

# 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.

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

```

display_residual_plot <- function(data, model, predictor_variable, predictor_variable_name) {
  fit = fitted(model)
  residuals = resid(model)
  ggplot(data = data, aes(x = predictor_variable, y = residuals)) +
    geom_point() +
    geom_hline(yintercept = 0) +
    labs(title = paste0("Residual Plot for Variable ", predictor_variable_name),
         x = predictor_variable_name, y = paste0("Residuals of ", predictor_variable_name))
}

```

## 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)
remaining_data = rearranged_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_columns = remaining_data[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
```

## 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 = remaining_data %>%  
  dplyr::select(-country)
```

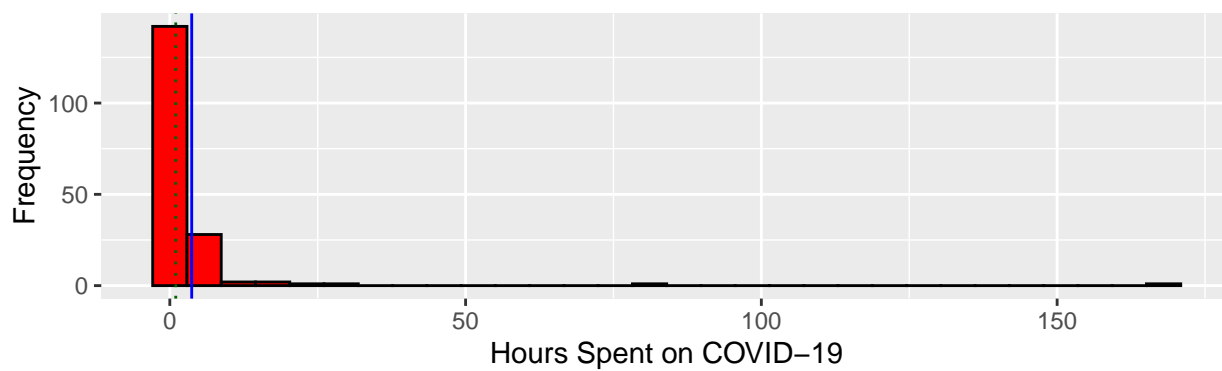
## Rows Filtered by Country

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

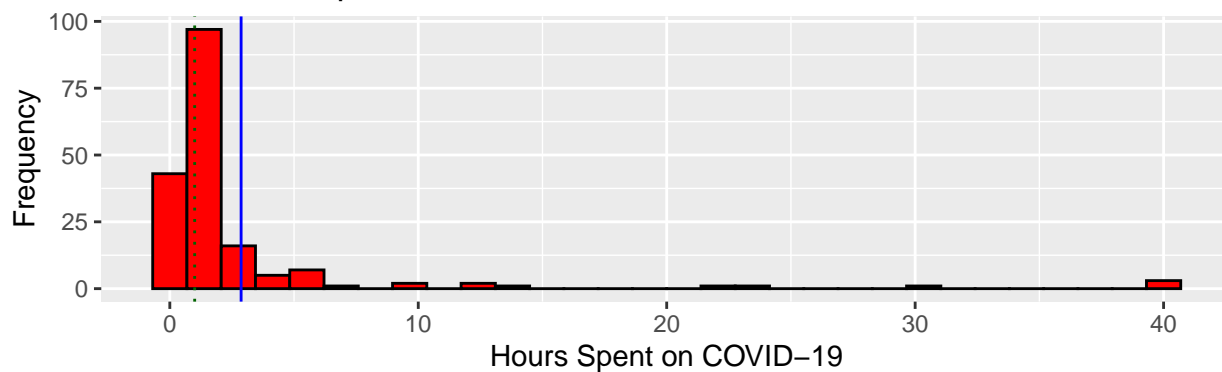
```
canada <- remaining_data %>%  
  filter(as.character(country) == "Canada") %>%  
  dplyr::select(-country)  
  
unknown <- remaining_data %>%  
  filter(is.na(as.character(country))) %>%  
  dplyr::select(-country)
```

##	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

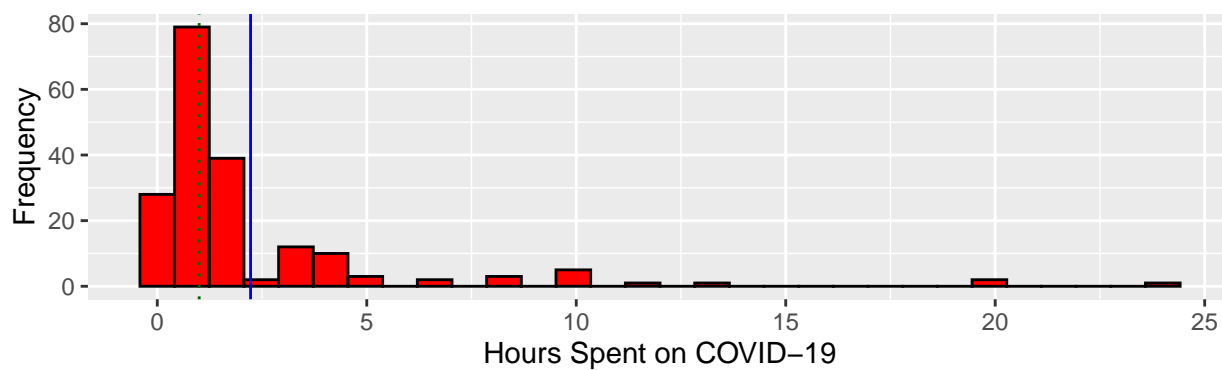
### Week 1 Time Spent on COVID-19



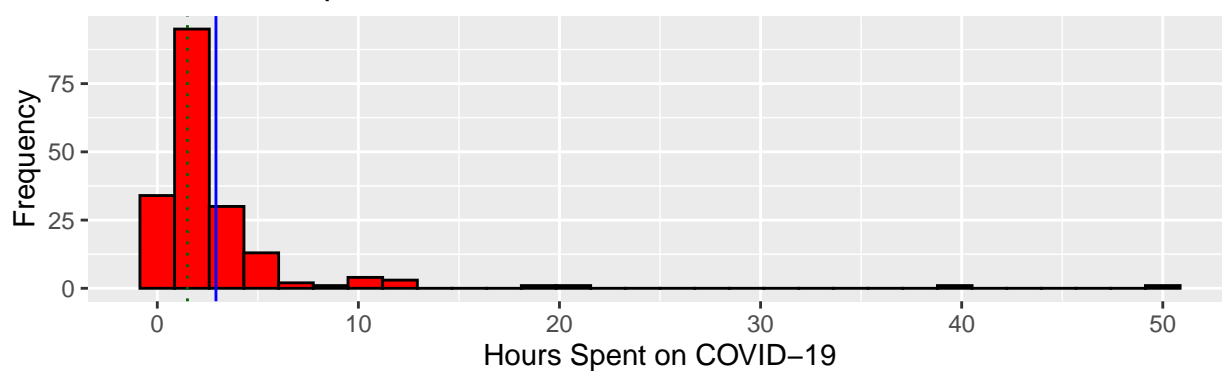
### Week 2 Time Spent on COVID-19

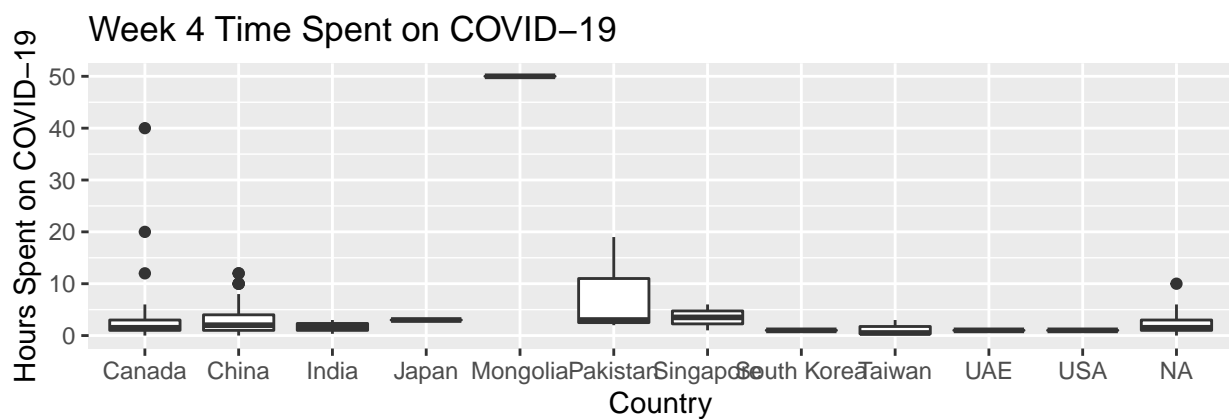
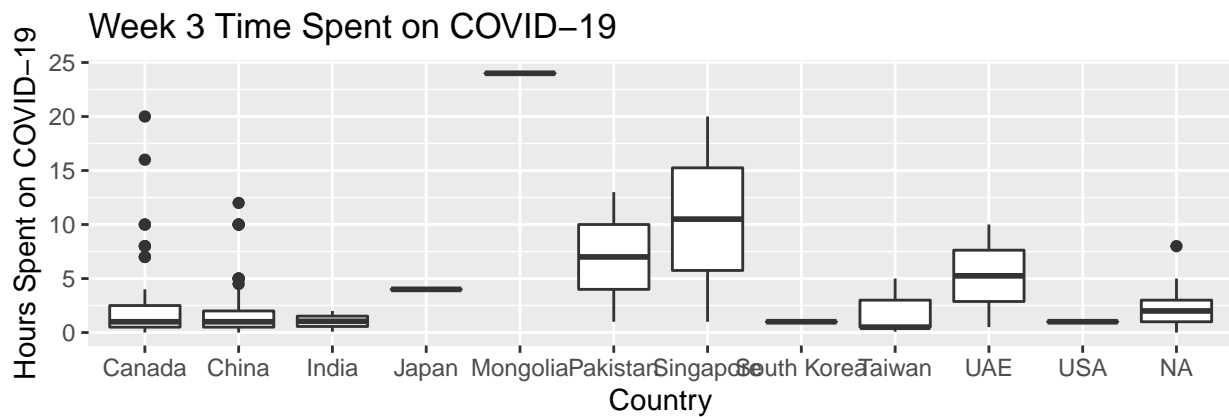
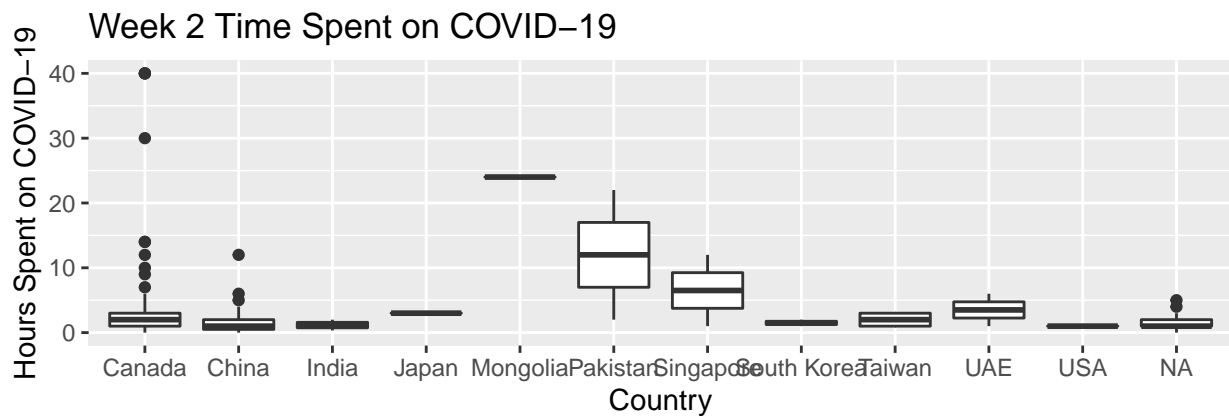
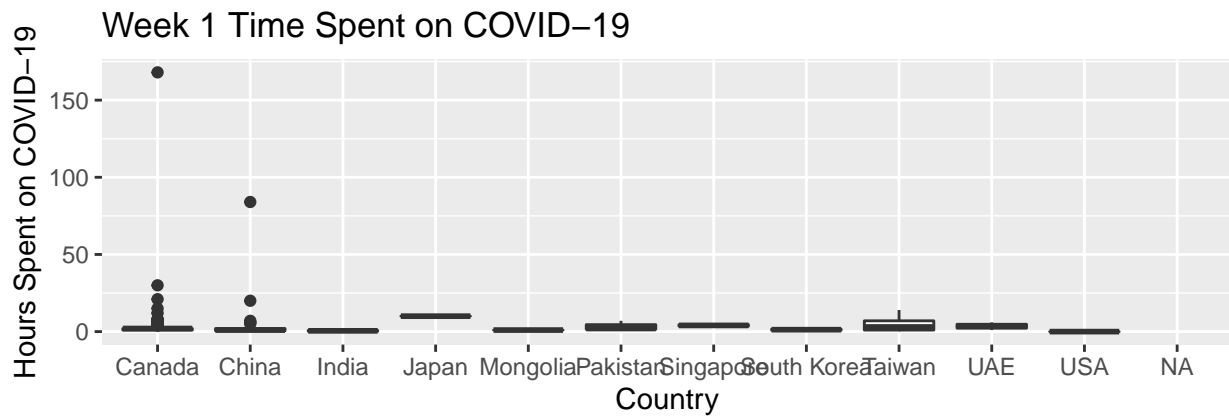


