

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 run 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/virginia_baskin/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")

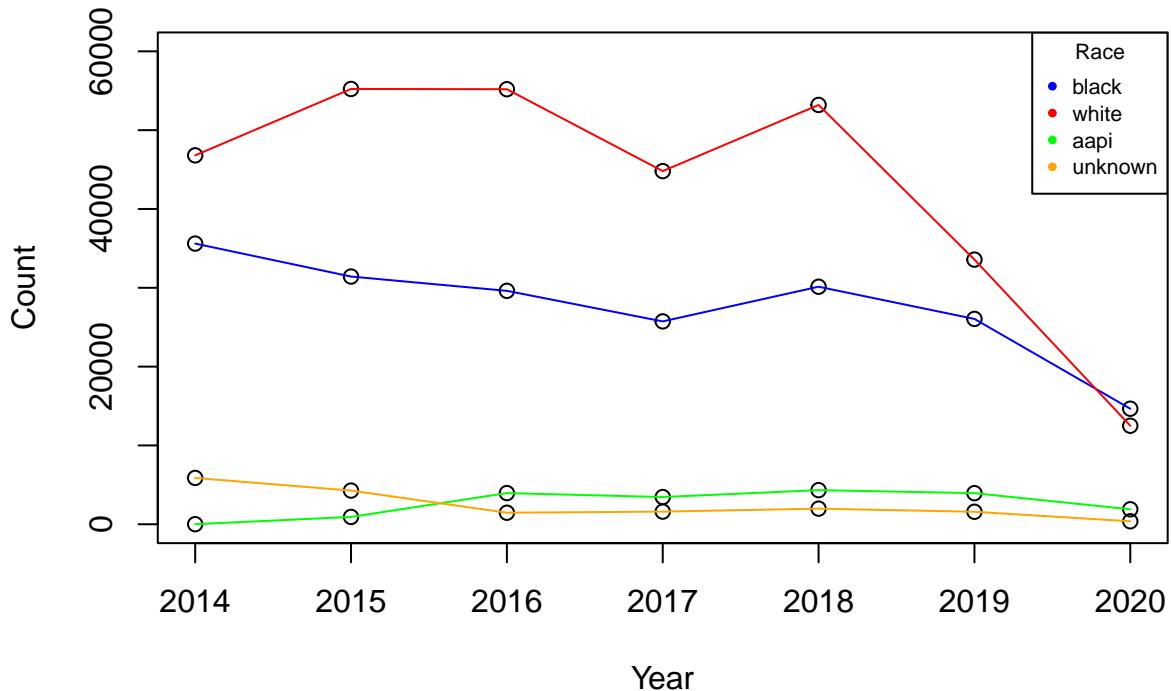
white <- race_date[race_date$subject_race=="white",]
white$year <- unlist(lapply(white$date, year_only))
points(sort(unique(white$year)), as.vector(table(as.factor(white$year))))
lines(sort(unique(white$year)), as.vector(table(as.factor(white$year))), type = "l", col = "red")

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 = "orange")

