EDA_mila.letirant

2024-09-25

```
hsb2 <- read.csv("/Users/milaletirant/Documents/Math 130/data/hsb2 .csv", header=TRUE)
head(hsb2)
```

```
##    id gender  race    ses schtyp       prog read write math science socst
## 1  70   male white    low public    general   57    52   41      47    57
## 2 121 female white middle public vocational   68    59   53      63    61
## 3  86   male white   high public    general   44    33   54      58    31
## 4 141   male white   high public vocational   63    44   47      53    56
## 5 172   male white middle public   academic   47    52   57      53    61
## 6 113   male white middle public   academic   44    52   51      63    61
```

Introduction:

In this exploratory data analysis, I will be looking into the relationship between student's reading scores (read) and socioeconomic status (ses) from the High School and Beyond data set.

My research question and what I'm going to explore is: "Do students with different socioeconomic status (low, middle, high) show differences in reading scores?"

Hypothesis: I believe students with a higher socioeconomic status will score higher on the reading test than students with a lower status because they tend to have a better access to educational resources and better home environments.

```
dim(hsb2)
```

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

The data we are going to look at has 200 observations which means it includes 200 students.

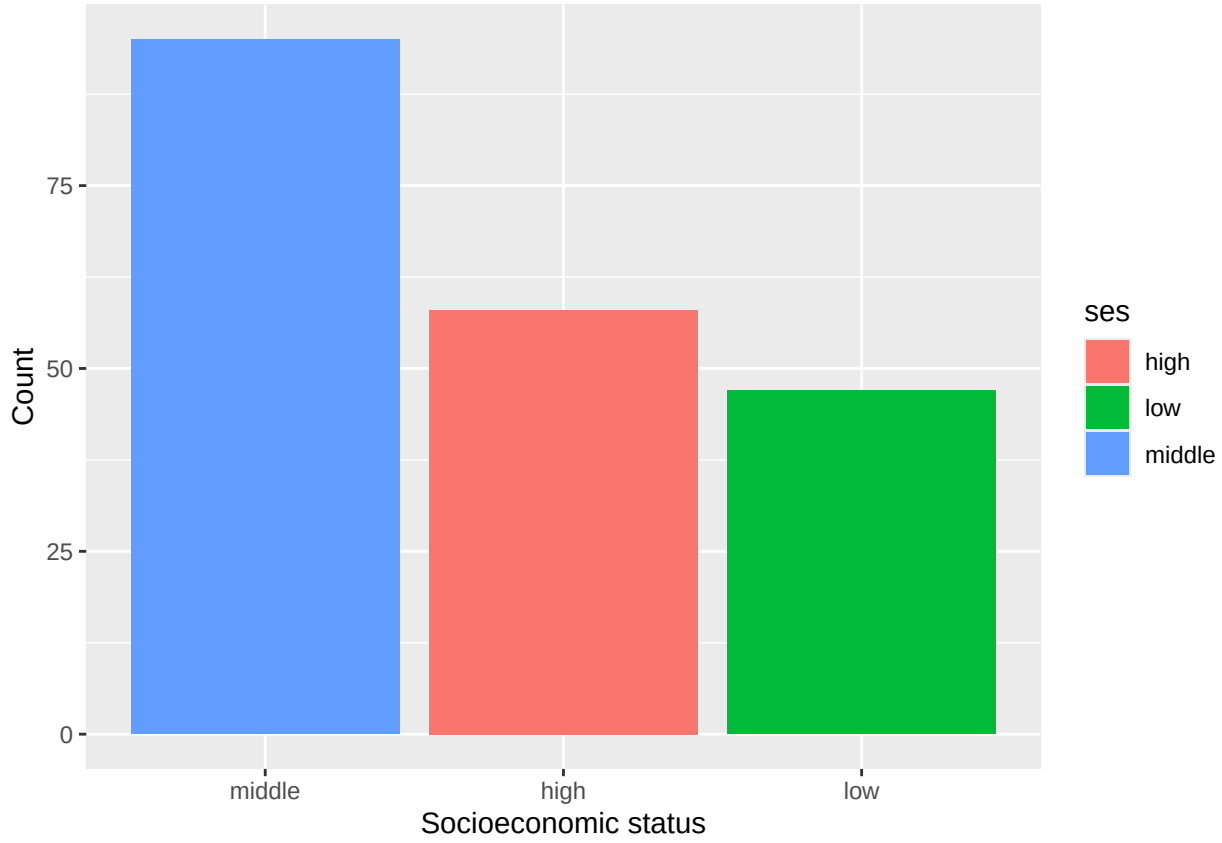
Socioeconomic status:

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

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

With this table we can see that 95 students have a middle socioeconomic status, 58 have a high socioeconomic status and 47 have a low socioeconomic status.

```
ggplot(hsb2, aes(x = forcats::fct_infreq(ses), fill = ses)) +
  geom_bar() +
  xlab("Socioeconomic status") +
  ylab("Count")
```



This bar graph also shows us how many students are in each socioeconomic status groups.

