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

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

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

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

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

```r
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")
}
```

```r
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)
}
```

```r
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)
}
```

```r
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
```
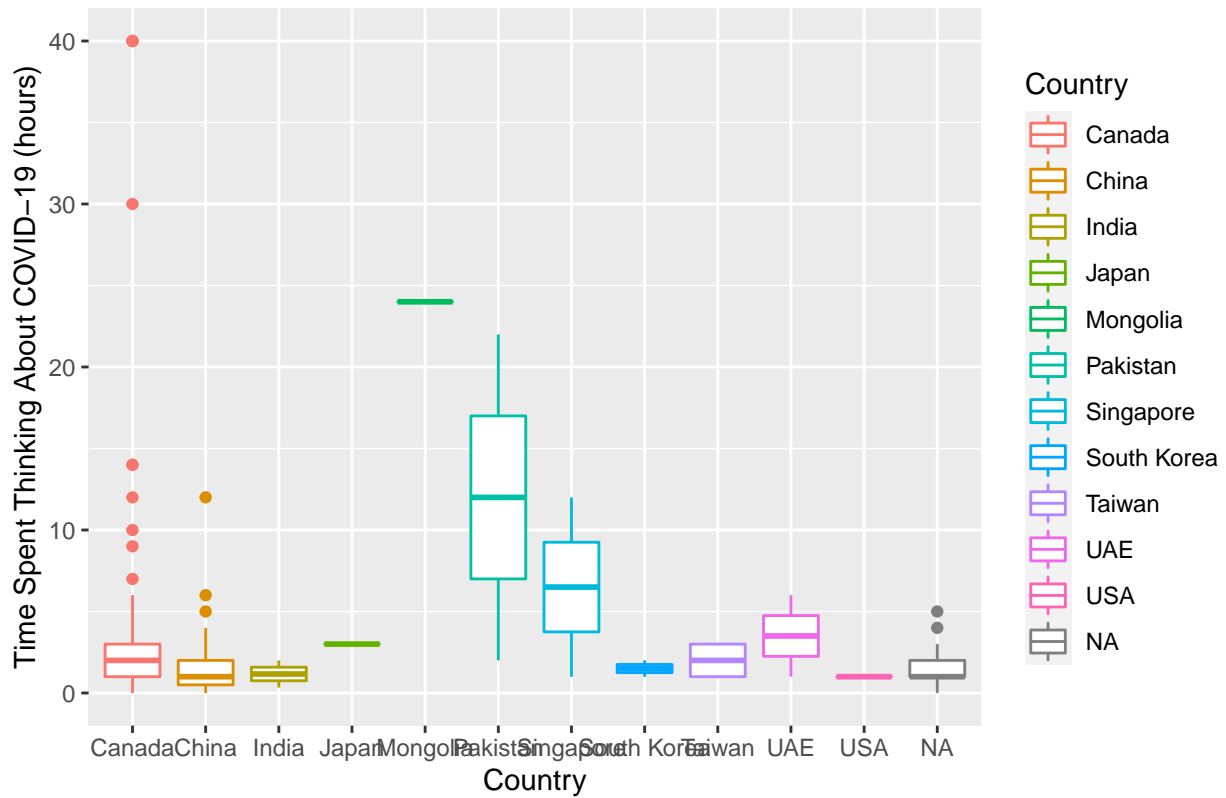
4

