NY Shooting Incidents Data Set

September 24, 2021

```
options(digits = 5)
```

Load and Read CSV file

```
dl <- tempfile()
download.file("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD", dl)
dat <- read_csv(dl)
file.remove(dl)
```

```
[1] FALSE
```

```
rm(dl)


# seeing and counting which columns have NA's
sum(is.na(dat$BORO))
```

```
[1] 0
```

```
sum(is.na(dat$PRECINCT))
```

```
[1] 0
```

```
sum(is.na(dat$JURISDICTION_CODE))
```

```
[1] 2
```

```
sum(is.na(dat$LOCATION_DESC))
```

```
[1] 13581
```

```
sum(is.na(dat$STATISTICAL_MURDER_FLAG))
```

```
[1] 0
```

```
sum(is.na(dat$PERP_AGE_GROUP))
```

```
[1] 8459
```

```
sum(is.na(dat$PERP_SEX))
```

```
[1] 8425
```

```
sum(is.na(dat$PERP_RACE))
```

```
[1] 8425
```

```
sum(is.na(dat$VIC_AGE_GROUP))
```

```
[1] 0
```

```
sum(is.na(dat$VIC_SEX))
```

```
[1] 0
```

```
sum(is.na(dat$VIC_RACE))
```

```
[1] 0
```

```
# proportion of NA's to the number of rows in the data set
sum(is.na(dat$JURISDICTION_CODE))/nrow(dat)
```

```
[1] 8.4861e-05
```

```
sum(is.na(dat$LOCATION_DESC))/nrow(dat)
```

```
[1] 0.57625
```

```
sum(is.na(dat$PERP_AGE_GROUP))/nrow(dat)
```

```
[1] 0.35892
```

```
sum(is.na(dat$PERP_SEX))/nrow(dat)
```

```
[1] 0.35748
```

```
sum(is.na(dat$PERP_RACE))/nrow(dat)
```

```
[1] 0.35748
```

```r
#remove the rows in jurisdiction code with NA because it has a small enough
# proportion to the data set, however, the columns location_desc, perp_age_group,
# perp_sex and perp_race have too high of a proportion to the dataset.
# removing na rows in jurisdiction code
dat <- dat[!is.na(dat$JURISDICTION_CODE),]
# make date into correct type
dat$OCCUR_DATE <- date(mdy(dat$OCCUR_DATE))
# remove a unnecessary columns
dat <- dat %>% select(-INCIDENT_KEY,
                      -Lon_Lat,
                      -X_COORD_CD,
                      -Y_COORD_CD,
                      -LOCATION_DESC,
                      -PERP_AGE_GROUP,
                      -PERP_SEX,
                      -PERP_RACE)
# make vic_race factor
dat$VIC_RACE <- factor(dat$VIC_RACE)

## CREATING WORKING AND VALIDATION SETS ##
library("caret")
y <- dat$VIC_RACE
set.seed(212, sample.kind = "Rounding")
validation_index <- createDataPartition(y, times = 1, p = 0.15, list = FALSE)
validation <- dat %>% slice(validation_index)
dat <- dat %>% slice(-validation_index)
```

Summary

Exploring the NYPD Shooting Incident (Historic) Data set. During the exploratory analysis, found a number of missing values in 4 columns: Location Description, Perpetrator Age Group, Perpetrator Sex, and Perpetrator Race. Due to their high proportion to the overall data set, I did not feel comfortable replacing the missing the values with the mean/mode nor was I comfortable with removing the entire row.After cleaning the data, I explored the data by looking at counts, proportions, and proportion of deaths in shooting incidents. I then visualized these insights in the following section before diving further into distributions and probabilities of the data set.

Exploratory Data Analysis

Preliminary Data Exploration

The overall NYPD Shooting Incident Data set has 23566 rows.

11 columns in the data:

- OCCUR_DATE `<date>` contains the date of the shooting incident.
- OCCUR_TIME `<time>` contains the time of the shooting incident.
- BOROUGH `<character>` contains the borough for where the shooting incident took place in New York City.
- PRECINCT `<numeric>` contains the NYPD precinct that responded to the shooting incident.
- JURISDICTION_CODE `<numeric>` contains the jurisdiction code with respect to the shooting incident.
- STATISTICAL_MURDER_FLAG `<logical>` contains TRUE for a shooting incident causing death and FALSE fora nonfatal shooting incident.
- VIC_AGE_GROUP `<character>` contains age ranges for which the victim of the shooting incident belongs to.
- VIC_SEX `<character>` contains genders for which the victim of the shooting incident belongs to.
- VIC_RACE `<factor>` contains races for which the victim of the shooting incident belongs to. This is the variable we are interested in predicting.
- Longitude `<numeric>` contains the longitudinal geographic coordinate for the shooting incident.
- Latitude `<numeric>` contains the latitudinal geographic coordinate for the shooting incident.

Explore some of the values in the columns:

- 2006-01-01 is the earliest shooting incident date as found in OCCUR_DATE.
- 2020-12-31 is the latest shooting incident date as found in OCCUR_DATE.
- QUEENS, BRONX, MANHATTAN, STATEN ISLAND, BROOKLYN are all the different boroughs in New York City under the BORO column.
- 103, 40, 23, 121, 46, 73, 81, 67, 101, 120, 75, 113, 78, 45, 49, 105, 61, 48, 47, 25, 44, 52, 114, 34, 71, 69, 102, 63, 30, 60, 77, 42, 41, 83, 79, 43, 88, 109, 26, 32, 110, 28, 108, 106, 62, 33, 9, 5, 70, 90, 84, 72, 13, 115, 112, 122, 7, 107, 100, 20, 50, 10, 104, 24, 123, 1, 94, 76, 14, 66, 68, 6, 18, 19, 111, 17, 22 are all the different precincts in New York City under the PRECINCT column.
- 0, 1, 2 are the jurisdiction codes in New York City under JURISDICTION_CODE.
- 25-44, 18-24, 45-64, <18, 65+, UNKNOWN are the different age groups related to victims of shooting incidents in VIC_AGE_GROUP.
- M, F, U are the different genders related to victims of shooting incidents in VIC_SEX.
- BLACK, BLACK HISPANIC, WHITE HISPANIC, WHITE, UNKNOWN, ASIAN / PACIFIC ISLANDER, AMERICAN INDIAN/ALASKAN NATIVE are the different races related to victims of shooting incidents in VIC_RACE.
- 0.19058 is the proportion of deaths caused by shooting incidents in STATISTICAL_MURDER_FLAG.

