EDA_JessicaLunsford

Jessica Lunsford

2024-03-01

```
knitr::opts_chunk$set(warning=FALSE, message=FALSE, fig.height=4, fig.width=5, fig.align='center')
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
```

```
library(forcats)
library(ggplot2)
hsb2<- read.table("../data/hsb2.txt", header=TRUE, sep = "\t")
```

Introduction:

The data set this project looks at is "High School and Beyond." The data set has 11 variables and 200 observations. The survey was conducted on high school seniors. The variables available in the study are number, gender, race, social-economic status, school type, program type, reading, writing, math, science and social studies. The variables we will look at and compare in this project are are gender, social-economic status and math scores. How do students of different socioeconomic status perform in math? How do students of different genders perform in math? There are certainly stereotypes that female students do not perform as well in math. There are also stereotypes that suggest students from lower socioeconomic status do not perform as well in math. This project will compare these variables graphically to allow for easy comparison.

Univariate description:

The variables compared today are socioeconomic status, gender and math scores. Let's begin by looking

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

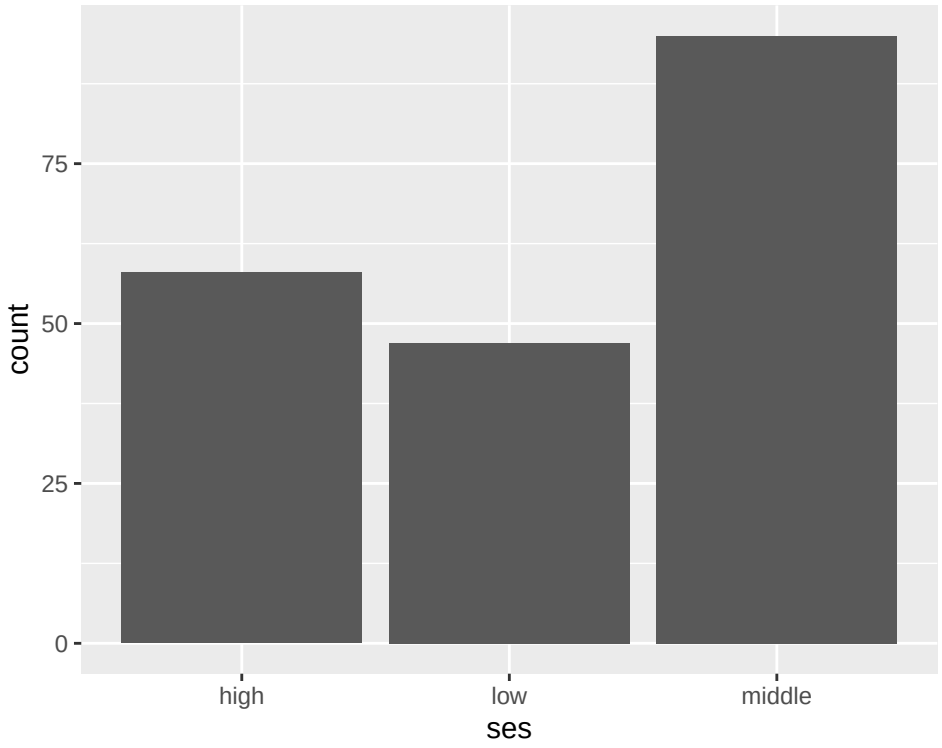
```
##
##   high    low middle
##     58     47     95
```

```
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
```

A graph allows for easy visual comparisons.

