

Project Problem [100 points]

This counts for presentation

P.(a) [25 points]

Explain what your project is about, share your research questions and justify why are these research questions are valid using existing research in the field? Share research papers as a list of references (See writing examples from assignments 6 and 7)

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.

very nice!

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. My research questions are:

1. What is the relationship between being stopped and the general demographics?

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.

Pierson et al. (2020) and Xu et al. (2024) also has a similar approach to this research area where they focus on possible racial biases in traffic stops including the effect of time of the day the stop occurred [Q1-3]. In my study, I employ a similar approach where I visualize and test if there is any racial bias in the traffic stops. Different than Pierson et al. (2020), I focus on San Francisco, rather than using all the data from the USA.

It is important to note that, in my literature review so far, I could not find any studies that worked on the locational aspect of the traffic stops [Q4]. At the same time, most of the studies focused on the racial bias, while I am also planning to examine other demographical factors and if there is any statistically significant relationship between them and any of the possible response variables.

P.(b) [20 points]

Share your final datasets and resources for these datasets. Explain stepwise all the work you have done for data wrangling to organize the dataset(s).

The original data I had from the [Stanford Open Policing Project](#) had traffic stop data from January 1st 2007 until June 30th 2016 in San Francisco. This 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. These 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.

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

First of all, to prevent any bias in my modeling or data due to having half of 2016 in my data, I removed the entries from January 2016 until June 2016

```
traffic.data$date = as.Date(traffic.data$date, format = "%Y-%m-%d")
traffic.data = traffic.data[traffic.data$date < as.Date("2016-01-01"),
]
```

Following this, I removed the columns I am not planning to use in my data analysis.

```
toRemove = c("location",
             "district",
             "type",
             "raw_search_vehicle_description",
             "raw_result_of_contact_description")
```

```
traffic.data = traffic.data[, !(names(traffic.data) %in% toRemove)]
```

I then converted the categorical variables like sex, race, and stop outcome into factors. It is important to note that, if the traffic stop does not have a significant outcome its outcome is saved as NA. I replaced the NA values with "No Action".

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

```
traffic.data$outcome = factor(traffic.data$outcome,
                              levels = c("citation",
                                          "warning",
                                          "arrest"),
```

```

                                "No Action"),
labels = c("Citation",
           "Warning",
           "Arrest",
           "No Action"))

```

Finally, I converted the date and time columns into date / time objects

```

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

```

I am also using a supplementary dataset from the US Census data which contains the proportions of each race from 2007 to 2016. I created this dataset manually by copying the values from a [3rd Party Website](#), which took the values from the US Census Data. This dataset is crucial to make a better analysis of the data as the races are not equally distributed.

```

race.data = read.csv("population_race_data.csv")

```

As this race proportions dataset is by year, I created a Year column in my original dataset to store the year the stops occurred in order to merge these two datasets. Following the merge, I removed this column.

```

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, as I had the proportion of race for each individual, I can normalize these values to have a weighted value for each traffic stop.

```

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

```

P.(c) [25 points]

Present exploratory data analysis (EDA) including summaries and data visualizations for one research goal. Make sure to add labels, titles etc. to make your tables and graphs informative. Explain how these numerical and graphical summaries address the research questions. Make sure you share your findings related to the problem not in terms of statistical terminology.

Proportion of Different Races in San Francisco

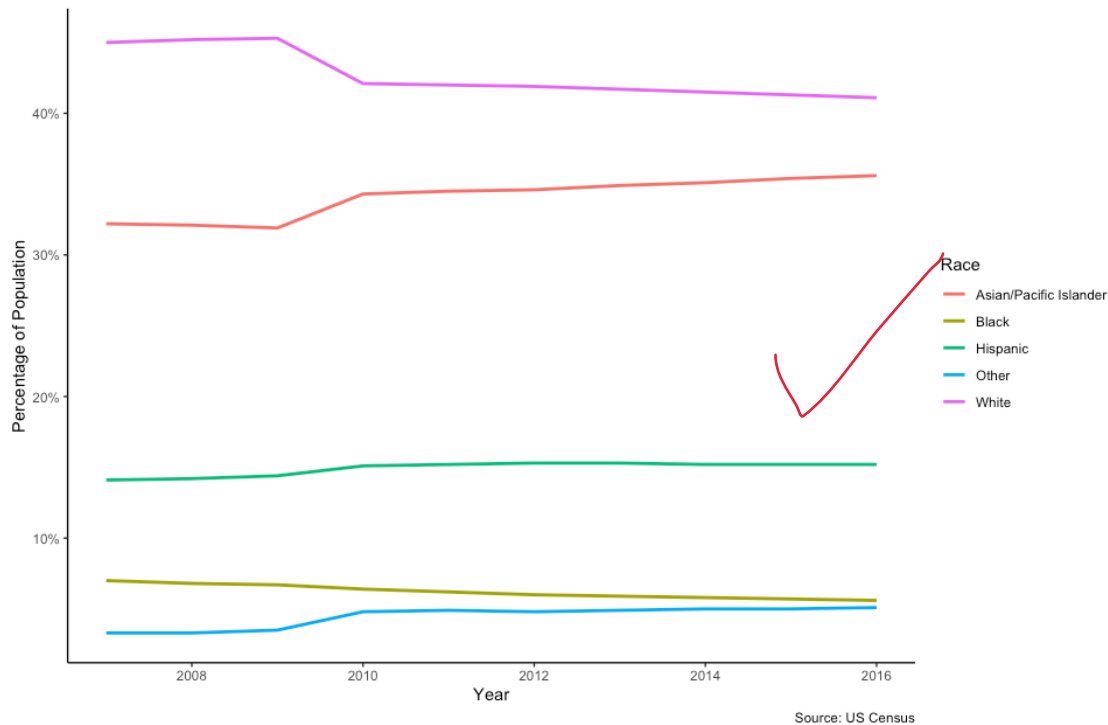
```
race.by.year = aggregate(race.data$Population_Percentage,
                          by = list(race.data$subject_race,
                                    race.data$Year),
                          FUN = sum)

