Visualizations

Virginia Baskin, Siddhi Narayan, Grace Wang

2/21/2023

Read in data

Created smaller dataset to work with initially.

Our computers were struggling to work with the entire dataset, so we created a mini dataset by randomly sampling 1,000 rows. We used this dataset for initially creating the graphs/brainstorming, then we will ran the code on the large dataset as a final step.

```
library(readr)
mini_htx <- read_csv("Datasets/mini_htx.csv")
```

```
## Rows: 1000 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (13): raw_row_number, location, geocode_source, beat, subject_race, sub...
## dbl   (5): lat, lng, district, speed, posted_speed
## lgl   (1): citation_issued
## date  (1): date
## time  (1): 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.
```

```
#final step, eval with the full dataset
mini_htx <- read_csv("/Users/virginiabaskin/Downloads/tx_houston_2023_01_26.csv")
```

```
## Rows: 2045972 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (13): raw_row_number, location, geocode_source, beat, subject_race, sub...
## dbl   (5): lat, lng, district, speed, posted_speed
## lgl   (1): citation_issued
## date  (1): date
## time  (1): 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.
```

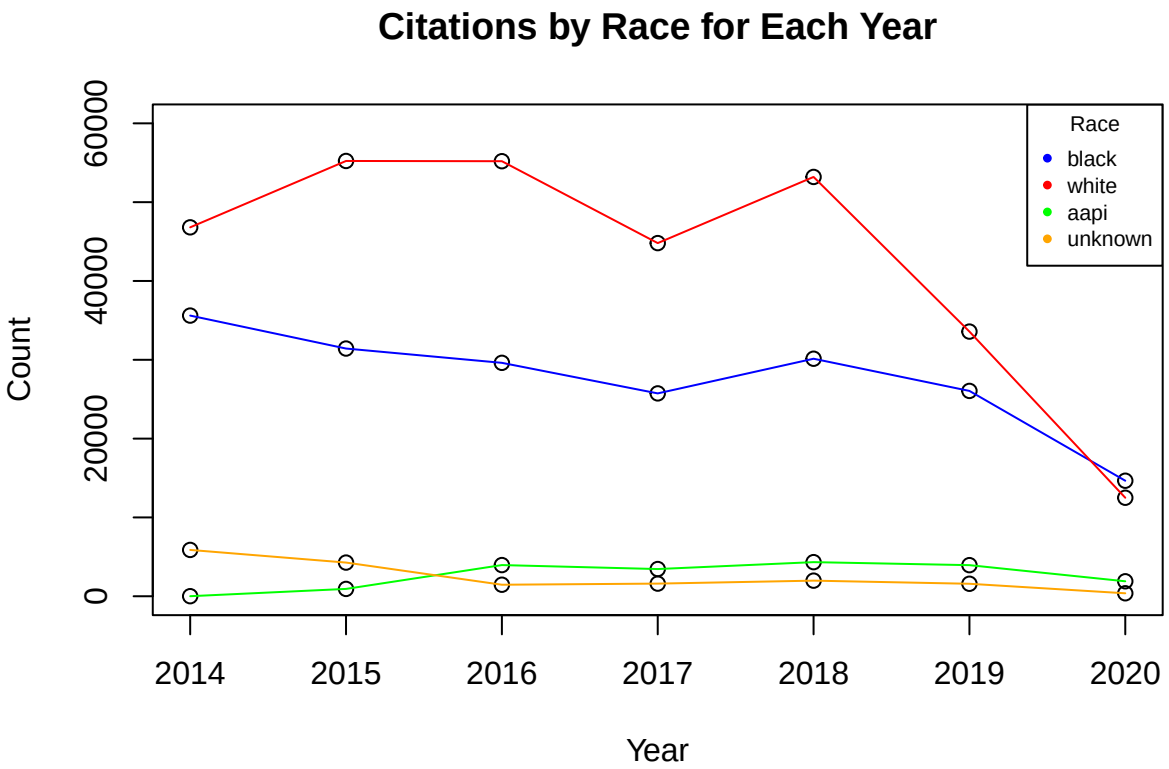
Visualizations

Before we go into visualizations, we would like to introduce our data. Our dataset is on police stops that result in citations in Houston, TX from 2015-2020. The data is from the Stanford Open Policing project and

contains a little over 2 million rows. The fields of the dataset include (non-exhaustively): the date, time, location, latitude, longitude of the stop, the police beat of the officer, district, the subject's race, sex, the type of violation they receive, and the vehicle make, model, and color.

