NYPD Shooting Cases Project

JJ.H

2022-06-06

NYPD Shooting Cases Project

This is a report to analysis the data of **NYPD Shooting Cases** with the data from source: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv.

Data Import

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
url_in = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
```

Data Transforming and Cleaning Up

First, cleaning the data set and some columns with less information are removed such as INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, Latitude, Longitude and changing the column for OCCUR_DATE to "date" type. Adding a column to identify and caculate the murder occurred in daily based.

```
shooting_cases = read_csv(url_in)
shooting_cases_clean <- shooting_cases %>% select(2,3,4,8,12,13,14)
shooting_cases_clean <- shooting_cases_clean %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))

shooting_agg <- shooting_cases_clean
shooting_agg <- shooting_agg %>% select(1)
shooting_agg$value <- 1
shooting_stat <- shooting_agg %>%
group_by(OCCUR_DATE) %>% summarise(cases = sum(value))
summary(shooting_stat)
```

```
##    OCCUR_DATE            cases
##  Min.   :2006-01-01   Min.   : 1.000
##  1st Qu.:2009-08-11   1st Qu.: 2.000
##  Median :2013-04-03   Median : 4.000
##  Mean   :2013-05-08   Mean   : 4.667
##  3rd Qu.:2017-01-05   3rd Qu.: 6.000
##  Max.   :2020-12-31   Max.   :47.000
```

```
shooting_area <- shooting_cases_clean %>% select(3)
shooting_area$case <- 1
shooting_area <- shooting_area %>% group_by(BORO) %>% summarise(cases = sum(case))
shooting_area
```

```
## # A tibble: 5 x 2
##   BORO          cases
##   <chr>         <dbl>
## 1 BRONX          6701
## 2 BROOKLYN       9734
## 3 MANHATTAN      2922
## 4 QUEENS         3532
## 5 STATEN ISLAND   696
```

```
shooting_vic_age <- shooting_cases_clean %>% select(5)
shooting_vic_age$case <- 1
shooting_vic_age <- shooting_vic_age %>%
group_by(VIC_AGE_GROUP) %>% summarise(cases = sum(case))
shooting_vic_age
```

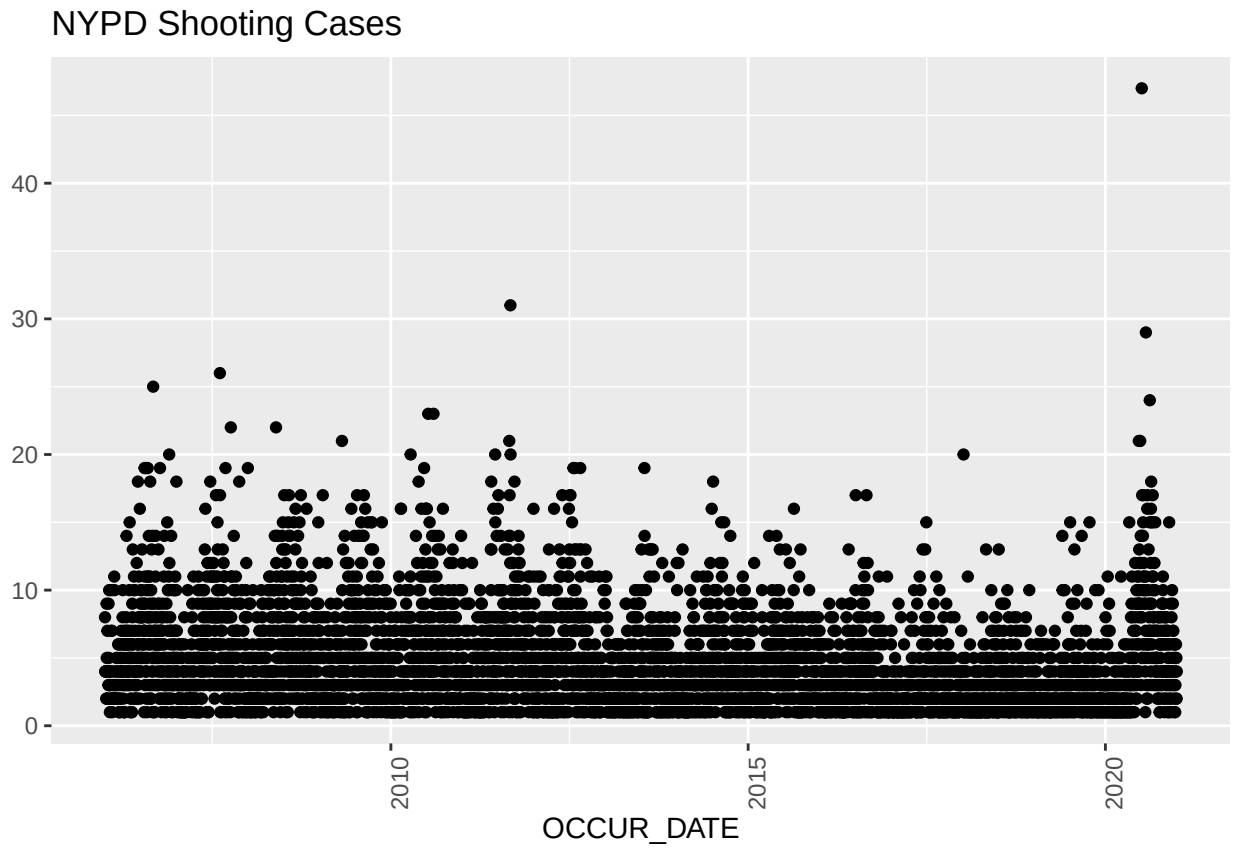
```
## # A tibble: 6 x 2
##   VIC_AGE_GROUP cases
##   <chr>         <dbl>
## 1 <18            2525
## 2 18-24          9003
## 3 25-44         10303
## 4 45-64          1541
## 5 65+             154
## 6 UNKNOWN          59
```

