EDA_rgrenard

Rene Grenard

9/27/2021

```
knitr::opts_chunk$set(warning=FALSE, message=FALSE, fig.height=4, fig.width=10,
  fig.align='center')
library(ggplot2)
library(dplyr)
```

Introduction

According to Dr.D's website: "The High School and Beyond (HS&B) Longitudinal Study was the second study conducted as part of NCES' National Longitudinal Studies Program. This program was established to study the educational, vocational, and personal development of young people, beginning with their elementary or high school years and following them over time as they take on adult roles and responsibilities." From this dataset, I will explore the relationship between reported math t-scores, socioeconomic status, and school type to determine if this dataset shows elevated scores among those of higher socioeconomic status and those attending private schools.

```
# This code reads in the data and assigns it the name 'school'
school <- read.table("/Users/rgrenard/Desktop/Math 130/Data/hsbeyond.txt",
  header=TRUE, sep="\t")
```

Univariate Plots

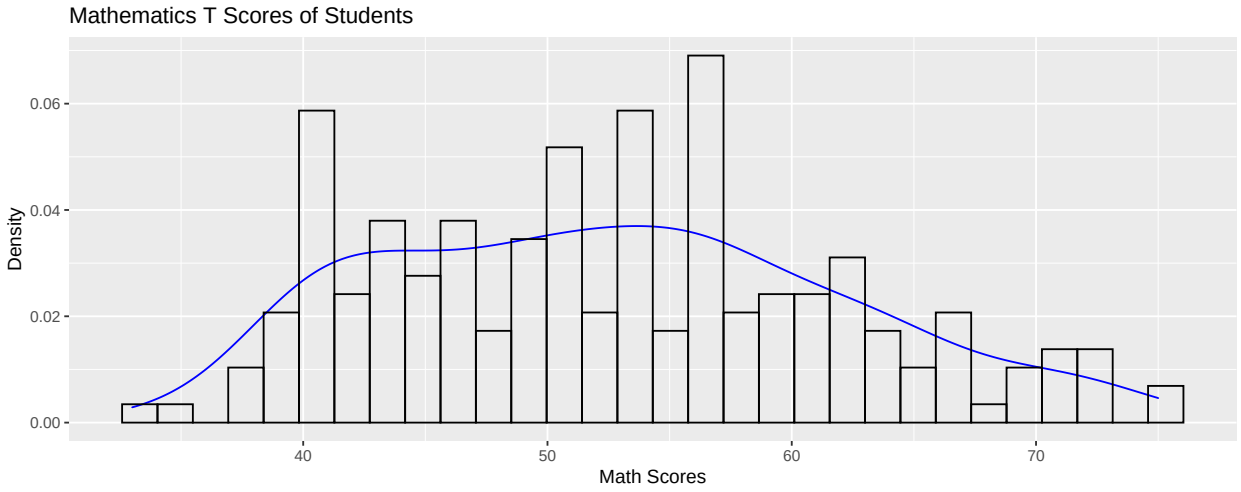
Math Variable

```
# This prints a summary table of the 'math' variable
summary(school$math)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33.00   45.00   52.00   52.65   59.00   75.00
```

This `math` variable is the reported t-scores of the participants. This continuous, numerical variable was chosen as a means of comparison.The higher the score, the greater the student improved upon test scores. This summary data above shows the mean of the scores to be 52.65, the median to be 52.00, and the IQR to be 14.00.

```
# Prints a density and histogram plot
ggplot(school, aes(x=math)) + geom_density(col="blue") +
  geom_histogram(aes(y=..density..), colour="black", fill=NA) +
  ggtitle("Mathematics T Scores of Students") + ylab("Density") + xlab("Math Scores")
```