Advanced Data Exploration and Analysis

Shooting Incidents grouped by Borough

Interested to see if there is a borough more likely to have shooting incidents and whether or not those shooting incidents are more likely to result in death.

```
boro_incidents <- dat %>%
  group_by(BORO) %>%
  summarize(count = n(),
            prop = count/nrow(dat),
            prop_death = mean(STATISTICAL_MURDER_FLAG))
boro_incidents %>%
  arrange(desc(count)) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                position = "center",
                font_size = 10,
                full_width = FALSE,
                latex_options = "HOLD_position")
```

BORO	count	prop	prop_death
BROOKLYN	8270	0.41292	0.19214
BRONX	5685	0.28385	0.18734
QUEENS	3008	0.15019	0.19914
MANHATTAN	2459	0.12278	0.18056
STATEN ISLAND	606	0.03026	0.19802

Top 10 Shooting Incidents grouped by Precinct

Now explore which precincts have the most shooting incidents in New York City? Are some precinct shooting incidents more likely to result in death than others? Observe the top 10 precincts invovled in shooting incidents.

```
precinct_incidents <- dat %>%
  group_by(PRECINCT) %>%
  summarize(count = n(),
            prop = count/nrow(dat),
            prop_death = mean(STATISTICAL_MURDER_FLAG))
precinct_incidents %>%
  arrange(desc(count)) %>%
  top_n(10) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                position = "center",
                font_size = 10,
                full_width = FALSE,
                latex_options = "HOLD_position")
```

PRECINCT	count	prop	prop_death
106	159	0.00794	0.29560
61	120	0.00599	0.28333
24	75	0.00374	0.28000
107	65	0.00325	0.26154
122	48	0.00240	0.39583
5	40	0.00200	0.30000
6	22	0.00110	0.27273
1	18	0.00090	0.33333
112	16	0.00080	0.37500
17	4	0.00020	0.50000

Shooting Incidents grouped by Jurisdiction Code

How are shooting incidents related to jurisdiction codes? Are certain jurisdiction codes more involved in shooting incidents?

```
jurisdiction_incidents <- dat %>%
  group_by(JURISDICTION_CODE) %>%
  summarize(count = n(),
            prop = count/nrow(dat),
            prop_death = mean(STATISTICAL_MURDER_FLAG))
jurisdiction_incidents %>%
  arrange(desc(count)) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                position = "center",
                font_size = 10,
                full_width = FALSE,
                latex_options = "HOLD_position")
```

JURISDICTION_CODE	count	prop	prop_death
0	16666	0.83214	0.19735
2	3312	0.16537	0.15640
1	50	0.00250	0.20000

Shooting Incidents grouped by Victim Age Group

What age groups are more likely to be involved in shooting incidents? Explore the age groups and the number of shooting incidents, proportion to total shooting incidents, and proportion to death.

```
victim_age_incidents <- dat %>%
  group_by(VIC_AGE_GROUP) %>%
  summarize(count = n(),
            prop = count/nrow(dat),
            prop_death = mean(STATISTICAL_MURDER_FLAG))
victim_age_incidents %>%
  arrange(desc(count)) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                position = "center",
```

```
                font_size = 10,
                full_width = FALSE,
                latex_options = "HOLD_position")
```

VIC_AGE_GROUP	count	prop	prop_death
25-44	8734	0.43609	0.21777
18-24	7647	0.38182	0.16386
<18	2160	0.10785	0.13056
45-64	1310	0.06541	0.24809
65+	125	0.00624	0.32800
UNKNOWN	52	0.00260	0.26923

Shooting Incidents grouped by Victim Sex

Is one gender more likely to be involved in shooting incidents? Look at all genders and their involvement in shooting incidents.

```
victim_sex_incidents <- dat %>%
  group_by(VIC_SEX) %>%
  summarize(count = n(),
            prop = count/nrow(dat),
            prop_death = mean(STATISTICAL_MURDER_FLAG))
victim_sex_incidents %>%
  arrange(desc(count)) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                position = "center",
                font_size = 10,
                full_width = FALSE,
                latex_options = "HOLD_position")
```

VIC_SEX	count	prop	prop_death
M	18141	0.90578	0.18990
F	1867	0.09322	0.19871
U	20	0.00100	0.05000

Shooting Incidents grouped by Victim Race

What is the relationship between victim race and shooting incidents? Are some races more likely to be involved in shooting incidents compared to others? Are some races more likely to die?

```
victim_race_incidents <- dat %>%
  group_by(VIC_RACE) %>%
  summarize(count = n(),
            prop = count/nrow(dat),
            prop_death = mean(STATISTICAL_MURDER_FLAG))
victim_race_incidents %>%
  arrange(desc(count)) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
```

```
        position = "center",
        font_size = 10,
        full_width = FALSE,
        latex_options = "HOLD_position")
```

