Math 748: Professor Tao He

Course Project: Progress Report I

Who Did We Stop This Time?

Gabrielle Salamanca

October 4, 2024

Contents

1. Introduction

There can be multiple reasons why a police officer would stop someone on the road. Perhaps the license plate has expired registration tags, someone was speeding, traffic violations, or someone was driving under the influence to name a few. And any of these can be issued a citation, meaning the culprit is being legally charged with violating the traffic law.

For this project, I want to find out if there's a pattern to receiving a citation, if certain characteristics or circumstances are more likely in receiving one.

2. The Data

a. Description of the Data

In the Rdatasets of Vincent Arel Bundock's Github, the dataset called Minneapolis Police Department 2017 Stop Data was found, where it holds results of nearly all stops made by the Minneapolis Police Department for the year 2017. This dataset is just a sample from http://opendata.minneapolismn.gov/datasets/police-stop-data, which actually begins from August 1, 2017 and is still being up to date to this day. The original holds 232,732 records as of October 4, 2024; while the Github one holds 51,920, only because it specifically extracted those from 2017.

The documentation lists 14 variables: idNum, date, problem, citationIssued, personSearch, vehicleSearch, preRace, race, gender, lat, long, policePrecinct, neighborhood, MDC. There was a 15th one when ran through R, and it was called rownames. Descriptions of these variables will be found in the next section.

b. Explanation of Variables

The following table contains a brief description of variables of the dataset, which were taken from the documentation provided by Vincent Arel-Bundock's Github projects. If there were other variables not included in the documentation, they will be added.

Variable	Description
rownames	Entry number of the incident
idNum	Character vector of incident identifiers
date	A POSIXlt date variable giving the date and time of the stop
problem	A factor with levels *suspicious* for suspicious vehicle or person stops and *traffic* for traffic stops
citationIssued	A factor with levels *no, yes* indicating if a citation was issued
personSearch	A factor with levels *no, yes* indicating if the stopped person was searched
	A factor with levels *no, yes* indicating if a vehicle was searched
preRace	A factor with levels *white, black, east african, latino, native american, asian, other, unknown* for the officer's assessment of race of the person stopped before speaking with the person stopped
race	A factor with levels *white, black, east african, latino, native american, asian, other, unknown*; officer's determination of race after the incident
gender	A factor with levels *female, male, unknown*; gender of person stopped
lat	Latitude of the location of the incident, somewhat rounded
long	Longitude of the location of the incident, somewhat rounded

policePrecinct	Minneapolis Police Precinct number
neighborhood	A factor with 84 levels giving the name of the Minneapolis neighborhood of the incident
MDC	A factor with levels *mdc* for data collected via in-vehicle computer, and for data submitted by officers not in a vehicle, either on foot, bicycle or horseback. Several of the variables above were recorded only in vehicle

c. Data Exploration & Cleaning

As stated, the original dimensions of the Github dataset were 51920 rows and 15 columns. When reading the dataset into R and running the summary function, a majority of the variables were characters. There were at least three variables I can convert into numbers: citationIssued, personSearch, and vehicleSearch. Those simply have yes and no as their inputs, so it's an easy conversion to 1's and 0's. I personally am debating if I should also do this to problem and gender, but this will have to be considered when I start doing regression tests.

To further explore the data, I did run the summary function to see if any NAs were within, but the output said all columns had false for all rows. I personally did not believe it, so I viewed the dataset myself, and was immediately greeted with empty rows in citationIssued. Because my main goal is to see if certain characteristics of the person or certain circumstances lead to being an issued a citation or not, I had to clear all the records that did not have that column filled. From 51,920 records, it was trimmed down to 19,110. And, if other columns had empty cells, I filled them in with Unknown.

Now, there were a few variables that I deemed unnecessary for my project: rownames, lat, long, and MDC. Rownames were unnecessary, because we already have a numbered system within R. Lat and long were specified coordinates of the incident, but I didn't find it relevant,

especially when neighborhood would make more sense when it came to grouping the points. Finally, MDC was removed, because it didn't seem relevant for what I want to answer.

d. Summary Results

When reading the dataset into R and running the summary function, a majority of the variables were characters; and there was an attempt to use as.numeric(), but it wouldn't give me counts when I tried running the summary function again. For example, I tried the problem variable and wrote summary(as.numeric(problem)), receiving an output that was just NA's all around. So, I needed another way to get a visual of how spread out the data was with each column, and thus, the table function provided what I need.