Reading Score:

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

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

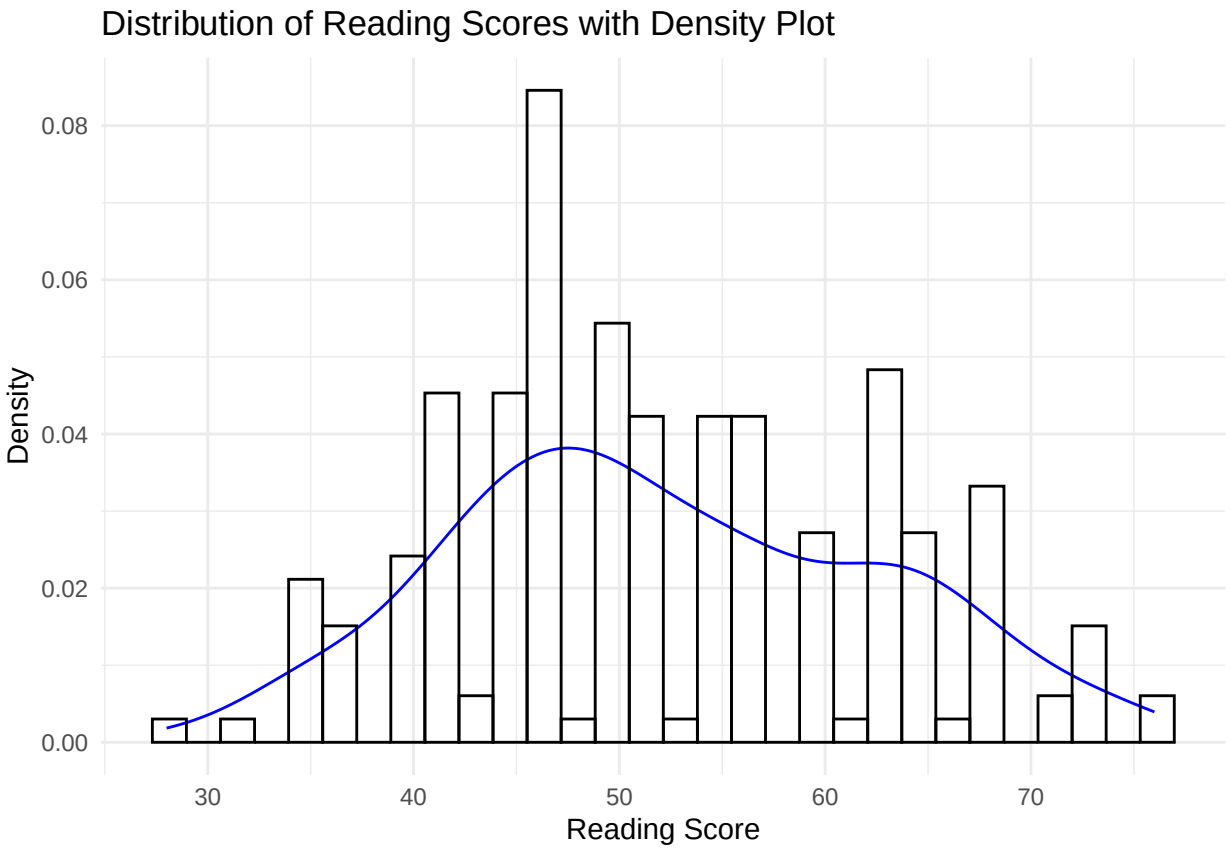
```
sd(hsb2$read)
```

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

These statistics tell us that the highest reading score is 76 while the lowest is 28, with an average score of 52.23. The median being 50 means that half of the scores are below 50, and half are above 50. A standard deviation of 10.25294 means that most reading scores are within about 10.25 points of the mean reading score

```
ggplot(hsb2, aes(x=read)) + geom_density(col="blue") +
  geom_histogram(aes(y=after_stat(density)), colour="black", fill=NA) +   labs(title = "Distribution of
       x = "Reading Score",
       y = "Density") +
  theme_minimal()
```

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



This graph provides a visual representation of the distribution of the reading scores.

