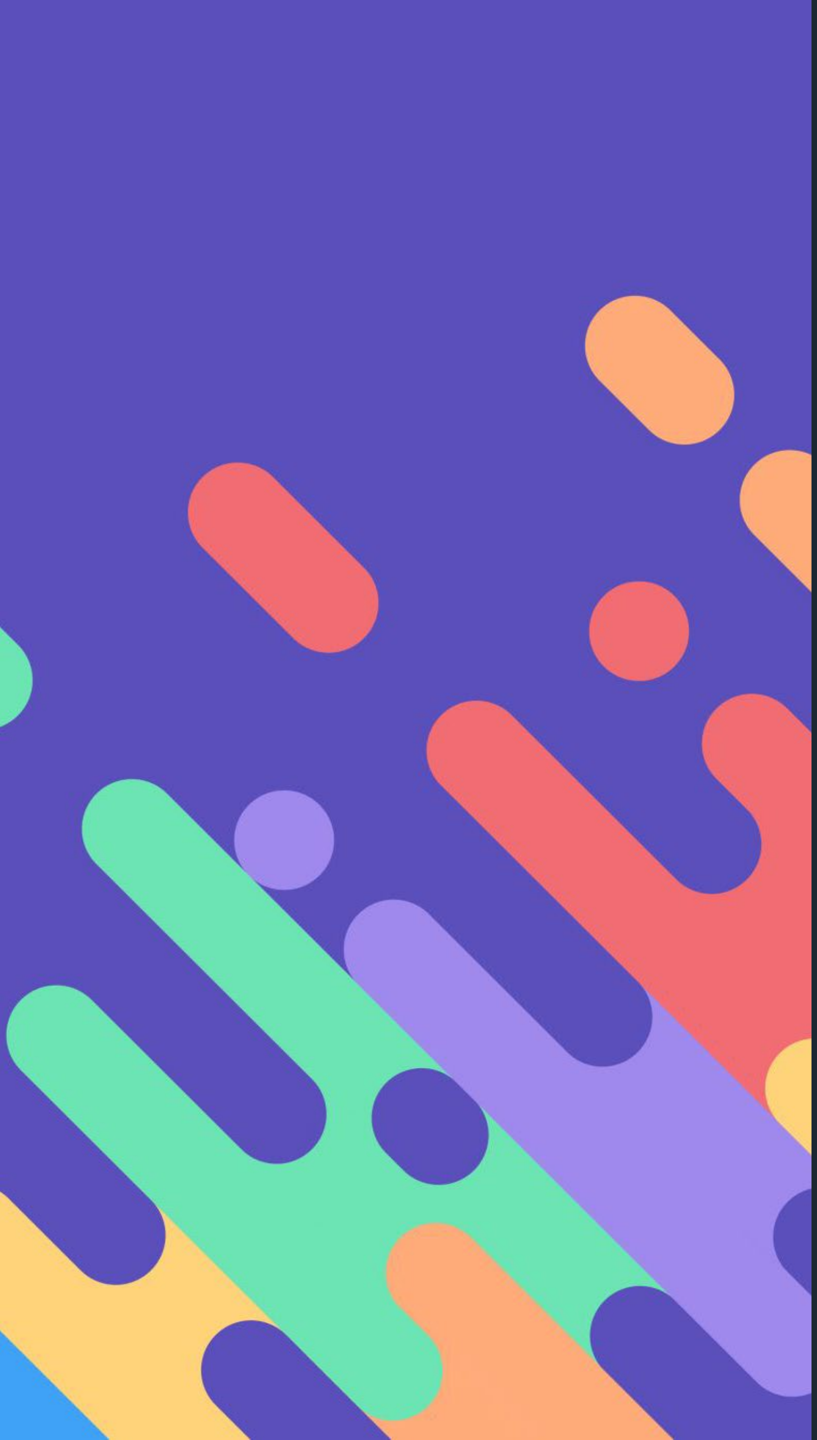


STUDENT CLINICAL EVALUATION AND PERFORMANCE ON COMPREHENSIVE EXAMINATION

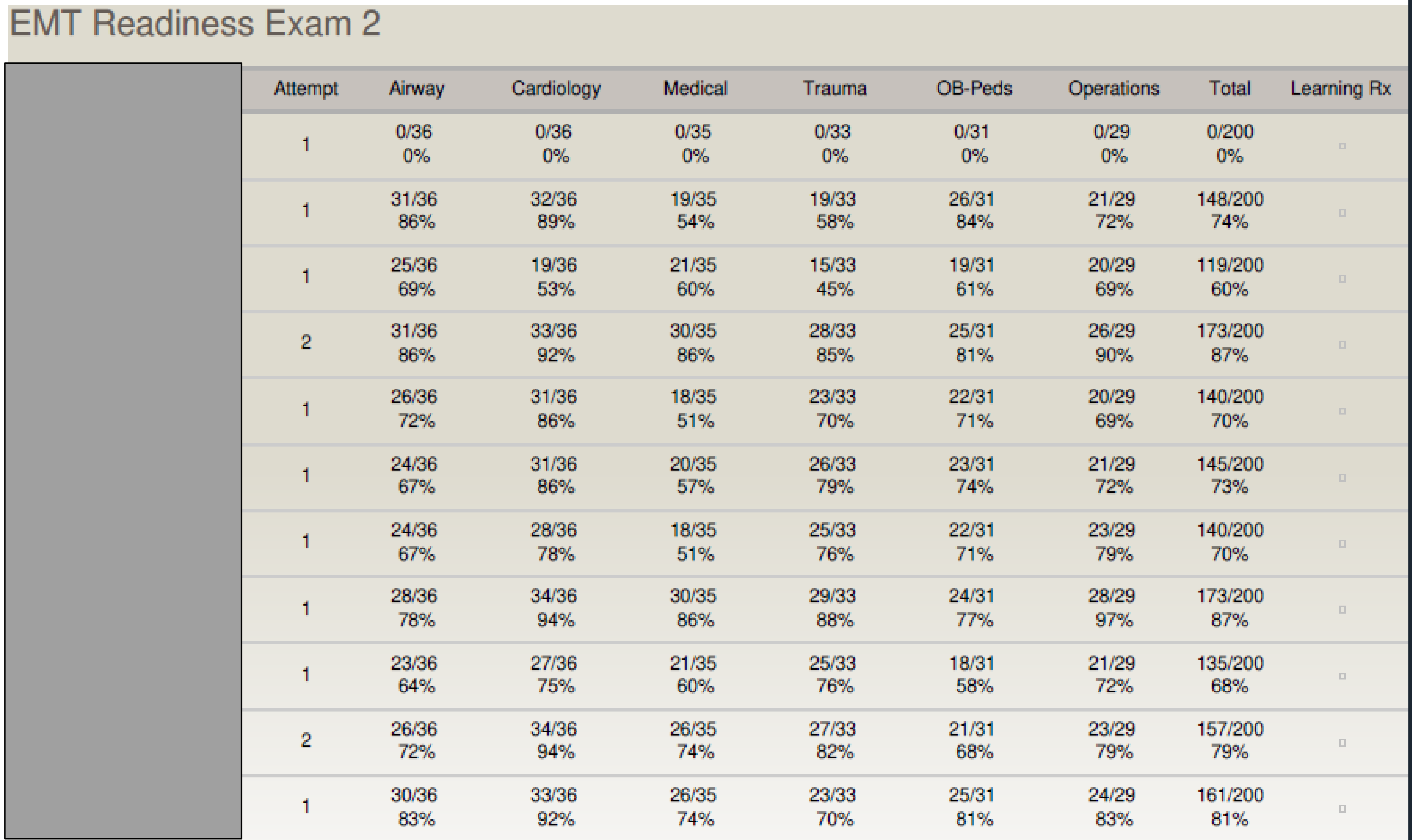
Otherwise known as….

The Trip into the Tidyverse of Madness.

WHAT WAS THE INITIAL PLAN?

“……..perform a practical exercise related to a real-world environment. Student survey information is currently contained on a platform that permits access through an API, while comprehensive testing results reside on a distinctly different platform that only allows for the data to be downloaded in PDF format. The challenge for this project is how to link the data from the two platforms to analyze student sentiment and ratings of clinical experience with their overall performance on the comprehensive examinations, while remaining consistent with protections afforded under federal law.”

