DATA 607 Final Project-Part I

Anthony Conrardy

2024-05-05

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 their site, and from a manual extraction of the testing data from the Platinum Educational Group 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 and 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 will 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, which 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 in this section below.

```
# 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
# 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 which 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 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 and then assign that identifier to the two imported data sets. Once done, we will remove the first and last names from both data sets and only keep the variables of interest assigned 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
# 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
```

```
## # A tibble: 1,379 x 12
##    anon_id Submission_ID       ‘Submission Date‘   Date
##      <int> <chr>               <dttm>              <dttm>
##  1      NA 5907591402068003805 2024-05-05 17:59:00 2024-05-05 00:00:00
## 2       NA 5906772441215069044 2024-05-04 19:14:04 2024-05-04 00:00:00
## 3       NA 5906764485969190836 2024-05-04 19:00:48 2024-05-04 00:00:00
## 4       NA 5906719430797533589 2024-05-04 17:45:43 2024-05-04 00:00:00
## 5       NA 5906711398539031245 2024-05-04 17:32:19 2024-05-04 00:00:00
## 6       NA 5906684657241122740 2024-05-04 16:47:45 2024-05-04 00:00:00
## 7       NA 5906017583159773844 2024-05-03 22:15:58 2024-05-03 00:00:00
## 8       NA 5905916316827384940 2024-05-03 19:27:11 2024-05-03 00:00:00
## 9       NA 5905902717181757927 2024-05-03 19:04:31 2024-05-03 00:00:00
## 10      NA 5905783144637126809 2024-05-03 15:45:14 2024-05-03 00:00:00
## # i 1,369 more rows
## # i 8 more variables: ‘Appropriate Orientation by your Preceptor:‘ <dbl>,
## #   ‘Responsibilities clearly defined by your Preceptor:‘ <dbl>,
## #   ‘Adequate Supervision on Ambulance:‘ <dbl>,
## #   ‘Preceptors responsiveness to clinical questionsby student:‘ <dbl>,
## #   ‘The hot wash afforded me the opportunity to discuss the patient encounter:‘ <dbl>,
## #   ‘Incorporated as member of crew:‘ <dbl>, ...
```

```
testing_export
```

```
## # A tibble: 134 x 3
##    anon_id final_exam final_exam_retest
##      <int> <chr>      <chr>
## 1        3 63         NA
## 2       NA 66         NA
## 3       NA 79         NA
## 4       NA 64         NA
## 5       NA 81         NA
## 6       55 80         NA
## 7       NA 65         NA
## 8       74 67         NA
## 9       15 86         NA
## 10      16 87         NA
## # i 124 more rows
```

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.

```
# 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")
```