Math 748: Professor Tao He

Course Project: Progress Report I

**Who Did We Stop This Time?**

Gabrielle Salamanca

October 4, 2024

**Contents**

1. **Introduction**
2. **The Data**
   1. **Description of the Data**
   2. **Explanation of Variables**
   3. **Data Exploration & Cleaning**
   4. **Summary Results**
   5. **Data Visualization**
3. **Code**

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.

1. **The Data**
   1. **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.

1. **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 |

1. **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.

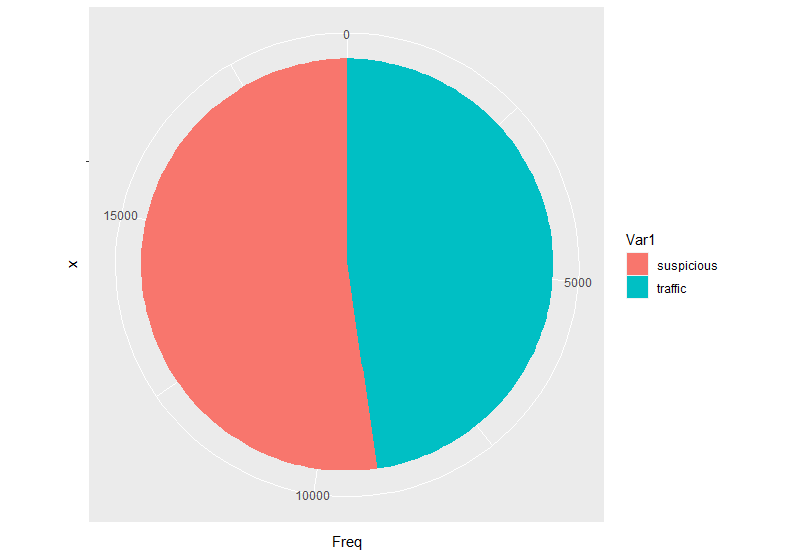
1. **Summary Results**

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.

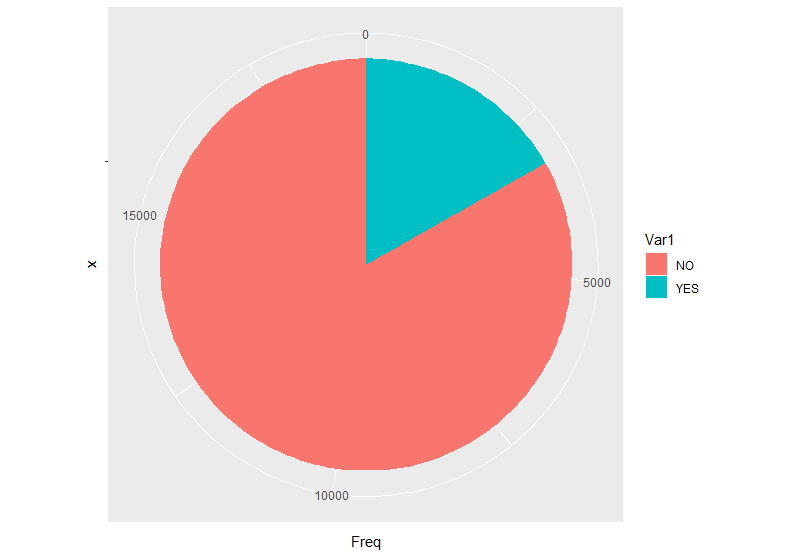
I will admit that I need to figure out how to extract the exact month from the date, because I believe it would be a better way to group results. As of right now, I have not, so I mainly have a month and day spread. There are dates that had triple digits, so it did bring a possibility of trying to group 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.