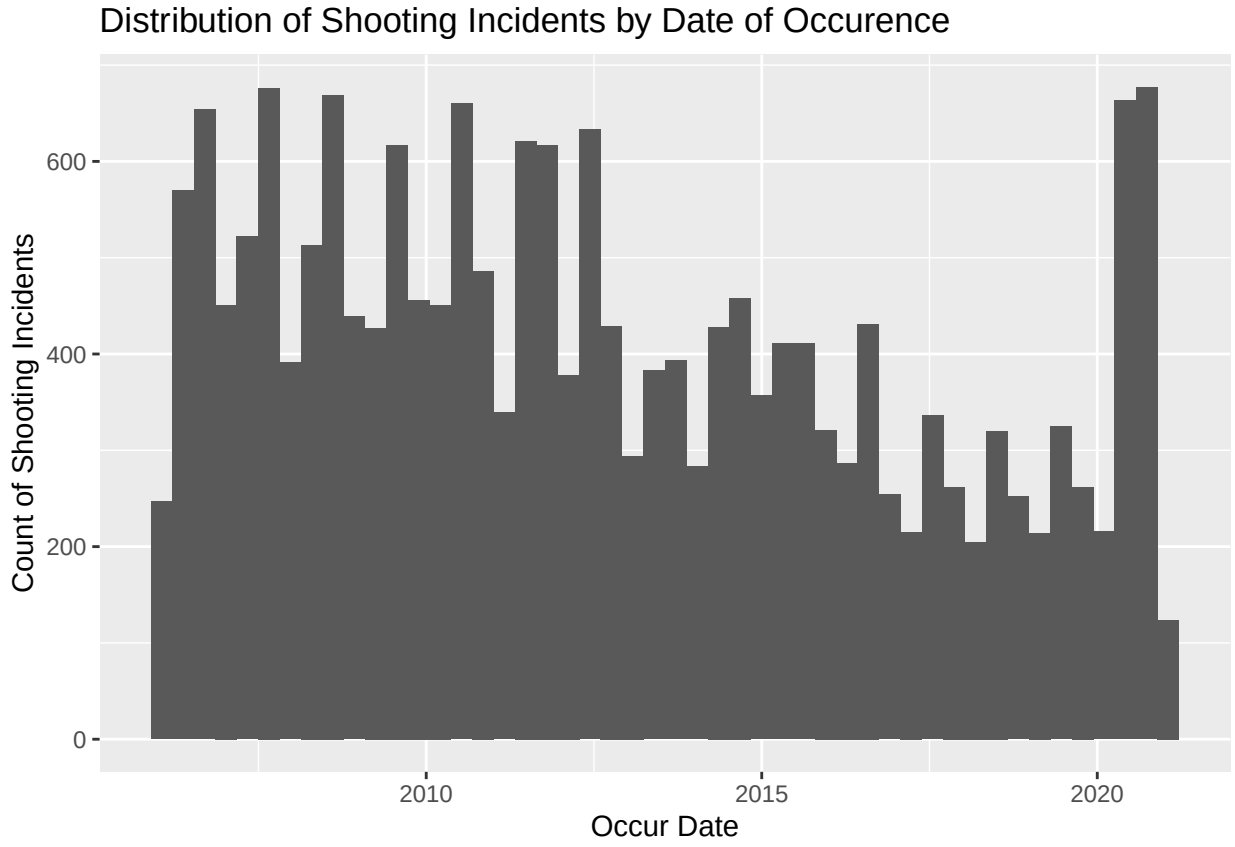
VIC_RACE	count	prop	prop_death
BLACK	14318	0.71490	0.18732
WHITE HISPANIC	2916	0.14560	0.21125
BLACK HISPANIC	1907	0.09522	0.15417
WHITE	522	0.02606	0.27395
ASIAN / PACIFIC ISLANDER	272	0.01358	0.25000
UNKNOWN	86	0.00429	0.16279
AMERICAN INDIAN/ALASKAN NATIVE	7	0.00035	0.00000

Data Visualization

Distribution Plots

Looking at the distribution of shooting incidents by occurrence date. This gives us a better idea of when shooting incidents were more likely to occur historically and if there is a sesonality effect.

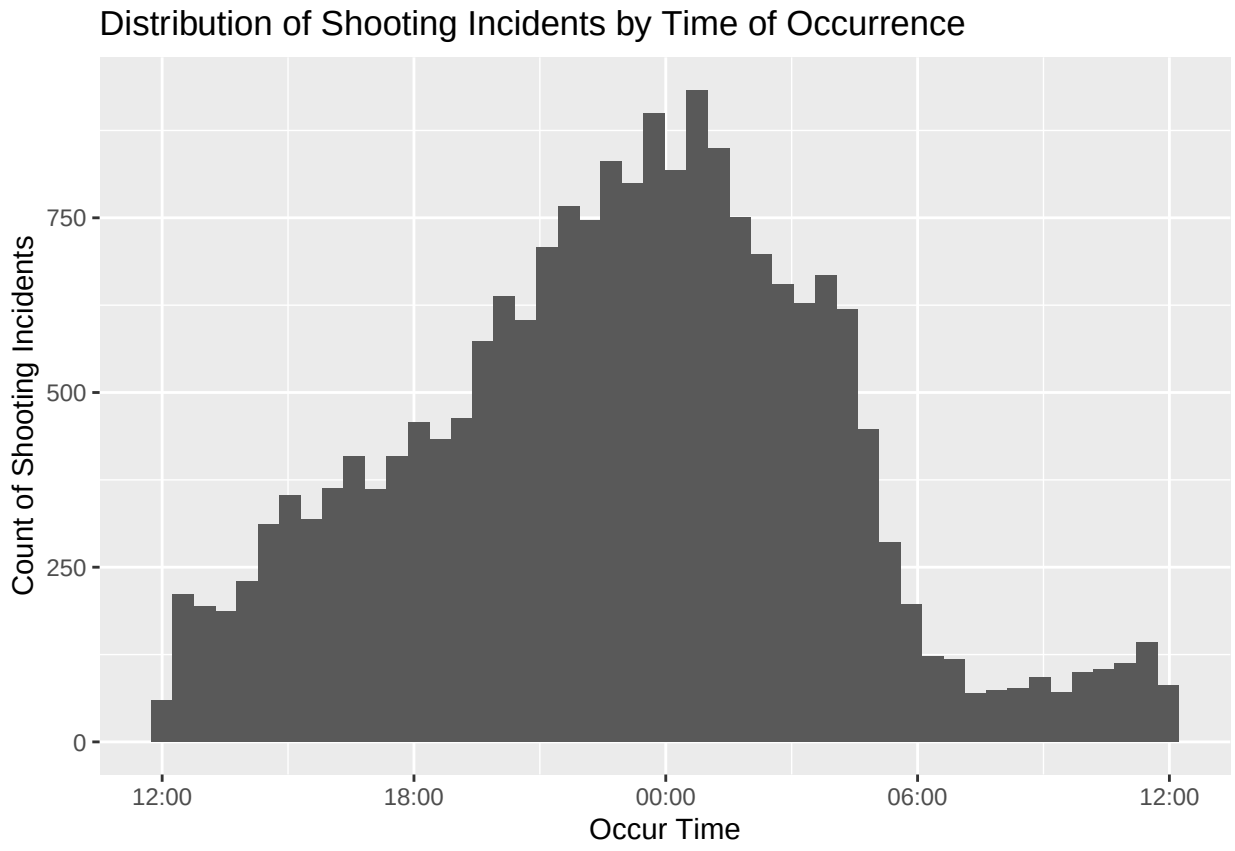
```
dat %>%
  ggplot(aes(x = OCCUR_DATE)) +
  geom_histogram(bins = 48) +
  xlab("Occur Date") +
  ylab("Count of Shooting Incidents") +
  ggtitle("Distribution of Shooting Incidents by Date of Occurence")
```



