

STA234: HW4

Derin Gezgin

2025-03-06

Excellent! 50/50

Problem 2 [50 points]

Write a summary of your project, including information about:

- **Introduction:** Tell us what problem of interest. Why is this problem interesting and important? 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.
- **Data:** Tell us about the data resource and explain dimensions of the data, variables in the data, and how does this data relate to your research questions.
- **Initial findings:** Perform exploratory data analysis (EDA) including summaries and data visualizations for one research goal. Show main plot(s) and findings. Make sure to add labels, titles etc. to make your tables and graphs informative.
- share one advanced data visualization with any required information on how this plot is useful for the study and what it tells about your project (interpretation). Show your R skills, creativity, and advanced work in R.

Note: This is an initial report.

Note: I restarted the figure numbers as this is a separate part.

Introduction

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

San Francisco Traffic Stops & Race Distribution Datasets

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.

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.

Initial Findings

Reading the Data

```
traffic.data = read.csv("ca_sf_vehicle_2007_2016.csv")
race.data = read.csv("population race data.csv")
```

[illegible]

```

        "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")

traffic.data$outcome[is.na(traffic.data$outcome)] = "No Action"

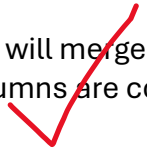
traffic.data$outcome = factor(traffic.data$outcome,
                              levels = c("citation",
                                           "warning",
                                           "arrest",
                                           "No Action"),
                              labels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

```



Merging Traffic Data and Race Data

To have the proportion of each race in our data, I will merge two data-frames here using the merge function. The subject_race and Year columns are common between these two datasets.



```

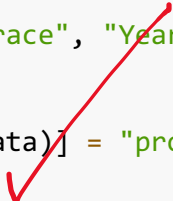
traffic.data$Year = format(traffic.data$date, "%Y")

traffic.data = merge(traffic.data,
                    race.data,
                    by = c("subject_race", "Year"),
                    all.x = TRUE)

colnames(traffic.data)[ncol(traffic.data)] = "proportioned_value"

traffic.data$Year = NULL

```




Now we can look at the traffic.data dataframe to see the new version of the dataframe

```

traffic.data[sample(nrow(traffic.data), 10), c("subject_race",
"proportioned_value", "date")]

```

##	subject_race	proportioned_value	date
## 242095	Black	0.060	2012-08-06
## 664352	White	0.421	2010-10-15



```
## 408624      Other      0.033 2007-09-23
## 577538      White      0.452 2008-09-17
## 805834      White      0.415 2014-03-28
## 813653      White      0.415 2014-06-22
## 323559      Hispanic    0.144 2009-06-25
## 801236      White      0.415 2014-04-29
## 318048      Hispanic    0.142 2008-10-23
## 583034      White      0.452 2008-11-05
```

We can see that we have a new value called `proportioned_value`. They represent the proportion of individuals from that race in that specific year. For example,

```
traffic.data[99999, c("subject_race", "proportioned_value", "date")]

##          subject_race proportioned_value      date
## 99999 Asian/Pacific Islander          0.346 2012-03-16
```

Asian/Pacific Islanders represented 34.6% of the population in 2012. As I showed in Figure 2 below, these proportions change over the years so we cannot have a global number for each year.

Normalizing the Proportions

While these proportions are valid for each year and race, as we duplicate them for each individual traffic stop, they do not have a clear use for our plots. At the same time, all the years do not have the same race distribution. We have to normalize these values so that we will have the weighted value of each stop.

As we are trying to show a **fair** analysis of data across the races, we have to take the multiplicative inverse of the weights. This ensures that groups with a smaller share in the population is equally represented.

Now the issue is we have bunch of weights but we did not normalize them in any way such that they have a fixed sum. We can ensure this by dividing each weight by the sum of all the weights in the dataset.

```
traffic.data$generalWeights = (1 / (traffic.data$proportioned_value)) /
sum(1 / (traffic.data$proportioned_value))
sum(traffic.data$generalWeights)

## [1] 1
```

We can see that the weights has a sum of 1 which shows that they have now have a normalized value.

