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.

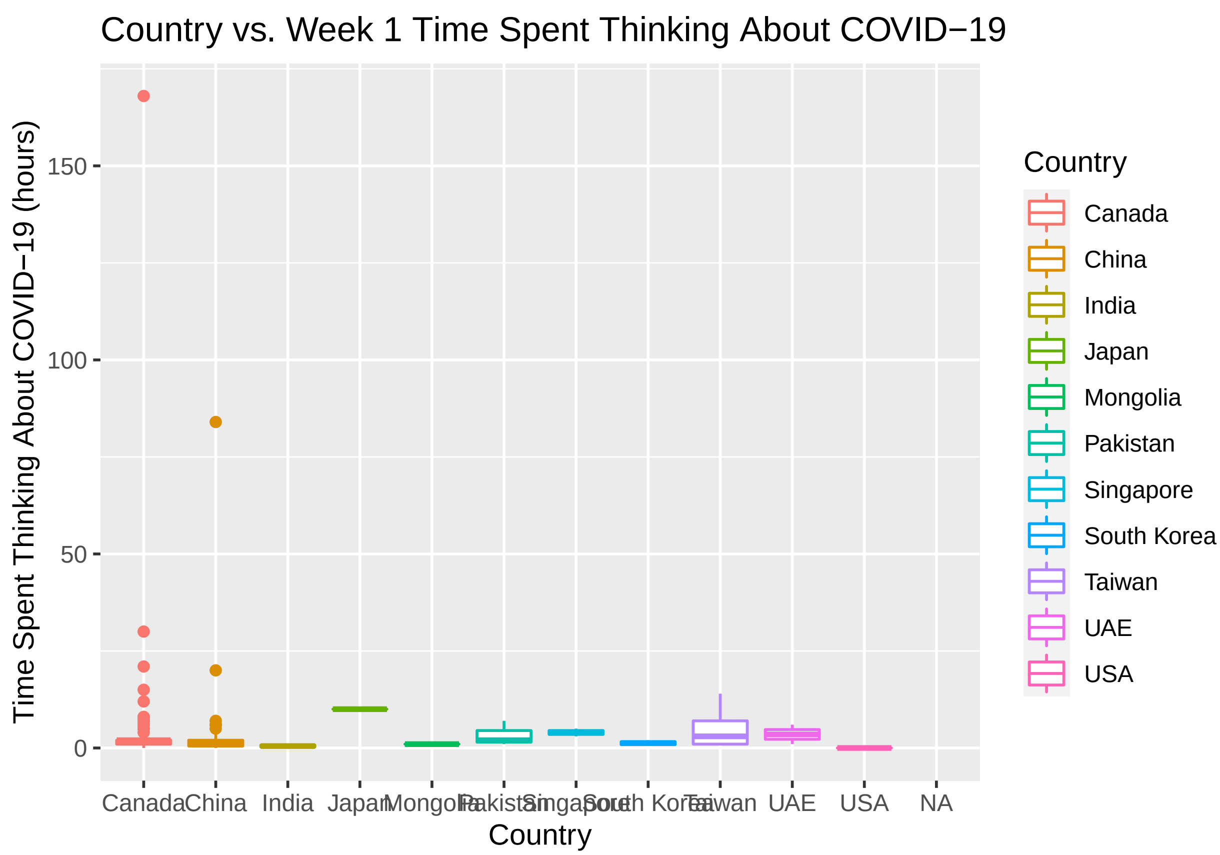
```
rows_with_no_country = cleaned_sta302_performance_data2 %>%
  dplyr::select(-country)
```

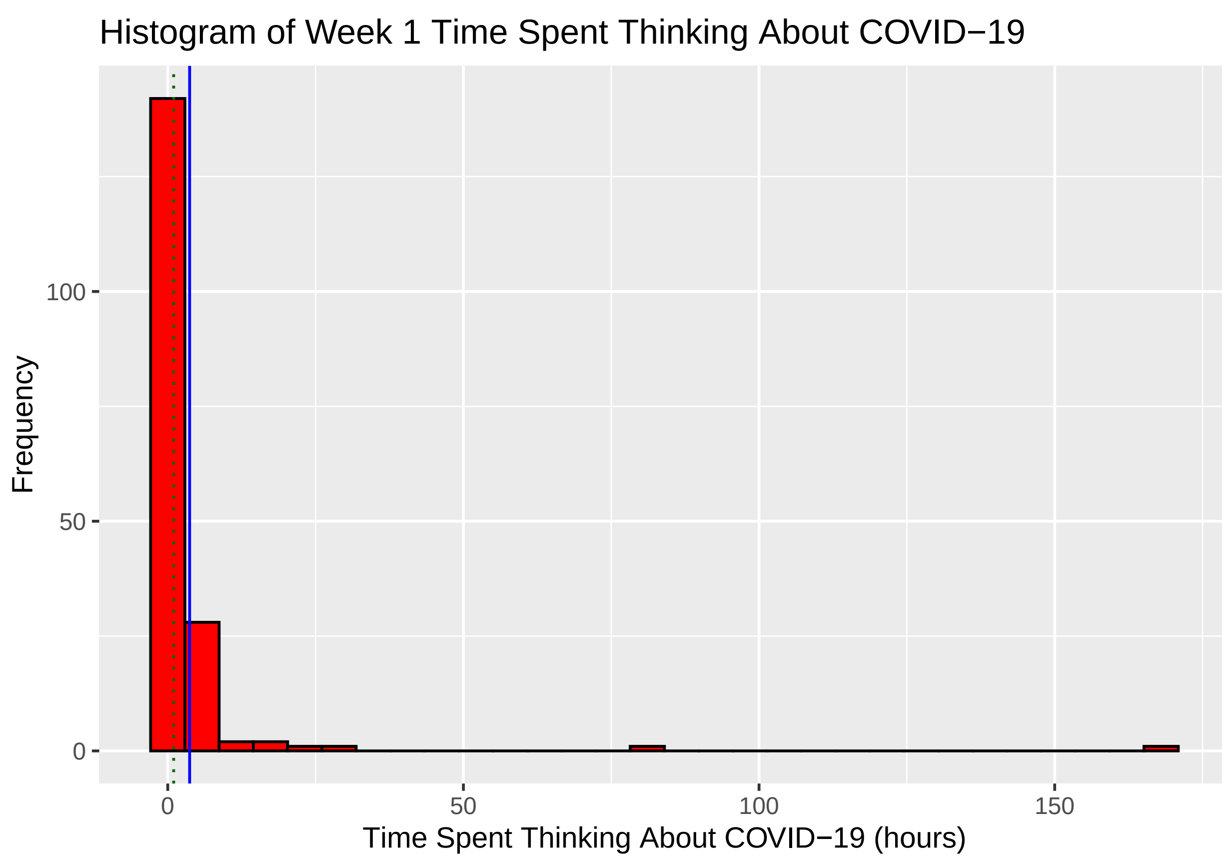
Rows Filtered by Country

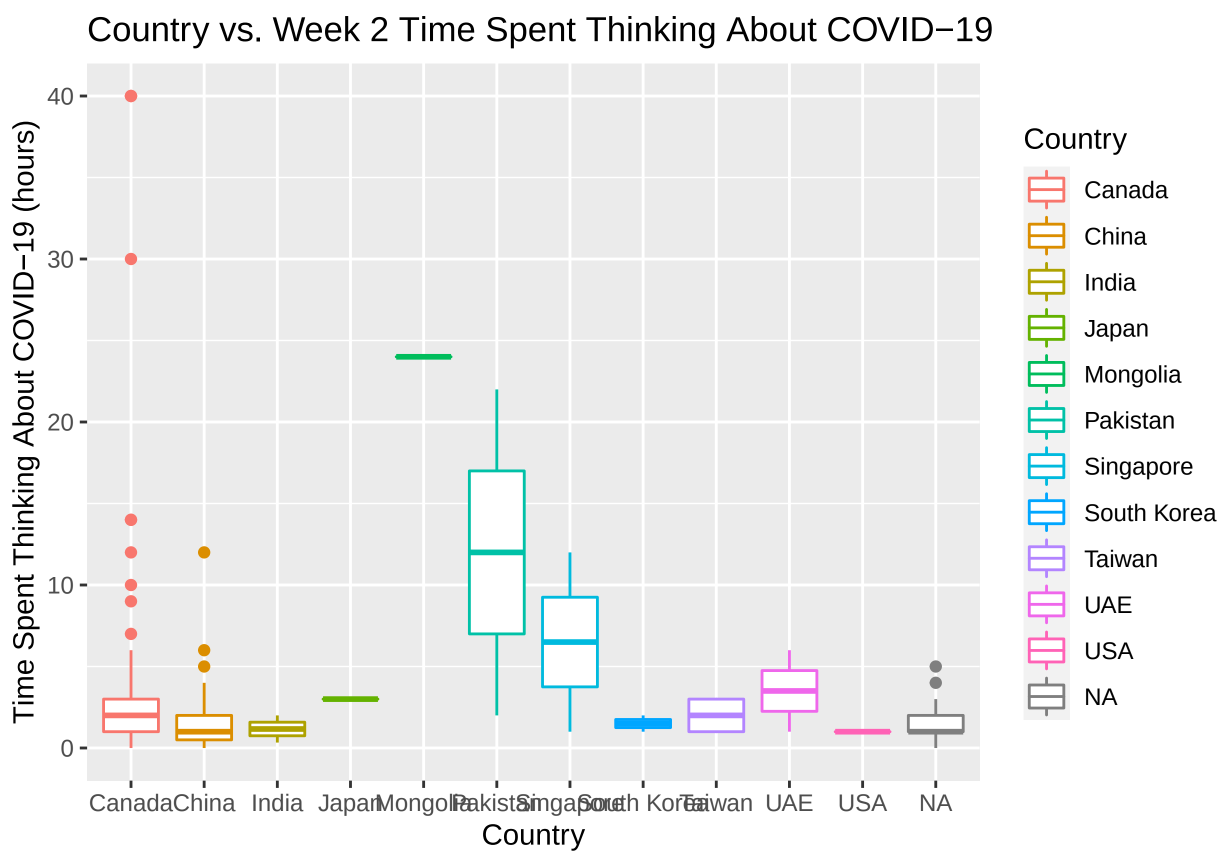
This is useful if we want data for individual countries.
Only the first and last code snippets are shown.

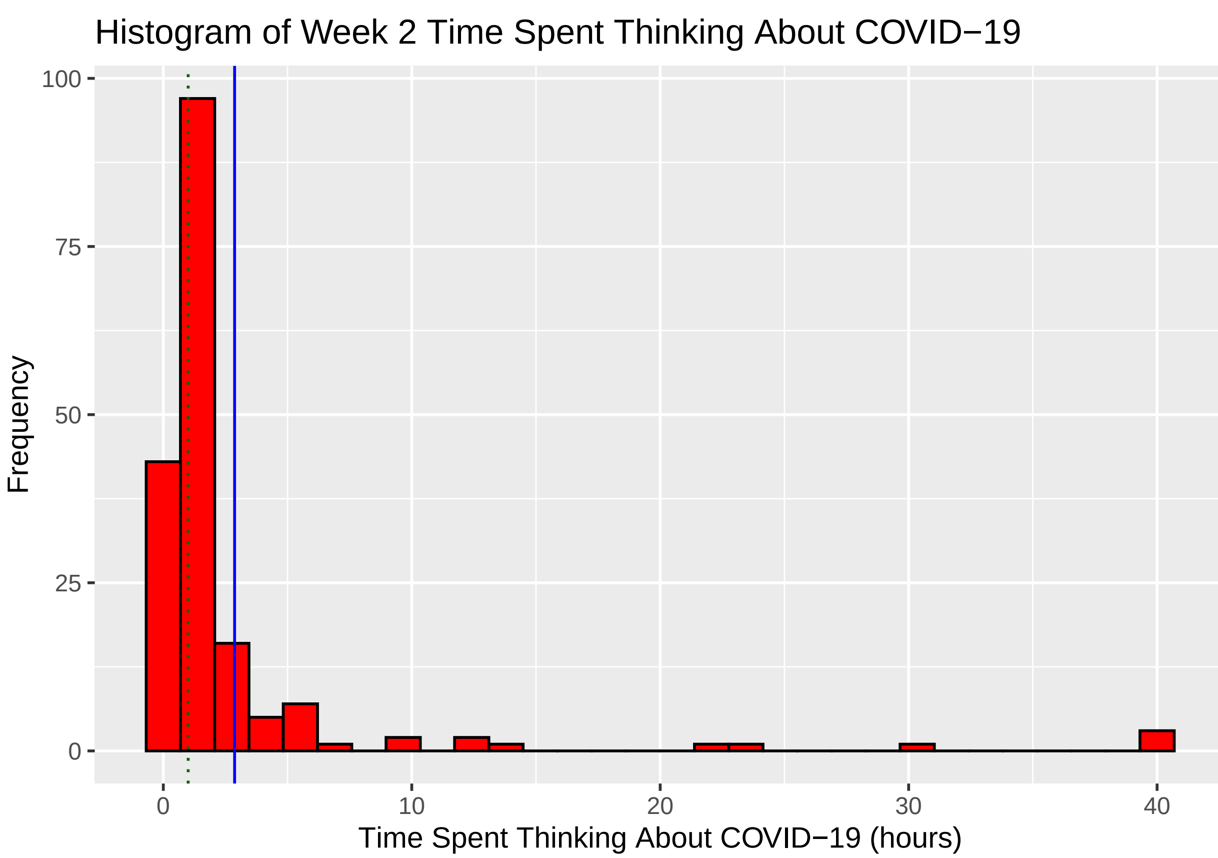
```
canada <- cleaned_sta302_performance_data2 %>%
  filter(as.character(country) == "Canada") %>%
  dplyr::select(-country)

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

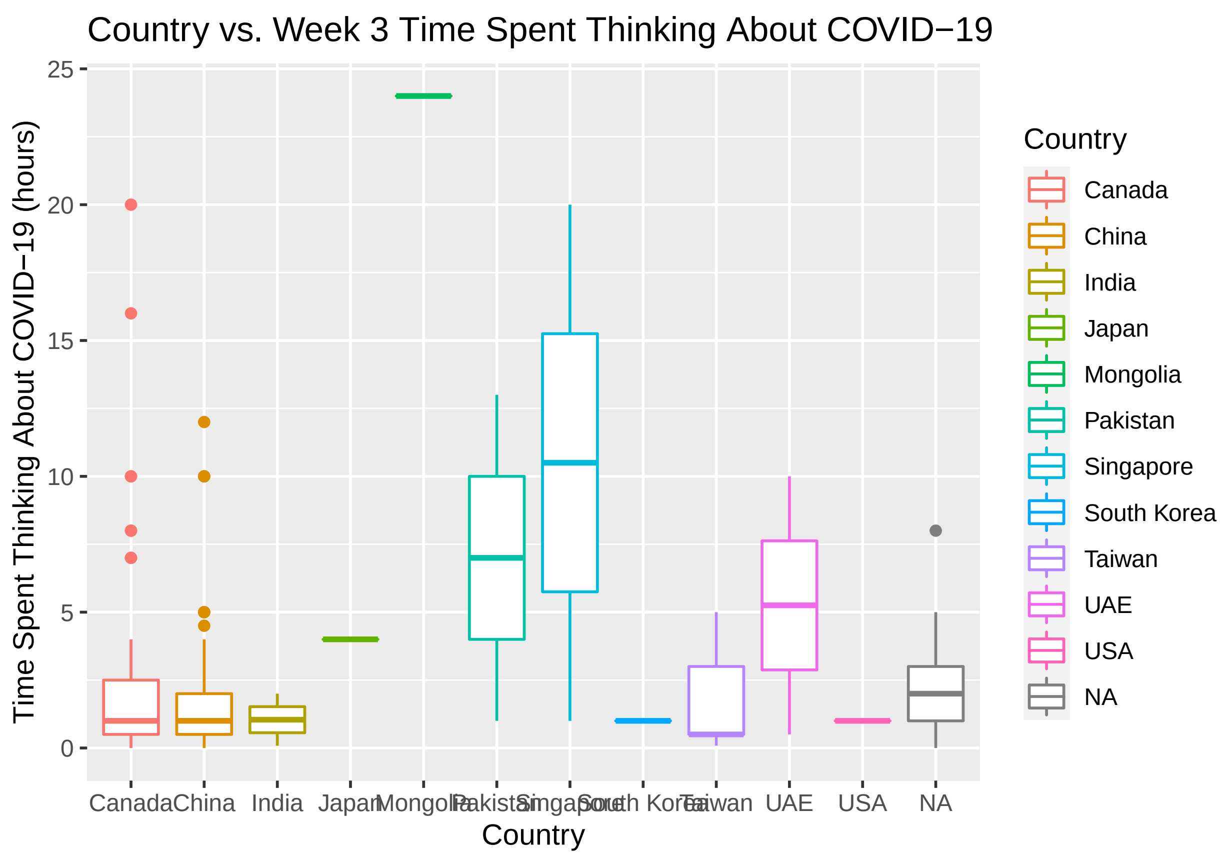
```
##             Country
## Canada           97
## China            63
## India             2
## Japan             1
## Mongolia          1
## Pakistan          3
## Singapore         2
## South_Korea       2
## Taiwan            3
## UAE               2
## USA               2
## Unknown          21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

