Exploratory Data Analysis

Christopher Portillo

2022-09-19

1. Introduction

For the exploratory data analysis project I will be going over the High school and Beyond. The data collected is a Longitudinal Study of High School and Sophomore classes. The class years is taken of 1980 and a follow up in 1992. The data set for High School and Beyond covers 200 observations and 11 variables. I will be going over race, gender, and test scores of all the students who were part of the data set. What subject has the highest test scores and which gender was the highest to score.How does gender effect which subject is more popular? I want to see how the variables correspond with the subjects picked by students compared to which type of students are deciding.

```
highschool <- read.csv("/Users/christopher/Desktop/math130/data/hsb2.csv", header=TRUE, sep = "\t")
str(highschool)
```

```
## 'data.frame':    200 obs. of  11 variables:
##  $ id     : int  70 121 86 141 172 113 50 11 84 48 ...
##  $ gender : chr  "male" "female" "male" "male" ...
##  $ race   : chr  "white" "white" "white" "white" ...
##  $ ses    : chr  "low" "middle" "high" "high" ...
##  $ schtyp : chr  "public" "public" "public" "public" ...
##  $ prog   : chr  "general" "vocational" "general" "vocational" ...
##  $ read   : int  57 68 44 63 47 44 50 34 63 57 ...
##  $ write  : int  52 59 33 44 52 52 59 46 57 55 ...
##  $ math   : int  41 53 54 47 57 51 42 45 54 52 ...
##  $ science: int  47 63 58 53 53 63 53 39 58 50 ...
##  $ socst  : int  57 61 31 56 61 61 61 36 51 51 ...
```

```
library("ggplot2")
library("knitr")
library("forcats")
library("dplyr")
```

```
##
## 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("RColorBrewer")
```

2. Univariate Exploration

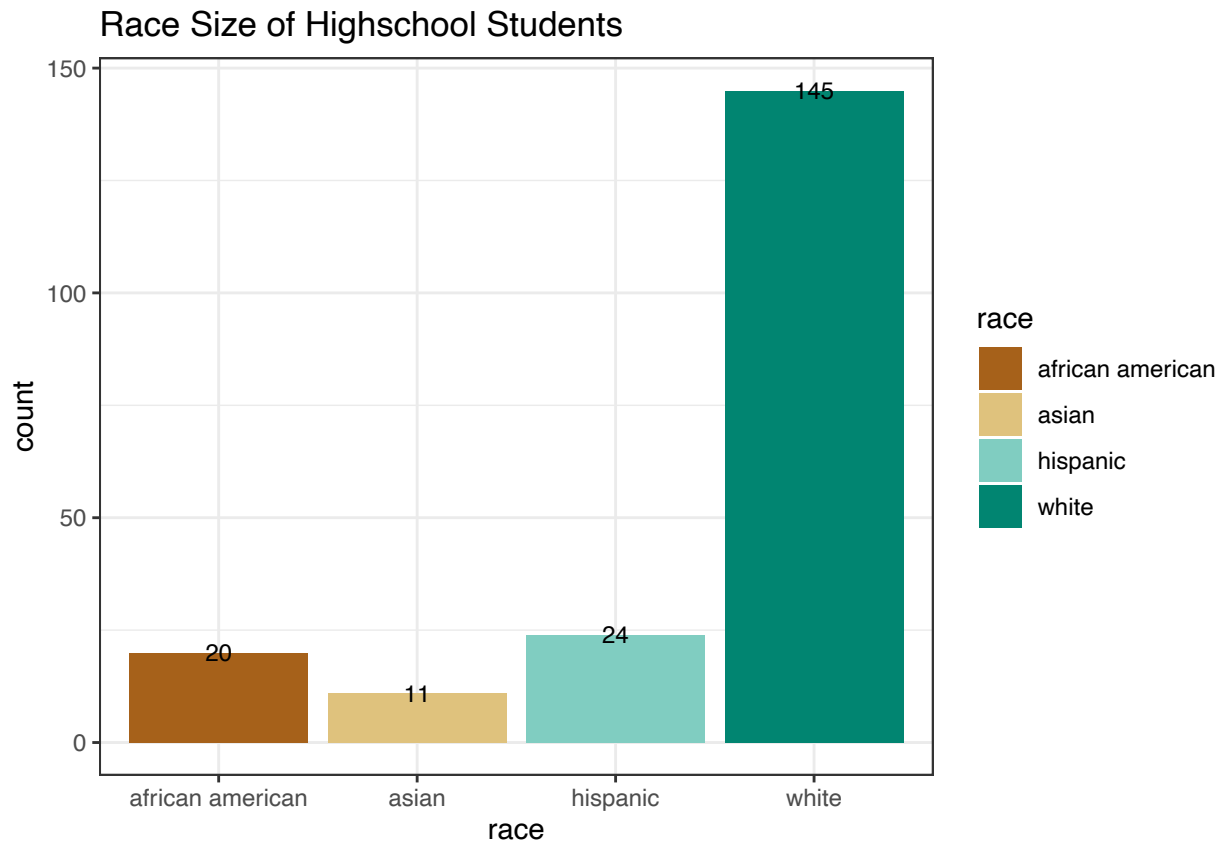
#lets first look at how many students there are in each race category.

