EDA_username

Jackson Klein

2025-06-26

```
library(ggplot2)
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
```

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

Introduction:

For my final project, I decided to use the High School and Beyond dataset from Dr. D's teaching course website. The High School and Beyond data set holds information on around 200 students (yes I counted) and includes different variables like gender, race, status, and even test scores for different subjects. I specifically chose this data set because I was interested in finding out whether testing scores differ between different school types and socioeconomic settings. With that being said, the variables I will be using are schtyp (school type), read (reading test scores), and ses (socioeconomic status). And my research question is: How do standardized reading scores vary between different school types and socioeconomic settings?

Univariate Exploration:

Socioeconomic Status:

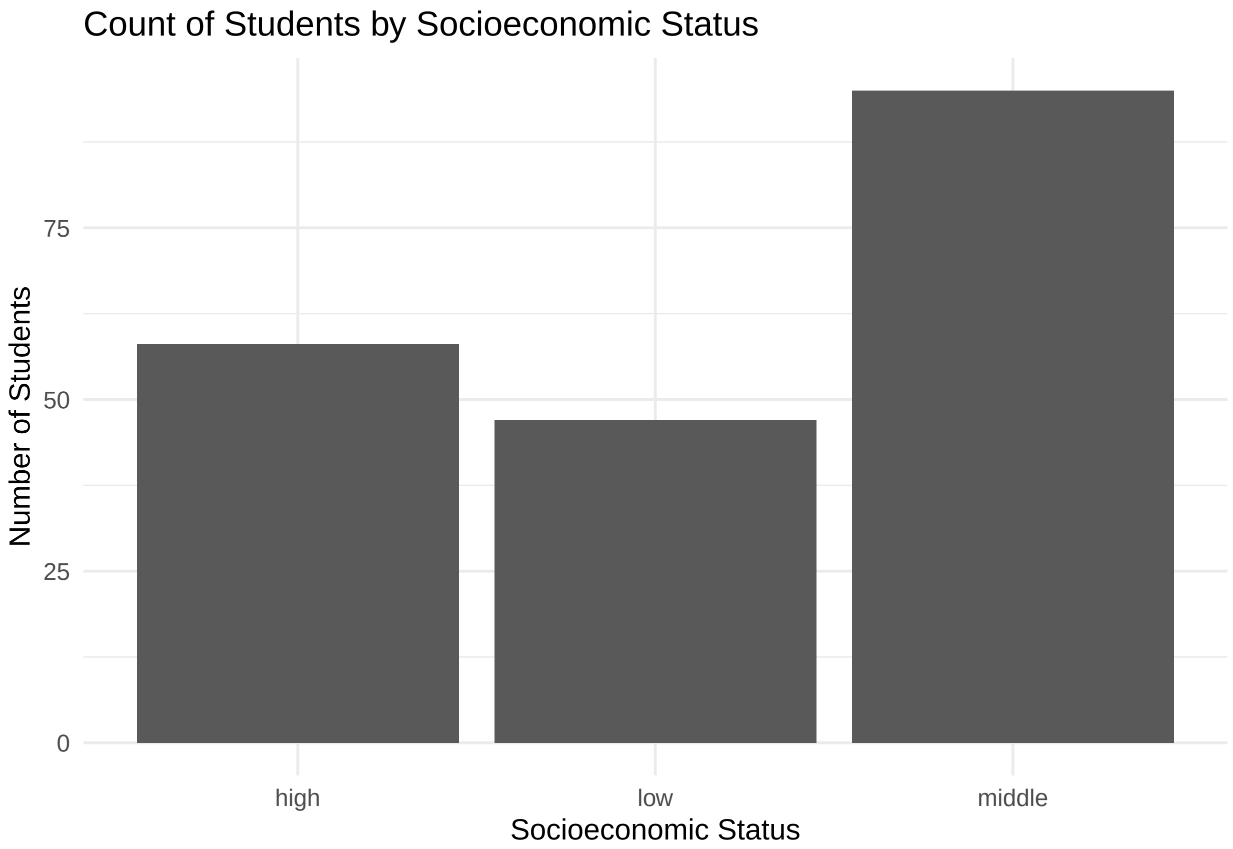
```
table(hsb2$ses)
```

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

```
prop.table(table(hsb2$ses))
```

```
##
##   high    low middle
##  0.290  0.235  0.475
```

```
ggplot(hsb2, aes(x = ses)) + geom_bar() + labs(title = "Count of Students by Socioeconomic Status", x =
```