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

Citations by Race for Each Year

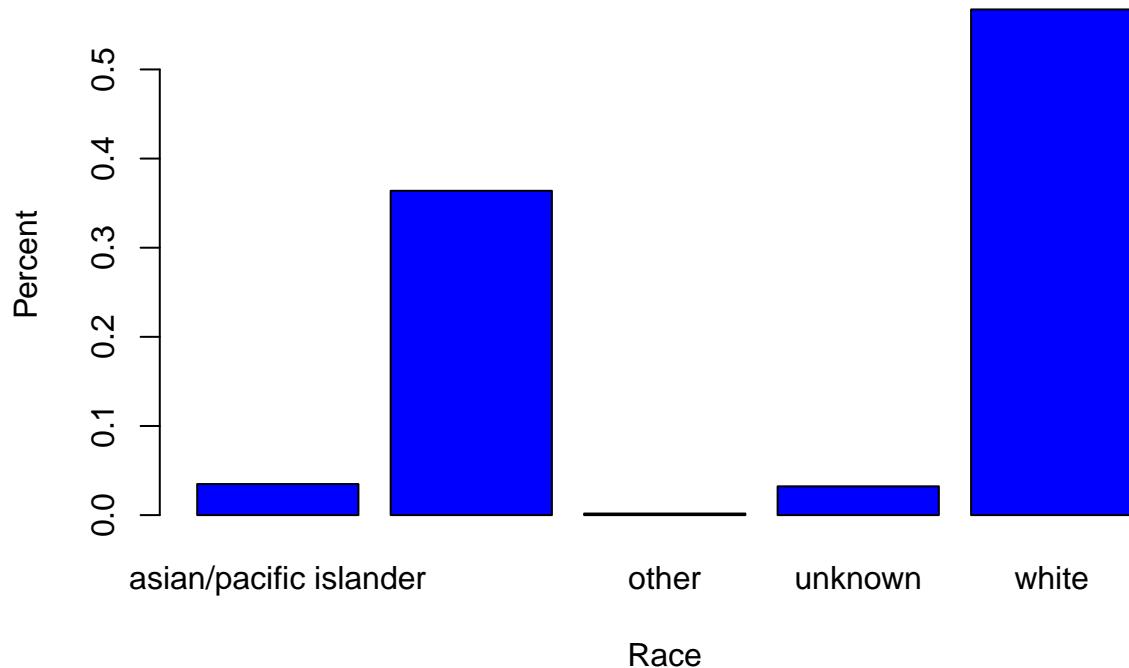


First, we turn our attention to the total individuals of each race issued citations in each year 2014-2020. From 2014-2018, the greatest number of individuals issued citations were white, but in 2019 and 2020, black individuals were issued the most citations of any racial group. Overall, individuals 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

```
# x axis: race
# y axis: %
counts <- table(mini_htx$citation_issued, mini_htx$subject_race)
counts <- na.omit(counts)
barplot(counts/sum(counts), main = "% Citation by Race", xlab = "Race", ylab = "Percent", col = "blue", ...)
```

