Exploratory Data Analysis

Naomi Jones

2/24/2022

```
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(ggplot2)
library(forcats)
library(RColorBrewer)
library( magrittr)
```

```
hsb2 <- read.table("C:/Users/nomer/Desktop/math130/data/hsb2.txt", header=TRUE, sep="\t")
dim(hsb2)
```

```
## [1] 200  11
```

INTRODUCTION:

I will be using the High School and Beyond data package (hsb2) for my Exploratory Data Analysis. This set includes 200 observations and 11 variables. According to the Data website provided by DR. D., this set was the second study conducted in a series as a part of NCES' National Longitudinal Studies Program established to study patterns in development of young people throughout their primary and secondary school years and as they transition into adulthood. In my analysis I will be looking at the variables of race and gender specifically to interpret which (if any) of the two variables contribute to higher scores in science. Does the gender of the student equate to a higher science score? Does the race? Or is there no difference in scientific achievement between the two variables?

UNIVARIATE EXPLORATION:

#First, let's summarize the observed scores in science to see what we will be comparing our variables of interest to:

```
summary(hsb2$science)
```

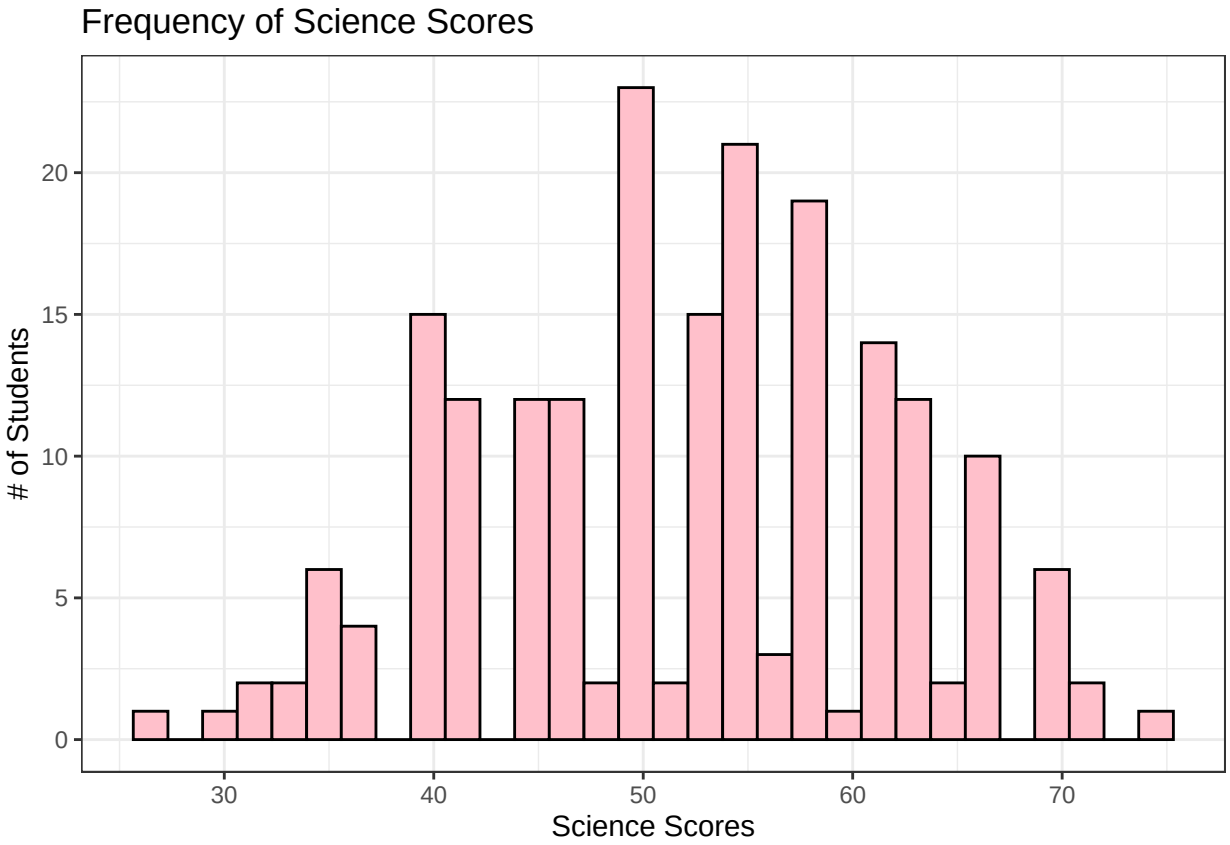
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   44.00   53.00   51.85   58.00   74.00
```

From this summary, we see a range of values that we can use to compare our later findings to. The data shows that the lowest score in science was a 26 whereas the highest was a 74. The mean is 53. When comparing our later findings, we will know that scores closer to 74 would indicate a correlation between science and the variable of interest.

#Next, let's create a histogram of these findings to better visualize the data:

```
ggplot(hsb2, aes(x=science, fill=science)) +
  geom_histogram(colour="black", fill="pink") + ggtitle("Frequency of Science Scores") +
  ylab("# of Students") + xlab("Science Scores") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This histogram shows the frequency of scores in science received by students/the number of students who received that score. The data almost follows a normal distribution pattern.

#Now lets examine gender and science to see if there are any obvious trends in scores:

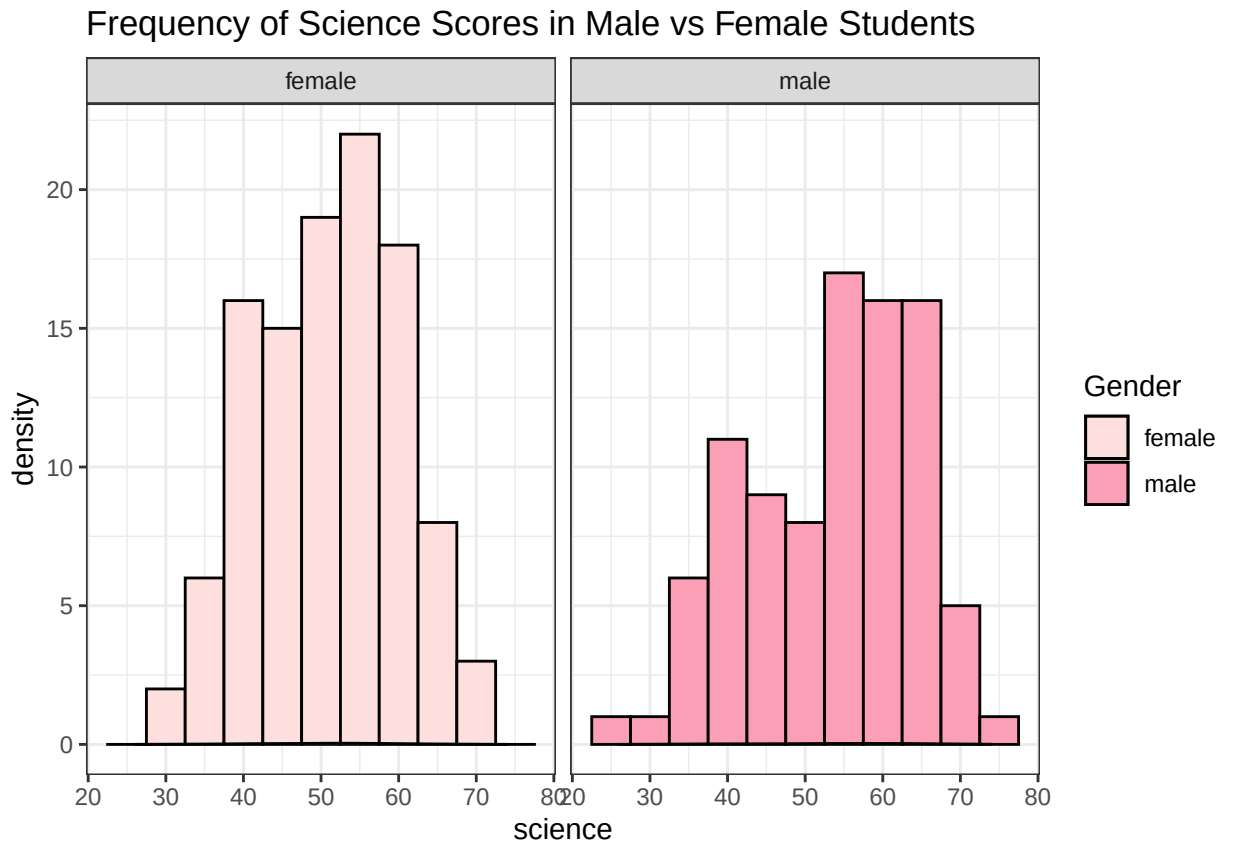
```
table(hsb2$gender, hsb2$science)
```

```
##
##          26 29 31 33 34 35 36 39 40 42 44 45 46 47 48 49 50 51 53 54 55 56 57
##   female  0  1  1  1  3  1  1  6  2  8  8  0  1  6  2  0 15  2  8  3 10  1  0
##   male    1  0  1  1  2  0  3  7  0  4  3  1  0  5  0  2  6  0  7  0  8  1  1
##
##          58 59 61 63 64 65 66 67 69 72 74
##   female 10  1  7  4  0  0  3  1  3  0  0
##   male    9  0  7  8  1  1  6  0  3  2  1
```

