week3 Project

Nadia Gontova

2023-11-30

DATA

The data used is an NYPD dataset covering a vareity of information on every recorded shooting from 2006 to 2021. The data includes information about the location, victim, and shooter. This is a very large dataset with almost 3000 entires however there is some missing information. Information on the perpetrator is often missing from entries in the dataset.

```
if (!require("dplyr")) install.packages("dplyr")
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
if (!require("ggplot2")) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```
library(dplyr)
library(ggplot2)

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

# Read the data
data <- read.csv(url_in)

# removed
columns_to_remove <- c("Latitude", "Longitude", "X_COORD_CD", "Y_COORD_CD", "Lon_Lat", "LOC_OF_OCCUR_DES

#data <- data[, !(names(data) %in% columns_to_remove)]
```

```r
#data <- data[data$PERP_AGE_GROUP != "UNKNOWN", ]


# Convert OCCUR_DATE to date object
data$OCCUR_DATE <- as.Date(data$OCCUR_DATE, format = "%m/%d/%Y")

# Convert OCCUR_TIME to time data type
data$OCCUR_TIME <- as.POSIXct(data$OCCUR_TIME, format = "%H:%M", tz = "UTC")


head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE          OCCUR_TIME     BORO LOC_OF_OCCUR_DESC
## 1    228798151 2021-05-27 2023-12-12 21:30:00   QUEENS
## 2    137471050 2014-06-27 2023-12-12 17:40:00    BRONX
## 3    147998800 2015-11-21 2023-12-12 03:56:00   QUEENS
## 4    146837977 2015-10-09 2023-12-12 18:30:00    BRONX
## 5     58921844 2009-02-19 2023-12-12 22:58:00    BRONX
## 6    219559682 2020-10-21 2023-12-12 21:36:00 BROOKLYN
##   PRECINCT JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC
## 1      105                0
## 2       40                0
## 3      108                0
## 4       44                0
## 5       47                0
## 6       81                0
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                   false                                          18-24
## 2                   false                                          18-24
## 3                    true                                          25-44
## 4                   false                                            <18
## 5                    true          25-44        M     BLACK         45-64
## 6                    true                                          25-44
##   VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M          BLACK    1058925   180924.0 40.66296 -73.73084
## 2       M          BLACK    1005028   234516.0 40.81035 -73.92494
## 3       M          WHITE    1007668   209836.5 40.74261 -73.91549
## 4       M WHITE HISPANIC    1006537   244511.1 40.83778 -73.91946
## 5       M          BLACK    1024922   262189.4 40.88624 -73.85291
## 6       M          BLACK    1004234   186461.7 40.67846 -73.92795
##                                     Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

```r
summary(data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :2023-12-12 00:00:00.00
##  1st Qu.: 63860880   1st Qu.:2009-07-18   1st Qu.:2023-12-12 03:27:00.00
```

```
##  Median : 90372218   Median :2013-04-29   Median :2023-12-12 15:11:00.00
##  Mean   :120860536   Mean   :2014-01-06   Mean   :2023-12-12 12:41:31.71
##  3rd Qu.:188810230   3rd Qu.:2018-10-15   3rd Qu.:2023-12-12 20:45:00.00
##  Max.   :261190187   Max.   :2022-12-31   Max.   :2023-12-12 23:59:00.00
##
##      BORO           LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  Length:27312       Length:27312       Min.   :  1.00   Min.   :0.0000
##  Class :character   Class :character   1st Qu.: 44.00   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median : 68.00   Median :0.0000
##                                        Mean   : 65.64   Mean   :0.3269
##                                        3rd Qu.: 81.00   3rd Qu.:0.0000
##                                        Max.   :123.00   Max.   :2.0000
##                                                         NA's   :2
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX           PERP_RACE          VIC_AGE_GROUP
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:27312       Length:27312       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000028   1st Qu.:182834
##  Mode  :character   Mode  :character   Median :1007731   Median :194487
##                                        Mean   :1009449   Mean   :208127
##                                        3rd Qu.:1016838   3rd Qu.:239518
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:27312
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :10      NA's   :10
```

Analysis

```
# Group by BORO and calculate # of shootings in each place
shootings_per_boro <- data %>%
  group_by(BORO) %>%
  summarise(NumberOfShootings = n())
```

