Project Part 1

Description of the data

Background Information

Interaction with police most commonly occurs at traffic stops. Police across the country pull over 50,000 drivers per day, and several states have passed laws requring police departments to record each traffic stop an officer initiates (1). Important demographic factors, such as race, gender, and age of the driver are often recorded, as well as details regarding the reason the driver was stopped, and what the outcome of the stop was. Many studies have revealed that people of color are disproportionately affected by traffic stop policing, and there are racial and gender differences in how people percieve these interactions with police (2). Fines and fees from stops can result in financial burdens for subjects, and minority groups who are disproportionately targeted by police have been found to suffer from increased stress and feelings of unfairness and second-class citizenship (2). As consequence, tensions between minority groups and law enforcement run high. This became evident to the wider US population in summer of 2020, as protests against police brutality and discrimination escalated following the killing of George Floyd. Given that traffic stop policing is key to many states' policing strategies, it is important to analyze the data to better understand the prevalance of discrimination and prevent harmful practices from continuing (3).

Data Collection

This data set was requested from the state of North Carolina by the Stanford Open Policing Project, an initaitive which has been studying racial disparities in traffic stop policing since 2014. The observations in this data were were collected by the Charlotte-Mecklenburg Police Department and the UNC Charlotte University Police, and together, form the full population of stops caused by stop-light/sign violations in Charlotte, NC from 2000-2015. Credit goes to the Stanford Open Policing Project for compiling the raw dataset, which I have further cleaned as specified below:

- 22 columns were removed from the orginal data set, reasons being that the additional variables were found to be repetitive, or irrelevant for non-government officials (e.g ID numbers, department codes).
- The data was subsetted by stops which occured in the year 2015, for reason of a Stop Light/Sign Violation, where search_conducted = TRUE. In addition, observations were subsetted by race. Asian drivers and drivers whose race was unknown made up less than 2% of the data, and were removed to reduce noise when caculating proportions. The data I am examining in this project is now a record of all police searches conducted at stop light/signs in Charlotte, NC, in 2015.

Explanation of rows and variables

Each row is a record of an individual traffic stop, and each column (variable name) describes an attribute corresponding to the given traffic stop which occured.

Variable Names:

- date: the date in which the stop occured, YYYY-MM-DD format
- subject age: age of the stopped subject
- subject_race: race of the stopped subject, values are standardized to white, black, hispanic
- subject_sex: sex of the stopped subject
- outcome: action taken, values are among arrest, citation, and warning
- countraband_found: indicates whether contraband was found

Potential Issues

Potential issues may arise from how the original population was subsetted. The data can be considered a severe undersample; stops recorded only in 2015 make up less than 10% of the overall population. It is difficult to determine whether 2015 searches during police stops are truly representative of the population, and thus whether or not conclusions can be generalized. Additionally, stops which occured because of Vehicle Regulatory Violations and Vehicle Equipment Violations were far more likely to occur in the orginal sample. Thus it is also difficult to guarantee that seaches caused by Stop Light/Sign Violations are represenative of the population. Other potential issues may have to do with how the original data was collected. Police officers do not monitor random traffic stops/signs, they intentionally place themselves in locations where traffic violations are expected, or where other

public crimes are expected to occur. As these expectations tend to be informed by personal bias and stereotypes against minority communities (college-age students, racial minorities), this practice can lead to over-policing in certain areas, creating harmful feedback loops and causing certain demographic groups to be over-represented in the data (4). This potential issue should prevent us from drawing the conclusion that individual police officers are biased against minority groups, but rather make us more critical of social norms and institutions.

Numeric and Graphic Representations

Table 1

```
Mean_Driver_Age <- data.frame(NC_stops %>%
  group_by(subject_race) %>%
  summarize(avg_age = round(mean(subject_age),2)))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
names(Mean_Driver_Age) <- c("Race", "Mean Age")
```

```
Mean_Driver_Age
```

```
##       Race Mean Age
## 1    black    28.08
## 2 hispanic    23.76
## 3    white    30.36
```

Graph 1

```
race_rename <- c(`black`="Black", `hispanic`="Hispanic", `white`="White")
```

```
ggplot(NC_stops, aes(x=outcome, y=prop)) +
  geom_bar(aes(y=..prop.., group=1)) +
  facet_grid(.~subject_race) +
  facet_grid(.~subject_race, labeller=as_labeller(race_rename)) +
  theme(strip.background=element_rect(fill="white", color="black"),
        strip.text.x = element_text(size=12)) +
```

```
  labs(title="Stop Outcomes by Race",
       x="Outcome",
       y="Proportion of Stops") +
  scale_x_discrete(labels=c("warning" = "Warning", "citation" = "Citation",
                            "arrest" = "Arrest")) +
  theme(plot.title = element_text(hjust = 0.5))
```

