DATA EXPLORATION PROJECT

James Pettigrew

2022-09-25

```
HS_and_Beyond <- read.table("E:/CSU CHICO FALL 2022/math130/DATA/hsb2.txt", header=TRUE, sep="\t")
```

Introduction

The data I have chosen for my project is titled High School and Beyond and was a longitudal study conducted by NCES National Longitudinal Studies Program. This data comes from the second study conducted by the NCES. The program was established to study the educational, vocational, and personal development of young people, beginning with elementary or high school years and following them over time as they take on adult roles and responsibilities. The three variables i have chosen to focus on in my project are titled science, race, and gender.

The variable gender records whether the individual was male or female. The race variable records the ethnicity of the individual and is a factor variable with four levels: African American, Asian, Hispanic, Caucasian. The Science variable records the corresponding individuals score on the science portion of the questionnaire. My first question is, did more females complete the questionnaire than males did? My second question is which ethnic group contained the least individuals that completed the questionnaire? My third research question is which ethnic group has the highest scores on the science portion of the questionnaire? My fourth question is for each ethnic group which gender has higher scores on the science section in the questionnaire?

My first predict is that more males completed the questionnaire than females. My second prediction is that Asians were the ethnic group with the lowest number of individuals to complete the questionnaire. My third prediction is that the Caucasians were the ethnic group with the highest scores on the science section. My fourth prediction is that females had higher scores on the science section for all ethnic groups.

Univariate Exploration

Gender

```
HS_and_Beyond$gender_recode<- fct_recode(HS_and_Beyond$gender,
                                         "Male" = "male",
                                         "Female" = "female") #converts gender variable to a factor var
```

```
table(HS_and_Beyond$gender_recode) %>% prop.table() * 100 #gives the percentage of males and females in
```

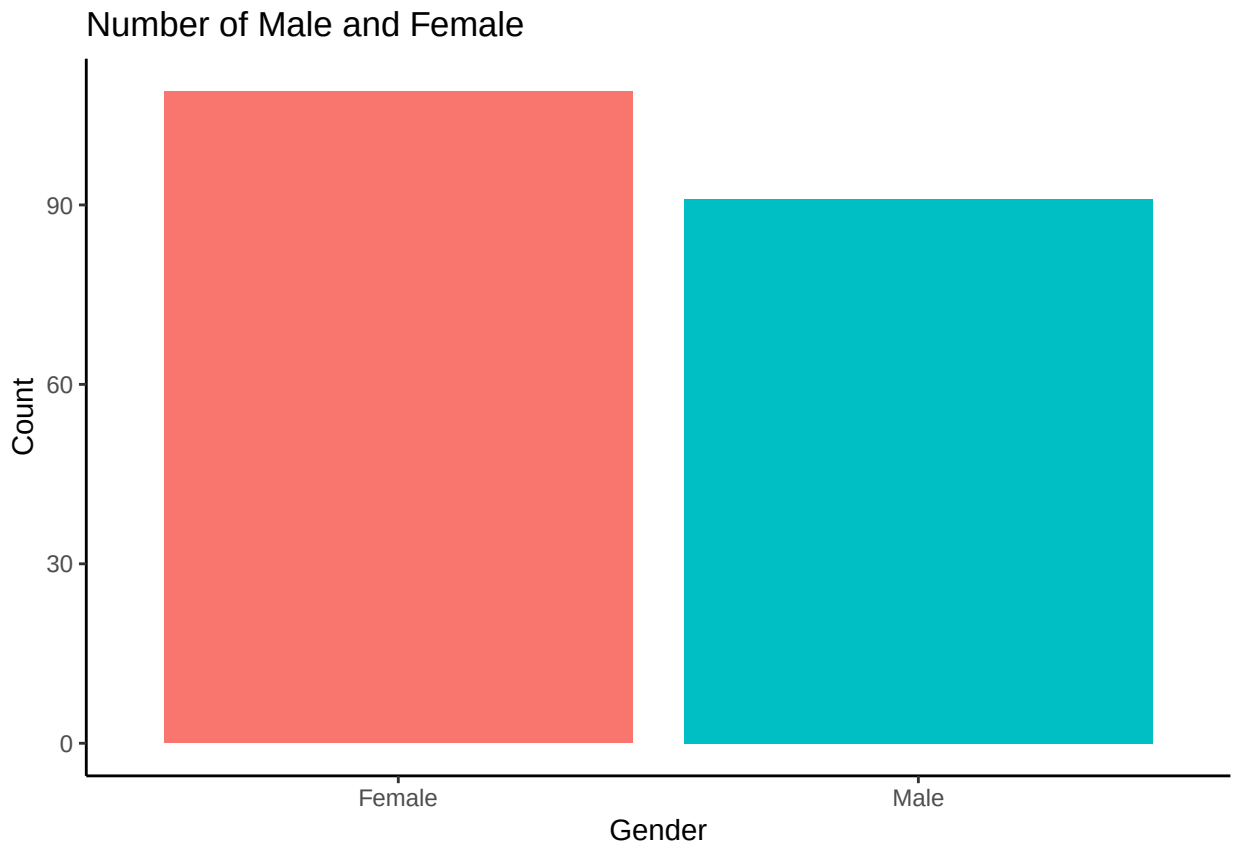
```
##
## Female   Male
##   54.5   45.5
```

```
HS_and_Beyond$gender %>% fct_count #returns the total count of males and females from the data.
```