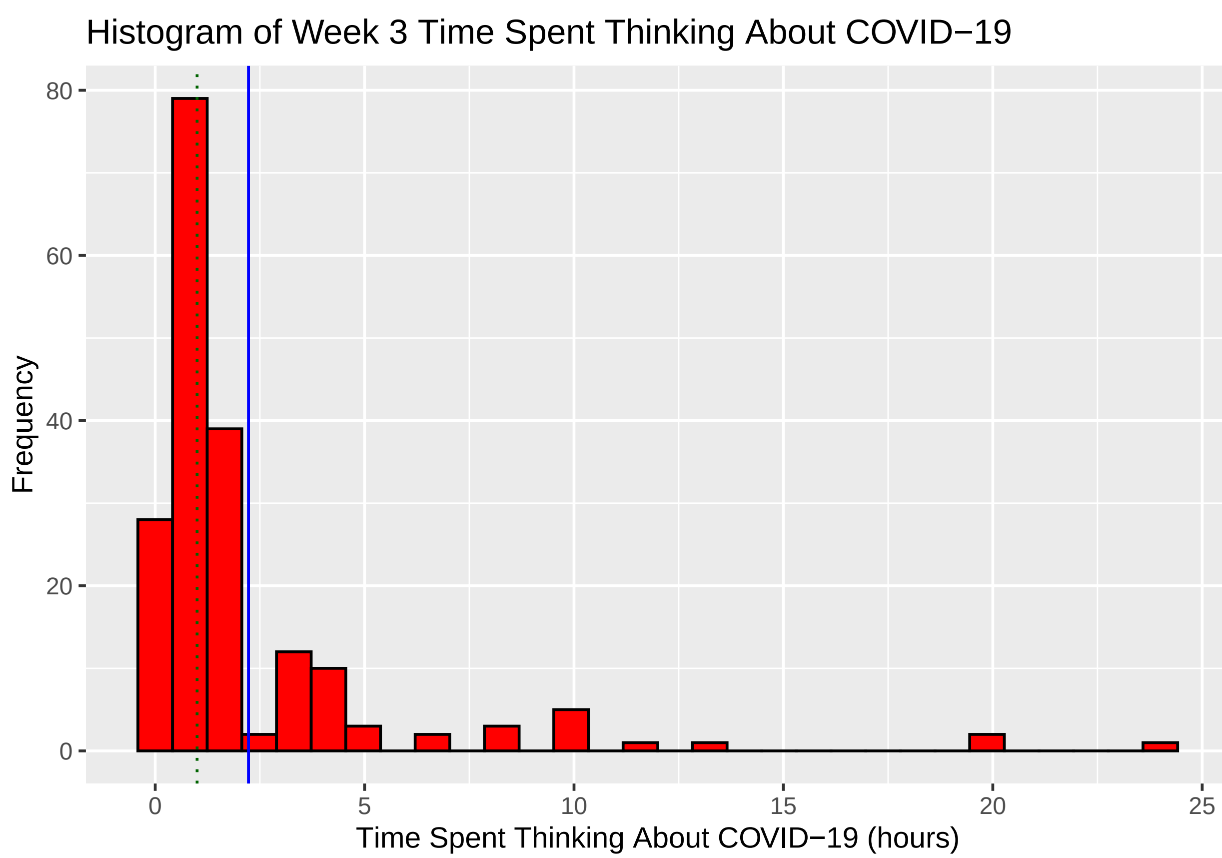

Country vs. Week 1 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
150
100
50
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA

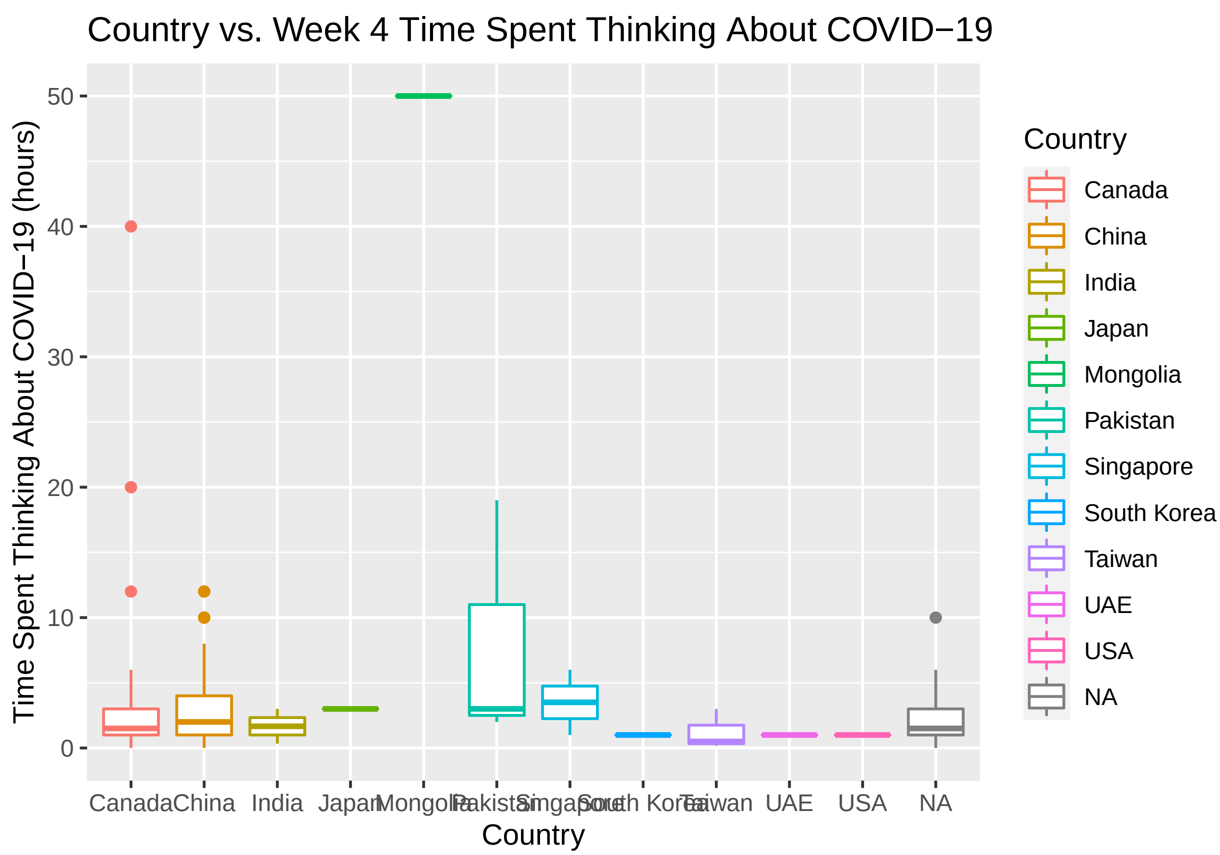

Histogram of Week 1 Time Spent Thinking About COVID−19
Frequency
100
50
0
0
50
100
150
Time Spent Thinking About COVID−19 (hours)

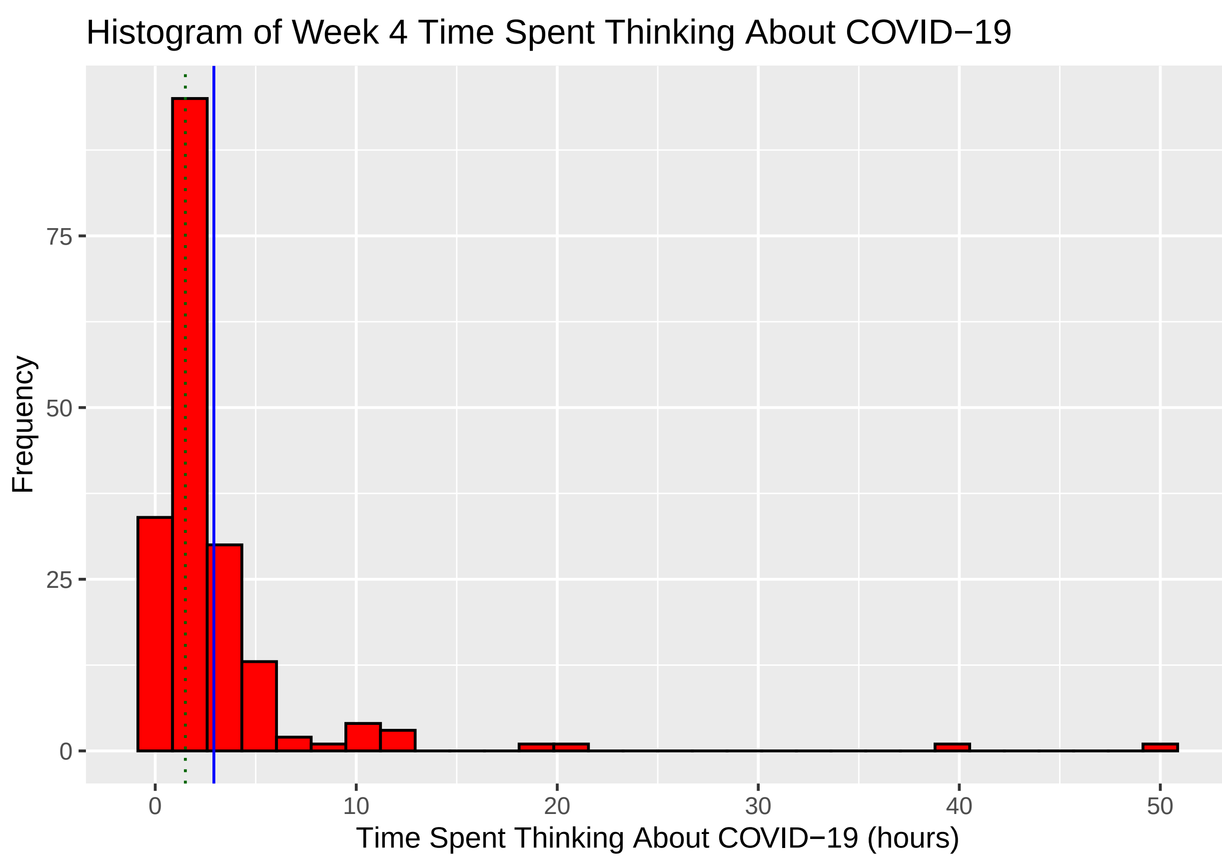

Country vs. Week 2 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
0
10
20
30
40
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

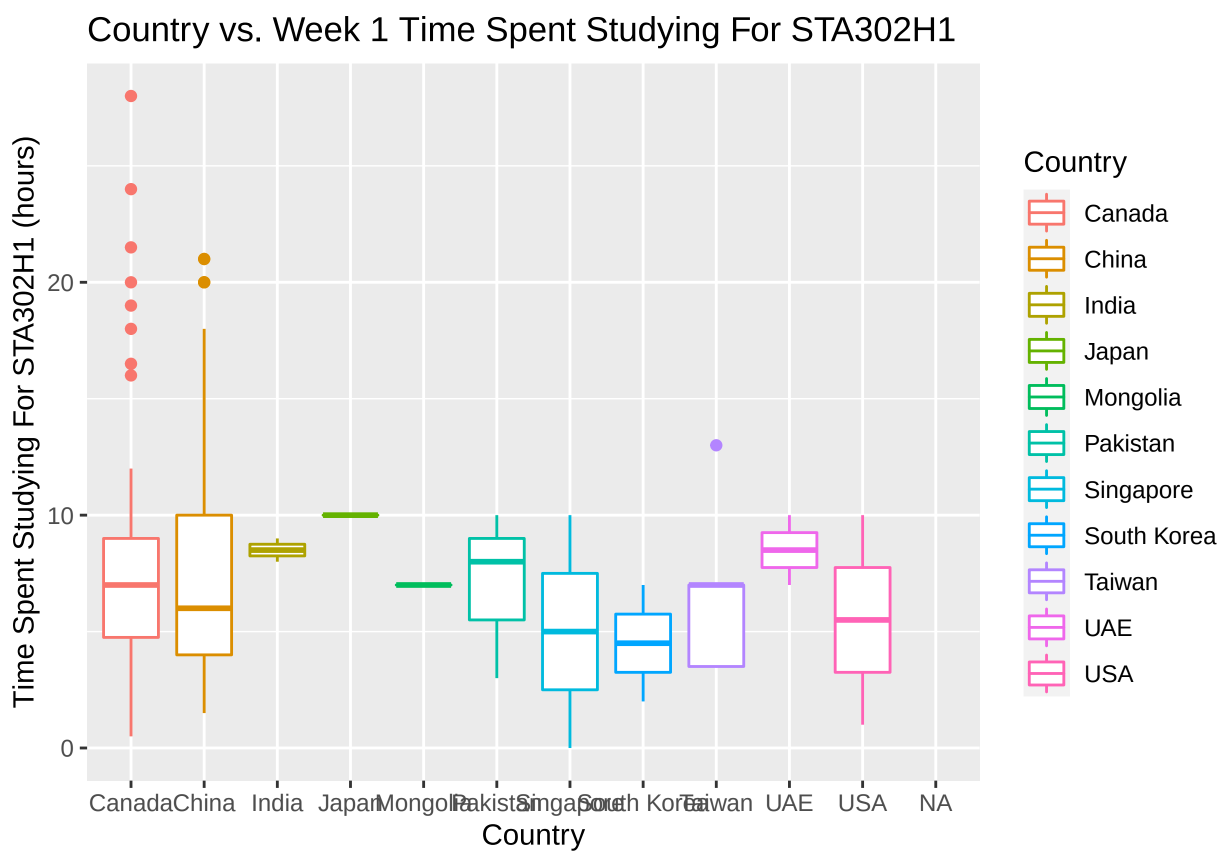

Histogram of Week 2 Time Spent Thinking About COVID−19
Frequency
0
25
50
75
100
0
10
20
30
40
Time Spent Thinking About COVID−19 (hours)