% Citation by Race

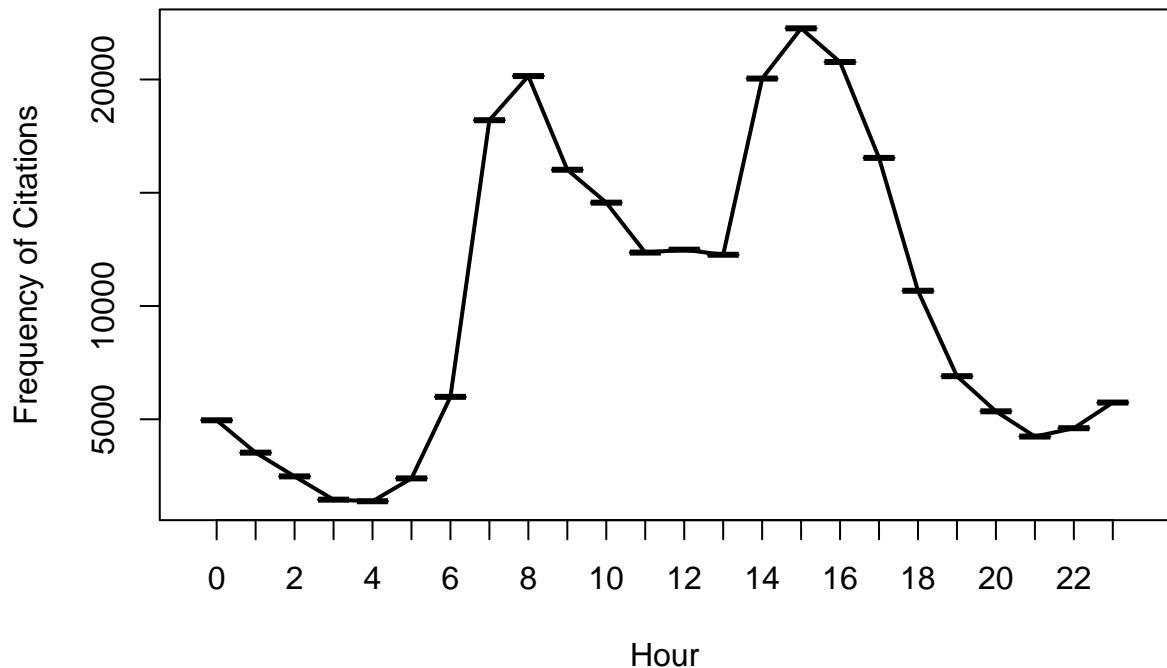


Now we will look at the Citation Count by Race. As we can see, white individuals have the highest percentage of citations, followed by black individuals, unknown races, and finally, Asian/pacific individuals. This is somewhat proportional to the actual demographics of Houston (according to the U.S. census) – the White and Black percentages of citations are slightly higher than their population in Houston, while the proportion of citations gotten by 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$Freq, type = "n", xlab = "Hour", ylab = "Frequency of Citations", main = "Frequency of Citations by Hour")
lines(count$Var1, count$Freq, type = "l", lwd = 2)
```

Frequency of Citations vs Hour in Day



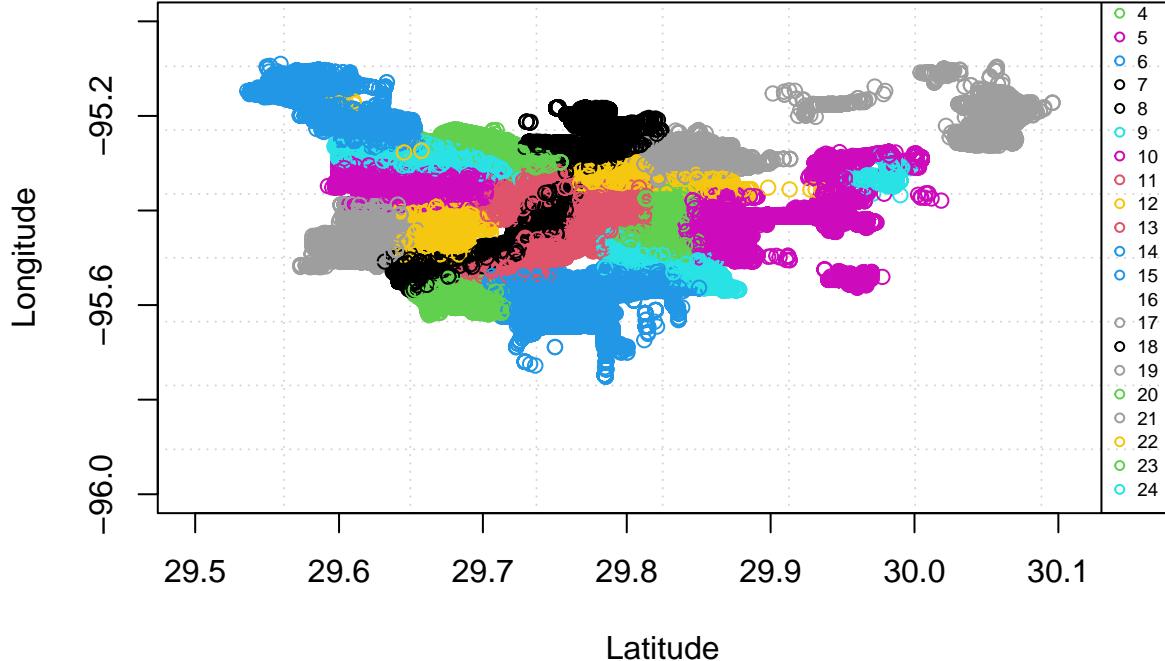
Now we will look at the Frequency of citations over the Hours of the Day. Citations are most common around 8 am, and around 3 pm. There may be a higher proportion of citations around 8 am because of high travel due to work. We are not quite sure why another peak is found at 3 pm, it may be people returning 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$district, panel.first = grid(8,8), xlim = c(29.5, 30.5), ylim = c(40.5, 41.5), asp = 1, legend("bottomright", legend=levels(mini_htx2$district), cex = 0.62, pch=1, col=unique(mini_htx2$district)))
```

