Analysis of Traffic Stops and Violations

By: Connor Knupp, Ethan Zuwiala, Joe Wenger

Introduction

Our Project deals with police traffic stops that happen every day on roadways across America. It is an essential part of law enforcement to stop drivers from speeding, carrying illegal substances, or to return stolen vehicles to their rightful owner. There are many different violations of traffic law that can initiate a traffic stop and the data regarding those stops is what we chose to look at. We thought that it would be interesting to analyze data from police traffic stops because there is a lot of information police are required to keep track of when they pull someone over such as the Date, Time of Day, Traffic Violation, Vehicle Make, Model and Year, Use of Alcohol, Race, Sex, Seatbelt Use, Accidents, Injuries, City, State, Etc. We sourced our data specifically from Montgomery County in Maryland. The data does not have information that can be used to uniquely identify the vehicle, the vehicle owner or the officer, but has plenty of other stuff for our use. The dataset we use is an excel file containing 10,000 rows, each being its own traffic stop. With all this data, we had to determine what questions we could ask that would be interesting and informative. We came up with the list below.

The data we used for our report can be found on: https://catalog.data.gov/dataset/traffic-violations

With the data analysis and graphs in this report we seek to answer the following questions:

1. What vehicle makes get pulled over the most often?
2. What times during the week are traffic stops more likely to happen?
3. What times during the year are traffic stops more likely to happen?
4. What times during the past decade did traffic stops happen the most?
5. Does using your seatbelt actually help prevent injury when in an accident?
6. How do race and gender apply to traffic violations?
7. Who is most likely to commit a traffic violation in Maryland? (Marylanders or Non-Marylanders?)

Here is our implementation of our R code:

Load Packages

```
#install.packages('rmarkdown')
#install.packages('readxl') -- Used for importing xlsx spreadsheets.
#install.packages('tidyverse')
#install.packages('ggrounded') -- Used for fancier bars.
library(rmarkdown)
library(readxl)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(ggrounded)
```

Load the Data

```
traff_violations <- read_excel('AttemptTwo.xlsx')
traff_violations
```

```
# A tibble: 10,000 x 25
   `Date Of Stop`      `Time Of Stop`      Description            Accident Belts
   <dttm>              <dttm>              <chr>                  <chr>    <chr>
 1 2015-10-20 00:00:00 1899-12-31 15:02:00 EXCEEDING MAXIMUM SPE~ No       No
 2 2013-12-02 00:00:00 1899-12-31 16:23:00 FAILURE TO DISPLAY RE~ No       No
 3 2013-08-20 00:00:00 1899-12-31 22:48:00 EXCEEDING THE POSTED ~ No       No
 4 2017-08-27 00:00:00 1899-12-31 16:39:00 DRIVER FAILURE TO OBE~ No       No
 5 2012-03-25 00:00:00 1899-12-31 13:16:00 DRIVING VEHICLE ON HI~ No       No
 6 2014-04-10 00:00:00 1899-12-31 03:44:00 DRIVING WHILE IMPAIRE~ No       No
```

```
 7 2023-11-17 00:00:00 1899-12-31 20:04:00 FAILURE TO ATTACH VEH~ No       No
 8 2018-10-15 00:00:00 1899-12-31 23:47:00 EXCEEDING POSTED MAXI~ No       No
 9 2013-04-17 00:00:00 1899-12-31 17:44:00 DRIVER FAILURE TO OBE~ No       No
10 2019-07-01 00:00:00 1899-12-31 09:08:00 DRIVER USING HANDS TO~ No       No
# i 9,990 more rows
# i 20 more variables: `Personal Injury` <chr>, `Property Damage` <chr>,
#   Fatal <chr>, `Commercial License` <chr>, HAZMAT <chr>,
#   `Commercial Vehicle` <chr>, Alcohol <chr>, `Work Zone` <chr>,
#   `Search Conducted` <chr>, VehicleType <chr>, Year <dbl>, Make <chr>,
#   Model <chr>, Color <chr>, `Contributed To Accident` <lgl>, Race <chr>,
#   Gender <chr>, `Driver City` <chr>, `Driver State` <chr>, ...
```

