NYPD_project

2025-03-18

The goal of this project is to explore the NYPD shooting data by creating a few visuals and a model.

Import Data:

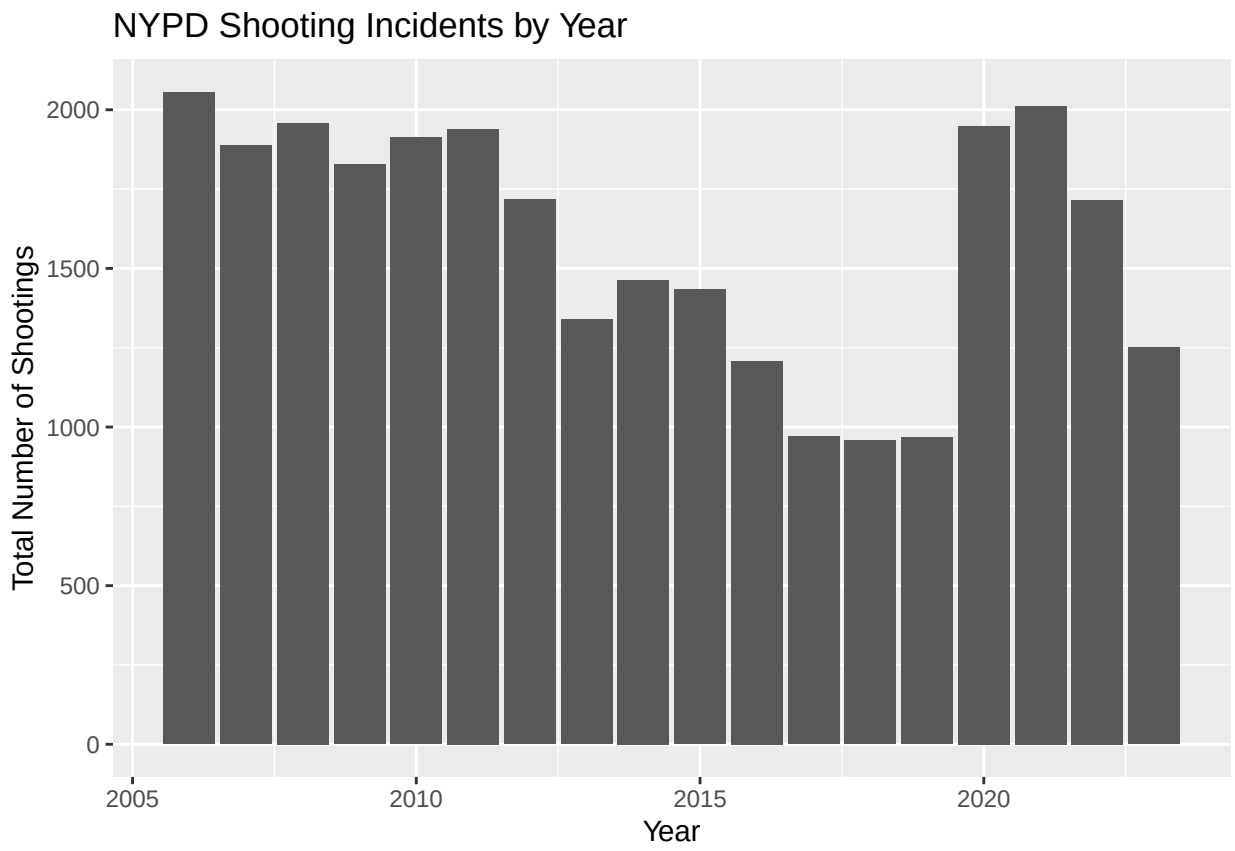
```
url_nypd <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read.csv(url_nypd)
#summary(nypd_data)
```

Clean the data by limiting to variables of interest, converting occur date and time to actual date and time variables.

