# Does a person's characteristics affect the amount of time they spend in jail?

By: Yogi Patel

## Introduction

All throughout the country, people of color face struggles whether it be with lack of opportunities or unfair treatment. This project will explore how race and other attributes of characteristics affect one's arrest details, specifically how people of color generally have different arrest details than white individuals. Overall, I will use modeling techniques and visualizations to display trends seen in society regarding inequality with local data from our college town of Champaign, IL.

# Project and Data

This project will be using the dataset from the Champaign County Sheriff Office that has data about the demographics of the jail bookings of those who were arrested in the area between the years 2011-2016. Although this dataset includes bookings from various states and cities, we will focus on arrest bookings throughout the years in the town of Champaign, IL.

# Data Cleaning

In this dataset, Champaign is misspelled several times. Since these are considered data errors that can provide valuable information to this project, I clean the data by correcting these misspellings. First, I filter the data to focus on bookings on Illinois and then use regex to to select mispellings and correct them using a condition. After this, I remove columns with missing values in the columns we will be testing for significance.

```
library(tidyverse)

CCSO <- read_csv("https://urldefense.com/v3/__https://uofi.box.com/shared/static/9elozjs
    col_types = cols(`BOOKING DATE` = col_date(format = "%m/%d/%Y"),
        `RELEASED DATE` = col_date(format = "%m/%d/%Y")))

CCSO <- CCSO %>%
  filter(STATE=='ILLINOIS') %>%
  mutate(CITYtemp=ifelse(str_detect(CITY, "^CH\\w+P|^CHA\\w+I"), "CHAMPAIGN", CITY)) %>%
  select(-CITY) %>%
  rename(CITY=CITYtemp) %>%
  filter(CITY=="CHAMPAIGN")

CCSO <- CCSO[!(is.na(CCSO$RACE)), ]
CCSO <- CCSO[!(is.na(CCSO$`EMPLOYMENT STATUS`)), ]
CCSO <- CCSO[!(is.na(CCSO$`Age at Arrest`)), ]
CCSO <- CCSO[!(is.na(CCSO$`Days in Jail`)), ]
CCSO <- CCSO[!(is.na(CCSO$SEX)), ]
```

# Data Modeling : Linear Regression and F tests

The first step in my project is to model the data and decide which linear regression model and variables are significant in predicting the outcome. Specifically, I explore which combination of variables creates the best model in predicting how many days an individual will spend in jail. I do this by using the lm() function with different variables.

In this case the test will focus on these hypotheses:
Null Hypothesis: The first model is adequate in predicting the response variable (Age at

Arrest is adequate in predicting the amount of time spent in jail)
Alternative Hypothesis: The additive model is better in predicting the response variable (The model with additional variables better predicts the amount of time spent in jail)

Then using multiple F tests, I compare each model with the previous one to see which regression model best predicts the days spent in jail. We expect the model with all the variables to be significant as we expect that race, age at arrest, sex, and employment status are all attributes that affects time spent in jail. I conduct these F tests by using the anova() function which tests how the dependent variable of "Days in Jail" changes according to the models with added independent variables.

```
mod1=lm(`Days in Jail`~ `Age at Arrest`, data=CCSO)
mod2=lm(`Days in Jail`~ `Age at Arrest` + RACE, data=CCSO)
mod3=lm(`Days in Jail`~ `Age at Arrest` + RACE + `EMPLOYMENT STATUS`, data=CCSO)
mod4=lm(`Days in Jail`~ `Age at Arrest` + RACE + `EMPLOYMENT STATUS` + `SEX`, data=CCSO)

anova(mod1, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: `Days in Jail` ~ `Age at Arrest`
## Model 2: `Days in Jail` ~ `Age at Arrest` + RACE
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  25961 58086095
## 2  25955 57250798  6    835296 63.114 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod2, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: `Days in Jail` ~ `Age at Arrest` + RACE
## Model 2: `Days in Jail` ~ `Age at Arrest` + RACE + `EMPLOYMENT STATUS`
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  25955 57250798
## 2  25949 56557927  6    692871 52.982 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod3, mod4)
```

```
## Analysis of Variance Table
##
## Model 1: `Days in Jail` ~ `Age at Arrest` + RACE + `EMPLOYMENT STATUS`
## Model 2: `Days in Jail` ~ `Age at Arrest` + RACE + `EMPLOYMENT STATUS` +
##     SEX
```

```
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  25949 56557927
## 2  25948 55701945  1    855982 398.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Data Modeling Results

From the analysis of variance tables, each P value for the F test between models explains which model is better. As shown, the P value for each model comparisons is less than alpha=0.05. This means that we are able to reject the null hypothesis and can say the additive model is better for each model. Overall, using backward elimination, the model with all variables is significant and the best fit in predicting the response of days spent in jail for each booking.

# Data Modeling Follow-up

Following the model selection, I conduct a Tukey Test which measures the significant difference between the attributes in predicting a specific outcome. By using the TukeyHSD() function, I can display how each race has a difference in predicting how many days one spends in jail.

