STA302H1 – Final Project Descriptive Statistics

Danny Chen

August 10, 2021

Import STA302H1 Study Time and COVID Contemplation Time vs. Quiz Performance Dataset

Data Cleaning

First, I'll clean my data.

```
cleaned_sta302_performance_data <- sta302_performance_data %>%
    # Create a new "country" column, which is just "Country" but whose entries are factors.
    mutate(country = as.factor(Country)) %>%

    # TODO: Replace quiz grades, covid hours, and sta302h1 study hours with their
    # TODO: median counterparts.

    # Remove the "X" column: it's simply the row number, which isn't very useful.
    # Remove the "Country" column: column "country" already exists
    dplyr::select(-X, -Country) %>%

    # Rearrange similar columns side-by-side.
    relocate(country,
             COVID.hours..W1., COVID.hours..W2., COVID.hours..W3., COVID.hours..W4.,
             STA302.hours..W1., STA302.hours..W2., STA302.hours..W3., STA302.hours..W4.,
             Quiz_1_score, Quiz_2_score, Quiz_3_score, Quiz_4_score)
```

Helper Functions

```
num_column_NAs = function(predictor_variable) {
  sum(is.na(predictor_variable))
}
```

```
row_nums_of_NA_columns = function(data, predictor_variable) {
  which(is.na(predictor_variable))
}
```

```
rows_with_num_NAs = function(data, num_NAs) {
  return (rowSums(is.na(data)) == num_NAs)
}
```

```
row_nums_of_NA_rows = function(data, num_NAs) {
  return (which(rows_with_num_NAs(data, num_NAs)))
}
```

```
display_histogram <- function(data, predictor_variable, histogram_title, x_axis_label) {
  ggplot(data = tibble(data), mapping = aes(x = predictor_variable)) +
    geom_histogram(col = "black", fill = "red", bins = 30) +
    labs(title = histogram_title, y = "Frequency", x = x_axis_label) +
    geom_vline(mapping = aes(xintercept = mean(predictor_variable, na.rm = TRUE)),
               color = "blue", linetype = "solid") +
    geom_vline(mapping = aes(xintercept = median(predictor_variable, na.rm = TRUE)),
               color = "dark green", linetype = "dotted")
}
```

```
display_boxplot <- function(data, predictor_variable, boxplot_title, y_axis_label) {
  ggplot(mapping = aes(x = Country, y = predictor_variable, color = Country)) +
    geom_boxplot(mapping = aes(x = Country, y = predictor_variable)) +
    labs(title = boxplot_title, x = "Country", y = y_axis_label)
}
```

```
get_row_nums_to_exclude <- function(data) {
  row_nums_with_3_NAs = which(rows_with_num_NAs(data, 3))
  row_nums_with_4_NAs = which(rows_with_num_NAs(data, 4))
  row_nums_to_exclude <- union(row_nums_with_3_NAs,
                               row_nums_with_4_NAs)
  return (row_nums_to_exclude)
}
```

```
display_correlation_by_country <- function(country_data) {
  colnames(country_data) <- c("W1COV", "W2COV", "W3COV", "W4COV",
                              "W1302", "W2302", "W3302", "W4302",
                              "Q1", "Q2", "Q3", "Q4")
  round(cor(country_data, use = "pairwise.complete.obs", method = "pearson"), 2)
}
```

Special Tables

Rows With At Least One NA

Rows with at least one NA deserve closer examination.
Some of the rows might only have 1 - 2 NAs and are therefore salvageable, which is OK.
Other rows may contain 3 or more NAs, and might indicate students who have dropped STA302H1. We'd like to exclude them from our analysis.

Here are the number of rows with 0 - 4 NAs.

```
##   nrows_0_NAs nrows_1_NAs nrows_2_NAs nrows_3_NAs nrows_4_NAs
## 1         143           9          16          19           1
```

Columns with NAs

```
##   week1_covid week2_covid week3_covid week4_covid
## 1          26          22          21          40
```

```
##   week1_sta302 week2_sta302 week3_sta302 week4_sta302
## 1           26           22           20           40
```

```
##   quiz1_score quiz2_score quiz3_score quiz4_score
## 1          13          36          31          34
```

Number of Missed Quizzes

```
##   miss_0_quizzes miss_1_quizzes miss_2_quizzes miss_3_quizzes miss_4_quizzes
## 1            176             20              3             24              4
```

Who to Exclude from the Dataset?

Identify rows with at least 3 missing quiz marks. These indicate students who have dropped STA302H1, and who should be excluded from the final data.

Notice that we didn't check the number of NAs for country of origin, COVID hours, and STA302H1 hours, since some students either forgot or abstained. So there's no reason to exclude these students from our final dataset.

```
row_nums_to_exclude <- get_row_nums_to_exclude(quiz_grades)
cleaned_sta302_performance_data2 =
  cleaned_sta302_performance_data[-row_nums_to_exclude,]
```

Rows with Mistyped Columns

Rows whose columns are mis-typed may need to be corrected via imputation.

```
rows_with_mistyped_columms = cleaned_sta302_performance_data2[c(38, 83, 84, 117),]
# row 83: Country -> "canada" -- DONE
# row 84: Country -> "canada" -- DONE

# row 117: COVID.hours..W4. -> 0.5 hours -- DONE

# row 38:  STA302.hours..W3. -> 5.5<U+00A0> -- DONE
# row 117: STA302.hours..W4. -> 7.5 hours -- DONE
```

```
# library(janitor)
# use it to clean up data.
```

Rows Without Country Entry

Taking out the country column can come in handy for functions like cor() where factors aren't allowed.

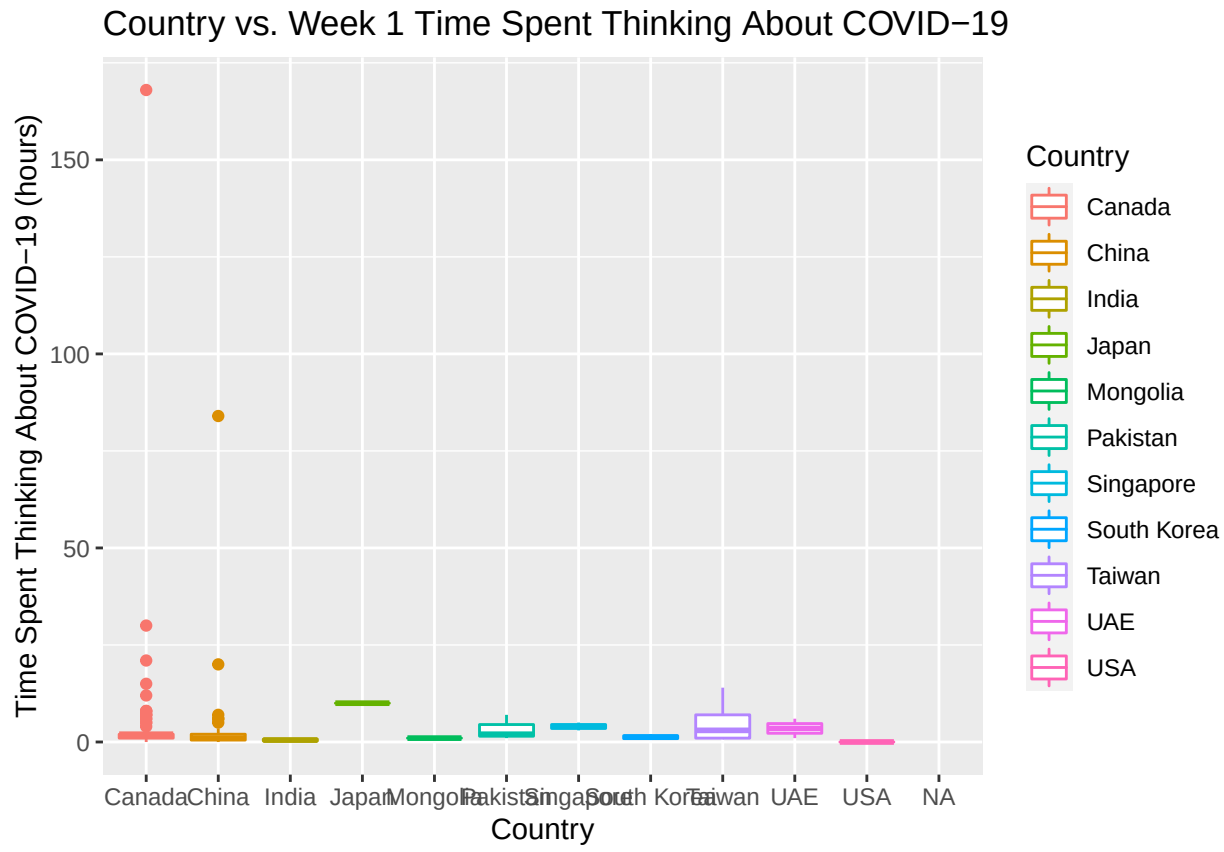
```
rows_with_no_country = cleaned_sta302_performance_data2 %>%
  dplyr::select(-country)
```

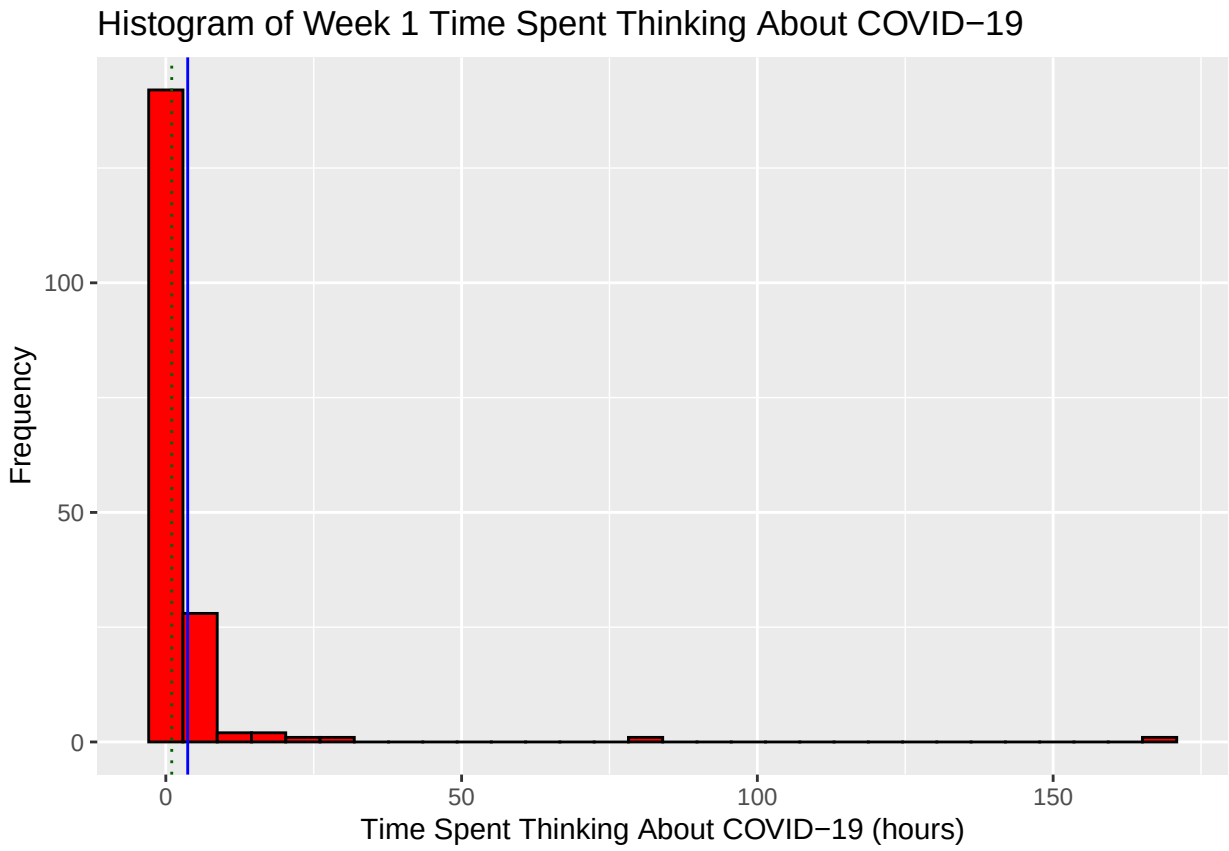
Rows Filtered by Country

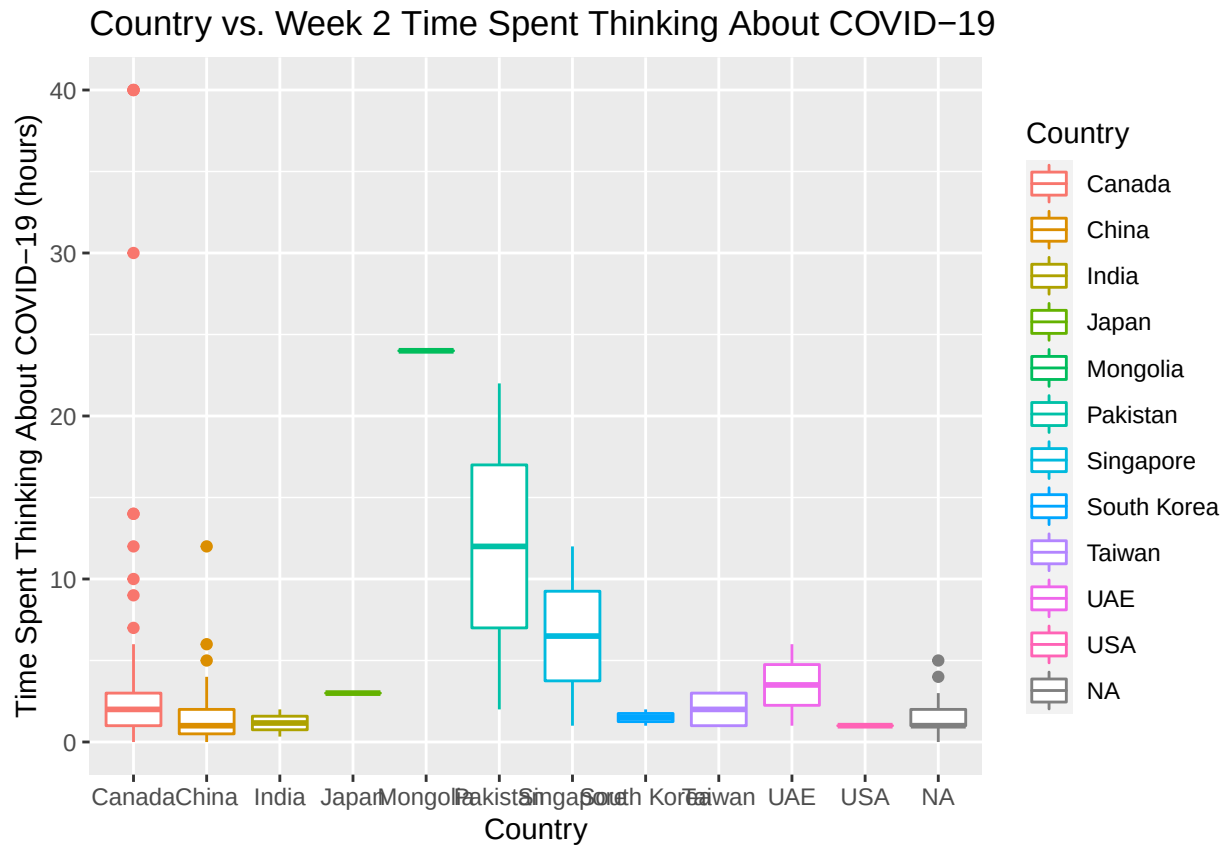
This is useful if we want data for individual countries.
Only the first and last code snippets are shown.

