DataMungingProject

Emma Horton, Jin Burgess, Hanna Butsko

2023-10-23

About

Our project aims to conduct a survey of the U.S. education system for the academic year 2021-2022. The goal is to provide transparency and empower students and families to make informed decisions about college choices. This project focuses on analyzing educational institution data, including factors like the number of institutions, predominant degree, demographic variables, Carnegie classification, and financial aspects.

Data Source

The data is sourced from the College Scorecard Project initiated by the U.S. Department of Education.

Content

1. **Data Preprocessing:**
    - Data preprocessing involves wrangling the data to filter and clean it for more precise analysis.
2. **Data Visualization:**
    - We will use bar plots, grouped bar plots, ridge plots, and heatmaps to visualize the data for better understanding.
3. **Analysis and Interpretation:**
    - We will provide analysis and interpretations for each visualization, discussing patterns, trends, and potential factors influencing the data.

Libraries

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(readr)
library(ggridges)
library(ggplotify)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(forcats)
library(viridis)
```

```
## Loading required package: viridisLite
```

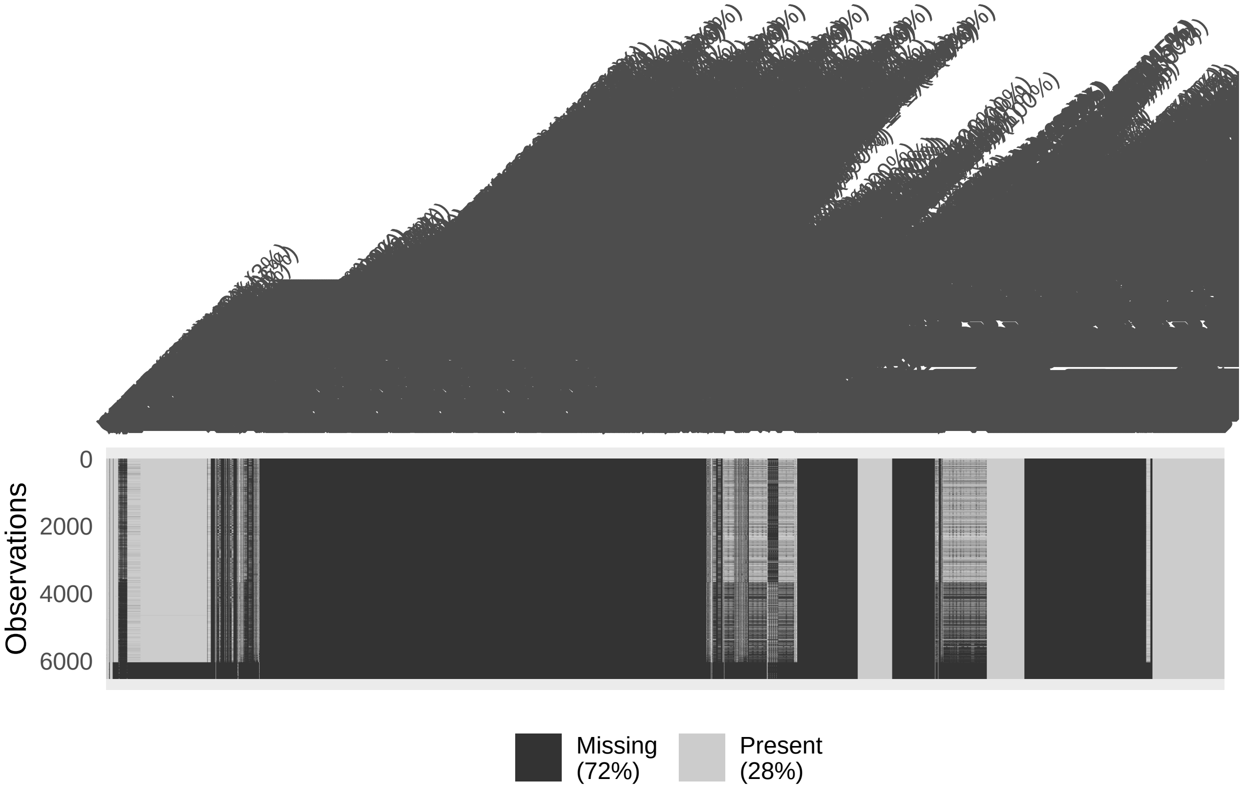
```
library(visdat)
```

Set Working Directory and Read Data

```
setwd('/Users/emmahorton/DataScience/DataMunging/DataSets')
df <- read_csv('MERGED2021_22_PP.csv')
```

```
## Rows: 6543 Columns: 3214
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3201): OPEID, OPEID6, INSTNM, CITY, STABBR, ZIP, ACCREDAGENCY, INSTURL,...
## dbl   (13): UNITID, HCM2, MAIN, NUMBRANCH, PREDDEG, HIGHDEG, CONTROL, ST_FIP...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df<-df %>%
    mutate_all(~ifelse(. == "NULL", NA, .))

vis_miss(df, warn_large_data = FALSE)
```



Function for Initial Data Filtering

1. **Handling NULL Values:**

    - Replaces all occurrences of “NULL” values with a standardized “NA” representation. This conversion ensures uniformity in missing data representation.

2. **Removing Bottom Rows:**

    - The function identifies and removes the bottom rows of the dataset. These rows are typically considered to contain less relevant or critical information for analysis. Eliminating them can improve computational efficiency and focus the analysis on more informative data points.

3. **Filtering Columns by Completeness:**

    - Filters out columns that do not meet a specified completeness threshold, which is set at 70% in our implementation. This function helps eliminate variables with significant missing data that could potentially skew analysis results.

4. **Converting Columns to Categorical Data:**

- Finally, the function incorporates the conversion of specific columns into categorical data. This conversion aids in ensuring that the dataset is appropriately formatted making it more suitable for further analysis.

```
df_filter <- function(df) {
  # Changing all NULL values into NA values
  df <- df %>%
    mutate_all(~ifelse(. == "NULL", NA, .))

  # Remove bottom rows since they do not contain important info for analysis
  df_test <- df[1:6049,]

  # Identify columns that are at least 70% complete
  names <- apply(df_test, 2, function(x) sum(!is.na(x)) / length(x) > 0.70)
  df <- as.data.frame(df)

  # Keep only the columns that meet the completeness threshold
  df <- df[, c(names)]

  # Convert selected columns to categorical data
  df <- df_categorical(df)

  return(df)
}
```

Function for Categorizing Columns

This function categorizes specific columns in the DataFrame into meaningful categories.

The columns being categorized are ‘Highest degree’, ‘Predominant degree’, and ‘Carnegie classification’.

```
df_categorical <- function(df){
  # Define categories for 'Predominant degree' column
  pred_deg_awarded <- c('Not classified', 'Certificate',
                        'Associate', 'Bachelor',
                        'Graduate')

  # Define categories for 'Highest degree' column
  high_deg_awarded <- c('Non-degree', 'Certificate',
                        'Associate', 'Bachelor',
                        'Graduate')
  # Define categories for 'Carnegie classification' column
  cc_basic_score <- c('Assoc.: High Transfer-High Traditional', 'Assoc.: High Transfer-Mixed Traditional
                      'Assoc.: High Transfer-High Nontraditional', 'Assoc.: Mixed Transfer/Career & Tech
                      'Assoc.: Mixed Transfer/Career & Technical-Mixed Traditional/Nontraditional','Asso
                      'Assoc.: High Career and Technical- High Traditional','Assoc.: High Career and Tec
                      'Assoc.: High Career and Technical-High Nontraditional', 'SF 2-Yrs- Health Profes
                      'SF 2-Yrs: Technical Professions','SF 2-Yrs: Arts and Design',
                      'SF 2-Yrs: Other Fields','Bacc./Assoc. Colleges: Assoc. Dominate',
                      'PhD UNI: Very High Research Activity','PhD UNI: High Research Activity',
                      'Doctoral/Professional UNI','Masters Colleges and UNI: Larger Programs',
                      'Masters Colleges and UNI: Medium Programs', 'Master Colleges and UNI: Small Prog
                      'Bacc. Colleges: Arts and Science Focus','Bacc. Colleges: Diverse Fields','Mixed 
                      'SF 4-Yrs: Faith-Related Institutions','SF 4-Years: Medical Schools and Centers',
                      'SF 4-Yrs: Other Health Profession Schools','SF 4-Yrs: Engineering Schools',
                      'SF 4-Yrs: Other Related-Technology Related Schools','SF 4-Yrs: Business and Mana
                      'SF 4-Yrs: Arts, Music, and Design School','SF4-Yrs: Law Schools',
                      'SF 4-Yrs: Other SF Institutions','Tribal Colleges')



  # Change numeric values to categorical values for specific columns

  # Identify rows where 'CCBASIC' column is not equal to -2 (numeric values)
  valid_indx <- which(df$CCBASIC != -2)

  # For the 'CCBASIC' column, transform numeric values to categorical values
  # We use 'cc_basic_score' to map the numeric values to corresponding categorical labels.
  # Adding 1 to the numeric values aligns them with the indices of the category vector.
  df$CCBASIC[valid_indx] <- cc_basic_score[as.numeric(df$CCBASIC[valid_indx]) + 1]

  # Similar transformations are applied to 'HIGHDEG' and 'PREDDEG' columns.
  # Numeric values are mapped to categorical labels using 'high_deg_awarded' and 'pred_deg_awarded' vec
  df$HIGHDEG <- high_deg_awarded[as.numeric(df$HIGHDEG) + 1]
  df$PREDDEG <- pred_deg_awarded[as.numeric(df$PREDDEG) + 1]

  return(df)
}
```

Function for Further Data Trimming

1. **Removing Highly Incomplete Rows:**
    - Identifies and removes rows that are considered highly incomplete. By eliminating these rows, the dataset becomes more concise and focused on relevant data points.
2. **Limiting the Number of Columns:**
    - Restricts the dataset to include only the first 450 columns. This is implemented to reduce computational complexity, speed up the analysis, and narrow the scope of investigation to the most relevant attributes.

```
df_shorten <- function(df){
  # Keep only the first 450 columns
  df <- df[, 1:450]

  # Remove rows that are 90% or more incomplete
  # Calculate the percentage of NA values for each row
  row_na_percent <- rowSums(is.na(df)) / ncol(df) * 100

  # Find rows with 90% or more NA values
  rows_with_90_percent_or_more_na <- which(row_na_percent >= 90)

  # Create a temporary DataFrame (na_df) containing rows with 90% or more NA values
  na_df <- df %>%
    filter(row_number() %in% rows_with_90_percent_or_more_na)

  # Create a filtered DataFrame (filtered_df) that excludes rows with 90% or more NA values
  filtered_df <- df %>%
    filter(!(row_number() %in% rows_with_90_percent_or_more_na))

  # Return the filtered DataFrame
  return(filtered_df)
}
```

Function for Filtering Data by State

This function filters the DataFrame to retain only rows associated with actual U.S. states.

```
state_isolation <- function(df){
  # List of two-letter state abbreviations for all U.S. states
  states <- c('AL', 'AK', 'AZ', 'AR', 'CA', 'CO',
              'CT', 'DE', 'FL', 'GA', 'HI', 'ID',
              'IL', 'IN', 'IA', 'KS', 'KY', 'LA',
              'ME', 'MD', 'MA', 'MI', 'MN', 'MS',
              'MO', 'MT', 'NE', 'NV', 'NH', 'NJ',
              'NM', 'NY', 'NC', 'ND', 'OH', 'OK',
              'OR', 'PA', 'RI', 'SC', 'SD', 'TN',
              'TX', 'UT', 'VT', 'VA', 'WA', 'WV',
              'WI', 'WY')

  # Remove any rows associated with territories or entities that are not considered U.S. states
  filtered_df <- df %>%
    filter(STABBR %in% states)

  # Return the filtered DataFrame containing only U.S. state data
  return(filtered_df)
}
```

Following four functions prepare dataset to facilitate future plotting functions.

They filter and aggregate data, ensuring that it is well-structured for creating plots and graphs for data visualizations.

Function for Data Preparation for Bar Plot

This function prepares the dataset to be used in the 'bar_plot' function by filtering and aggregating data related to institutions and their operational status at the state level.

```
instit_operations <- function(df){
  # Filter the DataFrame to retain only rows associated with actual U.S. states
  filtered_df <- state_isolation(df)

  # Count the number of institutions in each U.S. state
  state_count <- filtered_df %>%
    count(STABBR)

  # Filter and count the number of currently open institutions in each U.S. state
  open_instit <- filtered_df %>%
    filter(CURROPER == 1) %>%
    count(STABBR)

  # Filter and count the number of closed institutions in each U.S. state
  closed_instit <- filtered_df %>%
    filter(CURROPER == 0) %>%
    count(STABBR)

  # Return a list containing the three datasets for use in the 'bar_plot' function
  return(list(state_count = state_count, open_instit = open_instit, closed_instit = closed_instit))
}
```

Function for Data Preparation for Grouped Bar Plot

This function filters the dataset to be used in the 'group_bar_plot' function by extracting specific columns, counting occurrences based on degree and state, and ensuring that only states are considered.

```
group_bar_filter <- function(df){
  # Extract two columns 'STABBR' (state abbreviation) and 'PREDDEG' (predominant degree)
  state_class <- df[c('STABBR', 'PREDDEG')]

  # Count occurrences of unique combinations of 'PREDDEG' and 'STABBR' (state abbreviation)
  unique_combo <- state_class %>%
    group_by(STABBR, PREDDEG) %>%
    summarise(UniqueCount = n())

  # Ensure that only data related to actual U.S. states is considered
  unique_combo <- state_isolation(unique_combo)

  # Calculate and add a new column of z-scores
  unique_combo$zscore <- scale(unique_combo$UniqueCount)[,1] # Normalizing calculations using z-score

  return(unique_combo)
}
```

Function for Data Preparation for Ridge Plot

This function filters the dataset to be used in the 'ridge_plot' function by extracting specific columns related to student demographics and principal institution information.

```
ridge_plot_filter <- function(df){
  # Extract columns related to student demographic information (columns 264 to 273)
  demograph <- df[, 264:273]

  # Extract columns related to principal institution information including state, admission rate, compl
  pred_undergrad <- cbind(STABBR = df[, 6], df[, 14:15], CCBASIC = df[, 22])

  # Combine the demographic and principal institution information into a single DataFrame
  # This analysis focuses on special focus 4-year institutions and those considered baccalaureate
  bach_stud <- cbind(pred_undergrad, demograph)
  # Return the filtered dataset prepared for the 'ridge_plot' function
  return(bach_stud)
}
```

Function for Data Preparation for Expense vs. Profit Plot

This function filters the dataset to be used in the ‘exp__vs__profit__plot’ function by extracting specific columns related to institution finance, including state, degree information, tuition revenue, and instructional expenses.

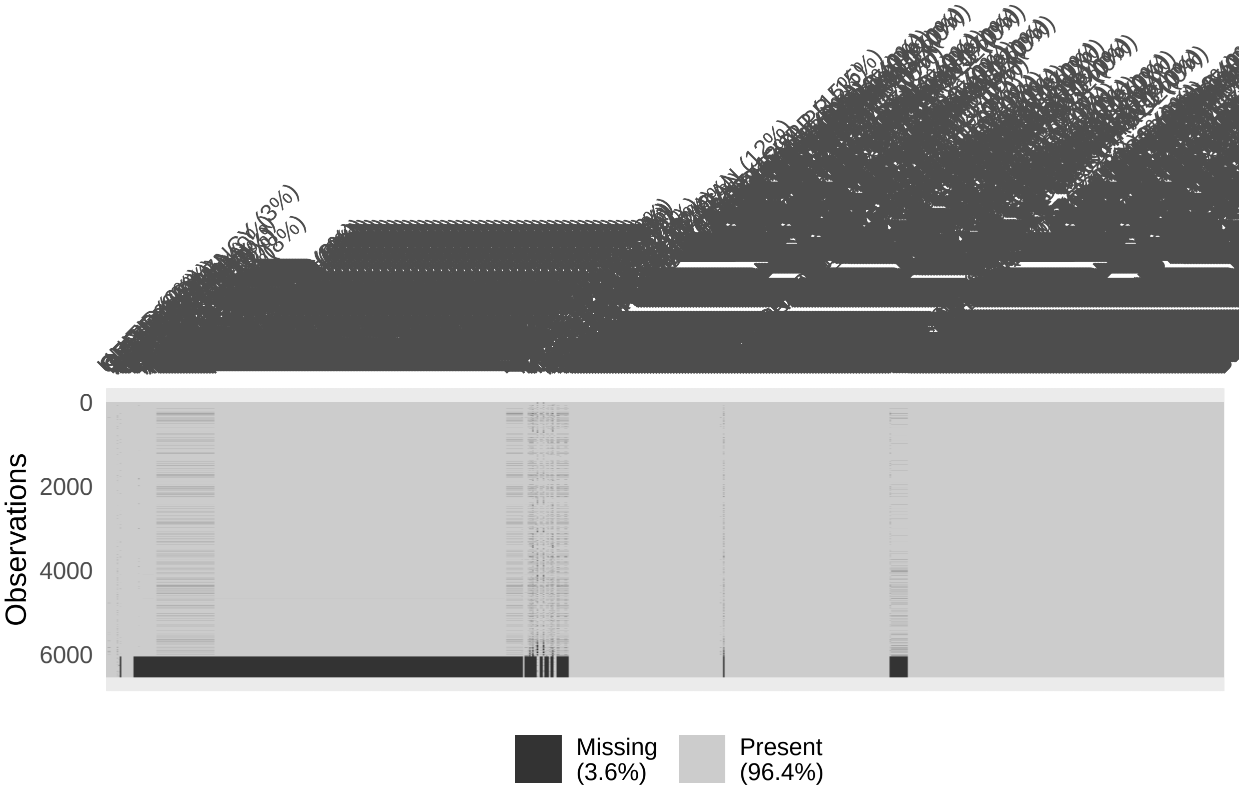
```
exp_vs_profit_filter <- function(df, colname){
  # Create a DataFrame that contains relevant columns (State, Predominant degree, Highest degree, Carne
  finance_df <- cbind(STABBR = df[, 6], df[, 14:15], CCBASIC = df[, 22], df[, 275:276])

