MyNYPDProject

A. Sorri

2024-07-24

Analyzing data from the NYPD Shooting Incident Data (Historic)

My Data Source: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic (https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic)

Description of Data: This data set contains every shooting recorded in New York from 2006 to 2023, at the time of this analysis. The data set includes data about the shooting location, date, and time, in addition to the demographics of the victims and perpetrators.

Libraries used include ggplot2, dplyr, and tidyverse.

Step 1 and 2: Start an RMD & Tidy/Transform Data

```
# Here I am reading in the data

library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(dplyr)
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWN
LOAD"
url_in
```

```
## [1] "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOA
D"
```

```
NYPD_data <- read_csv(url_in)
```

```
## Rows: 28562 Columns: 21
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
NYPD_data <- distinct(NYPD_data)

#Here I have removed several columns and made the date column a date type

NYPD_data = subset(NYPD_data, select = -c(JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD,
Latitude, Longitude, PRECINCT, Lon_Lat, LOC_OF_OCCUR_DESC, LOCATION_DESC, LOC_CLASSFC
TN_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE) )
NYPD_data$OCCUR_DATE <- as.Date(NYPD_data$OCCUR_DATE, format = "%m/%d/%y")

#Here I removed a single row with an uncategorized age group, the age group was "102
2".

NYPD_data <- subset(NYPD_data, NYPD_data$VIC_AGE_GROUP != "1022" )

print(summary(NYPD_data))
```

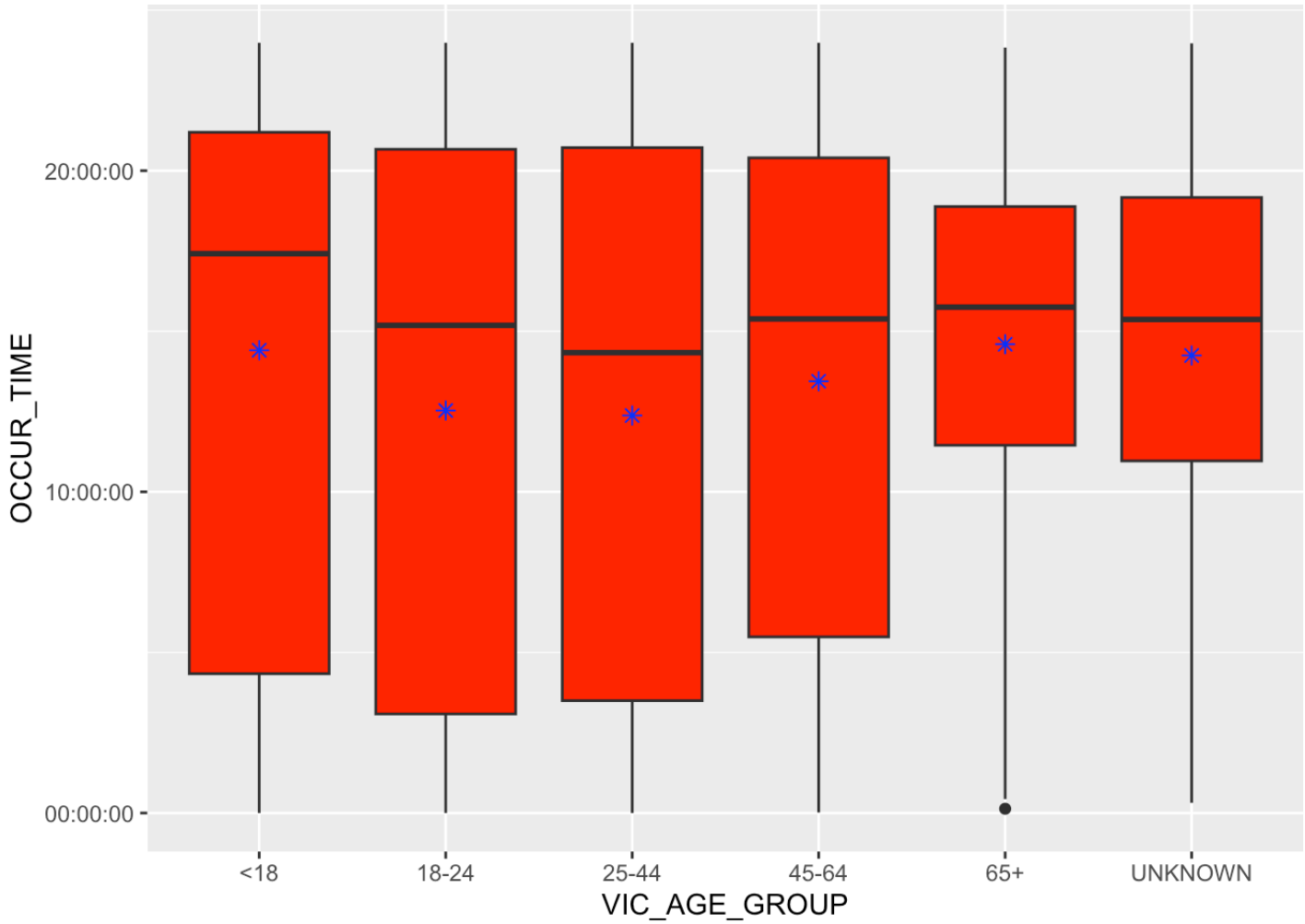
```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2020-01-01   Length:28561      Length:28561
##  1st Qu.: 65439914   1st Qu.:2020-05-01   Class1:hms        Class :character
##  Median : 92711253   Median :2020-07-13   Class2:difftime   Mode  :character
##  Mean   :127401585   Mean   :2020-07-10   Mode  :numeric
##  3rd Qu.:203083804   3rd Qu.:2020-09-23
##  Max.   :279758069   Max.   :2020-12-31
##  STATISTICAL_MURDER_FLAG VIC_AGE_GROUP        VIC_SEX
##  Mode :logical           Length:28561       Length:28561
##  FALSE:23035             Class :character   Class :character
##  TRUE :5526              Mode  :character   Mode  :character
##
##
##
##    VIC_RACE
##  Length:28561
##  Class :character
##  Mode  :character
##
##
##
```