Latitude & Longitude of Stops by District



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

```

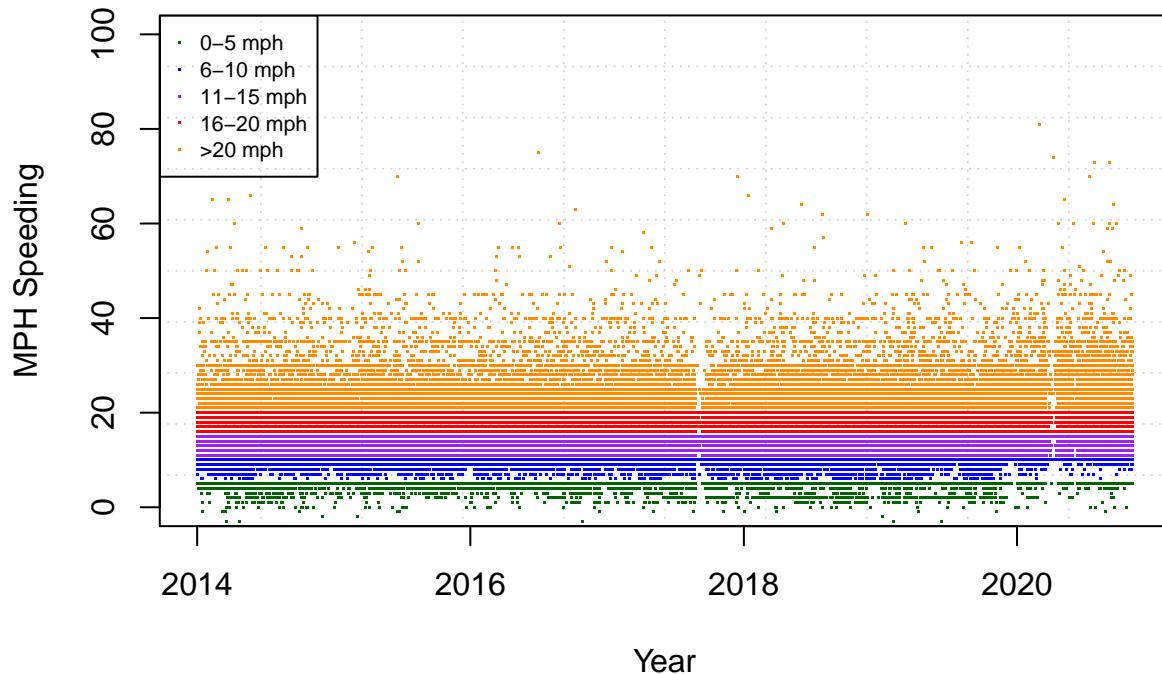
mini_htx2$diff <- (mini_htx2$speed - mini_htx2$posted_speed)
diff <- mini_htx2$diff

plot(mini_htx2$date, mini_htx2$diff, main = "MPH over Limit for Speeding Citations 2014-2020",
      xlab = "Year", ylab = "MPH Speeding", pch = ".", panel.first = grid(10,10), ylim = c(0,100),
      col = ifelse((diff <= 5), "dark green",
                  ifelse((diff <= 10), "blue",
                        ifelse((diff <= 15), "purple",
                              ifelse((diff <= 20), "red", "dark orange")))))

legend("topleft", legend = c("0-5 mph", "6-10 mph", "11-15 mph", "16-20 mph", ">20 mph"),
      col = c("dark green", "blue", "purple", "red", "dark orange"), pch = ".", cex = 0.7)