Distribution of shooting incidents grouped by occurrence time. Most values tend to center around midnight, below is a visualization of the findings.

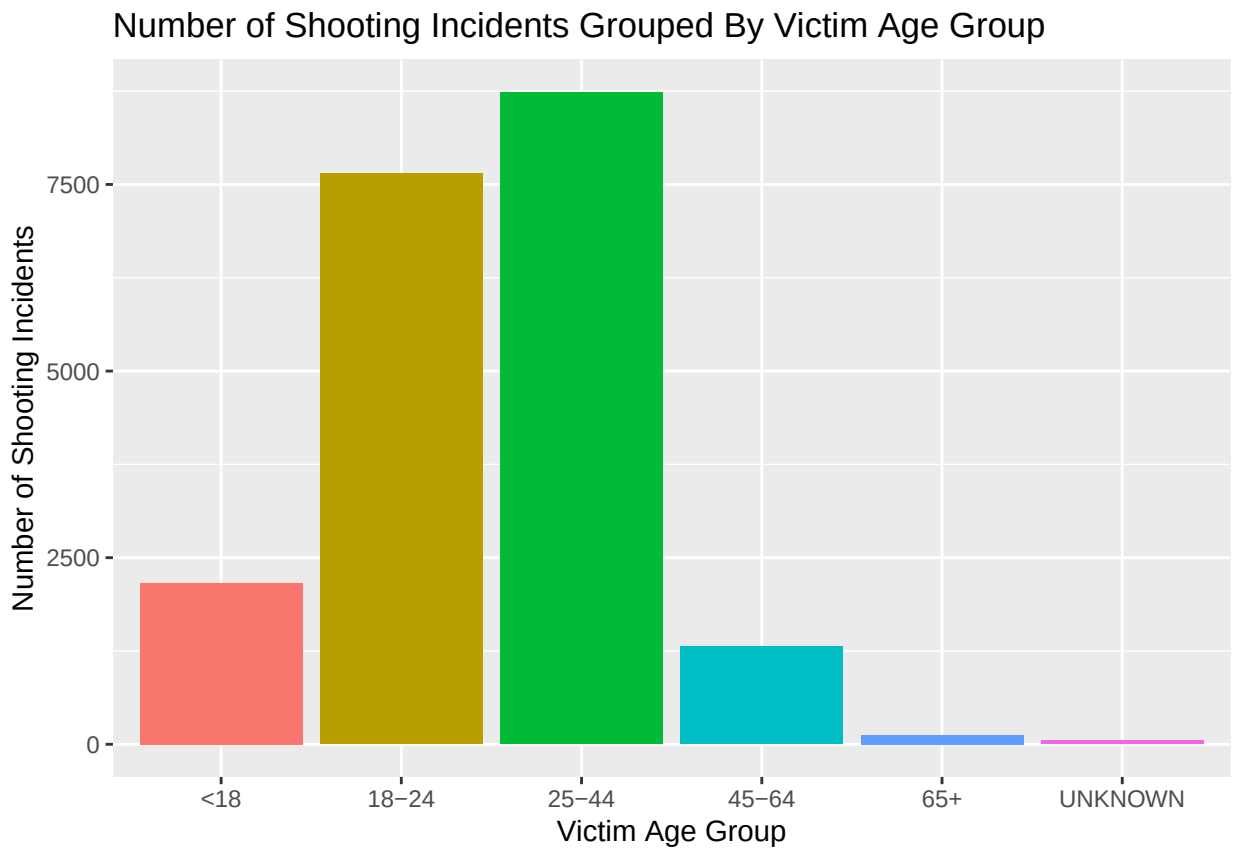
```
justtime <- function(x, split=12) {
  h <- as.numeric(strftime(x, "%H"))
  y <- as.POSIXct(paste(ifelse(h<split, "2015-01-02","2015-01-01"),strftime(x, "%H:%M:%S")))
}
dat %>%
  mutate(time = justtime(OCCUR_TIME)) %>%
  ggplot(aes(time)) +
```

```
  geom_histogram(bins = 48) +
  scale_x_datetime(labels = function(x) format(x, format = "%H:%M")) +
  xlab("Occur Time") +
  ylab("Count of Shooting Incidents") +
  ggtitle("Distribution of Shooting Incidents by Time of Occurrence")
```



