Homework 1

```
library(psych)
library(lme4)
library(knitr)
library(kableExtra)
library(qqplotr)
library(plyr)
library(tidyverse)
```

Question 1

Read in the HSB.csv file and save it in a dataframe called HSB_Data, excluding the variables, pracad and disclim. Verify you have done this correctly by printing the first several lines of the dataframe.

```
data_path <- "https://raw.githubusercontent.com/emoriebeck/homeworks/master/homework1/HSB.csv"
HSB <- read.csv(url(data_path), stringsAsFactors = F) %>%
  tbl_df() %>%
  select(-pracad, -disclim)

# print first few rows
head(HSB)
```

```
## # A tibble: 6 x 9
##   School minority female     ses mathach  size sector himnty meanses
##    <int>    <int>  <int>   <dbl>   <dbl> <int>  <int>  <int>   <dbl>
## 1   1224        0      1 -1.53      5.88   842      0      0  -0.430
## 2   1224        0      1 -0.590    19.7    842      0      0  -0.430
## 3   1224        0      0 -0.530    20.4    842      0      0  -0.430
## 4   1224        0      0 -0.670     8.78   842      0      0  -0.430
## 5   1224        0      0 -0.160    17.9    842      0      0  -0.430
## 6   1224        0      0  0.0200    4.58   842      0      0  -0.430
```

Question 2

Produce basic descriptive information for just these two variables: mathach and ses, using a single command.

```
describe(HSB %>% select(mathach, ses))
```

```
##         vars    n  mean   sd median trimmed  mad   min   max range  skew
## mathach    1 7185 12.75 6.88  13.13   12.92 8.12 -2.83 24.99 27.82 -0.18
## ses        2 7185  0.00 0.78   0.00    0.02 0.85 -3.76  2.69  6.45 -0.23
##         kurtosis   se
## mathach    -0.92 0.08
## ses        -0.38 0.01
```

Question 3

What is the overall correlation between mathach and ses?

```r
with(HSB, cor(mathach, ses))
```

```
## [1] 0.3607626
```

Question 4

Produce a cross-classification table for female and minority. Make sure that the rows and columns of the table have appropriate labels (not just numbers).

```r
HSB %>%
  mutate(female = mapvalues(female, 0:1, c("male", "female")),
         minority = mapvalues(minority, 0:1, c("non-minority", "minority"))) %>%
  group_by(female, minority) %>%
  summarize(n = n()) %>%
  spread(key = female, value = n) %>%
  kable(., "latex", booktabs = T) %>%
  kable_styling(latex_options = c("striped","repeat_header"), full_width = F)
```

minority	female	male
minority	1065	909
non-minority	2730	2481

Question 5

Are the two variables in Question 4 independent of each other?

```r
chi <- chisq.test(HSB$minority, HSB$female)
```

Yes, a χ^2 test of independence suggests that minority status and gender are independent, $\chi^2(1) = 1.34$, $p < 0.25$.

