```
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 be 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 ))
pdf_2_df$Cardiology <- substr(pdf_2_df$Cardiology, nchar(pdf_2_df$Cardiology) - 2,
nchar(pdf_2_df$Cardiology))
pdf_2_df$Medical <- substr(pdf_2_df$Medical, nchar(pdf_2_df$Medical) - 2,
nchar(pdf_2_df$Medical))
pdf_2_df$Trauma <- substr(pdf_2_df$Trauma, nchar(pdf_2_df$Trauma) - 2,
nchar(pdf_2_df$Trauma))
pdf_2_df$'OB-Peds' <- substr(pdf_2_df$'OB-Peds', nchar(pdf_2_df$'OB-Peds') - 2,
nchar(pdf_2_df$'OB-Peds'))
pdf_2_df$Operations <- substr(pdf_2_df$Operations, nchar(pdf_2_df$Operations) - 2,
nchar(pdf_2_df$Operations))
# Remove all the % signs
pdf_2_df[] <- lapply(pdf_2_df, function(k) gsub("%","",k))

# Working on the first 80 entries.
pdf_2a <- read_lines(pdf_exam_2)
# Remove lines that are empty
pdf_2a_clean <- pdf_2[grep("\\S", pdf_2a)]
# Drop the first two rows and all the rows after 80
pdf_2a_clean <- pdf_2a_clean[3:80]
# Drop all character strings that have percentage (%) in it.
pdf_2a_clean <- pdf_2a_clean[!grepl("/", pdf_2a_clean)]
# Now combine every two character strings to create an individual record
pdf_2a_combined <- c()
for (i in seq(1, length(pdf_2a_clean), by = 2)){
  pdf_2a_combined <- c(pdf_2a_combined, paste(pdf_2a_clean[i], pdf_2a_clean[i + 1], sep =
"  "))
}
# Split the vector by delimited whitespace of 2
split_vector1 <- strsplit(pdf_2a_combined, "\\s{2,}", perl = TRUE)
# Create Data Frame
pdf_2a_df <- as.data.frame(do.call(rbind, split_vector1))
# Rename the columns before splitting
colnames(pdf_2a_df)[colnames(pdf_2a_df) == "V2"] <- "V3"
# Split the last character of the name to identify exam attempt and then delete it from
the string
pdf_2a_df$V2 <- substr(pdf_2a_df$V1, nchar(pdf_2a_df$V1), nchar(pdf_2a_df$V1))
pdf_2a_df$V1 <- sub(".$","", pdf_2a_df$V1)
# Move the columns before the next split
pdf_2a_df <- pdf_2a_df |> select(V1, V2, V3)
# Split V3 into seven (7) columns
pdf_2a_df_split <- separate(pdf_2a_df, V3, into = c("V3", "V4", "V5", "V6", "V7", "V8",
"V9"))
# Place column names
colnames(pdf_2a_df_split) <- c(pdf2_header)

# Combine the two dataframes for one consolidated set
Comp_Exam_2_Combined <- rbind(pdf_2_df, pdf_2a_df_split)

# First Set of Exam Scores-Major Examination 4

# Pull the PDF file that was previously exported as TXT file
pdf_exam_4 <- "D:/Documents/R_Working_Directory/DATA 607 Final Project/EMT Readiness Exam
4 results.txt"
# Read the file into R
pdf_4 <- read_lines(pdf_exam_4)
# Remove lines that are empty
pdf_4_clean <- pdf_4[grep("\\S", pdf_4)]
# Take all the lines after 78, as the first 78 need to be tidy separately
pdf_4_clean <- pdf_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))
# Trim the whitespace from some of the values
pdf_4_df[] <- lapply(pdf_4_df, trimws)
# Drop the last row as having calculated sums which can be added later if necessary
pdf_4_df <- pdf_4_df[1:(nrow(pdf_4_df) - 1), ]
# Place the column names
colnames(pdf_4_df) <- c("Name", "Attempt", "Airway", "Cardiology", "Medical",
"Obstetrics", "Pediatrics", "Trauma", "Operations", "Total")
# Only keep the last three characters of each value in a column
pdf_4_df$Airway <- substr(pdf_4_df$Airway, nchar(pdf_4_df$Airway) - 2,
nchar(pdf_4_df$Airway ))
pdf_4_df$Cardiology <- substr(pdf_4_df$Cardiology, nchar(pdf_4_df$Cardiology) - 2,
nchar(pdf_4_df$Cardiology))
pdf_4_df$Medical <- substr(pdf_4_df$Medical, nchar(pdf_4_df$Medical) - 2,
nchar(pdf_4_df$Medical))
pdf_4_df$Obstetrics <- substr(pdf_4_df$Obstetrics, nchar(pdf_4_df$Obstetrics) - 2,
nchar(pdf_4_df$Obstetrics))
pdf_4_df$Pediatrics <- substr(pdf_4_df$Pediatrics, nchar(pdf_4_df$Pediatrics) - 2,
nchar(pdf_4_df$Pediatrics))
pdf_4_df$Trauma <- substr(pdf_4_df$Trauma, nchar(pdf_4_df$Trauma) - 2,
nchar(pdf_4_df$Trauma))
pdf_4_df$Operations <- substr(pdf_4_df$Operations, nchar(pdf_4_df$Operations) - 2,
nchar(pdf_4_df$Operations))
pdf_4_df$Total <- substr(pdf_4_df$Total, nchar(pdf_4_df$Total) - 2, nchar(pdf_4_df$Total))

# Remove all the % signs
pdf_4_df[] <- lapply(pdf_4_df, function(k) gsub("%","",k))

# Working on the first 78 lines.
pdf_4a <- read_lines(pdf_exam_4)
# Lines appear to be complicated to parse at this time.  It only includes a total of 24
students.
# Will work on this if we have time, but the data obtained is sufficient for out purposes.
Comp_Exam_4_Combined <- pdf_4_df

Exam_2_sep <- Comp_Exam_2_Combined |> separate(Name, into = c("Last Name", "First Name"),
sep = ",") |>
  select("Last Name", "First Name", Attempt, Total)
Exam_4_sep <- Comp_Exam_4_Combined |> separate(Name, into = c("Last Name", "First Name"),
sep = ",") |>
  select("Last Name", "First Name", Attempt, Total)
Exam_Results <- rbind(Exam_2_sep, Exam_4_sep)
``
#Drop all names to lower case for matching
Exam_Results$`Last Name` <- tolower(Exam_Results$`Last Name`)
Exam_Results$`First Name` <- tolower(Exam_Results$`First Name`)
df$`Last Name` <- tolower(df$`Last Name`)
df$`First Name` <- tolower(df$`First Name`)

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

url_pdf <- "D:/Documents/R_Working_Directory/DATA 607 Final Project/EMT Readiness Exam 2
results.pdf"

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