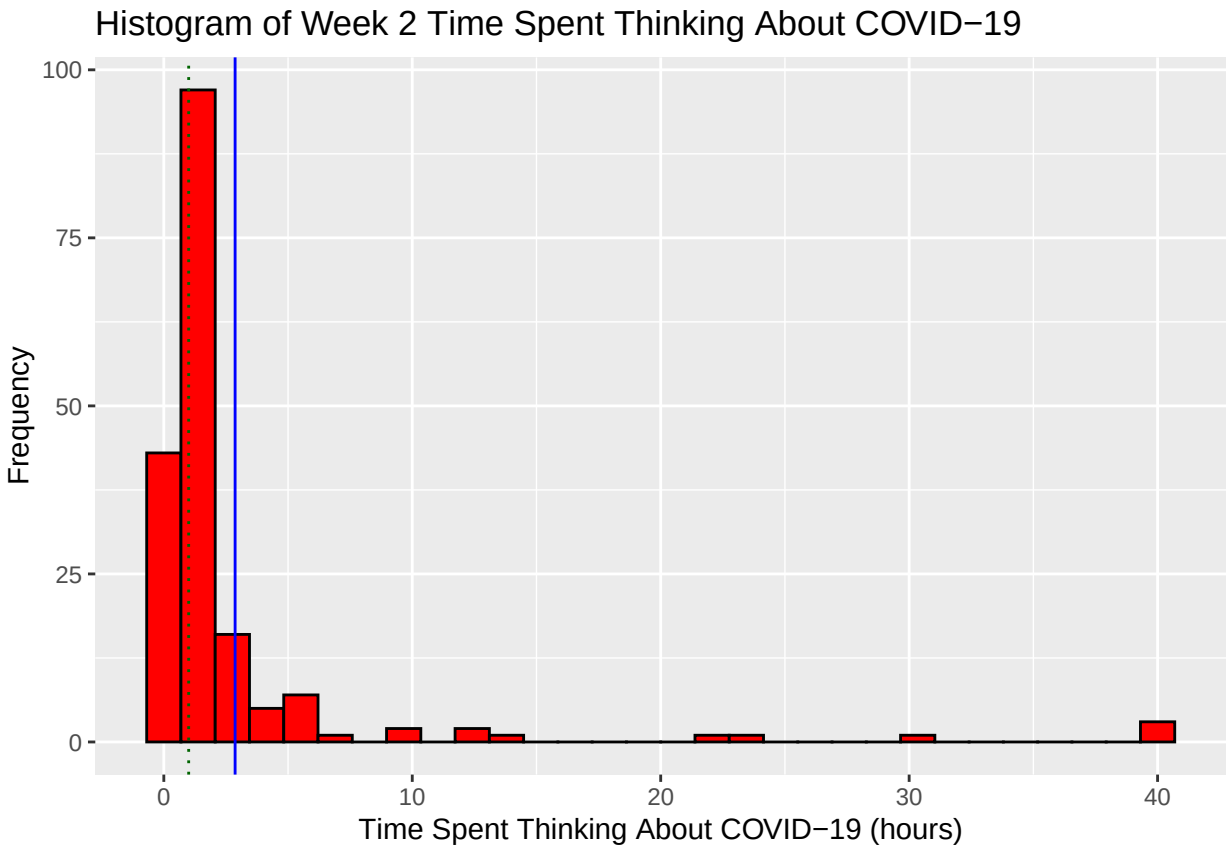
```
canada <- cleaned_sta302_performance_data2 %>%
  filter(as.character(country) == "Canada") %>%
  dplyr::select(-country)

unknown <- cleaned_sta302_performance_data2 %>%
  filter(is.na(as.character(country))) %>%
  dplyr::select(-country)
```

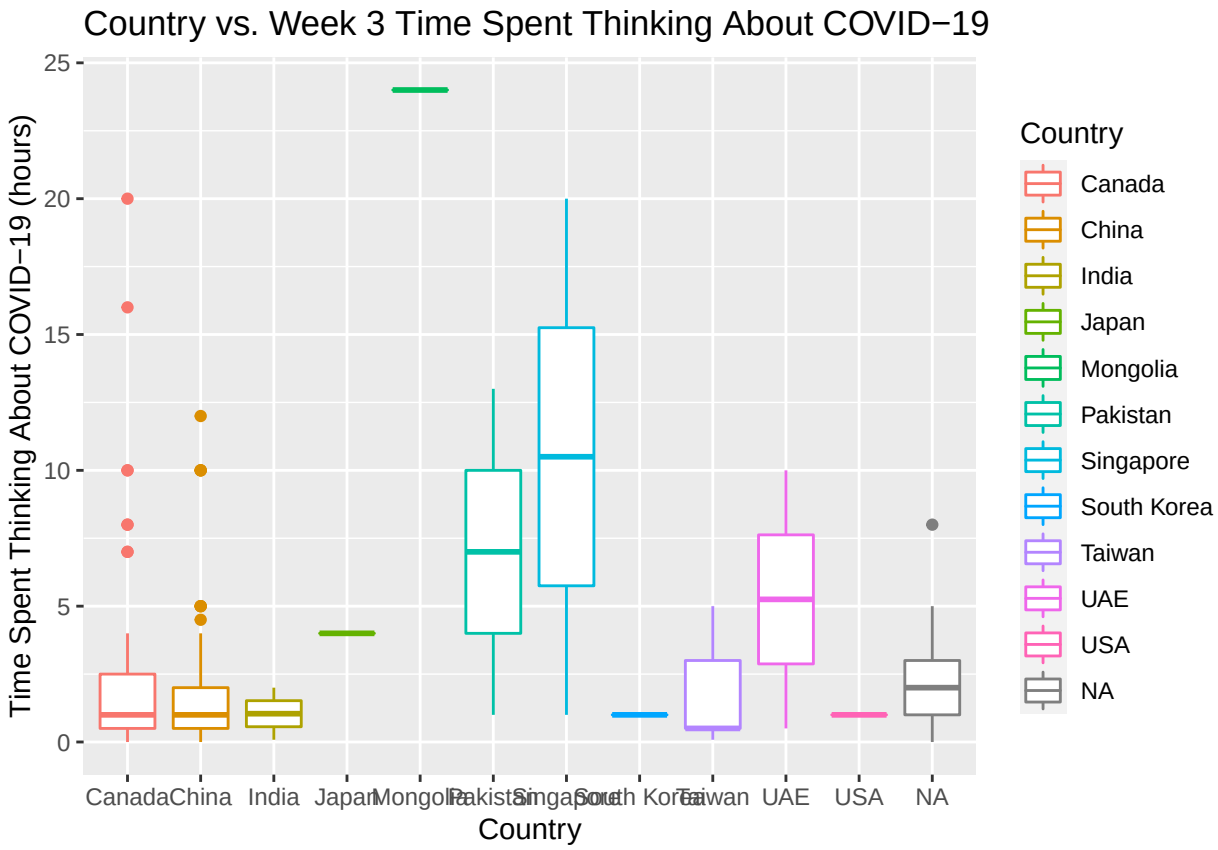
```
##             Country
## Canada           97
## China            63
## India             2
## Japan             1
## Mongolia          1
## Pakistan          3
## Singapore         2
## South_Korea       2
## Taiwan            3
## UAE               2
## USA               2
## Unknown          21
```

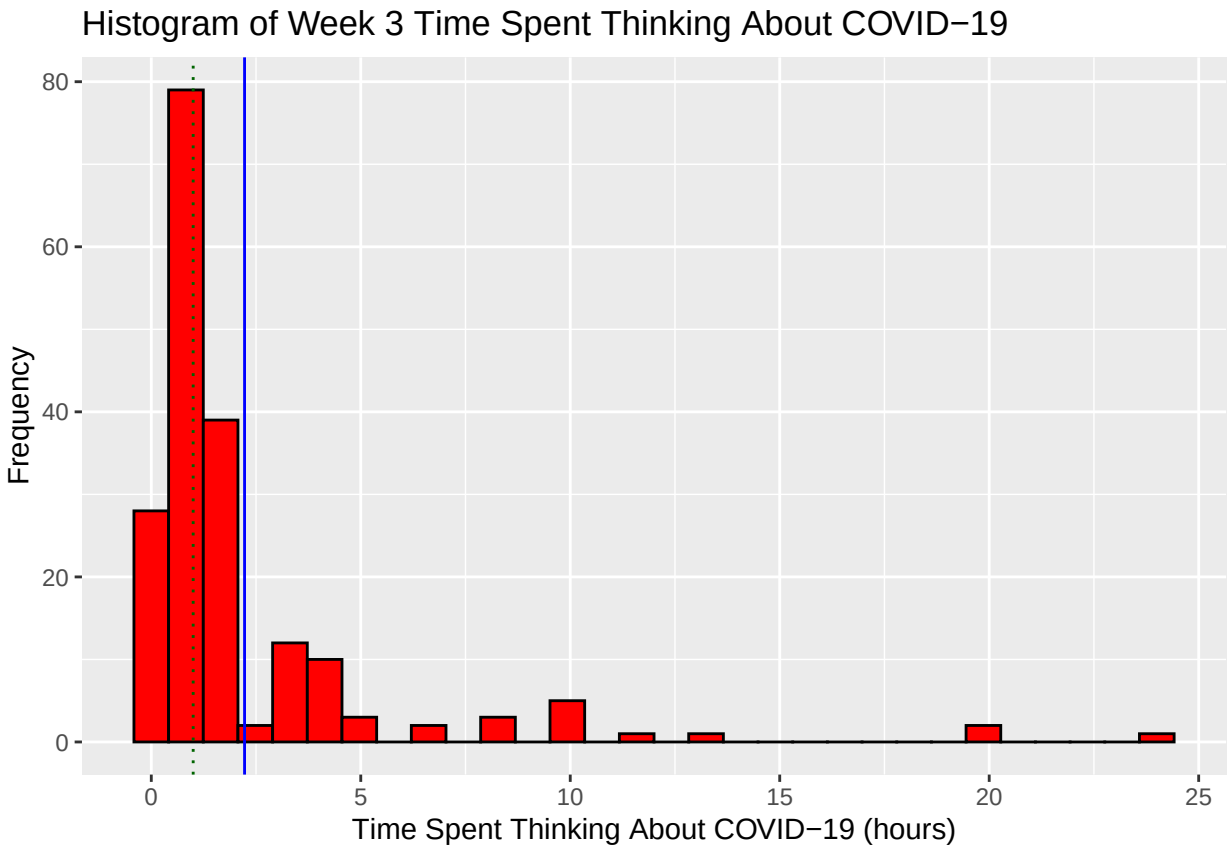

Country vs. Week 1 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
150
100
50
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA

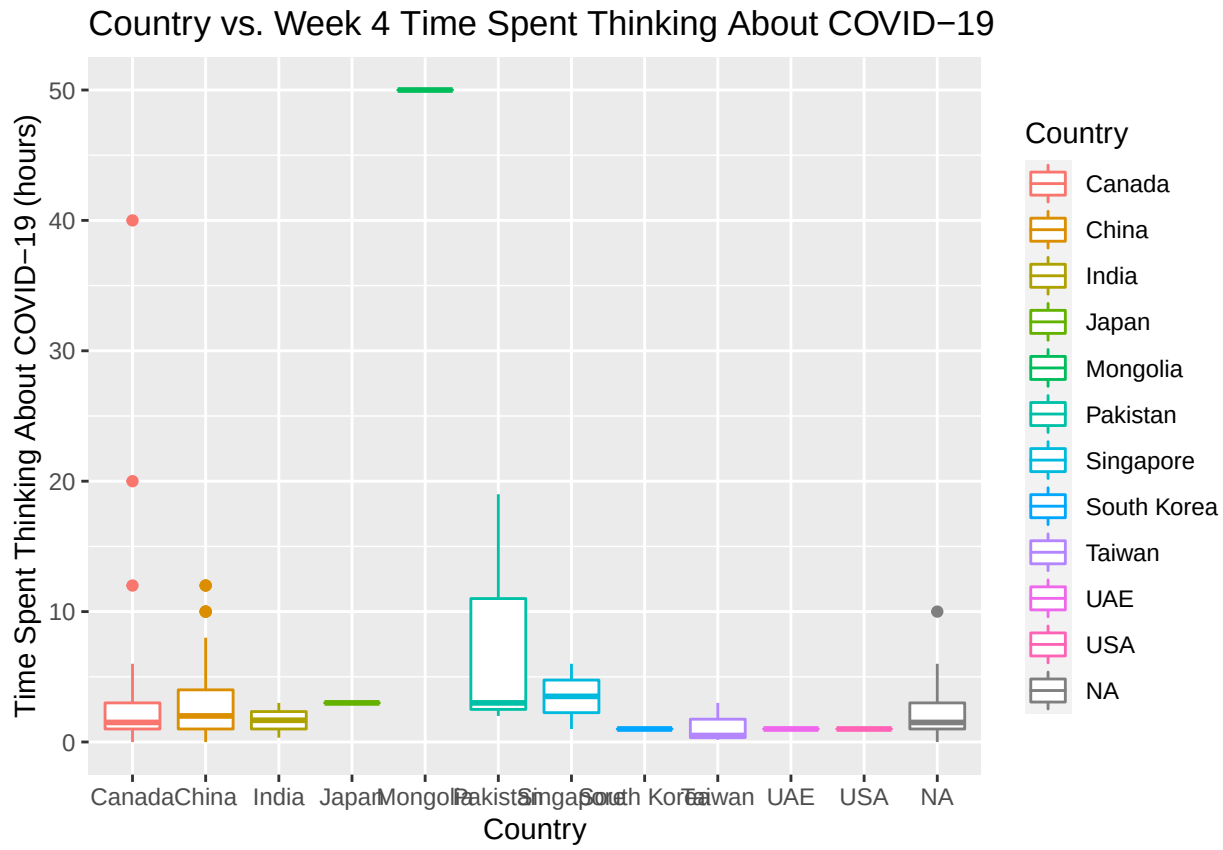

Histogram of Week 1 Time Spent Thinking About COVID−19
Frequency
100
50
0
0
50
100
150
Time Spent Thinking About COVID−19 (hours)

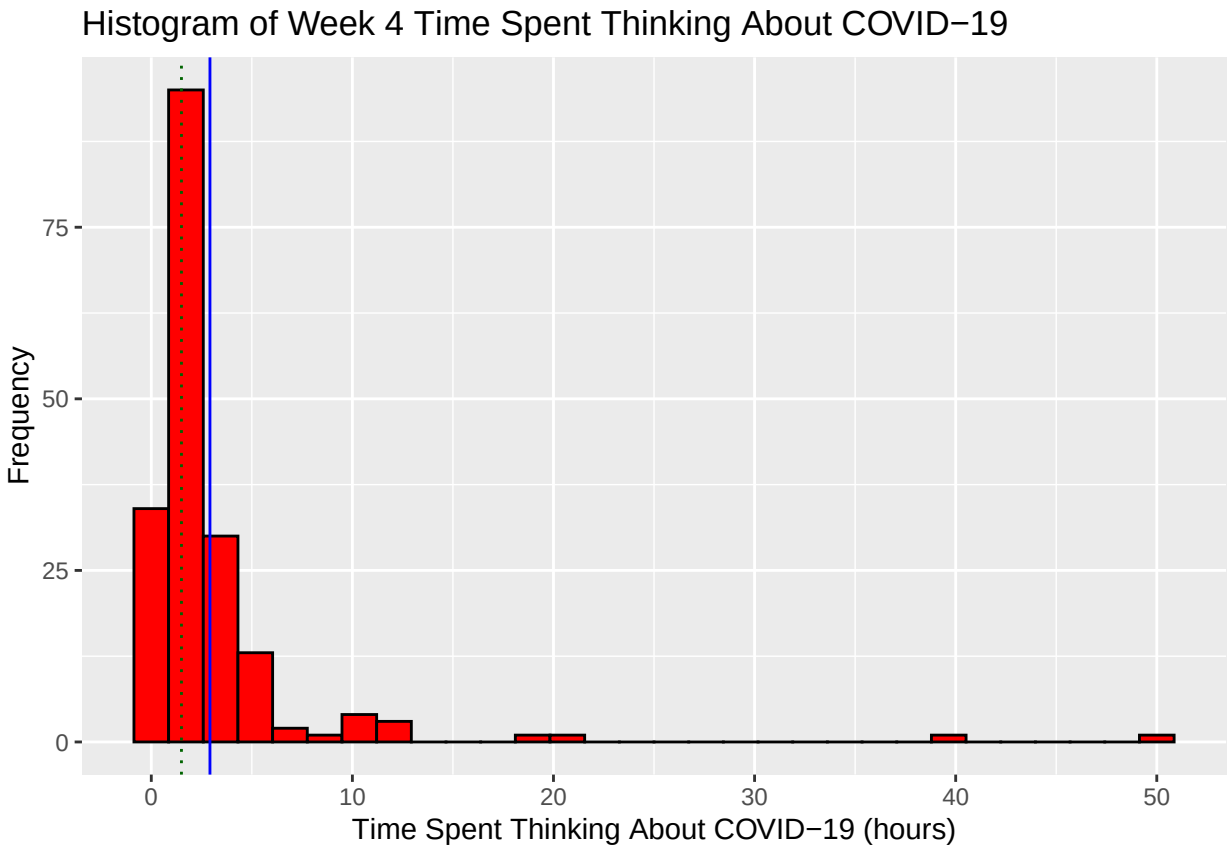

Country vs. Week 2 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
40
30
20
10
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

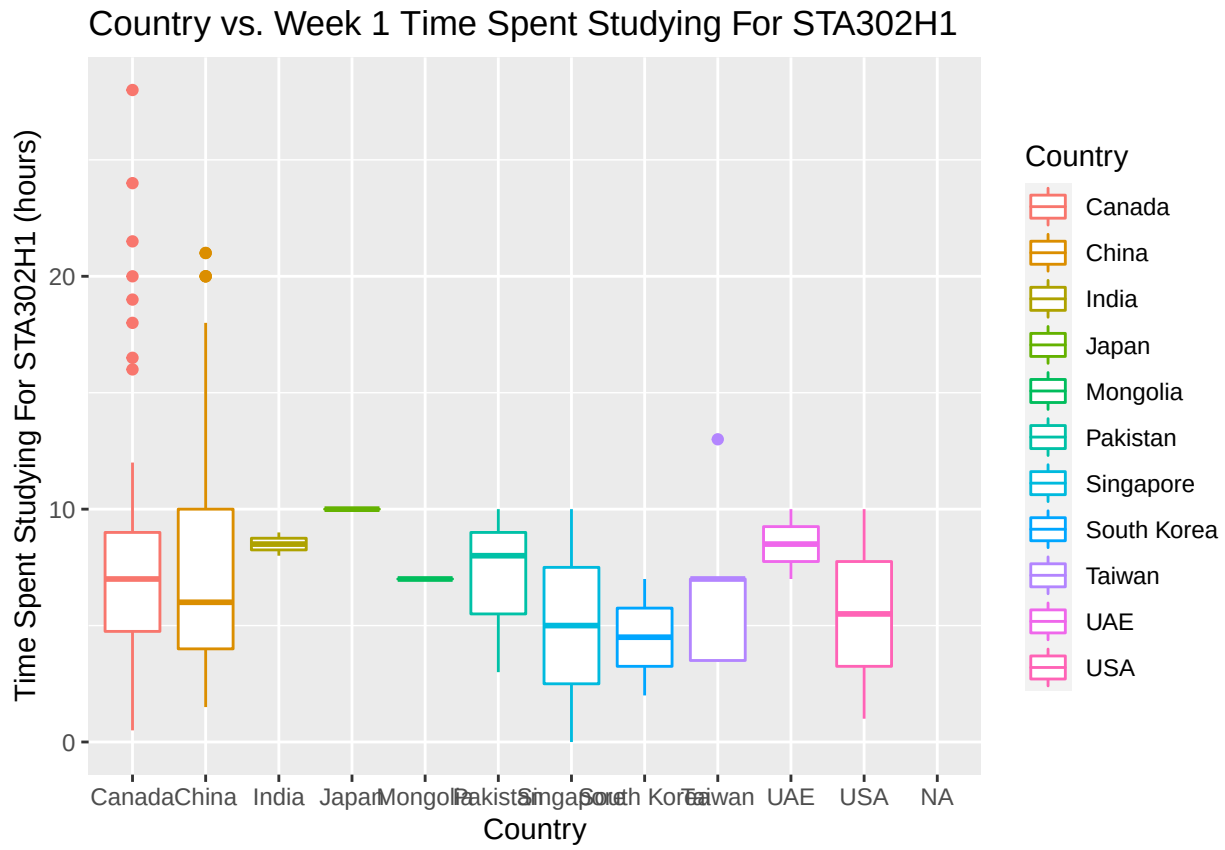

Histogram of Week 2 Time Spent Thinking About COVID−19
Frequency
100
75
50
25
0
0
10
20
30
40
Time Spent Thinking About COVID−19 (hours)