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

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

#the population of the highschool is diverse with a total of 200 observents.

```
ggplot(highschool, aes(x=race, fill=race))+ theme_bw() + geom_bar() + scale_fill_brewer(palette="BrBG")
```



#The Highschool data clealry shows that white is the majority of race being at 72.5% and lowest being asian at 5.5%

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

```
##
##                    female male
```

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

#Above shows many many of each race we have n the highschool then separated by their gender. The majority seems to be white females in the data set.

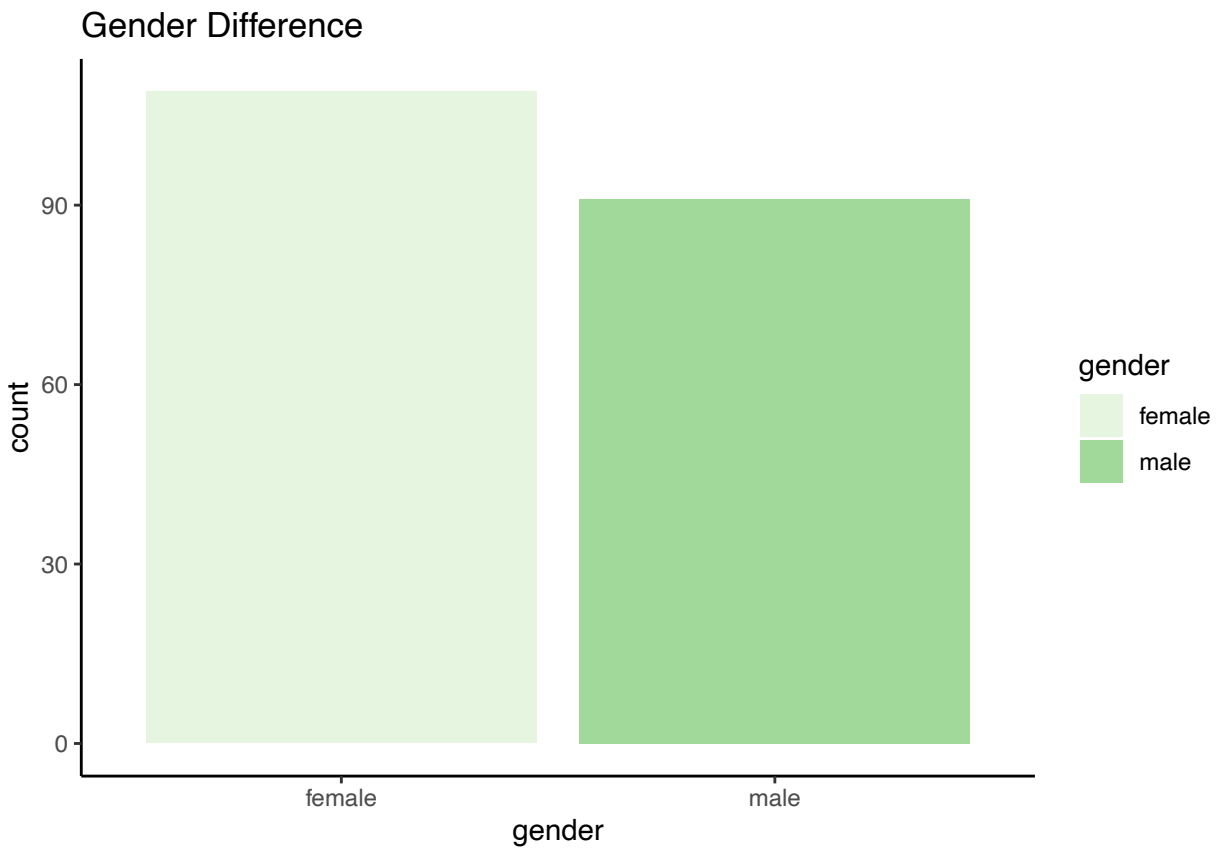
#Below is the amount of female and male students that are in the data recorded as their gender