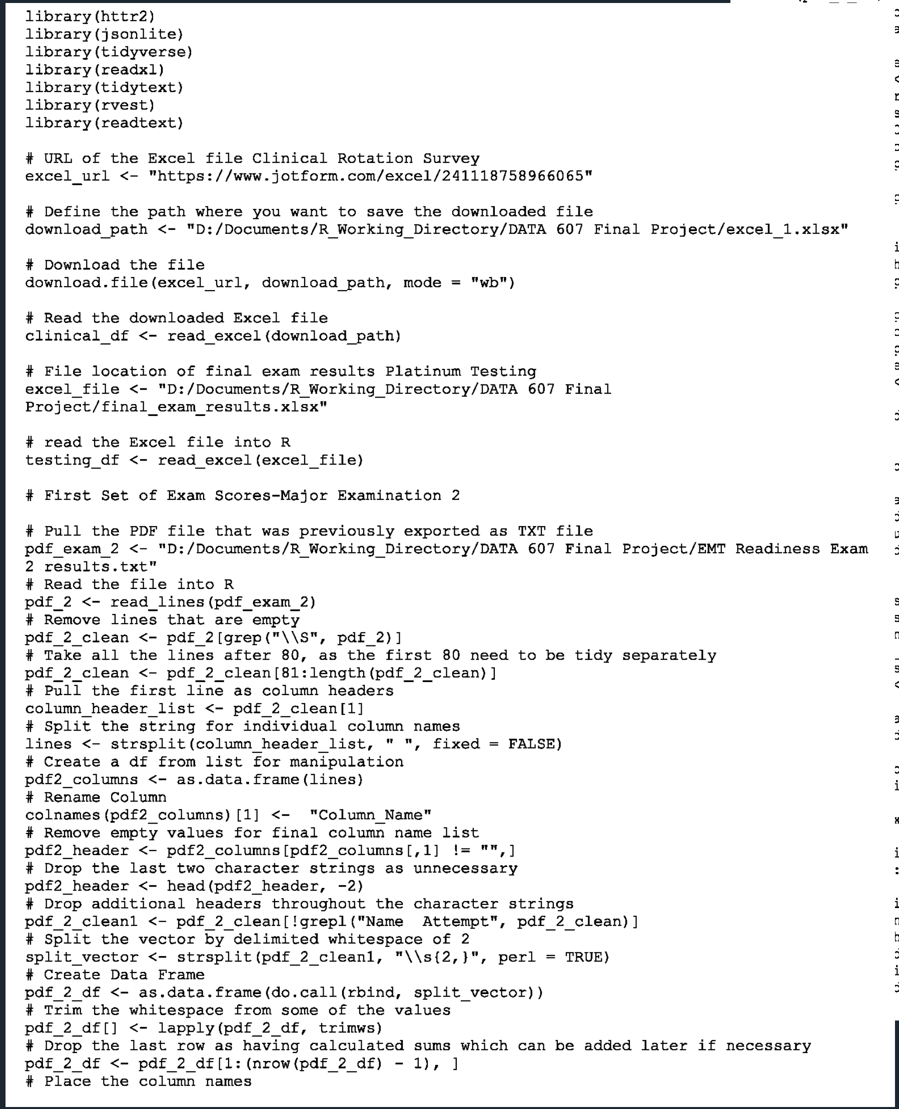
WHAT I GOT.....NOT SO MUCH



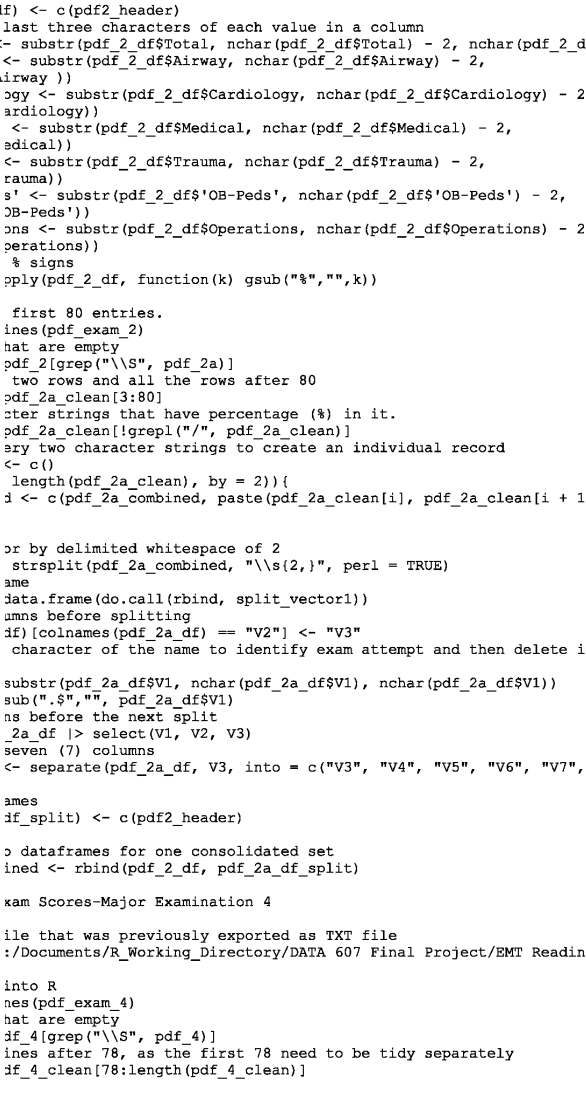
EMT Readiness Exam 2

[redacted]	Attempt	Airway	Cardiology	Medical	Trauma	OB-Peds	Operations	Total	Learning Rx
[redacted]	1	0/36 0%	0/36 0%	0/35 0%	0/33 0%	0/31 0%	0/29 0%	0/200 0%	□
[redacted]	1	31/36 86%	32/36 89%	19/35 54%	19/33 58%	26/31 84%	21/29 72%	148/200 74%	□
[redacted]	1	25/36 69%	19/36 53%	21/35 60%	15/33 45%	19/31 61%	20/29 69%	119/200 60%	□
[redacted]	2	31/36 86%	33/36 92%	30/35 86%	28/33 85%	25/31 81%	26/29 90%	173/200 87%	□
[redacted]	1	26/36 72%	31/36 86%	18/35 51%	23/33 70%	22/31 71%	20/29 69%	140/200 70%	□
[redacted]	1	24/36 67%	31/36 86%	20/35 57%	26/33 79%	23/31 74%	21/29 72%	145/200 73%	□
[redacted]	1	24/36 67%	28/36 78%	18/35 51%	25/33 76%	22/31 71%	23/29 79%	140/200 70%	□
[redacted]	1	28/36 78%	34/36 94%	30/35 86%	29/33 88%	24/31 77%	28/29 97%	173/200 87%	□
[redacted]	1	23/36 64%	27/36 75%	21/35 60%	25/33 76%	18/31 58%	21/29 72%	135/200 68%	□
[redacted]	2	26/36 72%	34/36 94%	26/35 74%	27/33 82%	21/31 68%	23/29 79%	157/200 79%	□
[redacted]	1	30/36 83%	33/36 92%	26/35 74%	23/33 70%	25/31 81%	24/29 83%	161/200 81%	□