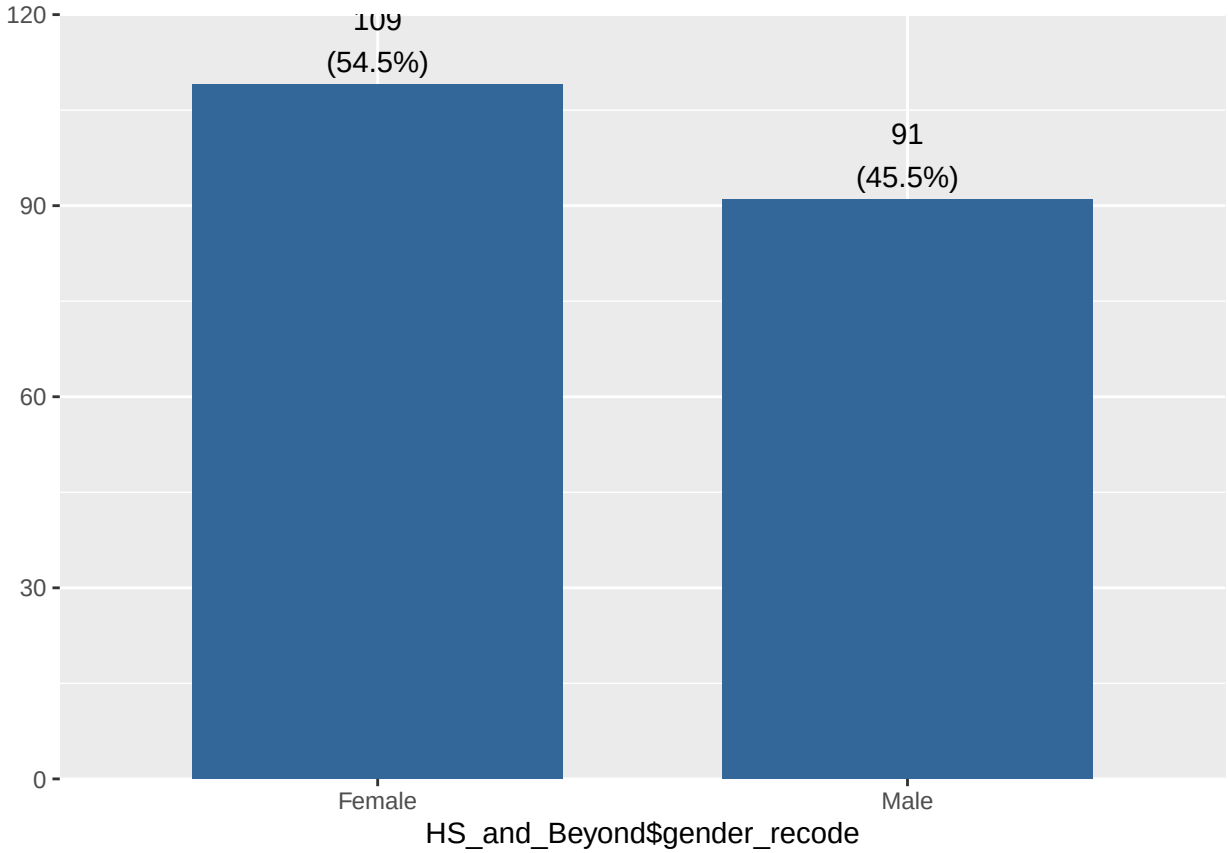
```
## # A tibble: 2 x 2
##   f          n
##   <fct>  <int>
## 1 female   109
## 2 male      91
```

```
ggplot(HS_and_Beyond, aes(x= gender_recode, fill= gender_recode)) + geom_bar() + theme_classic() + scal
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```




```
plot_frq(HS_and_Beyond$gender_recode) # bar graph showing percentages of males vs. females
```



RACE/ETHNICITY

```
HS_and_Beyond$race_recode<- fct_recode(HS_and_Beyond$race,
                                       "African American" ="african american",
                                       "Asian" = "asian",
                                       "Hispanic" = "hispanic",
                                       "Caucasian" = "white") #converts race to a factor variable.
```

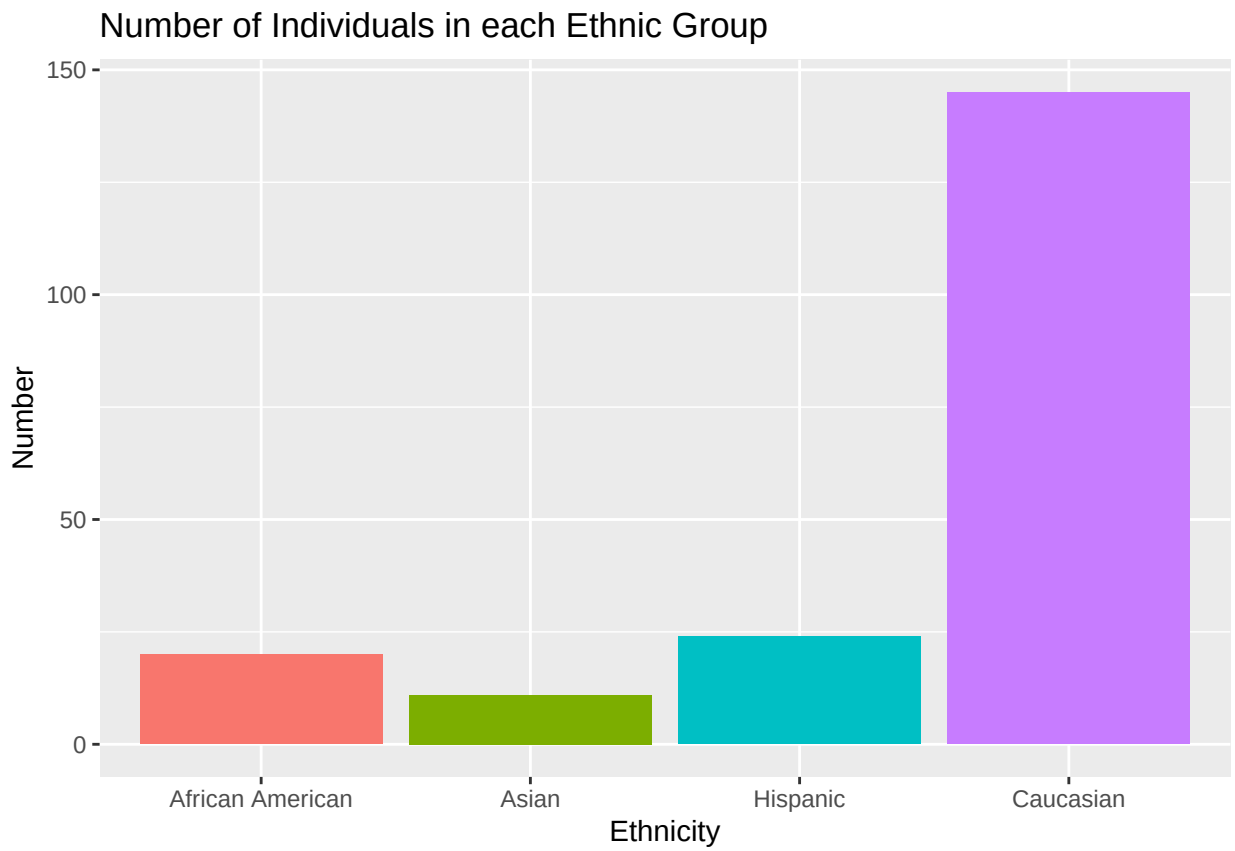
```
table(HS_and_Beyond$race_recode) #frequency table of ethnic groups
```

