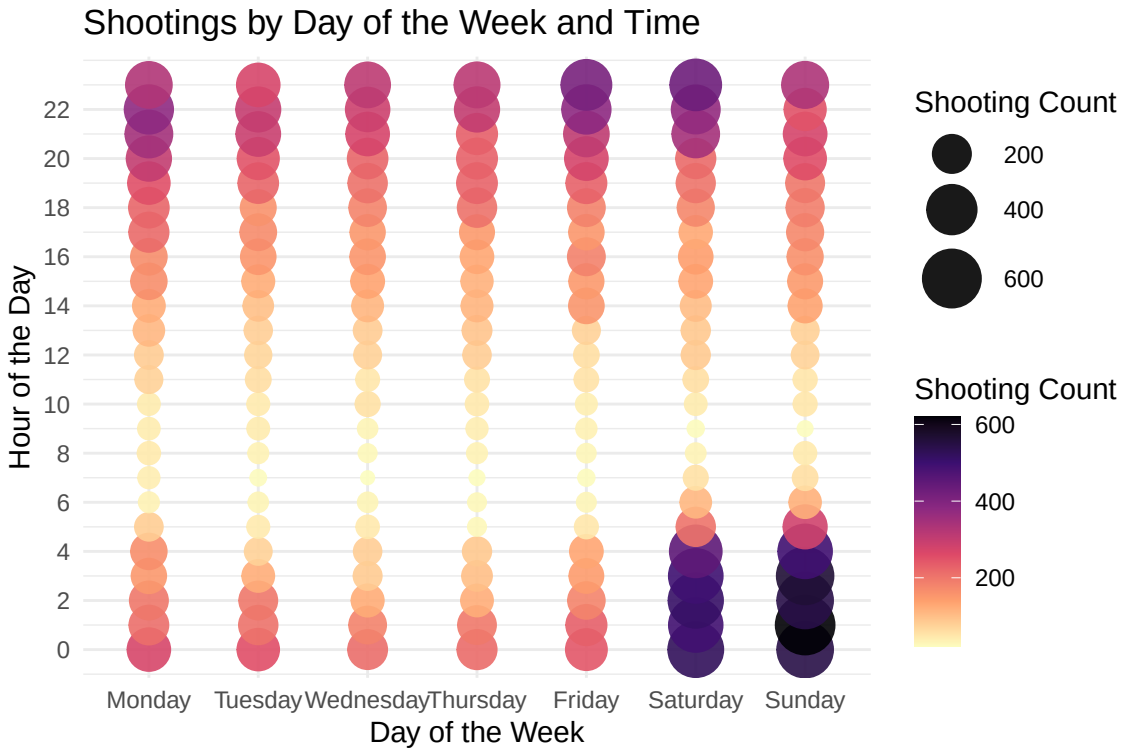
- Higher medians
- More variability (more outliers and wider boxes)
- During the weekdays, the shooting number is more stable but occurs

Observation:

Friday and Monday also shows more shootings which is consistent with the weekend proximity.

Hypothesis: Most of the shootings happens because of nightlife, to test it I'll:

- Plot the scatter day of week vs time of the shooting being the size of the marker the shooting count.
- This should show a clear pattern on the night during weekends and also Friday.



Insight: The crimes happens moslty during weekend during the night. Its associated with nightlife.
Future work: Check **where (using location)** the shootings happens the most to see if its close to bars. This can help the police to search for dangerous places.

POSSIBLE BIASES

As is a sensitive topic, its important to understand the potential risk of bias. For identifying them I'll:

- Train an explainable model using all the features when the perpetrator is identified
- Explain the values of the features that contribute the most.