Data Exploration

Exploring the Race Distribution of the San Francisco Population from 2007 to 2016

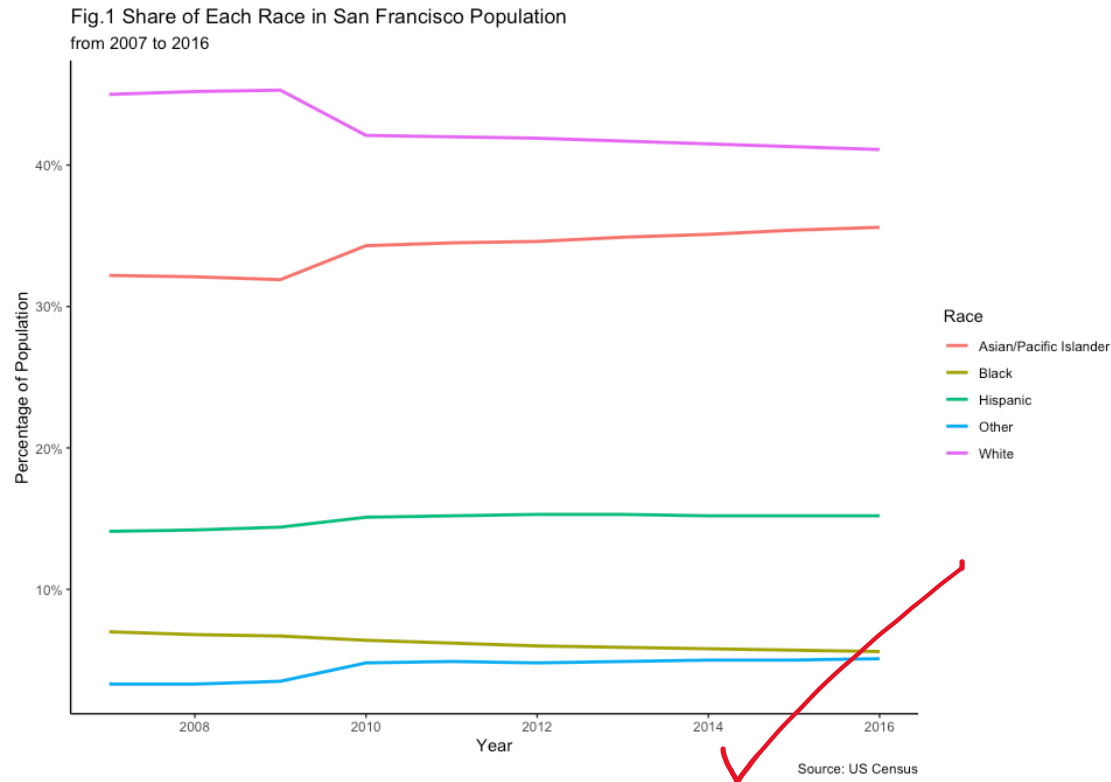
As I mentioned at the beginning of the previous section, the distribution of different races are not equal. To show their progression over time, we can look at their line graph.

```
race.by.year = aggregate(race.data$Population_Percentage,
                          by = list(race.data$subject_race,
                                    race.data$Year),
                          FUN = sum)

colnames(race.by.year) = c("subject_race", "Year", "Population_Percentage")

race.line.plot = ggplot(data = race.by.year,
                        aes(x = Year,
                            y = Population_Percentage,
                            color = subject_race)) +
  geom_line(size = 1) +
  labs(title = "Fig.1 Share of Each Race in San Francisco Population",
        subtitle = "from 2007 to 2016",
        x = "Year",
        y = "Percentage of Population",
        color = "Race",
        caption = "Source: US Census") +
  scale_y_continuous(labels = scales::percent) +
  theme_classic()

race.line.plot
```