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
```

Fig.1 Share of Each Race in San Francisco Population from 2007 to 2016



In the first part of my visualization, Figure 1, I showed that the distribution of races in San Francisco throughout the time-frame I work on, is not equal. This shows it clearly why we need to proportionate the traffic stops by the race. *Before interpretation of any further graphs, 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.*

Comparing the Race Proportions and the Stop Proportions

Add the year column

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

Calculate the stop proportions per year

```
stops_by_race_year = traffic.data %>%
  group_by(subject_race, Year) %>%
  summarise(Stop_Share = n(), .groups = "drop") %>%
  group_by(Year) %>%
  mutate(Stop_Share = Stop_Share / sum(Stop_Share)) %>%
  ungroup()
```

```

# Merge with the population data
merged_data = left_join(stops_by_race_year,
                        race.data,
                        by = c("subject_race", "Year")) %>%
  mutate(Stop_Share = round(Stop_Share * 100, 2),
         Population_Percentage = round(Population_Percentage * 100,
2))

# Stack the data of stop and population shares
merged_data = merged_data %>%
  mutate(Combined_Data = paste0(Stop_Share,
                                "% ", "\n(", Population_Percentage,
                                "%)"))

# Reshape the data
wide_table = merged_data %>%
  dplyr::select(subject_race, Year, Combined_Data) %>%
  pivot_wider(names_from = Year, values_from = Combined_Data) %>%
  arrange(subject_race)

# Output the table
race_table = wide_table %>%
  gt(rowname_col = "subject_race") %>%
  tab_header(title = "Stop % and Population % by Race and Year",
            subtitle = "Cells showing Stop Share and (Population
Share)") %>%
  fmt_markdown(everything()) %>%
  cols_align(align = "center") %>%
  gt_theme_nytimes()

race_table

```

Table 1: Stop % and Population % by Race and Year
Cells showing Stop Share and (Population Share)

For some reason, Table 1 is messed up?

.

A 1 1 1 1 1 1 1 1 1 1
si 6.7.7.5.7.7.8.8.
a 4 6 6 5 7 5 9 8 0
n 4 4 4 8 2 3 % 7 6
/ % % % % % % (% %
P (((((3 (((((3
a 3 3 3 3 3 3 4.3 3
ci 2.2.1.4.4.4.9.5.5.
fi 2 1 9 3 5 6 % 1 4
c % % % % % % % % % %
ls))))))))))
l
a
n
d
e
r

W 4 4 4 4 4 4 3 3 3
h 4.2.3.5.2.1.8.6.4.
it 2 9 2 7 5 7 2 5 7
e 1 7 3% 4 6 8 9 3
% % % (% % % % %
(((4 (((((4
4 4 4 2. 4 4 4 4 4
5 5.5.1 2 1.1.1.1.
% 2 3 % % 9 7 5 3
) % %)) % % % %
)))))))

From Table 1, we can see that there is a clear trend of over- and under-representation of different race groups. For example, while Black individuals make up only 6-7% of the population, their stop share is double or nearly triple their share in the population. on the

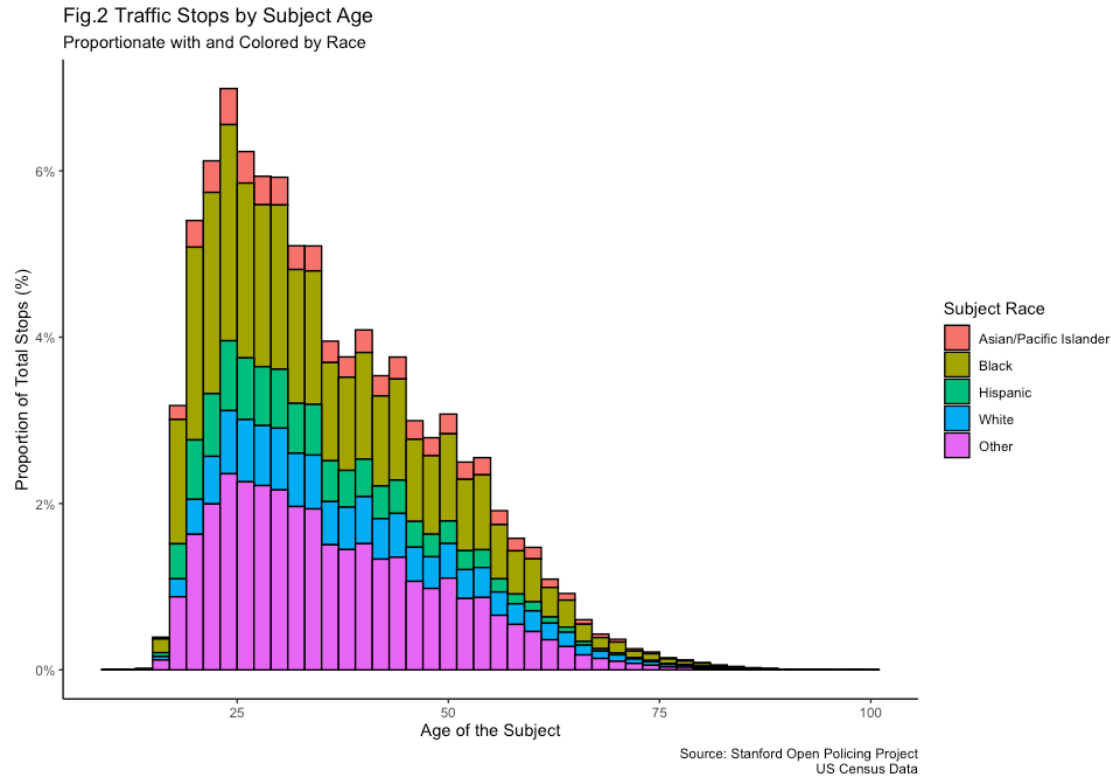
other hand, white individuals are consistently stopped less than their share in the population.

Distribution of Age of the Stopped Drivers