```
##
## African American            Asian         Hispanic        Caucasian
##               20               11               24              145
```

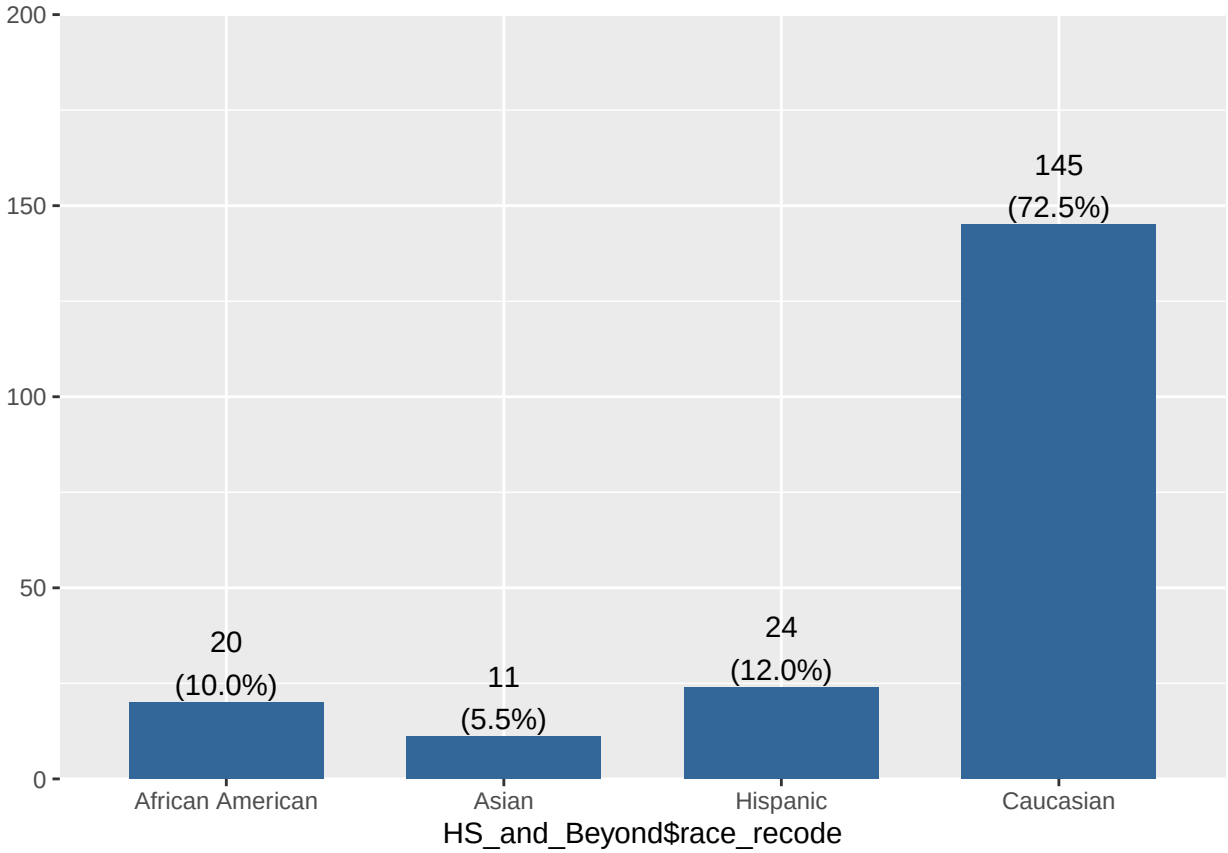
```
table(HS_and_Beyond$race_recode) %>% prop.table() * 100 #Percentage of Ethnicity included in the data
```

```
##
## African American            Asian         Hispanic        Caucasian
##             10.0              5.5             12.0             72.5
```

```
ggplot(HS_and_Beyond, aes(x= race_recode, fill= race_recode)) + geom_bar() + ggtitle("Number of Individ
```




```
plot_frq(HS_and_Beyond$race_recode) #Percentage of each Ethnic Group
```

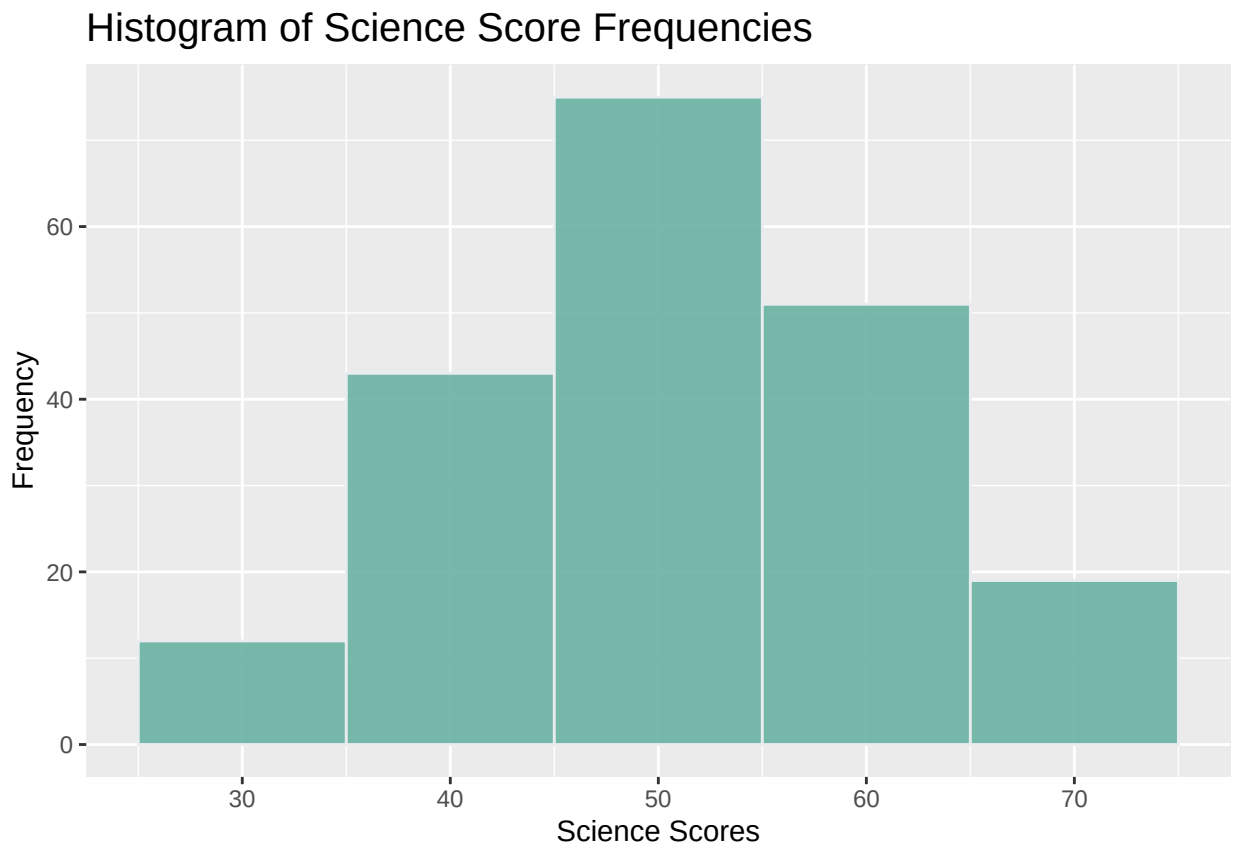


Science