I will admit that I need to figure out how to extract the exact month from the date, because I believe it would be a better way to group results. As of right now, I have not, so I mainly have a month and day spread. There are dates that had triple digits, so it did bring a possibility of trying to group my Y by the days of the week. However, that would be far more work than I can do right now, so that would need be shelved.

It was halfway through that I realized I could use stringsAsFactors to give a count when the summary function was run. However, it was only with the main dataset that worked, so the subset will need to settle for the table function for now. Results of the variables, minus the date and neighborhood due to being too large to include, will be shown down in data visualization.

e. Data Visualization

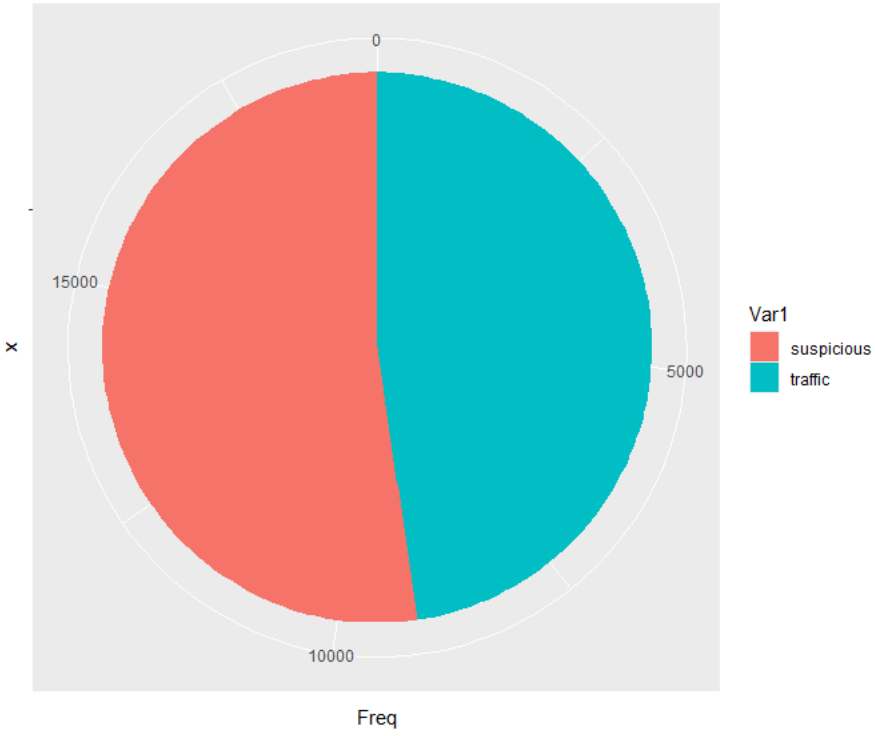

Figure 1: Type of Problem

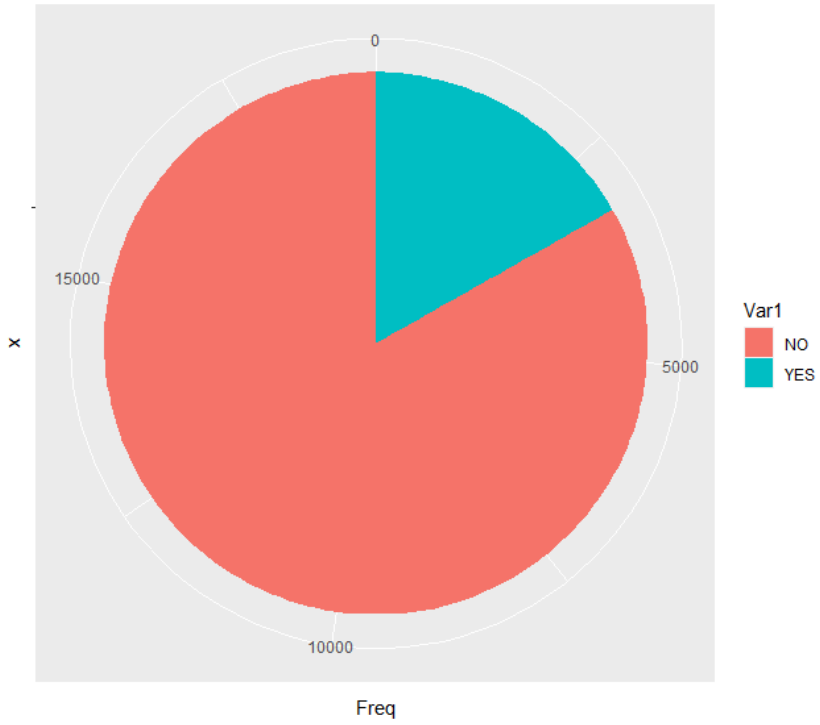

Figure 2: Was a Citation Issued?

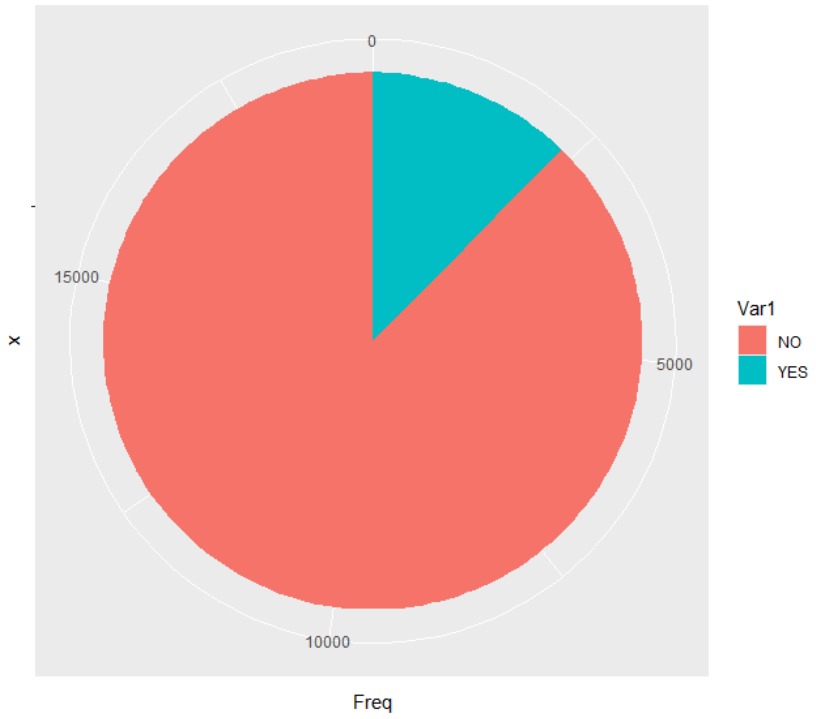

Figure 3: Was the Person Searched?

