Exploratory Data Analysis Project

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(forcats)
library(tinytex)
```

```
## Warning: package 'tinytex' was built under R version 4.0.5
```

Introduction: A short introduction/description of the data.Specifically mention the 2-3 variables you are going to explore.What is your research question? What are you interested in finding out more about?

The data set I choose to use for the exploratory analysis project is High School and beyond. This data set included 200 observations and 11 variables such as student race, gender, school type, program, and scores for math, writing, & science. I choose to focus my project on gender, race, and type of program.

```
highschool <-read.table("/Users/allisondanelia/Desktop/MATH130/data/hsb2  2.txt", header = TRUE, sep = "
head(highschool)
```

```
##    id gender  race    ses schtyp       prog read write math science socst
## 1  70   male white    low public    general   57    52   41      47    57
## 2 121 female white middle public vocational   68    59   53      63    61
## 3  86   male white   high public    general   44    33   54      58    31
## 4 141   male white   high public vocational   63    44   47      53    56
## 5 172   male white middle public   academic   47    52   57      53    61
## 6 113   male white middle public   academic   44    52   51      63    61
```

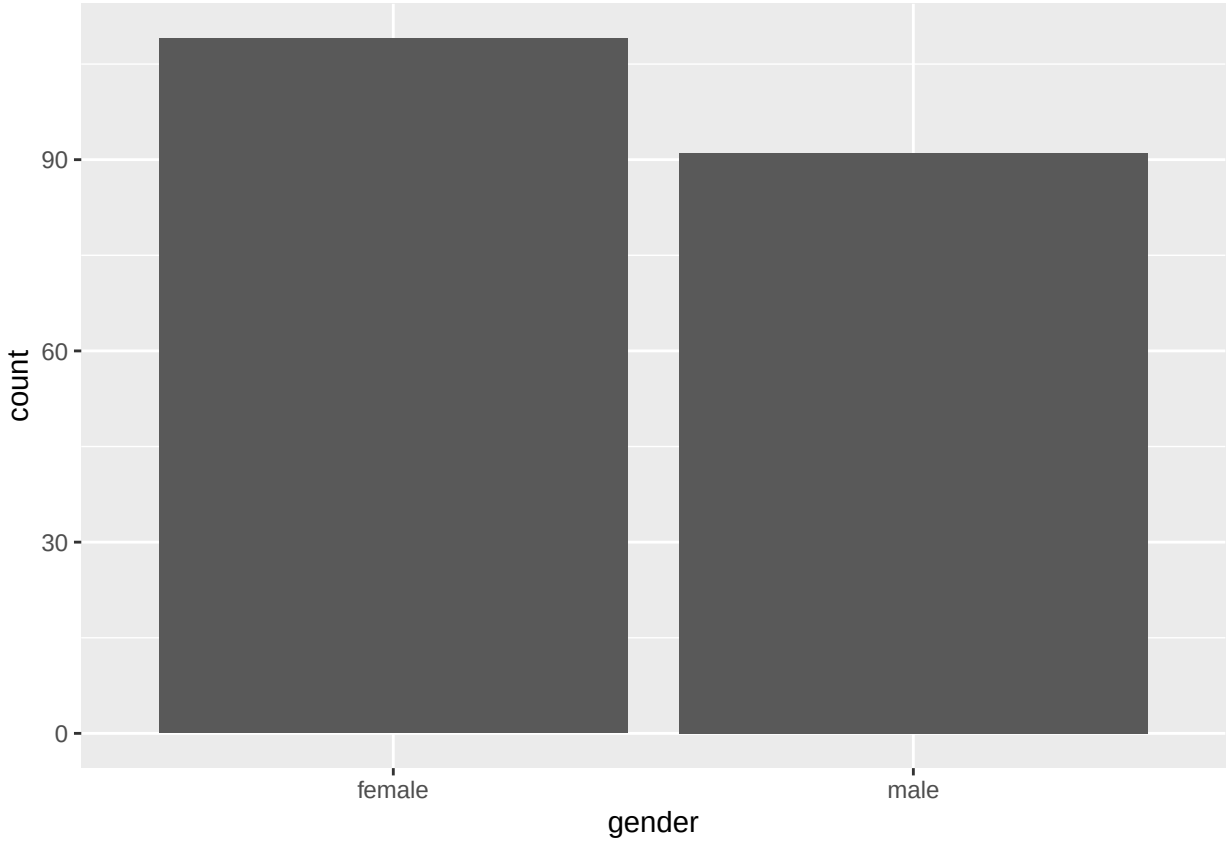
Univariate Exploration: Describe each of the variables under consideration.This means calculate some summary statistics (N(%) or mean(sd)) and make a graphic

Gender