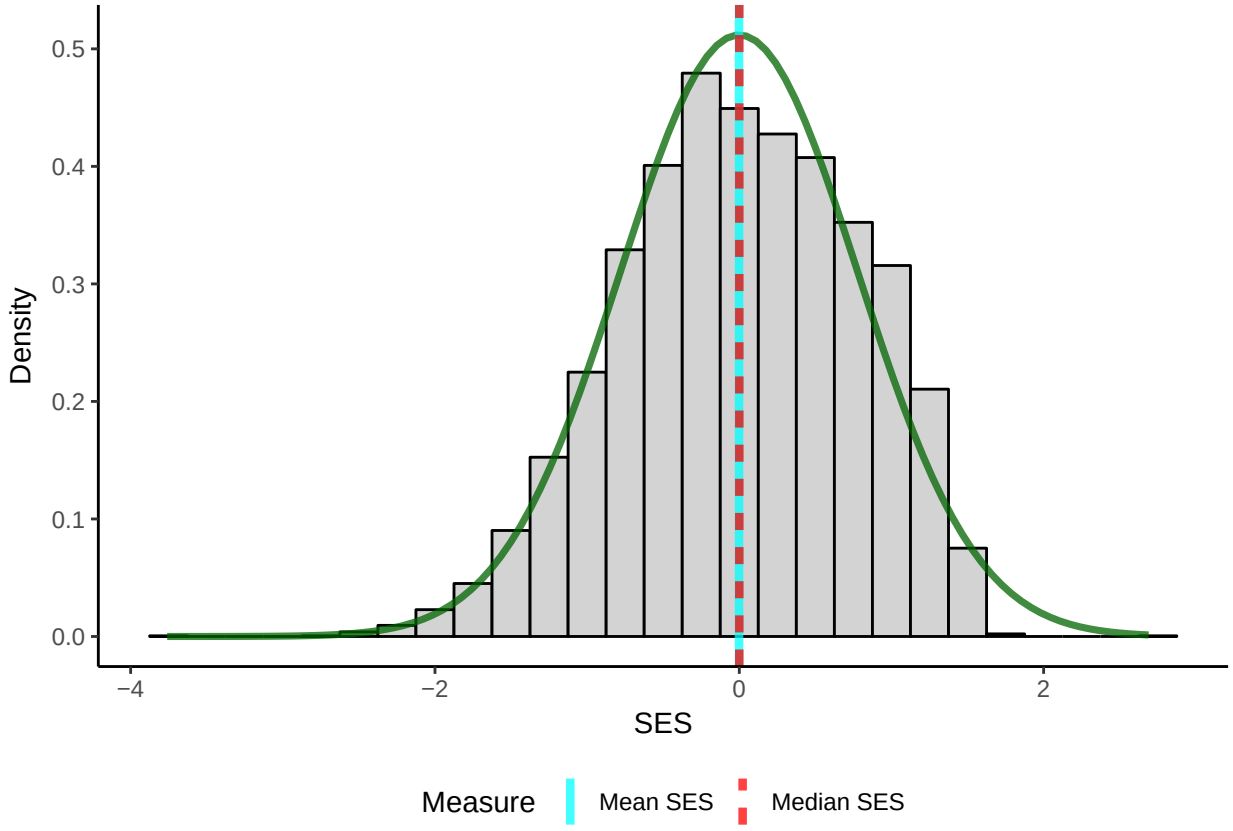
Question 6

Produce a histogram for ses. Include a blue vertical line indicating the mean, a red vertical line indicating the median, and the normal density curve (in green). Make sure the axes are appropriately labeled. Do the data seem to be normally distributed?

The data appear to be mostly normally distributed with a slight negative skew. However, because the median and the mean are nearly indistinguishable and the density curve appears normal, any non-normality is unlikely to greatly influence the results.

```r
tmp <- HSB %>%
  mutate(sdses = sd(ses, na.rm = T), `Median SES` = median(ses, na.rm = T),
         `Mean SES` = mean(ses, na.rm = T))
tmp2 <- tmp %>%
  gather(key = Measure, value = value, `Mean SES`, `Median SES`)
tmp %>%
  ggplot(aes(x = ses)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.25, color = "black",
                 fill = "lightgray") +
```

