STA302H1 – Final Project Descriptive Statistics

Danny Chen

August 10, 2021

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

```
country_factors = as.factor(c("canada", "china", "india", "japan",
                              "mongolia", "pakistan", "singapore",
                              "south korea", "taiwan", "usa", "uae", "na"))
```

Data Cleaning

First, I'll clean my data.

```
cleaned_sta302_performance_data <- sta302_performance_data %>%
    # Remove the "X" column: it's simply the row number, which isn't very useful.
    select(-X) %>%

    # Group student overall quiz 4 scores from highest to lowest.
    arrange(desc(Quiz_4_score)) %>%

    # 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)

    # TODO: Make sure all country names are lowercase.
    # e.g. "Canada" and "canada" are the same country.
    # 1. Consider running a for loop that makes all rows in column "Country" lowercase,
    # 2. Consider string replacement on "Canada" -> "canada"?

    # TODO: Make sure all STA302H1 hours and COVID contemplation hours are
    # all in numeric form.
    # 1. use as.numeric()?

    # Identify rows with no quiz 4.
    # These indicate students who have dropped STA302H1, and who
    # should be excluded from the final data.

head(cleaned_sta302_performance_data, n = 15)
```

```
##    Country COVID.hours..W1. COVID.hours..W2. COVID.hours..W3. COVID.hours..W4.
## 1   Canada              2.0            3.000              1.0                2
## 2    China              1.0            0.500              1.0                2
## 3    China              5.0            4.000              5.0               12
## 4    China              0.0            0.000              0.5              0.5
## 5   Canada              1.0            0.000              0.0             <NA>
## 6    China              0.5            0.500              0.0                2
## 7   Canada              2.0            1.000              0.5                2
## 8    China              0.5            1.000              0.0                1
## 9    China              2.0            2.000              1.5                2
## 10   China              0.1            0.000              1.0                1
## 11   china              3.0            2.000              1.0             <NA>
## 12   China              1.0            2.000              1.0                5
## 13   China              2.0            2.000              2.0                2
## 14   China              1.0            0.500              0.5              0.5
## 15   China              0.0            0.333               NA                1
##    STA302.hours..W1. STA302.hours..W2. STA302.hours..W3. STA302.hours..W4.
## 1                  3               7.0                 6                 6
## 2                  3               3.0                 3                 3
## 3                 18               6.0                12                15
## 4                  6               6.0                 3                 4
## 5                  5               4.0                 6              <NA>
## 6                  6               8.0                11                17
## 7                  9               9.0                15                 9
## 8                 20              11.0                10                 8
## 9                  8              10.0                11                12
## 10                 6               9.0                 8                14
## 11                 6               8.0                 7              <NA>
## 12                 8              10.0                10                16
## 13                10              14.0                14                24
## 14                 6               5.0                 8                18
## 15                 3               3.5              <NA>                20
##    Quiz_1_score Quiz_2_score Quiz_3_score Quiz_4_score
## 1            10          7.8            9           10
## 2             8          2.8            9           10
## 3             9          9.4            9           10
## 4             9         10.0            9           10
## 5             9         10.0            9           10
## 6             8          5.2           10           10
## 7             8          5.8            5           10
## 8             6         10.0            9           10
## 9             7          2.8            9           10
## 10            5          9.0            9           10
## 11            9           NA            8           10
## 12            9         10.0            9           10
## 13            6          8.2            8           10
## 14            7          8.2            9           10
## 15            6         10.0            9           10
```

Identifying Anomalous Data

Let's identify rows with at least one NA. Although some of the rows might only have 1 - 2 NAs and are therefore salvageable, other rows may contain 3 or more NAs, and might indicate students who have dropped STA302H1.

```
rows_with_some_NAs = cleaned_sta302_performance_data[
  rowSums(is.na(cleaned_sta302_performance_data)) >= 1,
]
head(rows_with_some_NAs, n = 10)
```

```
##    Country COVID.hours..W1. COVID.hours..W2. COVID.hours..W3. COVID.hours..W4.
## 5   Canada              1.0            0.000                0             <NA>
## 11   china              3.0            2.000                1             <NA>
## 15   China              0.0            0.333               NA                1
## 27  Canada              1.0            1.000                1             <NA>
## 28    <NA>               NA            2.000                3                3
## 29    <NA>               NA               NA                2                3
## 30    <NA>               NA               NA               NA             <NA>
## 31    <NA>               NA               NA               NA               10
## 36   China              0.5               NA                1                8
## 39  Canada              1.5               NA                1              1.5
##    STA302.hours..W1. STA302.hours..W2. STA302.hours..W3. STA302.hours..W4.
## 5                  5               4.0                 6              <NA>
## 11                 6               8.0                 7              <NA>
## 15                 3               3.5              <NA>                20
## 27                 6               5.0                 5              <NA>
## 28                NA               8.0                10                12
## 29                NA                NA                 4                 5
## 30                NA                NA              <NA>              <NA>
## 31                NA                NA              <NA>                10
## 36                 3                NA                 2                23
## 39                 7                NA               8.5                10
##    Quiz_1_score Quiz_2_score Quiz_3_score Quiz_4_score
## 5             9         10.0            9           10
## 11            9           NA            8           10
## 15            6         10.0            9           10
## 27           NA         10.0            9           10
## 28            7         10.0            9           10
## 29           10           NA            8           10
## 30           10         10.0           10           10
## 31           10         10.0           10           10
## 36            8          9.4           10            9
## 39           10          1.2            9            9
```