  # Calculate the mean of tuition revenue for each classification within each state
  tuition_rev_df <- finance_df %>%
    group_by(STABBR, !!sym(colname)) %>%
    summarise(state_tuition_rev_mean = mean(as.numeric(TUITFTE), na.rm = TRUE))

  # Calculate the mean of instructional expenses for each classification within each state
  expense_df <- finance_df %>%
    group_by(STABBR, !!sym(colname)) %>%
    summarise(state_expense_mean = mean(as.numeric(INEXPFTE), na.rm = TRUE))

  # Merge the information into one DataFrame
  exp_vs_profit_df  <- cbind(tuition_rev_df, expense_df[, 3])

  # Calculate the difference between revenue and expenses
  exp_vs_profit_df <- exp_vs_profit_df %>%
    mutate(difference = state_tuition_rev_mean - state_expense_mean)

  # Return the filtered dataset prepared for the 'exp_vs_profit_plot' function
  return(exp_vs_profit_df)
}
```

Data Filtering and Operations

```
# Applying the 'df_filter' function to filter the DataFrame and store the result in 'df_trimmed'
df_trimmed <- df_filter(df)

# Applying the 'df_shorten' function to further filter 'df_trimmed' and store the result in 'filtered_d
filtered_df <- df_shorten(df_trimmed)

# Applying the 'instit_operations' function to 'filtered_df' and store the results in 'institut_ops'
institut_ops <- instit_operations(filtered_df)

# Applying the 'group_bar_filter' function to 'filtered_df' and store the results in 'group_plot_filter
group_plot_filtered_df <- group_bar_filter(filtered_df)
```

```
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
```

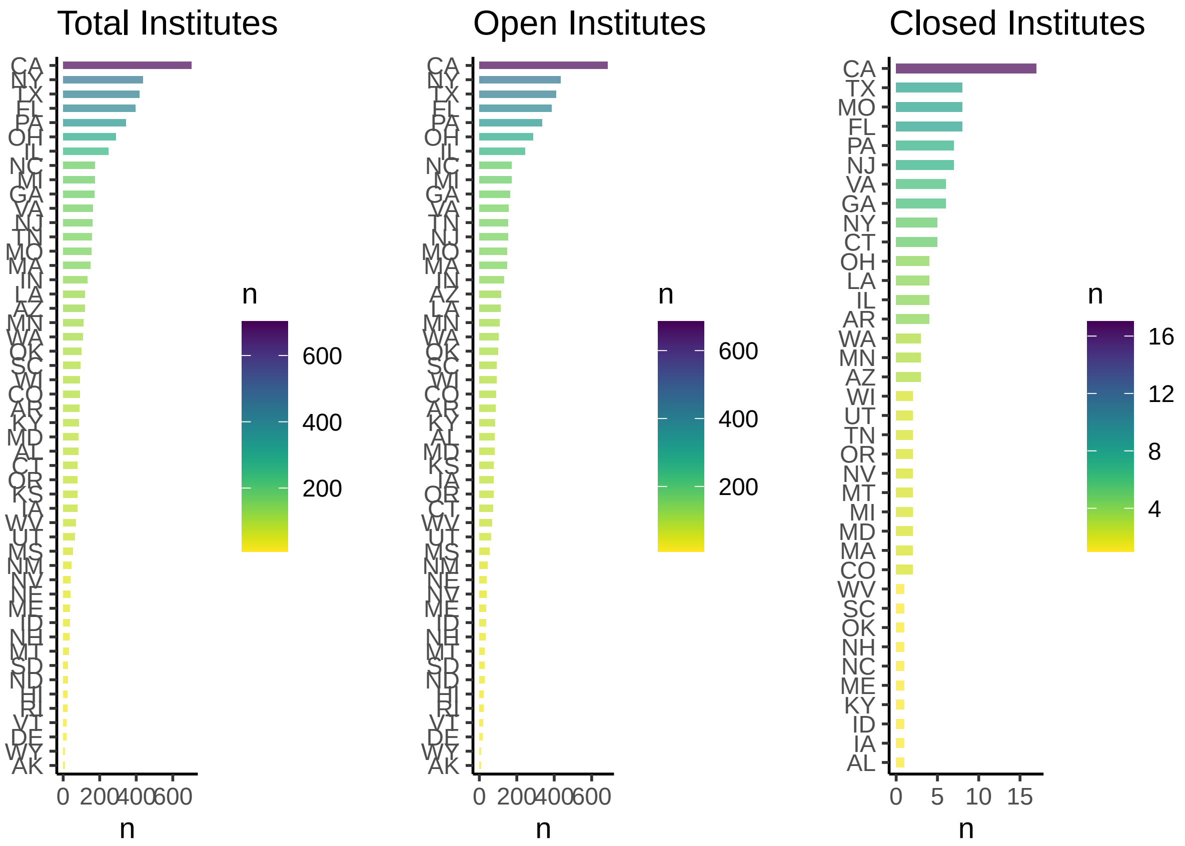
```
# Applying the 'ridge_plot_filter' function to 'filtered_df' and store the results in 'ridge_plot_df'
ridge_plot_df <- ridge_plot_filter(filtered_df)

vis_miss(df_trimmed, warn_large_data = FALSE)
```



Filter Data for Open Institutions

```
df_trimmed <- df_trimmed%>%
  filter(CURROPER == 1)

df_trimmed <- state_isolation(df_trimmed)
```

```
bar_plot <- function(df) {
  titles <- c('Total Institutes', 'Open Institutes', 'Closed Institutes') # main title
  ops_plots <- list() # list of 3 different graphs

  # looking thorough total, open, and closed institution's df
  for (ops in 1:length(df)) {
    data <- df[[ops]] # getting index from list
    title <- titles[ops] # retrieving associated title

    # order the bars in descending order
    data <- data %>%
      mutate(STABBR = fct_reorder(STABBR, n))

    ops_plot <- ggplot(data, aes(x = STABBR, y = n, fill = n)) +
      geom_bar(stat = 'identity', # determines how the data should be summarized (height based on value
               alpha = 0.7,
               width = 0.5) +
      coord_flip() +
      scale_fill_viridis(discrete = FALSE, # coloring system is continuous
                         option = 'viridis', # color palette used
                         direction = -1)+ # direction refers to the gradient direction
      xlab('') +
      ggtitle(title) +
      theme_classic() # background color of window

    ops_plots[[ops]] <- ops_plot # adding plot to a list
  }

  # Combine the plots into one
  combined_plots <- grid.arrange(grobs = ops_plots, ncol = length(ops_plots))
  # Return the combined plot
  return(combined_plots)
}

# Call the bar_plot function
combined_plot <- bar_plot(institut_ops)
```



ANALYSIS: Bar Plot for Total, Open, and Closed Institutions

The purpose of this analysis is to visualize the number of educational institutions in each U.S. state, categorized into three separate plots: one for total institutions, one for open institutions, and one for closed institutions. These plots offer insights into the distribution of institutions across states and highlight significant trends.

Key Observations:

1. **California Dominance:** It's immediately evident that the state of California consistently ranks at the top in all three plots, indicating the highest number of total, open, and closed institutions compared to other states. This dominance is likely due to California's substantial population and its attractiveness as a destination for education, drawing students from across the United States and around the world. Several other factors contributing to this dominance could also include: economic factors, cultural and

social factors, educational policies, geographical factors, education quality and more.

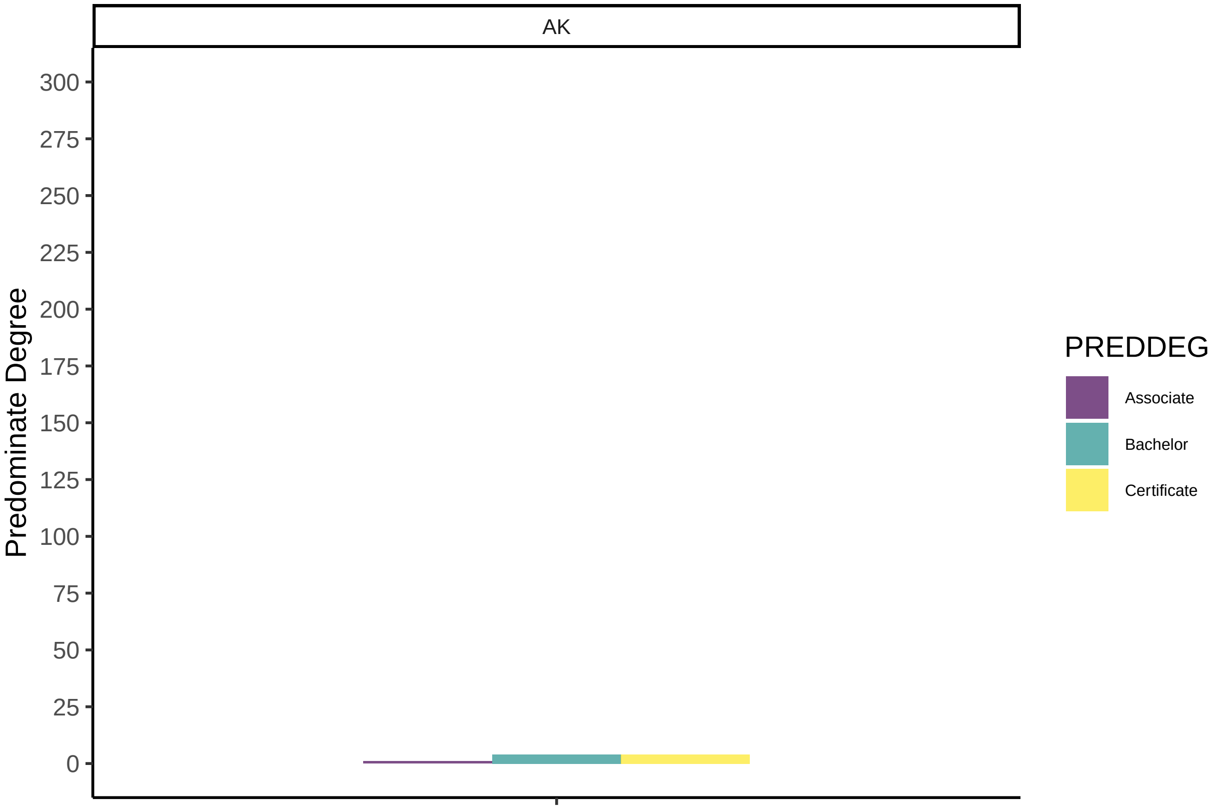
2. **New York, Texas, and Florida:** Following California, New York secures the second spot in terms of total and open institutions. This can be attributed to the high population density and the presence of numerous prestigious universities and colleges. Texas and Florida closely follow, mirroring similar population-driven demand for educational institutions.
3. **Population and Demand:** The observed trends are strongly correlated with population size and migration patterns. States with larger populations naturally exhibit a higher demand for educational opportunities, leading to a greater number of institutions. Conversely, at the bottom of the graph, states like Alaska and Wyoming have the fewest total and open institutions, in line with their lower populations. In areas with lower populations, the demand for educational institutions is notably lower.
4. **California's Closed Institutions:** An interesting observation is that in addition to having the highest number of open institutions, California also leads in the number of closed institutions. This phenomenon can be attributed to various factors, including intense competition, economic factors, educational innovations, and more. Importantly, this observation is not necessarily a negative reflection on California's educational system. Given California's size, population, and the sheer volume of institutions, it's expected that the number of closed institutions would also be higher compared to other states.

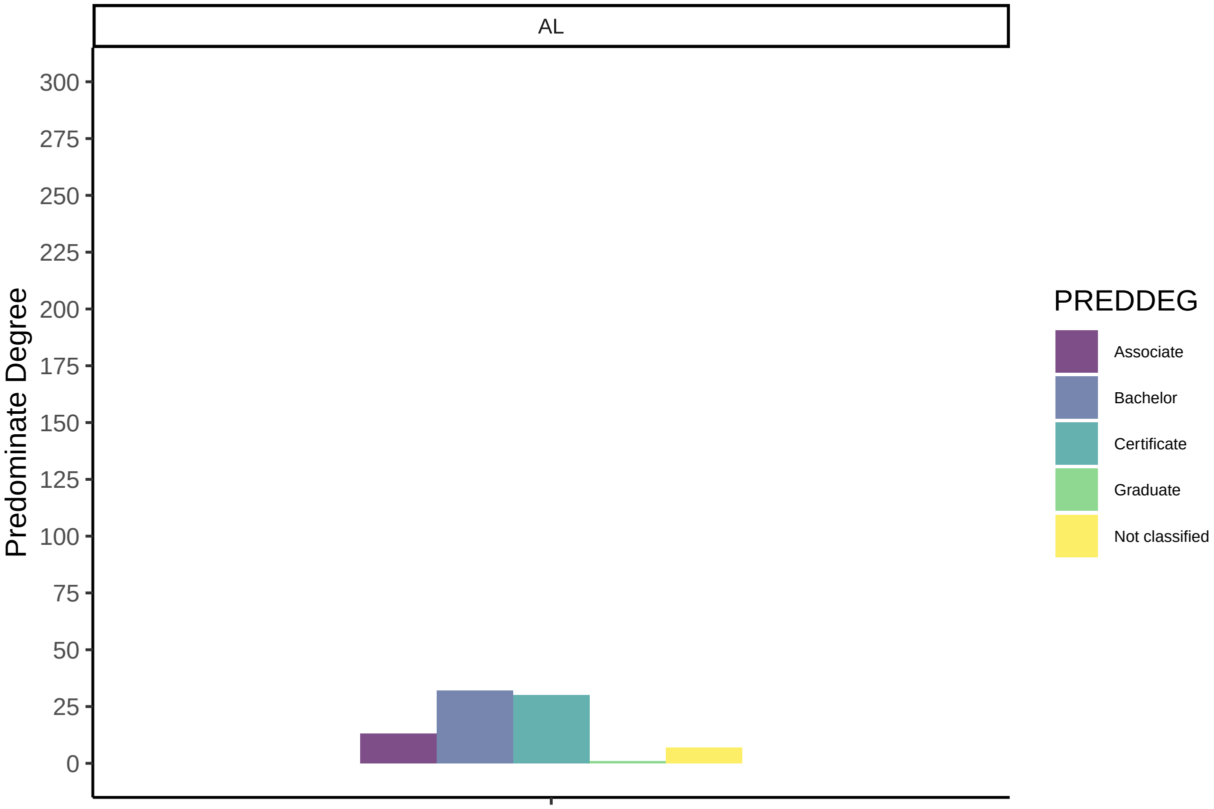
Conclusion:

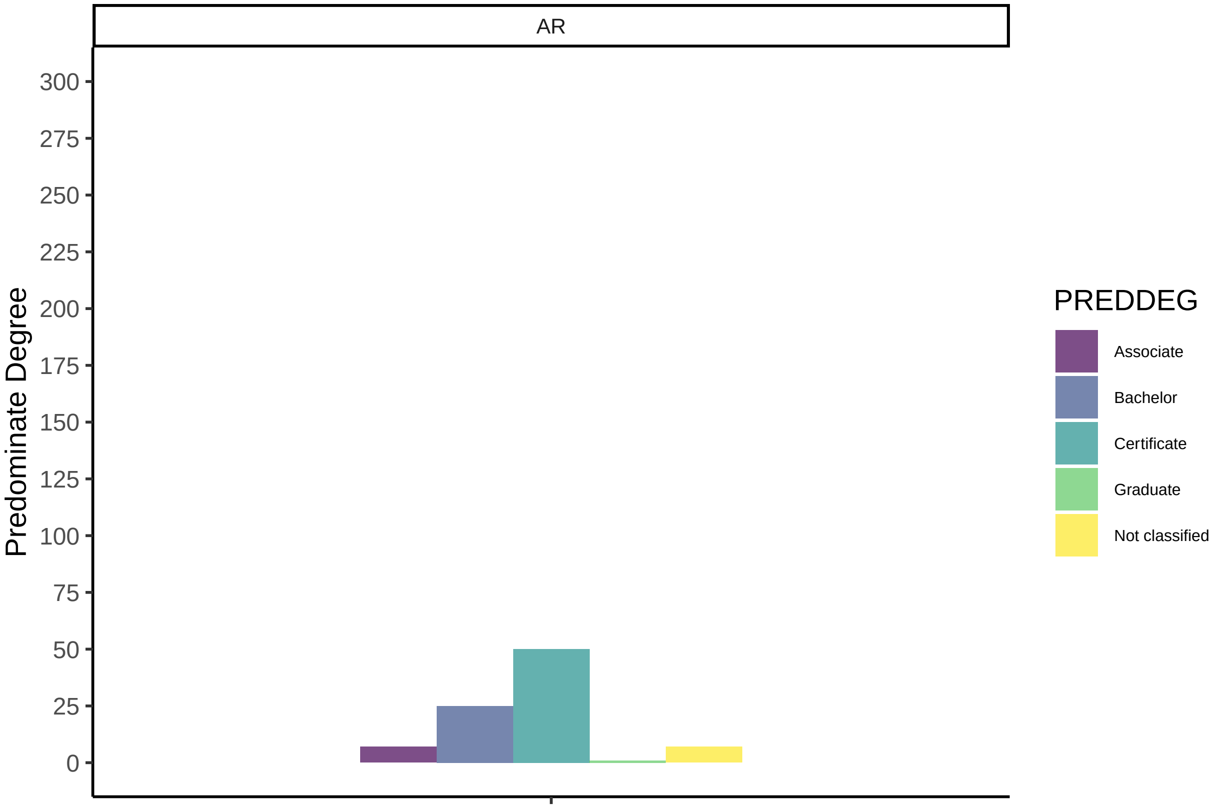
Understanding these patterns is valuable for parents when choosing schools for their children because it provides insights into the educational landscape of different states. These insights help make informed decisions about educational opportunities and policies.

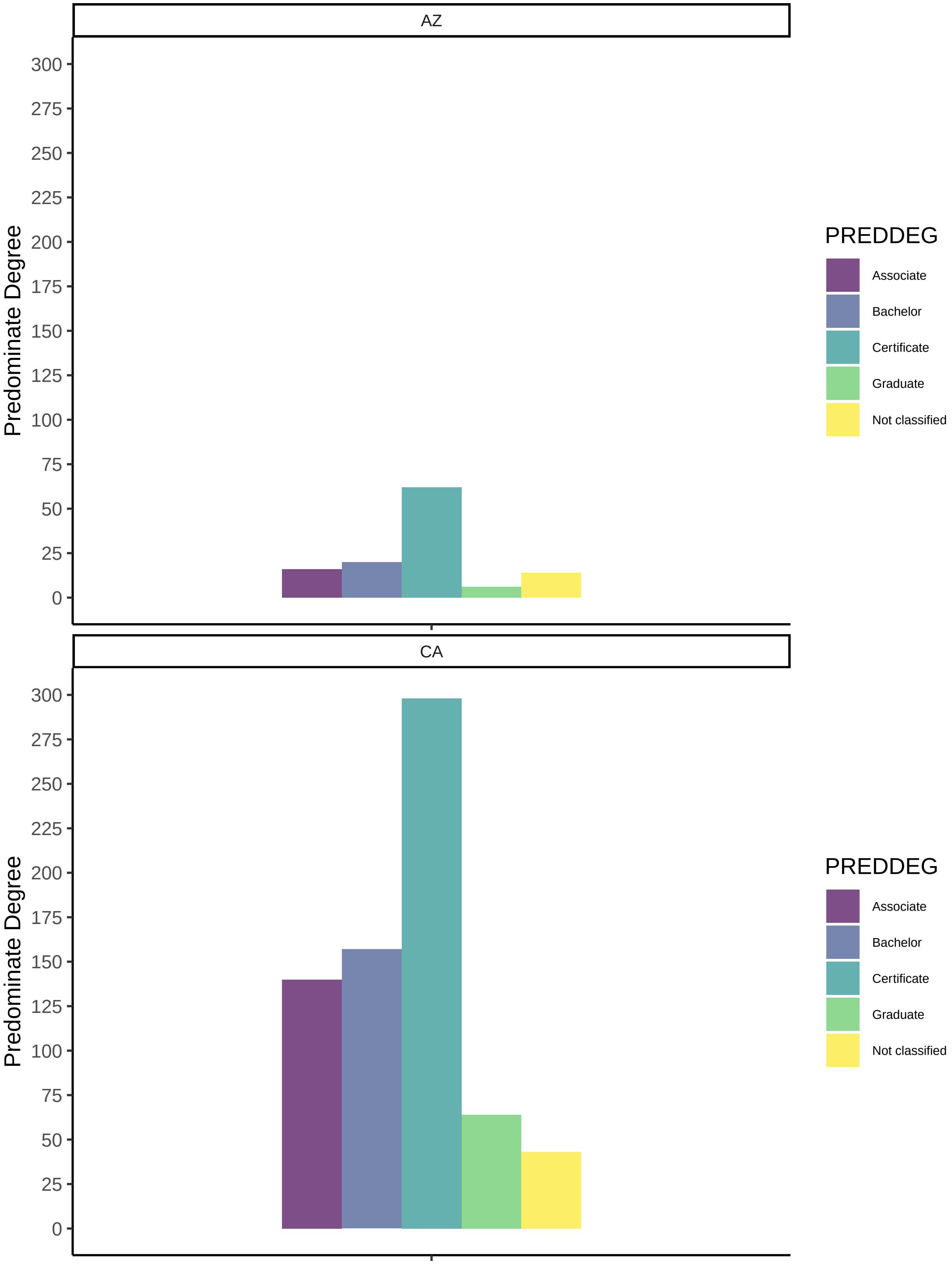
Grouped Bar Plot for Predominant Degree # creating a bar plot that looks at the frequency of predominate degree # within each state and returns a pdf file that contains a graph of each state

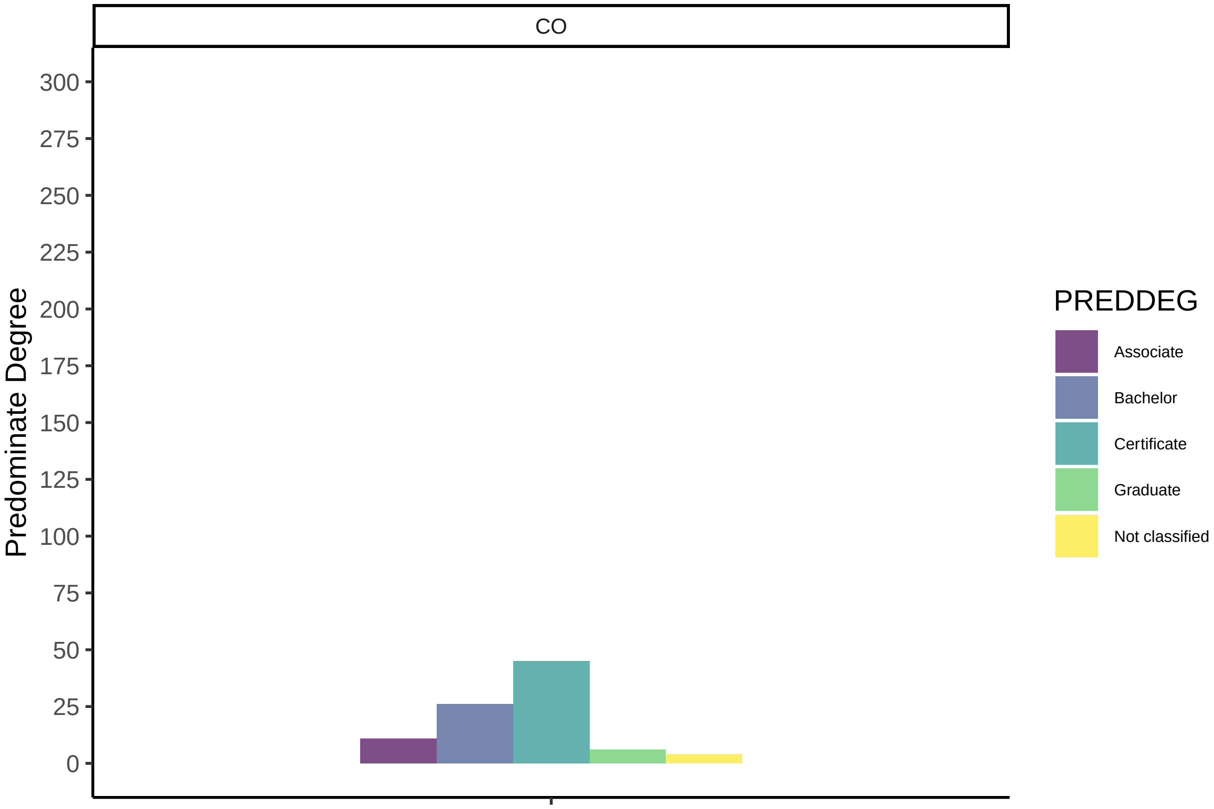
```
group_bar_plot <- function(df) {
  plots <- lapply(unique(df$STABBR), function(state) {

