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   -0.43
## 2   1224        0      1 -0.59   19.71   842      0      0   -0.43
## 3   1224        0      0 -0.53   20.35   842      0      0   -0.43
## 4   1224        0      0 -0.67    8.78   842      0      0   -0.43
## 5   1224        0      0 -0.16   17.90   842      0      0   -0.43
## 6   1224        0      0  0.02    4.58   842      0      0   -0.43
```

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?

```
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).

```
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?

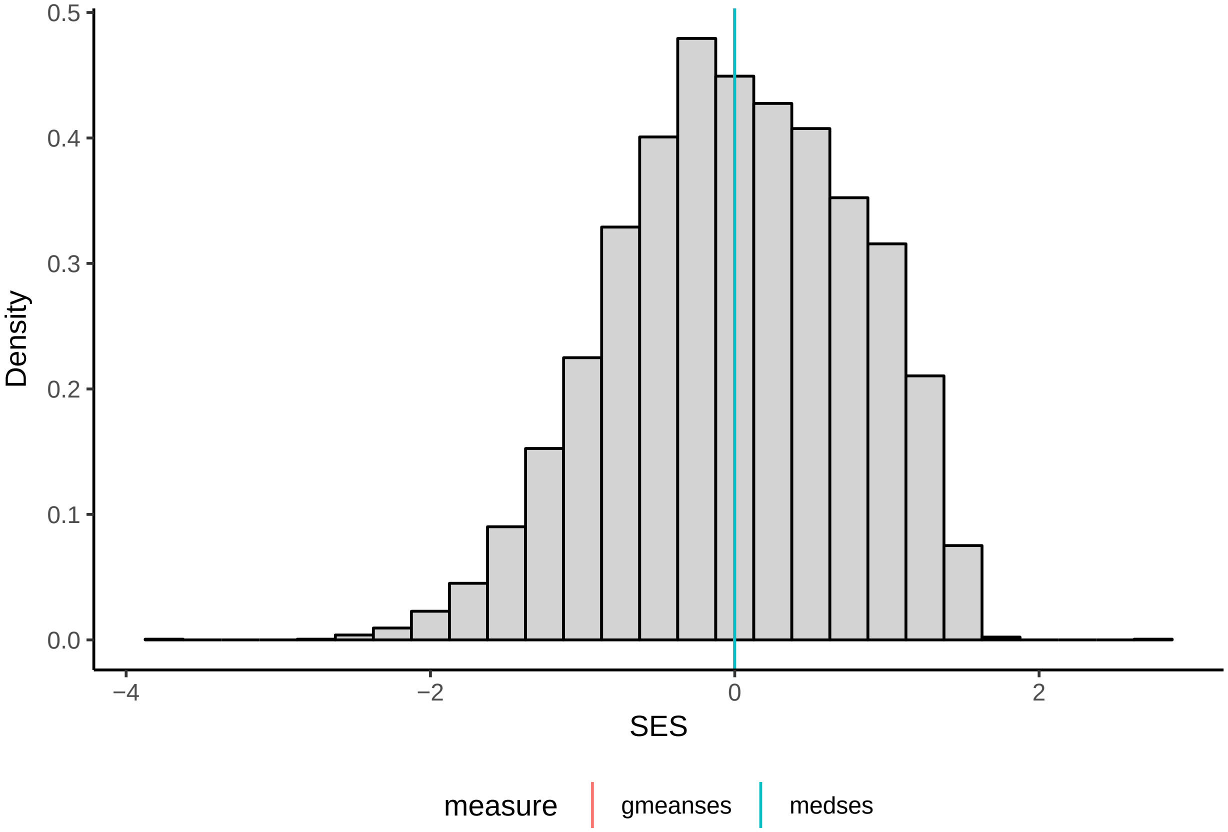
```
chi <- chisq.test(HSB$minority, HSB$female)
```

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?