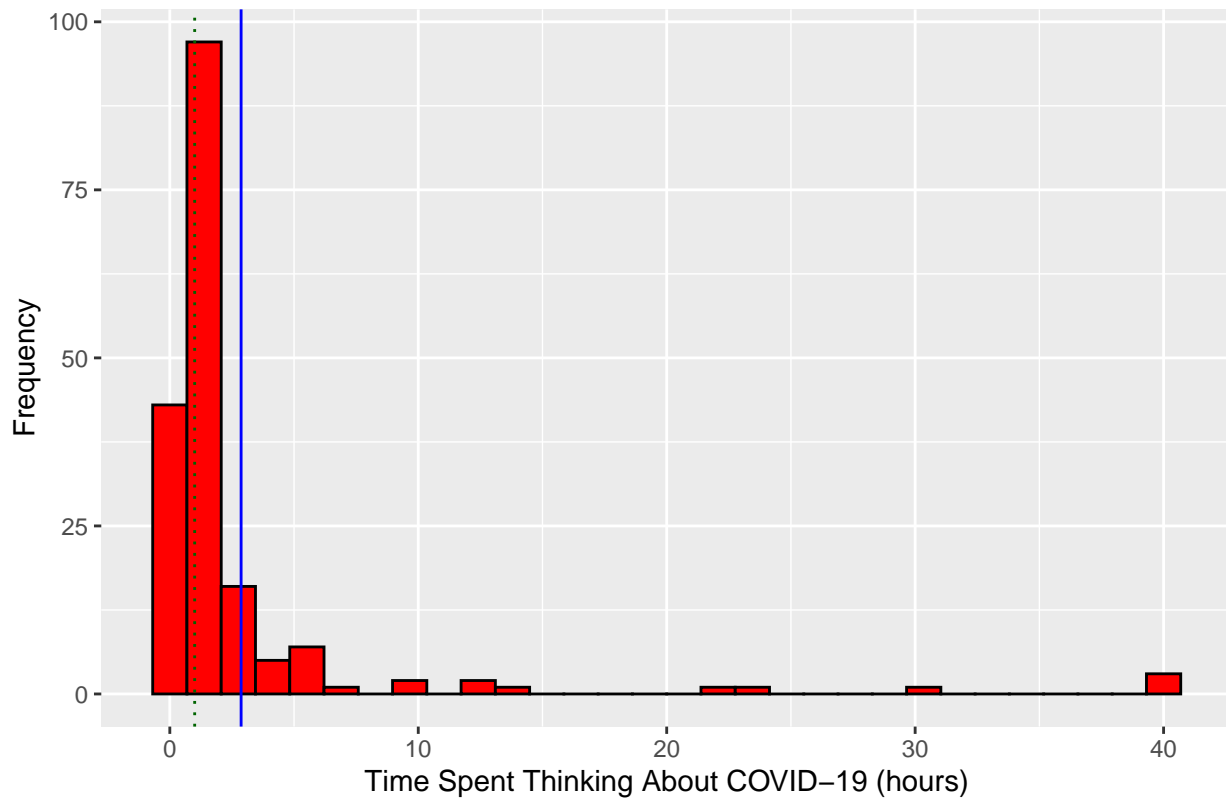## Country vs. Week 1 Time Spent Thinking About COVID−19
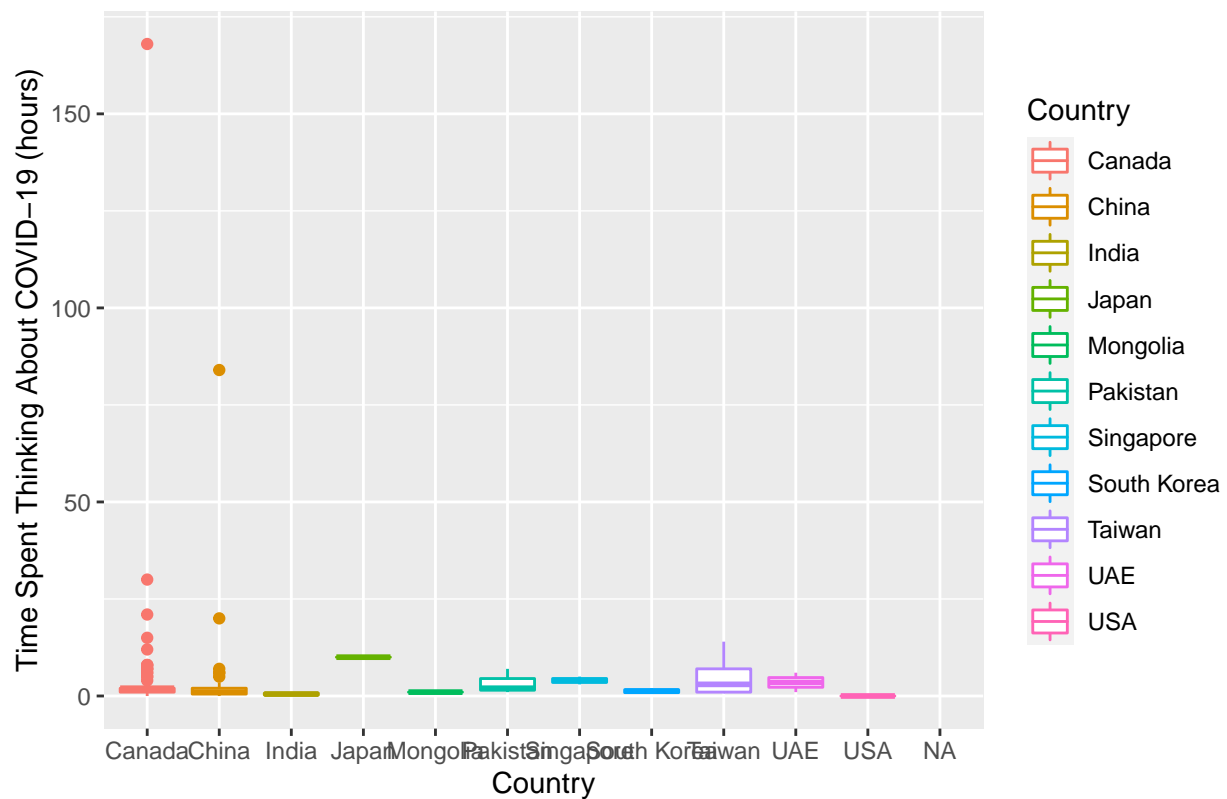
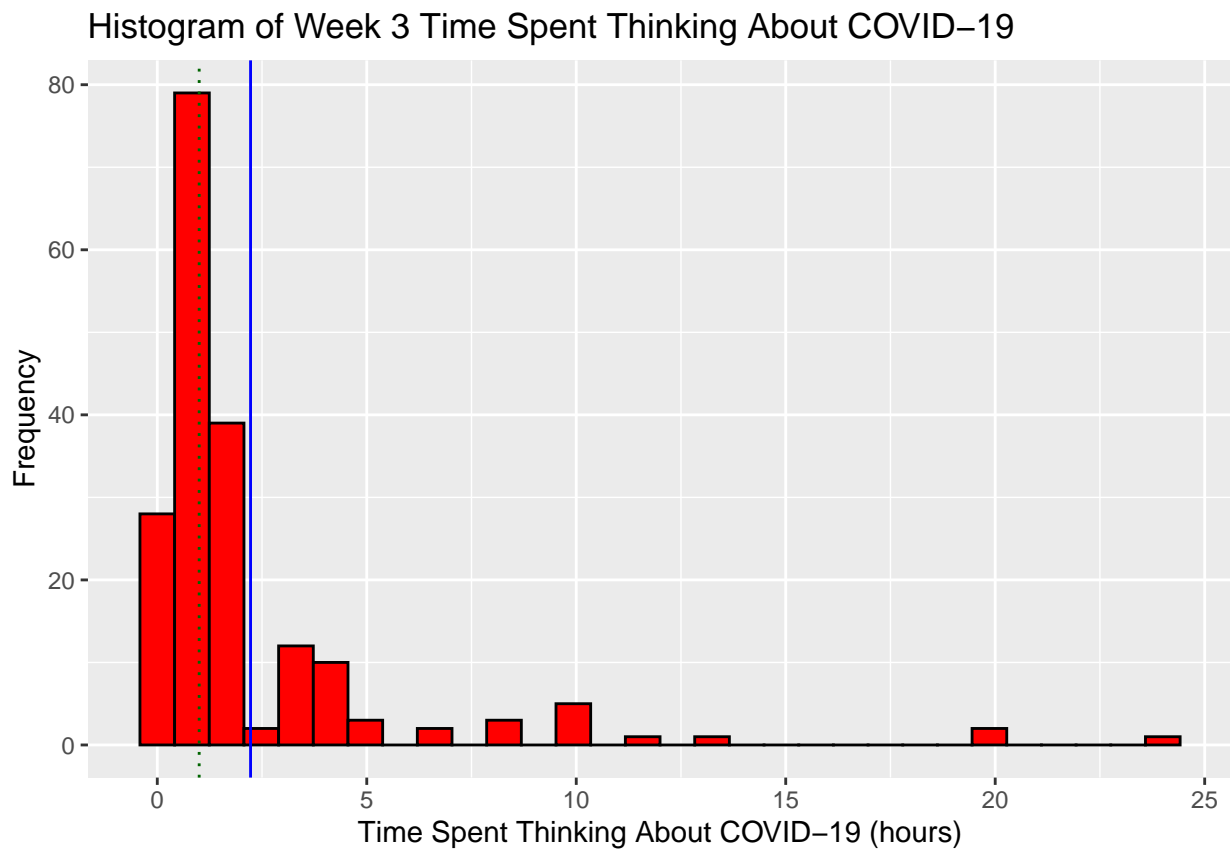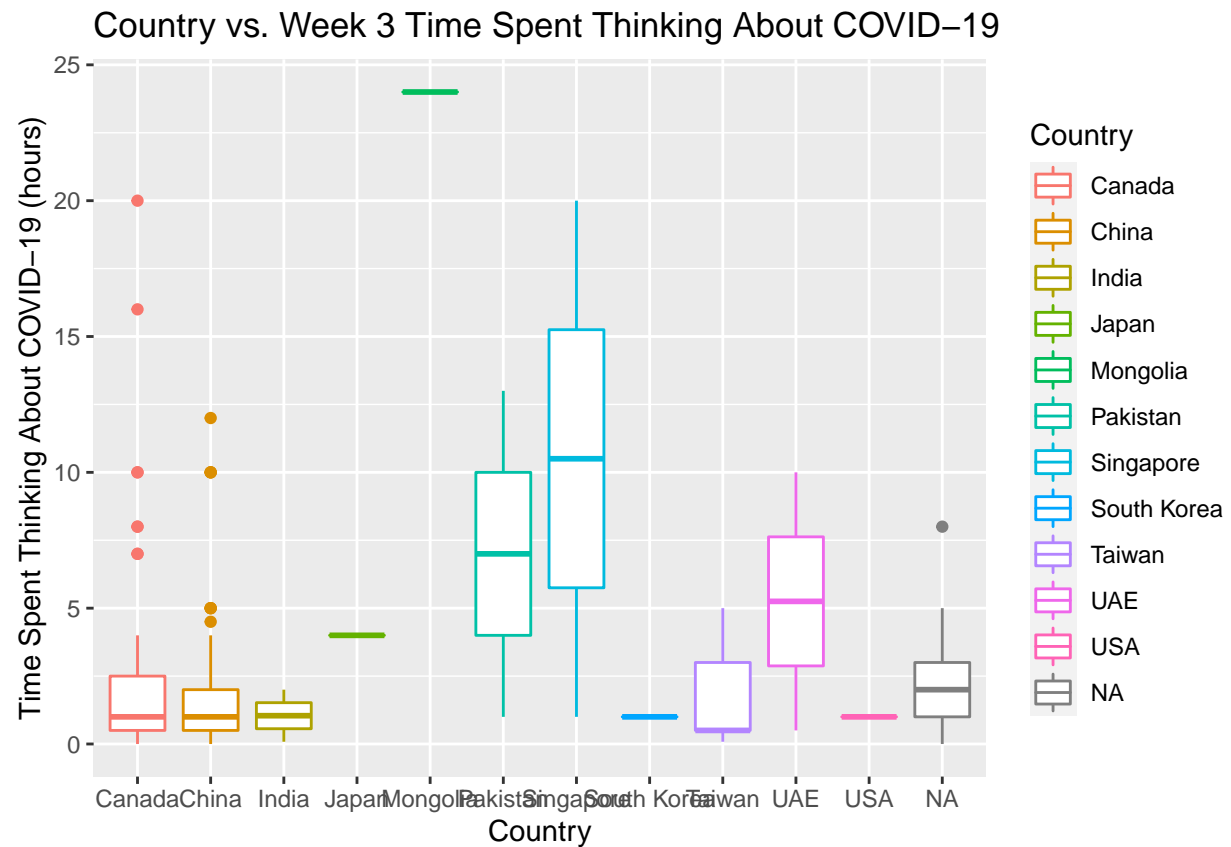## Histogram of Week 1 Time Spent Thinking About COVID−19
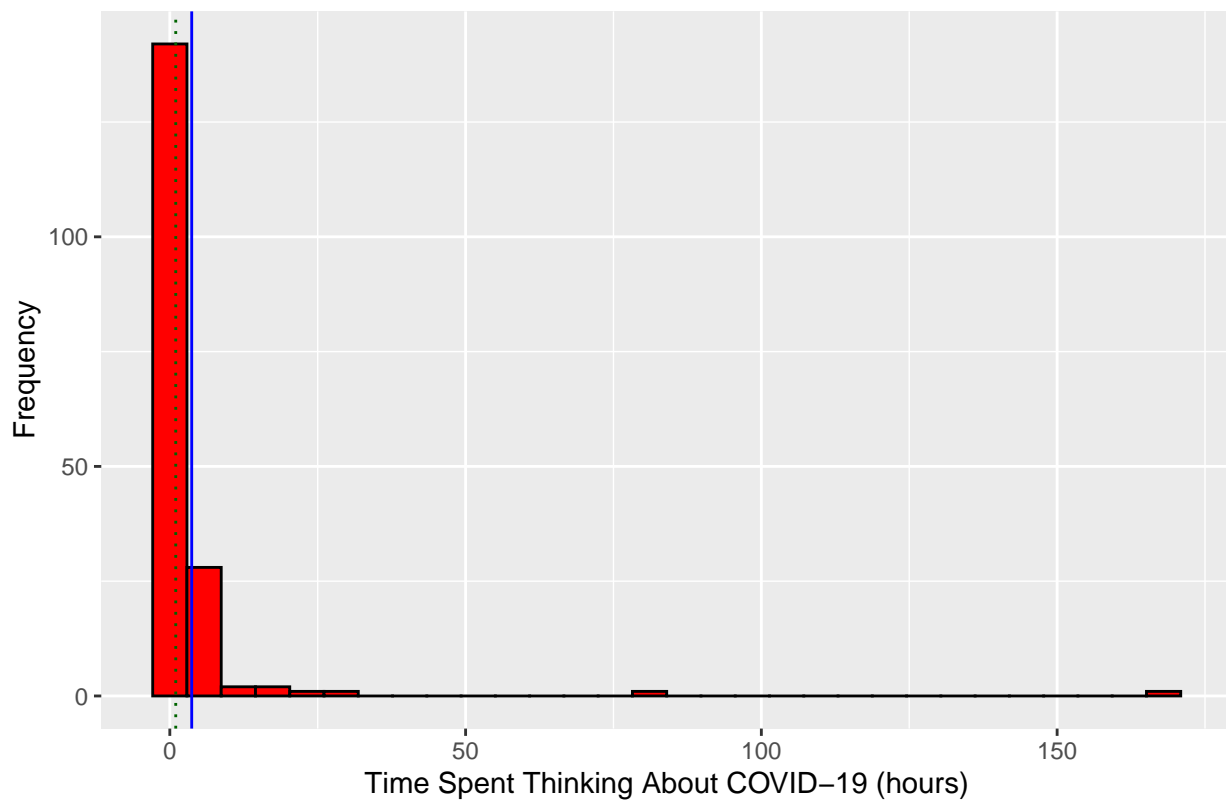
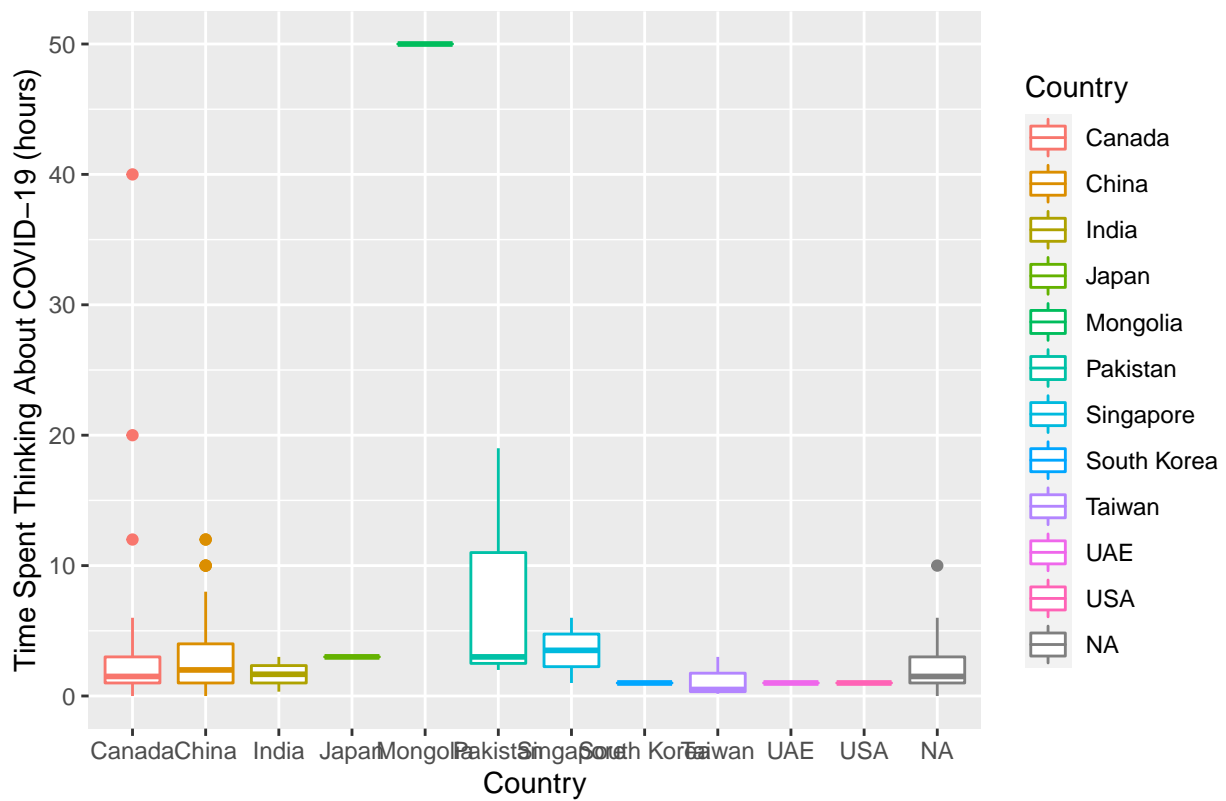Country vs. Week 2 Time Spent Thinking About COVID-19