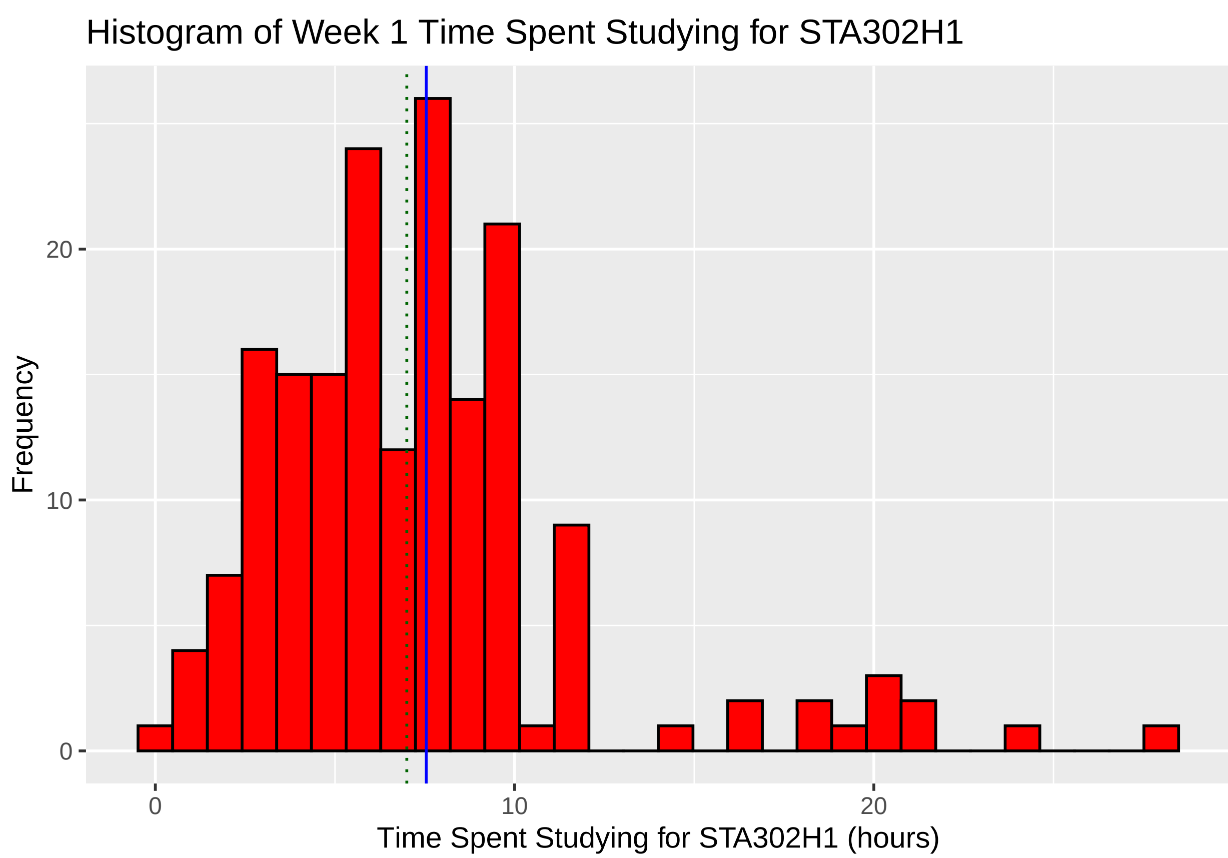

Country vs. Week 3 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
25
20
15
10
5
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

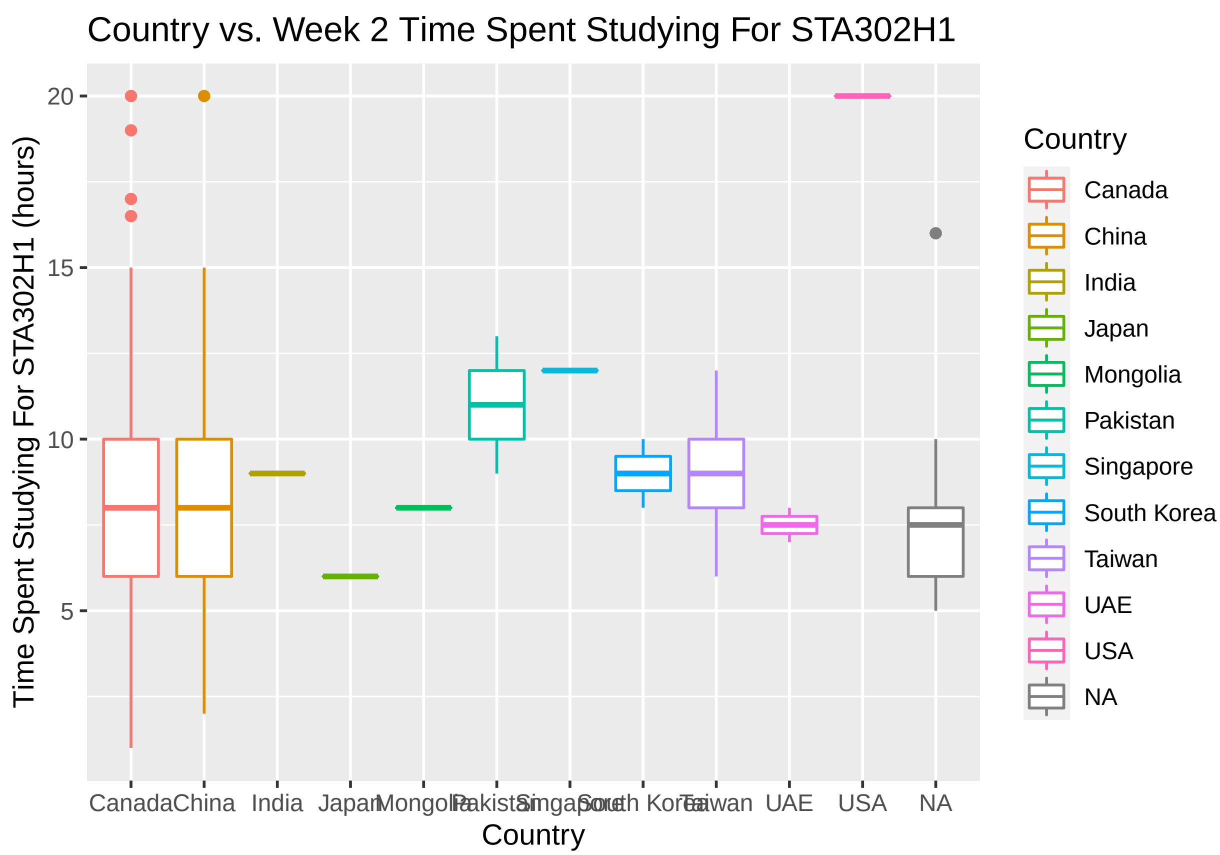

Histogram of Week 3 Time Spent Thinking About COVID−19
Frequency
80
60
40
20
0
0
5
10
15
20
25
Time Spent Thinking About COVID−19 (hours)

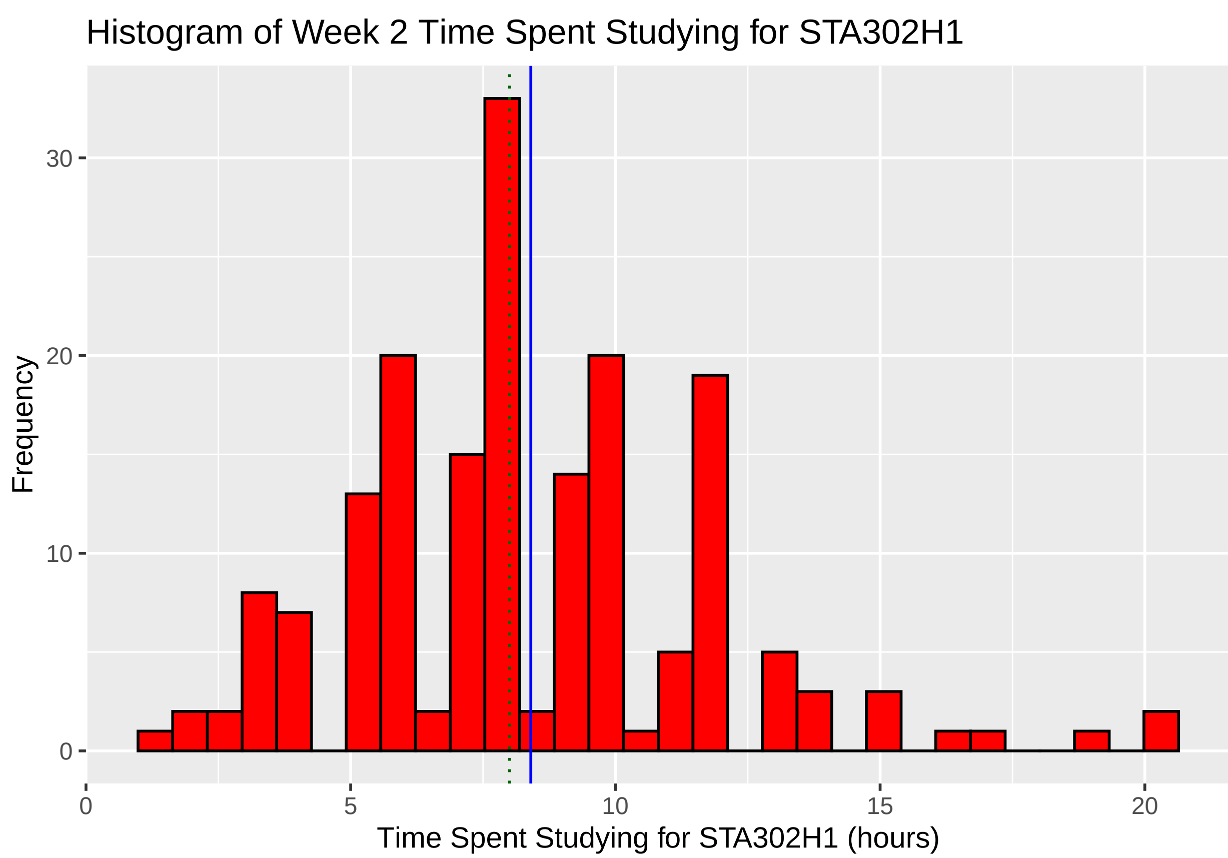

Country vs. Week 4 Time Spent Thinking About COVID−19
Time Spent Thinking About COVID−19 (hours)
0
10
20
30
40
50
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

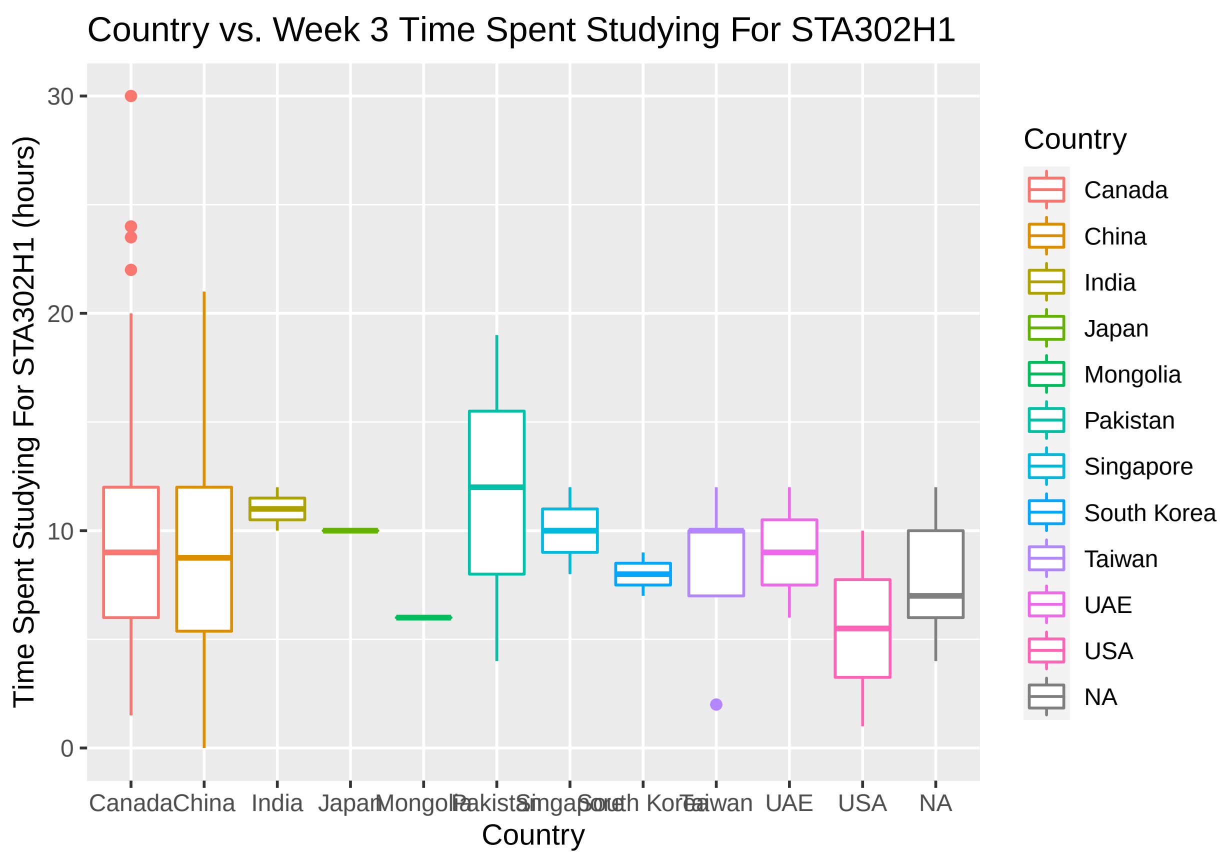

Histogram of Week 4 Time Spent Thinking About COVID−19
Frequency
0
25
50
75
0
10
20
30
40
50
Time Spent Thinking About COVID−19 (hours)

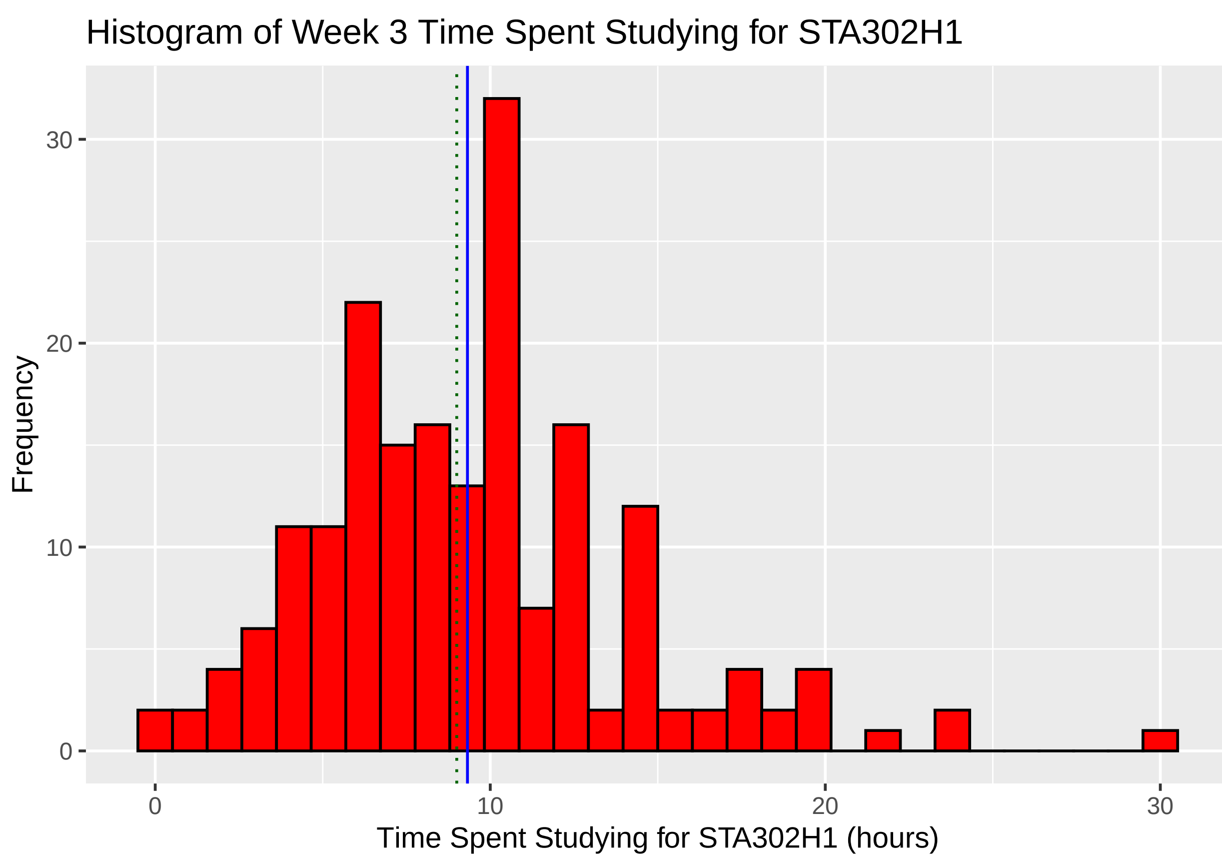

Country vs. Week 1 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
0
10
20
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA

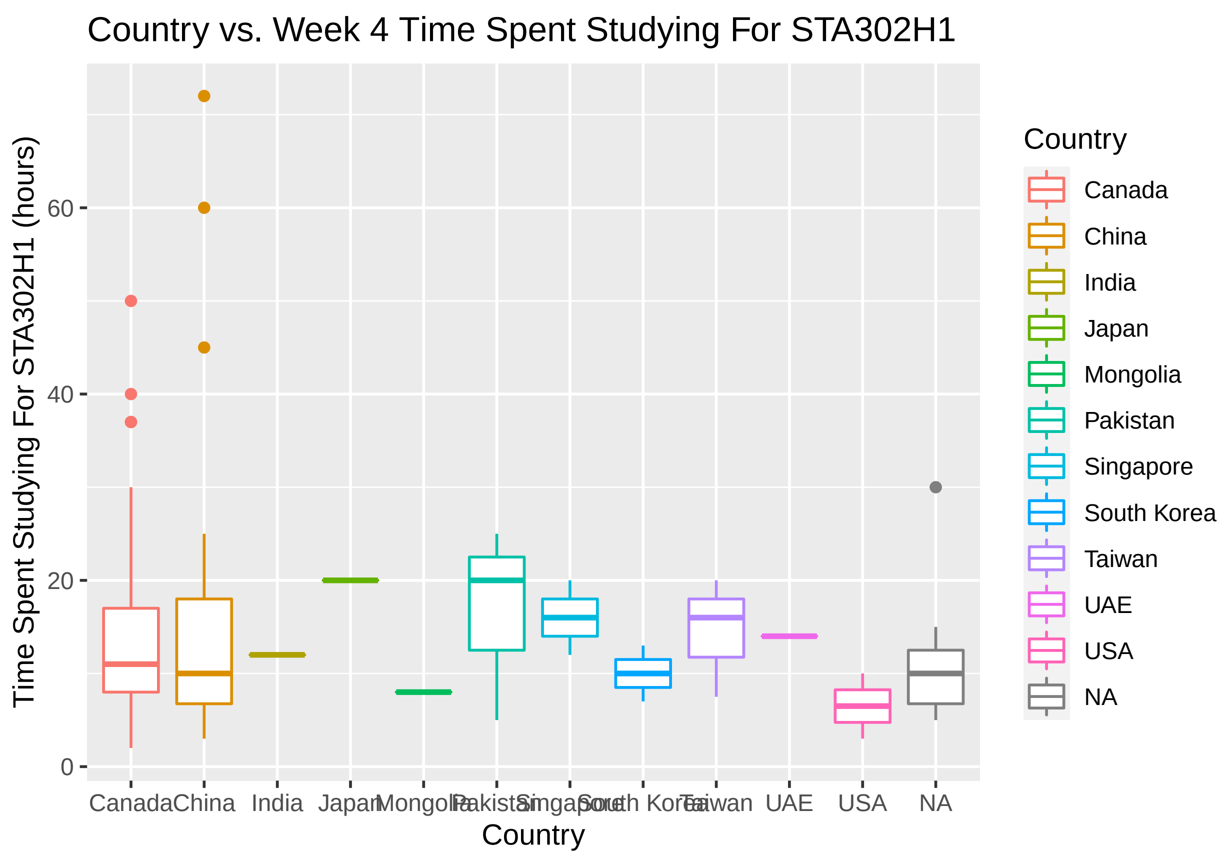

Histogram of Week 1 Time Spent Studying for STA302H1
Frequency
0
10
20
0
10
20
Time Spent Studying for STA302H1 (hours)

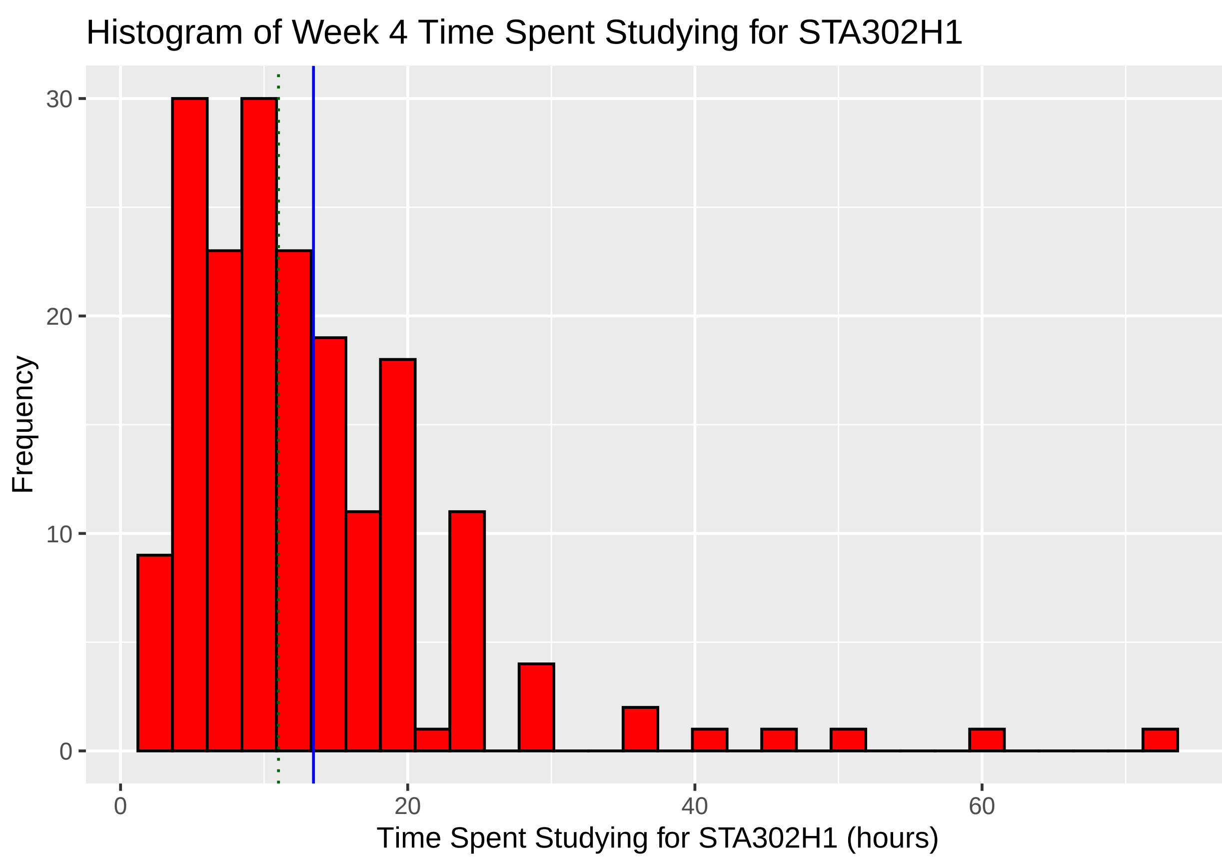

Country vs. Week 2 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
20
15
10
5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

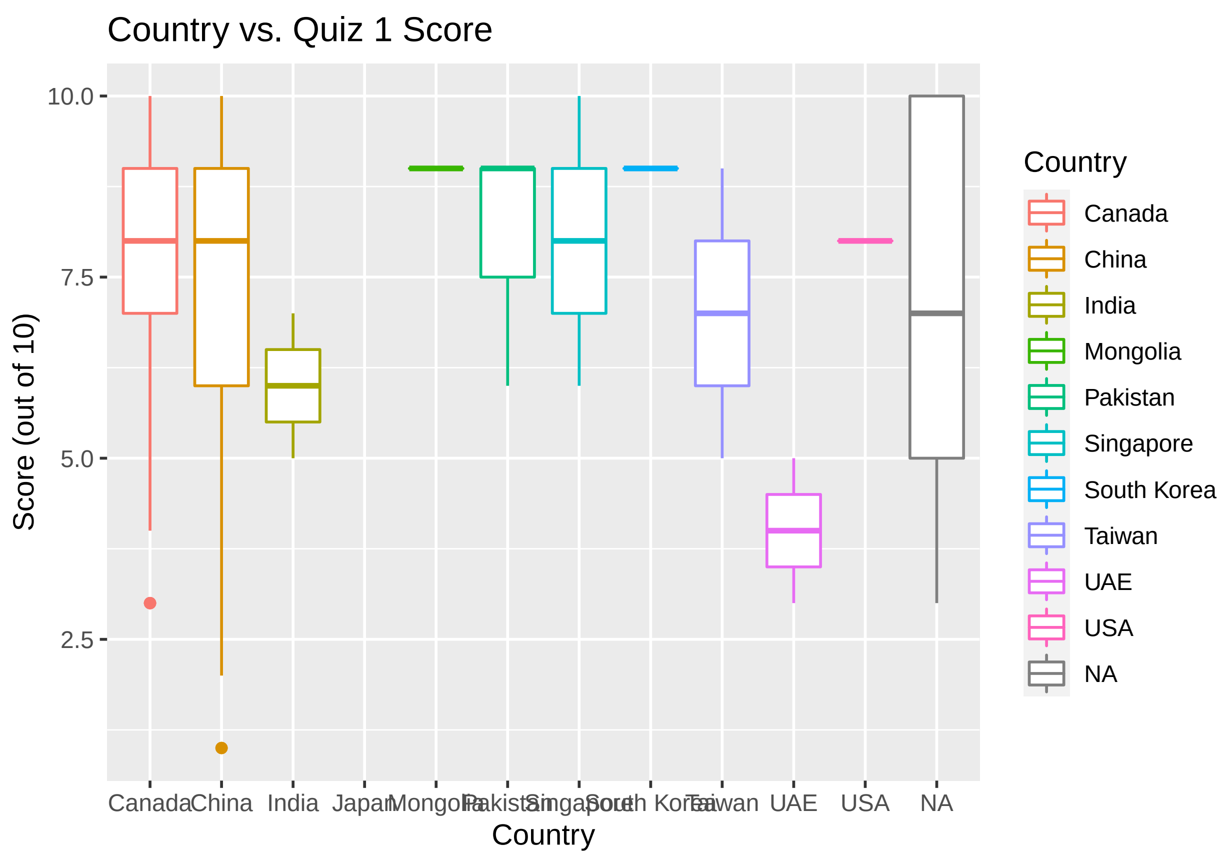

Histogram of Week 2 Time Spent Studying for STA302H1
Frequency
30
20
10
0
0
5
10
15
20
Time Spent Studying for STA302H1 (hours)

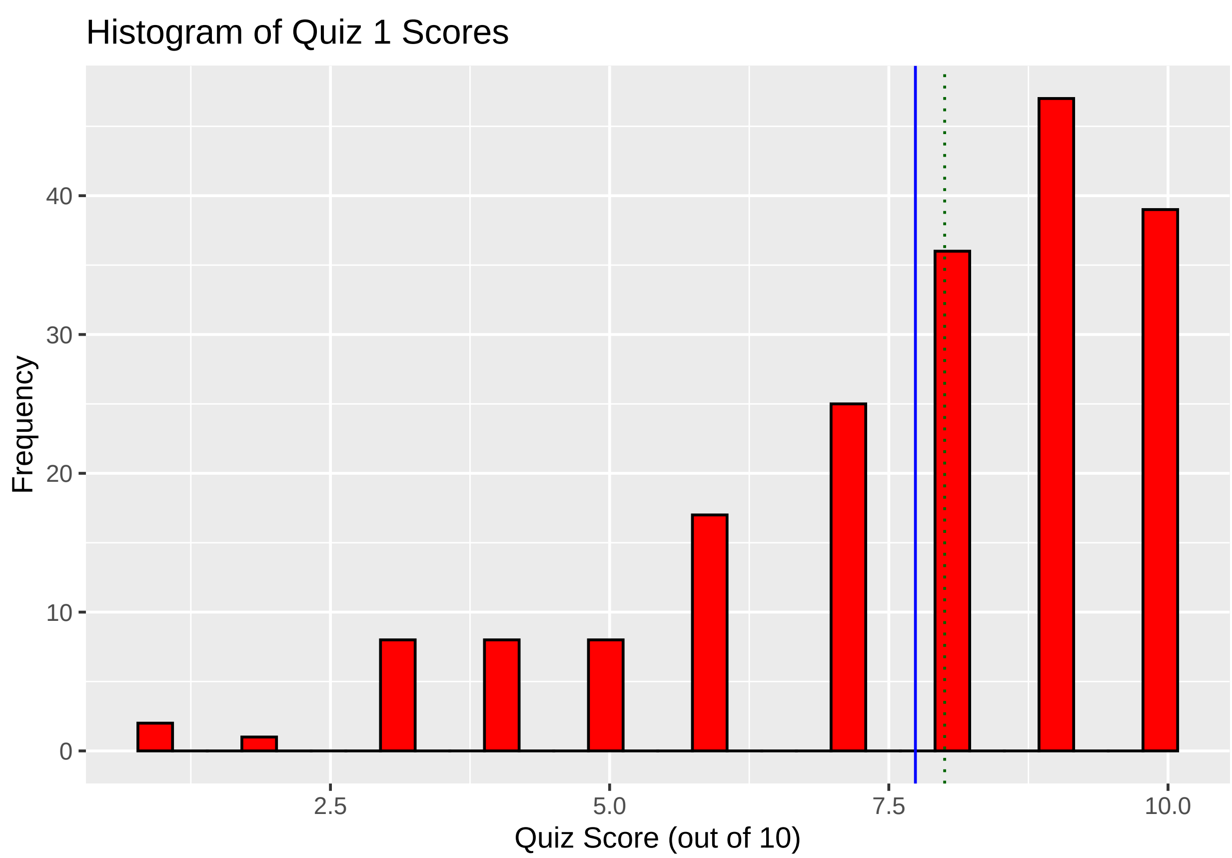

Country vs. Week 3 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
30
20
10
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