```r
TukeyHSD(aov(`Days in Jail`~RACE, data=CCSO))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = `Days in Jail` ~ RACE, data = CCSO)
##
## $RACE
##                                                 diff         lwr        upr
## Black-Asian/Pacific Islander               17.8883702    9.699415  26.077325
## Hispanic-Asian/Pacific Islander             7.0747358   -1.706617  15.856088
## Native American-Asian/Pacific Islander     11.4358835  -21.360960  44.232727
## Unknown-Asian/Pacific Islander             -1.4943224  -23.471101  20.482456
## White-Asian/Pacific Islander                6.2853315   -2.002145  14.572808
## White (Hispanic)-Asian/Pacific Islander    -4.0378007 -142.783468 134.707866
## Hispanic-Black                            -10.8136345  -14.323528  -7.303741
## Native American-Black                      -6.4524867  -38.246199  25.341226
## Unknown-Black                             -19.3826927  -39.832290   1.066904
## White-Black                               -11.6030387  -13.575215  -9.630862
## White (Hispanic)-Black                    -21.9261709 -160.438146 116.585804
## Native American-Hispanic                    4.3611478  -27.590274  36.312569
## Unknown-Hispanic                           -8.5690582  -29.262999  12.124882
## White-Hispanic                             -0.7894043   -4.523389   2.944581
```

