STA234: Homework 3

Derin Gezgin

2025-02-27

Part 1: Introduction

Tell me what problem you are working on? Why is this problem interesting and important. State specific research questions your group will work on. Introduce recent research done in area related to your problem. You can pack all this together to motivate us. Do keep it short, to the point, and interesting.

Traffic stops are a regular part of our lives, in fact, more than 20 million Americans are stopped each year in the traffic (Pierson et al., 2020). Traffic stops are one of the most common ways of public-police interaction. As police officers conduct these stops, the decision-making process comes down to human judgment, which certainly comes with a certain type of bias. There have been many research projects that focuses on possible bias factors in traffic stops. Most of these studies found that the race of the driver is an important factor influencing the likelihood of being stopped and the outcome of the stop.

According to Pierson et al. (2020) Black and Hispanic drivers are stopped and searched more often than White drivers. However, Black drivers are less likely to be stopped after sunset -compared to the rate of being stopped during the day- when the face of the driver is less visible. It is also pointed out that, the bar to search Black and Hispanic drivers is generally lower than White drivers. Lastly, the study also concludes that the success rates of searches is lower for Hispanic drivers compared to White and Black drivers who has comparable hit rates. Similarly Xu et al. (2024) points out that Black drivers are stopped at higher rates compared to their proportion in the traffic.

In my data-analysis project, I am planning to focus on the demographical analysis of the traffic stops conducted in San Francisco between 2007 and 2016. I already found a research project conducted by *San Francisco Bay Area Planning and Urban Research Association (SPUR)* on San Francisco traffic stop data which only covers the 2019 data. San Francisco Bay Area Planning and Urban Research Association (2023) also has similar findings of Pierson et al. (2020), as it shows Black and Hispanic drivers are stopped more than their share in the population while Black drivers have significantly lower citation rate compared to White and Hispanic drivers. In fact, according to the SPUR study, more than half of the Black drivers who are stopped do not end up with citation at all. At the same time, different than the previous studies I was able to find, I am planning to focus on the locational aspects of the traffic stops and analyze the relationship between the time, outcome, demographics, and the location of the stop.

I can list my possible research questions for my traffic stops data of San Francisco between 2007 and 2016 as the following:

1. What is the relationship between being stopped and the general demographics?
    - How does this relationship change when we split the data into separate times of the day?
2. How does the outcome of the traffic stop (warning, citation, search, arrest) relate to the racial demographics of the driver?
3. How does the amount of drivers stopped vary by time of the day and day of the year?
4. Are certain parts of SF have higher traffic-stop rates?
    - How does this relationship look like if we take race into account as well.
    - How does the outcome of the stop relates to the location

---

Part 2: Data

Tell me about the data resource and explain dimensions of the data, variables in the data, and how does this data relate to your research questions.

In my research project I am using the San Francisco police stop data which is a part of the Stanford Open Policing Project. The Open Policing Project is an ongoing project, that collects and organizes law enforcement stop data from different counties across the USA. San Francisco data includes details of 905,070 vehicle stops from January 1st, 2007 to June 30th, 2016. The dataset has 22 different variables on different detail of the stop such as date/time/location of the stop, age/race/sex of the driver that is stopped, and the outcome of the stop.

It is important to note that, I will use data from January 1st, 2007 to December 31st, 2015, inclusive. This is crucial as having 2016 until June can skew my data and lead me to incorrect conclusions. At the same time, some of the variables like the location/district/open address, raw data fields won't be used anywhere in my research project. The final dataset I will use (without the extra variables and half of 2016) will have 864,722 stops and 17 variables.

The variables in this dataset are helpful for my research questions as they provide important details about the demographic information of each traffic stop. At the same time, I am able to access information on what happened during the stop and how did the stop resulted. In the EDA and further parts of my project, these variables will provide me with much flexibility and potential to explore more.

Part 3: EDA

Use your dataset to make data visualizations that explain the variables of interest and how information through the graphics provides easy solution for your research questions. Explain your steps on how these visualizations help with your project.