Plot 1

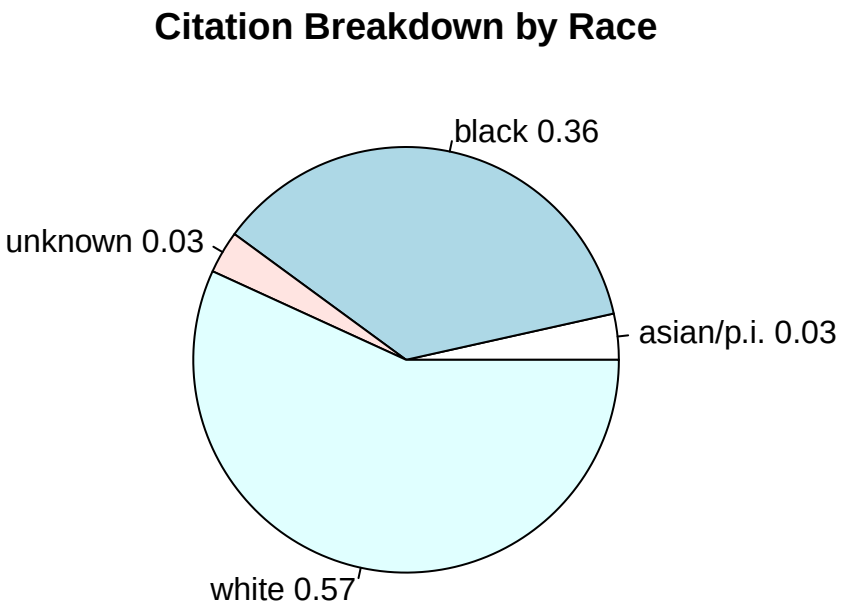
```
# Visualize totals for each race by year
# Grace

race_date <- mini_htx[,c("date", "subject_race")] # race and date
race_date <- na.omit(race_date) # omit na Values

year_only <- function(fulldate){
  year <- substring(fulldate, 1, 4) # only keep year
}

black <- race_date[race_date$subject_race=="black",]
black$year <- unlist(lapply(black$date, year_only))
plot(sort(unique(black$year)), as.vector(table((as.factor(black$year)))), ylim = c(0, 60000), main = "C
lines(sort(unique(black$year)), as.vector(table((as.factor(black$year)))), type = "l", col = "blue")

white <- race_date[race_date$subject_race=="white",]
white$year <- unlist(lapply(white$date, year_only))
points(sort(unique(white$year)), as.vector(table(as.factor(white$year))))
lines(sort(unique(white$year)), as.vector(table((as.factor(white$year)))), type = "l", col = "red")

aapi <- race_date[race_date$subject_race=="asian/pacific islander",]
aapi$year <- unlist(lapply(aapi$date, year_only))
points(sort(unique(aapi$year)), as.vector(table(as.factor(aapi$year))))
lines(sort(unique(aapi$year)), as.vector(table((as.factor(aapi$year)))), type = "l", col = "green")


unknown <- race_date[race_date$subject_race=="unknown",]
unknown$year <- unlist(lapply(unknown$date, year_only))
points(sort(unique(unknown$year)), as.vector(table(as.factor(unknown$year))))
lines(sort(unique(unknown$year)), as.vector(table((as.factor(unknown$year)))), type = "l", col = "orang

legend("topright", legend=c("black", "white", "aapi", "unknown"), col=c("blue", "red", "green", "orange
```



First, we turn our attention to the total individuals of each race issued citations in each year 2014-2020. From 2014-2019, the greatest number of individuals issued citations were white, but in 2020, black individuals narrowly surpassed white individuals as the racial group with the most citations of any racial group. Overall, individuals of unknown race and asian/pacific islanders received significantly fewer citations than white and black individuals. In 2019 and 2020, a decrease in citations issued to white individuals can be observed. This data is not sufficient for drawing conclusions of racial bias in the citation process; more investigation is necessary into confounding factors.