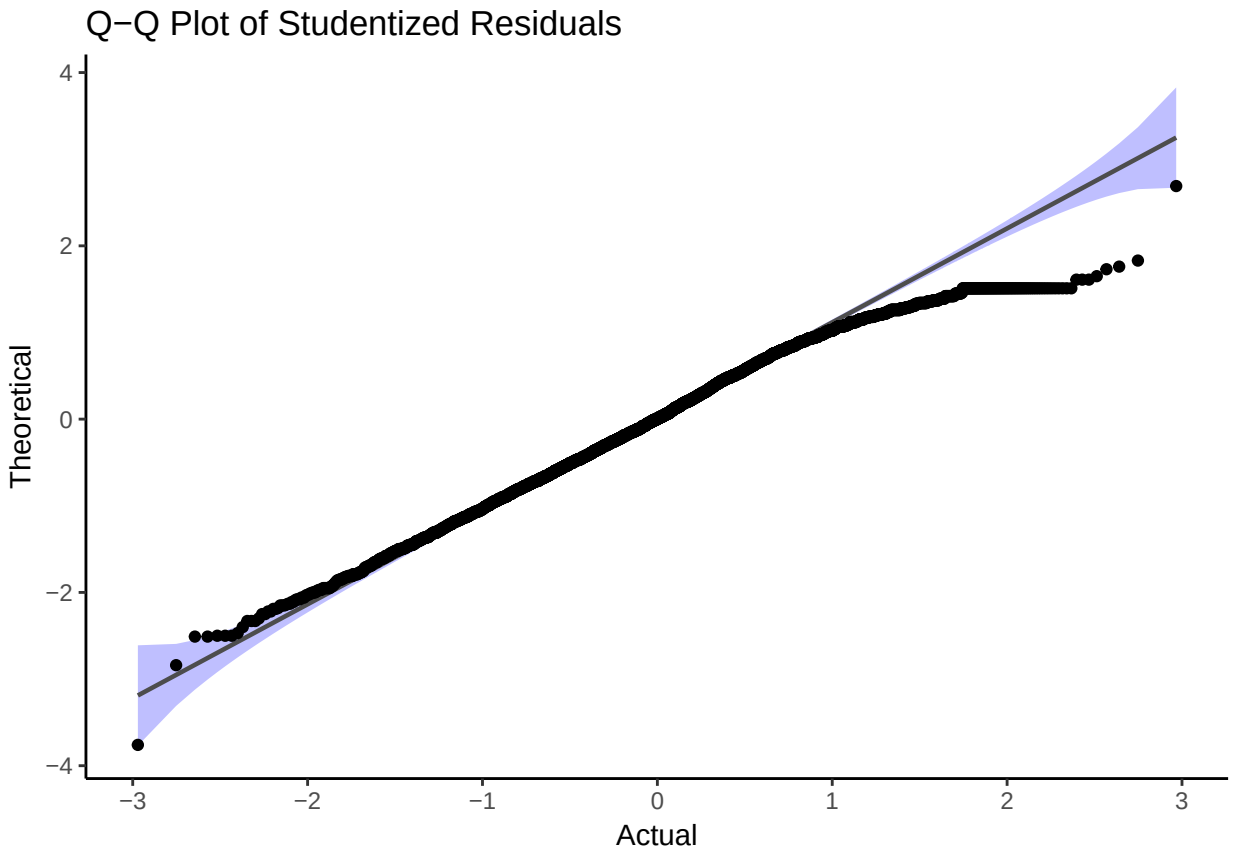
```
  stat_function(fun = dnorm, size = 1.25, color = "darkgreen", alpha = .75,
                args = list(mean = unique(tmp$`Mean SES`), sd = unique(tmp$sdses))) +
  geom_vline(data = tmp2, aes(xintercept = value,
      color = Measure, linetype = Measure), alpha = .75, size = 1.5) +
  scale_color_manual(values = c("cyan", "red")) +
  labs(x = "SES", y = "Density") +
  theme_classic() +
  theme(legend.position = "bottom")
```



Question 7

Produce a Q-Q plot for ses. Does this change your opinion regarding normality?
Looking at the Q-Q plot, earlier concerns about non-normality seem to be more justified. The departure of the data from the theoretic line at the positive extreme highlights the negative skew noted in the histogram.

```
HSB %>%
  ggplot(aes(sample=ses)) +
  stat_qq_band(fill = "blue", alpha = .25) +
  stat_qq_line() +
  stat_qq_point() +
  labs(x = "Actual", y = "Theoretical",
       title = "Q-Q Plot of Studentized Residuals") +
  theme_classic()
```