Tidy the Data

```r
# Modifications for proper time and day.
traff_violations <- traff_violations %>%
  mutate(
    TimeOnly = format(`Time Of Stop`, "%H:%M:%S"),
    FullDateTime = ymd_hms(paste(`Date Of Stop`, TimeOnly)),
    Hour = hour(FullDateTime),
    Day = wday(FullDateTime, label = TRUE, abbr = FALSE)
  )

#Description of Violation
most_common_desc <- traff_violations %>%
  count(Description) %>%
  arrange(desc(n)) %>%
  slice(1:5)

#Vehicle Year
most_common_vyear <- traff_violations %>%
  count(Year) %>%
  arrange(desc(n)) %>%
  slice(1:10)

#Vehicle Make
most_common_vmake <- traff_violations %>%
  mutate(Make = str_replace_all(Make, c(
    "CHEVY" = "CHEVROLET",
    "CHEV" = "CHEVROLET",
```

```
    "TOYT" = "TOYOTA",
    "HOND" = "HONDA",
    "HONDAA" = "HONDA",
    "CHEVROLETROLET" = "CHEVROLET",
    "VOLK" = "VOLKSWAGON",
    "VW" = "VOLKSWAGON",
    "HYUN" = "HYUNDAI",
    "TOTOTA" = "TOYOTA",
    "TOTYOA" = "TOYOTA",
    "TOYORA" = "TOYOTA",
    "HYUNDAIDAI" = "HYUNDAI",
    "MERZ" = "MERCEDES",
    "NISS" = "NISSAN",
    "DODG" = "DODGE",
    "MAZD" = "MAZDA",
    "MAZDAA" = "MAZDA",
    "DODGEE" = "DODGE",
    "ACUR" = "ACURA",
    "ACURAA" = "ACURA",
    "NISSANAN" = "NISSAN",
    "SUBA" = "SUBARU",
    "SUBARURU" = "SUBARU",
    "VOLKSWAGONSWAGEN" = "VOLKSWAGON",
    "VOLKSWAGONSWAGON" = "VOLKSWAGON",
    "INFI" = "INFINITI",
    "INFINITINITI" = "INFINITI",
    "MITS" = "MITSUBISHI",
    "MITSUBISHIUBISHI" = "MITSUBISHI",
    "VOLKSWAGONS" = "VOLKSWAGON",
    "TOYOT" = "TOYOTA",
    "TOYOTAAA" = "TOYOTA",
    "TOYO" = "TOYOTA",
    "TOYOTATA" = "TOYOTA",
    "TOYOTAA" = "TOYOTA",
    "HYUNDAID" = "HYUNDAI",
    "MAZADA" = "MAZDA",
    "CHRY" = "CRYSLER",
    "CHEVROLETORLET" = "CHEVROLET",
    "TOYOTAOA" = "TOYOTA",
    "VOLV" = "VOLVO",
    "VOLVOO" = "VOLVO",
    "MERCEDES" = "MERCEDES BENZ",
```