```

MPH over Limit for Speeding Citations 2014–2020



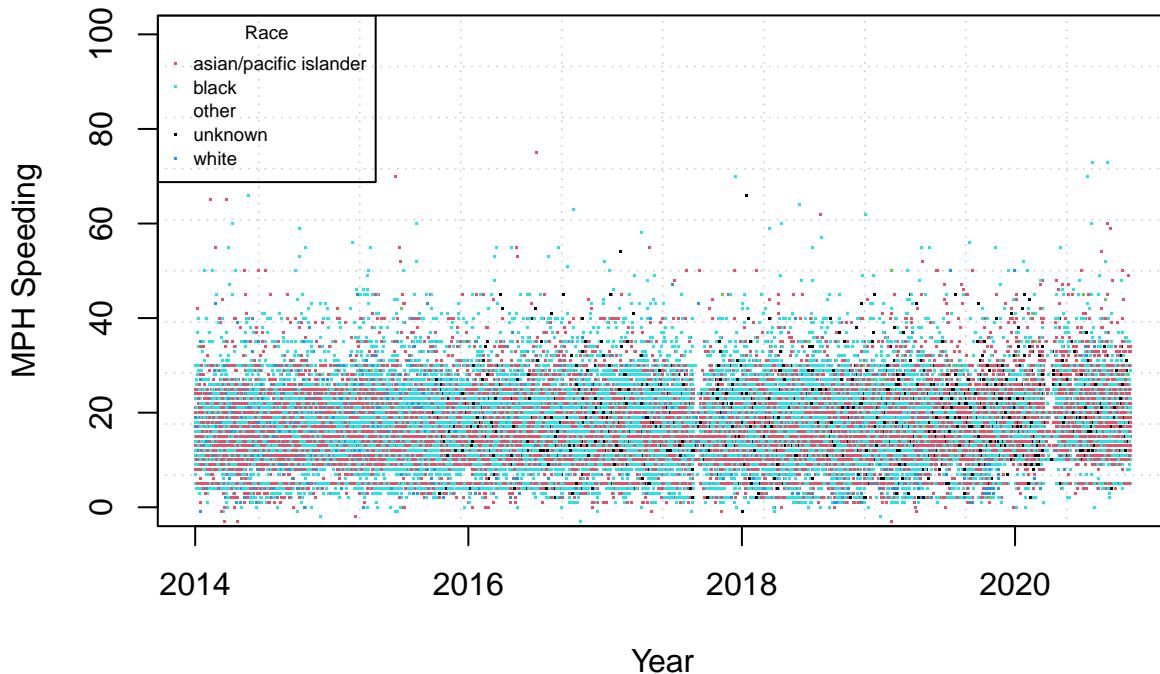
This plot investigates citations that are given for speeding. Namely, by how much were people speeding over the limit to have been issued a citation. The y axis is mph over the speed limit (as recorded on the citation) and we colored different instances of citation by levels of severity. The x axis is the year/general time from the citation occurred in, and there does seem to be some evidence of temporal trends. There is a vertical line before 2018 where it seems like there were less tickets given overall, which could be a number of things that we could look into. It could be something about how the data was collected, if there was some data lost at any point, or maybe HPD deprioritized patrolling for speeding during that time period. There is no way to tell from just looking at the graph. Another interesting feature is that there seems to be less citations given for speeding less than 10mph over the speed limit 2020 and onward.