```r
print(shootings_per_boro)
```

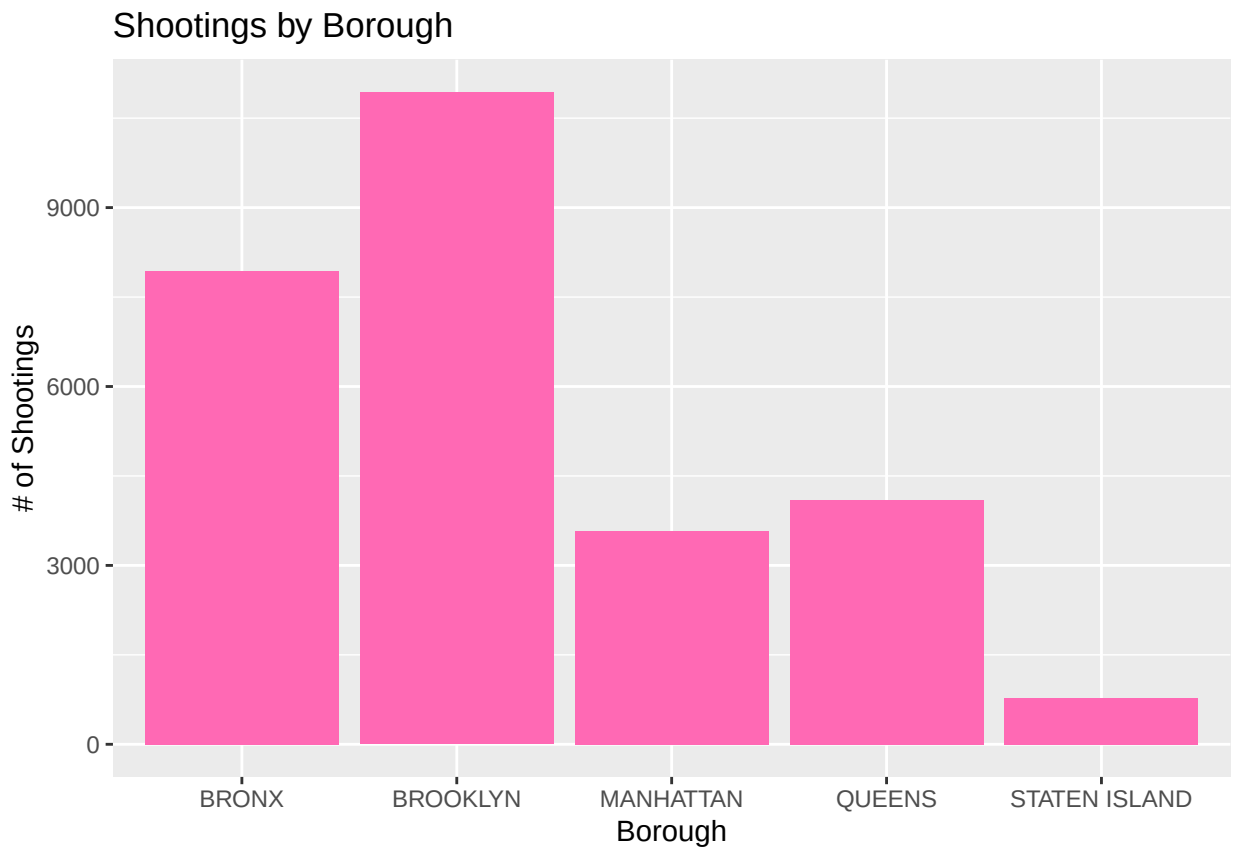
```
## # A tibble: 5 x 2
##   BORO          NumberOfShootings
##   <chr>                     <int>
## 1 BRONX                      7937
## 2 BROOKLYN                  10933
## 3 MANHATTAN                  3572
## 4 QUEENS                     4094
## 5 STATEN ISLAND               776
```

```r
# Group data by OCCUR_DATE
shootings_over_time <- data %>%
  group_by(OCCUR_DATE) %>%
  summarise(NumberOfShootings = n())