Plot 2

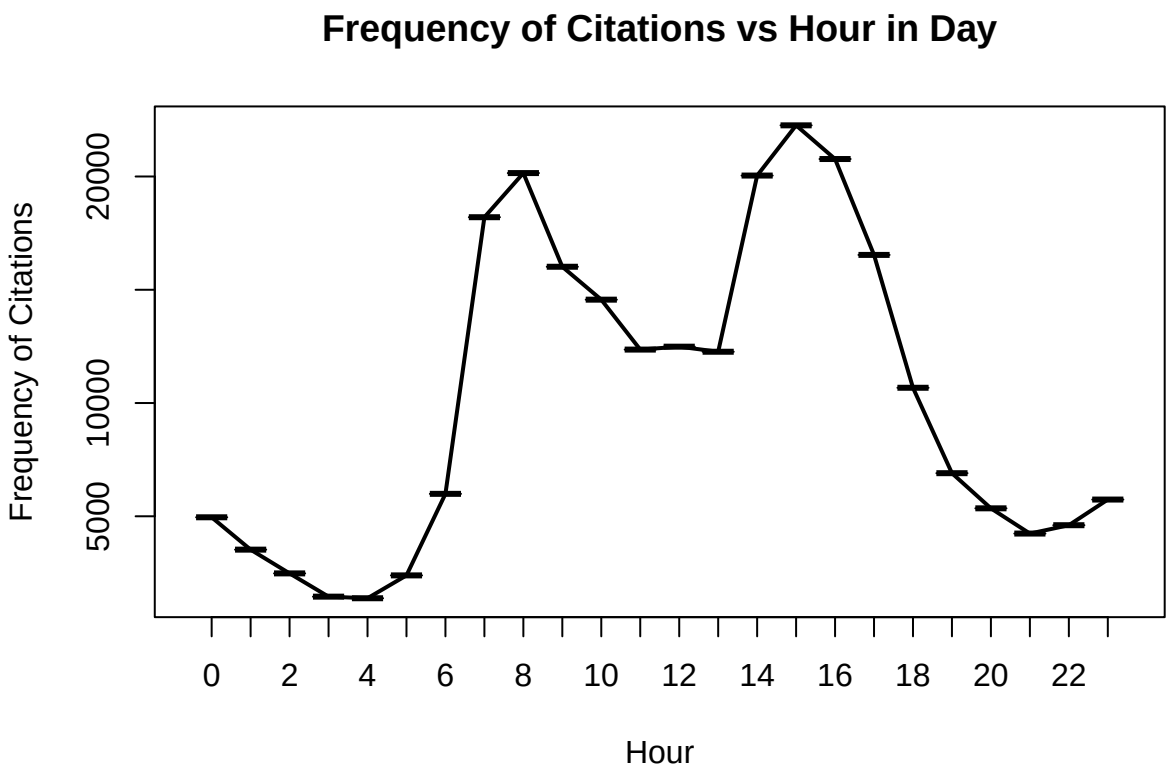
```
counts <- table(mini_htx$citation_issued, mini_htx$subject_race)
counts <- na.omit(counts)
frac <- counts/sum(counts)
pie(frac[-3], labels=paste0(c("asian/p.i. ", "black ", "unknown ", "white "), round(frac[-3], 2), sep="
```



Now we will look at the % of Citation by Race. As we can see, white individuals have the highest percentage of citations, followed by black individuals, and then Asian/pacific islanders and unknown races make up roughly 3% each. This is somewhat proportional to the actual demographics of Houston (according to the U.S. census) – the White and Black percentages of citations are slightly higher than their population in Houston, while the proportion of citations for Asian individuals is slightly lower than the proportion of Asian people in Houston.