This line graph shows us that, throughout the time-frame there had been a significant difference in the share of different races in San Francisco. We can also view the 2016 race distribution in a donut plot.

```

race.2016 = subset(race.data, Year == 2016)
race.2016$fraction = race.2016$Population_Percentage /
sum(race.2016$Population_Percentage)
race.2016$ymax = cumsum(race.2016$fraction)
race.2016$ymin = c(0, head(race.2016$ymax, n = -1))

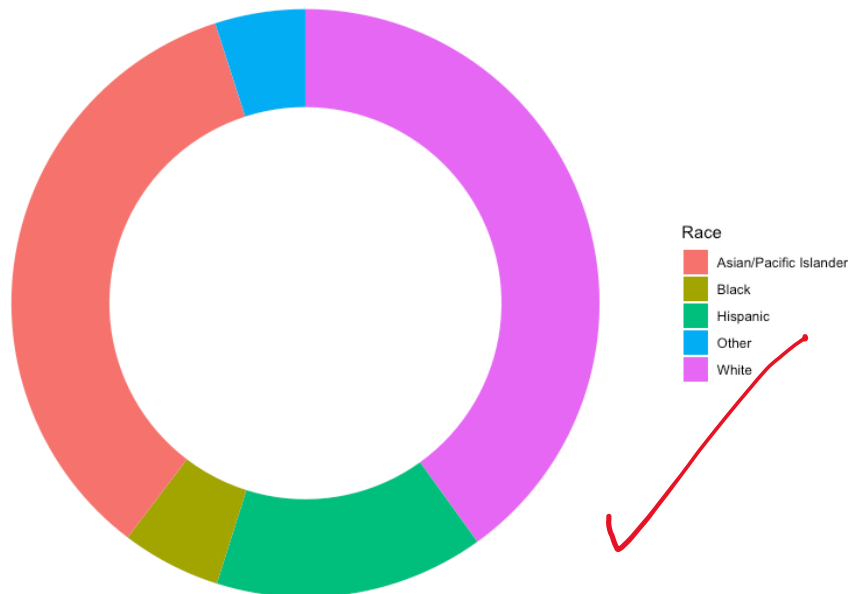
race.distribution.2016 = ggplot(race.2016,
                                aes(ymax = ymax,
                                    ymin = ymin,
                                    xmax = 4,
                                    xmin = 3,
                                    fill = subject_race)) +

  geom_rect() +
  coord_polar(theta = "y") +
  xlim(c(1, 4)) +
  theme_void() +
  labs(title = "Fig.2 Race Distribution in San Francisco (2016)",
       fill = "Race",
       caption = "Source: US Census")

race.distribution.2016

```

Fig.2 Race Distribution in San Francisco (2016)



Source: US Census

From Figure 2, we can see that the proportion of Black, Hispanic, and Other Race individuals in the San Francisco is significantly low while the large majority of individuals are White and Asian/Pacific Islander. This shows us the need for adjusting the data for race distribution as there is a significant gap between the proportion of races.