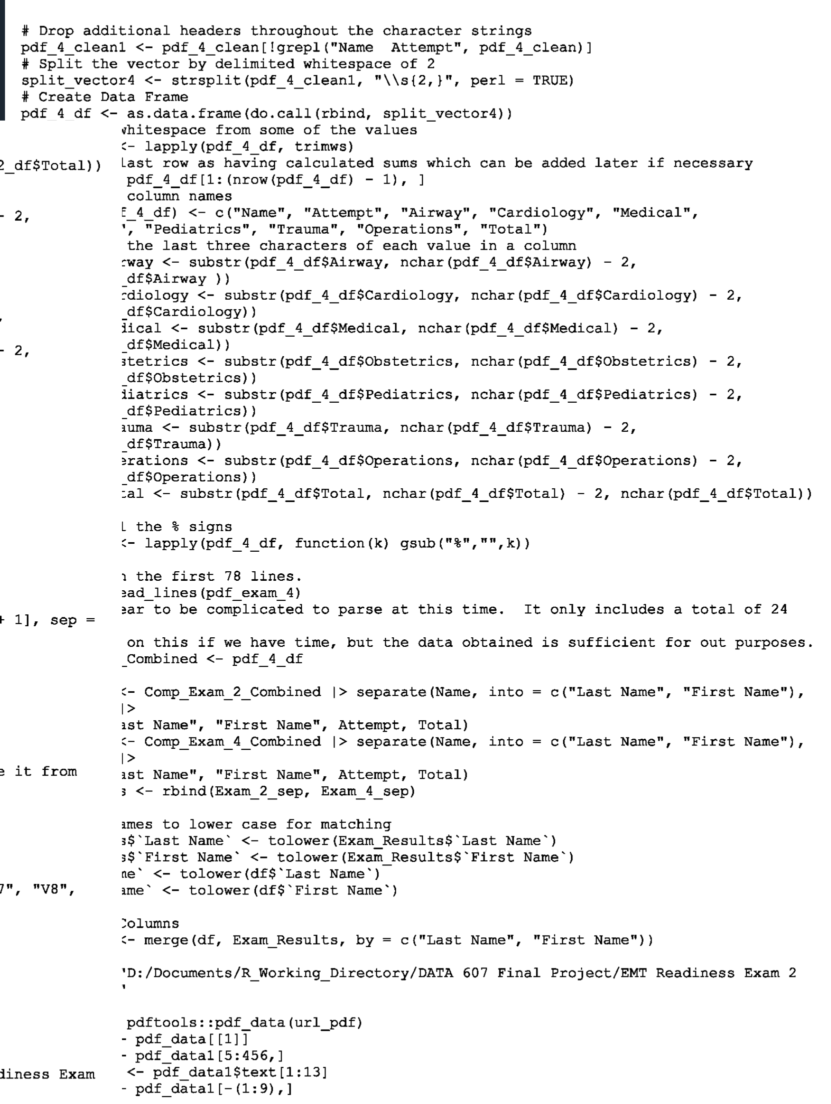
LET'S TIDY THIS UP A BIT…….



```
library(httr2)
library(jsonlite)
library(tidyverse)
library(readxl)
library(tidytext)
library(rvest)
library(readtext)

# URL of the Excel file Clinical Rotation Survey
excel_url <- "https://www.jotform.com/excel/241118758966065"

# Define the path where you want to save the downloaded file
download_path <- "D:/Documents/R_Working_Directory/DATA 607 Final Project/excel_1.xlsx"

# Download the file
download.file(excel_url, download_path, mode = "wb")

# Read the downloaded Excel file
clinical_df <- read_excel(download_path)

# File location of final exam results Platinum Testing
excel_file <- "D:/Documents/R_Working_Directory/DATA 607 Final
Project/final_exam_results.xlsx"

# read the Excel file into R
testing_df <- read_excel(excel_file)

# First Set of Exam Scores-Major Examination 2

# Pull the PDF file that was previously exported as TXT file
pdf_exam_2 <- "D:/Documents/R_Working_Directory/DATA 607 Final Project/EMT Readiness Exam
2 results.txt"
# Read the file into R
pdf_2 <- read_lines(pdf_exam_2)
# Remove lines that are empty
pdf_2_clean <- pdf_2[grep("\\S", pdf_2)]
# Take all the lines after 80, as the first 80 need to tidy separately
pdf_2_clean <- pdf_2_clean[81:length(pdf_2_clean)]
# Pull the first line as column headers
column_header_list <- pdf_2_clean[1]
# Split the string for individual column names
lines <- strsplit(column_header_list, " ", fixed = FALSE)
# Create a df from list for manipulation
pdf2_columns <- as.data.frame(lines)
# Rename Column
colnames(pdf2_columns)[1] <-  "Column_Name"
# Remove empty values for final column name list
pdf2_header <- pdf2_columns[pdf2_columns[,1] != "",]
# Drop the last two character strings as unnecessary
pdf2_header <- head(pdf2_header, -2)
# Drop additional headers throughout the character strings
pdf_2_clean1 <- pdf_2_clean[!grepl("Name  Attempt", pdf_2_clean)]
# Split the vector by delimited whitespace of 2
split_vector <- strsplit(pdf_2_clean1, "\\s{2,}", perl = TRUE)
# Create Data Frame
pdf_2_df <- as.data.frame(do.call(rbind, split_vector))
# Trim the whitespace from some of the values
pdf_2_df[] <- lapply(pdf_2_df, trimws)
# Drop the last row as having calculated sums which can be added later if necessary
pdf_2_df <- pdf_2_df[1:(nrow(pdf_2_df) - 1), ]
# Place the column names
```

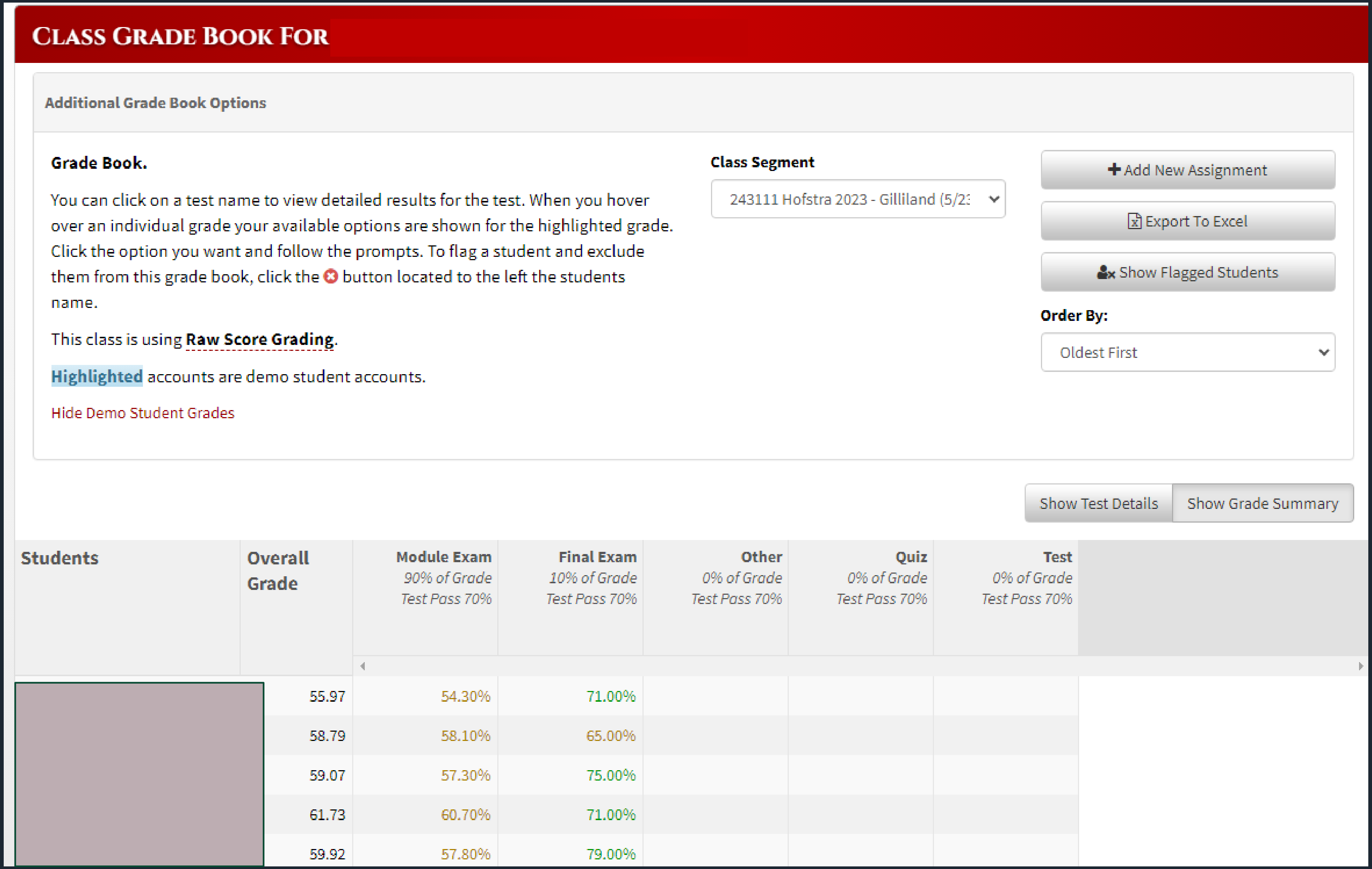


```
colnames(pdf_2_df) <- c(pdf2_header)
# Only keep the last three characters of each value in a column
pdf_2_df$Total <- substr(pdf_2_df$Total, nchar(pdf_2_df$Total) - 2, nchar(pdf_2_df$Total))
pdf_2_df$Airway <- substr(pdf_2_df$Airway, nchar(pdf_2_df$Airway) - 2,
nchar(pdf_2_df$Airway ))
                 ogy <- substr(pdf_2_df$Cardiology, nchar(pdf_2_df$Cardiology) - 2,
                 ardiology))
                  <- substr(pdf_2_df$Medical, nchar(pdf_2_df$Medical) - 2,
                 edical))
                 <- substr(pdf_2_df$Trauma, nchar(pdf_2_df$Trauma) - 2,
                 rauma))
                 s' <- substr(pdf_2_df$'OB-Peds', nchar(pdf_2_df$'OB-Peds') - 2,
                 OB-Peds'))
                 ons <- substr(pdf_2_df$Operations, nchar(pdf_2_df$Operations) - 2,
                 perations))
                  % signs
                 pply(pdf_2_df, function(k) gsub("%","",k))

                  first 80 entries.
                 ines(pdf_exam_2)
                 hat are empty
                 pdf_2[grep("\\S", pdf_2a)]
                  two rows and all the rows after 80
                 pdf_2a_clean[3:80]
                 cter strings that have percentage (%) in it.
                 pdf_2a_clean[!grepl("/", pdf_2a_clean)]
                 ery two character strings to create an individual record
                 <- c()
                  length(pdf_2a_clean), by = 2)){
                 d <- c(pdf_2a_combined, paste(pdf_2a_clean[i], pdf_2a_clean[i + 1], sep =


                 or by delimited whitespace of 2
                  strsplit(pdf_2a_combined, "\\s{2,}", perl = TRUE)
                 ame
                 data.frame(do.call(rbind, split_vector1))
                 umns before splitting
                 df)[colnames(pdf_2a_df) == "V2"] <- "V3"
                  character of the name to identify exam attempt and then delete it from

                 substr(pdf_2a_df$V1, nchar(pdf_2a_df$V1), nchar(pdf_2a_df$V1))
                 sub(".$","", pdf_2a_df$V1)
                 ns before the next split
                 _2a_df |> select(V1, V2, V3)
                 seven (7) columns
                 <- separate(pdf_2a_df, V3, into = c("V3", "V4", "V5", "V6", "V7", "V8",

                 ames
                 df_split) <- c(pdf2_header)

                 o dataframes for one consolidated set
                 ined <- rbind(pdf_2_df, pdf_2a_df_split)

                 xam Scores-Major Examination 4

                 ile that was previously exported as TXT file
                 :/Documents/R_Working_Directory/DATA 607 Final Project/EMT Readiness Exam

                 into R
                 nes(pdf_exam_4)
                 hat are empty
                 df_4[grep("\\S", pdf_4)]
                 ines after 78, as the first 78 need to be tidy separately
                 df_4_clean[78:length(pdf_4_clean)]
```



```
# Drop additional headers throughout the character strings
pdf_4_clean1 <- pdf_4_clean[!grepl("Name  Attempt", pdf_4_clean)]
# Split the vector by delimited whitespace of 2
split_vector4 <- strsplit(pdf_4_clean1, "\\s{2,}", perl = TRUE)
# Create Data Frame
pdf_4_df <- as.data.frame(do.call(rbind, split_vector4))
      whitespace from some of the values
      <- lapply(pdf_4_df, trimws)
      last row as having calculated sums which can be added later if necessary
       pdf_4_df[1:(nrow(pdf_4_df) - 1), ]
       column names
      f_4_df) <- c("Name", "Attempt", "Airway", "Cardiology", "Medical",
      ", "Pediatrics", "Trauma", "Operations", "Total")
       the last three characters of each value in a column
      rway <- substr(pdf_4_df$Airway, nchar(pdf_4_df$Airway) - 2,
       _df$Airway ))
      rdiology <- substr(pdf_4_df$Cardiology, nchar(pdf_4_df$Cardiology) - 2,
       _df$Cardiology))
      dical <- substr(pdf_4_df$Medical, nchar(pdf_4_df$Medical) - 2,
       _df$Medical))
      stetrics <- substr(pdf_4_df$Obstetrics, nchar(pdf_4_df$Obstetrics) - 2,
       _df$Obstetrics))
      diatrics <- substr(pdf_4_df$Pediatrics, nchar(pdf_4_df$Pediatrics) - 2,
       _df$Pediatrics))
      auma <- substr(pdf_4_df$Trauma, nchar(pdf_4_df$Trauma) - 2,
       _df$Trauma))
      erations <- substr(pdf_4_df$Operations, nchar(pdf_4_df$Operations) - 2,
       _df$Operations))
      tal <- substr(pdf_4_df$Total, nchar(pdf_4_df$Total) - 2, nchar(pdf_4_df$Total))

      l the % signs
      <- lapply(pdf_4_df, function(k) gsub("%","",k))

      n the first 78 lines.
      ead_lines(pdf_exam_4)
      ear to be complicated to parse at this time.  It only includes a total of 24

       on this if we have time, but the data obtained is sufficient for out purposes.
       _Combined <- pdf_4_df

      <- Comp_Exam_2_Combined |> separate(Name, into = c("Last Name", "First Name"),
      |>
      ast Name", "First Name", Attempt, Total)
      <- Comp_Exam_4_Combined |> separate(Name, into = c("Last Name", "First Name"),
      |>
      ast Name", "First Name", Attempt, Total)
      s <- rbind(Exam_2_sep, Exam_4_sep)

      ames to lower case for matching
      s$`Last Name` <- tolower(Exam_Results$`Last Name`)
      s$`First Name` <- tolower(Exam_Results$`First Name`)
      me` <- tolower(df$`Last Name`)
      ame` <- tolower(df$`First Name`)

      Columns
      <- merge(df, Exam_Results, by = c("Last Name", "First Name"))

      'D:/Documents/R_Working_Directory/DATA 607 Final Project/EMT Readiness Exam 2
      '

       pdftools::pdf_data(url_pdf)
      - pdf_data[[1]]
      - pdf_data1[5:456,]
       <- pdf_data1$text[1:13]
      - pdf_data1[-(1:9),]
```