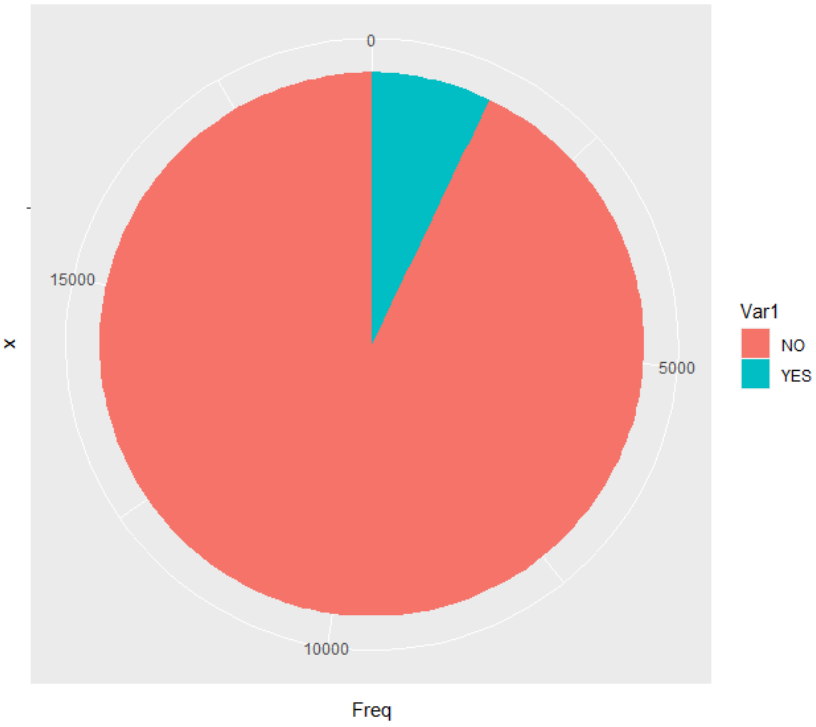

Figure 4: Was the Vehicle Searched?