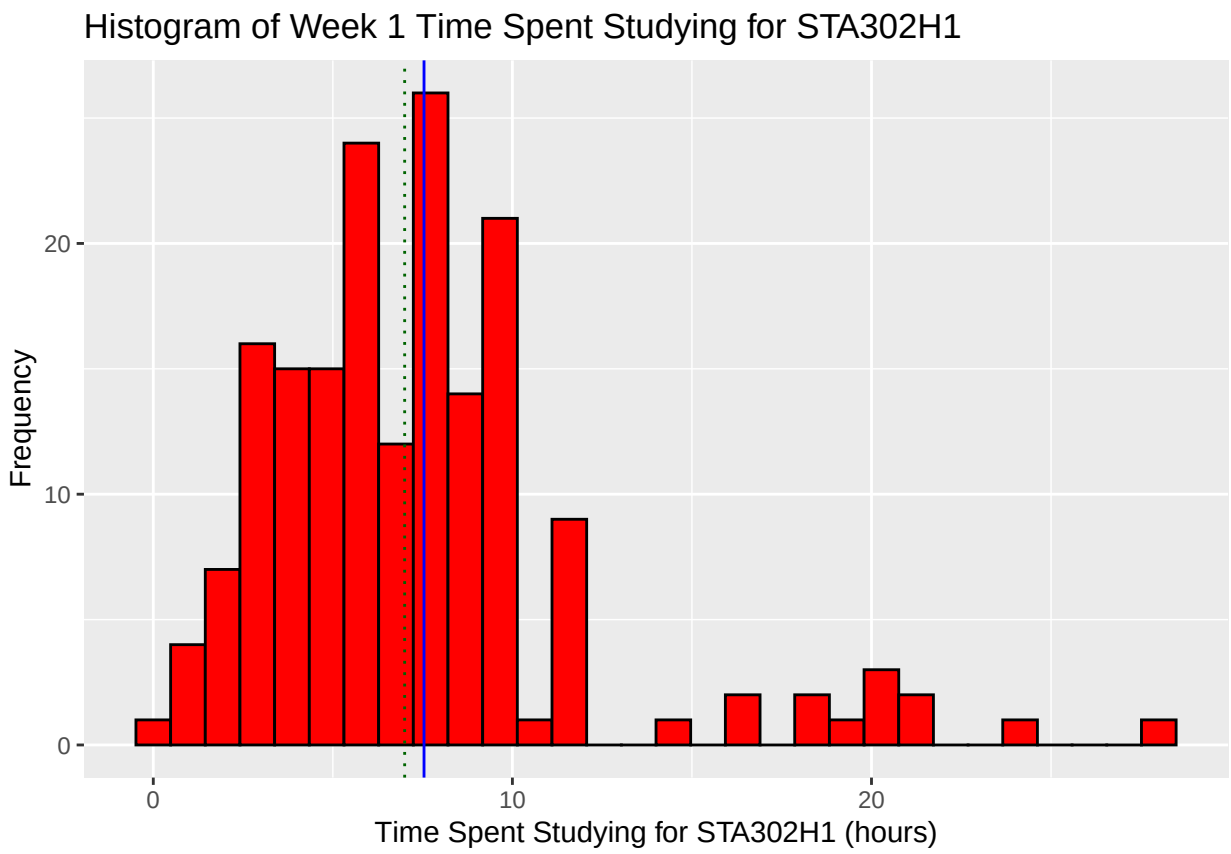

Country vs. Week 3 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
25
20
15
10
5
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

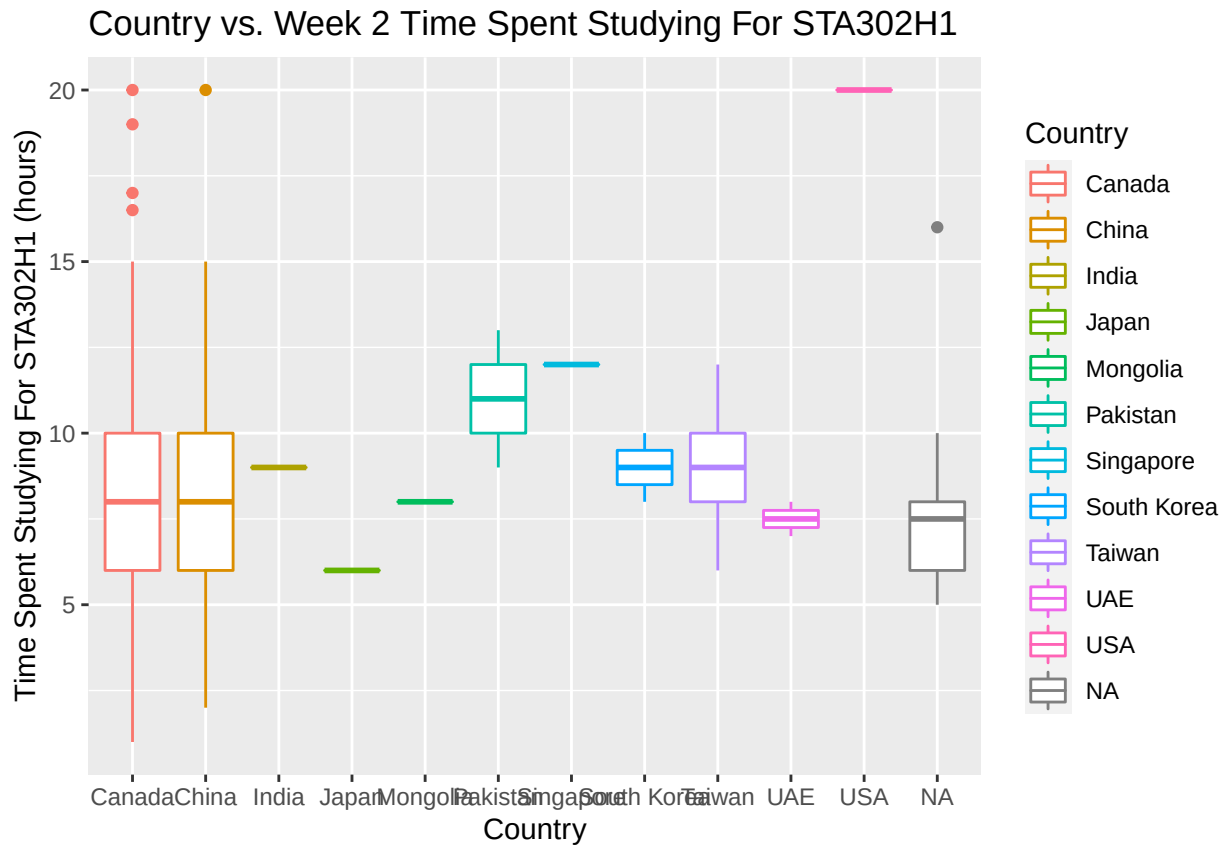

Histogram of Week 3 Time Spent Thinking About COVID−19
Frequency
80
60
40
20
0
0
5
10
15
20
25
Time Spent Thinking About COVID−19 (hours)

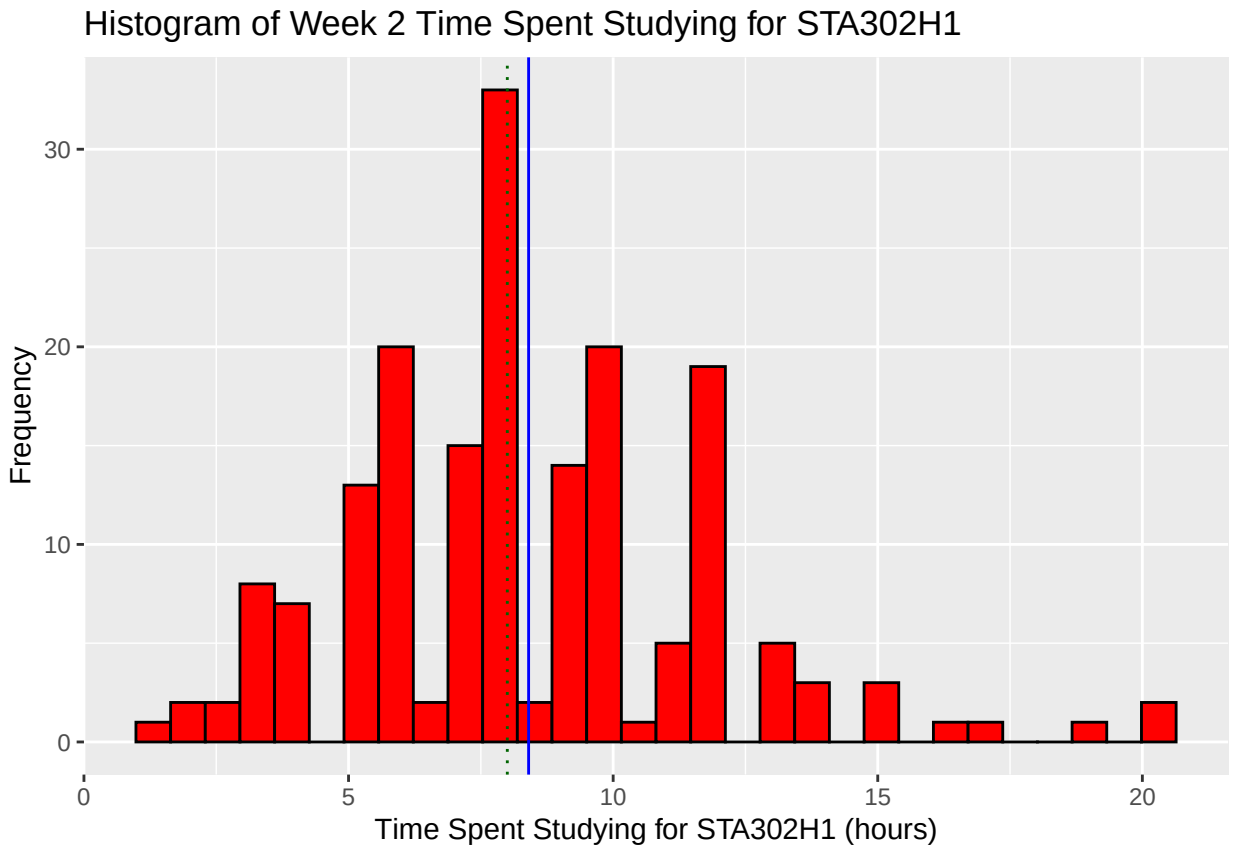

Country vs. Week 4 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
50
40
30
20
10
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

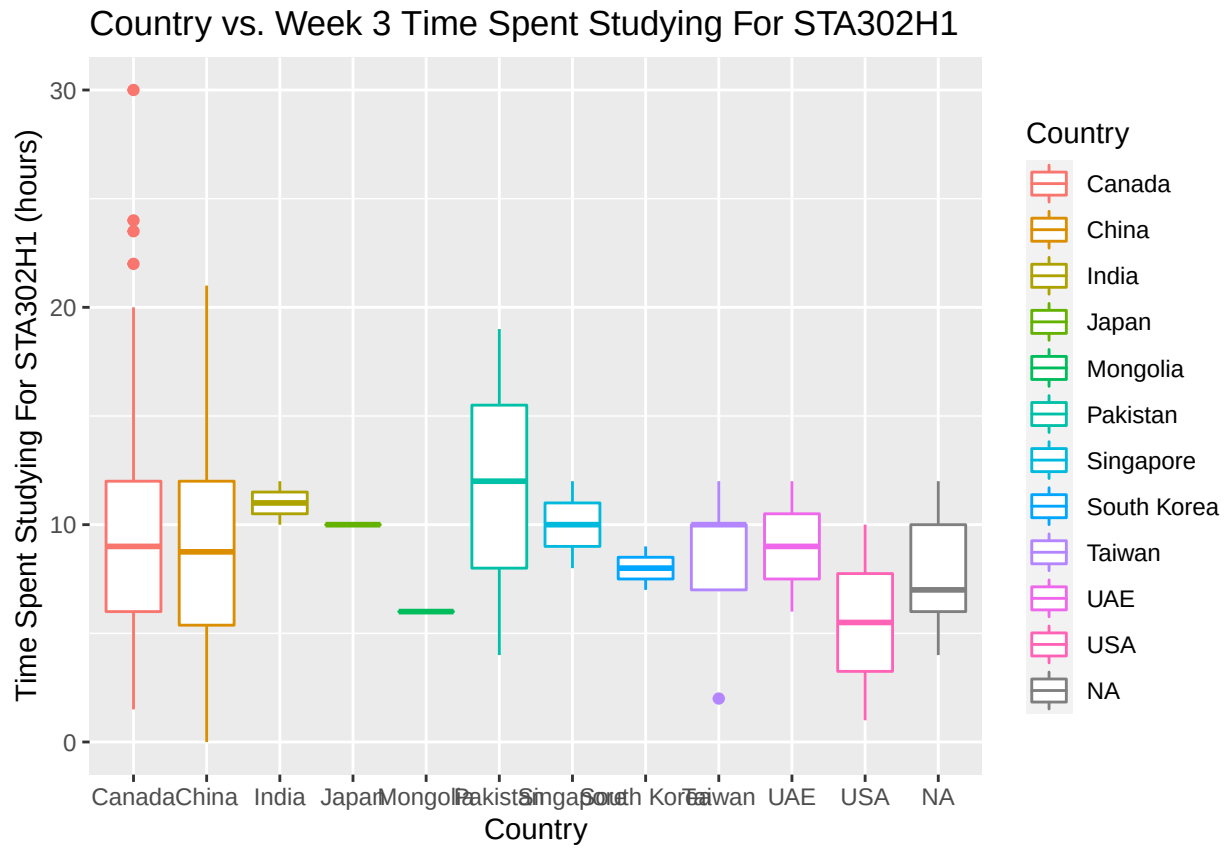

Histogram of Week 4 Time Spent Thinking About COVID−19
Frequency
75
50
25
0
0
10
20
30
40
50
Time Spent Thinking About COVID−19 (hours)

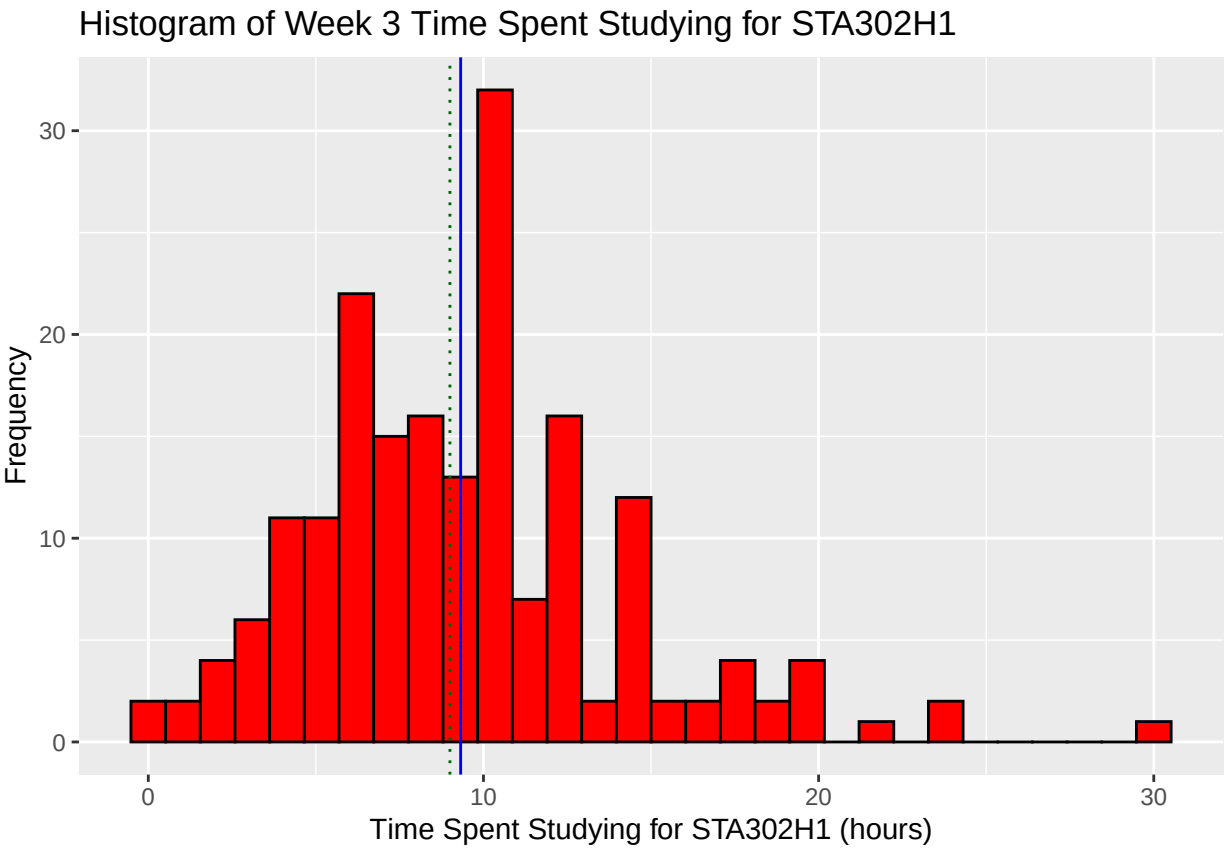

Country vs. Week 1 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
0
10
20
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA

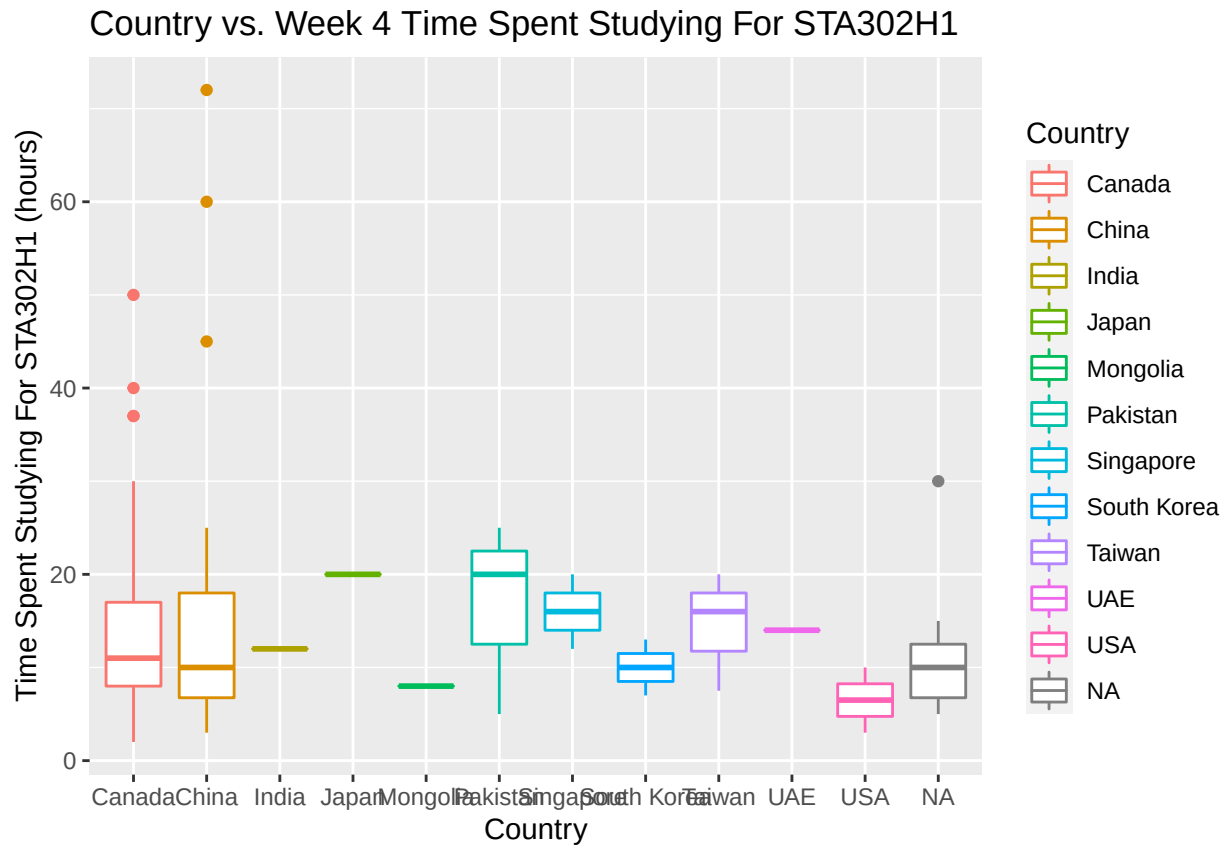

Histogram of Week 1 Time Spent Studying for STA302H1
Frequency
0
10
20
0
10
20
Time Spent Studying for STA302H1 (hours)

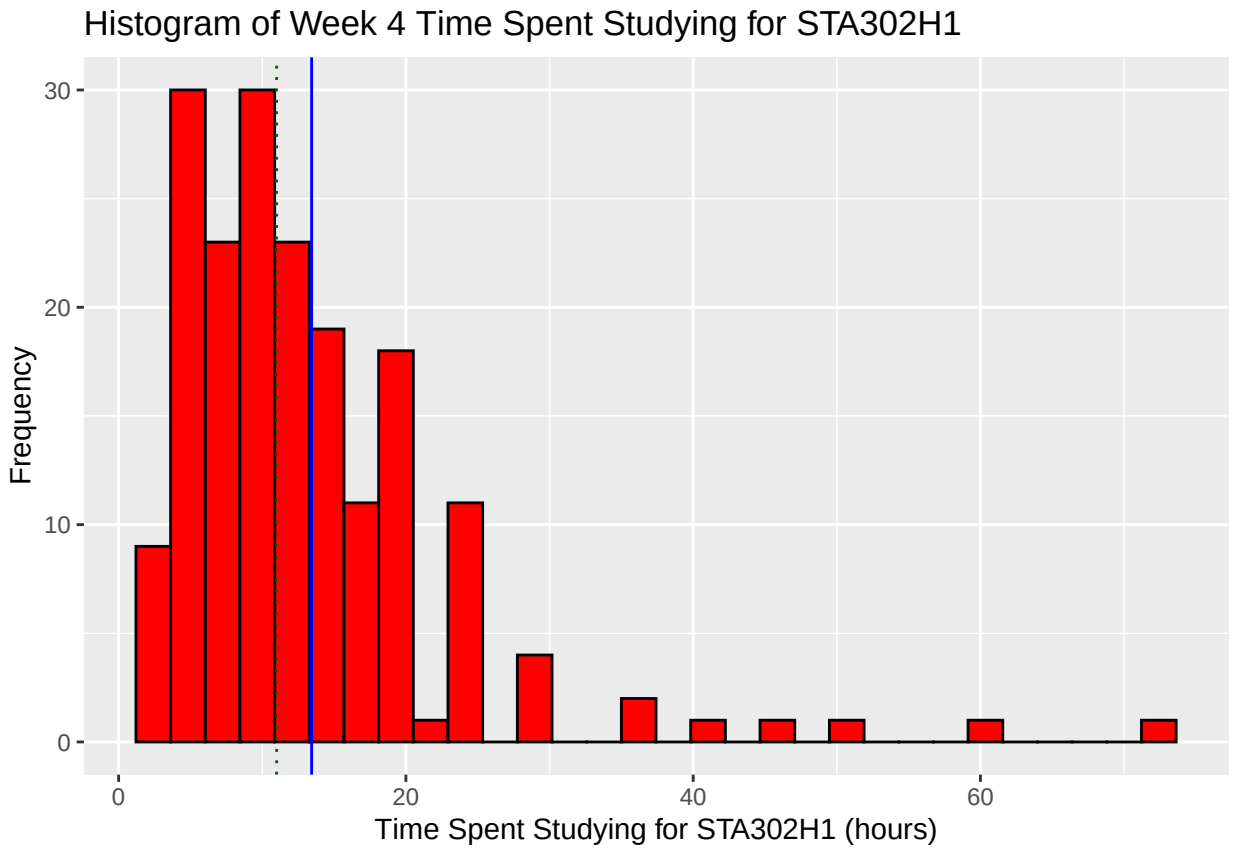

Country vs. Week 2 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
20
15
10
5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

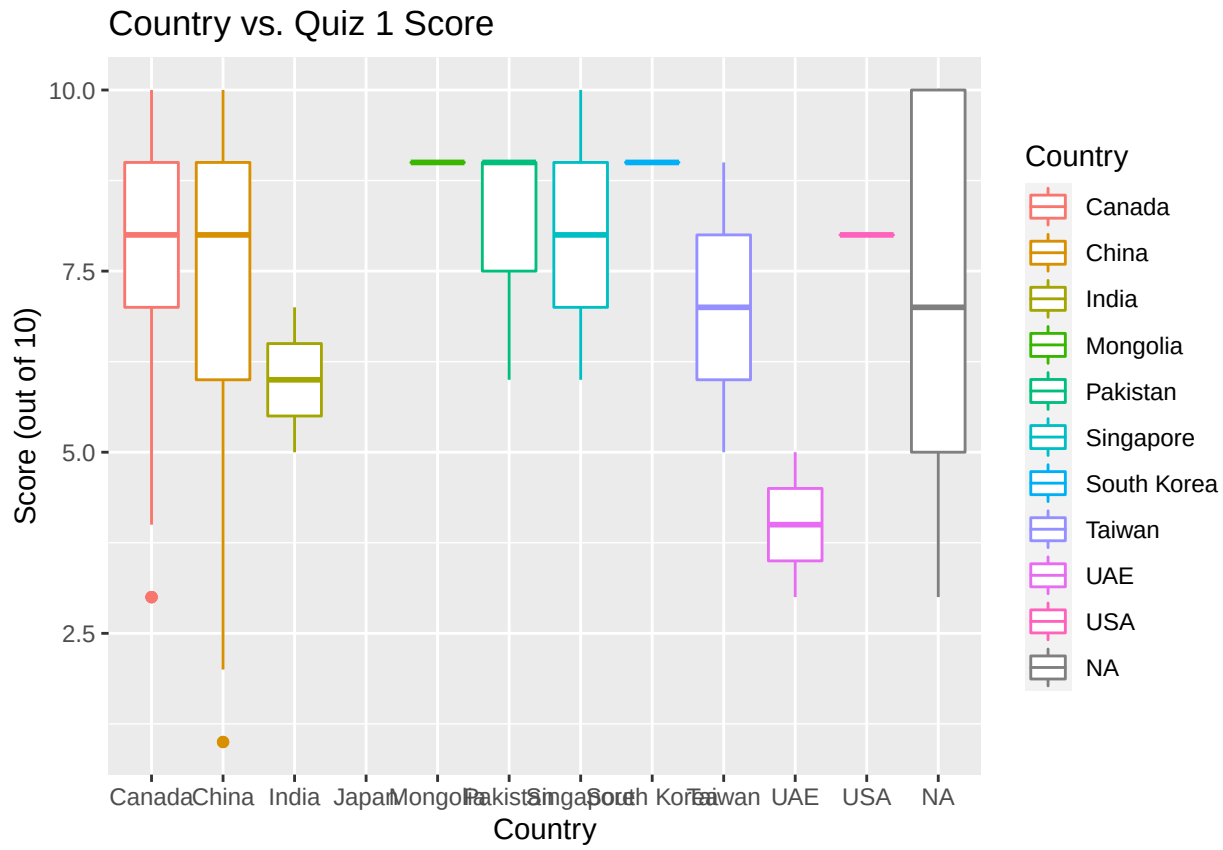

Histogram of Week 2 Time Spent Studying for STA302H1
Frequency
30
20
10
0
0
5
10
15
20
Time Spent Studying for STA302H1 (hours)

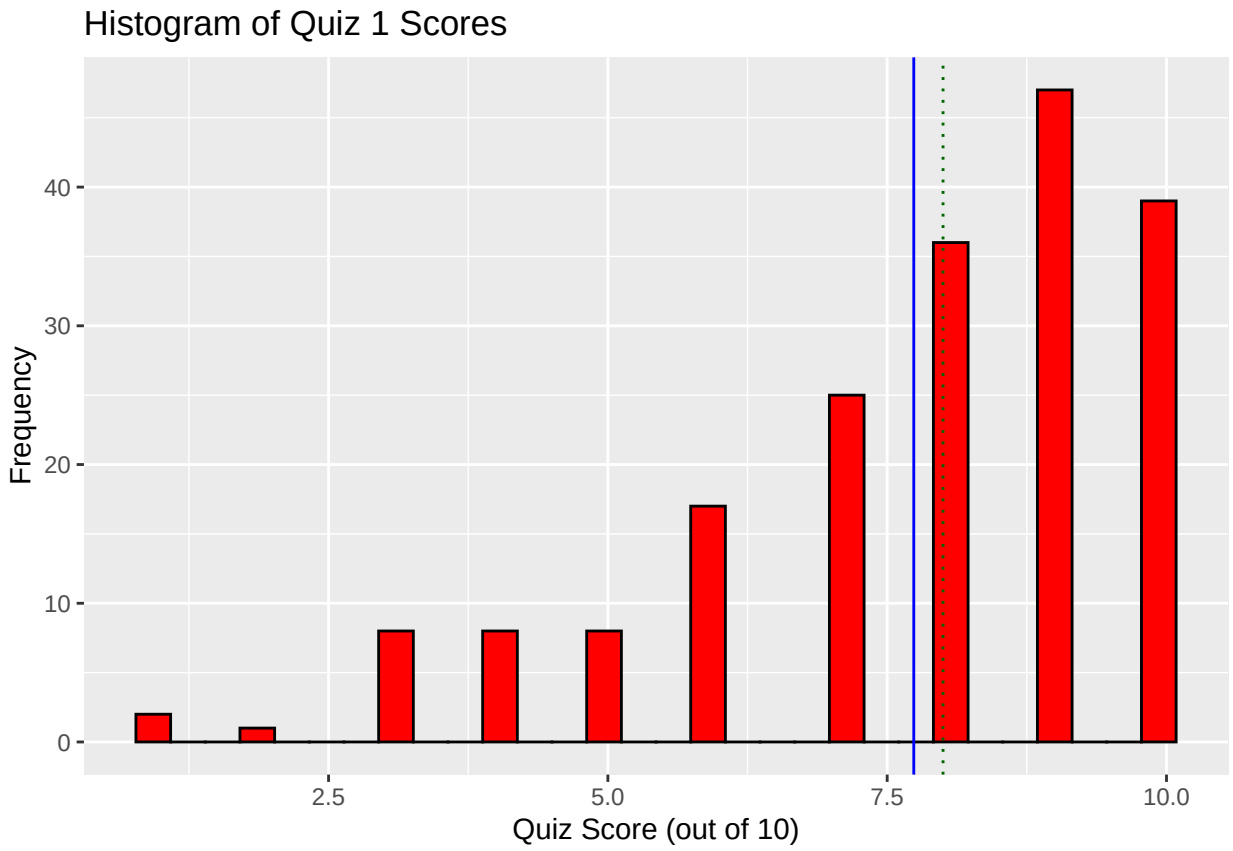

Country vs. Week 3 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
30
20
10
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

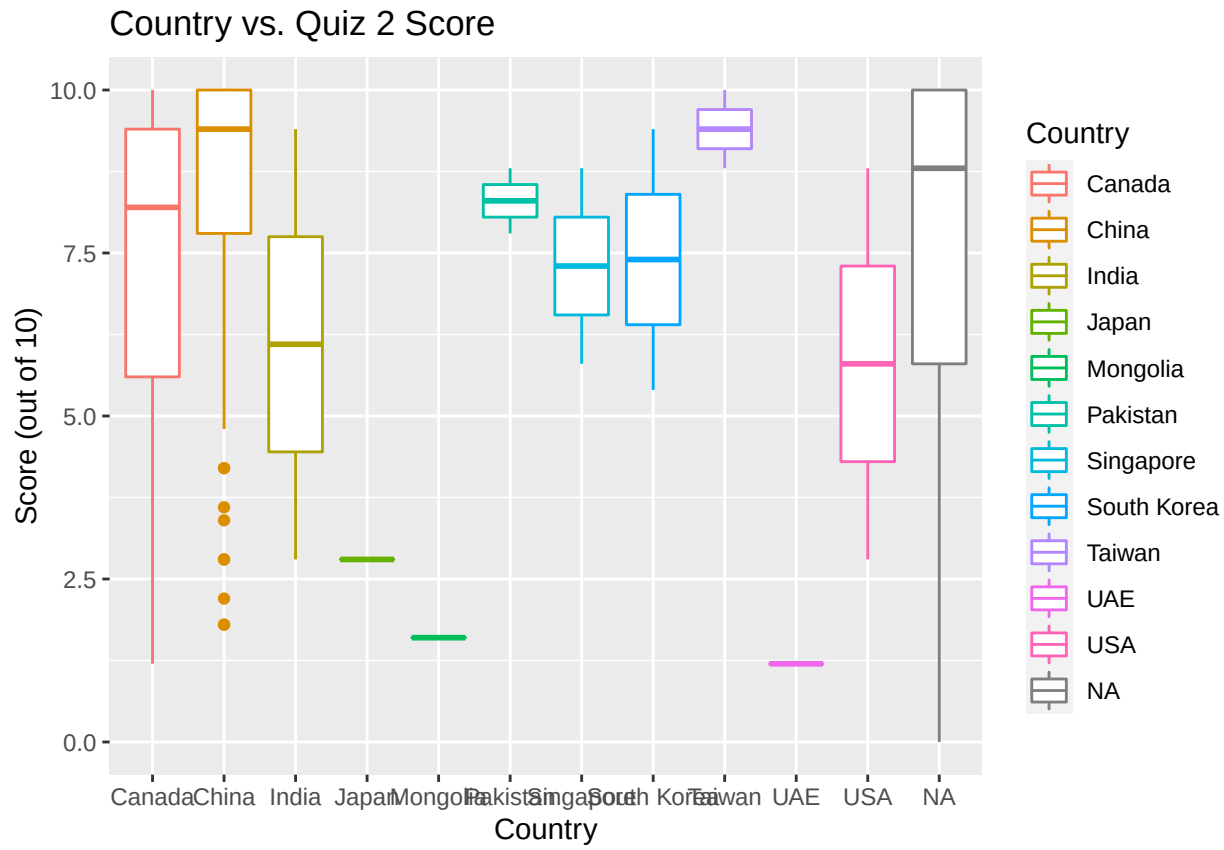

Histogram of Week 3 Time Spent Studying for STA302H1
Frequency
30
20
10
0
0
10
20
30
Time Spent Studying for STA302H1 (hours)