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

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

```
ggplot(highschool, aes(x=gender, fill=gender)) + geom_bar() + scale_fill_brewer(palette="Viridus") + gg
```

```
## Warning in pal_name(palette, type): Unknown palette Viridus
```



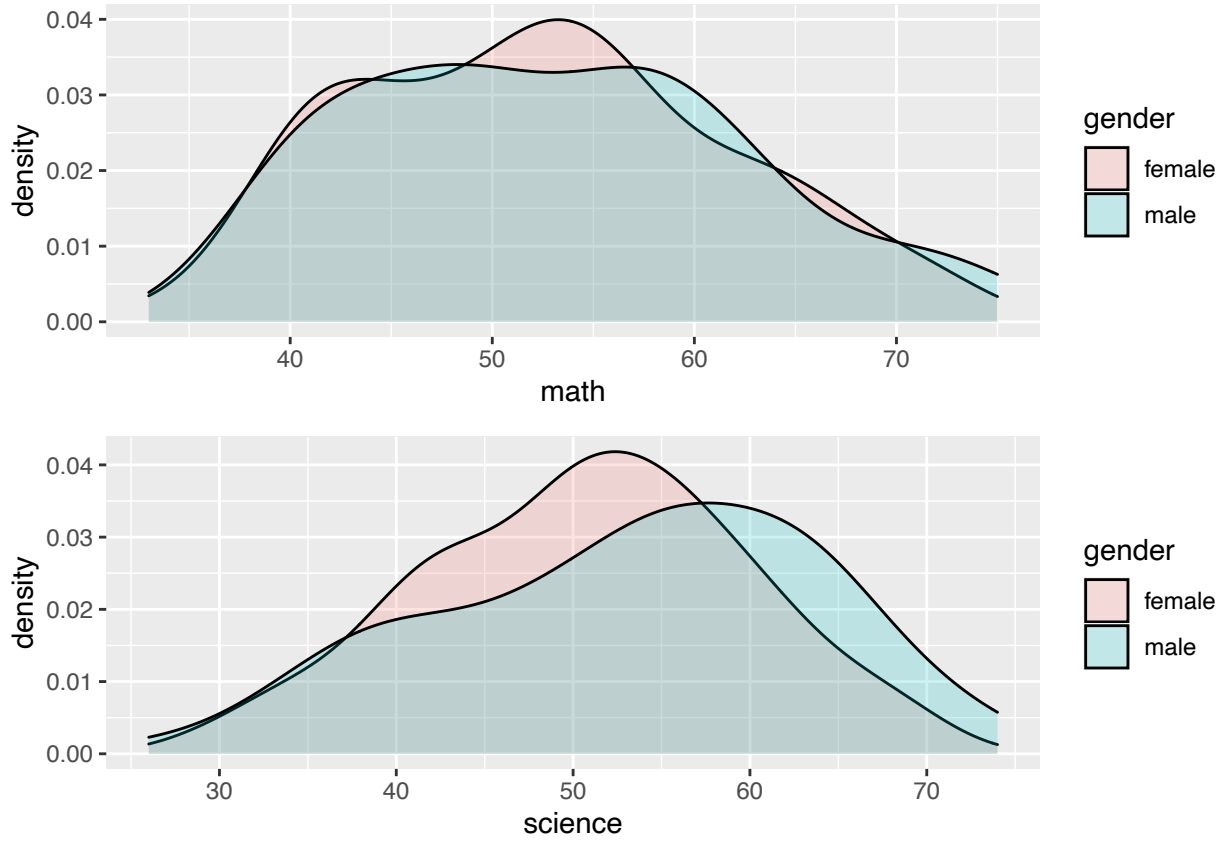
#Clearly this bar graph shows that there are more females in the data set then there are males totaling to 200.

