Historical NYC Shooting Analysis Project

Nicholas Cianci

2024-08-22

Historical NYC Shooting Project

Step 1

To begin, initialize the "tidyverse" package to facilitate importing and cleaning our data set.

```
library("tidyverse")
library("lubridate")
```

The next step in this project is to describe and import the historical NYC shooting data set.View the summary of the data set to help understand the values and types of each column.Preview the first few rows of the data to evaluate what fields should stay and what should be removed.

```
#Import the Historical NYC Shooting data set
nyc_shooting <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOA

#View the Summary of the NYC Shooting data set
summary(nyc_shooting)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class :character   Class :character
##  Median : 92711254   Mode  :character   Mode  :character   Mode  :character
##  Mean   :127405824
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Length:28562            Length:28562
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
```

```
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Before jumping into any analysis, I wanted to take a moment to identify some biases that I may have when doing this kind of project. To begin, I think that gun violence is a big issue in this country, especially in big cities like New York City. Because of this, I may be more inclined to show analysis and visualizations that are alarming or show how "bad" the number of shootings are. Another potential bias is around Perpetrator and Victim characteristics. The data set includes things like age, sex, and race of both the shooter and the victim. If somebody had an agenda for or against a certain age group, sex, or race, the data could be manipulated to support what you want to show.

To avoid these potential biases, I tried to analyze and visualize the data in a holistic view such as the total number of shootings and murders over time. I intentionally decided to not single out any specific age group, sex, or race of either the shooters or the victims as to avoid any bias against any particular subsection of people.

Step 2 - Tidy and Transform Data

Remove all unnecessary fields in the NYC Shooting Data set by using the "select" function.Convert "OCCUR_DATE" from "character" field to a "date" field and convert the "BORO", "VIC_AGE_GROUP", "VIC_SEX", and "VIC_RACE" columns from "character" to "factor" fields using the mutate function.Check for missing data and found 64 rows with "Unknown" listed in "VIC_AGE_GROUP" and 12 rows with "U" listed for "VIC_SEX".Considering the unknown data was very small relative to the 28,562 total rows of data, I decided to remove the rows with missing data via the "filter" function.

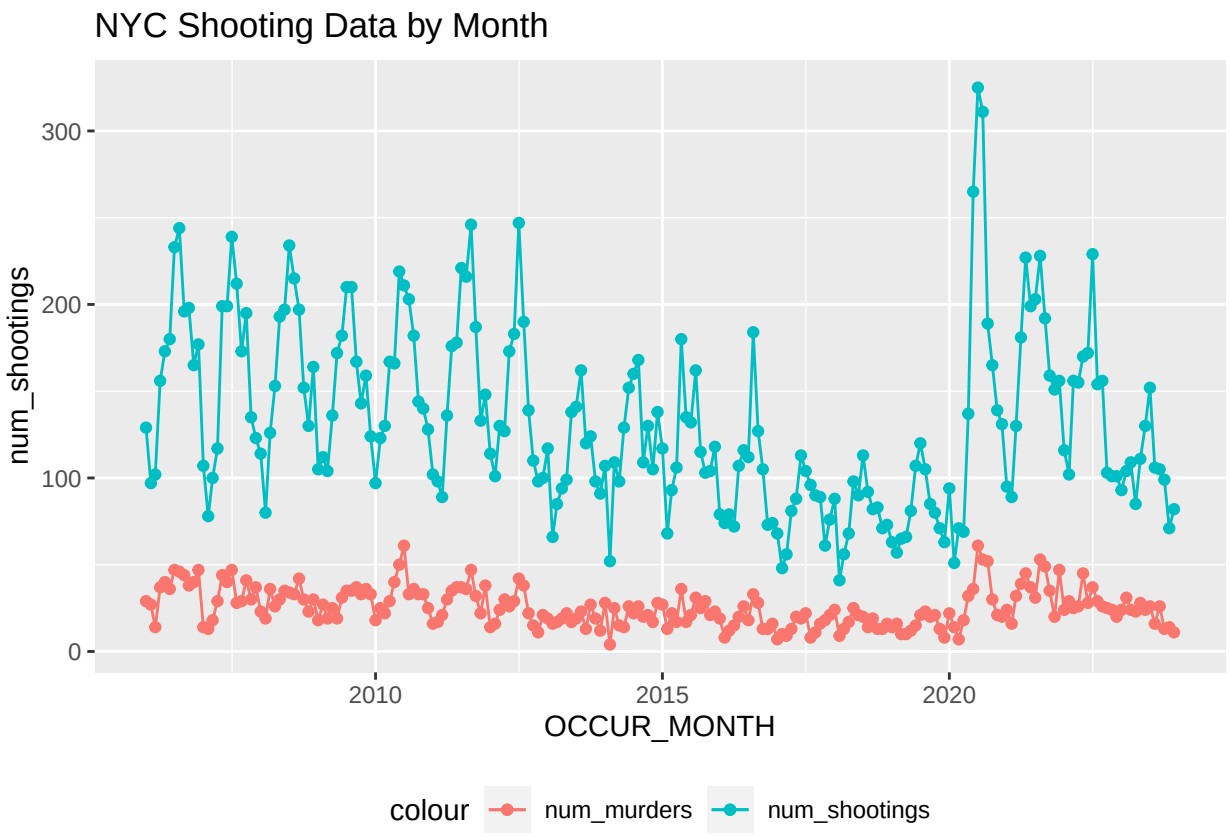
```
#Remove unnecessary fields, convert fields to date and factor as needed
nyc_shooting <- nyc_shooting %>%
      select(-c(INCIDENT_KEY,LOC_OF_OCCUR_DESC,LOC_CLASSFCTN_DESC,
```