### Week 3 Time Spent on COVID-19

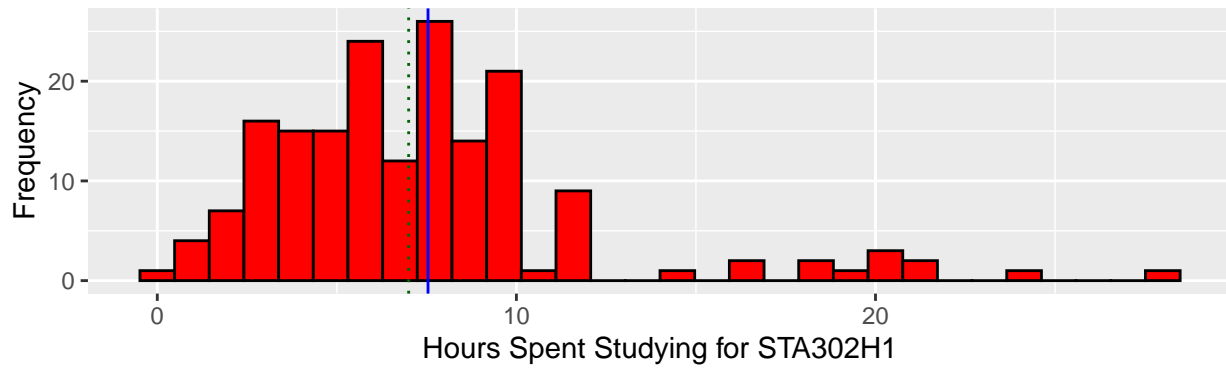


### Week 4 Time Spent on COVID-19

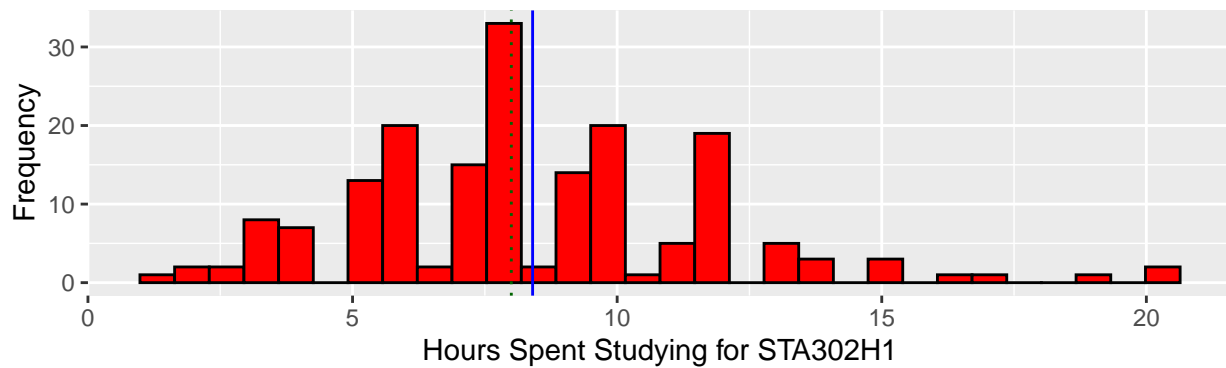




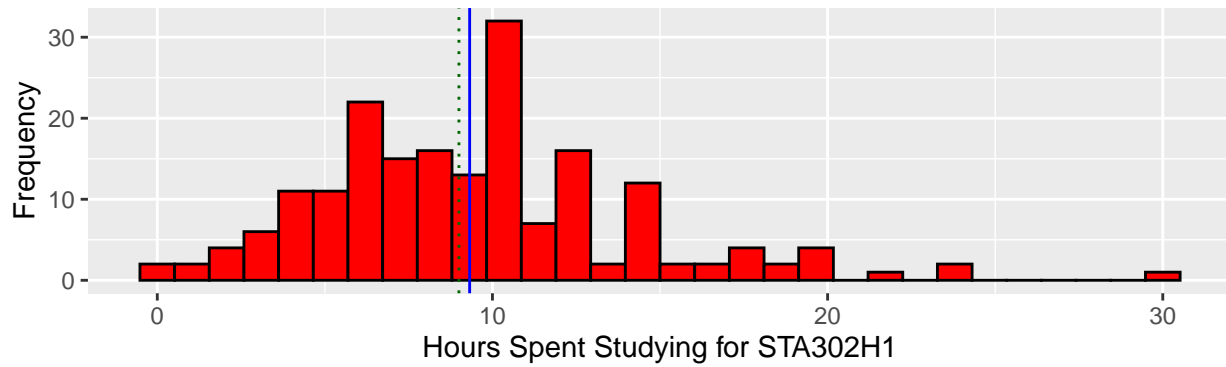
Week 1 Time Spent Studying for STA302H1



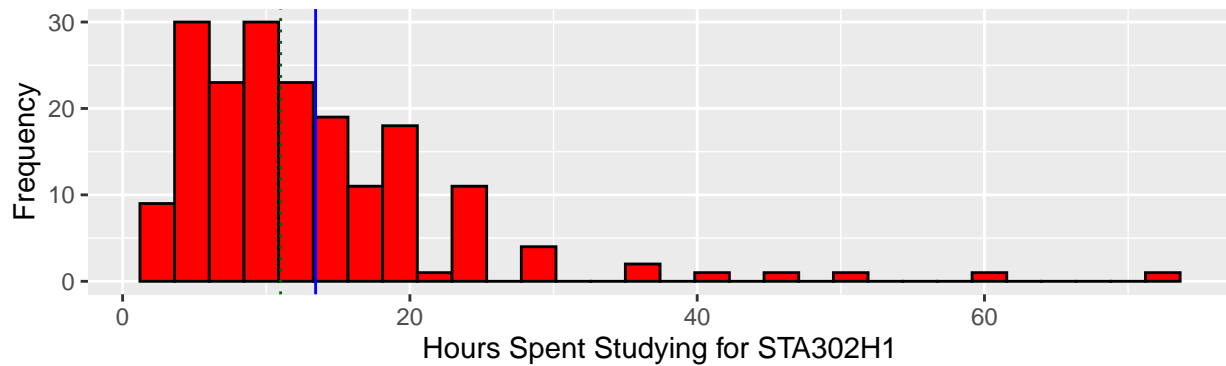
Week 2 Time Spent Studying for STA302H1



Week 3 Time Spent Studying for STA302H1

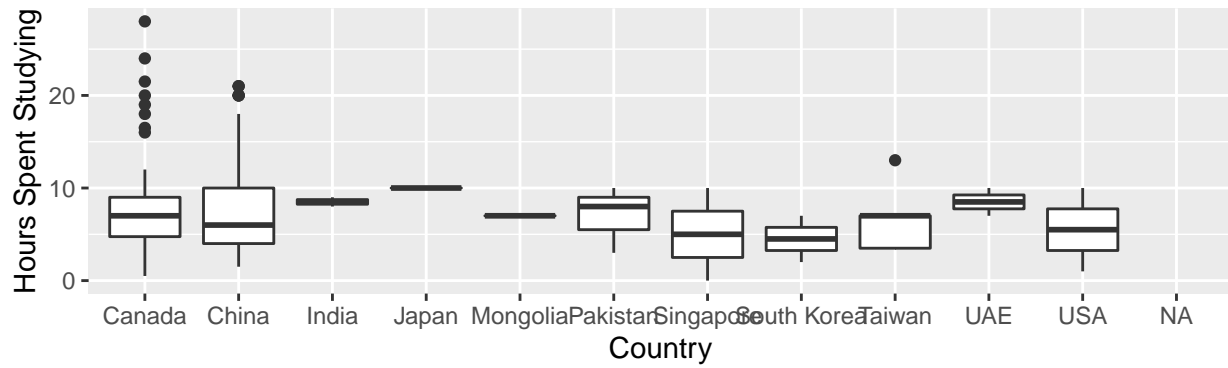


Week 4 Time Spent Studying for STA302H1

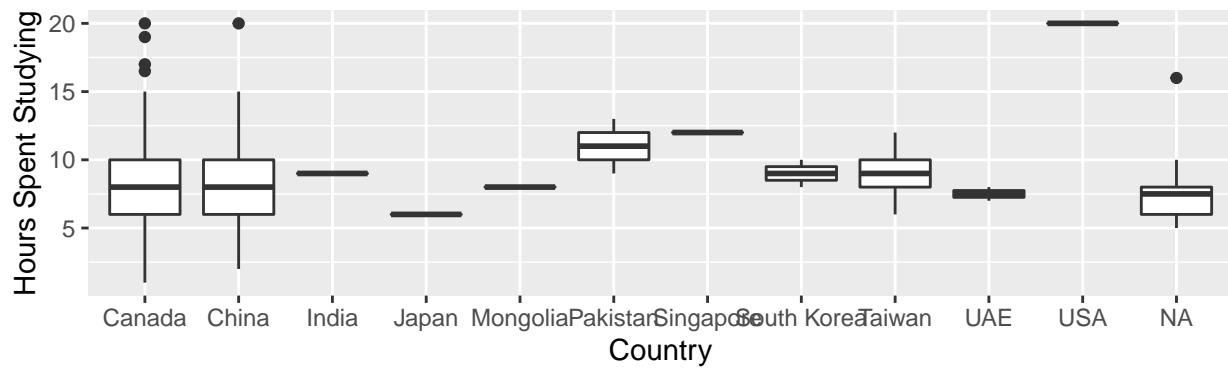




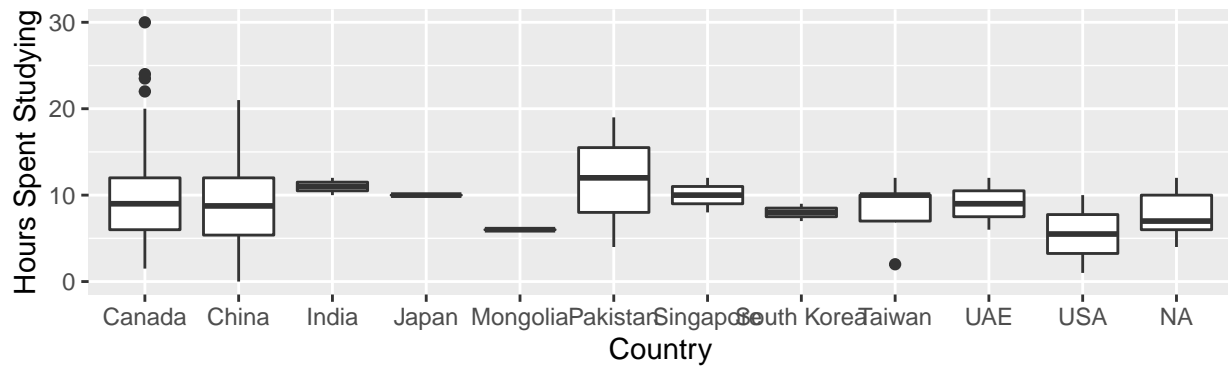
Week 1 Time Spent Studying For STA302H1



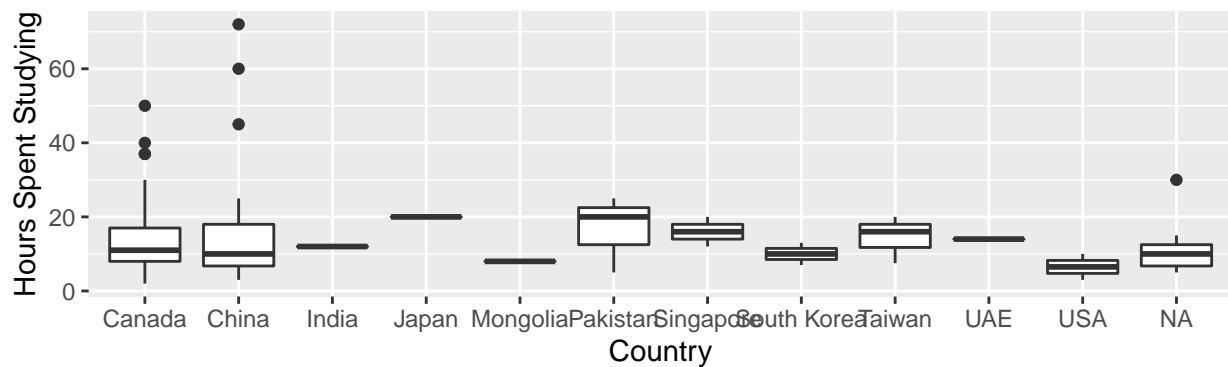
Week 2 Time Spent Studying For STA302H1

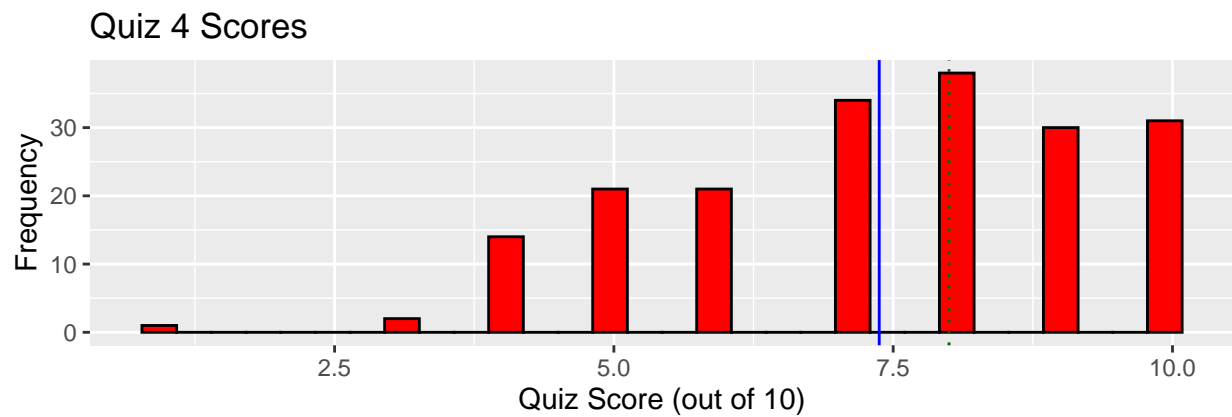
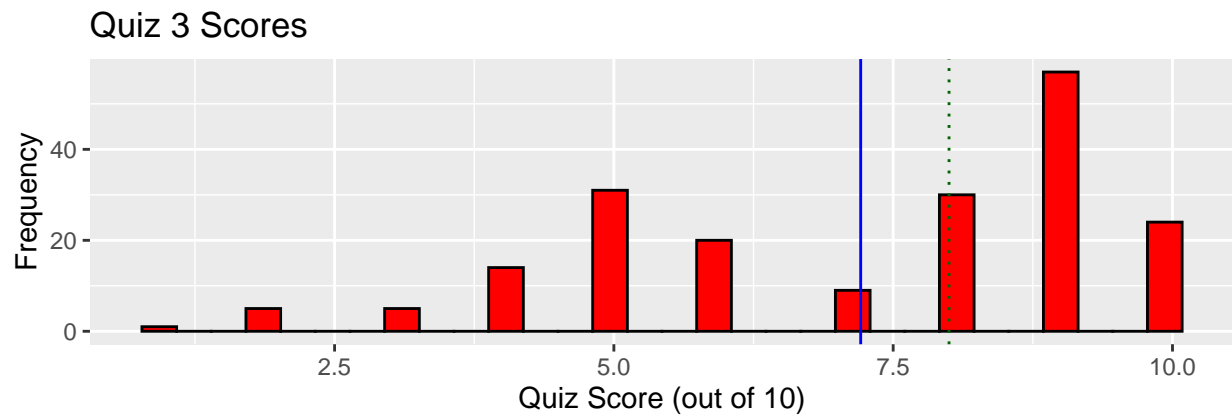
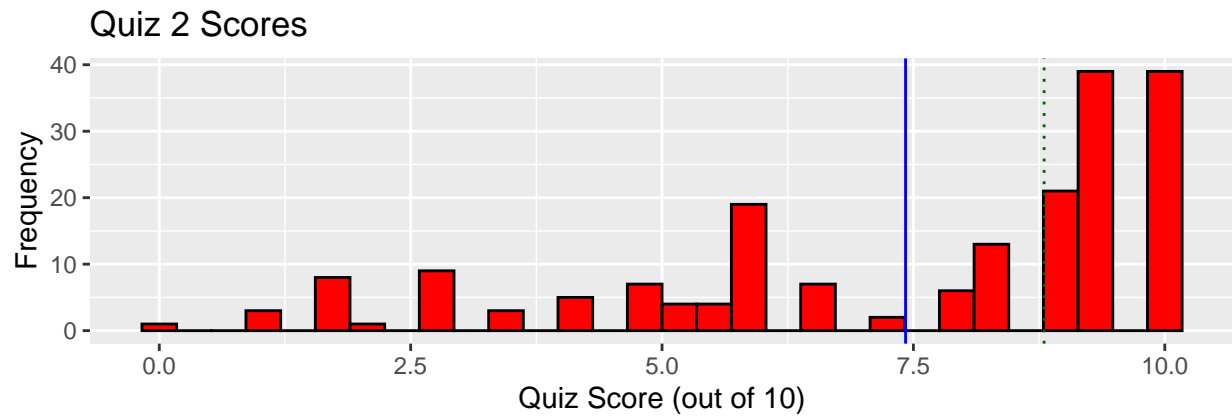
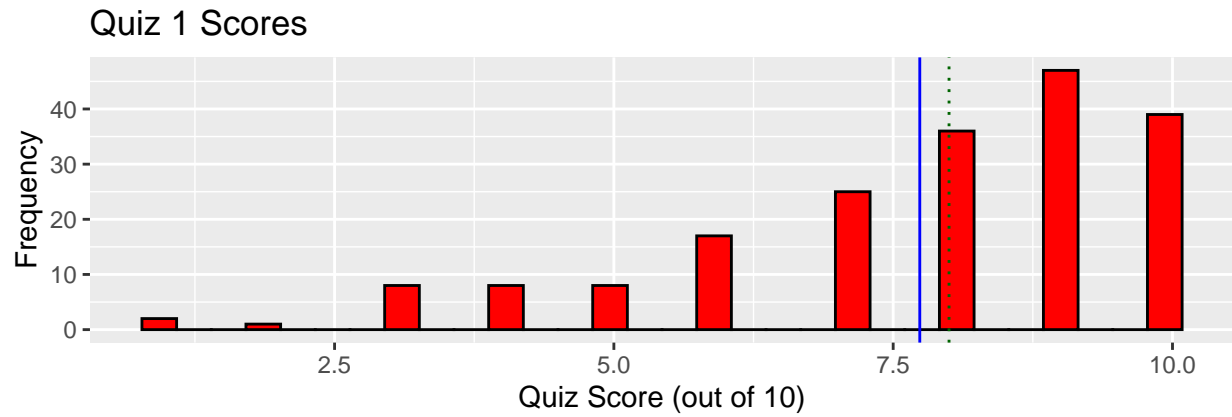


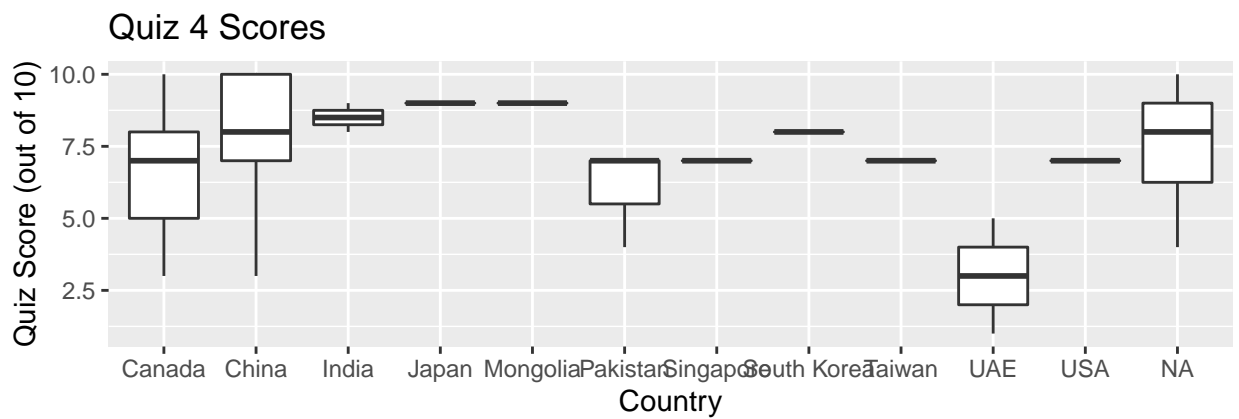
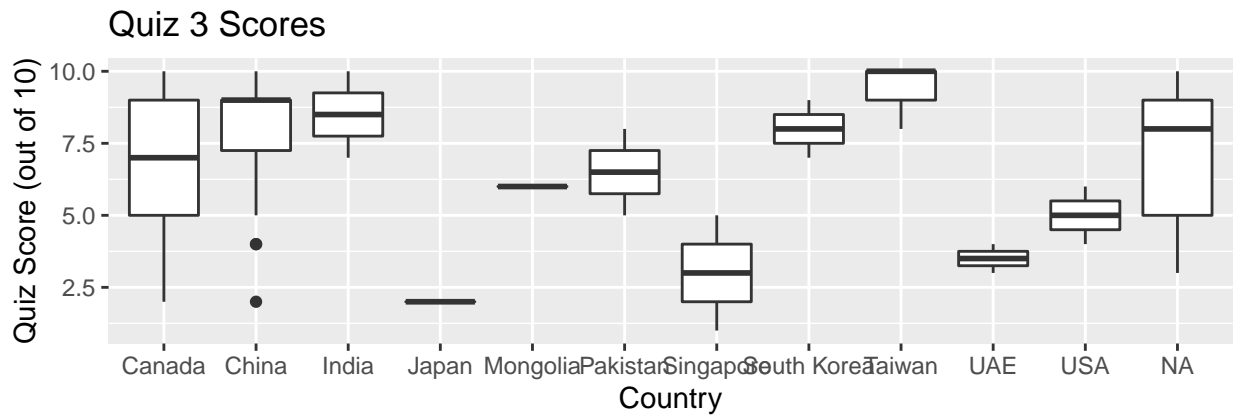
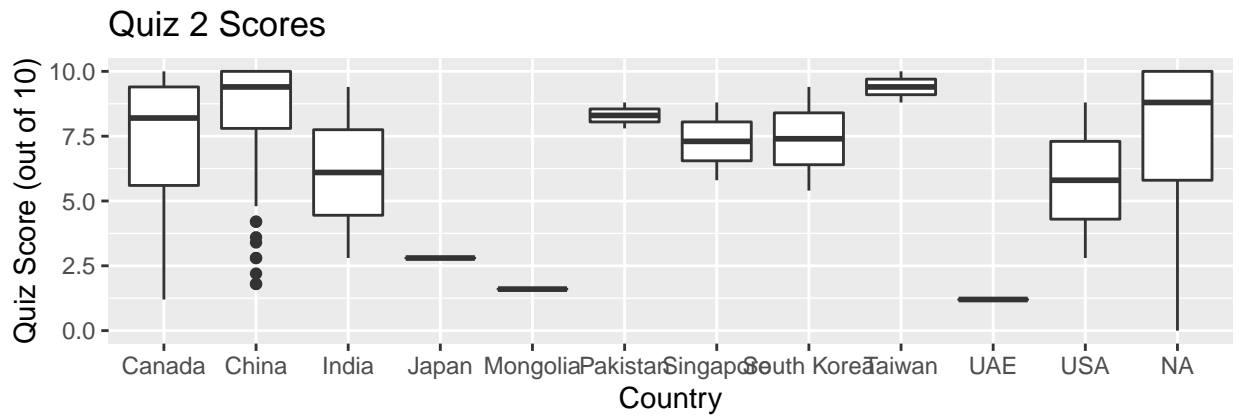
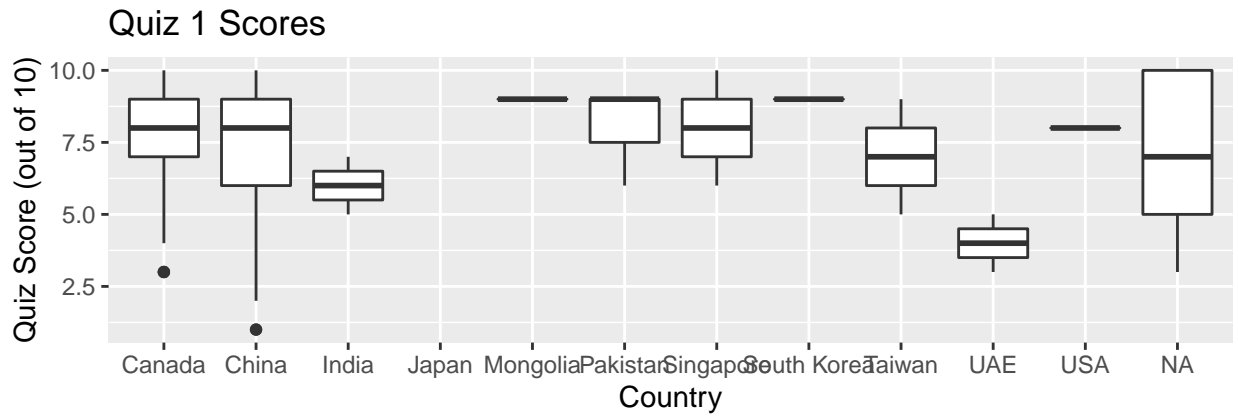
Week 3 Time Spent Studying



Week 4 Time Spent Studying







## 5-Number Summary Statistics

```
summary(remaining_data$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(remaining_data$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(remaining_data$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(remaining_data$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(remaining_data$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(remaining_data$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(remaining_data$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(remaining_data$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

```
summary(remaining_data$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
```

