

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)) %>%  
  
  # Remove the "X" column: it's simply the row number, which isn't very useful.  
  # Remove the "Country" column: column "country" already exists  
  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)  
  
  # Identify rows with no quiz 4.  
  # These indicate students who have dropped STA302H1, and who  
  # should be excluded from the final data.
```

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.

```
at_least_one_NA = function(data) {  
  return (rowSums(is.na(cleaned_sta302_performance_data)) >= 1)  
}
```

```
rows_with_some_NAs = cleaned_sta302_performance_data[  
  at_least_one_NA(cleaned_sta302_performance_data),  
]
```

Rows with Mistyped Columns

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

```
rows_with_mistyped_columns = cleaned_sta302_performance_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
```

```
# 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_data %>%  
  select(-country)
```

Find Significance Predictor Variables, Select Predictor Variables Based on Criterion

```
# 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.
```

Histograms

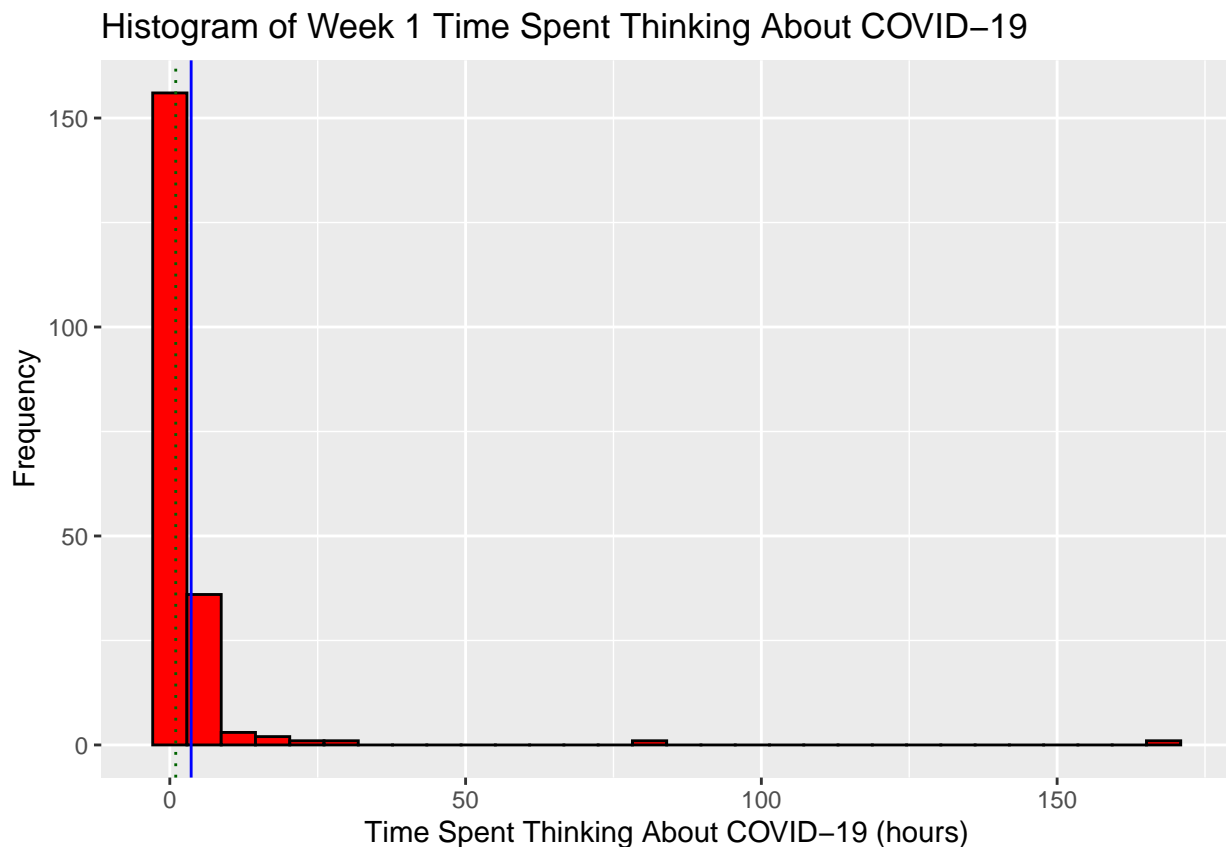
```
# TODO: See toy programs to learn how to create histograms with ggplot.
# TODO: See Demo 1 to figure out how to add histograms using grid.arrange.

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")
}
```

Histograms of COVID Hours

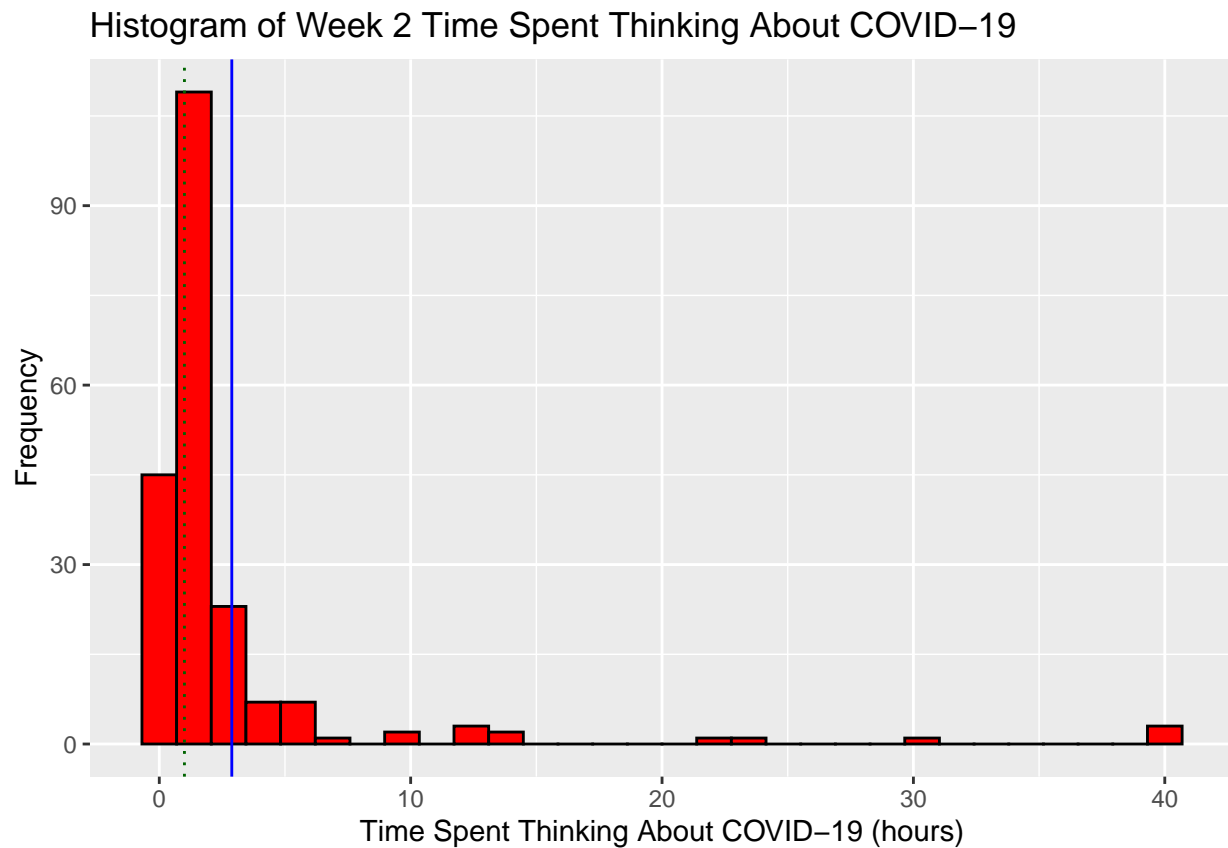
```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$COVID.hours..W1.,
                  "Histogram of Week 1 Time Spent Thinking About COVID-19",
                  "Time Spent Thinking About COVID-19 (hours)")
```

