DATA 607 Final Project-Part II

Anthony Conrardy

2024-05-05

Introduction

In Part II of the this final project, we will import the sterilized data sets from Part I that are located on the GitHub repository in the locations below. We will then merge the data sets into a unified set containing the required elements for further investigation and analysis. We will rename the variables to be something easier to work with, reorder the variables in the data frame, and create a survey_total variable. While the original proposal was to analyze the student sentiment in the survey comments, after importing and reviewing the data it became quickly apparent that the students did not avail themselves of the opportunity to provide reasonable commentary for us to undertake that analysis. Instead, I chose to analyze the number of clinical experiences the students participated against their scores on the program final examination. I also chose to look at the average survey scores for each of the clinical rotations the students participated with the overall examination score. Therefore, my hypotheses for this assignment are the following:

H1: The number of clinical experiences required to meet the ten (10) patient threshold correlated with student performance on the program final examination.

H2: The average survey score evaluating the clinical experience is correlated with student performance on the program final examination.

```
#  Pull in the data sets from the GitHub Repository
clinical_file <- "https://raw.githubusercontent.com/Aconrard/DATA607/main/Final%20Project/clinical.csv"
clinical_eval <- read.csv(clinical_file)
clinical_eval <- clinical_eval |> filter(!is.na(anon_id))

testing_file <- "https://raw.githubusercontent.com/Aconrard/DATA607/main/Final%20Project/testing.csv"
testing_results <- read.csv(testing_file)
testing_results <- testing_results |> filter(!is.na(anon_id))

# Match data sets on unique anonymous identifier
merged_df <- merge(clinical_eval, testing_results, by = "anon_id")

# Duplicate Check and Removal
duplicates <- merged_df$Submission_ID[duplicated(merged_df$Submission_ID)]
merged_df <- distinct(merged_df,Submission_ID, .keep_all = TRUE)

selected_data <- merged_df |>
  filter(!is.na(final_exam) & !is.na(anon_id)) |>
  select(anon_id, everything(), final_exam, final_exam_retest) |>
  rename(Orientation = Appropriate.Orientation.by.your.Preceptor.,
         Adequate = Adequate.Supervision.on.Ambulance.,
         Responsibility = Responsibilities.clearly.defined.by.your.Preceptor.,
         Hot_wash = The.hot.wash.afforded.me.the.opportunity.to.discuss.the.patient.encounter.,
         Preceptors = Preceptors.responsiveness.to.clinical.questionsby.student.,
         Objectives = Educational.objectives.accomplished.,
         Incorporated = Incorporated.as.member.of.crew.,
         Overall = Overall.educational.experience.,
         Submission_ID = Submission_ID,
         Submission_Date = Submission.Date) |>
  select(anon_id,
         Submission_ID,
         Submission_Date,
         final_exam,
         final_exam_retest,
         Orientation,
         Responsibility,
         Adequate,
         Preceptors,
         Hot_wash,
         Incorporated,
         Objectives,
         Overall)
selected_data$survey_total <- rowSums(selected_data[, (ncol(selected_data) - 7): ncol(selected_data)],
                                      na.rm = TRUE)

# Remove any observations where the survey total equals zero
selected_data1 <- selected_data |> filter(survey_total > 0)
```

Tidy and Transform

In this section we will once again tidy the data set and create some additional variables we want to use in the analysis. First, we will count the number of clinical experiences each student participated in and then put that into the existing data frame. We will then calculate the average survey rating of the clinical experience by each student. Each student was required to have at least ten (10) patient encounters. Some students were able to get that on one clinical rotation, and for others took several. For our purposes we felt the average of the number of survey totals was the fairest way to include it in our analysis. We perform those calculations and put them into the data frame with the other variables of interest. Finally, we extract the variables we will use for our analysis, which include anon_id(unique identifier), number_clinical (number of rotations), survey_average (average total of all surveys completed), and final exam score. We will call that data frame "summary_df".

```
# Count Clinicals for Each Participant
selected_data_count <- selected_data1 |>
  count(anon_id) |>
  rename(number_clinical = n)

# Assign back to the original data frame
selected_data_final <- left_join(selected_data1, selected_data_count, by = "anon_id")

# Calculate sum of all surveys completed by the participant
selected_data_final <- selected_data_final |>
  group_by(anon_id) |>
  mutate(total_survey = sum(survey_total)) |>
  ungroup()

# Average the sum of all surveys by the number fo clinical experiences completed.
```