Histogram of Week 2 Time Spent Thinking About COVID-19

Country vs. Week 3 Time Spent Thinking About COVID−19

Histogram of Week 3 Time Spent Thinking About COVID−19

Country vs. Week 4 Time Spent Thinking About COVID−19

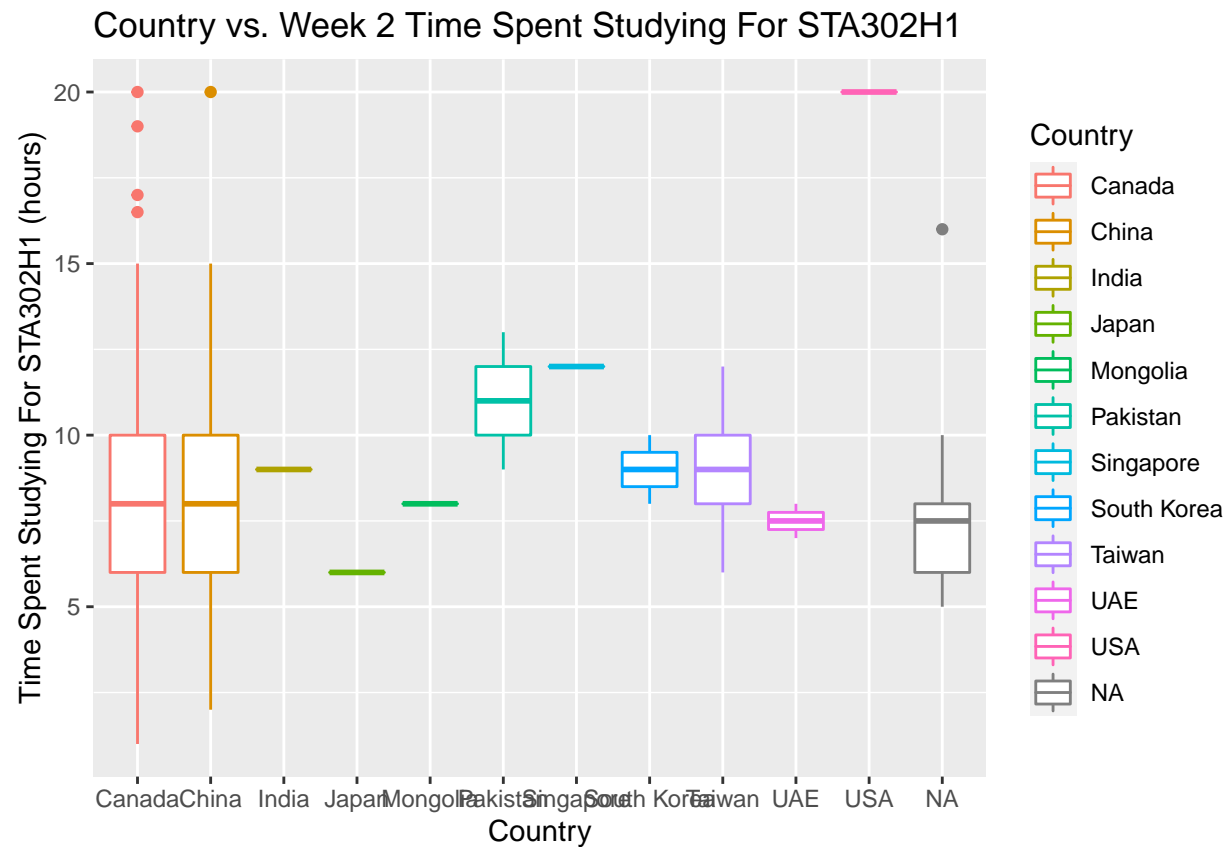# Histogram of Week 4 Time Spent Thinking About COVID−19

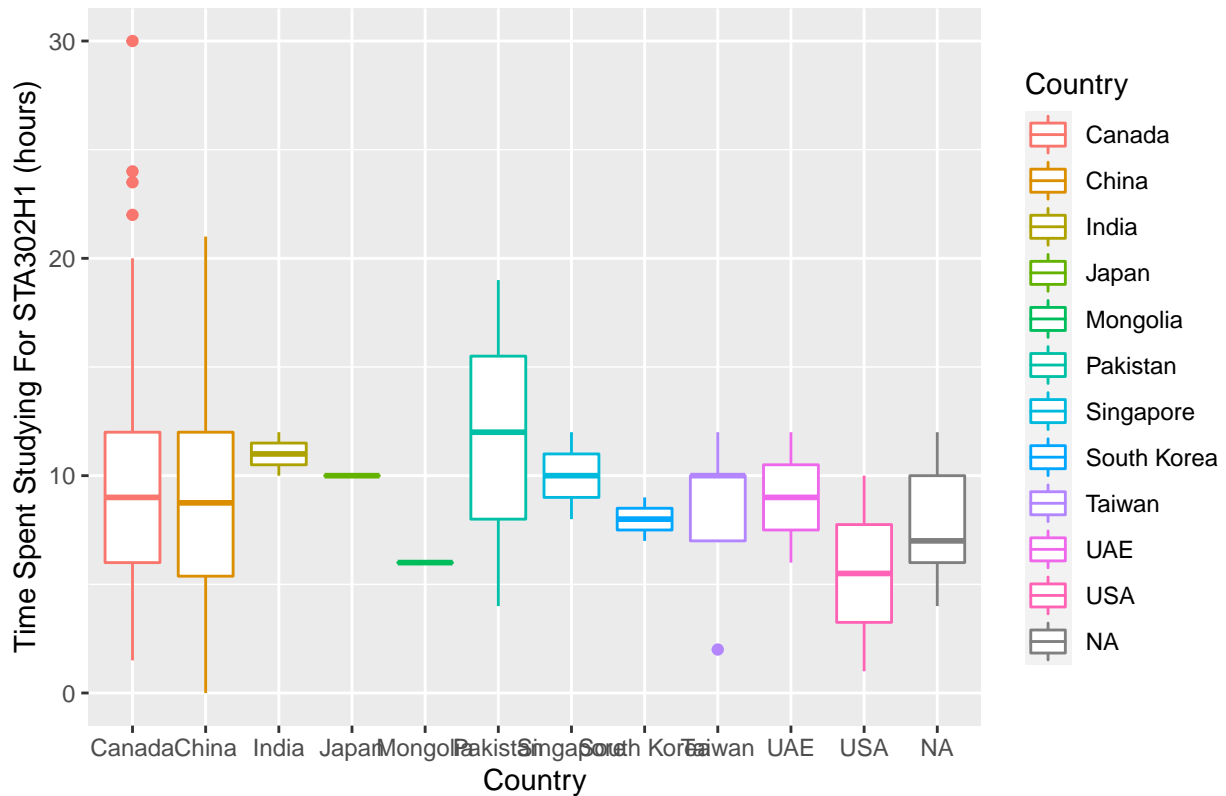# Country vs. Week 1 Time Spent Studying For STA302H1