```
## Warning: Removed 26 rows containing non-finite values (stat_bin).
```



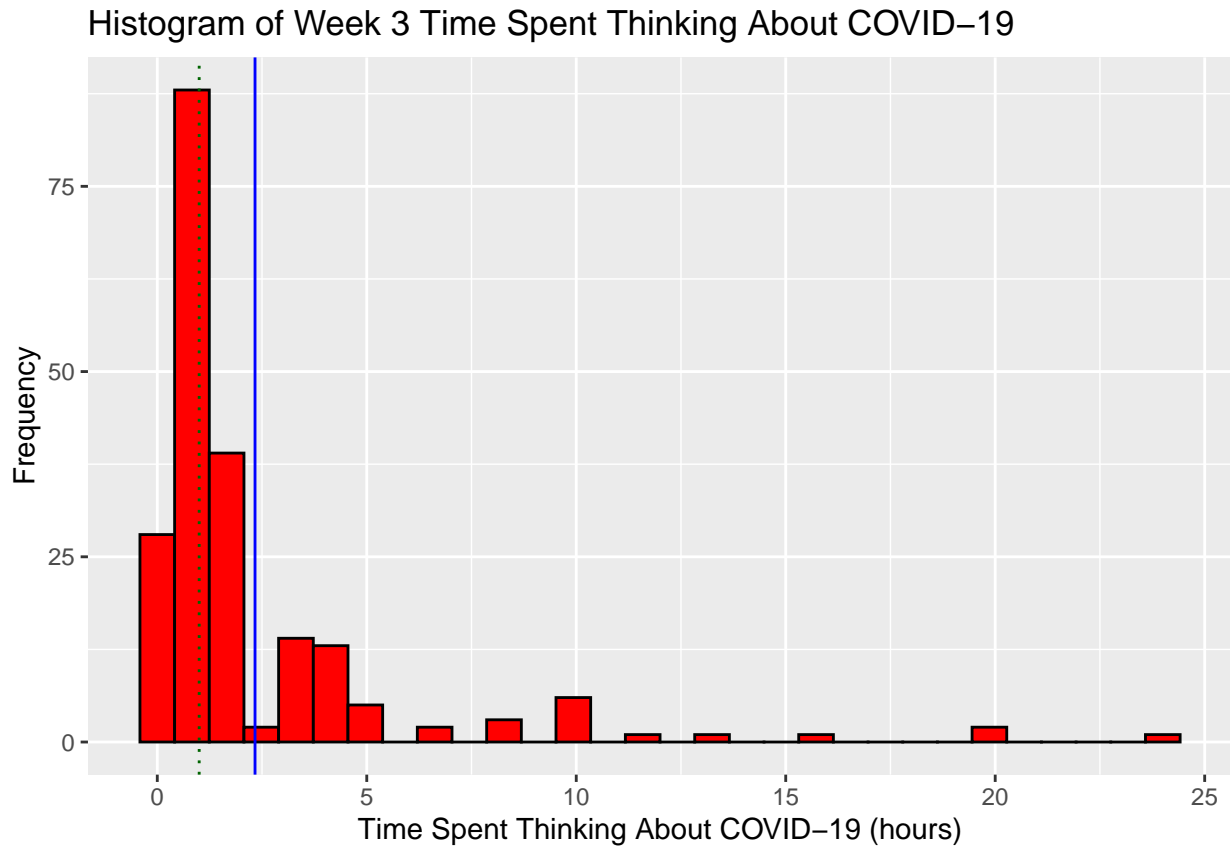
```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$COVID.hours..W2.,
                  "Histogram of Week 2 Time Spent Thinking About COVID-19",
                  "Time Spent Thinking About COVID-19 (hours)")
```

Warning: Removed 22 rows containing non-finite values (stat_bin).



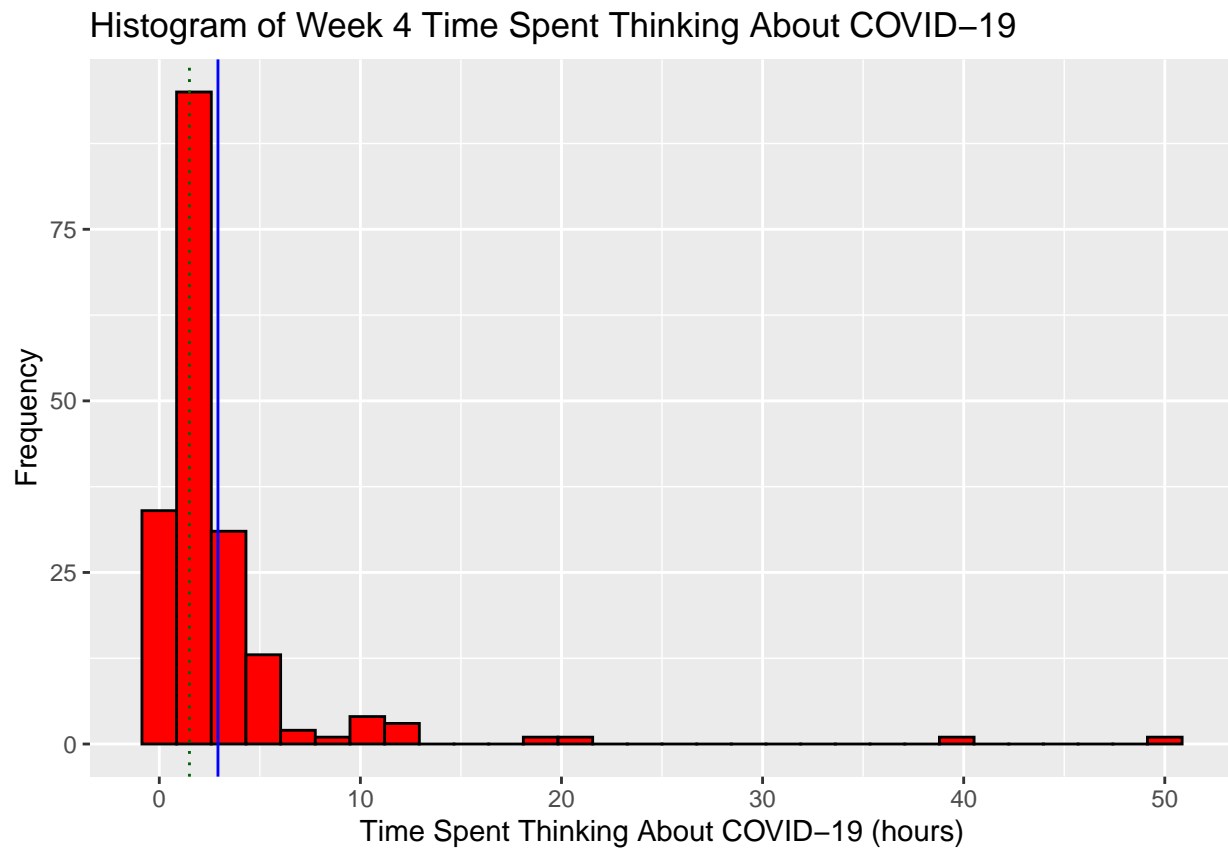
```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$COVID.hours..W3.,
                  "Histogram of Week 3 Time Spent Thinking About COVID-19",
                  "Time Spent Thinking About COVID-19 (hours)")
```

Warning: Removed 21 rows containing non-finite values (stat_bin).




```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$COVID.hours..W4.,
                  "Histogram of Week 4 Time Spent Thinking About COVID-19",
                  "Time Spent Thinking About COVID-19 (hours)")
```

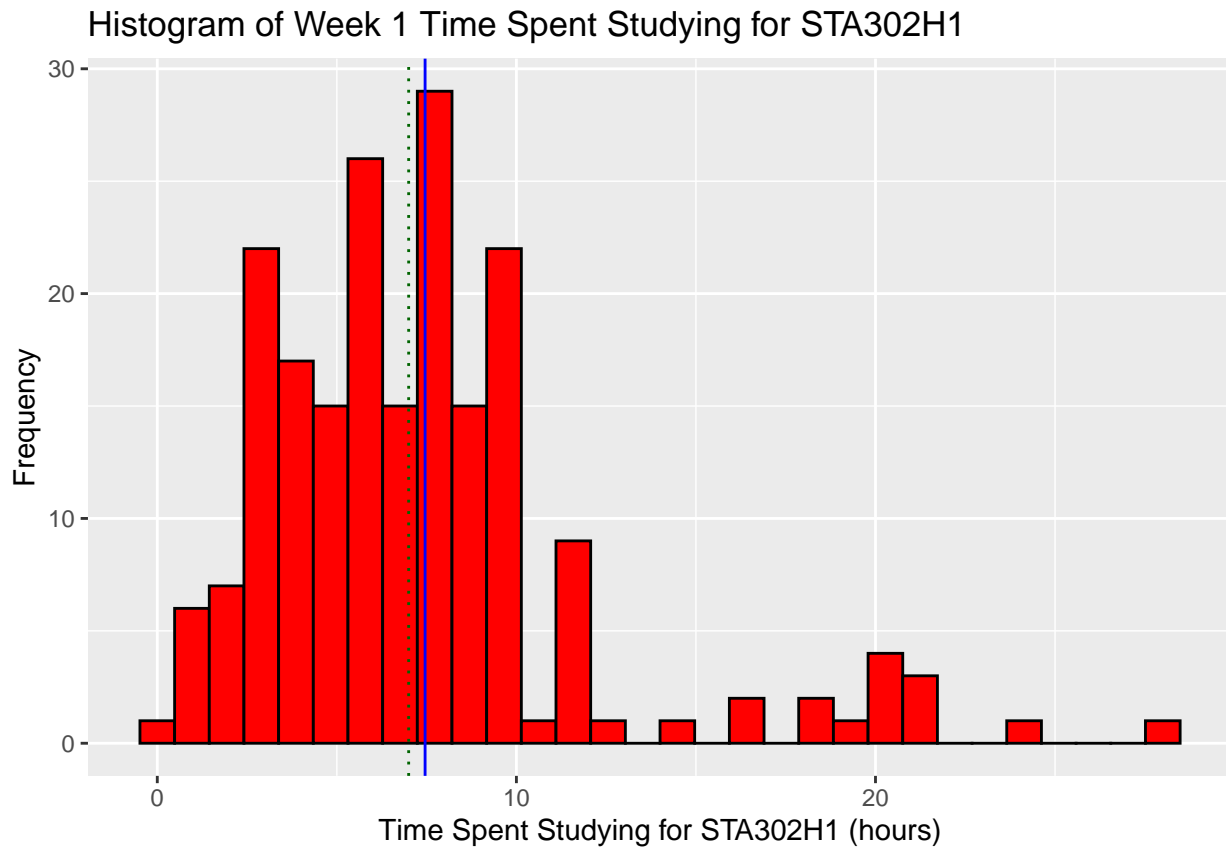
Warning: Removed 40 rows containing non-finite values (stat_bin).



Histograms of STA302H1 Hours

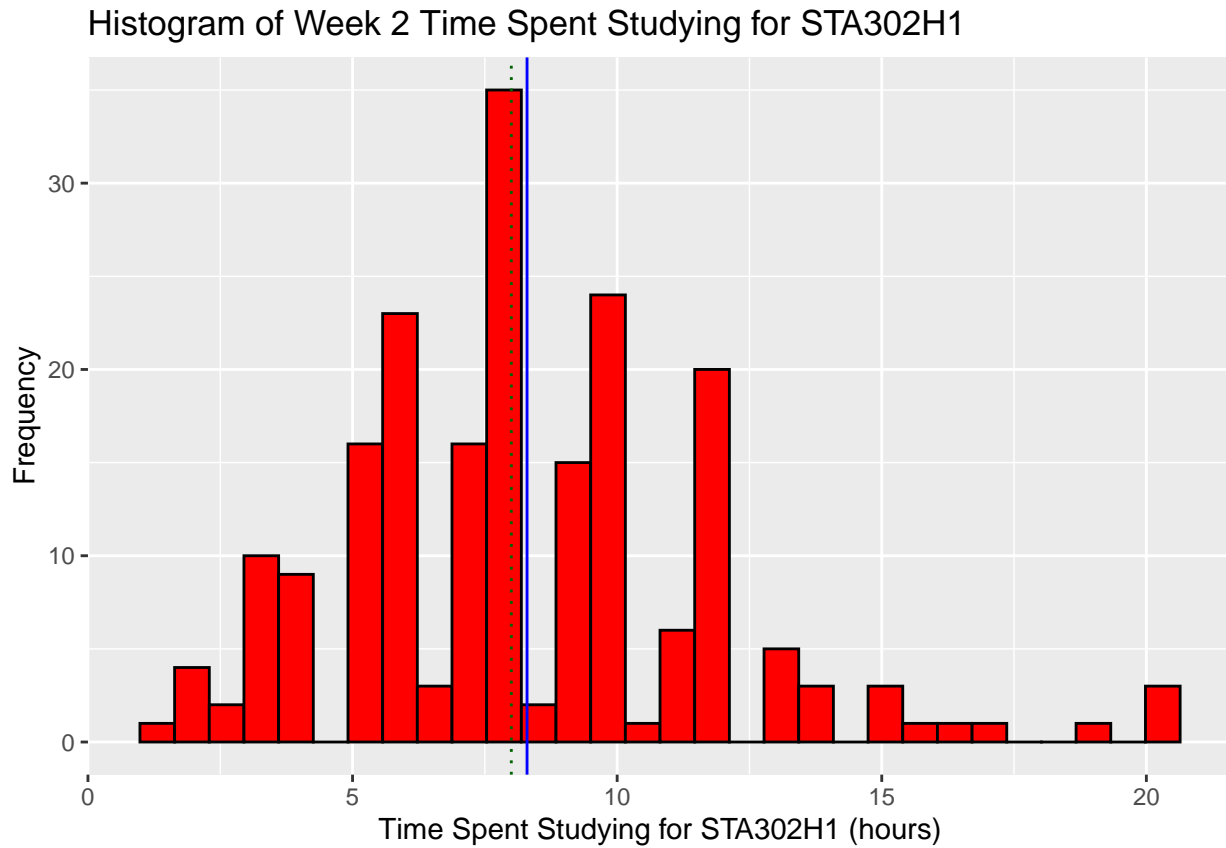
```
display_histogram(cleaned_sta302_performance_data,  
                  cleaned_sta302_performance_data$STA302.hours..W1.,  
                  "Histogram of Week 1 Time Spent Studying for STA302H1",  
                  "Time Spent Studying for STA302H1 (hours)")
```

Warning: Removed 26 rows containing non-finite values (stat_bin).



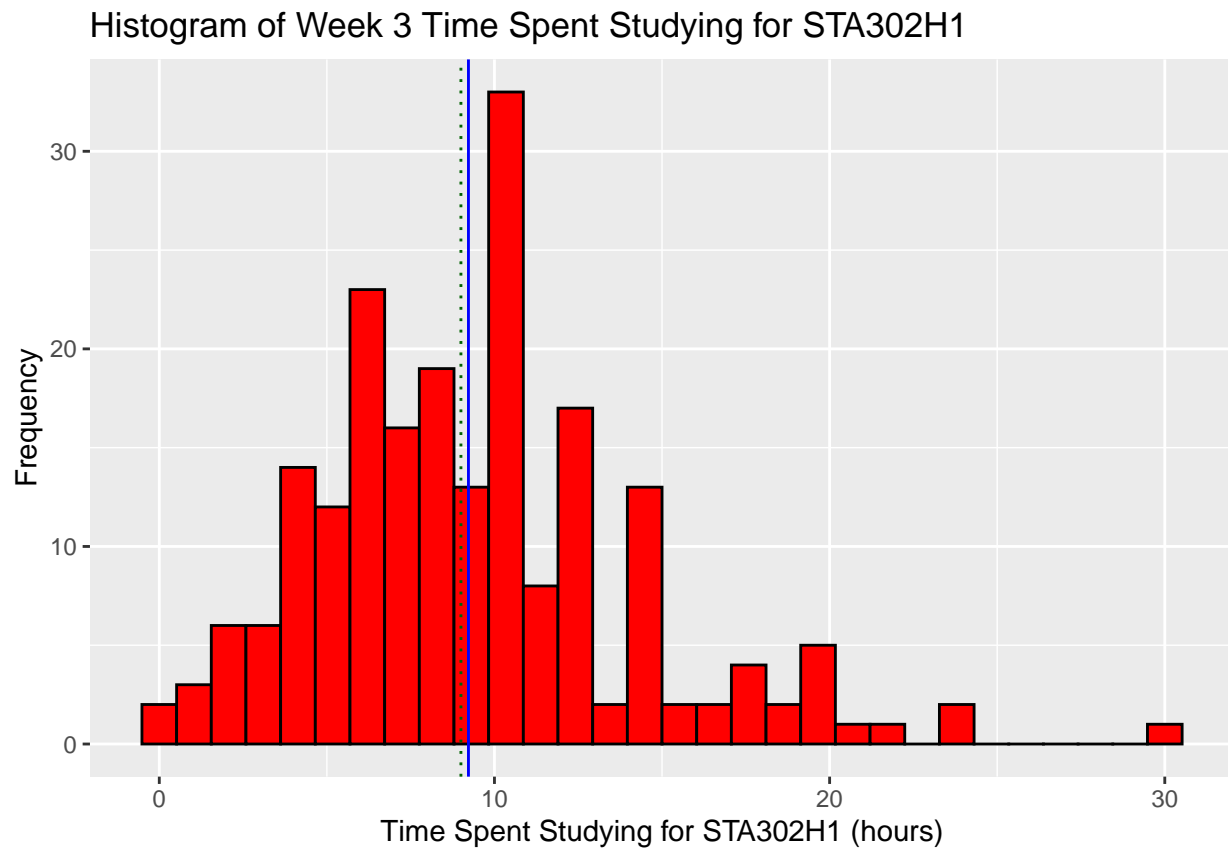
```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$STA302.hours..W2.,
                  "Histogram of Week 2 Time Spent Studying for STA302H1",
                  "Time Spent Studying for STA302H1 (hours)")
```

Warning: Removed 22 rows containing non-finite values (stat_bin).



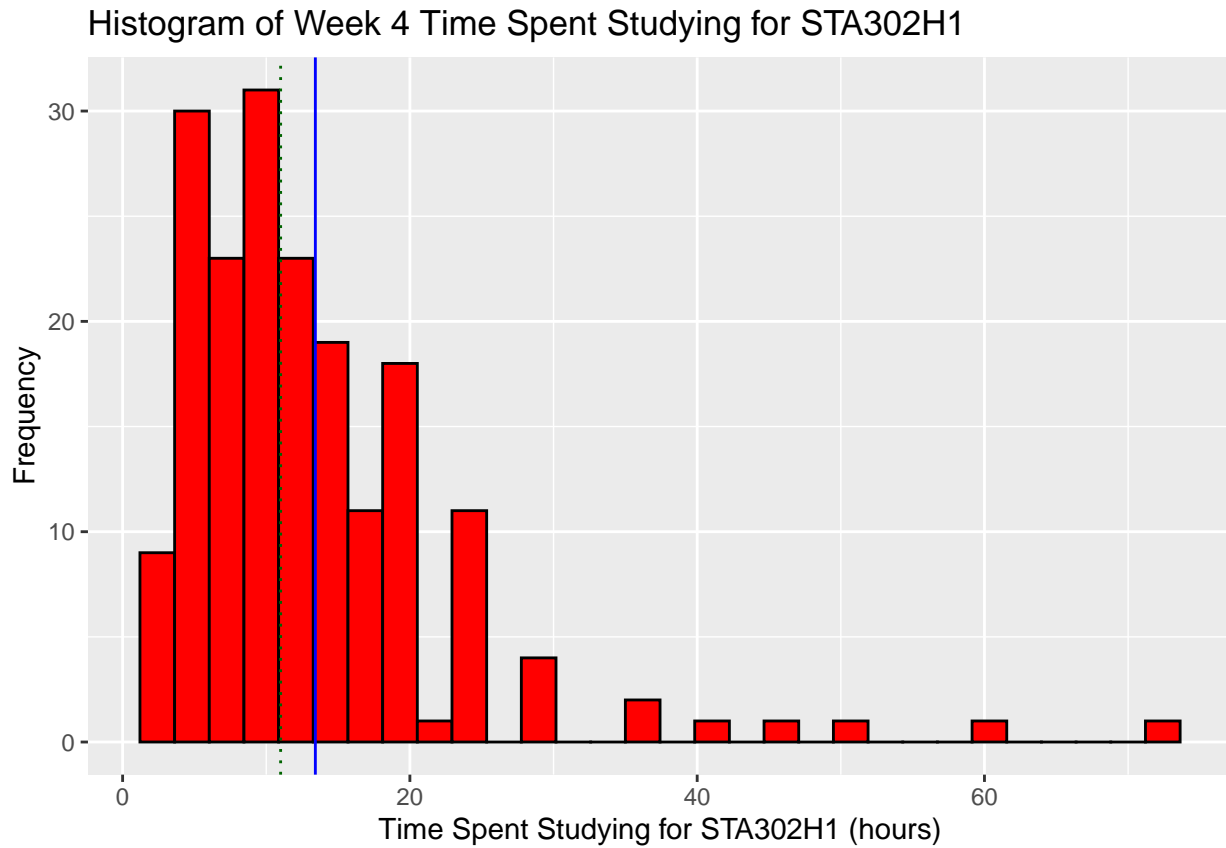
```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$STA302.hours..W3.,
                  "Histogram of Week 3 Time Spent Studying for STA302H1",
                  "Time Spent Studying for STA302H1 (hours)")
```

Warning: Removed 20 rows containing non-finite values (stat_bin).