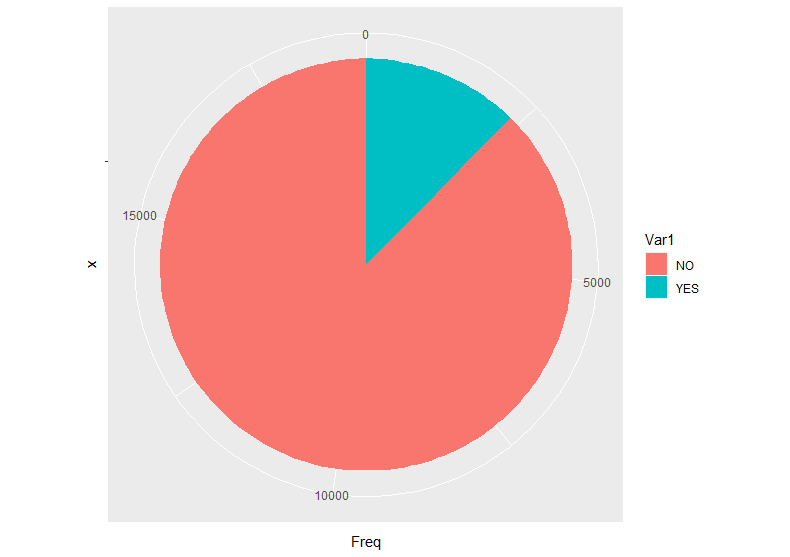
1. **Data Visualization**

****

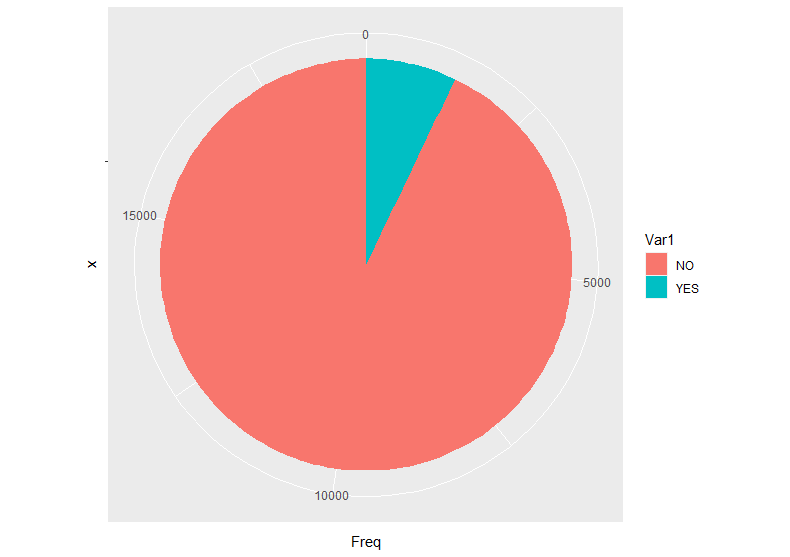
**Figure 1: Type of Problem**

****

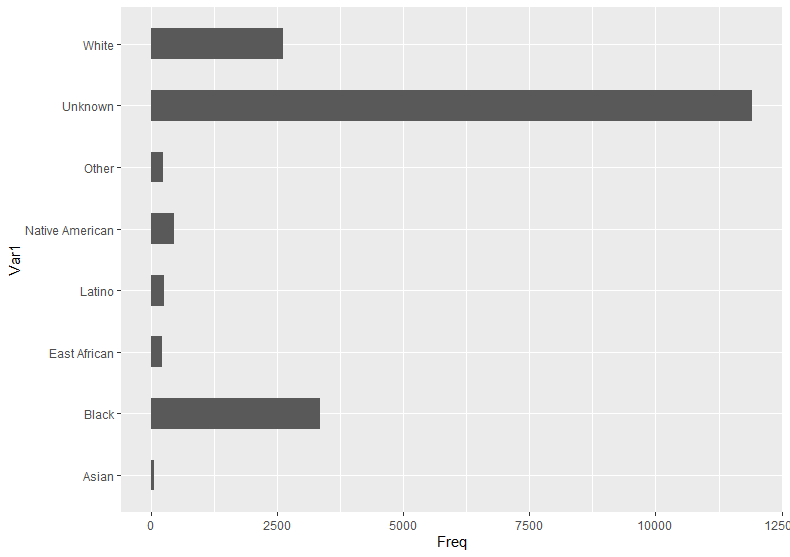
**Figure 2: Was a Citation Issued?**

****

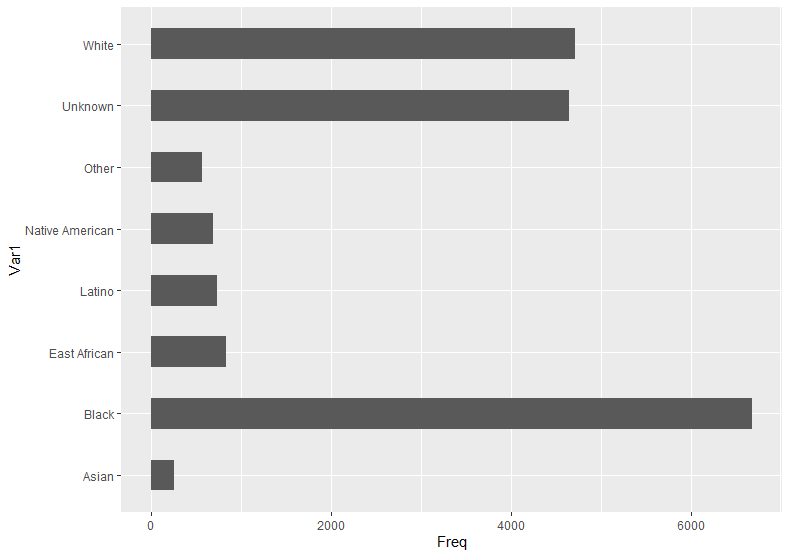
**Figure 3: Was the Person Searched?**

****

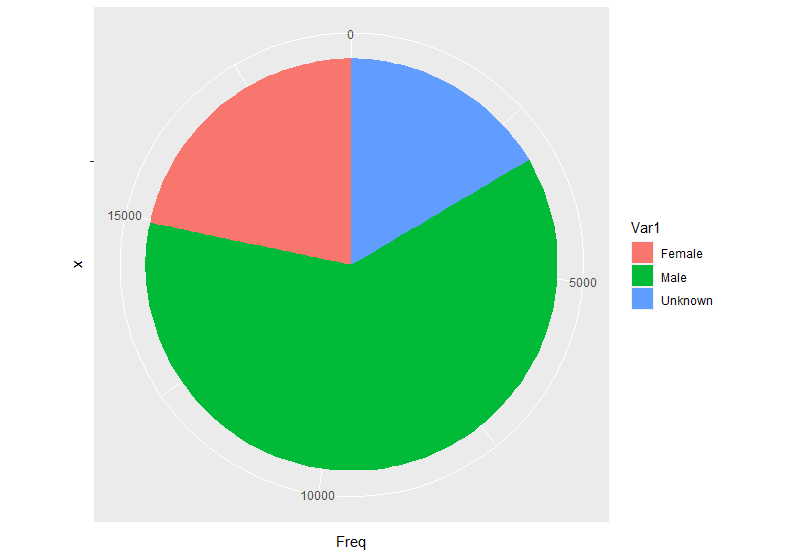