print(shootings_over_time)
```

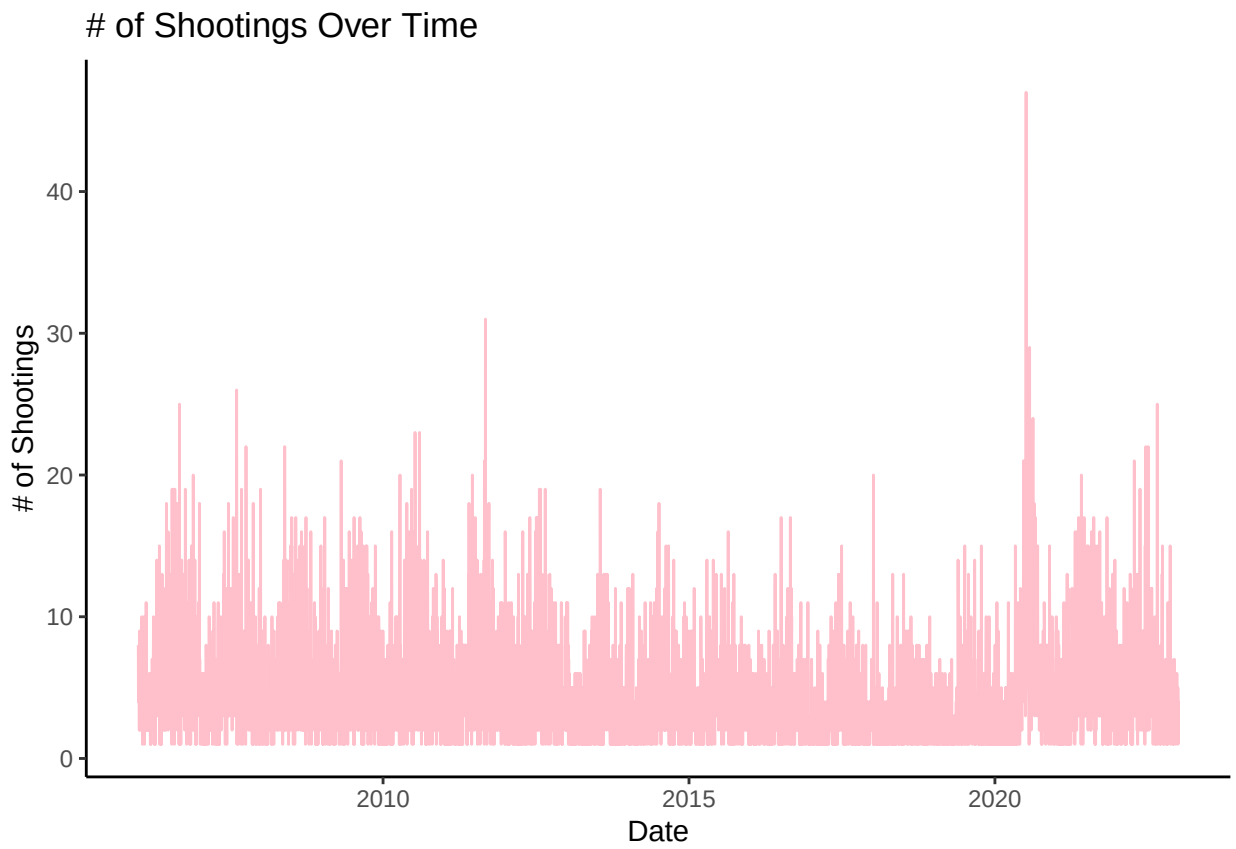
```
## # A tibble: 5,761 x 2
##    OCCUR_DATE NumberOfShootings
##    <date>                 <int>
##  1 2006-01-01                 8
##  2 2006-01-02                 4
##  3 2006-01-03                 4
##  4 2006-01-04                 4
##  5 2006-01-05                 4
##  6 2006-01-06                 4
##  7 2006-01-07                 2