```
## White (Hispanic)-Hispanic             -11.1125364 -149.660797 127.435724
## Unknown-Native American               -12.9302059  -50.702682  24.842271
## White-Native American                  -5.1505520  -36.969783  26.668679
## White (Hispanic)-Native American      -15.4736842 -157.579776 126.632407
## White-Unknown                           7.7796539  -12.709595  28.268902
## White (Hispanic)-Unknown               -2.5434783 -142.548789 137.461833
## White (Hispanic)-White                -10.3231322 -148.840967 128.194703
##                                           p adj
## Black-Asian/Pacific Islander          0.0000000
## Hispanic-Asian/Pacific Islander       0.2090052
## Native American-Asian/Pacific Islander  0.9476951
## Unknown-Asian/Pacific Islander        0.9999946
## White-Asian/Pacific Islander          0.2760775
## White (Hispanic)-Asian/Pacific Islander 1.0000000
## Hispanic-Black                        0.0000000
## Native American-Black                 0.9968804
## Unknown-Black                         0.0766718
## White-Black                           0.0000000
## White (Hispanic)-Black                0.9992326
## Native American-Hispanic              0.9996732
## Unknown-Hispanic                      0.8863210
## White-Hispanic                        0.9960893
## White (Hispanic)-Hispanic             0.9999856
## Unknown-Native American               0.9520886
## White-Native American                 0.9991282
## White (Hispanic)-Native American      0.9999126
## White-Unknown                         0.9223861
## White (Hispanic)-Unknown              1.0000000
## White (Hispanic)-White                0.9999907
```

# Data Modeling Follow-up Results

The tukey test results shows us the existence of differences between each race. From the results, it is revealed how the most significant differences are between groups Black-Asian/Pacific Islander, Hispanic-Black, and White-Black. There is a significant difference between Black and each race and we know this as the P value is less than an alpha value of 0.05. From this we can say that the difference supports that a booking of a individual of race Black typically predicts highers days in jail.

# Data Visualizations

Now that it is determined which variables are significant in determining the arrest details, I will display various visualizations that will support the model chosen. These visualizations will reveal how the effect of each variable changes the days spent in jail. In general, focusing on how inequality exists whether it comes to race, gender, age, employment status, and even more.

## Visualization 1:

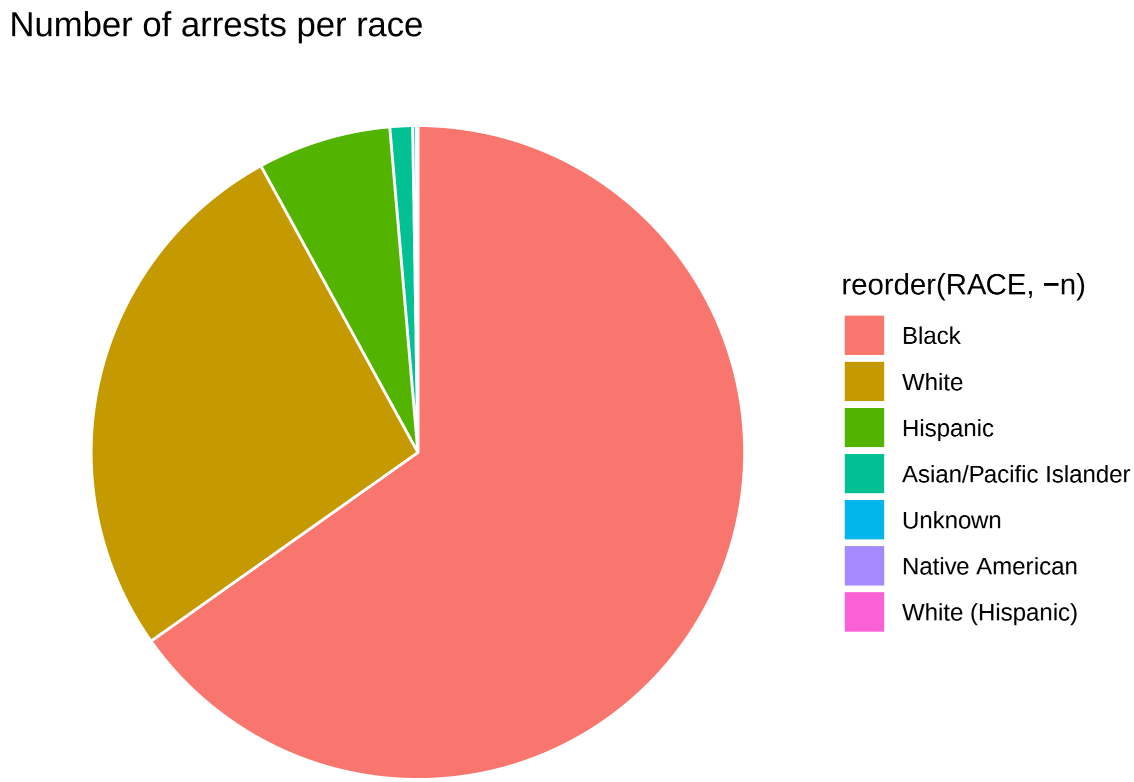
First, I begin by filtering the dataset to count the frequency of each race's bookings.

```
CCSOrace <- CCSO %>%
  count(RACE)

CCSOrace
```

```
## # A tibble: 7 x 2
##   RACE                       n
##   <chr>                  <int>
## 1 Asian/Pacific Islander   291
## 2 Black                  16931
## 3 Hispanic                1715
## 4 Native American           19
## 5 Unknown                   46
## 6 White                   6960
## 7 White (Hispanic)           1
```

**Pie chart:** This pie chart displays the number of arrests per race using the filtered dataset above.