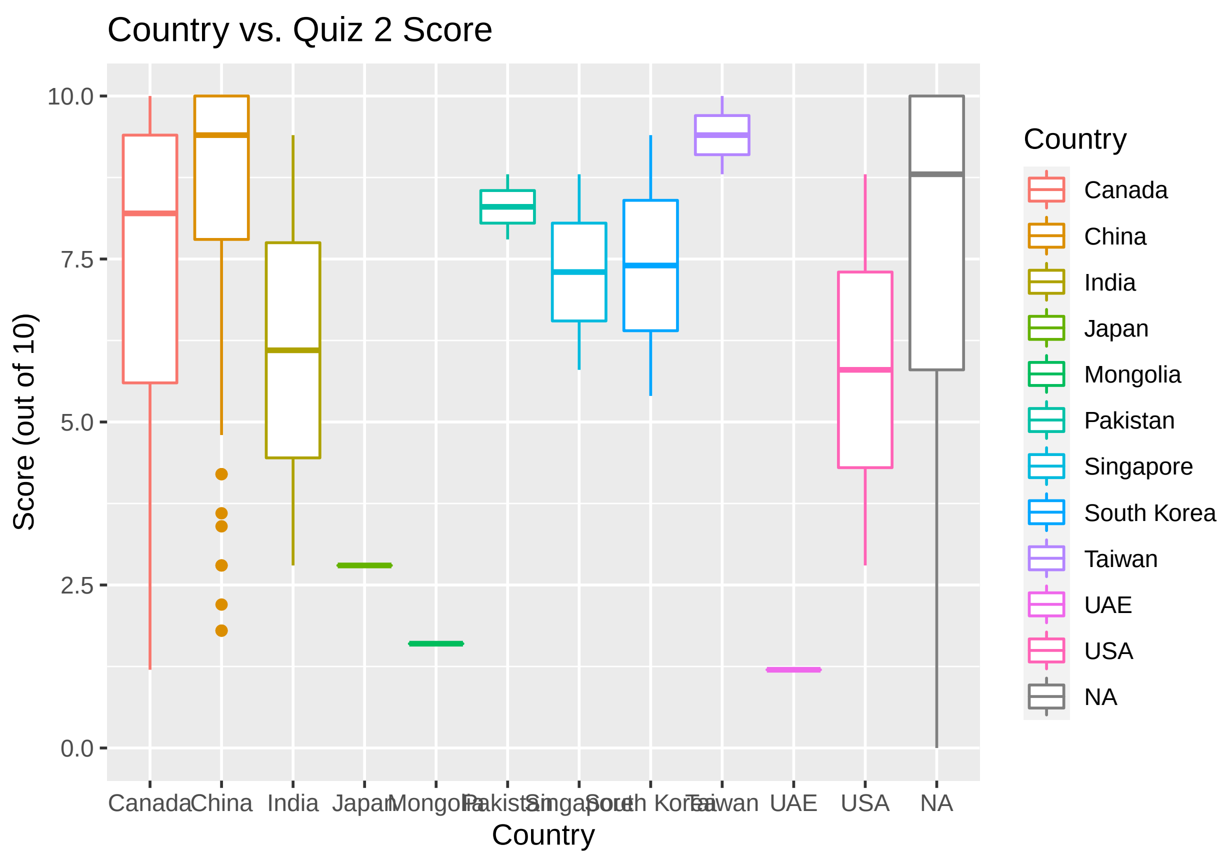

Histogram of Week 3 Time Spent Studying for STA302H1
Frequency
30
20
10
0
0
10
20
30
Time Spent Studying for STA302H1 (hours)

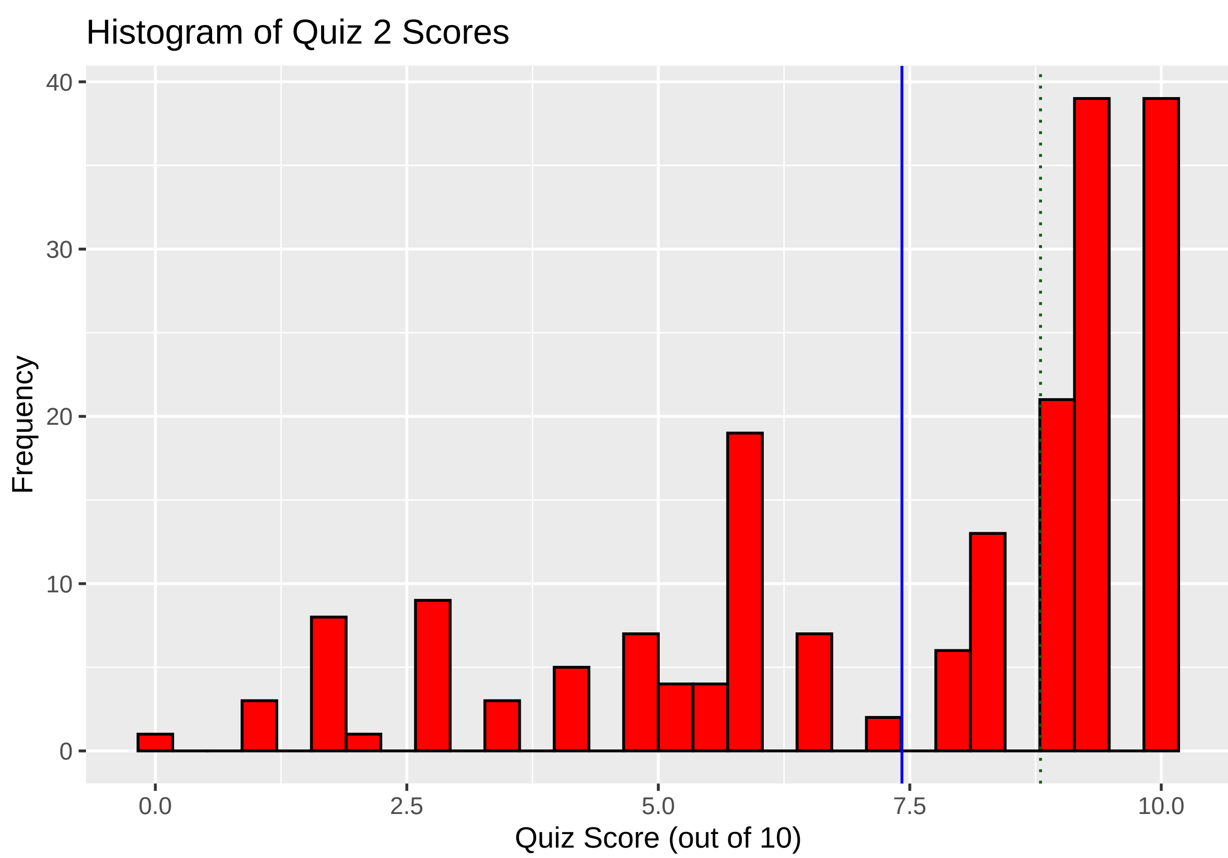

Country vs. Week 4 Time Spent Studying For STA302H1
Time Spent Studying For STA302H1 (hours)
0
20
40
60
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

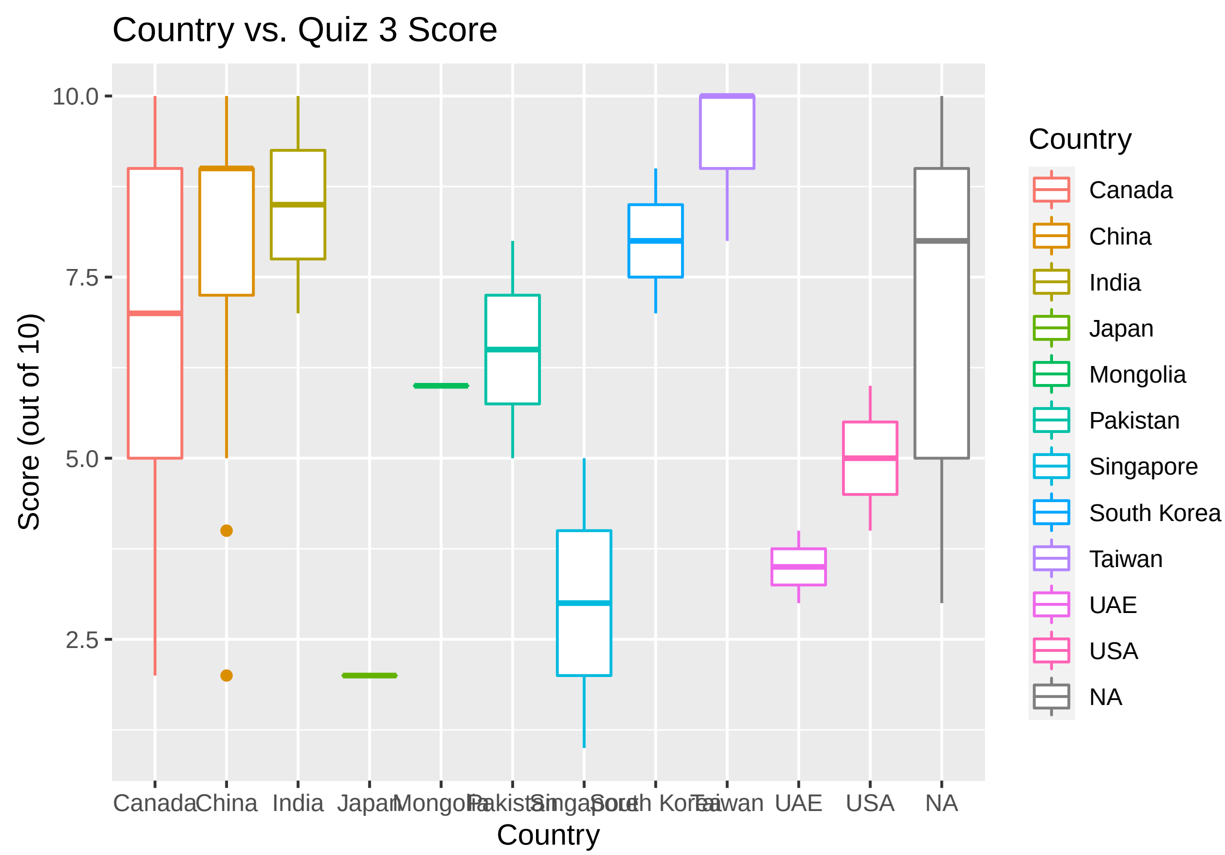

Histogram of Week 4 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
20
40
60
Time Spent Studying for STA302H1 (hours)

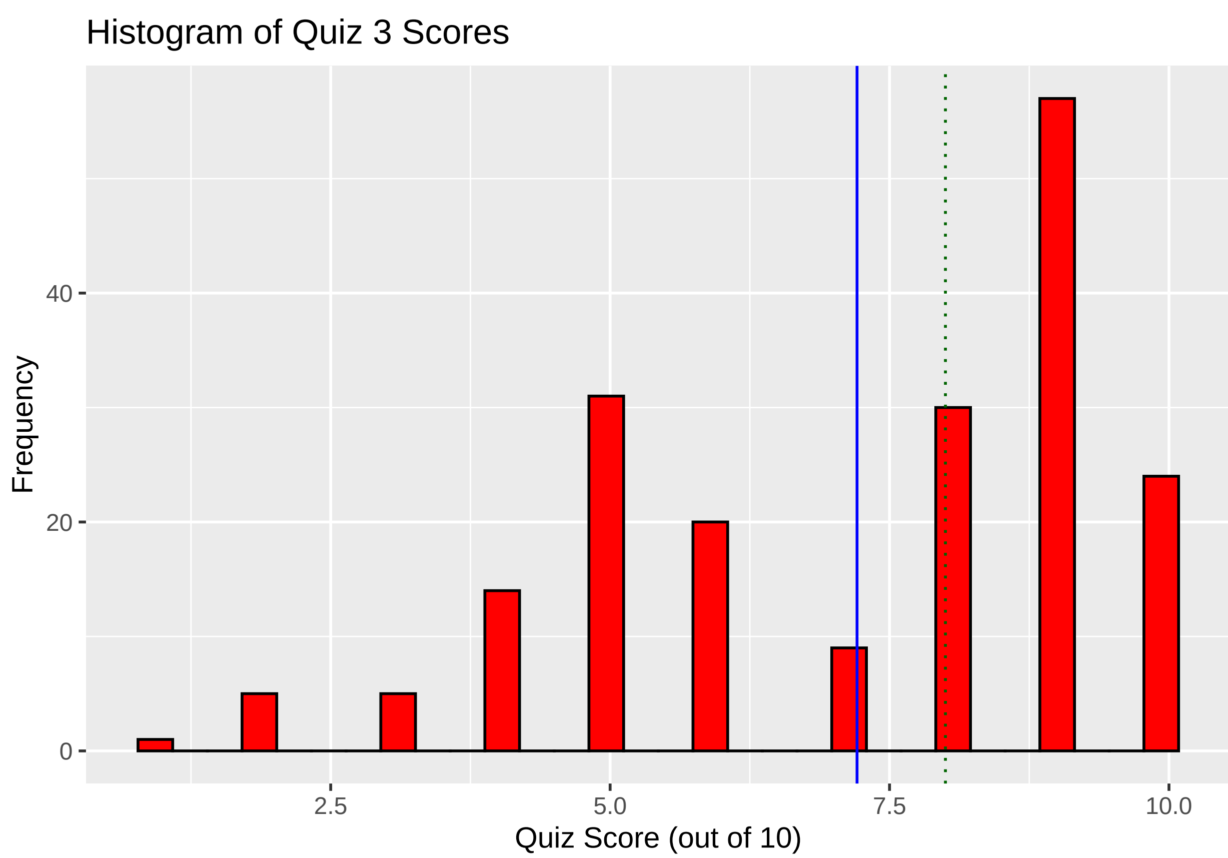

Country vs. Quiz 1 Score
Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA

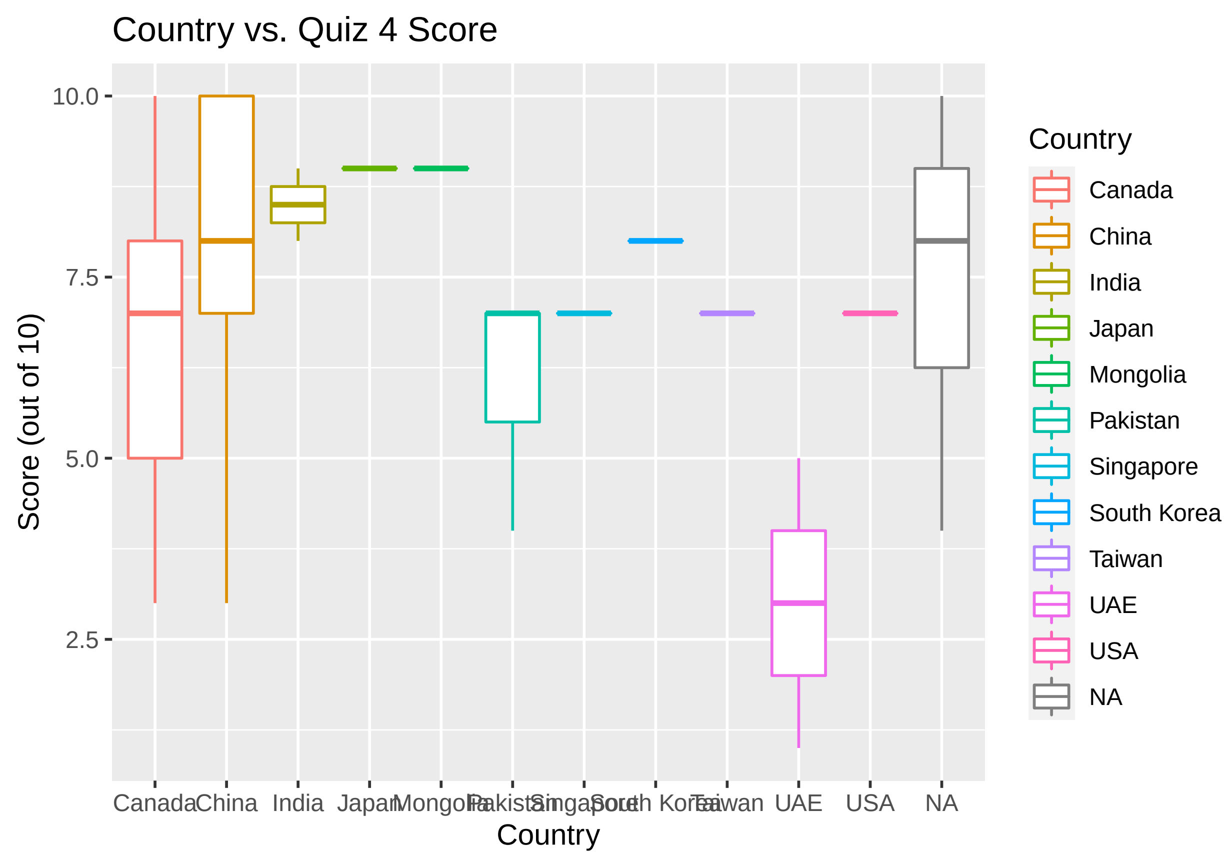

Histogram of Quiz 1 Scores
Frequency
40
30
20
10
0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)