##  8 2006-01-08                 4
##  9 2006-01-09                 9
## 10 2006-01-10                 5
## # i 5,751 more rows
```

Plots

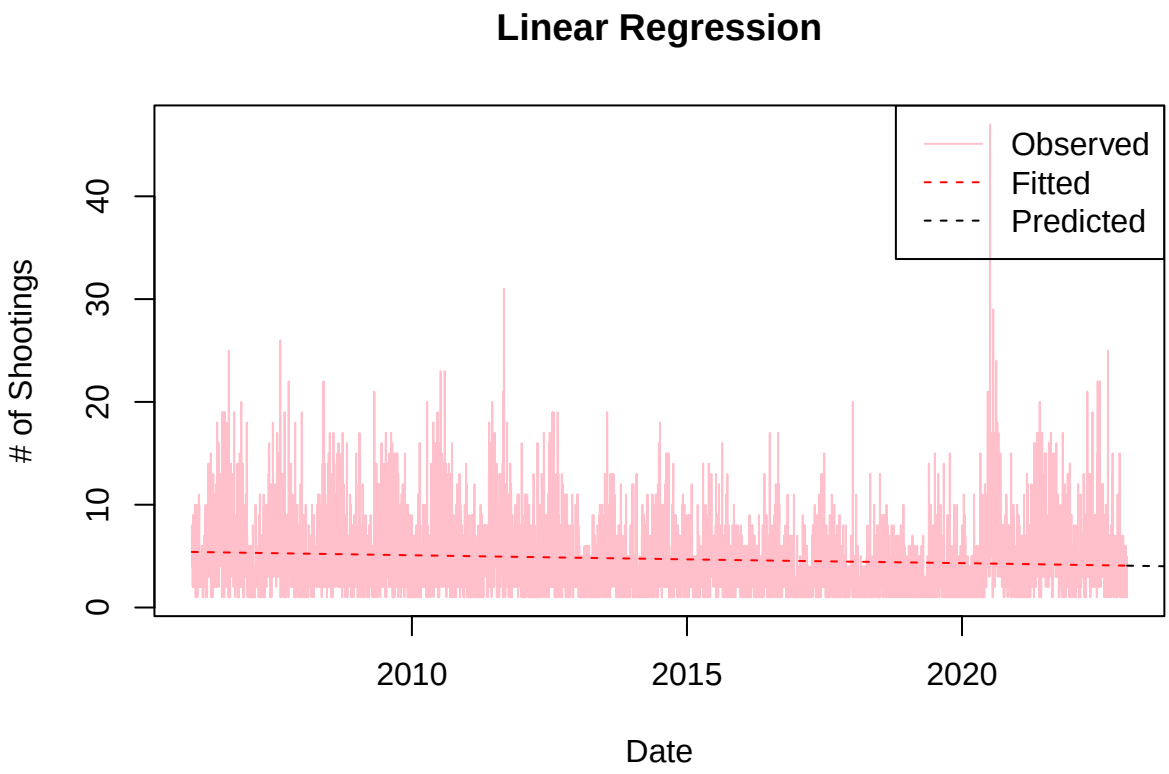
```r
# Bar plot
bar_plot <- ggplot(shootings_per_boro, aes(x = BORO, y = NumberOfShootings)) +
  geom_bar(stat = "identity", fill = "hotpink") +
  labs(title = "Shootings by Borough",
       x = "Borough",
       y = "# of Shootings")