```
summary(remaining_data$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
```

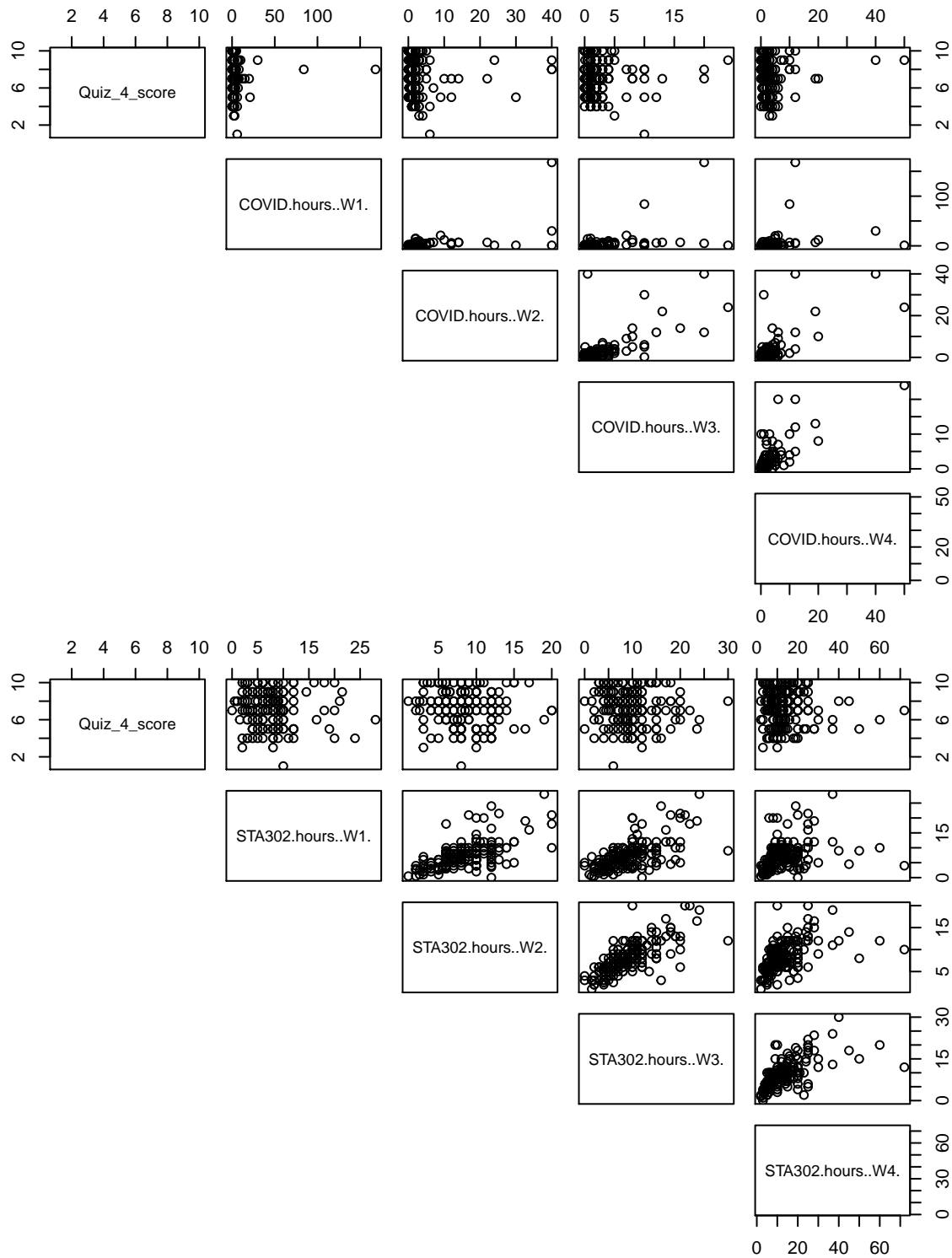
```
summary(remaining_data$Quiz_3_score)
```

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

```
summary(remaining_data$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
```





I wanted to group together all of the quiz scores, COVID times, and study times into separate pairwise scatterplots to see whether they had an individual relationship with quiz 4 scores.

## Correlation Matrix

```
display_correlation_matrix(all_countries)
```

```
##      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
```

We could also create separate correlation matrices for each country.

```
display_correlation_matrix(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_matrix(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_matrix(india)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      1      1      1      1     -1     NA     -1     NA -1  1 -1 -1
## W2COV      1      1      1      1     -1     NA     -1     NA -1  1 -1 -1
## W3COV      1      1      1      1     -1     NA     -1     NA -1  1 -1 -1
## W4COV      1      1      1      1     -1     NA     -1     NA -1  1 -1 -1
## W1302     -1     -1     -1     -1      1     NA      1     NA  1 -1  1  1
## W2302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W3302     -1     -1     -1     -1      1     NA      1     NA  1 -1  1  1
## W4302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q1        -1     -1     -1     -1      1     NA      1     NA  1 -1  1  1
## Q2         1      1      1      1     -1     NA     -1     NA -1  1 -1 -1
## Q3        -1     -1     -1     -1      1     NA      1     NA  1 -1  1  1
## Q4        -1     -1     -1     -1      1     NA      1     NA  1 -1  1  1
```

```
display_correlation_matrix(japan)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W2COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W3COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W4COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W1302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W2302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W3302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W4302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q1         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q2         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q3         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q4         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
```

```
display_correlation_matrix(mongolia)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W2COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W3COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W4COV      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W1302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W2302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W3302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W4302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q1         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q2         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q3         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## Q4         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
```

```
display_correlation_matrix(unknown)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302  Q1  Q2  Q3  Q4
## W1COV      NA     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
## W2COV      NA  1.00  0.70 -0.09     NA -0.19  0.30  0.41  0.41  0.06 -0.50  0.11
## W3COV      NA  0.70  1.00  0.48     NA -0.25  0.06  0.21  0.01  0.09 -0.23  0.16
## W4COV      NA -0.09  0.48  1.00     NA  0.06 -0.05 -0.09  0.07  0.12  0.12  0.39
## W1302      NA     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
## W2302      NA -0.19 -0.25  0.06     NA  1.00  0.06  0.38 -0.09 -0.24  0.14  0.17
## W3302      NA  0.30  0.06 -0.05     NA  0.06  1.00  0.48 -0.34  0.03 -0.22 -0.13
## W4302      NA  0.41  0.21 -0.09     NA  0.38  0.48  1.00  0.04  0.02 -0.35 -0.02
## Q1         NA  0.41  0.01  0.07     NA -0.09 -0.34  0.04  1.00  0.46  0.24  0.47
## Q2         NA  0.06  0.09  0.12     NA -0.24  0.03  0.02  0.46  1.00  0.57  0.47
## Q3         NA -0.50 -0.23  0.12     NA  0.14 -0.22 -0.35  0.24  0.57  1.00  0.58
## Q4         NA  0.11  0.16  0.39     NA  0.17 -0.13 -0.02  0.47  0.47  0.58  1.00
```

```
display_correlation_matrix(pakistan)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302  Q1 Q2 Q3  Q4
## W1COV  1.00     1     1  0.99  0.60     1  0.75  0.57 -0.99 -1 -1  0.36
## W2COV  1.00     1     1  1.00  1.00     1  1.00  1.00 -1.00 -1 -1     NA
## W3COV  1.00     1     1  1.00  1.00     1  1.00  1.00 -1.00 -1 -1     NA
## W4COV  0.99     1     1  1.00  0.68     1  0.82  0.65 -1.00 -1 -1  0.45
## W1302  0.60     1     1  0.68  1.00     1  0.98  1.00 -0.72 -1 -1  0.96
## W2302  1.00     1     1  1.00  1.00     1  1.00  1.00 -1.00 -1 -1     NA
## W3302  0.75     1     1  0.82  0.98     1  1.00  0.97 -0.85 -1 -1  0.88
## W4302  0.57     1     1  0.65  1.00     1  0.97  1.00 -0.69 -1 -1  0.97
## Q1     -0.99    -1    -1 -1.00 -0.72    -1 -0.85 -0.69  1.00  1  1 -0.50
## Q2     -1.00    -1    -1 -1.00 -1.00    -1 -1.00 -1.00  1.00  1  1     NA
## Q3     -1.00    -1    -1 -1.00 -1.00    -1 -1.00 -1.00  1.00  1  1     NA
## Q4       0.36    NA    NA  0.45  0.96     NA  0.88  0.97 -0.50 NA NA  1.00
```

```
display_correlation_matrix(singapore)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## W2COV      1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## W3COV      1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## W4COV      1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## W1302      1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## W2302      NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
## W3302      -1    -1    -1    -1    -1     NA     1     1 -1  1 -1 NA
## W4302      -1    -1    -1    -1    -1     NA     1     1 -1  1 -1 NA
## Q1         1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## Q2        -1    -1    -1    -1    -1     NA     1     1 -1  1 -1 NA
## Q3         1     1     1     1     1     NA     -1     -1  1 -1  1 NA
## Q4         NA     NA     NA     NA     NA     NA     NA     NA  NA NA NA NA
```

```
display_correlation_matrix(south_korea)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      1    -1    NA    NA      1    -1     1      1 NA -1 -1 NA
## W2COV     -1     1    NA    NA     -1     1    -1     -1 NA  1  1 NA
## W3COV      NA    NA    NA    NA      NA    NA    NA      NA NA NA NA NA
## W4COV      NA    NA    NA    NA      NA    NA    NA      NA NA NA NA NA
## W1302      1    -1    NA    NA      1    -1     1      1 NA -1 -1 NA
## W2302     -1     1    NA    NA     -1     1    -1     -1 NA  1  1 NA
## W3302      1    -1    NA    NA      1    -1     1      1 NA -1 -1 NA
## W4302      1    -1    NA    NA      1    -1     1      1 NA -1 -1 NA
## Q1         NA    NA    NA    NA      NA    NA    NA      NA NA NA NA NA
## Q2        -1     1    NA    NA     -1     1    -1     -1 NA  1  1 NA
## Q3        -1     1    NA    NA     -1     1    -1     -1 NA  1  1 NA
## Q4         NA    NA    NA    NA      NA    NA    NA      NA NA NA NA NA
```

```
display_correlation_matrix(taiwan)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV    1.00  1.00  0.99  0.99  0.50 -0.28  0.80  0.75 0.94 0.00 0.50 NA
## W2COV    1.00  1.00  0.99  0.99  0.50 -0.28  0.80  0.75 0.94 0.00 0.50 NA
## W3COV    0.99  0.99  1.00  1.00  0.61 -0.40  0.72  0.65 0.89 -0.13 0.38 NA
## W4COV    0.99  0.99  1.00  1.00  0.59 -0.38  0.73  0.67 0.90 -0.11 0.40 NA
## W1302    0.50  0.50  0.61  0.59  1.00 -0.97 -0.11 -0.20 0.19 -0.87 -0.50 NA
## W2302   -0.28 -0.28 -0.40 -0.38 -0.97  1.00  0.35  0.43 0.05  0.96  0.69 NA
## W3302    0.80  0.80  0.72  0.73 -0.11  0.35  1.00  1.00 0.95  0.60  0.92 NA
## W4302    0.75  0.75  0.65  0.67 -0.20  0.43  1.00  1.00 0.92  0.67  0.95 NA
## Q1       0.94  0.94  0.89  0.90  0.19  0.05  0.95  0.92 1.00  0.33  0.76 NA
## Q2       0.00  0.00 -0.13 -0.11 -0.87  0.96  0.60  0.67 0.33  1.00  0.87 NA
## Q3       0.50  0.50  0.38  0.40 -0.50  0.69  0.92  0.95 0.76  0.87  1.00 NA
## Q4        NA    NA    NA    NA      NA    NA    NA      NA NA    NA NA
```

```
display_correlation_matrix(uae)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      1     1     1    NA      1     1    -1     NA -1 NA -1 -1
## W2COV      1     1     1    NA      1     1    -1     NA -1 NA -1 -1
## W3COV      1     1     1    NA      1     1    -1     NA -1 NA -1 -1
## W4COV      NA    NA    NA    NA      NA    NA    NA     NA NA NA NA NA
## W1302      1     1     1    NA      1     1    -1     NA -1 NA -1 -1
## W2302      1     1     1    NA      1     1    -1     NA -1 NA -1 -1
## W3302     -1    -1    -1    NA     -1    -1     1     NA  1 NA  1  1
## W4302      NA    NA    NA    NA      NA    NA    NA     NA NA NA NA NA
## Q1        -1    -1    -1    NA     -1    -1     1     NA  1 NA  1  1
## Q2         NA    NA    NA    NA      NA    NA    NA     NA NA NA NA NA
## Q3        -1    -1    -1    NA     -1    -1     1     NA  1 NA  1  1
## Q4        -1    -1    -1    NA     -1    -1     1     NA  1 NA  1  1
```

