Exploratory Data Analysis

COLETTE COX

02/22/2021

1. Introduction/Description: High School and Beyond

```
hsb2 <- read.delim("/Users/colettecox/Downloads/MATH130/hsb2.txt", sep="\t")
dim(hsb2)
```

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

```
schsci <- select(hsb2, science, schtyp)
```

I chose the data set HS and Beyond for my Exploratory Data Analysis. This entire data set sought to study the educational, vocational and personal development of young people from elementary/highschool years and following them overtime as they become adults and assume adult roles and responsibilities. There are 200 observations of 11 different variables in the entire data set, but I will scale down my data analysis to focus on two variables instead. The two variables that I will be focused on for my particular analysis of the data will be school type (schtyp) and science aptitude (science).

2. Univariate Description of 2 variables:

Science

a. Summary Statistics for participants' Science Aptitude scores:

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

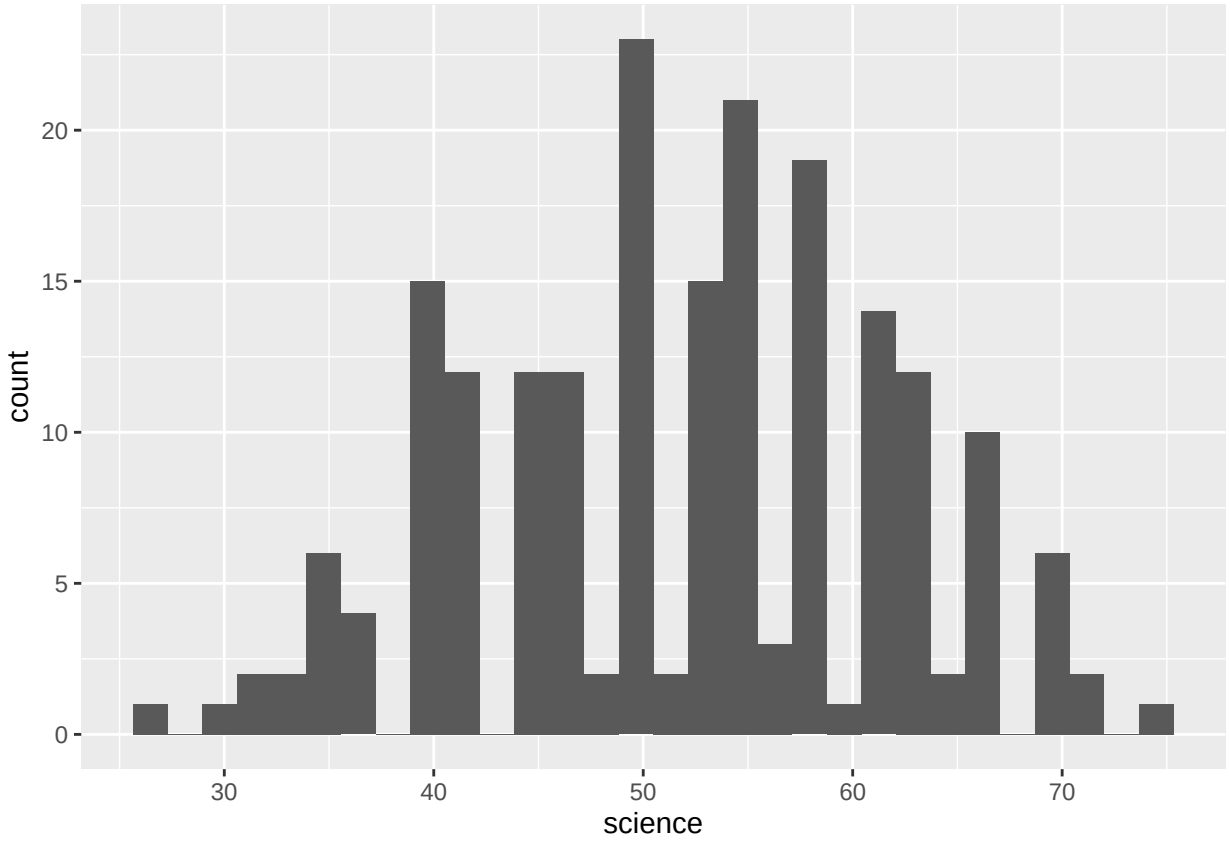
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   44.00   53.00   51.85   58.00   74.00
```

Above is the summary statistics for the Science Aptitude scores from 200 participants. The minimum score was 26 and the maximum score was 74. The mean score of this data is 51.85 (although this test was not scored based on fractional numbers) and the median score was 53. Since the median is slightly higher than the mean, this would suggest that perhaps the data is left-skewed, meaning there are some data points less than the median value that are dragging the average score to be lower. To explore further we would need a visual, as shown below.

b. Histogram for participants' Science Aptitude scores:

```
ggplot(schsci, aes(x=science)) + geom_histogram()
```

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

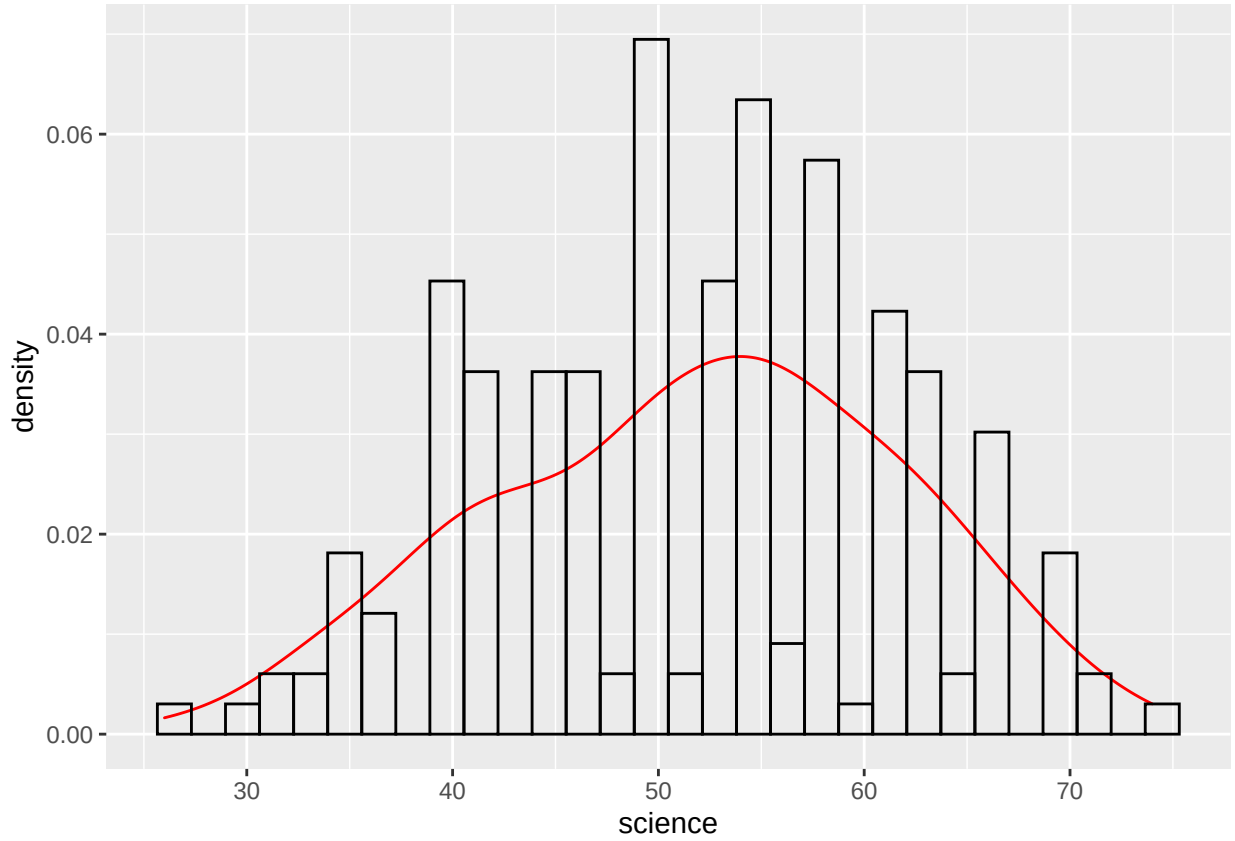


Now that we have the variable science aptitude scores visualized on our x-axis with corresponding counts on a histogram, we can better see how our data could be possibly skewed based on the greater concentration of bins on the right half of about 50 (near our median/mean scores). Yet, this visual representation alone does not paint a clear picture to suggest that our data is not normally distributed around our mean value. Another option to further explore this variable is denoted below.

c. Histogram + Density Plot for participants' Science Aptitude scores:

```
ggplot(schsci, aes(x=science)) + geom_density(col="red") + geom_histogram(aes(y=..density..), colour="bl
```

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



In this visual, we have a histogram of the same data with an added density plot (denoted by the traced red line) on top of the graph. Although the skew is not drastic, we can more clearly see how the right half of the density plot is steeper than the left half of the density plot. This graphic also illuminates how our median value is at about a score of 53, corresponding to the peak of the red density curve.

School Type

a. Table of School Types of the participants:

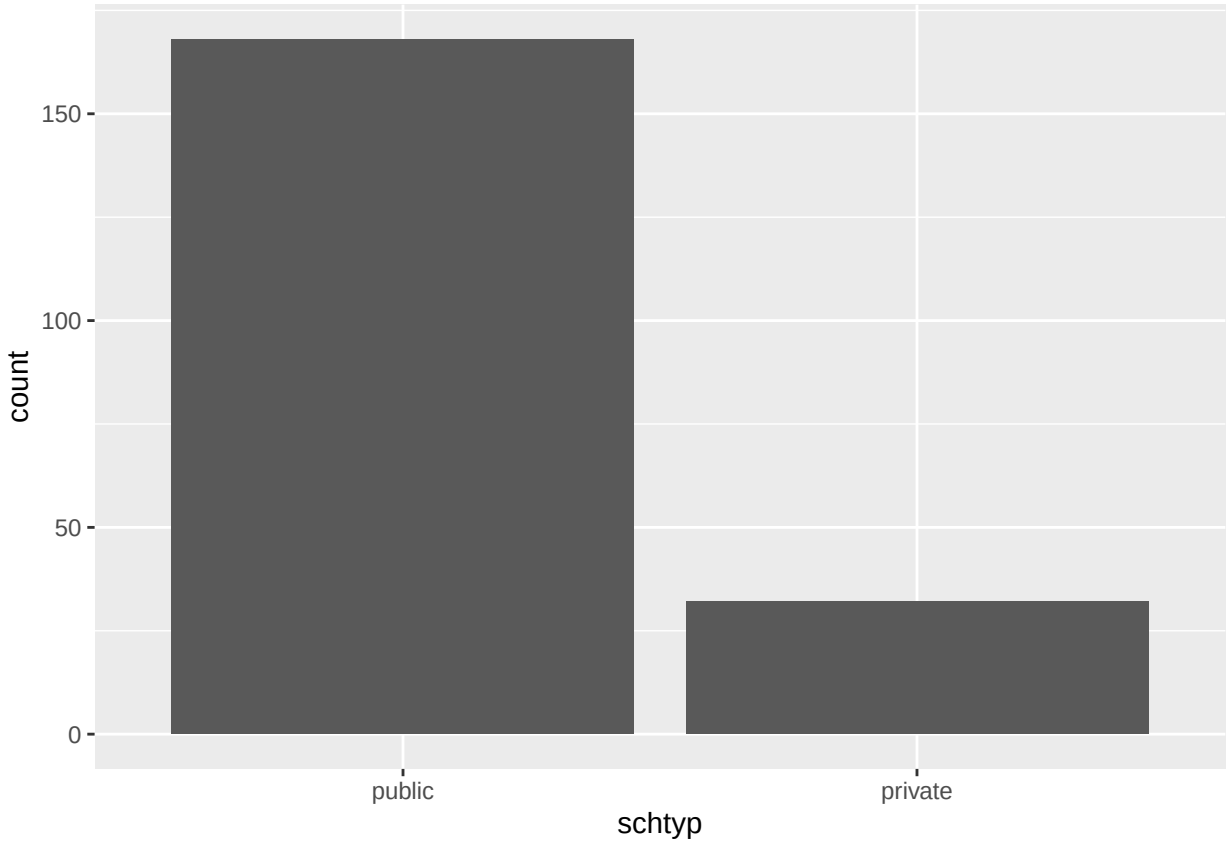
```
table(schsci$schtyp)
```