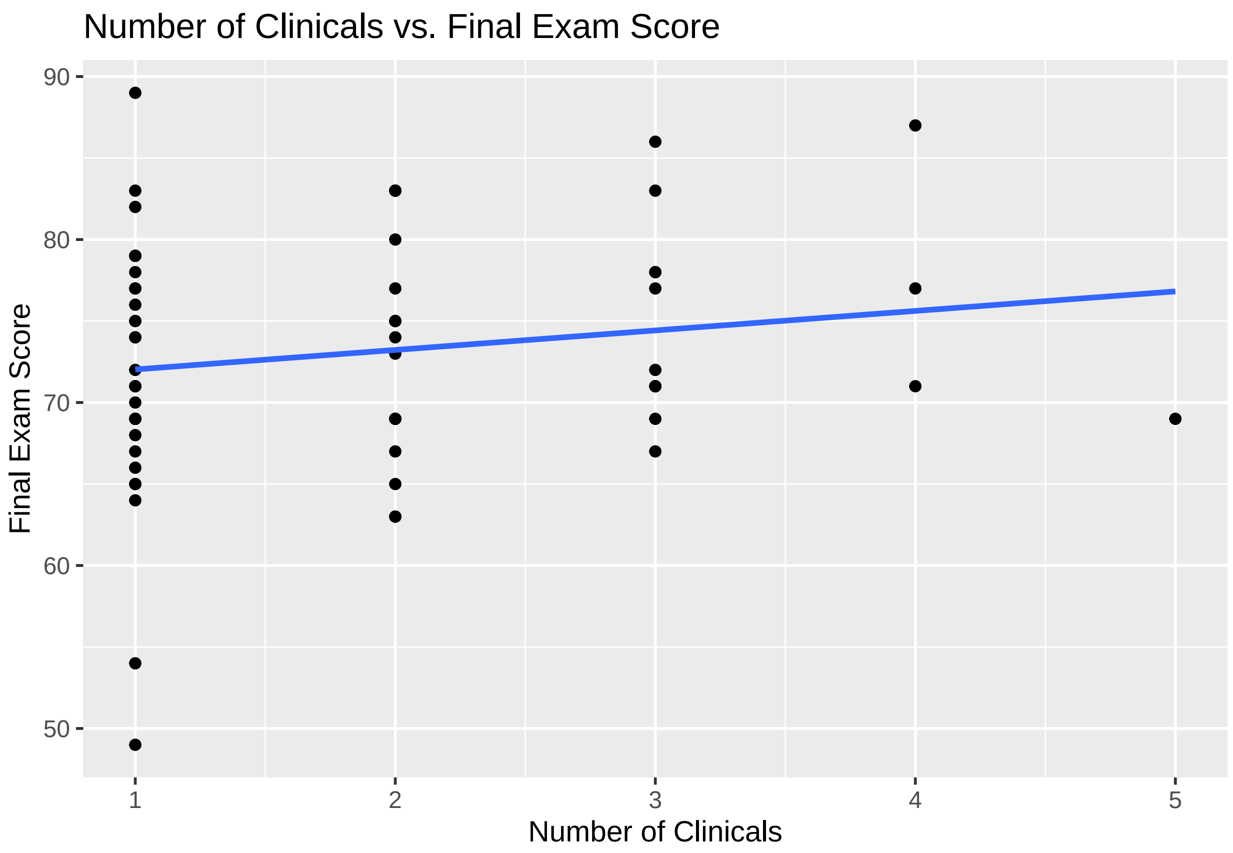
```r
selected_data_final <- selected_data_final |>
  group_by(anon_id) |>
  mutate(survey_average = total_survey/number_clinical) |>
  ungroup()

# Create the summary data frame with the necessary variables
summary_df <- selected_data_final |>
  group_by(anon_id) |>
  summarize(final_exam = first(final_exam),
            number_clinical = first(number_clinical),
            survey_average = first(survey_average)) |>
  ungroup()
summary_df
```