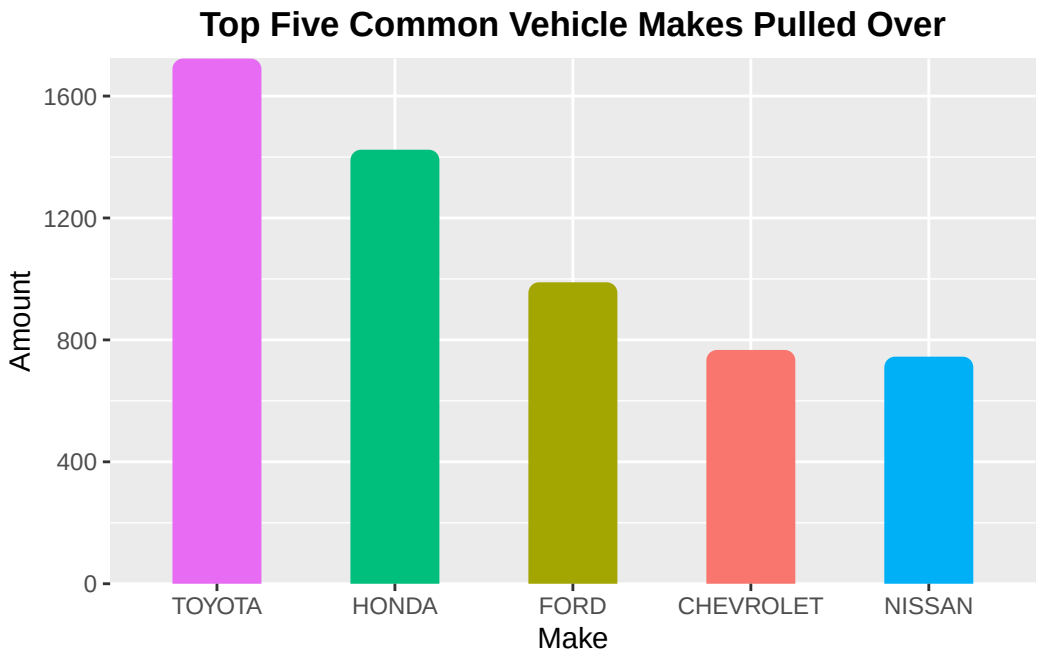
```
    "MERC" = "MERCEDES BENZ",
    "MERCEDES BENZURY" = "MERCEDES BENZ",
    "MERCEDES BENZEDES BENZ" = "MERCEDES BENZ",
    "MERCEDES BENZEDEZ" = "MERCEDES BENZ",
    "MERCEDES BENZ BENZ" = "MERCEDES BENZ",
    "MERCEDES BENZ-BENZ" = "MERCEDES BENZ",
    "CRYSLERSLER" = "CHRYSLER",
    "CRYSLER" = "CHRYSLER",
    "CHEVROLETE" = "CHEVROLET",
    "CHEVROLETY" = "CHEVROLET",
    "CHEVROLETROLET" = "CHEVROLET",
    "CHECY" = "CHEVROLET",
    "CVEVROLET" = "CHEVROLET",
    "HYUNDAII" = "HYUNDAI",
    "HYUNDAIIA" = "HYUNDAI",
    "HYUNDAIA" = "HYUNDAI",
    "TOY" = "TOYOTA",
    "TOYOTAOTA" = "TOYOTA",
    "HINDA" = "HONDA"
  ))) %>%
  count(Make) %>%
  arrange(desc(n)) %>%
  slice(1:5)


#Race of Driver
most_common_race <- traff_violations %>%
  count(Race) %>%
  arrange(desc(n))
  #%>% slice(1:5)
```

Analyze the Tidy Data

Plot 1 - Top Five Common Vehicle Makes Pulled Over

```
ggplot(data = most_common_vmake, aes(x= reorder(Make, -n), y= n, fill = Make)) +
 geom_col_rounded(width = 0.5) + scale_y_continuous(expand = c(0, 0))+
 labs(title = "Top Five Common Vehicle Makes Pulled Over") + xlab("Make") +
  ylab("Amount") + theme(legend.position = "none", plot.title =
                           element_text(hjust = 0.5, face = "bold"))
```