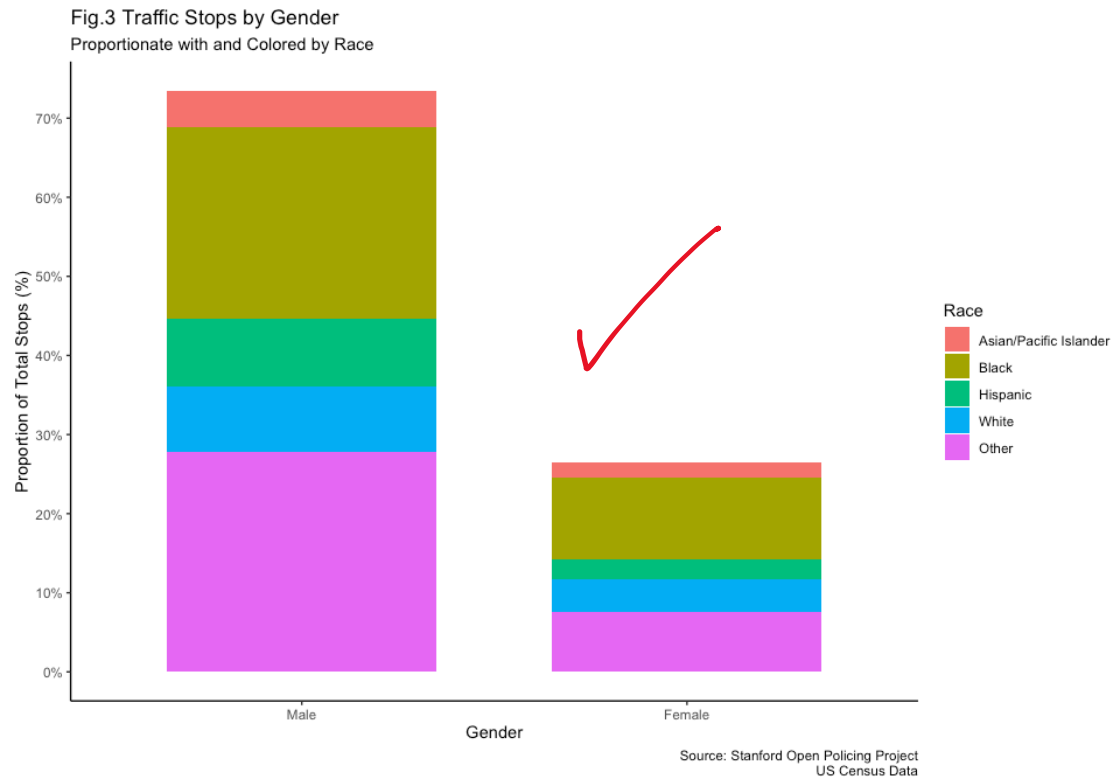
Distribution of the Traffic Stops by Gender

```
group.race.sex.total = aggregate(traffic.data$generalWeights,  
                                by = list(traffic.data$subject_race,  
                                           traffic.data$subject_sex),  
                                FUN = sum)  
  
colnames(group.race.sex.total) = c("Race", "Gender", "Count")  
  
stops.sex.race.bplot = ggplot(data = group.race.sex.total, aes(x = Gender,  
                                                                y = Count,  
                                                                fill = Race))  
  
+  
  geom_bar(stat = "identity", width = 0.7) +  
  scale_y_continuous(labels = scales::percent_format(scale = 100),  
                     breaks = seq(0, 1, 0.1)) +  
  xlab("Gender") +  
  ylab("Proportion of Total Stops (%)") +  
  labs(title = "Fig.3 Traffic Stops by Gender",  
       subtitle = "Proportionate with and Colored by Race",  
       caption = "Source: Stanford Open Policing Project\nUS Census Data")
```


+

```
theme_classic()
```

```
stops.sex.race.bplot
```



Before my interpretation of this graph, it is important to note that, the Y-Axis values do not represent the number of the stops anymore but they represent the proportion of the stops in all of the traffic stops.

We can still see that more than 70% of the stopped drivers were male compared to ~25% female drivers. As we adjusted our data by race, we can see that black drivers account for around 30% of the stops while covering less than 10% of the population. On the other hand, Asian/Pacific Islander drivers account for less than 10% of the stops while covering more than 30% of the traffic stops.

We can definitely confirm the findings of the previous studies where it was concluded that Black drivers were stopped significantly more than their share in the population.

