Exploratory Data Analysis Project

Jedidiah Hendry

2/24/2022

```
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(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
hsb2 <- read.table("C:/Users/JedHe/OneDrive/Documents/MATH130/Data/hsb2.txt", header=TRUE, sep=
"\t")
```

INTRO:

For my project, I plan to do research on the data set involving the development of elementary and high school students through their education (High School and Beyond). The variables I will build my report from will be the type of school (schtyp) and gender which could affect the grades of writing and math of each student. I am interested in learning if the grades of different students is directly proportional or independent to their school type and gender. I expect that the grades of each student would be dependent on the type of school and independent of gender.

Univariate Exploration:

```
mean(hsb2$write)
```

```
## [1] 52.775
```

```
mean(hsb2$math)
```

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

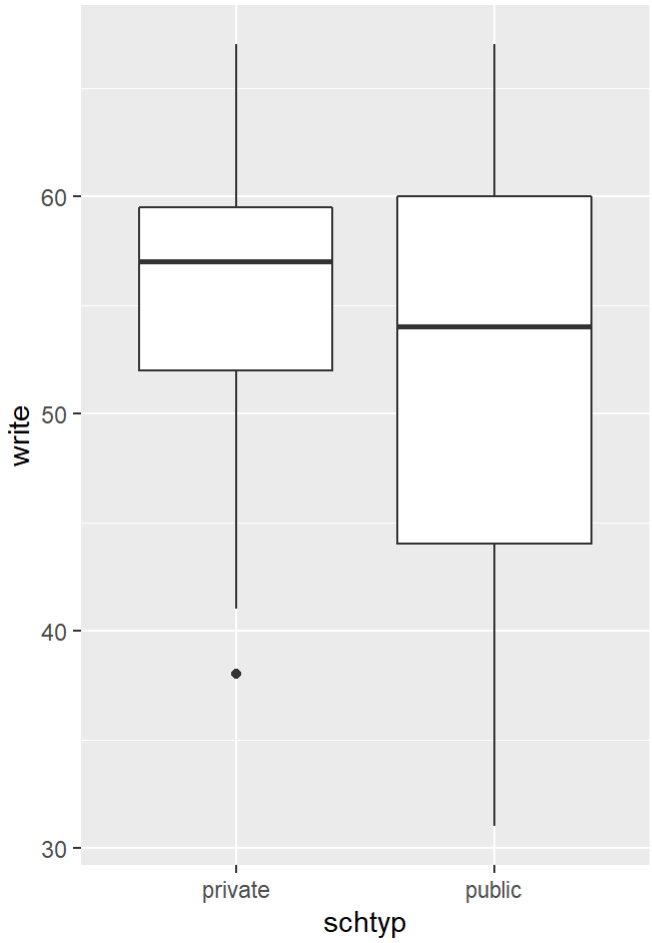
```
by_schtyp <- group_by(hsb2, schtyp)
summarise(by_schtyp, avg_write = mean(write, na.rm=TRUE))
```

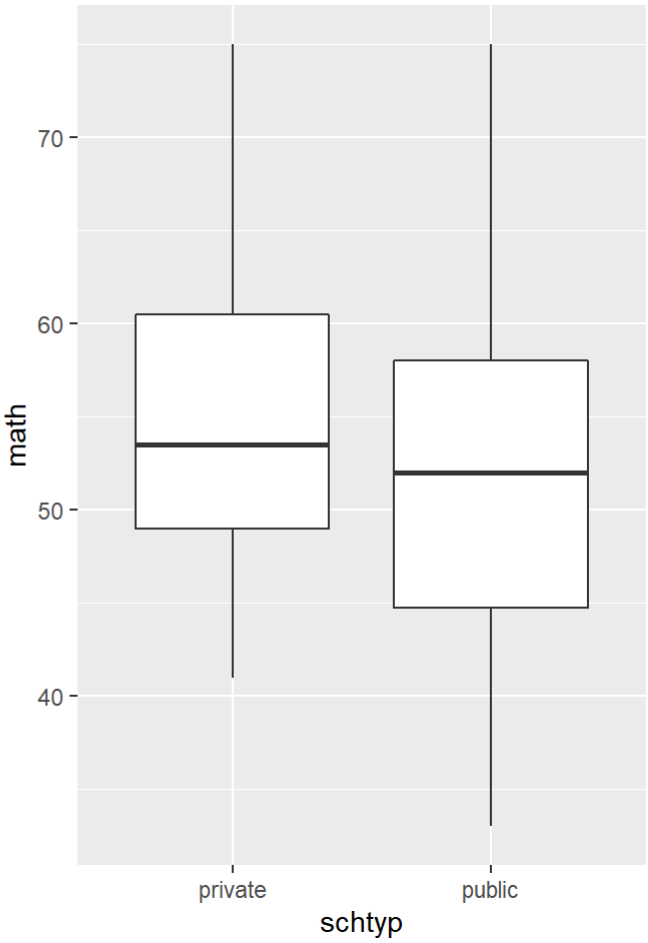
```
## # A tibble: 2 x 2
##   schtyp  avg_write
##   <chr>       <dbl>
## 1 private      55.5
## 2 public       52.2
```

```
by_schtyp <- group_by(hsb2, schtyp)
summarise(by_schtyp, avg_math = mean(math, na.rm=TRUE))
```

```
## # A tibble: 2 x 2
##   schtyp  avg_math
##   <chr>      <dbl>
## 1 private     54.8
## 2 public      52.2
```

```
schtyp1 <- ggplot(hsb2, aes(x=schtyp, y=write)) + geom_boxplot()
schtyp2 <- ggplot(hsb2, aes(x=schtyp, y=math)) + geom_boxplot()
grid.arrange(schtyp1, schtyp2, ncol=2)
```