```
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.2 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
```





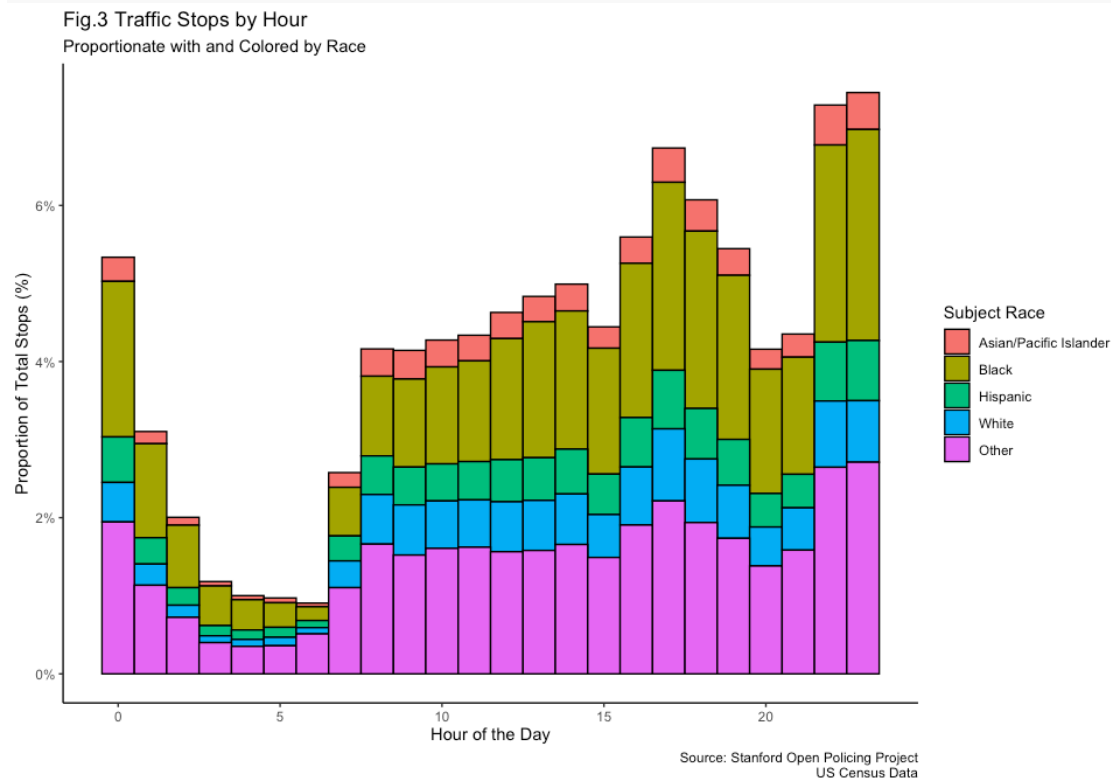
When we check the age distribution in Figure 2, we can see the high share of black drivers across all the ages groups that were stopped. It is important to note that percentage of Hispanic drivers that are stopped are very similar to White drivers (while Hispanic drivers have a lower share in the population). Lastly, it can be seen that the age of the stopped driver do not show a balanced distribution and young drivers are stopped more often than the old drivers.

Distribution of Traffic Stops Throughout 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.3 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
```



When we look at the traffic stops by time in Figure 3, 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).

Racial Distribution for Possible Outcomes of Traffic Stops

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

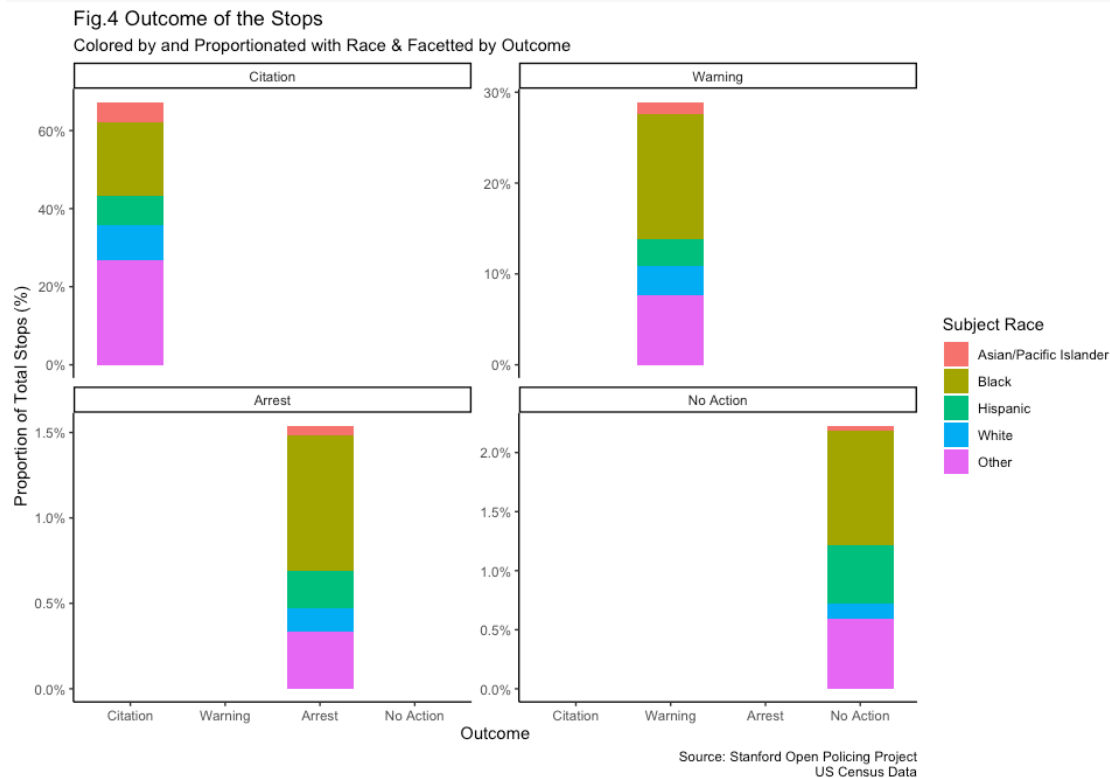
colnames(stop.race.total) = c("Race", "Outcome", "Count")
```

```

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.4 Outcome of the Stops",
       subtitle = "Colored by and Proportionated 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 4, 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.