```
shooting_murder <- shooting_cases_clean %>% select(1,4)
shooting_murder$case <- 1
shooting_murder$murder <- ifelse(shooting_murder$STATISTICAL_MURDER_FLAG, 1, 0)
shooting_murder_stat <- shooting_murder %>%
group_by(OCCUR_DATE) %>%
summarise(cases = sum(case), murders = sum(murder))
summary(shooting_murder_stat)
```

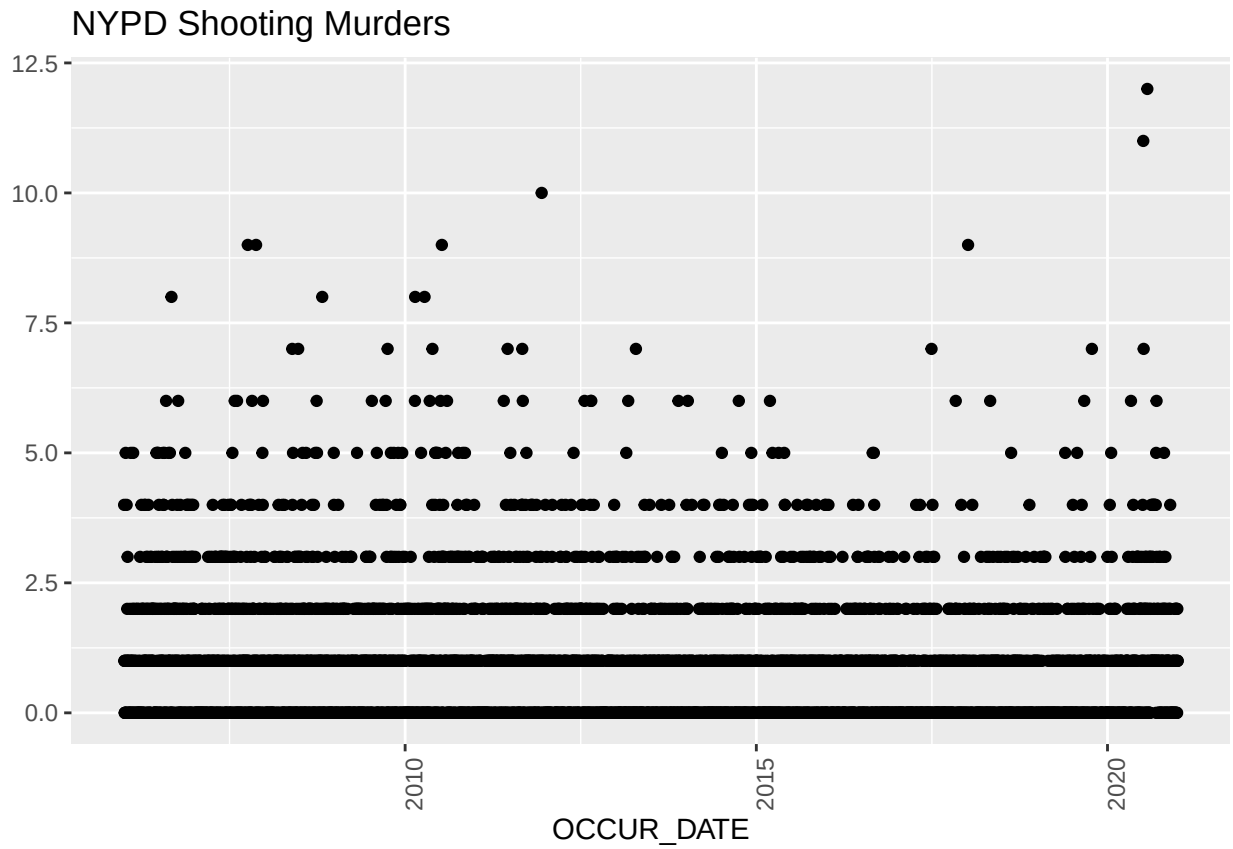
```
##    OCCUR_DATE             cases           murders
##  Min.   :2006-01-01   Min.   : 1.000   Min.   : 0.0000
##  1st Qu.:2009-08-11   1st Qu.: 2.000   1st Qu.: 0.0000
##  Median :2013-04-03   Median : 4.000   Median : 0.0000
##  Mean   :2013-05-08   Mean   : 4.667   Mean   : 0.8904
##  3rd Qu.:2017-01-05   3rd Qu.: 6.000   3rd Qu.: 1.0000
##  Max.   :2020-12-31   Max.   :47.000   Max.   :12.0000
```

Visualizations & Analysis