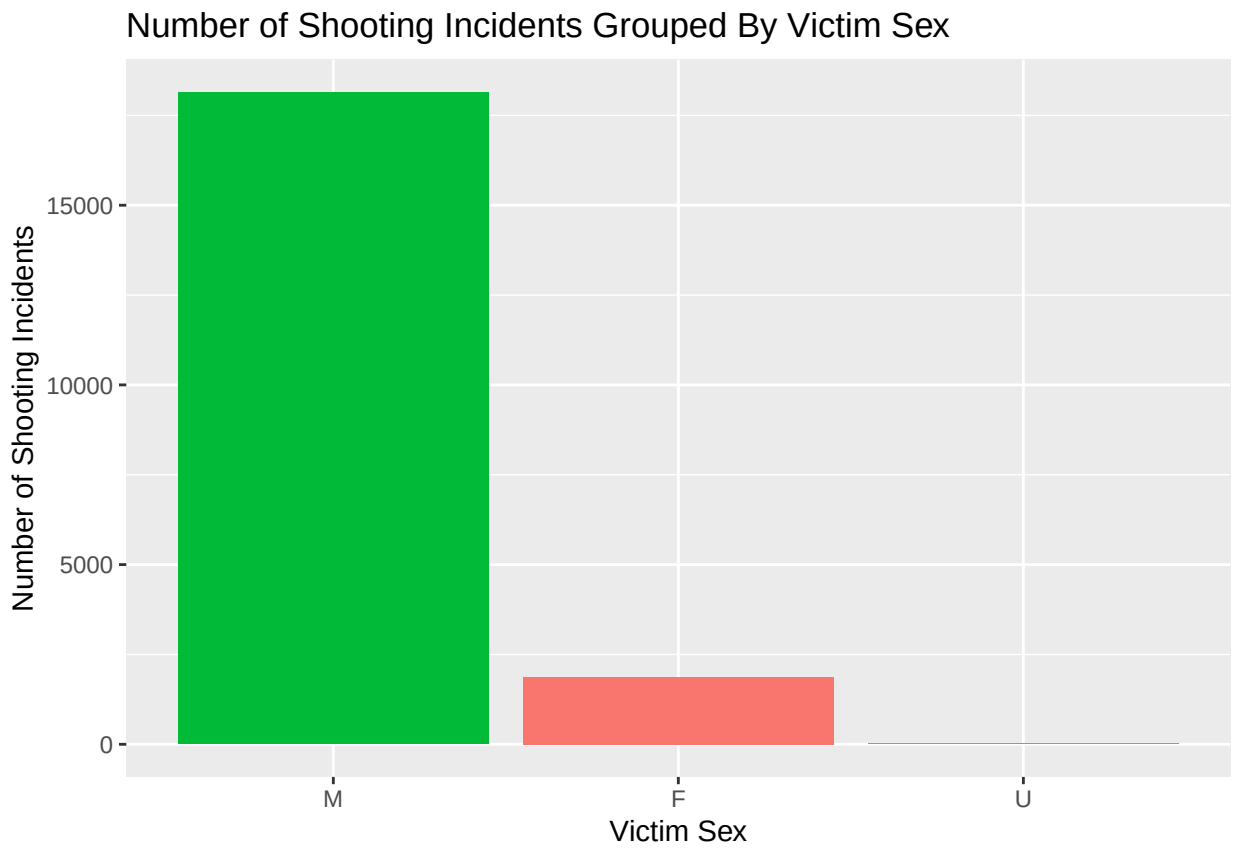
Visualizing the findings by victim Age group.

```
victim_age_incidents %>%
  ggplot(aes(VIC_AGE_GROUP, y = count, fill = VIC_AGE_GROUP)) +
  geom_bar(stat = "identity") +
  ylab("Number of Shooting Incidents") +
  xlab("Victim Age Group") +
  ggtitle("Number of Shooting Incidents Grouped By Victim Age Group") +
  theme(legend.position = "none")
```



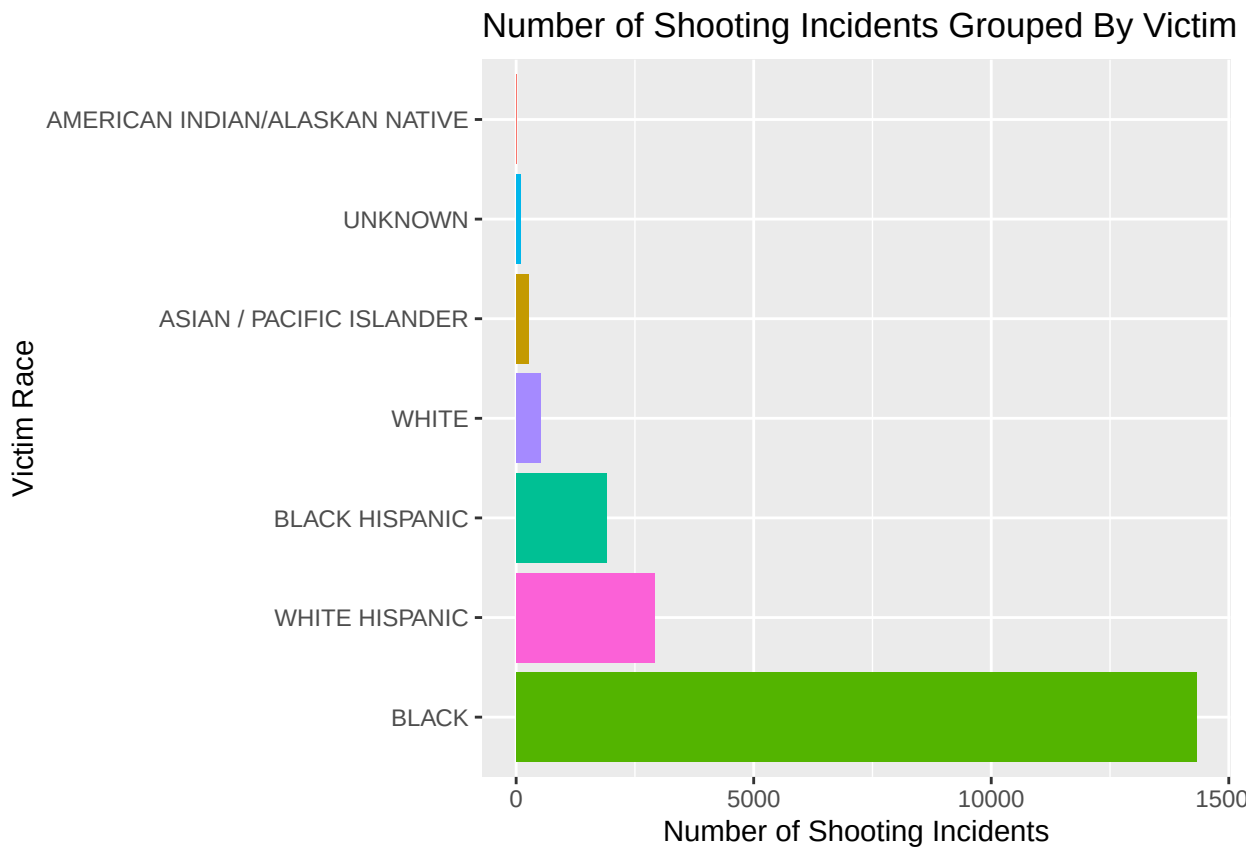
Visualizing the findings by victim Sex.

```
victim_sex_incidents %>%
  ggplot(aes(reorder(VIC_SEX, -count), y = count, fill = VIC_SEX)) +
  geom_bar(stat = "identity") +
  ylab("Number of Shooting Incidents") +
  xlab("Victim Sex") +
  ggtitle("Number of Shooting Incidents Grouped By Victim Sex") +
  theme(legend.position = "none")
```