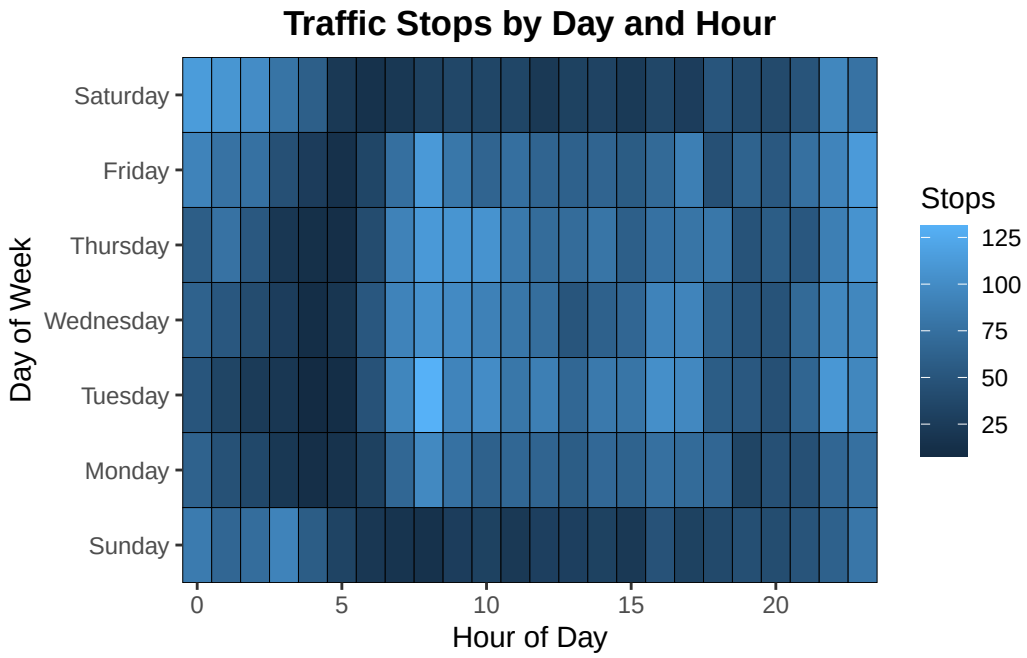



```
print(most_common_vmake)
```

```
# A tibble: 5 x 2
  Make           n
  <chr>      <int>
1 TOYOTA      1723
2 HONDA       1424
3 FORD         989
4 CHEVROLET    767
5 NISSAN       745
```

Plot 2 - Traffic Stops by Day and Hour

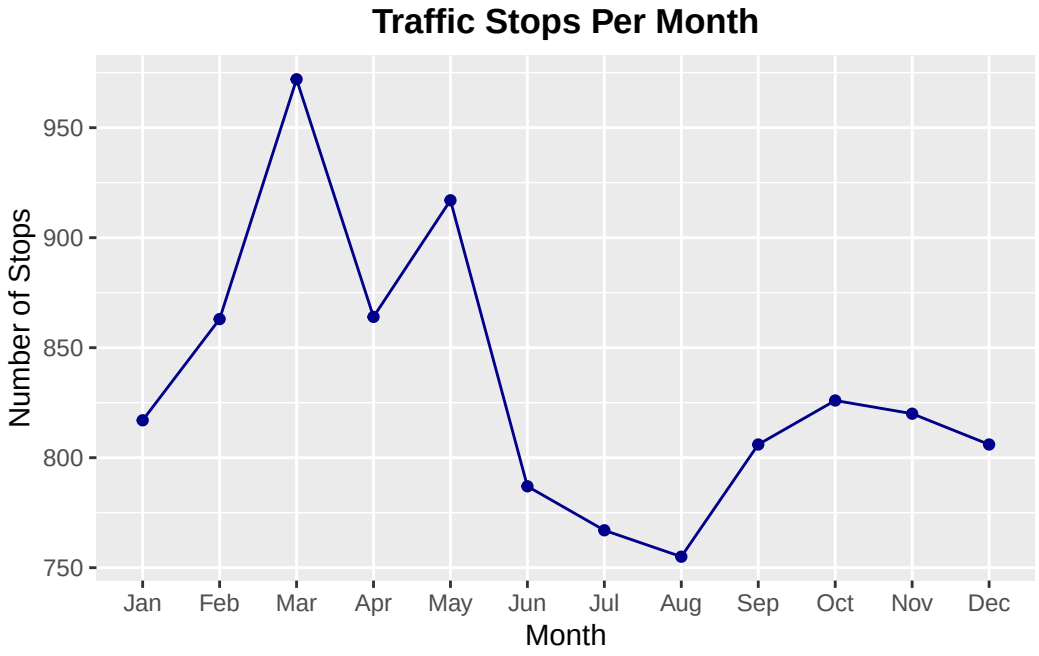
```
traff_violations %>%
  count(Hour, Day) %>%
  ggplot(aes(Hour, Day, fill = n)) + geom_tile(color = 'black') +
  scale_x_continuous(expand = c(0, 0)) + scale_y_discrete(expand = c(0, 0)) +
  labs(title = "Traffic Stops by Day and Hour", x = "Hour of Day",
       y = "Day of Week", fill = "Stops") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Plot 3 - Traffic Stops Per Month