```
display_histogram(cleaned_sta302_performance_data,  
                  cleaned_sta302_performance_data$STA302.hours..W4.,  
                  "Histogram of Week 4 Time Spent Studying for STA302H1",  
                  "Time Spent Studying for STA302H1 (hours)")
```

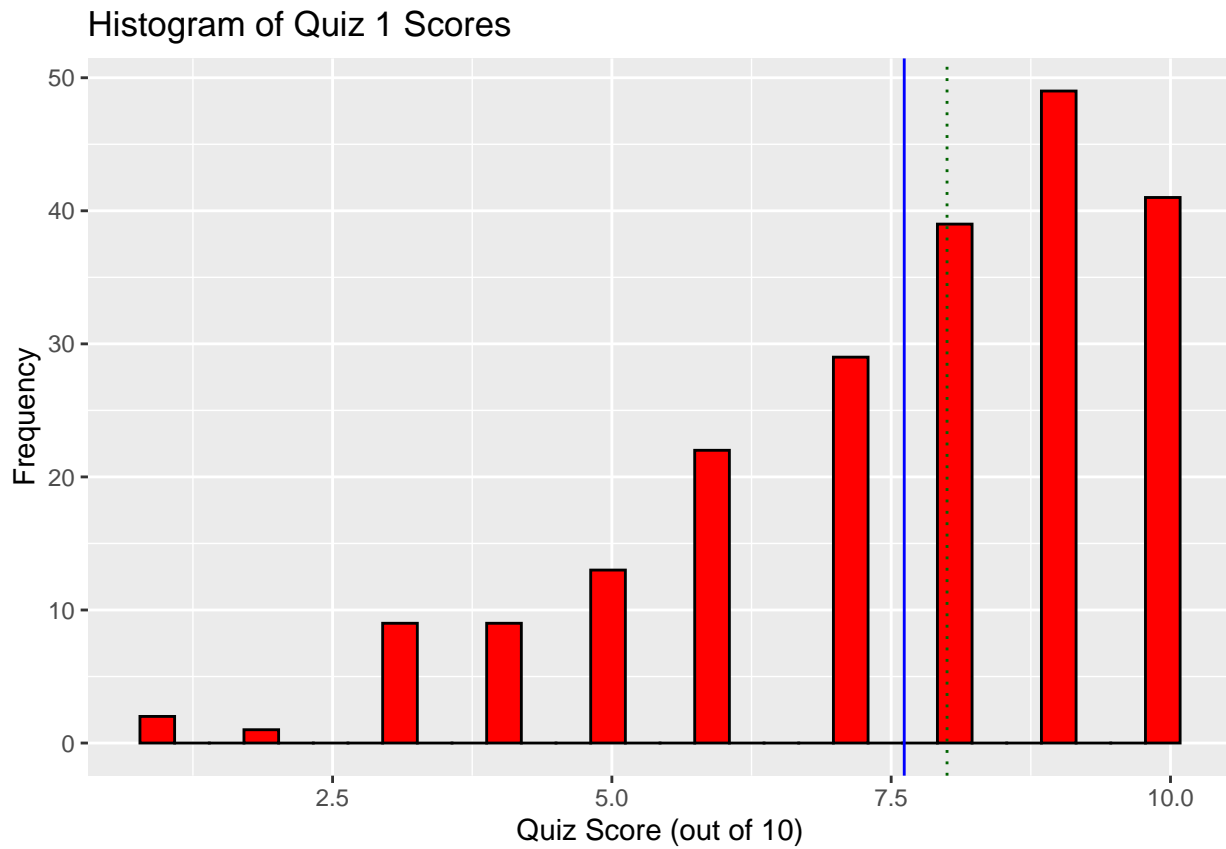
Warning: Removed 40 rows containing non-finite values (stat_bin).



Histograms of Quiz Scores

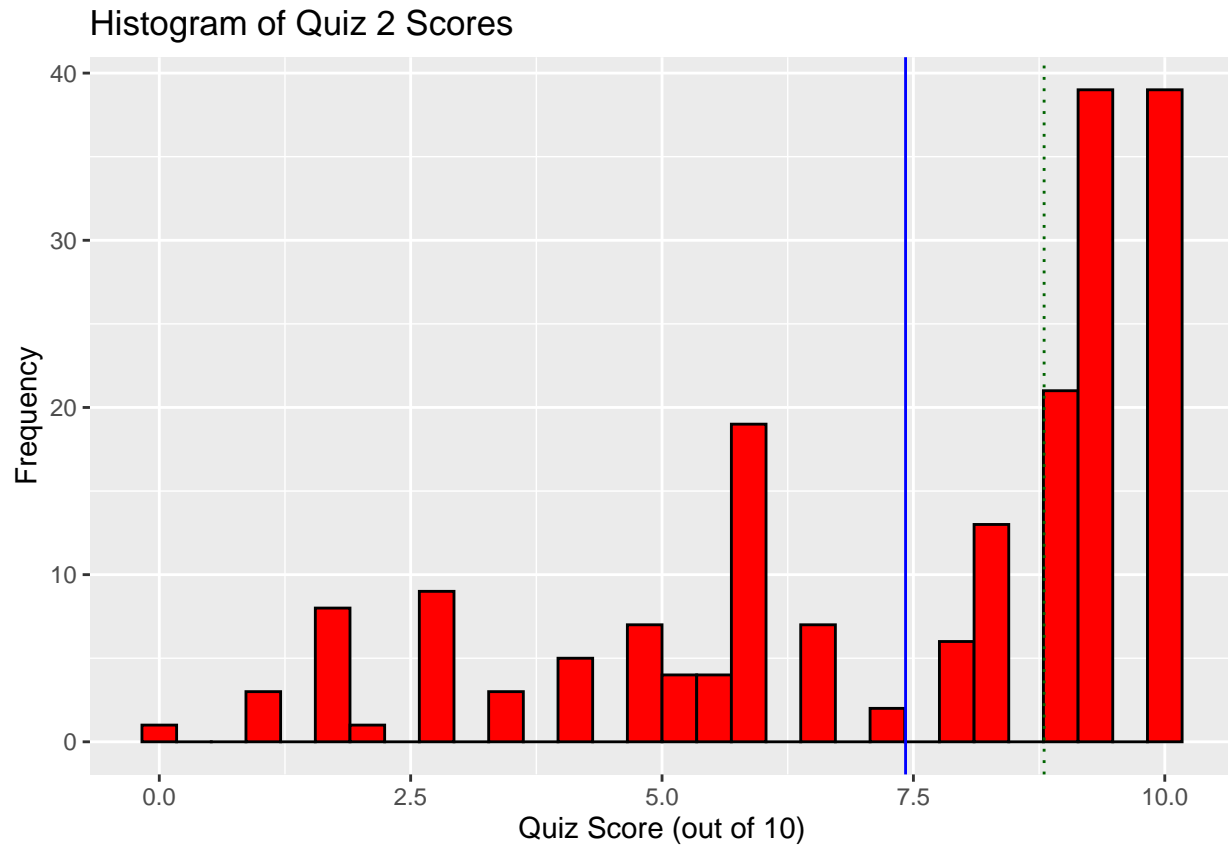
```
display_histogram(cleaned_sta302_performance_data,  
                  cleaned_sta302_performance_data$Quiz_1_score,  
                  "Histogram of Quiz 1 Scores",  
                  "Quiz Score (out of 10)")
```

Warning: Removed 13 rows containing non-finite values (stat_bin).