Histogram of Week 1 Time Spent Studying for STA302H1
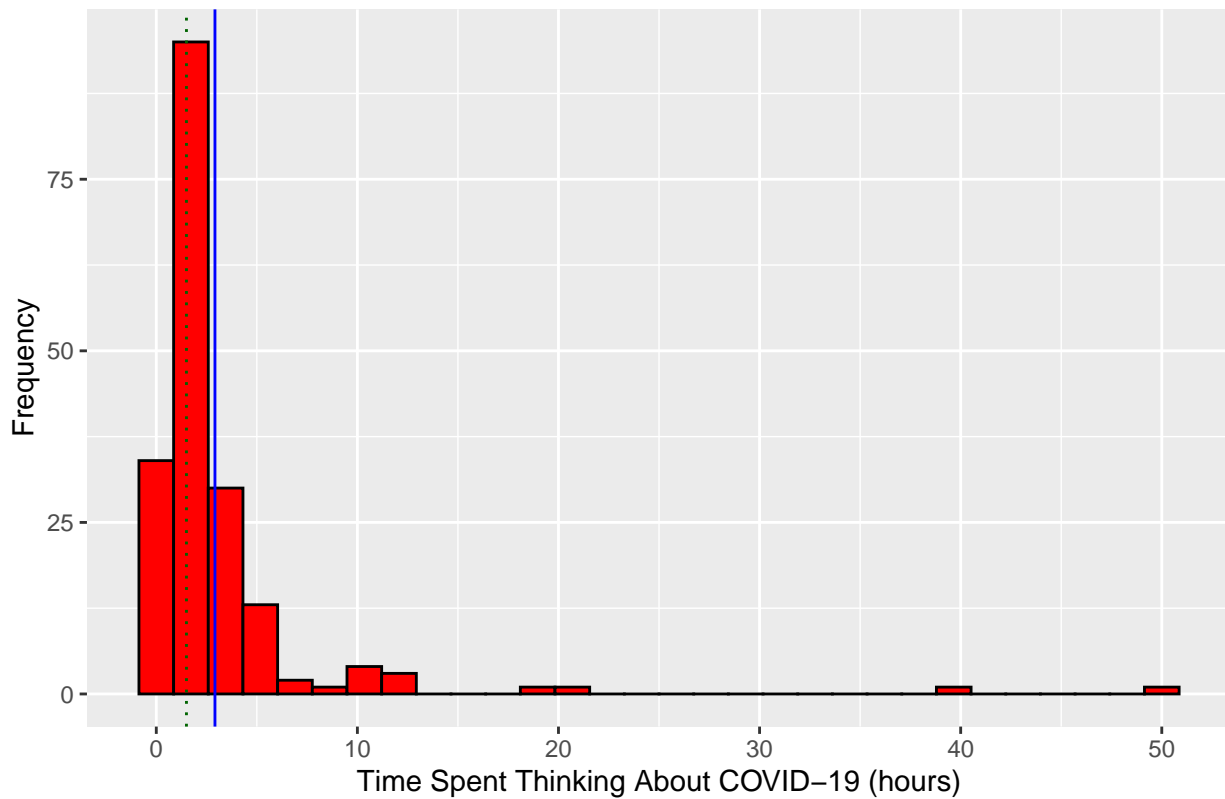
Country vs. Week 2 Time Spent Studying For STA302H1
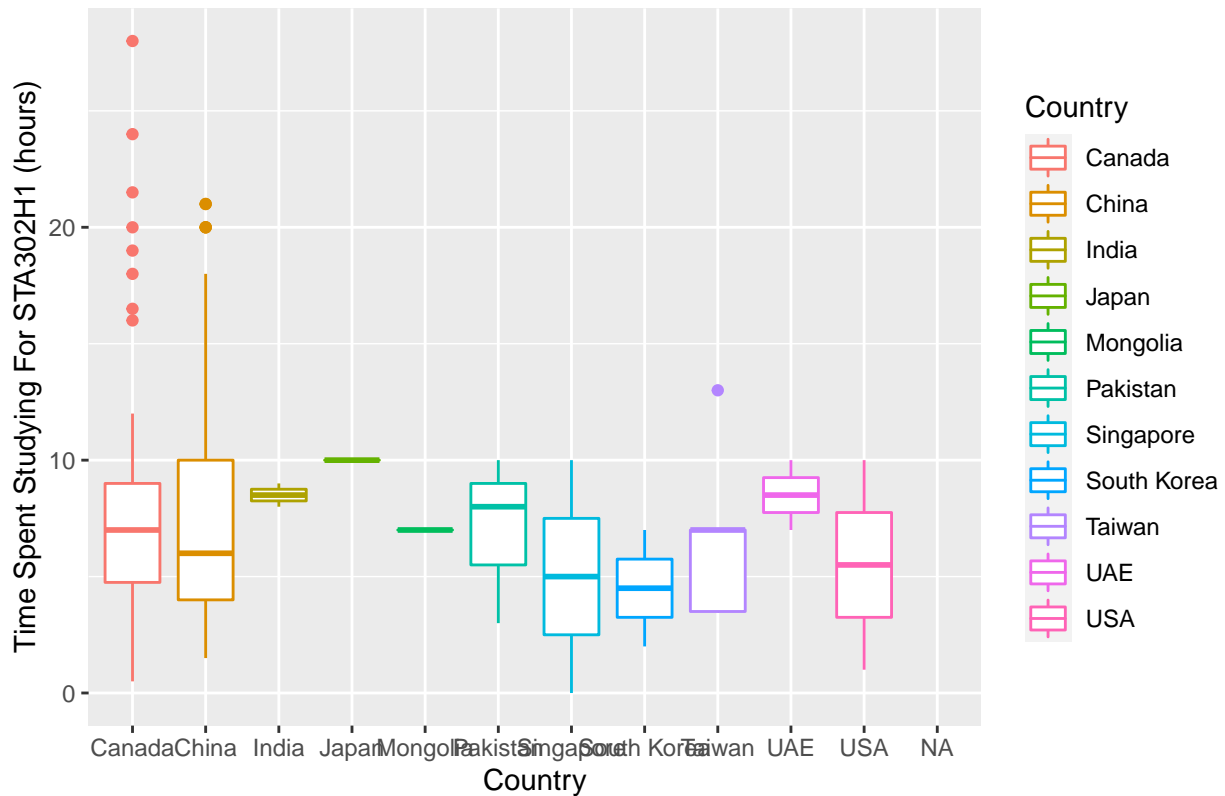