    # uses built-in state.abb to group graph based on state
    data_subset <- subset(df, STABBR == state)

    cc_dist <- ggplot(data_subset, aes(x = STABBR, y = UniqueCount, fill = PREDDEG)) +
      geom_bar(stat = 'identity',
               alpha = 0.7, # transparency of bars
               width = 0.5, # bar width
               position = 'dodge') + # each distinct bar is positioned next to each other
      facet_wrap(~STABBR, nrow = 10) +
      scale_fill_viridis(discrete = TRUE,
                         option = 'viridis', # color palette being used in viridis library
                         direction = 1)+ # direction of gradient color
      # setting parameters for y-axis
      scale_y_continuous(name = 'Predominate Degree', # y title
                         limits = c(0, 300), # the min and max of y-axis
                         breaks = seq(0,300, by = 25))+ # the incrementation markers of y-axis
      xlab('') +
      theme_classic()+
      theme(legend.text = element_text(size = 6), # font text of variables in legend
            legend.position = 'right', #
            axis.text.x = element_blank(),
            strip.text = element_text(size = 8))# the text on top of each graph

    return(cc_dist)
  })

  for (i in 1:length(plots)) {
    print(plots[[i]])
  } # Close PDF device

  return(plots)

}

# Call the group_bar_plot function
preddeg_plots <- group_bar_plot(group_plot_filtered_df)
```

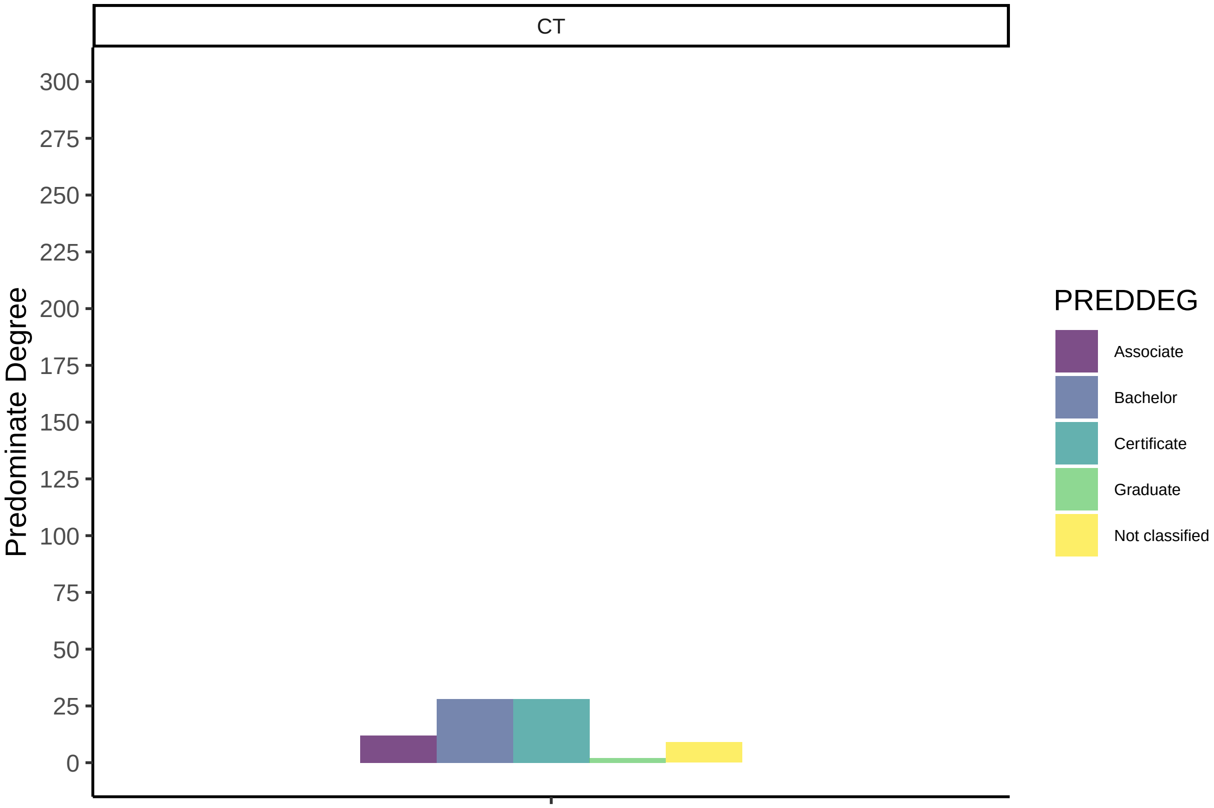


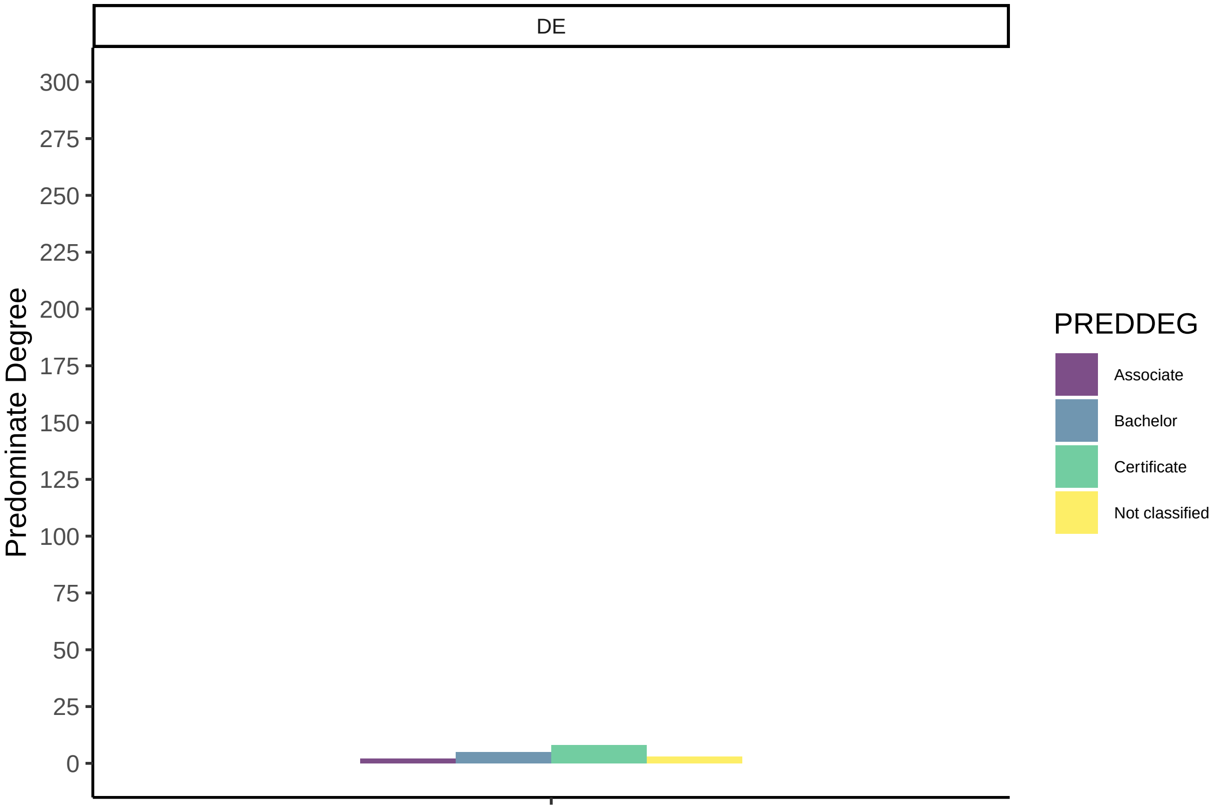

AL
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

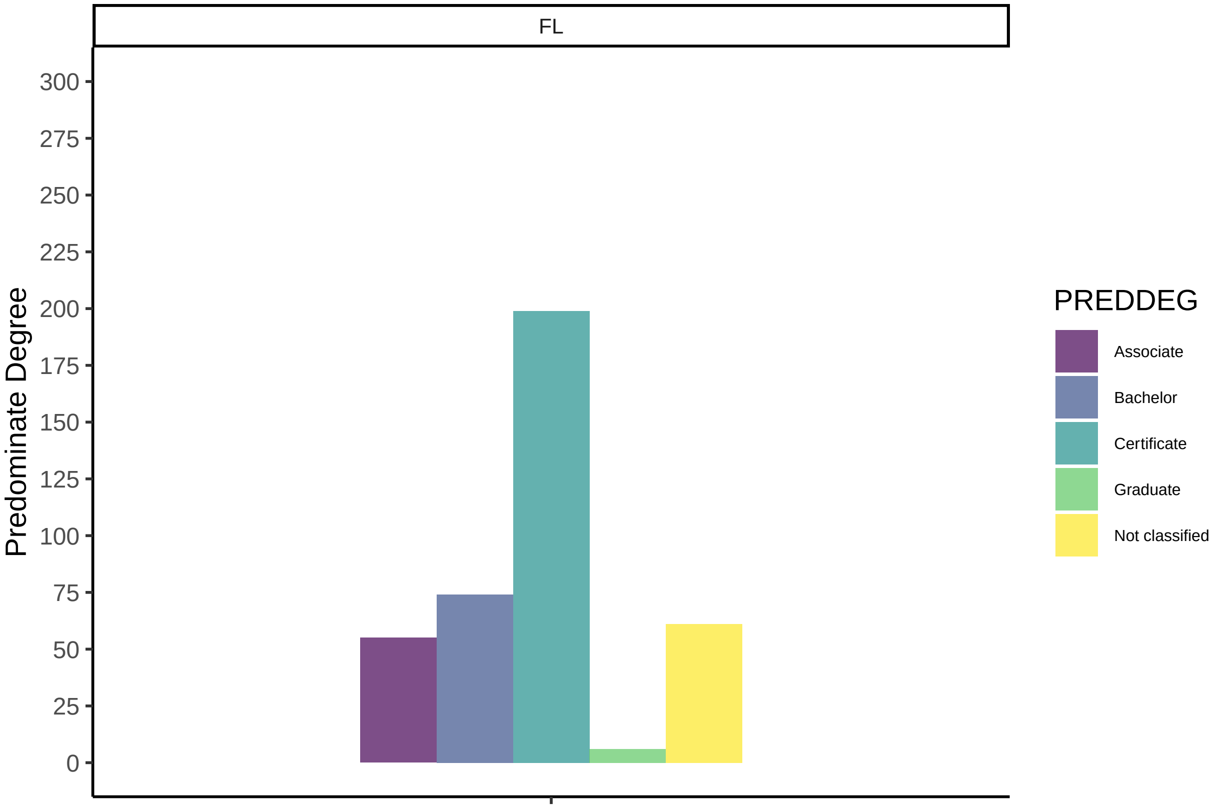

AR
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