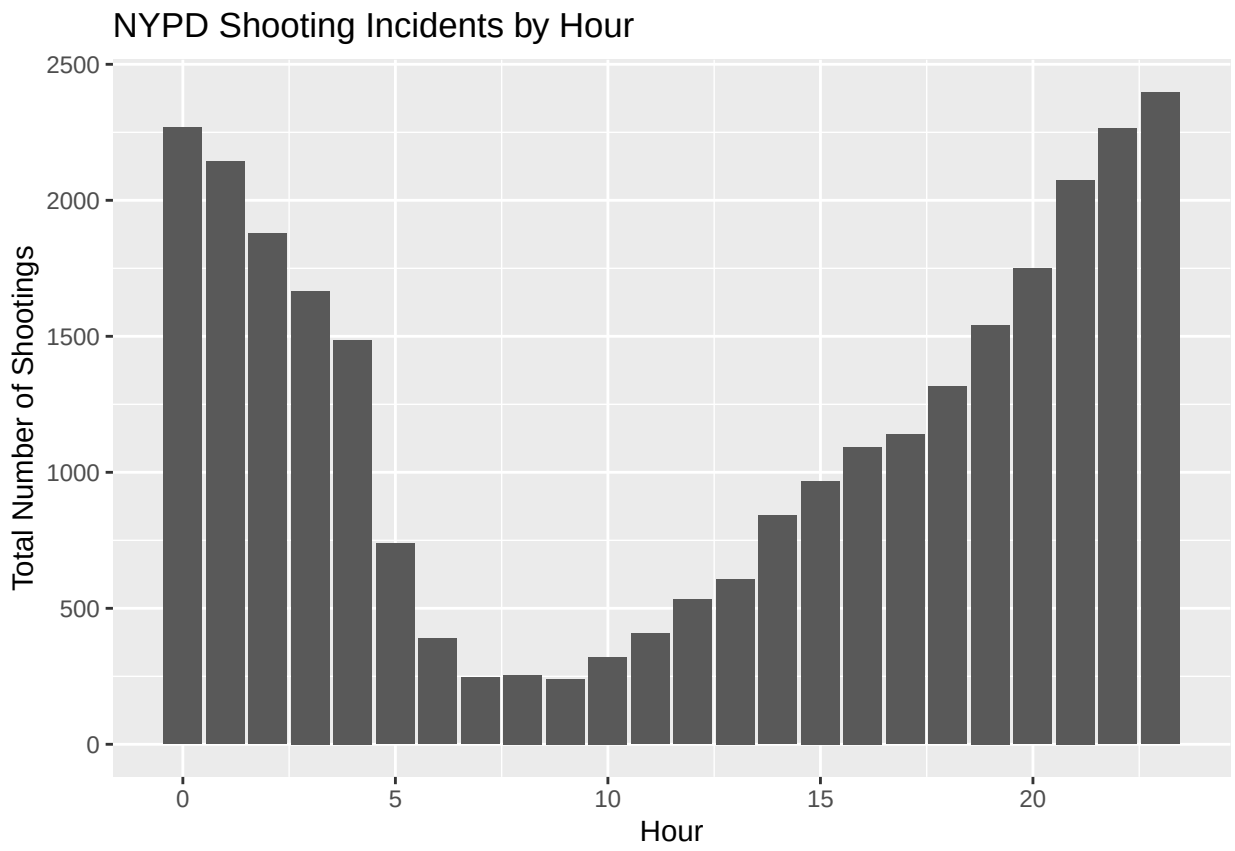
```
nypd_clean <- nypd_data %>%
  select(c("OCCUR_DATE","OCCUR_TIME", "BORO","PRECINCT",
           "STATISTICAL_MURDER_FLAG","VIC_AGE_GROUP","VIC_SEX","VIC_RACE"
           , "PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE" )) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME),
         STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
         Shootings = 1,
         Year = year(OCCUR_DATE),
         Month = month(OCCUR_DATE),
         Hour = hour(OCCUR_TIME))
```

Various graphs looking at the nubmer of shootings by year, hour, boroughs, precincts, perp data, and killings.

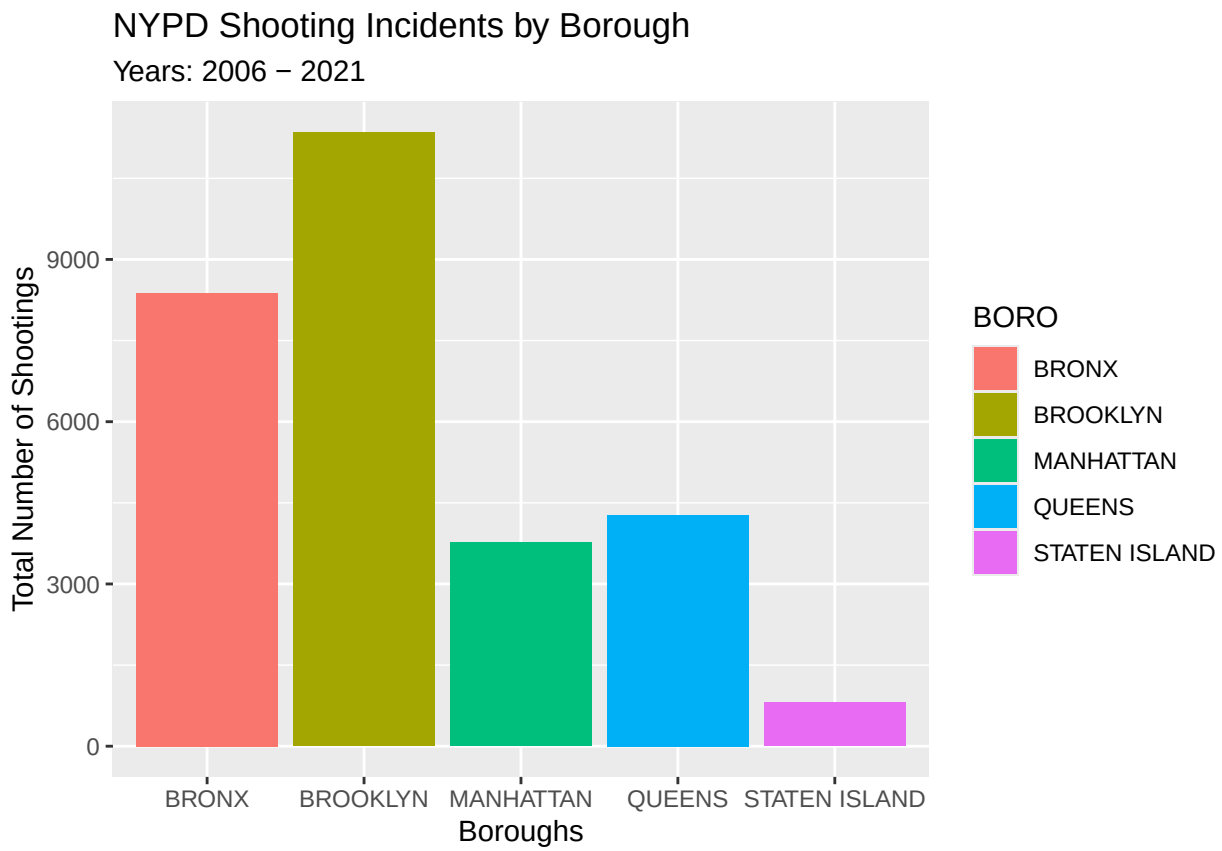
```
nypd_clean %>%
  ggplot(aes(x = Year)) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Year",
       x = "Year",
       y = "Total Number of Shootings")
```