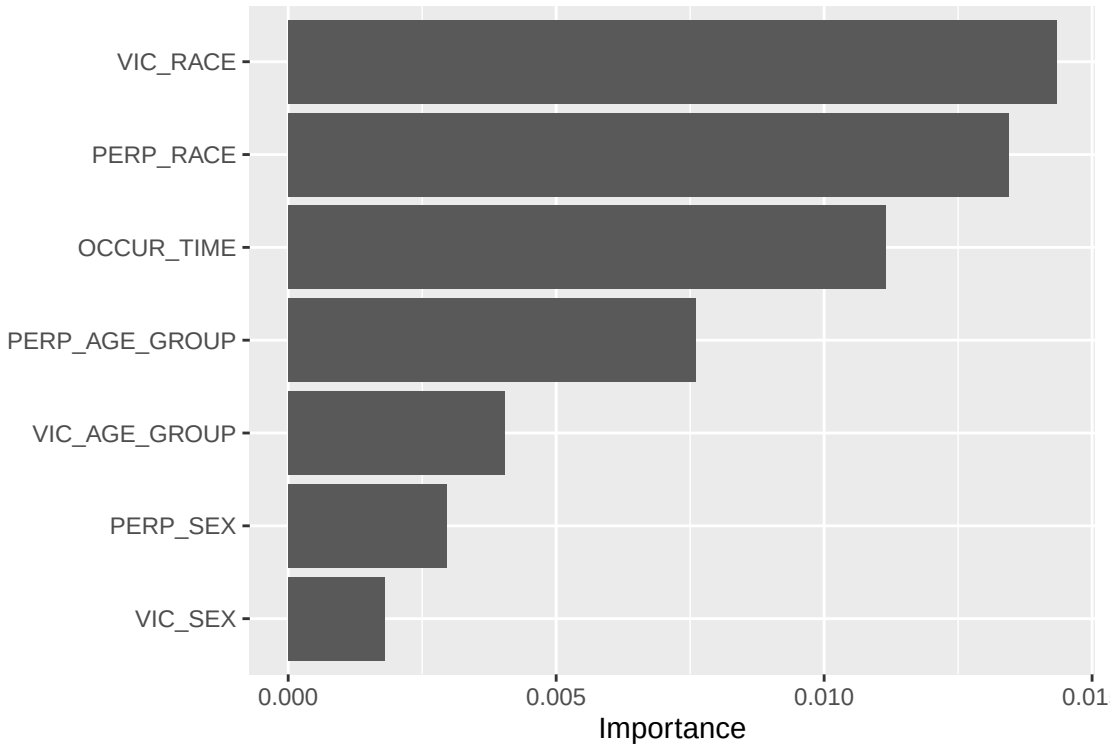
This should show the potential biases if only this dataset is used for training.

Modelization I'll train a Random Forest, the split will be 70% training and 30% test

```
### FEATURE SELECTION AND REMOVE ALL NANs
    crime_data <- nypd_shootings %>%
      select(PRECINCT, OCCUR_TIME, PERP_SEX, PERP_AGE_GROUP, PERP_RACE, VIC_SEX, VIC_AGE_GROUP, VIC_RACE
      drop_na()

### TRANSFORM THE CATEGORICAL VARIABLES INTO TRAINABLE ONES
  crime_data <- crime_data %>%
  mutate(across(c(PRECINCT, PERP_SEX, PERP_AGE_GROUP, PERP_RACE, VIC_SEX, VIC_AGE_GROUP, VIC_RACE), as.

### SPLIT THE DATASET
  set.seed(123)
  crime_split <- initial_split(crime_data, prop = 0.70)
  crime_train <- training(crime_split)
  crime_test <- testing(crime_split)

### TRAIN THE RANDOM FOREST
  model <- rand_forest(mode = "classification") %>%
    set_engine("ranger", importance="permutation") %>%
    fit(PRECINCT ~ ., data = crime_train)
```

```
vip(model)
```