```
shooting_stat %>%
ggplot(x = OCCUR_DATE, y = cases) +
geom_point(aes(y = cases, x = OCCUR_DATE)) + theme(legend.position = "bottom", axis.text.x = element_te
labs(title = "NYPD Shooting Cases", y = NULL)
```




```
shooting_murder_stat %>%
ggplot(x = OCCUR_DATE, y = murders  ) +
geom_point(aes(y = murders , x = OCCUR_DATE)) + theme(legend.position = "bottom", axis.text.x = element_
labs(title = "NYPD Shooting Murders", y = NULL)
```

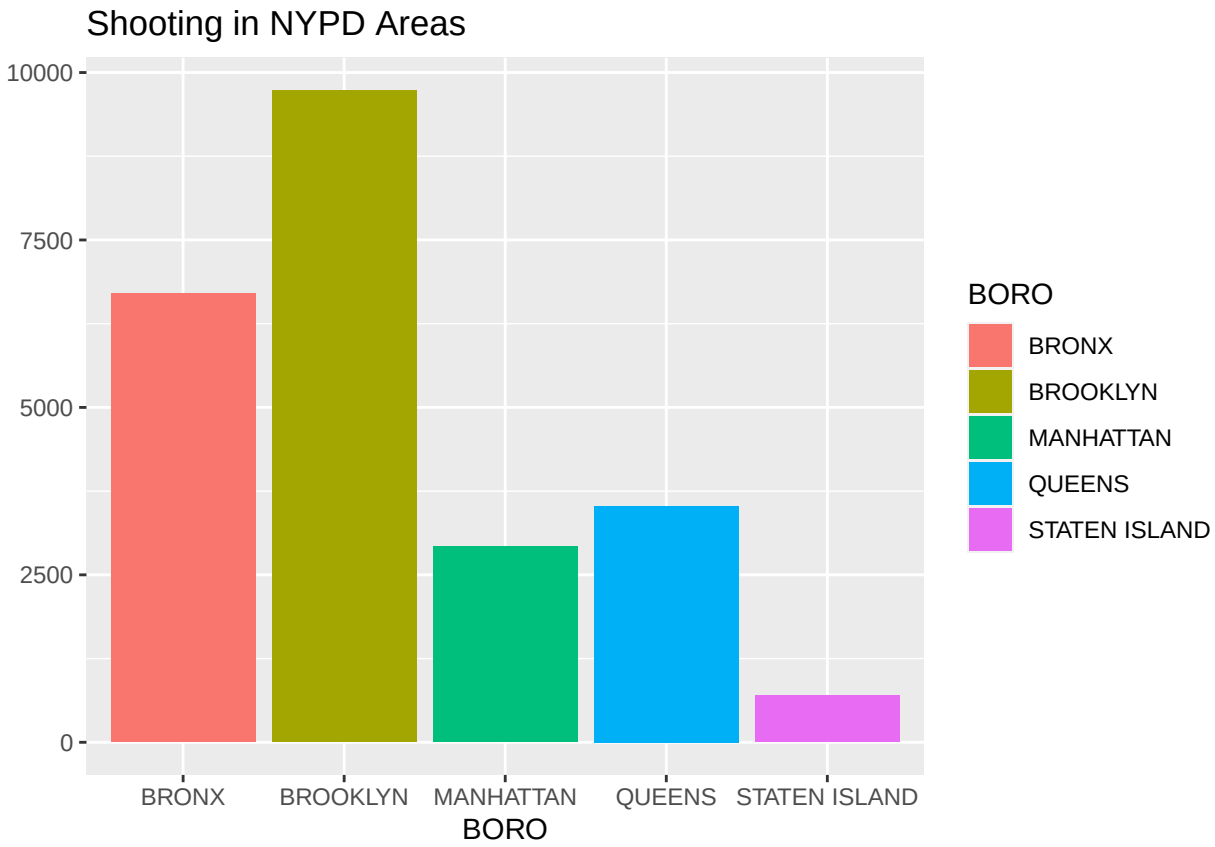


From above two graphs, we can see the numbers of both the shooting cases and murder cases are decreasing from 2006 to 2019, the overall decreasing trend is obvious, however, the numbers in year 2020 increase to the level of year 2006.