```
display_correlation_matrix(usa)
```

```
##          W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302 Q1 Q2 Q3 Q4
## W1COV      NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
## W2COV      NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
## W3COV      NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
## W4COV      NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
## W1302      NA     NA     NA     NA      1     NA      1      1 NA  1  1 NA
## W2302      NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
## W3302      NA     NA     NA     NA      1     NA      1      1 NA  1  1 NA
## W4302      NA     NA     NA     NA      1     NA      1      1 NA  1  1 NA
## Q1         NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
## Q2         NA     NA     NA     NA      1     NA      1      1 NA  1  1 NA
## Q3         NA     NA     NA     NA      1     NA      1      1 NA  1  1 NA
## Q4         NA     NA     NA     NA     NA     NA     NA     NA NA NA NA NA
```

Since every country other than China, Canada, or Unknown doesn't have enough entries, it does not make sense to look at the correlation of those countries individually.

Therefore, we'll only be working with the correlation of all covariates, regardless of country.

## Full Model

```
# single variable per term = additive model
first_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(first_model)
```

```
tbl_regression(first_model, exponentiate = FALSE)
```

Characteristic	Beta	95% CI	p-value
quiz1	0.03	-0.13, 0.20	0.7
quiz2	0.05	-0.07, 0.17	0.4
quiz3	0.48	0.32, 0.63	<0.001
covid1	0.18	-0.07, 0.43	0.2
I(covid1^2)	0.02	0.00, 0.03	0.029
covid2	0.29	-0.09, 0.67	0.13
I(covid2^2)	-0.02	-0.05, 0.00	0.046
covid3	-0.05	-0.30, 0.20	0.7
covid4	-0.25	-0.55, 0.06	0.11
I(covid4^2)	0.02	-0.01, 0.05	0.2
I(covid1 * covid2)	-0.07	-0.14, -0.01	0.029
I(covid2 * covid3)	0.05	-0.01, 0.11	0.12
I(covid2 * covid4)	0.04	-0.01, 0.09	0.093
I(covid3 * covid4)	-0.08	-0.18, 0.02	0.13
I(study1 * study2)	-0.02	-0.03, 0.00	0.018
I(study1 * study3)	0.01	0.00, 0.02	0.14
I(study2 * study3)	0.01	0.00, 0.02	0.095
I(study3 * study4)	0.00	0.00, 0.00	0.14
country			
Canada			
China	0.59	-0.10, 1.3	0.092
India	0.87	-1.5, 3.2	0.5
Mongolia	-13	-51, 26	0.5
Pakistan	-0.15	-3.3, 3.0	>0.9
Singapore	1.2	-2.1, 4.5	0.5
South Korea	-0.02	-2.3, 2.3	>0.9
Taiwan	-1.2	-3.5, 1.1	0.3
UAE	-0.63	-3.9, 2.6	0.7
USA	1.5	-2.0, 5.0	0.4

```
stepAIC(first_model, direction = "forward")$anova
```

```
stepAIC(first_model, direction = "backward")$anova
```

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

## Final Model

```
final_model = lm(
  quiz4 ~ quiz3
  + I(covid1 ^ 2)    # don't remove, else all other terms become insignificant
  + I(covid1 * covid2)
  + I(covid2 * covid3) # don't remove, else all other terms become insignificant
  + I(study1 * study2)
  + I(study1 * study3) # maybe don't remove?
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(final_model)
```

```
tbl_regression(final_model, exponentiate = FALSE)
```

Characteristic	Beta	95% CI	p-value
quiz3	0.50	0.38, 0.62	<0.001
I(covid1^2)	0.00	0.00, 0.01	0.046
I(covid1 * covid2)	-0.02	-0.04, 0.00	0.043
I(covid2 * covid3)	0.00	0.00, 0.01	0.066
I(study1 * study2)	-0.02	-0.03, 0.00	0.014
I(study1 * study3)	0.01	0.00, 0.02	0.2
I(study2 * study3)	0.01	0.00, 0.02	0.029
I(study3 * study4)	0.00	0.00, 0.00	0.035

```
stepAIC(final_model, direction = "forward")$anova
```

```
stepAIC(final_model, direction = "backward")$anova
```

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

## Final Model with Some Terms I Pruned Myself

```
# I decide to remove more terms for simplicity.
third_model = lm(
  quiz4 ~ quiz3
  # + I(covid1 ^ 2) # this lone quadratic term add a lot of complexity for negligible change in R^2 and
  # + 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 than
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(third_model)
```

```
tbl_regression(third_model, exponentiate = FALSE)
```

Characteristic	Beta	95% CI	p-value
quiz3	0.48	0.36, 0.61	<0.001
I(study1 * study2)	-0.01	-0.01, 0.00	0.055
I(study2 * study3)	0.01	0.00, 0.02	0.10
I(study3 * study4)	0.00	0.00, 0.00	0.2

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

```
stepAIC(third_model, direction = "forward")$anova
```

```
stepAIC(third_model, direction = "backward")$anova
```



## Simplistic Model

```
fourth_model = lm(quiz4 ~ quiz3)
summary(fourth_model)
```

```
tbl_regression(fourth_model, exponentiate = FALSE)
```

Characteristic	Beta	95% CI	p-value
quiz3	0.46	0.34, 0.59	<0.001

```
stepAIC(fourth_model, direction = "forward")$anova
```

```
stepAIC(fourth_model, direction = "backward")$anova
```

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

## Linear Model Only

```
additive_model = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + covid1 + covid2 + covid3 + covid4
  + study1 + study2 + study3 + study4
  + country
)
summary(additive_model)
```