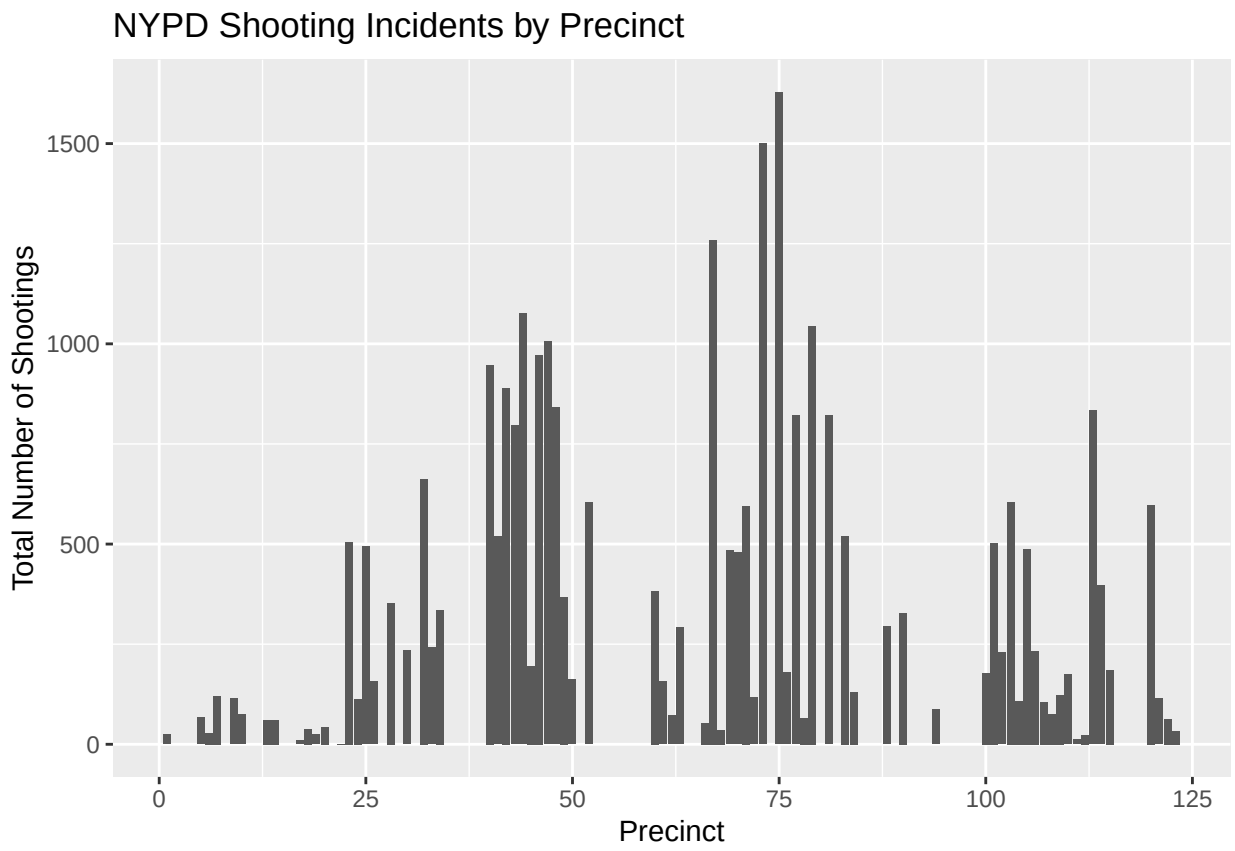

```
nypd_clean %>%
  ggplot(aes(x = Hour)) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Hour",
       x = "Hour",
       y = "Total Number of Shootings")
```