**Figure 4: Was the Vehicle Searched?**

****

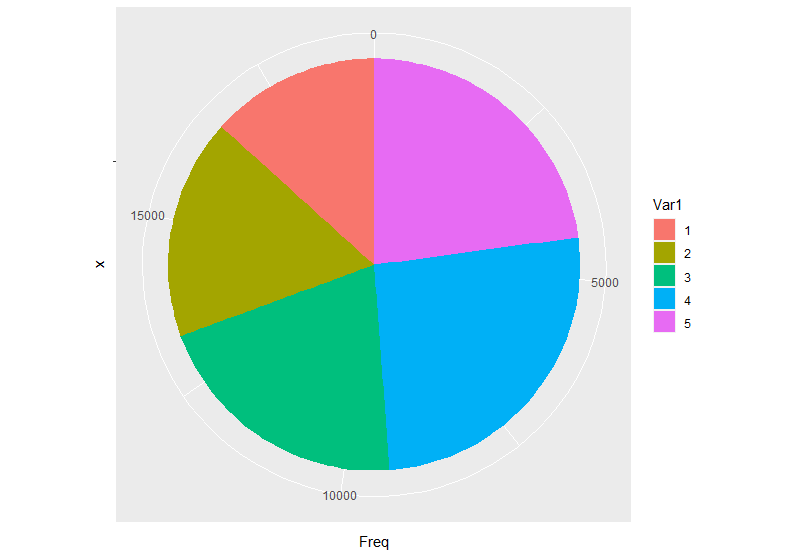
**Figure 5: Assumed Race of the Person Before Speaking with Them**

****

**Figure 6: Determined Race after the Incident**

****

**Figure 7: Gender**

****

**Figure 8: Police Precinct**

1. **Code**

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

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(psych)