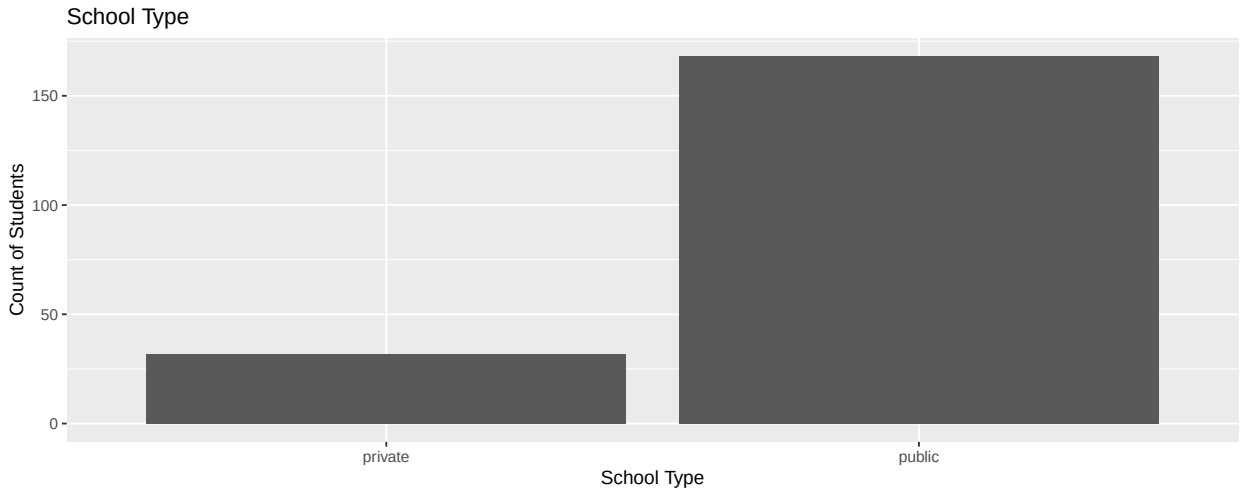
The above graph is a histogram overlaid with a density plot of Math Scores. We see there are a few extended scores near 40, 50, 54, and 56.

School Type

```
# Prints a table showing the total of students in each school type
table(school$schtyp)
```

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

```
# Prints a bar chart of School Types
ggplot(school, aes(x=schtyp)) + geom_bar() + ggtitle("School Type") +
  ylab("Count of Students") + xlab("School Type")
```



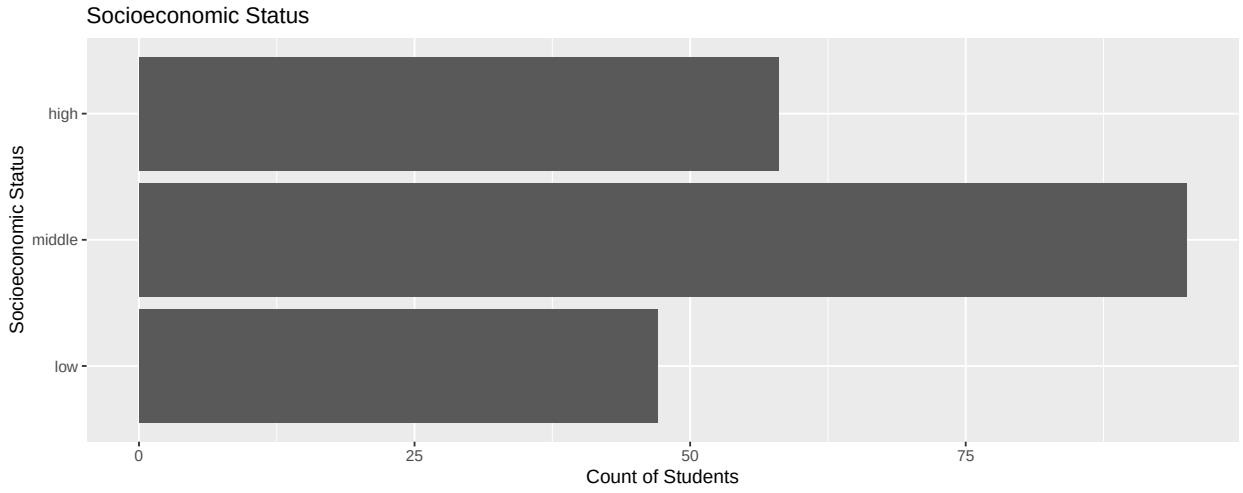
The `School Type` variable contains 2 levels: private and public. There from the table above, we see there are 32 students who attended private schools and 168 who attended public schools. The above bar chart gives a visual representation of the significant difference between students.