. . . and rows whose columns are mis-typed and in need of correcting.

Rows with Mistyped Columns

```
rows_with_mistyped_columms = cleaned_sta302_performance_data[c(38, 83, 84, 117),]
rows_with_mistyped_columms
```

```
##     Country COVID.hours..W1. COVID.hours..W2. COVID.hours..W3. COVID.hours..W4.
## 38    China                0              0.5              1.0              0.5
## 83   canada              168             40.0             20.0               12
## 84   canada                1              1.0              2.0                1
## 117  Taiwan                1              1.0              0.5         0.5 hour
##     STA302.hours..W1. STA302.hours..W2. STA302.hours..W3. STA302.hours..W4.
## 38                  4               5.5       5.5<U+00A0>                 6
## 83                  8               6.0                 6                20
## 84                  9               8.0                12                15
## 117                 7               8.0                 7         7.5 hours
##     Quiz_1_score Quiz_2_score Quiz_3_score Quiz_4_score
## 38             9         10.0           10            9
## 83            10          9.4            9            8
## 84             9          5.4            9            8
## 117            6          8.8            8            7
```

```
# row 83: Country -> "canada"
# row 84: Country -> "canada"

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

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

Select Predictor Variables, Find Their Significance

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


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

Histograms

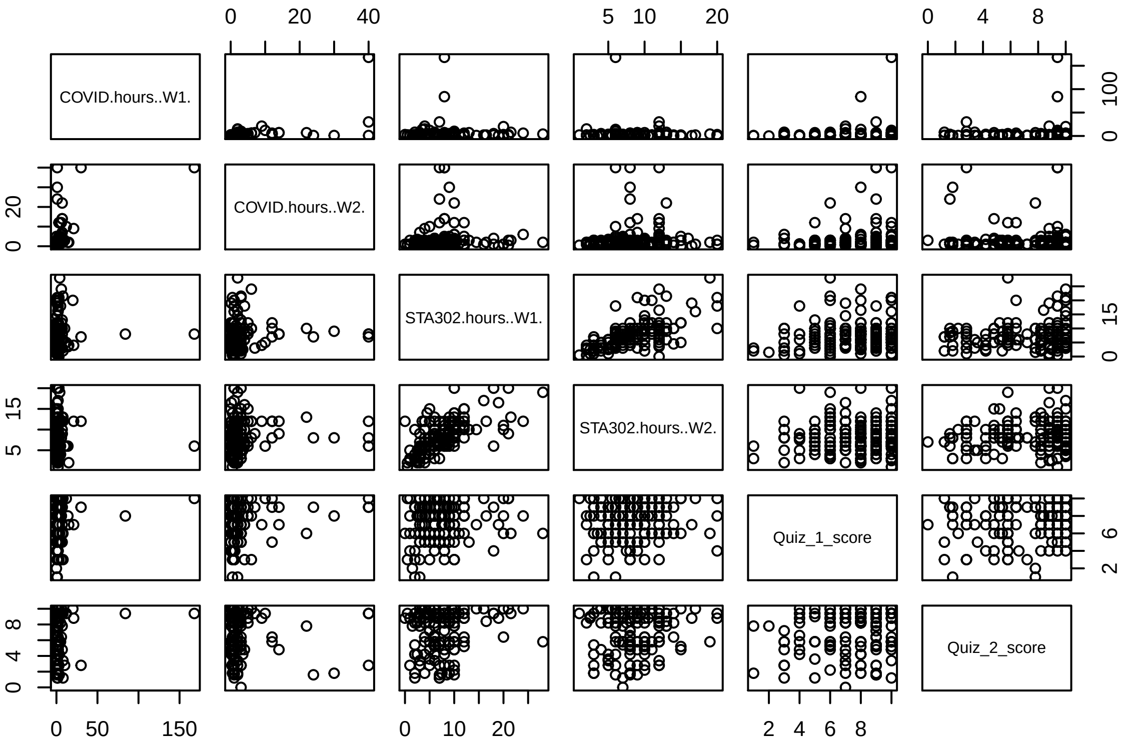
```
# TODO: See Demo 1 to figure out how to add histograms in a matrix format.
```

Boxplots