```
tbl_regression(additive_model, exponentiate = FALSE)
```

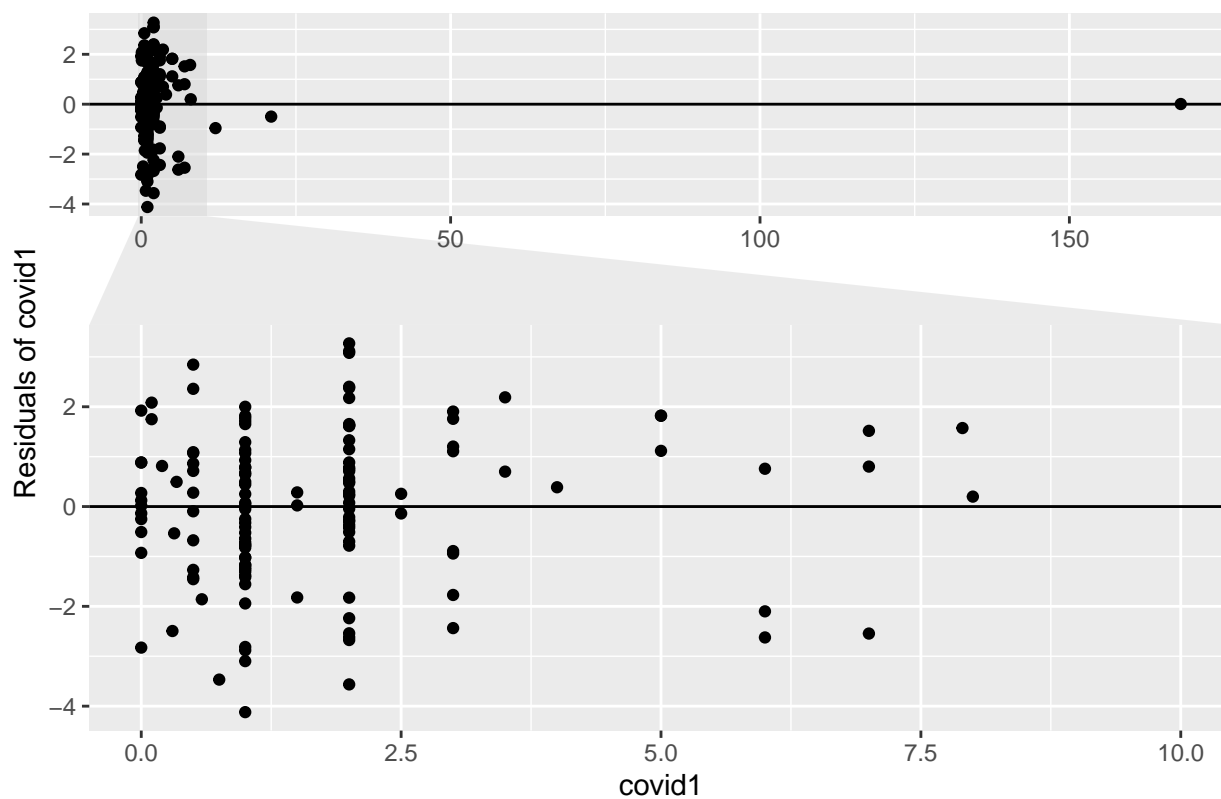
Characteristic	Beta	95% CI	p-value
quiz1	0.05	-0.11, 0.21	0.5
quiz2	0.03	-0.09, 0.15	0.6
quiz3	0.46	0.31, 0.61	<0.001
covid1	0.01	-0.02, 0.04	0.5
covid2	0.03	-0.09, 0.14	0.7
covid3	-0.07	-0.24, 0.09	0.4
covid4	-0.08	-0.21, 0.04	0.2
study1	-0.04	-0.12, 0.04	0.3
study2	-0.06	-0.19, 0.07	0.3
study3	0.10	0.01, 0.19	0.038
study4	-0.02	-0.06, 0.01	0.2
country			
Canada			
China	0.50	-0.16, 1.2	0.13
India	0.77	-1.5, 3.1	0.5
Mongolia	7.8	1.7, 14	0.013
Pakistan	0.87	-1.6, 3.3	0.5
Singapore	2.5	-0.10, 5.2	0.059
South Korea	0.30	-2.0, 2.6	0.8
Taiwan	-1.0	-3.3, 1.3	0.4
UAE	-0.79	-4.1, 2.5	0.6
USA	1.0	-2.5, 4.4	0.6

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