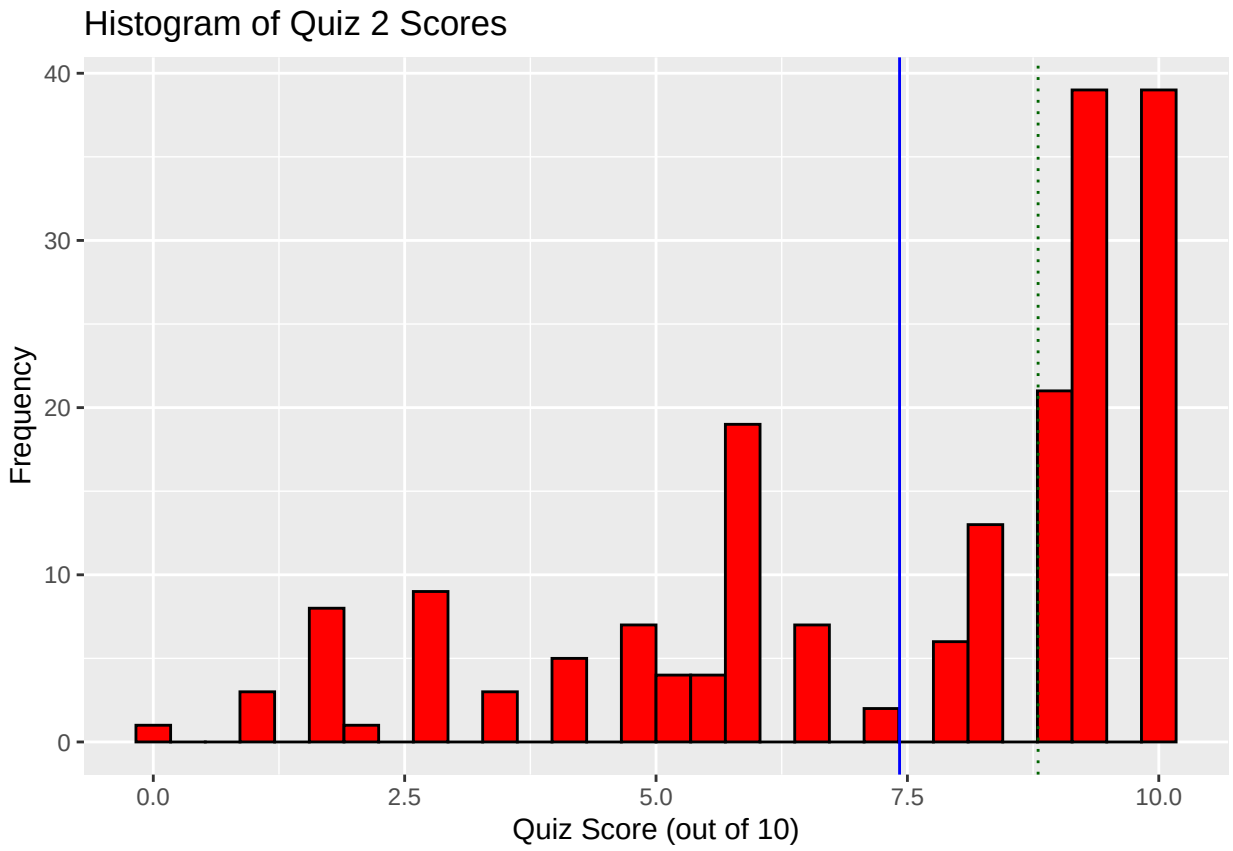

Country vs. Week 4 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
0
20
40
60
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

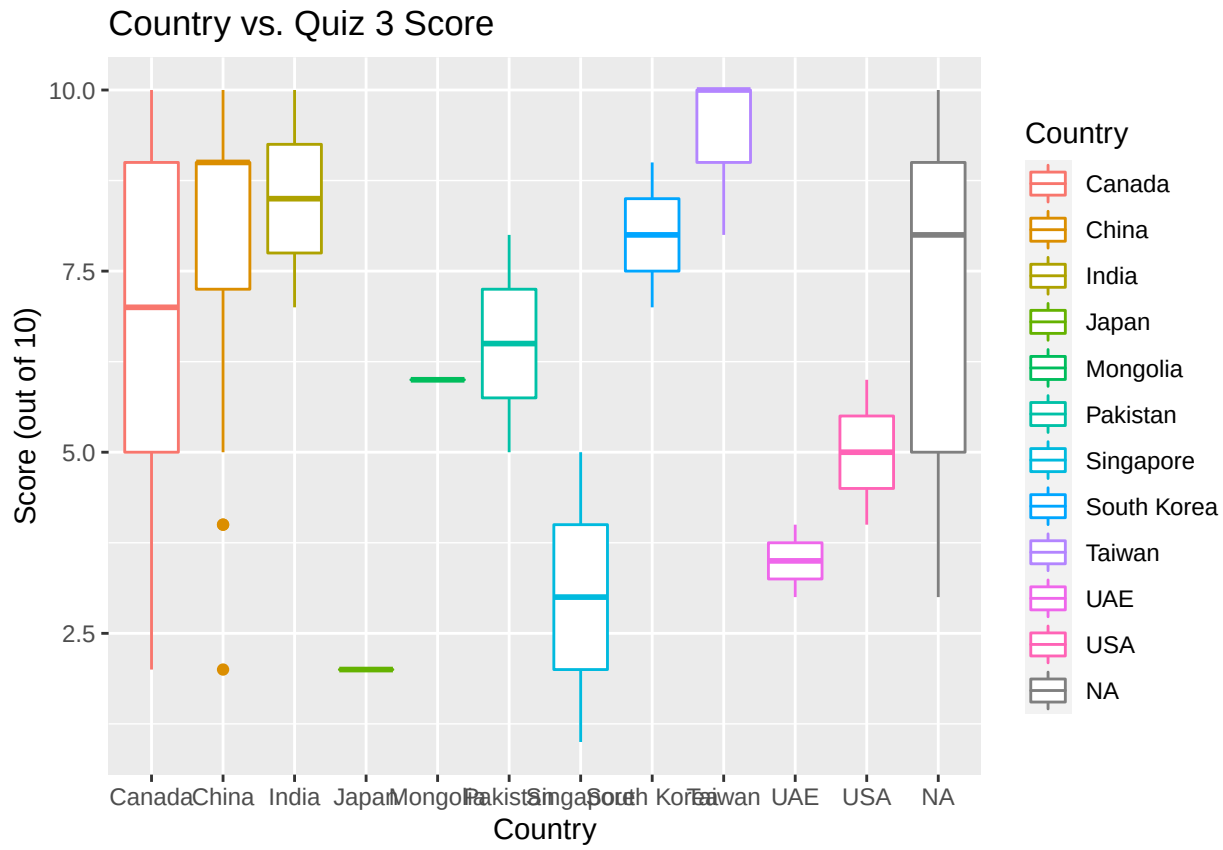

Histogram of Week 4 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
20
40
60
Time Spent Studying for STA302H1 (hours)

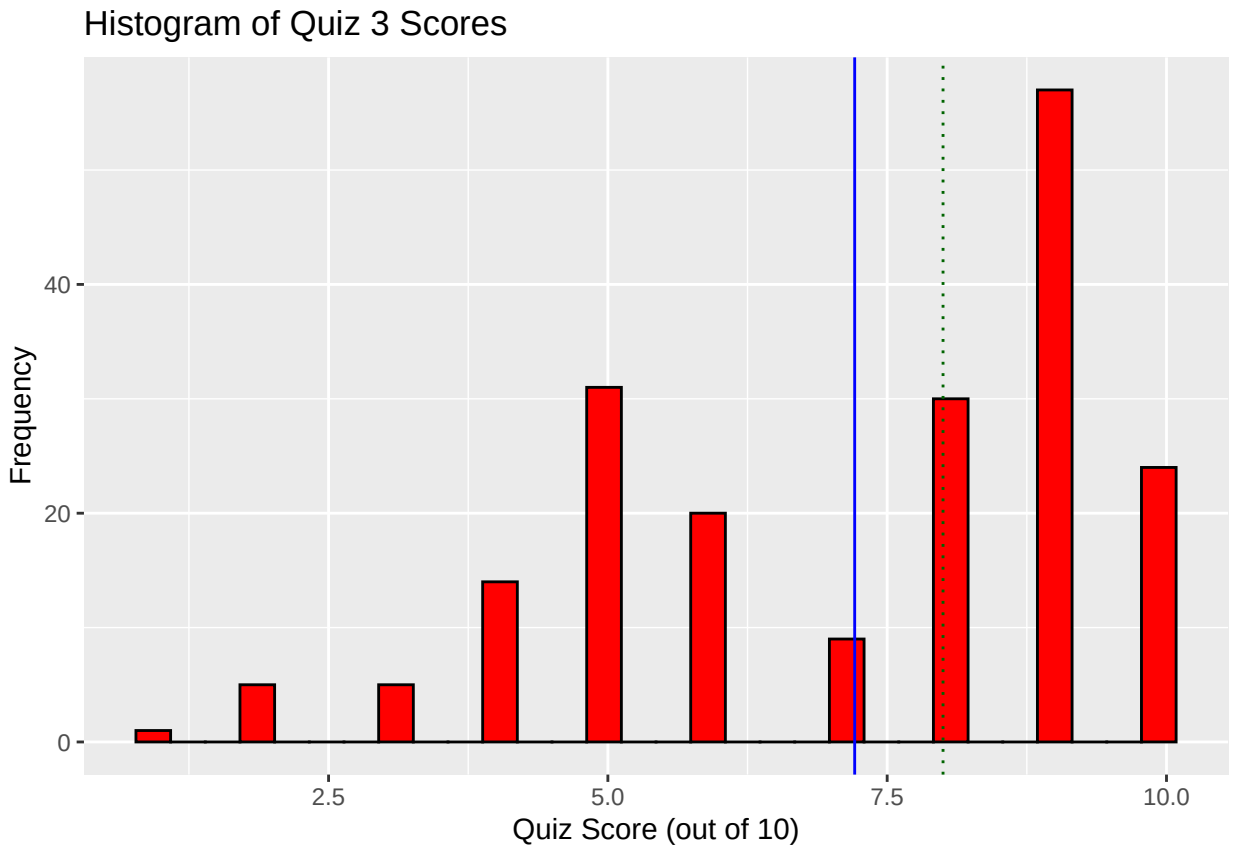

Country vs. Quiz 1 Score
Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

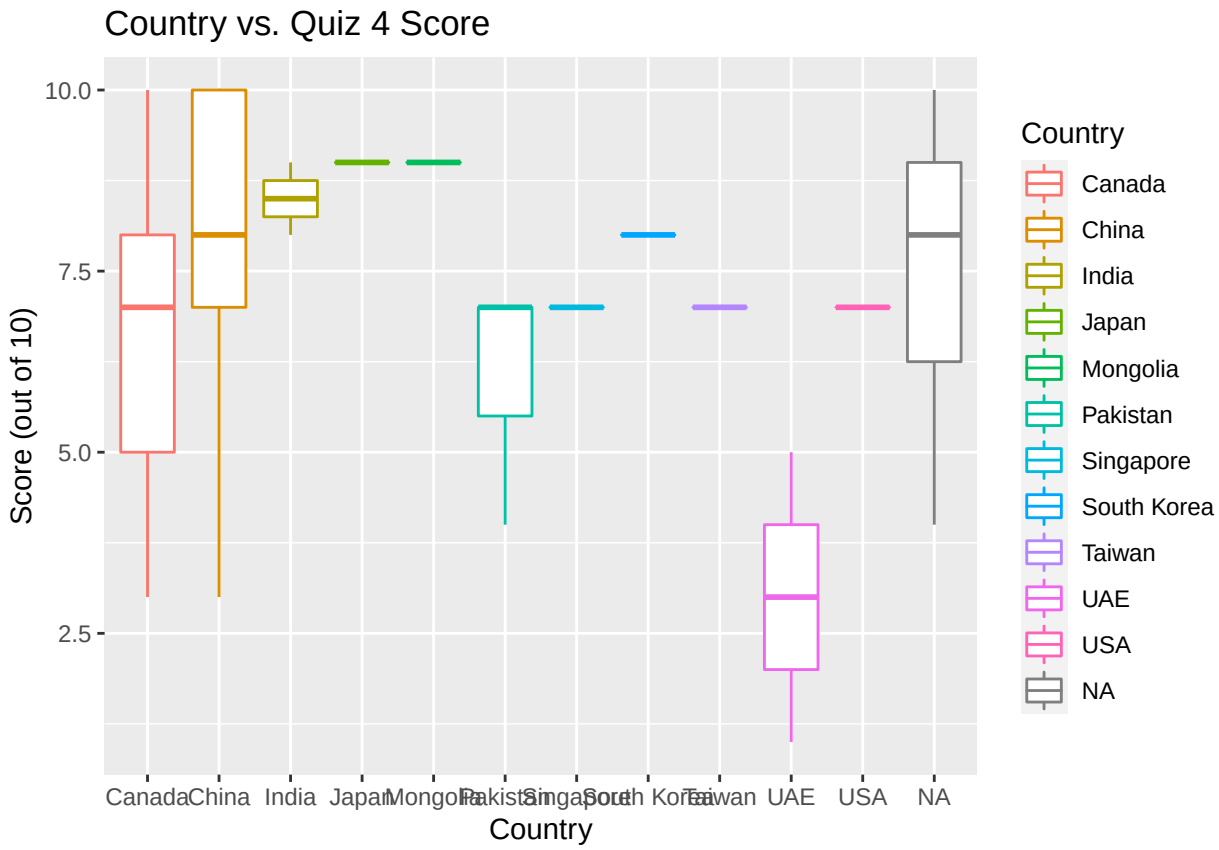

Histogram of Quiz 1 Scores
Frequency
40
30
20
10
0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)

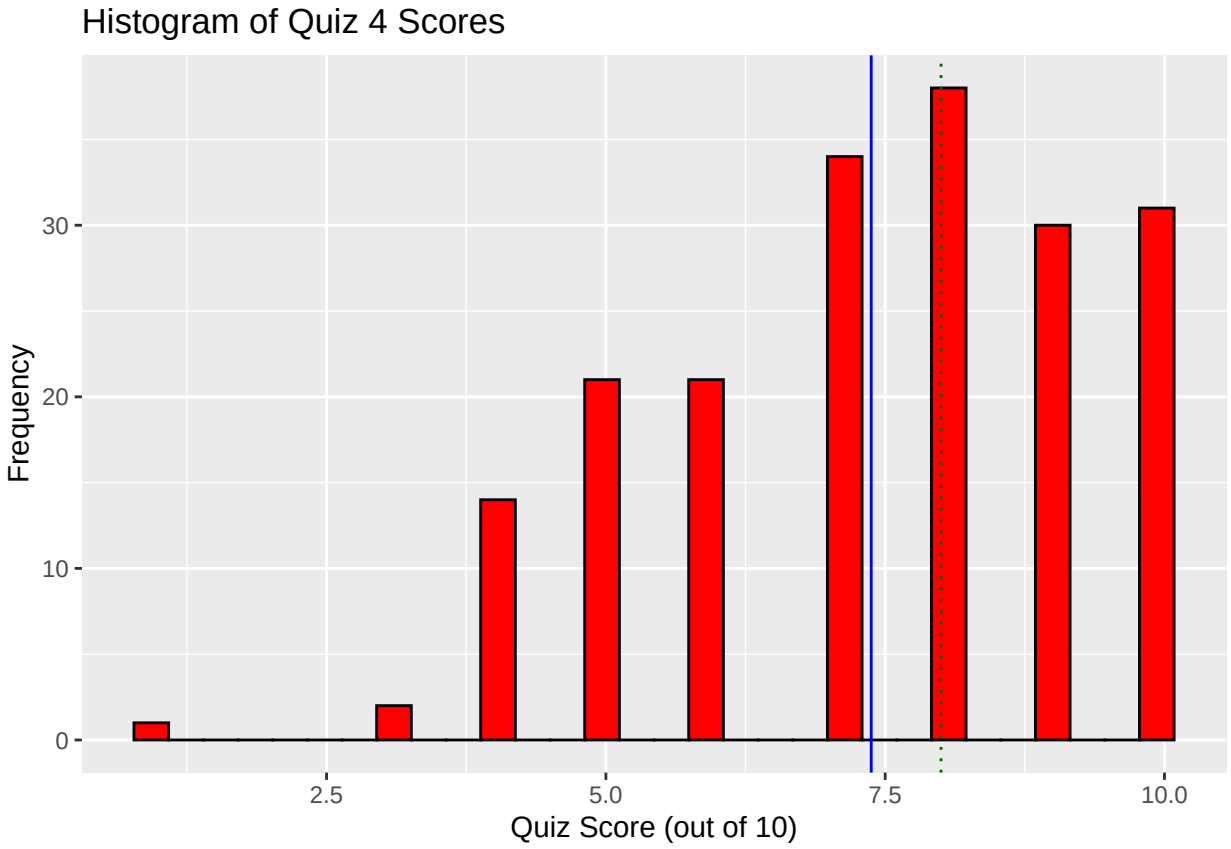

Country vs. Quiz 2 Score
Score (out of 10)
10.0
7.5
5.0
2.5
0.0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA


Histogram of Quiz 2 Scores
Frequency
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)


Country vs. Quiz 3 Score
Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA


Histogram of Quiz 3 Scores
Frequency
40
20
0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)


Country vs. Quiz 4 Score
Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA


Histogram of Quiz 4 Scores
Frequency
30
20
10
0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)

5-Number Summary Statistics

```
summary(cleaned_sta302_performance_data2$COVID.hours..W1.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     1.0     1.0     3.7     2.0   168.0      21
```

```
summary(cleaned_sta302_performance_data2$COVID.hours..W2.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   1.000   2.869   2.000  40.000      19
```

```
summary(cleaned_sta302_performance_data2$COVID.hours..W3.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.500   1.000   2.227   2.000  24.000      11
```

```
summary(cleaned_sta302_performance_data2$COVID.hours..W4.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   1.500   2.917   3.000  50.000      13
```

```
summary(cleaned_sta302_performance_data2$STA302.hours..W1.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   5.000   7.000   7.539   9.000  28.000      21
```

```
summary(cleaned_sta302_performance_data2$STA302.hours..W2.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   8.000   8.403  10.000  20.000      19
```

```
summary(cleaned_sta302_performance_data2$STA302.hours..W3.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    6.00    9.00    9.32   12.00   30.00      10
```

```
summary(cleaned_sta302_performance_data2$STA302.hours..W4.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00    7.00   11.00   13.44   16.00   72.00      13
```

```r
summary(cleaned_sta302_performance_data2$Quiz_1_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   7.000   8.000   7.738   9.000  10.000       8
```

```r
summary(cleaned_sta302_performance_data2$Quiz_2_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   5.800   8.800   7.422   9.400  10.000       8
```

```r
summary(cleaned_sta302_performance_data2$Quiz_3_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   5.000   8.000   7.209   9.000  10.000       3
```

```r
summary(cleaned_sta302_performance_data2$Quiz_4_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   8.000   7.375   9.000  10.000       7
```

```r
cleaned_sta302_performance_data2 = na.omit(cleaned_sta302_performance_data2)

quiz1 = cleaned_sta302_performance_data2$Quiz_1_score
quiz2 = cleaned_sta302_performance_data2$Quiz_2_score
quiz3 = cleaned_sta302_performance_data2$Quiz_3_score
quiz4 = cleaned_sta302_performance_data2$Quiz_4_score

covid1 = cleaned_sta302_performance_data2$COVID.hours..W1.
covid2 = cleaned_sta302_performance_data2$COVID.hours..W2.
covid3 = cleaned_sta302_performance_data2$COVID.hours..W3.
covid4 = cleaned_sta302_performance_data2$COVID.hours..W4.

study1 = cleaned_sta302_performance_data2$STA302.hours..W1.
study2 = cleaned_sta302_performance_data2$STA302.hours..W2.
study3 = cleaned_sta302_performance_data2$STA302.hours..W3.
study4 = cleaned_sta302_performance_data2$STA302.hours..W4.

country = cleaned_sta302_performance_data2$country
```

Full Model (Without Splitting by Country)

```r
additive_model = lm(
  quiz4 ~
    quiz1  # scatterplot seems to have no relationship
  + quiz2  # scatterplot seems to have no relationship
  + quiz3  # scatterplot looks more linear
  + covid1 # must add this linear term b/c i have a quadratic term
  + I(covid1 ^ 2)  # scatterplot looks more quadratic
  + covid2 # must add this linear term b/c i have a quadratic term
  + I(covid2 ^ 2)  # scatterplot looks more quadratic
  + covid3
  # + I(covid3 ^ 2) # scatterplot looks less quadratic
  + covid4  # must add this linear term b/c i have a quadratic term
  + I(covid4 ^ 2) # scatterplot looks more quadratic

  + I(covid1 * covid2) # first impressions from correlation matrix
  + I(covid2 * covid3) # correlation = 0.67
  + I(covid2 * covid4) # discard: correlation = 0.71
  + I(covid3 * covid4) # correlation = 0.72

  + I(study1 * study2)  # correlation = 0.61
  + I(study1 * study3)  # correlation = 0.58
  + I(study2 * study3)  # correlation = 0.70
  + I(study3 * study4)  # correlation = 0.62

  + country  # for simplicity, but backwards process shows this term is not significant
)
summary(additive_model)
```

```
##
## Call:
```

```
## lm(formula = quiz4 ~ quiz1 + quiz2 + quiz3 + covid1 + I(covid1^2) +
##     covid2 + I(covid2^2) + covid3 + covid4 + I(covid4^2) + I(covid1 *
##     covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 *
##     covid4) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4) + country)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5884 -0.8610  0.1800  0.8824  3.2815
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.184400   0.797065   3.995 0.000114 ***
## quiz1                 0.034421   0.081102   0.424 0.672054
## quiz2                 0.047852   0.061074   0.784 0.434941
## quiz3                 0.477087   0.079290   6.017 2.16e-08 ***
## covid1                0.178659   0.126969   1.407 0.162094
## I(covid1^2)           0.016115   0.007279   2.214 0.028818 *
## covid2                0.289324   0.192110   1.506 0.134802
## I(covid2^2)          -0.023657   0.011719  -2.019 0.045850 *
## covid3               -0.053594   0.125976  -0.425 0.671317
## covid4               -0.248941   0.154339  -1.613 0.109497
## I(covid4^2)           0.020698   0.014617   1.416 0.159476
## I(covid1 * covid2)   -0.074201   0.033661  -2.204 0.029489 *
## I(covid2 * covid3)    0.050008   0.031997   1.563 0.120826
## I(covid2 * covid4)    0.040835   0.024083   1.696 0.092671 .
## I(covid3 * covid4)   -0.076459   0.050768  -1.506 0.134798
## I(study1 * study2)   -0.016578   0.006879  -2.410 0.017537 *
## I(study1 * study3)    0.007613   0.005076   1.500 0.136424
## I(study2 * study3)    0.007761   0.004604   1.686 0.094568 .
## I(study3 * study4)   -0.001958   0.001328  -1.474 0.143221
## countryChina          0.585571   0.344768   1.698 0.092127 .
## countryIndia          0.873927   1.174061   0.744 0.458175
## countryMongolia     -12.901734  19.426608  -0.664 0.507938
## countryPakistan      -0.148747   1.593692  -0.093 0.925800
## countrySingapore      1.191079   1.651695   0.721 0.472296
## countrySouth Korea   -0.015750   1.146622  -0.014 0.989064
## countryTaiwan        -1.213168   1.161154  -1.045 0.298309
## countryUAE           -0.631273   1.649231  -0.383 0.702598
## countryUSA            1.456298   1.765878   0.825 0.411256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.582 on 115 degrees of freedom
## Multiple R-squared:  0.4211, Adjusted R-squared:  0.2851
## F-statistic: 3.098 on 27 and 115 DF,  p-value: 1.436e-05
```

```
stepAIC(additive_model, direction = "both")$anova
```

```
## Start:  AIC=156.09
## quiz4 ~ quiz1 + quiz2 + quiz3 + covid1 + I(covid1^2) + covid2 +
##     I(covid2^2) + covid3 + covid4 + I(covid4^2) + I(covid1 *
##     covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 *
##     covid4) + I(study1 * study2) + I(study1 * study3) + I(study2 *
```

```
##     study3) + I(study3 * study4) + country
##
##                       Df Sum of Sq    RSS    AIC
## - country              9    17.220 305.16 146.39
## - quiz1                1     0.451 288.39 154.31
## - covid3               1     0.453 288.39 154.31
## - quiz2                1     1.537 289.48 154.85
## <none>                             287.94 156.09
## - covid1               1     4.957 292.90 156.53
## - I(covid4^2)          1     5.020 292.96 156.56
## - I(study3 * study4)   1     5.440 293.38 156.76
## - I(study1 * study3)   1     5.632 293.57 156.86
## - covid2               1     5.679 293.62 156.88
## - I(covid3 * covid4)   1     5.679 293.62 156.88
## - I(covid2 * covid3)   1     6.116 294.05 157.09
## - covid4               1     6.514 294.45 157.28
## - I(study2 * study3)   1     7.115 295.05 157.58
## - I(covid2 * covid4)   1     7.198 295.14 157.62
## - I(covid2^2)          1    10.203 298.14 159.06
## - I(covid1 * covid2)   1    12.167 300.11 160.00
## - I(covid1^2)          1    12.271 300.21 160.05
## - I(study1 * study2)   1    14.543 302.48 161.13
## - quiz3                1    90.647 378.59 193.22
##
## Step:  AIC=146.39
## quiz4 ~ quiz1 + quiz2 + quiz3 + covid1 + I(covid1^2) + covid2 +
##     I(covid2^2) + covid3 + covid4 + I(covid4^2) + I(covid1 *
##     covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 *
##     covid4) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - quiz1                1     0.013 305.17 144.40
## - covid3               1     0.516 305.67 144.63
## - covid2               1     2.908 308.07 145.75
## - covid1               1     2.914 308.07 145.75
## - quiz2                1     3.587 308.75 146.06
## - covid4               1     3.867 309.03 146.19
## - I(covid4^2)          1     4.193 309.35 146.34
## <none>                             305.16 146.39
## - I(study1 * study3)   1     5.244 310.40 146.83
## - I(study3 * study4)   1     5.879 311.04 147.12
## - I(covid2 * covid4)   1     8.357 313.52 148.26
## - I(covid3 * covid4)   1     8.439 313.60 148.29
## - I(study2 * study3)   1     8.640 313.80 148.38
## - I(covid1 * covid2)   1    10.319 315.48 149.15
## - I(covid1^2)          1    10.436 315.60 149.20
## - I(covid2 * covid3)   1    12.174 317.33 149.99
## - I(covid2^2)          1    12.626 317.79 150.19
## - I(study1 * study2)   1    15.842 321.00 151.63
## + country              9    17.220 287.94 156.09
## - quiz3                1   133.023 438.18 196.13
##
## Step:  AIC=144.4
```

```
## quiz4 ~ quiz2 + quiz3 + covid1 + I(covid1^2) + covid2 + I(covid2^2) +
##     covid3 + covid4 + I(covid4^2) + I(covid1 * covid2) + I(covid2 *
##     covid3) + I(covid2 * covid4) + I(covid3 * covid4) + I(study1 *
##     study2) + I(study1 * study3) + I(study2 * study3) + I(study3 *
##     study4)
##
##                       Df Sum of Sq    RSS    AIC
## - covid3               1     0.519 305.69 142.64
## - covid2               1     2.895 308.07 143.75
## - covid1               1     2.905 308.08 143.75
## - quiz2                1     3.644 308.82 144.10
## - covid4               1     3.871 309.04 144.20
## - I(covid4^2)          1     4.183 309.35 144.34
## <none>                             305.17 144.40
## - I(study1 * study3)   1     5.231 310.40 144.83
## - I(study3 * study4)   1     5.868 311.04 145.12
## - I(covid2 * covid4)   1     8.369 313.54 146.27
## - I(covid3 * covid4)   1     8.430 313.60 146.29
## - I(study2 * study3)   1     8.629 313.80 146.39
## + quiz1                1     0.013 305.16 146.39
## - I(covid1 * covid2)   1    10.307 315.48 147.15
## - I(covid1^2)          1    10.425 315.60 147.20
## - I(covid2 * covid3)   1    12.209 317.38 148.01
## - I(covid2^2)          1    12.665 317.84 148.21
## - I(study1 * study2)   1    15.850 321.02 149.64
## + country              9    16.782 288.39 154.31
## - quiz3                1   147.206 452.38 198.69
##
## Step:  AIC=142.64
## quiz4 ~ quiz2 + quiz3 + covid1 + I(covid1^2) + covid2 + I(covid2^2) +
##     covid4 + I(covid4^2) + I(covid1 * covid2) + I(covid2 * covid3) +
##     I(covid2 * covid4) + I(covid3 * covid4) + I(study1 * study2) +
##     I(study1 * study3) + I(study2 * study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - covid2               1     2.653 308.34 141.88
## - covid1               1     2.820 308.51 141.95
## - quiz2                1     3.471 309.16 142.25
## <none>                             305.69 142.64
## - I(study1 * study3)   1     5.017 310.71 142.97
## - I(covid4^2)          1     5.296 310.99 143.10
## - covid4               1     5.314 311.00 143.11
## - I(study3 * study4)   1     5.697 311.39 143.28
## + covid3               1     0.519 305.17 144.40
## - I(study2 * study3)   1     8.239 313.93 144.44
## + quiz1                1     0.016 305.67 144.63
## - I(covid3 * covid4)   1    10.263 315.95 145.36
## - I(covid1 * covid2)   1    10.616 316.31 145.52
## - I(covid2 * covid4)   1    10.762 316.45 145.59
## - I(covid1^2)          1    10.814 316.50 145.61
## - I(covid2 * covid3)   1    11.692 317.38 146.01
## - I(covid2^2)          1    12.258 317.95 146.26
## - I(study1 * study2)   1    15.364 321.05 147.65
## + country              9    16.843 288.85 152.54
```

```
## - quiz3               1   147.526 453.22 196.95
##
## Step:  AIC=141.88
## quiz4 ~ quiz2 + quiz3 + covid1 + I(covid1^2) + I(covid2^2) +
##     covid4 + I(covid4^2) + I(covid1 * covid2) + I(covid2 * covid3) +
##     I(covid2 * covid4) + I(covid3 * covid4) + I(study1 * study2) +
##     I(study1 * study3) + I(study2 * study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - covid4               1     3.598 311.94 141.54
## - covid1               1     3.718 312.06 141.59
## - quiz2                1     3.720 312.06 141.59
## - I(covid4^2)          1     4.291 312.63 141.85
## <none>                             308.34 141.88
## - I(study1 * study3)   1     5.110 313.45 142.23
## + covid2               1     2.653 305.69 142.64
## - I(study3 * study4)   1     6.430 314.77 142.83
## - I(covid1 * covid2)   1     8.219 316.56 143.64
## - I(covid1^2)          1     8.251 316.59 143.65
## - I(covid3 * covid4)   1     8.431 316.77 143.74
## - I(covid2 * covid4)   1     8.457 316.80 143.75
## + covid3               1     0.276 308.07 143.75
## + quiz1                1     0.000 308.34 143.88
## - I(study2 * study3)   1     9.129 317.47 144.05
## - I(covid2^2)          1     9.860 318.20 144.38
## - I(covid2 * covid3)   1    11.044 319.39 144.91
## - I(study1 * study2)   1    15.548 323.89 146.91
## + country              9    14.020 294.32 153.22
## - quiz3                1   145.086 453.43 195.02
##
## Step:  AIC=141.54
## quiz4 ~ quiz2 + quiz3 + covid1 + I(covid1^2) + I(covid2^2) +
##     I(covid4^2) + I(covid1 * covid2) + I(covid2 * covid3) + I(covid2 *
##     covid4) + I(covid3 * covid4) + I(study1 * study2) + I(study1 *
##     study3) + I(study2 * study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - I(covid4^2)          1     2.711 314.65 140.77
## - quiz2                1     3.034 314.97 140.92
## - covid1               1     4.063 316.00 141.39
## <none>                             311.94 141.54
## + covid4               1     3.598 308.34 141.88
## - I(study1 * study3)   1     5.551 317.49 142.06
## + covid3               1     1.353 310.59 142.91
## + covid2               1     0.937 311.00 143.11
## - I(covid3 * covid4)   1     8.048 319.99 143.18
## - I(study3 * study4)   1     8.556 320.50 143.41
## + quiz1                1     0.005 311.94 143.53
## - I(covid2^2)          1     9.046 320.99 143.62
## - I(covid2 * covid4)   1     9.180 321.12 143.68
## - I(study2 * study3)   1    10.167 322.11 144.12
## - I(covid2 * covid3)   1    10.834 322.78 144.42
## - I(covid1 * covid2)   1    14.592 326.53 146.07
## - I(covid1^2)          1    15.537 327.48 146.49
```

```
## - I(study1 * study2)  1    16.267 328.21 146.81
## + country             9    14.516 297.43 152.72
## - quiz3               1   145.118 457.06 194.16
##
## Step:  AIC=140.77
## quiz4 ~ quiz2 + quiz3 + covid1 + I(covid1^2) + I(covid2^2) +
##     I(covid1 * covid2) + I(covid2 * covid3) + I(covid2 * covid4) +
##     I(covid3 * covid4) + I(study1 * study2) + I(study1 * study3) +
##     I(study2 * study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - quiz2                1     2.945 317.60 140.11
## - covid1               1     3.910 318.56 140.54
## <none>                             314.65 140.77
## - I(study1 * study3)   1     5.793 320.45 141.38
## + I(covid4^2)          1     2.711 311.94 141.54
## + covid3               1     2.060 312.59 141.83
## + covid4               1     2.018 312.63 141.85
## - I(covid2 * covid4)   1     7.249 321.90 142.03
## - I(covid2^2)          1     7.496 322.15 142.14
## - I(covid3 * covid4)   1     8.020 322.67 142.37
## + covid2               1     0.706 313.95 142.45
## - I(study3 * study4)   1     8.306 322.96 142.50
## + quiz1                1     0.004 314.65 142.77
## - I(study2 * study3)   1    10.415 325.07 143.43
## - I(covid2 * covid3)   1    10.841 325.49 143.62
## - I(covid1 * covid2)   1    13.803 328.46 144.91
## - I(covid1^2)          1    14.749 329.40 145.32
## - I(study1 * study2)   1    17.243 331.90 146.40
## + country              9    15.728 298.92 151.44
## - quiz3                1   145.685 460.34 193.18
##
## Step:  AIC=140.11
## quiz4 ~ quiz3 + covid1 + I(covid1^2) + I(covid2^2) + I(covid1 *
##     covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 *
##     covid4) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - covid1               1     3.516 321.11 139.68
## <none>                             317.60 140.11
## + quiz2                1     2.945 314.65 140.77
## - I(study1 * study3)   1     6.076 323.67 140.82
## + I(covid4^2)          1     2.623 314.97 140.92
## - I(covid2 * covid4)   1     7.003 324.60 141.22
## + covid4               1     1.559 316.04 141.40
## + covid3               1     1.544 316.05 141.41
## - I(covid2^2)          1     7.498 325.10 141.44
## - I(covid3 * covid4)   1     7.653 325.25 141.51
## - I(study3 * study4)   1     8.030 325.63 141.68
## + covid2               1     0.920 316.68 141.69
## + quiz1                1     0.168 317.43 142.03
## - I(covid2 * covid3)   1     9.966 327.56 142.52
## - I(study2 * study3)   1    10.228 327.83 142.64
```

```
## - I(covid1 * covid2)  1    12.608 330.21 143.67
## - I(covid1^2)         1    13.517 331.11 144.06
## - I(study1 * study2)  1    17.109 334.71 145.61
## + country             9    17.021 300.58 150.23
## - quiz3               1   157.581 475.18 195.72
##
## Step:  AIC=139.68
## quiz4 ~ quiz3 + I(covid1^2) + I(covid2^2) + I(covid1 * covid2) +
##     I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 * covid4) +
##     I(study1 * study2) + I(study1 * study3) + I(study2 * study3) +
##     I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - I(covid2 * covid4)  1     3.650 324.76 139.30
## - I(covid3 * covid4)  1     4.211 325.33 139.54
## <none>                            321.11 139.68
## - I(covid2^2)         1     4.857 325.97 139.83
## + covid1              1     3.516 317.60 140.11
## - I(study1 * study3)  1     6.198 327.31 140.41
## + quiz2               1     2.551 318.56 140.54
## + I(covid4^2)         1     2.486 318.63 140.57
## - I(covid2 * covid3)  1     6.560 327.67 140.57
## + covid4              1     1.823 319.29 140.87
## + covid3              1     1.402 319.71 141.05
## + covid2              1     1.378 319.74 141.06
## - I(study3 * study4)  1     7.695 328.81 141.07
## + quiz1               1     0.205 320.91 141.59
## - I(study2 * study3)  1     9.490 330.60 141.84
## - I(covid1 * covid2)  1    12.332 333.45 143.07
## - I(covid1^2)         1    12.474 333.59 143.13
## - I(study1 * study2)  1    16.166 337.28 144.70
## + country             9    16.861 304.25 149.97
## - quiz3               1   154.066 475.18 193.72
##
## Step:  AIC=139.3
## quiz4 ~ quiz3 + I(covid1^2) + I(covid2^2) + I(covid1 * covid2) +
##     I(covid2 * covid3) + I(covid3 * covid4) + I(study1 * study2) +
##     I(study1 * study3) + I(study2 * study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - I(covid3 * covid4)  1     0.700 325.46 137.60
## - I(covid2^2)         1     1.738 326.50 138.06
## - I(covid2 * covid3)  1     3.190 327.95 138.69
## <none>                            324.76 139.30
## + I(covid2 * covid4)  1     3.650 321.11 139.68
## - I(study1 * study3)  1     5.613 330.38 139.75
## + covid3              1     3.454 321.31 139.77
## + quiz2               1     2.616 322.15 140.14
## + covid4              1     2.474 322.29 140.20
## + I(covid4^2)         1     0.854 323.91 140.92
## - I(study3 * study4)  1     8.966 333.73 141.19
## + covid1              1     0.163 324.60 141.22
## + quiz1               1     0.056 324.71 141.27
## + covid2              1     0.037 324.73 141.28
```

```
## - I(covid1 * covid2)  1     9.459 334.22 141.40
## - I(covid1^2)         1     9.536 334.30 141.44
## - I(study2 * study3)  1    10.509 335.27 141.85
## - I(study1 * study2)  1    15.455 340.22 143.94
## + country             9    14.961 309.80 150.55
## - quiz3               1   152.570 477.33 192.37
##
## Step:  AIC=137.6
## quiz4 ~ quiz3 + I(covid1^2) + I(covid2^2) + I(covid1 * covid2) +
##     I(covid2 * covid3) + I(study1 * study2) + I(study1 * study3) +
##     I(study2 * study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## - I(covid2^2)          1     1.050 326.51 136.06
## <none>                             325.46 137.60
## - I(study1 * study3)   1     5.152 330.62 137.85
## - I(covid2 * covid3)   1     6.340 331.80 138.36
## + covid4               1     2.632 322.83 138.44
## + quiz2                1     2.454 323.01 138.52
## + covid3               1     1.795 323.67 138.81
## - I(study3 * study4)   1     8.451 333.91 139.27
## + I(covid3 * covid4)   1     0.700 324.76 139.30
## + I(covid4^2)          1     0.504 324.96 139.38
## + I(covid2 * covid4)   1     0.139 325.33 139.54
## + quiz1                1     0.094 325.37 139.56
## + covid1               1     0.069 325.39 139.57
## + covid2               1     0.067 325.40 139.57
## - I(covid1^2)          1     9.205 334.67 139.59
## - I(covid1 * covid2)   1     9.235 334.70 139.60
## - I(study2 * study3)   1    10.038 335.50 139.95
## - I(study1 * study2)   1    14.771 340.23 141.95
## + country              9    14.925 310.54 148.89
## - quiz3                1   152.957 478.42 190.69
##
## Step:  AIC=136.06
## quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 *
##     covid3) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4)
##
##                       Df Sum of Sq    RSS    AIC
## <none>                             326.51 136.06
## - I(study1 * study3)   1     5.058 331.57 136.26
## + quiz2                1     2.723 323.79 136.87
## + I(covid2^2)          1     1.050 325.46 137.60
## - I(covid2 * covid3)   1     8.366 334.88 137.68
## + covid4               1     0.642 325.87 137.78
## + covid3               1     0.586 325.93 137.81
## + covid2               1     0.408 326.11 137.89
## + covid1               1     0.222 326.29 137.97
## + quiz1                1     0.120 326.39 138.01
## + I(covid2 * covid4)   1     0.080 326.43 138.03
## + I(covid3 * covid4)   1     0.013 326.50 138.06
## + I(covid4^2)          1     0.001 326.51 138.06
## - I(covid1^2)          1     9.902 336.42 138.34
```

```
## - I(covid1 * covid2)  1    10.151 336.66 138.44
## - I(study3 * study4)  1    11.051 337.57 138.82
## - I(study2 * study3)  1    11.803 338.32 139.14
## - I(study1 * study2)  1    14.973 341.49 140.48
## + country             9    15.898 310.62 146.93
## - quiz3               1   154.675 481.19 189.52

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## quiz4 ~ quiz1 + quiz2 + quiz3 + covid1 + I(covid1^2) + covid2 +
##     I(covid2^2) + covid3 + covid4 + I(covid4^2) + I(covid1 *
##     covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 *
##     covid4) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4) + country
##
## Final Model:
## quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 *
##     covid3) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4)
##
##
##                     Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                                          115   287.9384 156.0860
## 2             - country  9 17.2202459       124   305.1587 146.3922
## 3               - quiz1  1  0.0132315       125   305.1719 144.3984
## 4              - covid3  1  0.5185280       126   305.6904 142.6411
## 5              - covid2  1  2.6530372       127   308.3435 141.8769
## 6              - covid4  1  3.5979031       128   311.9414 141.5358
## 7         - I(covid4^2)  1  2.7110239       129   314.6524 140.7732
## 8               - quiz2  1  2.9452731       130   317.5977 140.1055
## 9              - covid1  1  3.5163942       131   321.1141 139.6801
## 10 - I(covid2 * covid4)  1  3.6497154       132   324.7638 139.2962
## 11 - I(covid3 * covid4)  1  0.7000985       133   325.4639 137.6042
## 12        - I(covid2^2)  1  1.0504757       134   326.5143 136.0650
```

```
# I decide to remove more terms for simplicity.
additive_model2 = lm(
quiz4 ~ quiz3
  # + I(covid1 ^ 2)  # this lone quadratic term add a lot of complexity for negligible change in R^2 an
  # + I(covid1 * covid2) + I(covid2 * covid3) # these terms alone add complexity -- harder to interpret

  + I(study1 * study2)
# + I(study1 * study3)  # make weeks consecutive: "want to see correlation from week to week", rather t
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(additive_model2)
```

```
##
## Call:
## lm(formula = quiz4 ~ quiz3 + I(study1 * study2) + I(study2 *
```

```
##     study3) + I(study3 * study4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9789 -0.8534  0.2102  1.0730  3.4523
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.937808   0.513776   7.664 2.88e-12 ***
## quiz3               0.483867   0.062342   7.762 1.69e-12 ***
## I(study1 * study2) -0.006539   0.003378  -1.936   0.0549 .
## I(study2 * study3)  0.006867   0.004117   1.668   0.0976 .
## I(study3 * study4) -0.001531   0.001093  -1.401   0.1634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.573 on 138 degrees of freedom
## Multiple R-squared:  0.3135, Adjusted R-squared:  0.2936
## F-statistic: 15.76 on 4 and 138 DF,  p-value: 1.207e-10
```

```
# Doing stepAIC on a well-fitted model produces the same model.
# The model is already in a "steady state."
stepAIC(additive_model2, direction = "both")$anova
```