## Warning: package 'psych' was built under R version 4.2.3

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(readr)

## Warning: package 'readr' was built under R version 4.2.3

library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

library(tidyr)

## Warning: package 'tidyr' was built under R version 4.2.3

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

## rownames idNum date   
## Min. : 6823 17-000003: 1 2017-01-10T18:05:28Z: 2   
## 1st Qu.:20379 17-000007: 1 2017-01-22T23:33:37Z: 2   
## Median :33864 17-000073: 1 2017-01-28T04:40:21Z: 2   
## Mean :33861 17-000092: 1 2017-01-28T05:21:45Z: 2   
## 3rd Qu.:47387 17-000098: 1 2017-01-30T20:26:45Z: 2   
## Max. :60838 17-000111: 1 2017-02-03T19:35:57Z: 2   
## (Other) :51914 (Other) :51908   
## problem MDC citationIssued personSearch vehicleSearch  
## suspicious:25822 MDC :43699 :32810 : 8221 : 8221   
## traffic :26098 other: 8221 NO :15899 NO :38462 NO :40579   
## YES: 3211 YES: 5237 YES: 3120   
##   
##   
##   
##   
## preRace race gender lat   
## Unknown :28337 Black :15220 : 8282 Min. :44.89   
## : 8221 White :11703 Female :10015 1st Qu.:44.95   
## Black : 6805 Unknown : 9219 Male :27131 Median :44.98   
## White : 6004 : 8221 Unknown: 6492 Mean :44.97   
## Native American: 908 East African: 2188 3rd Qu.:45.00   
## Latino : 528 Latino : 1858 Max. :45.05   
## (Other) : 1117 (Other) : 3511   
## long policePrecinct neighborhood   
## Min. :-93.33 Min. :1.000 Downtown West: 4409   
## 1st Qu.:-93.29 1st Qu.:2.000 Whittier : 3328   
## Median :-93.28 Median :3.000 Near - North : 2256   
## Mean :-93.27 Mean :3.257 Lyndale : 2154   
## 3rd Qu.:-93.25 3rd Qu.:4.000 Jordan : 2075   
## Max. :-93.20 Max. :5.000 Hawthorne : 2031   
## (Other) :35667

summary(is.na(stop))

## rownames idNum date problem   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:51920 FALSE:51920 FALSE:51920 FALSE:51920   
## MDC citationIssued personSearch vehicleSearch   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:51920 FALSE:51920 FALSE:51920 FALSE:51920   
## preRace race gender lat   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:51920 FALSE:51920 FALSE:51920 FALSE:51920   
## long policePrecinct neighborhood   
## Mode :logical Mode :logical Mode :logical   
## FALSE:51920 FALSE:51920 FALSE:51920

dim(stop)

## [1] 51920 15

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

## [1] 19110 15

# 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)

## idNum date problem   
## 17-251489: 1 2017-07-31T18:20:03Z: 2 suspicious:10004   
## 17-251517: 1 2017-08-27T04:32:00Z: 2 traffic : 9106   
## 17-251588: 1 2017-09-17T03:32:09Z: 2   
## 17-251645: 1 2017-09-21T01:40:37Z: 2   
## 17-251668: 1 2017-09-22T04:21:51Z: 2   
## 17-251673: 1 2017-09-22T16:42:15Z: 2   
## (Other) :19104 (Other) :19098   
## citationIssued personSearch vehicleSearch preRace   
## : 0 : 0 : 0 Unknown :11916   
## NO :15899 NO :16747 NO :17751 Black : 3349   
## YES: 3211 YES: 2363 YES: 1359 White : 2611   
## Native American: 461   
## Latino : 257   
## Other : 228   
## (Other) : 288   
## race gender policePrecinct neighborhood   
## Black :6685 : 0 Min. :1.000 Downtown West: 1365   
## White :4711 Female : 4149 1st Qu.:2.000 Whittier : 1032   
## Unknown :4645 Male :11810 Median :3.000 Jordan : 878   
## East African : 829 Unknown: 3151 Mean :3.277 Near - North : 844   
## Latino : 731 3rd Qu.:4.000 Hawthorne : 800   
## Native American: 690 Max. :5.000 Lyndale : 731   
## (Other) : 819 (Other) :13460