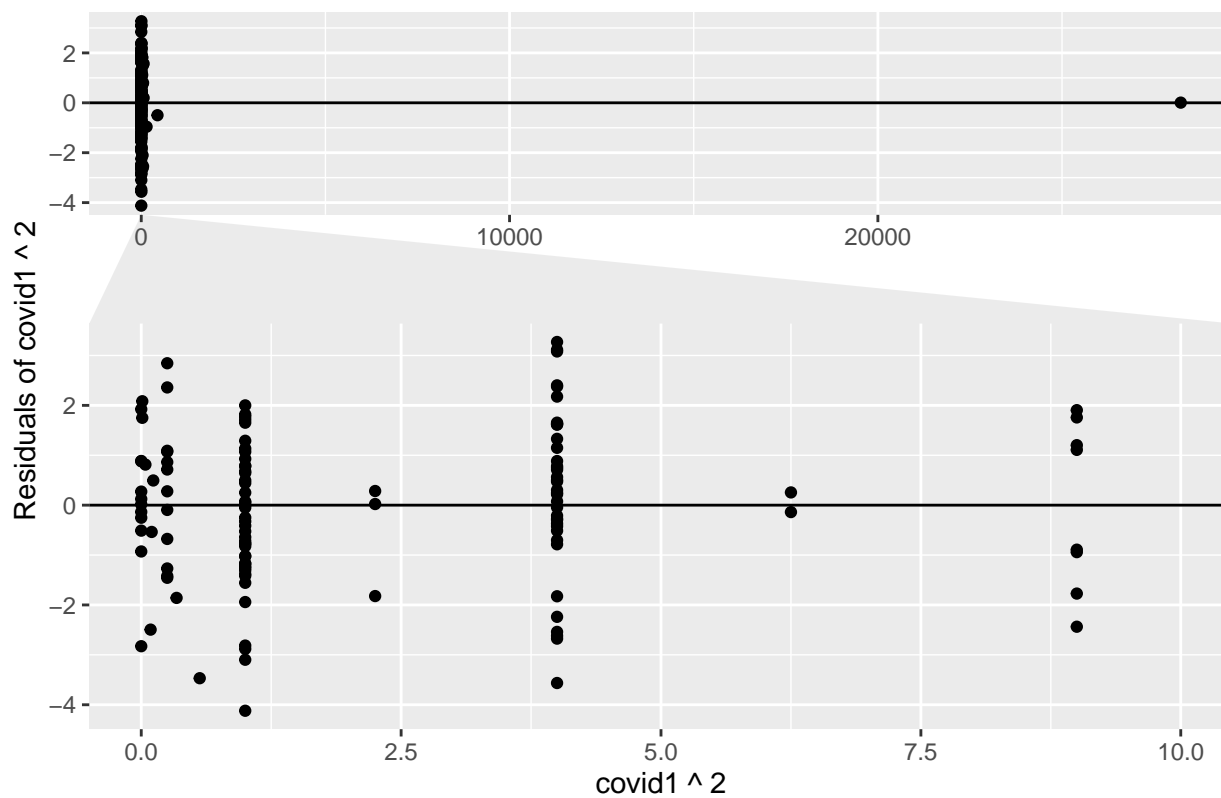
```
stepAIC(additive_model, direction = "backward")$anova
```

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

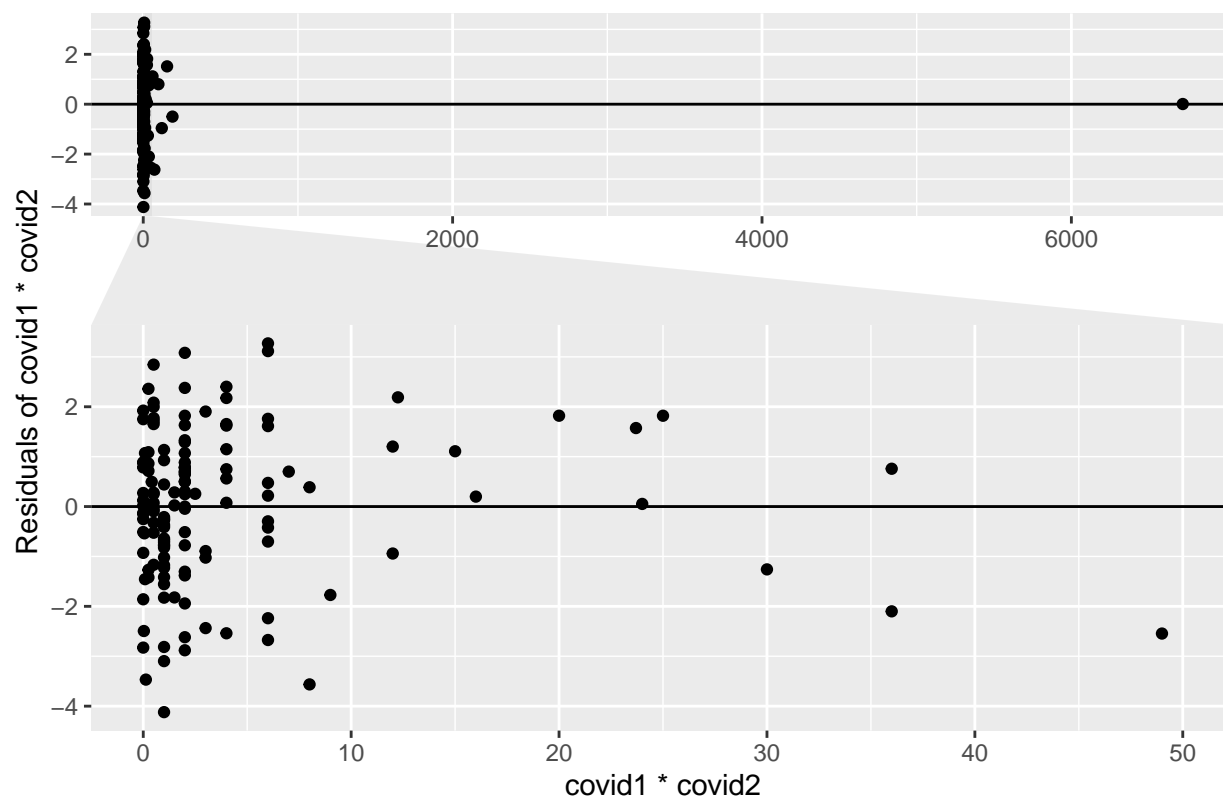
Residual Plot for Variable covid1



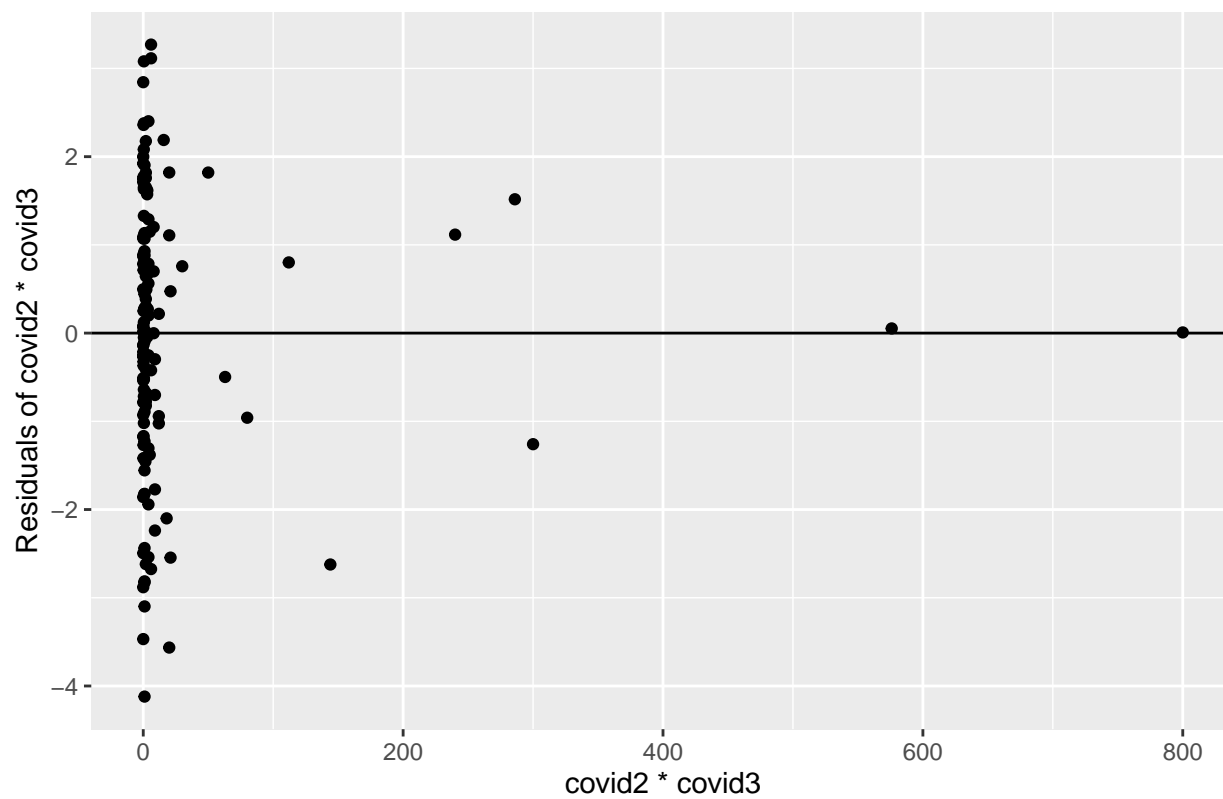
Residual Plot for Variable covid1 ^ 2



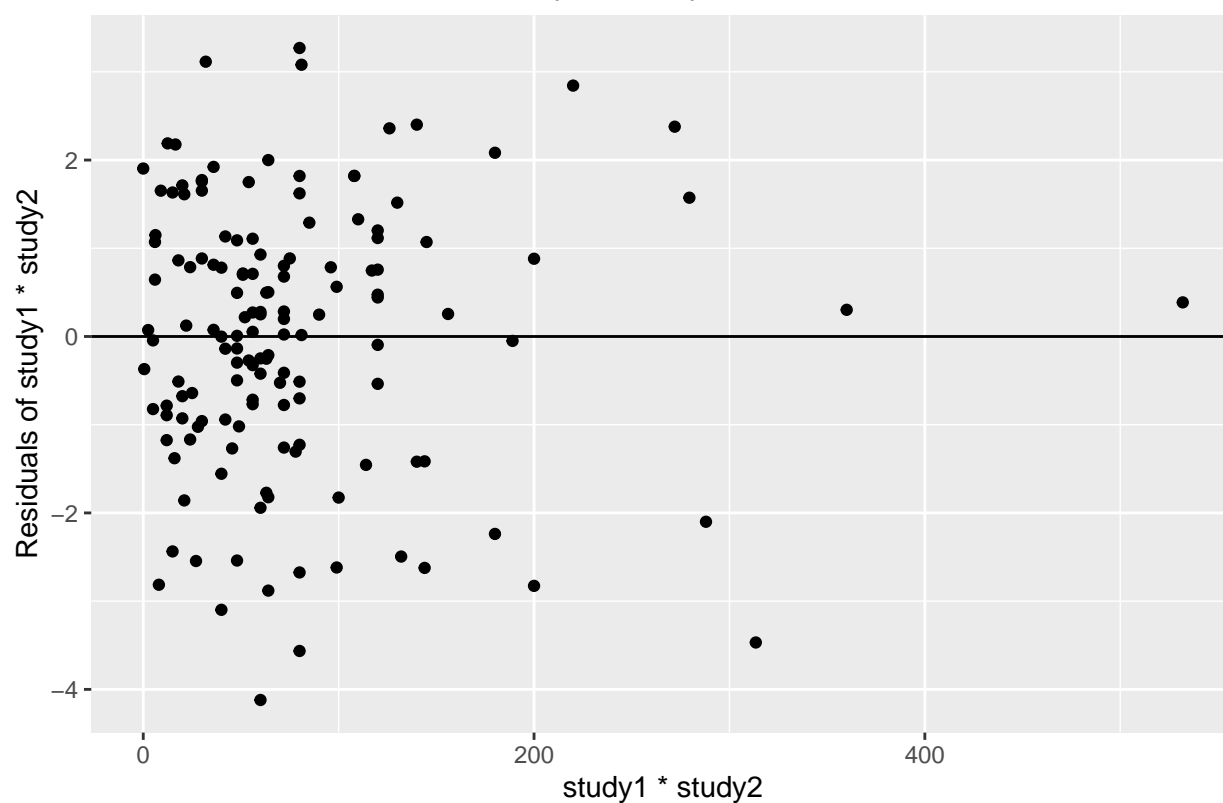
Residual Plot for Variable covid1 \* covid2



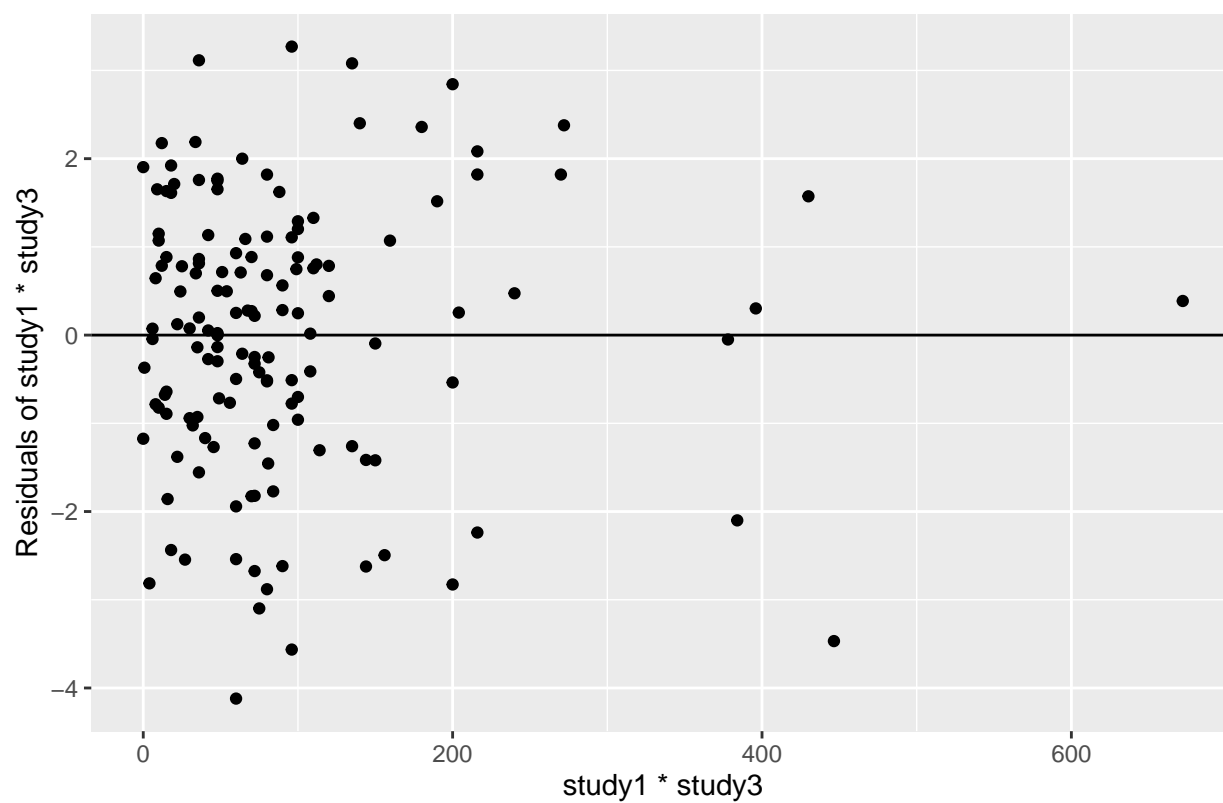
Residual Plot for Variable covid2 \* covid3



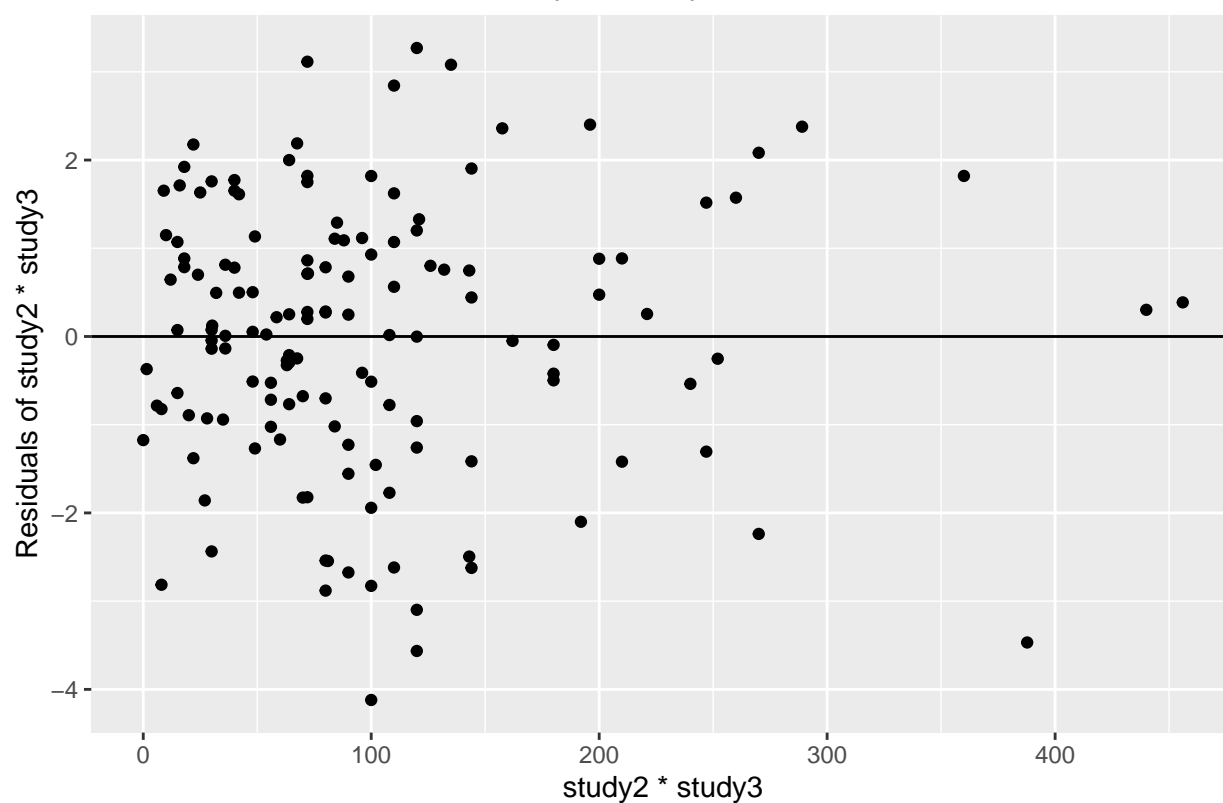
Residual Plot for Variable study1 \* study2



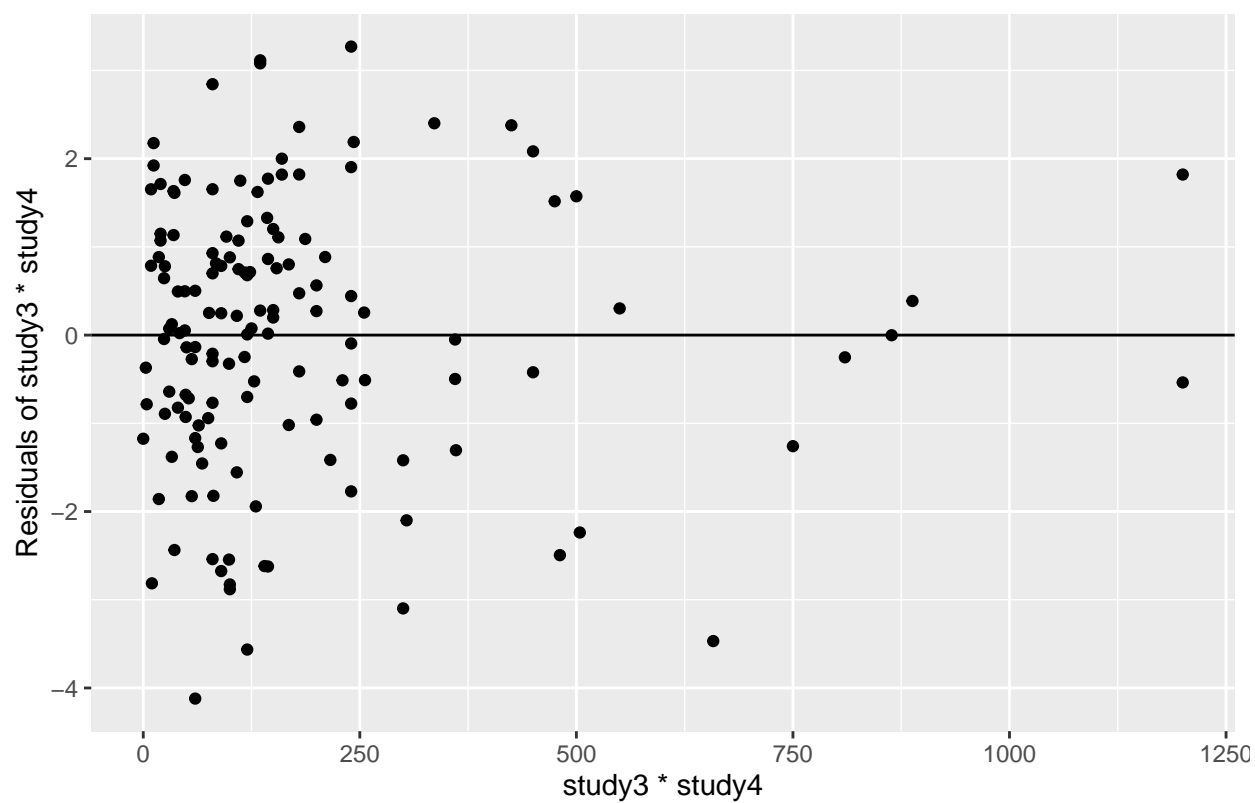
Residual Plot for Variable study1 \* study3

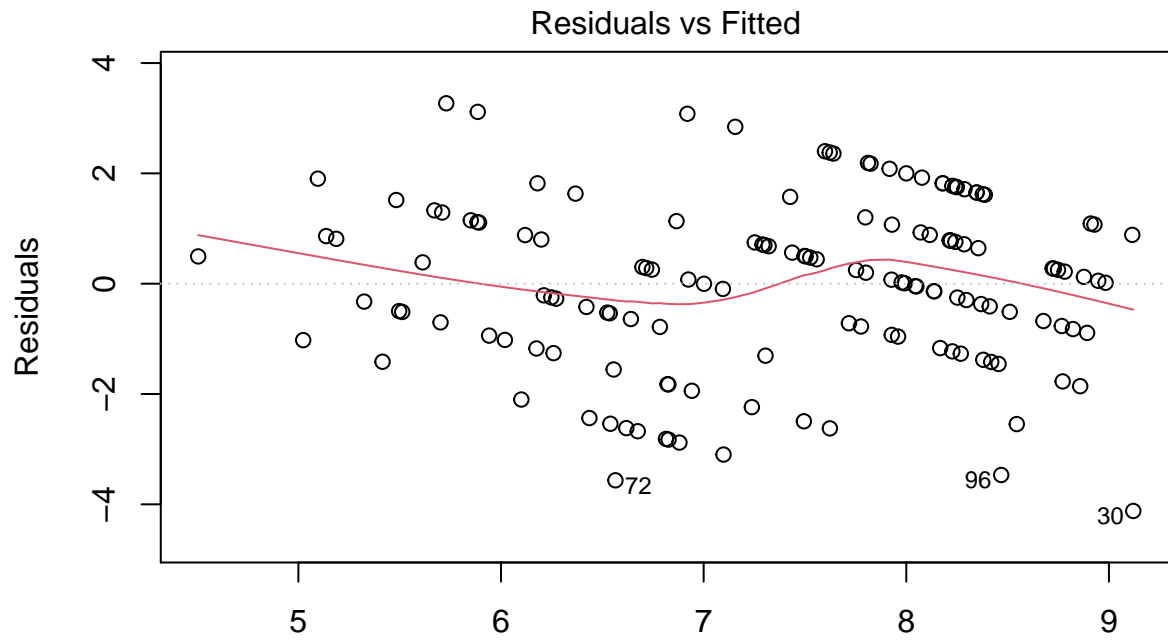


Residual Plot for Variable study2 \* study3

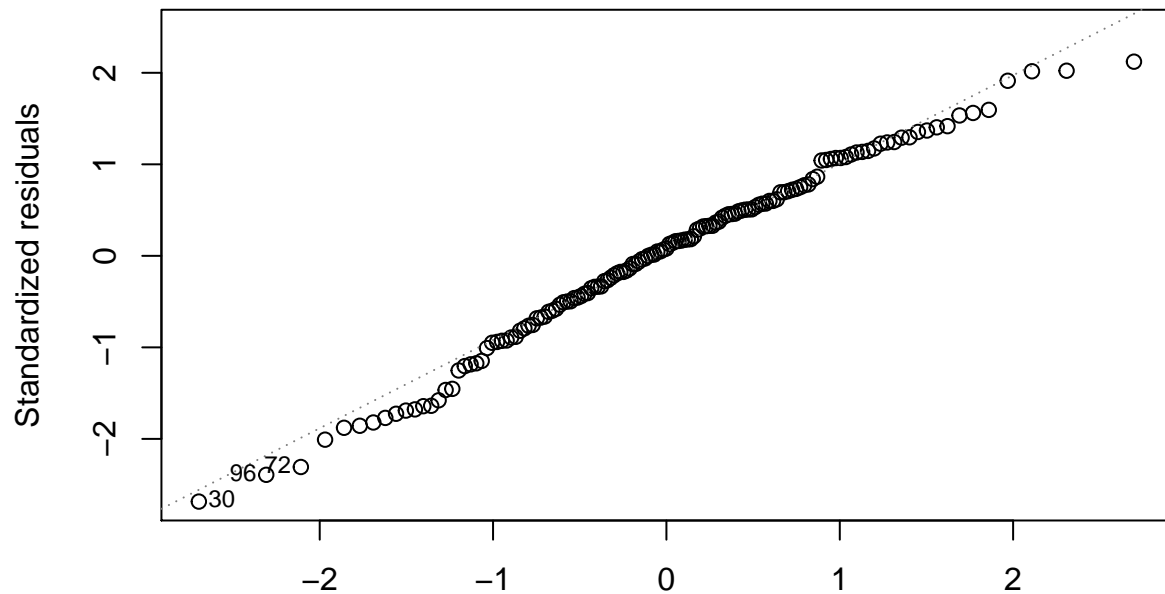