```
                LOCATION_DESC,PRECINCT,JURISDICTION_CODE,
                PERP_AGE_GROUP:PERP_RACE ,X_COORD_CD:Lon_Lat)) %>%
      mutate(OCCUR_DATE=mdy(OCCUR_DATE)) %>%
      mutate(BORO=factor(BORO),VIC_AGE_GROUP=factor(VIC_AGE_GROUP),
             VIC_SEX=factor(VIC_SEX),VIC_RACE=factor(VIC_RACE)) %>%
      mutate(STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG))

#Limited number of unknown values, 64 rows with "Unknown" and 12 rows with "U",
#out of a total of 28,562 rows
nyc_shooting <- nyc_shooting %>%
  filter(VIC_AGE_GROUP!="UNKNOWN") %>%
  filter(VIC_SEX != "U")
```

Step 3 - Analyze and Visualize the Data

The first thing I investigated was the total number of shootings and the total number of murders that occurred over the time span captured in this data set. To do this, I grouped the dates of each shooting by month and created two fields to count the number of shootings and number of murders for each month. Next I plotted the number of shootings and murders over time on the same plot, adding color and labels where necessary.

```
#Analyze data by grouping the individual dates into months and creating two
#new fields that count the total number of shootings and the total number
#of murders for each month

nyc_total <- nyc_shooting %>%
  group_by(OCCUR_MONTH = floor_date(OCCUR_DATE, 'month')) %>%
  summarise(num_shootings = n(), num_murders = sum(STATISTICAL_MURDER_FLAG))

#Visualize both the total shootings and total murders as line graphs on the
#same plot along with appropriate legend, and title.