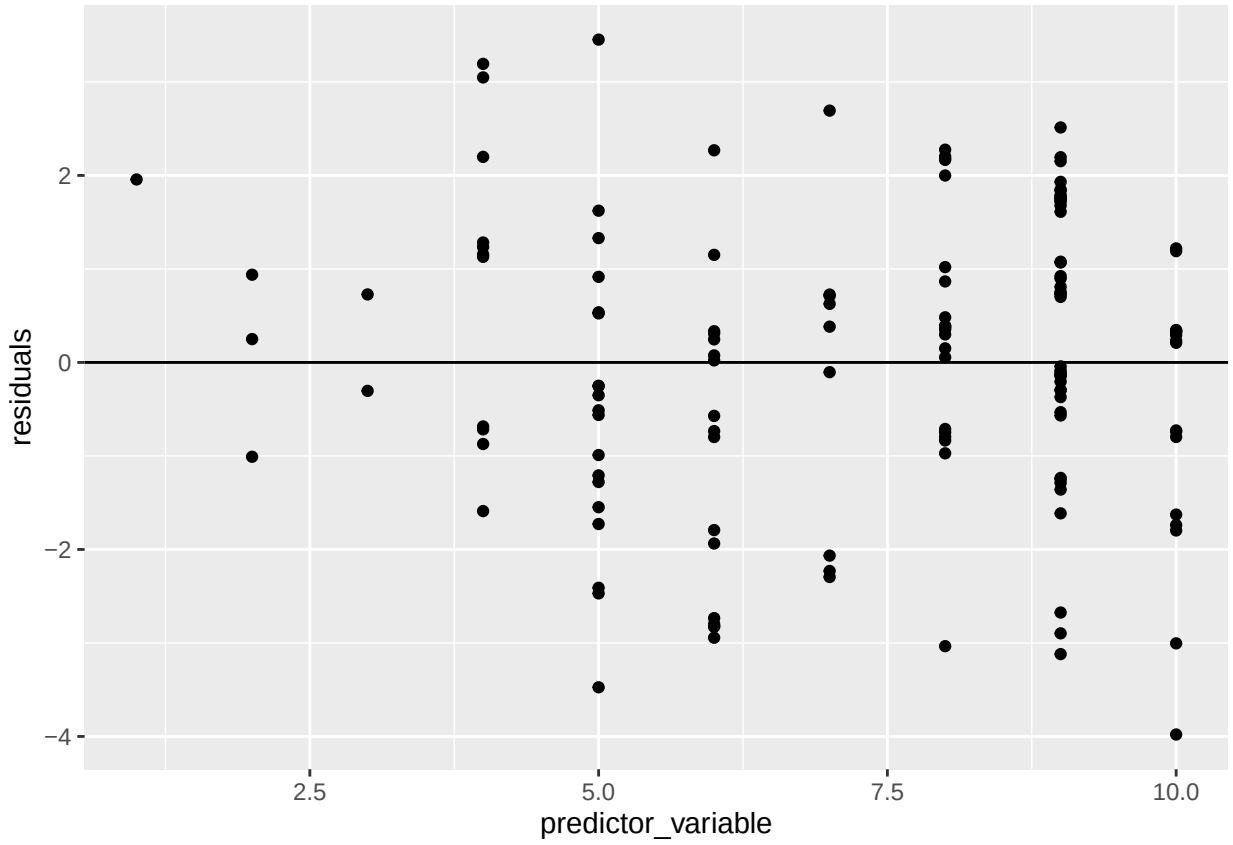
```
## Start:  AIC=134.45
## quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 *
##     study4)
##
##                       Df Sum of Sq    RSS    AIC
## <none>                              341.42 134.45
## - I(study3 * study4)  1     4.857 346.28 134.47
## - I(study2 * study3)  1     6.883 348.31 135.30
## - I(study1 * study2)  1     9.271 350.69 136.28
## - quiz3               1   149.041 490.46 184.25

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 *
##     study4)
##
## Final Model:
## quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 *
##     study4)
##
##
##   Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                       138   341.4222 134.4493
```

```
display_residual_plot <- function(data, model, predictor_variable) {
  fit = fitted(model)
  residuals = resid(model)
```

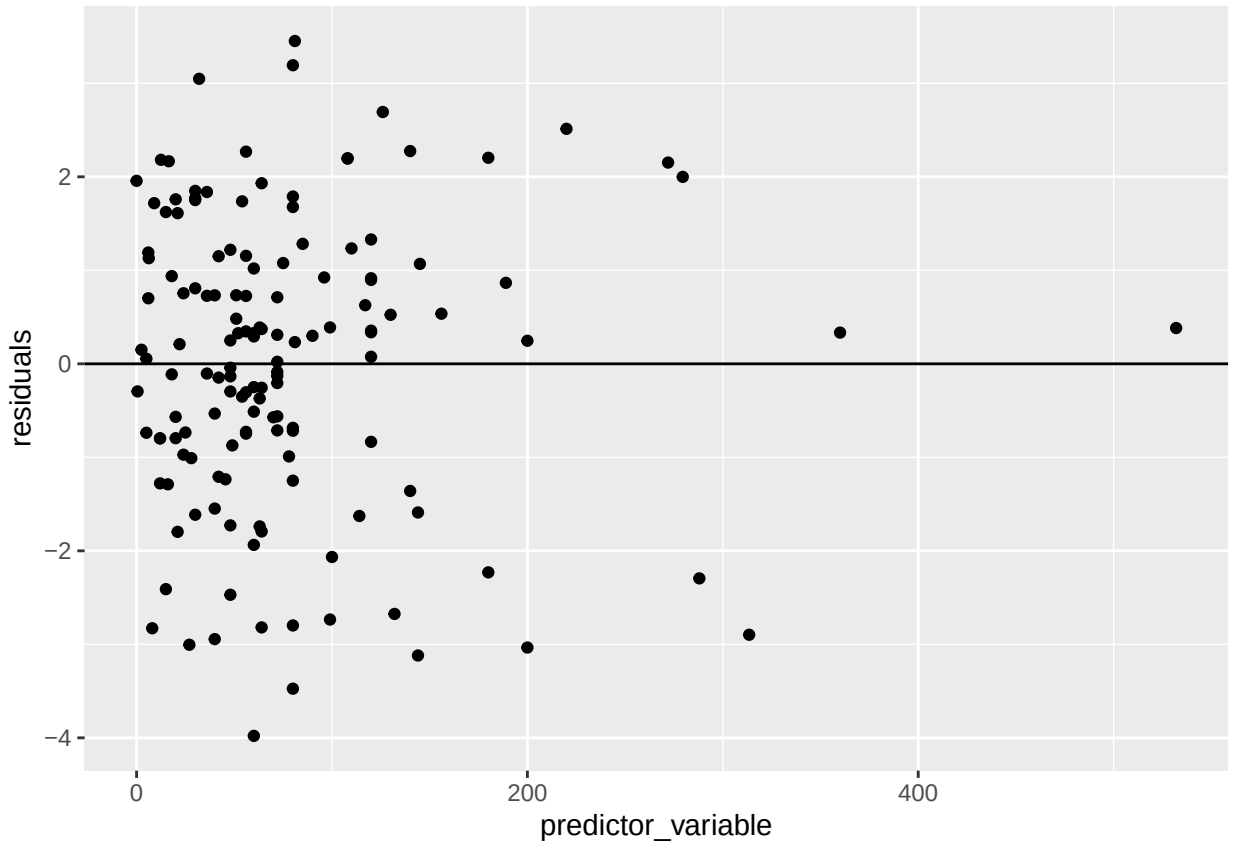
```
  ggplot(data = data, aes(x = predictor_variable, y = residuals)) +
    geom_point() +
    geom_hline(yintercept = 0)
}
```