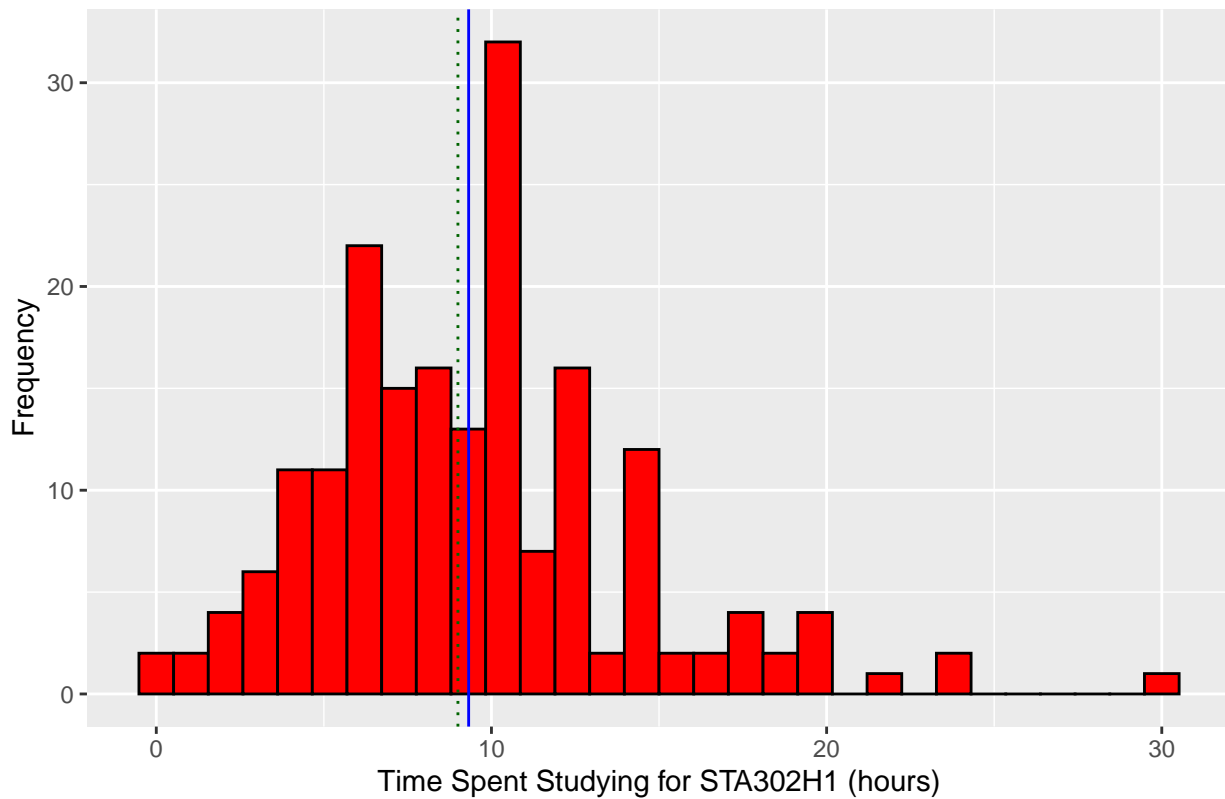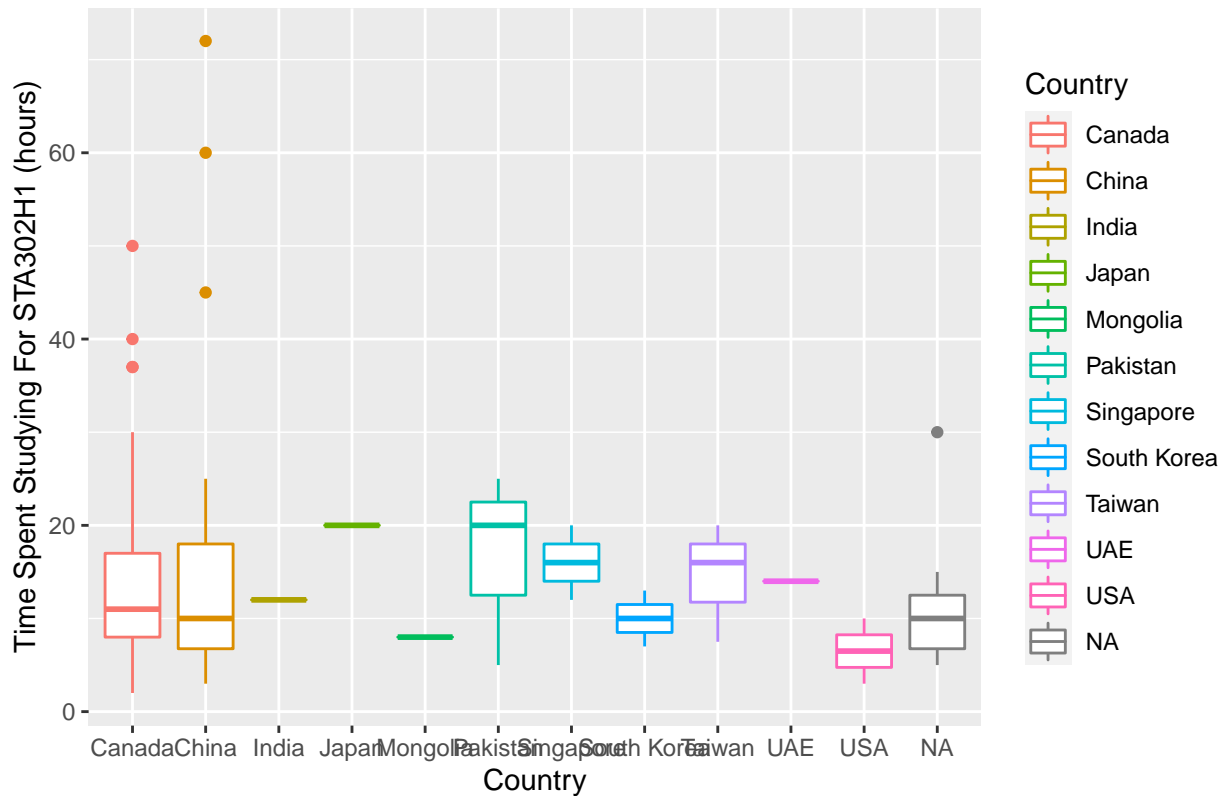Histogram of Week 2 Time Spent Studying for STA302H1
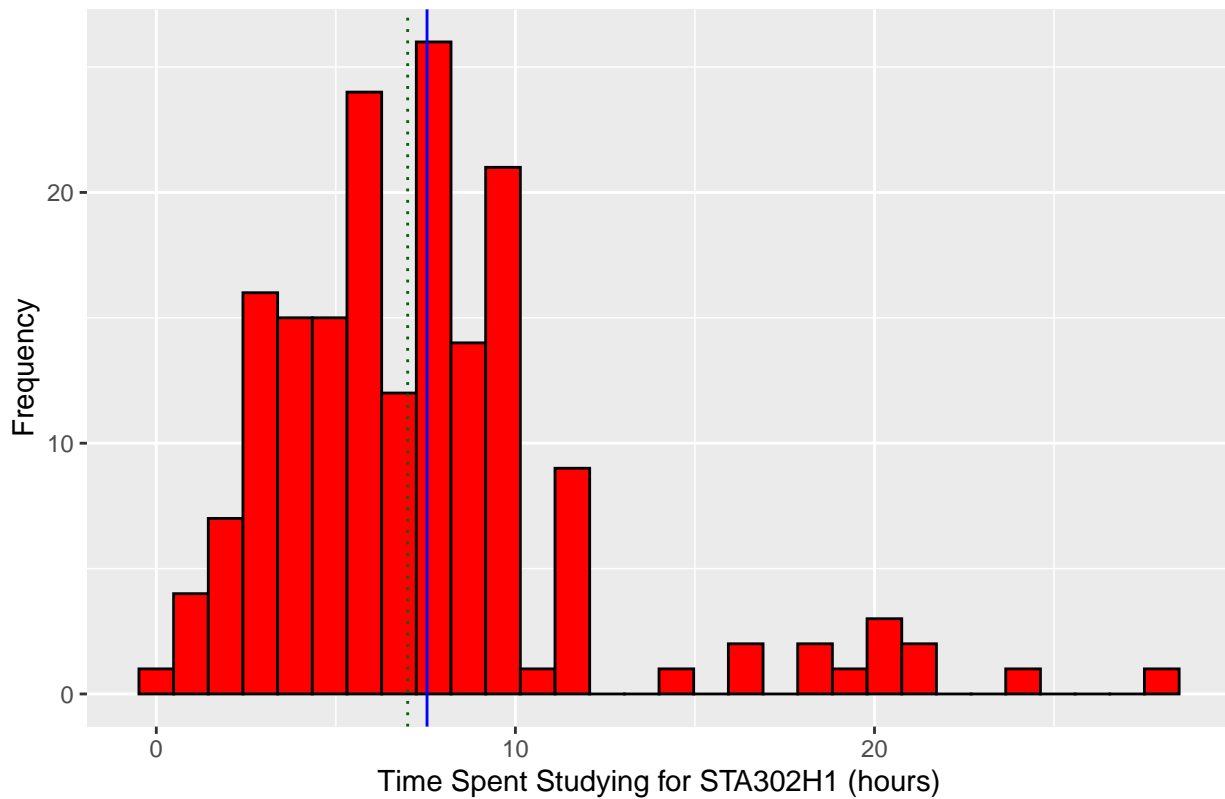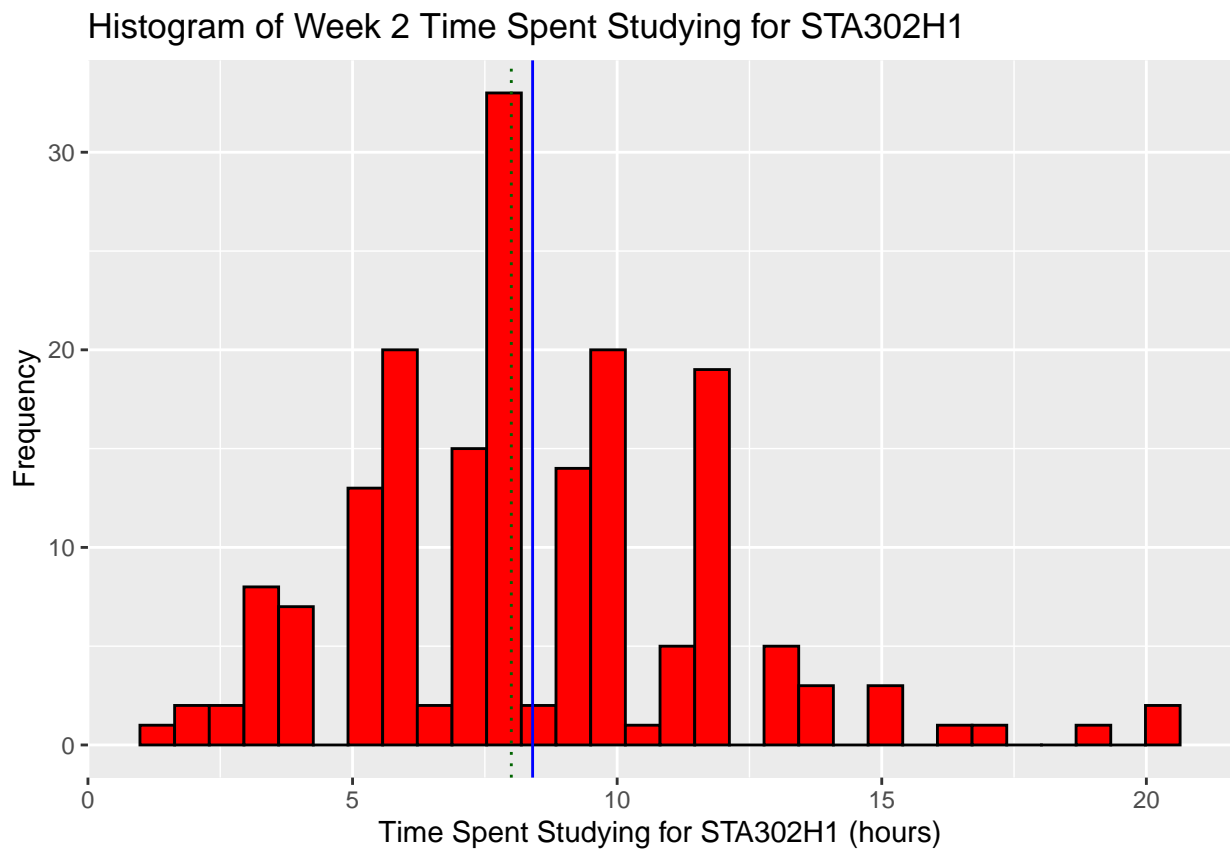