```
display_histogram(cleaned_sta302_performance_data,
                  cleaned_sta302_performance_data$Quiz_2_score,
                  "Histogram of Quiz 2 Scores",
                  "Quiz Score (out of 10)")
```

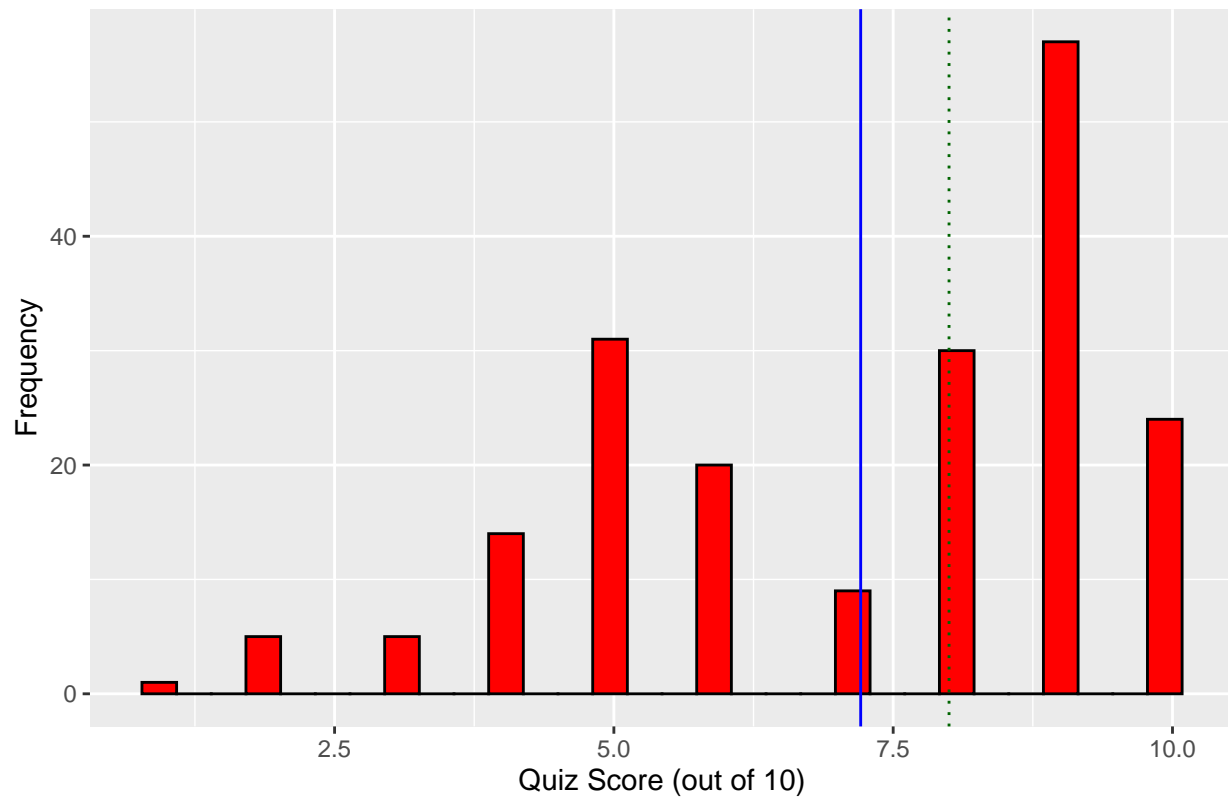
Warning: Removed 36 rows containing non-finite values (stat_bin).



```
display_histogram(cleaned_sta302_performance_data,  
                  cleaned_sta302_performance_data$Quiz_3_score,  
                  "Histogram of Quiz 3 Scores",  
                  "Quiz Score (out of 10)")
```

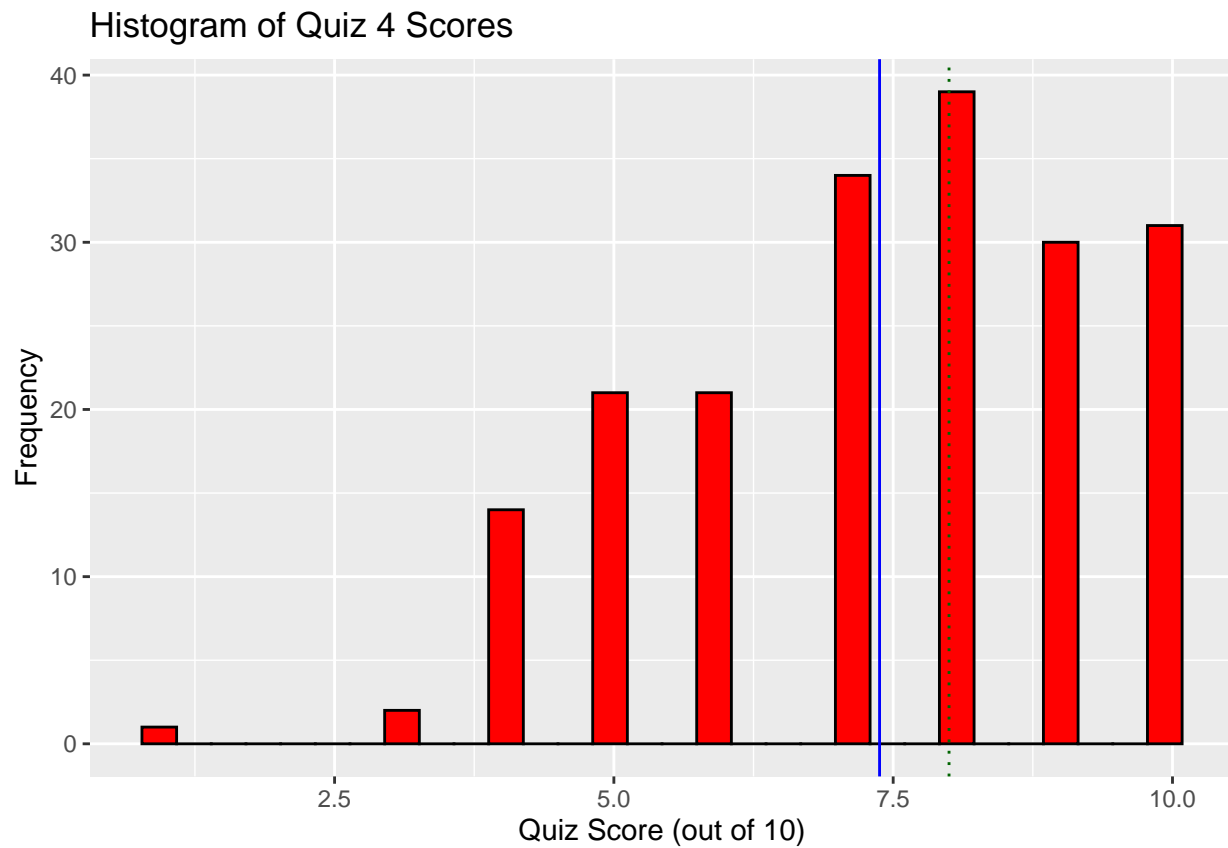
Warning: Removed 31 rows containing non-finite values (stat_bin).

Histogram of Quiz 3 Scores