print(bar_plot)
```



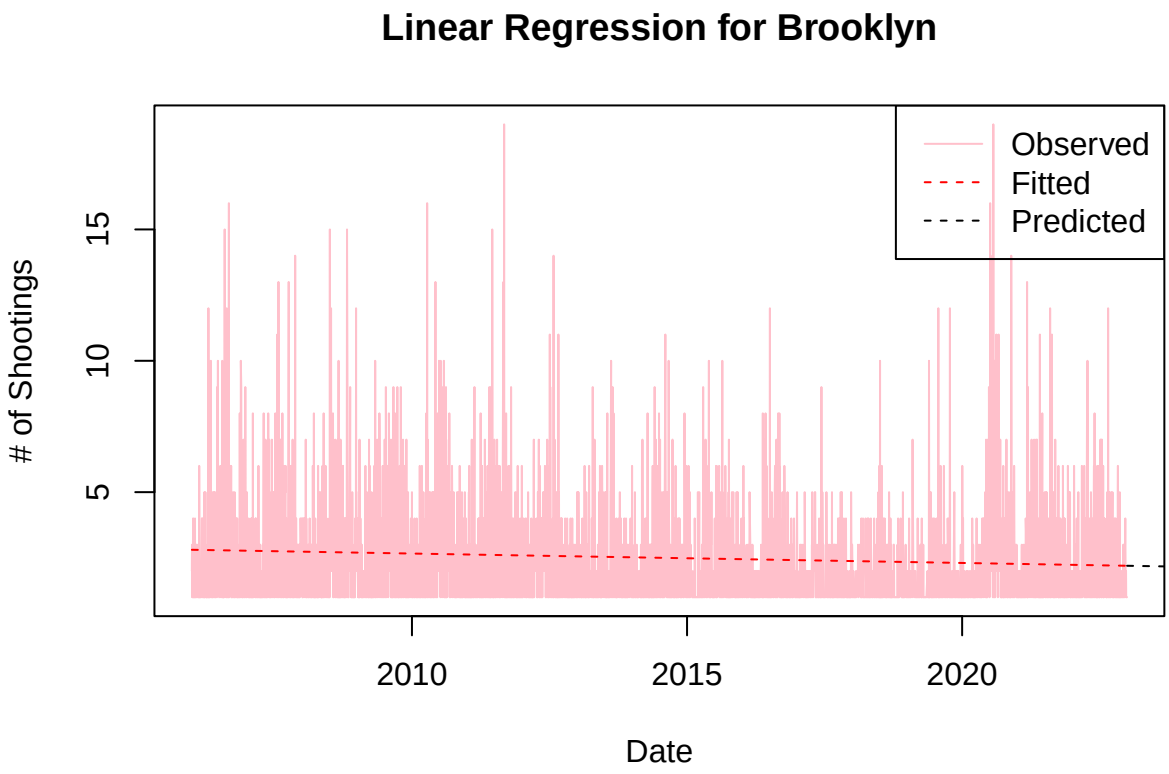

```
# line plot
line_plot <- ggplot(shootings_over_time, aes(x = OCCUR_DATE, y = NumberOfShootings)) +
  geom_line(color = "pink") +
  labs(title = "# of Shootings Over Time",
       x = "Date",
       y = "# of Shootings") +
  theme_classic()

print(line_plot)
```



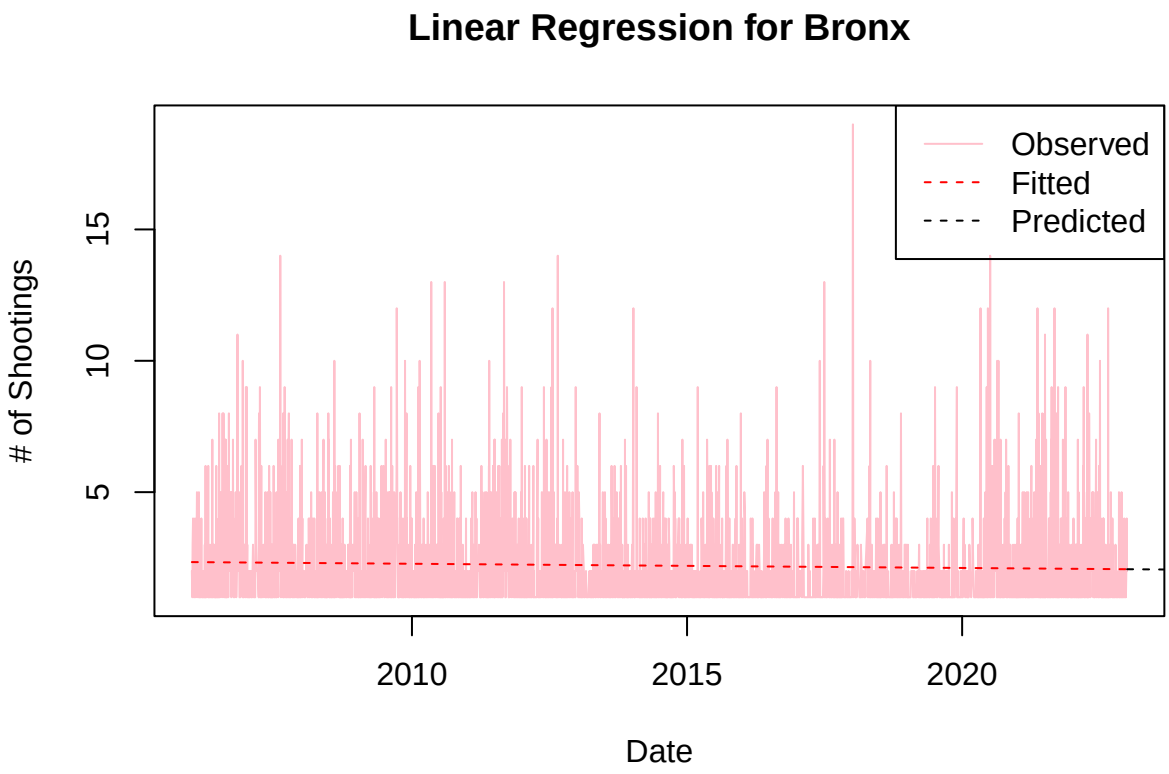

```
shootings_over_time$Time <- 1:nrow(shootings_over_time)

# linear regression model
linear_model <- lm(NumberOfShootings ~ Time, data = shootings_over_time)

# Prediction
futuretime <- seq(max(shootings_over_time$Time) + 1, length.out = 365, by = 1)

futuredata <- data.frame(Time=futuretime)

predicted_values <- predict(linear_model, newdata=futuredata)