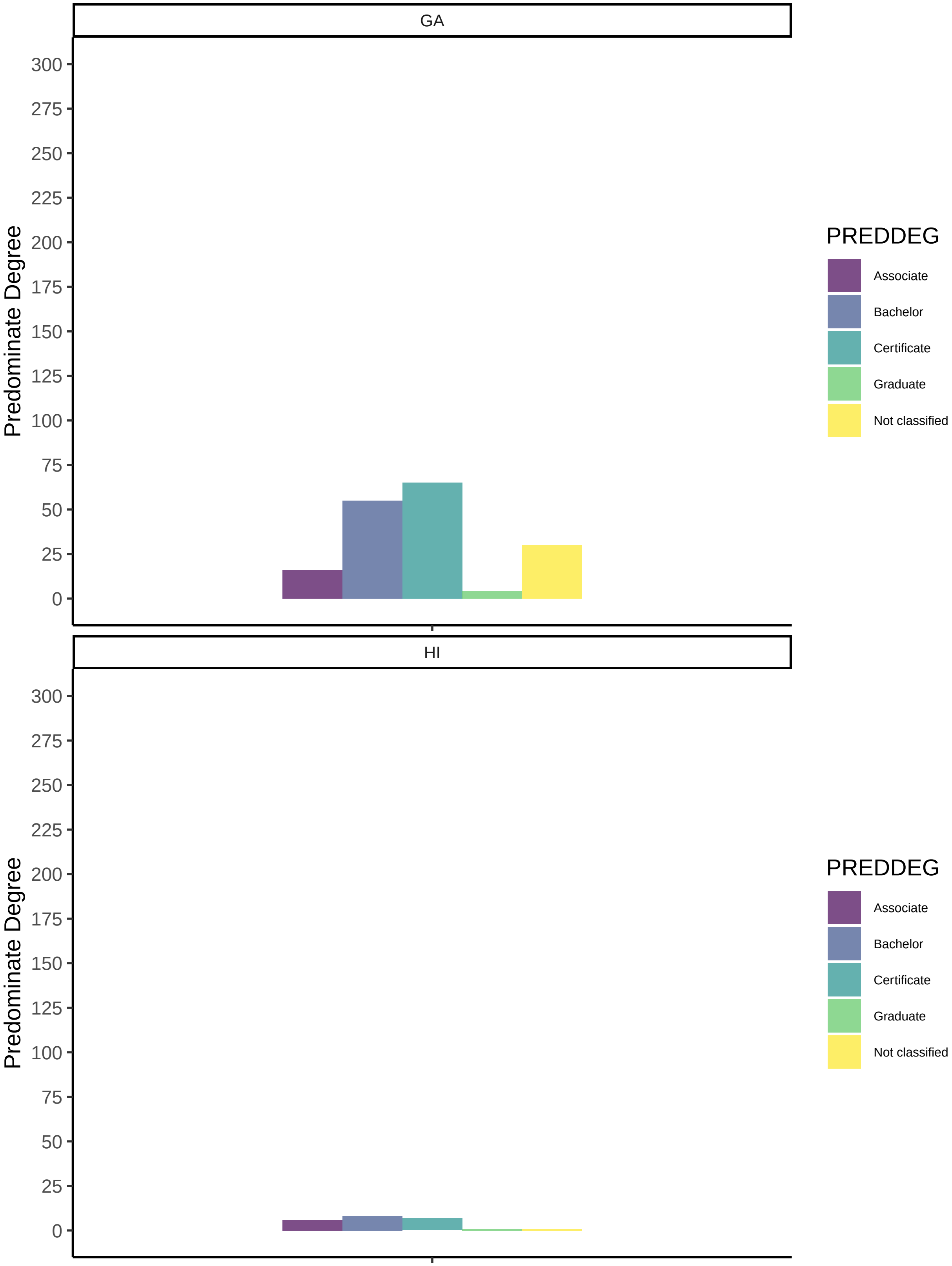

AZ
CA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

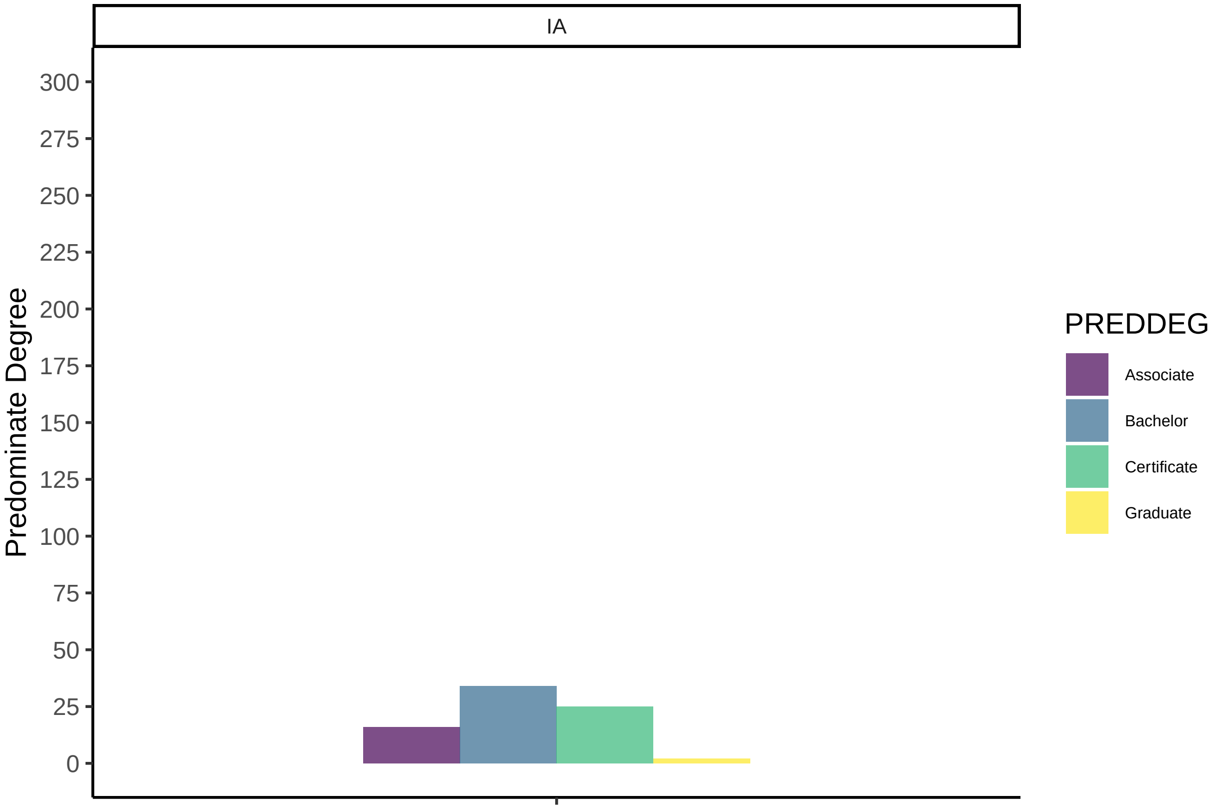

CO
300
275
250
225
200
175
150
125
100
75
50
25
0
Predominate Degree
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

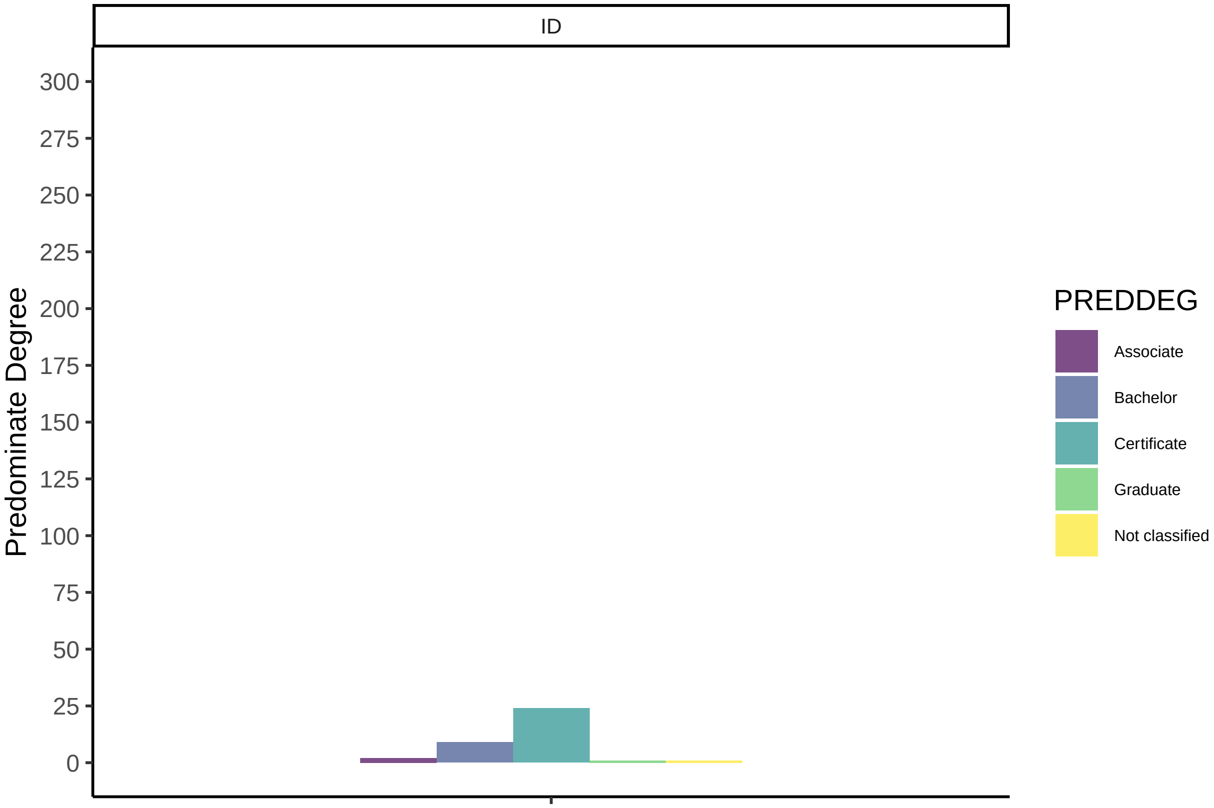

CT
300
275
250
225
200
175
150
125
100
75
50
25
0
Predominate Degree
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

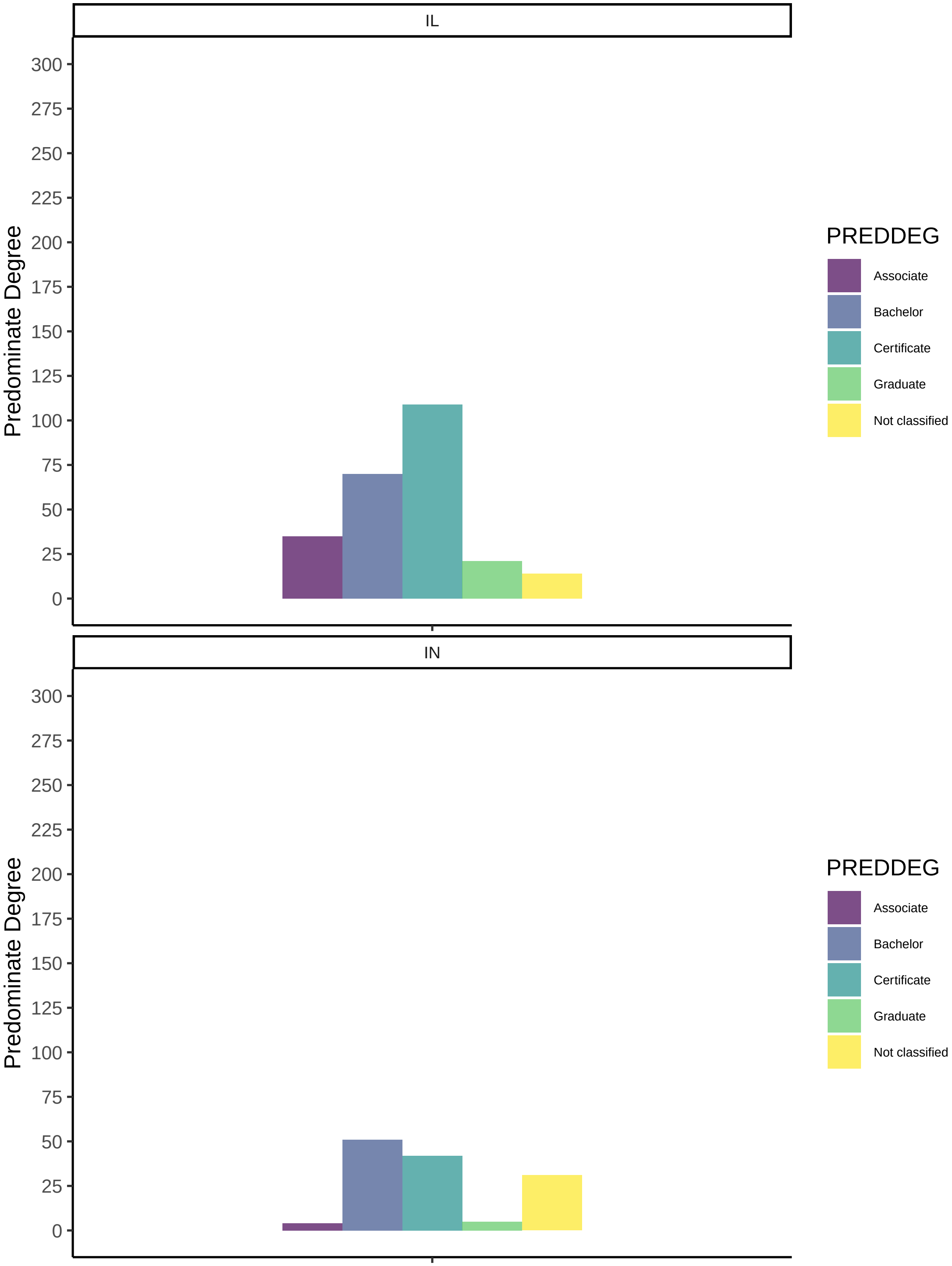

DE
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Not classified

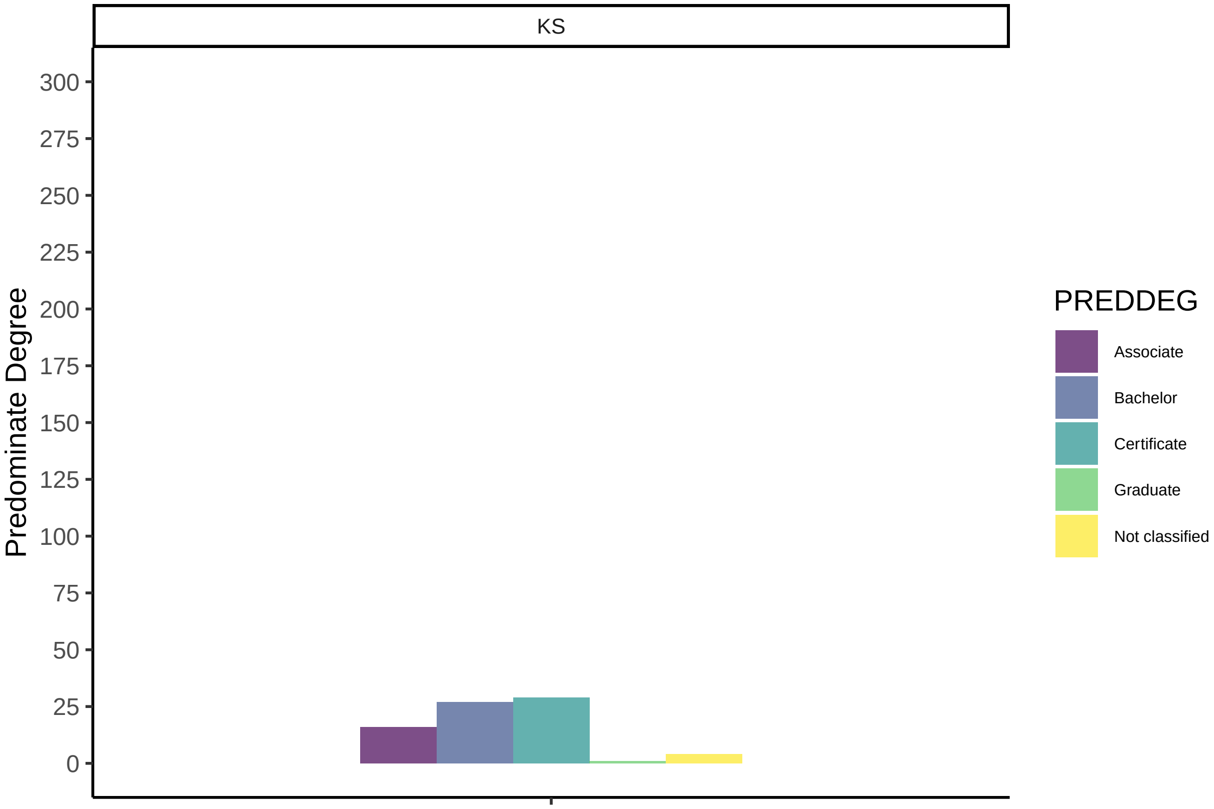

FL
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

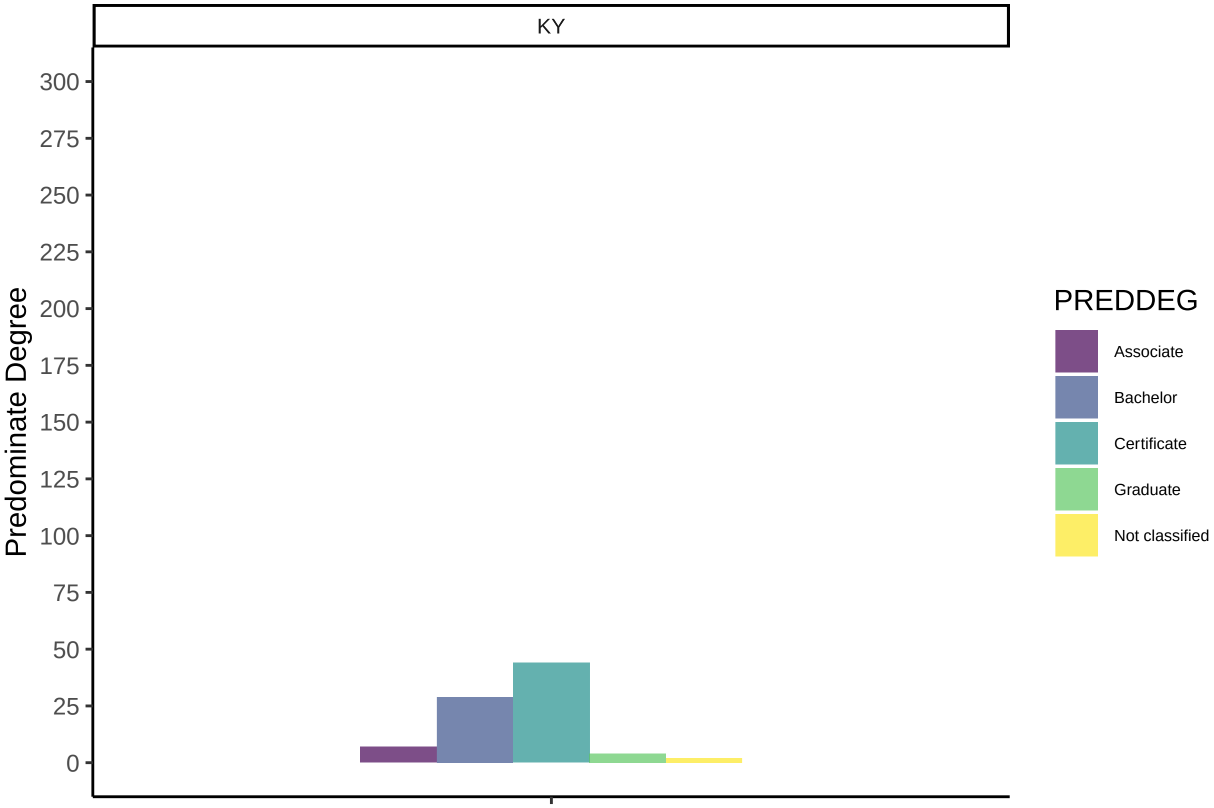

GA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified
HI
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

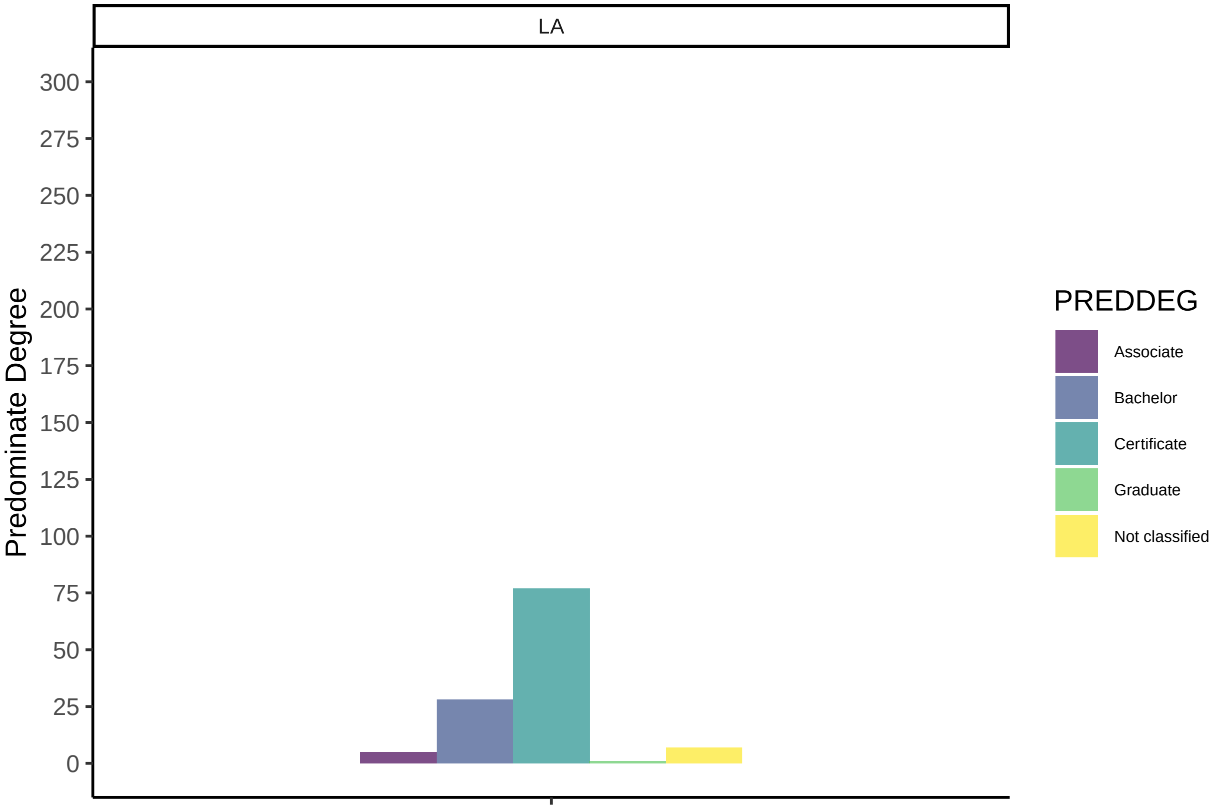