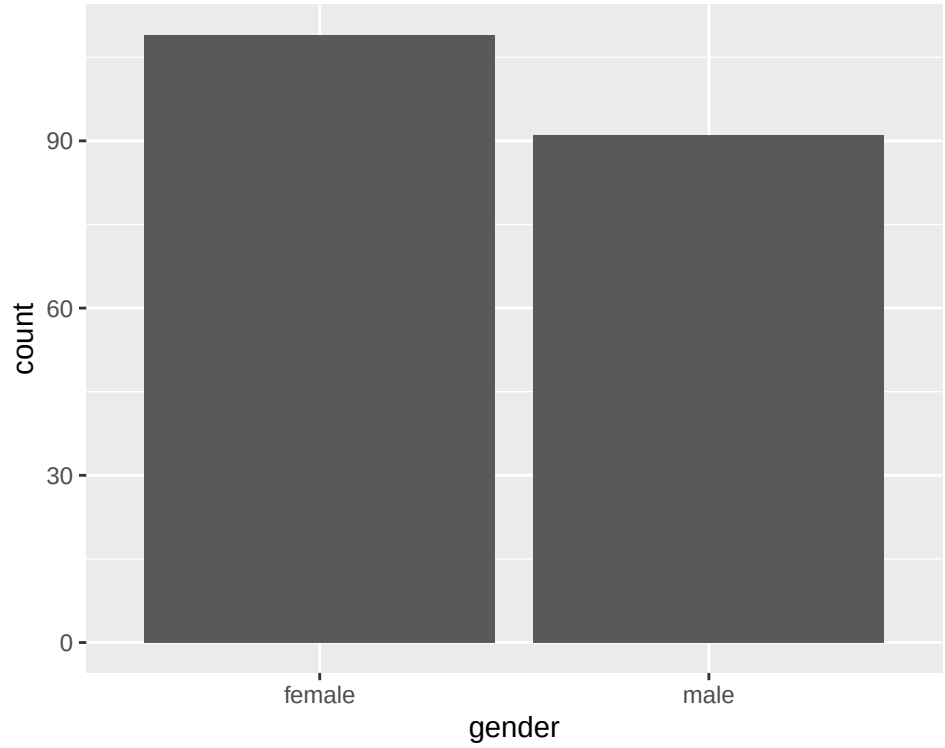
```
ggplot(hsb2,aes(ses)) +geom_bar()
```




```
table(hsb2$gender)
```

```
##
## female   male
##    109     91
```

```
ggplot(hsb2, aes(gender))+geom_bar()
```



Here is a summary of the standardized math scores.

```
summary(hsb2$math, na.rm=TRUE)
```

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

The math data set has a mean value of...

```
hsb2$math%>% mean()
```

```
## [1] 52.645
```

Bivariate exploration:

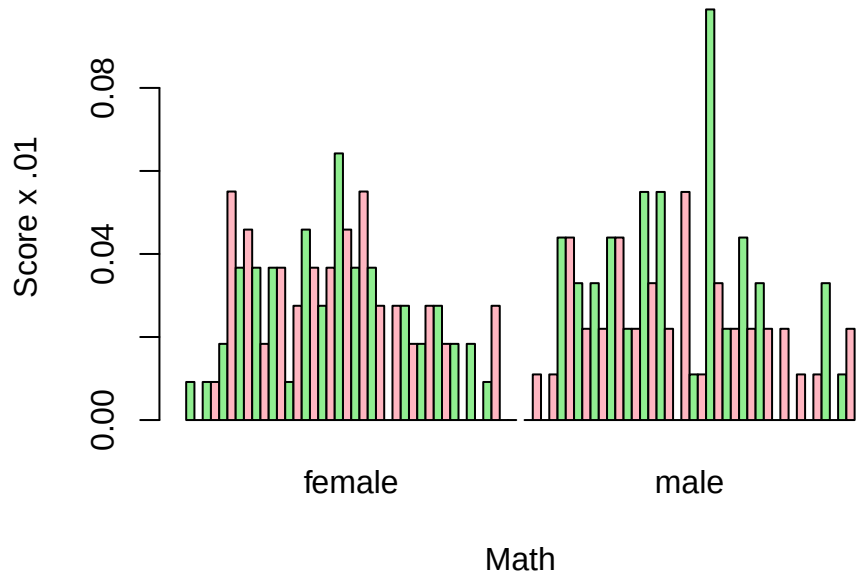
Let's compare these variables against each other. By graphing the data against one another we may find preconceived notions to be refuted. Let's begin by looking at math scores vs. gender.

```
two_way_table<-table(hsb2$math, hsb2$gender)
two_way_table
```

```
##
##      female male
##   33      1    0
##   35      0    1
##   37      1    0
##   38      1    1
```

```
##   39      2    4
##   40      6    4
##   41      4    3
##   42      5    2
##   43      4    3
##   44      2    2
##   45      4    4
##   46      4    4
##   47      1    2
##   48      3    2
##   49      5    5
##   50      4    3
##   51      3    5
##   52      4    2
##   53      7    0
##   54      5    5
##   55      4    1
##   56      6    1
##   57      4    9
##   58      3    3
##   59      0    2
##   60      3    2
##   61      3    4
##   62      2    2
##   63      2    3
##   64      3    2
##   65      3    0
##   66      2    2
##   67      2    0
##   68      0    1
##   69      2    0
##   70      0    1
##   71      1    3
##   72      3    0
##   73      0    1
##   75      0    2
```

```r
proportion_table <- prop.table(two_way_table, margin=2)
barplot(proportion_table, beside = TRUE, legend= FALSE,
main="Standardized Math Scores by Gender",
xlab="Math", ylab="Score x .01",
col=c("lightgreen","lightpink"))
```



