# Intermediate Progress

Khanh Do, Joyce Gill, Matthew Billings-Chiu

2025-05-03

## Introduction

This project explores higher education data from U.S. colleges to uncover trends related to **admissions selectivity, student body diversity, and institutional outcomes** such as graduation rates. The goal is to build an interactive Shiny app to help users — such as students and parents compare colleges based on customizable filters like state, admission rates, SAT scores, demographics, and more.

Data Source - Most Recent Institution-Level Data

## Data Cleaning

```
# Load data
data <- read.csv("https://raw.githubusercontent.com/khanhdo05/stats-final-230/refs/heads/main/Most-Recer
```

### Feature 1 Joyce

Users will be able to select an institution and view its racial composition through a Plotly pie chart, using variables like `UGDS_WHITE`, `UGDS_BLACK`, `UGDS_HISP`, `UGDS_ASIAN`, and more. The chart will display raw percentages when hovering over each slice, giving a quick and clear breakdown of the student body.

```
# Data cleaning
feature1data <- data %>%
  dplyr::select(INSTNM, UGDS_WHITE, UGDS_BLACK, UGDS_HISP, UGDS_ASIAN, UGDS_AIAN, UGDS_NHPI, UGDS_2MOR,
  filter(!(is.na(UGDS_WHITE) & is.na(UGDS_BLACK) & is.na(UGDS_HISP) & is.na(UGDS_ASIAN) & is.na(UGDS_AI

# Write to csv file
write_csv(feature1data, "cleaned_feature1data.csv")

# Sketch Visualization
selected_inst <- "Grinnell College"

selected_data <- feature1data %>%
  filter(INSTNM == selected_inst)

race_labels <- c("White", "Black", "Hispanic", "Asian",
                 "American Indian/Alaska Native", "Native Hawaiian/Pacific Islander",
                 "Two or More", "Non-Resident Alien", "Unknown")
```

```r
race_columns <- c("UGDS_WHITE", "UGDS_BLACK", "UGDS_HISP", "UGDS_ASIAN",
                  "UGDS_AIAN", "UGDS_NHPI", "UGDS_2MOR", "UGDS_NRA", "UGDS_UNKN")

race_values <- as.numeric(selected_data[1, race_columns])

plot_ly(
  labels = race_labels,
  values = race_values,
  type = "pie"
) %>%
  layout(title = paste("Racial Composition of", selected_inst))
```