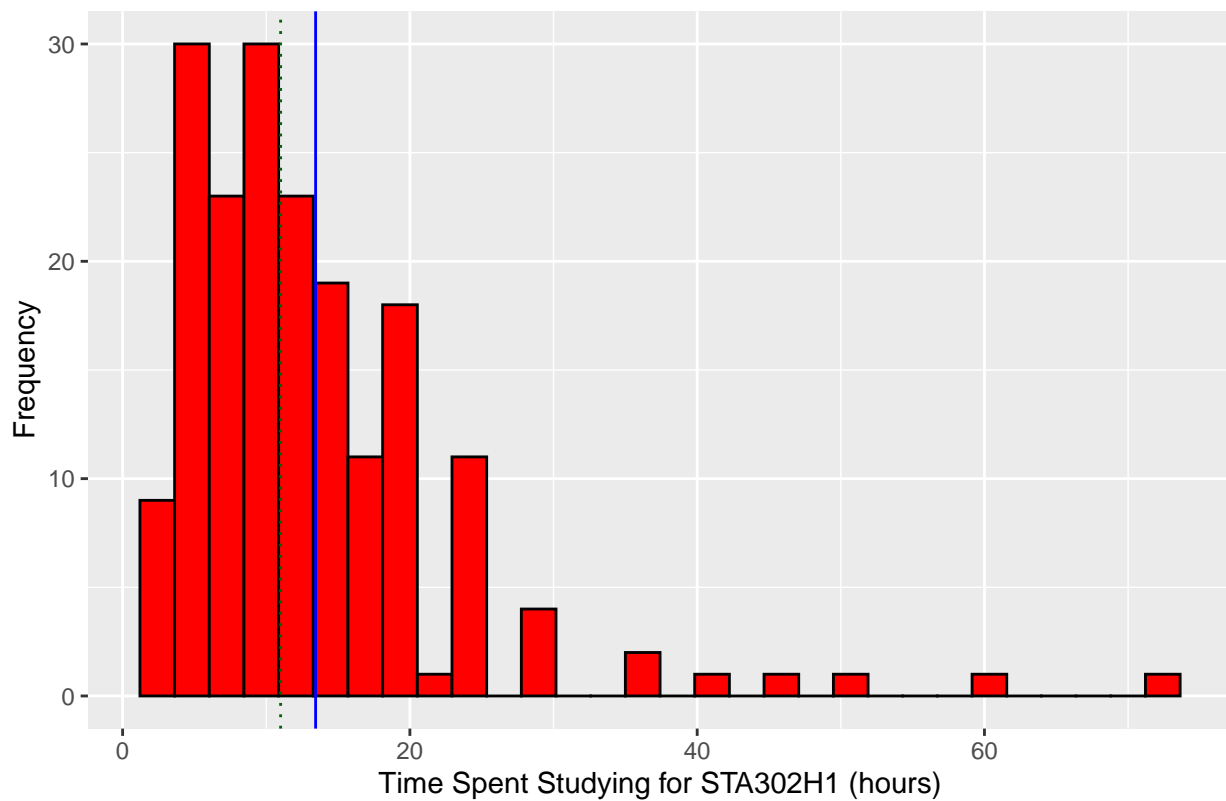
Country vs. Week 3 Time Spent Studying For STA302H1
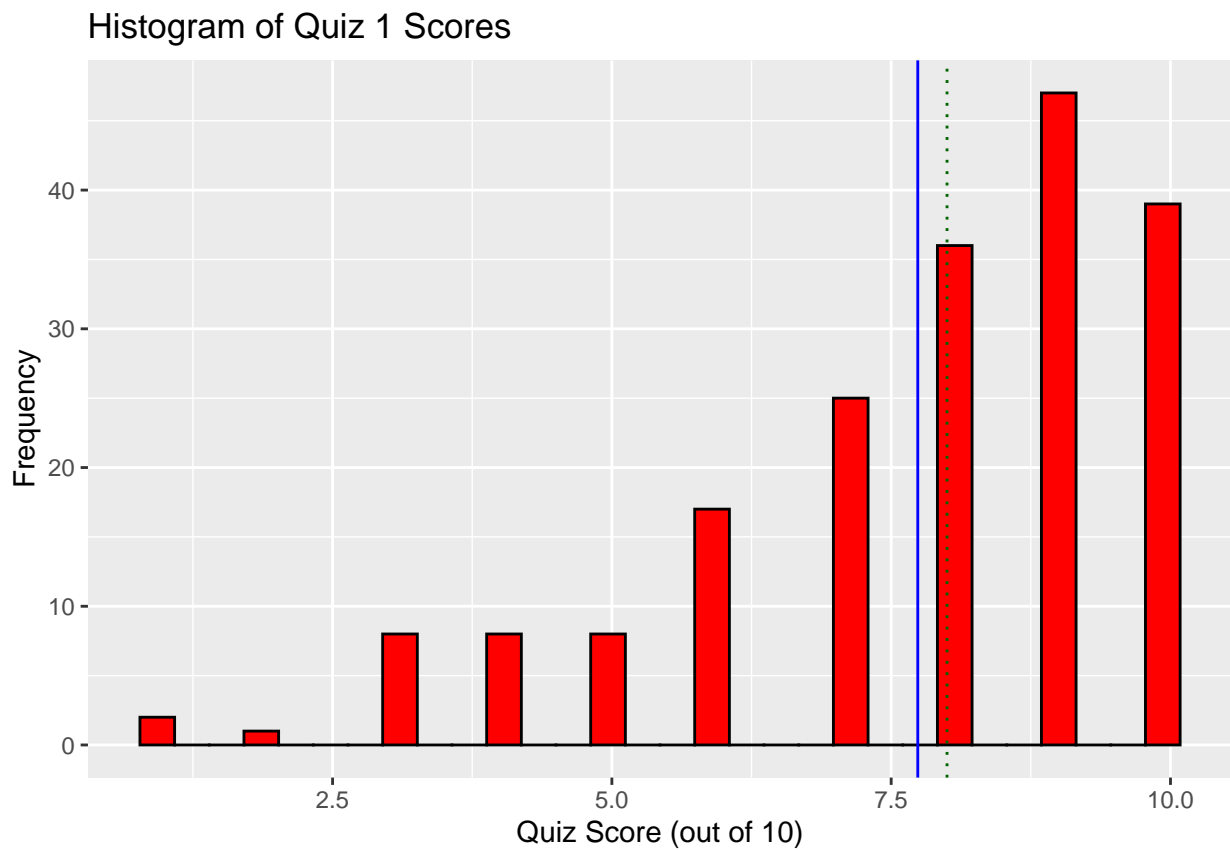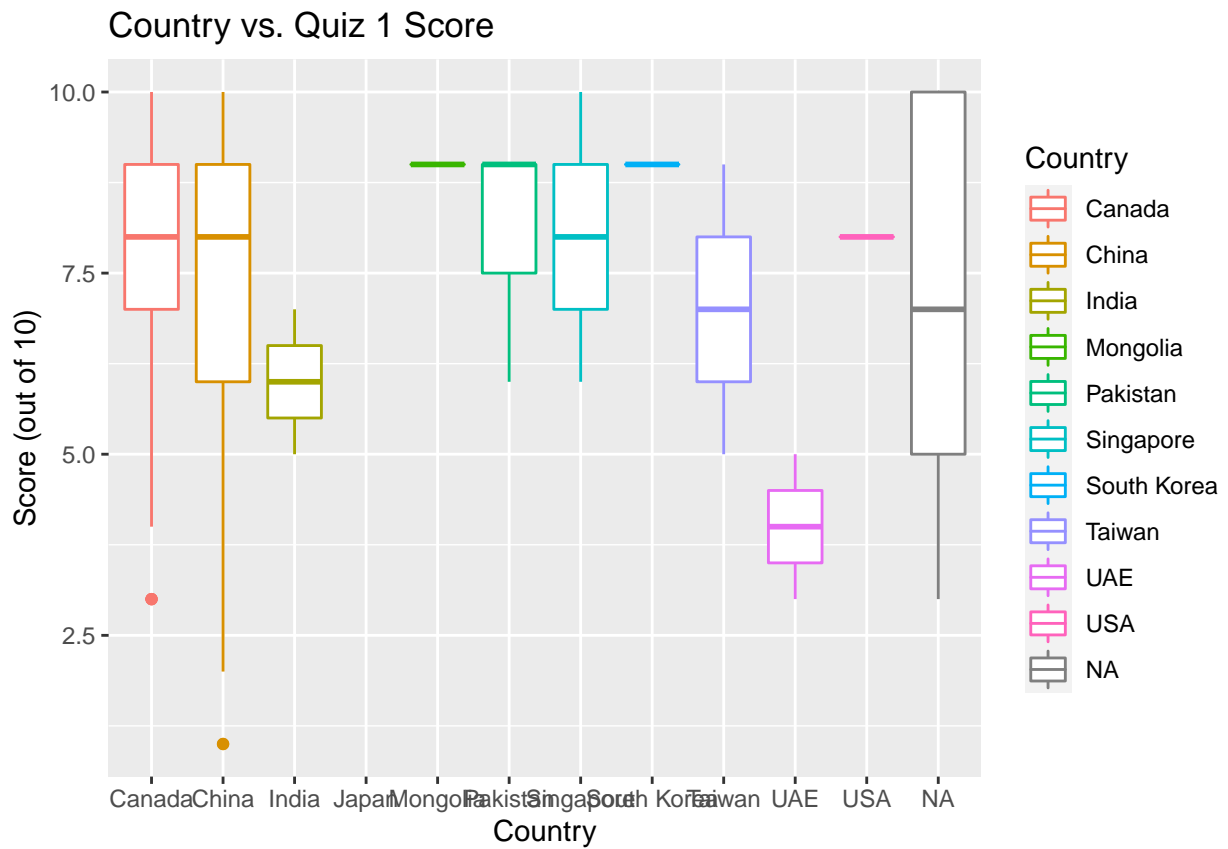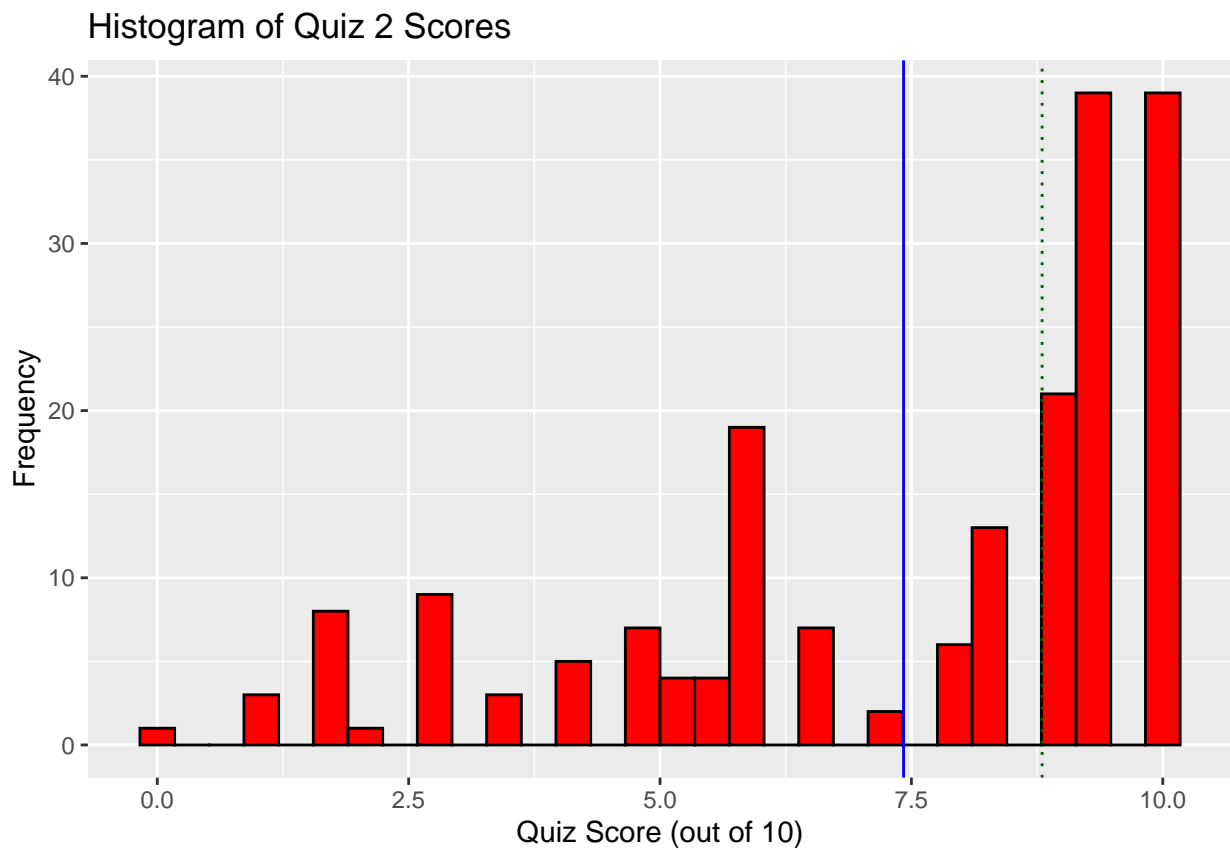