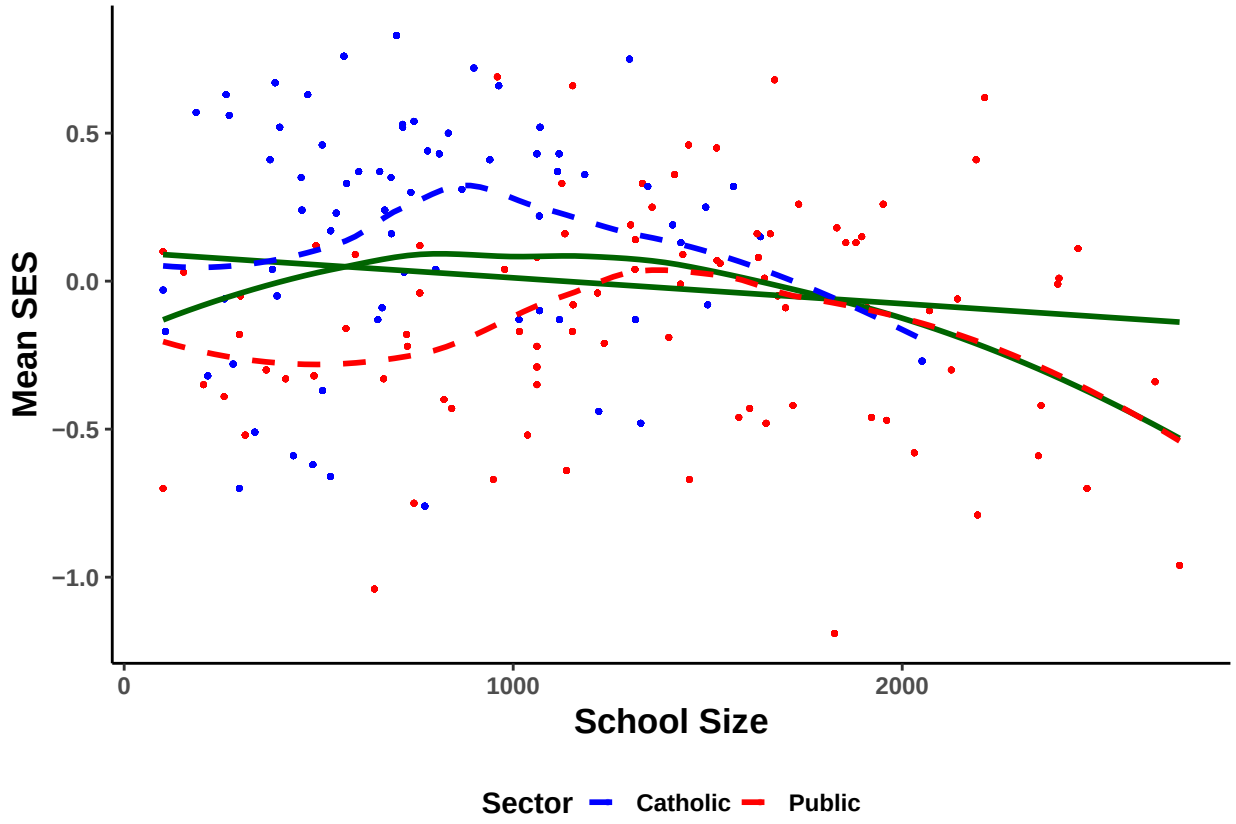
Question 8

Produce a scatterplot of meanses (y axis) versus size (x axis). Make sure the axes are appropriately labeled. Add the best-fitting linear regression line as well as a loess (nonlinear) fit line. Color the plot symbols so that public schools are red and Catholic schools are blue. What does this figure tell you?

Without grouping the best fitting line by sector, the relationship between school size and SES appears to be non-linear – that is, the largest and smallest schools tend to have lower SES, while the middle-sized schools have higher SES. However, closer examination reveals (e.g. the dashed loess lines grouped by sector) sugget that this relationship depends on sector. Catholic schools tend to be smaller than public schools overall and have have higher SES, regardless of size. SES Does not appear to be a function of school size for Catholic school students, but it does appear to be related to the SES of public school students, with middle sized schools having the highest SES.

```
HSB %>%
  mutate(sector = mapvalues(sector, 0:1, c("Public", "Catholic"))) %>%
  ggplot(aes(x = size, y = meanses)) +
  scale_color_manual(values = c("blue", "red")) +
  geom_point(aes(color = sector), alpha = .5, size = .5) +
  geom_smooth(method = "lm", se = F, color = "darkgreen") + #aes(color = sector),
  geom_smooth(method = "loess", se = F, color = "darkgreen") + #aes(color = sector),
  geom_smooth(aes(color = sector), method = "loess", se = F, linetype = "dashed") + #
  labs(x = "School Size", y = "Mean SES", color = "Sector") +
  theme_classic() +
  theme(axis.text = element_text(face = "bold"),
        axis.title = element_text(face = "bold", size = rel(1.2)),
```

```
        legend.text = element_text(face = "bold"),
        legend.title = element_text(face = "bold"),
        legend.position = "bottom")
```



