graphs

Helen Yuan – 106145727

2024-04-27

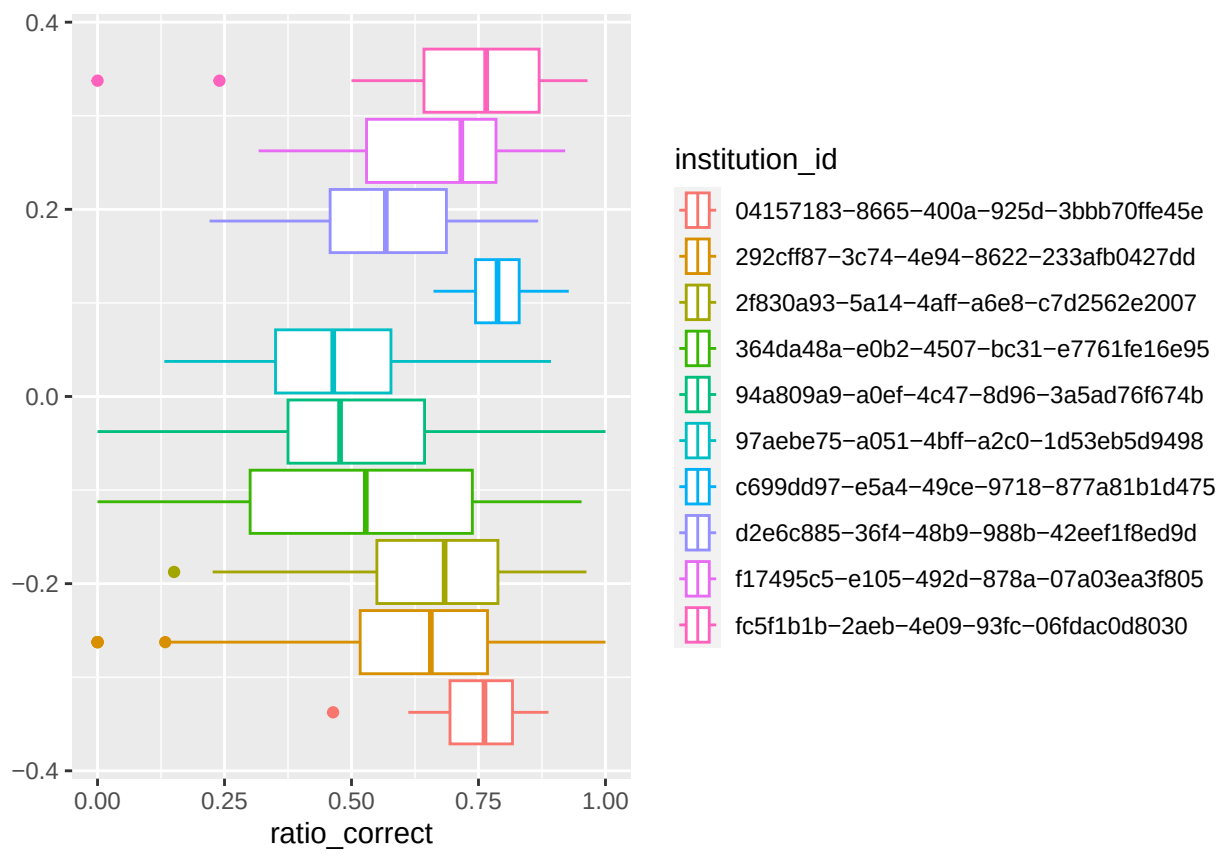
Load stuff

EOC as students progress through the books

In general students' EOC scores get lower and lower as they progress through the book.

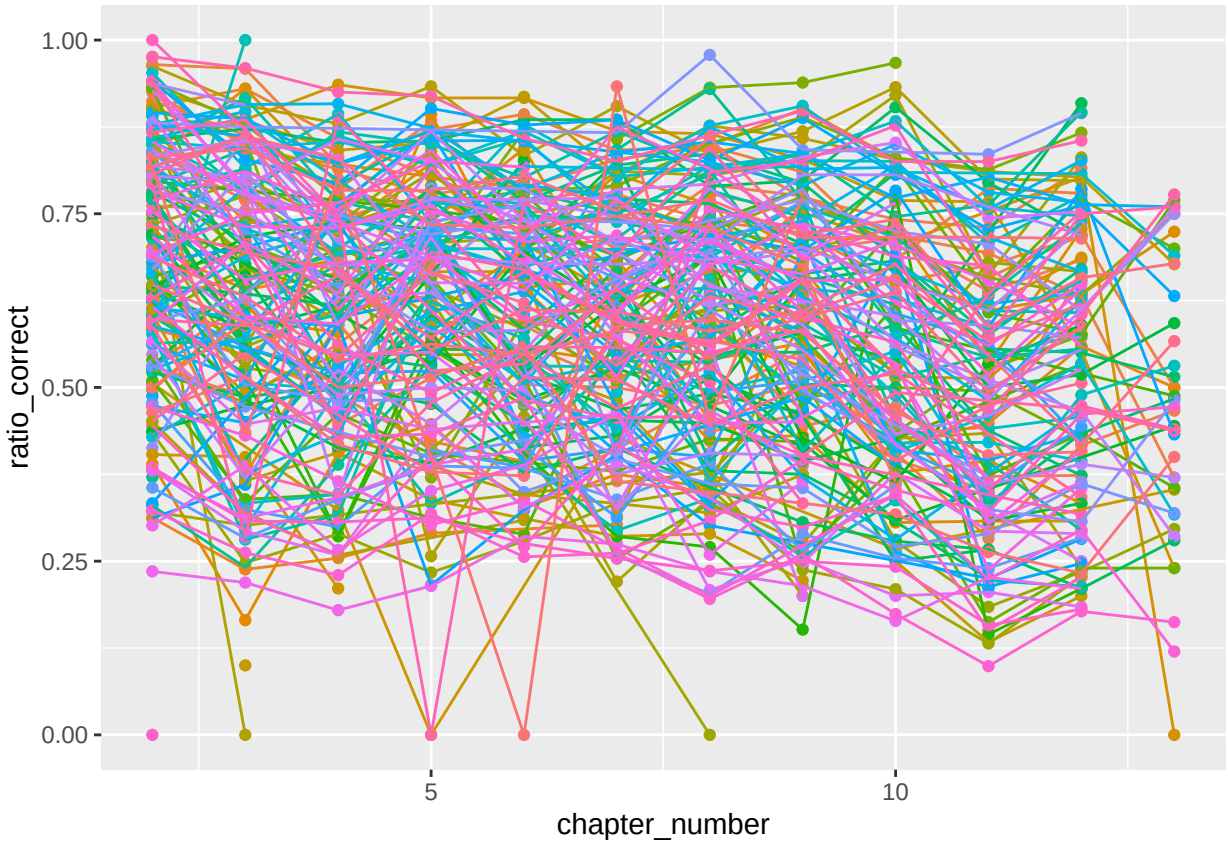
levels(big$book) [1] "College / Advanced Statistics and Data Science (ABCD)"
[2] "College / Statistics and Data Science (ABC)"
[3] "High School / Advanced Statistics and Data Science I (ABC)"

```
big |> ggplot(aes(x = ratio_correct, color = institution_id)) +
  geom_boxplot() # ratio_correct by institution
```

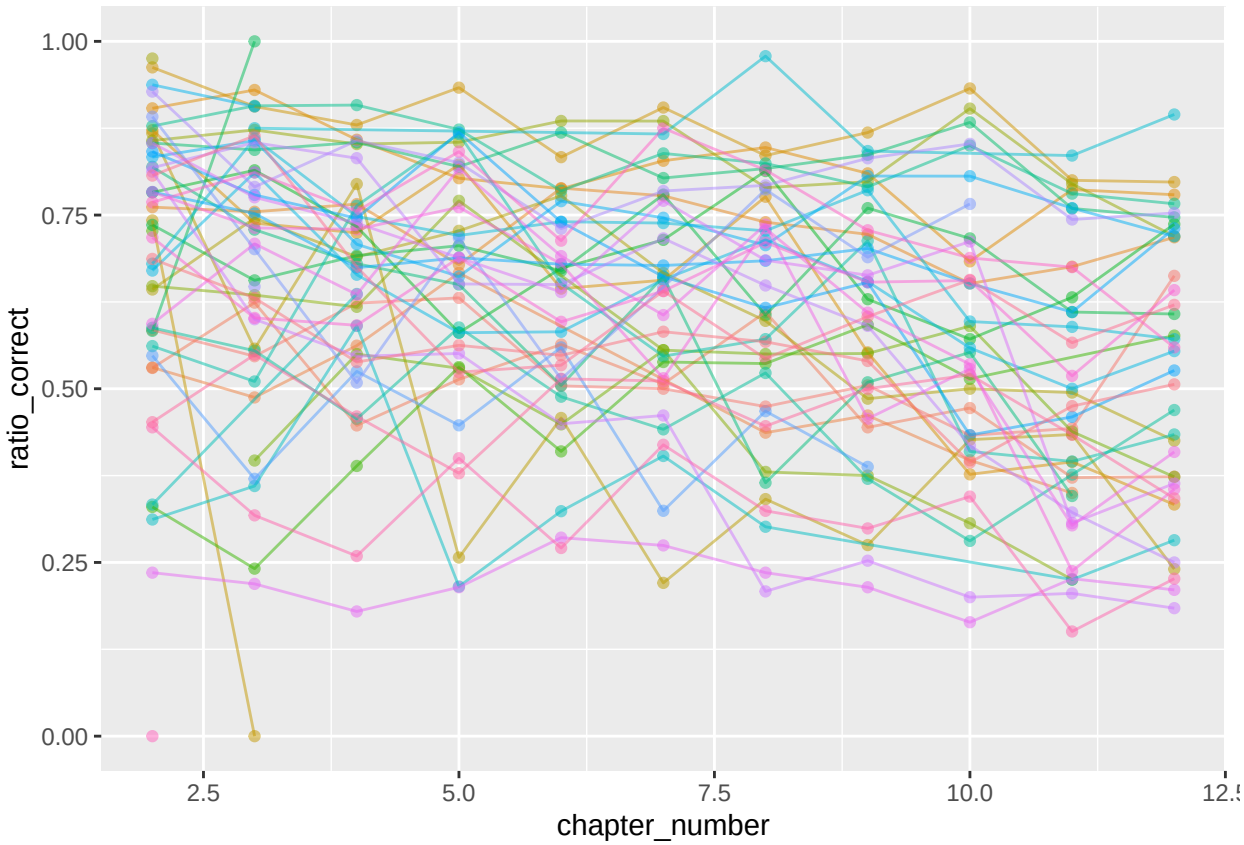



```
# Convert book column to factor
big$book <- factor(big$book)
```