3. Bivariate Exploration

```
library(gridExtra)
```

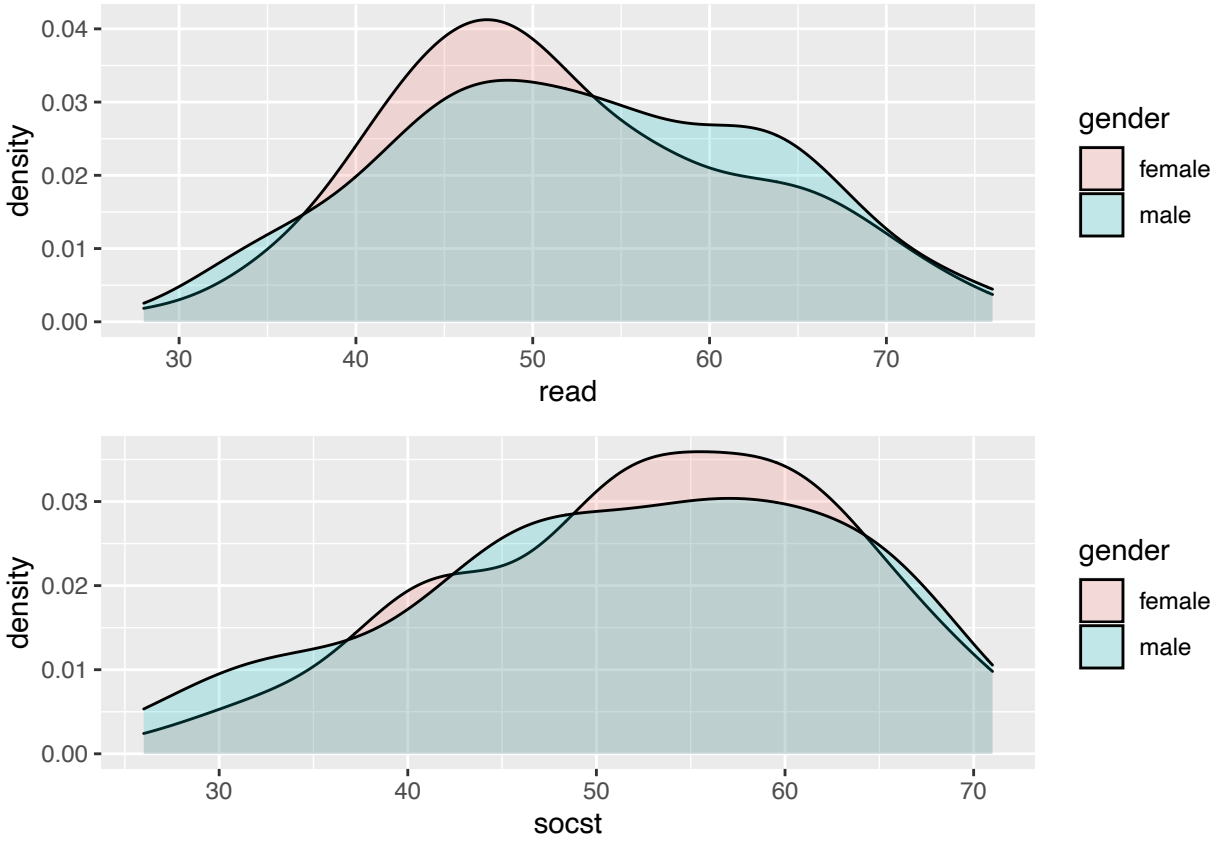
```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
plot1 <- ggplot(highschool, aes(x=math, fill=gender)) + geom_density(alpha=.2)
plot2 <- ggplot(highschool, aes(x=science, fill=gender)) + geom_density(alpha=.2)
grid.arrange(plot1,plot2, ncol=1)
```



#We can compare math and science test scores and see the difference that is shown from the males of the highschool and the females. Both density plots are shown on the same graph of each school categories recorded in the data set. With the test scores shown we can see the majority of scores in Math which was 54 for females and 59 for males.Then take a look at science and see that the majority scored 53 for females and 59 for males.

```
library(gridExtra)
plot1 <- ggplot(highschool, aes(x=read, fill=gender)) + geom_density(alpha=.2)
plot2 <- ggplot(highschool, aes(x=socst, fill=gender)) + geom_density(alpha=.2)
grid.arrange(plot1,plot2, ncol=1)
```