Visualizing the findings by victim race.

```
victim_race_incidents %>%
  ggplot(aes(count, y = reorder(VIC_RACE, - count), fill = VIC_RACE)) +
  geom_bar(stat = "identity") +
  ylab("Victim Race") +
  xlab("Number of Shooting Incidents") +
  ggtitle("Number of Shooting Incidents Grouped By Victim Race") +
  theme(legend.position = "none")
```


Number of Shooting Incidents Grouped By Victim
AMERICAN INDIAN/ALASKAN NATIVE
UNKNOWN
ASIAN / PACIFIC ISLANDER
WHITE
BLACK HISPANIC
WHITE HISPANIC
BLACK
Victim Race
0
5000
10000
15000
Number of Shooting Incidents

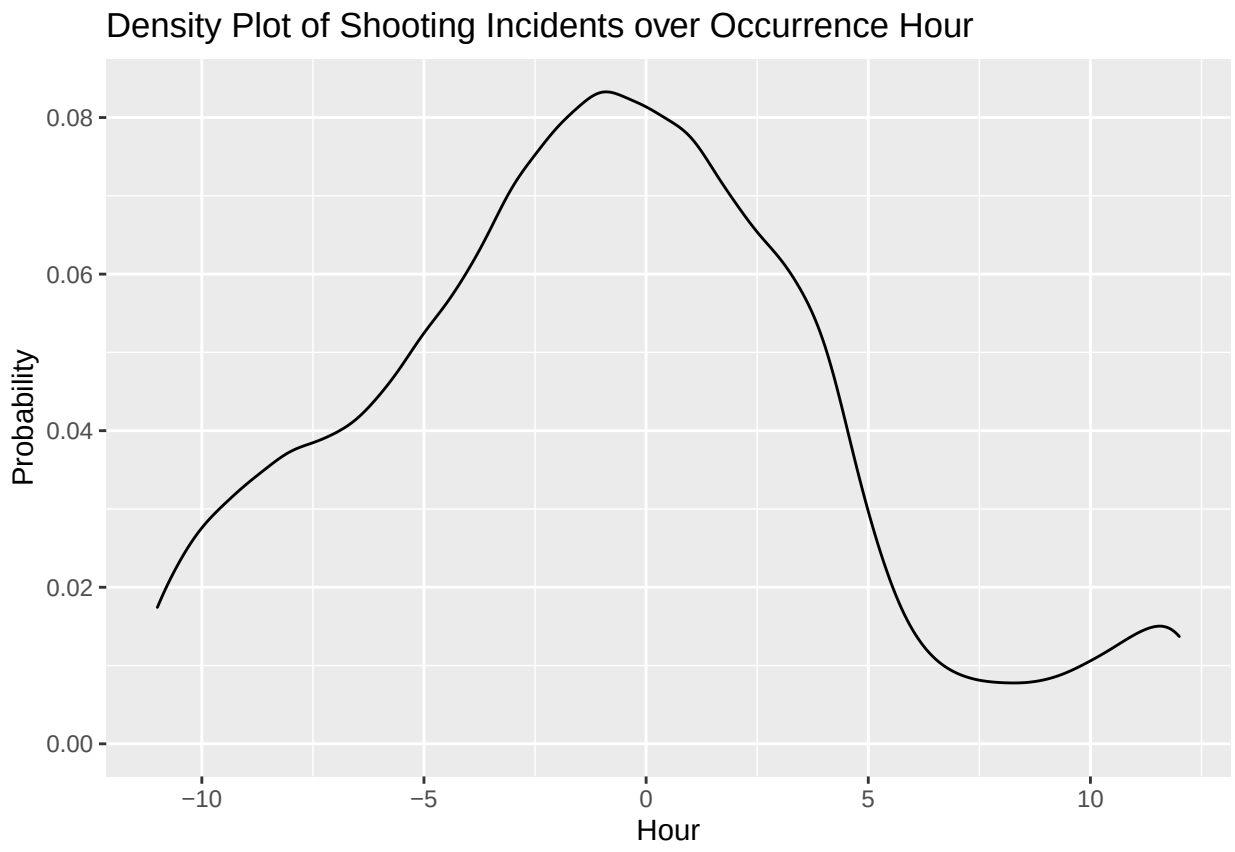
Distribution & Probability Analysis

Density Distributions

Shooting Incidents over Occurrence Hour

We will now stratify the occurrence time into occurrence hour centered around midnight and then we plot a density plot to see what times shooting incidents most likely happen. We see that most shootings happen at or before midnight and shootings rarely occur past 5 am.