Distribution of Traffic Stops by Age

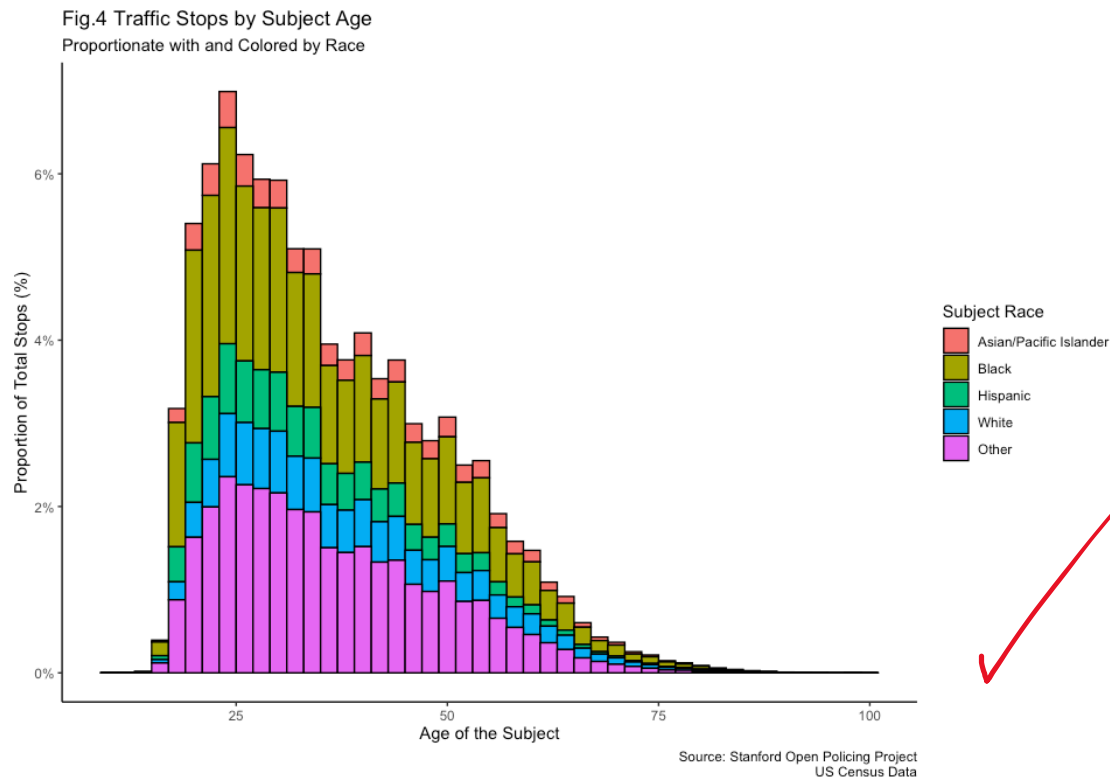
```
stops.age.hist = ggplot(traffic.data,
                        aes(x = subject_age,
                           fill = subject_race,
                           weight = generalWeights)) +
  geom_histogram(binwidth = 2, color = "black") +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
```

```

xlab("Age of the Subject") +
ylab("Proportion of Total Stops (%)") +
labs(title = "Fig.4 Traffic Stops by Subject Age",
      subtitle = "Proportionate with and Colored by Race",
      caption = "Source: Stanford Open Policing Project\nUS Census Data",
      fill = "Subject Race") +
theme_classic()

```

stops.age.hist



When we check the age distribution, we can again see the high share of black drivers across all the ages groups that were stopped. It is important to note that percentage of Hispanic drivers that are stopped are very similar to White drivers (while Hispanic drivers have a lower share in the population). Lastly, it can be seen that the age of the stopped driver do not follow a bell-curve distribution, and obviously right-skewed.