```
display_histogram(cleaned_sta302_performance_data,  
                  cleaned_sta302_performance_data$Quiz_4_score,  
                  "Histogram of Quiz 4 Scores",  
                  "Quiz Score (out of 10)")
```

Warning: Removed 34 rows containing non-finite values (stat_bin).



Boxplots

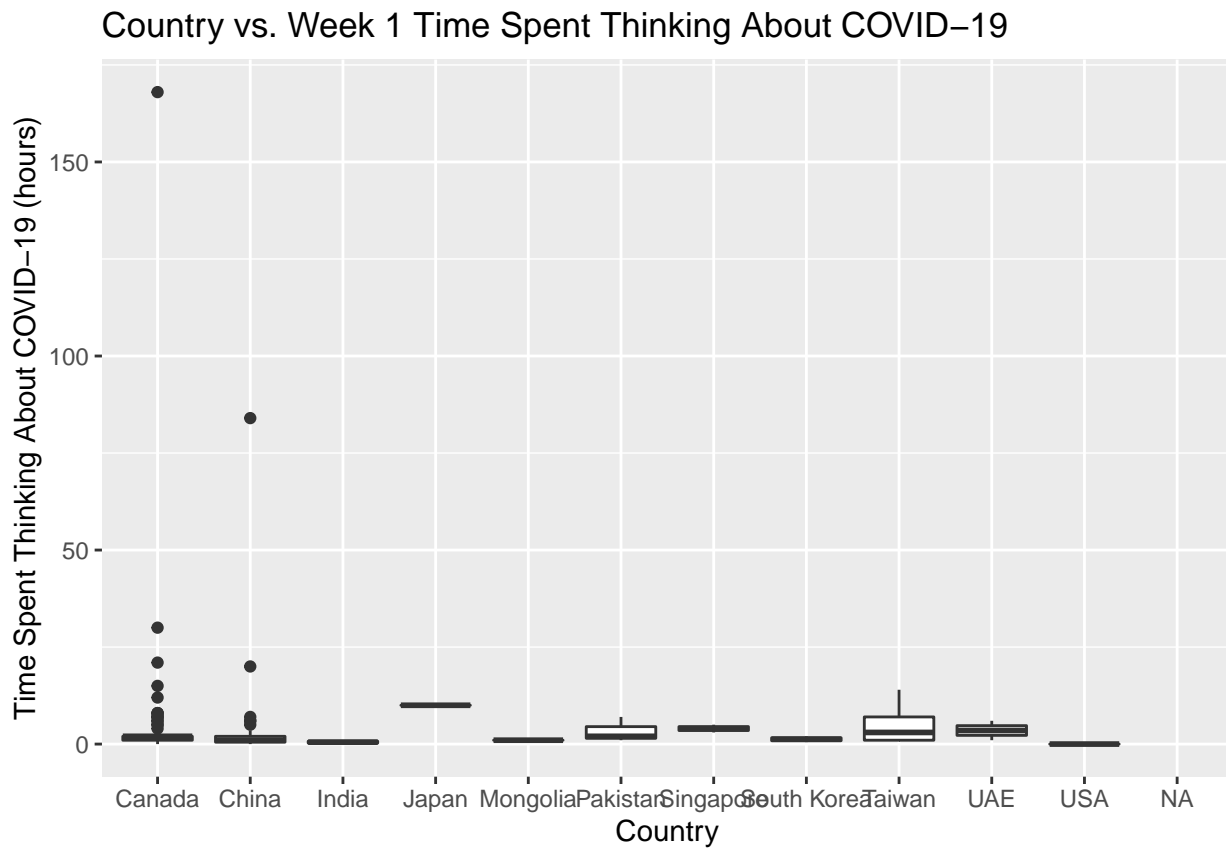
```
# TODO: See STA248H1 notes to figure out how to create boxplots. -- DONE
# TODO: See toy program of boxplots to see how to color them by factor

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

Boxplots of COVID Hours

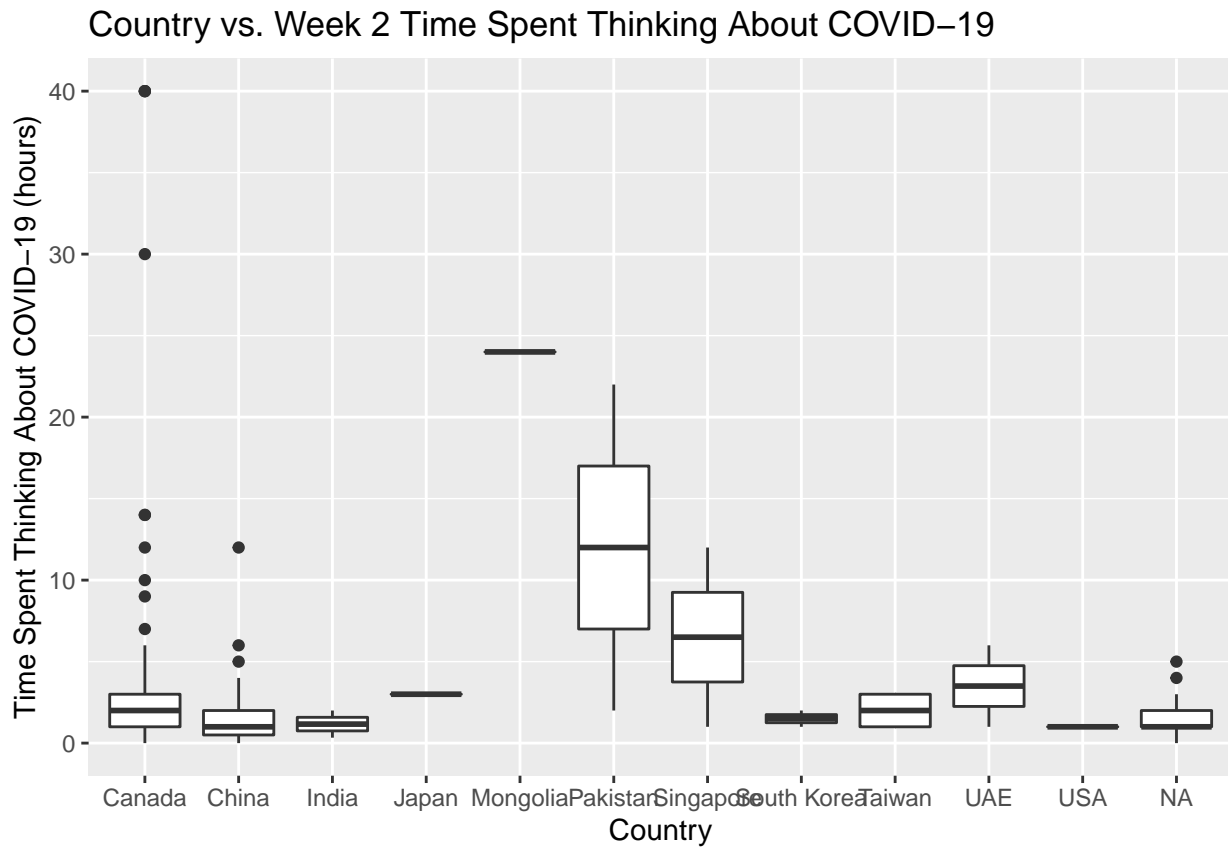
```
display_boxplot(cleaned_sta302_performance_data, COVID.hours..W1.,
  "Country vs. Week 1 Time Spent Thinking About COVID-19",
  "Time Spent Thinking About COVID-19 (hours)")
```

```
## Warning: Removed 26 rows containing non-finite values (stat_boxplot).
```



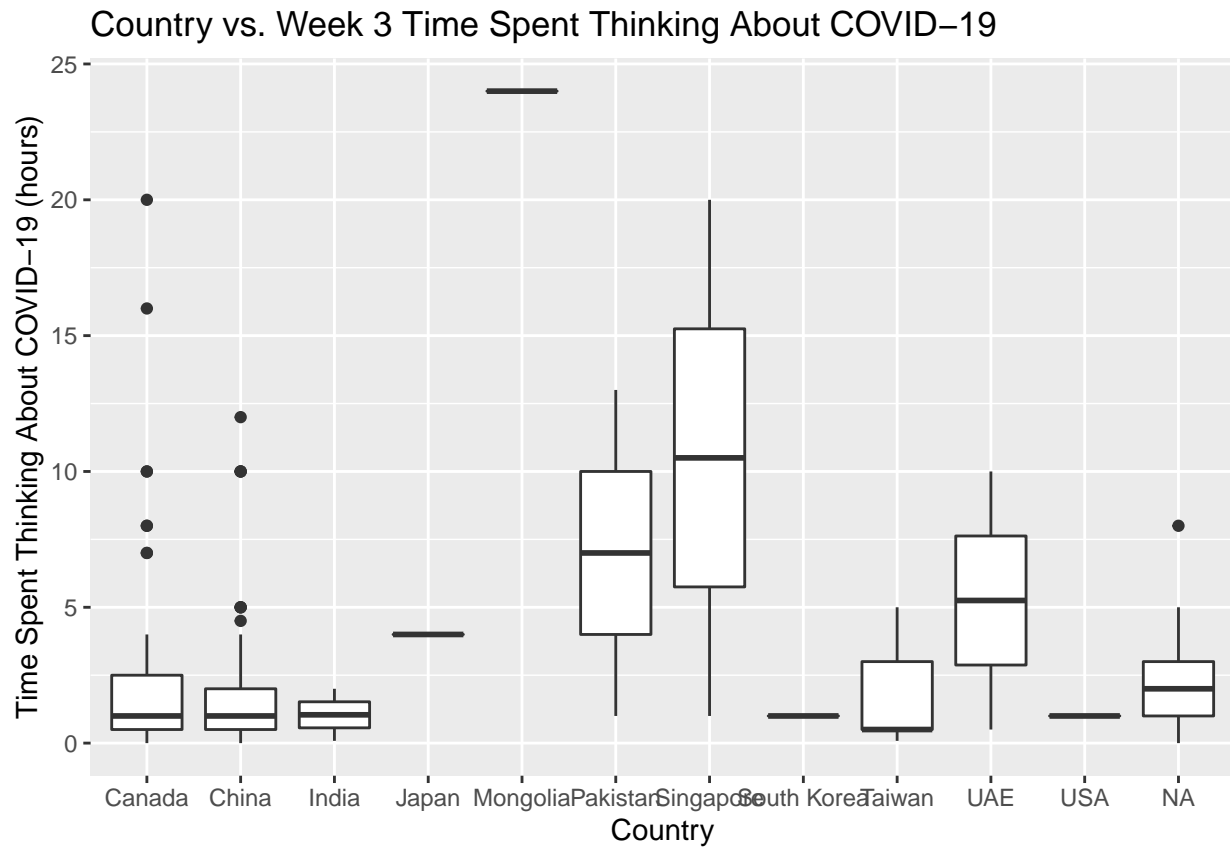
```
display_boxplot(cleaned_sta302_performance_data, COVID.hours..W2.,
                "Country vs. Week 2 Time Spent Thinking About COVID-19",
                "Time Spent Thinking About COVID-19 (hours)")
```

```
## Warning: Removed 22 rows containing non-finite values (stat_boxplot).
```