Step 3 Vizualization and Analysis

Which Borough and Age Group is at the greatest risk of a Shooting Incident? When do Shootings Occur on Average? Is there a significant relationship between Victim Age and Murder?

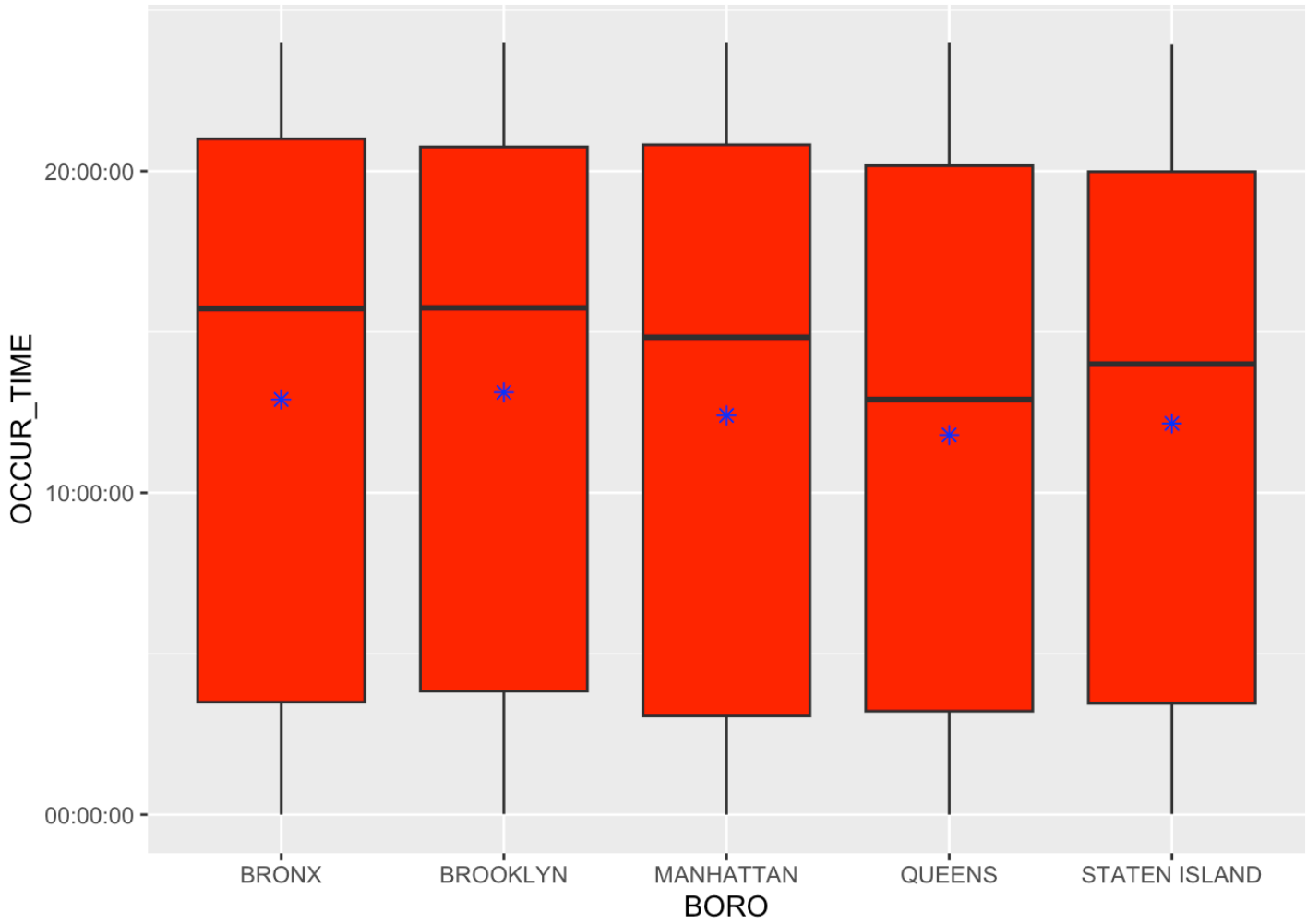
```
# A. Here I have created box plot to visualize the Victim Age Group against the Occur Time of the Shooting Incident. The mean is shown visually with the blue star on the b ox plot. It appears that the mean Occur Time is very similar across age groups, most shootings occurred in the late afternoon, however, for the 65+ and Unknown categories the range of the occurrence time is much narrower compared to the other groups.

library(ggplot2)
ggplot(NYPD_data, aes(x= VIC_AGE_GROUP, y = OCCUR_TIME)) +  geom_boxplot(fill='red')
+ stat_summary(fun=mean, colour="blue", geom="point", shape=8, size=2, show.legend=FA
LSE)
```