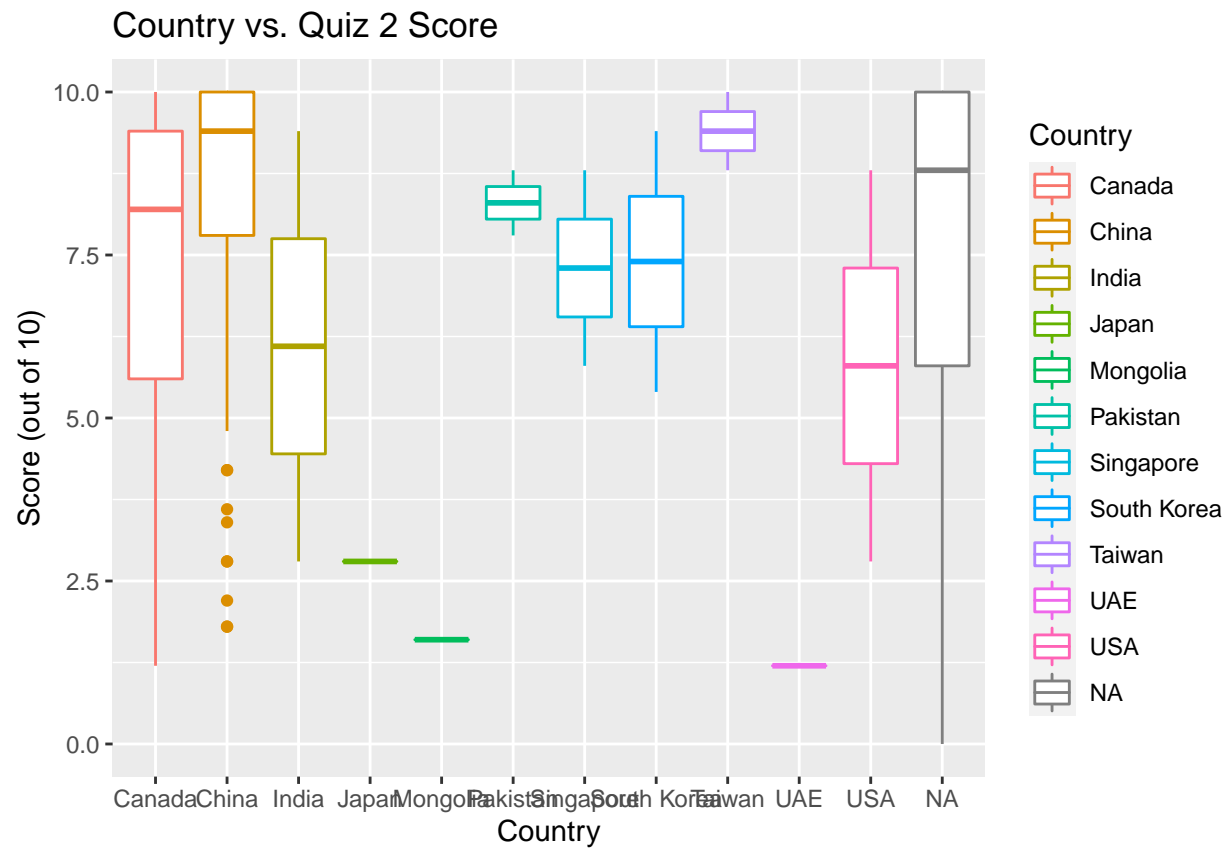Histogram of Week 3 Time Spent Studying for STA302H1