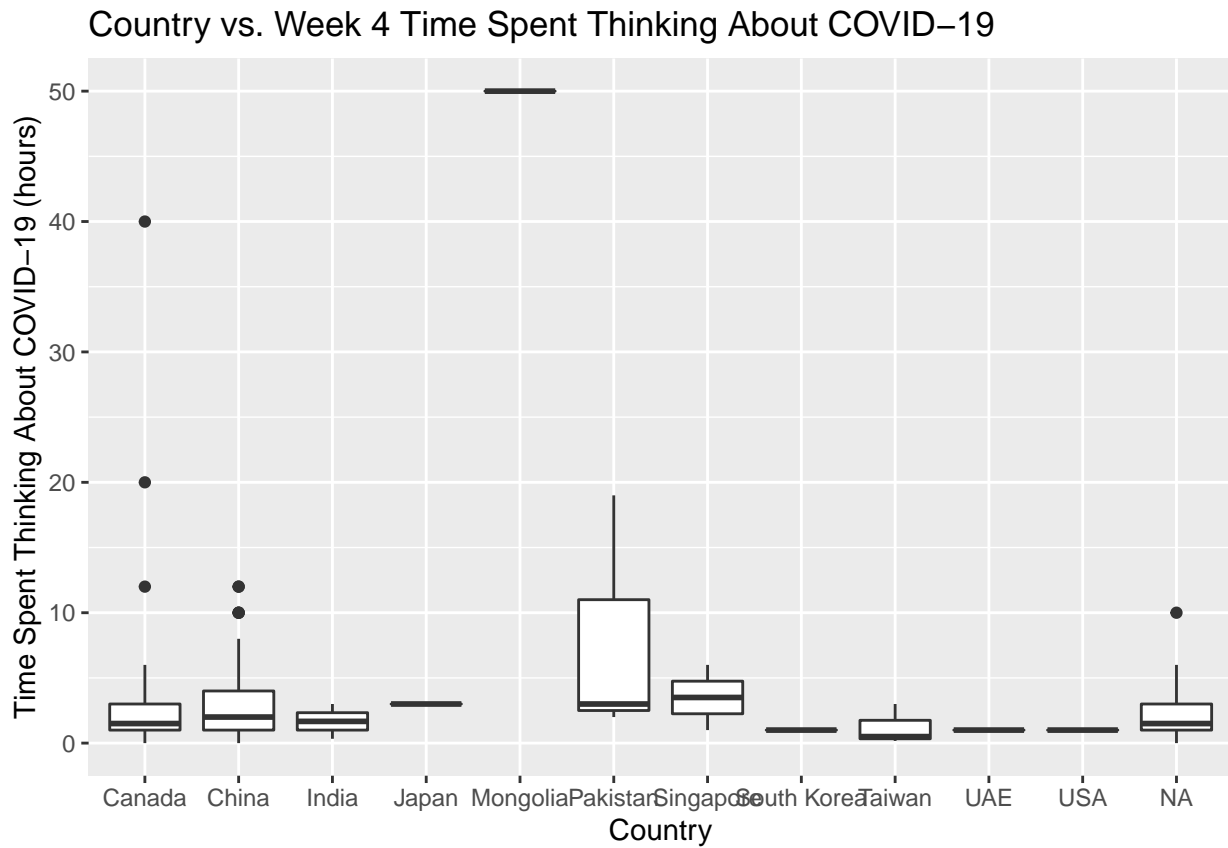
```
display_boxplot(cleaned_sta302_performance_data, COVID.hours..W3.,
                "Country vs. Week 3 Time Spent Thinking About COVID-19",
                "Time Spent Thinking About COVID-19 (hours)")
```

```
## Warning: Removed 21 rows containing non-finite values (stat_boxplot).
```



```
display_boxplot(cleaned_sta302_performance_data, COVID.hours..W4.,
                "Country vs. Week 4 Time Spent Thinking About COVID-19",
                "Time Spent Thinking About COVID-19 (hours)")
```

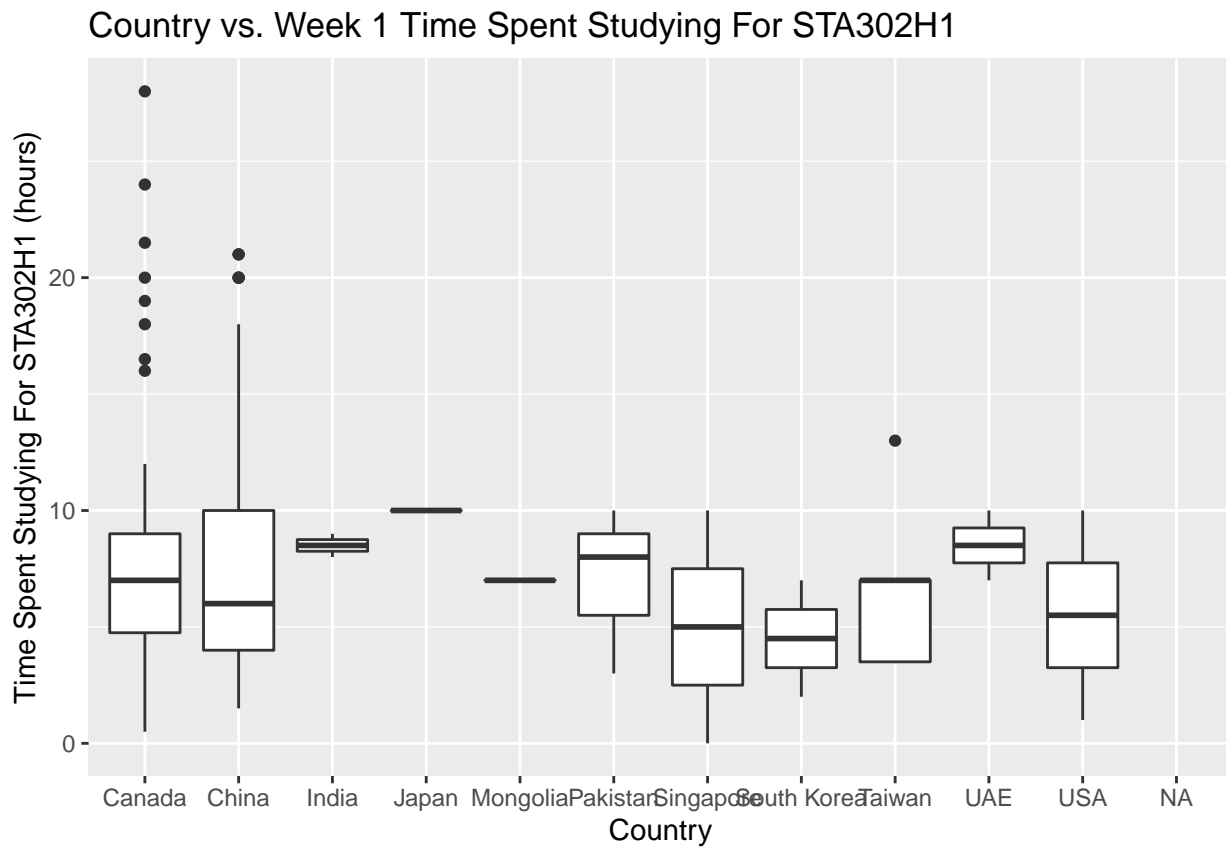
```
## Warning: Removed 40 rows containing non-finite values (stat_boxplot).
```



Boxplots of STA302H1 Hours

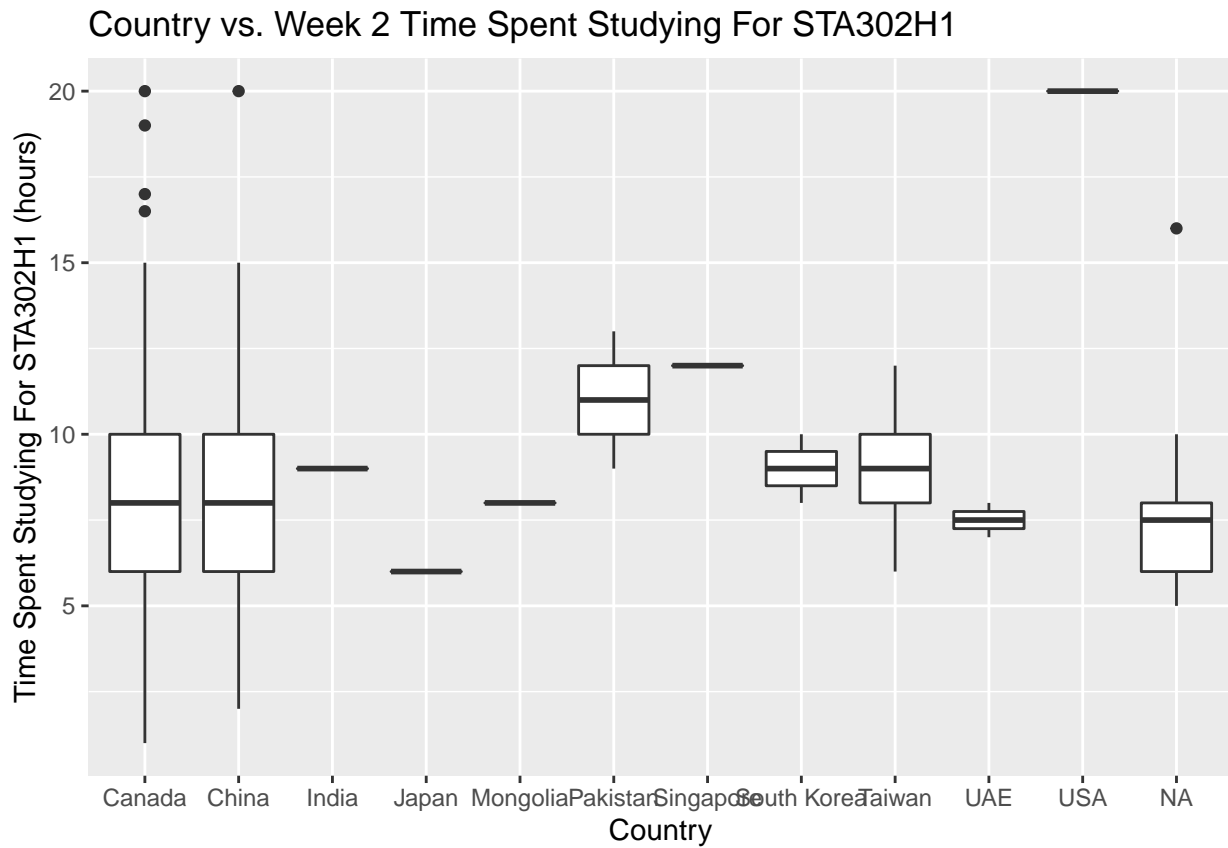
```
display_boxplot(cleaned_sta302_performance_data, STA302.hours..W1.,  
                "Country vs. Week 1 Time Spent Studying For STA302H1",  
                "Time Spent Studying For STA302H1 (hours)")
```

```
## Warning: Removed 26 rows containing non-finite values (stat_boxplot).
```



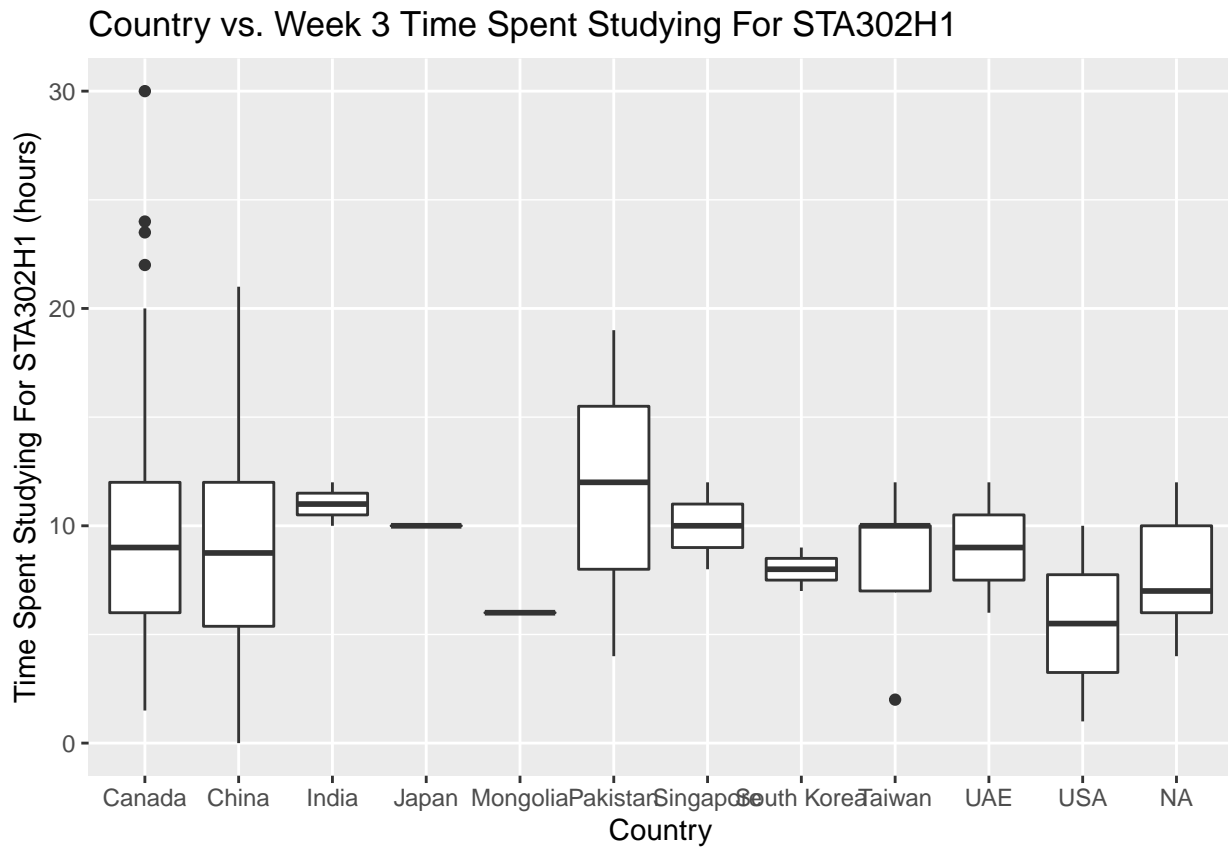
```
display_boxplot(cleaned_sta302_performance_data, STA302.hours..W2.,
                "Country vs. Week 2 Time Spent Studying For STA302H1",
                "Time Spent Studying For STA302H1 (hours)")
```

Warning: Removed 22 rows containing non-finite values (stat_boxplot).