For the individual data concerning the school type with the two variables, writing and math scores, the means and data were different to what I predicted. When compparing the original means to the means of the writing and math scores based on school type, there isn't a significant difference between the private and public mean scores. For both, the lower mean score is from public schools, but private schools were only 2 points above the original mean for math and 3 points for writing. Looking at the boxplots and table portrays a possibe explaination, the large population of sample school students prevents the few outliers from affecting the mean of public school students too drastically. This highlights the importance of considering two or more variables per observation as it can reduce the chances of the data being skewed by special circumstances.

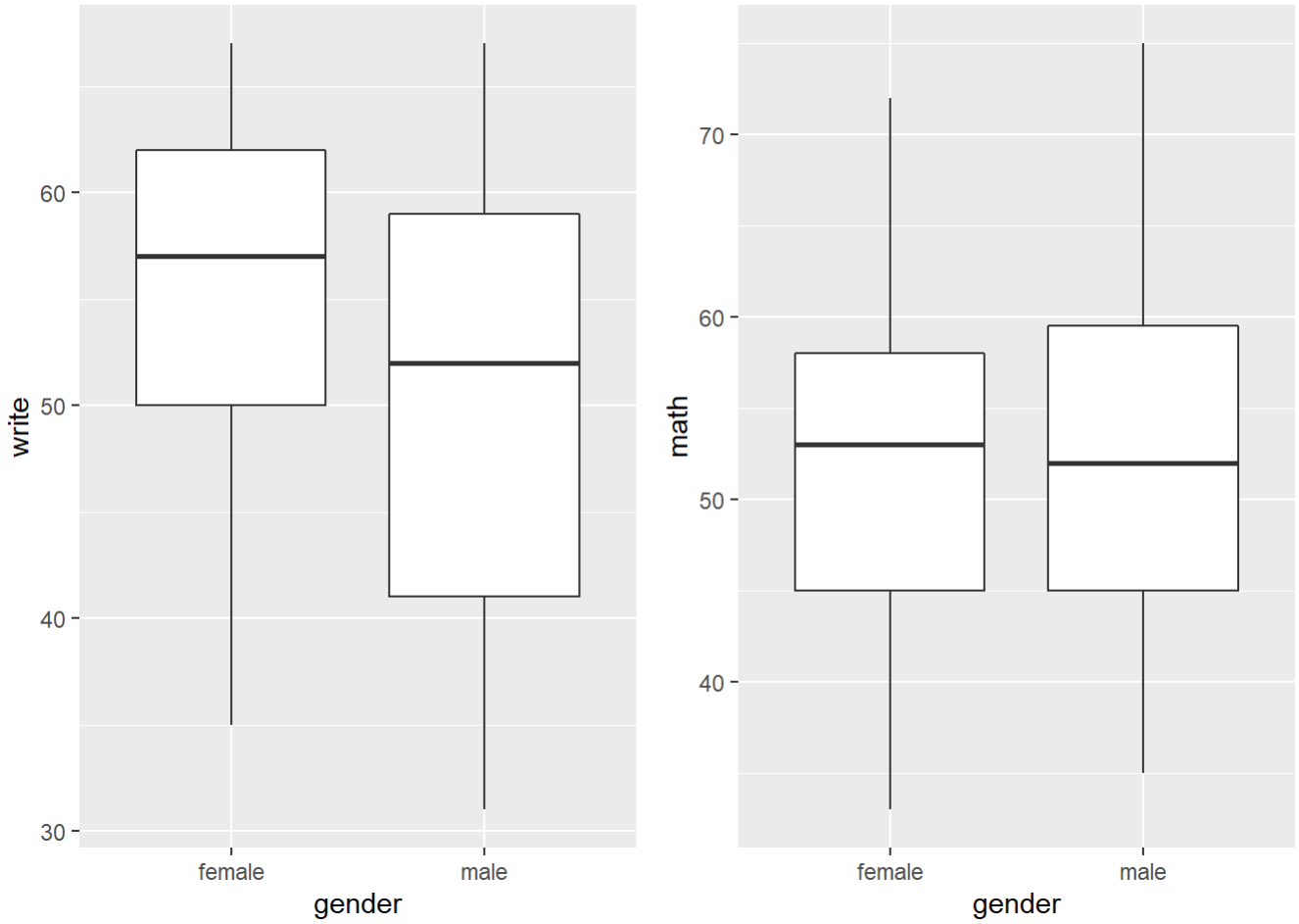
```
by_gender <- group_by(hsb2, gender)
summarise(by_gender, avg_writing = mean(write, na.rm=TRUE))
```

```
## # A tibble: 2 x 2
##   gender avg_writing
##   <chr>        <dbl>
## 1 female        55.0
## 2 male          50.1
```

```
by_gender <- group_by(hsb2, gender)
summarise(by_gender, avg_mathematics = mean(math, na.rm=TRUE))
```

```
## # A tibble: 2 x 2
##   gender avg_mathematics
##   <chr>            <dbl>
## 1 female            52.4
## 2 male              52.9
```