```
table(highschool$gender)
```

```
##
## female   male
##    109     91
```

```
ggplot(highschool, aes(x= gender))+ geom_bar()
```

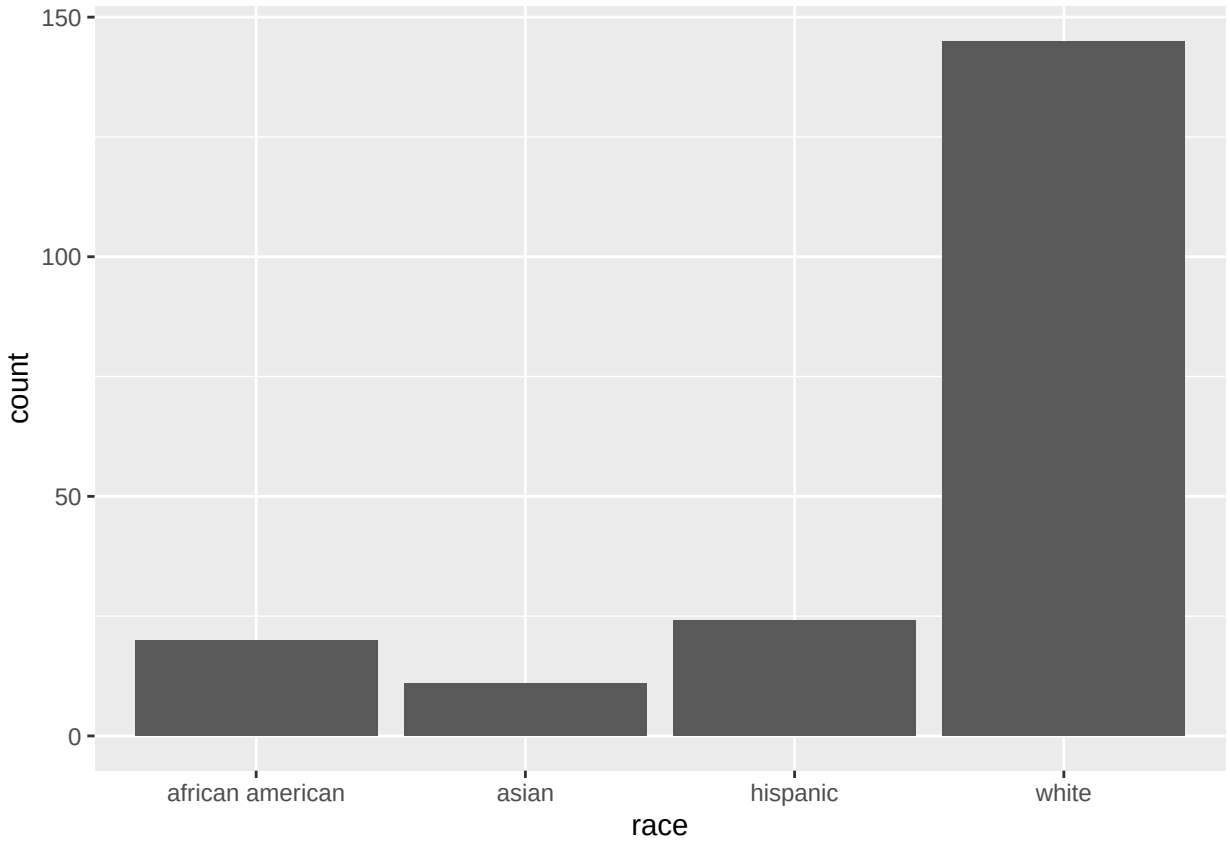


As you can visually see, this data set contains more females than males. #Race

```
table(highschool$race)
```

```
##
## african american            asian         hispanic            white
##               20               11               24              145
```