```
summary(HS_and_Beyond$science) #summary statistics for science scores
```

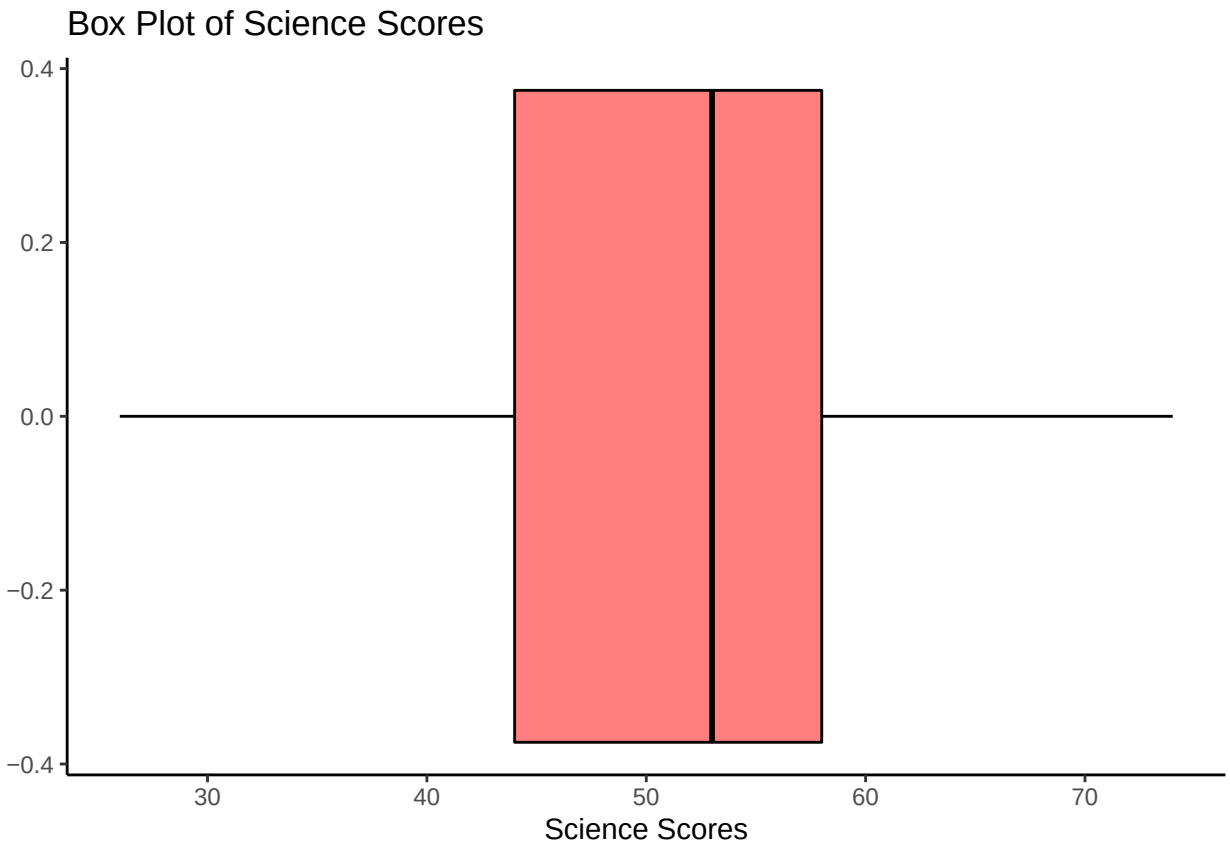
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   44.00   53.00   51.85   58.00   74.00
```

```
HS_and_Beyond %>%
  ggplot( aes(x=science)) +
    geom_histogram( binwidth=10, fill="#69b3a2", color="#e9ecef", alpha=0.9) +
    ggtitle("Histogram of Science Score Frequencies") +
    theme(
      plot.title = element_text(size=15)) + xlab("Science Scores") + ylab("Frequency")
```



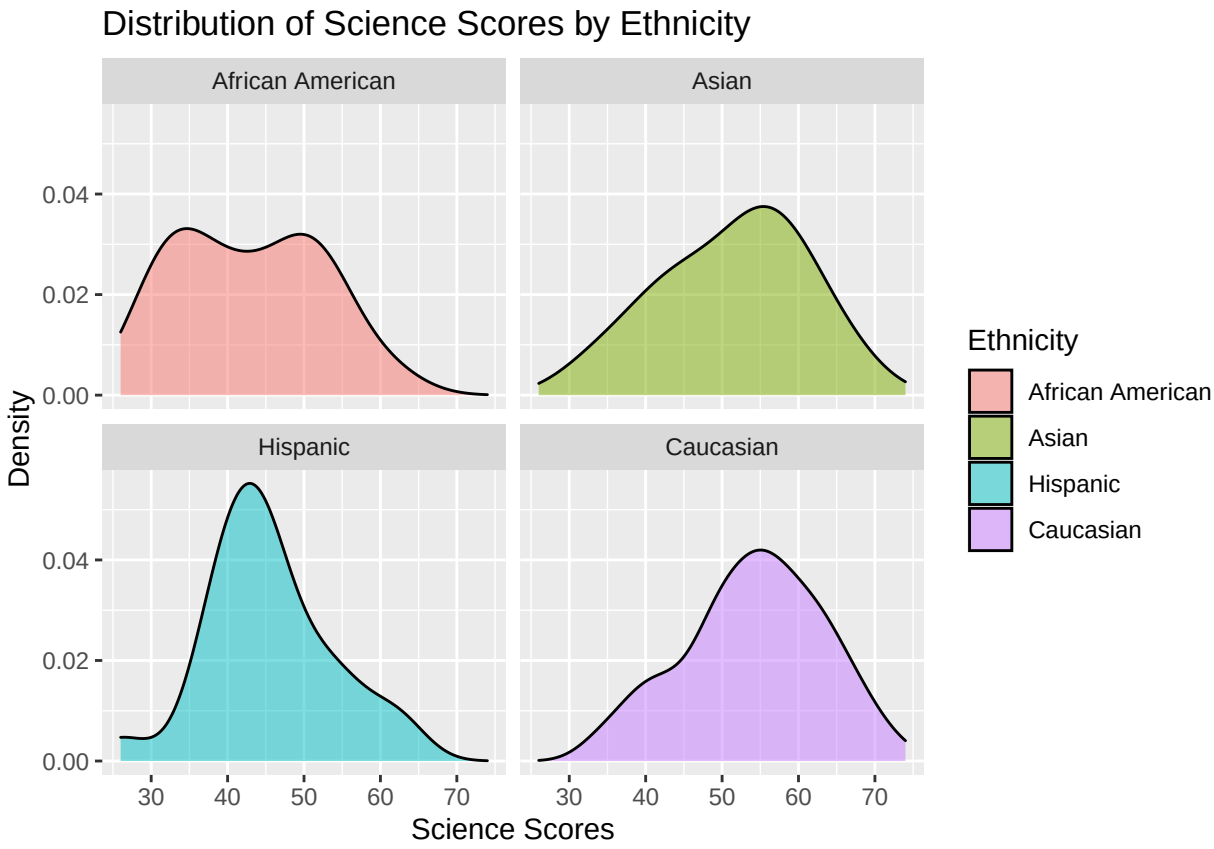
From the histogram we can see that the majority of the scores fall into the 50 percent range

```
ggplot(HS_and_Beyond, aes(x=science)) + geom_boxplot(col= "black", fill= "red", alpha=0.5) + ggtitle("B
```