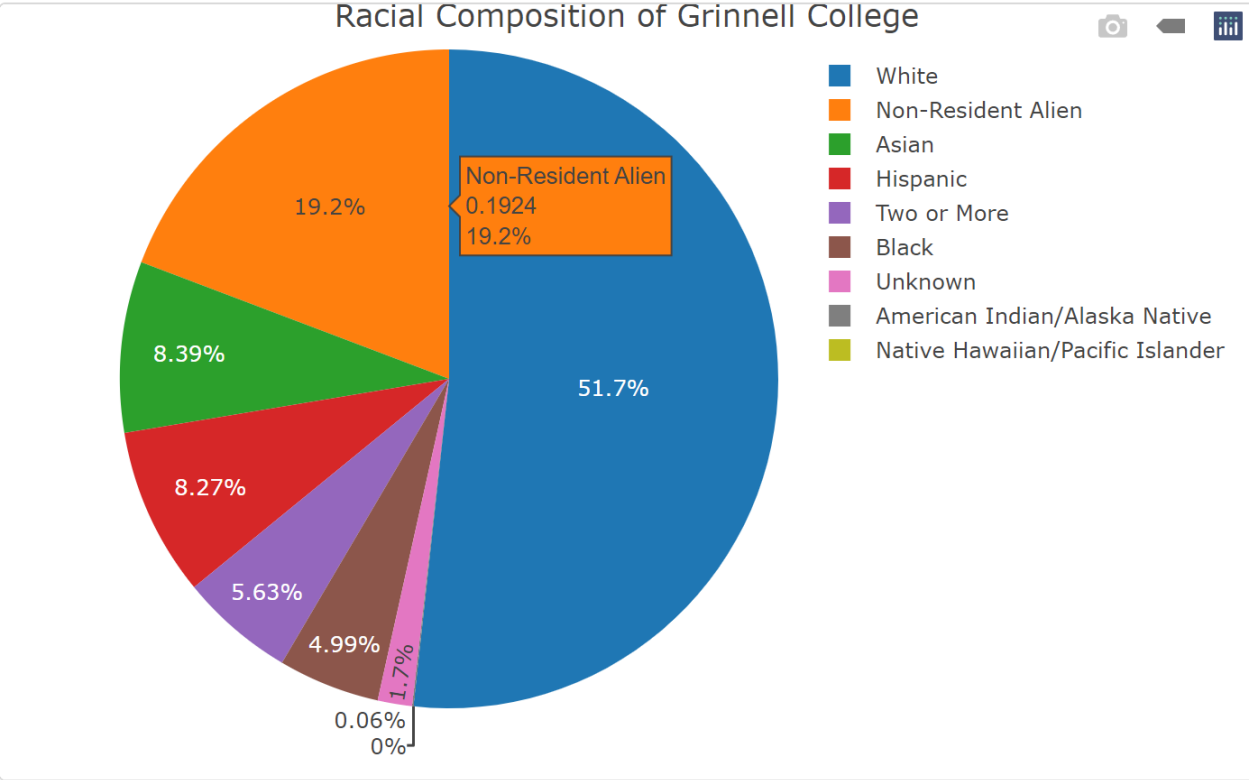



Figure 1: Interactive pie chart

## Feature 2

The app will feature sliders for SAT average and admission rate (in ranges), along with filters for state and institution `CONTROL` type (Public = 1, Private = 2 or 3), Locale (by rural & town, suburb & city) . Based on the user's selections, the app will return a list of colleges that meet the criteria in the form of a scatter plot with x-axis is the SAT average and y-axis is the Admission rate, where on hover the point, it shows the information, on click, user can click on the school website link. User will also have the option to tick whether they want to include criteria SAT, Admission rate. For Locale and IS_PUBLIC, they can also not tick any box or all boxes. The sliders for SAT and Admission rate would have a start and end pointer to show the range in which they want to filter for.

```r
# Data cleaning
feature2data <- data %>%
  # Select only relevant variables
  dplyr::select(INSTNM, INSTURL, CONTROL, SAT_AVG, ADM_RATE, LOCALE) %>%
  # Drop rows that have NA values for all SAT_AVG, ADM_RATE, and LOCALE
  filter(!(is.na(SAT_AVG) & is.na(ADM_RATE) & is.na(LOCALE))) %>%
  # Binary variable for colleges that are public and is in city or suburb
  mutate(
    IS_PUBLIC = case_when(
      CONTROL == 1 ~ 1,
      CONTROL %in% c(2, 3) ~ 0,
      TRUE ~ NA_real_
    ),
    IS_CITY = case_when(
      LOCALE >= 11 & LOCALE <= 23 ~ 1,
      LOCALE >= 31 & LOCALE <= 43 ~ 0,
      TRUE ~ NA_real_
    )) %>%
  # Remove column that is no longer needed
  select(!c("CONTROL", "LOCALE"))

# Writing to csv file
write_csv(feature2data, "cleaned_feature2data.csv")

df <- feature2data %>%
  mutate(
    SAT_AVG = as.numeric(SAT_AVG),
    ADM_RATE = as.numeric(ADM_RATE)
  ) %>%
  filter(!is.na(SAT_AVG), !is.na(ADM_RATE))

# Sketch visualization
plot_ly(
  data = df,
  x = ~SAT_AVG,
  y = ~ADM_RATE,
  type = 'scatter',
  mode = 'markers',
  text = ~paste(
    "School:", INSTNM,
    "<br>SAT:", SAT_AVG,
    "<br>Adm Rate:", round(ADM_RATE, 3)
  ),
  hoverinfo = 'text',
  marker = list(color = 'rgba(0, 102, 204, 0.6)', size = 10)
) %>%
layout(
  title = "Colleges: SAT vs Admission Rate",
  xaxis = list(title = "Average SAT Score"),
  yaxis = list(title = "Admission Rate"),
  annotations = list(
    x = 0.5, y = 1,  # Positioning the subtitle
    text = "Hover over and click on a point to see details about the school.",
```

```
    showarrow = FALSE,
    font = list(size = 14),
    align = 'center',
    xref = 'paper', yref = 'paper'
  )
)
```