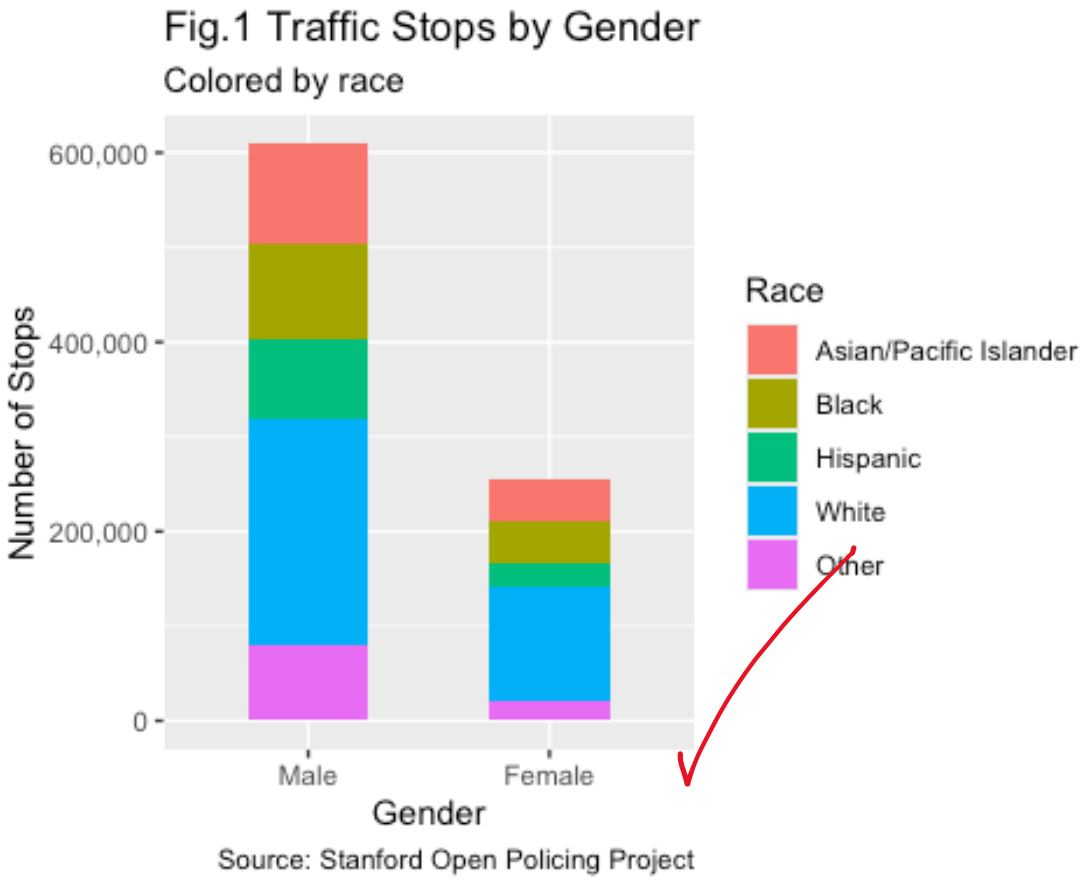
Before starting, I can read my data, convert specific columns into Date objects and also factors.

```
traffic.data = read.csv("../../DATA/ca_sf_vehicle_2007_2016.csv")

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

traffic.data$subject_race = factor(traffic.data$subject_race,
                                   levels = c("asian/pacific islander",
                                              "black",
                                              "hispanic",
                                              "white",
                                              "other"),
                                   labels = c("Asian/Pacific Islander",
                                              "Black",
                                              "Hispanic",
                                              "White",
                                              "Other"))

traffic.data$date = as.Date(traffic.data$date)
traffic.data$time = strptime(traffic.data$time, format="%H:%M:%S")

# Outcome is the most severe action taken among (warning, citation arrest).
# If no action is taken, it is NA. I made this NA values No Action so that
# I can make it a factor
traffic.data$outcome[is.na(traffic.data$outcome)] = "No Action"

traffic.data$outcome = factor(traffic.data$outcome,
                              levels = c("citation",
                                         "warning",
                                         "arrest",
                                         "No Action"),
                              labels = c("Citation",
                                         "Warning",
                                         "Arrest",
                                         "No Action"))
```

As the prompt suggests, in the initial EDA, I worked on getting straightforward answers from my dataset without anything advanced.