nyc_total %>%
  ggplot(aes(x=OCCUR_MONTH, y = num_shootings))+
  geom_line(aes(color = "num_shootings")) +
  geom_point(aes(color = "num_shootings")) +
  geom_line(aes(y = num_murders, color = "num_murders")) +
  geom_point(aes(y = num_murders, color = "num_murders")) +
  theme(legend.position = "bottom") +
  labs(title = "NYC Shooting Data by Month")
```

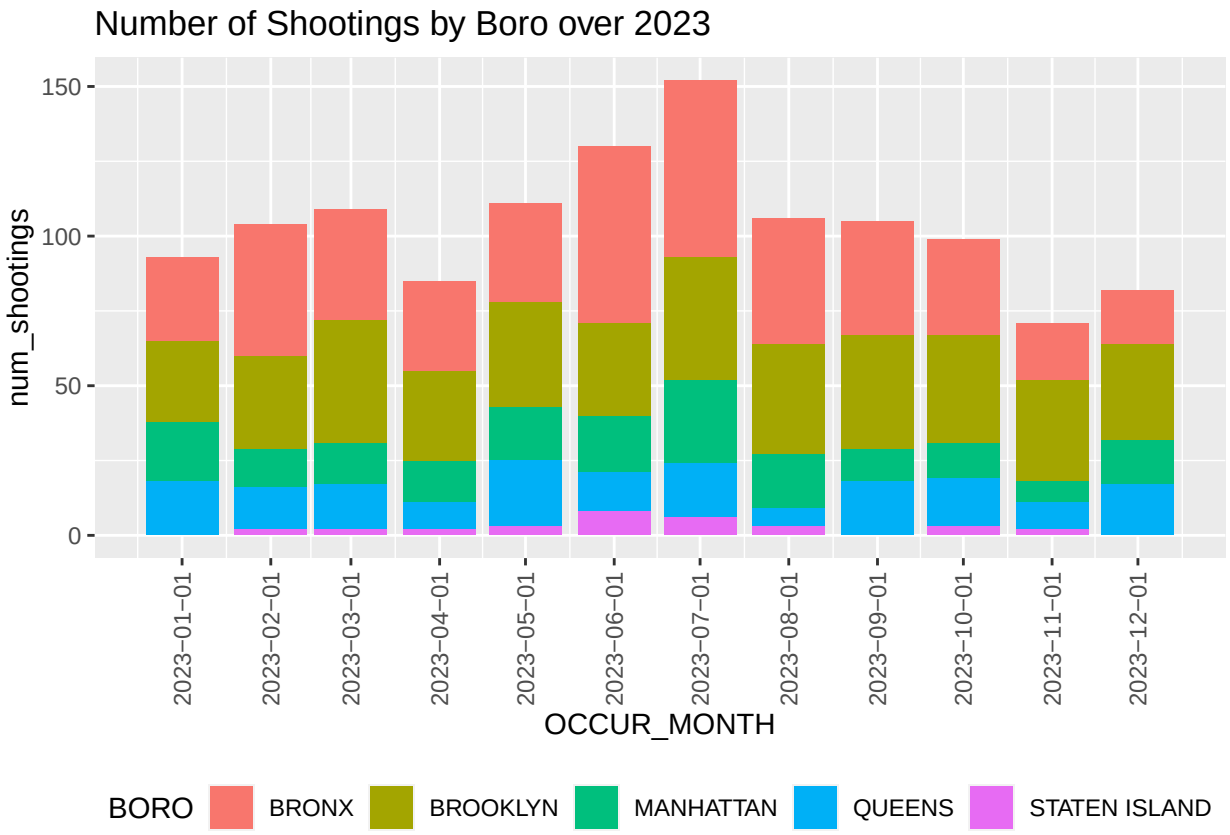


Next I was curious about the breakdown of shootings across the different Boros. To address this question I grouped the data by both "Boro" and "OCCUR_MONTH" and again counted the total number of shootings. At first I did this for the entirety of the data set but the resulting plot was too crowded to make sense of anything. So I decided to filter the view down to just the year of 2023 and visualize that by way of a stacked bar chart.