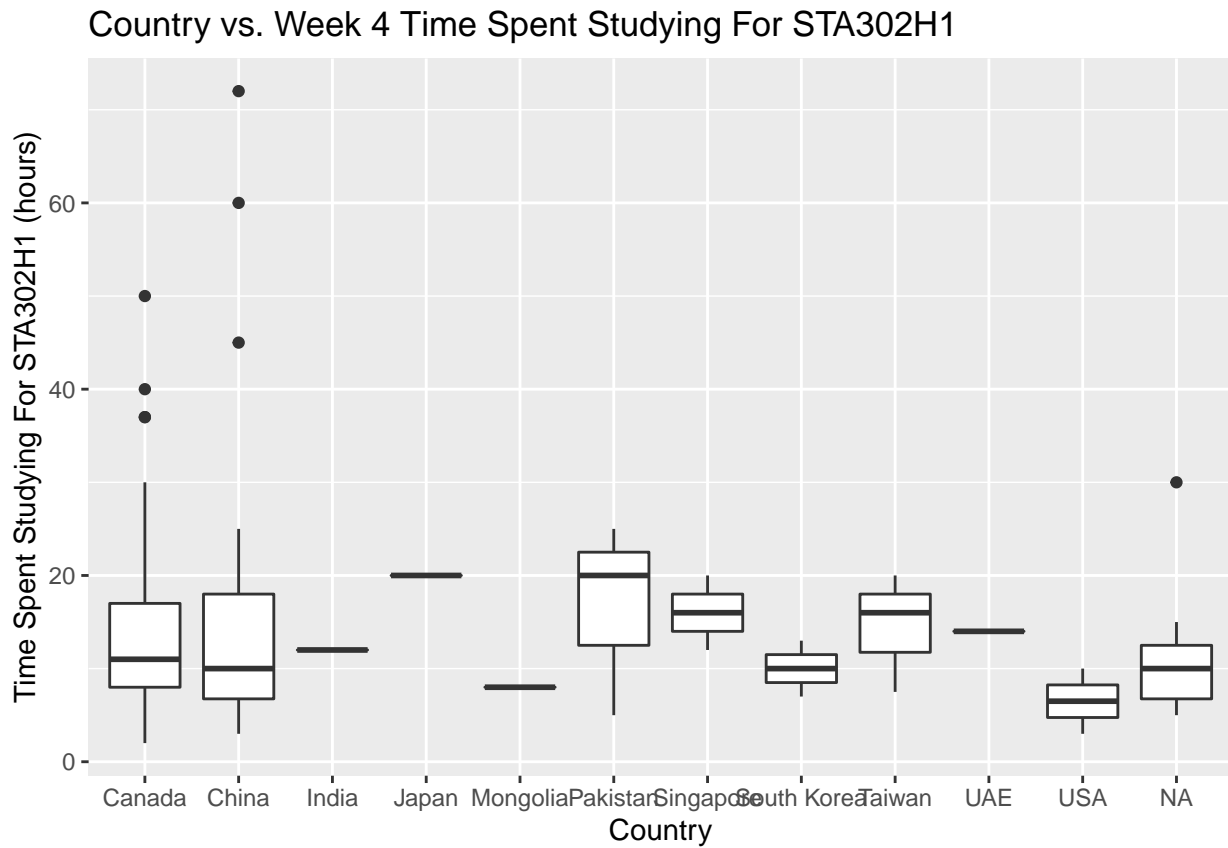
```
display_boxplot(cleaned_sta302_performance_data, STA302.hours..W3.,
                "Country vs. Week 3 Time Spent Studying For STA302H1",
                "Time Spent Studying For STA302H1 (hours)")
```

Warning: Removed 20 rows containing non-finite values (stat_boxplot).




```
display_boxplot(cleaned_sta302_performance_data, STA302.hours..W4.,
                "Country vs. Week 4 Time Spent Studying For STA302H1",
                "Time Spent Studying For STA302H1 (hours)")
```

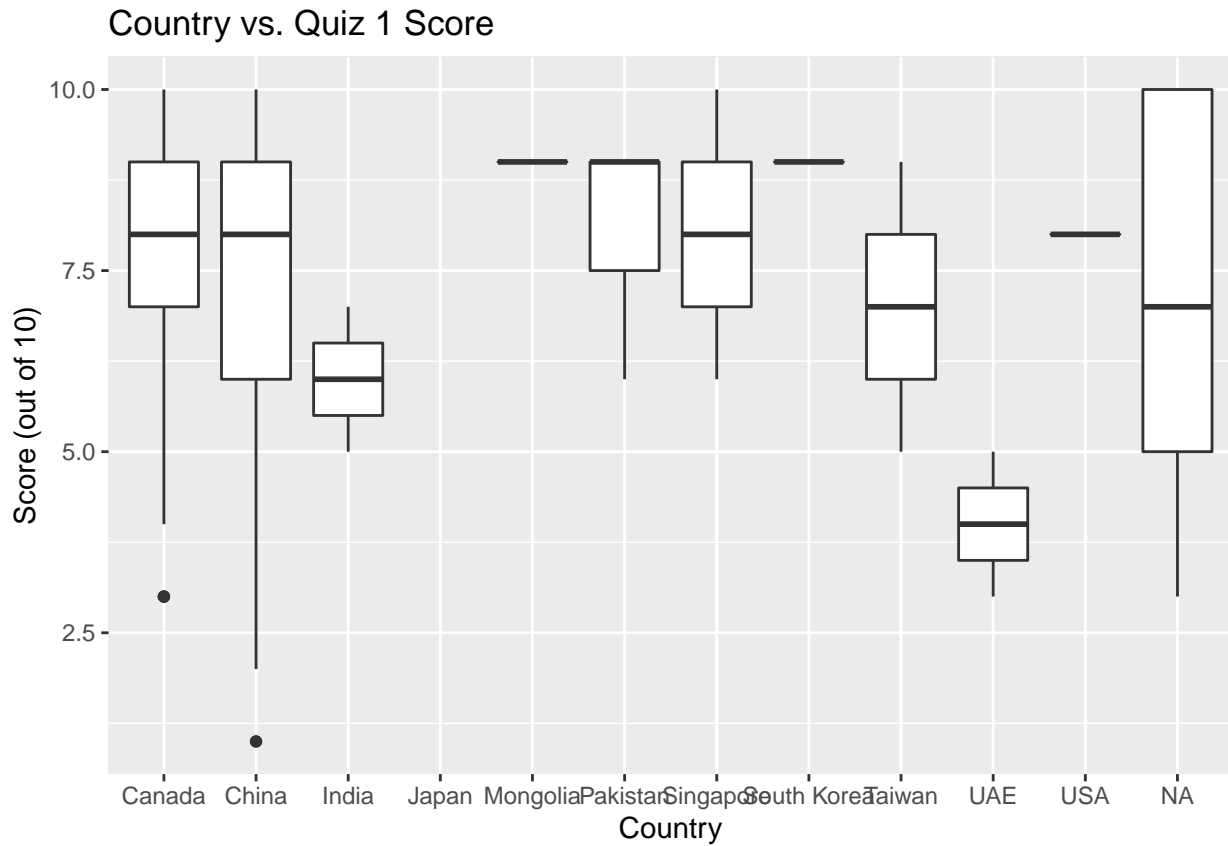
Warning: Removed 40 rows containing non-finite values (stat_boxplot).



Boxplots of Quiz Scores

```
display_boxplot(cleaned_sta302_performance_data, Quiz_1_score,  
                "Country vs. Quiz 1 Score", "Score (out of 10)")
```

```
## Warning: Removed 13 rows containing non-finite values (stat_boxplot).
```



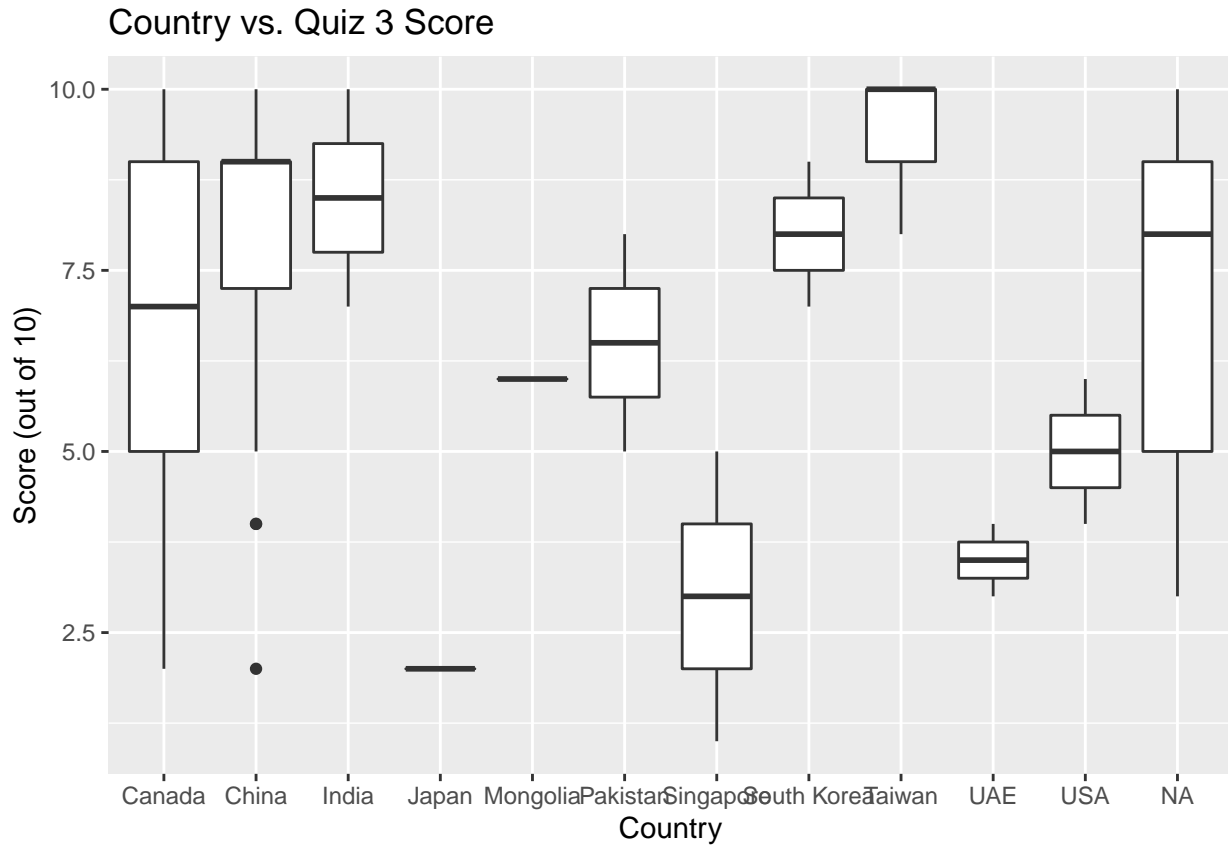
```
display_boxplot(cleaned_sta302_performance_data, Quiz_2_score,
                "Country vs. Quiz 2 Score", "Score (out of 10)")
```

Warning: Removed 36 rows containing non-finite values (stat_boxplot).