LET'S TRY THIS AGAIN….



Class Grade Book For

Additional Grade Book Options

Grade Book.

You can click on a test name to view detailed results for the test. When you hover over an individual grade your available options are shown for the highlighted grade. Click the option you want and follow the prompts. To flag a student and exclude them from this grade book, click the ✖ button located to the left the students name.

This class is using **Raw Score Grading**.

Highlighted accounts are demo student accounts.

Hide Demo Student Grades

Class Segment

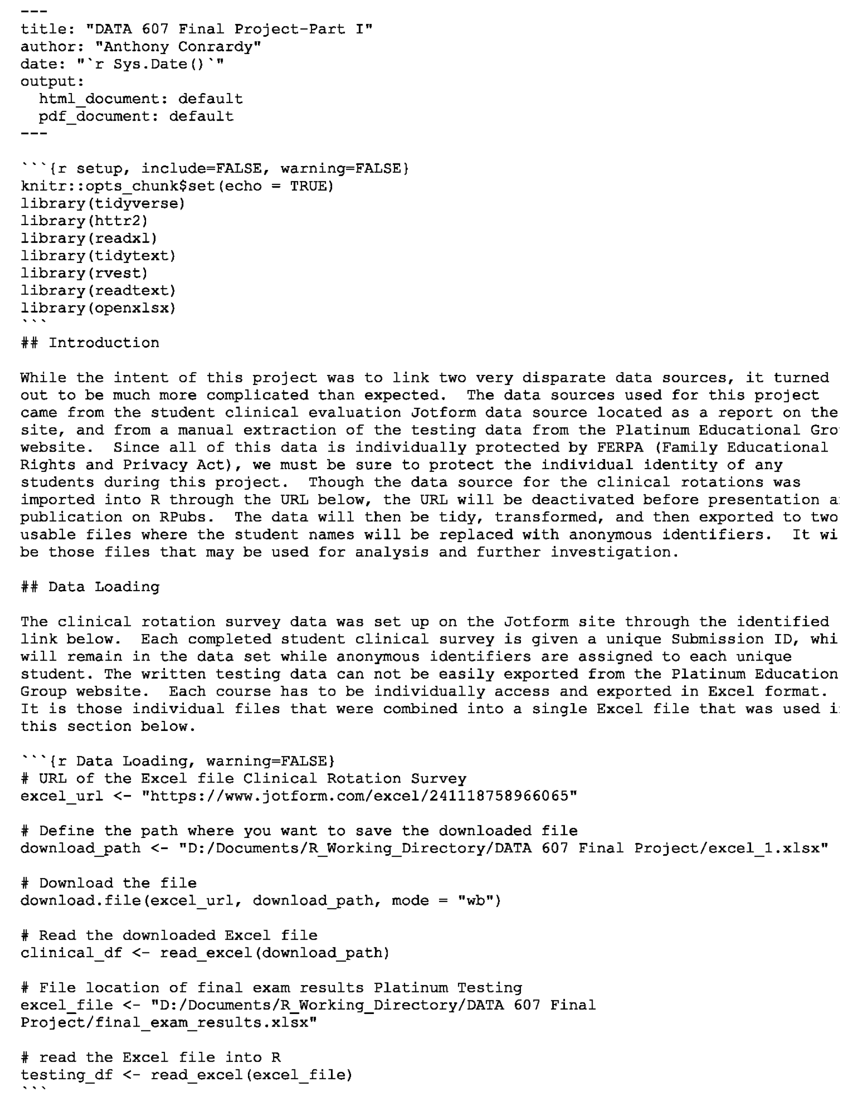
243111 Hofstra 2023 - Gilliland (5/2:

Order By:

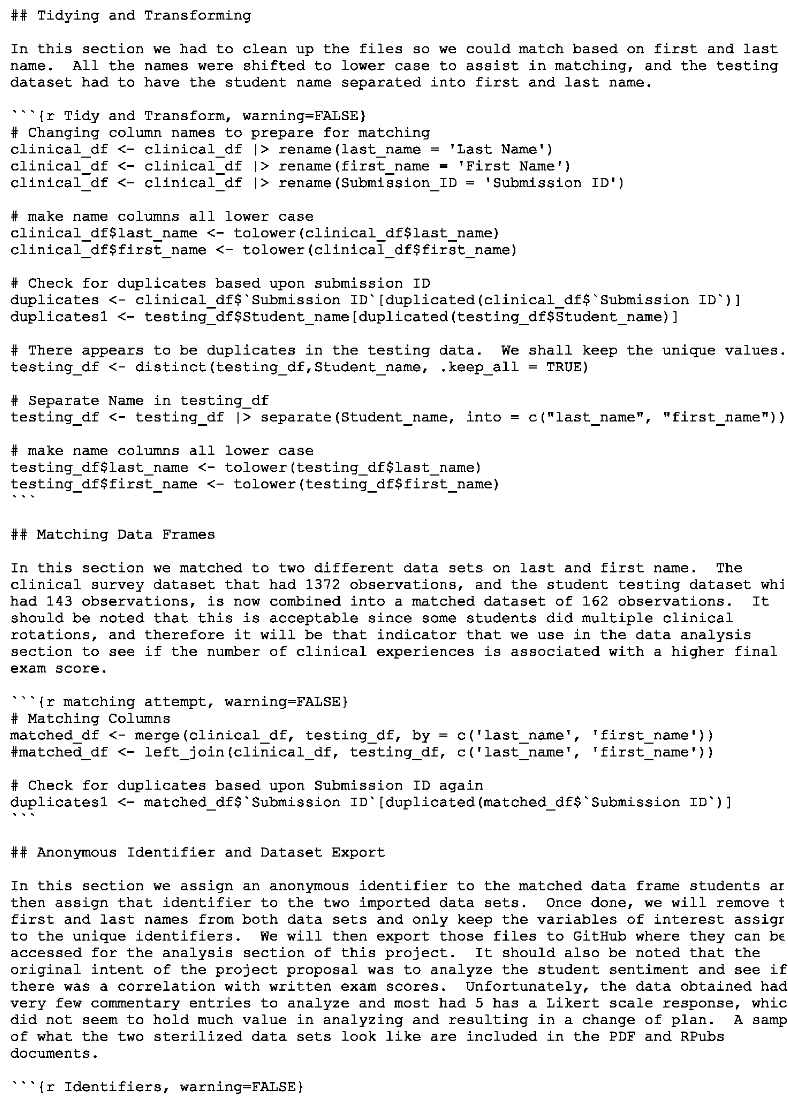
Oldest First

Students	Overall Grade	Module Exam *90% of Grade* *Test Pass 70%*	Final Exam *10% of Grade* *Test Pass 70%*	Other *0% of Grade* *Test Pass 70%*	Quiz *0% of Grade* *Test Pass 70%*	Test *0% of Grade* *Test Pass 70%*
[redacted]	55.97	54.30%	71.00%			
[redacted]	58.79	58.10%	65.00%			
[redacted]	59.07	57.30%	75.00%			
[redacted]	61.73	60.70%	71.00%			
[redacted]	59.92	57.80%	79.00%			

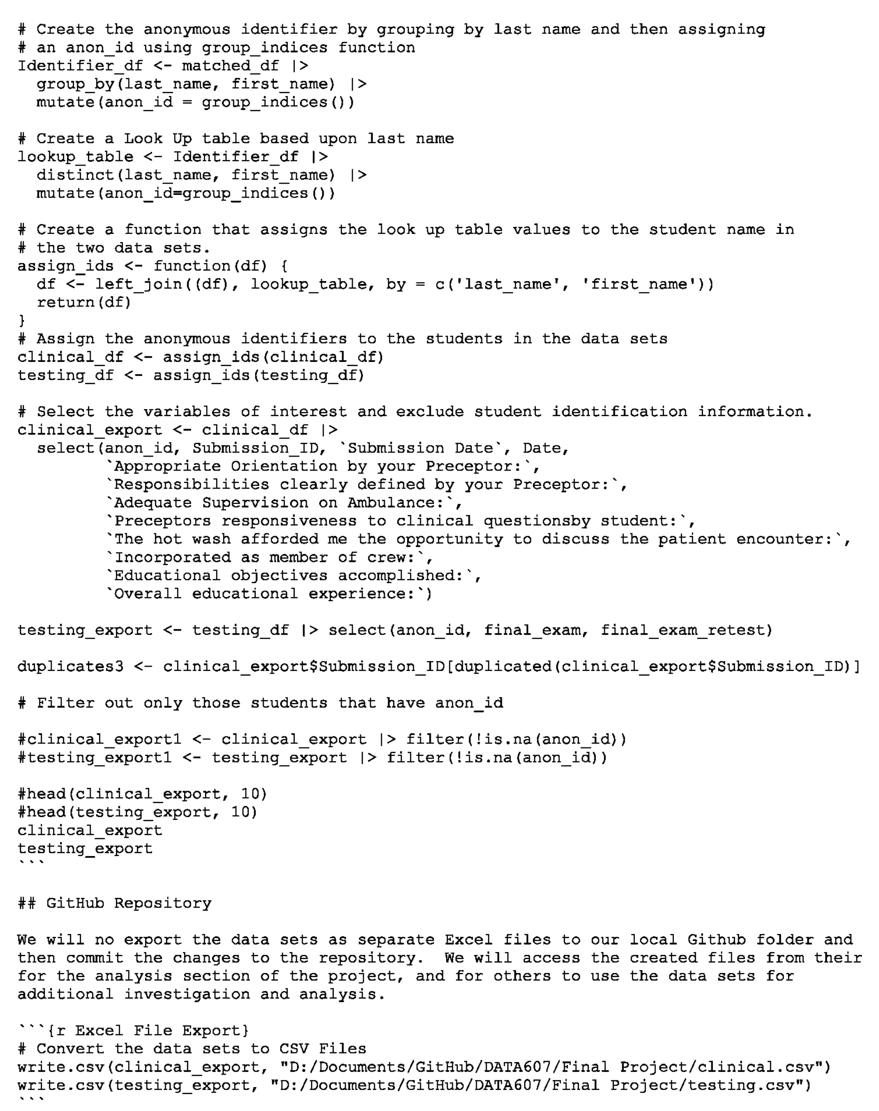
MAKING SOME PROGRESS……



~~~
---
title: "DATA 607 Final Project-Part I"
author: "Anthony Conrardy"
date: "`r Sys.Date()`"
output:
  html_document: default
  pdf_document: default
---

```{r setup, include=FALSE, warning=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(httr2)
library(readxl)
library(tidytext)
library(rvest)
library(readtext)
library(openxlsx)
```
## Introduction

While the intent of this project was to link two very disparate data sources, it turned
out to be much more complicated than expected.  The data sources used for this project
came from the student clinical evaluation Jotform data source located as a report on the
site, and from a manual extraction of the testing data from the Platinum Educational Gro
website.  Since all of this data is individually protected by FERPA (Family Educational
Rights and Privacy Act), we must be sure to protect the individual identity of any
students during this project.  Though the data source for the clinical rotations was
imported into R through the URL below, the URL will be deactivated before presentation a
publication on RPubs.  The data will then be tidy, transformed, and then exported to two
usable files where the student names will be replaced with anonymous identifiers.  It wi
be those files that may be used for analysis and further investigation.

## Data Loading

The clinical rotation survey data was set up on the Jotform site through the identified
link below.  Each completed student clinical survey is given a unique Submission ID, whi
will remain in the data set while anonymous identifiers are assigned to each unique
student. The written testing data can not be easily exported from the Platinum Education
Group website.  Each course has to be individually access and exported in Excel format.
It is those individual files that were combined into a single Excel file that was used i
this section below.