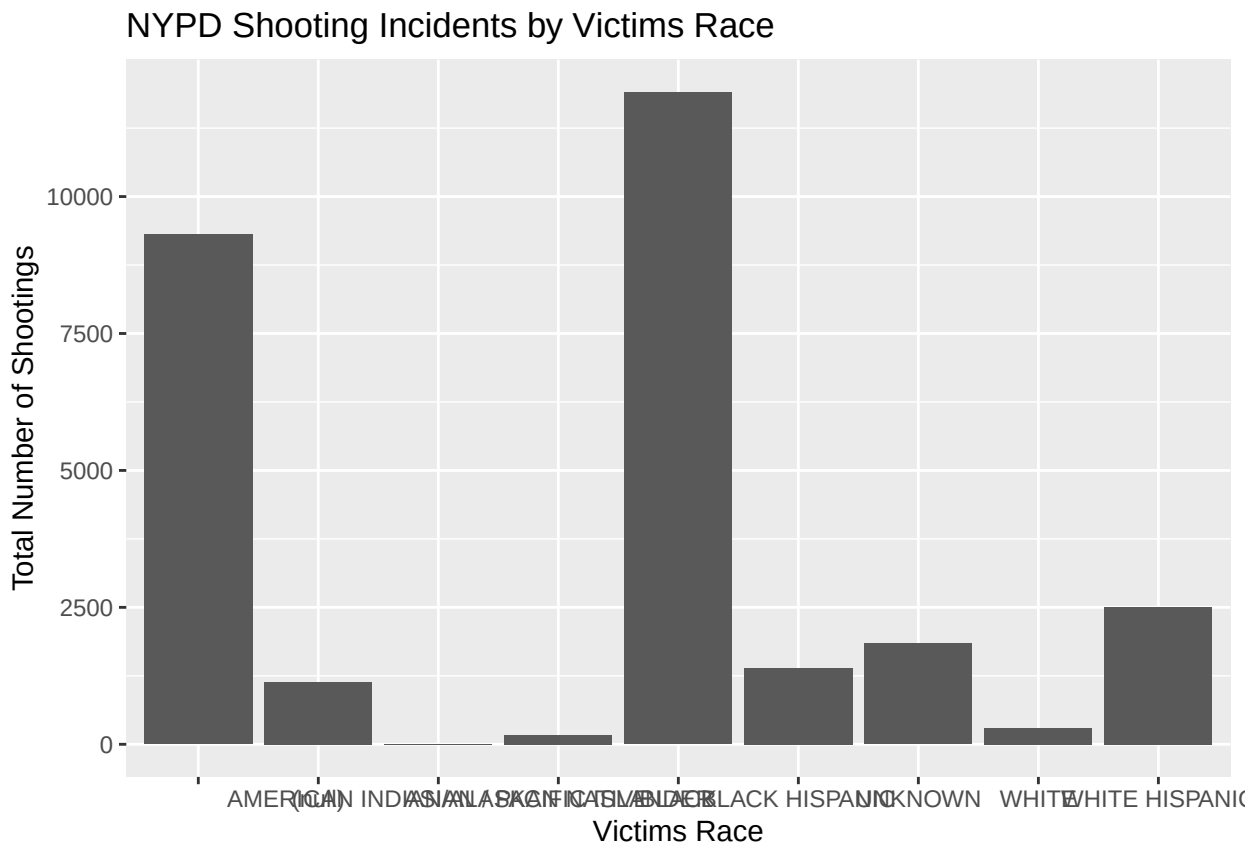

```
nypd_clean %>%
  ggplot(aes(x = BORO, fill = BORO)) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Borough",
       subtitle = "Years: 2006 - 2021",
       x = "Boroughs",
       y = "Total Number of Shootings")
```