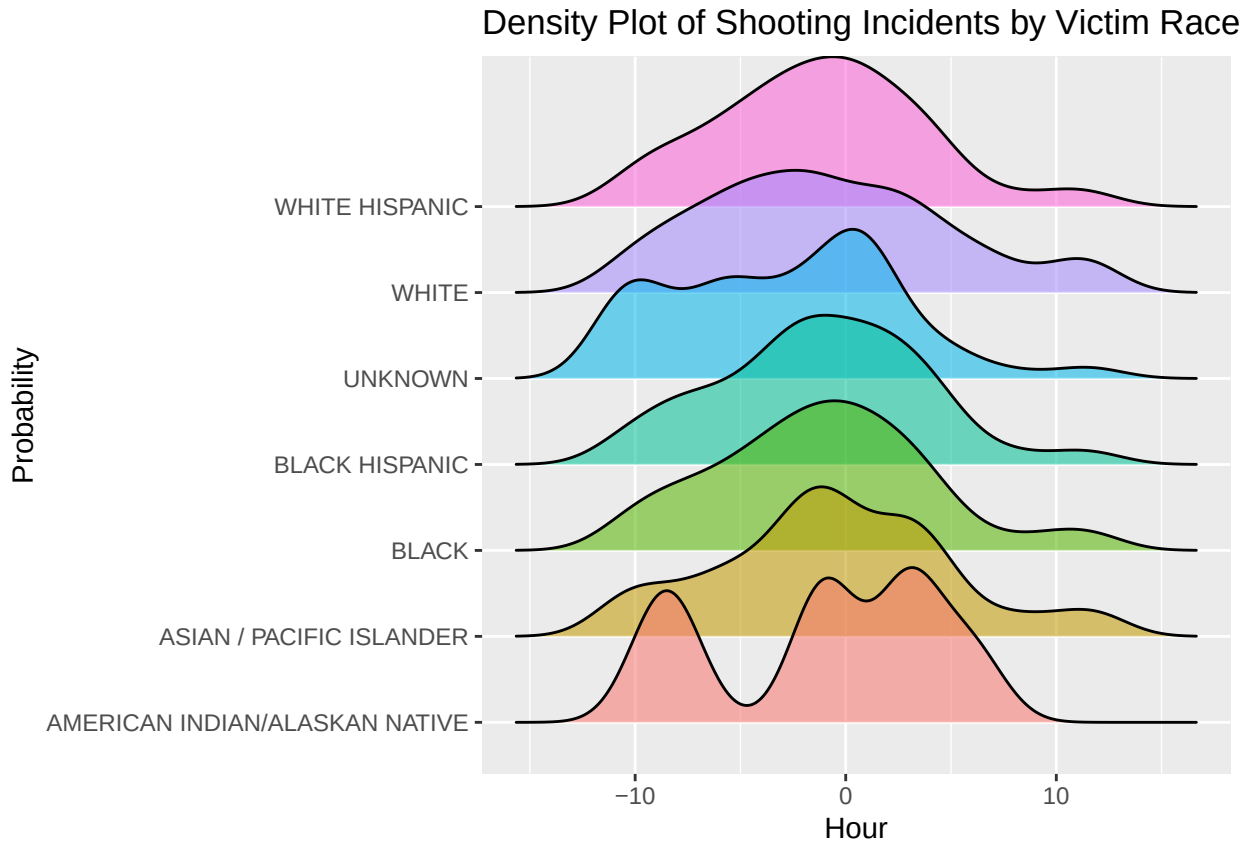
```
dat %>%
  mutate(hour = ifelse(hour(OCCUR_TIME) > 12, hour(OCCUR_TIME) - 24, hour(OCCUR_TIME))) %>%
  group_by(hour) %>%
  ggplot(aes(x = hour)) +
  geom_density() +
  xlab("Hour") +
  ylab("Probability") +
  ggtitle("Density Plot of Shooting Incidents over Occurrence Hour")
```



Shooting Incidents over Occurence Hour split by Victim Race

Now let's explore if any one race tends to have a more distinct time for when a shooting incident is to occur. As we can see visually, there tends to be no difference between races, however, we can note that `AMERICAN INDIAN/ALASKAN NATIVE` has a lower likelihood at around evening time.

```
dat %>%
  mutate(hour = ifelse(hour(OCCUR_TIME) > 12, hour(OCCUR_TIME) - 24, hour(OCCUR_TIME))) %>%
  group_by(hour, VIC_RACE) %>%
  ggplot(aes(x = hour, y = VIC_RACE)) +
  geom_density_ridges(aes(fill = VIC_RACE), alpha = 0.55) +
  xlab("Hour") +
  ylab("Probability") +
  ggtitle("Density Plot of Shooting Incidents by Victim Race over Occurence Hour") +
  theme(legend.position = "none")
```