Socioeconomic Status

```
# Prints a table showing the total number of students in each socioeconomic status
table(school$ses)
```

```
##
##   high    low middle
##     58     47     95
```

```
# Prints a horizontal bar chart for Socioeconomic Status
soci <- factor(school$ses, levels=c("low", "middle", "high"))
ggplot(school, aes(x=soci)) + geom_bar() + ggtitle("Socioeconomic Status") +
  ylab("Count of Students") + xlab("Socioeconomic Status") + coord_flip()
```

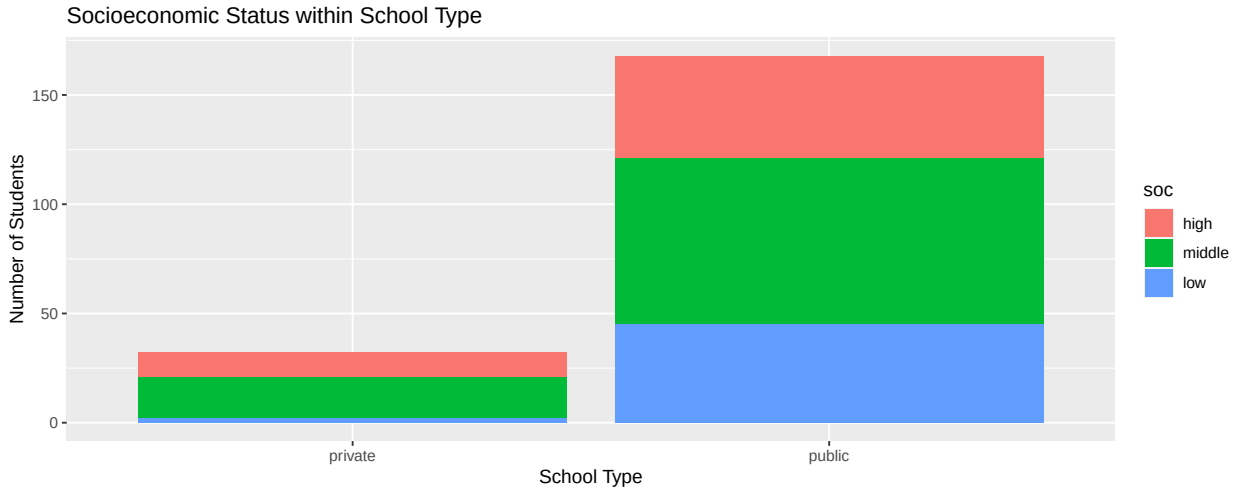


The `Socioeconomic Status` variable contains 3 levels: low, middle, and high. There from the table above, we see there are 47 low status, 95 middle status, and 58 high status students. The above bar chart gives a visual comparison of each. We see the majority of students fall in the middle category.