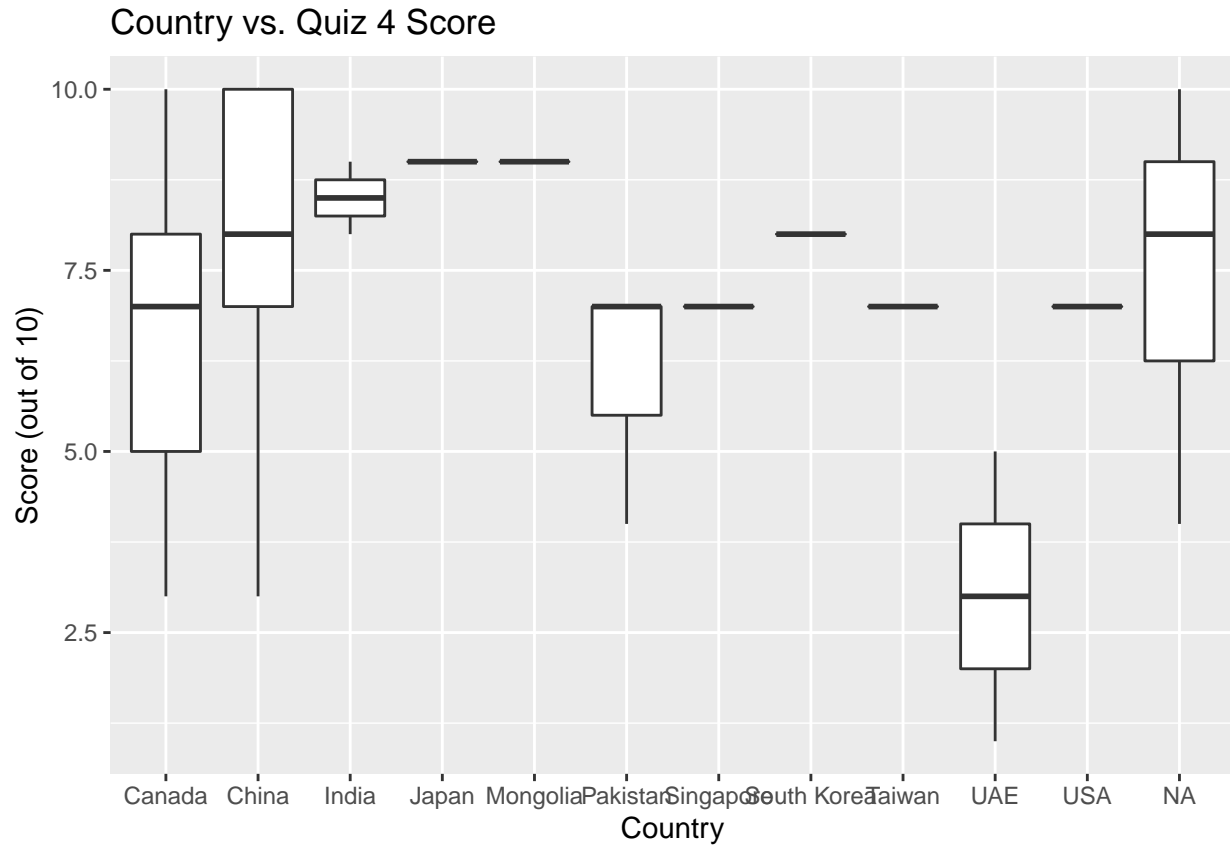
```
display_boxplot(cleaned_sta302_performance_data, Quiz_3_score,
                "Country vs. Quiz 3 Score", "Score (out of 10)")
```

Warning: Removed 31 rows containing non-finite values (stat_boxplot).



```
display_boxplot(cleaned_sta302_performance_data, Quiz_4_score,
                "Country vs. Quiz 4 Score", "Score (out of 10)")
```

Warning: Removed 34 rows containing non-finite values (stat_boxplot).

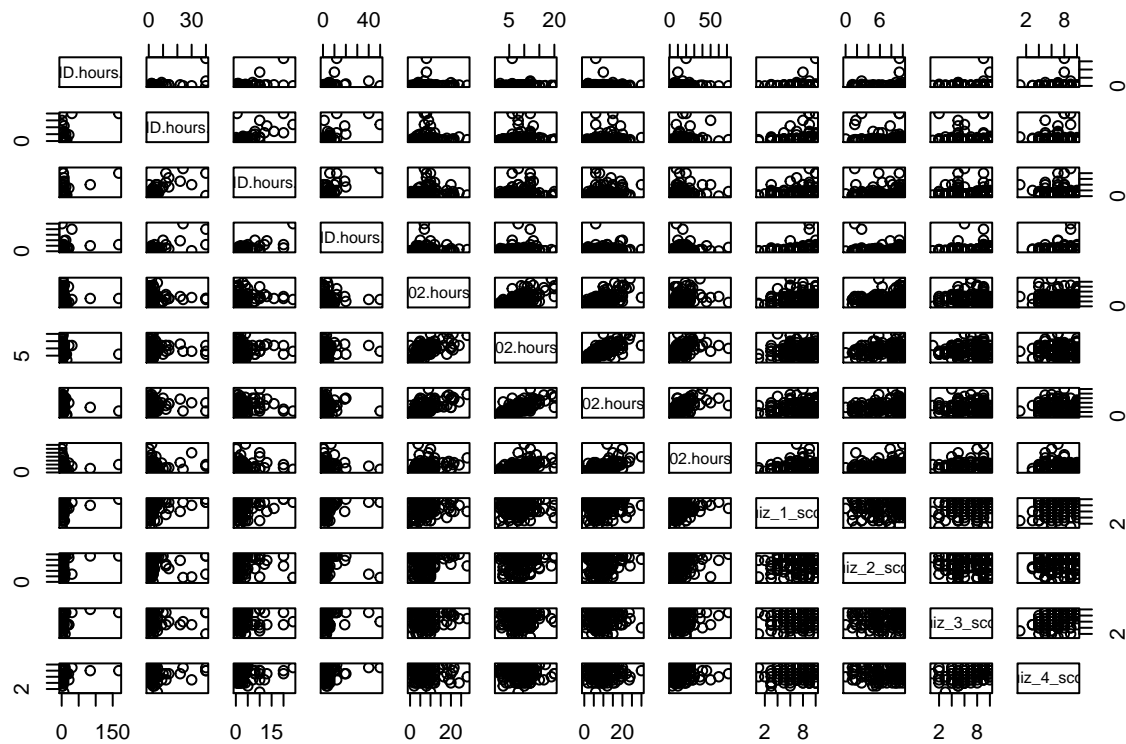


Scatterplots

TODO: How do I know which scatterplots are most important to focus on?

pairwise scatterplot

```
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_data)
```



Correlation Matrix

All Countries

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

```
# library(GGally)
colnames(rows_with_no_country) <- c("W1COV", "W2COV", "W3COV", "W4COV",
                                     "W1302", "W2302", "W3302", "W4302",
                                     "Q1", "Q2", "Q3", "Q4")
# ggcorr(rows_with_no_country, label = TRUE, label_round = 2)
round(cor(rows_with_no_country, use = "complete.obs"), 2) # TODO: na.rm = true
```

```
##      W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302  Q1   Q2   Q3   Q4
## W1COV  1.00  0.66  0.46  0.20  0.02 -0.04 -0.02  0.06  0.10  0.07  0.05  0.01
## W2COV  0.66  1.00  0.82  0.60  0.06  0.05  0.13  0.21  0.11 -0.10 -0.08 -0.06
## W3COV  0.46  0.82  1.00  0.73  0.06  0.09  0.14  0.13  0.13 -0.10 -0.11 -0.06
## W4COV  0.20  0.60  0.73  1.00  0.02  0.04  0.09  0.07  0.10 -0.09 -0.03  0.01
## W1302  0.02  0.06  0.06  0.02  1.00  0.61  0.57  0.31  0.02  0.11  0.03 -0.07
## W2302 -0.04  0.05  0.09  0.04  0.61  1.00  0.70  0.49 -0.04  0.08 -0.09 -0.12
## W3302 -0.02  0.13  0.14  0.09  0.57  0.70  1.00  0.62 -0.07  0.08 -0.14 -0.09
## W4302  0.06  0.21  0.13  0.07  0.31  0.49  0.62  1.00 -0.07  0.02 -0.05 -0.11
## Q1     0.10  0.11  0.13  0.10  0.02 -0.04 -0.07 -0.07  1.00  0.22  0.33  0.21
## Q2     0.07 -0.10 -0.10 -0.09  0.11  0.08  0.08  0.02  0.22  1.00  0.22  0.16
## Q3     0.05 -0.08 -0.11 -0.03  0.03 -0.09 -0.14 -0.05  0.33  0.22  1.00  0.54
## Q4     0.01 -0.06 -0.06  0.01 -0.07 -0.12 -0.09 -0.11  0.21  0.16  0.54  1.00
```

By Individual Country

```
# TODO: You could also create separate correlation matrices for each country.
```

5-Number Summary Statistics

STA302H1 Hours 5-Number Summary

```
summary(sta302_performance_data$STA302.hours..W1.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	4.000	7.000	7.458	9.000	28.000	26

```
summary(sta302_performance_data$STA302.hours..W2.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	1.000	6.000	8.000	8.298	10.000	20.000	22

```
summary(sta302_performance_data$STA302.hours..W3.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	6.000	9.000	9.225	11.500	30.000	20

```
summary(sta302_performance_data$STA302.hours..W4.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	2.00	7.00	11.00	13.42	16.00	72.00	40

COVID Hours 5-Number Summary

```
summary(sta302_performance_data$COVID.hours..W1.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	1.000	1.000	3.607	2.000	168.000	26

```
summary(sta302_performance_data$COVID.hours..W2.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	1.000	1.000	2.884	2.000	40.000	22

```
summary(sta302_performance_data$COVID.hours..W3.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	0.500	1.000	2.333	2.000	24.000	21

```
summary(sta302_performance_data$COVID.hours..W4.)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	1.000	1.500	2.918	3.000	50.000	40

Quiz Scores 5-Number Summary

```
summary(sta302_performance_data$Quiz_1_score)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	1.000	6.000	8.000	7.617	9.000	10.000	13

```
summary(sta302_performance_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	36

```
summary(sta302_performance_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	31

```
summary(sta302_performance_data$Quiz_4_score)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	1.000	6.000	8.000	7.378	9.000	10.000	34