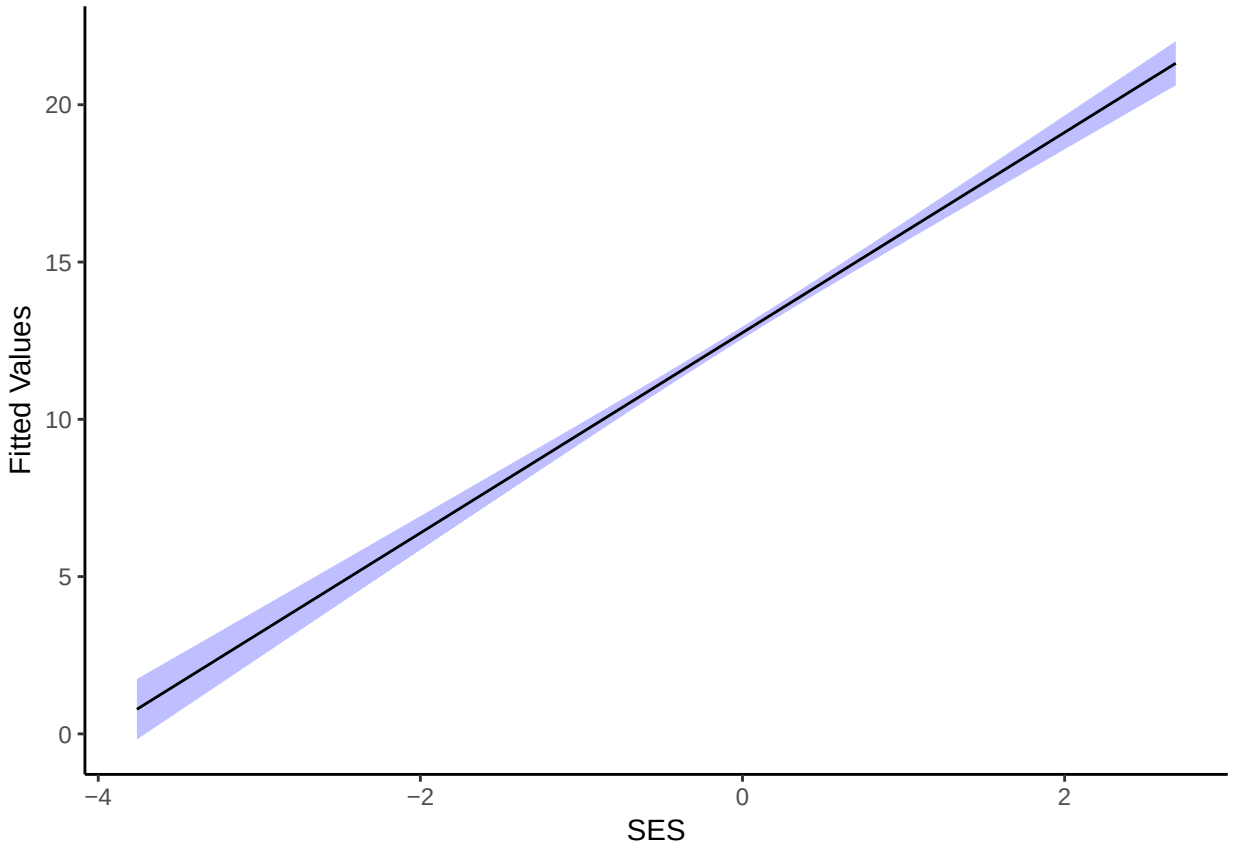
This figure seems to suggest

Question 9

Plot the best-fitting line relating mathach to ses and include the 99% confidence interval around the line.

```
m <- lm(mathach ~ ses, data = HSB)

cbind(HSB, predict(m, interval = "conf", level = .99)) %>%
  ggplot(aes(x = ses, y = fit)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = .25, fill = "blue") +
  geom_line() +
  labs(x = "SES", y = "Fitted Values") +
  theme_classic()
```