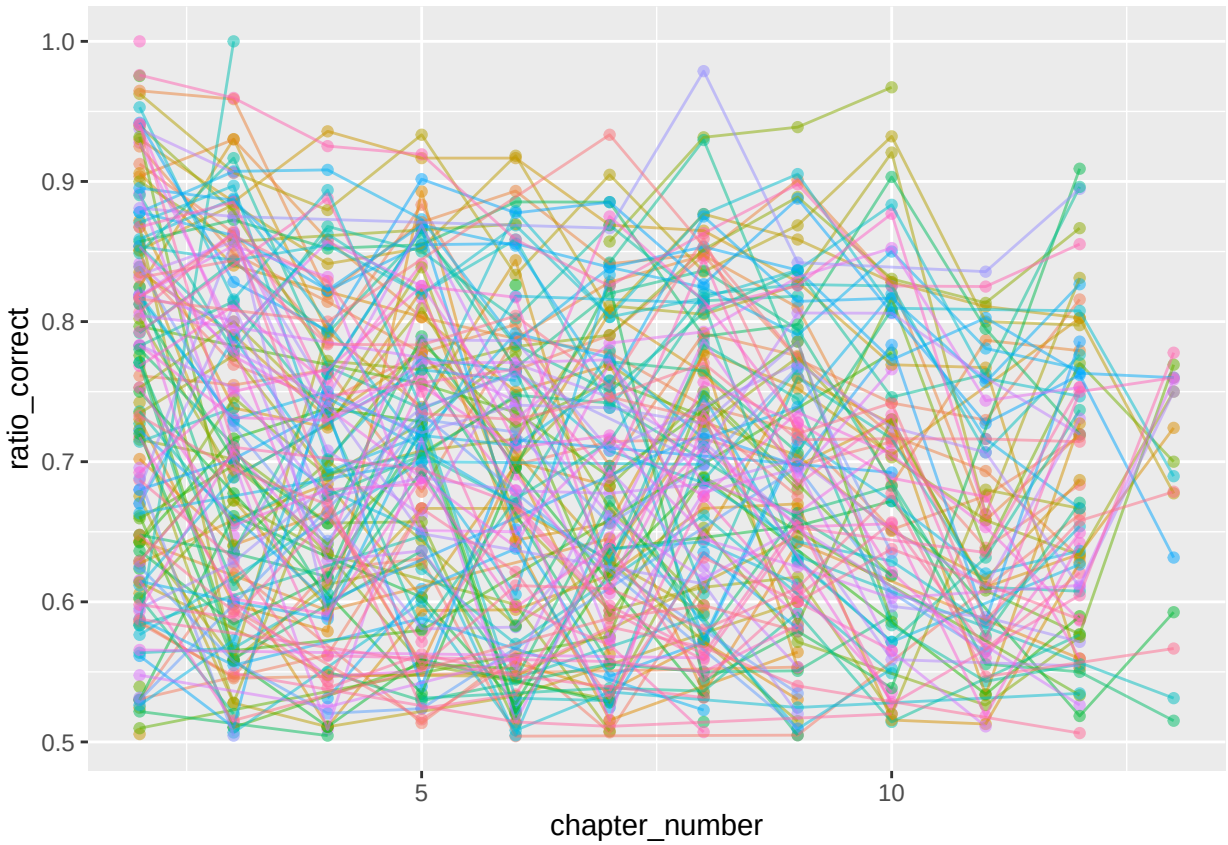
```r
big |> ggplot(aes(x = chapter_number, y = ratio_correct), group = student_id) +
  geom_point(aes(color = student_id), show.legend = FALSE) +
  geom_line(aes(color = student_id), show.legend = FALSE) # this is too cluttered
```




```r
big[as.numeric(big$book) == 1,] |>
  ggplot(aes(x = chapter_number, y = ratio_correct), group = student_id) +
  geom_point(aes(color = student_id), show.legend = FALSE, alpha = 0.5) +
  geom_line(aes(color = student_id), show.legend = FALSE, alpha = 0.5) # slightly less so
```

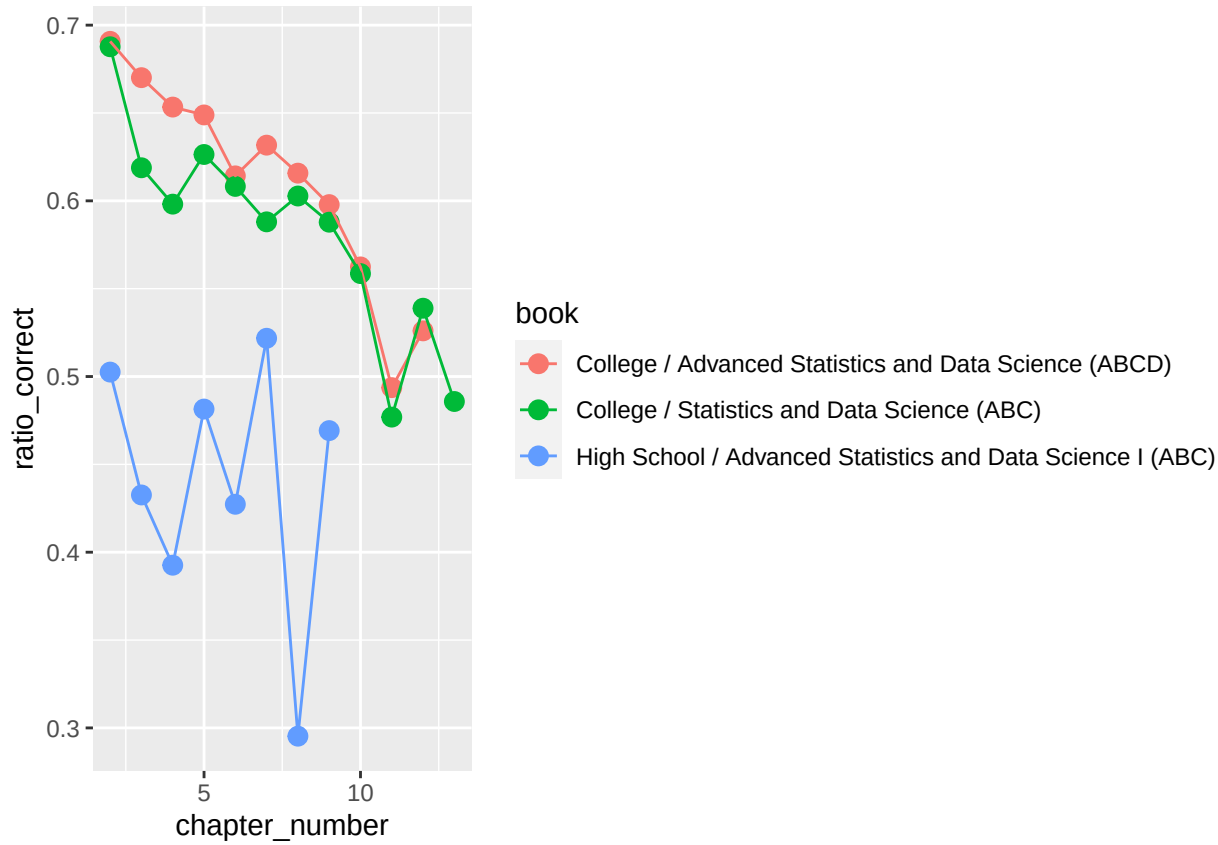



```
big[big$ratio_correct > 0.5,] |>
  ggplot(aes(x = chapter_number, y = ratio_correct), group = student_id) +
  geom_point(aes(color = student_id), show.legend = FALSE, alpha = 0.5) +
  geom_line(aes(color = student_id), show.legend = FALSE, alpha = 0.5)
```




```
big |> group_by(book, chapter_number) |> summarize(ratio_correct = mean(ratio_correct, na.rm = TRUE)) |>
  geom_point(cex = 3) +
  geom_line() # chapter_number vs ratio_correct by book
```

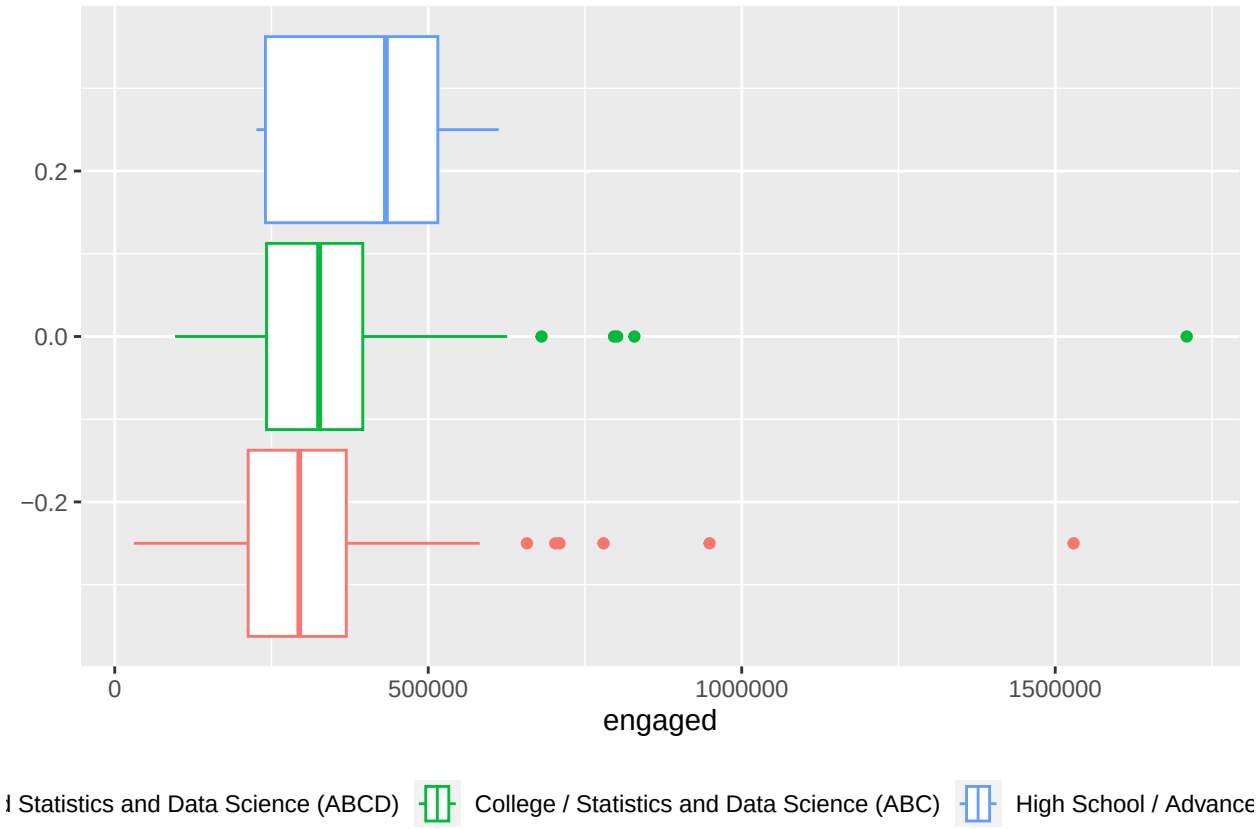
```
## `summarise()` has grouped output by 'book'. You can override using the
## `.groups` argument.
```