```
shooting_area
```

```
## # A tibble: 5 x 2
##   BORO          cases
##   <chr>         <dbl>
## 1 BRONX          6701
## 2 BROOKLYN       9734
## 3 MANHATTAN      2922
## 4 QUEENS         3532
## 5 STATEN ISLAND   696
```

```
shooting_area %>%
ggplot(x = BORO , y = cases, fill = BORO) +
geom_col(aes( x = BORO, y = cases, fill = BORO )) +
labs(title = "Shooting in NYPD Areas", y = NULL)
```

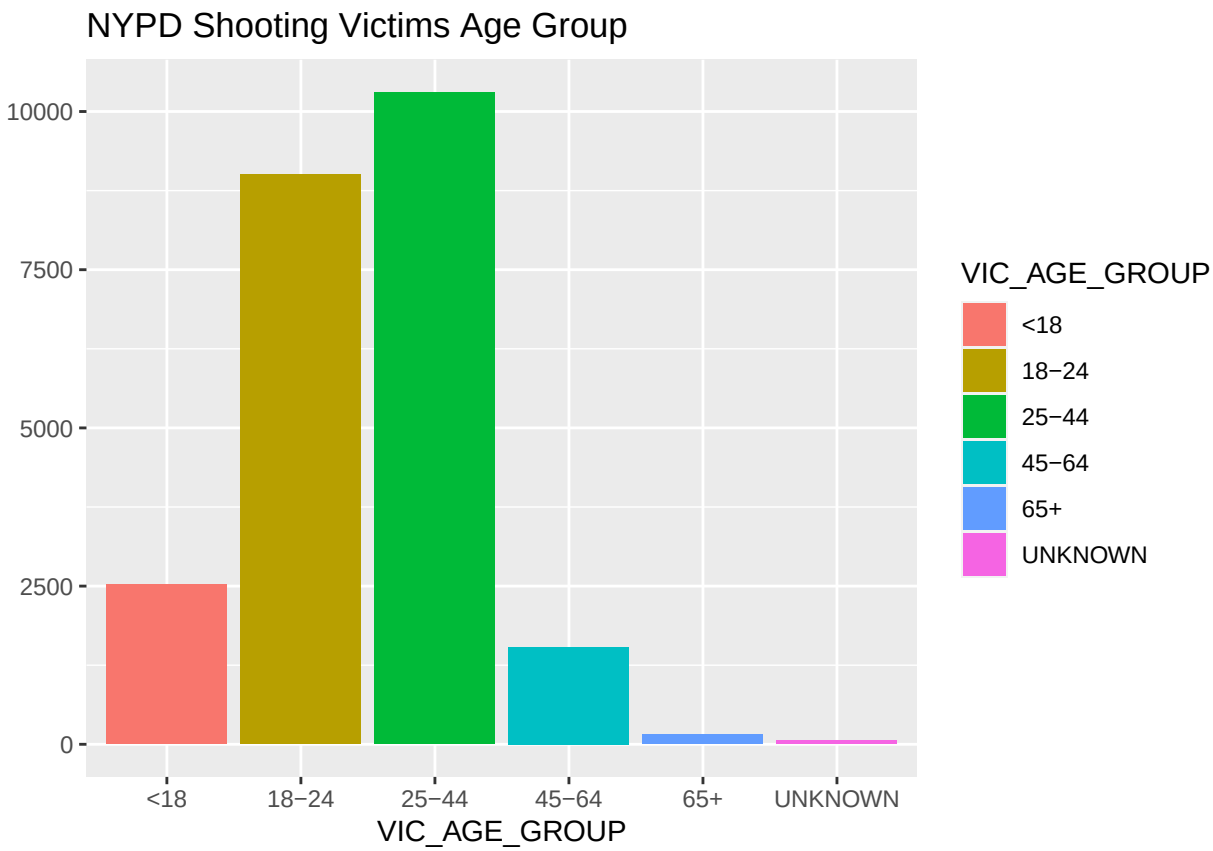


The above data & bar graph shows the shooting cases in different NYPD areas. Most of shooting cases are occurred in Brooklyn with the highest number of shooting incidences occurred from 2006 to 2020.