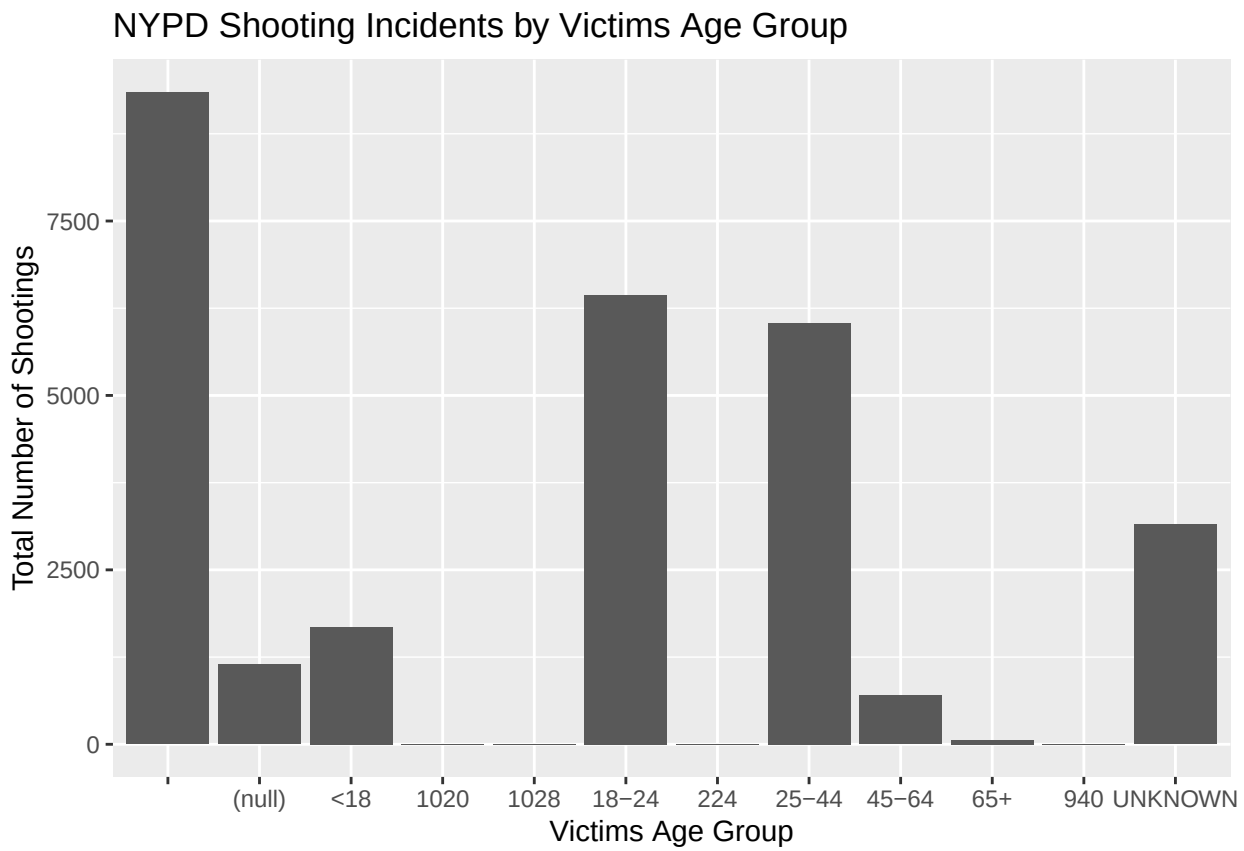

```
nypd_clean %>%
  ggplot(aes(x = PRECINCT)) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Precinct",
       x = "Precinct",
       y = "Total Number of Shootings")
```



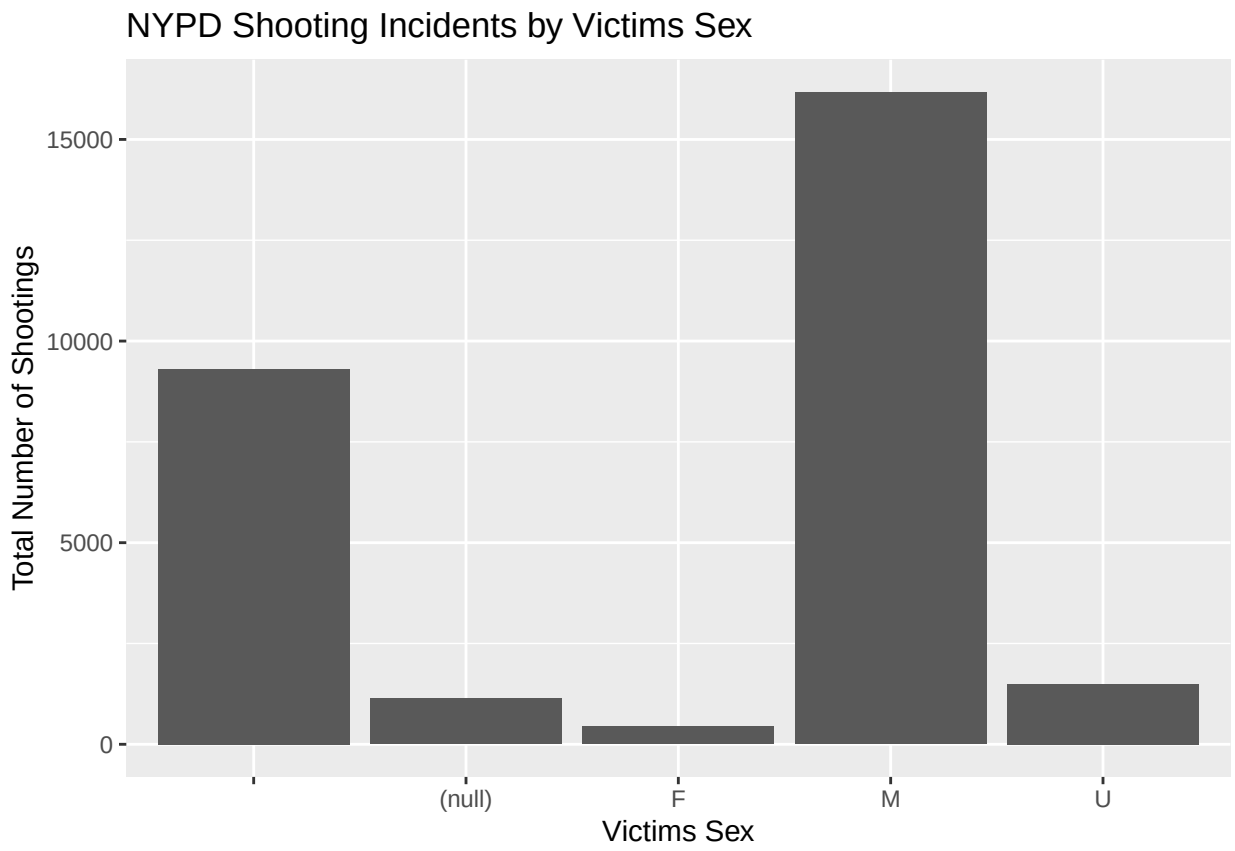

```
nypd_clean %>%
  ggplot(aes(x = PERP_RACE)) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Victims Race",
       x = "Victims Race",
       y = "Total Number of Shootings")
```