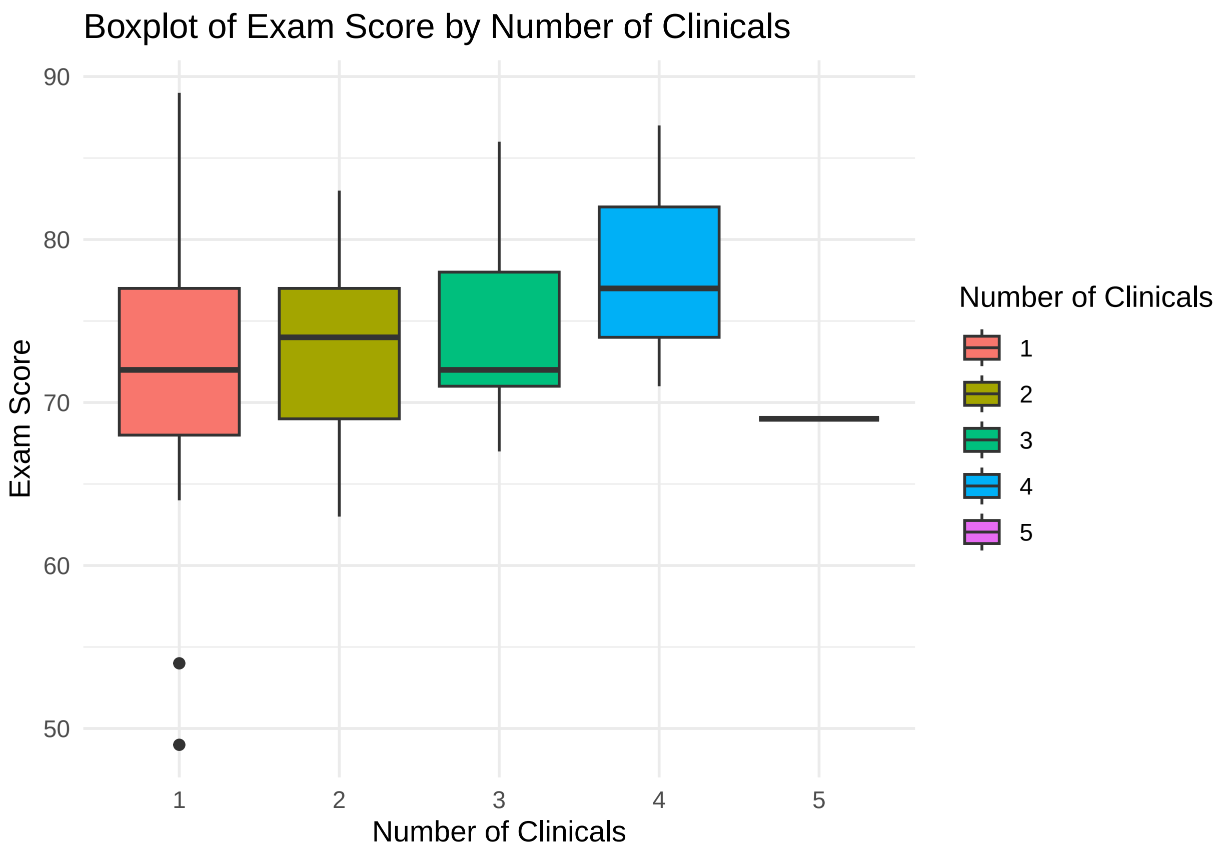
```
## # A tibble: 61 x 4
##    anon_id final_exam number_clinical survey_average
##      <int>      <int>           <int>          <dbl>
##   1       1         69               1           35
##   2       3         63               2           37.5
##   3       4         71               3           40
##   4       5         89               1           40
##   5       6         74               1           40
##   6       7         66               1           40
##   7       8         65               1           40
##   8       9         69               2           40
##   9      10         67               1           40
## 10      11         69               5           39.2
## # i 51 more rows
```

