STA234: Homework 2

Derin Gezgin

2025-02-18

Project Problem

Good work, Derin!

Do the following for the approved dataset(s):

Part (A) Read the data here in R.

```
traffic.data = read.csv("../../DATA/ca_san_francisco_2020_04_01.csv")
```

Part (B) Show the structure of data.

```
str(traffic.data)
```

```
## 'data.frame':    905070 obs. of  22 variables:
##  $ raw_row_number                  : chr  "869921" "869922" "869923"
"869924" ...
##  $ date                            : chr  "2014-08-01" "2014-08-01"
"2014-08-01" "2014-08-01" ...
##  $ time                            : chr  "00:01:00" "00:01:00"
"00:15:00" "00:18:00" ...
##  $ location                        : chr  "MASONIC AV & FELL ST"
"GEARY&10TH AV" "SUTTER N OCTAVIA ST" "3RD ST & DAVIDSON" ...
##  $ lat                             : num  37.8 37.8 37.8 37.7 37.8 ...
##  $ lng                             : num  -122 -122 -122 -122 -122 ...
##  $ district                        : chr  NA NA NA NA ...
##  $ subject_age                     : int  NA NA NA NA NA NA NA NA NA NA
...
##  $ subject_race                    : chr  "asian/pacific islander"
"black" "hispanic" "hispanic" ...
##  $ subject_sex                     : chr  "female" "male" "male" "male"
...
##  $ type                            : chr  "vehicular" "vehicular"
"vehicular" "vehicular" ...
##  $ arrest_made                     : logi  FALSE FALSE FALSE FALSE FALSE
FALSE ...
##  $ citation_issued                 : logi  FALSE TRUE TRUE FALSE TRUE
TRUE ...
##  $ warning_issued                  : logi  TRUE FALSE FALSE TRUE FALSE
FALSE ...
##  $ outcome                         : chr  "warning" "citation" "citation"
"warning" ...
##  $ contraband_found                : logi  NA NA NA NA NA NA ...
```

```
##  $ search_conducted                  : logi  FALSE FALSE FALSE FALSE FALSE
FALSE ...
##  $ search_vehicle                    : logi  FALSE FALSE FALSE FALSE FALSE
FALSE ...
##  $ search_basis                      : chr  NA NA NA NA ...
##  $ reason_for_stop                   : chr  "Mechanical or Non-Moving
Violation (V.C.)" "Mechanical or Non-Moving Violation (V.C.)" "Mechanical or
Non-Moving Violation (V.C.)" "Mechanical or Non-Moving Violation (V.C.)" ...
##  $ raw_search_vehicle_description    : chr  "No Search" "No Search" "No
Search" "No Search" ...
##  $ raw_result_of_contact_description: chr  "Warning" "Citation" "Citation"
"Warning" ...
```

This output shows us the variable names and the way they are recognized by R. This shows me that in the future parts of my project, I might have to convert some of the columns to appropriate data types as date/time are recognized as strings (they should be date and time objects, some of the categorical variables (such as subject sex/age/race) are not recognized a factors, etc.

Part (C) What is the dimension of your data?

```
dim(traffic.data)

## [1] 905070     22
```

From this output, we can see that the dataset has 22 columns and 905070 rows.

Part (D) Show names of variables in the data.

```
colnames(traffic.data)

##  [1] "raw_row_number"                    "date"
##  [3] "time"                              "location"
##  [5] "lat"                               "lng"
##  [7] "district"                          "subject_age"
##  [9] "subject_race"                      "subject_sex"
## [11] "type"                              "arrest_made"
## [13] "citation_issued"                   "warning_issued"
## [15] "outcome"                           "contraband_found"
## [17] "search_conducted"                  "search_vehicle"
## [19] "search_basis"                      "reason_for_stop"
## [21] "raw_search_vehicle_description"
"raw_result_of_contact_description"
```

Similar to the structure, this part gives us the variable names of the dataset.

Part (E) Find easy answers to your research question (one of them) using the data

One of my research questions was: ***How do traffic stop rates vary by driver demographic characteristics?***

Before tackling the question and plotting the results, there are two data preparation steps we have to complete:

```
traffic.data$subject_sex = factor(traffic.data$subject_sex,
                                  levels = c("male", "female"),
                                  labels = c("Male", "Female"))

traffic.data$date = as.Date(traffic.data$date)
summary(traffic.data$date)

##         Min.      1st Qu.       Median         Mean      3rd Qu.
Max.
## "2007-01-01" "2009-02-02" "2011-03-16" "2011-06-13" "2013-11-16" "2016-06-
30"
```

I converted the subject_sex column to a R-factor. At the same time, I converted the date into a real date column in order to extract the year data from each entry.

Lastly, I checked the summary of the date column to see the starting and the end date of the data.

In this question -as an easy answer is asked- I am going to focus on the gender information of the traffic stops.

At first step, I am going to look into the gender information of stopped drivers for the whole time frame.

```
group.gender.total = aggregate(traffic.data$raw_row_number,
                               by = list(traffic.data$subject_sex),
                               FUN = length)
colnames(group.gender.total) = c("Gender", "Count")

group.gender.total

##   Gender  Count
## 1   Male 639219
## 2 Female 265851
```

From this output, we can see that the amount male drivers stopped is much higher than (approximately 2.4 times higher) than the female drivers.

At the second stage of this problem, I grouped the data by year and also gender. This allowed me to see the gender distribution of the traffic stops throughout the years.

```
group.gender.year.total = aggregate(x = traffic.data$raw_row_number,
                                    by = list(traffic.data$subject_sex,
                                              format(traffic.data$date,
"%Y")),
                                    FUN = length)

colnames(group.gender.year.total) = c("Gender", "Year", "Count")

group.gender.year.total
```

```
##    Gender Year Count
## 1    Male 2007 72339
## 2  Female 2007 29857
## 3    Male 2008 79638
## 4  Female 2008 33487
## 5    Male 2009 76942
## 6  Female 2009 33670
## 7    Male 2010 72194
## 8  Female 2010 32576
## 9    Male 2011 68768
## 10 Female 2011 30739
## 11   Male 2012 57758
## 12 Female 2012 24633
## 13   Male 2013 52997
## 14 Female 2013 21165
## 15   Male 2014 65666
## 16 Female 2014 26295
## 17   Male 2015 63182
## 18 Female 2015 22816
## 19   Male 2016 29735
## 20 Female 2016 10613
```

From this table, we can see that throughout the years the trend of “male drivers being stopped more then female drivers” continued. It is important to note that 2016 data is available only until June which explains the less than usual amount of data points. In the future parts of my data exploration and modeling, I might consider to remove 2016 data from the dataset when comparing years.

From both of these simple data explorations, we can see that male drivers have been stopped significantly more compared to the female drivers in San Francisco between January 1st 2007 and June 30th 2016 in both total number of stops through the time period and also the total number of stops annually.

Another demographic information I would like to explore is the relationship between the age and the number of stops. As the question asked for easy answers I limited myself only to the gender part of this dataset. In the further assignments I can include subject age, and race to the demographic data exploration part of the project.