Plot 3

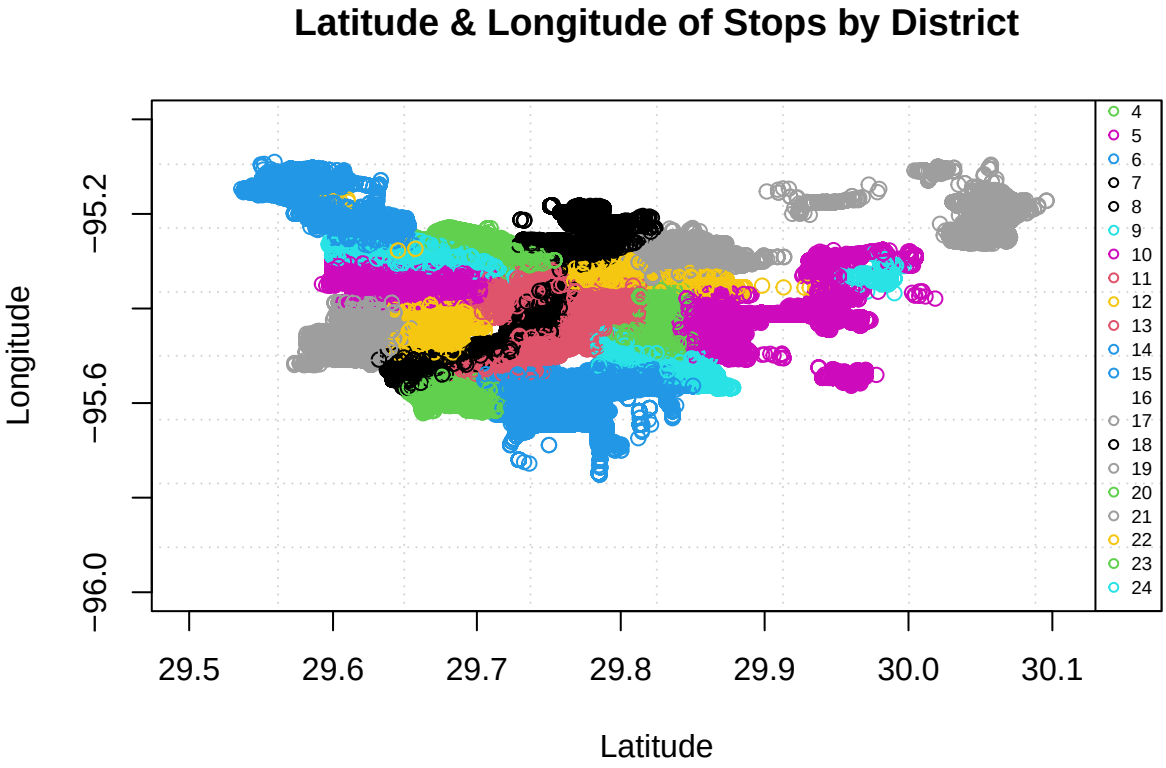
```
# Observe citations over time and any time-associate patterns
time <- as.POSIXct(mini_htx$time, format = "%H:%M:%S")
hour <- as.numeric(format(time, "%H"))
count <- data.frame(table(na.omit(hour)))
plot(count$Var1, count$Freq, type = "n", xlab = "Hour", ylab = "Frequency of Citations", main = "Freque
lines(count$Var1, count$Freq, type = "l", lwd = 2)
```



Now we will look at the Frequency of citations over the Hours of the Day.Citations are most common around 8 am, and around 3 pm. There may be a higher proportion of citations around 8 am because of high travel due to work. We are not quite sure why another peak is found at 3 pm, it may be people returning from school and/or work. We also see less citations during the super late/super early hours – likely due to less travel as a whole.