Residual Plot for Variable study3 \* study4

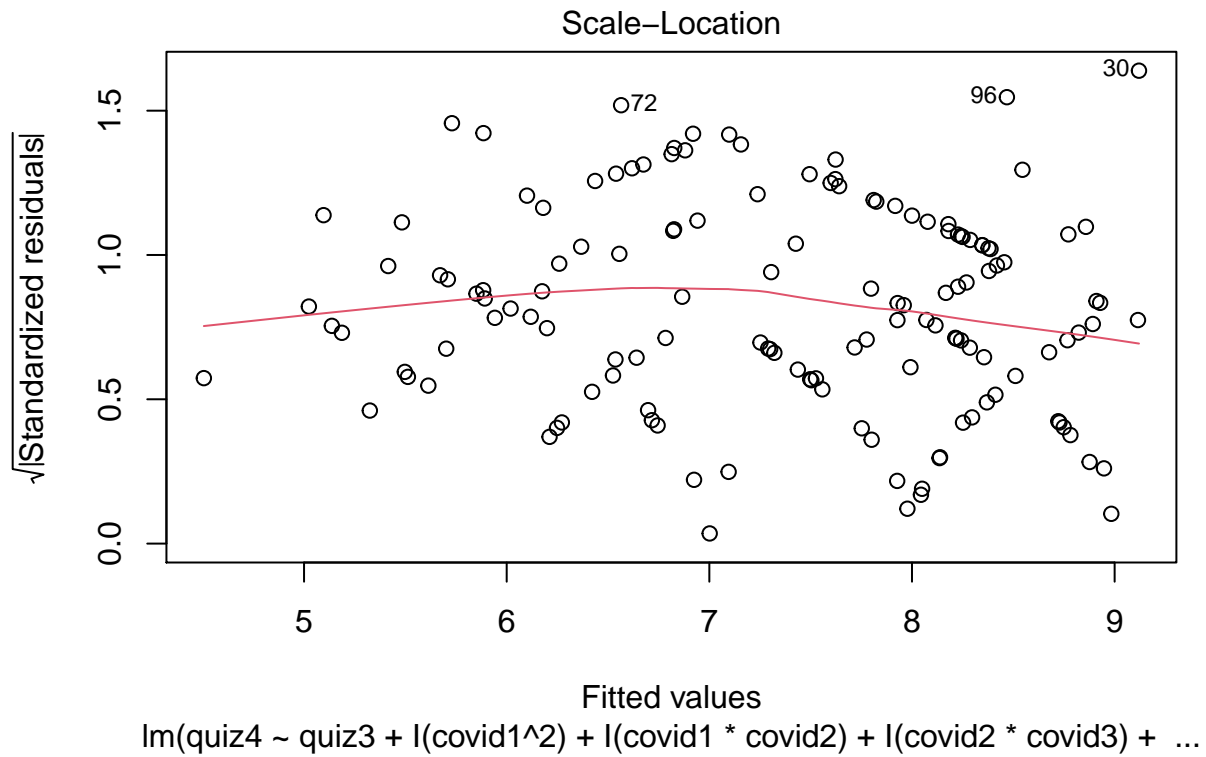




Fitted values  
 $\text{lm}(\text{quiz4} \sim \text{quiz3} + \text{I}(\text{covid1}^2) + \text{I}(\text{covid1} * \text{covid2}) + \text{I}(\text{covid2} * \text{covid3}) + \dots$   
 Normal Q-Q

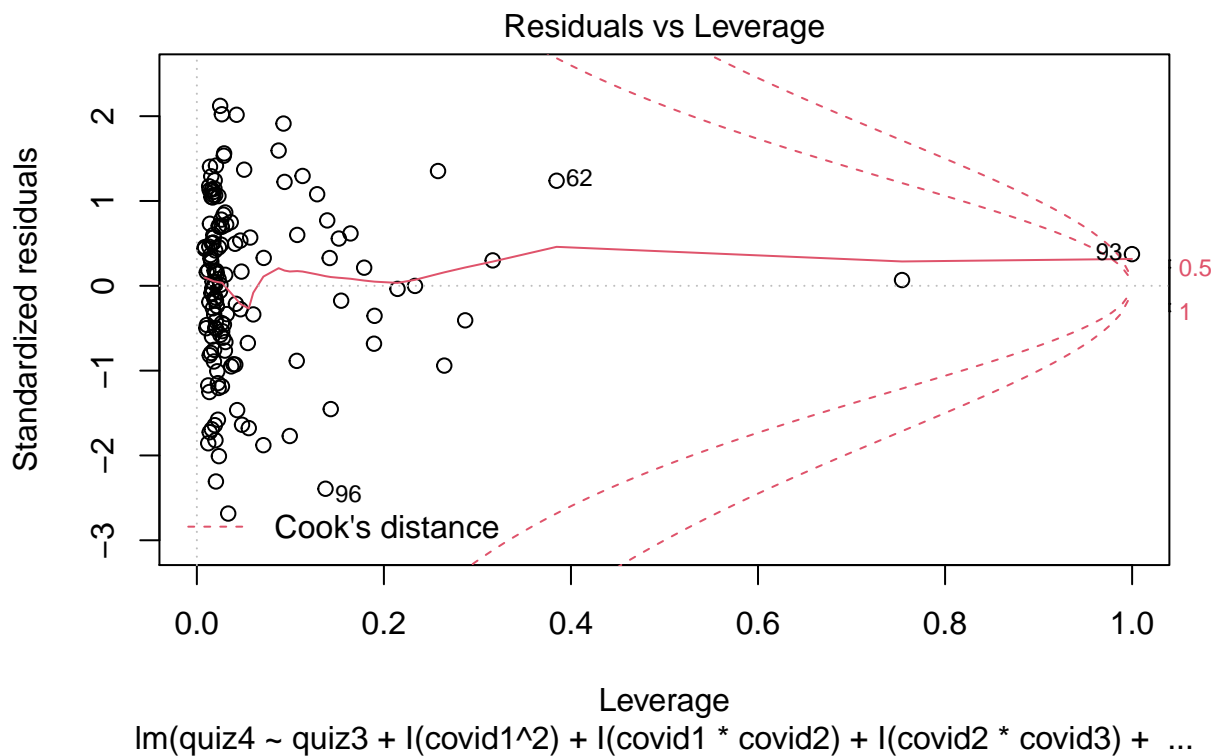


Theoretical Quantiles  
 $\text{lm}(\text{quiz4} \sim \text{quiz3} + \text{I}(\text{covid1}^2) + \text{I}(\text{covid1} * \text{covid2}) + \text{I}(\text{covid2} * \text{covid3}) + \dots$



## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced

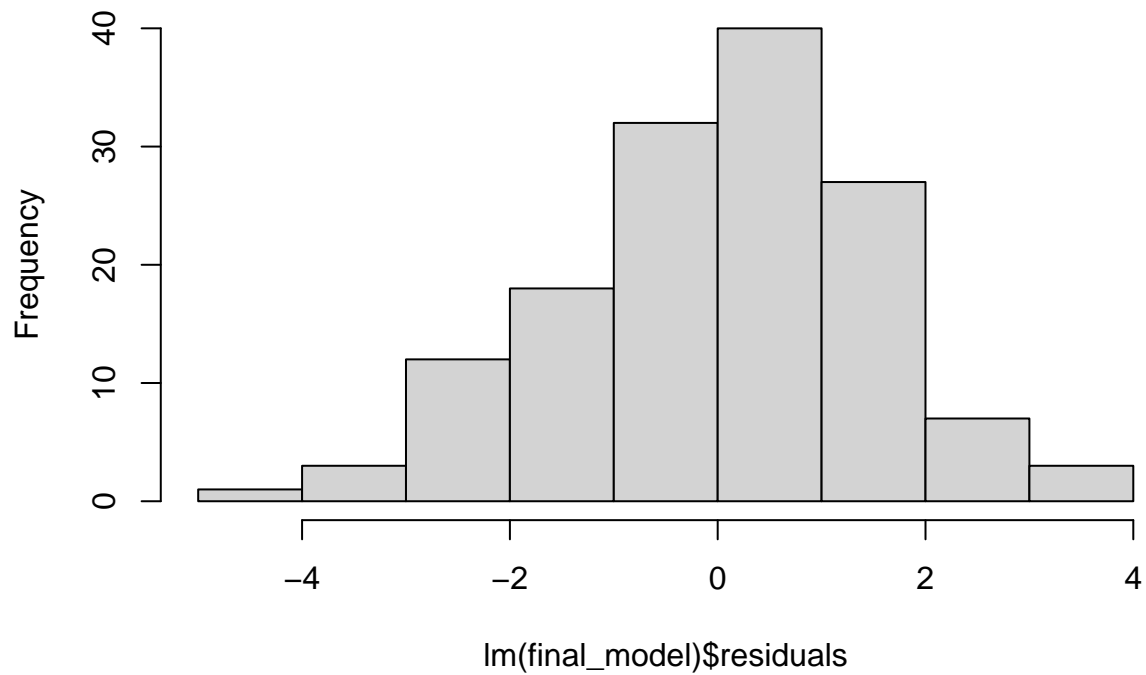
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced





## Goodness of Current Model

### Histogram of `lm(final_model)$residuals`