```
ggplot(highschool, aes(x=race))+ geom_bar()
```



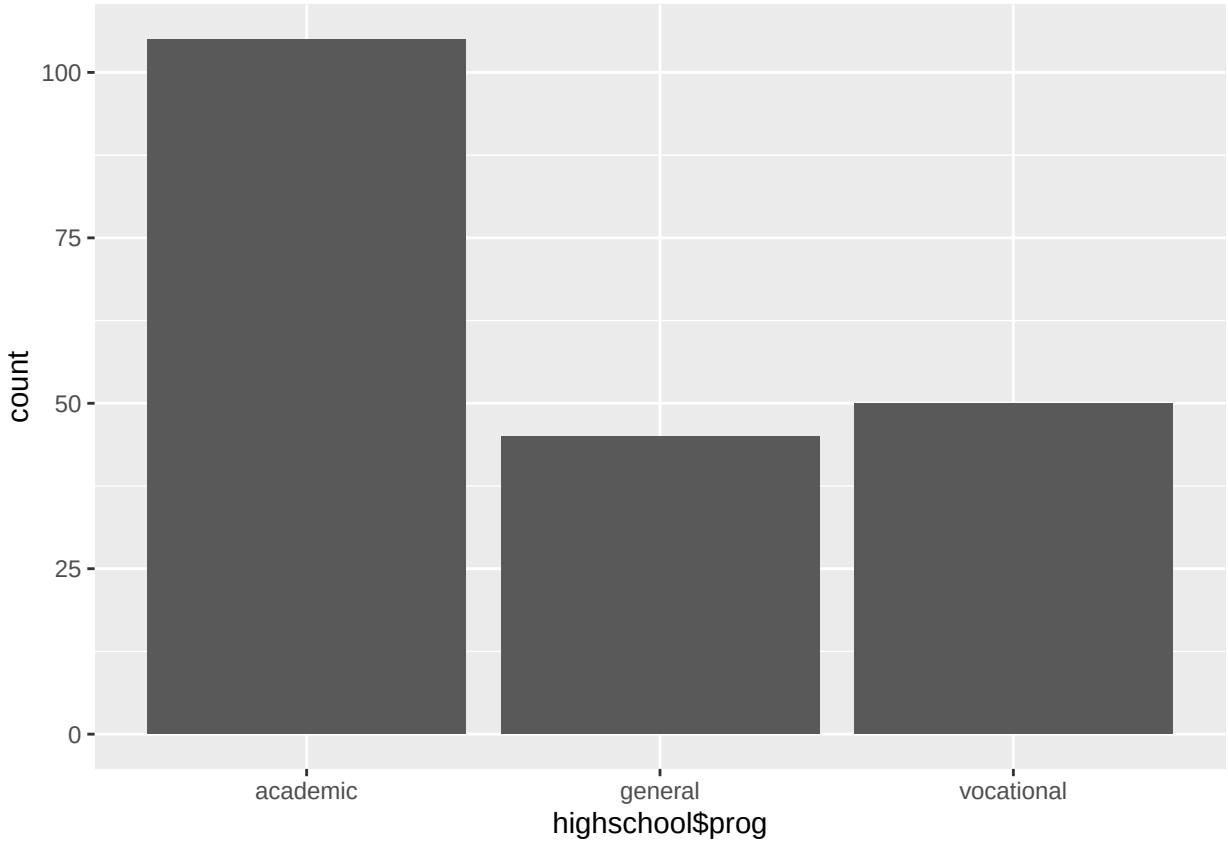
The histogram shows the majority of the students are white versus any other student race group. #Type of program

```
table(highschool$prog)
```

```
##
##   academic    general vocational
##        105         45         50
```

```
ggplot(highschool, aes(x=highschool$prog))+geom_bar()
```

```
## Warning: Use of ‘highschool$prog‘ is discouraged. Use ‘prog‘ instead.
```



The graph shows that the majority of the high school students have attended academic programs while it is about the same amount of students enrolled in general or vocational programs.

##Bivariate Exploration: Comparison between two variables of interest.Calculate grouped summary statistics as appropriate. This is often the most often forgotten partYou can go further and explore more than two variables at a time using paneling, but be sure to explain what you learn from each graph. ##### Race Vs. Program