Plot 4

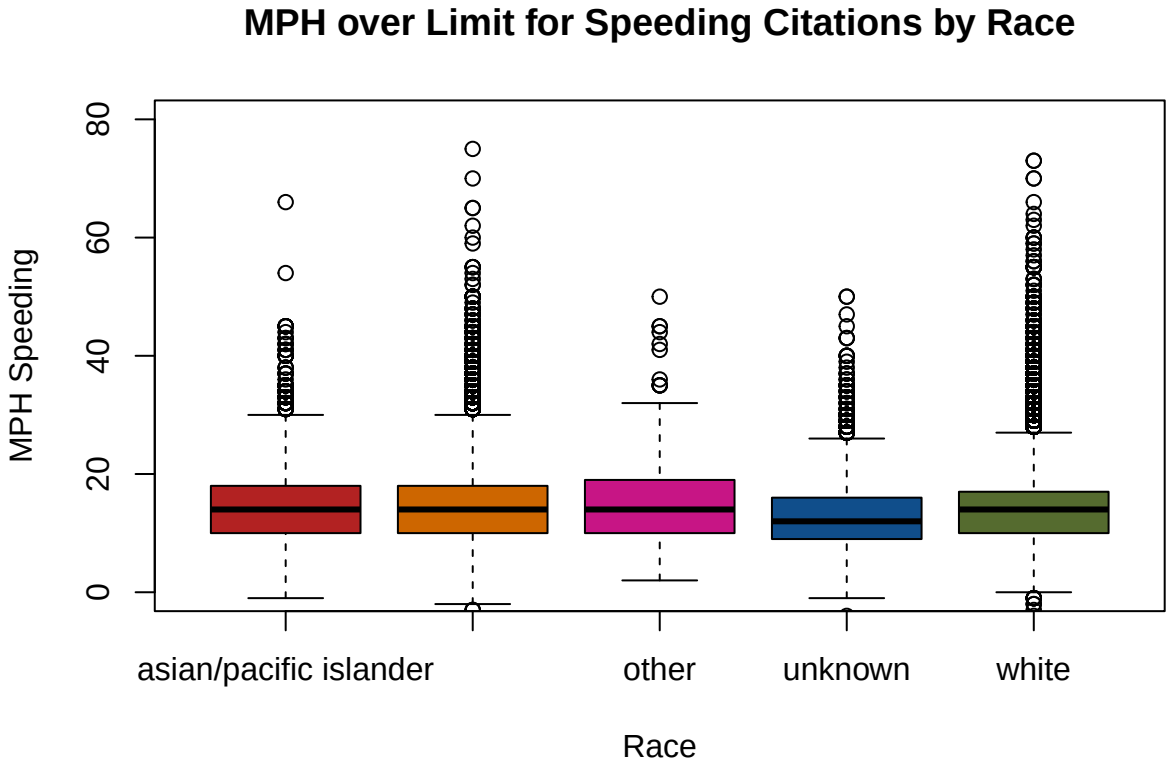
```
# Verifying the hubs for citation and locations of districts
mini_htx2 <- mini_htx
mini_htx2$district <- factor(mini_htx2$district)

plot(mini_htx2$lat, mini_htx2$lng, col = mini_htx2$district, panel.first = grid(8,8),  xlim = c(29.5, 3
legend("bottomright", legend=levels(mini_htx2$district), cex = 0.62, pch=1, col=unique(mini_htx2$distri
```



Next, we want to explore the spatial features of our data. We have both the latitude and the longitude of each stop, along with the district and beat. We knew that beat meant a motorized police unit that patrols a specific territory, but "district" in this context is not as clear, as it could mean multiple things. We investigated this by creating a scatterplot of the latitude versus the longitude of stops and colored it by district. We see very clear spatial grouping for each district, which means districts are not a police-defined feature but a location feature. From this plot, we can ascertain that district refers to Houston's subdistricts. [As seen here: https://www.houstontx.gov/police/pdfs/hpd_beat_map.pdf]

Plot 5

```
racediff <- mini_htx2[, c("subject_race", "diff")]
racediff <- na.omit(racediff)

colors = c("firebrick", "darkorange3", "mediumvioletred", "dodgerblue4", "darkolivegreen")

# change the colors to looks better
boxplot(diff~subject_race, data = racediff, ylim = c(0,80), xlab="Race", ylab="MPH Speeding", col = colo
```