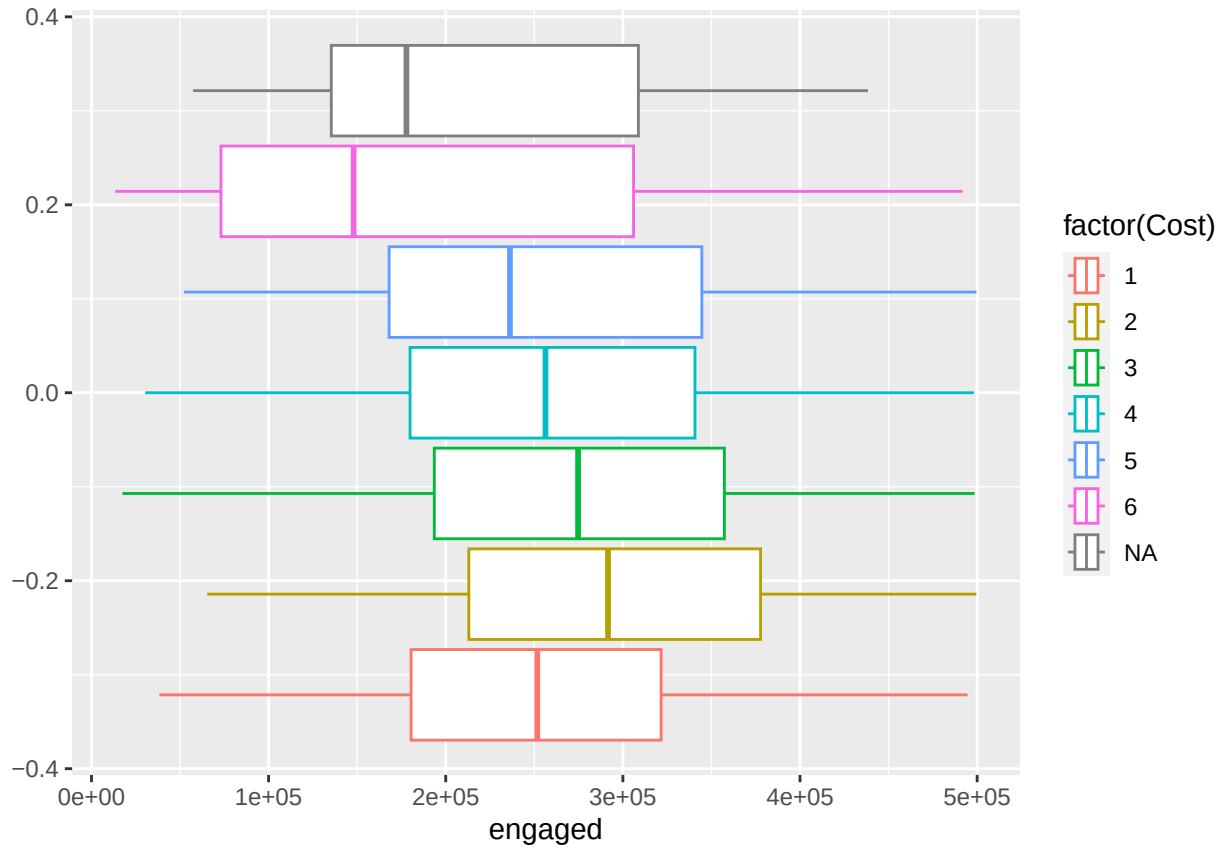
Why are high schoolers scoring so low? Seems like they don't differ too much in time engaged/idle/off page or anything actually maybe they're just stupid

```
# Remove outliers
big |> group_by(book, student_id) |> summarize(engaged = mean(engaged, na.rm = TRUE), ratio_correct = me
  ggplot(aes(x = engaged, color = book)) +
  geom_boxplot() +
  theme(legend.position = "bottom")
```

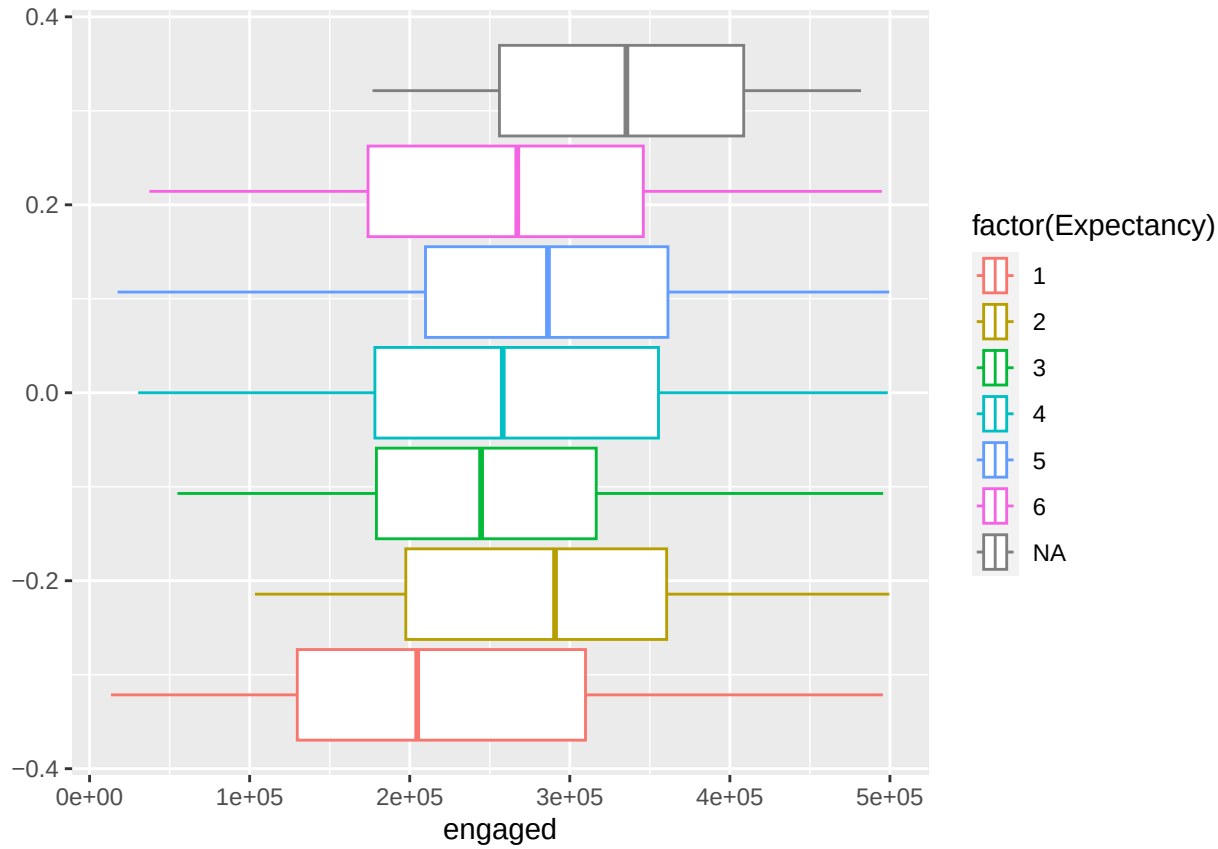
```
## `summarise()` has grouped output by 'book'. You can override using the
## `.groups` argument.
```




```
big |> filter(engaged < 5e+05) |> ggplot(aes(x = engaged)) +
  geom_boxplot(aes(color = factor(Cost))) # engagement faceted by Cost
```

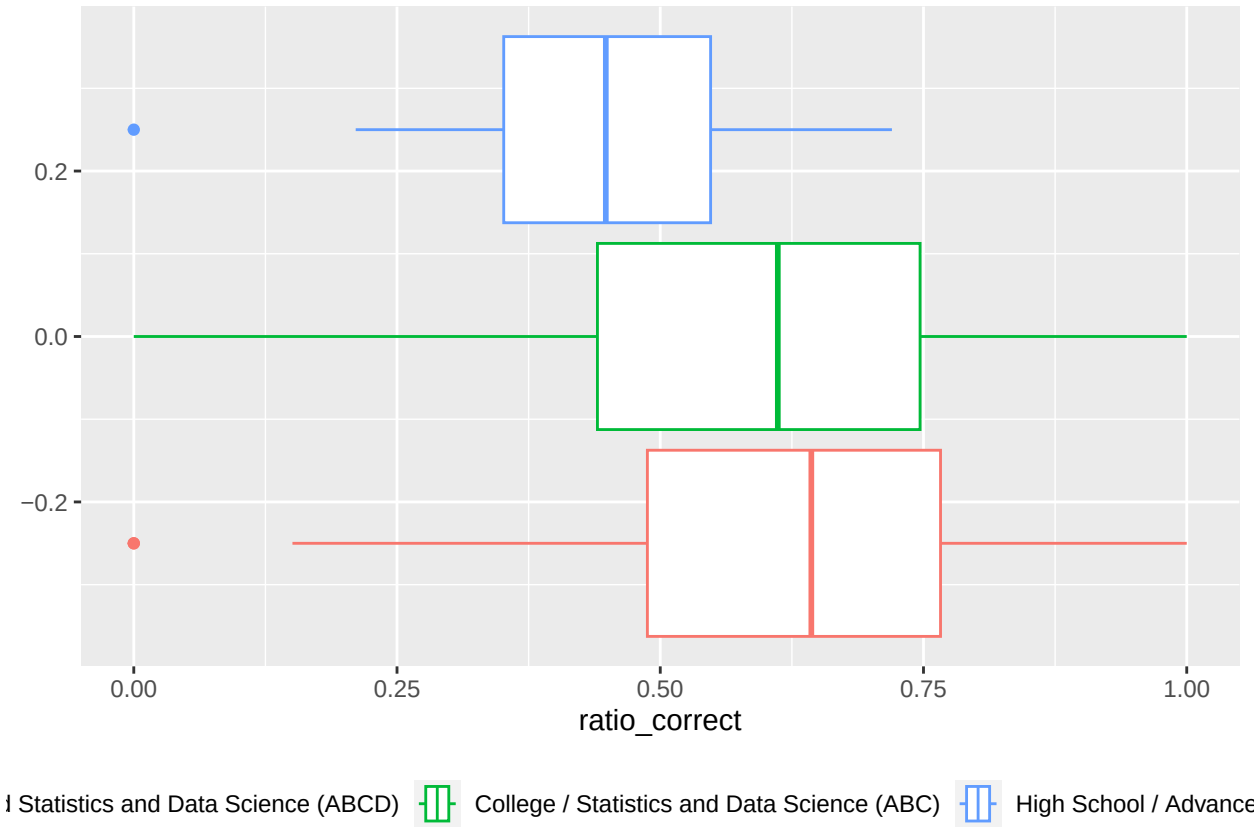



```
big |> filter(engaged < 5e+05) |> ggplot(aes(x = engaged)) +
  geom_boxplot(aes(color = factor(Expectancy))) # engagement faceted by Expectancy
```



