# M130 Exploratory Data Analysis Project - High School and Beyond

Celine Riehle

2023-09-20

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

## Introduction

The High School and Beyond Longitudinal Study was conducted to study personal, educational, and vocational developments of young individuals, following them into adult life. The two groups of focus were high school seniors and sophomores, with a base-year survey then four follow-up surveys. The data set contains 200 observations with 11 variables. This project will be focusing on socioeconomic status compared to their math scores. Examining these two variables will compare their performance in test scores throughout the years and how each group was affected differently.

```
hsb2 <- read.delim("/Users/celineriehle/Desktop/math130/data/hsb2.txt", sep="\t")
dim(hsb2)
```

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

## Univariate Exploration

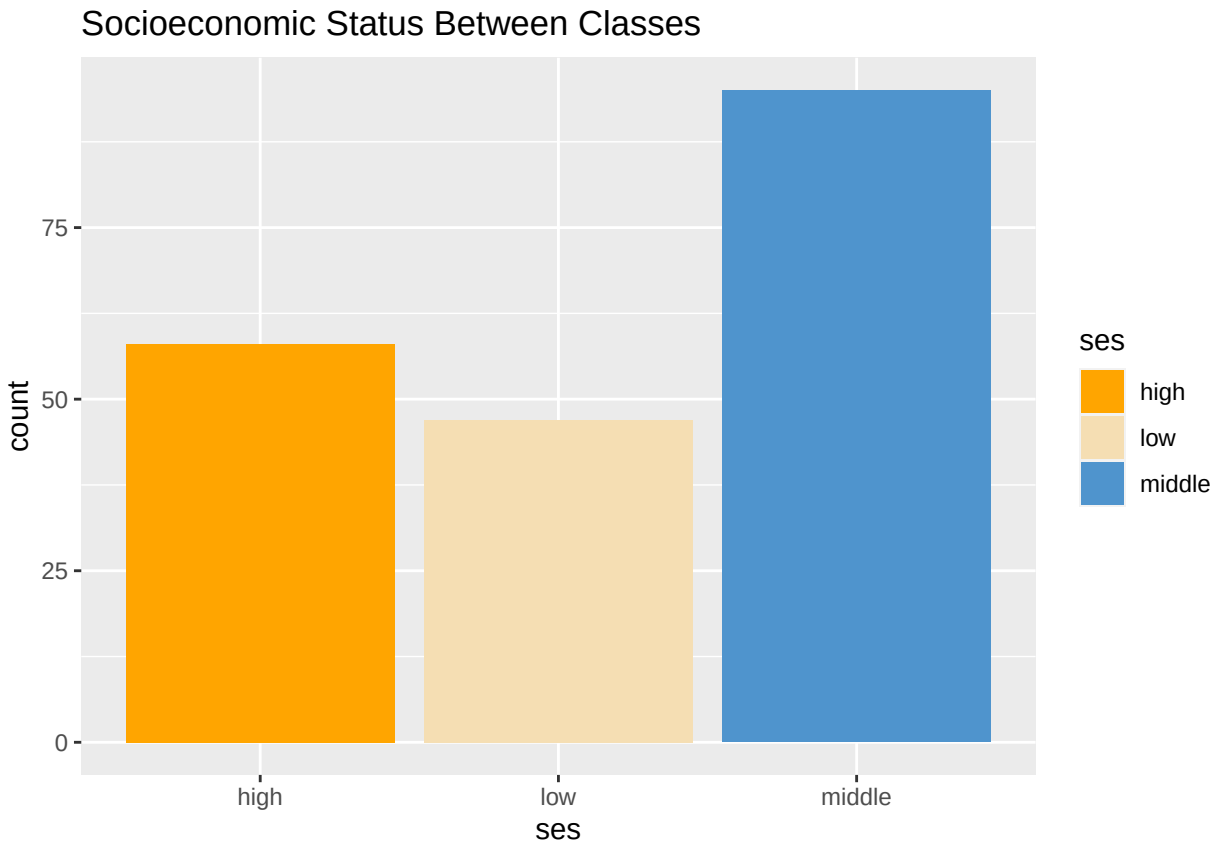
### Socioceonomic Status

```
table(hsb2$ses)
```

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

This table is of socioeconomic status of high, middle, and low class, majority making up the middle.

```
ggplot(hsb2, aes(x=ses, fill=ses)) + geom_bar() + xlab("ses") + ggtitle("Socioeconomic Status Between C
```




The bar chart gives a better visual of each component in decreasing frequency. The category with the least is the low socioeconomic class.