Distribution of Traffic Stops by the Time of the Day

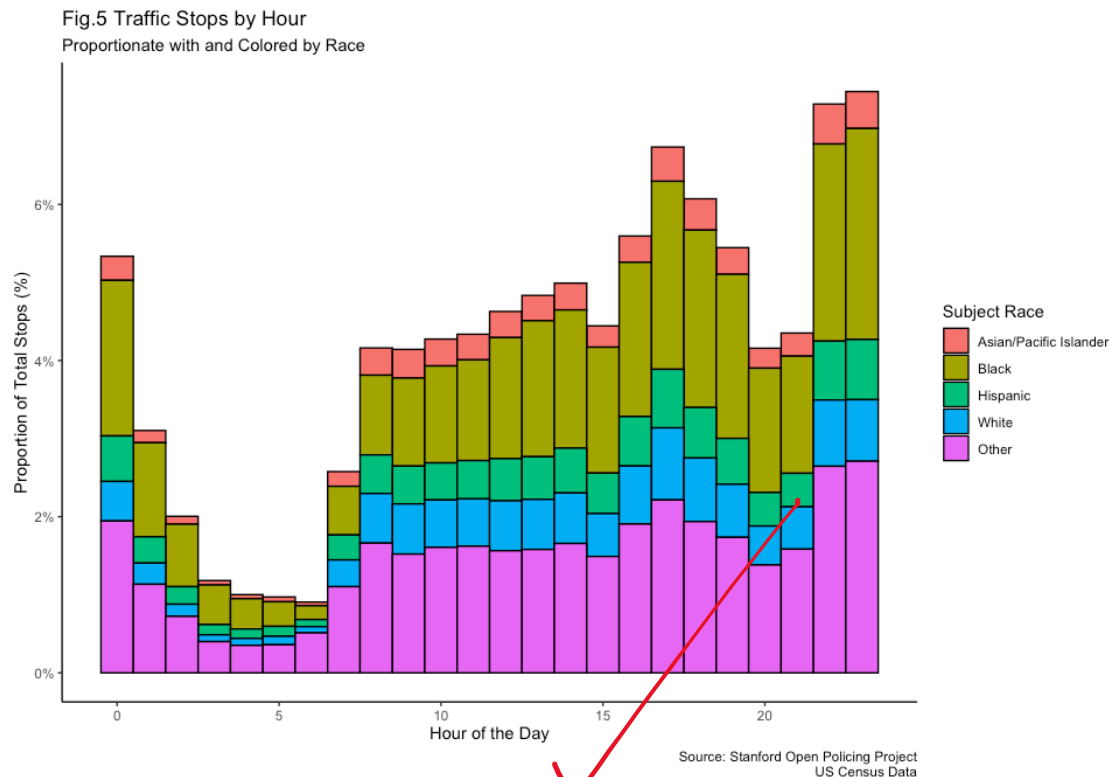
```

stops.time.hist = ggplot(traffic.data, aes(x = as.numeric(format(time,
"%H"))),
                        fill = subject_race,
                        weight = generalWeights)) +
  geom_histogram(binwidth = 1, color = "black") +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Hour of the Day") +
  ylab("Proportion of Total Stops (%)") +

```

```
labs(title = "Fig.5 Traffic Stops by Hour",
      subtitle = "Proportionate with and Colored by Race",
      caption = "Source: Stanford Open Policing Project\nUS Census Data",
      fill = "Subject Race") +
theme_classic()
```

stops.time.hist



In Figure 5, it can be seen that after midnight, the amount of stops significantly decline and start to rise again at morning hours around 7AM. Interestingly, we cannot see that the proportion of Black drivers stopped declining after sunset which was claimed by Pierson et al. (2020). I will investigate this further in the upcoming parts of the project.