```
ggplot(data = CCSOrace, aes(x = "", y = -n, fill = reorder(RACE, -n))) +
  geom_bar(stat = "identity", color="white") +
  coord_polar("y", start=0) + theme_void() + ggtitle("Number of arrests per race")
```

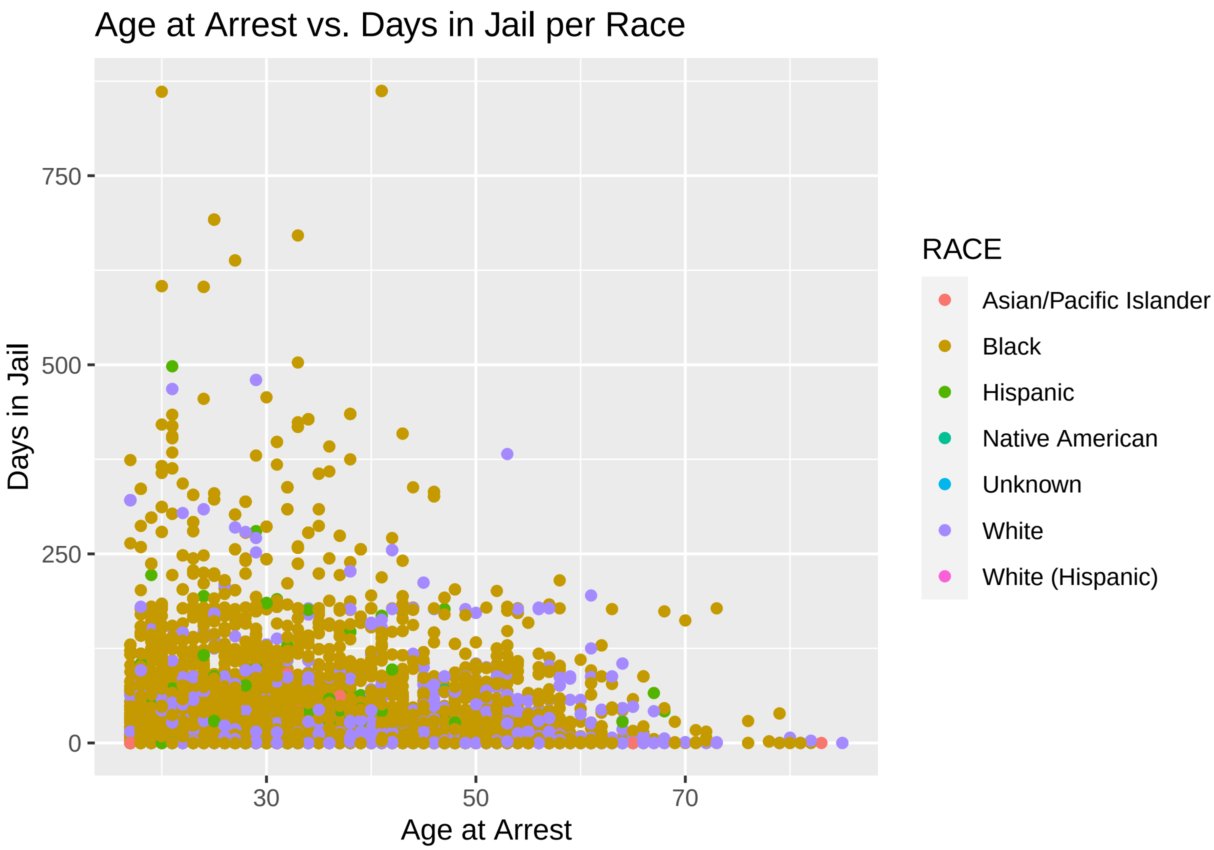



**Pie Chart Analysis:** With over 65% of the data, the race with the highest number of arrests is Black. This supports the idea that race is a main variable that affects one's arrest details.

### Visualization 2:

**Scatterplot:** This scatter plot displays the age at arrest vs the days in jail for each booking colored by race.