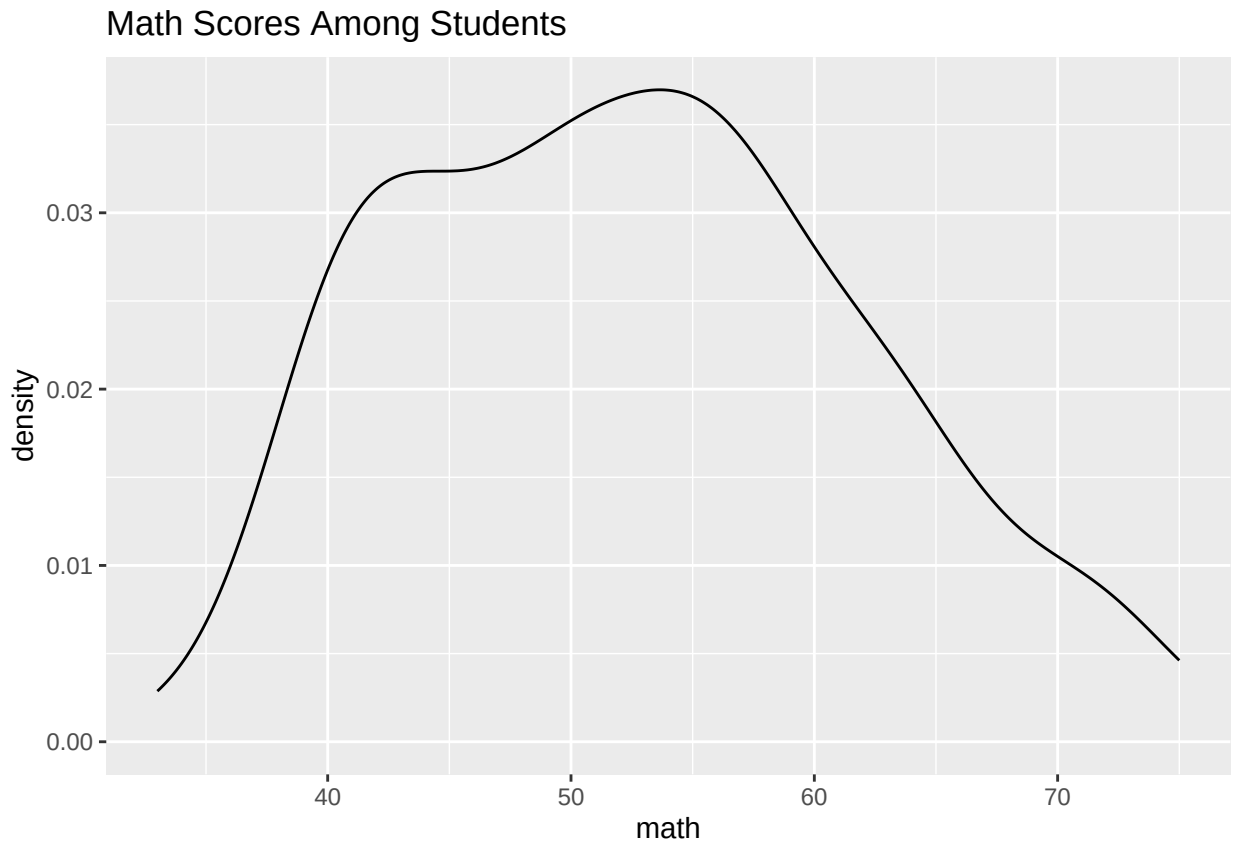
### Math Scores

```
table(hsb2$math)
```

```
##
## 33 35 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
##  1  1  1  2  6 10  7  7  7  4  8  8  3  5 10  7  8  6  7 10  5  7 13  6  2  5
## 61 62 63 64 65 66 67 68 69 70 71 72 73 75
##  7  4  5  5  3  4  2  1  2  1  4  3  1  2
```

The table compares students different math test scores. The math scores range from 33 to 75.

```
ggplot(hsb2, aes(x=math)) + geom_density() + ggtitle("Math Scores Among Students")
```




The density plot is comparing students math scores over all of the surveys taken. The density curve is at its peak around 50-60, averaging the math scores.

```
summary(hsb2$math)
```

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

Students math scores averaged about 52.65% and their median averaged 52%.The third qu. scores were better than the first qu.