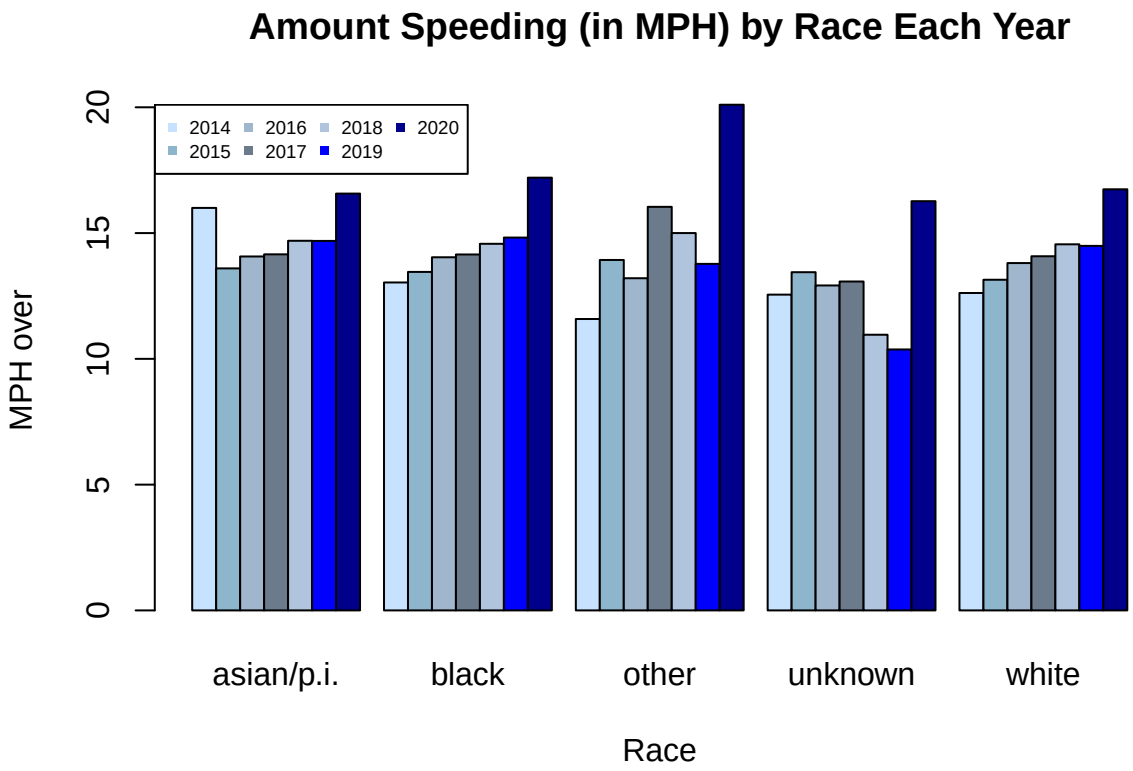


To further explore the possibility of racial bias in citations, we created a box plot of miles per hour over the speed limit for speeding citations by race. We aimed to investigate whether any racial group(s) demonstrated a distribution of speed differences that was significantly lower than any other race, which might indicate bias in the citation process. However, from the plot, we found that all races had a fairly similar median and overall distribution of MPH over limit. We also found that several races had minimums and outliers that were negative, indicating that these individuals were pulled over when driving under the speed limit; these citations may have been issued for reasons other than speeding, but further investigation into these data points is warranted.

Plot 6

```
#Virginia
mini_htx2$subject_race <- factor(mini_htx2$subject_race)
date <- as.POSIXct(mini_htx$date, format = "%Y/%m/%d")
year <- as.numeric(format(date, "%Y"))
mini_htx2$year <- year
uni_year <- sort(unique(year))

agg_tx <- aggregate(mini_htx2$diff~mini_htx2$year + mini_htx2$subject_race, data=mini_htx2, FUN=mean)

# RESHAPE

data <- reshape(agg_tx, idvar = "mini_htx2$year",
                timevar = "mini_htx2$subject_race", direction = "wide")
row.names(data) <- data$`mini_htx2$year`
data <- data[ , 2:ncol(data)]
colnames(data) <- c( "asian/p.i.", "black", "other", "unknown", "white")
data <- as.matrix(data)
```

```r
data <- data[-8,]

colors <- c("slategray1", "lightskyblue3", "slategray3", "slategray4", "lightsteelblue", "blue", "darkb
barplot(height = data, beside = TRUE, col = colors, main = "Amount Speeding (in MPH) by Race Each Year"
legend("topleft", legend = uni_year[-1], cex = 0.6, ncol = 4, pch = 15, col = colors)
```