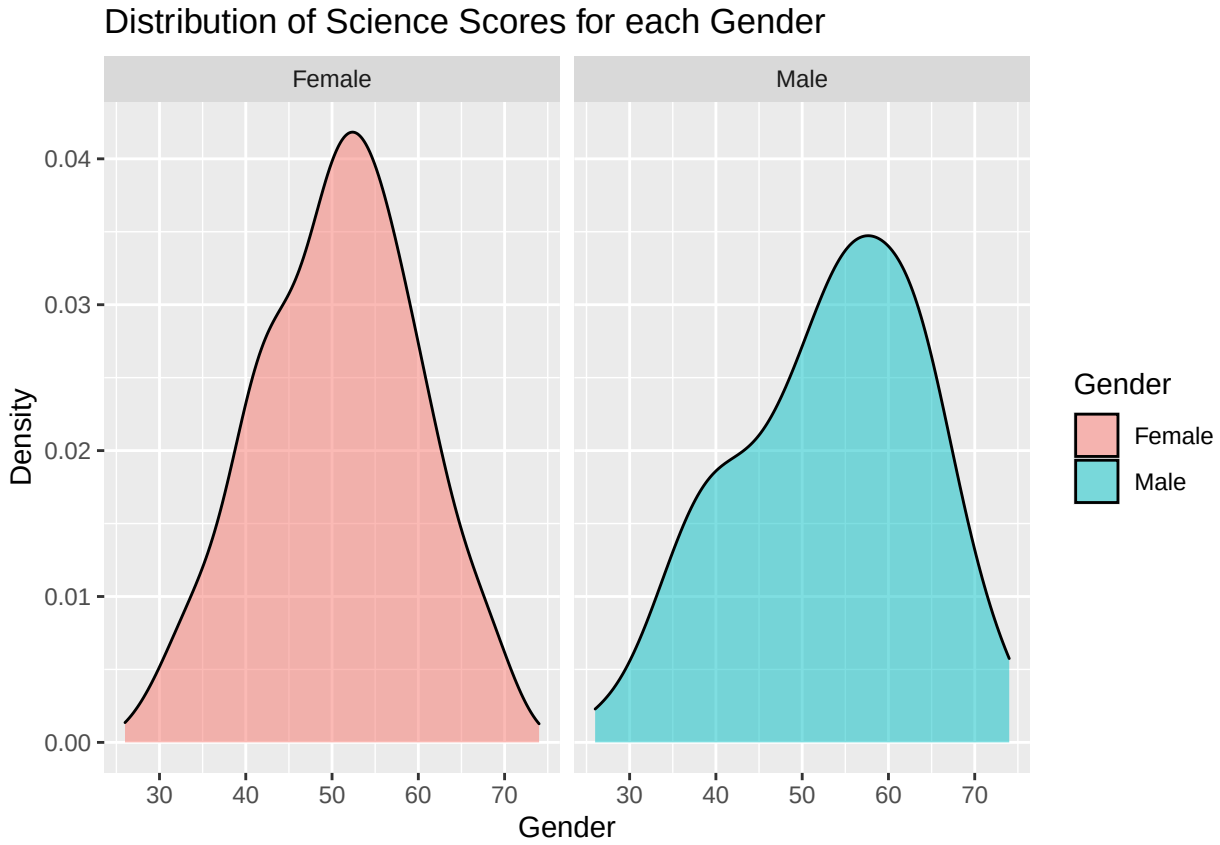


Bivariate Exploration

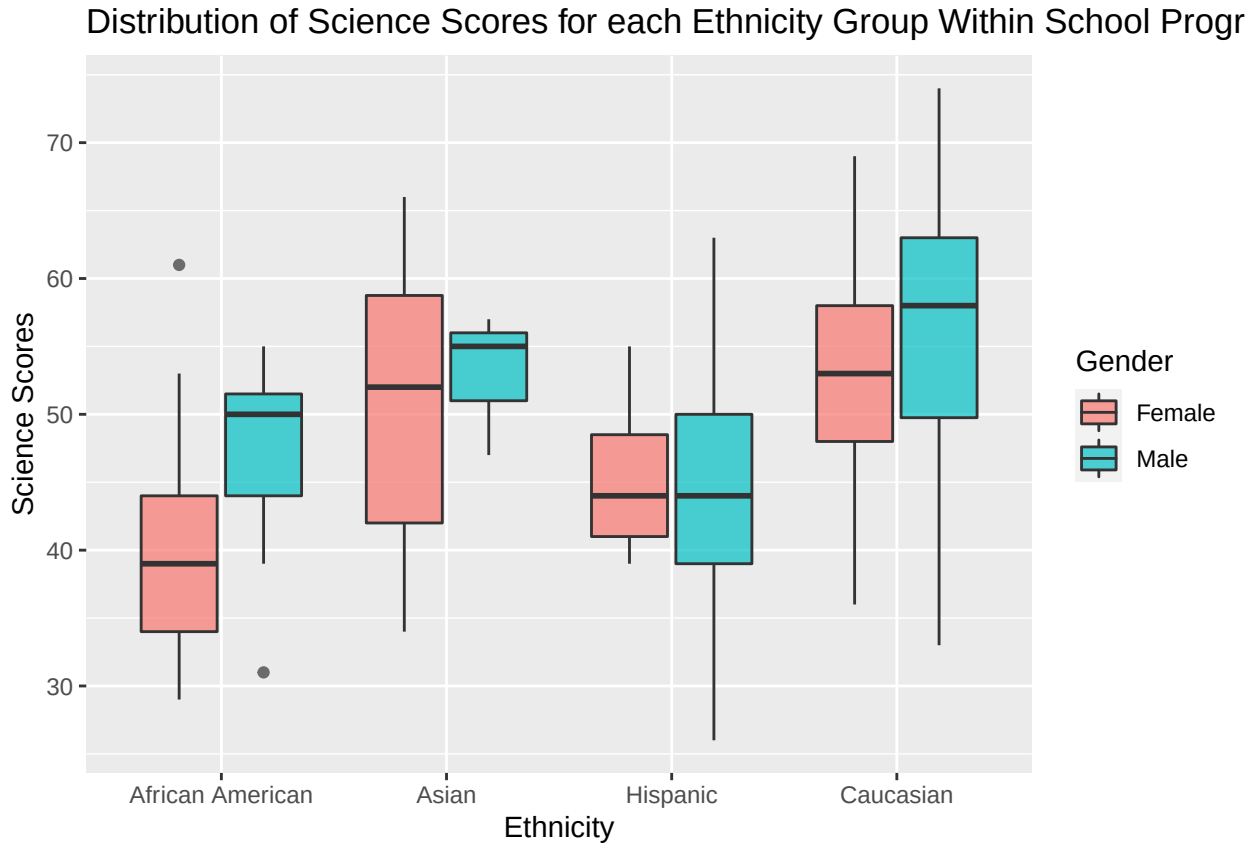
```
ggplot(HS_and_Beyond, aes(x=science, fill= race_recode)) + geom_density(alpha=0.5) + ggtitle("Distributi
```




```
ggplot(HS_and_Beyond, aes( x= science, fill= gender_recode)) +
        geom_density(alpha= .5) + ggtitle("Distribution of Science Scores for each Gender") + facet_wra
```