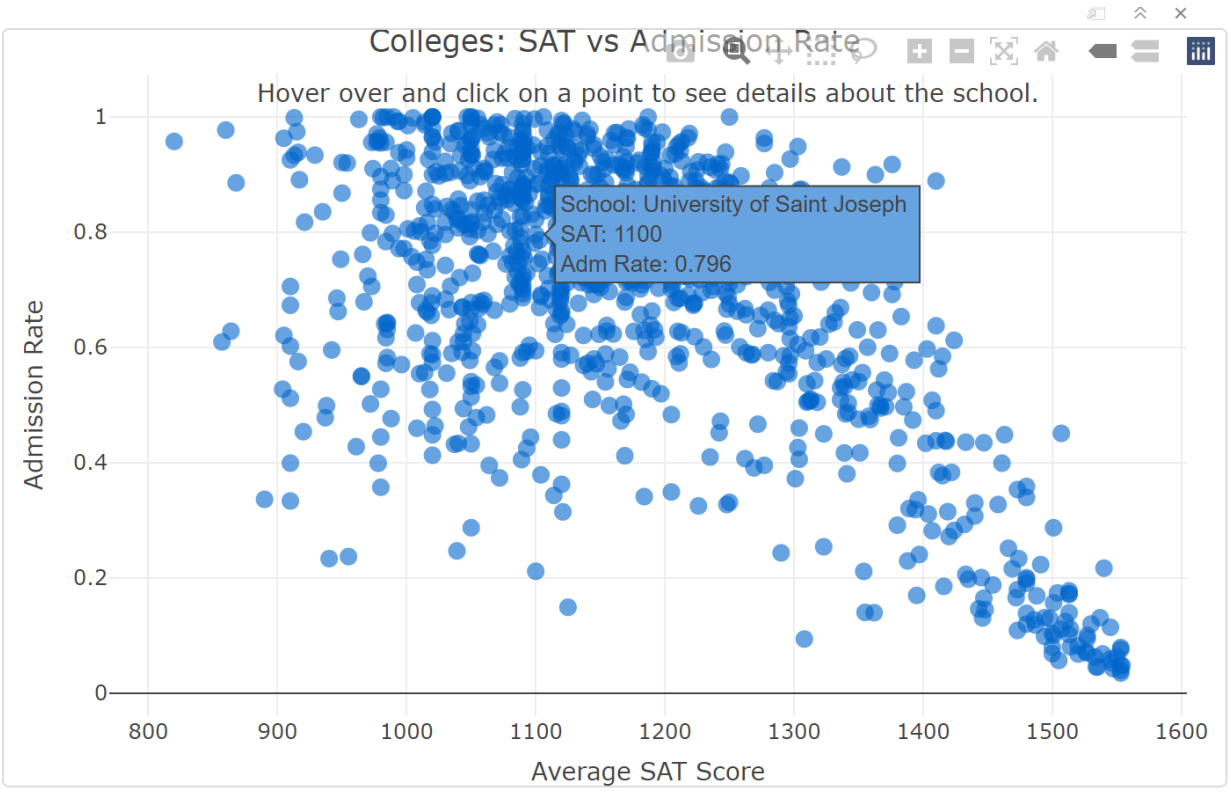



Figure 2: Interactive scatter plot

## Feature 3

The app will include a plot that allows users to filter colleges based on median student debt, median earnings, undergraduate enrollment, admission rate, and ACT average (calculated from subject scores using data manipulation). After selecting a college of interest, the app will use clustering techniques to identify and visualize similar institutions, helping users explore comparable schools based on these key financial and academic attributes.

```
# Data cleaning
feature3data <- data %>%
  mutate(ACT_MEDIAN = ACTWRMID + ACTMTMID) %>%
  select(INSTNM, UGDS, ADM_RATE, ACT_MEDIAN, GRAD_DEBT_MDN) %>%
  mutate(GRAD_DEBT_MDN = na_if(GRAD_DEBT_MDN, "PrivacySuppressed")) %>%
  mutate(GRAD_DEBT_MDN = as.numeric(GRAD_DEBT_MDN)) %>%
  drop_na()
```

```
## Warning: There was 1 warning in `mutate()`.
```

```
## i In argument: `GRAD_DEBT_MDN = as.numeric(GRAD_DEBT_MDN)`.
## Caused by warning:
## ! NAs introduced by coercion
```