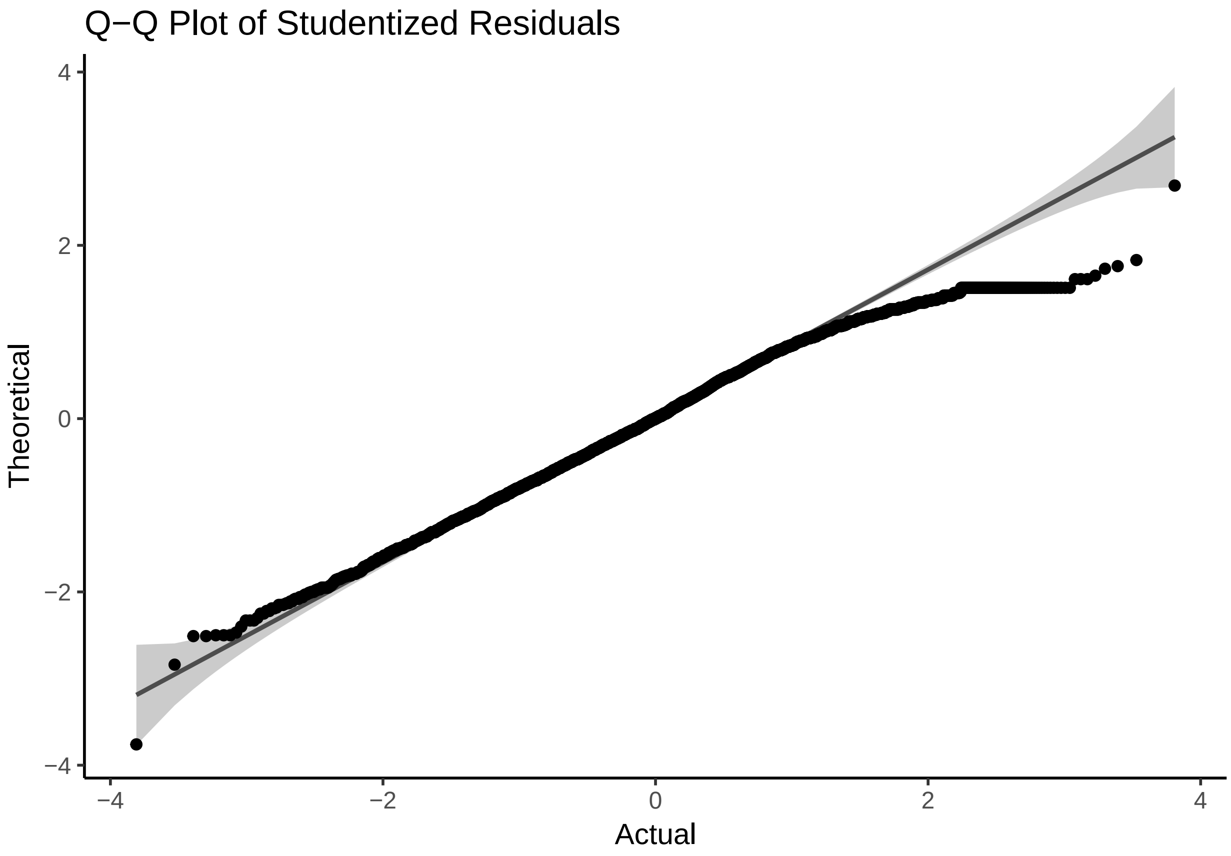
```
tmp <- HSB %>%
  mutate(sdses = sd(ses, na.rm = T), medses = median(ses, na.rm = T),
         gmeanses = mean(ses, na.rm = T))
tmp2 <- tmp %>%
  gather(key = measure, value = value, gmeanses, medses)
tmp %>%
  ggplot(aes(x = ses)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.25, color = "black",
                 fill = "lightgray") +
  stat_function(fun = dnorm, size = 1.25, color = "darkgreen",
                args = list(mean = HSB$gmeanses, sd = HSB$sdses)) +
  geom_vline(data = tmp2, aes(xintercept = value, color = measure)) +
  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?