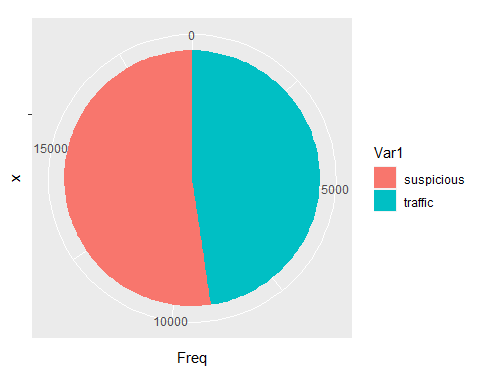
# yes/no --> 1/0  
police$citationIssued <- ifelse(police$citationIssued == "YES",1,0)  
police$personSearch <- ifelse(police$personSearch == "YES",1,0)  
police$vehicleSearch <- ifelse(police$vehicleSearch == "YES",1,0)  
  
# counts  
table(as.Date(police$date))

##   
## 2017-07-06 2017-07-07 2017-07-08 2017-07-09 2017-07-10 2017-07-11 2017-07-12   
## 26 113 77 70 123 121 149   
## 2017-07-13 2017-07-14 2017-07-15 2017-07-16 2017-07-17 2017-07-18 2017-07-19   
## 127 135 115 86 131 108 166   
## 2017-07-20 2017-07-21 2017-07-22 2017-07-23 2017-07-24 2017-07-25 2017-07-26   
## 139 167 114 103 105 135 89   
## 2017-07-27 2017-07-28 2017-07-29 2017-07-30 2017-07-31 2017-08-01 2017-08-02   
## 130 115 102 79 116 93 110   
## 2017-08-03 2017-08-04 2017-08-05 2017-08-06 2017-08-07 2017-08-08 2017-08-09   
## 102 103 95 77 95 119 98   
## 2017-08-10 2017-08-11 2017-08-12 2017-08-13 2017-08-14 2017-08-15 2017-08-16   
## 96 109 100 79 110 147 101   
## 2017-08-17 2017-08-18 2017-08-19 2017-08-20 2017-08-21 2017-08-22 2017-08-23   
## 97 103 105 99 122 130 118   
## 2017-08-24 2017-08-25 2017-08-26 2017-08-27 2017-08-28 2017-08-29 2017-08-30   
## 144 162 89 89 116 114 117   
## 2017-08-31 2017-09-01 2017-09-02 2017-09-03 2017-09-04 2017-09-05 2017-09-06   
## 103 87 125 130 76 117 132   
## 2017-09-07 2017-09-08 2017-09-09 2017-09-10 2017-09-11 2017-09-12 2017-09-13   
## 153 120 95 98 97 110 134   
## 2017-09-14 2017-09-15 2017-09-16 2017-09-17 2017-09-18 2017-09-19 2017-09-20   
## 130 120 120 110 91 105 132   
## 2017-09-21 2017-09-22 2017-09-23 2017-09-24 2017-09-25 2017-09-26 2017-09-27   
## 130 144 135 94 84 102 145   
## 2017-09-28 2017-09-29 2017-09-30 2017-10-01 2017-10-02 2017-10-03 2017-10-04   
## 141 130 100 92 93 127 189   
## 2017-10-05 2017-10-06 2017-10-07 2017-10-08 2017-10-09 2017-10-10 2017-10-11   
## 151 101 88 75 108 123 168   
## 2017-10-12 2017-10-13 2017-10-14 2017-10-15 2017-10-16 2017-10-17 2017-10-18   
## 100 155 104 81 91 105 128   
## 2017-10-19 2017-10-20 2017-10-21 2017-10-22 2017-10-23 2017-10-24 2017-10-25   
## 121 150 110 67 98 112 107   
## 2017-10-26 2017-10-27 2017-10-28 2017-10-29 2017-10-30 2017-10-31 2017-11-01   
## 118 107 117 101 119 78 94   
## 2017-11-02 2017-11-03 2017-11-04 2017-11-05 2017-11-06 2017-11-07 2017-11-08   
## 111 99 90 93 68 94 82   
## 2017-11-09 2017-11-10 2017-11-11 2017-11-12 2017-11-13 2017-11-14 2017-11-15   
## 115 119 99 100 76 82 87   
## 2017-11-16 2017-11-17 2017-11-18 2017-11-19 2017-11-20 2017-11-21 2017-11-22   
## 106 110 99 66 100 110 107   
## 2017-11-23 2017-11-24 2017-11-25 2017-11-26 2017-11-27 2017-11-28 2017-11-29   
## 137 79 84 131 91 145 155   
## 2017-11-30 2017-12-01 2017-12-02 2017-12-03 2017-12-04 2017-12-05 2017-12-06   
## 118 96 121 96 98 42 69   
## 2017-12-07 2017-12-08 2017-12-09 2017-12-10 2017-12-11 2017-12-12 2017-12-13   
## 91 71 81 105 76 62 88   
## 2017-12-14 2017-12-15 2017-12-16 2017-12-17 2017-12-18 2017-12-19 2017-12-20   
## 114 111 131 131 90 82 100   
## 2017-12-21 2017-12-22 2017-12-23 2017-12-24 2017-12-25 2017-12-26 2017-12-27   
## 95 100 100 69 40 83 90   
## 2017-12-28 2017-12-29 2017-12-30 2017-12-31 2018-01-01   
## 110 106 64 68 29