## Bivariate Exploration

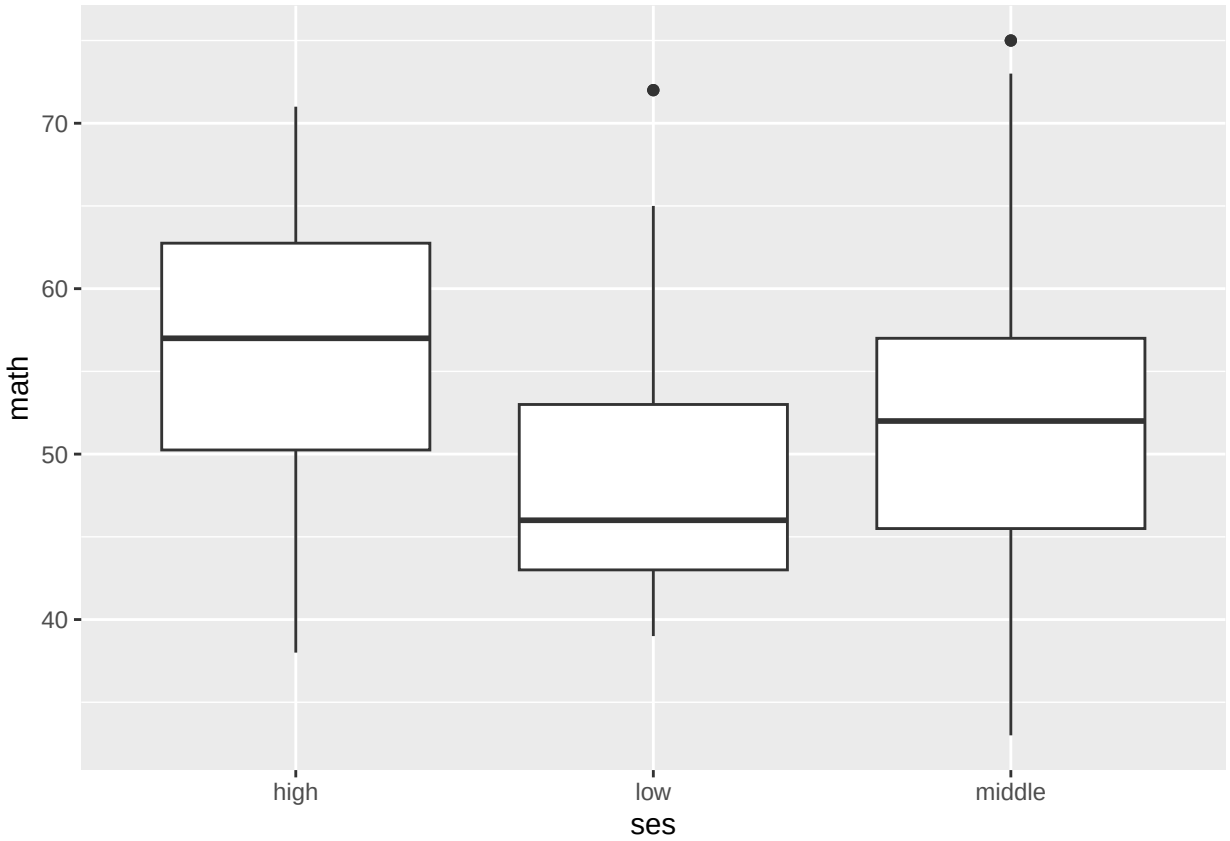
```
table(hsb2$ses, hsb2$math) %>% prop.table(margin=1) %>% round (3)
```

```
##
##              33    35    37    38    39    40    41    42    43    44    45
##   high    0.000 0.000 0.000 0.017 0.034 0.000 0.000 0.052 0.017 0.017 0.017
##   low     0.000 0.000 0.000 0.000 0.043 0.085 0.085 0.021 0.106 0.064 0.043
##   middle 0.011 0.011 0.011 0.011 0.021 0.063 0.032 0.032 0.011 0.000 0.053
##
```

```
##              46    47    48    49    50    51    52    53    54    55    56
##   high   0.000 0.017 0.017 0.017 0.052 0.052 0.000 0.000 0.069 0.017 0.086
##   low    0.064 0.021 0.021 0.085 0.043 0.021 0.021 0.043 0.021 0.000 0.000
##   middle 0.053 0.011 0.032 0.053 0.021 0.042 0.053 0.053 0.053 0.042 0.021
##
##              57    58    59    60    61    62    63    64    65    66    67
##   high   0.103 0.034 0.000 0.017 0.052 0.052 0.017 0.052 0.034 0.034 0.034
##   low    0.000 0.021 0.021 0.021 0.021 0.000 0.021 0.043 0.021 0.000 0.000
##   middle 0.074 0.032 0.011 0.032 0.032 0.011 0.032 0.000 0.000 0.021 0.000
##
##              68    69    70    71    72    73    75
##   high   0.017 0.034 0.000 0.034 0.000 0.000 0.000
##   low    0.000 0.000 0.000 0.000 0.043 0.000 0.000
##   middle 0.000 0.000 0.011 0.021 0.011 0.011 0.021
```