Bivariate

Socioeconomic Status vs School Type

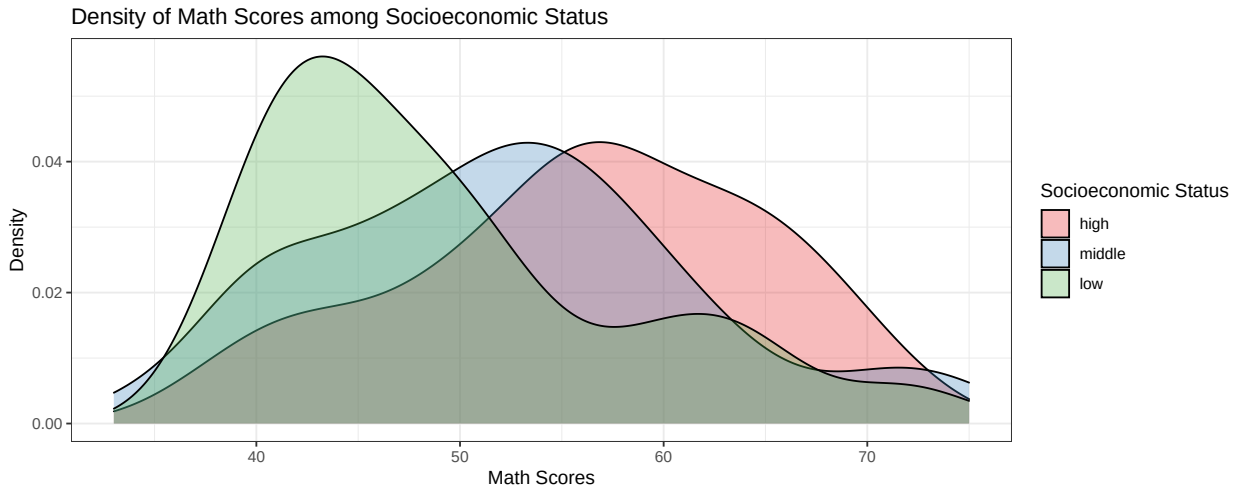
```
# Bar Plot comparing socioeconomic status within school type
soc <- factor(school$ses, levels=c("high", "middle", "low"))
ggplot(school, aes(x=schtyp, fill=soc)) + geom_bar() +
  ggtitle("Socioeconomic Status within School Type") +
  ylab("Number of Students") + xlab("School Type")
```



The above bar chart compares the number of low, middle, and high status students within both private and public schools. Notice the distribution of student in public schools are fairly similar. However, the low status students within the private school sector in this data set is disproportionately low.

Socioeconomic Status vs Math T Scores

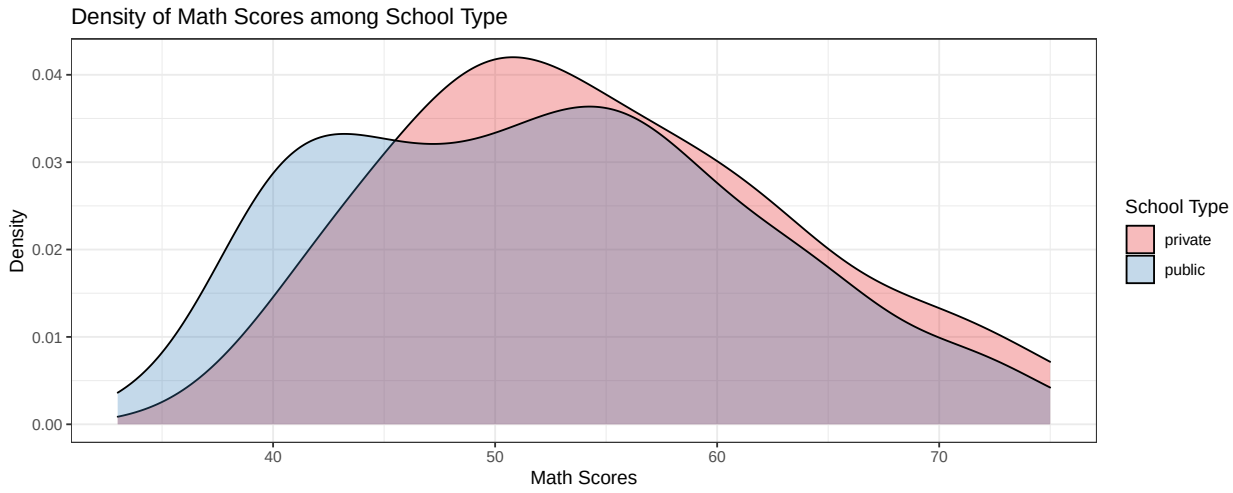
```
# Density graph comparing Socioeconomic Status and Math Scores
ggplot(school, aes(x=math, fill=soc)) + geom_density(alpha=.3) + theme_bw() + scale_fill_brewer(palette=
```



As we can see, the low income scores have an elevated density near the 43 score. This is low as compared to the middle and high status peaks.