School Type:

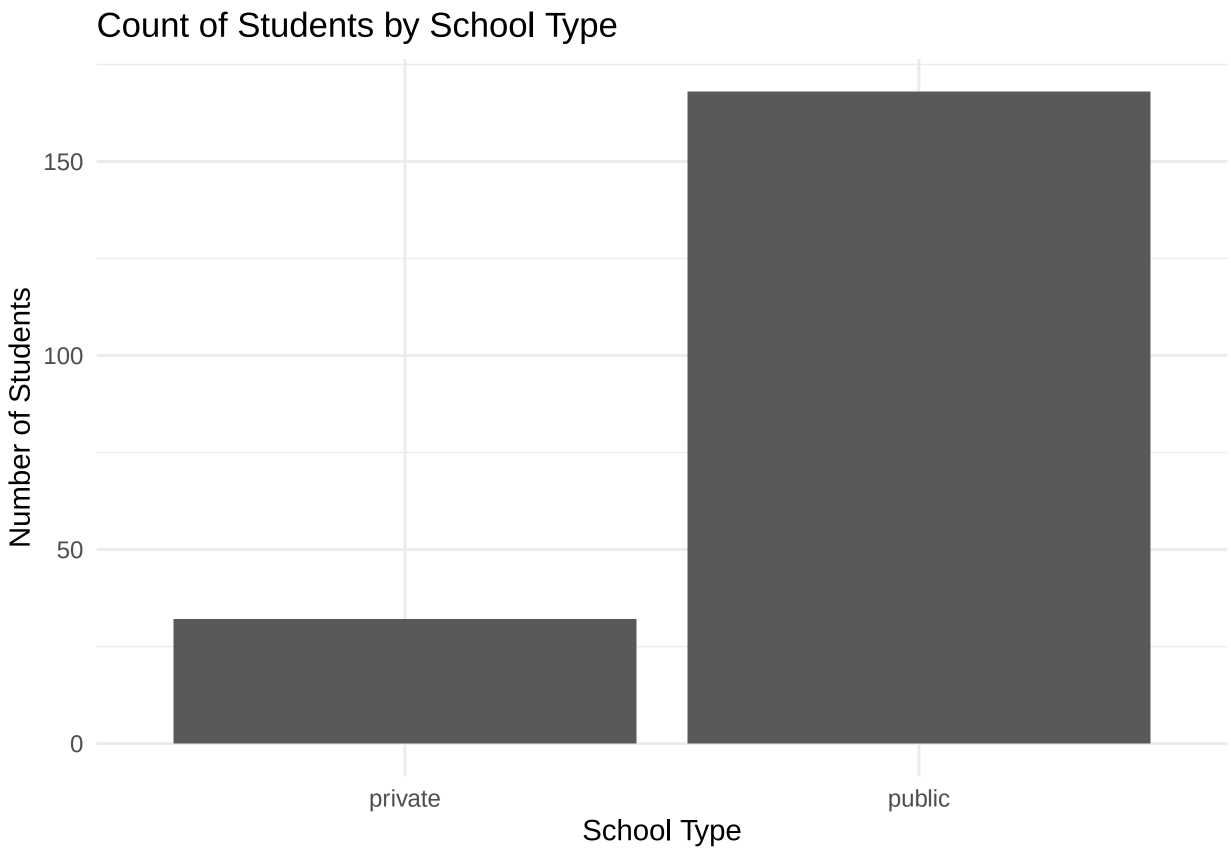
```
table(hsb2$schtyp)
```

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

```
prop.table(table(hsb2$schtyp))
```

```
##
## private  public
##    0.16    0.84
```

```
ggplot(hsb2, aes(x = schtyp)) + geom_bar() + labs(title = "Count of Students by School Type", x = "Schoo
```