IA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate

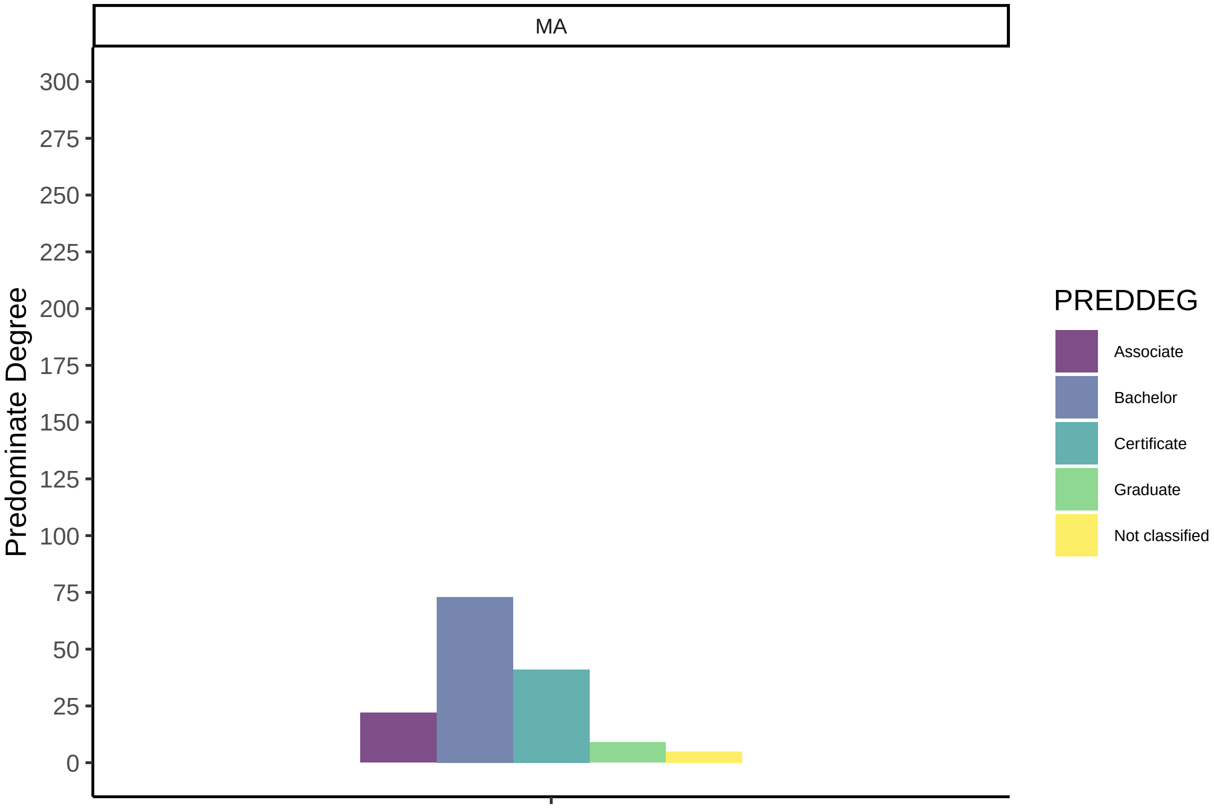

ID
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

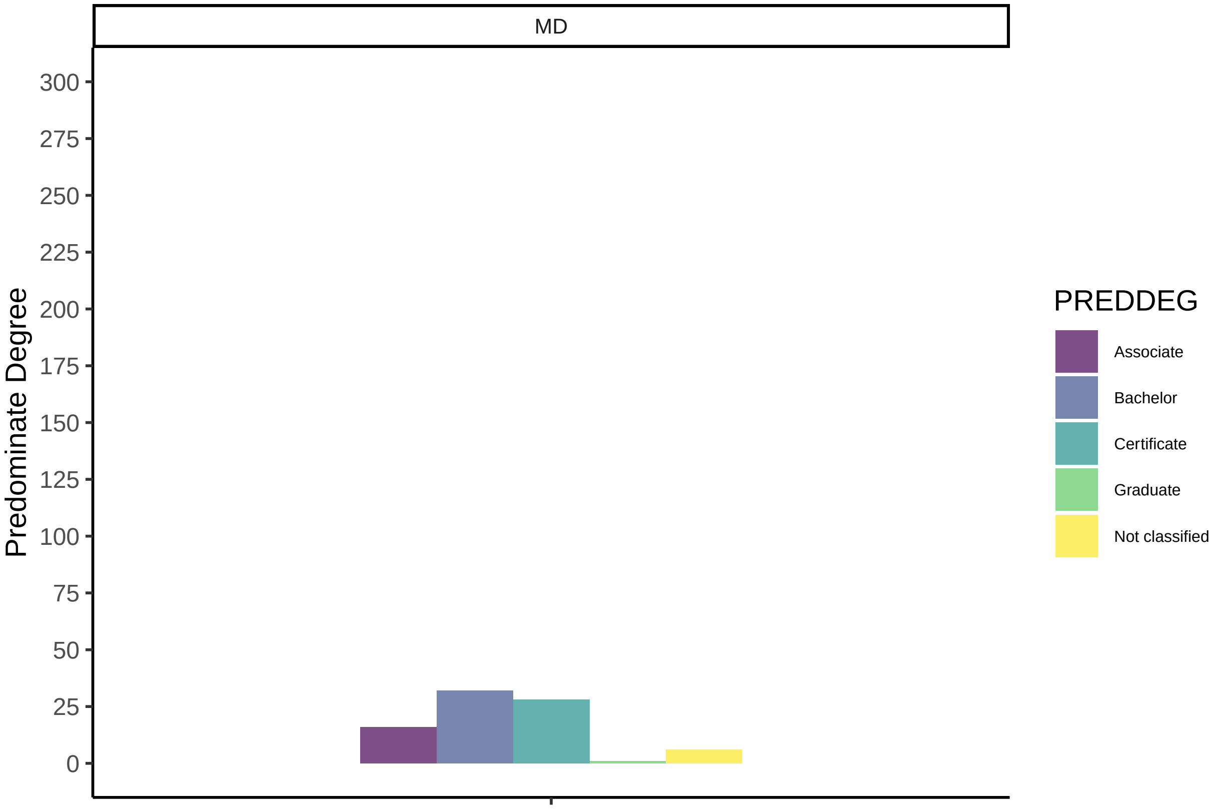

IL
300
275
250
225
200
175
150
125
100
75
50
25
0
Predominate Degree
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified
IN
300
275
250
225
200
175
150
125
100
75
50
25
0
Predominate Degree
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

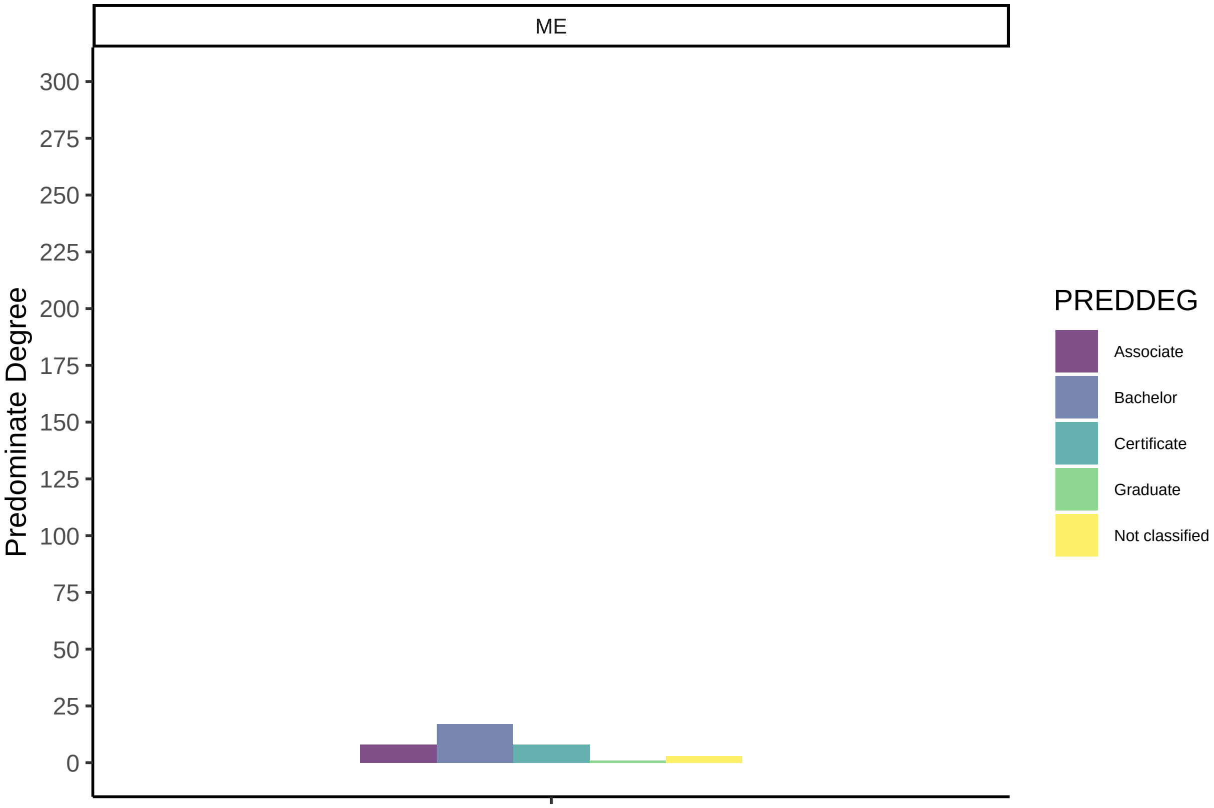

KS
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

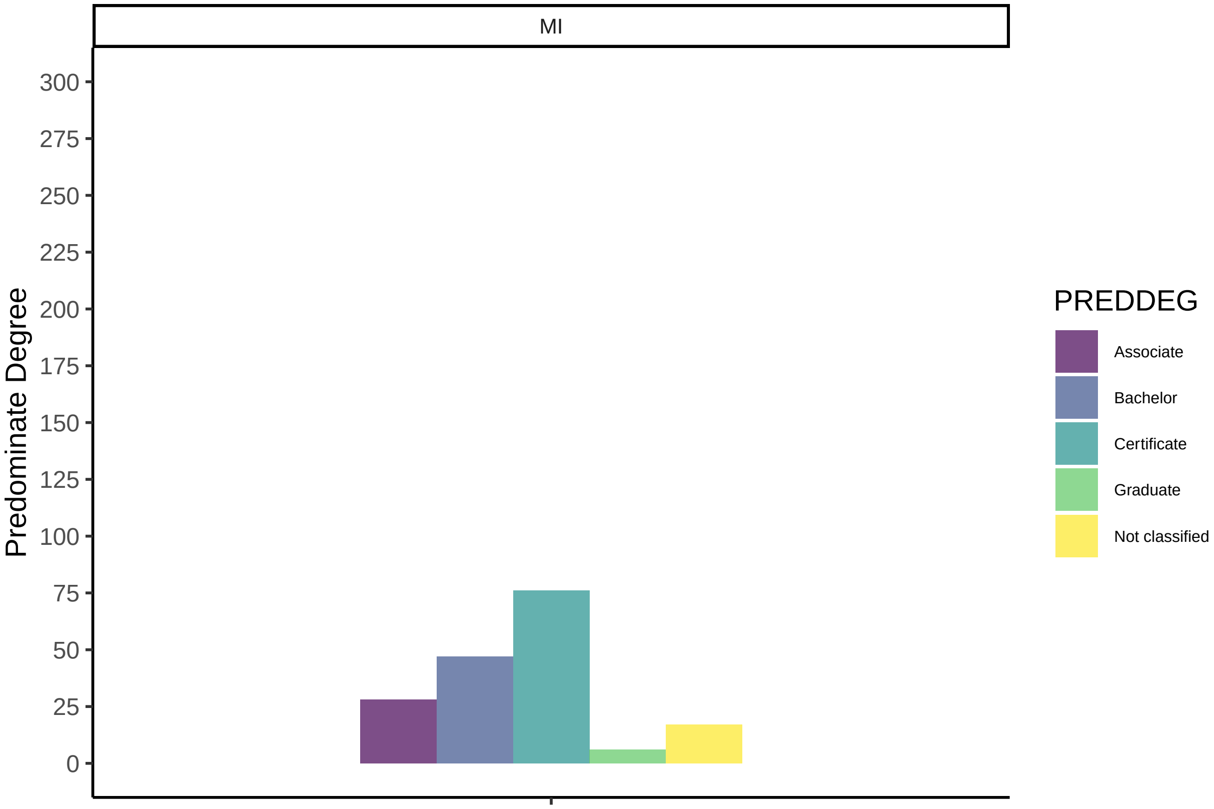

KY
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

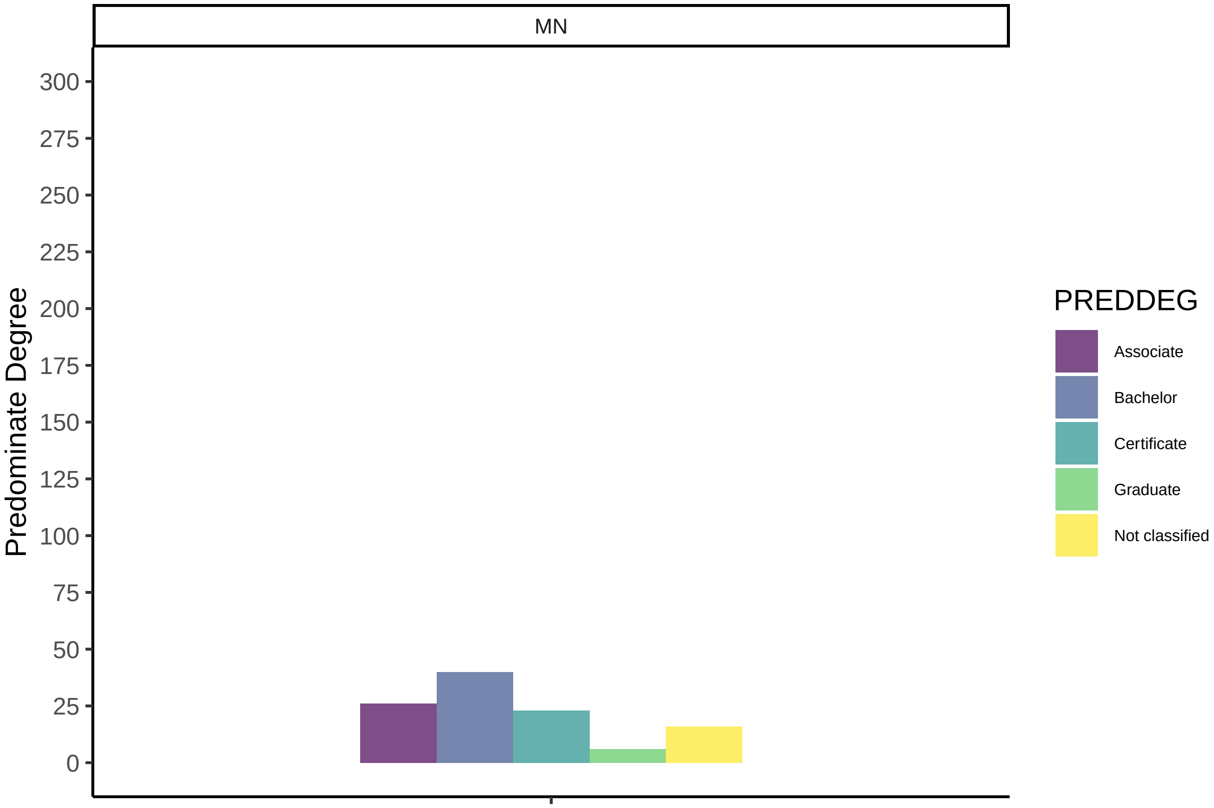

LA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

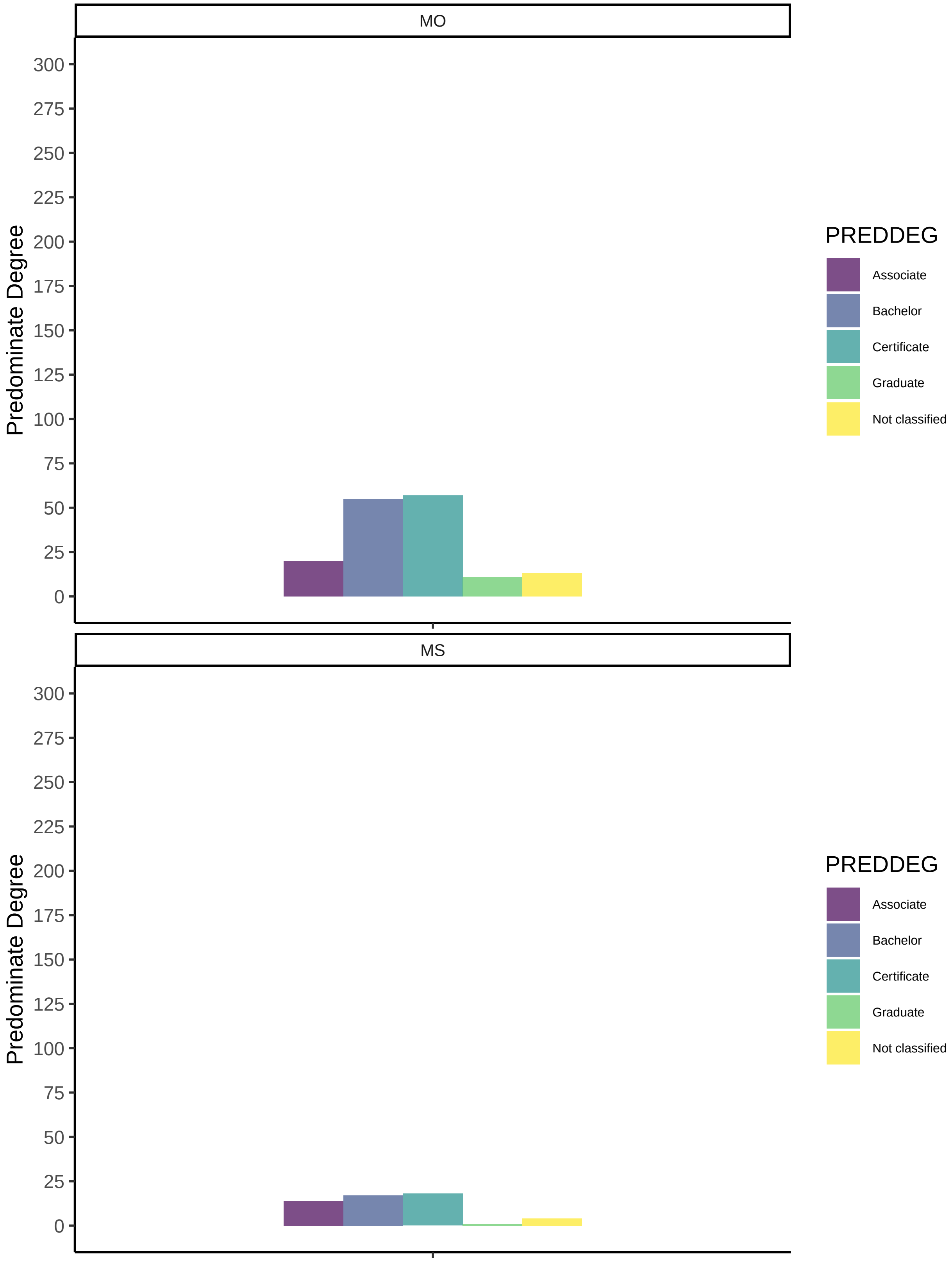

MA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