```
ggplot(data=CCSO, aes(x=`Age at Arrest`, y=`Days in Jail`, col=RACE)) + geom_point() +
  ggtitle("Age at Arrest vs. Days in Jail per Race")
```




**Scatterplot Analysis:** The data displays how Black individuals not only account for most of the arrests, but typically spend more days in jail than White individuals or other races. This exemplifies how race has an impact on the arrest attributes such as what age they are arrested and how long they spend in jail.

## Visualizations 3 & 4:

Here I filter the dataset to find the average days spent in jail and the average age at arrest for each race.

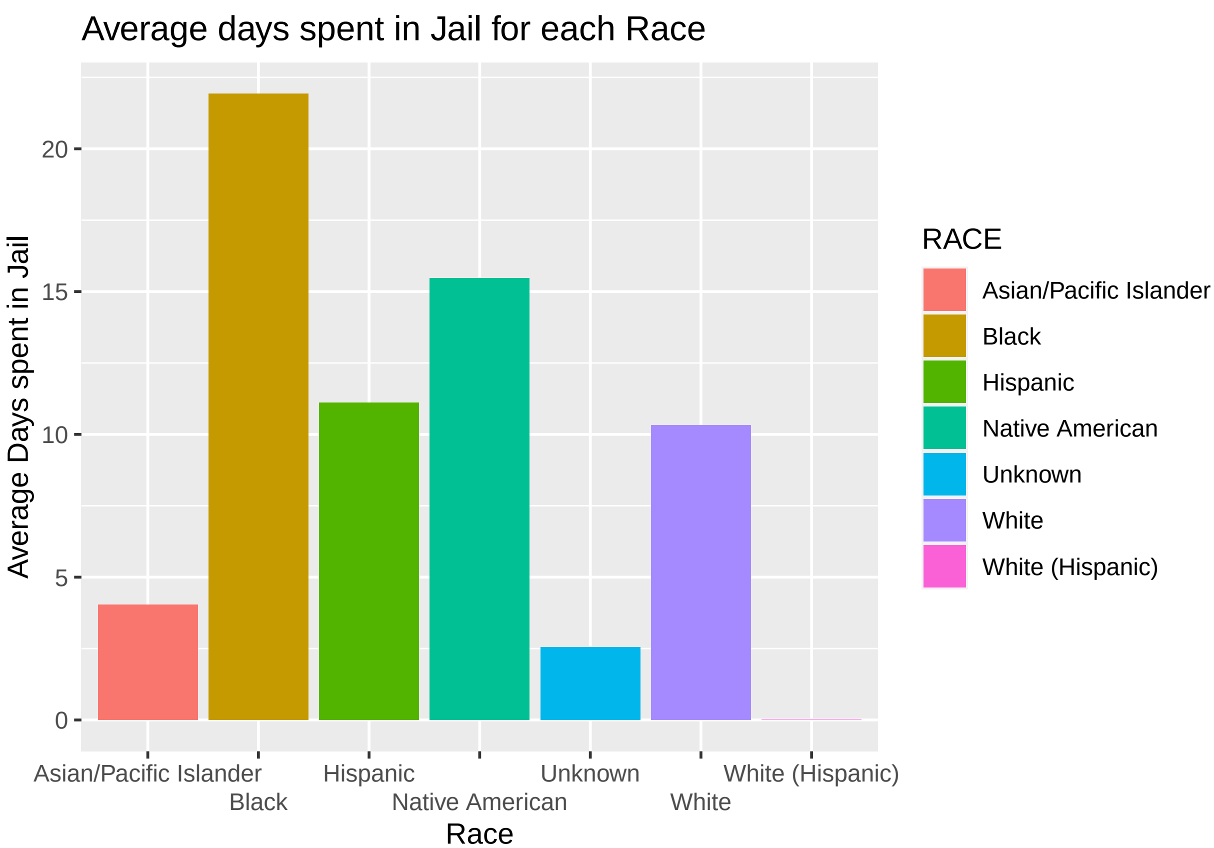
```
CCSOav <- CCSO %>%
  group_by(RACE) %>%
  summarise(avDays = mean(`Days in Jail`), avAge = mean(`Age at Arrest`))

CCSOav
```

```
## # A tibble: 7 x 3
##   RACE                   avDays avAge
##   <chr>                   <dbl> <dbl>
## 1 Asian/Pacific Islander   4.04  25.7
## 2 Black                   21.9   29.9
## 3 Hispanic                11.1   28.8
## 4 Native American         15.5   33.8
```