Math Scores vs School Type

```
# Density graph comparing School Type and Math Scores
ggplot(school, aes(x=math, fill=schtyp)) + geom_density(alpha=.3) + theme_bw() + scale_fill_brewer(pale
```

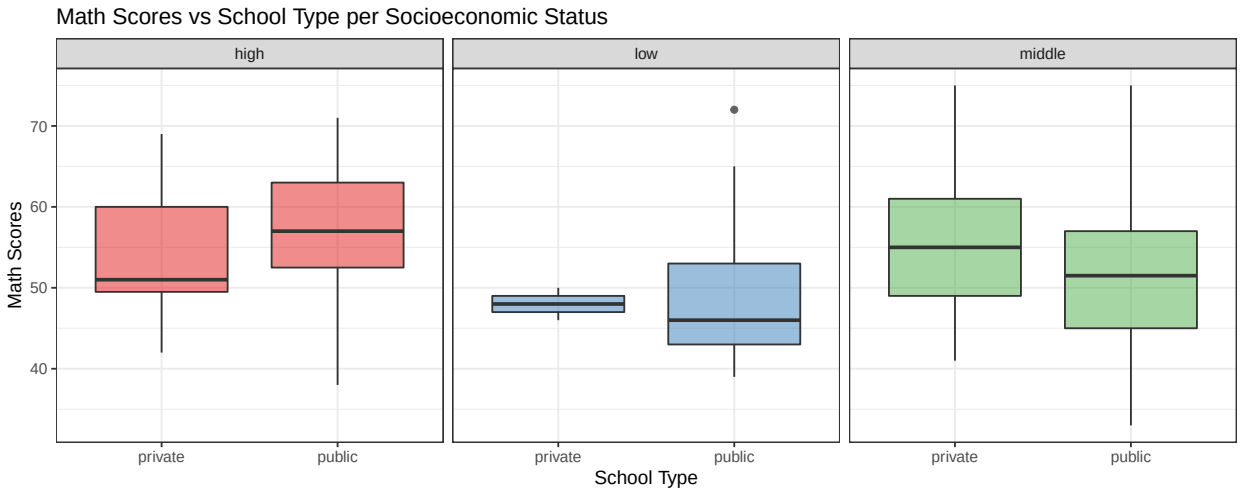


Both public and private schools appear to have very similar density distributions. The private school has a slightly higher peak near the 50 score range.

Multivariate

Math Scores, School Type, and Socioeconomic Status

```
# Wrapped boxplot comparing all 3 variables
ggplot(school, aes(x=math, y=schtyp, fill=ses)) + geom_boxplot(alpha=.5) +
  facet_wrap(~ses,scales="free_x") + theme_bw() +
  scale_fill_brewer(palette="Set1", guide="none") + coord_flip() +
  ggtitle("Math Scores vs School Type per Socioeconomic Status ") +
  xlab("Math Scores") + ylab("School Type")
```



The above boxplot gives an excelent visual representation of the differences in distributions of math scores between public and private schools per socioeconomic status. We see that students of low socioeconomic status peak with scores just under 50. This is the lowest of the three statuses. Also, the public school provides the lowest scores for the socioeconomic status, but is higer than the private schools for the high socioeconomic status. It could be worth looking into the reason for this. A possible question to ask could be: Are public schools in higher socioeconomic neighborhoods providing a higher quality education than those in a lower socioeconomic neighborhood.