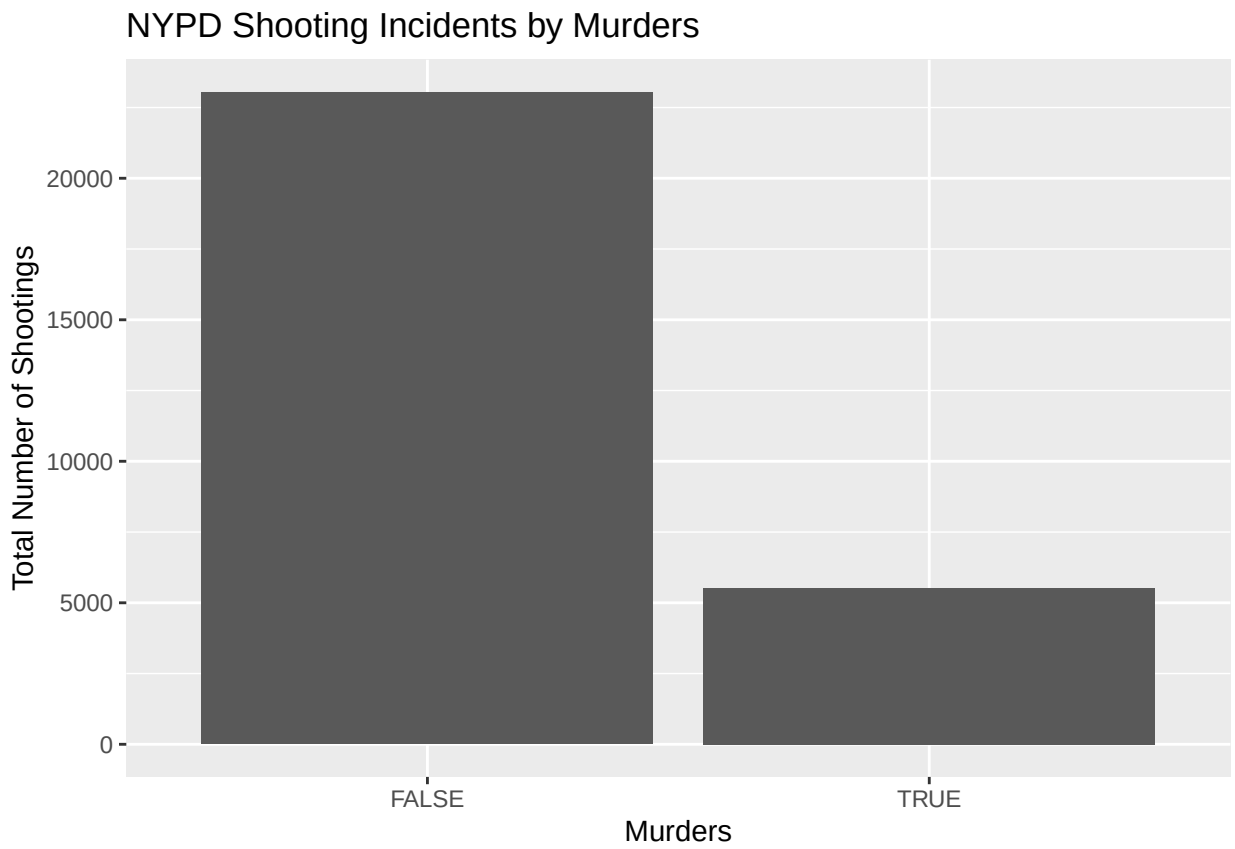

```
nypd_clean %>%
  ggplot(aes(x = PERP_AGE_GROUP )) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Victims Age Group",
       x = "Victims Age Group",
       y = "Total Number of Shootings")
```




```
nypd_clean %>%
  ggplot(aes(x = PERP_SEX )) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Victims Sex",
       x = "Victims Sex",
       y = "Total Number of Shootings")
```




```
nypd_clean %>%
  ggplot(aes(x = STATISTICAL_MURDER_FLAG  )) +
  geom_bar() +
  labs(title = "NYPD Shooting Incidents by Murders",
       x = "Murders",
       y = "Total Number of Shootings")
```



As we saw with hour there is a quadradic relationship. So, let's try to a simple model of shooting by hour.

```
nypd_hour_mod <- nypd_clean %>%
     group_by(Hour, Shootings) %>%
     summarize(Shootings = sum(Shootings),
               STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG))
```

```
## `summarise()` has grouped output by 'Hour'. You can override using the
## `.groups` argument.
```

```
mod <-lm(data=nypd_hour_mod, Shootings ~ Hour )
summary(mod)
```