Distribution of Traffic Stops by Outcome

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

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

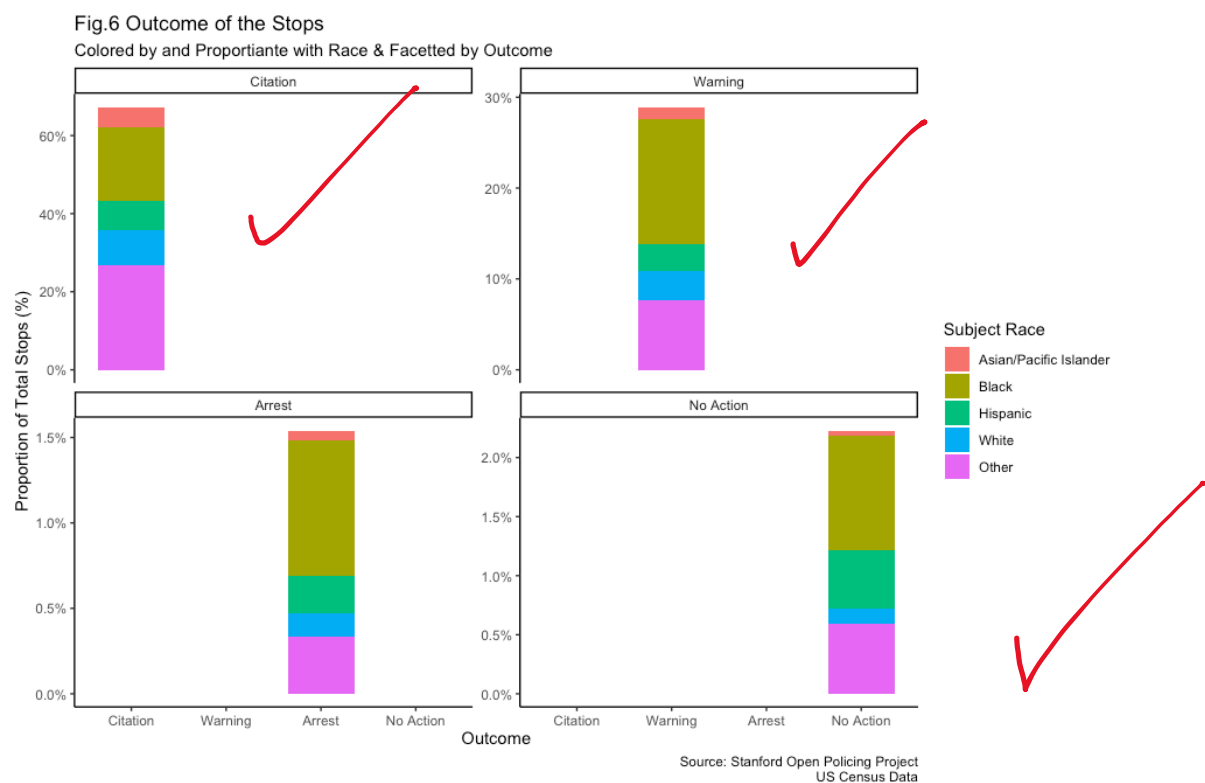
stops.outcome.bplot = ggplot(data = stop.race.total,
                              aes(x = Outcome,
                                   y = Count,
                                   fill = Race)) +
```