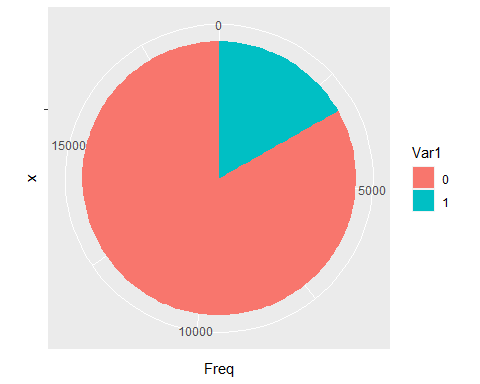
problem <- table(police$problem)  
problem <- as.data.frame(problem)  
  
cite <- table(police$citationIssued)  
cite <- as.data.frame(cite)  
  
person <- table(police$personSearch)  
person <- as.data.frame(person)  
  
vehicle <- table(police$vehicleSearch)  
vehicle <- as.data.frame(vehicle)  
  
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)

##   
## Armatage Audubon Park   
## 45 181   
## Bancroft Beltrami   
## 61 84   
## Bottineau Bryant   
## 118 42   
## Bryn - Mawr Camden Industrial   
## 57 19   
## CARAG Cedar - Isles - Dean   
## 216 51   
## Cedar Riverside Central   
## 277 307   
## Cleveland Columbia Park   
## 169 59   
## Como Cooper   
## 137 49   
## Corcoran Diamond Lake   
## 148 75   
## Downtown East Downtown West   
## 103 1365   
## East Harriet East Isles   
## 66 190   
## East Phillips ECCO   
## 508 99   
## Elliot Park Ericsson   
## 231 80   
## Field Folwell   
## 45 505   
## Fulton Hale   
## 58 24   
## Harrison Hawthorne   
## 149 800   
## Hiawatha Holland   
## 125 364   
## Howe Humboldt Industrial Area   
## 88 6   
## Jordan Keewaydin   
## 878 52   
## Kenny Kenwood   
## 26 48   
## King Field Lind - Bohanon   
## 331 137   
## Linden Hills Logan Park   
## 119 129   
## Longfellow Loring Park   
## 229 265   
## Lowry Hill Lowry Hill East   
## 73 541   
## Lyndale Lynnhurst   
## 731 51   
## Marcy Holmes Marshall Terrace   
## 543 119   
## McKinley Mid - City Industrial   
## 276 107   
## Midtown Phillips Minnehaha   
## 439 51   
## Morris Park Near - North   
## 24 844   
## Nicollet Island - East Bank North Loop   
## 560 303   
## Northeast Park Northrop   
## 127 91   
## Page Phillips West   
## 17 283   
## Powderhorn Park Prospect Park - East River Road   
## 433 161   
## Regina Seward   
## 58 179   
## Sheridan Shingle Creek   
## 99 53   
## St. Anthony East St. Anthony West   
## 82 166   
## Standish Steven's Square - Loring Heights   
## 94 330   
## Sumner - Glenwood Tangletown   
## 50 218   
## University of Minnesota Ventura Village   
## 30 396   
## Victory Waite Park   
## 226 102   
## Webber - Camden Wenonah   
## 288 36   
## West Calhoun Whittier   
## 24 1032   
## Willard - Hay Windom   
## 491 115   
## Windom Park   
## 152