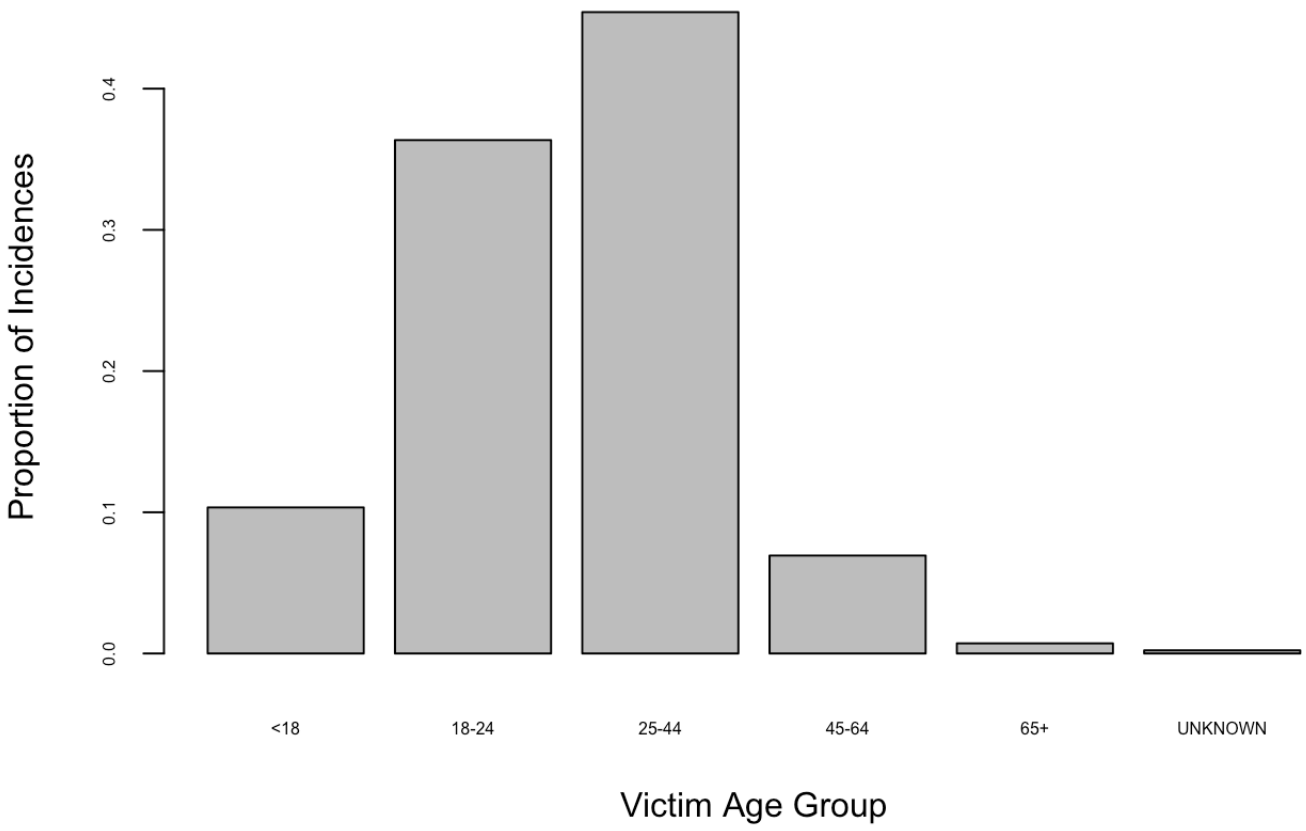
 # B. Here I have created a box plot to visualize the Borough against the Occur Time o f the Shooting Incidents. The mean is shown visually with the blue star on the box pl ot. The range and mean appear very similar to each other for each borough, but Queens has a slightly earlier average shooting occurrence time than the other boroughs. The Bronx and Brooklyn appear to have a slightly later mean occurrence time.

```
ggplot(NYPD_data, aes(x= BORO, y = OCCUR_TIME)) +  geom_boxplot(fill='red') + stat_su
mmary(fun=mean, colour="blue", geom="point", shape=8, size=2, show.legend=FALSE)
```