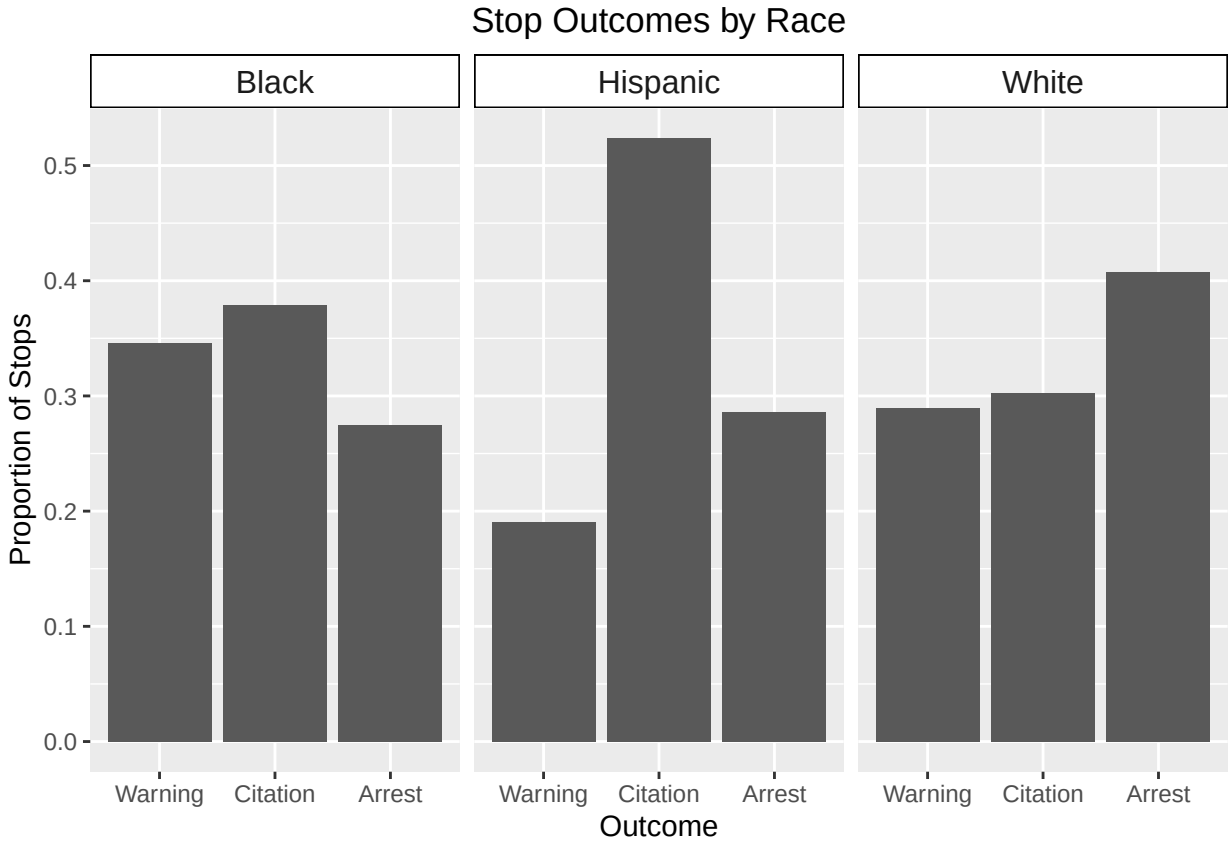


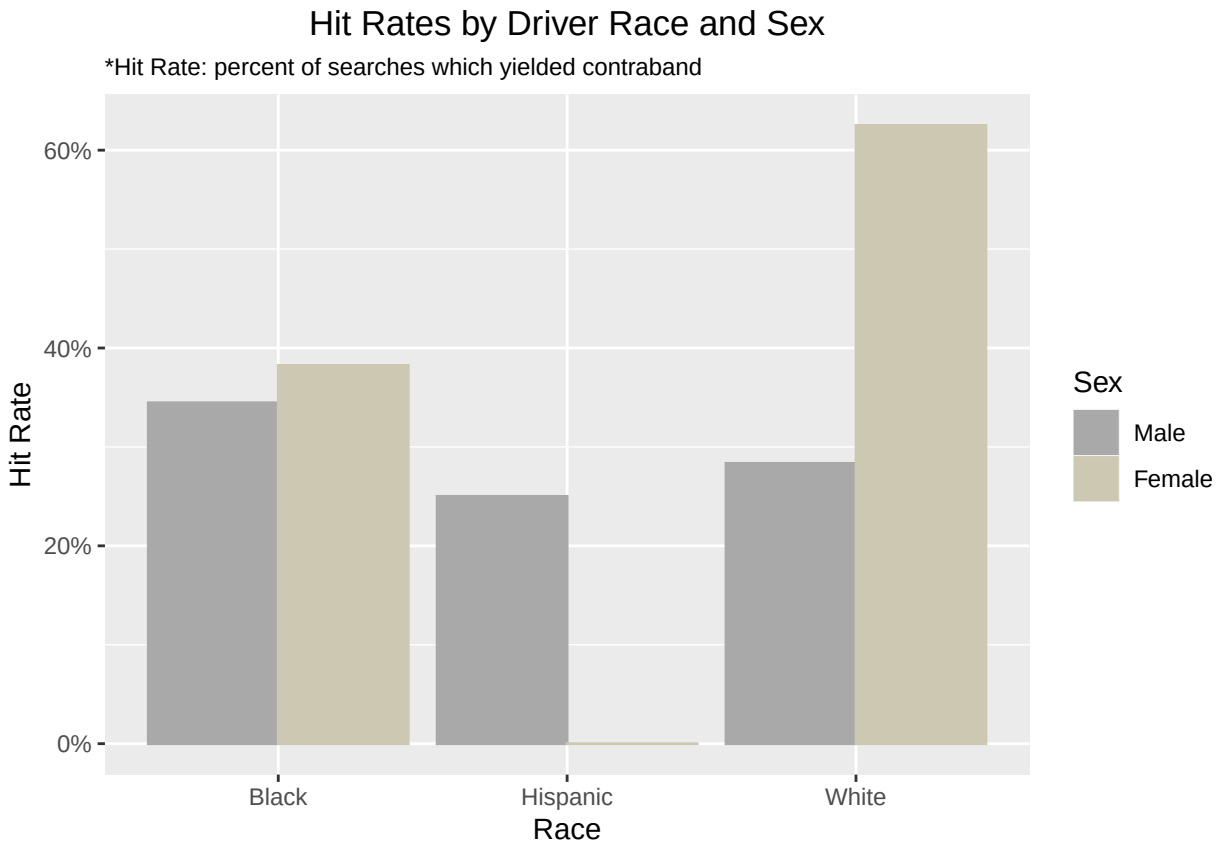
Table 2

```
# Reference 5
Hit_rates <- data.frame(NC_stops %>%
  group_by(subject_race) %>%
  summarize(hit_rate = round(mean(contraband_found),2)))

## `summarise()` ungrouping output (override with `.groups` argument)

names(Hit_rates) <- c("Race", "Hit Rate")

Hit_rates
```

```
##       Race Hit Rate
## 1    black     0.35
## 2 hispanic     0.24
## 3    white     0.36
```