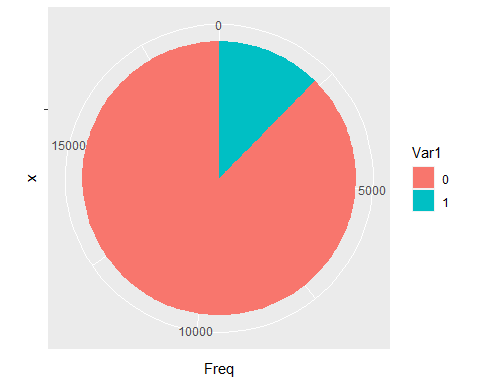
# visuals  
  
ggplot(problem, aes(x="", y = Freq, fill = Var1)) +  
 geom\_bar(stat = "identity", width=1) +  
 coord\_polar("y", start=0)



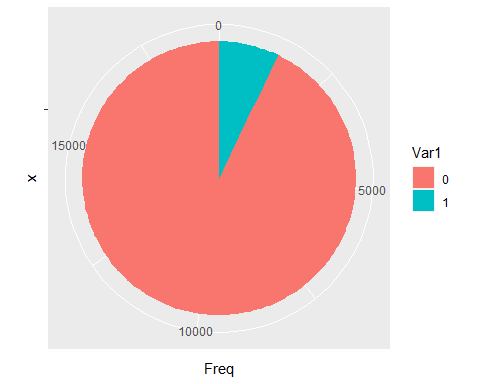
ggplot(cite, aes(x="", y = Freq, fill = Var1)) +  
 geom\_bar(stat = "identity", width=1) +  
 coord\_polar("y", start=0)



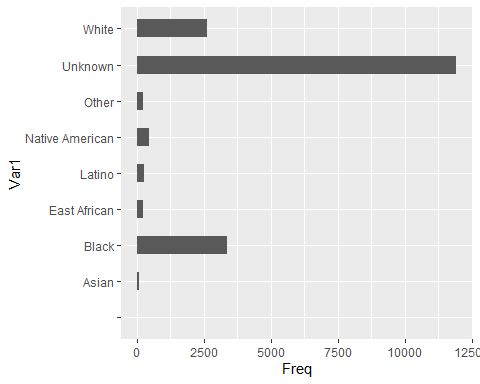
ggplot(person, aes(x="", y = Freq, fill = Var1)) +  
 geom\_bar(stat = "identity", width=1) +  
 coord\_polar("y", start=0)



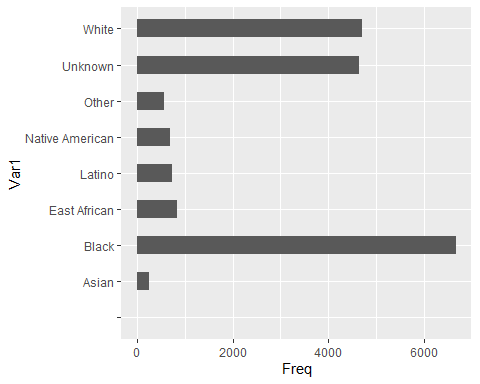
ggplot(vehicle, aes(x="", y = Freq, fill = Var1)) +  
 geom\_bar(stat = "identity", width=1) +  
 coord\_polar("y", start=0)



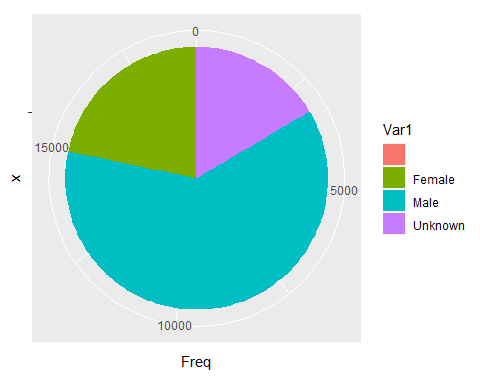
ggplot(preRace, aes(x= Var1, y = Freq)) +  
 geom\_bar(stat = "identity", width = 0.5) +  
 coord\_flip()



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)