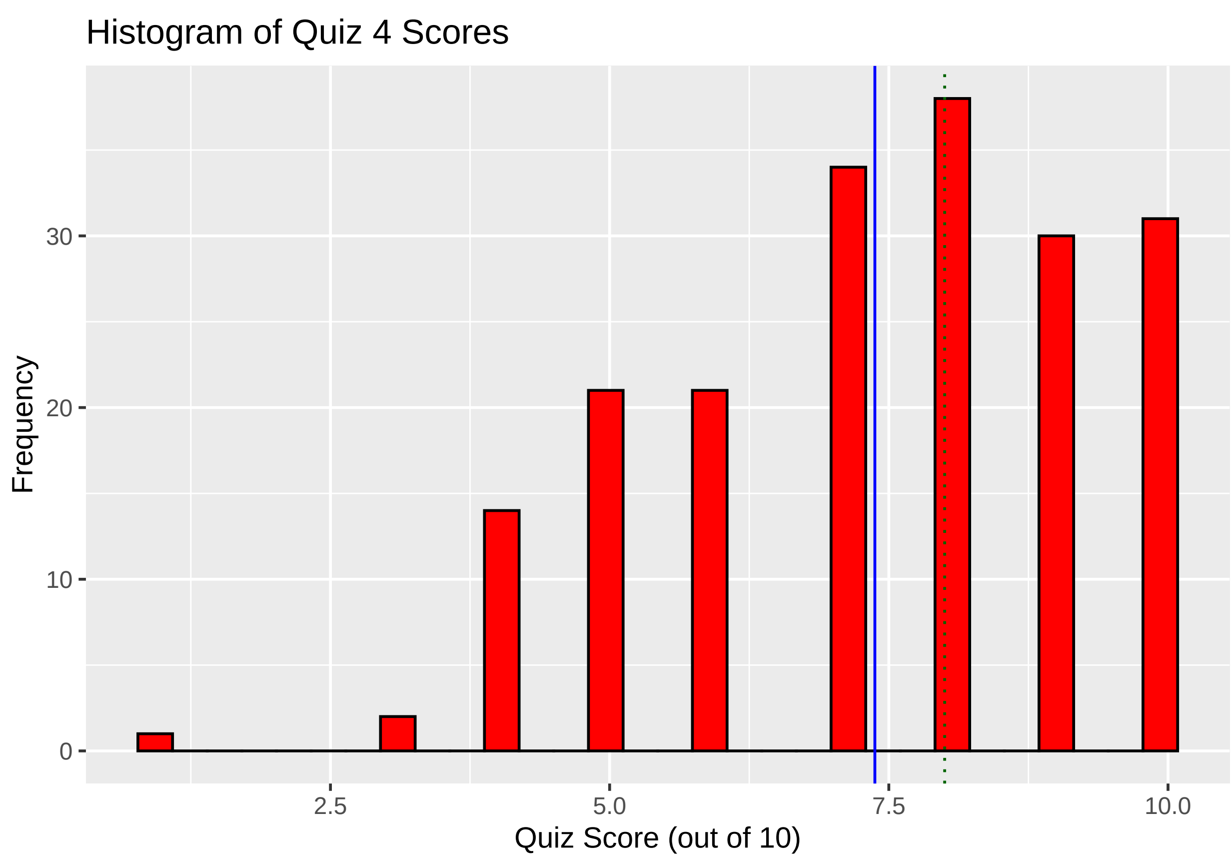

Country vs. Quiz 2 Score
Score (out of 10)
10.0
7.5
5.0
2.5
0.0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country
Country
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA


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
10.0
7.5
5.0
2.5
Score (out of 10)
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
30
20
10
0
Frequency
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
```

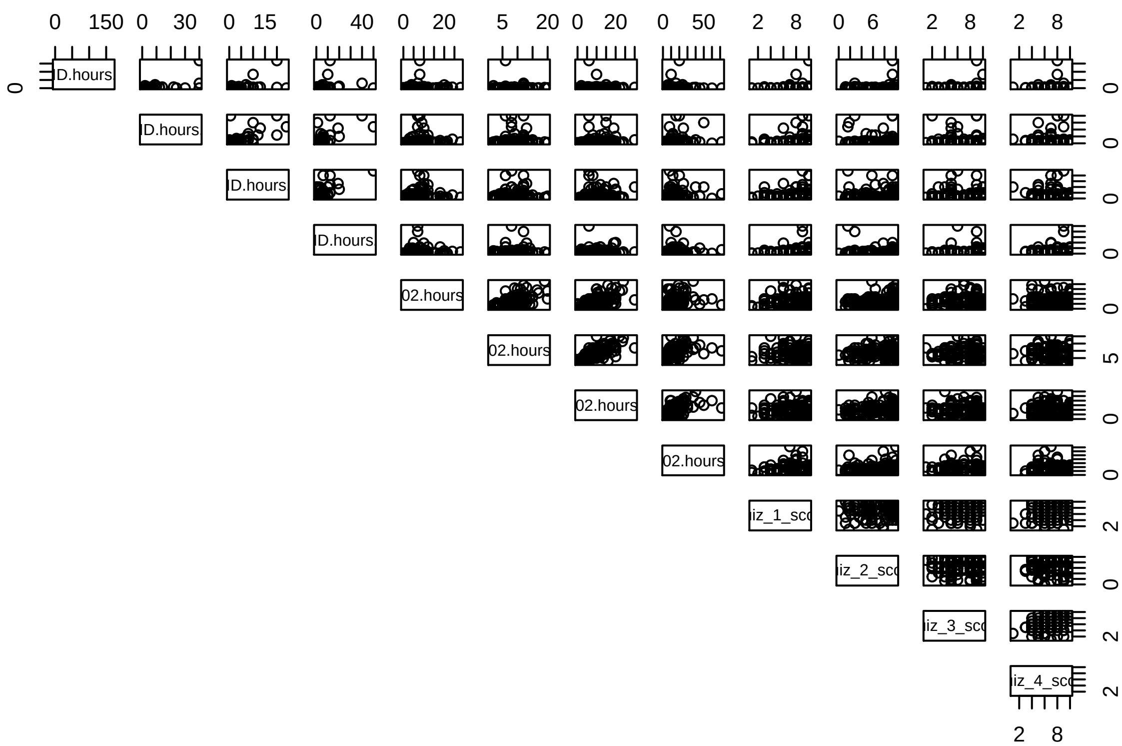
```r
summary(cleaned_sta302_performance_data2$Quiz_4_score)
```

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

Scatterplots

Comprehensive pairwise scatterplot

```r
pairs(~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,
      data = cleaned_sta302_performance_data2, lower.panel = NULL)
```

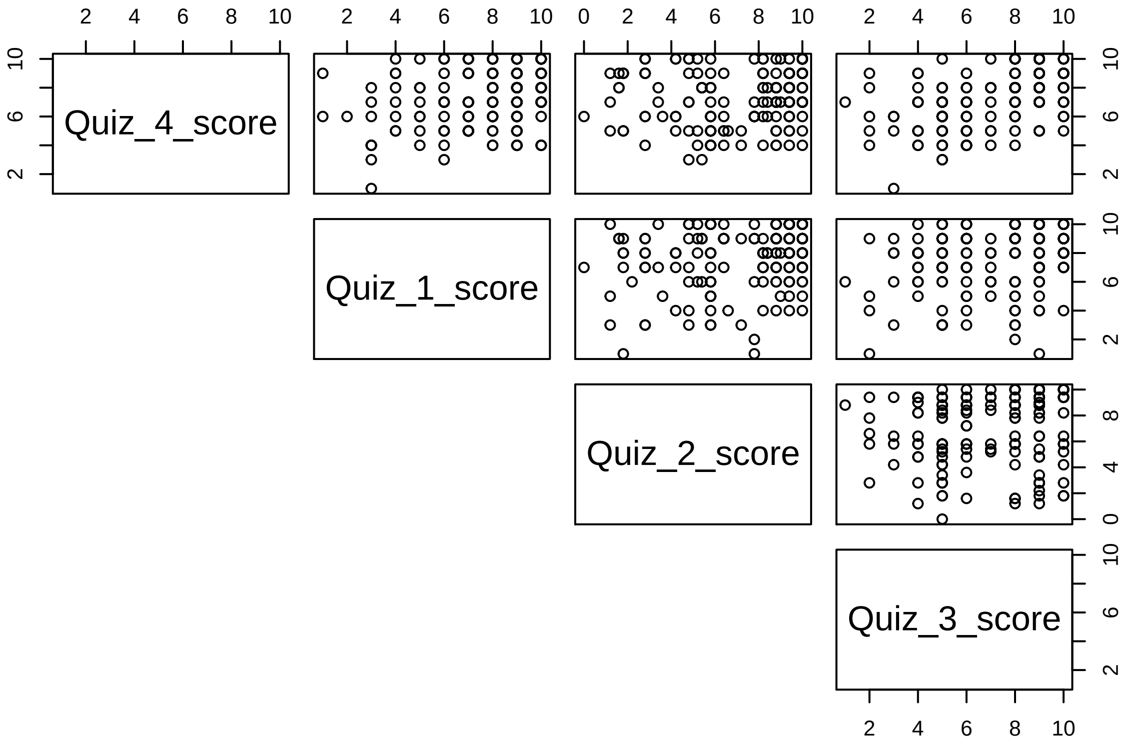


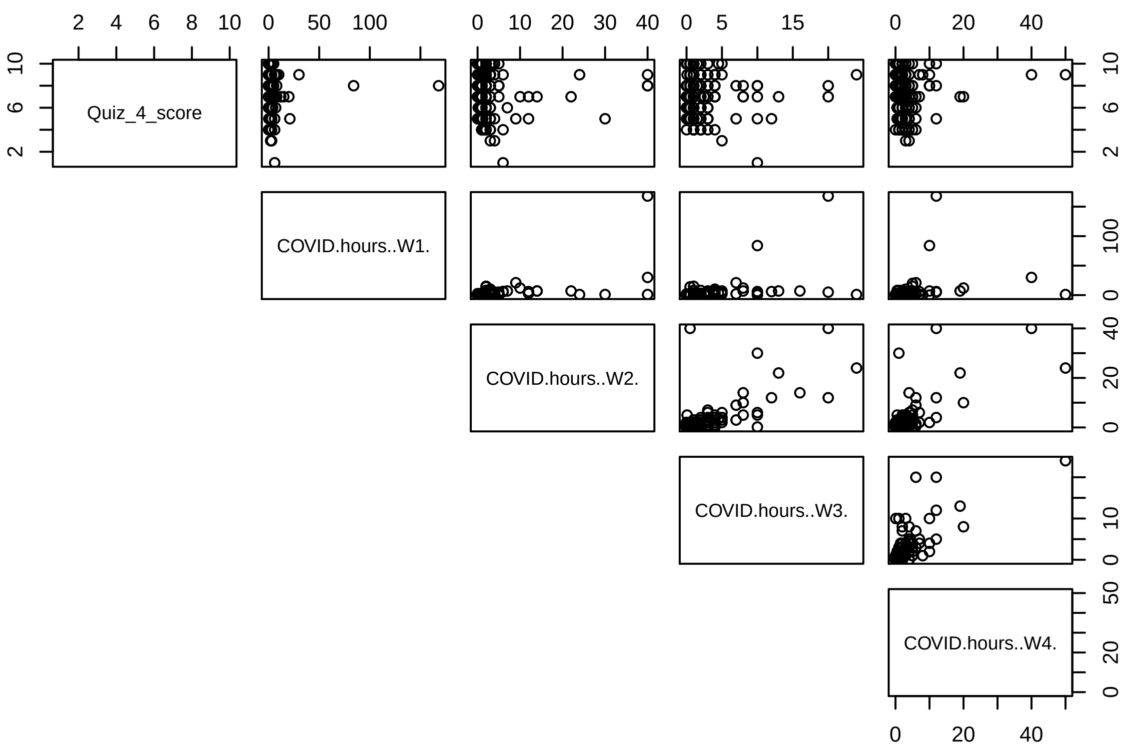