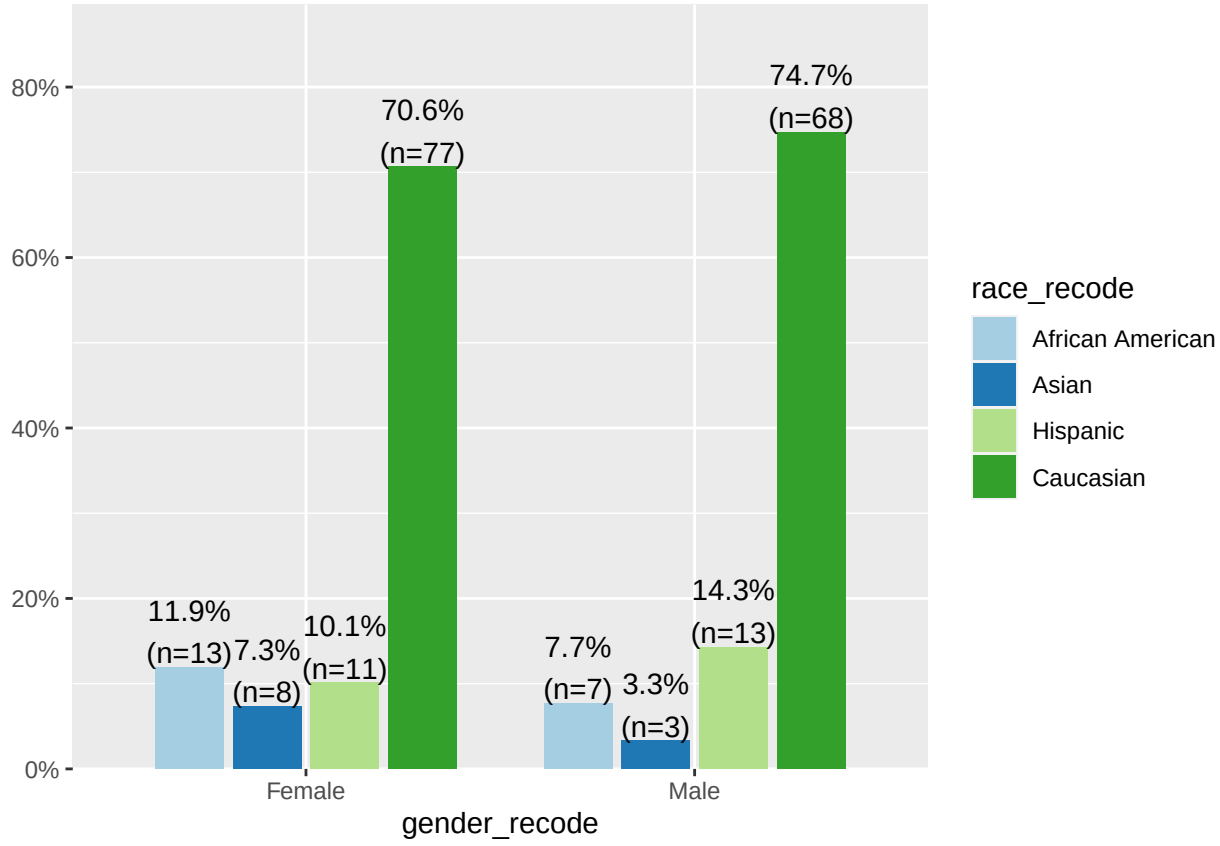

```
ggplot(HS_and_Beyond, aes( x= race_recode, y= science, fill= gender_recode)) +
        geom_boxplot(alpha= .7) + ggtitle("Distribution of Science Scores for each Ethnicity Group Wit
```




```
table(HS_and_Beyond$gender,HS_and_Beyond$race_recode) # number of males and females for each ethnic gro
```