Reading Scores:

```
summary(hsb2$read)
```

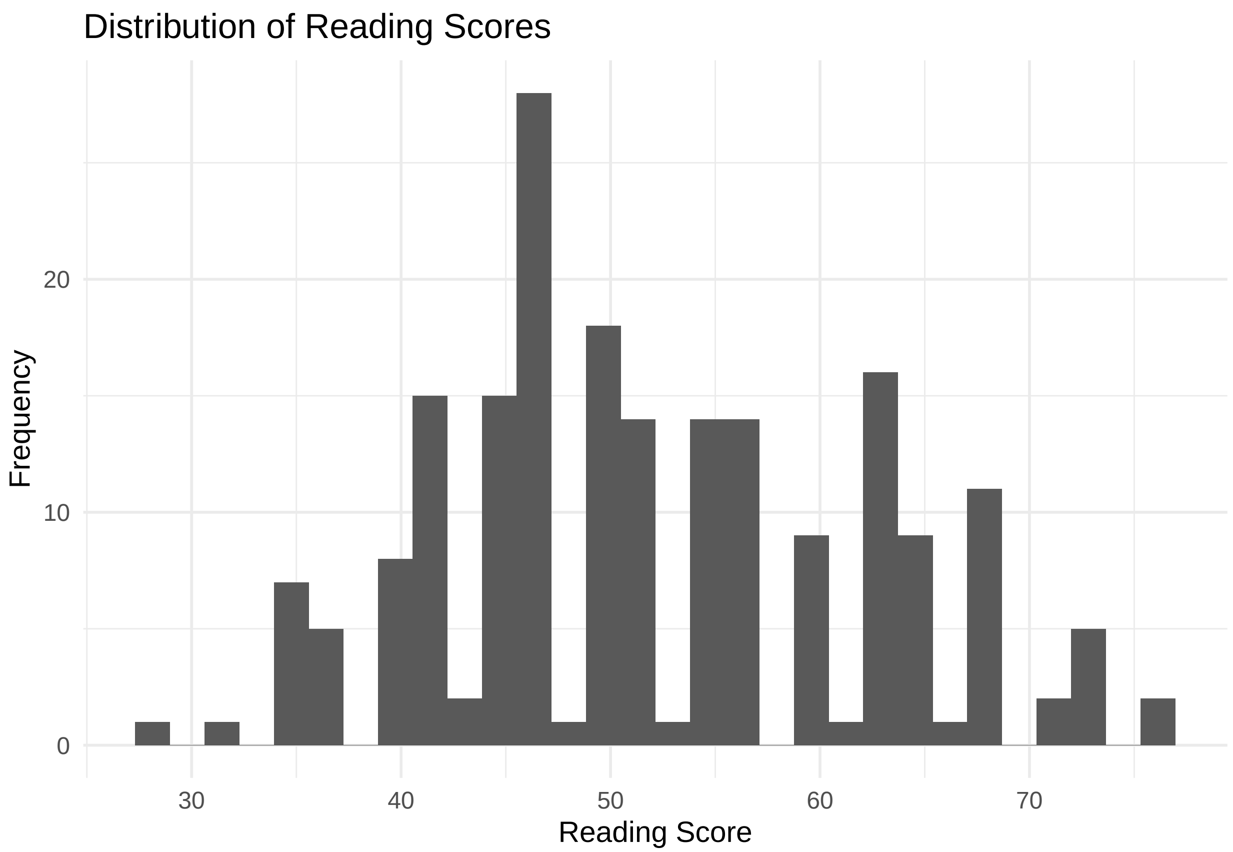
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   44.00   50.00   52.23   60.00   76.00
```