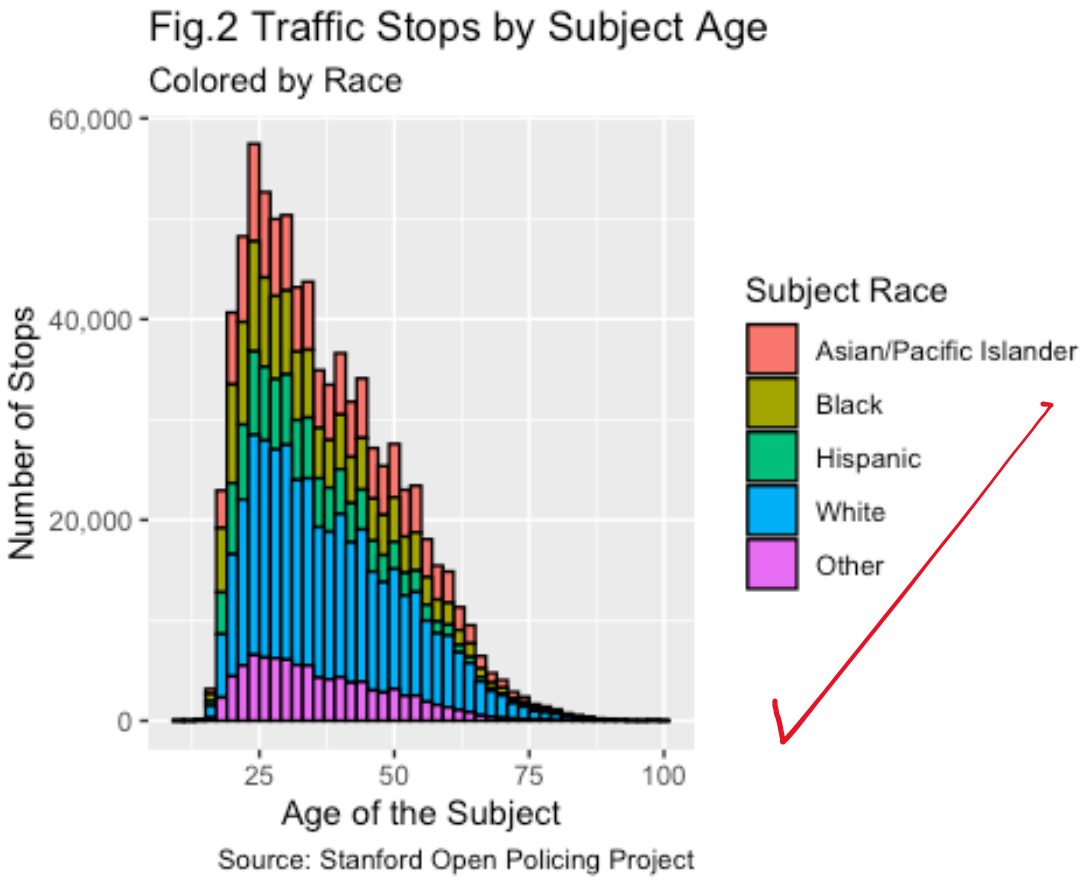
First of all, I wanted to look into the total number of traffic stops by gender and race throughout the time period.

```
group.race.sex.total = aggregate(traffic.data$raw_row_number,
                                 by = list(traffic.data$subject_race,
                                           traffic.data$subject_sex),
                                 FUN = length)

colnames(group.race.sex.total)= c("Race", "Gender", "Count")

g = ggplot(data = group.race.sex.total, aes(x = Gender,
                                            y = Count,
                                            fill = Race)) +
    geom_bar(stat = "identity", width = 0.5) +
    xlab("Gender") +
    ylab("Number of Stops") +
    labs(title = "Fig.1 Traffic Stops by Gender",
         subtitle = "Colored by race",
         caption = "Source: Stanford Open Policing Project") +

    scale_y_continuous(labels = scales::comma)

g
```