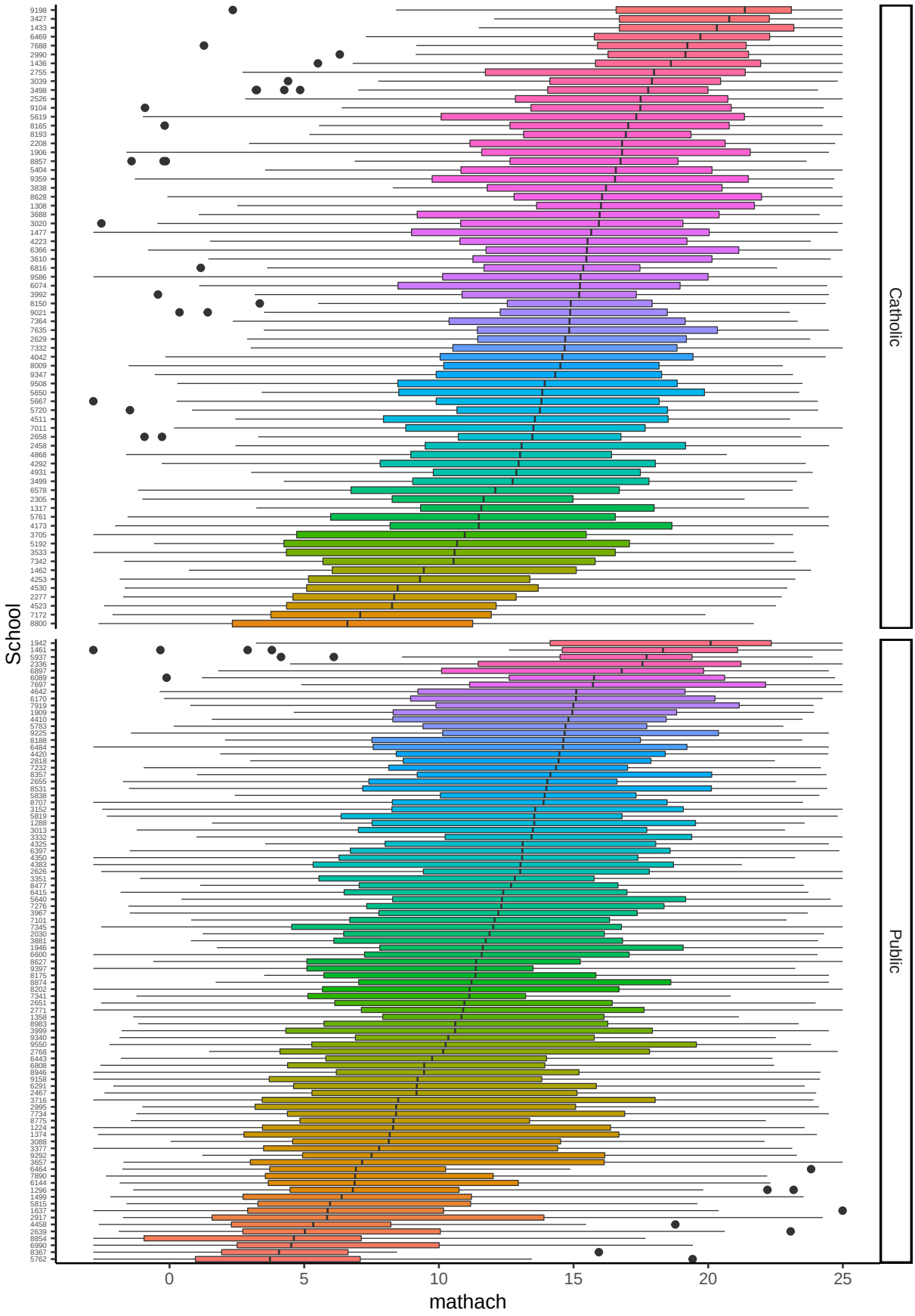


Question 10

Produce a two-panel plot. In the upper panel, show the boxplots of mathach separately for each public school. In the lower panel, show the boxplots of mathach separately for each Catholic school.

```
orders <- HSB %>%
  group_by(School) %>%
  summarize(median = median(mathach, na.rm = T)) %>%
  arrange(median)

HSB %>%
  mutate(sector = mapvalues(sector, 0:1, c("Public", "Catholic")),
         School = factor(School, levels = orders$School)) %>%
  ggplot(aes(x = School, y = mathach, fill = School)) +
  geom_boxplot(size = .25) +
  coord_flip() +
  facet_grid(sector~., scales = "free_y") +
  theme_classic() +
  theme(legend.position = "none",
        axis.text.y = element_text(size = rel(.5)))
```


School
Catholic
Public
mathach
0
5
10
15
20
25