```
table(highschool$race, highschool$prog)
```

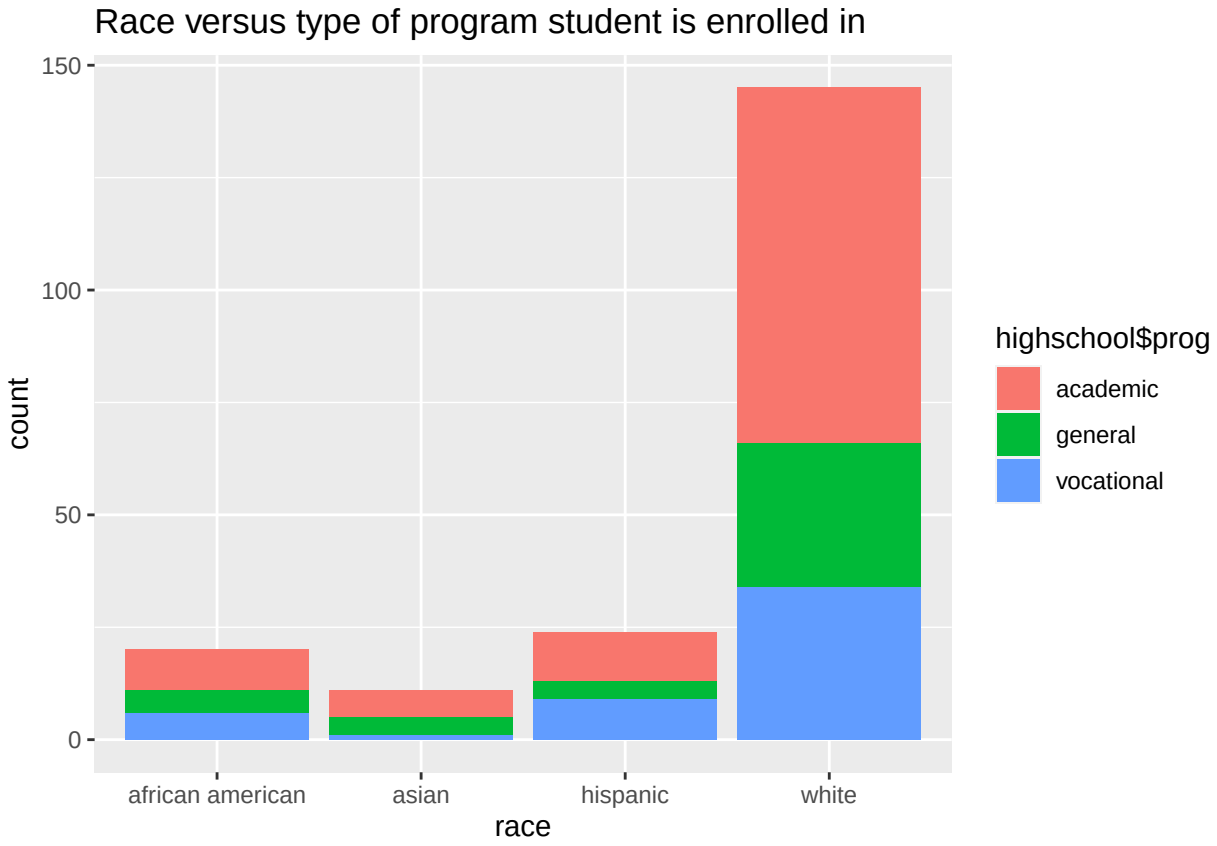
```
##
##                   academic general vocational
##   african american        9       5          6
##   asian                   6       4          1
##   hispanic               11       4          9
##   white                  79      32         34
```

```
prop.table(table(highschool$race, highschool$prog))
```

```
##
##                   academic general vocational
##   african american    0.045   0.025      0.030
##   asian               0.030   0.020      0.005
##   hispanic            0.055   0.020      0.045
##   white               0.395   0.160      0.170
```

```
ggplot(highschool, aes(x=race, fill=highschool$prog))+geom_bar()+ggtitle("Race versus type of program s
```

```
## Warning: Use of ‘highschool$prog‘ is discouraged. Use ‘prog‘ instead.
```



The histogram shows what type of program students are enrolled in based on thier race. You can see that the majority of white individuals are involved in academic programs. While Hispanics and African American involved in the three types of schools are about the same for all three. It shows that Asians are only involved in general or academic programs.