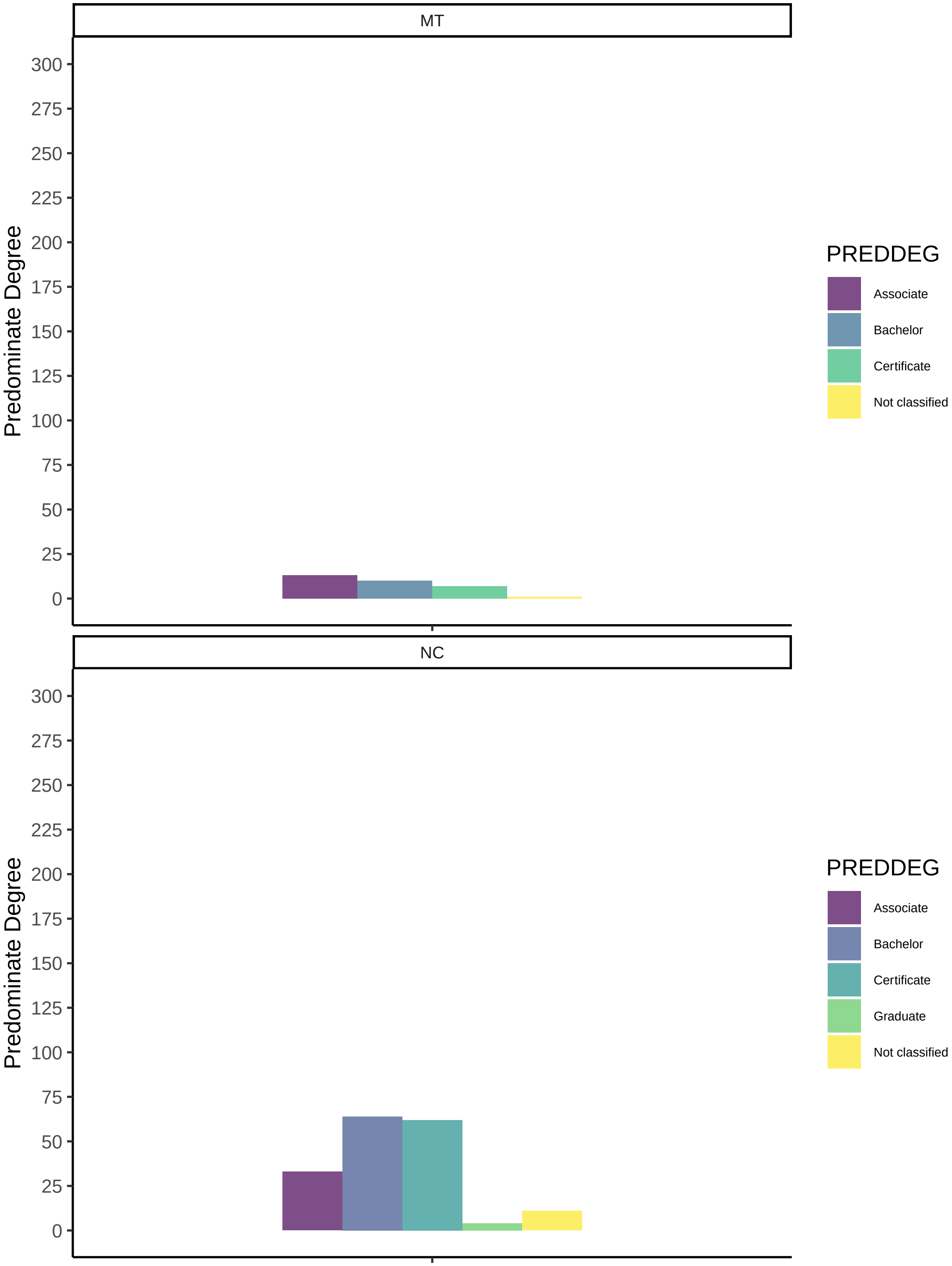

MD
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

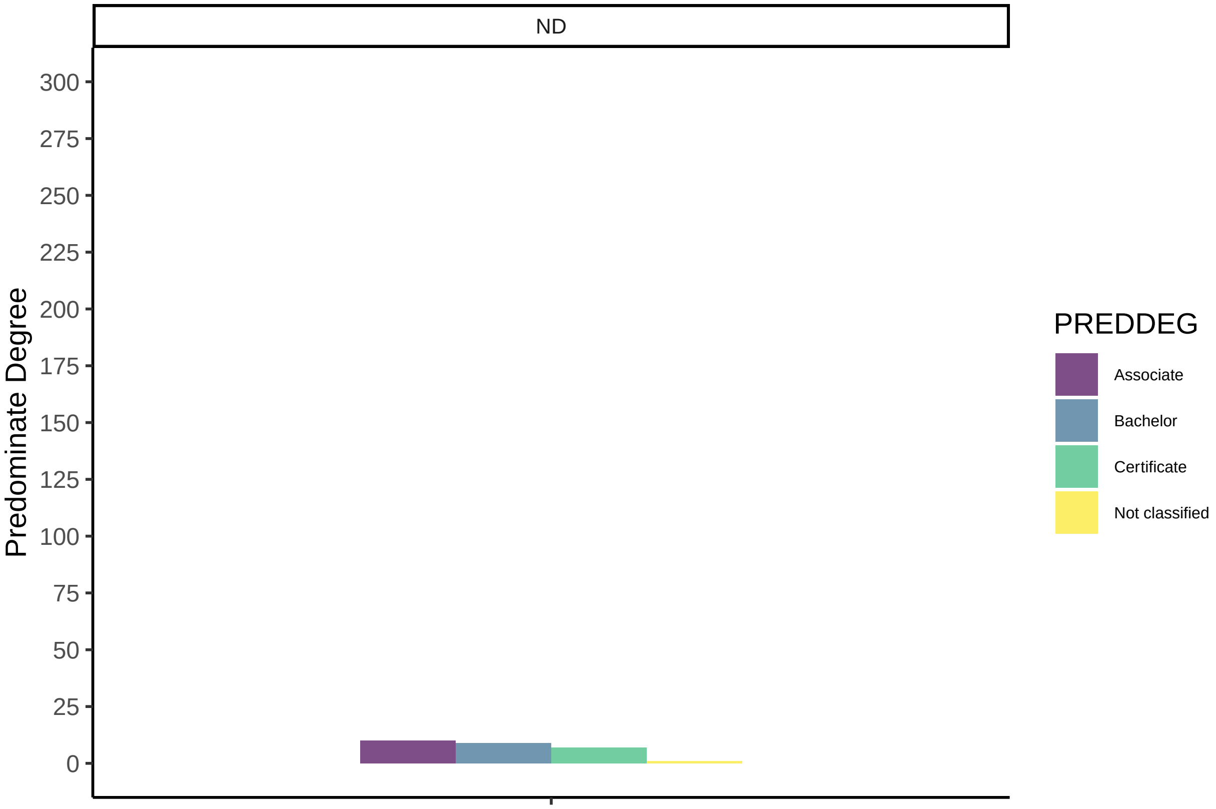

ME
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

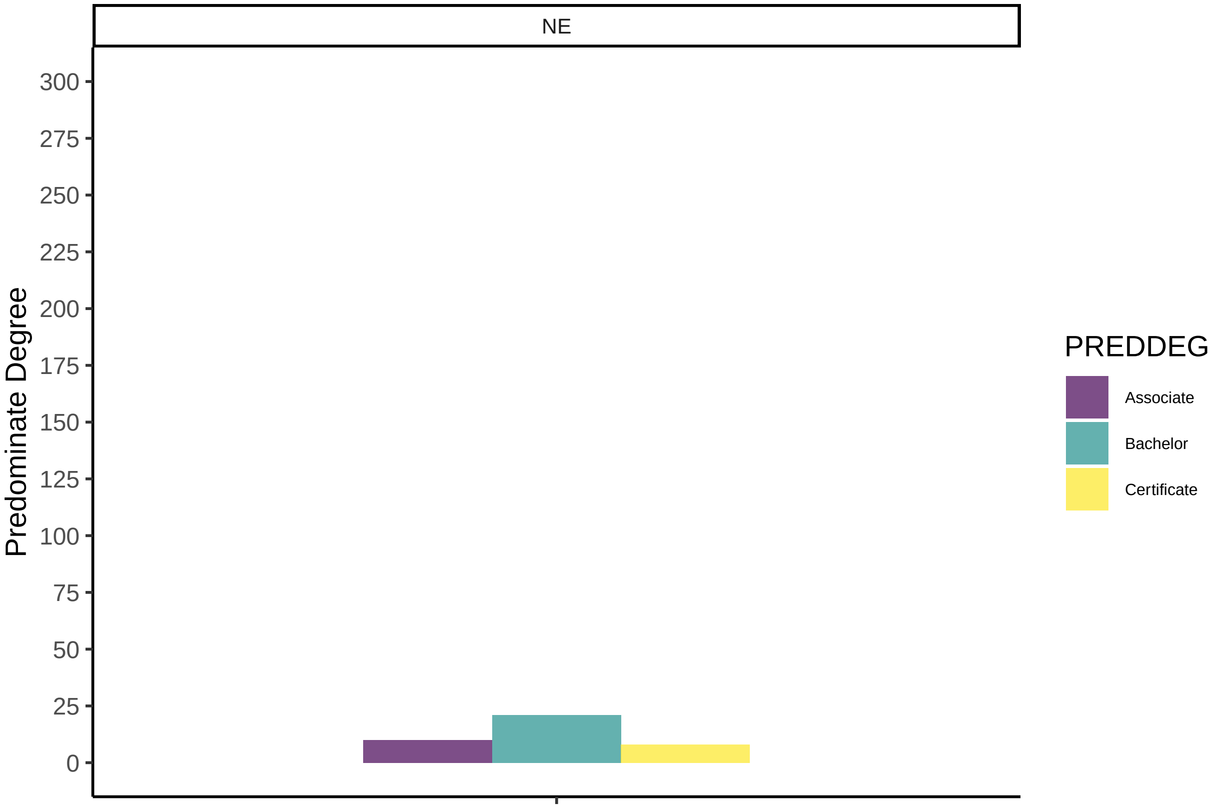

MI
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

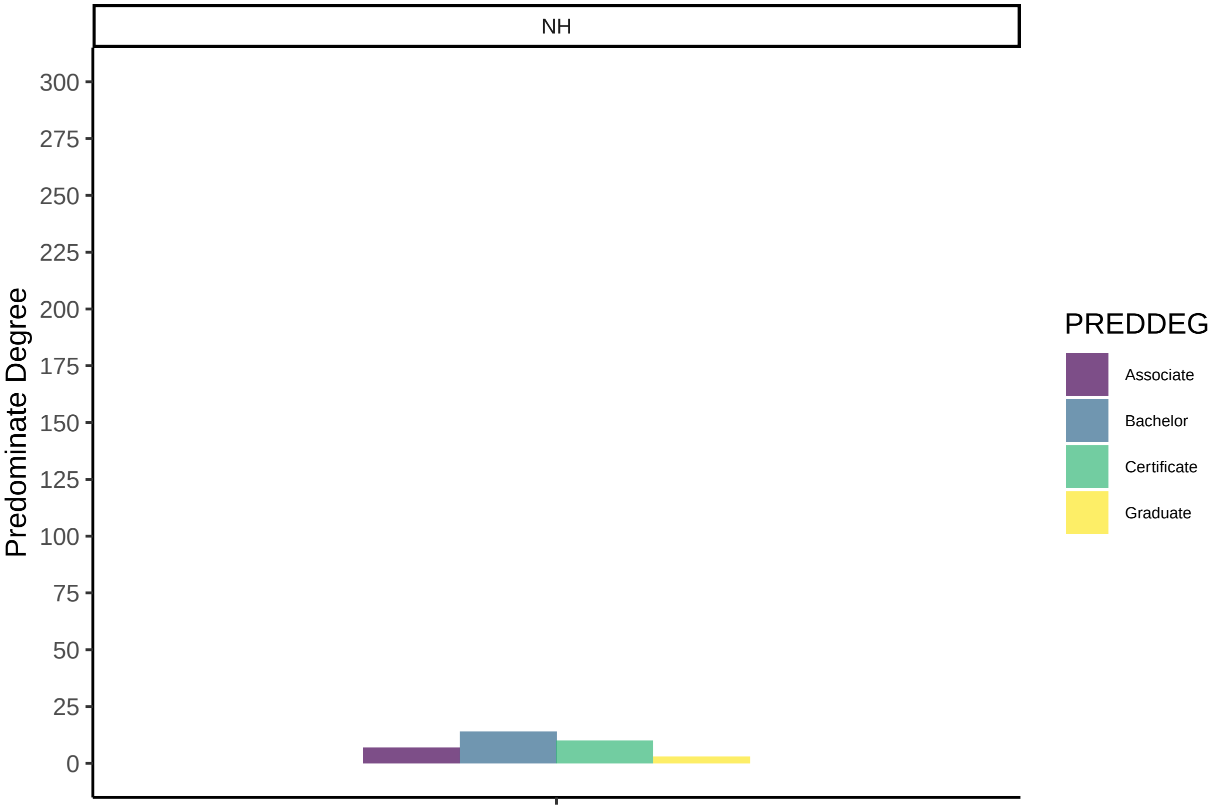

MN
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

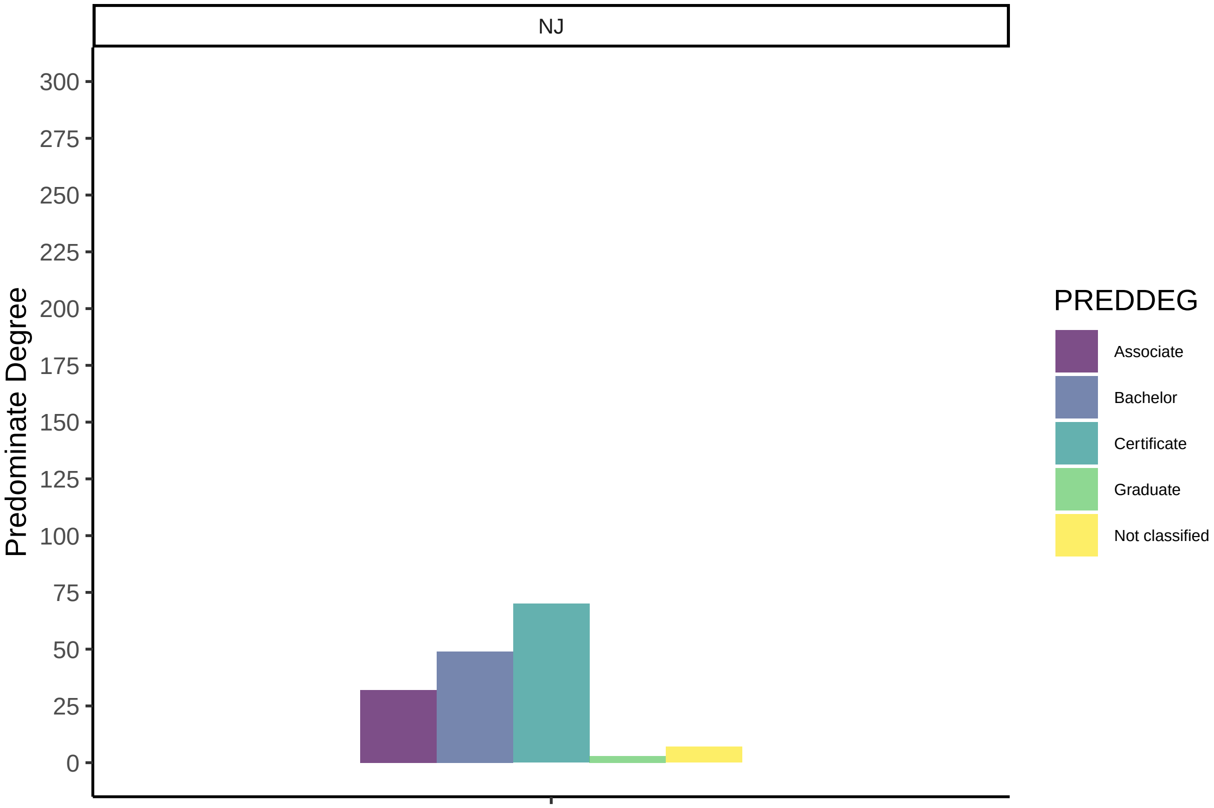

MO
MS
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

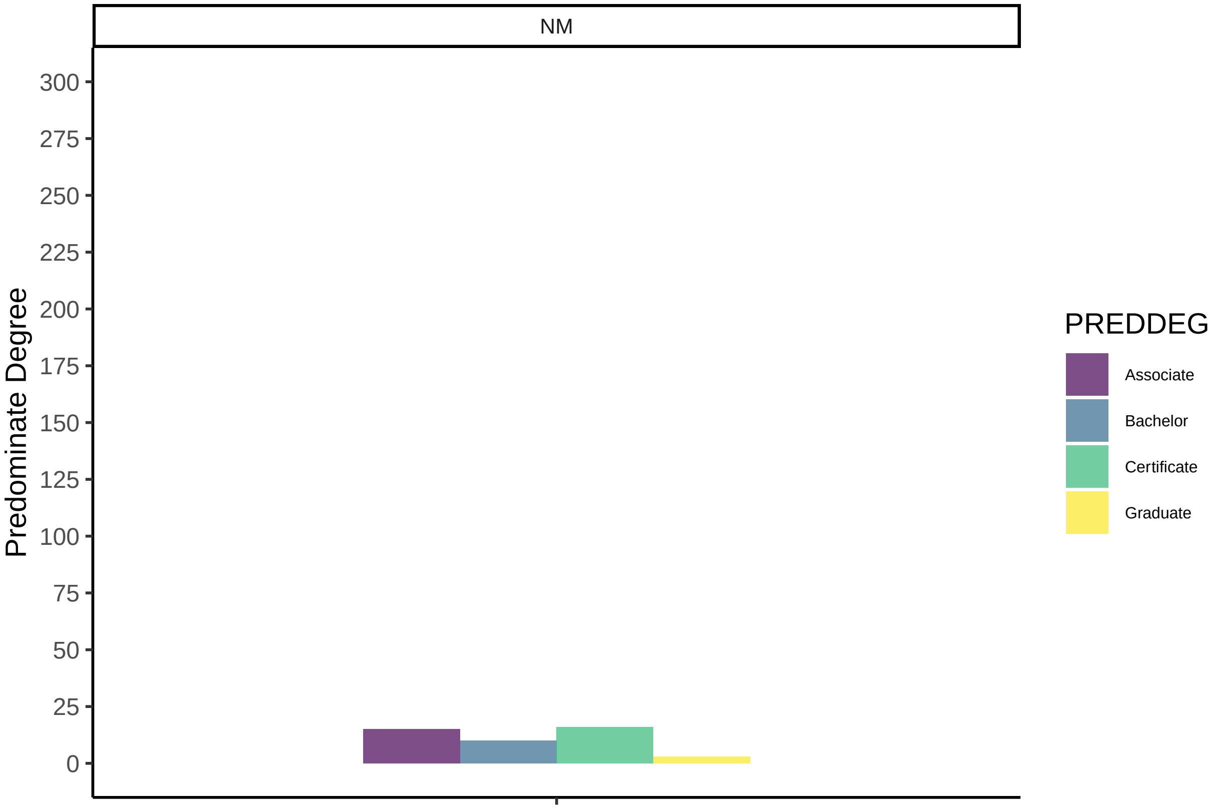

MT
300
275
250
225
200
175
150
125
100
75
50
25
0
Predominate Degree
PREDDEG
Associate
Bachelor
Certificate
Not classified
NC
300
275
250
225
200
175
150
125
100
75
50
25
0
Predominate Degree
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

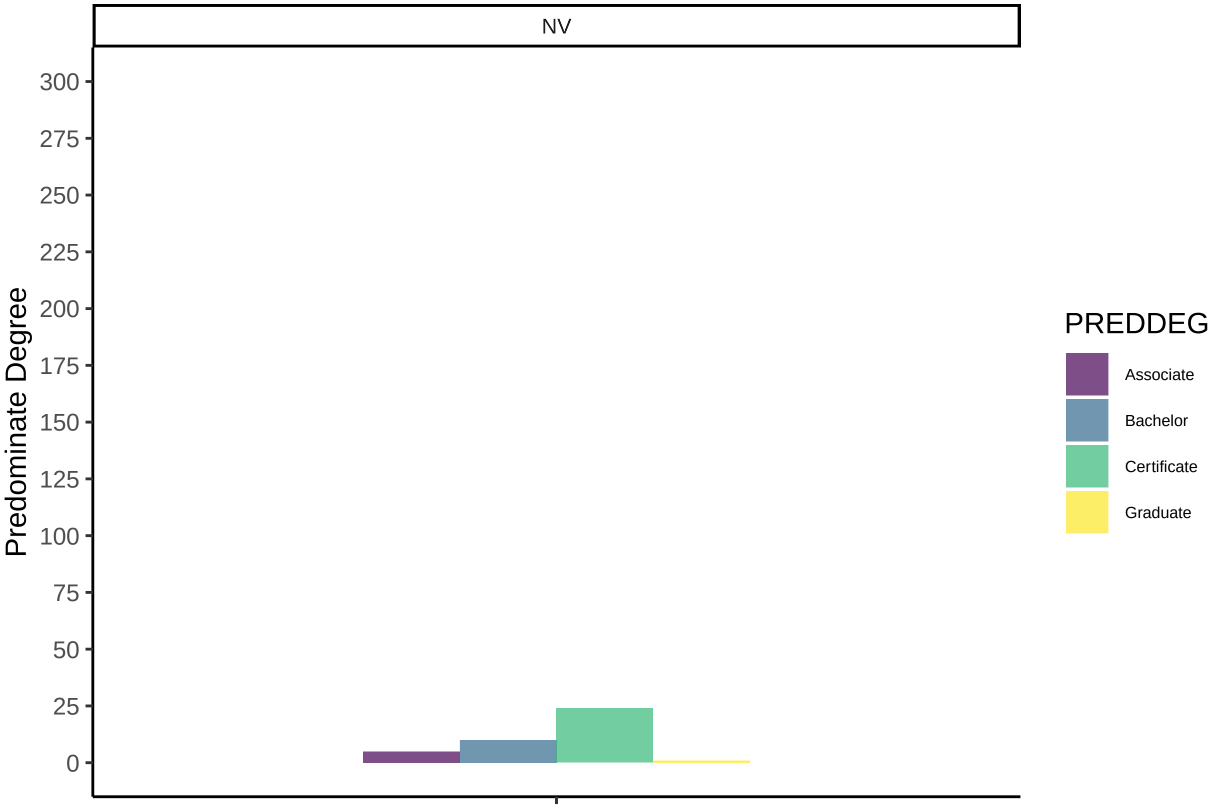

ND