Spatial Aspect of the Traffic Stops by Race

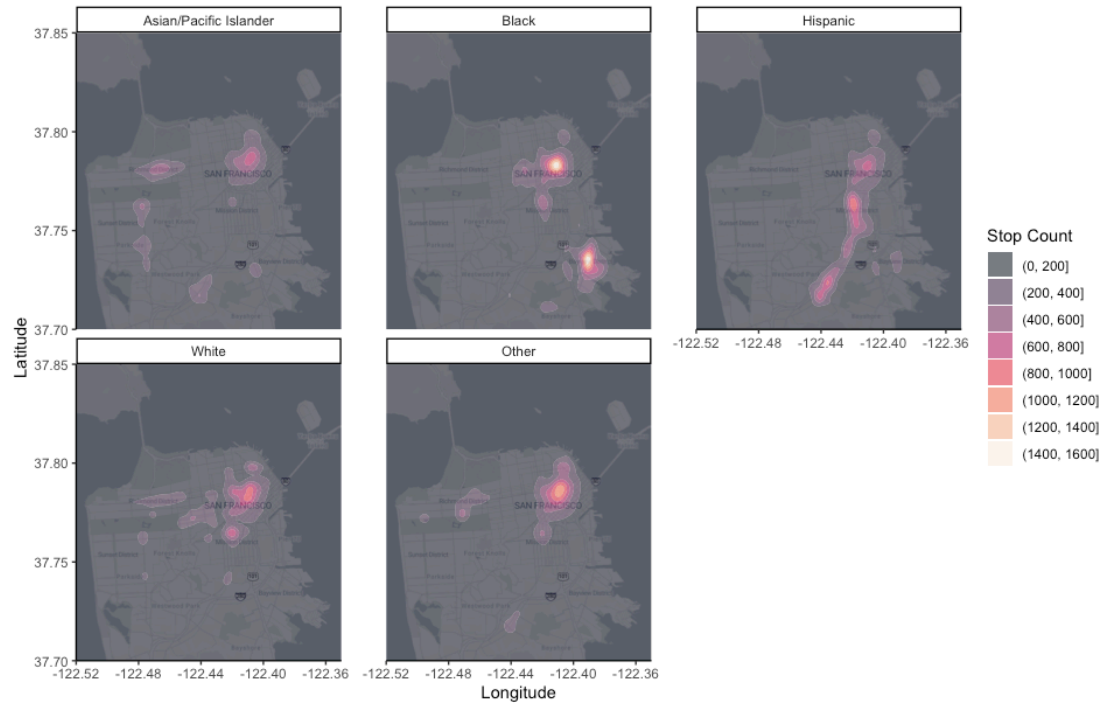
```
sf_map = get_stadiamap(bbox = c(left = -122.52,
                                bottom = 37.70,
                                right = -122.35,
                                top = 37.85),
                      zoom = 12,
                      maptype = "alidade_smooth")

stop.map.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.5 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
```

Fig.5 Police Stops in San Francisco

Facetted by Race



Source: Stanford Open Policing Project

In the final figure, Figure 5, I showed the locational aspect of the traffic stops. 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).

P.(d) [30 points]

Using the statistical techniques learned in class, perform the statistical analysis to find answers for your research questions. What conclusions can you draw from this analysis? Make sure you share your findings related to the problem not in terms of statistical terminology.

Hit (Search Success) Rate Test

Filtering the data by searches


```
searches = traffic.data %>%
  filter(search_conducted == 1) %>%
  mutate(subject_race = factor(subject_race),
         found_contraband = as.integer(contraband_found))
```

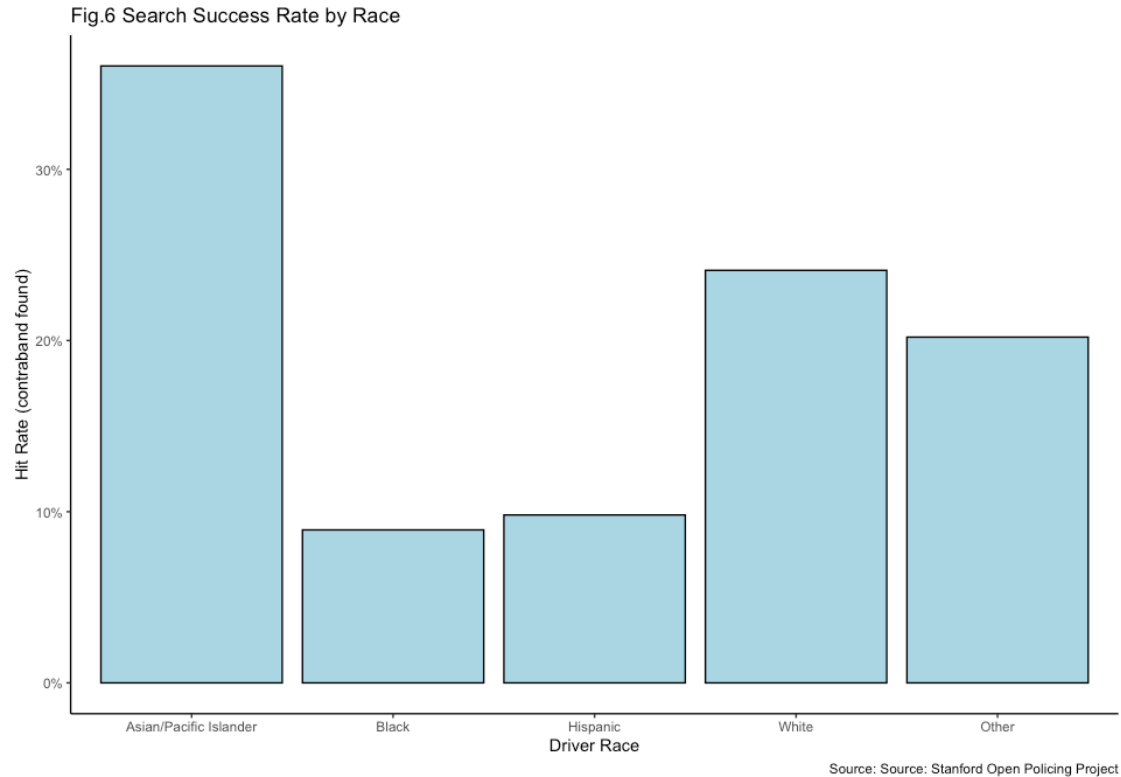


```
# Calculating the success rates
hit_rates = searches %>%
  group_by(subject_race) %>%
  summarise(n_searches = n(),
            n_hits = sum(found_contraband),
            hit_rate = n_hits / n_searches)

hitRatePlot = ggplot(data = hit_rates,
                     aes(x = subject_race,
                         y = hit_rate)) +
  geom_col(fill = "lightblue",
           color = "black") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Driver Race",
       y = "Hit Rate (contraband found)",
       title = "Fig.6 Search Success Rate by Race",
       caption = "Source: Source: Stanford Open Policing Project") +
  theme_classic()

hitRatePlot
```