```

geom_bar(stat = "identity", width = 0.7) +
scale_y_continuous(labels = scales::percent_format(scale = 100)) +
xlab("Outcome") +
ylab("Proportion of Total Stops (%)") +
labs(title = "Fig.6 Outcome of the Stops",
      subtitle = "Colored by and Proportiantie with Race & Facetted by
Outcome",
      caption = "Source: Stanford Open Policing Project\nUS Census Data",
      fill = "Subject Race") +
facet_wrap(~ Outcome, scales = "free_y") +
theme_classic()

```

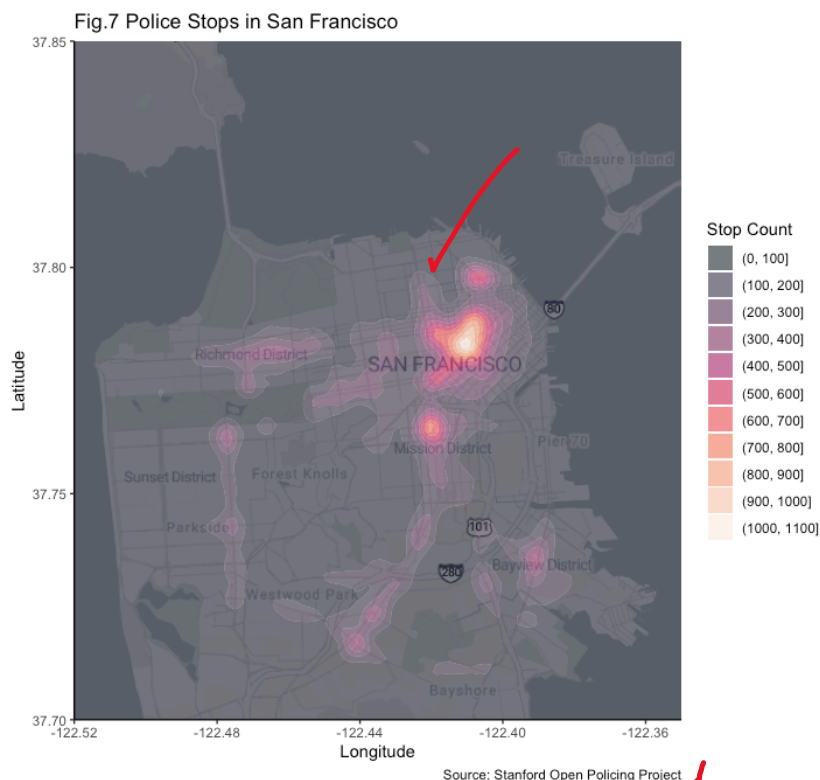
stops.outcome.bplot



From the facets in Figure 6, we can see that most of the arrests resulted with citation followed by warning, no action and arrest. While citation has a race distribution relatively equal, the arrests were dominated by Black drivers. At the same time, while there are nearly no White drivers whose stop results with No Action, 1% of the total stops, which are all Black drivers, result in No action, which raises the concern about the legitimacy of the stops of Black drivers. In the upcoming parts of the project, I can examine the reason for stop variable and see the correlation between the driver race, reason and outcome of the traffic stops.

Spatial Distribution of Traffic Stops in San Francisco

```
sf_map = get_stadiamap(bbox = c(left = -122.52,  
                                bottom = 37.70,  
                                right = -122.35,  
                                top = 37.85),  
                      zoom = 12,  
                      maptype = "alidade_smooth")  
  
stop.map = 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.7 Police Stops in San Francisco",  
       fill = "Stop Count",  
       caption = "Source: Stanford Open Policing Project") +  
  xlab("Longitude") +  
  ylab("Latitude") +  
  theme_classic()  
  
stop.map
```



We can see that most of the traffic stops accumulated around the same part of the city and following some paths which can point to major roads or generally used paths, etc.

Examining Spatial Distribution by Race

```
stop.map.race = 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.8 Police Stops in San Francisco",  
    subtitle = "Facetted by Race",  
    fill = "Stop Count",  
    caption = "Source: Stanford Open Policing Project") +  
  xlab("Longitude") +  
  ylab("Latitude") +  
  facet_wrap(~ subject_race, nrow = 2) +  
  theme_classic() +  
  theme(panel.spacing.x = unit(2, "lines"))
```

stop.map.race



As an addition to the figure 7, I split it to facets of different races. Interestingly, we can see that traffic stops of Black drivers have accumulated in specific places in the city while Hispanic drivers' traffic stops have followed a path (which can point to a major road). In the future parts of the project, I will examine the reasoning of this relationship as White, Asian and Hispanic drivers do not show this peak pattern and have more linear distribution around the map.

Conclusion

In our analysis of the San Francisco traffic stops data from 2007 to 2016, we saw a clear racial bias. Black drivers were stopped for more than their share in the population while Asian/Pacific Islander drivers were stopped the least. Black drivers faced significantly high arrest rates, while significant amount of their stops resulted in no action which question the legitimacy of these stops.

Moving forward, I plan to examine the stop reasons, location-based biases and the general flow of stop (maybe with an alluvial plot). At the same time, I will look at how the stop rates change under different conditions like time of the year.

Resources

- How to plot a line graph: [R-Graph Gallery](#)
- How to add spacing between facets: [StackOverflow Question](#)
- Modifying tick frequency in axis: [Source](#)

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>