```
shooting_vic_age
```

```
## # A tibble: 6 x 2
##   VIC_AGE_GROUP cases
##   <chr>         <dbl>
## 1 <18            2525
## 2 18-24          9003
## 3 25-44         10303
## 4 45-64          1541
## 5 65+             154
## 6 UNKNOWN          59
```

```
shooting_vic_age %>%
ggplot(x = VIC_AGE_GROUP,y = cases)+
geom_col(aes( x = VIC_AGE_GROUP, y = cases, fill = VIC_AGE_GROUP ))+
labs(title = "NYPD Shooting Victims Age Group", y = NULL)
```



The above data & bar graph shows the shooting cases in different age groups of victims. From the bar graph above, we see that the victims in 25-44 age group with the highest number of shootings, followed by victims in 18-24 age group.

Model

```
mod1 <- lm (murders ~ cases, data = shooting_murder_stat)
shooting_murder_model <- shooting_murder_stat%>% mutate(pred = predict(mod1))
mod1
```

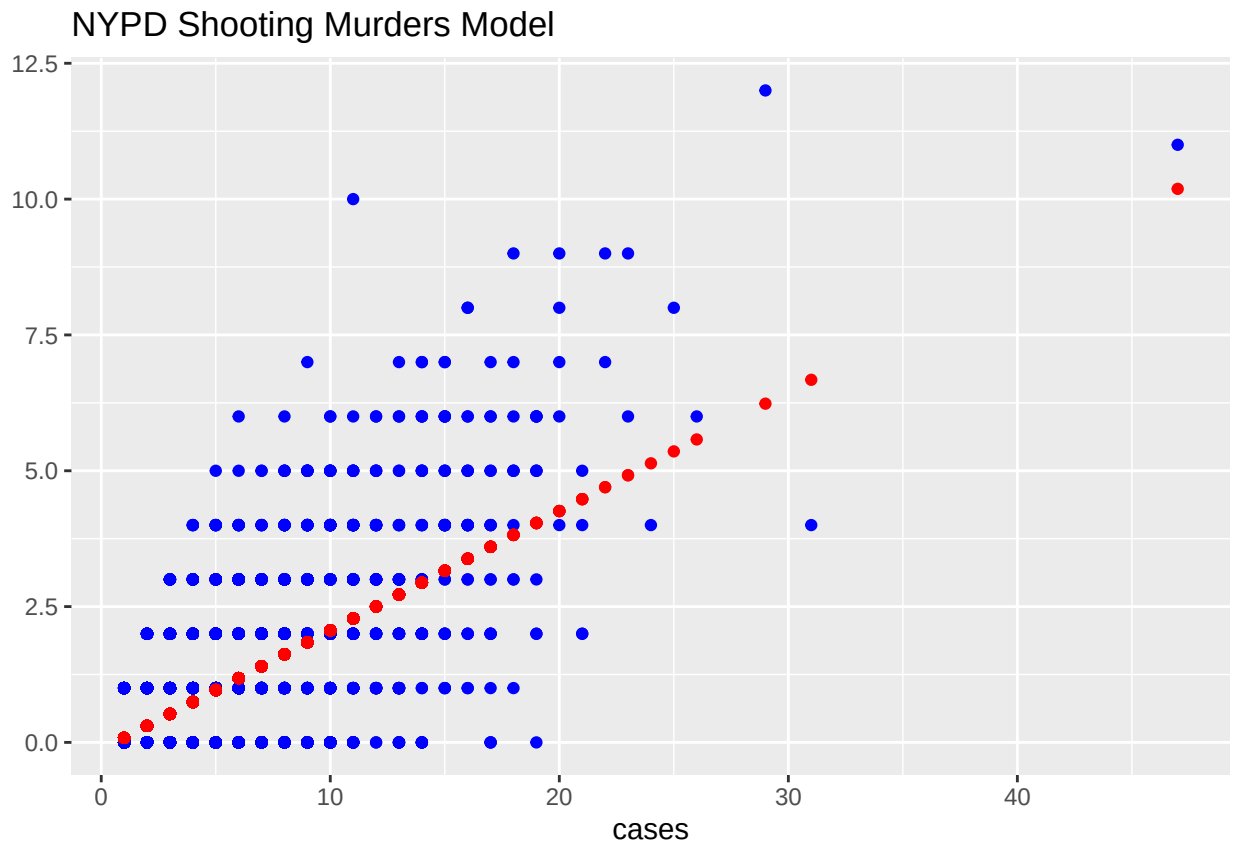
```
##
## Call:
## lm(formula = murders ~ cases, data = shooting_murder_stat)
##
## Coefficients:
## (Intercept)        cases
##     -0.1346       0.2196
```

```
summary(shooting_murder_model)
```

```
##    OCCUR_DATE             cases           murders             pred
##  Min.   :2006-01-01   Min.   : 1.000   Min.   : 0.0000   Min.   : 0.08503
##  1st Qu.:2009-08-11   1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.: 0.30468
##  Median :2013-04-03   Median : 4.000   Median : 0.0000   Median : 0.74397
```

```