SHOW FEATURE IMPORTANCE TO IDENTIFY THE BIASES



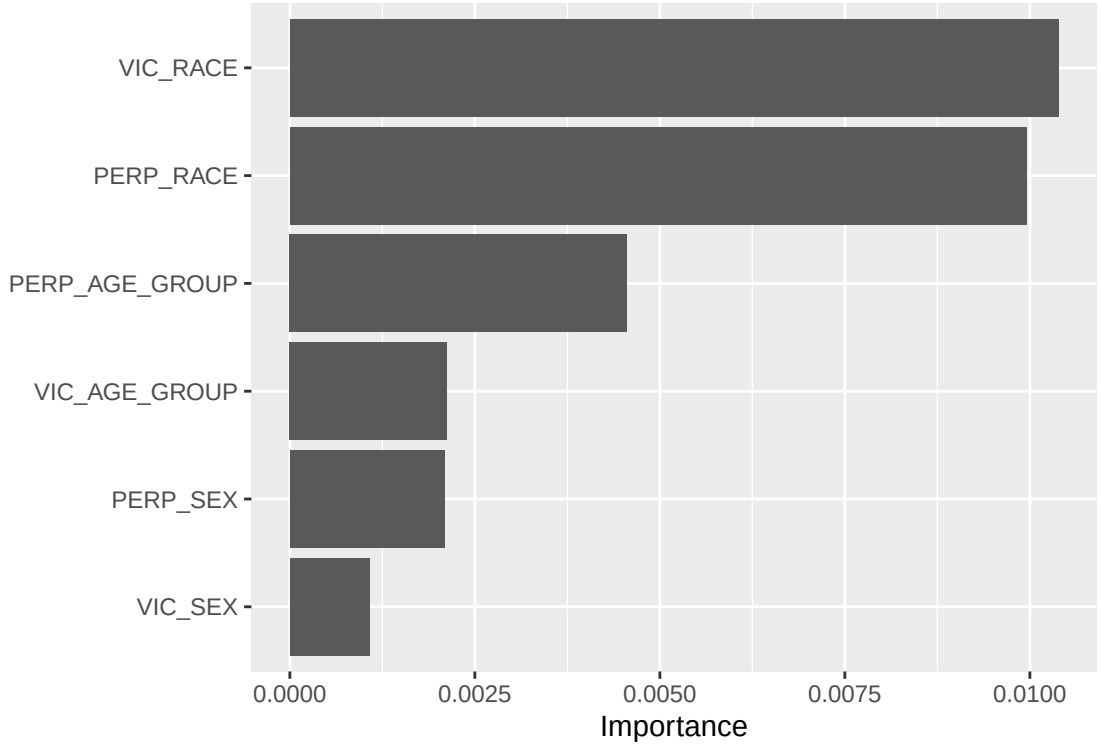
MODEL CORRECTION TO IDENTIFY BIASES

- We already estabilshed that the time its important, but not useful to confirm the bias that its appearing (RACE and PERPETRATOR age group)

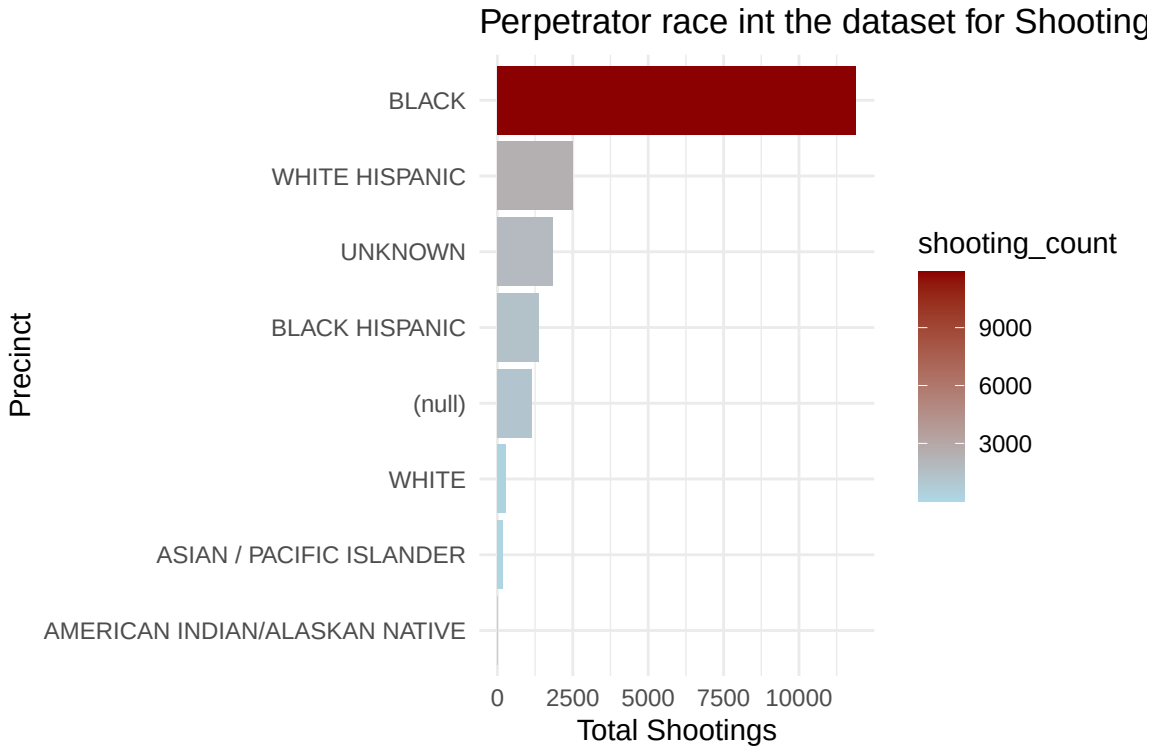
```r
### FEATURE SELECTION AND REMOVE ALL NANs
    crime_data <- nypd_shootings %>%
      select(PRECINCT, PERP_SEX, PERP_AGE_GROUP, PERP_RACE, VIC_SEX, VIC_AGE_GROUP, VIC_RACE) %>%
      drop_na()

### TRANSFORM THE CATEGORICAL VARIABLES INTO TRAINABLE ONES
  crime_data <- crime_data %>%
  mutate(across(c(PRECINCT, PERP_SEX, PERP_AGE_GROUP, PERP_RACE, VIC_SEX, VIC_AGE_GROUP, VIC_RACE),

### SPLIT THE DATASET
  set.seed(123)
  crime_split <- initial_split(crime_data, prop = 0.70)
  crime_train <- training(crime_split)
  crime_test <- testing(crime_split)

### TRAIN THE RANDOM FOREST
  corrected_model <- rand_forest(mode = "classification") %>%
    set_engine("ranger", importance="permutation") %>%
    fit(PRECINCT ~ ., data = crime_train)
```

```r
vip(corrected_model)
```