```
sd(hsb2$read)
```

```
## [1] 10.25294
```

```
ggplot(hsb2, aes(x = read)) + geom_histogram() + labs(title = "Distribution of Reading Scores", x = "Rea
```

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



Bivariate Exploration:

Reading Scores by Socioeconomic Status:

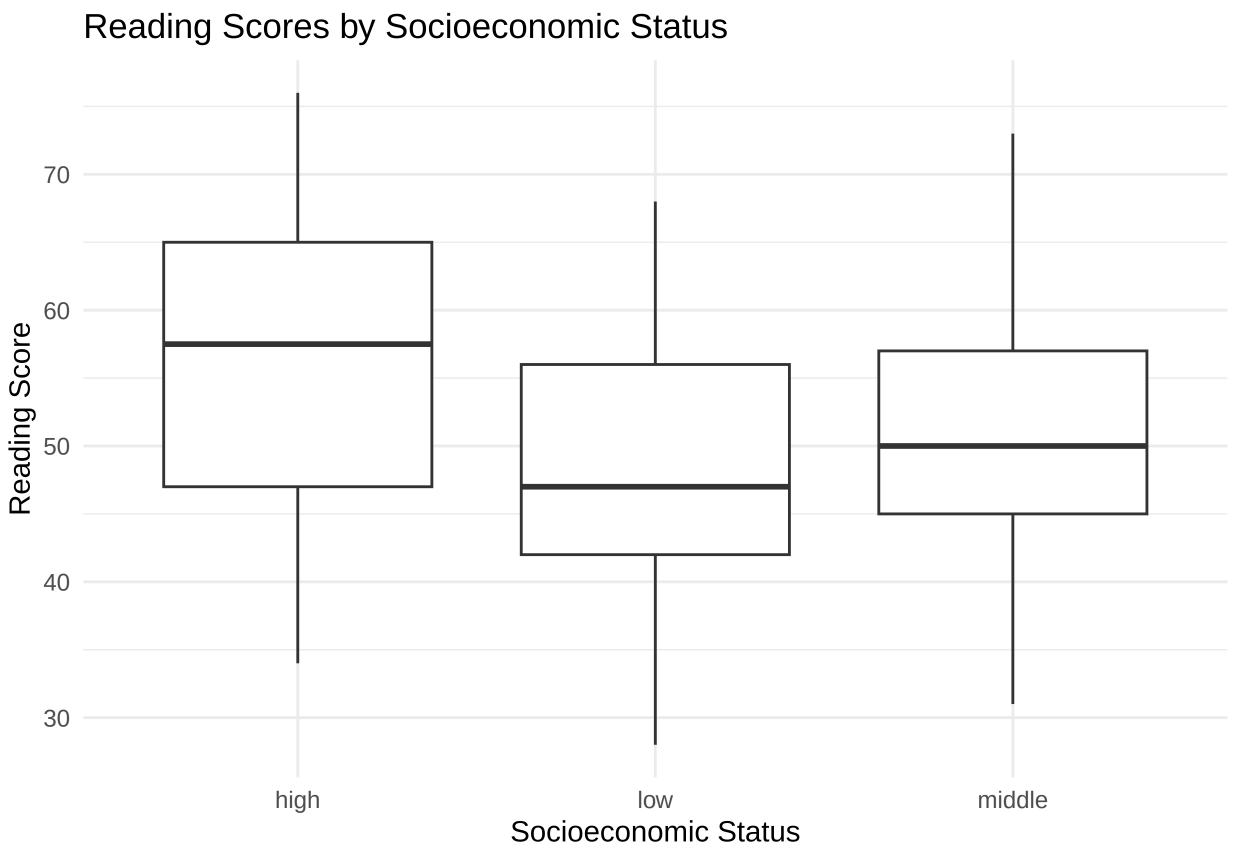
```
tapply(hsb2$read, hsb2$ses, mean, na.rm = TRUE)
```