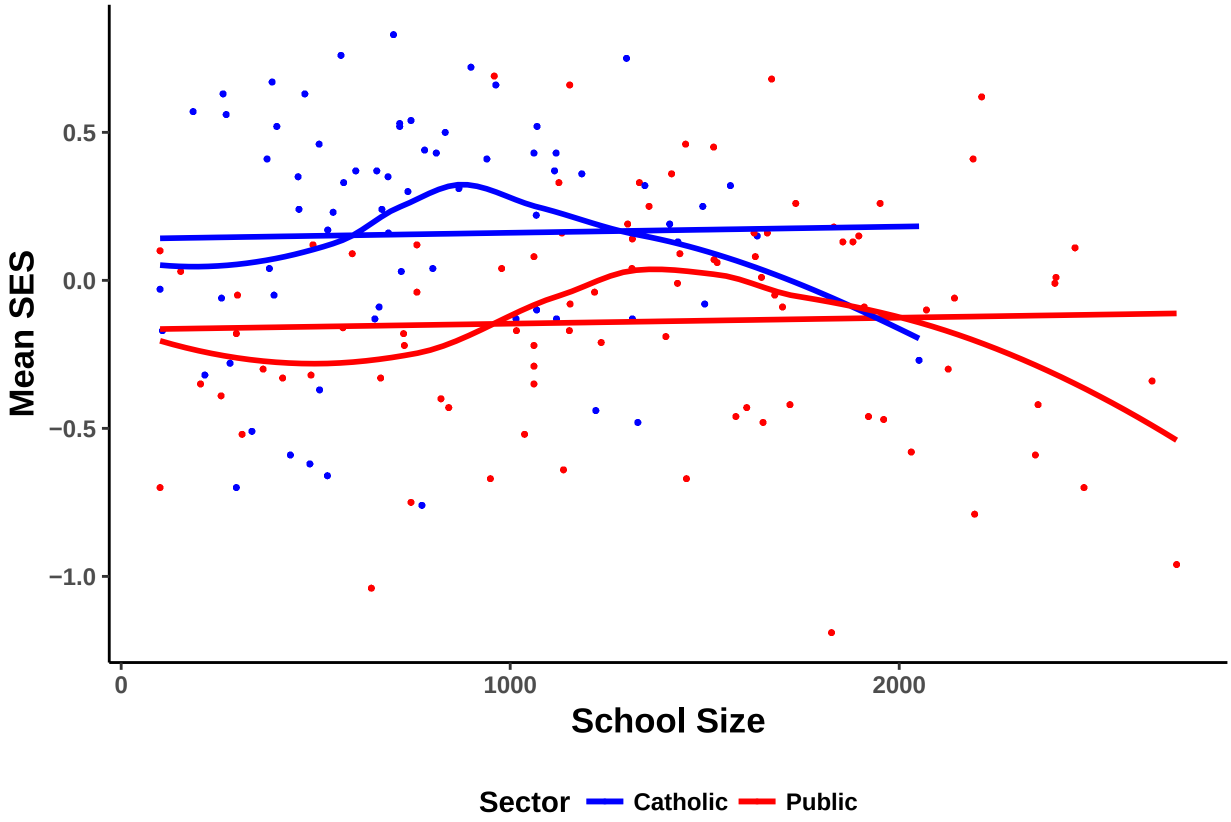
```
HSB %>%
  ggplot(aes(sample=ses)) +
  stat_qq_band() +
  stat_qq_line() +
  stat_qq_point() +
  labs(x = "Actual", y = "Theoretical",
       title = "Q-Q Plot of Studentized Residuals") +
  theme_classic()
```



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?