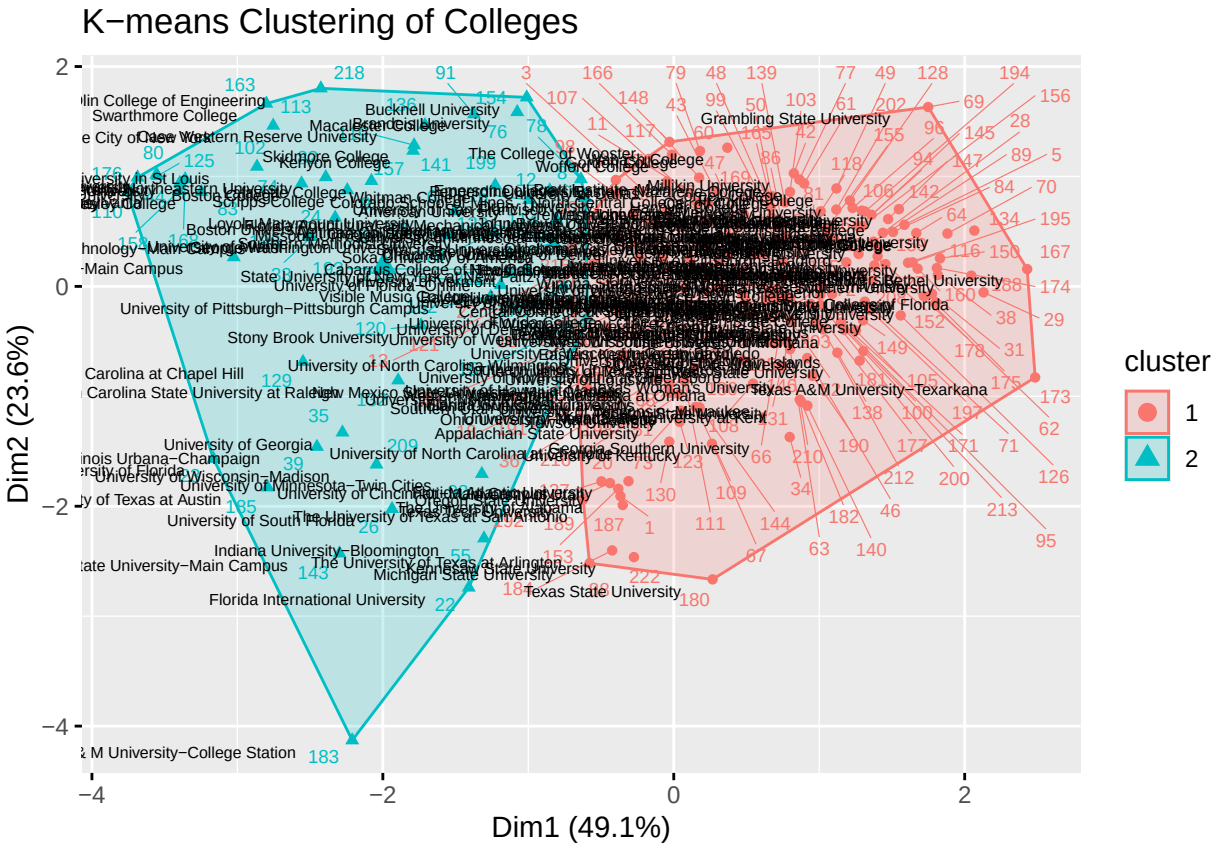
```r
# Write to csv file
write_csv(feature3data, "cleaned_feature3data.csv")

# Scale features (exclude school name)
features_scaled <- feature3data %>%
  select(-INSTNM) %>%
  scale()

# K-means clustering
k2 <- kmeans(features_scaled, centers = 2, nstart = 25)

# Extract numeric features and scale
school_names <- feature3data$INSTNM
features_only <- feature3data %>% select(-INSTNM)
scaled_features <- scale(features_only)

# Sketch visualization no.1
# Obviously, there is still modifications to come to make this easier to read
fviz_cluster(k2, data = scaled_features,
             labelsize = 7,
             main = "K-means Clustering of Colleges",
             repel = TRUE) +
  geom_text(aes(label = school_names), size = 2, vjust = 1.5, hjust = 1.2)
```




```
# PCA for 2D visualization
pca <- prcomp(features_scaled)
pca_data <- as.data.frame(pca$x[, 1:2]) # PC1 and PC2
pca_data$School <- feature3data$INSTNM
pca_data$Cluster <- as.factor(k2$cluster)

# Sketch visualization no.2
plot_ly(pca_data,
        x = ~PC1,
        y = ~PC2,
        type = 'scatter',
        mode = 'markers',
        color = ~Cluster,
        text = ~paste("School:", School,
                      "<br>Cluster:", Cluster),
        hoverinfo = 'text') %>%
  layout(
    title = "Interactive K-means Cluster Plot (PCA)",
    xaxis = list(title = "Principal Component 1"),
    yaxis = list(title = "Principal Component 2"),
    annotations = list(
      x = 0.5, y = 1,  # Positioning the subtitle
      text = "Hover over on a point to see which school it is.",
      showarrow = FALSE,
      font = list(size = 14),
      align = 'center',
```

```
      xref = 'paper', yref = 'paper'
  ))
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette
```