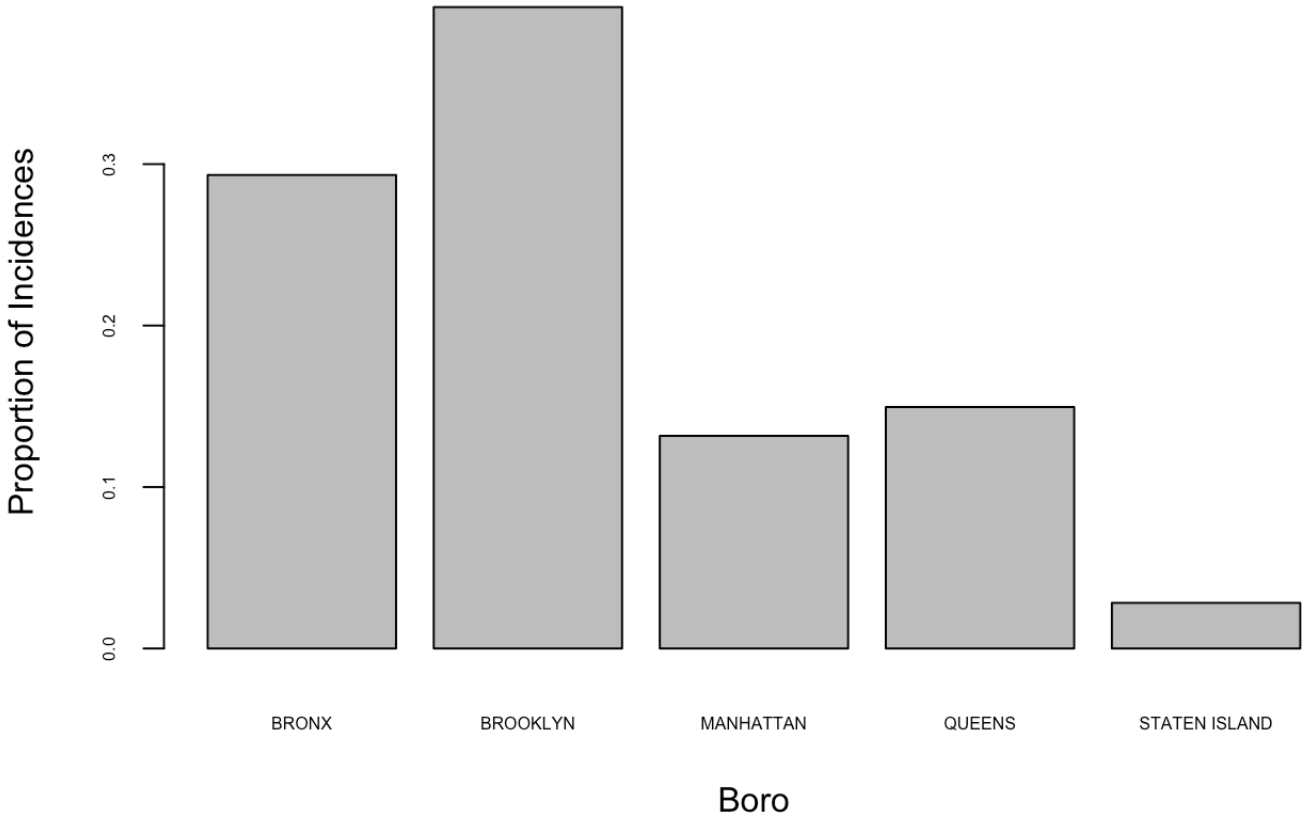

```
# C. The Bar Plot shows the proportion of people in each Victim Age Group, it appears that the category of people that experienced the least number of shooting are in the 65+ and unknown categories. The age group that had the highest number of shooting were the 25-44 year old age group.

barplot(prop.table(table(NYPD_data$VIC_AGE_GROUP)), cex.names=.5, cex.axis=.5, xlab = "Victim Age Group", ylab = "Proportion of Incidences")
```




```
#D. In the following bar plot Brooklyn has the highest proportion of incidents and St
aten Island has the least.

barplot(prop.table(table(NYPD_data$BORO)), width=c(.1,.1,.1,.1,.1), cex.names=.5, ce
x.axis=.5, xlab = "Boro", ylab = "Proportion of Incidences")
```




```
#E. Here I calculated the total number of incidences by Borough, Brooklyn has the mos
t shootings.

NYPD_data %>%
group_by(BORO) %>%
summarise(Total = n())
```

```
## # A tibble: 5 × 2
##   BORO          Total
##   <chr>         <int>
## 1 BRONX          8376
## 2 BROOKLYN      11346
## 3 MANHATTAN      3761
## 4 QUEENS         4271
## 5 STATEN ISLAND   807
```

```
#F. Here I calculated the total number of incidences by Victim Age Group, the 18-24 a
nd 25-44 age groups experienced the most shootings.

NYPD_data %>%
group_by(VIC_AGE_GROUP) %>%
summarise(Total = n())
```

```
## # A tibble: 6 × 2
##   VIC_AGE_GROUP Total
##   <chr>         <int>
## 1 18-24         10384
## 2 25-44         12973
## 3 45-64          1981
## 4 65+             205
## 5 <18            2954
## 6 UNKNOWN          64
```

```
#G. My linear model looks at the variable Victim Age Group, on the response, Statisti
cal Murder Flag. Assuming an alpha of .05, based on the F-test (F-statistic: 39.91) a
nd corresponding p-value (p-value: < 2.2e-16), there is a statistically significant r
elationship between the two variables. This linear model shows that there is a statis
tical significant relationship between the victim's age group and if the shooting res
ulted in that victim's death (murder).

lm_vic_age <- lm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, data = NYPD_data)
summary(lm_vic_age)
```