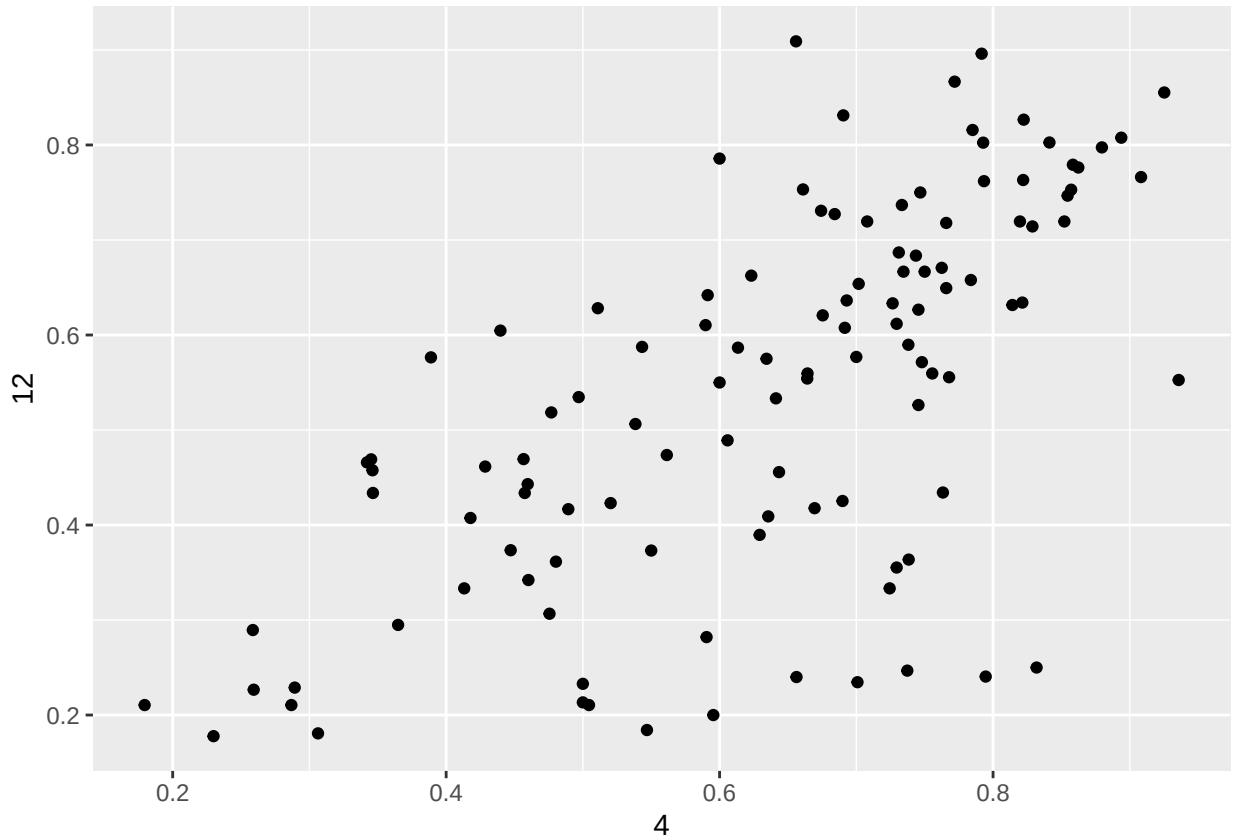
Are students who start out struggling still struggling?

```
# Facet by book
big |> ggplot(aes(x = ratio_correct)) +
  geom_boxplot(aes(color = book)) +
  theme(legend.position = "bottom")
```




```
# Summarize by student for joining purposes
summarized <- big |> group_by(student_id) |> summarize(engaged = mean(engaged, na.rm = TRUE), Intrinsic

# Chapter 4 avg scores against Chapter 10
big |> pivot_wider(id_cols = c(institution_id, student_id), names_from = chapter_number, values_from = 
  geom_point()
```

```
## Warning: Removed 77 rows containing missing values (`geom_point()`).
```

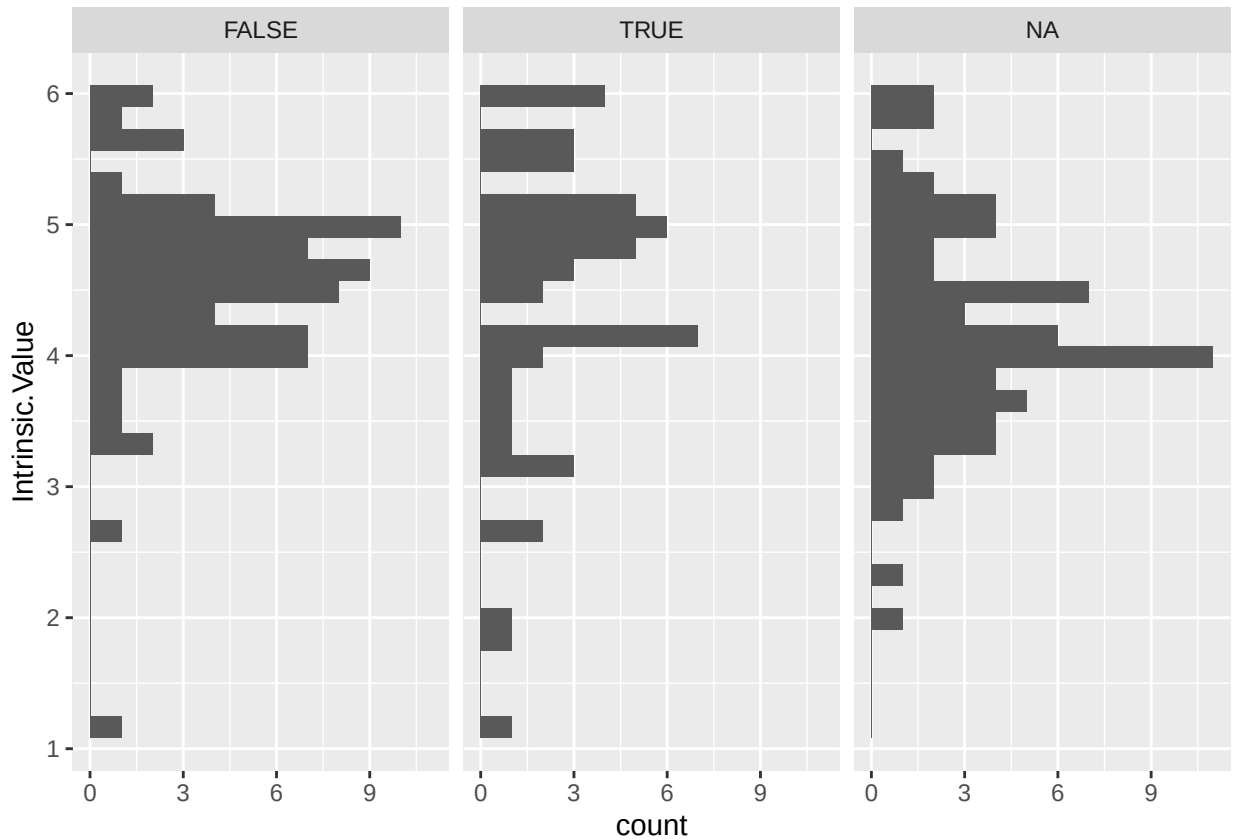



```r
# Do students who do well differ in some way??
summary(big$engaged) # A broad view of the data
```

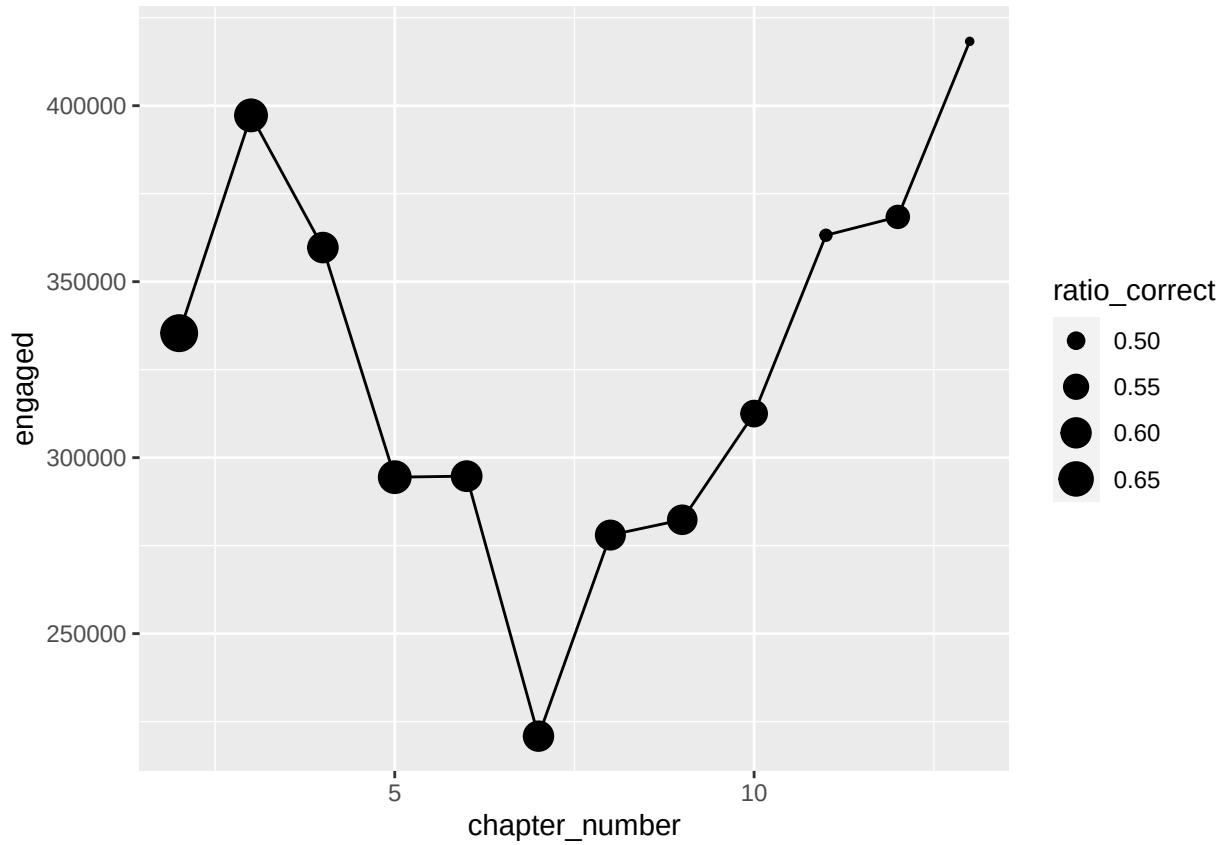
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   13401  198589  294728  342333  402051 5296517       1
```