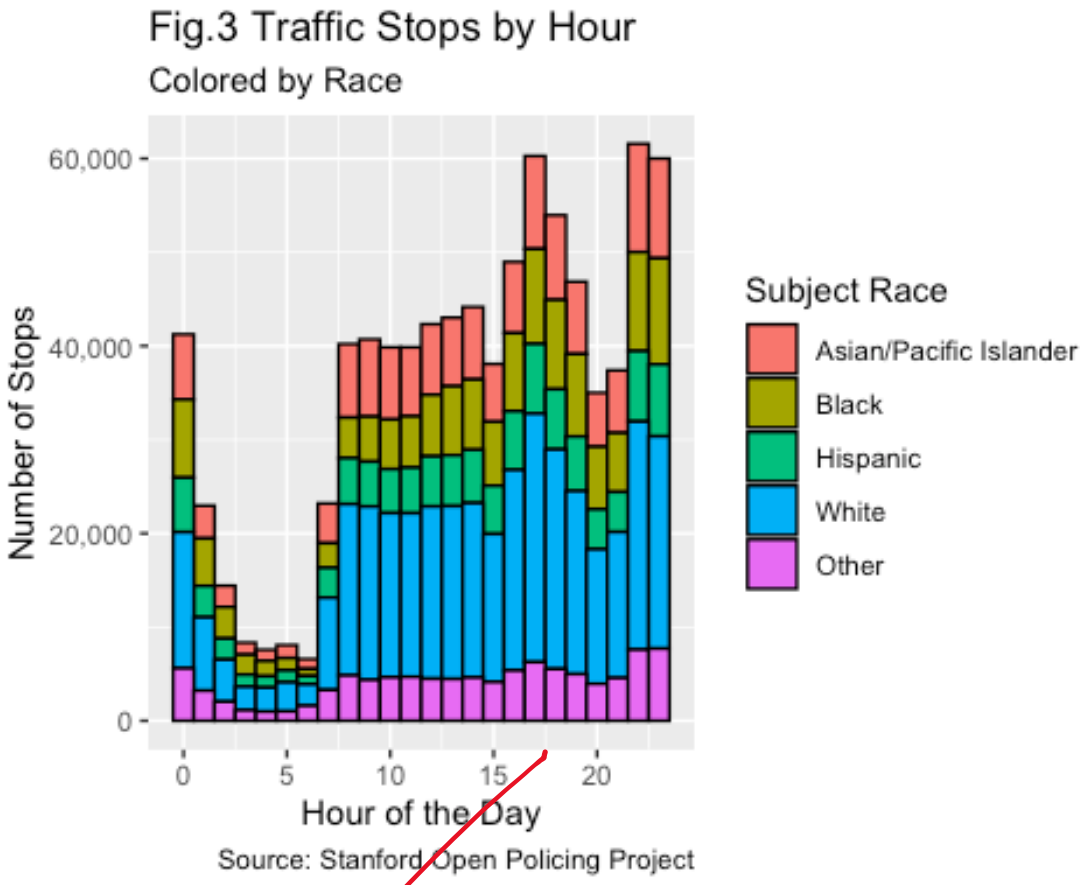
As another demographical data exploration, I checked the distribution of the subject age

```
g = ggplot(traffic.data, aes(x = subject_age,
                             fill = subject_race)) +
    geom_histogram(binwidth = 2,
                   color = "black") +
    xlab("Age of the Subject") +
    ylab("Number of Stops") +
    labs(title = "Fig.2 Traffic Stops by Subject Age",
         subtitle = "Colored by Race",
         caption="Source: Stanford Open Policing Project") +
    scale_y_continuous(labels = scales::comma) +
    labs(fill="Subject Race")

g
```