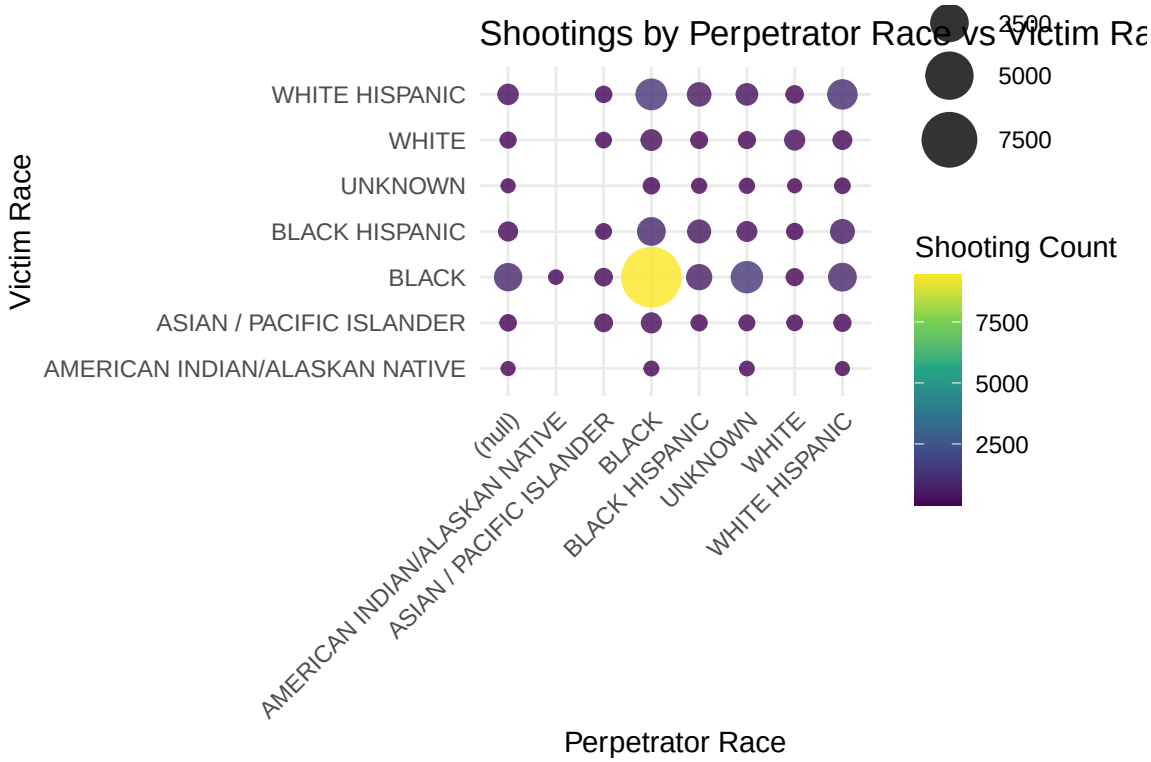
CORRECTED FEATURE IMPORTANCE



CHECK PERPETRATOR RACE SHOOTINGS DISTRIBUTION



FINALLY LETS BRIEFLY ANALYZE THE RELATIONSHIP BETWEEN PERPETRATOR RACE AND VICTIM RACE



CONCLUSIONS

- There is an increase of shootings during the **weekends**
- Most of the crimes happens during the **night**
- Most of the shootings APPEARS to happen **between same race,** in this case BLACK, but **WARNING**
  THE DATA IS **NOT NORMALIZED** by population, therefore, ITS AN ERRPR TO CONCLUDE THAT THE ANY RACE COMMITS MORE CRIMES - if certain race is more abundant, then the dataset is biased.
- The dataset may be biased, therefore, the model **IS BIASED**
- This dataset cannot be used without normalization because:
  - Training a model with this dataset, as is, will lead to incorrect predictions, as will be enough to be black have a guilty tag.
  - Given that most of the crimes happens between same race (according to this dataset), its highly likely that if the accused is black the victim is also black, increasing the chances to be incorrectly accused.
  - If the perpetrator is an American Indian or Alaskan Native and guilty, your chances to get free are very high
- Even if the model is normalized, **there is still posibility of bias by race** as, in the hypotesis that certain race is commits more crimes, the estimation will be biased to this race. Therefore, using machine learning for guiltiness estimation **is risky** and potentially racist.

FUTURE WORK

- Normalize the dataset by census (out of the scope of this work)
- Study where the crimes happens