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, data = NYPD_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3220 -0.2188 -0.1666 -0.1303  0.8697
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.130332   0.007239  18.005  < 2e-16 ***
## VIC_AGE_GROUP18-24    0.036271   0.008204   4.421 9.85e-06 ***
## VIC_AGE_GROUP25-44    0.088430   0.008021  11.026  < 2e-16 ***
## VIC_AGE_GROUP45-64    0.118028   0.011425  10.331  < 2e-16 ***
## VIC_AGE_GROUP65+      0.191619   0.028415   6.744 1.58e-11 ***
## VIC_AGE_GROUPUNKNOWN 0.104043   0.049708   2.093   0.0363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3934 on 28555 degrees of freedom
## Multiple R-squared:  0.008309,   Adjusted R-squared:  0.008135
## F-statistic: 47.85 on 5 and 28555 DF,  p-value: < 2.2e-16
```

Step 4: Bias and Conclusion

Conclusion: In conclusion the range of occurrence times for shootings across boroughs didn't seem to drastically deviate, Queens appears to have the earliest mean shooting occurrence time, but all mean occurrence times occur in the afternoon. Occurrence time of shooting by Victim Age Group did have more variety. The 65+ and Unknown age groups had a much smaller range in occurrence times but all mean shooting occurrences were in the afternoon as well. The total number of shootings for each Borough and Victim Age Group were calculated and visualized in a bar plot. Age Groups 18-24 and 25-44 experienced the highest proportion of shooting and the 65+ and Unknown age groups experienced the smallest proportion of shootings. Brooklyn had the highest number of shootings and Staten Island had the least. I created a linear model between Victim Age Group and Statistical Murder Flag. According to the NYPD Data Shooting Data report, Statistical Murder Flag is marked as True, if the shooting resulted in the victim's death (a murder). According to the linear model, assuming an alpha of .05, there was a statistically significant relationship or correlation between the two variables.

Bias: Since there were a total of 64 incidents in which the victims' age was unknown, this could have lead to some bias; had their ages been known this could have changed the analysis done on the victim age groups. The NYPD Shooting data also only encompasses shooting that were recorded by the police, there could be bias in the data but we would have to know more about how the data was collected to determine that. Bias in policing or the recording of data could impact the data set but it cannot be said with certainty if there is such bias in this data set without further investigation. I don't believe that any personal bias affected the analyses of data in this report, but I do believe that I would have assumed older age groups to be less likely involved in shooting incidents. I don't have a grasp of the differences between the boroughs of New York, so I wouldn't really have any bias or assumptions there.

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.
0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapac
k.dylib;  LAPACK version 3.11.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## 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] sass_0.4.8        utf8_1.2.4        generics_0.1.3    stringi_1.8.3
##  [5] hms_1.1.3         digest_0.6.33     magrittr_2.0.3    evaluate_0.23
##  [9] grid_4.3.2        timechange_0.2.0  fastmap_1.1.1     jsonlite_1.8.8
## [13] fansi_1.0.6       scales_1.3.0      jquerylib_0.1.4   cli_3.6.2
## [17] rlang_1.1.2       crayon_1.5.2      bit64_4.0.5       munsell_0.5.0
## [21] withr_2.5.2       cachem_1.0.8      yaml_2.3.8        tools_4.3.2
## [25] parallel_4.3.2    tzdb_0.4.0        colorspace_2.1-0  curl_5.2.0
## [29] vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4   bit_4.0.5
## [33] vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0      bslib_0.6.1
## [37] gtable_0.3.4      glue_1.6.2        highr_0.10        xfun_0.41
## [41] tidyselect_1.2.0  rstudioapi_0.15.0 knitr_1.45        farver_2.1.1
## [45] htmltools_0.5.7   rmarkdown_2.25    labeling_0.4.3    compiler_4.3.2
```