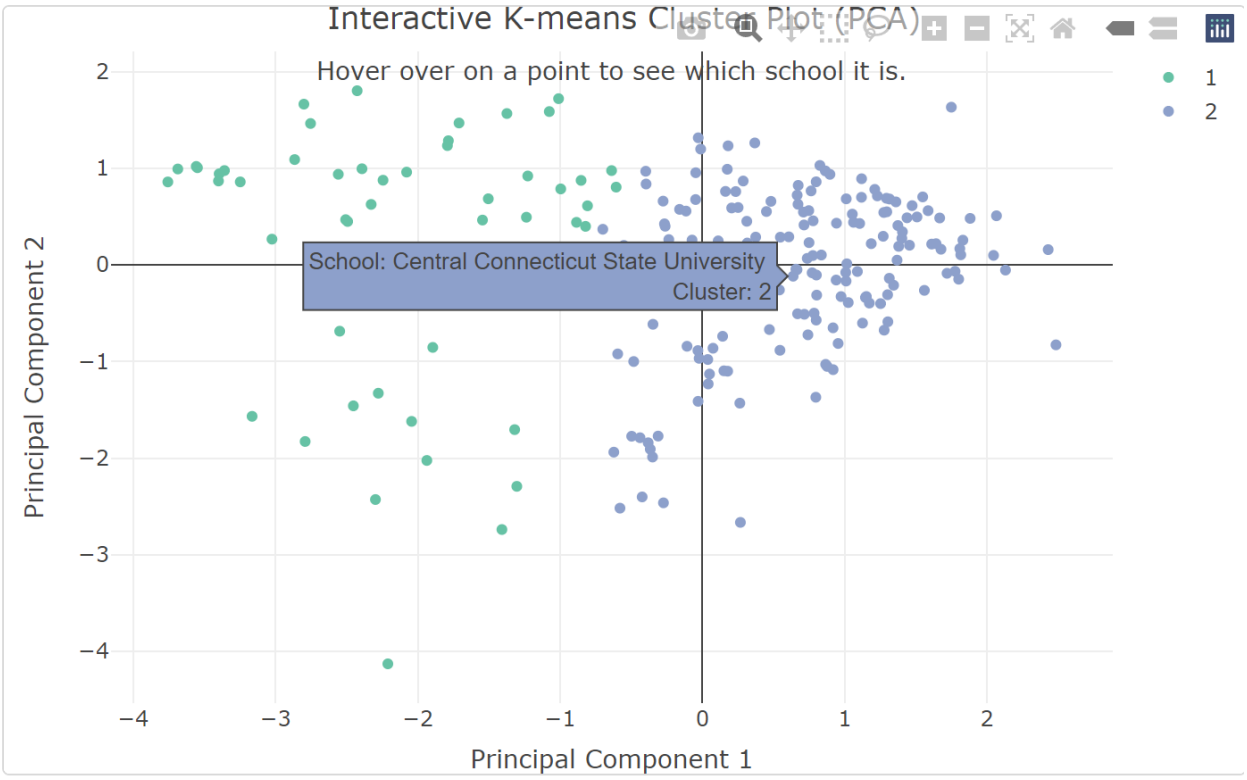



Figure 3: Interactive cluster scatter