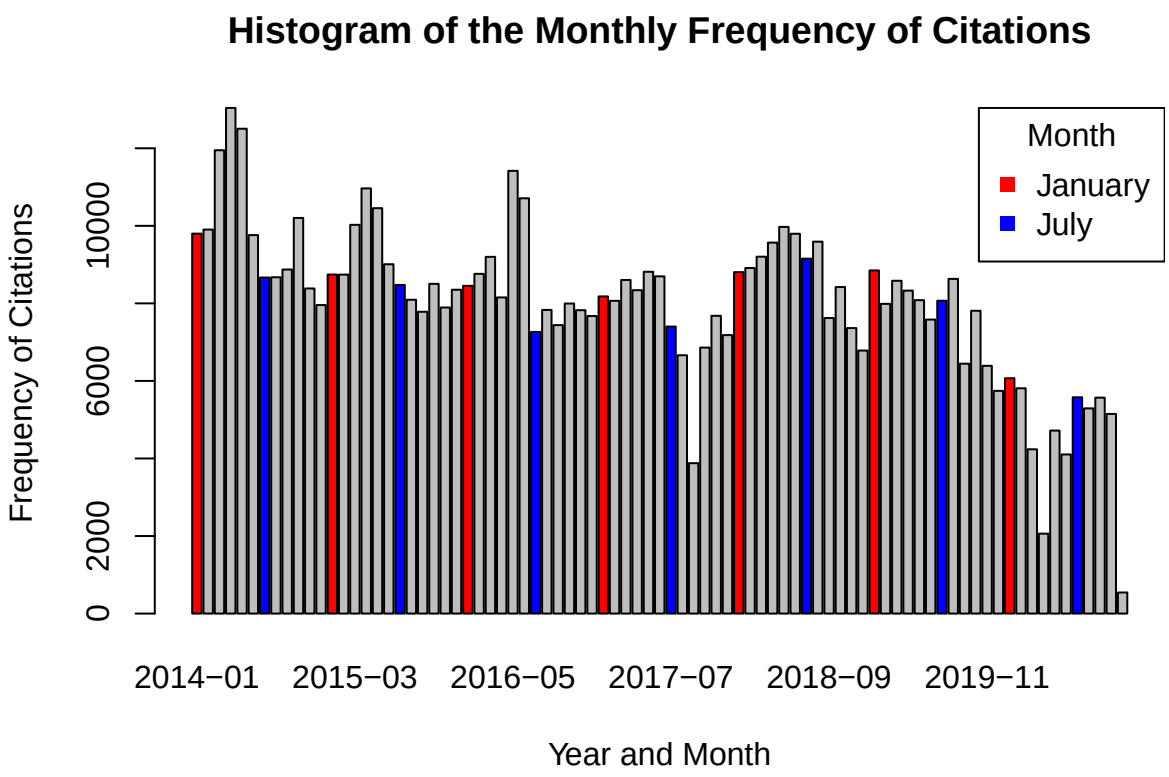


With this plot, we sought to investigate the average amount of speeding that resulted in a citation and how these values varied across years. We added further complexity to our investigation by separating these averages by race. In terms of race, our findings were consistent with the pie chart: we saw that the average amount of speeding that warranted a ticket was fairly similar between races, although individuals of unknown race had the lowest average amount of speeding that resulted in a citation across all years. Separating the data by year also gave us new insights as we saw that in 2020, the average amount of speeding that resulted in a ticket was the highest for all racial groups. This is interesting when considering our previous data which indicated a drop off in total citations in 2020. Perhaps the decrease in citations in 2020 is related to the increase in the amount of speeding an individual must perform in order to receive one.

Plot 7

```r
# x axis is the time of day
# y is frequency
# Histogram

dates <- mini_htx$date # get the dates column

yearmonth <- function(fulldate){ # function to extract year and month
  fulldate <- substring(fulldate, 1, 7) # delete day
}
```

```r
ym_dates <- unlist(lapply(dates, yearmonth)) # apply function
ym_dates <- as.factor(ym_dates)              # save unique dates

barplot(table(ym_dates), xlab = "Year and Month", ylab = "Frequency of Citations", col = rep(c("red", "
     main = "Histogram of the Monthly Frequency of Citations")

legend("topright", legend=c("January", "July"), col=c("red", "blue"), title="Month", pch=15)
```



Finally, we evaluated the frequency of citations for each month in the dataset in order to observe any monthly citation patterns over the years. This histgram gives the number of citations given in each month in the dataset, with January and July of every year colored red and blue, respectively. From 2014-2016, citations increased from January into the early months of Spring, but this pattern did not consistently persist into the following years. No singular month stood out as consistently yielding the most or least citations between years. We also observed that 2020 had the least number of total citations overall and began with the least number of citations in January for all years.