```r
## GGally
# ggpairs -- removes bottom half of pairs plot
# ggpairs(data = cleaned_sta302_performance_data2)
```

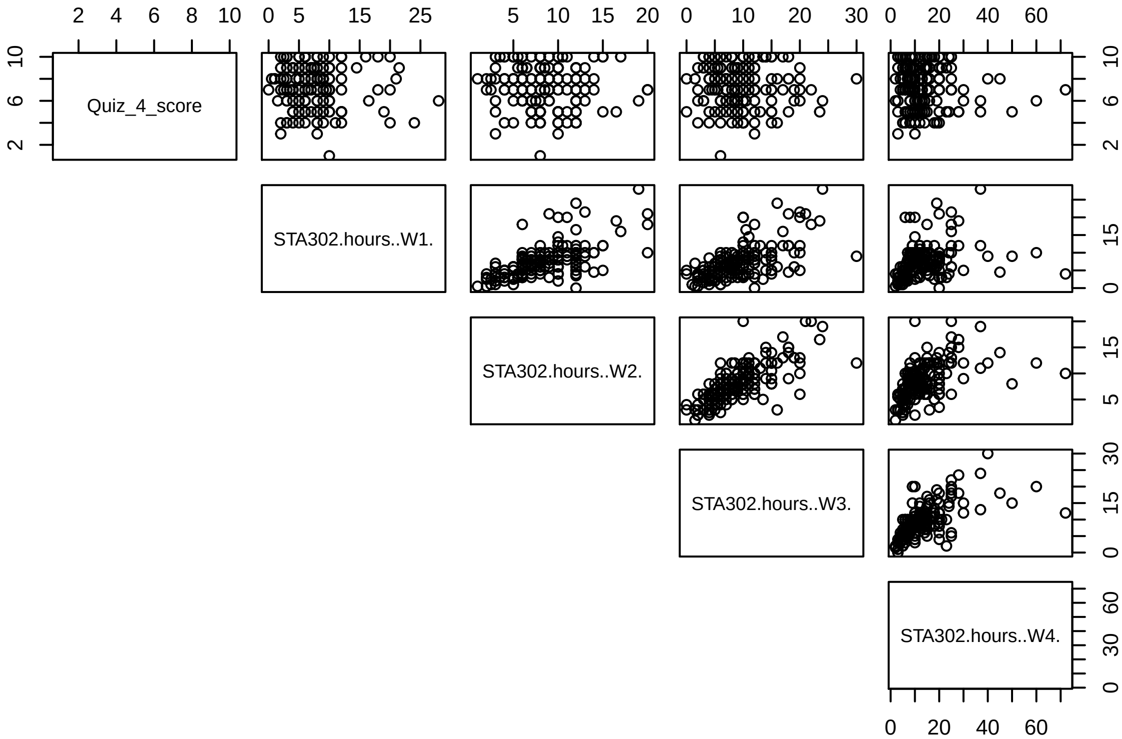
Slightly Zoomed In Pairwise Scatterplots

We can zoom in a bit by creating 3 - 4 pairs() functions:

- quiz4 ~ quiz 1, 2, 3
- quiz4 ~ covid 1, 2, 3, 4
- quiz4 ~ sta302h1 1, 2, 3, 4






Quiz_4_score
STA302.hours..W1.
STA302.hours..W2.
STA302.hours..W3.
STA302.hours..W4.

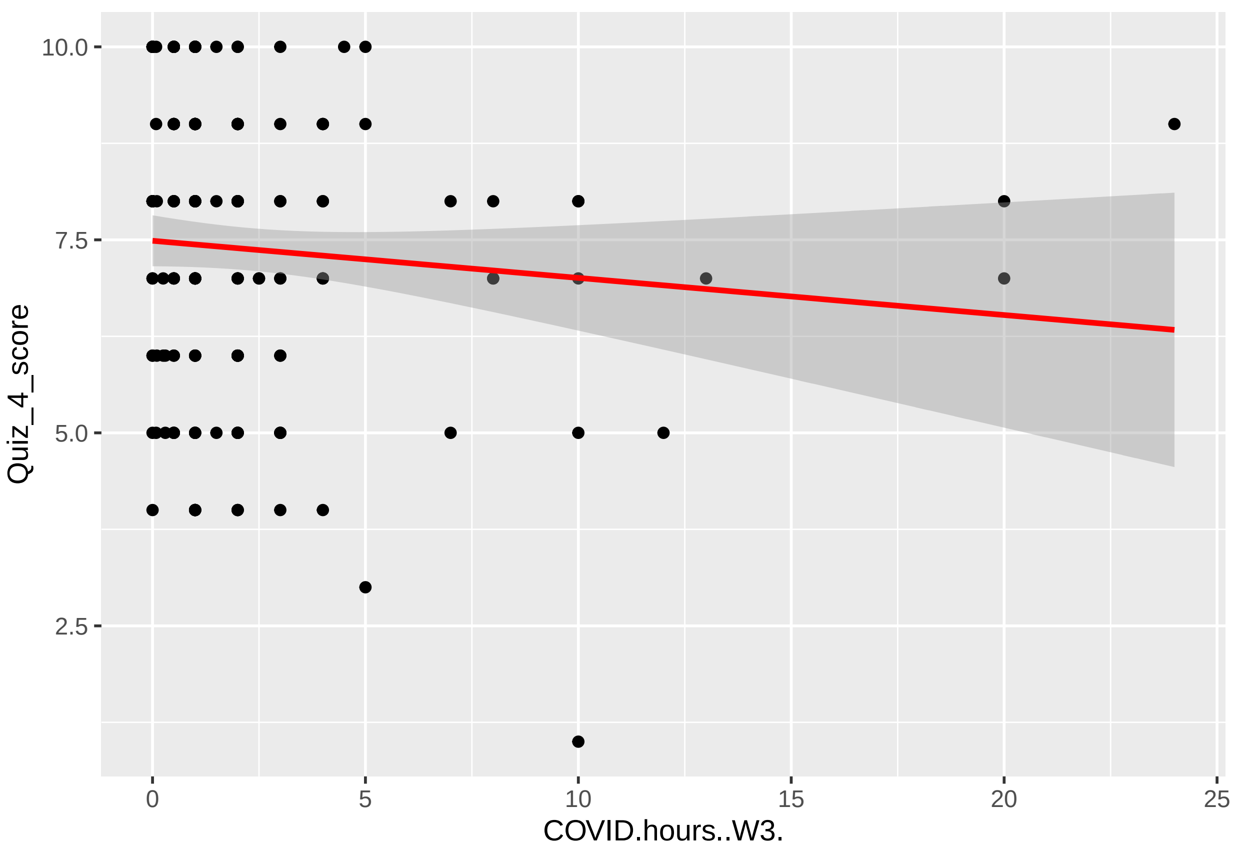
Top 4 - 5 Interesting Scatterplots

On page 28 “Correlation Matrix,” when I look at the row labelled “Q4” (Quiz 4), I obtained the top 4 - 5 predictor variables based on their correlation:

- Quiz 3 Score ($R = 0.55$), significant
- Quiz 1 Score ($R = 0.29$), significant
- Quiz 2 Score ($R = 0.19$), moderately significant
- Week 2 STA302H1 Hours ($R = -0.11$), not significant at all
- Week 3 COVID Hours ($R = -0.09$), not significant at all

Here are all of their scatterplots, linear regressions, and p-values.

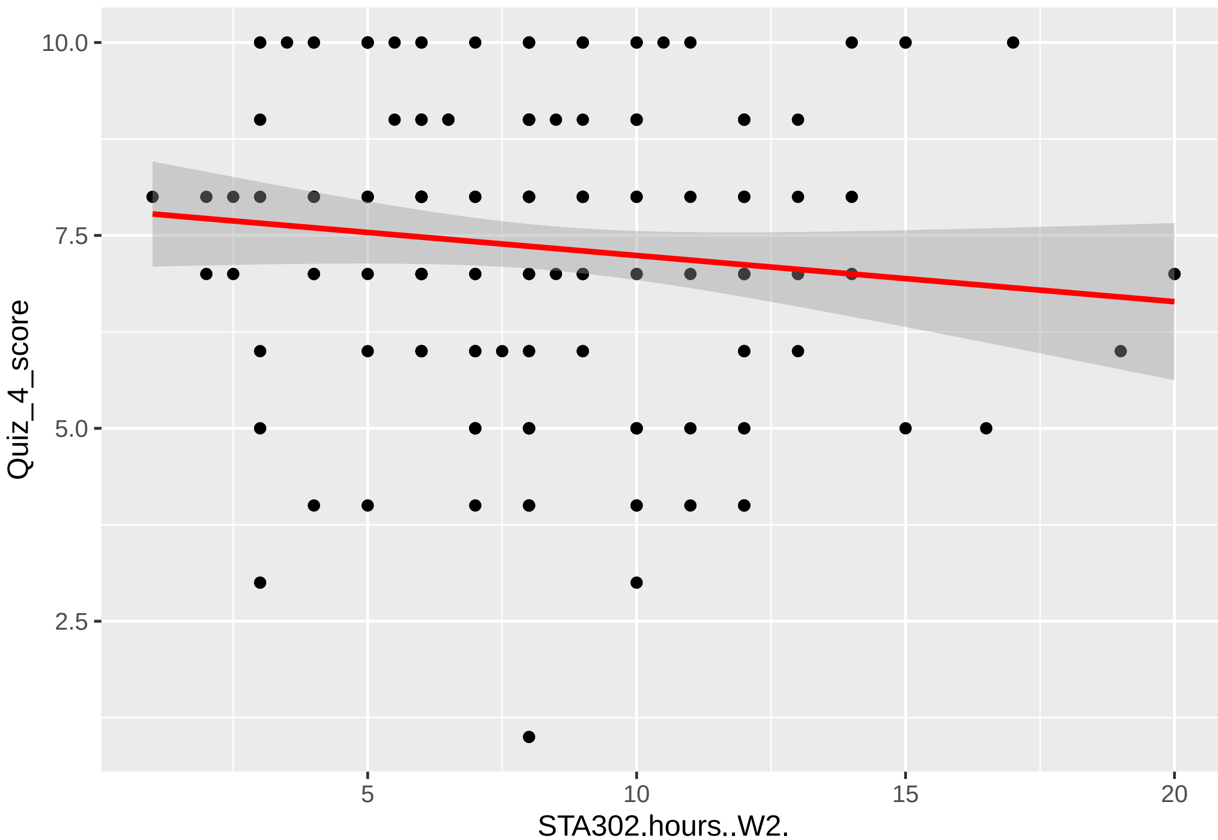
```r
ggplot(data = cleaned_sta302_performance_data2, mapping = aes(x = COVID.hours..W3., y = Quiz_4_score)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", formula = y ~ x)
```




```r
summary(lm(formula = Quiz_4_score ~ COVID.hours..W3.))$coefficients
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       7.4878594 0.16726635 44.766083 1.713615e-99
## COVID.hours..W3. -0.0480875 0.04092781 -1.174935 2.415728e-01
```

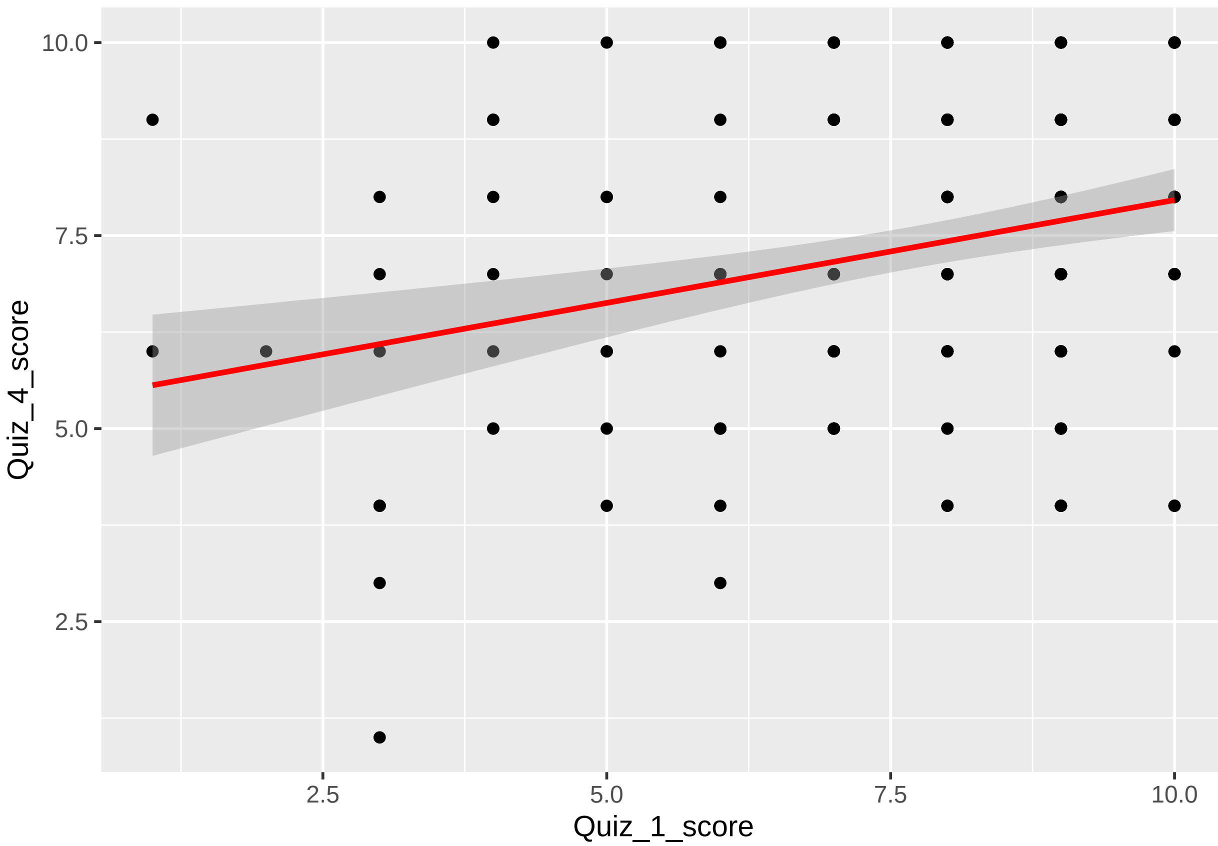
```r
ggplot(data = cleaned_sta302_performance_data2, mapping = aes(x = STA302.hours..W2., y = Quiz_4_score))
  geom_point() +
  geom_smooth(method = "lm", col = "red", formula = y ~ x)
```




```r
summary(lm(formula = Quiz_4_score ~ STA302.hours..W2.))$coefficients
```

```
##                      Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        7.83737079 0.38459115 20.378448 7.564391e-48
## STA302.hours..W2. -0.05979501 0.04256155 -1.404907 1.618420e-01
```

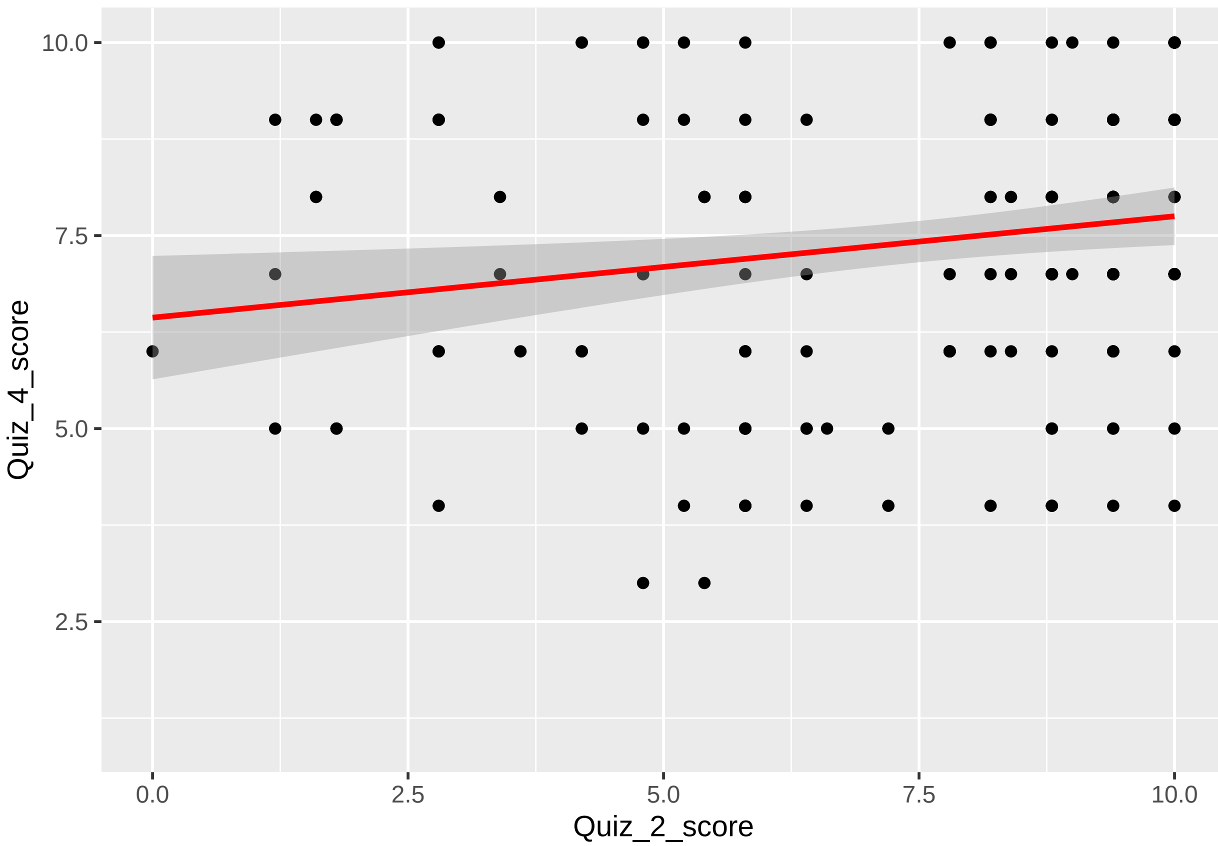
```r
ggplot(data = cleaned_sta302_performance_data2, mapping = aes(x = Quiz_1_score, y = Quiz_4_score)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", formula = y ~ x)
```