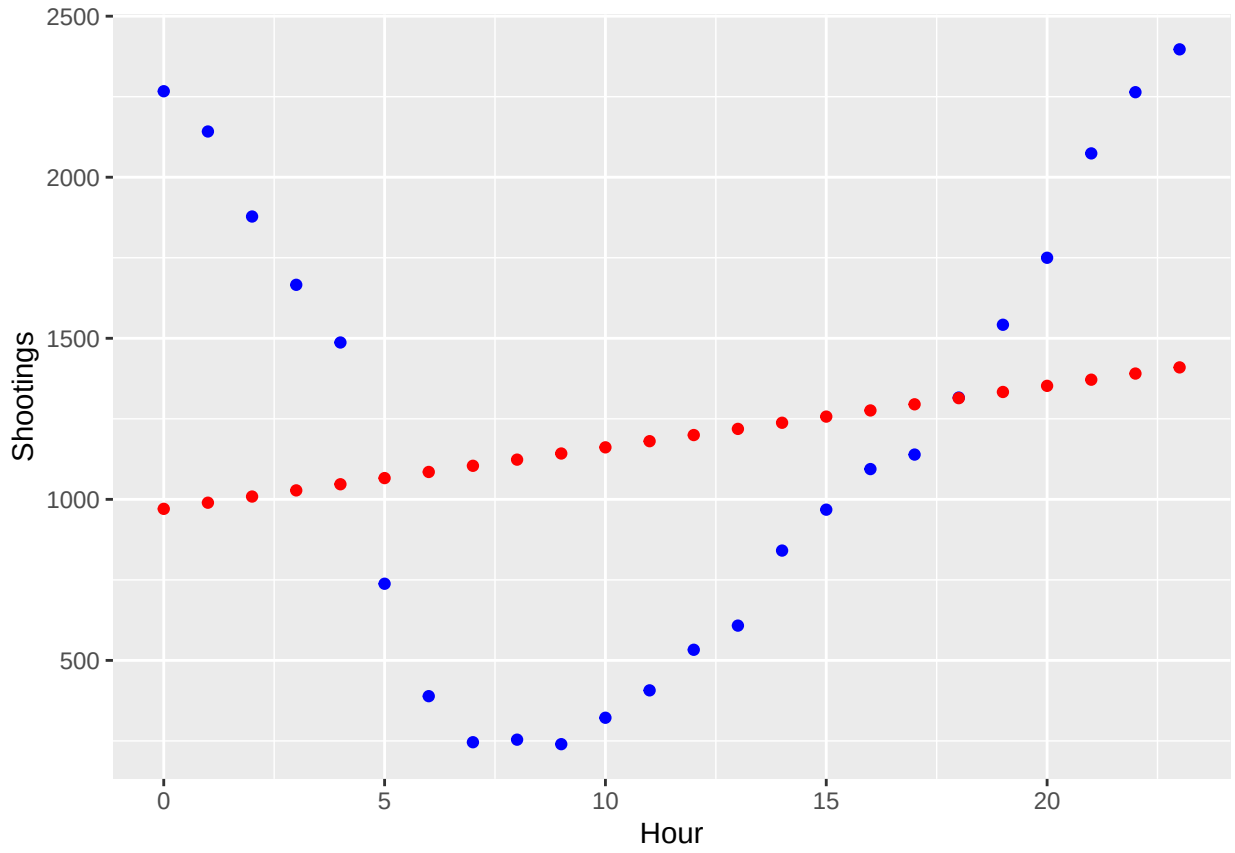
```
##
## Call:
## lm(formula = Shootings ~ Hour, data = nypd_hour_mod)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -902.4 -674.0 -169.0  654.3 1296.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   970.52     293.73   3.304  0.00323 **
## Hour           19.09      21.88   0.872  0.39239
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 742.1 on 22 degrees of freedom
## Multiple R-squared:  0.03344,    Adjusted R-squared:  -0.01049
## F-statistic: 0.7612 on 1 and 22 DF,  p-value: 0.3924
```

```
nypd_hour_mod <- nypd_hour_mod %>%
    ungroup() %>%
    mutate(pred = predict(mod, newdata = nypd_hour_mod))
 nypd_hour_mod <- nypd_hour_mod %>%
    rowwise() %>%
    mutate(pred = predict(mod, newdata = cur_data()))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `pred = predict(mod, newdata = cur_data())`.
## i In row 1.
## Caused by warning:
## ! `cur_data()` was deprecated in dplyr 1.1.0.
## i Please use `pick()` instead.
```

```
nypd_hour_mod  %>%
  ggplot()+
  geom_point(aes(x=Hour, y=Shootings), color = "blue") +
  geom_point(aes(x=Hour, y=pred), color = "red")
```



As we saw with model one, a linear relationship doesn't model shootings well. Let's try adding a quadratic hour variable and see if that gets a better model.

```
nypd_hour_mod <- nypd_clean %>%
    group_by(Hour, Shootings) %>%
    summarize(Shootings = sum(Shootings),
              STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
    mutate(hour_sq = Hour^2)
```