Bivariate Exploration:

Mean of reading scores of 3 groups : low, middle and high socioeconomic status:

```
aggregate(read ~ ses, data = hsb2, FUN = mean)
```

```
##      ses     read
## 1   high 56.50000
## 2    low 48.27660
## 3 middle 51.57895
```

Standard deviation of reading scores of 3 groups : low, middle and high socioeconomic status:

```
aggregate(read ~ ses, data = hsb2, FUN = sd)
```

```
##      ses      read
## 1   high 10.858338
## 2    low  9.342987
## 3 middle  9.425609
```

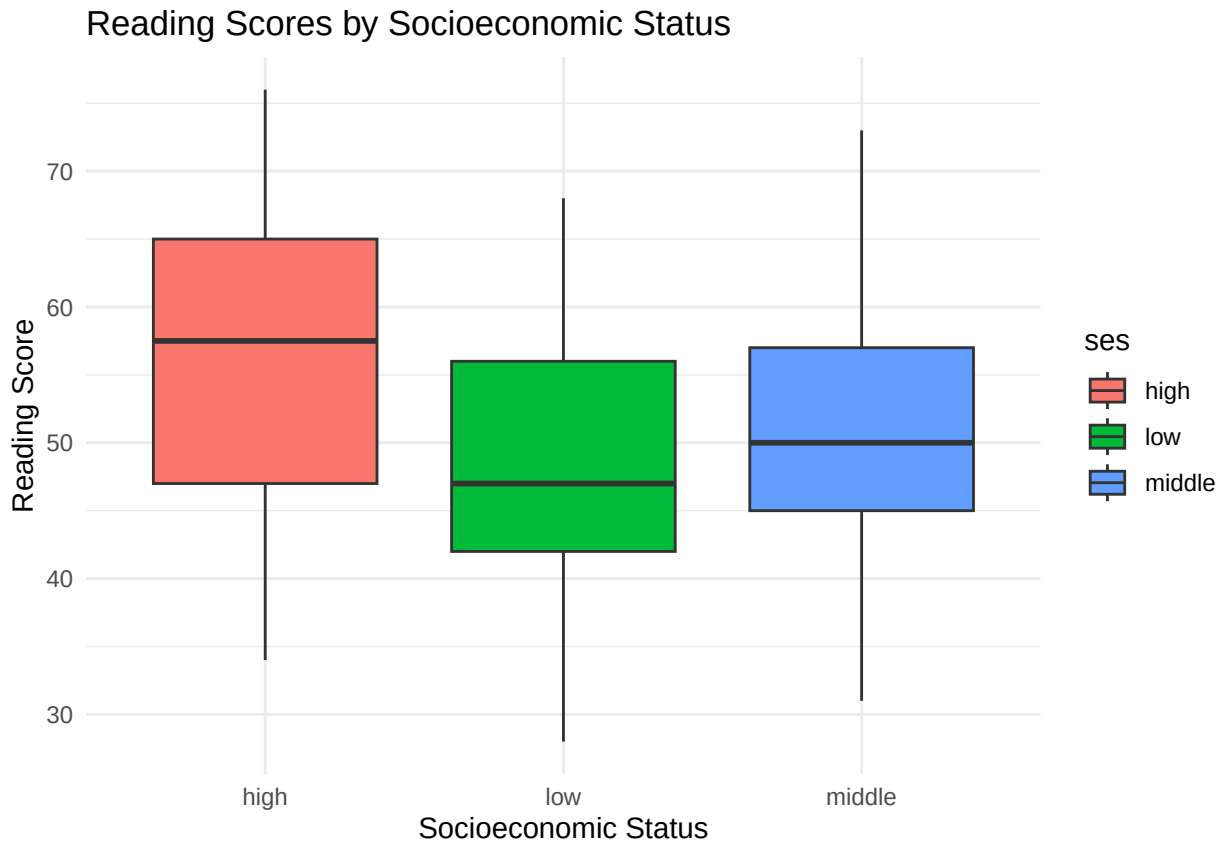
Median of reading scores of 3 groups : low, middle and high socioeconomic status:

```r
aggregate(read ~ ses, data = hsb2, FUN = median)
```

```
##      ses read
## 1   high 57.5
## 2    low 47.0
## 3 middle 50.0
```

These 3 set of statistics shows that the low socioeconomic status group has the lowest median, standard deviation and mean score on the test, and the high socioeconomic status group has the highest. Which tell us that the higher group tend to score better on the test than lower groups.