Country vs. Week 4 Time Spent Studying For STA302H1

Histogram of Week 4 Time Spent Studying for STA302H1

Country vs. Quiz 1 Score

Histogram of Quiz 1 Scores

Country vs. Quiz 2 Score

Histogram of Quiz 2 Scores

Country vs. Quiz 3 Score

Histogram of Quiz 3 Scores

Country vs. Quiz 4 Score

Histogram of Quiz 4 Scores
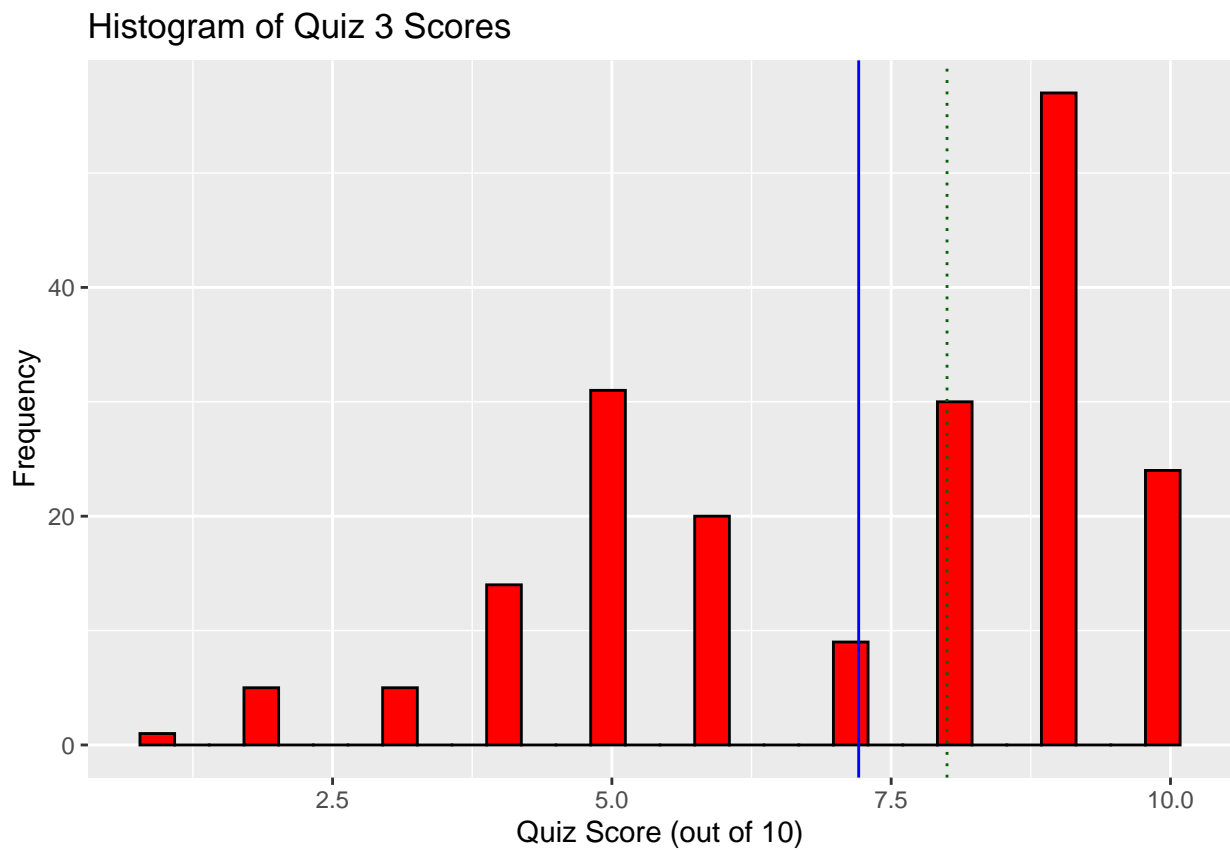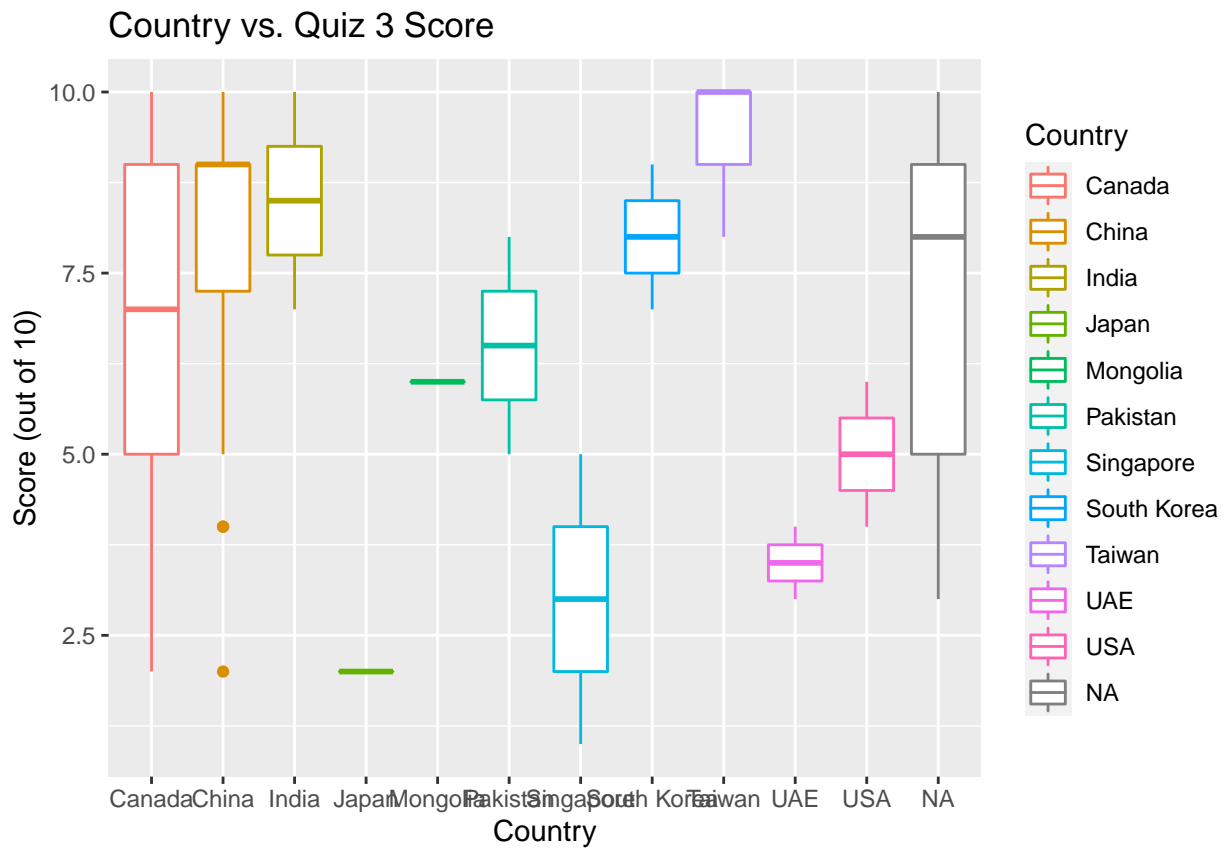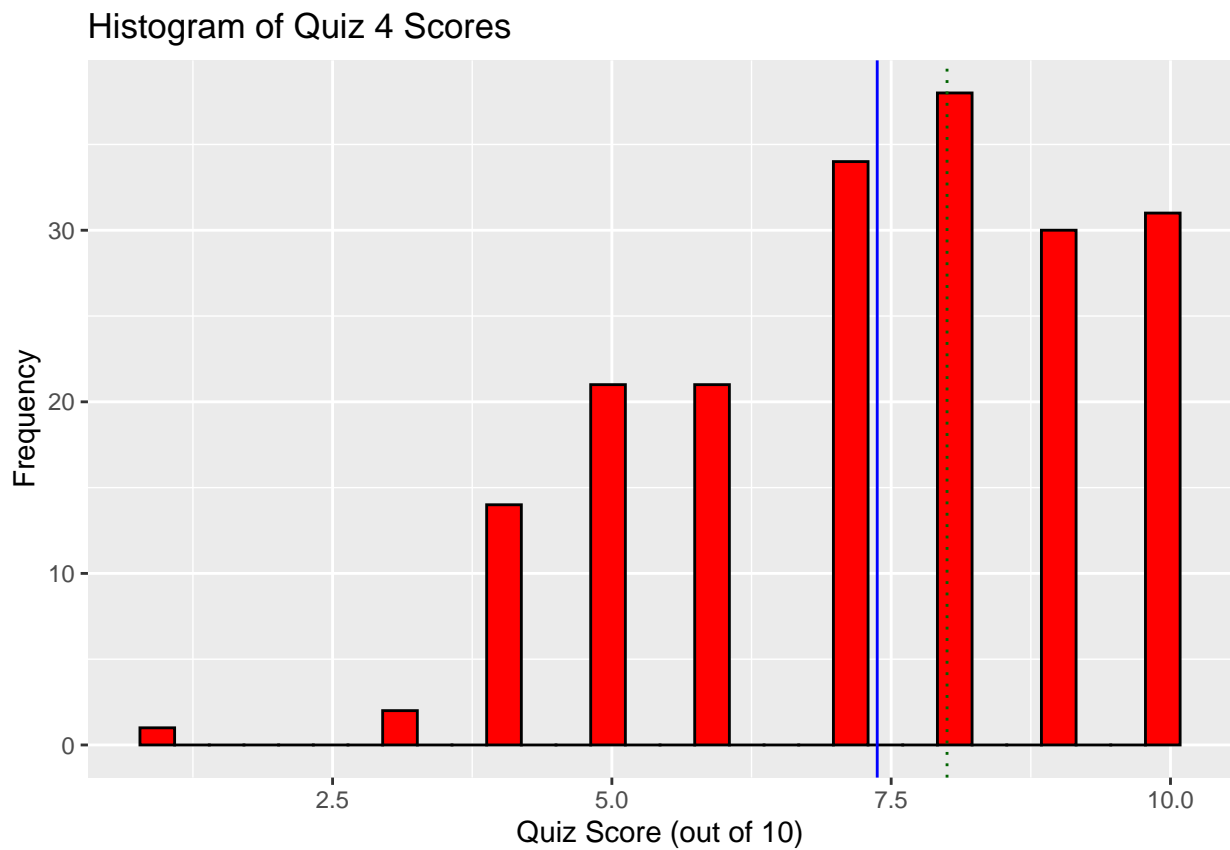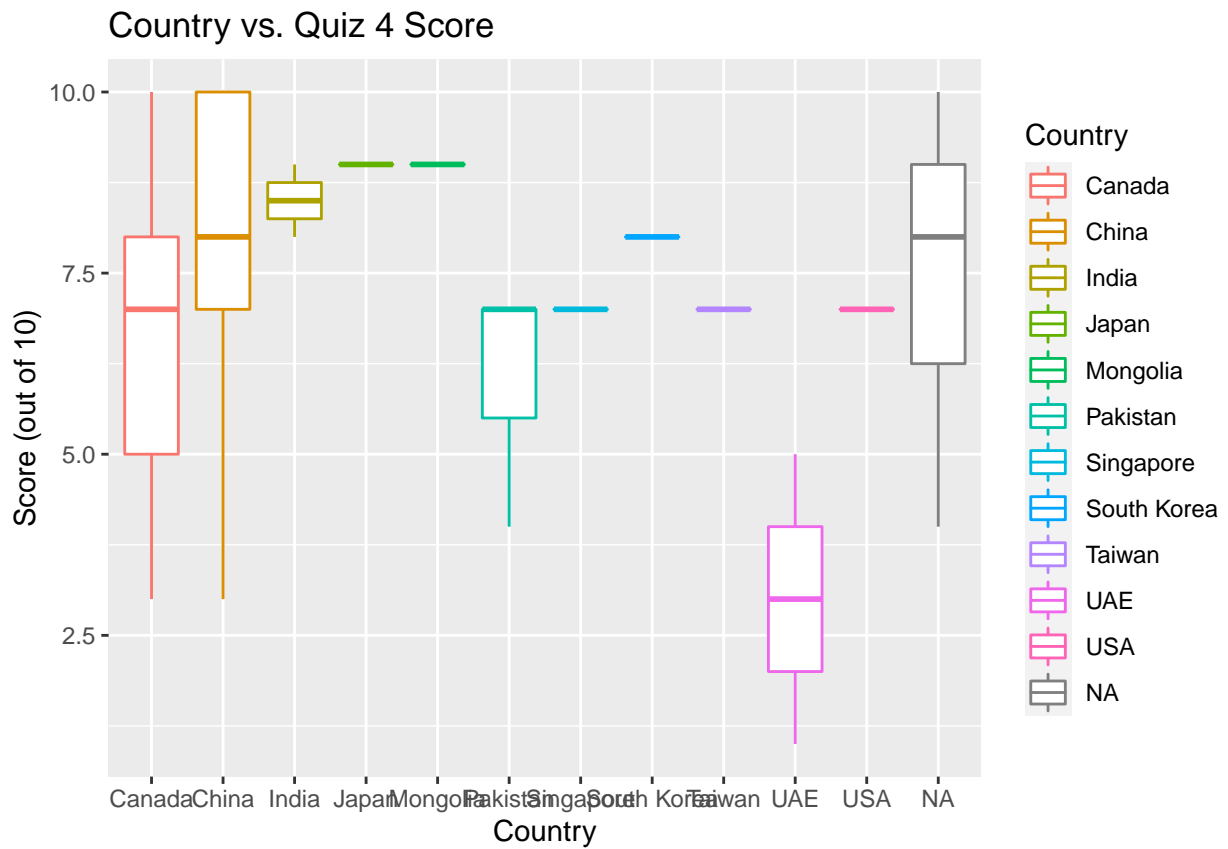
5-Number Summary Statistics

```
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
```