```
monthly_stops <- traff_violations %>%
  mutate(Month = month(`Date Of Stop`, label = TRUE, abbr = TRUE)) %>%
  count(Month)

ggplot(monthly_stops, aes(Month, n))+geom_line(group = 1, color = 'darkblue')+
  geom_point(color ='darkblue') + labs(title = "Traffic Stops Per Month")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))+
  ylab("Number of Stops")
```



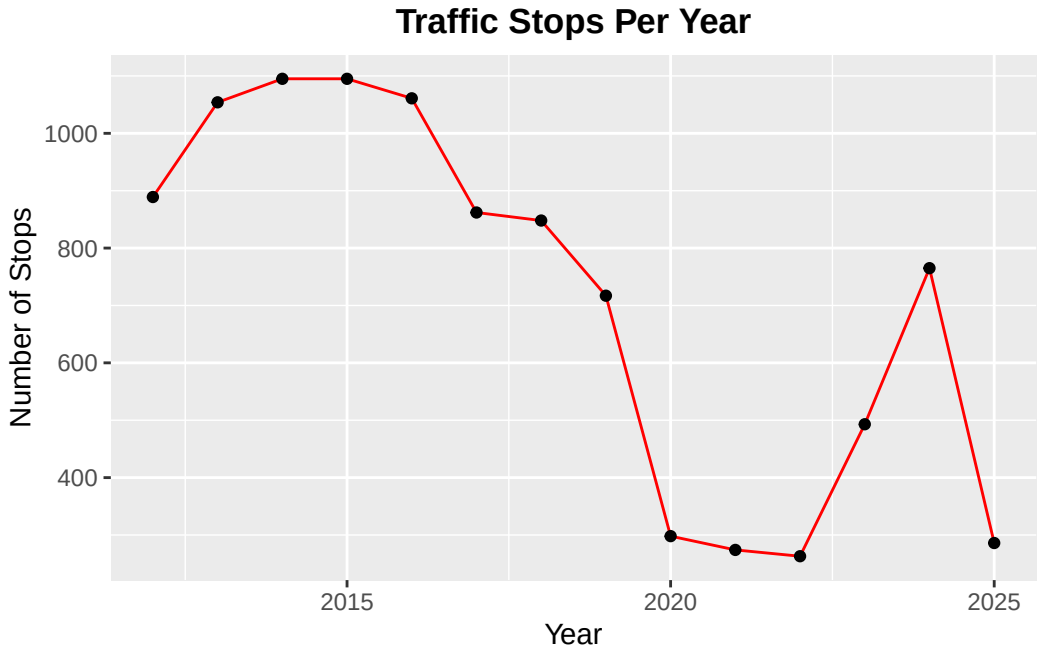

```
monthly_stops
```

```
# A tibble: 12 x 2
   Month     n
   <ord> <int>
 1 Jan     817
 2 Feb     863
 3 Mar     972
 4 Apr     864
 5 May     917
 6 Jun     787
 7 Jul     767
 8 Aug     755
 9 Sep     806
10 Oct     826
11 Nov     820
12 Dec     806
```

Plot 4 - Traffic Stops Per Year

```r
yearly_stops <- traff_violations %>%
  mutate(Year = year(`Date Of Stop`)) %>%
  count(Year)