```
HSB %>%
  mutate(sector = mapvalues(sector, 0:1, c("Public", "Catholic"))) %>%
  ggplot(aes(x = size, y = meanses, color = sector)) +
  scale_color_manual(values = c("blue", "red")) +
  geom_point(aes(color = sector), alpha = .5, size = .5) +
  geom_smooth(method = "lm", se = F) +
  geom_smooth(method = "loess", se = F) +
  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")
```

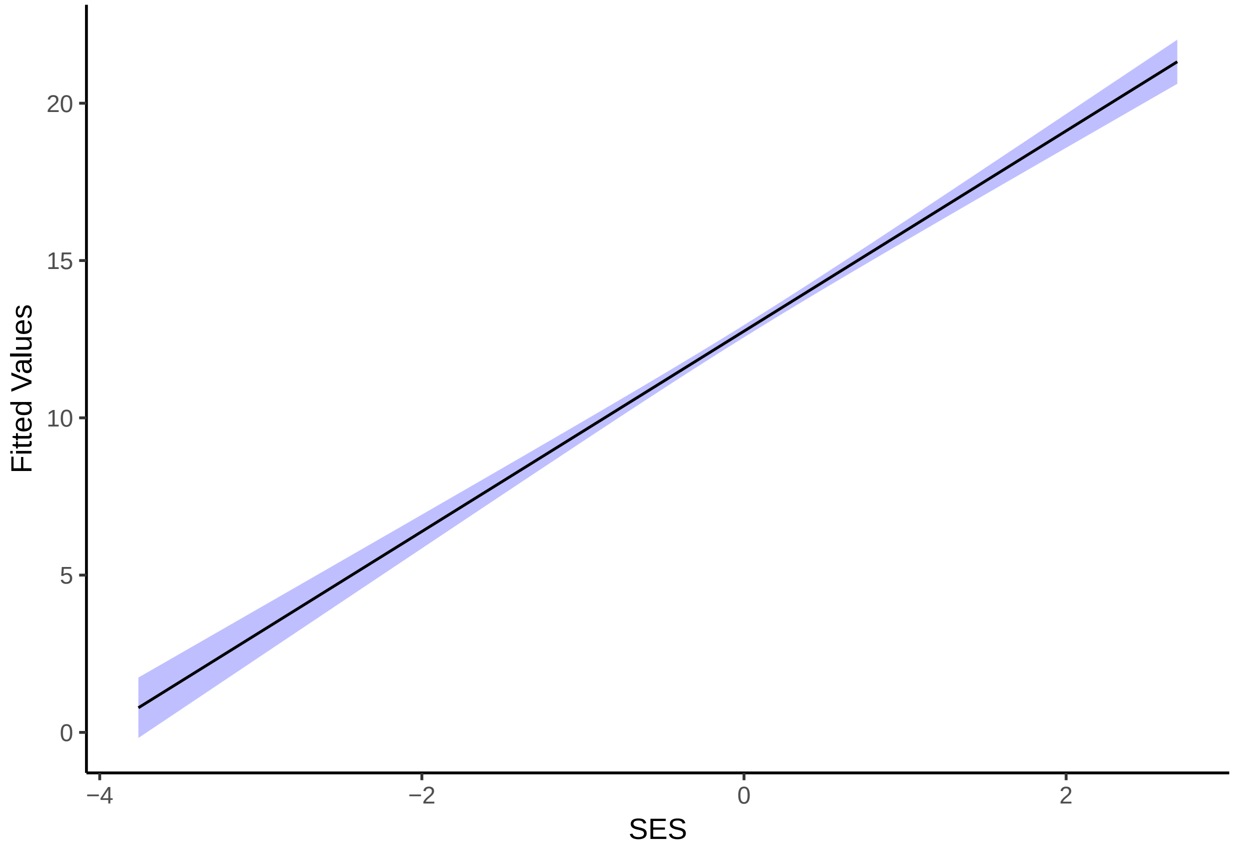


Question 9

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()
```

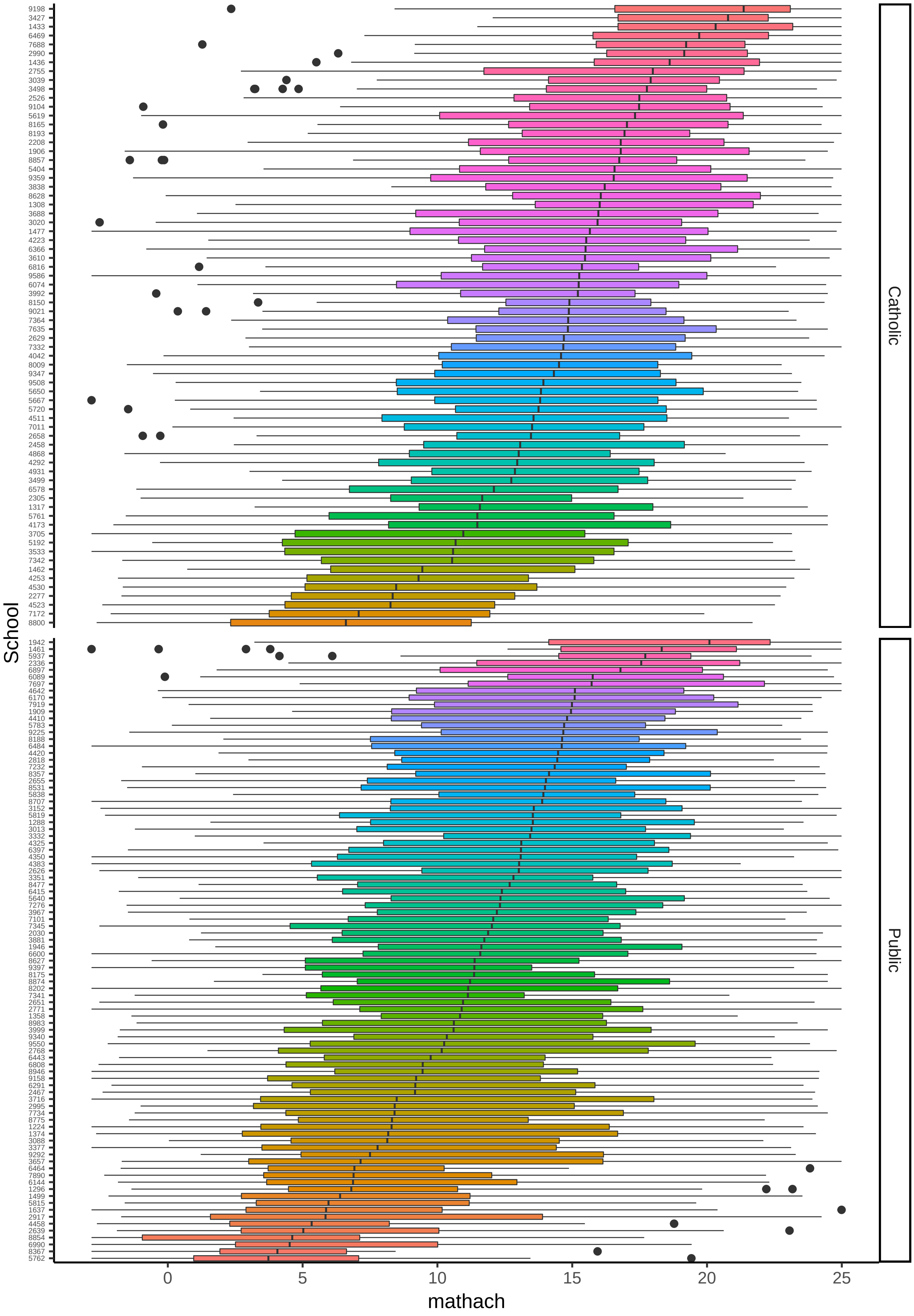


Question 10

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
9198
3427
1433
6469
7688
2990
1436
2755
3039
3498
2526
9104
5619
8165
8193
2208
1906
8857
5404
9359
3838
8628
1308
3688
3020
1477
4223
6366
3610
6816
9586
6074
3992
8150
9021
7364
7635
2629
7332
4042
8009
9347
9508
5650
5667
5720
4511
7011
2658
2458
4868
4292
4931
3499
6578
2305
1317
5761
4173
3705
5192
3533
7342
1462
4253
4530
2277
4523
7172
8800
Catholic
1942
1461
5937
2336
6897
6089
7697
4642
6170
7919
1909
4410
5783
9225
8188
6484
4420
2818
7232
8357
2655
8531
5838
8707
3152
5819
1288
3013
3332
4325
6397
4350
4383
2626
3351
8477
6415
5640
7276
3967
7101
7345
2030
3881
1946
6600
8627
9397
8175
8874
8202
7341
2651
2771
1358
8983
3999
9340
9550
2768
6443
6808
8946
9158
6291
2467
3716
2995
7734
8775
1224
1374
3088
3377
9292
3657
6464
7890
6144
1296
1499
5815
1637
2917
4458
2639
8854
6990
8367
5762
Public
0
5
10
15
20
25
mathach