```
#Group by Boro and filter for 2023 and beyond
nyc_by_boro <- nyc_shooting %>%
  filter(OCCUR_DATE >= "2023-1-01") %>%
  group_by(BORO, OCCUR_MONTH= floor_date(OCCUR_DATE, 'month')) %>%
  summarise(num_shootings = n())
```

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

```
#Visualize stacked bar chart that shows each boro's # of shootings for each
#month of 2023

nyc_by_boro %>%
  ggplot(aes(fill=BORO, y=num_shootings, x=OCCUR_MONTH)) +
  geom_bar(position = "stack", stat="identity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle=90, vjust=0.5))+
  scale_x_date(breaks="1 month")+
  labs(title = "Number of Shootings by Boro over 2023")
```



Step 4 - Model the Data

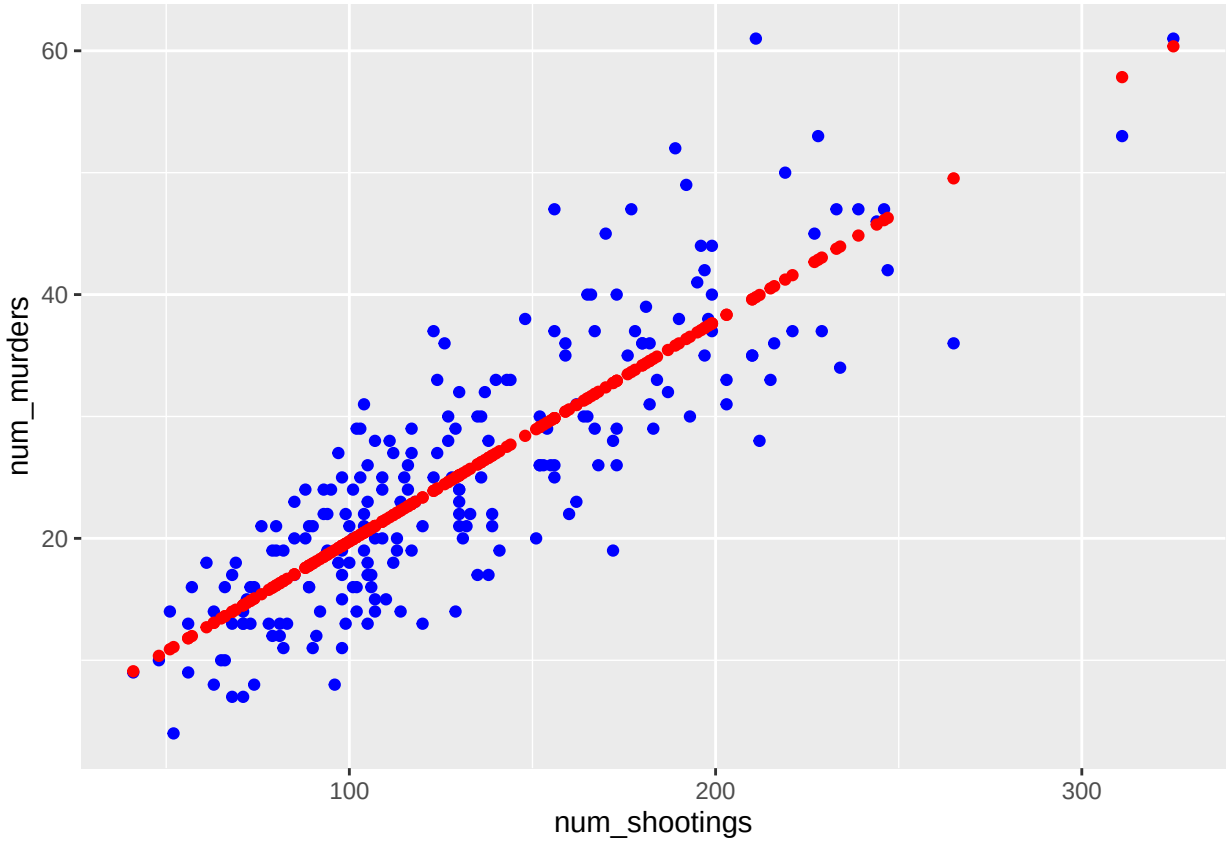
During this step I created a simple linear regression model that uses the number of shootings to predict the number of murders using the nyc_total data set we created in Step 3. Using summary() we were able to see that our model is in fact statistically significance as evidence by an extremely small p-value. Next, we add a new field "pred" to the data set that is predictions of the number of murders based on the number of deaths. I then plotted the actual number of murders against the number of shootings AND overlay-ed that with predicted number of murders against the number of shootings.