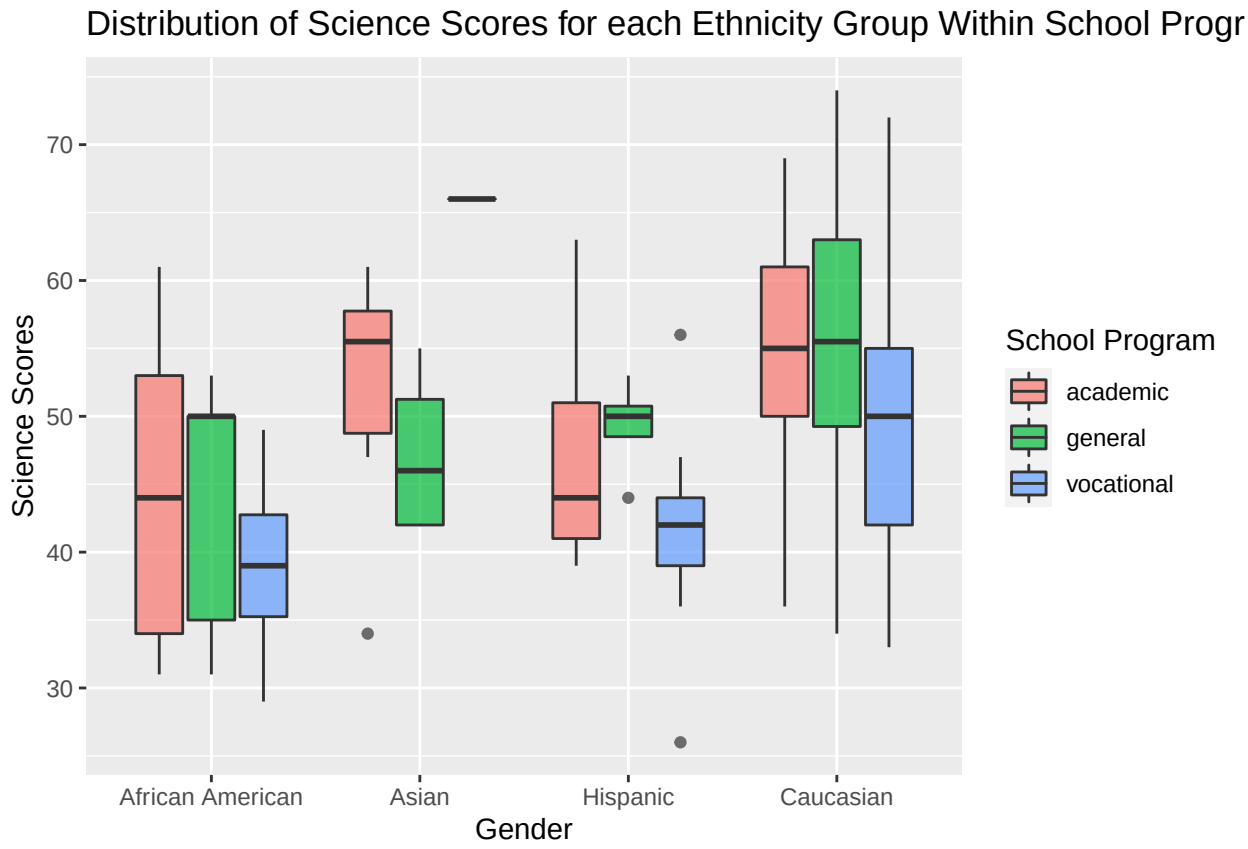
```
##
##          African American Asian Hispanic Caucasian
##   female               13     8       11        77
##   male                  7     3       13        68
```

```
plot_xtab(HS_and_Beyond$gender_recode, HS_and_Beyond$race_recode, margin='row', show.total = FALSE) + tl
```