```r
big |> pivot_wider(id_cols = c(institution_id, student_id), names_from = chapter_number, values_from = 
  ggplot(aes(y = Intrinsic.Value)) +
  geom_histogram() +
  facet_wrap((`12`< 0.5)~., nrow = 1) +
  theme(legend.position = "bottom") # facet_wrapped histograms
```

```
## Joining with `by = join_by(student_id)`
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



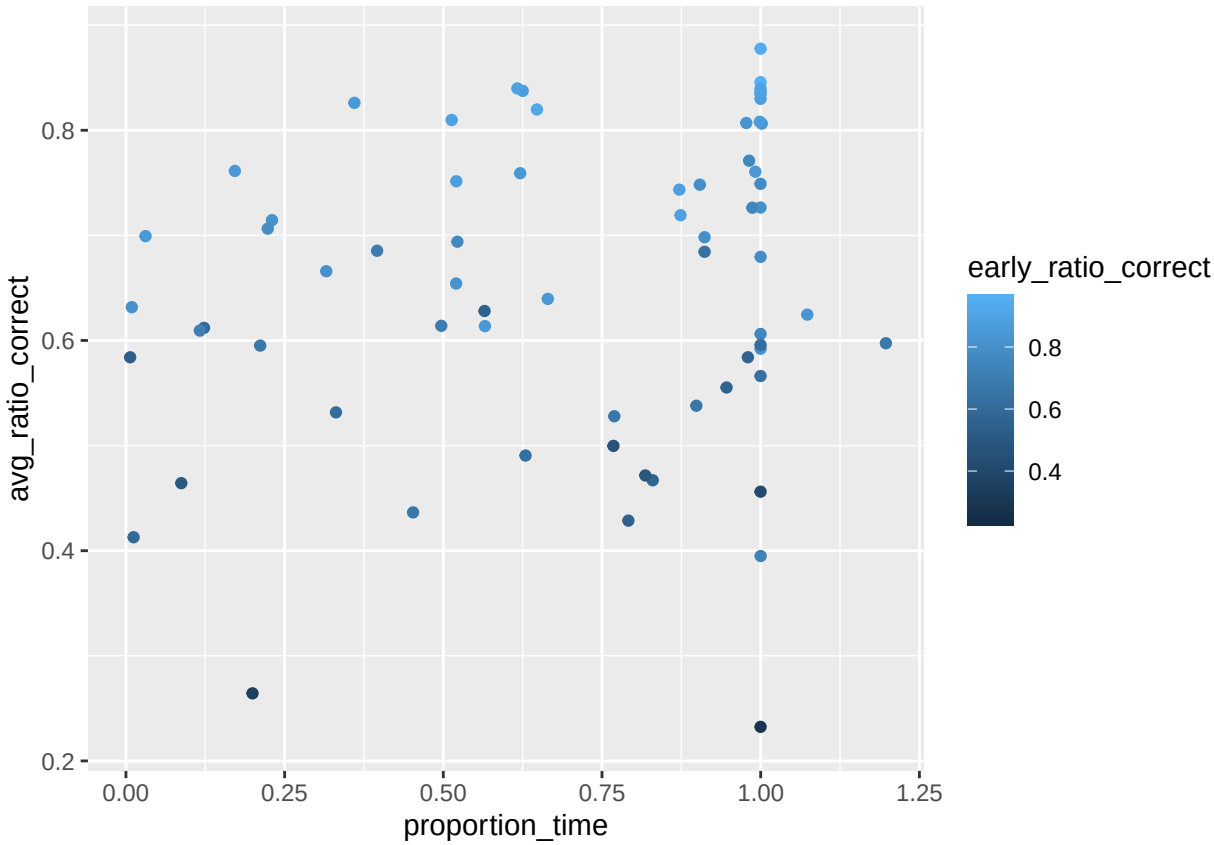

```
# Engagement vs chapter_number
big |> filter(engaged < 1e+06) |> group_by(chapter_number) |> summarize(engaged = mean(engaged, na.rm =
  geom_point(aes(size = ratio_correct)) +
  geom_line()
```




```
early_ratio_correct <- big |> pivot_wider(id_cols = c(institution_id, student_id),
                                          names_from = chapter_number,
                                          values_from = ratio_correct) |>
  mutate(early_ratio_correct = (`2` + `3`)/2,
         avg_ratio_correct = (`2`+`3`+`4`+`5`+`6`+`7`+`8`+`9`+`10`+`11`+`12`)/11)

big |> group_by(student_id) |> summarize(proportion_time = mean(proportion_time, na.rm = TRUE)) |> righ
  ggplot(aes(x = proportion_time, y = avg_ratio_correct)) +
  geom_point(aes(color = early_ratio_correct))
```

```
## Joining with `by = join_by(student_id)`
```

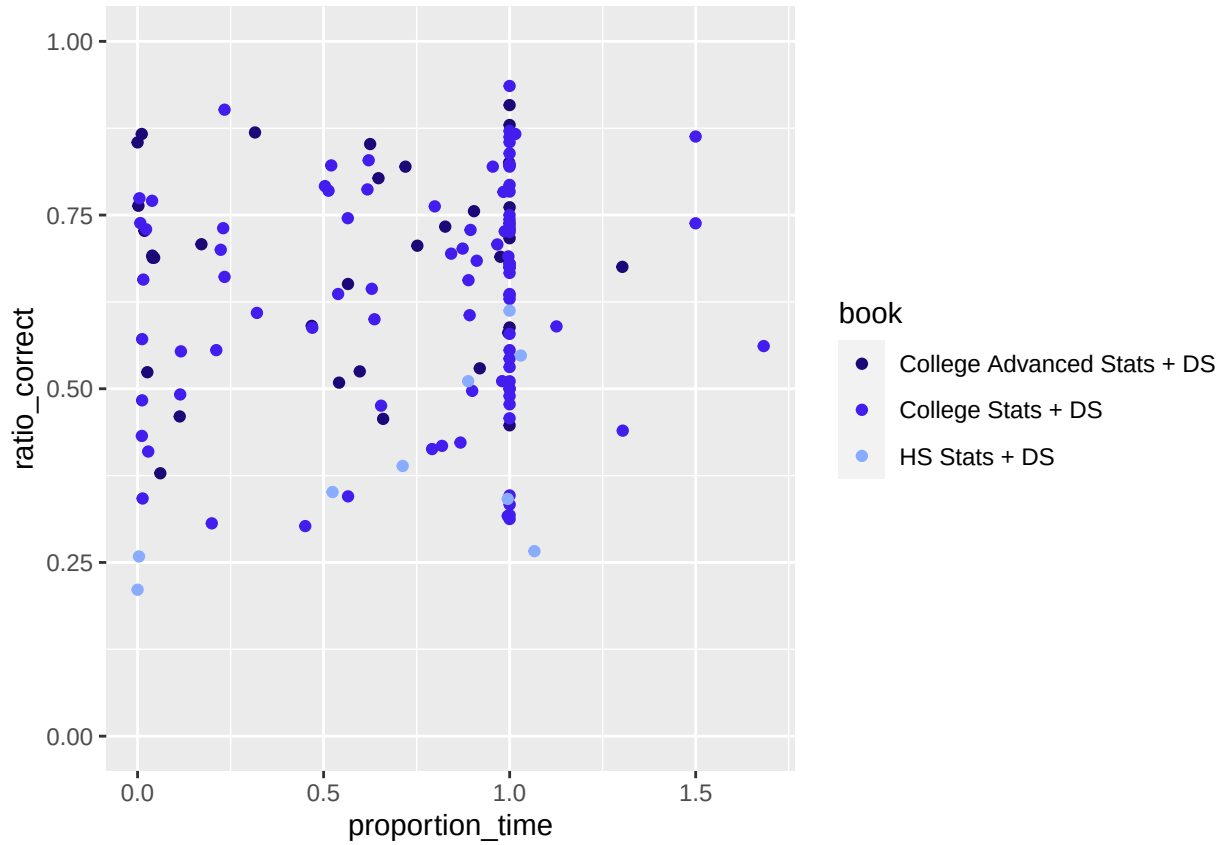
```
## Warning: Removed 125 rows containing missing values (`geom_point()`).
```