```
##     high      low   middle
## 56.50000 48.27660 51.57895
```

```
tapply(hsb2$read, hsb2$ses, sd, na.rm = TRUE)
```

```
##      high       low    middle
## 10.858338  9.342987  9.425609
```

```
ggplot(hsb2, aes(x = ses, y = read)) + geom_boxplot() + labs(title = "Reading Scores by Socioeconomic S
```



Reading Scores by School Type:

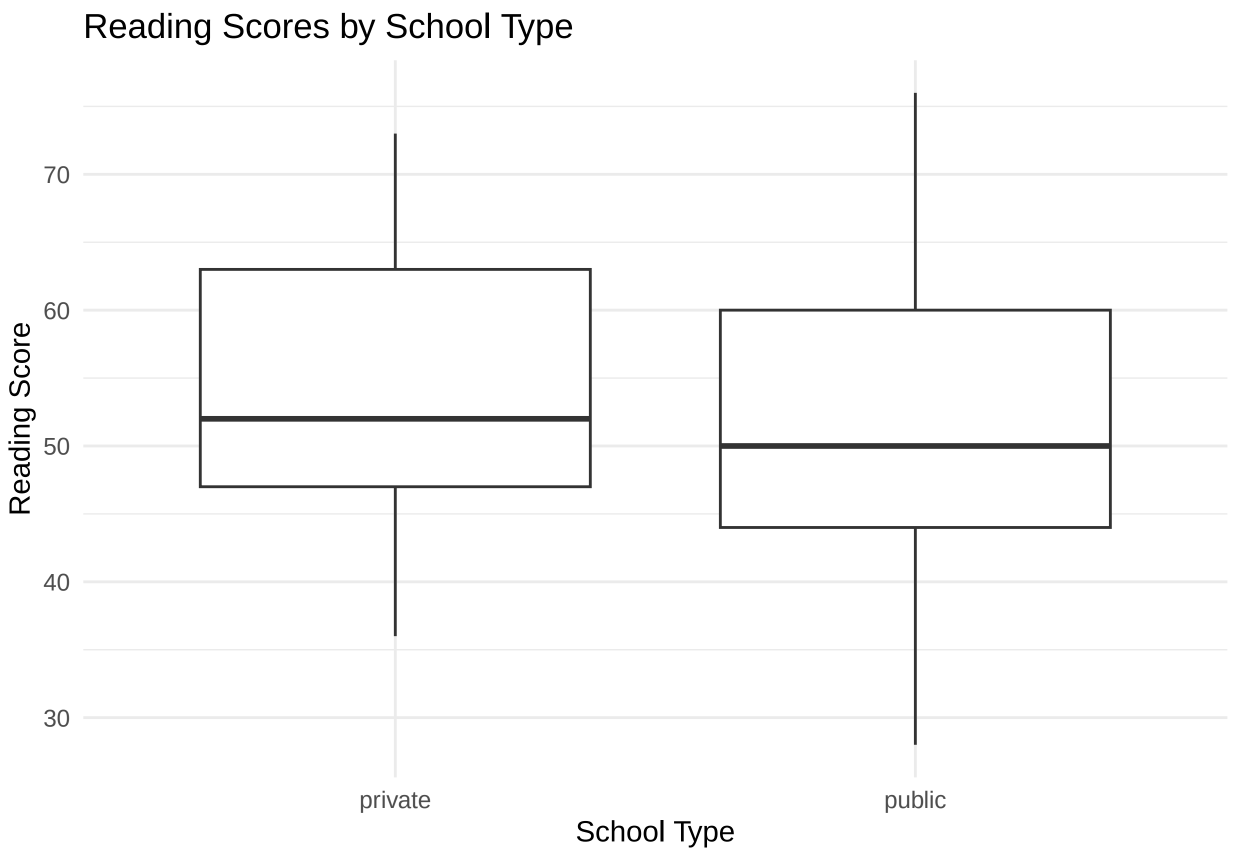
```
tapply(hsb2$read, hsb2$schtyp, mean, na.rm = TRUE)
```