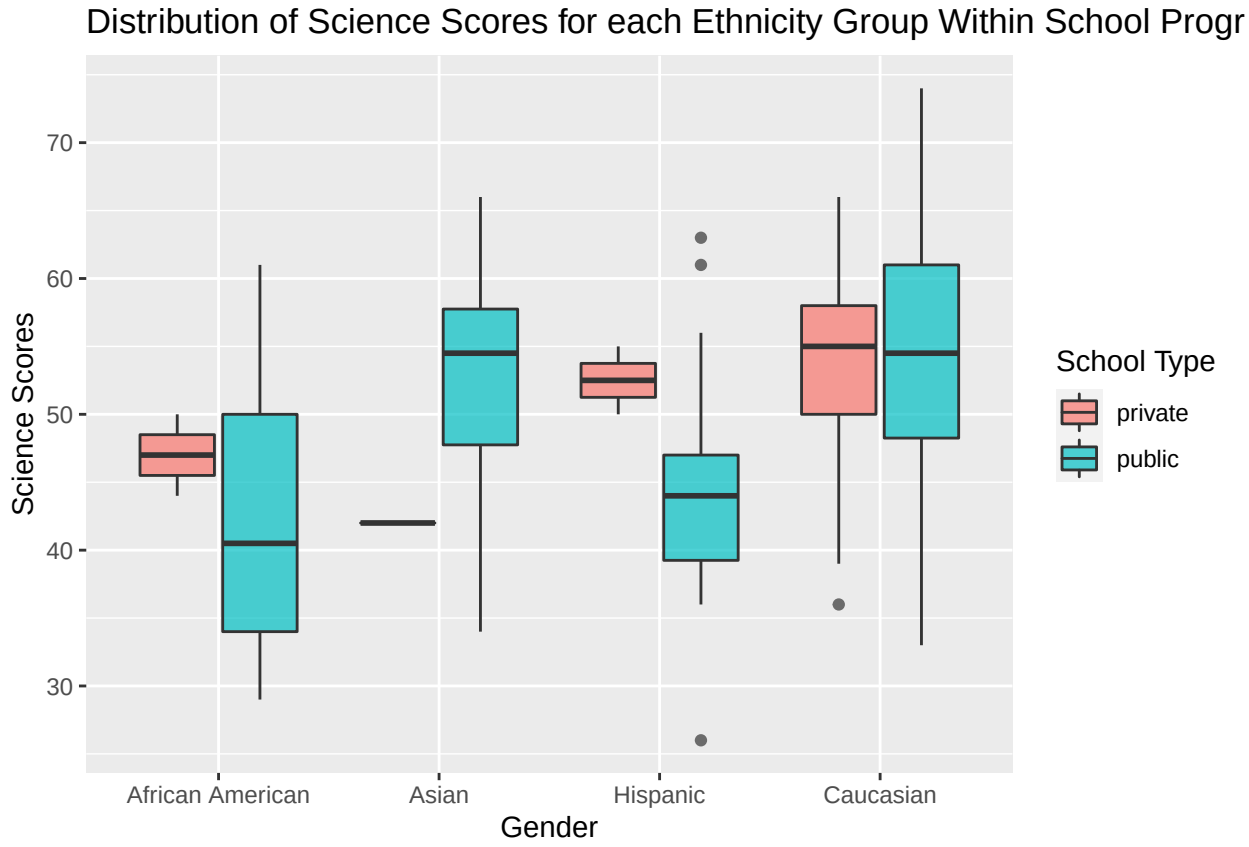


Graph above shows the percentage of males and females for each ethnic group

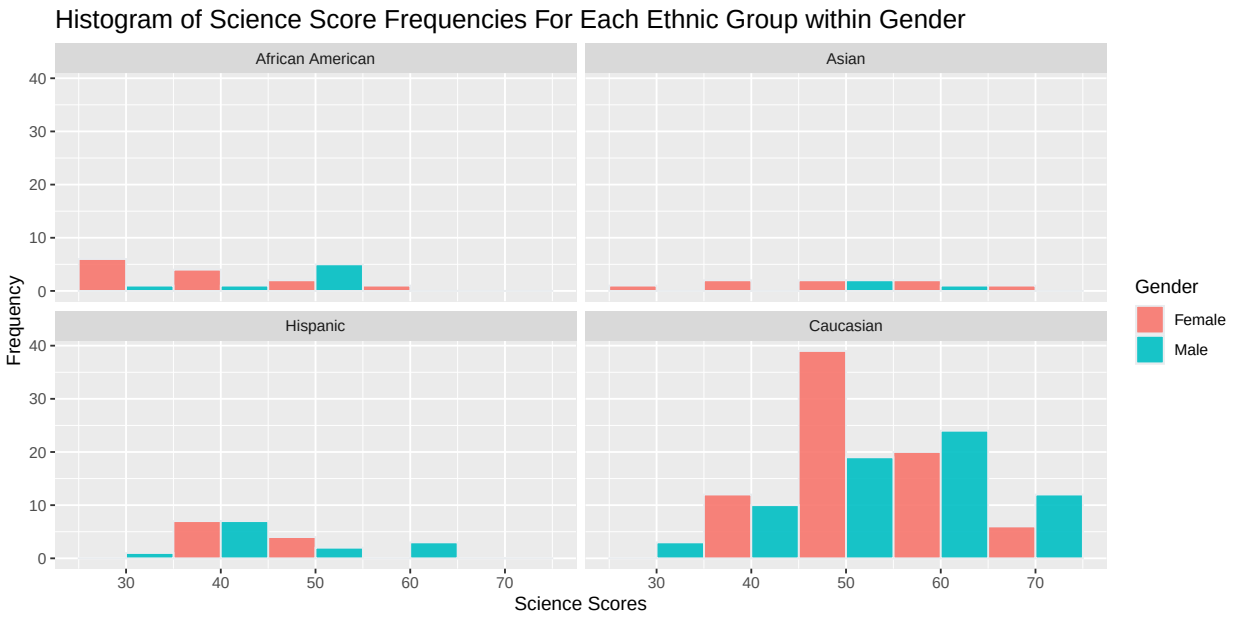
```
ggplot(HS_and_Beyond, aes( x= race_recode, y= science, fill= prog)) +
        geom_boxplot(alpha= .7) + ggtitle("Distribution of Science Scores for each Ethnicity Group Wit
```



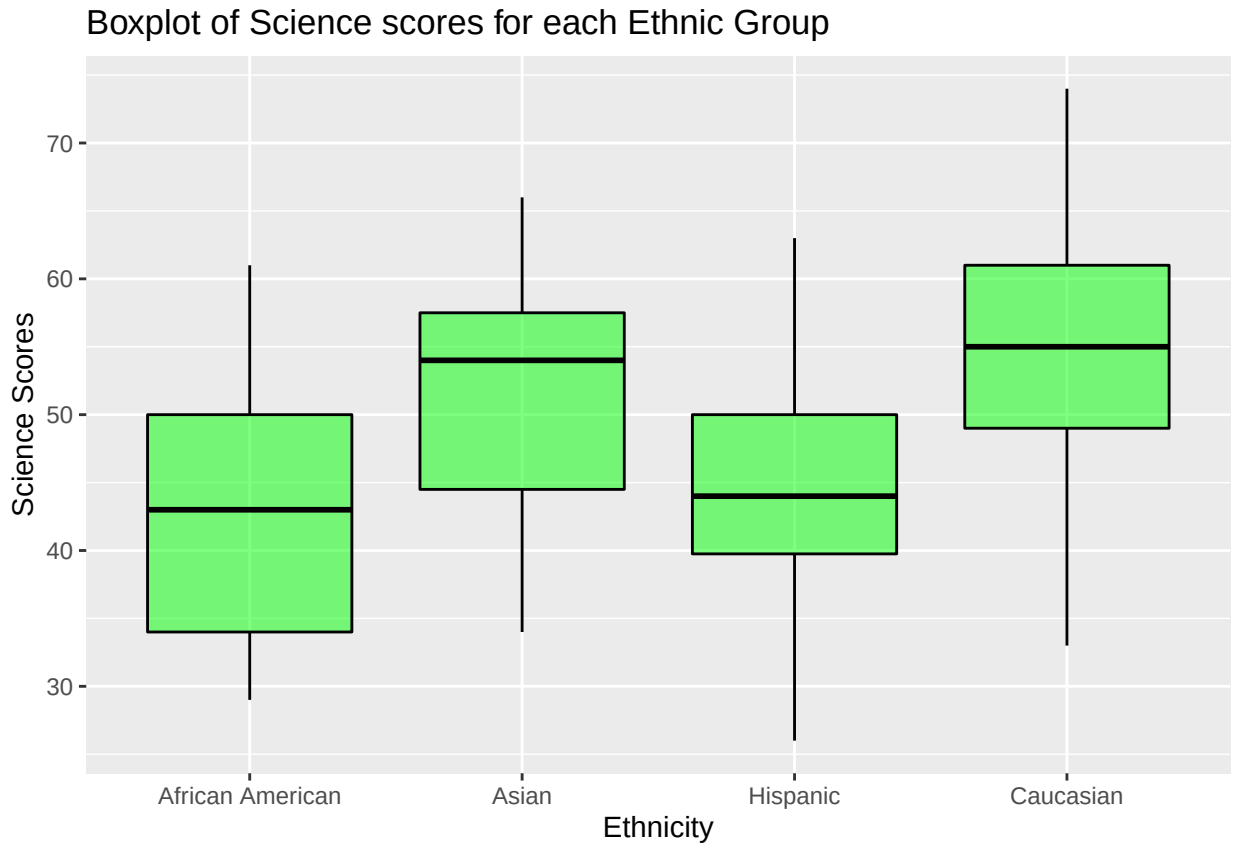

```
ggplot(HS_and_Beyond, aes( x= race_recode, y= science, fill= schtyp)) +
        geom_boxplot(alpha= .7) + ggtitle("Distribution of Science Scores for each Ethnicity Group Wit
```




```
HS_and_Beyond %>%
  ggplot( aes(x=science, fill= gender_recode)) +
    geom_histogram( binwidth=10, color="#e9ecef", alpha=0.9, position = "dodge") +
    ggtitle("Histogram of Science Score Frequencies For Each Ethnic Group within Gender") +
    theme(
      plot.title = element_text(size=15)) + xlab("Science Scores") + ylab("Frequency") + facet_wrap(~ra
```




```
ggplot(HS_and_Beyond, aes(x= race_recode, y= science)) + geom_boxplot(col="black", fill= "green", alpha=
```



Conclusion After conducting my analysis on the three variables i have chosen from my data set I can now answer the questions I had about my data and determine if my hypotheses are valid. 1. For my first

question, did more females complete the questionnaire than males did? The answer is: there was a larger number of females that completed the questionnaire. Therefore hypothesis 1 is invalid. 2. For my second question, which ethnic group contained the least individuals that completed the questionnaire? Asians have the lowest percentage of individuals who completed the questionnaire, thus my hypothesis is valid. 3.For my third question, which ethnic group has the highest scores on the science portion of the questionnaire? The Caucasians were the ethnic group with the highest scores on the science section, therefore my hypothesis is true. 4. For my final question, for each ethnic group which gender has higher scores on the science section in the questionnaire? Males had the highest scores on the science section for African Americans, Caucasians, and Hispanics. However females had the highest scores for the Asian's. Thus, my prediction was false.