ggplot(yearly_stops, aes(Year, n))+geom_line(group = 1, color = 'red')+
  geom_point(color ='black') + labs(title = "Traffic Stops Per Year")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))+ ylab("Number of Stops")
```



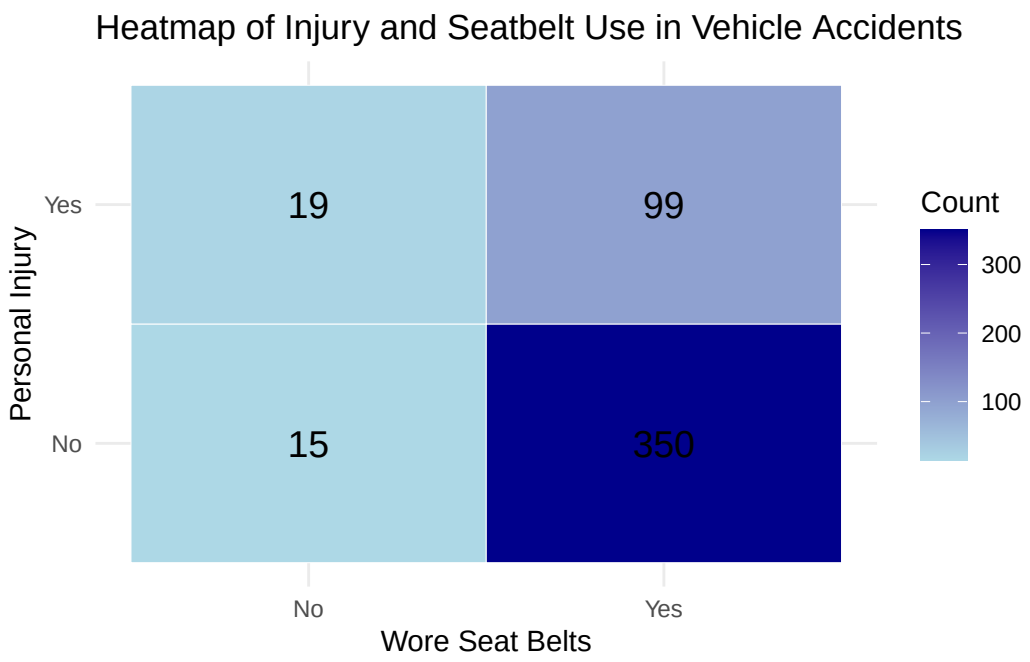

```r
yearly_stops
```

```
# A tibble: 14 x 2
    Year     n
   <dbl> <int>
 1  2012   889
 2  2013  1054
 3  2014  1095
 4  2015  1095
 5  2016  1061
 6  2017   862
 7  2018   848
 8  2019   717
 9  2020   298
```

```
10   2021    274
11   2022    263
12   2023    493
13   2024    765
14   2025    286
```

Plot 5 - Correlation of Injury and Seatbelt Use in Accidents

```
traff_violations_filtered <- traff_violations %>% filter(Accident == "Yes")
traff_violations_filtered <- traff_violations_filtered %>%
  mutate(Belts = ifelse(Belts == "Yes", "N", Belts))
traff_violations_filtered <- traff_violations_filtered %>%
  mutate(Belts = ifelse(Belts == "No", "Yes", Belts))
traff_violations_filtered <- traff_violations_filtered %>%
  mutate(Belts = ifelse(Belts == "N", "No", Belts))
traff_violations_filtered %>%
  count(`Belts`, `Personal Injury`) %>%
  ggplot(aes(x = `Belts`, y = `Personal Injury`, fill = n)) +
  geom_tile(color = "white") +
  geom_text(aes(label = n), color = "black", size = 5) +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Injury and Seatbelt Use in Vehicle Accidents",
       x = "Wore Seat Belts", y = "Personal Injury", fill = "Count") +
  theme_minimal()
```



Plot 6 - Traffic Violations by Race and Gender

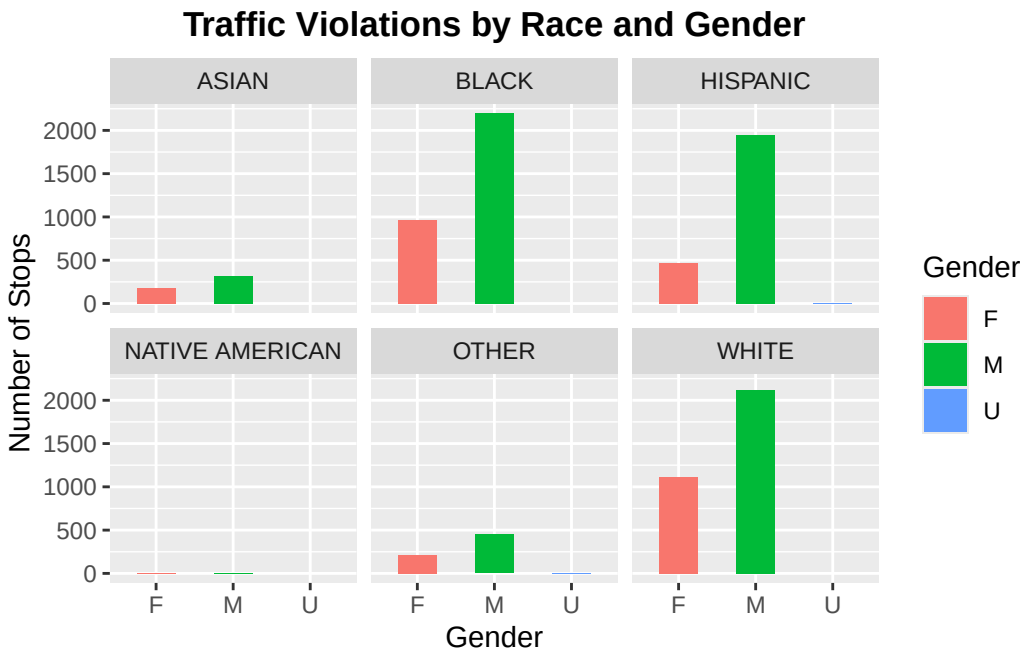
```
race_gender_counts <- traff_violations %>%
  count(Race, Gender)