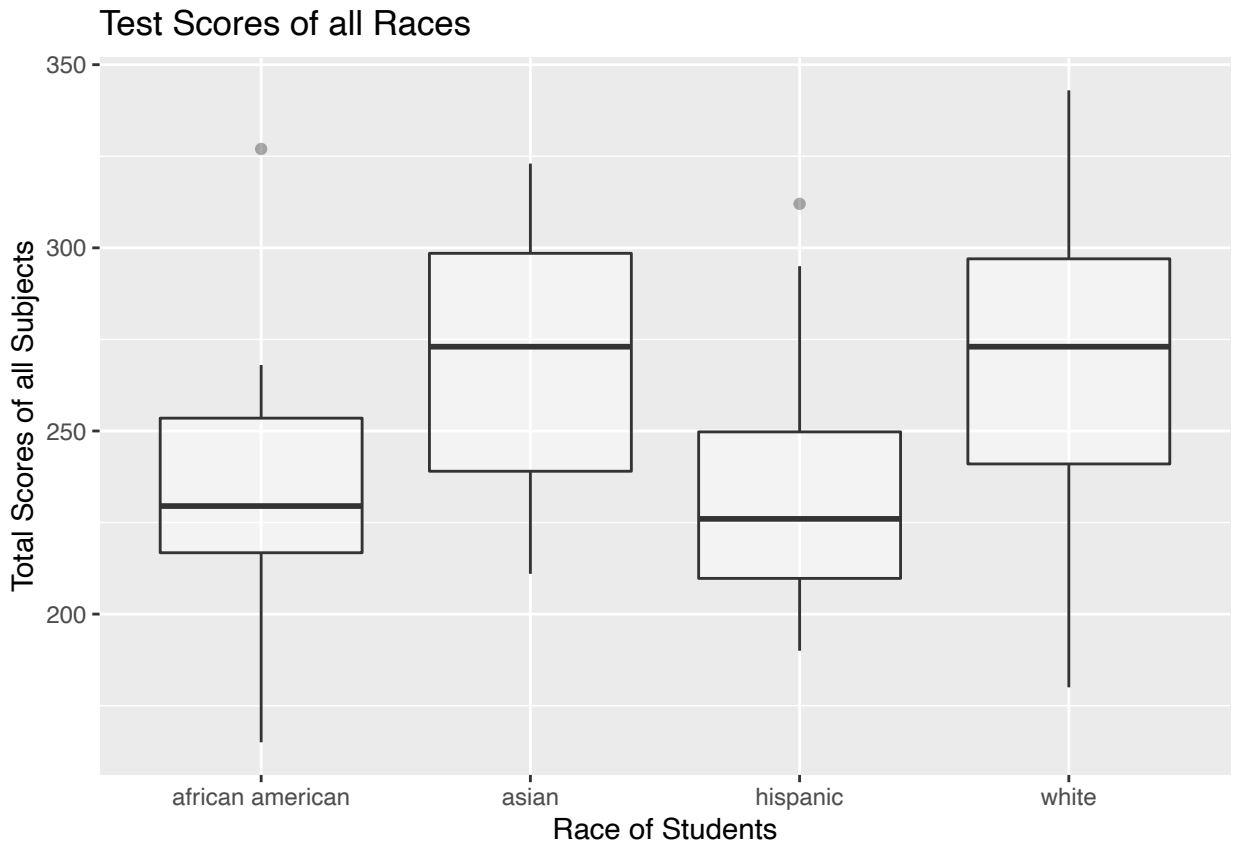


#The majority of reading scores for females were 46. The density plot for the males had a weird shape but its highest was at 47. Under Socst females scored consistantly at scores 52-60 and males scores consistantly at scores 48-63.

```
highschool$subjects <- highschool$read + highschool$write + highschool$math + highschool$science + highs
```

#Now all the test scores are added together.

```
ggplot(highschool, aes(x=subjects, y=race,  fill = subjects)) + geom_boxplot(alpha=0.4) +
        theme(legend.position="none") + ylab ("Race of Students") + xlab ("Total Scores of all Subjects
```



#The boxplot above shows the different Races and their overall test scores of this data set. Looks like the (white) column shows the max of 340 which is the highest out of all 4 Races that were acounted for. We can also see that column (white) and column (asain) both have the same median in scores, while column (hispanic) and column (african American) roughly share the same median also

4. Conclusion

#In conclusion with all the data and graphs shown we can conclude that a high amount of male white students will overall show high scores in testing. The highest race was white and highest gender was male. Clearly this reflects off of the scores for overall population of both male and female participants. This can indicate that your scores might differ if you are not white or male. It was interesting to see the amount of females averaging around a score of 50 throught all subjects while males have scored higher. We can agree that with an increase in gender, wont increase test score as this data set had a higer amount of females than males.