residual plot

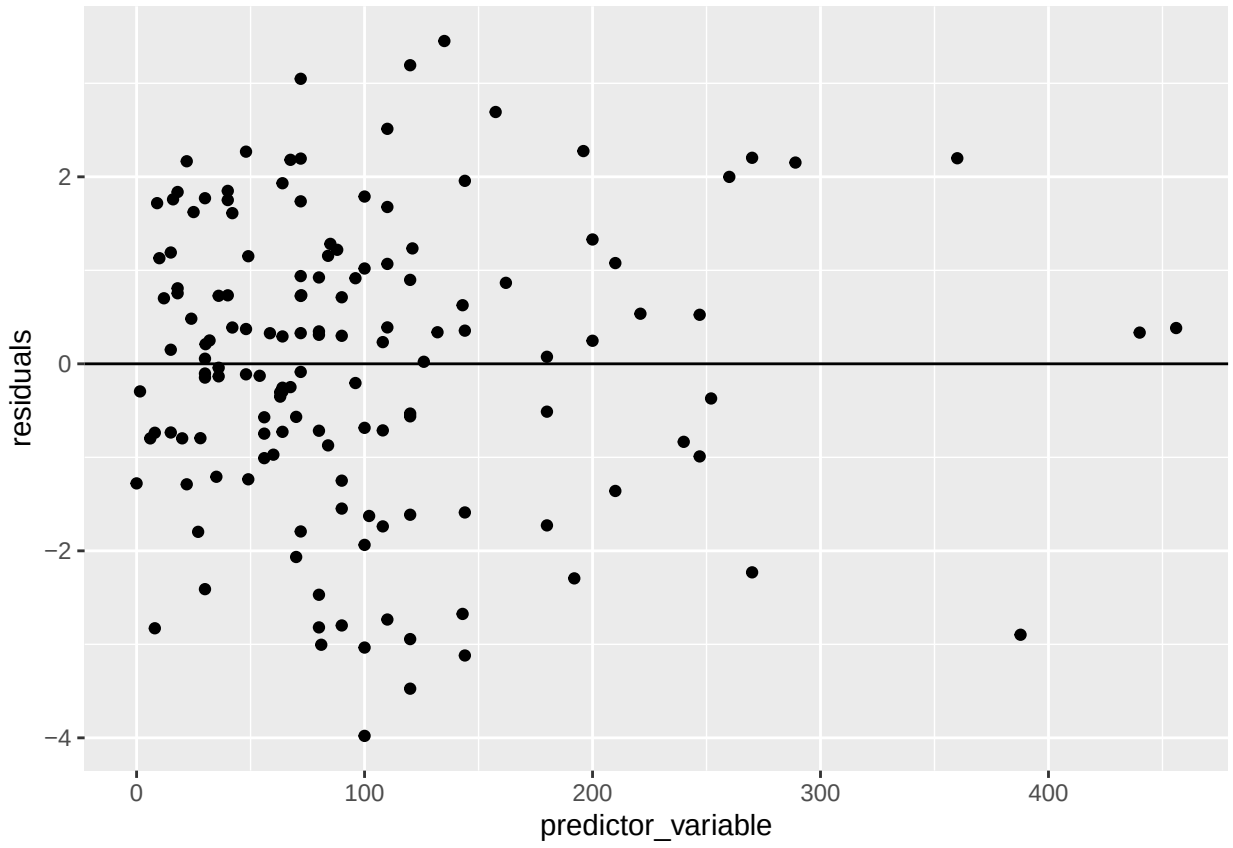
```
display_residual_plot(cleaned_sta302_performance_data2, additive_model2, quiz3)
```



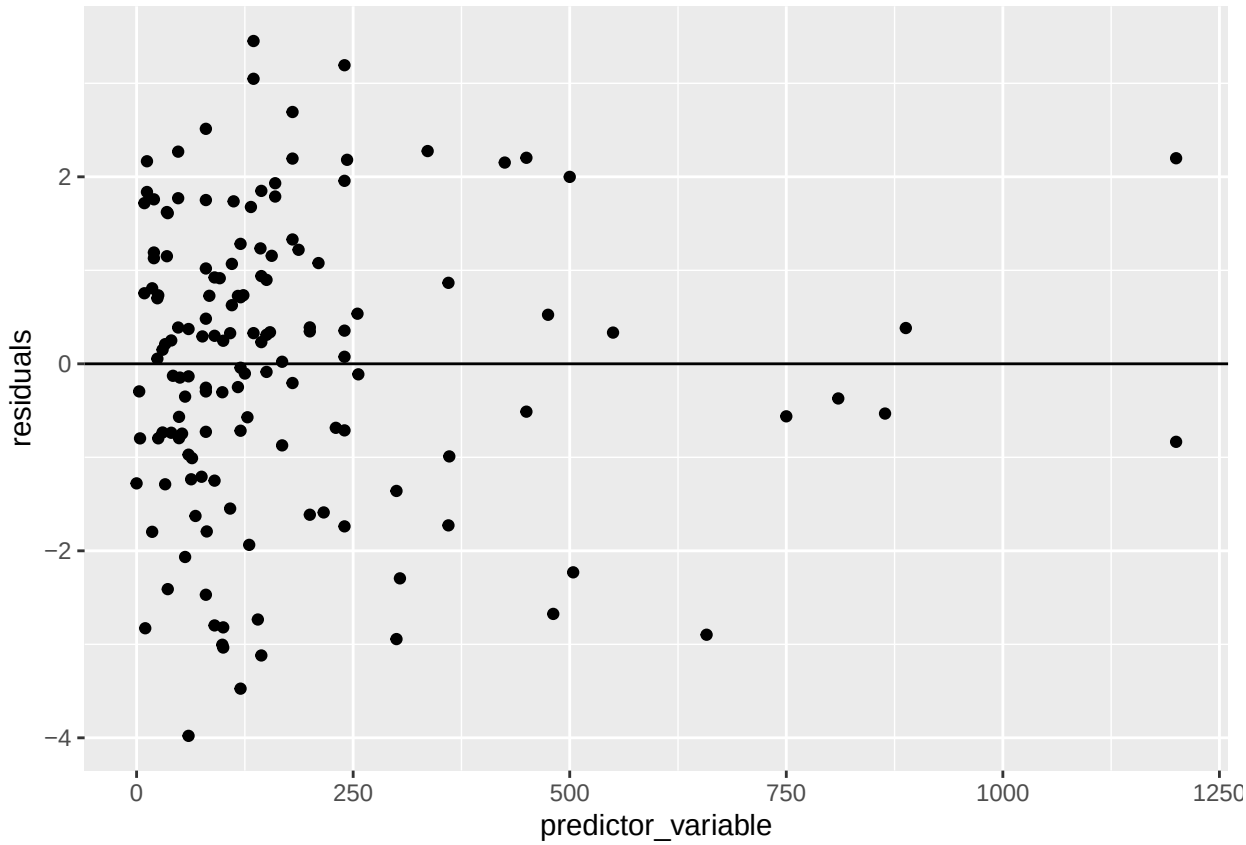

```
display_residual_plot(cleaned_sta302_performance_data2, additive_model2, study1 * study2)
```




```
display_residual_plot(cleaned_sta302_performance_data2, additive_model2, study2 * study3)
```

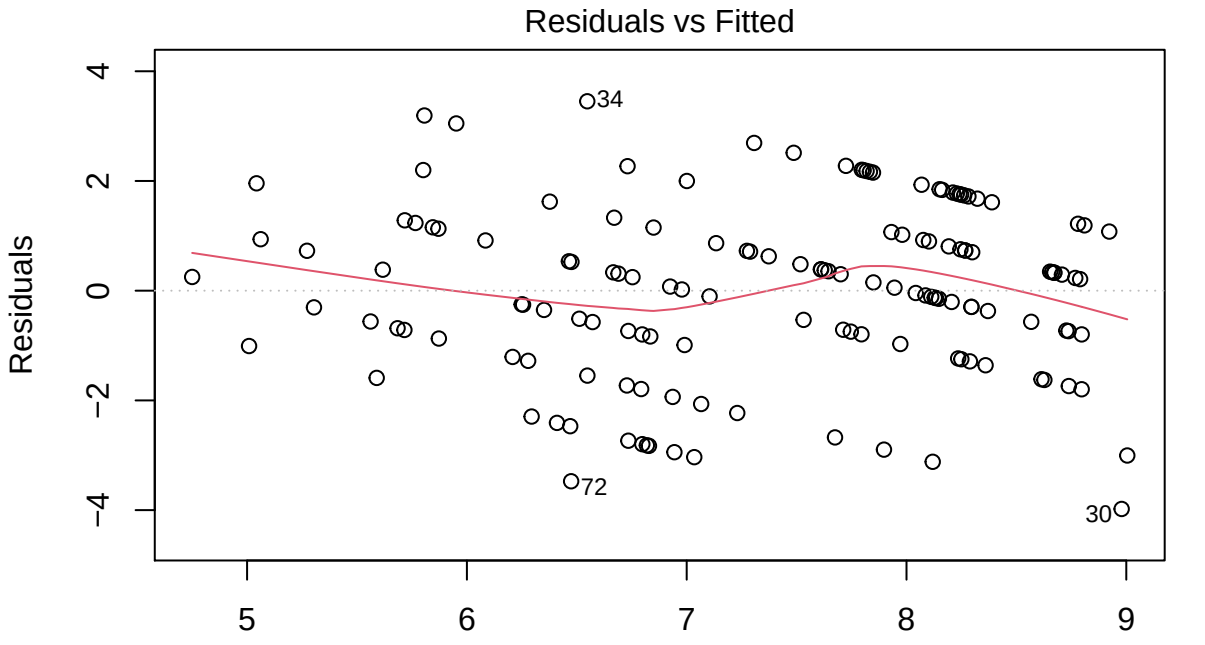


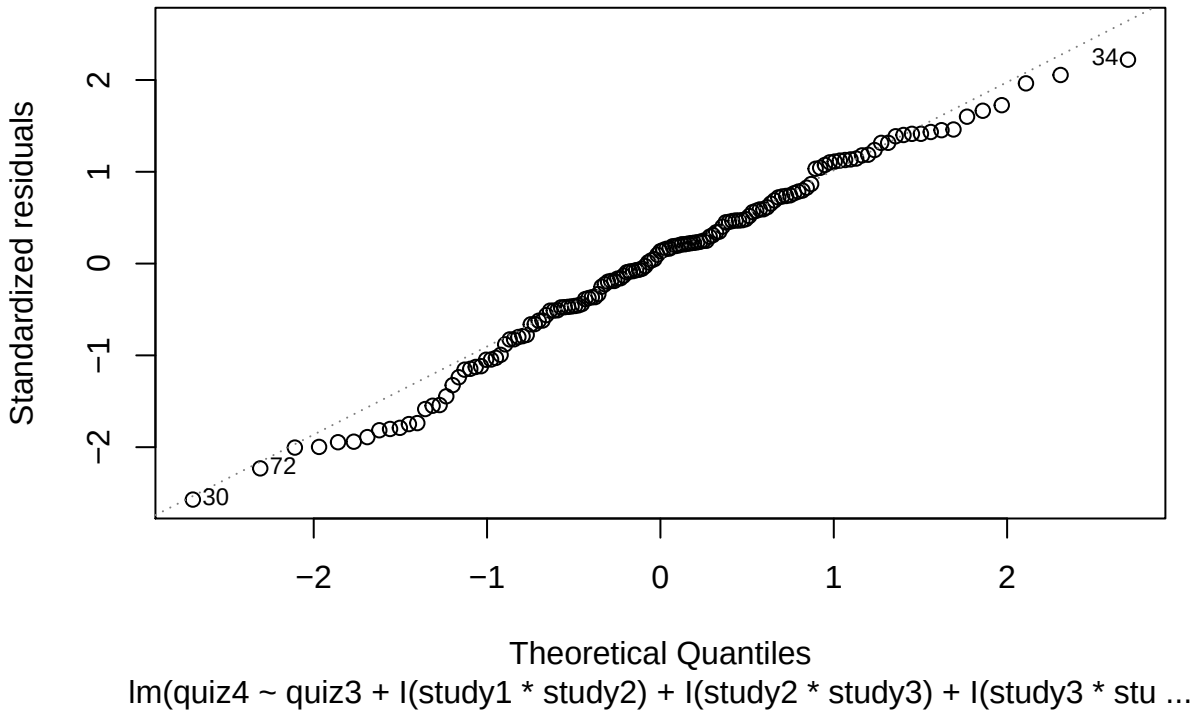

```
display_residual_plot(cleaned_sta302_performance_data2, additive_model2, study3 * study4)
```

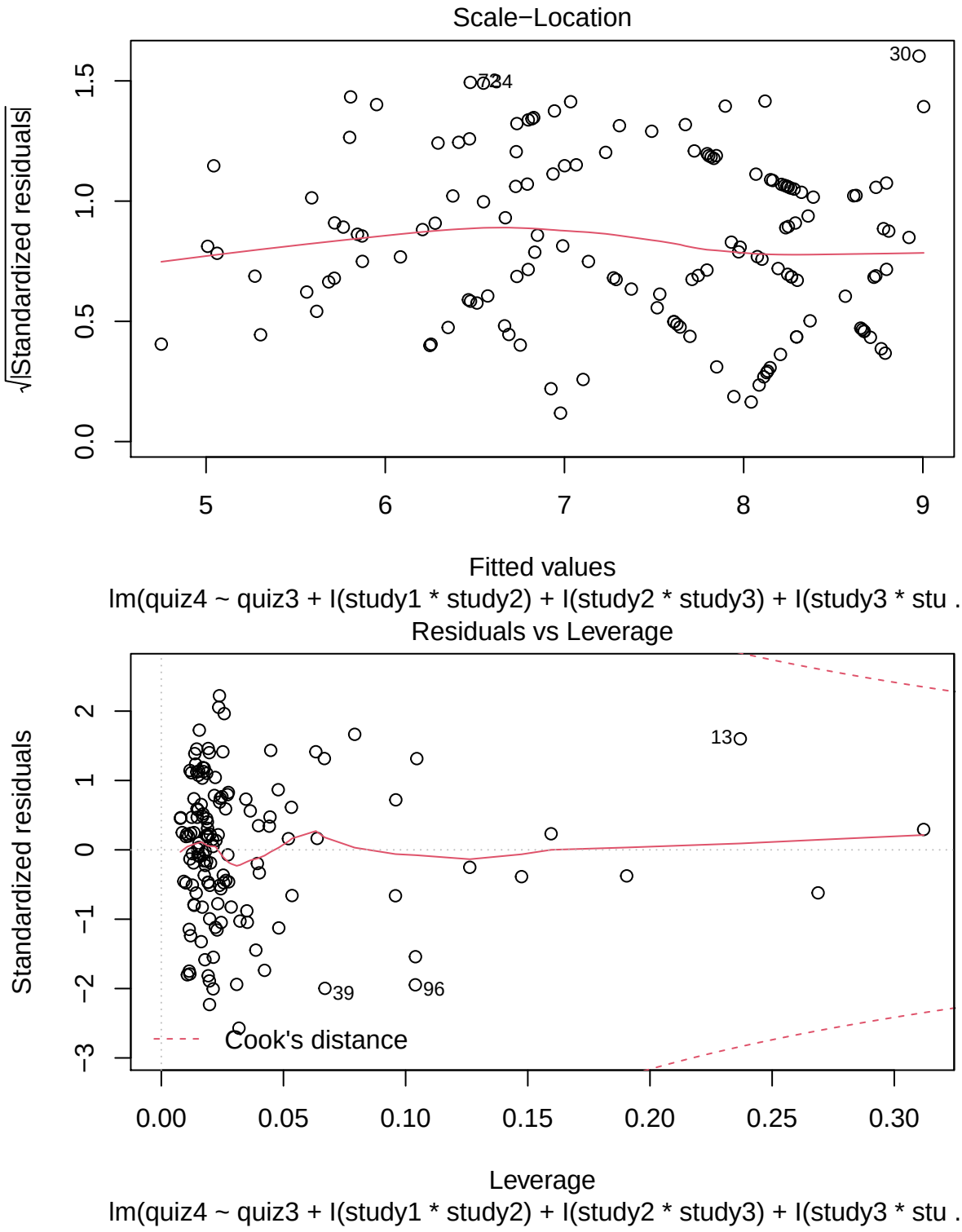


residual vs. fit, qqplot, scale-location, and residual vs. leverage

```
plot(additive_model2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 * stu ...
34
72
30


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 * stu ...
34
72
30


Scale−Location
√|Standardized residuals|
Fitted values
lm(quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 * stu ...
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(quiz4 ~ quiz3 + I(study1 * study2) + I(study2 * study3) + I(study3 * stu ...