race_gender_counts
```

```
# A tibble: 14 x 3
   Race            Gender     n
   <chr>           <chr>  <int>
 1 ASIAN           F        177
 2 ASIAN           M        316
 3 BLACK           F        962
 4 BLACK           M       2199
 5 HISPANIC        F        469
 6 HISPANIC        M       1950
 7 HISPANIC        U          1
 8 NATIVE AMERICAN F          5
 9 NATIVE AMERICAN M          7
10 OTHER           F        217
11 OTHER           M        450
```

```
12 OTHER           U          7
13 WHITE           F       1118
14 WHITE           M       2122
```

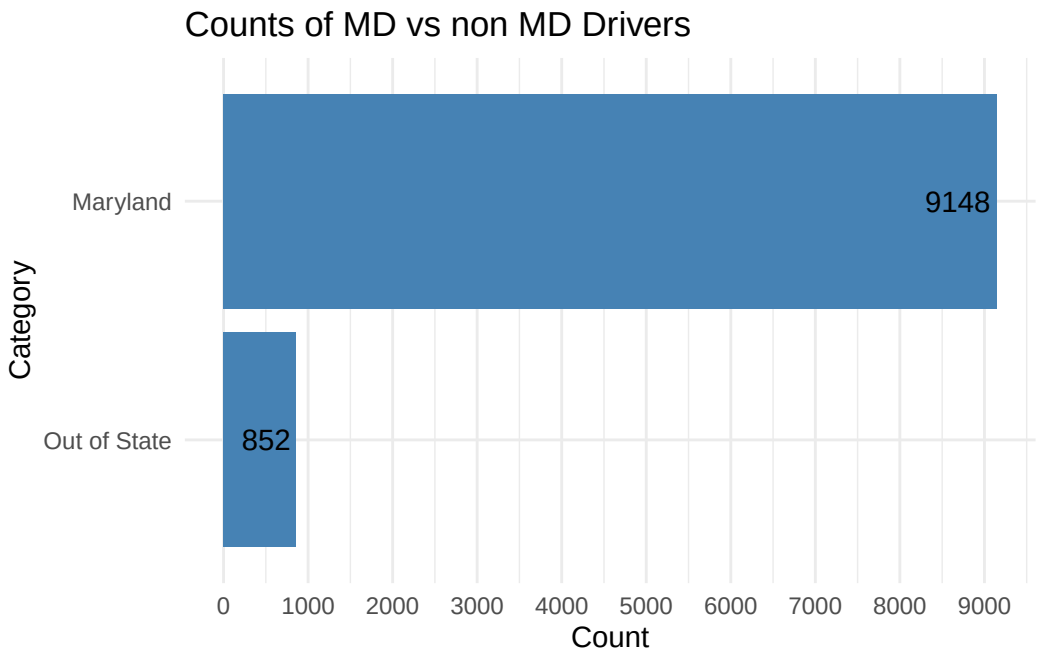
```
ggplot(race_gender_counts, aes(x = Gender, y = n, fill = Gender)) +
  geom_col(width = 0.5) +
  facet_wrap(~ Race)+ labs(title = "Traffic Violations by Race and Gender")+
  ylab("Number of Stops")+theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Plot 7 – Traffic Violations per Geographic Area

```
traff_violations_filtered_2 <- traff_violations %>%
  mutate(`Driver State` = ifelse(`Driver State` != "MD", "Out of State", `Driver State`))
traff_violations_filtered_2 <- traff_violations_filtered_2 %>%
  mutate(`Driver State` = ifelse(`Driver State` == "MD", "Maryland", `Driver State`))
traff_violations_filtered_2 %>%
  count(`Driver State`) %>%
  ggplot(aes(x = reorder(`Driver State`, n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = n), hjust = 1.1, size = 4) +
  coord_flip() +
```

```
  labs(title = "Counts of MD vs non MD Drivers",
       x = "Category",
       y = "Count") +
  #scale_y_continuous(expand = expansion(mult = c(0, 0.1))) +
  scale_y_continuous(breaks = seq(0, max(9000), by = 1000)) +
  theme_minimal()
```