Graphic Presentation

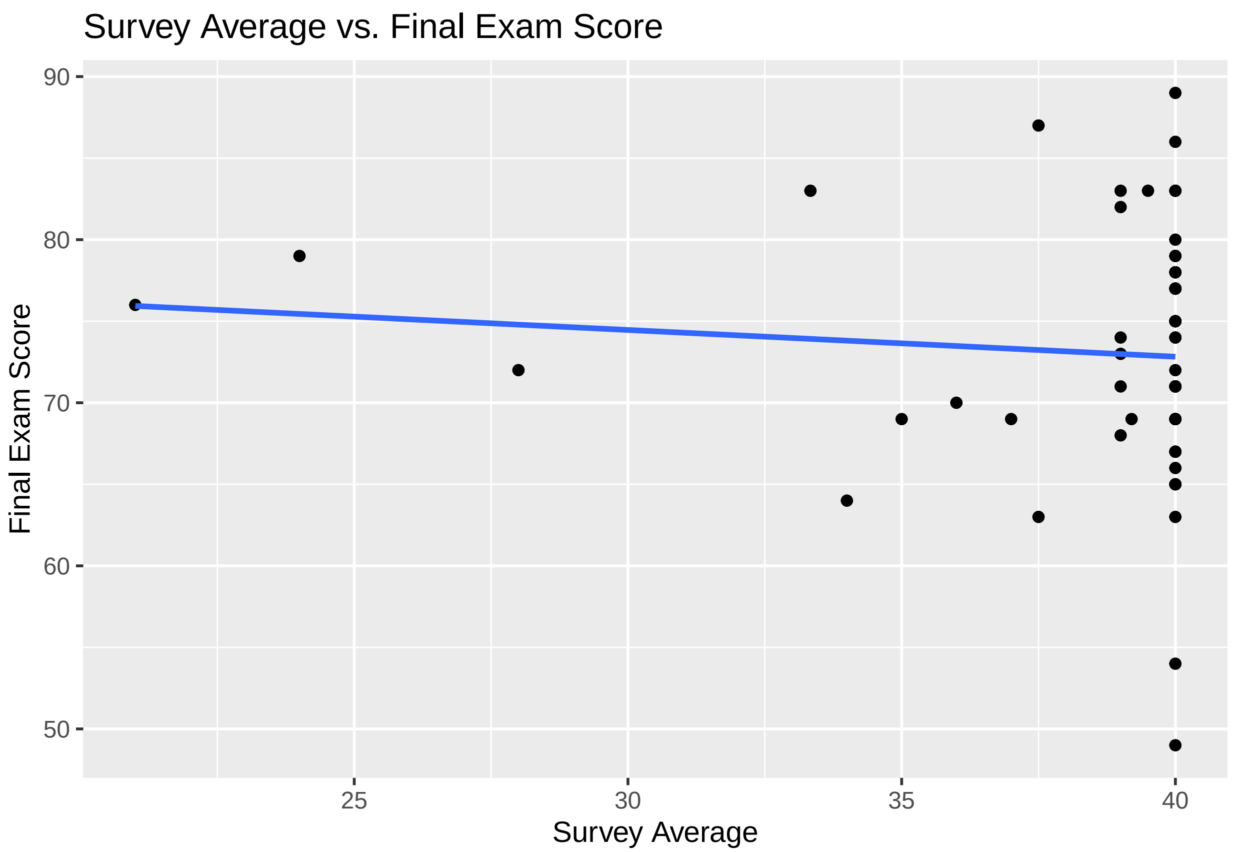
We provide three (3) plots to sense if there is indeed a direction associated with the number of clinical rotations and/or the survey_average. While the number of clinical experiences appear to increase the possibility of a higher exam score, the survey average evaluating the clinical experience seems to be not related with performance on the examination. However, we will conduct a linear regression to see if anything mathematically comes out of the analysis.