Hit Rate Test aims to look at if the searches had positive outcome at a similar rate across different races. Pierson et al. (2020) applied the Hit Rate test to multiple counties and looked at the general picture across the USA, while as my research project focuses on the San Francisco data, I will apply the hit rate test only to my dataset.

From this plot, we can see that Black and Hispanic drivers have a significantly less hit rate compared to the White and Asian/Pacific Islander drivers. This means that for every 100 Black drivers that were stopped and searched, less than 10 of them actually had an illegal possession while this number is over 30 for Asian Drivers. This shows that officers do not necessarily have a similar threshold (in other words, min. amount of suspicion) for drivers from any race and they employ a negative bias towards Black and Hispanic drivers in terms of searching the stopped drivers.

Modeling Outcome Likelihoods Using Demographics

While Xu et al. (2024) used binary logistic regression to guess if the traffic stop would be considered as legit or not, I will use binary logistic regression to see if there is a relationship between the stops that involve a search and the drivers race, sex, and age.

In the second part, I will examine the relationship between if the traffic stop is resulted in an arrest or not is related to the race, sex, and / or age of the driver.

Before starting the modeling, I will remove the NA values of different variables.

```
model.data = traffic.data %>%  
  filter(!is.na(subject_race), !is.na(subject_sex), !  
is.na(subject_age))
```

At the same time, this is a simple function I created that plots the ROC and show the AUC value for the logistic regression models. nice!

```
plotROC = function(model,  
  data,  
  responseVar,  
  figNum) {  
  predicted_probs = predict(model, type = "response")  
  roc_obj = roc(data[[responseVar]], predicted_probs)  
  auc_value = auc(roc_obj)  
  
  ggroc(data = roc_obj,  
    size = 1.5) +  
    labs(title = paste0('Fig.',  
      figNum,  
      ' ROC Curve (AUC = ',  
        round(auc_value, 3), ')'),  
      subtitle = paste0('for Logistic Regression Predicting ',  
        responseVar),  
      x = "Specificity",  
      y = "Sensetivity") +  
    theme_classic()  
}
```

Modeling Search Likelihood Using Subject Demographics

I can fit a logistic regression model where race, sex and age are predictor variables while search_conducted (1/0) is the response variable

```
search_model = glm(search_conducted ~  
  subject_race + subject_sex + subject_age,  
  data = model.data,  
  family = binomial)
```

```
summary(search_model)
```

```
##
## Call:
## glm(formula = search_conducted ~ subject_race + subject_sex +
##      subject_age, family = binomial, data = model.data)
##
## Coefficients:
##              Estimate Std. Error z value      Pr(>|
z|)
## (Intercept)      -2.5644228   0.0243909  -105.14
<0.00000000000000002 ***
## subject_raceBlack    2.2195404   0.0211630   104.88
<0.00000000000000002 ***
## subject_raceHispanic  1.6557689   0.0223232    74.17
<0.00000000000000002 ***
## subject_raceWhite    0.6043566   0.0220378    27.42
<0.00000000000000002 ***
## subject_raceOther    0.6193931   0.0263268    23.53
<0.00000000000000002 ***
## subject_sexFemale   -0.7702888   0.0126126   -61.07
<0.00000000000000002 ***
## subject_age         -0.0337612   0.0004242   -79.58
<0.00000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 367945  on 805839  degrees of freedom
## Residual deviance: 326183  on 805833  degrees of freedom
## AIC: 326197
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(search_model))
```

```
##              (Intercept)  subject_raceBlack subject_raceHispanic
##              0.07696359      9.20310028      5.23710519
## subject_raceWhite  subject_raceOther  subject_sexFemale
##              1.83007430      1.85780019      0.46287938
```

```
##          subject_age
##          0.96680238
```

When we look at the odds ratios for this model, we can see valuable insights for both race and the other demographics. When we look at the odds ratios for the races

- Black drivers have a 9.2 times, complete the sentences 9.2 times chance of what?
- Hispanic drivers have a 5.2 times;
- White drivers have a 1.8 times;
- and drivers from other races have 1.8 times higher odds to be searched compared to the Asian/Pacific Islander drivers.

When we look at sex, we can see that female drivers have 46% of the odds of the male drivers, which means that male drivers are 2.17 times more likely to get searched compared to the female drivers.

Lastly, when we look at the odds ratio for age, we can see that for each additional year in age, the odds of getting searched decreases 3.3%, which is around 30% over a decade.