```{r Data Loading, warning=FALSE}
# URL of the Excel file Clinical Rotation Survey
excel_url <- "https://www.jotform.com/excel/241118758966065"

# Define the path where you want to save the downloaded file
download_path <- "D:/Documents/R_Working_Directory/DATA 607 Final Project/excel_1.xlsx"

# Download the file
download.file(excel_url, download_path, mode = "wb")

# Read the downloaded Excel file
clinical_df <- read_excel(download_path)

# File location of final exam results Platinum Testing
excel_file <- "D:/Documents/R_Working_Directory/DATA 607 Final
Project/final_exam_results.xlsx"

# read the Excel file into R
testing_df <- read_excel(excel_file)
```
~~~

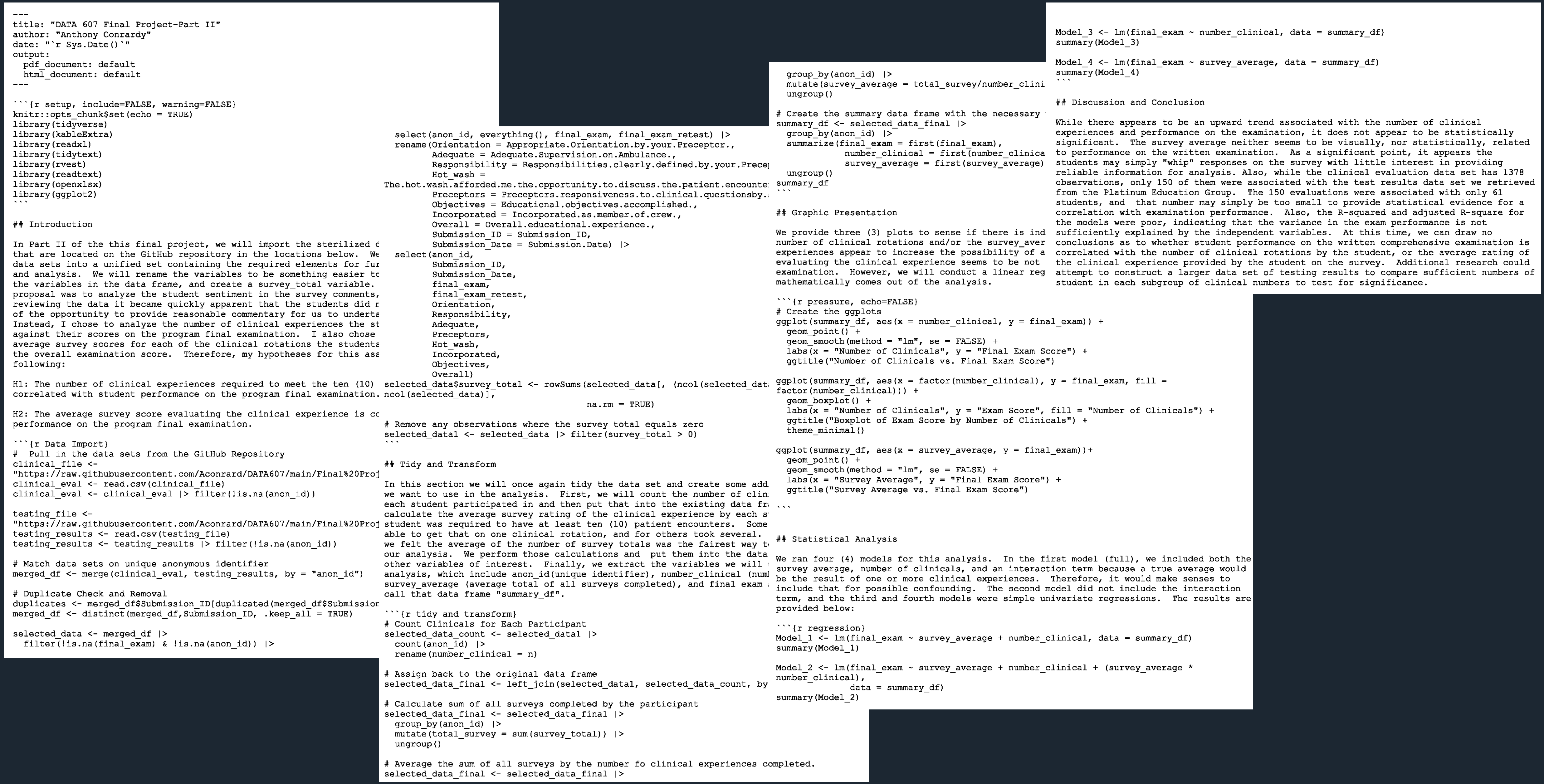


~~~
## Tidying and Transforming

In this section we had to clean up the files so we could match based on first and last
name.  All the names were shifted to lower case to assist in matching, and the testing
dataset had to have the student name separated into first and last name.

```{r Tidy and Transform, warning=FALSE}
# Changing column names to prepare for matching
clinical_df <- clinical_df |> rename(last_name = 'Last Name')
clinical_df <- clinical_df |> rename(first_name = 'First Name')
clinical_df <- clinical_df |> rename(Submission_ID = 'Submission ID')

# make name columns all lower case
clinical_df$last_name <- tolower(clinical_df$last_name)
clinical_df$first_name <- tolower(clinical_df$first_name)

# Check for duplicates based upon submission ID
duplicates <- clinical_df$`Submission ID`[duplicated(clinical_df$`Submission ID`)]
duplicates1 <- testing_df$Student_name[duplicated(testing_df$Student_name)]

# There appears to be duplicates in the testing data.  We shall keep the unique values.
testing_df <- distinct(testing_df,Student_name, .keep_all = TRUE)

# Separate Name in testing_df
testing_df <- testing_df |> separate(Student_name, into = c("last_name", "first_name"))

# make name columns all lower case
testing_df$last_name <- tolower(testing_df$last_name)
testing_df$first_name <- tolower(testing_df$first_name)
```

## Matching Data Frames

In this section we matched to two different data sets on last and first name.  The
clinical survey dataset that had 1372 observations, and the student testing dataset whi
had 143 observations, is now combined into a matched dataset of 162 observations.  It
should be noted that this is acceptable since some students did multiple clinical
rotations, and therefore it will be that indicator that we use in the data analysis
section to see if the number of clinical experiences is associated with a higher final
exam score.

```{r matching attempt, warning=FALSE}
# Matching Columns
matched_df <- merge(clinical_df, testing_df, by = c('last_name', 'first_name'))
#matched_df <- left_join(clinical_df, testing_df, c('last_name', 'first_name'))

# Check for duplicates based upon Submission ID again
duplicates1 <- matched_df$`Submission ID`[duplicated(matched_df$`Submission ID`)]
```

## Anonymous Identifier and Dataset Export

In this section we assign an anonymous identifier to the matched data frame students ar
then assign that identifier to the two imported data sets.  Once done, we will remove t
first and last names from both data sets and only keep the variables of interest assigr
to the unique identifiers.  We will then export those files to GitHub where they can be
accessed for the analysis section of this project.  It should also be noted that the
original intent of the project proposal was to analyze the student sentiment and see if
there was a correlation with written exam scores.  Unfortunately, the data obtained had
very few commentary entries to analyze and most had 5 has a Likert scale response, which
did not seem to hold much value in analyzing and resulting in a change of plan.  A sample
of what the two sterilized data sets look like are included in the PDF and RPubs
documents.

```{r Identifiers, warning=FALSE}
~~~



~~~
# Create the anonymous identifier by grouping by last name and then assigning
# an anon_id using group_indices function
Identifier_df <- matched_df |>
  group_by(last_name, first_name) |>
  mutate(anon_id = group_indices())

# Create a Look Up table based upon last name
lookup_table <- Identifier_df |>
  distinct(last_name, first_name) |>
  mutate(anon_id=group_indices())

# Create a function that assigns the look up table values to the student name in
# the two data sets.
assign_ids <- function(df) {
  df <- left_join((df), lookup_table, by = c('last_name', 'first_name'))
  return(df)
}
# Assign the anonymous identifiers to the students in the data sets
clinical_df <- assign_ids(clinical_df)
testing_df <- assign_ids(testing_df)

# Select the variables of interest and exclude student identification information.
clinical_export <- clinical_df |>
  select(anon_id, Submission_ID, `Submission Date`, Date,
         `Appropriate Orientation by your Preceptor:`,
         `Responsibilities clearly defined by your Preceptor:`,
         `Adequate Supervision on Ambulance:`,
         `Preceptors responsiveness to clinical questionsby student:`,
         `The hot wash afforded me the opportunity to discuss the patient encounter:`,
         `Incorporated as member of crew:`,
         `Educational objectives accomplished:`,
         `Overall educational experience:`)

testing_export <- testing_df |> select(anon_id, final_exam, final_exam_retest)

duplicates3 <- clinical_export$Submission_ID[duplicated(clinical_export$Submission_ID)]

# Filter out only those students that have anon_id

#clinical_export1 <- clinical_export |> filter(!is.na(anon_id))
#testing_export1 <- testing_export |> filter(!is.na(anon_id))

#head(clinical_export, 10)
#head(testing_export, 10)
clinical_export
testing_export
```

## GitHub Repository

We will no export the data sets as separate Excel files to our local Github folder and
then commit the changes to the repository.  We will access the created files from their
for the analysis section of the project, and for others to use the data sets for
additional investigation and analysis.