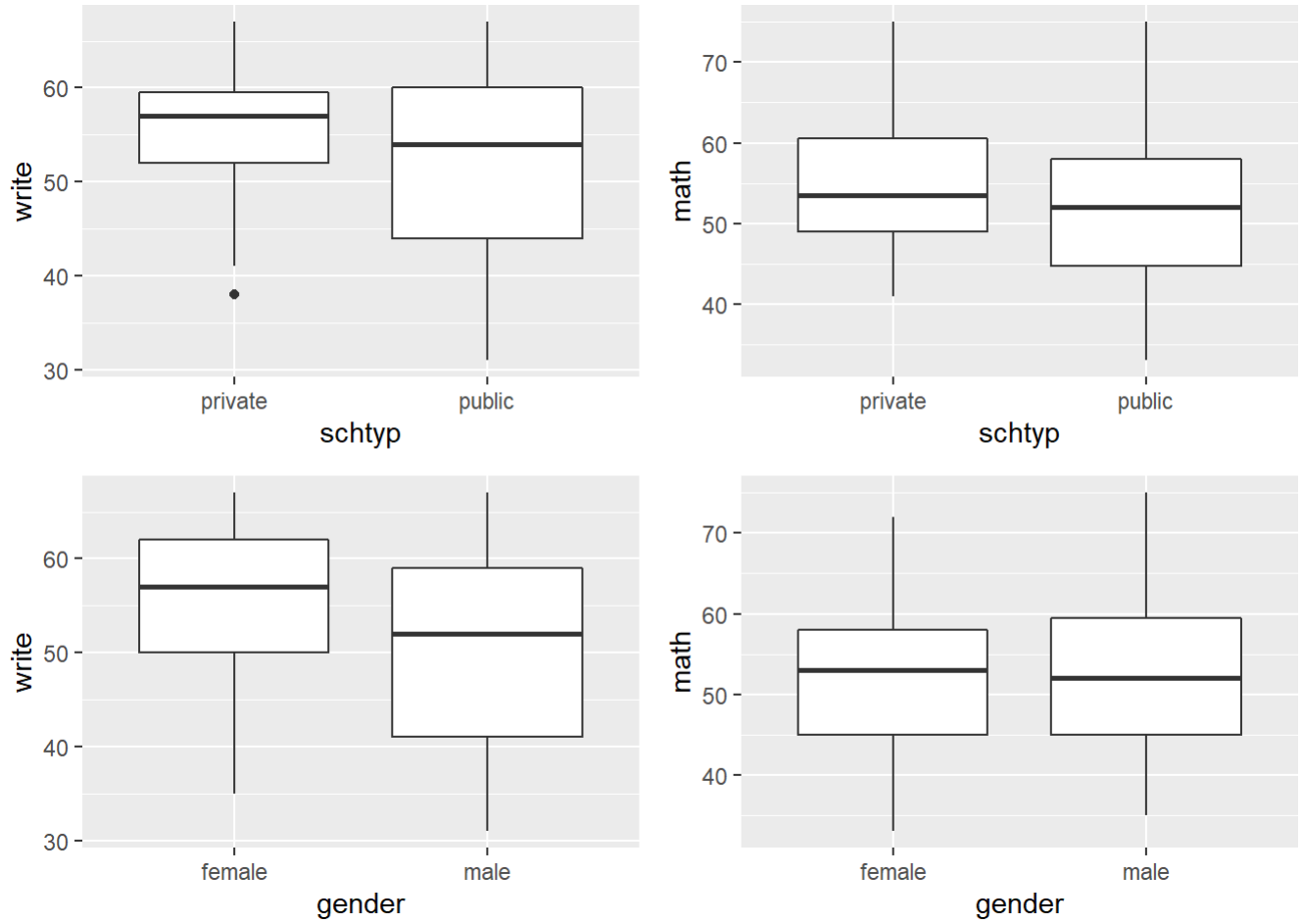
```
gender1 <- ggplot(hsb2, aes(x=gender, y=write)) + geom_boxplot()
gender2 <- ggplot(hsb2, aes(x=gender, y=math)) + geom_boxplot()
grid.arrange(gender1, gender2, ncol=2)
```