Insights from media views

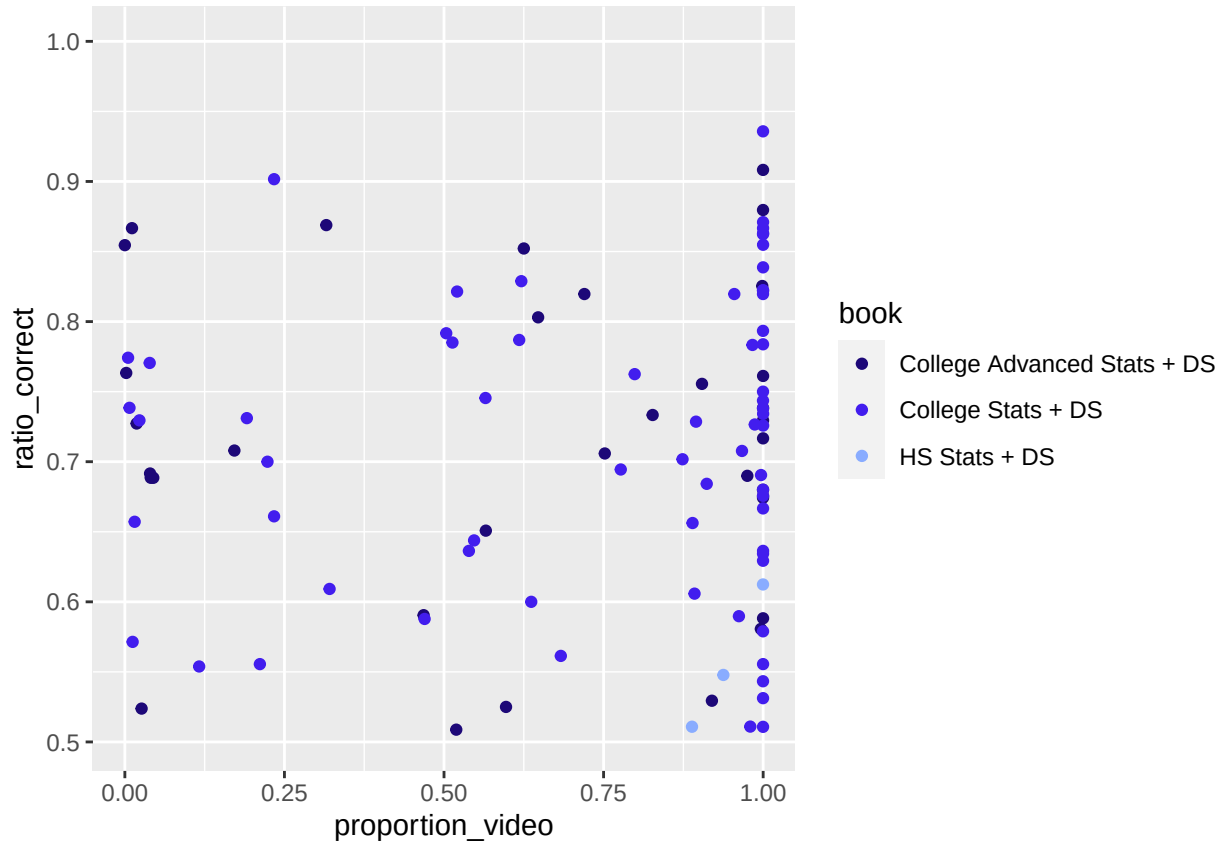
```
big |> ggplot(aes(x = proportion_time, y = ratio_correct)) +
  geom_point(aes(color = book)) +
  scale_color_manual(values = c("#1D0878", "#421DEE", "#89ACFF"),
                     labels = c("College Advanced Stats + DS", "College Stats + DS",
                                "HS Stats + DS")) # proportion_TIME vs ratio_correct
```

```
## Warning: Removed 1573 rows containing missing values (`geom_point()`).
```



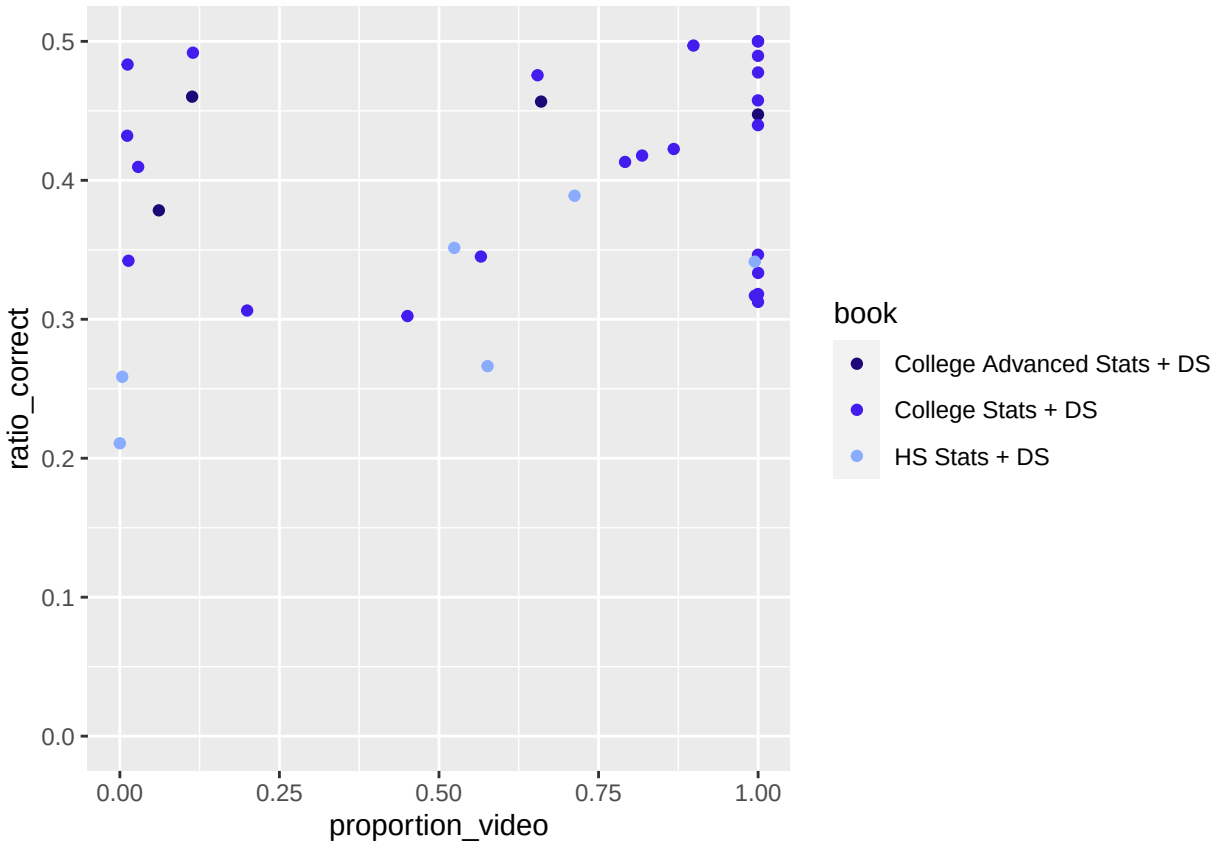

```
# If we divide ratio_correct into high scorers and low scorers

#High scorers
big |> filter(ratio_correct > 0.5) |> ggplot(aes(x = proportion_video, y = ratio_correct)) + geom_point
  scale_color_manual(values = c("#1D0878", "#421DEE", "#89ACFF"),
                     labels = c("College Advanced Stats + DS", "College Stats + DS",
                                "HS Stats + DS"))
```

```
## Warning: Removed 1054 rows containing missing values (`geom_point()`).
```



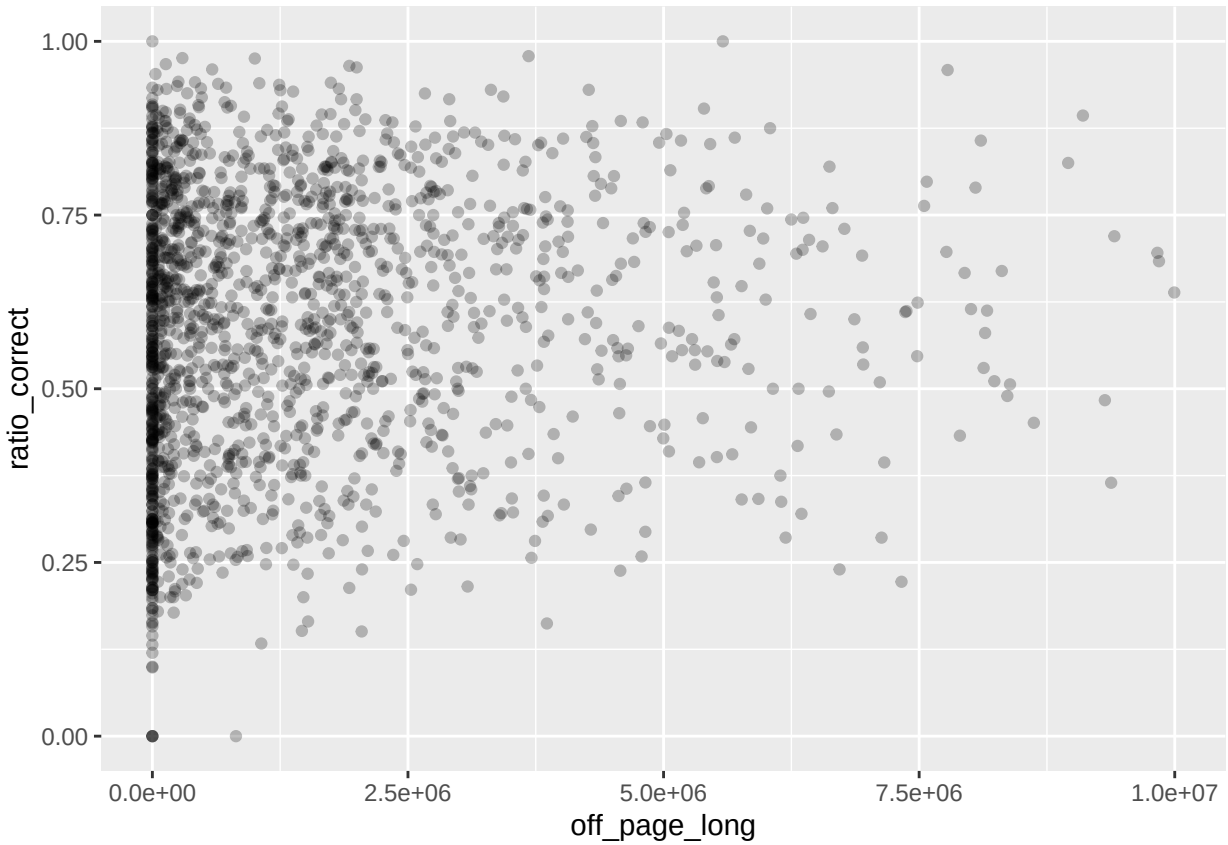

```
#Low scorers
big |> filter(ratio_correct <= 0.5) |> ggplot(aes(x = proportion_video, y = ratio_correct)) + geom_poin
  scale_color_manual(values = c("#1D0878", "#421DEE", "#89ACFF"),
                     labels = c("College Advanced Stats + DS", "College Stats + DS",
                                "HS Stats + DS"))
```

```
## Warning: Removed 519 rows containing missing values (`geom_point()`).
```