```r
summary(lm(formula = Quiz_4_score ~ Quiz_1_score))$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  5.294618 0.52656215 10.05507 3.389243e-19
## Quiz_1_score 0.266566 0.06581747  4.05008 7.569858e-05
```

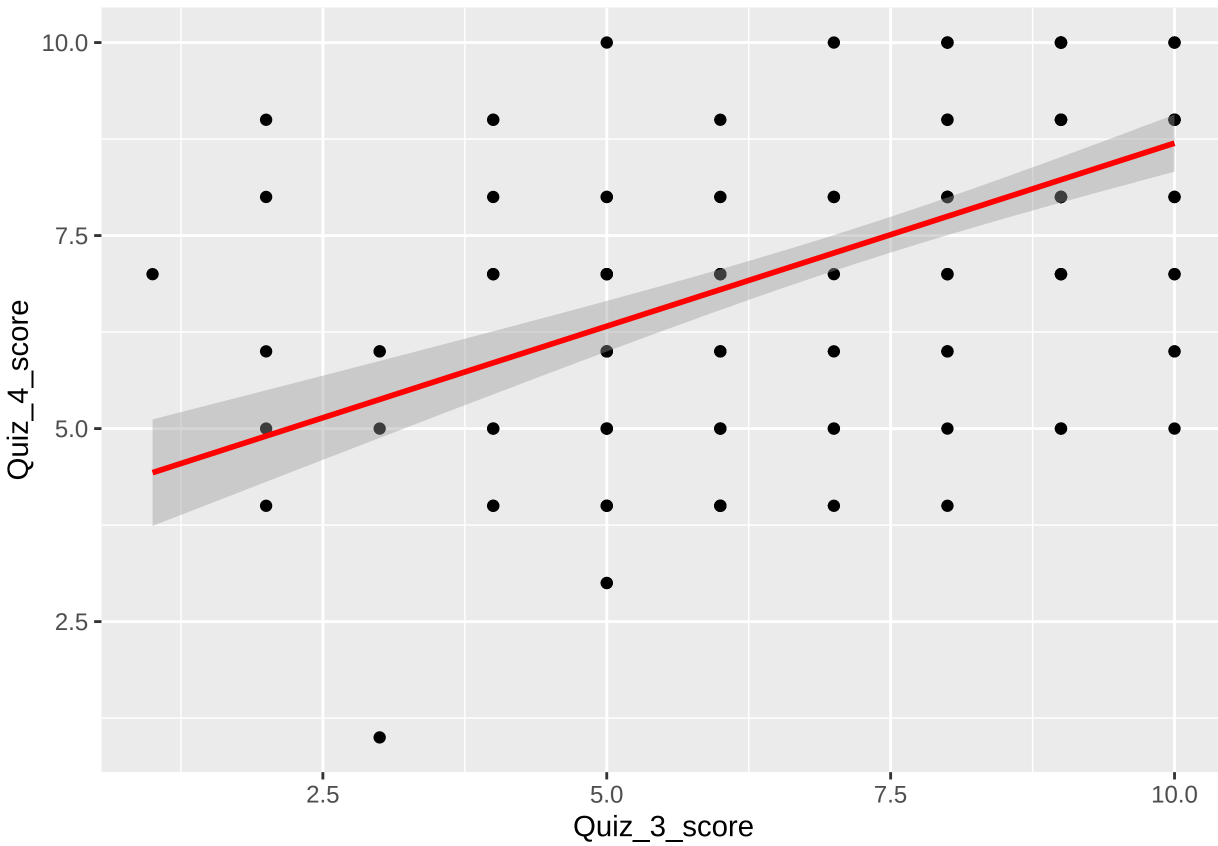
```r
ggplot(data = cleaned_sta302_performance_data2, mapping = aes(x = Quiz_2_score, y = Quiz_4_score)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", formula = y ~ x)
```




```r
summary(lm(formula = Quiz_4_score ~ Quiz_2_score))$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   6.4369424 0.40505781 15.891417 3.115470e-36
## Quiz_2_score  0.1312884 0.05134437  2.557017 1.137315e-02
```

```r
ggplot(data = cleaned_sta302_performance_data2, mapping = aes(x = Quiz_3_score, y = Quiz_4_score)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", formula = y ~ x)
```




```r
summary(lm(formula = Quiz_4_score ~ Quiz_3_score))$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  3.9548096 0.40042387 9.876558 8.520117e-19
## Quiz_3_score 0.4742257 0.05296413 8.953715 3.324360e-16
```

I'll back up my choices with their correlations (R value).

Correlation Matrix

All Countries

We can find the correlation matrix to determine candidate significant predictor values.

```
##       W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302    Q1    Q2    Q3    Q4
## W1COV  1.00  0.56  0.48  0.27  0.04 -0.03 -0.01  0.04  0.08  0.06  0.07  0.02
## W2COV  0.56  1.00  0.67  0.71  0.05  0.08  0.17  0.19  0.13 -0.10 -0.12 -0.01
## W3COV  0.48  0.67  1.00  0.72  0.08  0.08  0.14  0.13  0.09 -0.07 -0.11 -0.09
## W4COV  0.27  0.71  0.72  1.00  0.02  0.07  0.09  0.07  0.12 -0.10  0.02  0.06
## W1302  0.04  0.05  0.08  0.02  1.00  0.61  0.58  0.30  0.05  0.13 -0.04 -0.08
## W2302 -0.03  0.08  0.08  0.07  0.61  1.00  0.70  0.48  0.00  0.06 -0.05 -0.11
## W3302 -0.01  0.17  0.14  0.09  0.58  0.70  1.00  0.62 -0.01  0.08 -0.12 -0.08
## W4302  0.04  0.19  0.13  0.07  0.30  0.48  0.62  1.00 -0.01  0.04 -0.05 -0.06
## Q1     0.08  0.13  0.09  0.12  0.05  0.00 -0.01 -0.01  1.00  0.25  0.29  0.29
## Q2     0.06 -0.10 -0.07 -0.10  0.13  0.06  0.08  0.04  0.25  1.00  0.23  0.19
## Q3     0.07 -0.12 -0.11  0.02 -0.04 -0.05 -0.12 -0.05  0.29  0.23  1.00  0.55
## Q4     0.02 -0.01 -0.09  0.06 -0.08 -0.11 -0.08 -0.06  0.29  0.19  0.55  1.00
```

By Individual Country

We could also create separate correlation matrices for each country.

```
display_correlation_by_country(canada)
```

```
##       W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302    Q1    Q2    Q3    Q4
## W1COV  1.00  0.60  0.72  0.39 -0.01 -0.05 -0.05  0.10  0.10  0.06  0.11  0.06
## W2COV  0.60  1.00  0.67  0.72  0.00  0.02  0.15  0.29  0.13 -0.07 -0.06  0.05
## W3COV  0.72  0.67  1.00  0.61 -0.01 -0.01  0.23  0.31  0.09  0.03 -0.08 -0.07
## W4COV  0.39  0.72  0.61  1.00 -0.04  0.04  0.16  0.02  0.10 -0.04  0.10  0.04
## W1302 -0.01  0.00 -0.01 -0.04  1.00  0.70  0.64  0.55 -0.09  0.09 -0.02 -0.16
## W2302 -0.05  0.02 -0.01  0.04  0.70  1.00  0.73  0.62 -0.15 -0.09 -0.01 -0.11
## W3302 -0.05  0.15  0.23  0.16  0.64  0.73  1.00  0.68 -0.09  0.05 -0.13 -0.12
## W4302  0.10  0.29  0.31  0.02  0.55  0.62  0.68  1.00 -0.12 -0.08 -0.09 -0.12
## Q1     0.10  0.13  0.09  0.10 -0.09 -0.15 -0.09 -0.12  1.00  0.16  0.38  0.35
## Q2     0.06 -0.07  0.03 -0.04  0.09 -0.09  0.05 -0.08  0.16  1.00  0.07  0.14
## Q3     0.11 -0.06 -0.08  0.10 -0.02 -0.01 -0.13 -0.09  0.38  0.07  1.00  0.50
## Q4     0.06  0.05 -0.07  0.04 -0.16 -0.11 -0.12 -0.12  0.35  0.14  0.50  1.00
```

```
display_correlation_by_country(china)
```

```
##       W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302   Q1    Q2    Q3    Q4
## W1COV  1.00  0.70  0.51  0.40  0.12  0.08  0.05 -0.07 0.06  0.09  0.10 -0.01
## W2COV  0.70  1.00  0.67  0.67  0.25  0.34  0.27  0.08 0.14 -0.05  0.05 -0.12
## W3COV  0.51  0.67  1.00  0.62  0.15  0.28  0.19  0.04 0.14  0.04  0.20 -0.19
## W4COV  0.40  0.67  0.62  1.00  0.13  0.25  0.19  0.28 0.24  0.07  0.08 -0.10
## W1302  0.12  0.25  0.15  0.13  1.00  0.56  0.49  0.02 0.21  0.28 -0.06  0.07
## W2302  0.08  0.34  0.28  0.25  0.56  1.00  0.76  0.42 0.15  0.27 -0.02 -0.08
## W3302  0.05  0.27  0.19  0.19  0.49  0.76  1.00  0.57 0.11  0.14 -0.07  0.00
## W4302 -0.07  0.08  0.04  0.28  0.02  0.42  0.57  1.00 0.07  0.15  0.04 -0.03
## Q1     0.06  0.14  0.14  0.24  0.21  0.15  0.11  0.07 1.00  0.44  0.36  0.25
## Q2     0.09 -0.05  0.04  0.07  0.28  0.27  0.14  0.15 0.44  1.00  0.19  0.15
## Q3     0.10  0.05  0.20  0.08 -0.06 -0.02 -0.07  0.04 0.36  0.19  1.00  0.61
## Q4    -0.01 -0.12 -0.19 -0.10  0.07 -0.08  0.00 -0.03 0.25  0.15  0.61  1.00
```

```
display_correlation_by_country(other)
```

```
##       W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302    Q1    Q2    Q3    Q4
## W1COV  1.00  0.33  0.37  0.10  0.43 -0.20  0.29  0.60 -0.26 -0.11 -0.54 -0.15
## W2COV  0.33  1.00  0.90  0.87  0.29  0.17  0.24  0.18  0.14 -0.19 -0.24  0.03
## W3COV  0.37  0.90  1.00  0.82  0.36  0.09  0.03  0.10  0.12 -0.21 -0.26 -0.02
## W4COV  0.10  0.87  0.82  1.00  0.18  0.04  0.04  0.01  0.12 -0.25 -0.10  0.22
## W1302  0.43  0.29  0.36  0.18  1.00  0.01  0.40  0.35 -0.16 -0.16  0.02  0.00
## W2302 -0.20  0.17  0.09  0.04  0.01  1.00  0.34  0.19  0.09  0.23 -0.02 -0.01
## W3302  0.29  0.24  0.03  0.04  0.40  0.34  1.00  0.72 -0.21  0.15 -0.06  0.08
## W4302  0.60  0.18  0.10  0.01  0.35  0.19  0.72  1.00 -0.05  0.10 -0.28  0.02
## Q1    -0.26  0.14  0.12  0.12 -0.16  0.09 -0.21 -0.05  1.00  0.26  0.29  0.45
## Q2    -0.11 -0.19 -0.21 -0.25 -0.16  0.23  0.15  0.10  0.26  1.00  0.49  0.24
## Q3    -0.54 -0.24 -0.26 -0.10  0.02 -0.02 -0.06 -0.28  0.29  0.49  1.00  0.49
## Q4    -0.15  0.03 -0.02  0.22  0.00 -0.01  0.08  0.02  0.45  0.24  0.49  1.00
```

Select Predictor Variables Based on Criterion

Here are our tentative models:

- quiz4 ~ quiz 1, 2, 3 grades (out of 10)
- quiz4 ~ weeks 1, 2, 3, 4 covid19 hours
- quiz4 ~ weeks 1, 2, 3, 4 sta302h1 hours

This R code snippet below produces the model quiz4 = quiz3. This model seems to under-fit though.

```
quiz_grades.fit = lm(quiz4 ~ quiz1 + quiz2 + quiz3)
step = stepAIC(quiz_grades.fit, direction = "both")
step$anova
```

```
summary(quiz_grades.fit)$r.squared
summary(quiz_grades.fit)$adj.r.squared
```

This R code snippet below produces the model

quiz4 = quiz1 + I(quiz2 ^ 2) + I(quiz3 ^ 2) + I(quiz1 quiz3) + I(quiz2 quiz3).

This model seems to be a good fit.

```
quiz_grades.fit2 = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + I(quiz1 ^ 2) + I(quiz2 ^ 2) + I(quiz3 ^ 2)
  + I(quiz1 * quiz2) + I(quiz1 * quiz3) + I(quiz2 * quiz3))
step2 = stepAIC(quiz_grades.fit2, direction = "both")
step2$anova
```

```
summary(quiz_grades.fit2)$r.squared
summary(quiz_grades.fit2)$adj.r.squared
```

This R code snippet below produces the model quiz4 = 1. This model seems to under-fit though.

```
study_hours.fit = lm(quiz4 ~ study1 + study2 + study3 + study4)
step3 = stepAIC(study_hours.fit, direction = "both")
step3$anova
```

```
summary(study_hours.fit)$r.squared
summary(study_hours.fit)$adj.r.squared
```

This R code snippet below produces the model quiz4 = 1. This model seems to be a good fit.

```
study_hours.fit2 = lm(
  quiz4 ~ study1 + study2 + study3 + study4 +
    I(study1 ^ 2) + I(study2 ^ 2) + I(study3 ^ 2) + I(study4 ^ 2) +
    I(study1 * study2) + I(study1 * study3) + I(study1 * study4) +
    I(study2 * study3) + I(study2 * study4) + I(study3 * study4))
step4 = stepAIC(study_hours.fit2, direction = "both")
step4$anova
```

```r
summary(study_hours.fit2)$r.squared
summary(study_hours.fit2)$adj.r.squared
```

This R code snippet below produces the model quiz4 = 1. This model seems to under-fit though.

```
covid_hours.fit = lm(quiz4 ~ covid1 + covid2 + covid3 + covid4)
step5 = stepAIC(covid_hours.fit, direction = "both")
step5$anova
```

```
summary(covid_hours.fit)$r.squared
summary(covid_hours.fit)$adj.r.squared
```

This R code snippet below produces the model

quiz4 = covid3 + I(covid3 ˆ 2).

This model seems to be a good fit.

```
covid_hours.fit2 = lm(
  quiz4 ~ covid1 + covid2 + covid3 + covid4 +
    I(covid1 ^ 2) + I(covid2 ^ 2) + I(covid3 ^ 2) + I(covid4 ^ 2) +
    I(covid1 * covid2) + I(covid1 * covid3) + I(covid1 * covid4) +
    I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 * covid4))
step6 = stepAIC(covid_hours.fit2, direction = "both")
step6$anova
```

```
summary(covid_hours.fit2)$r.squared
summary(covid_hours.fit2)$adj.r.squared
```

This R code snippet below produces the model quiz4 = quiz3. This model seems to under-fit though.

```
full_model.fit = lm(quiz4 ~
     quiz1 + quiz2 + quiz3 +
     covid1 + covid2 + covid3 + covid4 +
     study1 + study2 + study3 + study4)
step7 = stepAIC(full_model.fit, direction = "both")
step7$anova
```

```
summary(full_model.fit)$r.squared
summary(full_model.fit)$adj.r.squared
```

This R code snippet below produces the model

quiz4 = quiz3 + I(study1 study2) + I(study2 study3) + I(study3 study4).

This model seems to be a good fit.

```
full_model.fit2 = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + I(covid1 * covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 * covid4)
  + I(study1 * study2) + I(study1 * study3) + I(study2 * study3) + I(study3 * study4))
step8 = stepAIC(full_model.fit2, direction = "both")
step8$anova
```

```r
summary(full_model.fit2)$r.squared
summary(full_model.fit2)$adj.r.squared
```

With just Canada, we get the model

quiz4 ~ quiz3 + I(covid1 covid2) + I(covid2 covid3) + I(study2 study3) + I(study2 study4)

```
full_model_canada.fit = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + I(covid1 * covid2) + I(covid1 * covid3) + I(covid2 * covid3) + I(covid2 * covid4)
  + I(covid3 * covid4)

  + I(study1 * study2) + I(study1 * study3) + I(study1 * study4) + I(study2 * study3)
  + I(study2 * study4) + I(study3 * study4))
step9 = stepAIC(full_model_canada.fit, direction = "both")
step9$anova
```

```
summary(full_model_canada.fit)$r.squared
summary(full_model_canada.fit)$adj.r.squared
```

With just China, we get the model

quiz4 ~ quiz3 + I(covid1 covid2) + I(covid2 covid3) + I(study2 study3) + I(study3 study4)

```
full_model_china.fit = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + I(covid1 * covid2) + I(covid1 * covid3) + I(covid2 * covid3) + I(covid2 * covid4)
  + I(covid3 * covid4)

  + I(study1 * study2) + I(study2 * study3) + I(study3 * study4))
step10 = stepAIC(full_model_china.fit, direction = "both")
step10$anova
```

```
summary(full_model_china.fit)$r.squared
summary(full_model_china.fit)$adj.r.squared
```

With all other countries, we get the model

quiz4 ~ quiz3 + I(covid2 covid3)

```
full_model_other.fit = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 * covid4)
  + I(study3 * study4))
step11 = stepAIC(full_model_other.fit, direction = "both")
step11$anova
```

```
summary(full_model_other.fit)$r.squared
summary(full_model_other.fit)$adj.r.squared
```

Find Significance Predictor Variables

```
# use week 5b slides -- choose model selection criterion to pick predictor variables.


# use lm() on a bunch of predictor variables to determine significant
# predictor variables.
```