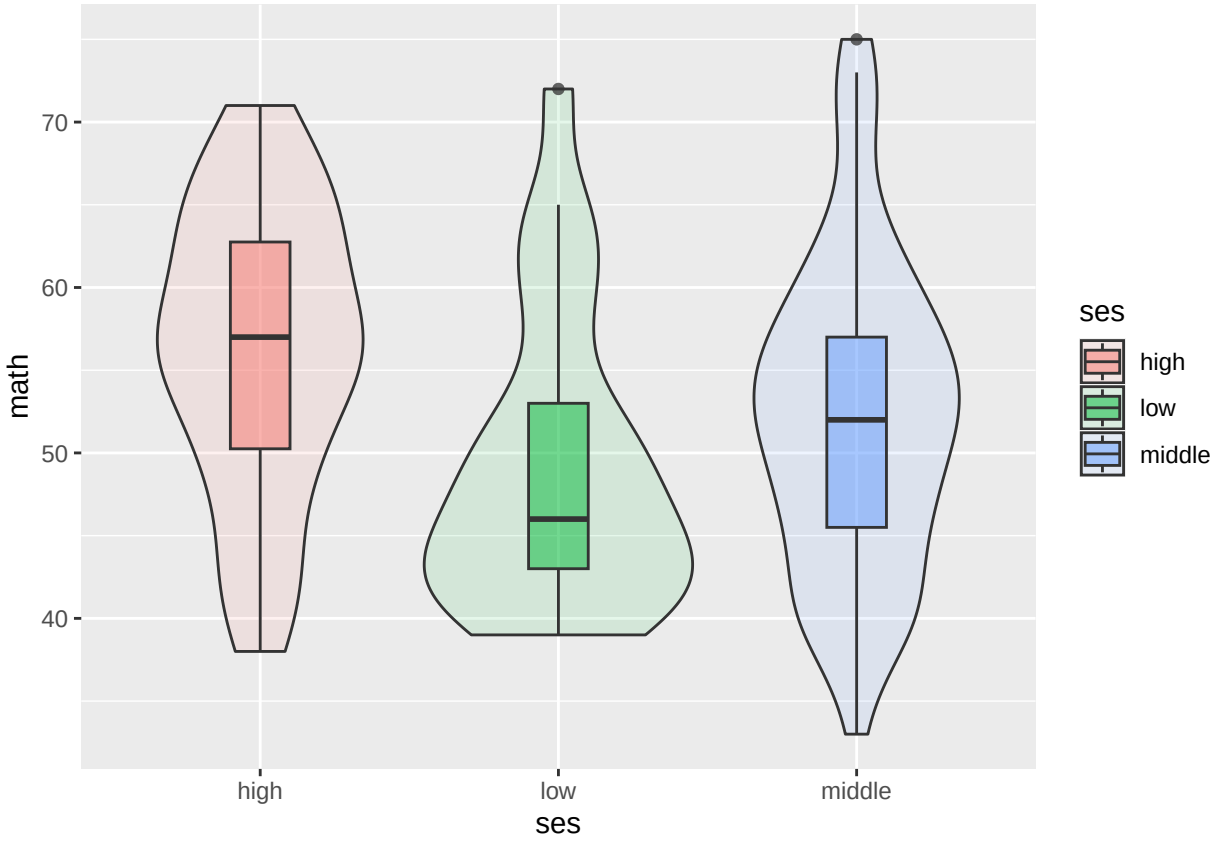
Comparing the two variables students who were in the middle socioeconomic class had the worst and best test scores. Averaging the test scores, students who had 54 were in higher socioeconomic class with 69%.

```
ggplot(hsb2, aes(x=ses, y=math)) + geom_boxplot()
```




In the box plot math scores were overall better in students who were in higher socioeconomic class. In lower socioeconomic status students did worse, also compared to the middle class.

```
ggplot(hsb2, aes(x=ses, y=math, fill=ses)) +
        geom_violin(alpha=.1) +
        geom_boxplot(alpha=.5, width=.2)
```




High socioeconomic class had more equal distribution throughout the math test scores. Students in low socioeconomic class had more low test scores, but not the lowest. The lowest test scores were the middle class, but also had the highest test scores.

## Conclusion

My prior hypothesis would suggest that the higher socioeconomic class would have the highest math scores. The data does not support this hypothesis. Students with a higher socioeconomic status overall did the best in math scores, when compared to lower and middle classes. The distribution between math scores fluctuated though, with the middle class having the lowest and highest math test scores.