```
## 5 Unknown              2.54   27.5
## 6 White               10.3    32.0
## 7 White (Hispanic)     0      22
```

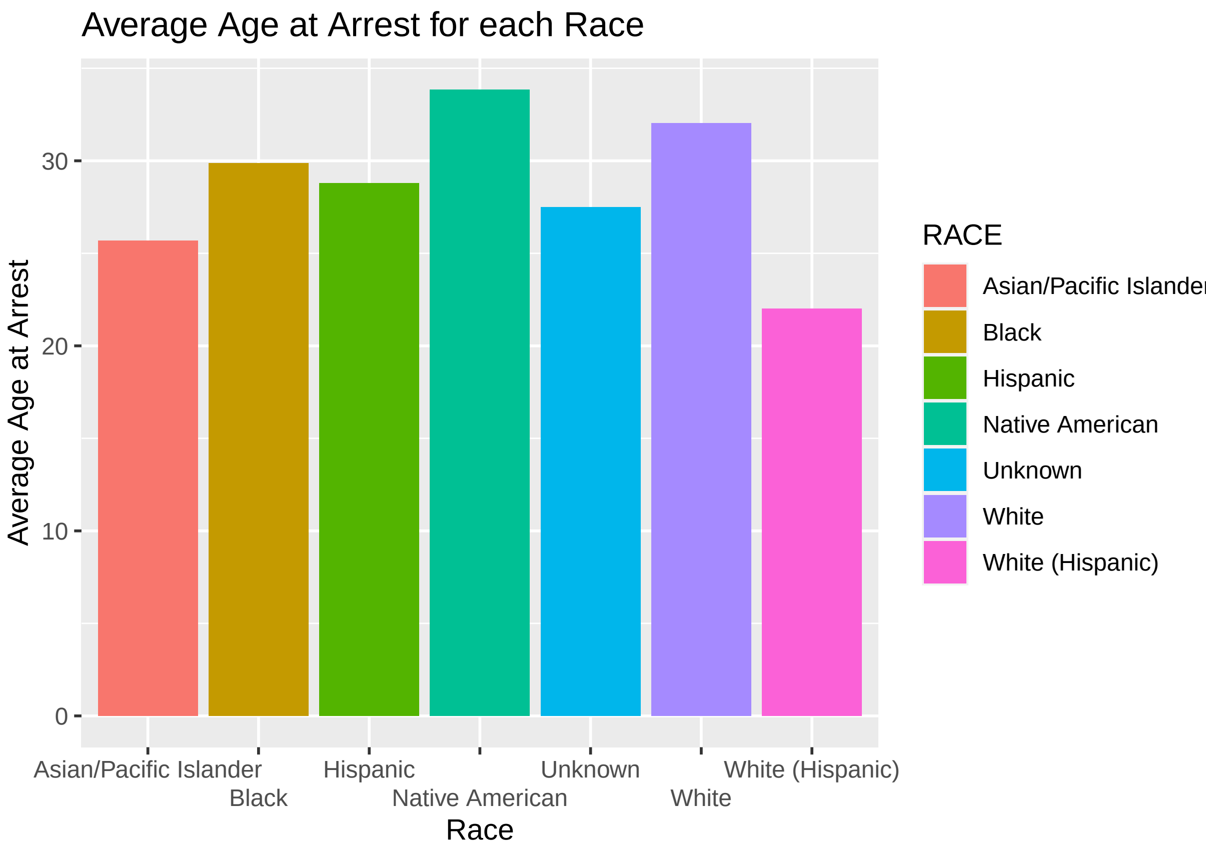
**Bar Charts:** This chart displays the race of an individual along with the average number of days that particular race spent in jail.

```
ggplot(data=CCSOav, aes(x=RACE, y=avDays, fill=RACE)) +
    geom_col() + ggtitle("Average days spent in Jail for each Race") + xlab("Race") +
  ylab("Average Days spent in Jail") + scale_x_discrete(guide = guide_axis(n.dodge=2))
```




This chart displays the race of an individual along with the average age that particular race was arrested.