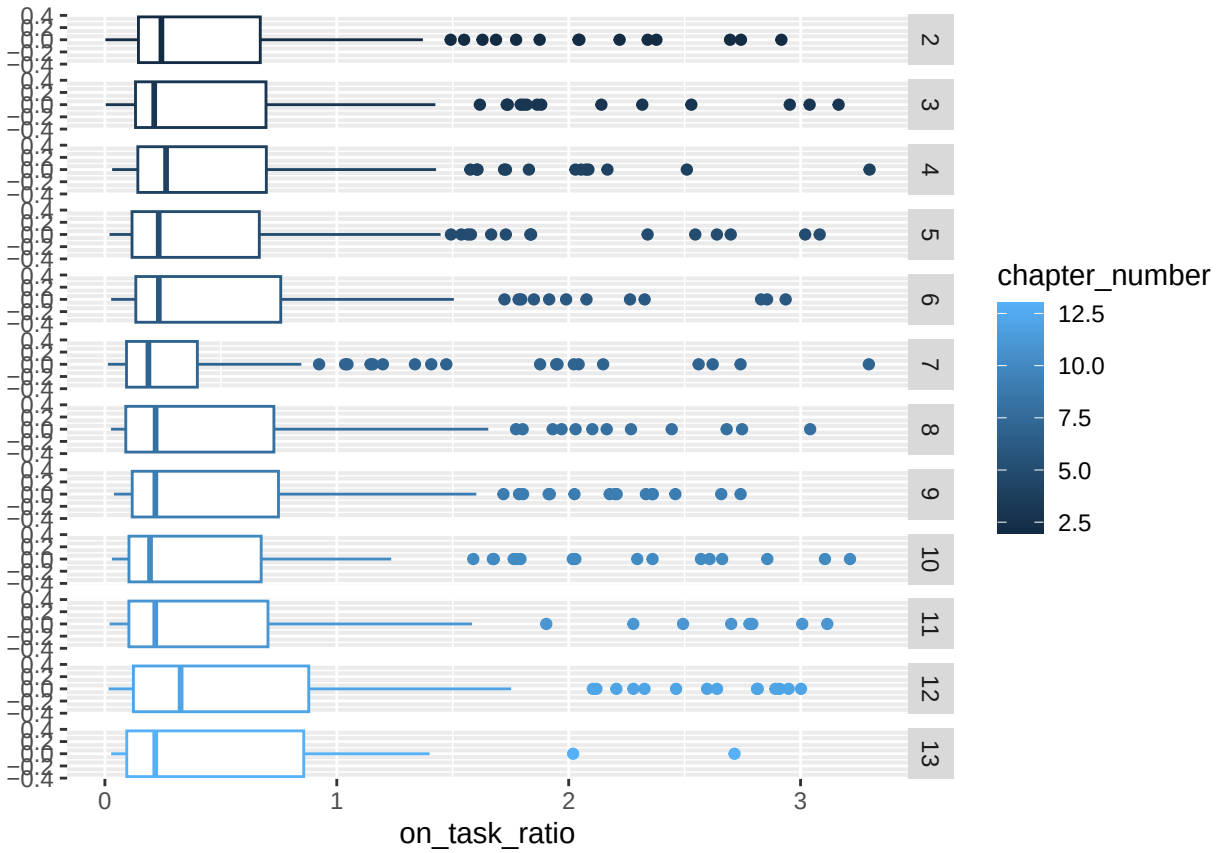



```
big[big$off_page_long < 1e+07,] |> ggplot(aes(x = off_page_long, y = ratio_correct)) +
  geom_point(alpha = 0.25) # off_page_long vs ratio_correct
```

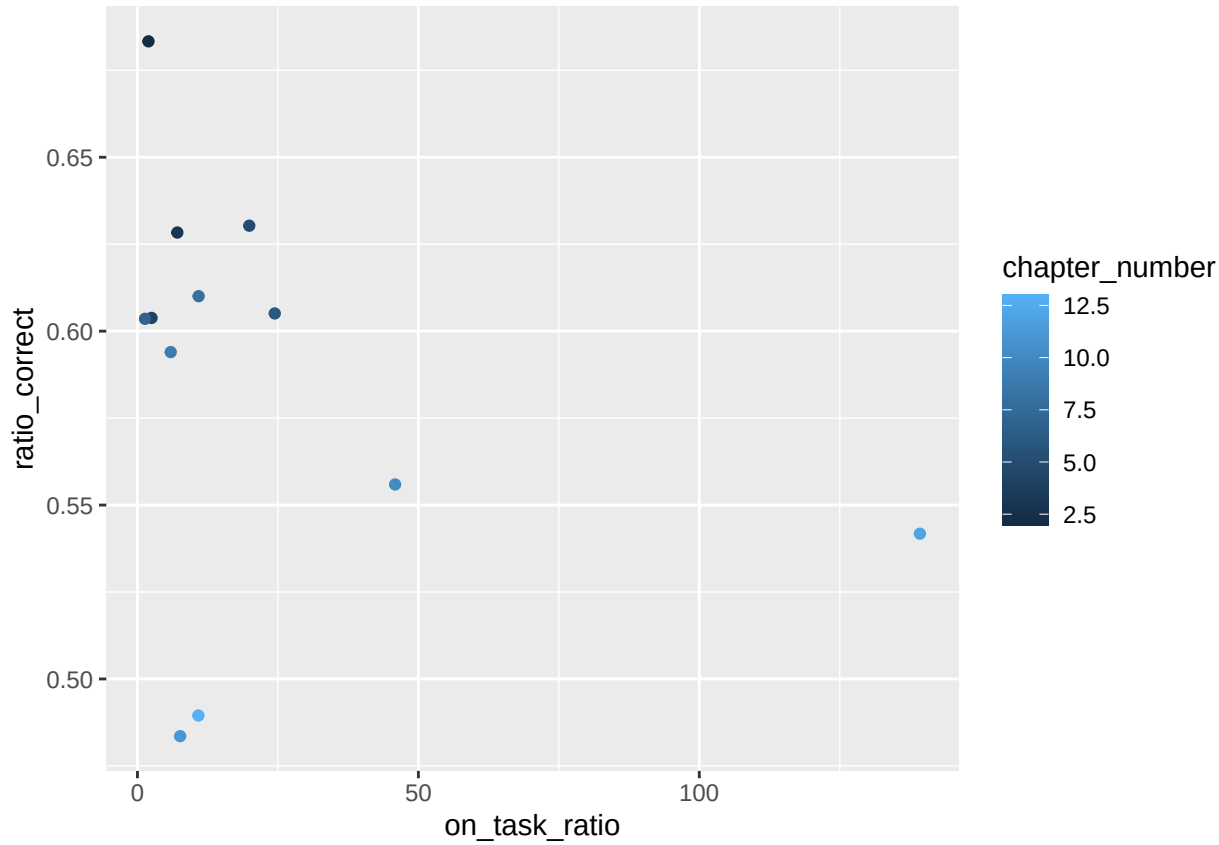
```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```r
# Engagement ratio??
big <- big |> mutate(on_task_ratio = engaged/(off_page_brief+off_page_long))
big |> mutate(on_task_ratio = engaged/(off_page_brief+off_page_long)) |>
  filter(on_task_ratio < 3.318226) |>
  ggplot(aes(x = on_task_ratio, color = chapter_number)) +
  geom_boxplot() +
  facet_grid(chapter_number~.) # comparing engagement across chapters
```



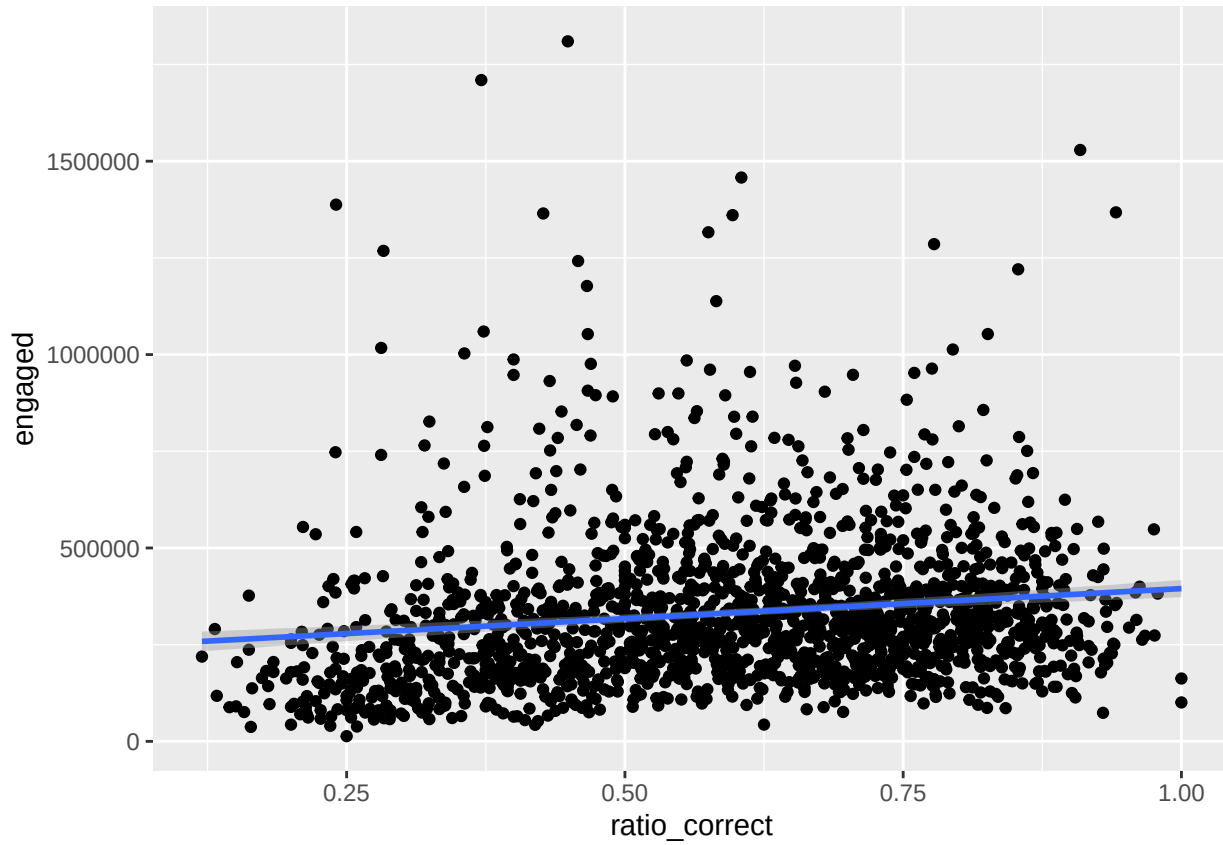

```
big[is.finite(big$on_task_ratio),] |>
  group_by(chapter_number) |>
  summarize(on_task_ratio = mean(on_task_ratio, na.rm = TRUE), ratio_correct =
              mean(ratio_correct, na.rm = TRUE)) |>
  ggplot(aes(x = on_task_ratio, y = ratio_correct)) +
  geom_point(aes(color = chapter_number)) # on_task_ratio vs ratio_correct by chapter #
```




```r
summary(big$engaged/(big$off_page_brief+big$off_page_long)) # outliers at 3.318226
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## 0.001851 0.133704 0.325917      Inf 1.407513      Inf        1
```