```
##  private   public
## 54.25000 51.84524
```

```
tapply(hsb2$read, hsb2$schtyp, sd, na.rm = TRUE)
```

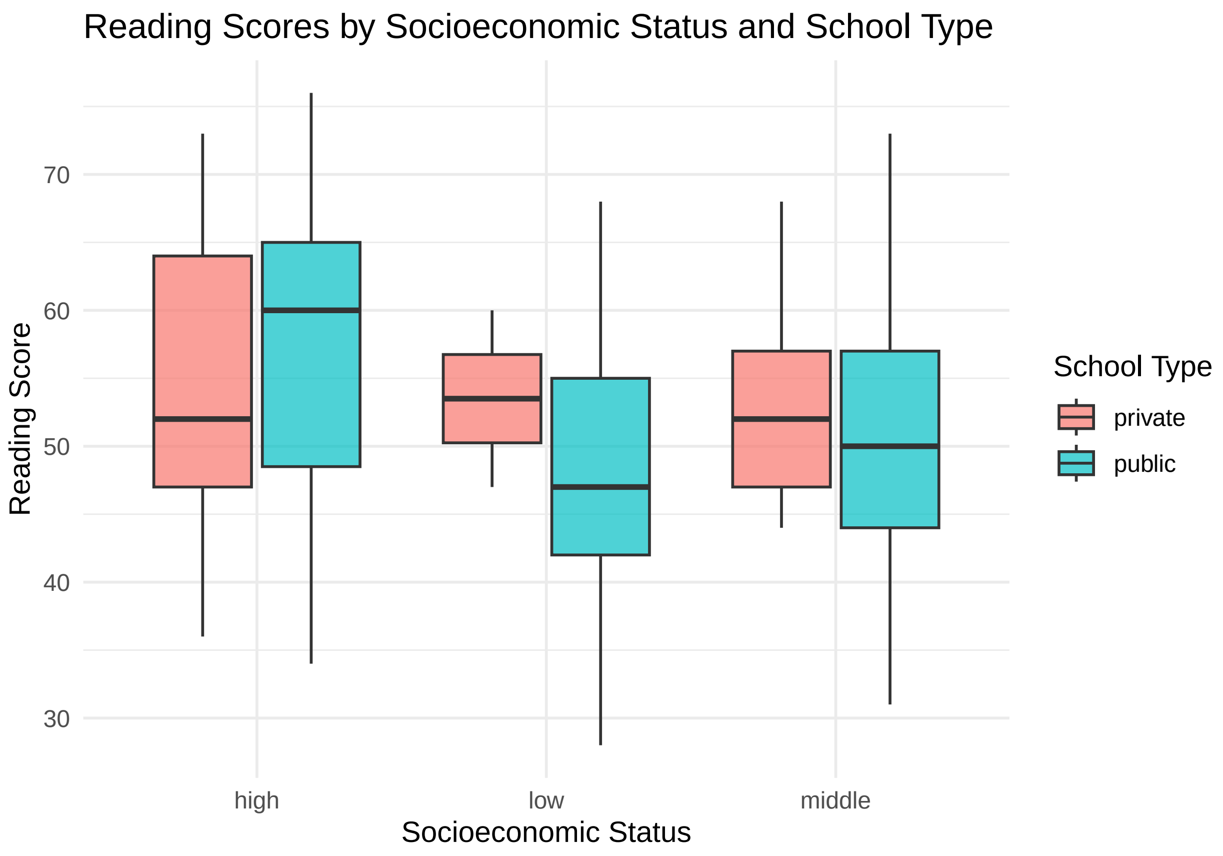
```
##   private    public
##  9.196774 10.422792
```

```
ggplot(hsb2, aes(x = schtyp, y = read)) + geom_boxplot() + labs(title = "Reading Scores by School Type"
```



Relationship between the two:

```r
ggplot(hsb2, aes(x = ses, y = read, fill = schtyp)) + geom_boxplot(alpha = 0.7) + labs(title = "Reading
```



Conclusion:

In conclusion, reading scores did differ by socioeconomic status, with students from higher status generally scoring higher on their reading tests. Additionally, students attending private schools scored higher than their public school counterparts. While this is not a formal statistical analysis, the graphs and averages show meaningful differences in reading scores in the different socioeconomic settings and school types.