```{r Excel File Export}
# Convert the data sets to CSV Files
write.csv(clinical_export, "D:/Documents/GitHub/DATA607/Final Project/clinical.csv")
write.csv(testing_export, "D:/Documents/GitHub/DATA607/Final Project/testing.csv")
```
~~~

WE FINALLY GOT SOMETHING……



~~~
---
title: "DATA 607 Final Project-Part II"
author: "Anthony Conrardy"
date: "`r Sys.Date()`"
output:
  pdf_document: default
  html_document: default
---

```{r setup, include=FALSE, warning=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(kableExtra)
library(readxl)
library(tidytext)
library(rvest)
library(readtext)
library(openxlsx)
library(ggplot2)
~~~

Introduction

In Part II of the this final project, we will import the sterilized c
that are located on the GitHub repository in the locations below. We
data sets into a unified set containing the required elements for fur
and analysis. We will rename the variables to be something easier tc
the variables in the data frame, and create a survey_total variable.
proposal was to analyze the student sentiment in the survey comments,
reviewing the data it became quickly apparent that the students did r
of the opportunity to provide reasonable commentary for us to underta
Instead, I chose to analyze the number of clinical experiences the st
against their scores on the program final examination. I also chose
average survey scores for each of the clinical rotations the students
the overall examination score. Therefore, my hypotheses for this ass
following:

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

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

```{r Data Import}
#  Pull in the data sets from the GitHub Repository
clinical_file <-
"https://raw.githubusercontent.com/Aconrard/DATA607/main/Final%20Proj
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%20Proj
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
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.Prece
         Hot_wash =
The.hot.wash.afforded.me.the.opportunity.to.discuss.the.patient.encounte
         Preceptors = Preceptors.responsiveness.to.clinical.questionsby.
         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
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 add
we want to use in the analysis. First, we will count the number of clin
each student participated in and then put that into the existing data fr
calculate the average survey rating of the clinical experience by each st
student was required to have at least ten (10) patient encounters. Some
able to get that on one clinical rotation, and for others took several.
we felt the average of the number of survey totals was the fairest way t
our analysis. We perform those calculations and put them into the data
other variables of interest. Finally, we extract the variables we will
analysis, which include anon_id(unique identifier), number_clinical (num
survey_average (average total of all surveys completed), and final exam
call that data frame "summary_df".

```{r tidy and transform}
# 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

# 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.
selected_data_final <- selected_data_final |>
```

```
  group_by(anon_id) |>
  mutate(survey_average = total_survey/number_clini
  ungroup()

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

Graphic Presentation

We provide three (3) plots to sense if there is ind
number of clinical rotations and/or the survey_aver
experiences appear to increase the possibility of a
evaluating the clinical experience seems to be not
examination. However, we will conduct a linear reg
mathematically comes out of the analysis.

```{r pressure, echo=FALSE}
# Create the ggplots
ggplot(summary_df, aes(x = number_clinical, y = final_exam)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Number of Clinicals", y = "Final Exam Score") +
  ggtitle("Number of Clinicals vs. Final Exam Score")

ggplot(summary_df, aes(x = factor(number_clinical), y = final_exam, fill =
factor(number_clinical))) +
  geom_boxplot() +
  labs(x = "Number of Clinicals", y = "Exam Score", fill = "Number of Clinicals") +
  ggtitle("Boxplot of Exam Score by Number of Clinicals") +
  theme_minimal()

ggplot(summary_df, aes(x = survey_average, y = final_exam))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Survey Average", y = "Final Exam Score") +
  ggtitle("Survey Average vs. Final Exam Score")

```

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:

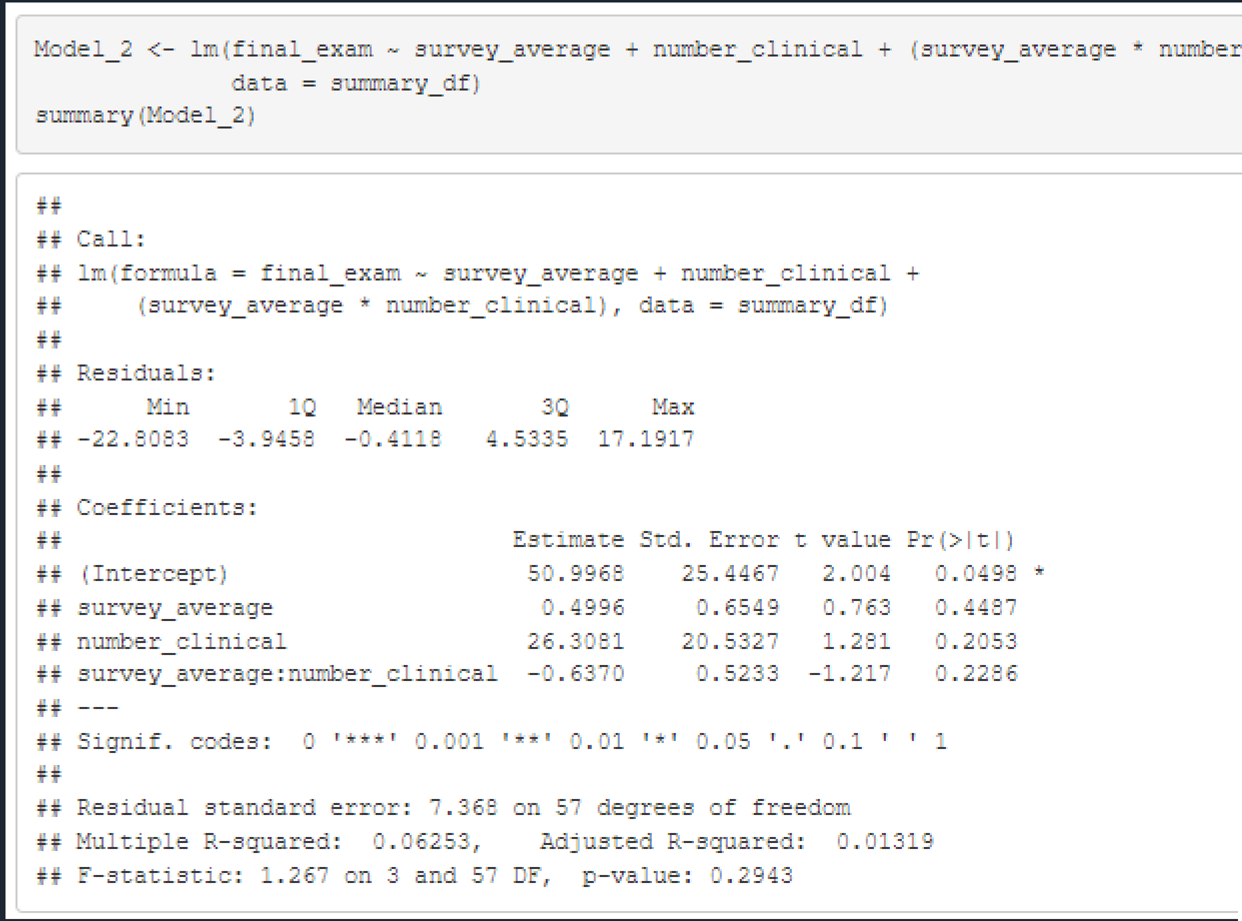
```{r regression}
Model_1 <- lm(final_exam ~ survey_average + number_clinical, data = summary_df)
summary(Model_1)