This table shows the number of males and females that received each score in science. The minimum score for females was 29 and the maximum was 69. The minimum score for males was 26 and the maximum was 74.

#We can also put this table into histograms:

```
ggplot(hsb2, aes(x=science, fill=gender)) + geom_histogram(colour="black", binwidth=5) +
  ggtitle("Frequency of Science Scores in Male vs Female Students") +
  geom_density(alpha=.5) +
  facet_wrap(~gender) +
  theme_bw() + scale_fill_brewer(name="Gender", palette = "RdPu")
```



This histogram of frequency of science scores in male vs female students allows for easier visualization of the data from the earlier table.

#Now lets examine race and science to see if there are any obvious trends in scores:

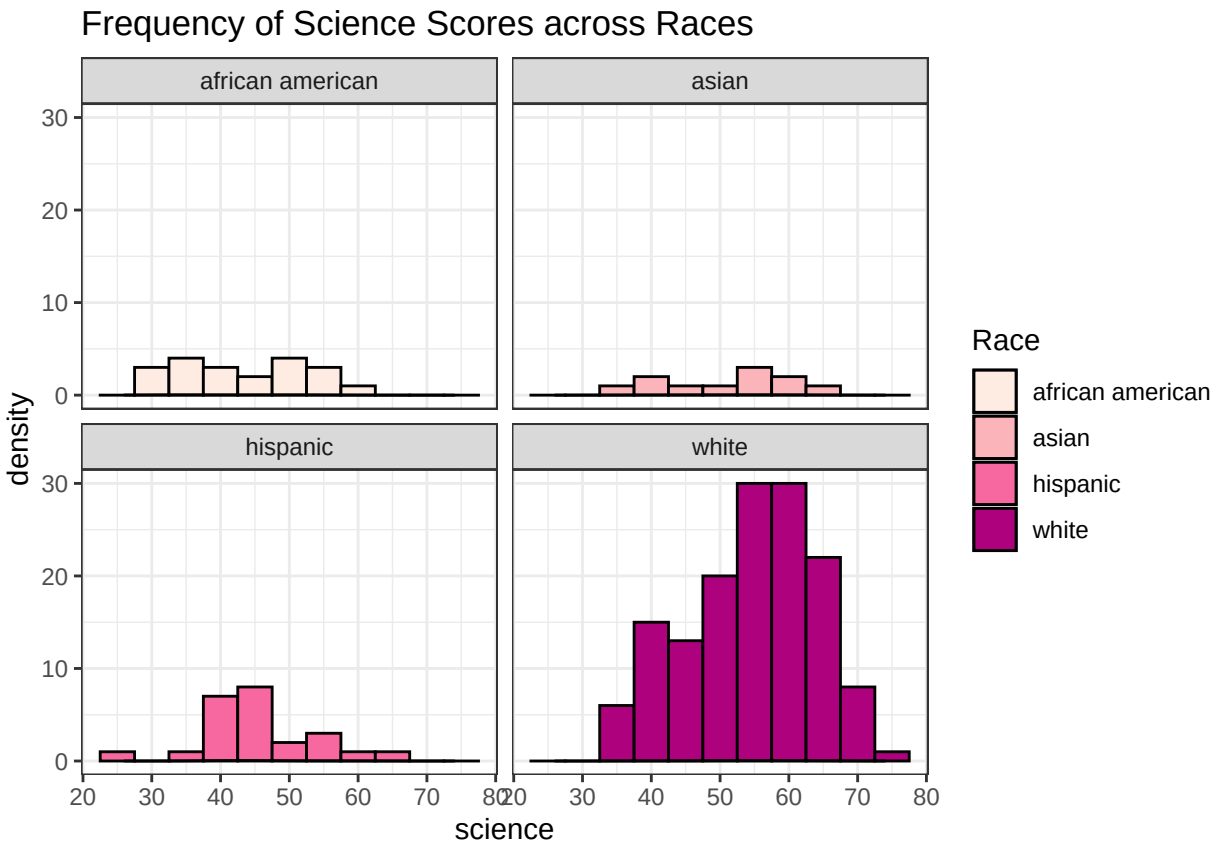
```
table(hsb2$race, hsb2$science)
```

```
##
##                   26 29 31 33 34 35 36 39 40 42 44 45 46 47 48 49 50 51 53 54
##   african american  0  1  2  1  2  1  0  2  0  1  2  0  0  0  0  1  3  0  2  0
##   asian             0  0  0  0  1  0  0  0  0  2  0  0  0  1  0  0  1  0  0  1
##   hispanic          1  0  0  0  0  0  1  4  1  2  5  1  0  2  0  0  2  0  1  0
##   white             0  0  0  1  2  0  3  7  1  7  4  0  1  8  2  1 15  2 12  2
##
##                   55 56 57 58 59 61 63 64 65 66 67 69 72 74
##   african american  1  0  0  0  0  1  0  0  0  0  0  0  0  0
##   asian             1  0  1  1  0  1  0  0  0  1  0  0  0  0
##   hispanic          1  1  0  0  0  1  1  0  0  0  0  0  0  0
##   white            15  1  0 18  1 11 11  1  1  8  1  6  2  1
```

This table shows the number of each race that received each score in science. African American students received a minimum score of 29 and a maximum score of 64. Asian students received a minimum score of 34 and a maximum of 66. Hispanic students received a minimum score of 26 and a maximum score of 63. White students received a minimum score of 33 and a maximum score of 74.

#We can also put this table into histograms:

```
ggplot(hsb2, aes(x=science, fill=race)) + geom_histogram(colour="black", binwidth=5) +
  ggtitle("Frequency of Science Scores across Races") +
  geom_density(alpha=.5) +
  facet_wrap(~race) +
  theme_bw() + scale_fill_brewer(name="Race", palette = "RdPu")
```

