**Data Wrangling Summary**

The most important steps taken to clean up the data were performed in R on the data frame (enroll\_original) to transform it into a format that can be analyzed. The techniques used, but were not limited to were: Installing library packages, analyzing the dataset after each major manipulation, added new columns, removed existing columns, replaced missing/blank values with numeric or string values, replaced existing values with new values, changed column names, added data functions (i.e. calculated the mean for each row), and exported to .csv file format.

**Installed Necessary Library Packages and Analyzed the Dataset:**

Before anything could happen in R studio, several library packages needed to be installed before and throughout the cleaning process as well. Library packages such as: dplyr, base, plyr, tidyverse, data.table were installed. Once all packages were updated/installed, R calls such as: “view”, “colnames”, and “head” were used to determine if the dataset was missing values, needed additional columns, name changes, columns removed, variables assigned for easy coding, etc.

**Add or Removed Columns/ Changed Column Names:**

The original data frame included 32 columns. Some of the column variables were not necessary for the purposes of this interaction. Three columns (“PIDM\_KEY”, “ENROLLMENT\_AR\_IND”, and “ETHN\_DESC”) were removed, while three columns were added (“LEVEL\_IND”, “STATUS”, and “ETHNIICTY”). Two columns were renamed for ease of coding (“MAJR\_DESC\_CONCENTRATION\_1” to “CONCENTRATION" and "FULL\_PART\_TIME\_IND"="STATUS"). When the data was imported into R Studio, some variables names were too long, and others did not accurately describe the particular data. Therefore, the names of the columns were changed to shorter, concise names that accurately describe the data in the column. Some of the columns needed partnering variables (i.e. "STATUS" and “STATUS\_IND”) to transition easily into the interaction, if needed.

**Added Values to Missing/Blank or Existing Values:**

One column in particular (“ETHN\_DESC”) was missing values. Those rows with missing values needed to be transformed to be included with the value “Other”. A new column (“ETHNICITY”) was created and the data in the existing column (“ETHN\_DESC”) was copied into this new column. Then a IF-ELSE statement was used to record those rows with the blank/missing value to “Other”.

**Data Function:**

The MEAN function was used to capture the average of four columns (“ROUND\_1”, “ROUND\_2”, “ROUND\_3”, and “ROUND\_4”) for each row. A new column (“PERCENTAGE”) was created in order to capture the row average.

**Export:**

The original file was in .csv format and exported to the local drive in a .csv format.

**Data Wrangling Code**

getwd()

setwd("C:/Users/tkinsey/Desktop/Updates")

read.csv(file = "enroll\_original.csv", header = FALSE)

enroll\_original <- read.csv(file = "enroll\_original.csv")

View(enroll\_original)

enroll <- enroll\_original

colnames(enroll)

library (dplyr)

library (base)

library (plyr)

library(tidyverse)

enroll$ETHN\_DESC

enroll["ETHNICITY"]<- NA

enroll$ETHNICITY <- enroll$ETHN\_DESC

colnames(enroll)

enroll$ETHNICITY

enroll$ETHNICITY <- ifelse(enroll$ETHNICITY=="Black or African American", "Black or African American",

ifelse(enroll$ETHNICITY=="American Indian/Alaskan Native","American Indian/Alaskan Native",

ifelse(enroll$ETHNICITY=="Asian or Pacific Islander","Asian or Pacific Islander",

ifelse(enroll$ETHNICITY=="Hispanic","Hispanic",

ifelse(enroll$ETHNICITY=="White/Non-Hispanic","White/Non-Hispanic",

ifelse(enroll$ETHNICITY=="Other","Other","Other"))))))

enroll$ETHNICITY

colnames(enroll)

enroll <- enroll[ -c(1,8,13) ]

colnames(enroll)

enroll$LEVL\_DESC

enroll["LEVEL\_IND"] <- NA

enroll$LEVEL\_IND <- enroll$LEVL\_DESC

colnames(enroll)

enroll$LEVEL\_IND

head (LEVEL\_IND)

enroll$LEVEL\_IND <- ifelse(enroll$LEVEL\_IND=="Undergraduate", 1,

ifelse(enroll$LEVL\_DESC=="Graduate Masters",0, NA))

enroll$LEVEL\_IND

head (enroll)

colnames(enroll)

enroll$FULL\_PART\_TIME\_IND

enroll["STATUS\_IND"] <- NA

enroll$STATUS\_IND <- enroll$FULL\_PART\_TIME\_IND

colnames(enroll)

enroll$STATUS\_IND

enroll$STATUS\_IND <- ifelse(enroll$STATUS\_IND=="FT", 1,

ifelse(enroll$FULL\_PART\_TIME\_IND=="PT",0, NA))

enroll$STATUS\_IND

head (enroll)

library(data.table)

names(enroll)

rename(enroll, c("MAJR\_DESC\_CONCENTRATION\_1"="CONCENTRATION", "FULL\_PART\_TIME\_IND"="STATUS"))

enroll["PERCENTAGE"] <- NA

enroll$PERCENTAGE <- apply(enroll[,2:5],1, mean)

colnames(enroll)

enroll$PERCENTAGE

write.csv(enroll, file = "enroll\_capstone.csv")