Following this initial exploration on gender, race and age; I focused on the distribution of time of the stops throughout the time-frame.

```
g = ggplot(traffic.data, aes(x = as.numeric(format(time, "%H")),
                             fill = subject_race)) +
    geom_histogram(binwidth = 1, color = "black") +
    xlab("Hour of the Day") +
    ylab("Number of Stops") +
    scale_y_continuous(labels = scales::comma) +
    labs(title = "Fig.3 Traffic Stops by Hour",
         subtitle = "Colored by Race",
         caption = "Source: Stanford Open Policing Project",
         fill = "Subject Race")

g
```



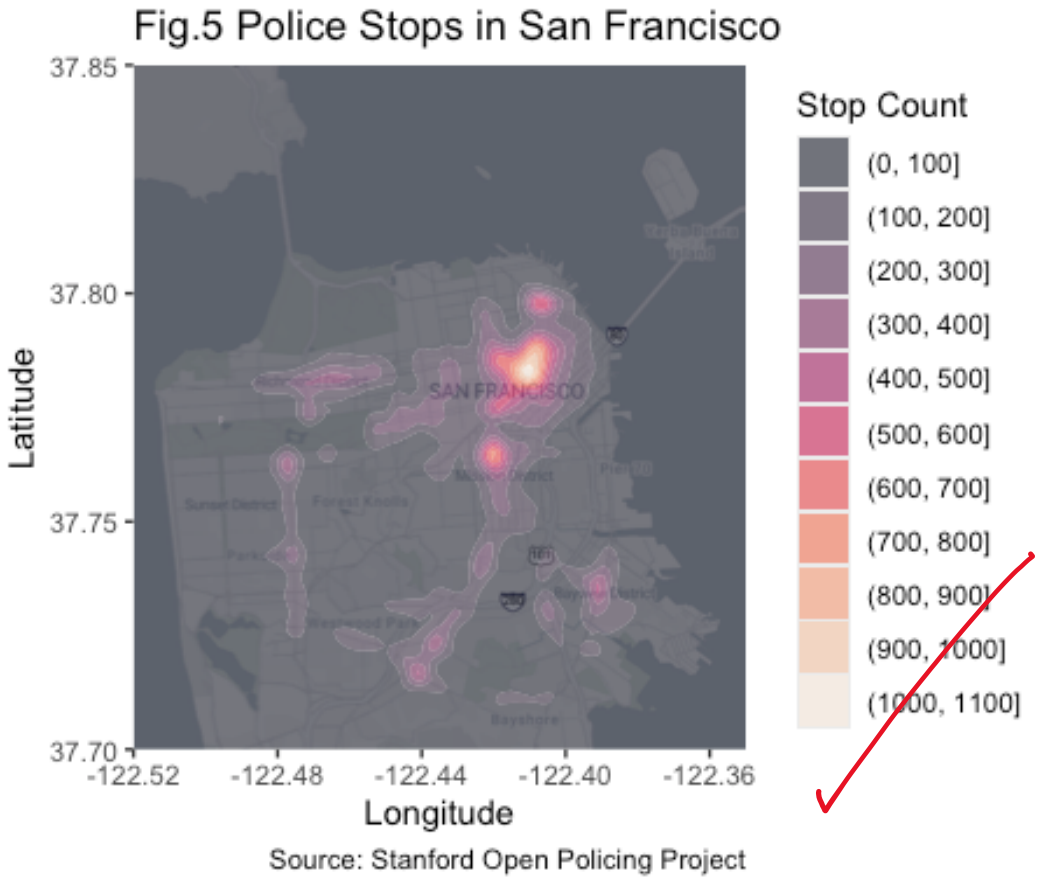
One of the other important parts of the traffic stops is the outcome of the stop. In this part, I plotted the outcome of the stops divided by the race. As the initial graph was hard to read, I applied log transformation to Y-Axis.

```
stop.race.total = aggregate(traffic.data$raw_row_number,
                            by = list(traffic.data$subject_race,
                                      traffic.data$outcome),
                            FUN = length)

colnames(stop.race.total)= c("Race", "Outcome", "Count")

g = ggplot(data = stop.race.total, aes(x = Outcome,
                                       y = Count,
                                       fill = Race)) +
    geom_bar(stat = "identity", width = 0.5) +
    xlab("Outcome") +
    ylab("Log of Number of Stops") +
    labs(title = "Fig.4 Outcome of the Stops",
         subtitle = "Colored by Race",
         caption = "Source: Stanford Open Policing Project") +
    scale_y_continuous(trans = "log10")

g
```