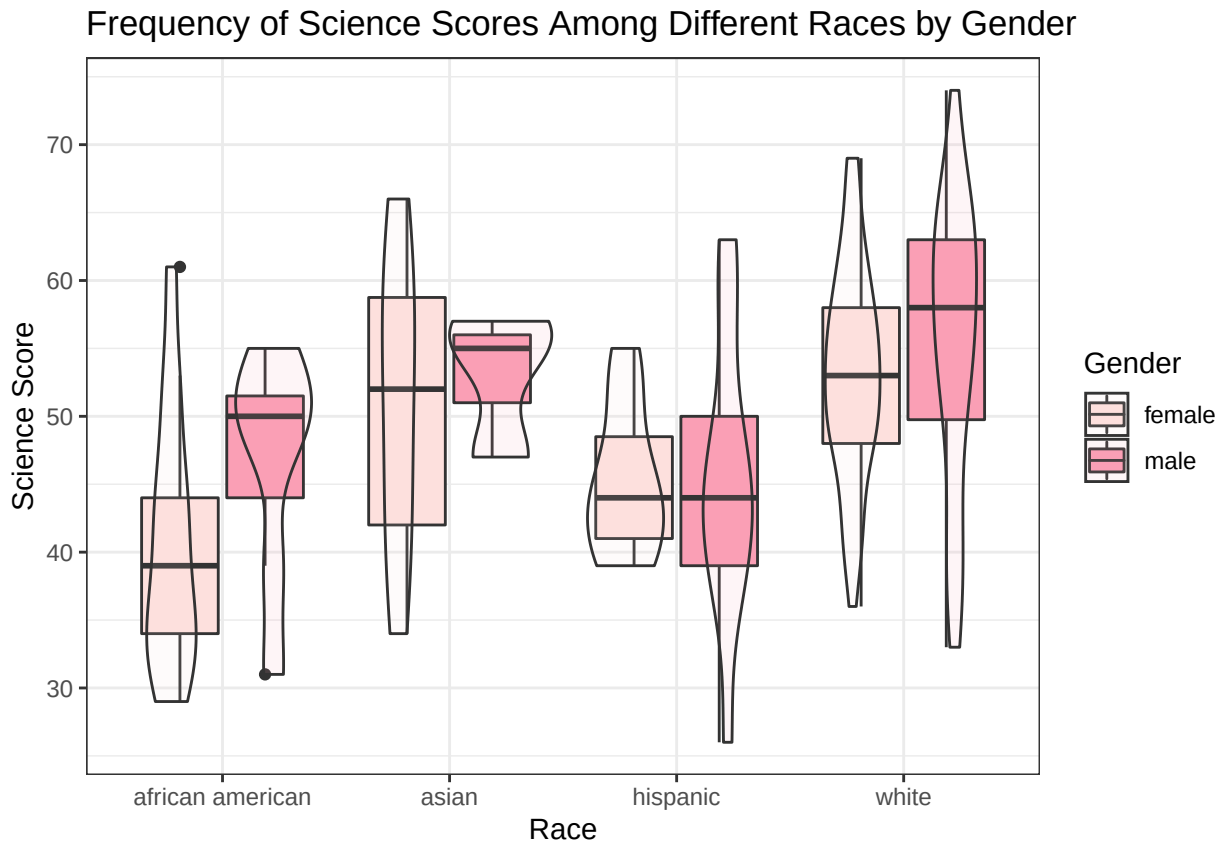


In looking at this histogram, it is immediately obvious that there are far more data plots for white students than any other racial category. This makes comparison of the values more difficult as there is more sample data for one race which leads to the data being visually skewed and harder to come to an accurate conclusion.

BIVARIATE EXPLORATION:

#Now lets combine the three variables into a boxplot for easier comparison:

```
ggplot(hsb2, aes(y=science, x=race, fill=gender)) +
  ggtitle("Frequency of Science Scores Among Different Races by Gender") +
  geom_boxplot() + geom_violin(alpha=.1) +
  theme_bw() + scale_fill_brewer(name="Gender", palette = "RdPu") +
  ylab("Science Score") + xlab("Race")
```



This boxplot shows the frequency of scores among different racial groups coupled with gender. Upon first look of the boxplots, one may come to the conclusion that white males seemed to score the highest in science. That being said, this would overall be an inaccurate conclusion for a number of reasons. Firstly, there were far more white participants than any other race, skewing the data. There is also major overlap in the boxplots with no clearly distinguishable leader.

#Lets look at the number of students for each race:

```
table(hsb2$race)
```

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

As was alluded to in previous graphs/plots, the number of white students far outweighed the other races with whites accounting for 145 participants while the other three categories did not even make 3 digits.

Conclusion

In conclusion, the results of the analysis are inconclusive and from the data set provided, it appears that neither race nor gender are a better indicator of how a student will score in science. There is far too much overlap betweem the boxplots to truly conclude that a singular race or gender equated to a high score in science. In order to come to a solid/significant conclusion, the sample data would have to be more evenly distributed across all races and other factors would have to be considered.