```
ggplot(data=CCSOav, aes(x=RACE, y=avAge, fill=RACE)) +
    geom_col() + ggtitle("Average Age at Arrest for each Race") + xlab("Race") +
  ylab("Average Age at Arrest") + scale_x_discrete(guide = guide_axis(n.dodge=2))
```




**Bar Charts Analysis:**
Bar Chart 1: The data displays how on average, black individuals are one of the races that spent the most amount of days in jail. Not only blacks, but other people of color also spend a greater average of time in jail compared to white individuals. This chart reinforces how race impacts arrest attributes such as affecting how long they spend in jail.
Bar Chart 2: The data displays how on average, people of color (black, hispanic, asian, etc.) were arrested at a younger age than White individuals. This reinforces how race impacts arrest attributes such as affecting what age they are arrested.

## Visualization 5:

For this visualization, I filter the dataset to find the average days spent in jail and the average age at arrest for each sex.

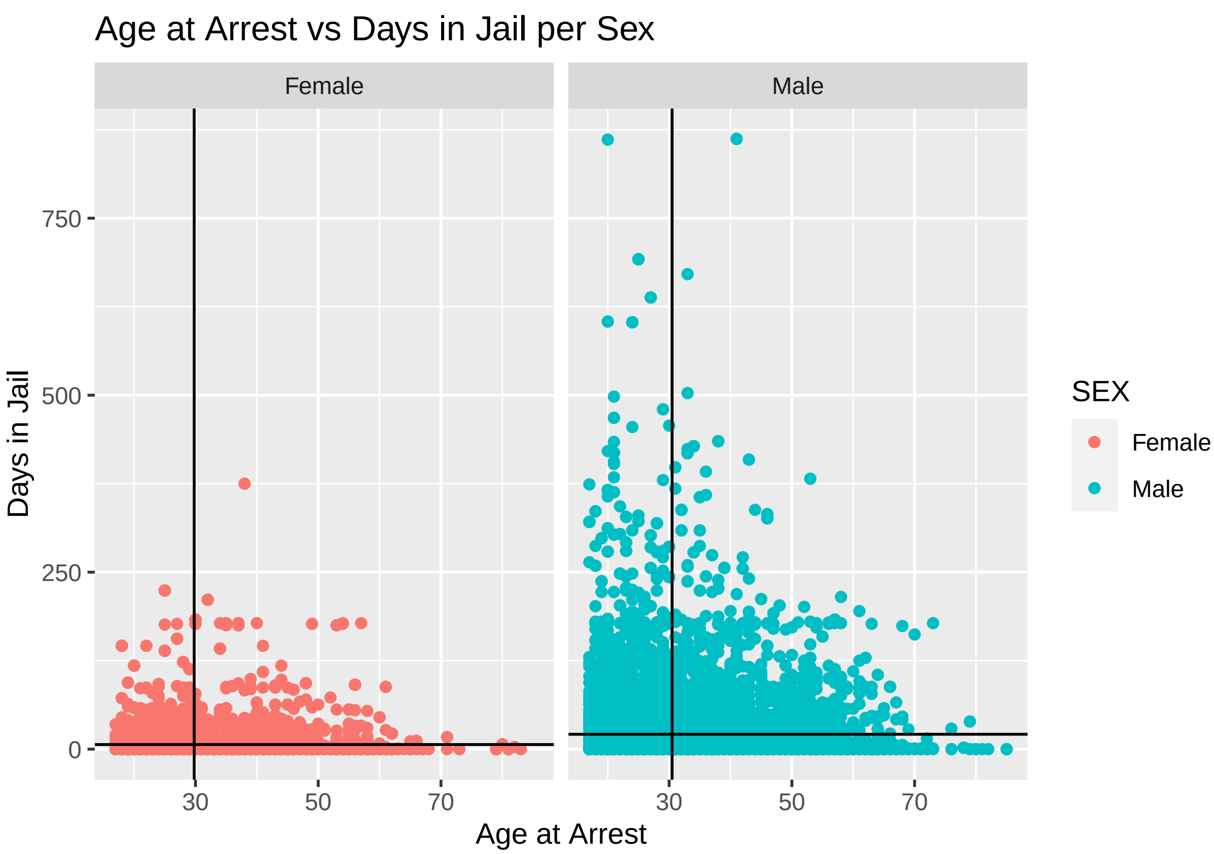
```
CCSOsex <- CCSO %>%
  group_by(SEX) %>%
  summarise(avDays = mean(`Days in Jail`), avAge = mean(`Age at Arrest`))

CCSOsex
```

```
## # A tibble: 2 x 3
##   SEX     avDays avAge
##   <chr>    <dbl> <dbl>
```

```
## 1 Female   6.50  29.8
## 2 Male    21.0   30.5
```

**Facet Scatterplots:** This scatterplot shows age at arrest versus days in jail for different facets of sex. It also displays vertical and horizontal lines for the average of each attribute per group.