```
table(highschool$race, highschool$gender)
```

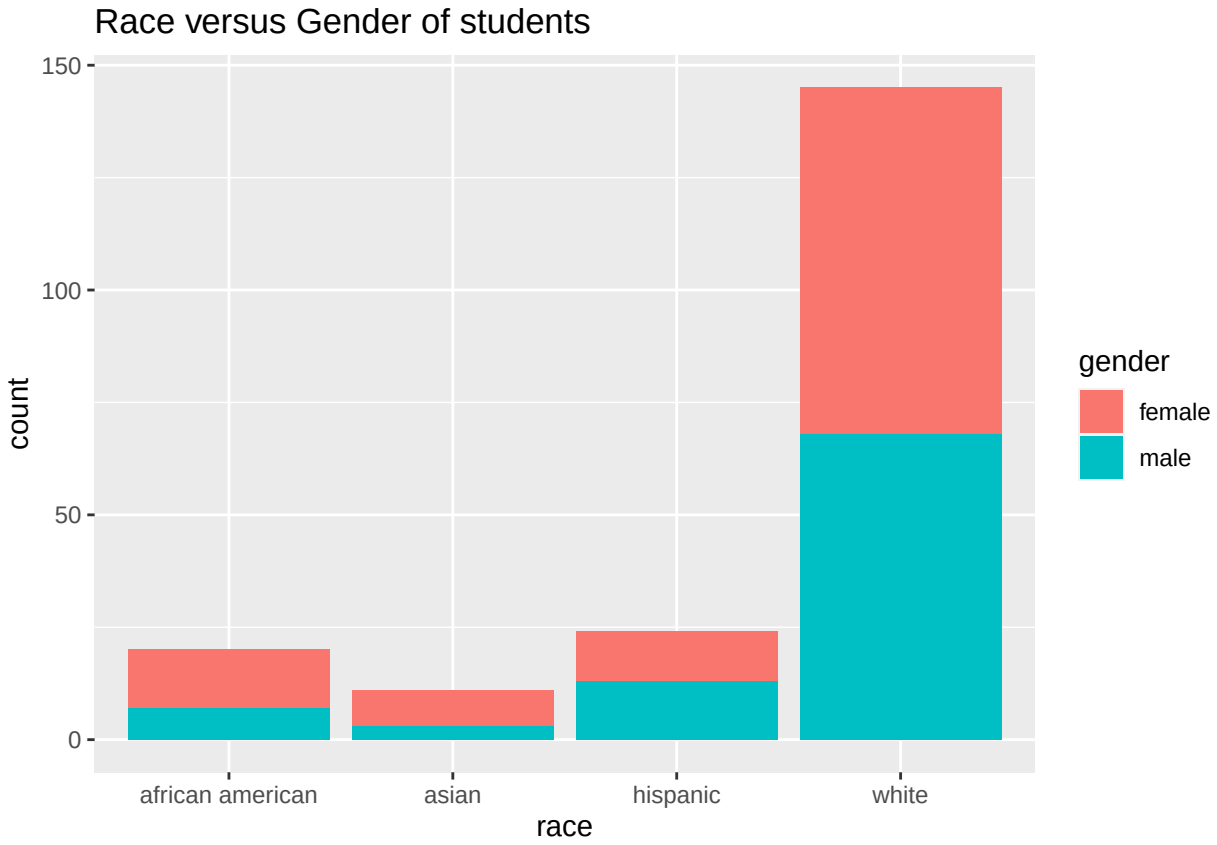
Race Vs Gender

```
##
##                   female male
##   african american     13    7
##   asian                 8    3
##   hispanic             11   13
##   white                77   68
```

```
prop.table(table(highschool$race, highschool$gender))
```

```
##
##                    female  male
##   african american  0.065 0.035
##   asian             0.040 0.015
##   hispanic          0.055 0.065
##   white             0.385 0.340
```

```
ggplot(highschool, aes(x=race, fill=gender))+geom_bar()+ggtitle("Race versus Gender of students")
```



The graph shows that the amount of white and Hispanic students is about half male, half female. There are more female African American students than males. There are also more Asian females involved in programs versus that of males.

##Conclusion: What did you find? If you had a prior hypothesis, does the data seem to support it? Remember this is NOT a statistical analysis. In conclusion, it was really interesting to work with this dataset and compare race, gender and type of program. I assumed that the majority of students involved in higher education would be white. This anaylsis showed that my hypothesis was in deed correct.The data supports my claim. I found it interesting that for African American and Asian students there is quite a larger number of females versus males. It was also proven that the number of white students involved in higher education programs is much higher than that of African American, Hispanic or Asian.