Murder and Victim Race

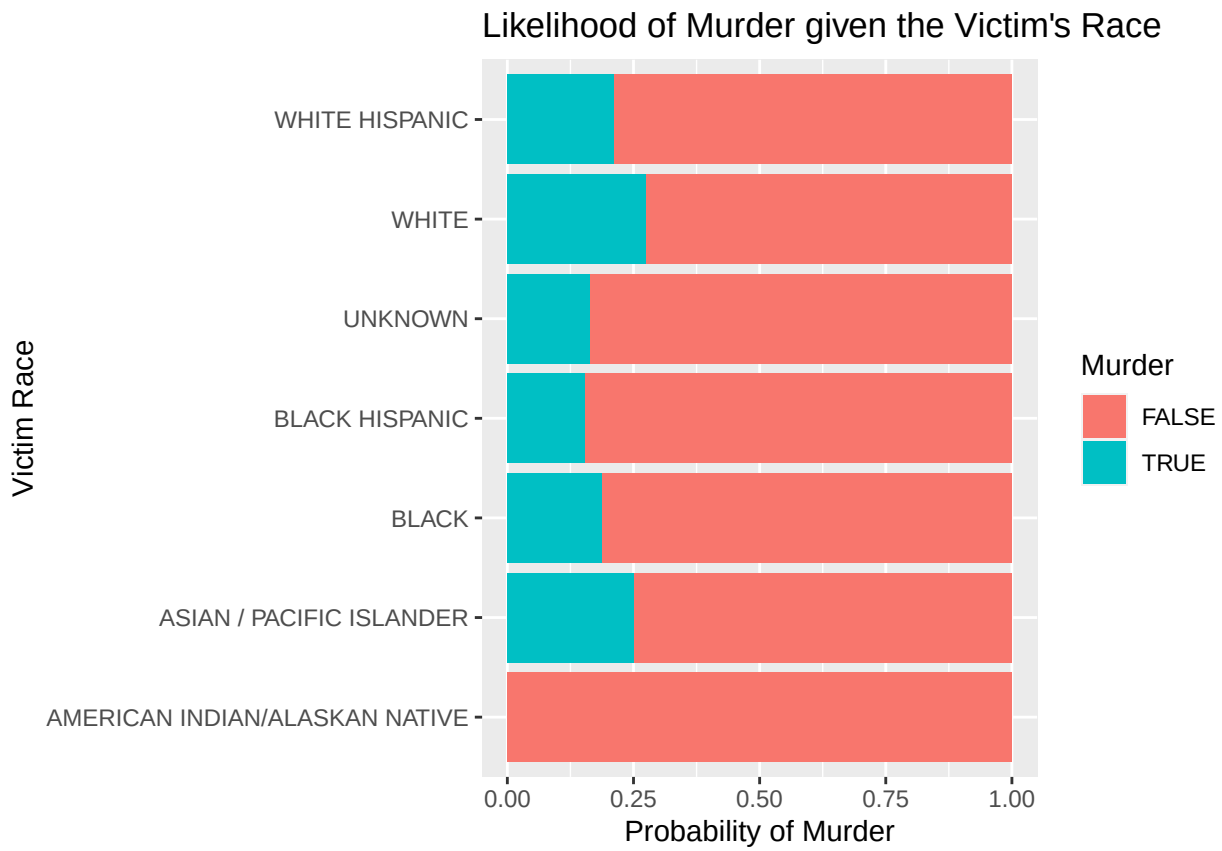
Is one race more likely to be murdered in the even of a shooting? We group the data by victim race to find out. Across all races except `AMERICAN INDIAN/ALASKAN NATIVE`, murder rates tend to be similar. `AMERICAN INDIAN/ALASKAN NATIVE` is the only victim race to have no murders from shootings.

```
race_by_murder <- dat %>%
  group_by(VIC_RACE, STATISTICAL_MURDER_FLAG) %>%
```

```
  summarize(count = n()) %>%
  mutate(prob = count / sum(count))
race_by_murder %>%
  arrange(desc(prob)) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                position = "center",
                font_size = 10,
                full_width = FALSE,
                latex_options = "HOLD_position")
```

VIC_RACE	STATISTICAL_MURDER_FLAG	count	prob
AMERICAN INDIAN/ALASKAN NATIVE	FALSE	7	1.00000
BLACK HISPANIC	FALSE	1613	0.84583
UNKNOWN	FALSE	72	0.83721
BLACK	FALSE	11636	0.81268
WHITE HISPANIC	FALSE	2300	0.78875
ASIAN / PACIFIC ISLANDER	FALSE	204	0.75000
WHITE	FALSE	379	0.72605
WHITE	TRUE	143	0.27395
ASIAN / PACIFIC ISLANDER	TRUE	68	0.25000
WHITE HISPANIC	TRUE	616	0.21125
BLACK	TRUE	2682	0.18732
UNKNOWN	TRUE	14	0.16279
BLACK HISPANIC	TRUE	294	0.15417

```
race_by_murder %>%
  ggplot(aes(prob, VIC_RACE, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(stat = "identity") +
  xlab("Probability of Murder") +
  ylab("Victim Race") +
  ggtitle("Likelihood of Murder given the Victim's Race") +
  guides(fill = guide_legend(title = "Murder"))
```


Likelihood of Murder given the Victim's Race
Victim Race
WHITE HISPANIC
WHITE
UNKNOWN
BLACK HISPANIC
BLACK
ASIAN / PACIFIC ISLANDER
AMERICAN INDIAN/ALASKAN NATIVE
0.00
0.25
0.50
0.75
1.00
Probability of Murder
Murder
FALSE
TRUE

Conclusion

To conclude, the data showed us that most shootings happen at or before midnight and shootings rarely occur past 5 am. We also see shooting incidents rise in 2020 and 0.19058 is the proportion of deaths caused by shooting incidents. When looking at victims' demographics 0.90578 are male and 0.71490 are Black. I was not able to explore the location description since the proportion of NA's was high but it would have been interesting to see where the shooting incidents are occuring such as public spaces or private homes, etc.

Bias

My personal bias regarding this topic was influenced by news and I thought most shooting incidents would occur in Queens or the Bronx along with thinking most shooting incidents would be male. What I did to mitigate this prior bias is to let the data speak for itself and exploring the data from all angles. Therefore, with that in mind I was actually able to find that majority of the shooting incidents occur in Brooklyn as opposed to what I thought. It was great letting the data be able to tell a story and prove some of my bias wrong.