Graph 2

```
# Reference 6
# Reference 7

ggplot(NC_stops, aes(x=subject_race,
                     y=as.numeric(contraband_found),
                     color=as.factor(subject_sex), fill=as.factor(subject_sex))) +
  stat_summary(fun="mean", geom="bar",position="dodge") +
  scale_y_continuous(labels=percent_format(accuracy=1)) +
  scale_color_manual(values=c("dark gray", "cornsilk3"),
                     labels=c("Male", "Female")) +
  scale_fill_manual(values=c("dark gray", "cornsilk3"),
                    labels=c("Male", "Female")) +
  labs(title="Hit Rates by Driver Race and Sex",
       subtitle="*Hit Rate: percent of searches which yielded contraband",
       x="Race",
       y="Hit Rate",
       color="Sex",
       fill="Sex") +
  scale_x_discrete(labels=c("black" = "Black", "hispanic" = "Hispanic",
                            "white" = "White")) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(size=9))
```



Conclusion

Based on the graphs and tables, we can conclude there is no obvious disparity between black and white drivers stopped at stop lights/signs. Drivers from both groups are about the same age, as seen in Table 1, and in Graph 1, we can see that both groups recieve warnings, citations, and arrests at similar rates. (Whites are arrested at a slightly higher rate, buth is does not seem especially significant in context.) Additionally, when examining the percentage of searches which yielded contraband, we can see in Table 2 that about 35% of searches are successful among both black and white drivers. This suggests that police officers target suspect behavior in both groups approximately equally. If the hit rate of one group was higher than the other, we could have concluded that the group with the higher hit rate is targeted only when officers are relatively certain contraband can be found. However, because this is not the case between blacks and whites, we can conclude that both groups are treated with relatively equal levels of suspicion. However, when examining the breakdown of hit rates by gender in Graph 2, it appears that females have higher hit rates. This leads us to conclude that officers may generally be less suspicious of females compared to males, across all racial groups. Note that white females have the highest hit rate by far, indicating that

officers may percieve them as least suspect. Hispanics, however, seem to be treated slightly differently, according to the data. The hit rate of this group is approximately 10 points lower than blacks and whites, and the average age of stopped hispanic drivers is approximately 7 years younger than blacks and whites. This could be because young hispanics are over-represented in the population of Charlotte, NC. If this is true, it could also indicate that age is correlated to hit-rates, but this is yet to be explored. At the moment, we cannot conclude whether or not police officers discriminate against Hispanics.

References

1. https://openpolicing.stanford.edu/
2. https://fbaum.unc.edu/books/SuspectCitizens/SuspectCitizens-Overview.pdf
3. https://www.policingproject.org/news-main/2019/9/27/its-time-to-start-collecting-stop-data-a-case-for-comprehensive-statewide-legislation
4. *Weapons of Mass Destruction*, Cathy O'Neil
5. https://gist.github.com/gadenbuie/c83e078bf8c81b035e32c3fc0cf04ee8
6. https://stackoverflow.com/questions/43634136/barchart-of-count-of-true-false-values-by-group-dodged-graphs
7. http://www.sthda.com/english/wiki/ggplot2-axis-ticks-a-guide-to-customize-tick-marks-and-labels#change-tick-mark-labels