For the gender variable in High School and Beyond, there are some patterns that it shares with school type. Firstly, the writing variable seems to vary greatly in range like in school type while the math variablle is more consistent in its range according to the boxplots. For gender, the mean of writing for females is 4 points higher than males, with females being 2 above the overall mean and men being 2 below the mean. Also similarly to school type, the means of math are closer than writing, with females and males within 1 point difference for the mean. With the data combined, it is more likely that the school subjects being scored has a greater influence over the means rather than the variables like school type and gender.

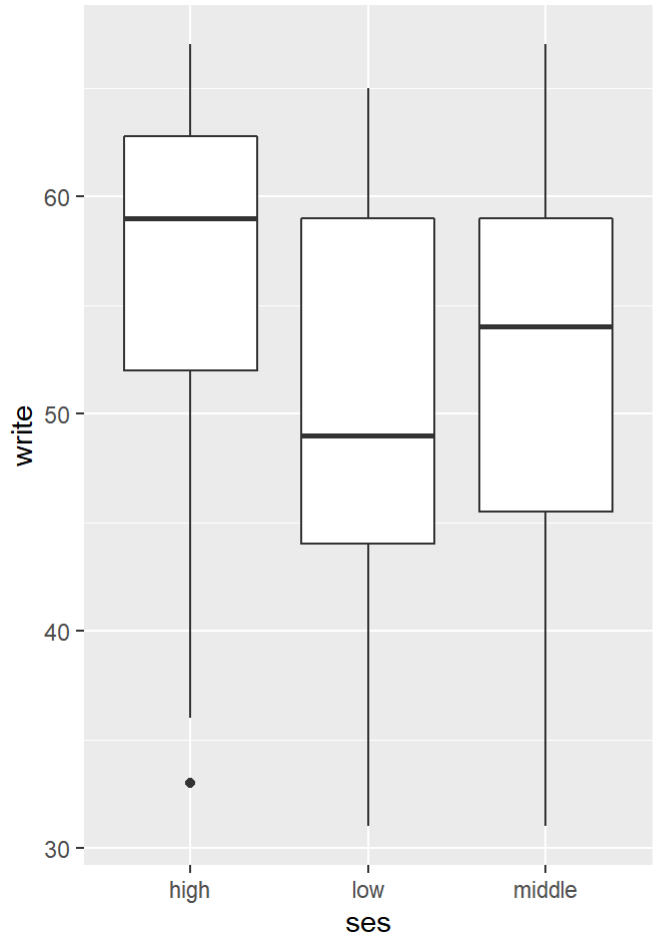
Bivariate Exploration:

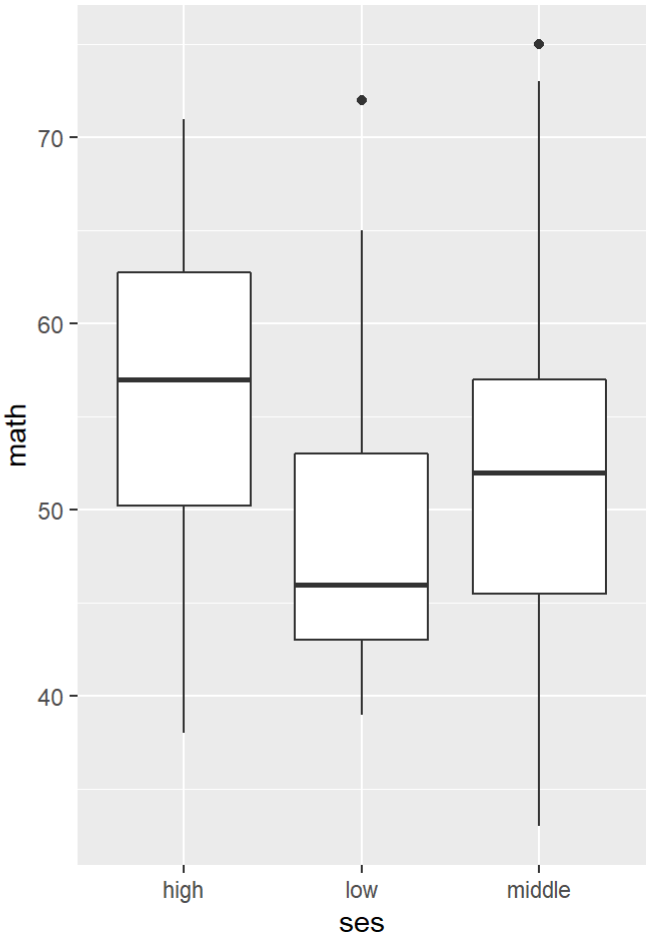
```
combined <- group_by(hsb2, gender, schtyp)
grid.arrange(schtyp1, schtyp2, gender1, gender2, ncol=2)
```



According to the complete set of data between the independent variables of school type and gender and the depdendent variables of writing and math, there seems to be a pattern between the subjects being scored rather than the independent variables. To test this, I want to observe one more independent variable: session type. This will help me get a better understanding of the patterns between the original variables I selected without diverting too far from my original question.

```
prog1 <- ggplot(hsb2, aes(x=ses, y=write)) + geom_boxplot()
prog2 <- ggplot(hsb2, aes(x=ses, y=math)) + geom_boxplot()
grid.arrange(prog1, prog2, ncol=2)
```






```
by_ses <- group_by(hsb2, ses)
summarise(by_ses, avg_write = mean(write, na.rm=TRUE))
```

```
## # A tibble: 3 x 2
##   ses    avg_write
##   <chr>      <dbl>
## 1 high        55.9
## 2 low         50.6
## 3 middle      51.9
```

```
summarise(by_ses, avg_math = mean(math, na.rm=TRUE))
```

```
## # A tibble: 3 x 2
##   ses    avg_math
##   <chr>     <dbl>
## 1 high       56.2
## 2 low        49.2
## 3 middle     52.2
```

According to my new set of data involving session type, the means of each set are more consistent than most of the variables. The boxplots and the tables demonstrate how the variation in mean scores for each independent variable is not influenced by the subject being scored like the school type and gender. This supports the idea that

the dependent variables affect the variation in scores for my original variables more than the independent variables.

Conclusion:

Through these tests, I found out that my two independent variables, school type and gender, have varying means in the two subjects of writing and math, but the source of the variation is most likely not from the independent variables. Both writing and math have different patterns between school type and gender, like a greater variation in writing means with math means staying closer together between both variables, which supports the idea that the results are independent from the school type and gender. I also ran a different test based on session which demonstrated what the data would look like if the variation was caused by the independent variable. There is no pattern difference between writing and math, but the three variables of session have more consistent values of the mean which supports the idea that the session type has a significant role in determining average success in overall subjects. This was different from my original hypothesis where I theorized that there would be a variation caused by the school type but it supported the idea that there would be no difference due to gender.