In general, we can see that male, young, drivers from specific age groups show a high likelihood of arrest.

```
search_0 = glm(search_conducted ~ 1,
               data = model.data,
               family = binomial)
```

```
anova(search_model, search_0, test = "Chisq")
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: search_conducted ~ subject_race + subject_sex +
subject_age
```

```
## Model 2: search_conducted ~ 1
```

```
##   Resid. Df Resid. Dev Df Deviance              Pr(>Chi)
```

```
## 1      805833      326183
```

```
## 2      805839      367945 -6    -41762 < 0.000000000000000022 ***
```

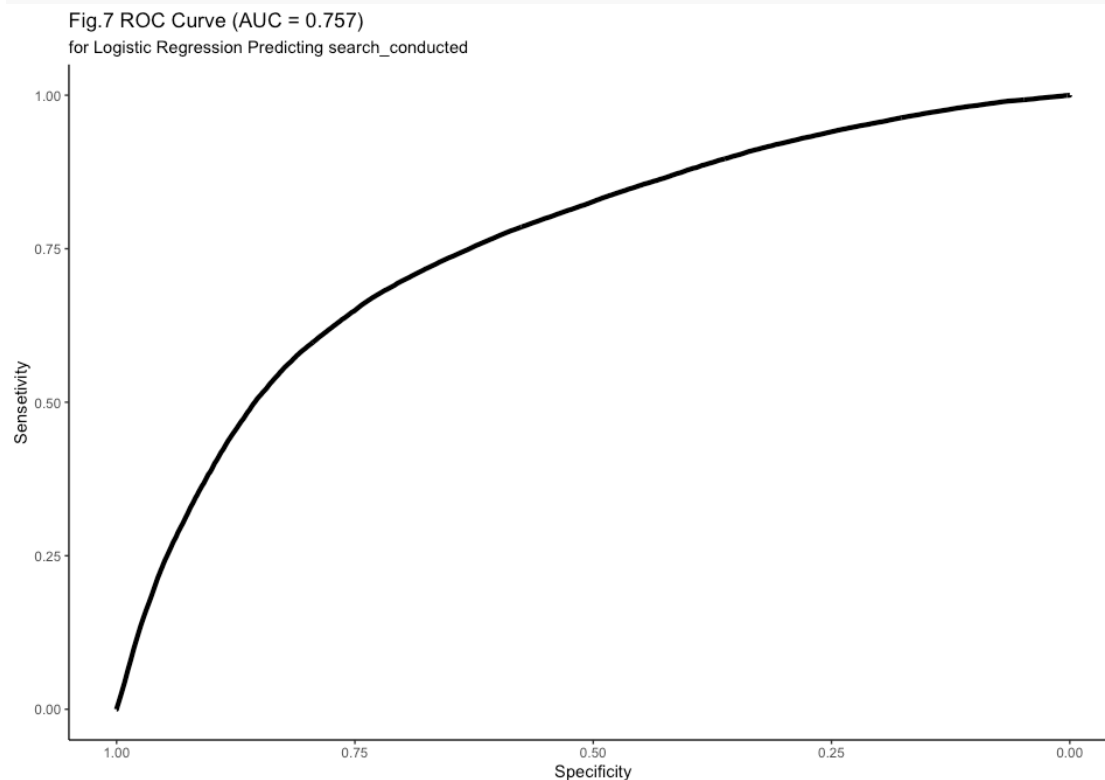
```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this Chi-square test, I compared my full model (with race, sex, and age as predictors) against a null model. We can see that p-value is less than 2×10^{-16} , which means that the full model explains significantly more variability in search outcomes than randomly guessing. In other words, we can say they *at least* some of my predictors are significant.

Following the interpretation of the coefficients, we can plot the ROC curve and also see the AUC value.

```
plotROC(search_model, model.data, "search_conducted", 7)
```



The AUC value shows how well my model can separate these binary classes of Search conducted and not conducted. An AUC value of 0.757 means that the model is able to make a correct prediction for 75.7% of the data, which is considered an acceptable prediction accuracy.

Modeling Arrest Likelihood Using Subject Demographics

In this part, I fitted a logistic regression model where race, sex and age are predictor variables while arrest_made (1/0) is the response variable

```

arrest_model = glm(arrest_made ~
                    subject_race + subject_sex + subject_age,
                    data = model.data,
                    family = binomial)

summary(arrest_model)

##
## Call:
## glm(formula = arrest_made ~ subject_race + subject_sex +
##      subject_age,
##      family = binomial, data = model.data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|
z|)
## (Intercept)      -4.0704790   0.0413006  -98.557
<0.00000000000000002 ***
## subject_raceBlack    1.0175940   0.0348754   29.178
<0.00000000000000002 ***
## subject_raceHispanic  0.8490327   0.0371797   22.836
<0.00000000000000002 ***
## subject_raceWhite    0.3249509   0.0340696    9.538
<0.00000000000000002 ***
## subject_raceOther    0.0447622   0.0458833    0.976
0.329
## subject_sexFemale   -0.4564452   0.0237338  -19.232
<0.00000000000000002 ***
## subject_age         -0.0164734   0.0007826  -21.051
<0.00000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 115529  on 805839  degrees of freedom
## Residual deviance: 112892  on 805833  degrees of freedom
## AIC: 112906
##
## Number of Fisher Scoring iterations: 7

exp(coef(arrest_model))

```

```
##           (Intercept)      subject_raceBlack subject_raceHispanic
##           0.01706921          2.76653034          2.33738491
##      subject_raceWhite      subject_raceOther      subject_sexFemale
##           1.38396267          1.04577918          0.63353170
##           subject_age
##           0.98366158
```

When we look at the odds-ratios, we can see that Asian/Pacific Islander is the base category. If we analyze the values for the other races,

- Black drivers are 2.8 times;
- Hispanic drivers are 2.3 times;
- White drivers are 1.4 times;
- and Other drivers are 1.05 times more likely to get arrested compared to the Asian\Pacific islander drivers.

It is important to note that Other category in race is not statistically significant ($p \approx 0.33$), indicating there is not meaningful difference in the odds of arrest for Other category compared to the Asian/Pacific Islander drivers. All other categories are statistically significant showing that race is a strong predictor of arrest odds.

We can see that sex and age also have an effect on the odds of getting arrested. When we look at the odds ratio for female, we can see that female drivers have about the 63% of odds of getting arrested of the male drivers ($1/0.63=1.58$) who are 1.58 times more likely to get arrested compared to the female drivers.

Lastly, each additional year of age multiplies the odds by 0.98 which is a nearly 2% decrease in odds per year per year. Over a 10-year period, this is a 14% reduction in the odds.

In general, this results shows that younger, male drivers of certain race groups show a higher likelihood of arrest.