```
ggplot(data=CCSO, aes(x=`Age at Arrest`, y=`Days in Jail`, col=SEX)) + geom_point() +
  facet_wrap(~SEX) + geom_hline(data=CCSOsex, aes(yintercept=avDays)) +
  geom_vline(data= CCSOsex, aes(xintercept=avAge)) +
  ggtitle("Age at Arrest vs Days in Jail per Sex")
```

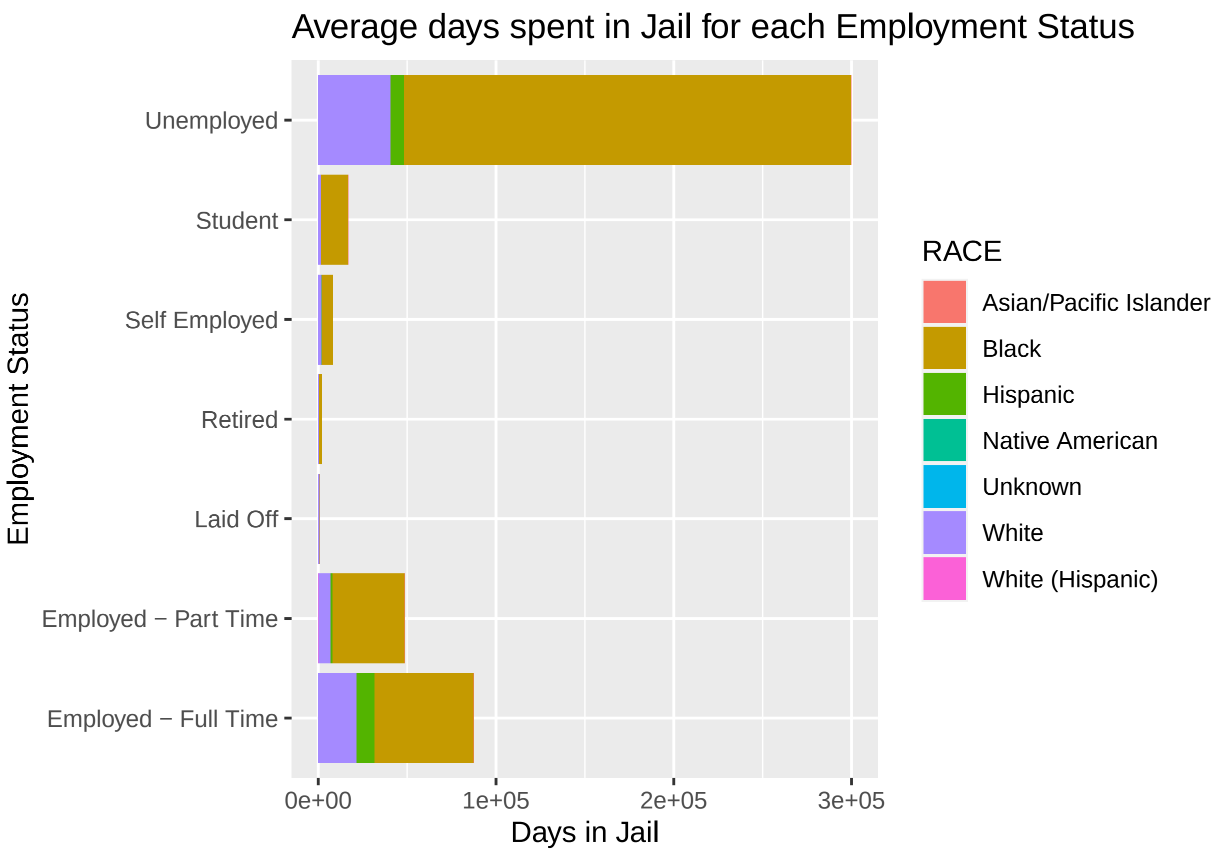



**Facet Scatterplots Analysis:** The plots reveal how females not only have less arrests but also spend an average of less days in jail. Females spend an average of 6.50115 days in jail vs 21.02285 days spent be males. These graphs support how sex is another characteristic that affects arrest details similar to how race does.

## Visualization 6:

**Stacked Row Chart:** This row chart displays the amount of days spent in jail for each type of employment status, stacked by race.

```r
ggplot(CCSO, aes(fill=`RACE`, y=`Days in Jail`, x=`EMPLOYMENT STATUS`)) +
    geom_bar(position="stack", stat="identity") + coord_flip() +
  ggtitle("Average days spent in Jail for each Employment Status") +
  xlab("Employment Status")
```




**Stacked Row Chart Analysis:** This visualization displays how those who are unemployed spend the most time in jail. Also, people of the race black account for most of this group. Race and employment status both affect the days spent in jail significantly as students, employed, retired, and other individuals typically spend less time and account for less of the bookings as well.

## Conclusion:

Through cleaning, modeling, visualizing, and summarizing data, this project exhibits the relationship between a person's characteristics and their arrest details. As seen, the variables that are significant in predicting the amount of days one spends in jail include age at arrest, race, sex, and employment status. Being from specific minorities and groups largely changes the circumstances of arrest and society typically follows these same trends as well. By analyzing this issue in our local college campus town, we can see how inequality exists where we live similar to how we see in the news all over the country.