# Plot
plot(shootings_over_time$OCCUR_DATE, shootings_over_time$NumberOfShootings,
     type = "l", col = "pink", xlab = "Date", ylab = "# of Shootings",
     main = "Linear Regression")
lines(shootings_over_time$OCCUR_DATE, fitted(linear_model), col = "red", lty = 2)
lines(seq(max(shootings_over_time$OCCUR_DATE) + 1, length.out = 365, by = 1), predicted_values, col = "
legend("topright", legend = c("Observed", "Fitted", "Predicted"), col = c("pink", "red", "black"), lty =
```




```r
brooklyn_data <- data %>%
  filter(BORO == "BROOKLYN")

# Group data by OCCUR_DATE for Brooklyn
shootings_over_time_brooklyn <- brooklyn_data %>%
  group_by(OCCUR_DATE) %>%
  summarise(NumberOfShootings = n())

# Linear regression model for Brooklyn
linear_model_brooklyn <- lm(NumberOfShootings ~ as.numeric(OCCUR_DATE - min(shootings_over_time_brooklyn
                            data = shootings_over_time_brooklyn)

# Prediction for Brooklyn
futuretime_brooklyn <- seq(max(shootings_over_time_brooklyn$OCCUR_DATE) + 1, length.out = 365, by = 1)
futuredata_brooklyn <- data.frame(OCCUR_DATE = futuretime_brooklyn)
predicted_values_brooklyn <- predict(linear_model_brooklyn, newdata = futuredata_brooklyn)

# Plot for Brooklyn
plot(shootings_over_time_brooklyn$OCCUR_DATE, shootings_over_time_brooklyn$NumberOfShootings,
     type = "l", col = "pink", xlab = "Date", ylab = "# of Shootings",
     main = "Linear Regression for Brooklyn")
lines(shootings_over_time_brooklyn$OCCUR_DATE, fitted(linear_model_brooklyn), col = "red", lty = 2)
lines(futuretime_brooklyn, predicted_values_brooklyn, col = "black", lty = 2)
legend("topright", legend = c("Observed", "Fitted", "Predicted"), col = c("pink", "red", "black"), lty =
```




```r
bronx_data <- data %>%
  filter(BORO == "BRONX")

# Group data by OCCUR_DATE for the Bronx
shootings_over_time_bronx <- bronx_data %>%
  group_by(OCCUR_DATE) %>%
  summarise(NumberOfShootings = n())

# Linear regression model for the Bronx
linear_model_bronx <- lm(NumberOfShootings ~ as.numeric(OCCUR_DATE - min(shootings_over_time_bronx$OCCU
                         data = shootings_over_time_bronx)

# Prediction for the Bronx
futuretime_bronx <- seq(max(shootings_over_time_bronx$OCCUR_DATE) + 1, length.out = 365, by = 1)
futuredata_bronx <- data.frame(OCCUR_DATE = futuretime_bronx)
predicted_values_bronx <- predict(linear_model_bronx, newdata = futuredata_bronx)

# Plot for the Bronx
plot(shootings_over_time_bronx$OCCUR_DATE, shootings_over_time_bronx$NumberOfShootings,
     type = "l", col = "pink", xlab = "Date", ylab = "# of Shootings",
     main = "Linear Regression for Bronx")
lines(shootings_over_time_bronx$OCCUR_DATE, fitted(linear_model_bronx), col = "red", lty = 2)
lines(futuretime_bronx, predicted_values_bronx, col = "black", lty = 2)
legend("topright", legend = c("Observed", "Fitted", "Predicted"), col = c("pink", "red", "black"), lty =
```



#Conclusion

In my analysis I found that Brooklyn and Bronx had the most shootings over the other regions. Staten Island had the least shootings from the five locations. In the graph of shootings over time I did not see any trend so I decided to preform a linear regression on the data. The linear model did not show an increase or decrease in shootings over time and based on this pattern the predicted pattern does not continue to increase or decrease. This aligns with what I visually saw when looking at the data. It is possible that the shootings data follows a different kind of trend and is therefore not well fit to a linear model.

#biases

One source of bias could be from bias in the data collection. There is no way for me to verify how accuratly the data is collected and if certain entires might be missing. There might be a bais in how the regions are grouped and a small area of crime might not be representative of an entire region. Frthermore shootings in one region could possibly skew the data in a neighboring region due to how the borders are determined. These typed of biases can be reduced by looking into how the data is collected and how accurate the dataset is.

There are also personal biases as an analyist that might skew the data. For example, knowing certain regions might have higher crime could impact the analysis someone chooses. To reduce this bias I tried to stick to analyzing the dataset as a whole and sticking to know statisical methods from class.