```
arrest_0 = glm(arrest_made ~ 1,
               data = model.data,
               family = binomial)

anova(arrest_model, arrest_0, test = "Chisq")

## Analysis of Deviance Table
##
```

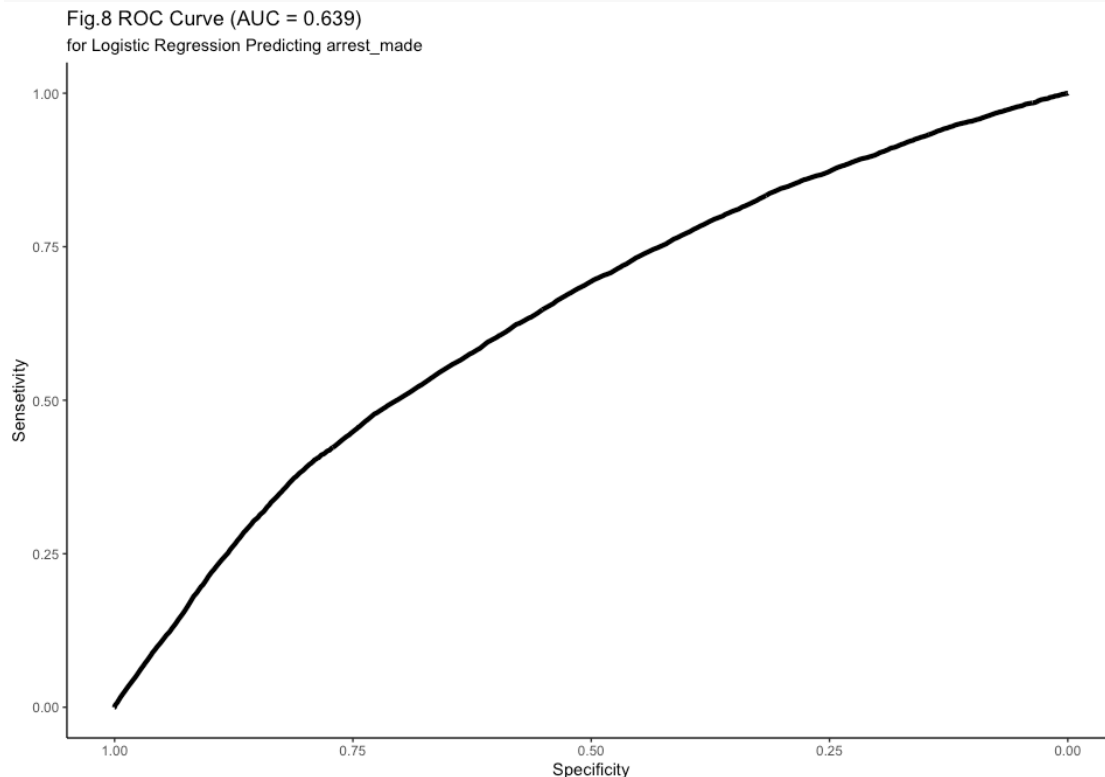


```
## Model 1: arrest_made ~ subject_race + subject_sex + subject_age
## Model 2: arrest_made ~ 1
##   Resid. Df Resid. Dev Df Deviance          Pr(>Chi)
## 1      805833      112892
## 2      805839      115529 -6  -2636.5 < 0.000000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the Chi-square test, I compared my full model (with race, sex, and age as predictors) against a null model. We can see that p-value is less than 2.2×10^{-16} , which means that the full model explains significantly more variability in arrest outcomes than randomly guessing. In other words, we can say they *at least* some of my predictors are significant.

After the interpretation of the coefficients, we can plot the ROC curve and also see the AUC value.

```
plotROC(arrest_model, model.data, "arrest_made", 8)
```



The AUC value shows how well my model can separate these binary classes of Arrest Made and not made. An AUC value of 0.639 means that the model is able to make a

correct prediction for 63.9% of the data, which is considered poor/fair prediction accuracy.

Veil-of-Darkness Test

This test was originally developed by Grogger & Ridgeway (2005) to detect racial bias in the traffic stops. It comes from the assumption that if an officer is racially biased, it would not be able to do so after a certain time of the day. I was not able to get any significant results showing bias as a result of this test, thus I am keeping this outside of the original flow of narrative.

I used a dataset I generated using a Python script that utilizes the [Astral](#) library. I created a dataset of Sunrise and Sunset times using this library. The following cell is the python script I used. It is not going to run during knitting.

```
"""
Script for generation a dataset of sunset times in San Francisco
from Janury 1st, 2007 to December 31st, 2015
"""

import csv
from datetime import date, timedelta
from astral import LocationInfo
from astral.sun import sun
import pytz

# Setting the city information
city = LocationInfo("San Francisco", "USA", "America/Los_Angeles",
37.7749, -122.4194)

# Set the date range
start_date = date(2007, 1, 1)
end_date = date(2015, 12, 31)

# Prepare the CSV file
with open('san_francisco_sunrise_sunset_2007_2015.csv', 'w',
newline='') as csvfile:
    fieldnames = ['Date', 'Sunrise', 'Sunset']
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
    writer.writeheader()

    current_date = start_date
    while current_date <= end_date:
```

```

    # Get the sunrise and sunset times
    s = sun(city.observer, date=current_date,
tzinfo=pytz.timezone(city.timezone))
    sunrise_time = s['sunrise']
    sunset_time = s['sunset']

    writer.writerow({
        'Date': current_date.strftime('%Y-%m-%d'),
        'Sunrise': sunrise_time.strftime('%H:%M:%S'),
        'Sunset': sunset_time.strftime('%H:%M:%S')
    })

    current_date += timedelta(days=1)

print("CSV file 'san_francisco_sunrise_sunset_2007_2015.csv' has been
created.")