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


Number of Clinicals vs. Final Exam Score
Final Exam Score
90
80
70
60
50
1
2
3
4
5
Number of Clinicals



```
## `geom_smooth()` using formula = 'y ~ x'
```



Statistical Analysis

We ran four (4) models for this analysis. In the first model (full), we included both the survey average, number of clinicals, and an interaction term because a true average would be the result of one or more clinical experiences. Therefore, it would make senses to include that for possible confounding. The second model did not include the interaction term, and the third and fourth models were simple univariate regressions. The results are provided below:

```
Model_1 <- lm(final_exam ~ survey_average + number_clinical, data = summary_df)
summary(Model_1)
```

```
##
## Call:
## lm(formula = final_exam ~ survey_average + number_clinical, data = summary_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5871  -4.5871  -0.1614   4.0677  17.4129
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      79.4056    10.1774   7.802 1.32e-10 ***
## survey_average   -0.2291     0.2666  -0.859    0.394
## number_clinical   1.3452     0.9731   1.382    0.172
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.399 on 58 degrees of freedom
## Multiple R-squared:  0.03816,    Adjusted R-squared:  0.004995
## F-statistic: 1.151 on 2 and 58 DF,  p-value: 0.3236
```

```
Model_2 <- lm(final_exam ~ survey_average + number_clinical + (survey_average * number_clinical),
              data = summary_df)
summary(Model_2)
```

```
##
## Call:
## lm(formula = final_exam ~ survey_average + number_clinical +
##     (survey_average * number_clinical), data = summary_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8083  -3.9458  -0.4118   4.5335  17.1917
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     50.9968    25.4467   2.004   0.0498 *
## survey_average                   0.4996     0.6549   0.763   0.4487
## number_clinical                 26.3081    20.5327   1.281   0.2053
## survey_average:number_clinical  -0.6370     0.5233  -1.217   0.2286
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.368 on 57 degrees of freedom
## Multiple R-squared:  0.06253,    Adjusted R-squared:  0.01319
## F-statistic: 1.267 on 3 and 57 DF,  p-value: 0.2943
```

```
Model_3 <- lm(final_exam ~ number_clinical, data = summary_df)
summary(Model_3)
```

```
##
## Call:
## lm(formula = final_exam ~ number_clinical, data = summary_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0286  -4.2258  -0.0286   4.9714  16.9714
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      70.8313     2.0068  35.295   <2e-16 ***
## number_clinical   1.1973     0.9557   1.253    0.215
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.382 on 59 degrees of freedom
## Multiple R-squared:  0.02591,    Adjusted R-squared:  0.009403
## F-statistic:  1.57 on 1 and 59 DF,  p-value: 0.2152
```

```r
Model_4 <- lm(final_exam ~ survey_average, data = summary_df)
summary(Model_4)
```

```
##
## Call:
## lm(formula = final_exam ~ survey_average, data = summary_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.8236  -3.9547   0.0125   4.1764  16.1764
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     79.3793    10.2557    7.74 1.51e-10 ***
## survey_average  -0.1639     0.2644   -0.62    0.538
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.456 on 59 degrees of freedom
## Multiple R-squared:  0.006471,   Adjusted R-squared:  -0.01037
## F-statistic: 0.3843 on 1 and 59 DF,  p-value: 0.5377
```

Discussion and Conclusion

While there appears to be an upward trend associated with the number of clinical experiences and performance on the examination, it does not appear to be statistically significant. The survey average neither seems to be visually, nor statistically, related to performance on the written examination. As a significant point, it appears the students may simply “whip” responses on the survey with little interest in providing reliable information for analysis. Also, while the clinical evaluation data set has 1378 observations, only 150 of them were associated with the test results data set we retrieved from the Platinum Education Group. The 150 evaluations were associated with only 61 students, and that number may simply be too small to provide statistical evidence for a correlation with examination performance. Also, the R-squared and adjusted R-square for the models were poor, indicating that the variance in the exam performance is not sufficiently explained by the independent variables. At this time, we can draw no conclusions as to whether student performance on the written comprehensive examination is correlated with the number of clinical rotations by the student, or the average rating of the clinical experience provided by the student on the survey. Additional research could attempt to construct a larger data set of testing results to compare sufficient numbers of student in each subgroup of clinical numbers to test for significance.