Phase One: Exploratory Analysis

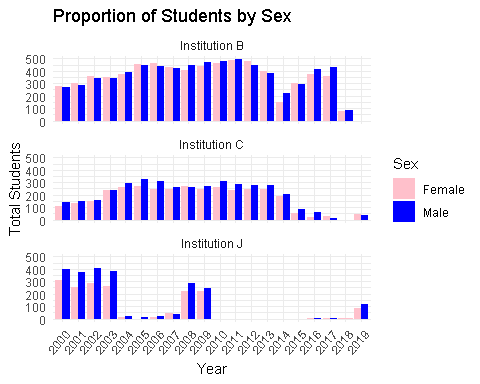
Mohamed Elmanzalawi

## Introduction

### The aim of this project is to examine higher education data from institutions within the MidField (Multiple-Institution Database for Investigating Engineering Longitudinal Development) database. We explore student demographics through descriptive summary statistics and visualizations. Additionally, we train and evaluate predictive models guided by educational domain insights. Collectively, this analysis aims to understand the predictors/ features that are most significant to a student’s graduation from a four-year bachelor’s degree program.

## Student Demographics

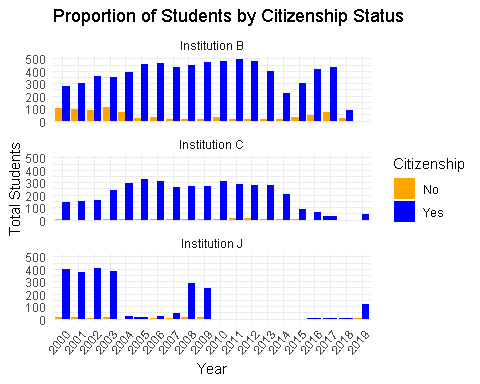
### In this section, we aim to explore the student demographics in the MidField Data. In broad strokes, who are the students in these data, and how have they changed over time?



### This graph displays the graduation volume at the three MidField institutions over the past 20 years. Throughout this time, Institution B has seen a steady influx of graduates from both sexes. Contrary to Institution B, Institution C has seen a steady decline in graduates from both sexes. Institution J has experienced sharp declines in students overall.

### 

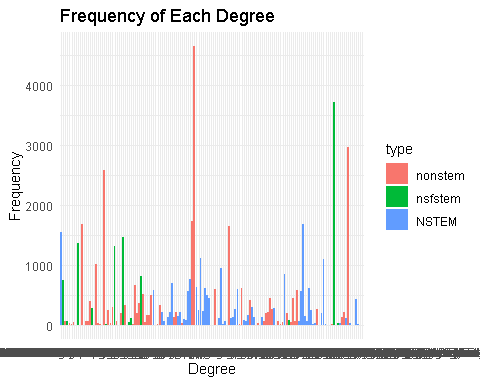
### Similar to the previous graph, this Bar Chart displays the age categories of graduates at the three Institutions. While the shape of this graph is similar, there is a striking difference in the proportion of students. The overwhelming majority of students at all institutions are under 25. This underrepresentation could be a potential area of bias in this sample.



### This graph shows the Citizenship Status of graduates at the three Institutions. Once again, the distribution of this Bar Chart is similar to the previous one. This discrepancy suggests that international students are underrepresented in this sample, hinting toward another potential bias in the data.

## Education & Degrees

**In this section, we want to identify the trends in continued education. By examining the majors which the most students earn their degrees in, we can infer the more sought-after fields.**



### Based on the frequency table above, we can see that some of the most popular majors are non-Stem fields such as business administration, Journalism, and Psychology. These findings offer an intriguing insight into the most sought-after fields.

## Conclusion

### Understanding student demographics is a crucial aspect of this analysis. We will continue to build off these findings in future phases.

# Code Appendix

# Setting Document Options  
knitr::opts\_chunk$set(  
 echo = FALSE,  
 warning = FALSE,  
 message = FALSE,  
 fig.align = "center",  
 cache = FALSE   
)  
# Imports  
#install.packages("tidyverse")  
#install.packages("plotly")  
install.packages("midfielddata",  
 repos = "https://MIDFIELDR.github.io/drat/",  
 type = "source"  
)  
library(dplyr)  
library(ggplot2)  
library(plotly)  
# Load the Data  
data(student, course, package = "midfielddata")  
data(term, degree, package = "midfielddata")  
# Data Processing  
# Select relevant columns from student and degree tables  
student1 <- student %>%  
 select("mcid", "institution", "race", "sex", "age\_desc", "us\_citizen")  
  
degree1 <- degree %>%  
 select("mcid", "term\_degree")  
  
# Combine the Tables   
demographics <- right\_join(student1, degree1, by = "mcid")  
  
# Convert to integer  
demographics$term\_degree <- as.integer(demographics$term\_degree)  
  
# Drop Last number of each term\_degree  
demographics$term\_degree <- substr(as.character(demographics$term\_degree), 1, nchar(as.character(demographics$term\_degree)) - 1)  
# Creating Glyph Ready Data   
  
# Glyph ready table for a Bar chart  
demo\_glyph <- demographics %>%  
 group\_by(institution, sex, term\_degree, age\_desc, race, us\_citizen) %>%  
 summarise(total = n()) %>%  
 group\_by(institution, term\_degree) %>%  
 mutate(Proportion = total / sum(total))  
  
# Explore the more recent subset of years  
demo\_glyph1 <- demo\_glyph %>%  
 filter(term\_degree >= 2000 & term\_degree <= 2020)  
# Proportion of Students by Sex  
ggplot(demo\_glyph1, aes(x = term\_degree, y = total, fill = sex)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~institution, ncol = 1) +   
 labs(x = "Year", y = "Total Students", fill = "Sex") +  
 scale\_fill\_manual(values = c("pink","blue")) +   
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 50, hjust = 1)) +  
 ggtitle("Proportion of Students by Sex")  
  
# Proportion of Students by Age  
ggplot(demo\_glyph1, aes(x = term\_degree, y = total, fill = age\_desc)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~institution, ncol = 1) +   
 labs(x = "Year", y = "Total Students", fill = "Age Description") +  
 scale\_fill\_manual(values = c("orange","blue")) +   
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 50, hjust = 1)) +  
 ggtitle("Proportion of Students by Age")  
  
# Proportion of Students by Citizenship Status  
ggplot(demo\_glyph1, aes(x = term\_degree, y = total, fill = us\_citizen)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~institution, ncol = 1) +   
 labs(x = "Year", y = "Total Students", fill = "Citizenship") +  
 scale\_fill\_manual(values = c("orange","blue")) +   
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 50, hjust = 1)) +  
 ggtitle("Proportion of Students by Citizenship Status")  
  
# Data Processing 2  
cip <- read.csv("cipCodings.csv")  
  
# Select relevant columns from degree & cip tables  
cip2 <- cip %>%  
 select("cip6", "category")  
  
degree2 <- degree %>%  
 select("mcid", "term\_degree", "cip6", "degree")  
degree2$cip6 <- as.integer(degree2$cip6)  
  
majors <- right\_join(cip2, degree2, by = "cip6")  
  
# Drop rows with missing category value   
majors <- majors %>%  
 filter(!is.na(majors$category))  
# Creating Glyph Ready Data   
degree\_freq <- majors %>%  
 group\_by(cip6, degree) %>%  
 summarise(type = first(category), frequency = n())  
  
ggplot(data = degree\_freq, aes(x = degree, y = frequency, fill = type)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Frequency of Each Degree",  
 x = "Degree",  
 y = "Frequency") +  
 theme\_minimal()