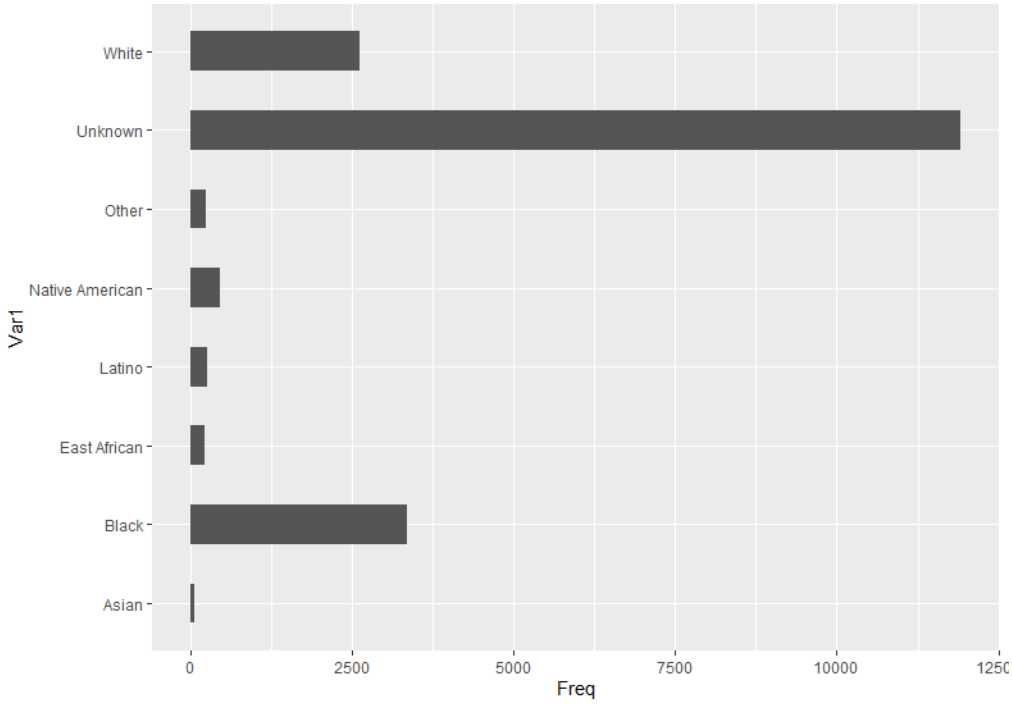

Figure 5: Assumed Race of the Person Before Speaking with Them

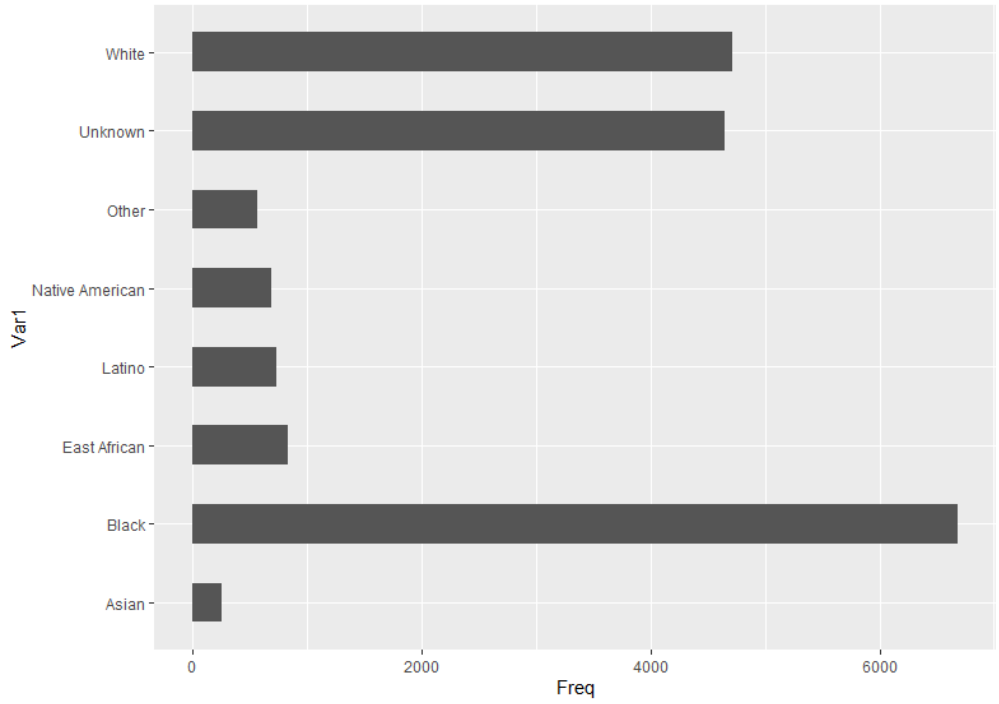

Figure 6: Determined Race after the Incident

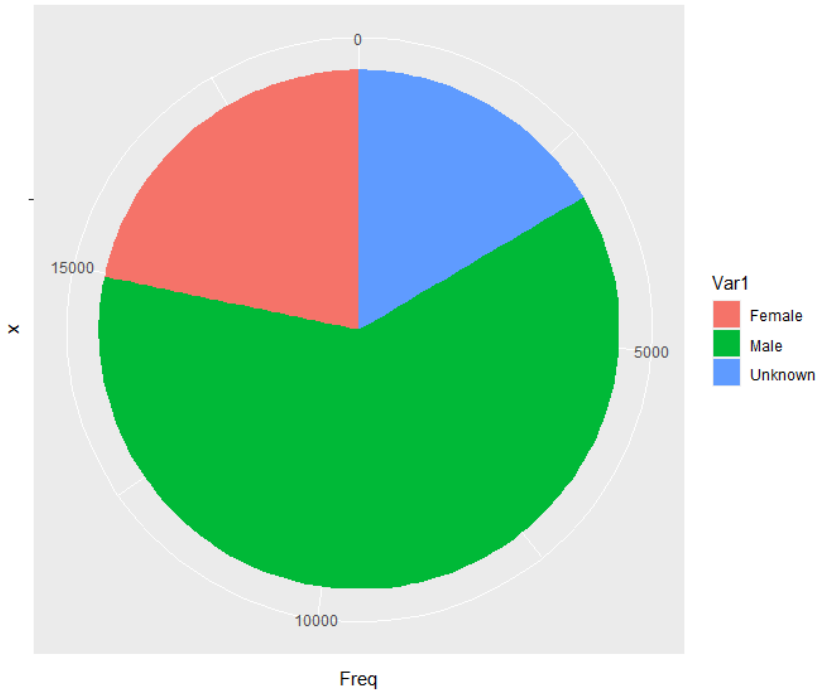

Figure 7: Gender

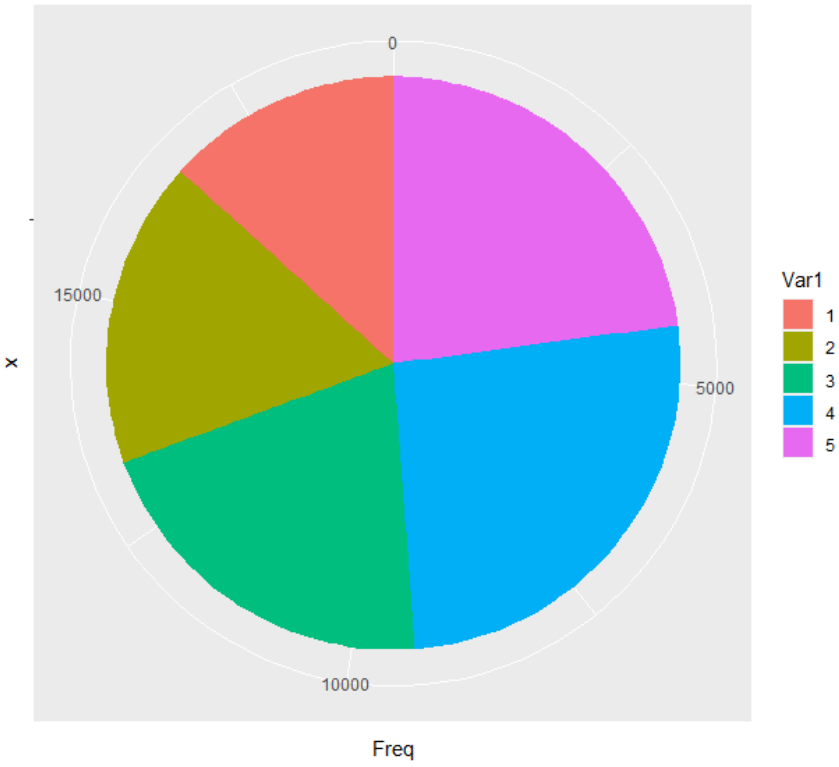

Figure 8: Police Precinct

3. Code

```
############
 # Libraries #
 ##########
 library(dplyr)

library(ggplot2)

library(psych)

library(readr)

library(readxl)

library(tidyr)

##########
 # Dataset #
 #########
 stop <- read.csv("D:/Coding/R Storage/M748/Project/Data/MplsStops.csv", stringsAsFactors =
TRUE)
 summary(stop)

summary(is.na(stop))

dim(stop)

# NA omission
 mlp <- stop[!(is.na(stop$citationIssued) | stop$citationIssued == ""), ]
 dim(mlp)

# write-in NA
 mlp[mlp == ''] <- "Unknown"
 # vars
 num <- mlp$rownames
 lat <- mlp$lat
 long <- mlp$long
 mdc <- mlp$MDC
 # column removal
 police <- select(mlp, -c(1,5,12,13))