```
##
## private  public
##      32     168
```

Based on the data illustrated in the table above, we can conclude that 168 participants went to public school while only 32 participants went to private school for their education.

b. Barchart of School Types of the participants:

```
ggplot(schsci, aes(x=forcats::fct_infreq(schtyp)))+geom_bar()+xlab("schtyp")
```



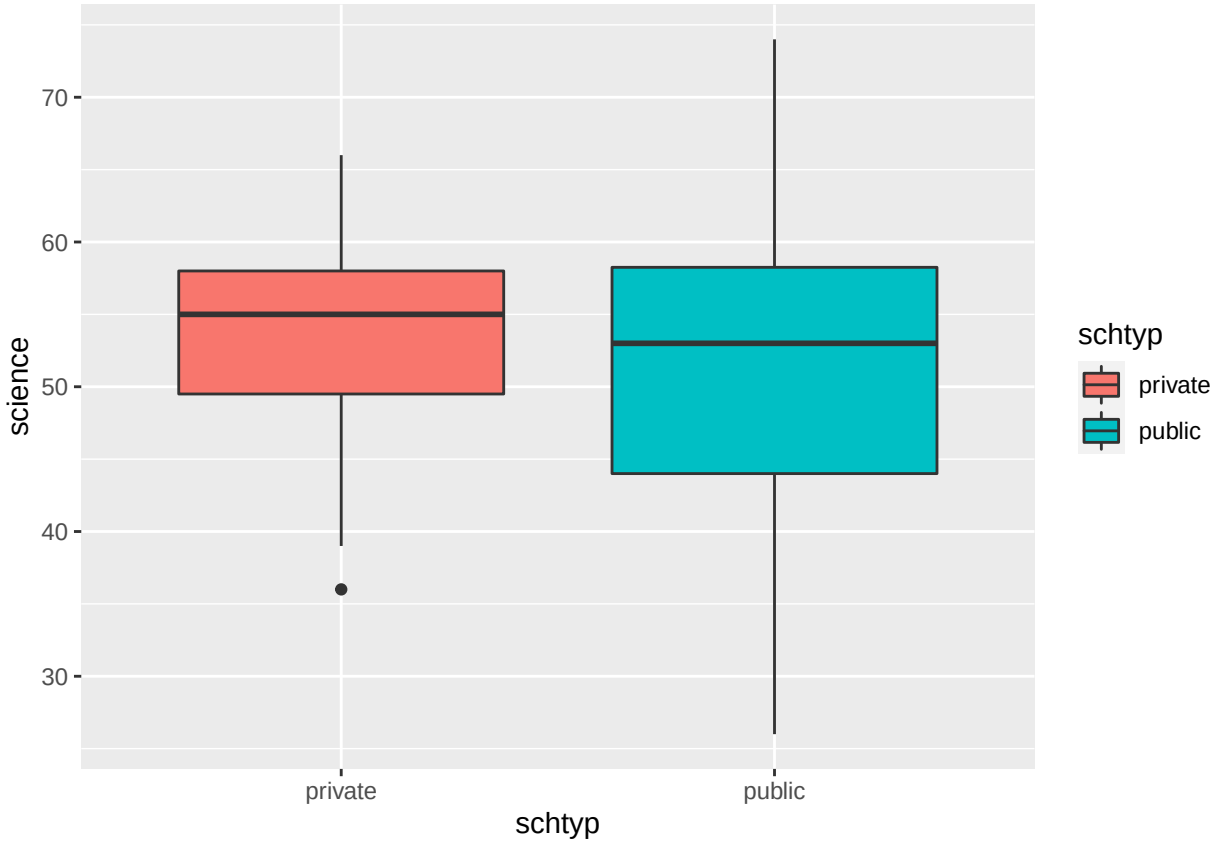
In the barchart above, we can see the variable School Type (denoted through code as schtyp) separated into private and public schooling, with their respective participant counts. This helps us visualize how much more of the 200 sampled participants went to public school compared to the much fewer who went to private school.

3. Bivariate Comparison of variables:

School Type vs. Science Aptitude Score

a. Grouped Boxplot

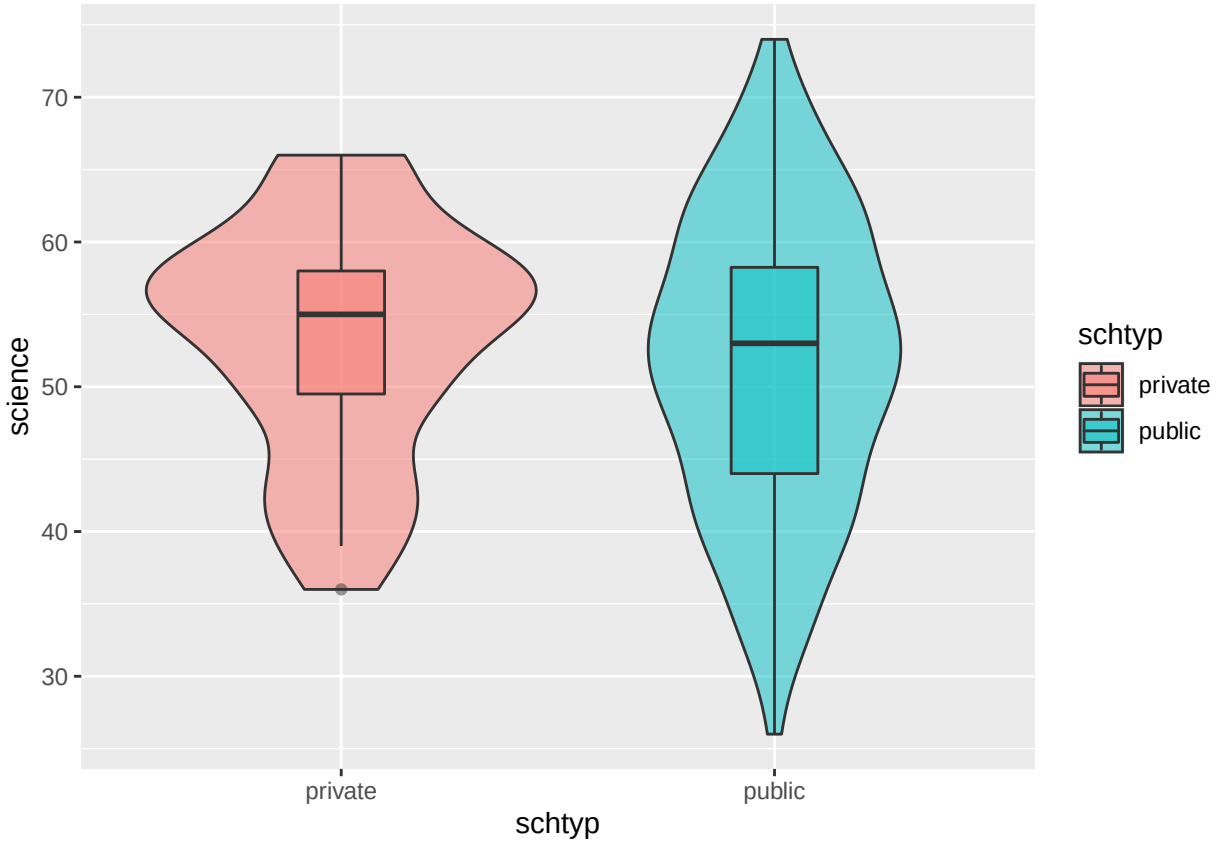
```
ggplot(schsci, aes(x=schtyp, y=science, fill=schtyp))+geom_boxplot()
```