```

At the first part, I read the dataset and processed the Sunrise Sunset and Date columns.

```

time.data = read.csv("san_francisco_sunrise_sunset_2007_2015.csv")

time.data = time.data %>%
  mutate(Sunrise = as.POSIXct(paste(Date, Sunrise)),
         Sunset = as.POSIXct(paste(Date, Sunset)),
         date = as.Date(Date))

```

Following this, I merged this smaller dataset with the traffic data I have and created a variable that determines if it is currently dark or not (time between sunset and the sunrise).

```

traffic.data = traffic.data %>%
  left_join(time.data, by = "date") %>%
  mutate(datetime = as.POSIXct(paste(format(date, "%Y-%m-%d"),
                                       format(time, "%H:%M:%S")),
                                       format = "%Y-%m-%d %H:%M:%S"),
         isDark = datetime < Sunset | datetime > Sunset)

```

I then created a new dataframe that focuses on the 120 minutes around the sunset, and created bins of 10 minutes.

```

dark.data = traffic.data %>%
  mutate(sunset_diff = as.numeric(difftime(datetime, Sunset, units =
"mins")) %>%
  filter(sunset_diff >= -90, sunset_diff <= 60) %>%
  mutate(bin = floor(sunset_diff / 5) * 5) %>%
  group_by(bin) %>%
  summarize(pct_black = (sum(subject_race == "Black", na.rm =
TRUE) / n()) * 100,
            n = n(),
            .groups = "drop")

```

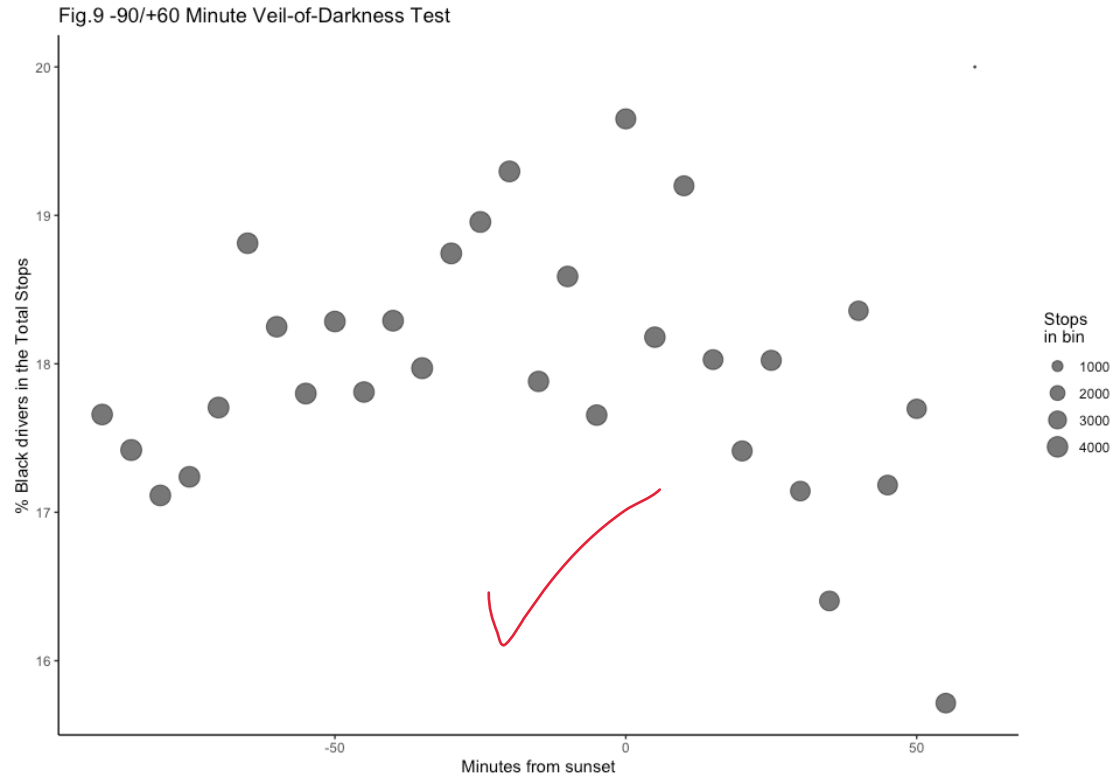
Lastly, I plotted the stops

```

vod_plot = ggplot(data = dark.data,
                  aes(x = bin,
                     y = pct_black)) +
  geom_point(aes(size = n),
            alpha = 0.6) +
  scale_size_area("Stops\nin bin") +
  labs(title = "Fig.9 -90/+60 Minute Veil-of-Darkness Test",
       x = "Minutes from sunset",
       y = "% Black drivers in the Total Stops") +
  theme_classic()

```

vod_plot



From this plot, it is not possible to see the exact trend in the data in terms of the decreasing amount of percentage of black drivers stopped among the total number of drivers are stopped. I am unsure about the reason but this might be because of my data being limited to San Francisco, and not enough for catching a general trend.

Ok! good attempt

Resources

How to make tables in R → [Source](#)

How to reshape the data from long to wide → [Source](#)

Date and time in R → [Source](#)

References

Grogger, J., & Ridgeway, G. (2005). *Testing for racial profiling in traffic stops from behind a veil of darkness* (Working Papers 0507). Harris School of Public Policy Studies, University of Chicago. <https://EconPapers.repec.org/RePEc:har:wpaper:0507>

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>



Excellent!