As a bonus, I created a heat-map of the stops in order to see which parts o the city have higher traffic-stop rates.

```
sf_map = get_stadiamap(bbox = c(left = -122.52,
                                bottom = 37.70,
                                right = -122.35,
                                top = 37.85),
                       zoom = 12,
                       maptype = "alidade_smooth")

ggmap(sf_map) +
    geom_density2d_filled(data = traffic.data,
                          aes(x = lng, y = lat),
                          alpha = 0.6) +
    scale_fill_viridis_d(option = "rocket") +
    labs(title = "Fig.5 Police Stops in San Francisco",
         fill = "Stop Count",
         caption = "Source: Stanford Open Policing Project") +
    xlab("Longitude") +
    ylab("Latitude")
```

Very nice!



From these simple EDA steps, I was able to have a general grasp of the dataset. One of my main and strongest finding was, according to Figure 1, male drivers had been stopped significantly more than the female drivers throughout the time period we work on. At the same time, shockingly, in Figure 2, the age-distribution of the stopped drivers does not follow a normal distribution but is significantly right-skewed. On the other hand, the distribution of the stops during the day had peaks at 17.00/22.00/23.00 PM and it significantly drops after 12AM which can be seen in Figure 3. When we check the outcome of the stops in Figure 4, we can see that it seems like pretty balanced among the races for all possible outcomes.

One of the main points I will work on in the upcoming parts of this project is to proportionate my data. Despite having the stop counts for specific races for all the graphs, it is hard to comment on the relationship between the races as we do not have information about their share in the larger population. In the upcoming parts of the project, I can access the census data for San Francisco and compare the subject race counts with the share of that specific race in the population. This can give us a more valuable insight.

Lastly, the density map of the traffic stops (Figure 5) shows us that the traffic stops really accumulate in the center area of the city while following through some major roads in the map. This plot is directly helpful to answer my 3rd research question.

In general, these simple EDA steps are helpful for me to answer some of my research questions I presented in the previous part. With few additions to the dataset and modifications on how I present the data, it is possible to have further information and visualizations from this dataset. As I mentioned, one of the crucial part of my data presentation would be normalizing the data by race distribution in order to have a better comparison.

Resources

- How to extract time from the HH:MM. Source
- How to create a heat-map of the San Francisco police stops: ggmap / Stadia Documentation / Density Map

References

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, *4*(7), 736–745. https://doi.org/10.1038/s41562-020-0858-1

San Francisco Bay Area Planning and Urban Research Association. (2023). *Putting an end to biased traffic stops in san francisco*. https://www.spur.org/news/2023-02-21/putting-end-biased-traffic-stops-san-francisco

Xu, W., Smart, M., Tilahun, N., Askari, S., Dennis, Z., Li, H., & Levinson, D. (2024). The racial composition of road users, traffic citations, and police stops. *Proceedings of the National Academy of Sciences*, *121*(24). https://doi.org/10.1073/pnas.2402547121