In the graphic above, we see a boxplot illustrating how the different school types (public and private school) compared in their student participants' science aptitude scores. Although each school had approximately similar means, the greatest range in scores was observed in the public school science scores. The private school scores had a much smaller range, even denoting an outlier within the lower bounds of their data. One would think that private school students would score highest on a science aptitude test, the data actually shows the highest score belonging to the public school participants, despite how most private school students scored within the 25th percentile of public school scores.

b. Grouped Boxplot with Violins

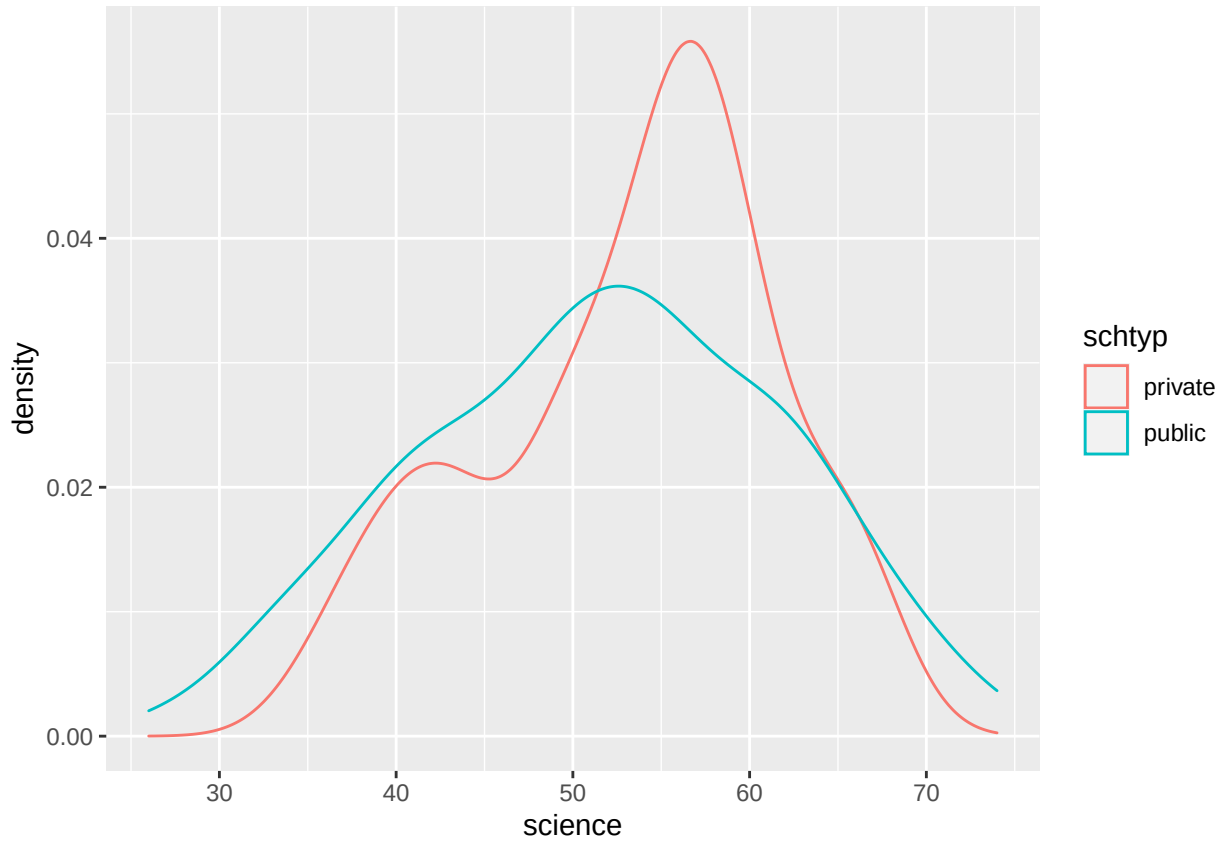
```
ggplot(schsci, aes(x=schtyp, y=science, fill=schtyp))+geom_violin(alpha=0.5)+geom_boxplot(alpha=0.5, wi
```



This type of boxplot allows us to better see the density of the data presented visually, and therefore get a better sense of the distribution of these two variables. The tighter range of the private school science aptitude scores is better seen through the overlaid violin plots, as well as the more spread distribution of public school scores. The private school scores also illustrate more of a skewed distribution, while public school scores illustrate an approximately normal distribution, aside from how flat it appears.

c. Density Plot of School Type vs. Science Aptitude Score

```
ggplot(schsci, aes(x=science, col=schtyp))+geom_density(alpha=0.5)
```



This final graph depicts a denisty plot of science aptitude scores, based on whether they were earned from a participant from private or public school. This model display would suggests that while public school participant scores followed a farily unimodal distribution, the private school participant scores appeared to have an almost bimodal distribution at the lower and upper bounds of the data.