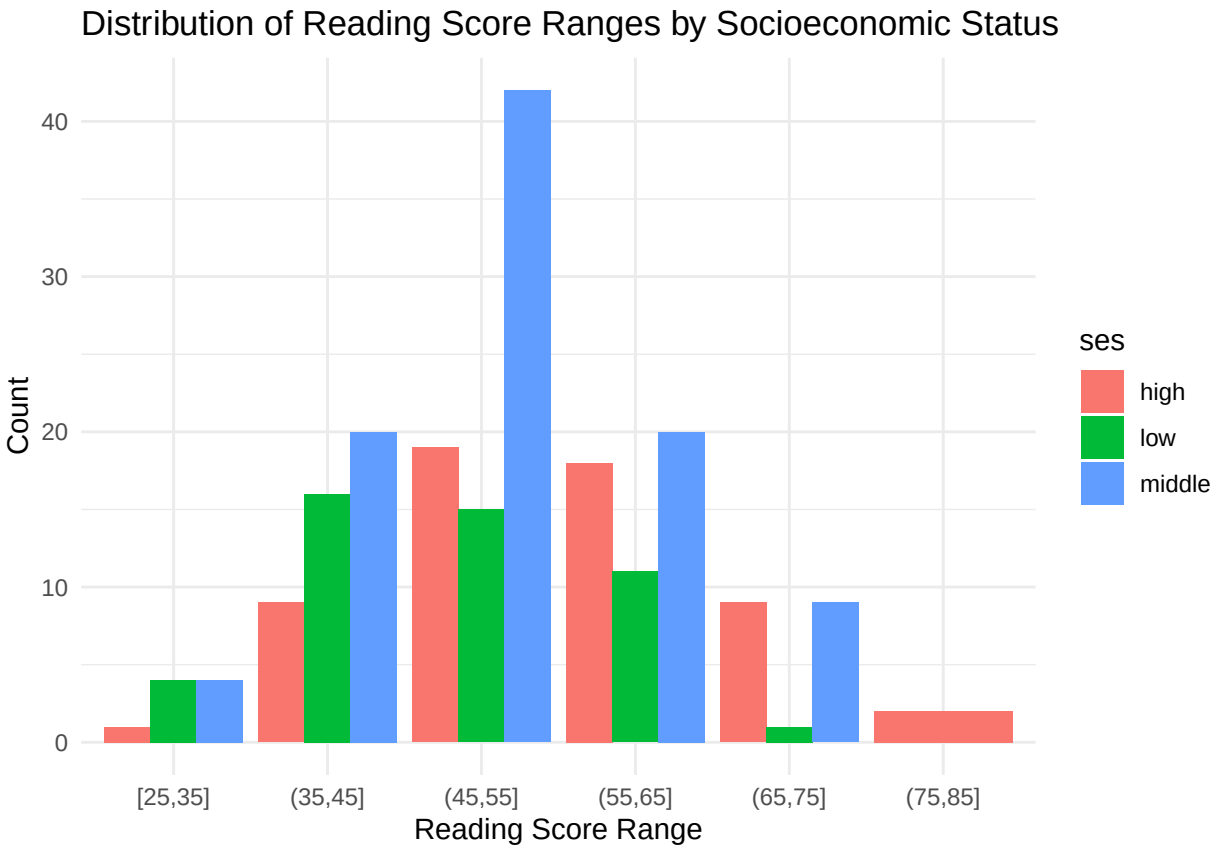
```r
ggplot(hsb2, aes(y=read, x=ses, fill=ses)) +
        geom_boxplot() + labs(title = "Reading Scores by Socioeconomic Status",
       x = "Socioeconomic Status",
       y = "Reading Score") +
  theme_minimal()
```



This boxplot graph clearly shows us that the group with the highest scores on the reading test is the high socioeconomic status group, followed by the middle and then the low socioeconomic status group.

```r
hsb2$read_range <- cut_width(hsb2$read, width=10)

ggplot(hsb2, aes(x=read_range, fill=ses)) + geom_bar(position = "dodge") +  labs(title = "Distribution
       x = "Reading Score Range",
       y = "Count") +
  theme_minimal()
```



High Socioeconomic Status: This group is the only one with scores in the 75-85 range, indicating high performance on the reading test. Additionally, very few students from this group have scores in the 25-35 range, suggesting that only a small number of students perform poorly.

Low Socioeconomic Status: This group has a small number of students scoring in the 65-75 range, with the majority scoring between 35-65 points. This indicates that students from the lowest socioeconomic status tend to have lower reading scores.

Middle Socioeconomic Status: Most students in this group scored between 45-55 points, reflecting an average performance.

Conclusion:

With the help of the graphs and summary statistics we observe that students with higher socioeconomic status tend to score better on the reading test than students with lower socioeconomic status. Which support the hypothesis.