 summary(police)

# yes/no --> 1/0
 police$citationIssued <- ifelse(police$citationIssued == "YES",1,0)
 police$personSearch <- ifelse(police$personSearch == "YES",1,0)
 police$vehicleSearch <- ifelse(police$vehicleSearch == "YES",1,0)
```

```
 # counts
 table(as.Date(police$date))

problem <- table(police$problem)
 problem <- as.data.frame(problem)

 cite <- table(police$citationIssued)
 cite <- as.data.frame(cite)
 person <- table(police$personSearch)
 person <- as.data.frame(person)
 vehicle <- table(police$vehicleSearch)
 vehicle <- as.data.frame(vehicle)
 preRace <- table(police$preRace)
 preRace <- as.data.frame(preRace)
 race <- table(police$race)
 race <- as.data.frame(race)
 gender <- table(police$gender)
 gender <- as.data.frame(gender)
 precinct <- table(police$policePrecinct)
 precinct <- as.data.frame(precinct)
 table(police$neighborhood)

# visuals
 ggplot(problem, aes(x="", y = Freq, fill = Var1)) +
   geom_bar(stat = "identity", width=1) +
   coord_polar("y", start=0)

ggplot(cite, aes(x="", y = Freq, fill = Var1)) +
   geom_bar(stat = "identity", width=1) +
   coord_polar("y", start=0)

ggplot(person, aes(x="", y = Freq, fill = Var1)) +
   geom_bar(stat = "identity", width=1) +
   coord_polar("y", start=0)

ggplot(vehicle, aes(x="", y = Freq, fill = Var1)) +
   geom_bar(stat = "identity", width=1) +
   coord_polar("y", start=0)

ggplot(preRace, aes(x= Var1, y = Freq)) +
   geom_bar(stat = "identity", width = 0.5) +
   coord_flip()

ggplot(race, aes(x=Var1, y = Freq)) +
   geom_bar(stat = "identity", width=0.5) +
   coord_flip()
```

```
ggplot(gender, aes(x = "", y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", width=1) +
  coord_polar("y", start=0)

ggplot(precinct, aes(x="", y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", width=1) +
  coord_polar("y", start=0)
```