```
mean(lm(final_model)$residuals)
```

```
## [1] -1.651627e-17
```

```
median(lm(final_model)$residuals)
```

```
## [1] 0.07546203
```

## Try Predicting on the Fitted Values

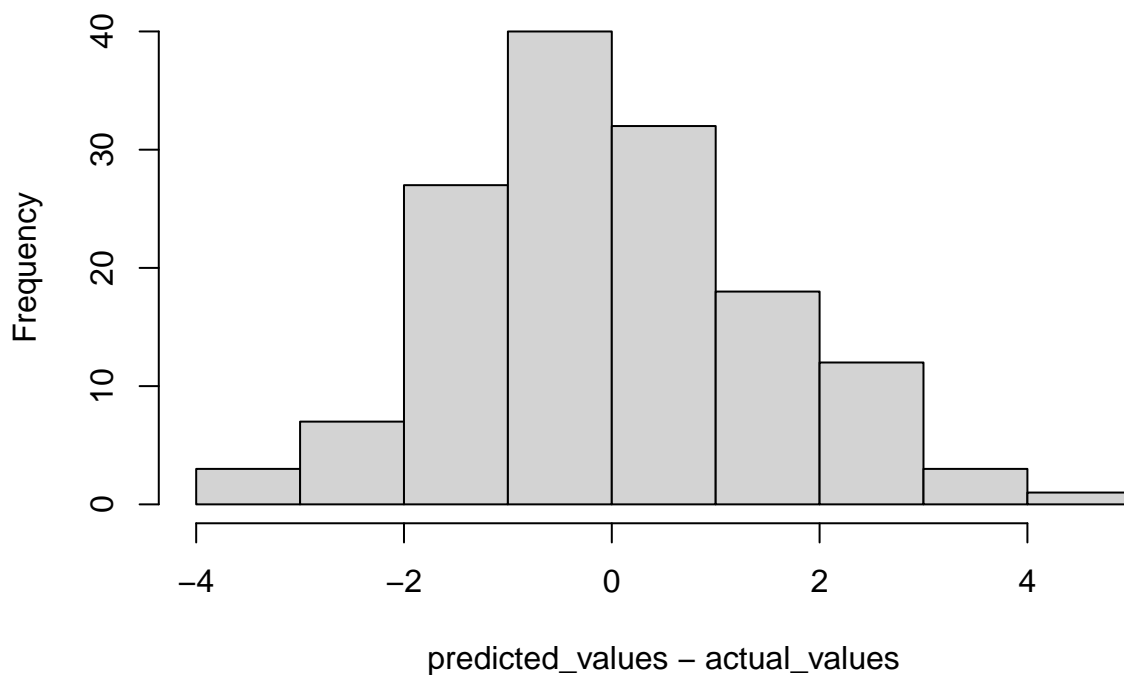
```
mean(predicted_values - actual_values)
```

```
## [1] 5.664449e-15
```

```
median(predicted_values - actual_values)
```

```
## [1] -0.07546203
```

### Histogram of predicted\_values – actual\_values



```
t.test(predicted_values - actual_values)
```

```
##  
## One Sample t-test  
##  
## data: predicted_values - actual_values  
## t = 4.467e-14, df = 142, p-value = 1  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.250671 0.250671  
## sample estimates:  
## mean of x  
## 5.664449e-15
```

## 50/50 Training/Testing

### Partitioning Phase

Source: <https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function>

```
set.seed(888)
sample = sample.split(remaining_data_no_NAs, SplitRatio = 0.55)
training_data = subset(remaining_data_no_NAs, sample == TRUE)
testing_data = subset(remaining_data_no_NAs, sample == FALSE)
```

### Training Phase

```
final_model = lm(
  quiz4 ~ quiz3
  + I(covid1 ^ 2) # don't remove, else all other terms become insignificant
  + I(covid1 * covid2)
  + I(covid2 * covid3) # don't remove, else all other terms become insignificant
  + I(study1 * study2)
  + I(study1 * study3) # maybe don't remove?
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(final_model)
```

### Testing Phase

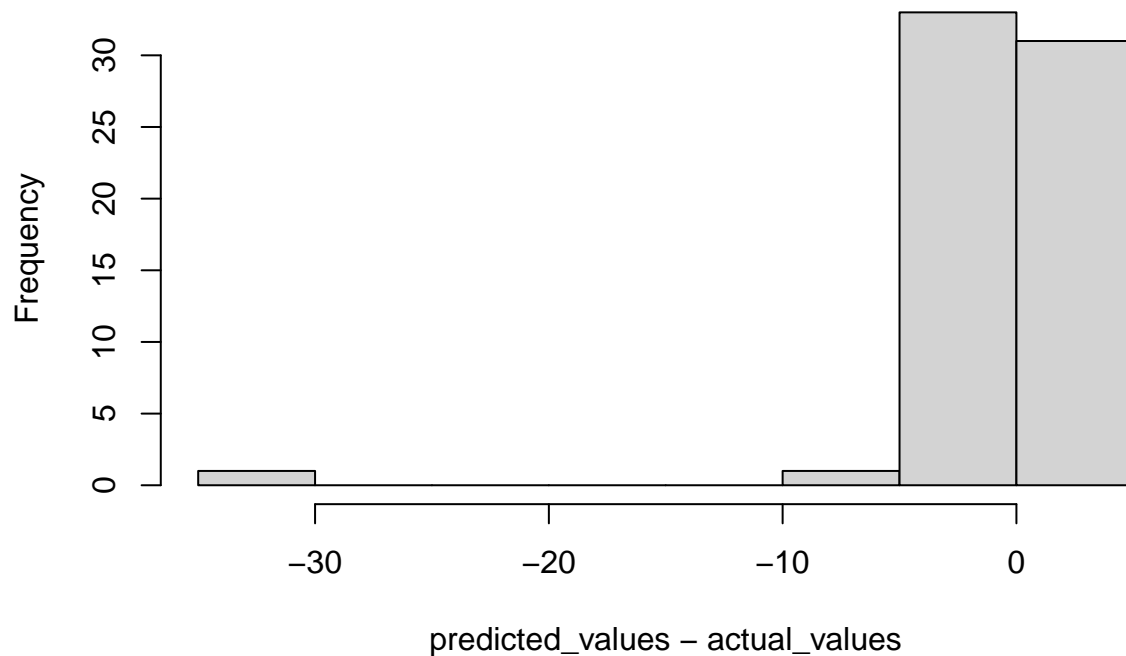
```
mean(predicted_values - actual_values)
```

```
## [1] -0.7223123
```

```
median(predicted_values - actual_values)
```

```
## [1] -0.1126939
```

## Histogram of predicted\_values – actual\_values



## One sample t-test on Mean

- $H_0 : \mu_{residuals} = 0$
- $H_1 : \mu_{residuals} \neq 0$
- is -0.427222 statistically different from 0?
- the p-value should be small.
- $n = 77$ , so by CLT sample mean is approximately normal

```
t.test(predicted_values - actual_values)
```

```
##  
## One Sample t-test  
##  
## data: predicted_values - actual_values  
## t = -1.263, df = 65, p-value = 0.2111  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -1.8644405 0.4198159  
## sample estimates:  
## mean of x  
## -0.7223123
```

- p-value = 0.9693
- t-value = -0.038679