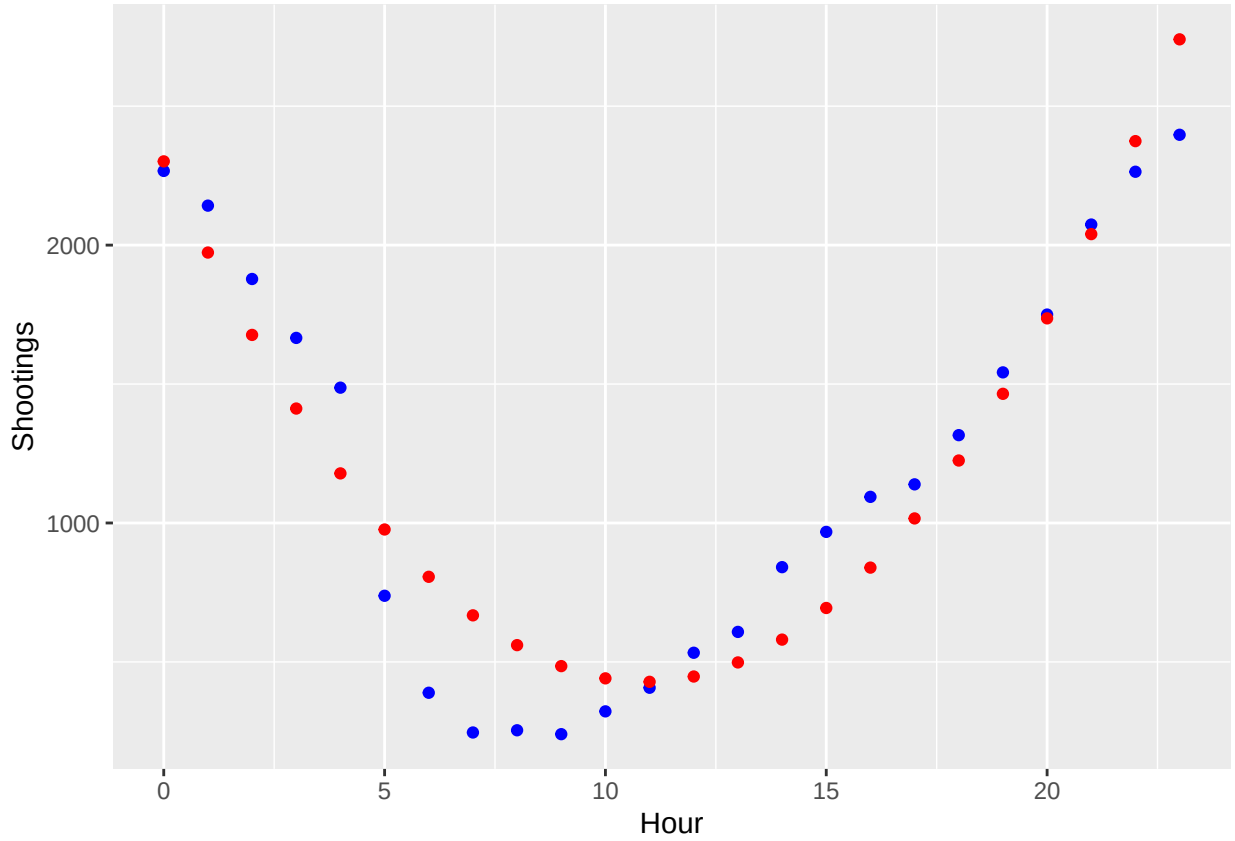
```
## `summarise()` has grouped output by 'Hour'. You can override using the
## `.groups` argument.
```

```
mod <-lm(data=nypd_hour_mod, Shootings ~ Hour + hour_sq )
summary(mod)
```

```
##
## Call:
## lm(formula = Shootings ~ Hour + hour_sq, data = nypd_hour_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -421.58 -148.74   55.86  176.86  308.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2301.344    134.713   17.08 8.55e-14 ***
## Hour        -343.859     27.131  -12.67 2.64e-11 ***
## hour_sq       15.780      1.139   13.85 4.95e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 238.6 on 21 degrees of freedom
## Multiple R-squared:  0.9046, Adjusted R-squared:  0.8956
## F-statistic: 99.62 on 2 and 21 DF,  p-value: 1.919e-11
```

```
nypd_hour_mod <- nypd_hour_mod %>%
    ungroup() %>%
    mutate(pred = predict(mod, newdata = nypd_hour_mod))
 nypd_hour_mod <- nypd_hour_mod %>%
    rowwise() %>%
    mutate(pred = predict(mod, newdata = cur_data()))
```

```
nypd_hour_mod  %>%
  ggplot()+
  geom_point(aes(x=Hour, y=Shootings), color = "blue") +
  geom_point(aes(x=Hour, y=pred), color = "red")
```



Conclusion

Looking at the charts, we see shootings were going down until 2019 and then there is a spike and potential re-normalization in 2020 with COVID. There is a clear relationship between time (hour) and shootings. From the other charts there are some other metrics that could have a relationship with a shoot, such as which borough and sex of perp.

My model is a simple linear model using time (hour) to predict number of shootings. The relationship isn't linear, so hour is squared. Then a really strong model is produced.