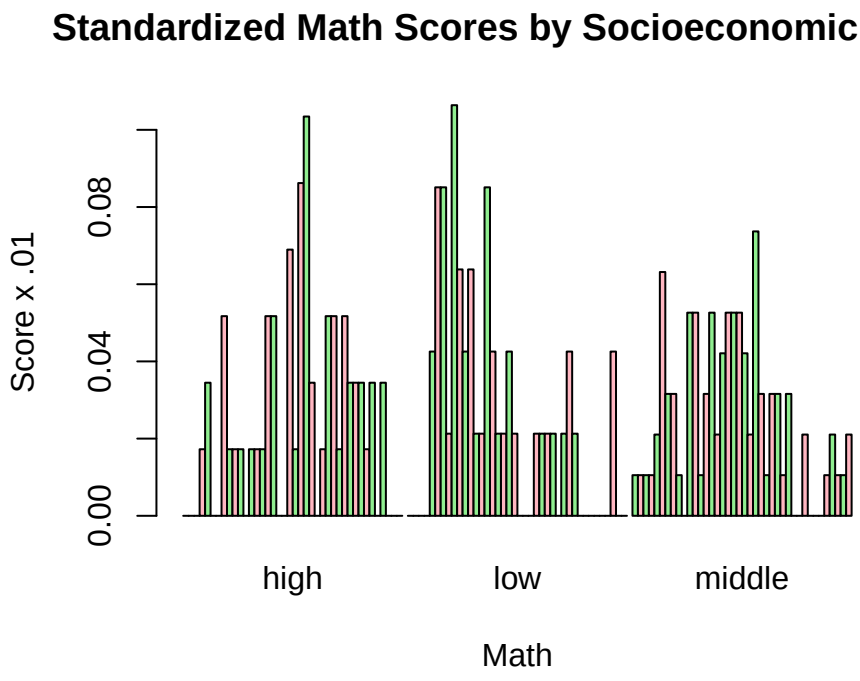
This graph shows that the scores are well distributed. Some values are not present and some are very high. One could not rightfully infer that high or low social economic status affects math scores.

```
two_way_table<-table(hsb2$math, hsb2$ses)
two_way_table
```

```
##
##      high low middle
##   33    0   0      1
##   35    0   0      1
##   37    0   0      1
##   38    1   0      1
##   39    2   2      2
##   40    0   4      6
##   41    0   4      3
##   42    3   1      3
##   43    1   5      1
##   44    1   3      0
##   45    1   2      5
##   46    0   3      5
##   47    1   1      1
##   48    1   1      3
##   49    1   4      5
##   50    3   2      2
##   51    3   1      4
##   52    0   1      5
##   53    0   2      5
##   54    4   1      5
```

```
##    55    1    0      4
##    56    5    0      2
##    57    6    0      7
##    58    2    1      3
##    59    0    1      1
##    60    1    1      3
##    61    3    1      3
##    62    3    0      1
##    63    1    1      3
##    64    3    2      0
##    65    2    1      0
##    66    2    0      2
##    67    2    0      0
##    68    1    0      0
##    69    2    0      0
##    70    0    0      1
##    71    2    0      2
##    72    0    2      1
##    73    0    0      1
##    75    0    0      2
```

```
proportion_table <- prop.table(two_way_table, margin=2)
barplot(proportion_table, beside = TRUE, legend= FALSE,
main="Standardized Math Scores by Socioeconomic Status",
xlab="Math", ylab="Score x .01",
col=c("lightgreen","lightpink"))
```



The graph of socioeconomic status vs.math scores shows that students of high and low status got higher

Conclusion:

This project looked at three variables and compared them against each other to gain a greater understanding of the data relates to each other. The variable "socioeconomic status" divided students into high, middle and low categories.The sample contained mostly middle class students. The variable gender was fairly even with a slightly higher female count. The math value was a range of individual scores with a mean value of 52.645.The ability to graph these data sets against each other makes it easier to visualize the comparison.