```
#Create linear regression model that uses the total number of shootings to
#predict the total number of murders
model <- lm(num_murders ~ num_shootings, data = nyc_total)

#View the summary to determine how good (or bad) the model is
summary(model)
```

```
##
## Call:
## lm(formula = num_murders ~ num_shootings, data = nyc_total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7469  -4.1240  -0.0699   3.7757  21.2128
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.697047   1.087611    1.56     0.12
```

```
## num_shootings 0.180522   0.007681   23.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.813 on 214 degrees of freedom
## Multiple R-squared:  0.7208, Adjusted R-squared:  0.7195
## F-statistic: 552.4 on 1 and 214 DF,  p-value: < 2.2e-16
```

```r
#Add a new field that uses the model to make predictions
nyc_total_w_pred <- nyc_total %>% mutate(pred = predict(model))

#visually compare the actuals against the predictions
nyc_total_w_pred %>% ggplot()+
  geom_point(aes(x=num_shootings, y=num_murders), color = "blue")+
  geom_point(aes(x=num_shootings, y=pred), color = "red")
```



Step 5 - Conclusion

In conclusion, I was able to load in the historical NYC shooting data set, remove fields that I did not want to use, clean up rows with missing data, analyze and visualize the data with intention, create a simply linear regression model, and identified some potential sources of bias in myself regarding this project.

At the start of the project, I found myself curious to learn about the data set and found the summary() function to be very helpful. The summary view informed my next set of decisions which was what fields I wanted to keep/remove and what kind of NA/Unknown data was in our data set. After I decided on the

fields to keep, I decided to remove the rows with missing data. I came to this decision because the number of rows with missing data were so few compared to the total number of rows that I didn't think it would affect the results.

I then came up with some general questions I wanted to investigate. These questions motivated my initial analysis and visualization which lead me to more questions and ultimately more analysis and more visualization.

It was very interesting to see how quickly new questions came up once I started exploring the data set. There were countless paths I wanted to investigate but found myself having to take a step back and re-focus for the sake of the assignment.

Ultimately, I was able to uncover some interesting findings. We can see from the line graphs that shootings tend to follow a similar pattern of high's and low's year after year based on the month. The summer months tend to have the highest number of shootings while the winter months tend to have lower numbers of shootings. This is supported by the stacked bar chart which shows the same trend but specifically for the year 2023. The bar chart also gives us some insight into which Boro's had the most shootings in the year of 2023. Staten Island consistently had the lowest and Brooklyn, and the Bronx typically had the most. I was also able to investigate the obvious theory that the number of shootings is a good predictor of the number of murders that will occur. This was accomplished by creating a linear regression model which was used to create predicted values. Finally, I plotted both the actual murders and the predicted murders on the same graph which clearly visualized a linear relationship.

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22631)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
##  [9] ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.4      highr_0.10        compiler_4.3.2    tidyselect_1.2.0
##  [5] scales_1.3.0      yaml_2.3.8        fastmap_1.1.1     R6_2.5.1
##  [9] labeling_0.4.3    generics_0.1.3    knitr_1.45        munsell_0.5.0
## [13] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.2       utf8_1.2.4
```

```
## [17] stringi_1.8.3     xfun_0.41         timechange_0.2.0  cli_3.6.2
## [21] withr_2.5.2       magrittr_2.0.3    digest_0.6.33     grid_4.3.2
## [25] rstudioapi_0.15.0 hms_1.1.3         lifecycle_1.0.4   vctrs_0.6.5
## [29] evaluate_0.23     glue_1.6.2        farver_2.1.1      fansi_1.0.6
## [33] colorspace_2.1-0  rmarkdown_2.25    tools_4.3.2       pkgconfig_2.0.3
## [37] htmltools_0.5.7
```