What are our Findings?

Within our analyses, we considered the most common and least common findings we had at hand. The plot of car makes and how many times throughout the data a specific make was pulled over helped our group come up with some interesting results from our first question.

The bar graph we created had a significant portion of Toyota makes taking the top place as the most pulled over vehicle. As one might imagine, this is not surprising, being one of the top selling car makes in the United States of America. Considering the reputation of other cars such as a Honda or Ford, they were also unsurprisingly among the most pulled over as well. Our least pulled over vehicle was the make of Nissan. They most likely take up a smaller portion of the U.S. market and therefore have a smaller presence in most U.S. states.

For findings in relation to time, the first question was "What times during the week are traffic stops most likely to happen?". We used a heat map to look into patterns among the traffic stops and specifically, we saw that weekdays and weekends had inverse times for them. During

weekdays, the most traffic stops happened during the hours from 7:00 AM to 6:00 PM. For the weekends, the majority of stops happened around late night to early mornings. The least common times are respective to the other hours for weekdays and weekends.

For our questions regarding what times during the months of a year, we found a couple of things: first, we noticed that a majority of the traffic stops happened slightly after the beginning of the year but dipped to its minimum in the summer months. The highest point here, which is truly a point, was during March.

Going forwards, the data for different years had some inherently interesting trends. Over the years of 2012 to 2013, there seems to be a trend upwards until a near peak in 2014. Then from 2015 to 2019 we have a sharp decline. Of course, after this point, 2020 marks the pandemic so we have an all time low throughout the data. More or less the trend picks back up to a more normal trend.

For our 4th question, the correlation of injury to seat belts was examined, and found that the obvious assumption of seatbelts being a useful instrument of human life protection, held true. Once again utilizing a heat map, we found, through a less complex system, that the data showed fewer injuries with more seat belt usage.

For a comparison of traffic stops and race plus sex of an individual, we faceted several graphs together based on separate races and found a majority of the data showed men being stopped more than women. Additionally, Black individuals were stopped most, while other groups were stopped less frequently.

In our final graph, we did a simple bar graph with two options: Marylanders and non-Marylanders, and who within Montgomery County where each driver was from during their stop. This was a harder dataset to tidy and combine into any graph, so we made do with a numerical count of each. Here we find an overwhelming amount of traffic stops involved Marylanders.

Limitations:

While our project provided some insight into various variables related to traffic stops/violations, it also presented some limitations. Much of the data consisted of binary yes and no responses (1s and 0s), which limited the complexity of our statistical modeling and visualizations.

Our findings apply specifically to traffic stops that occurred throughout the years in Maryland. Additionally, since the data-set is unique to Maryland, the findings may not be the same as in other regions that may have different policies, demographics, or procedures. In the future it would be helpful to see how the data is different or even similar in these other regions.