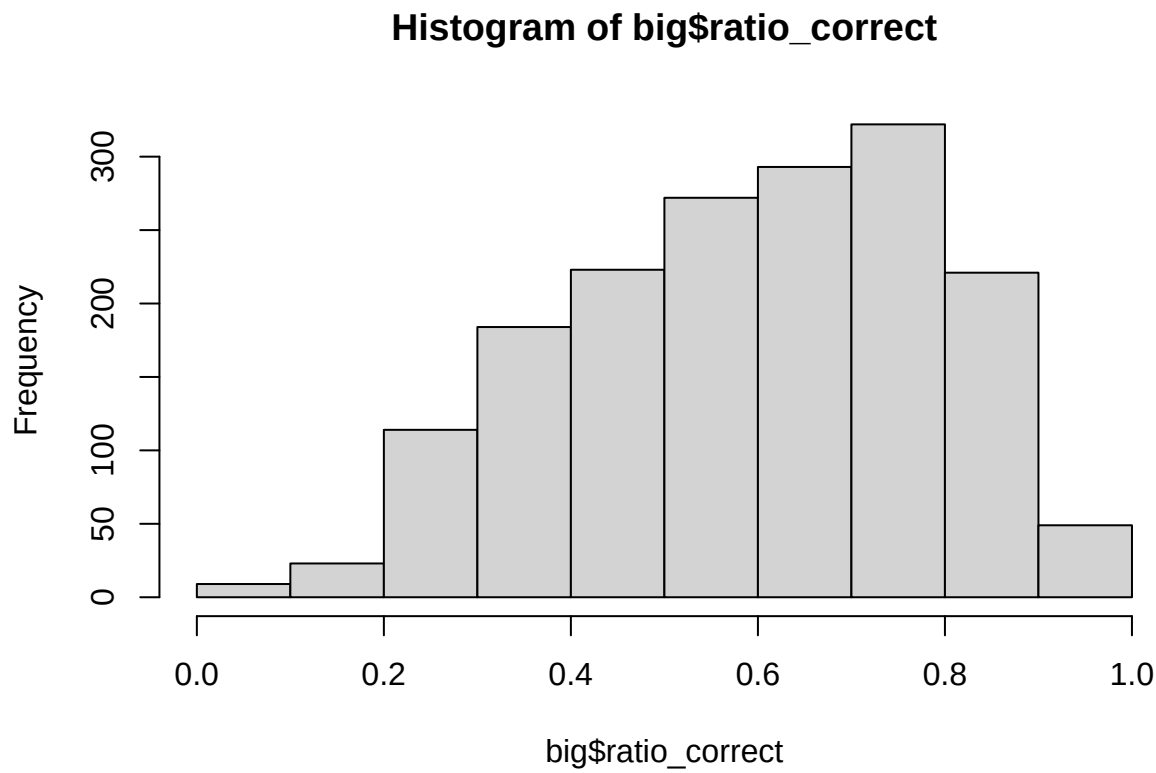
```r
big |> filter(ratio_correct > 0.1, engaged < 2e+06) |> ggplot(aes(y = engaged, x = ratio_correct)) +
  geom_point() +
  stat_smooth(method = "lm",  formula = y ~ x, geom = "smooth")
```




```r
summary(lm(ratio_correct~engaged, data = big))
```

```
##
## Call:
## lm(formula = ratio_correct ~ engaged, data = big)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59501 -0.14822  0.01959  0.15449  0.42047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.739e-01  7.582e-03  75.687  < 2e-16 ***
## engaged     5.635e-08  1.735e-08   3.249  0.00118 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1949 on 1707 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.006144,   Adjusted R-squared:  0.005562
## F-statistic: 10.55 on 1 and 1707 DF,  p-value: 0.001182
```

```r
hist(big$ratio_correct)
```



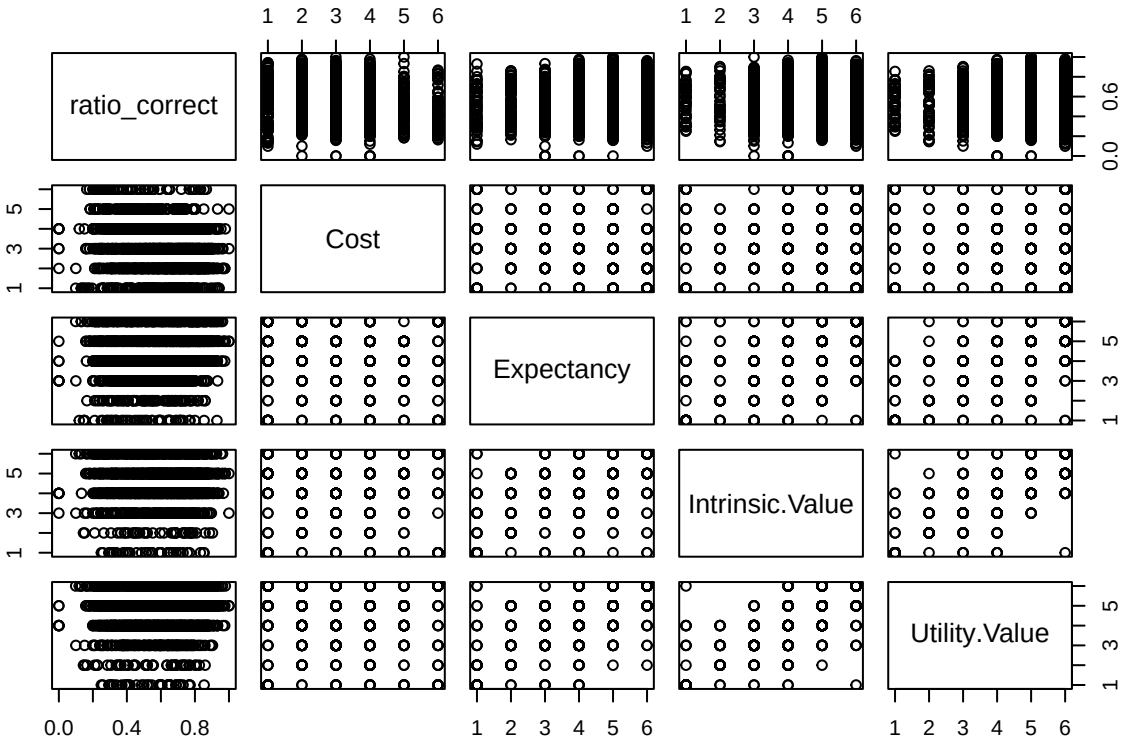

```
big$ratio_correct_log <- log(big$ratio_correct + 1)
```

Datetime stuff

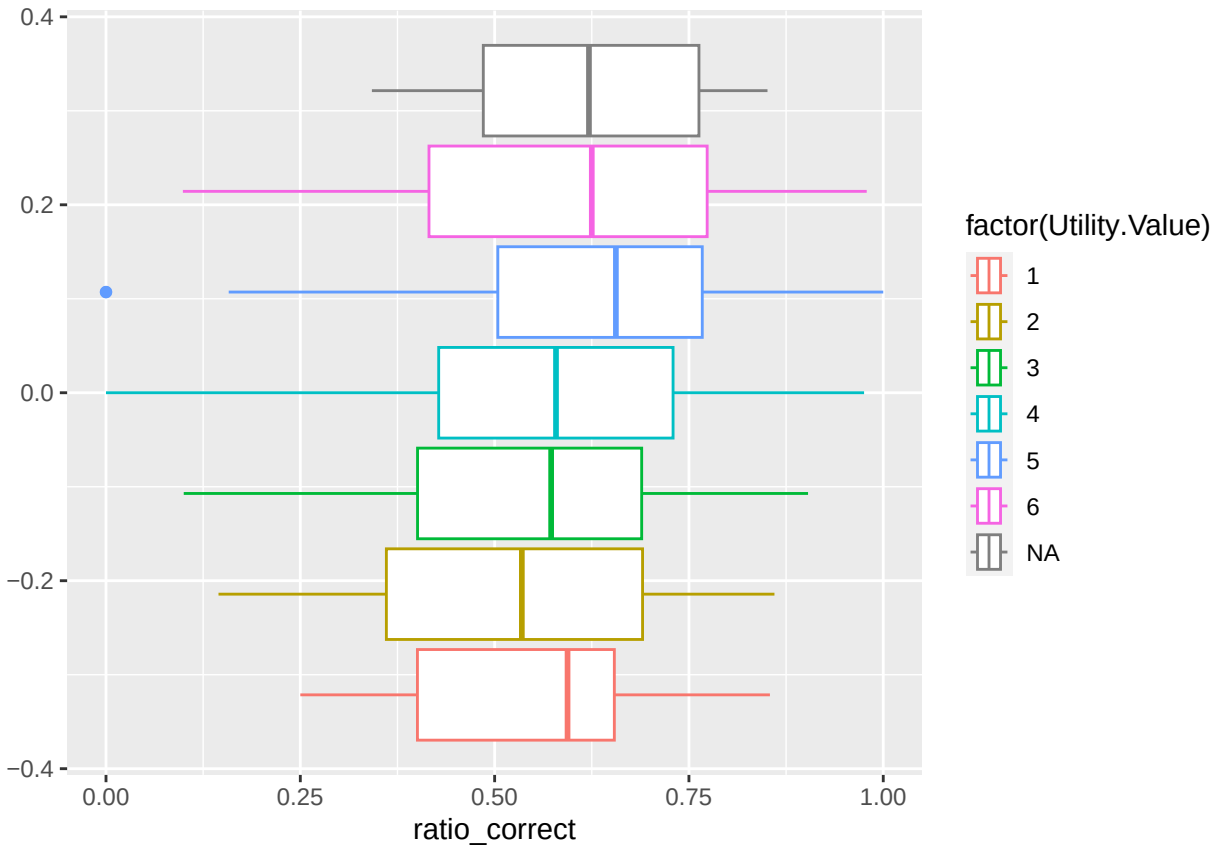
```
# Timezone is GMT/UTC but the majority of users are in LA so i would use PST to understand this graph..
```

Pulse

```
pairs(~ratio_correct+Cost+Expectancy+Intrinsic.Value+Utility.Value, big) # LOL
```




```
big |> ggplot(aes(x = ratio_correct)) +
  geom_boxplot(aes(color = factor(Utility.Value)))
```


0.4
0.2
0.0
−0.2
−0.4
0.00
0.25
0.50
0.75
1.00
ratio_correct
factor(Utility.Value)
1
2
3
4
5
6
NA