Plot 5.5

```
plot(mini_htx2$date, mini_htx2$diff, main = "MPH over Limit for Speeding Citations 2014–2020 by Race",
      xlab = "Year", ylab = "MPH Speeding", pch = ".", panel.first = grid(10,10), col = factor(mini_htx2$subject_race))

legend("topleft", legend=levels(factor(mini_htx2$subject_race)), cex = 0.62, pch=".",
      col=unique(factor(mini_htx2$subject_race)))
```

MPH over Limit for Speeding Citations 2014–2020 by Race



Just for fun, I colored the same points as the previous graph by using the person's race to investigate if there were any observable affects there.

For some reason, visually it looks like more asian/pacific islanders were given speeding citations past 2019 than from 2014-2018. This pattern needs additional investigation before we can decern if it is a true pattern or simply an anomaly from the data (or lack there of).

Plot 6

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

dates <- mini_htx$date # get the dates column

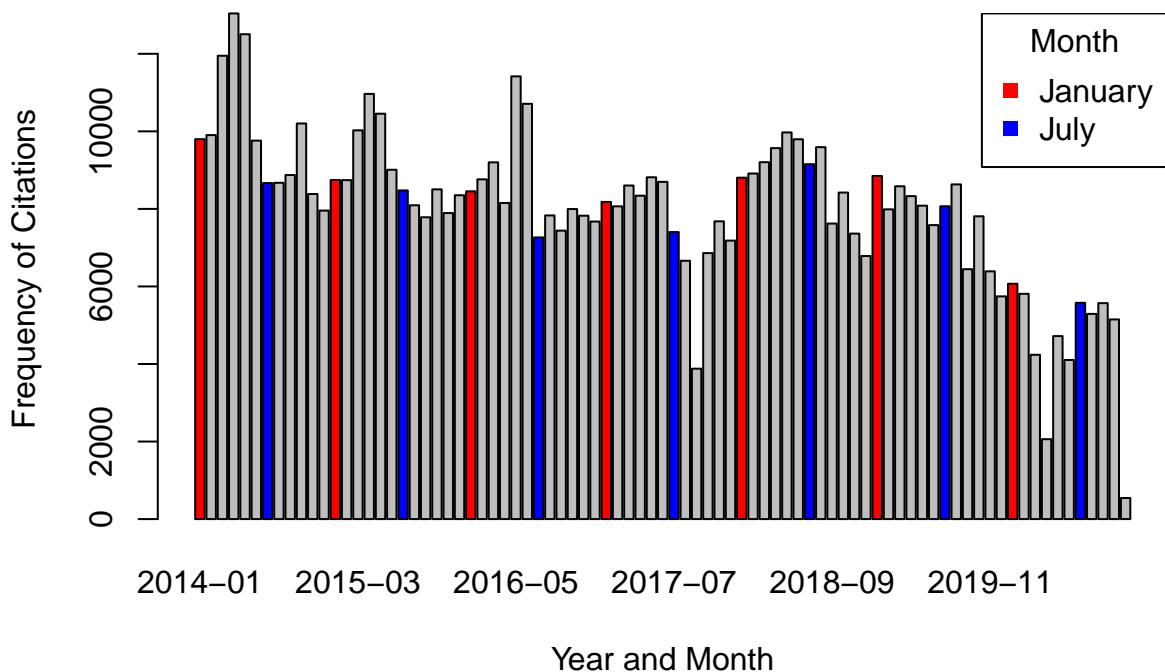
yearmonth <- function(fulldate){ # function to extract year and month
  fulldate <- substring(fulldate, 1, 7) # delete day
}

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", "blue"), 12),
       main = "Frequency of Citations by Month")

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

Frequency of Citations by Month



Finally, we evaluated the frequency of citations for each month in the dataset in order to observe any monthly citation patterns over the years. From 2014-2016, citations experienced a mid-year dip in July, but this pattern did not consistently persist. No singular month stood out as consistently yielding the most or least citations between years. We also observed that 2019 and 2020 had the least number of total citations overall and began with the least number of citations in January for all years.