##  Mean   :2013-05-08   Mean   : 4.667   Mean   : 0.8904   Mean   : 0.89038
##  3rd Qu.:2017-01-05   3rd Qu.: 6.000   3rd Qu.: 1.0000   3rd Qu.: 1.18326
##  Max.   :2020-12-31   Max.   :47.000   Max.   :12.0000   Max.   :10.18873
```

```
shooting_murder_model %>%
ggplot() +
geom_point(aes(x= cases, y= murders),color = "blue")+ geom_point(aes(x= cases, y= pred),color = "red")+
labs(title = "NYPD Shooting Murders Model", y = NULL)
```



Overall, from the relationship between the numbers of shootings occurred and numbers of shooting murders occurred, seems there are about half of the shooting cases occurred in NYPD are murder cases.

Conclusion

The report analysis the shooting incidents occurred in New York city during Jan 2006 and Dec 2020. There is a decreasing trend of murder cases and shooting case between 2006 and 2019, and the numbers are increase sharply in 2020. Moreover, from the model we can see, there are nearly about half of the shooting cases are murder cases.

Bias Identification

When cleaning and transforming the data, the data of cases and murders are managed as daily based. The daily based data may give more detail in visualization, but also make the trend of data not visual. Moreover, the predicted murder cases in the prediction model may not as accurate as it should be. There are many missing values in the data set, I remove many observations in order to clean and analysis effectively. The

missing data could make me ignore much details in cleaning and analyzing, and it could be potential bias source.