```
# TODO: See STA248H1 notes to figure out how to create boxplots by class.
```

Scatterplots

```
# pairwise scatterplot
pairs(~COVID.hours..W1. + COVID.hours..W2. +
      # COVID.hours..W3. + COVID.hours..W4. +  # TODO: Clean columns.
      STA302.hours..W1. + STA302.hours..W2. +
      # STA302.hours..W3. + STA302.hours..W4. +  # TODO: Clean columns.
      Quiz_1_score + Quiz_2_score,
      # Quiz_3_score + Quiz_4_score,  # TODO: Clean columns.
      data = cleaned_sta302_performance_data)
```



Correlation Matrix

```r
# TODO: Find correlation matrix to determine candidate significant
# TODO: predictor values.
# cor(cleaned_sta302_performance_data)  # TODO: How to make numeric?
```

Summary Statistics

Mean STA302H1 study time

```
mean_STA302H1_study_times <- data.frame(
  week1 = mean(sta302_performance_data$STA302.hours..W1., na.rm = TRUE),
  week2 = mean(sta302_performance_data$STA302.hours..W2., na.rm = TRUE),
  week3 = mean(sta302_performance_data$STA302.hours..W3., na.rm = TRUE), # TODO: Clean column.
  week4 = mean(sta302_performance_data$STA302.hours..W4., na.rm = TRUE)  # TODO: Clean column.
)
```

```
## Warning in mean.default(sta302_performance_data$STA302.hours..W3., na.rm =
## TRUE): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(sta302_performance_data$STA302.hours..W4., na.rm =
## TRUE): argument is not numeric or logical: returning NA
```

```
mean_STA302H1_study_times
```

```
##      week1    week2 week3 week4
## 1 7.457711 8.297561    NA    NA
```

Mean COVID contemplation time

```
mean_COVID_contemplation_times <- data.frame(
  week1 = mean(sta302_performance_data$COVID.hours..W1., na.rm = TRUE),
  week2 = mean(sta302_performance_data$COVID.hours..W2., na.rm = TRUE),
  week3 = mean(sta302_performance_data$STA302.hours..W3., na.rm = TRUE),  # TODO: Clean column.
  week4 = mean(sta302_performance_data$STA302.hours..W4., na.rm = TRUE)   # TODO: Clean column.
)
```

```
## Warning in mean.default(sta302_performance_data$STA302.hours..W3., na.rm =
## TRUE): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(sta302_performance_data$STA302.hours..W4., na.rm =
## TRUE): argument is not numeric or logical: returning NA
```

```
mean_COVID_contemplation_times
```

```
##      week1    week2 week3 week4
## 1 3.607163 2.884312    NA    NA
```

Median STA302H1 study time

```
median_STA302H1_study_times <- data.frame(
  week1 = median(sta302_performance_data$STA302.hours..W1., na.rm = TRUE),
  week2 = median(sta302_performance_data$STA302.hours..W2., na.rm = TRUE),
  week3 = as.double(median(sta302_performance_data$STA302.hours..W3., na.rm = TRUE)),
  week4 = as.double(median(sta302_performance_data$STA302.hours..W4., na.rm = TRUE))
)
median_STA302H1_study_times
```

```
##   week1 week2 week3 week4
## 1     7     8     3    20
```

Median COVID contemplation time

```
median_COVID_contemplation_times <- data.frame(
  week1 = median(sta302_performance_data$COVID.hours..W1., na.rm = TRUE),
  week2 = median(sta302_performance_data$COVID.hours..W2., na.rm = TRUE),
  week3 = median(sta302_performance_data$COVID.hours..W3., na.rm = TRUE),
  week4 = as.double(median(sta302_performance_data$COVID.hours..W4., na.rm = TRUE))
)
median_COVID_contemplation_times
```

```
##   week1 week2 week3 week4
## 1     1     1     1   1.5
```

Country summary statistics

```
length(which(cleaned_sta302_performance_data$Country == "Canada")) + 2
```

```
## [1] 112
```

```
length(which(is.na(cleaned_sta302_performance_data$Country)))
```

```
## [1] 26
```

Study hours summary statistics

```
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.)
```

```
##    Length     Class      Mode
##       227 character character
```

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

```
##    Length     Class      Mode
##       227 character character
```

COVID hours summary statistics

```
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.)
```

```
##    Length     Class      Mode
##       227 character character
```