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

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

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

Discussion and Conclusion

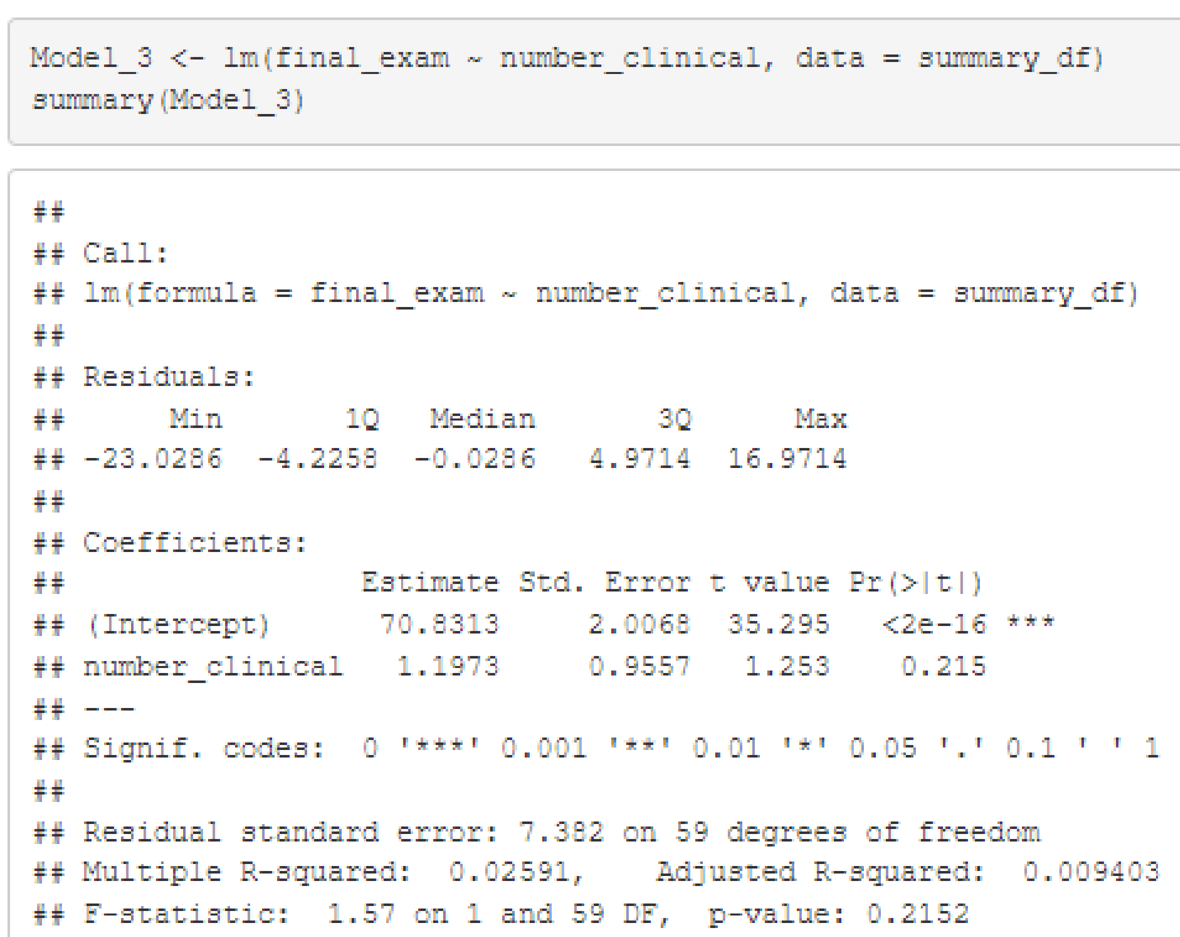
While there appears to be an upward trend associated with the number of clinical
experiences and performance on the examination, it does not appear to be statistically
significant. The survey average neither seems to be visually, nor statistically, related
to performance on the written examination. As 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.
```
```

BUT IT'S CRAP…….



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

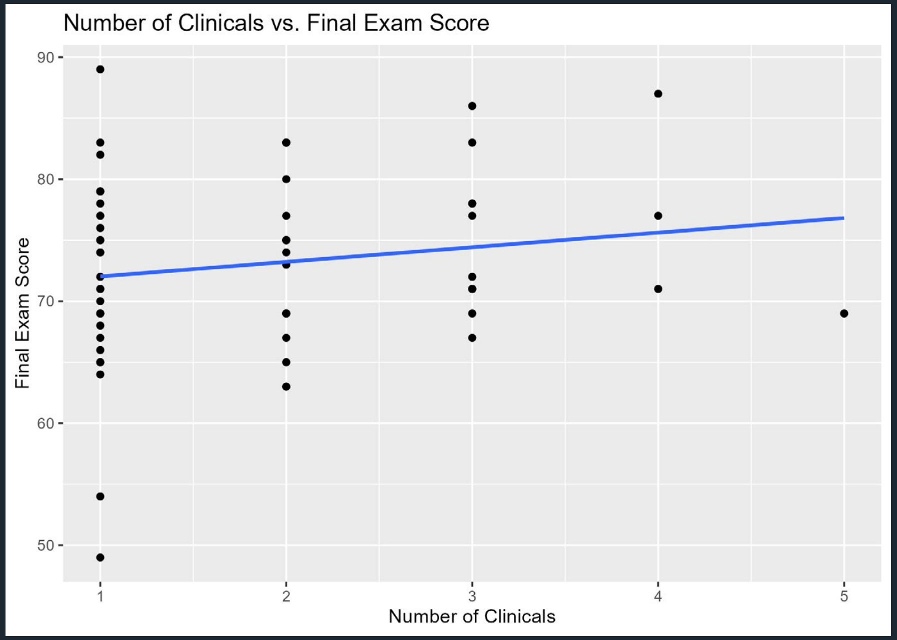
```
##
## Call:
## lm(formula = final_exam ~ survey_average + number_clinical +
##     (survey_average * number_clinical), data = summary_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8083  -3.9458  -0.4118   4.5335  17.1917
##
## Coefficients:
##                                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
```

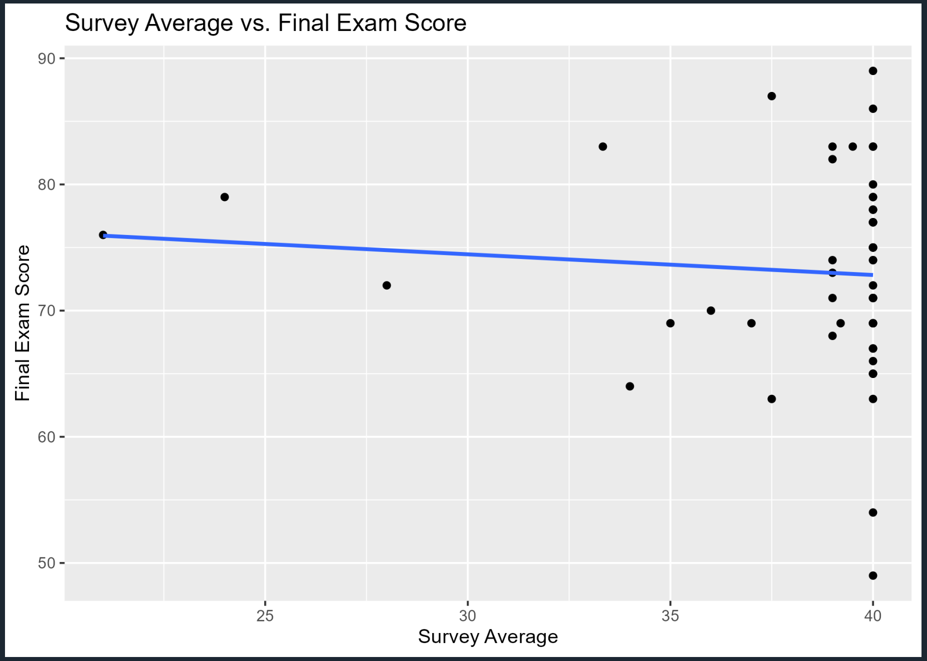


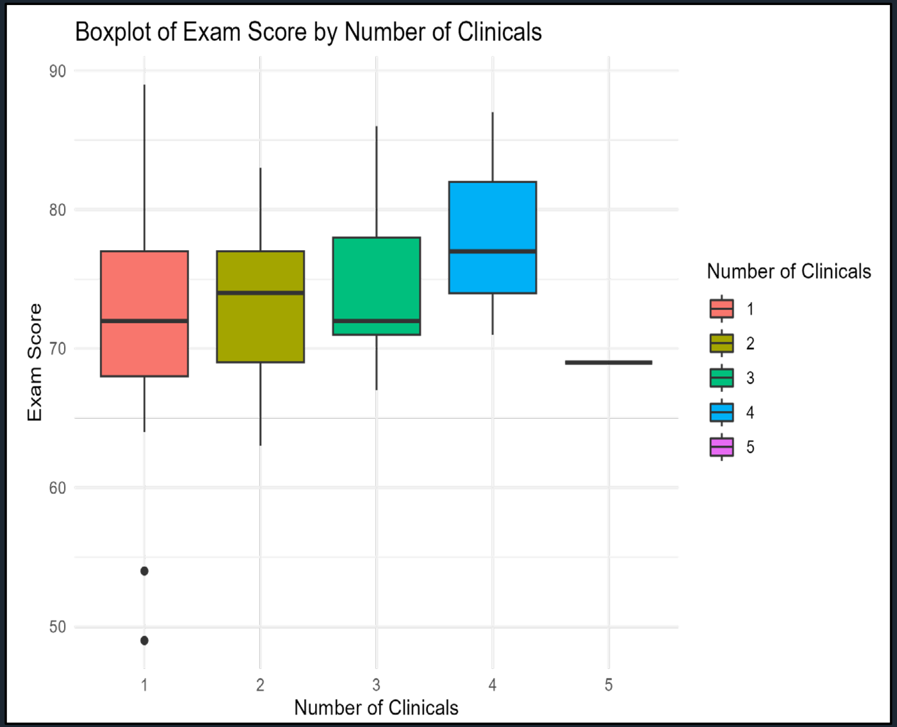
```
Model_3 <- lm(final_exam ~ number_clinical, data = summary_df)
summary(Model_3)
```

```
##
## 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
```

STILL MORE CRAP.....







WHAT ARE THE TAKEAWAYS…..

- Check your data and make sure you are looking at what you think your looking at. (rookie mistake) .
- This was a proof of concept to see if it could be done.
- Something was found, but we don’t know if it is something important.
- Data sources are messy and hard to manage.
- Summer is almost here…

THANK YOU

?