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Not classified

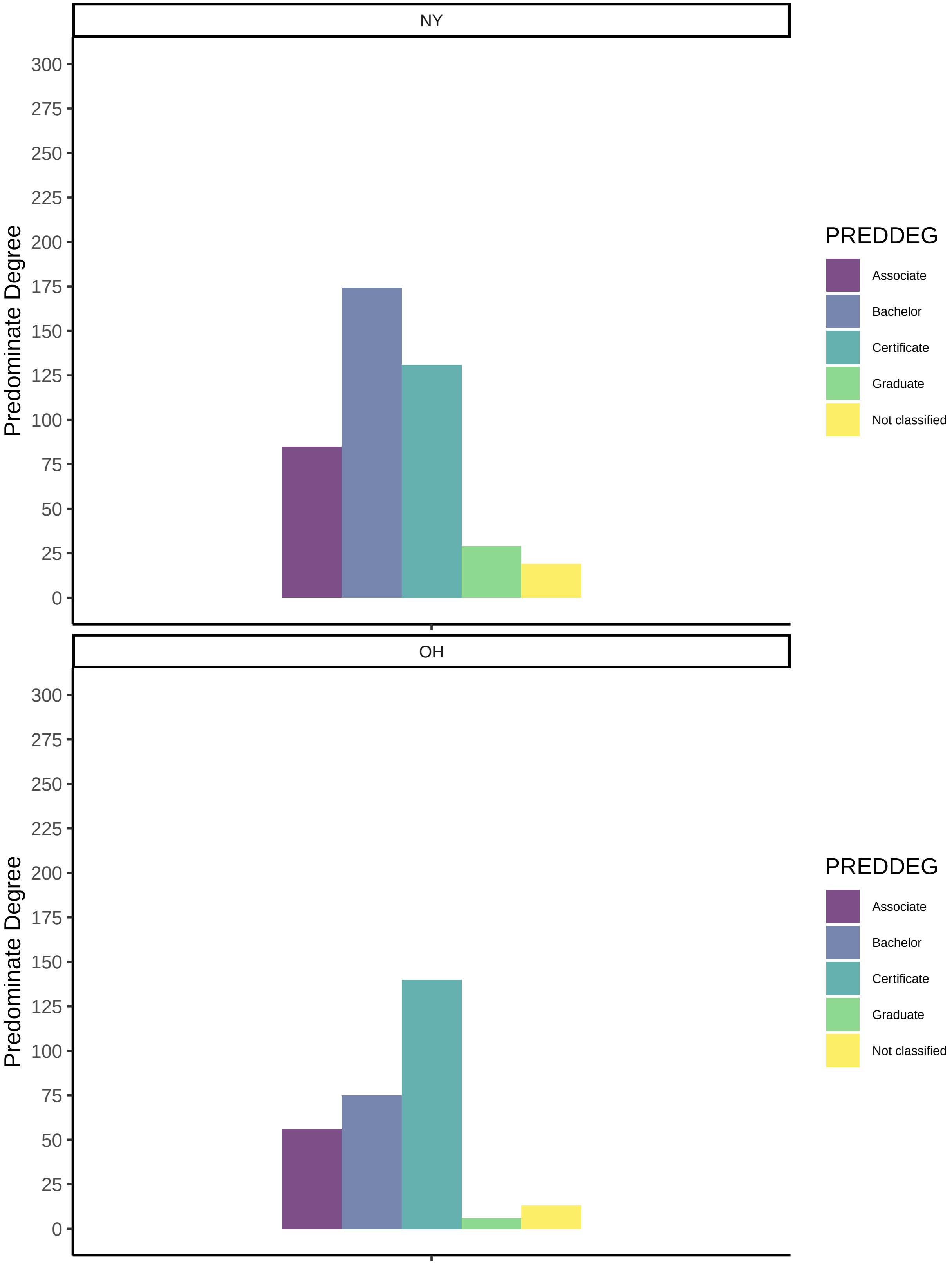

NE
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate

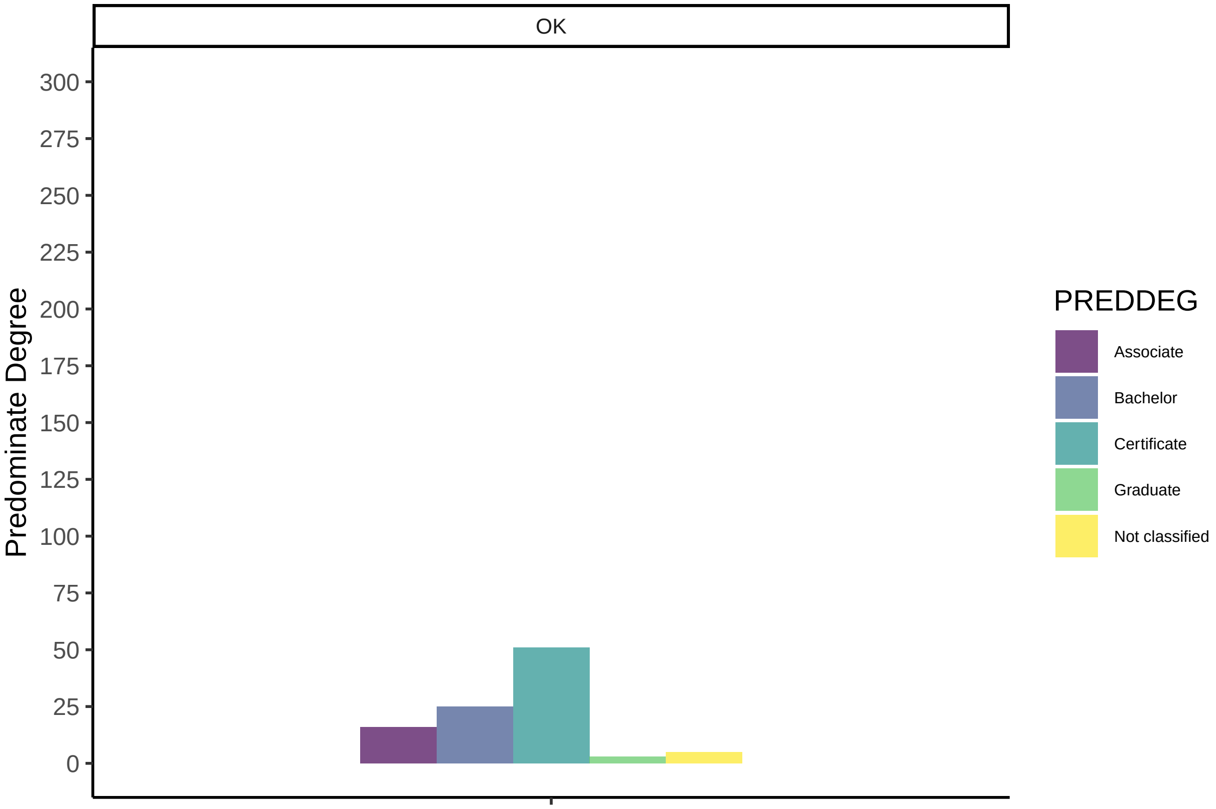

NH
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate

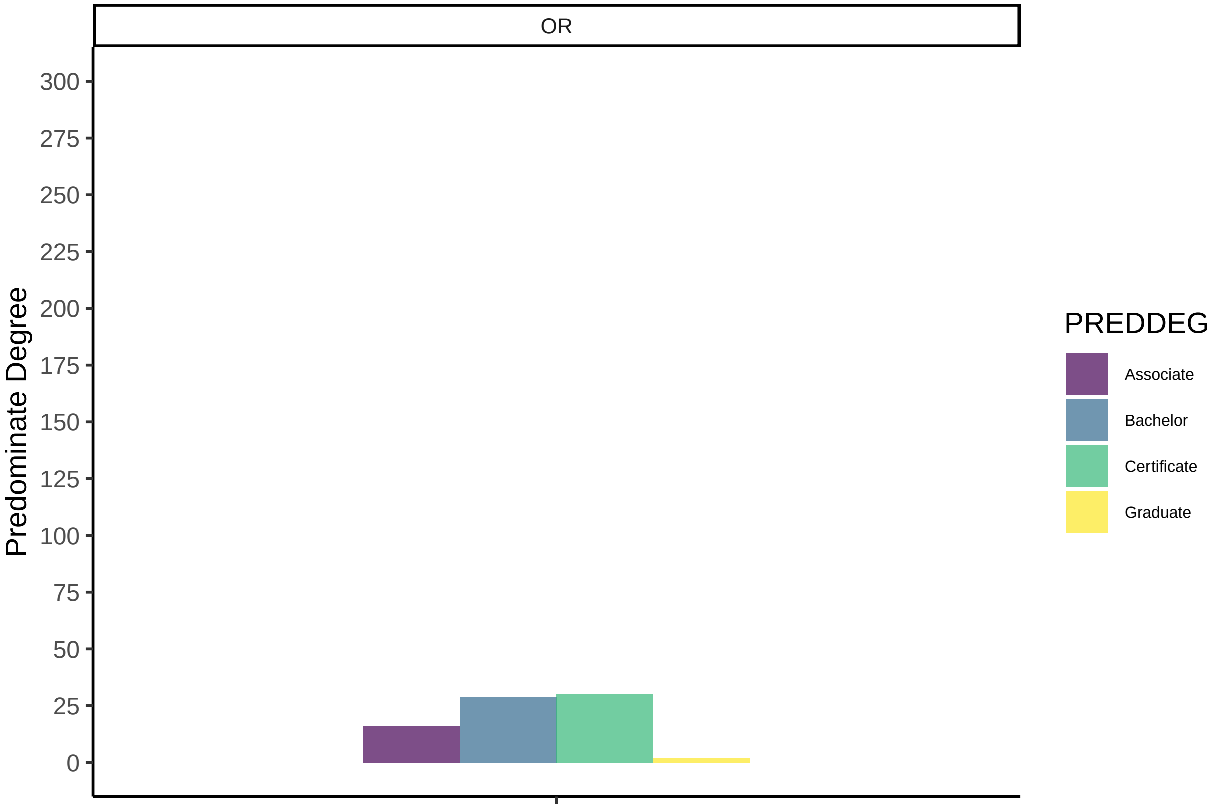

NJ
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

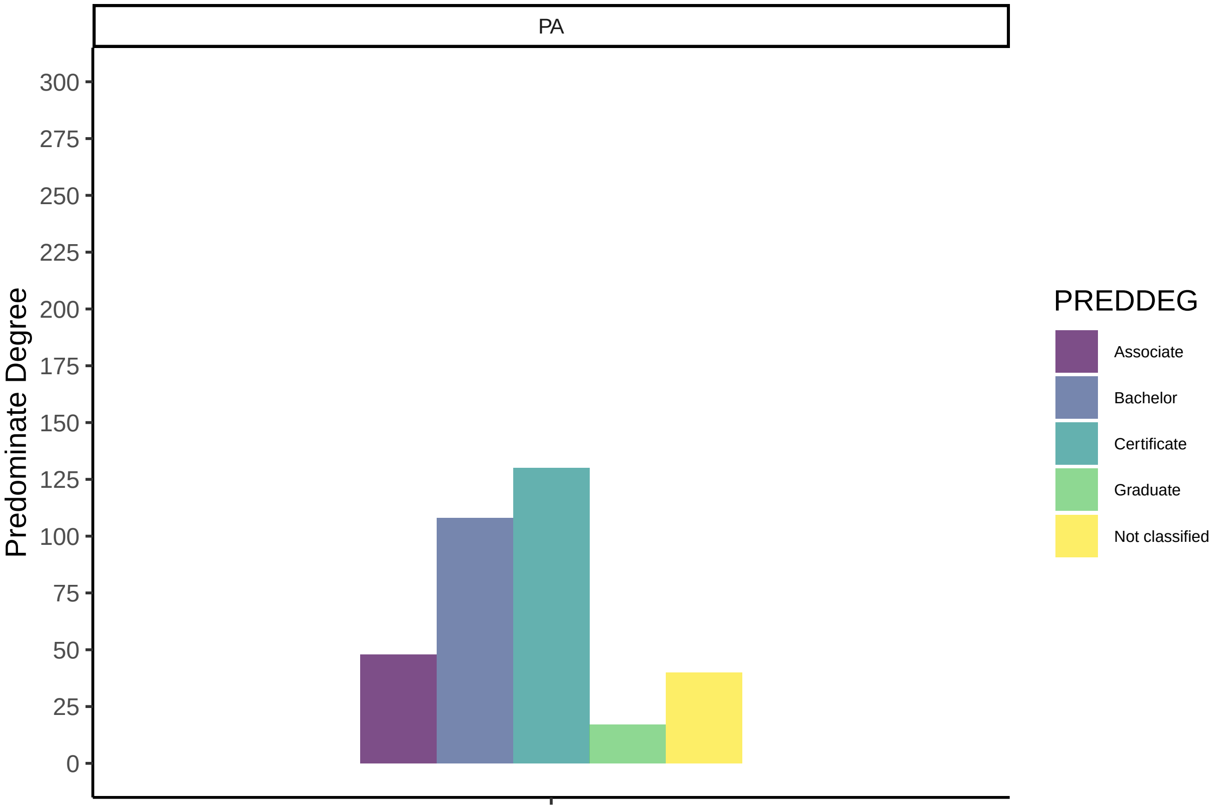

NM
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate

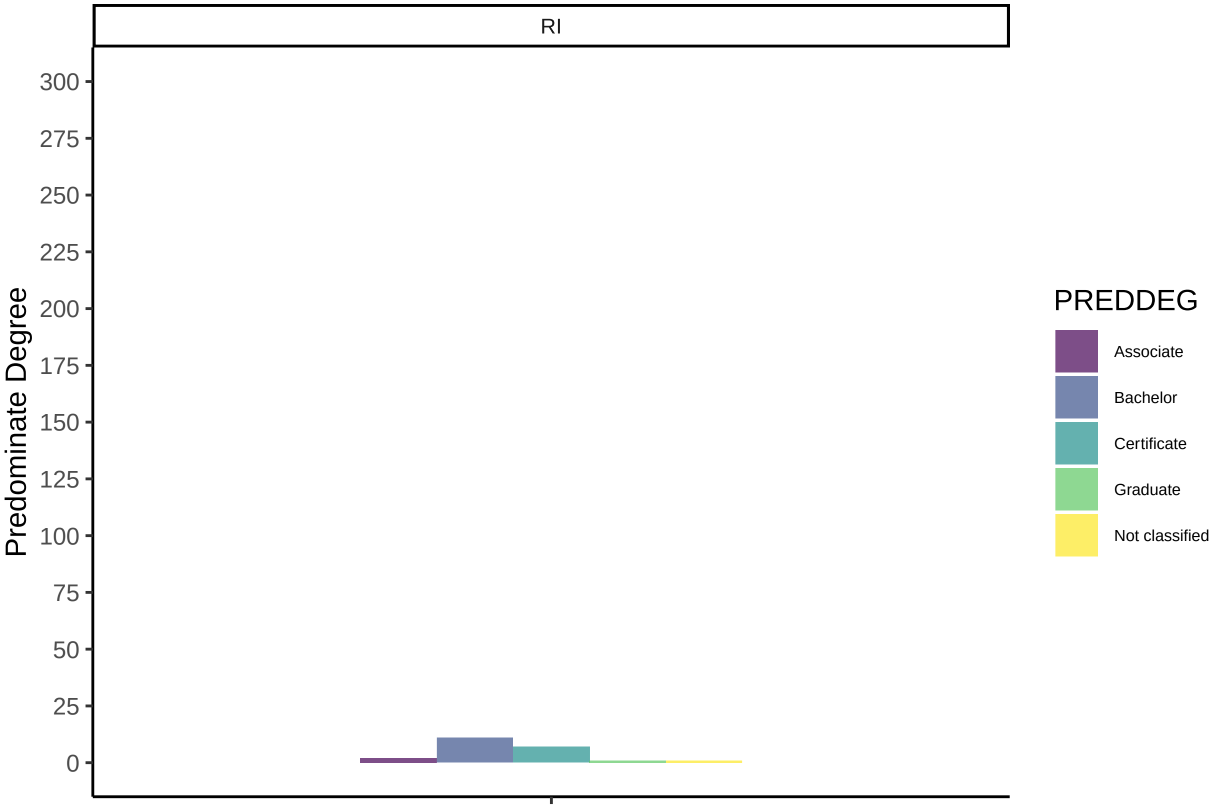

NV
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate

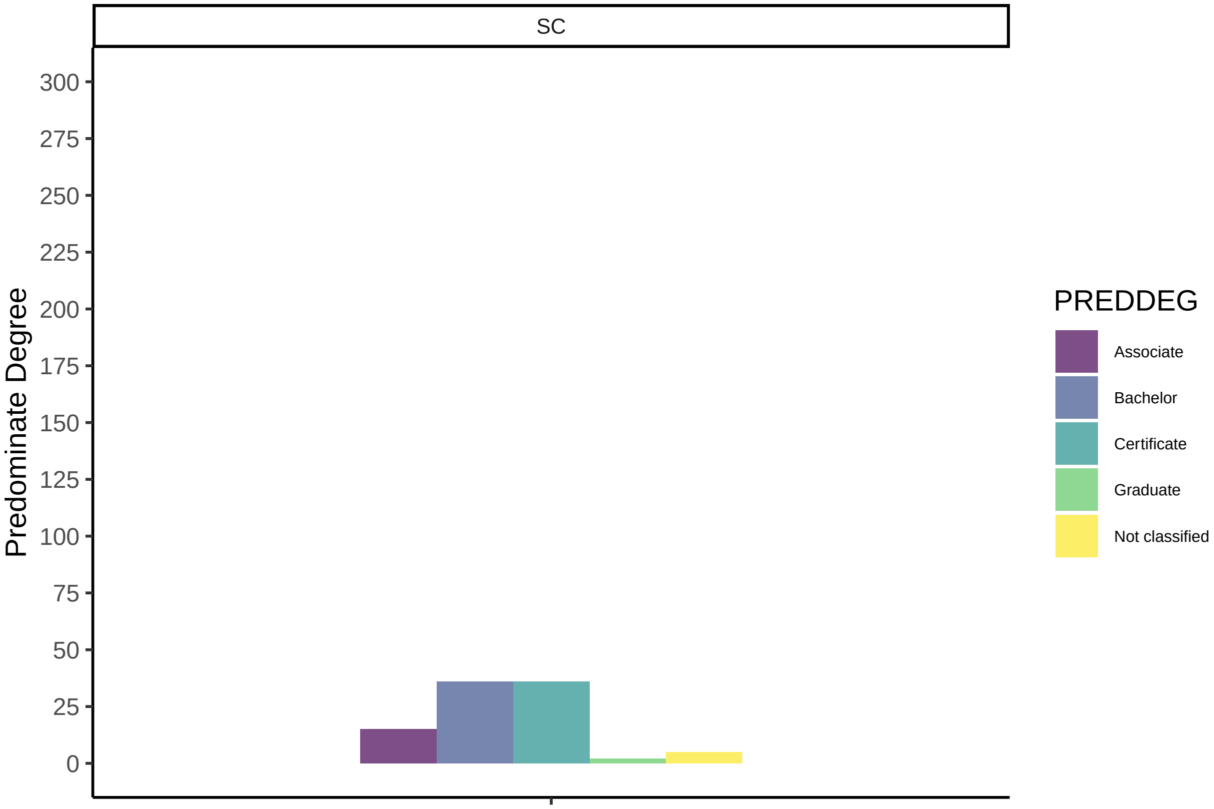

NY
OH
Predominate Degree
0
25
50
75
100
125
150
175
200
225
250
275
300
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

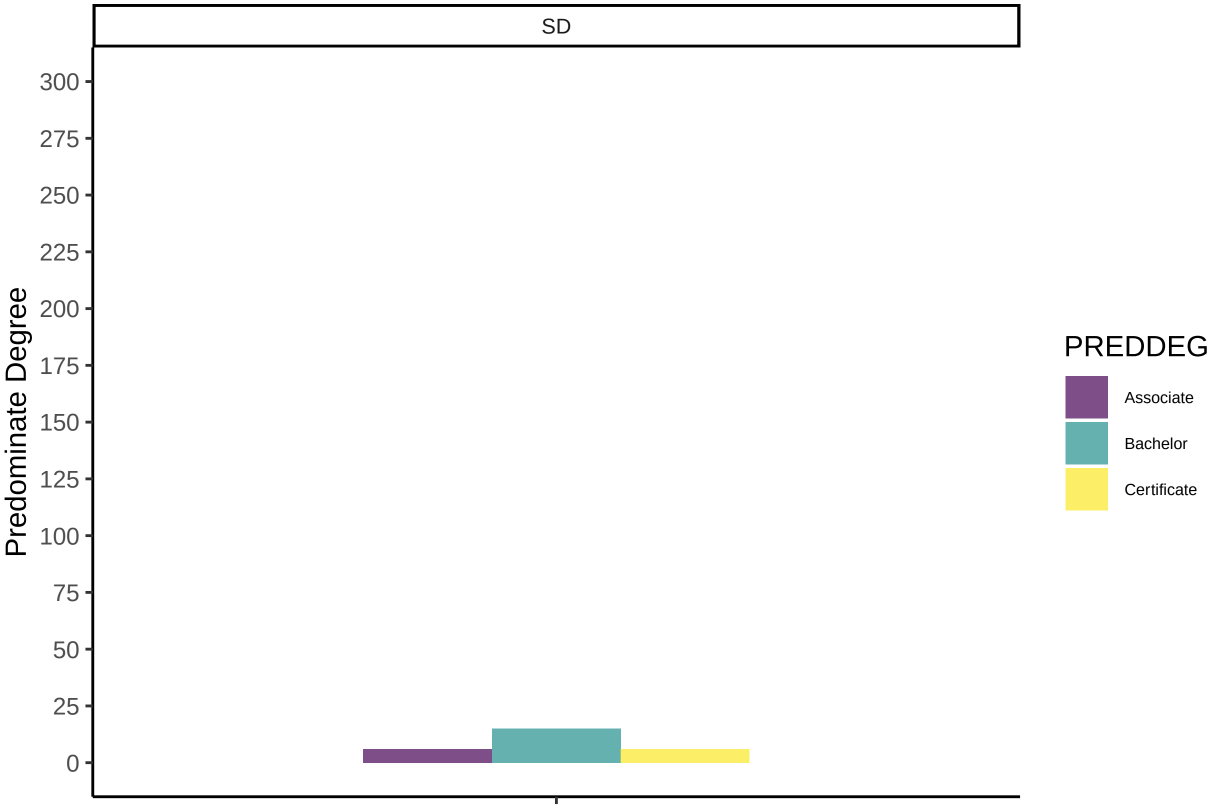

OK
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

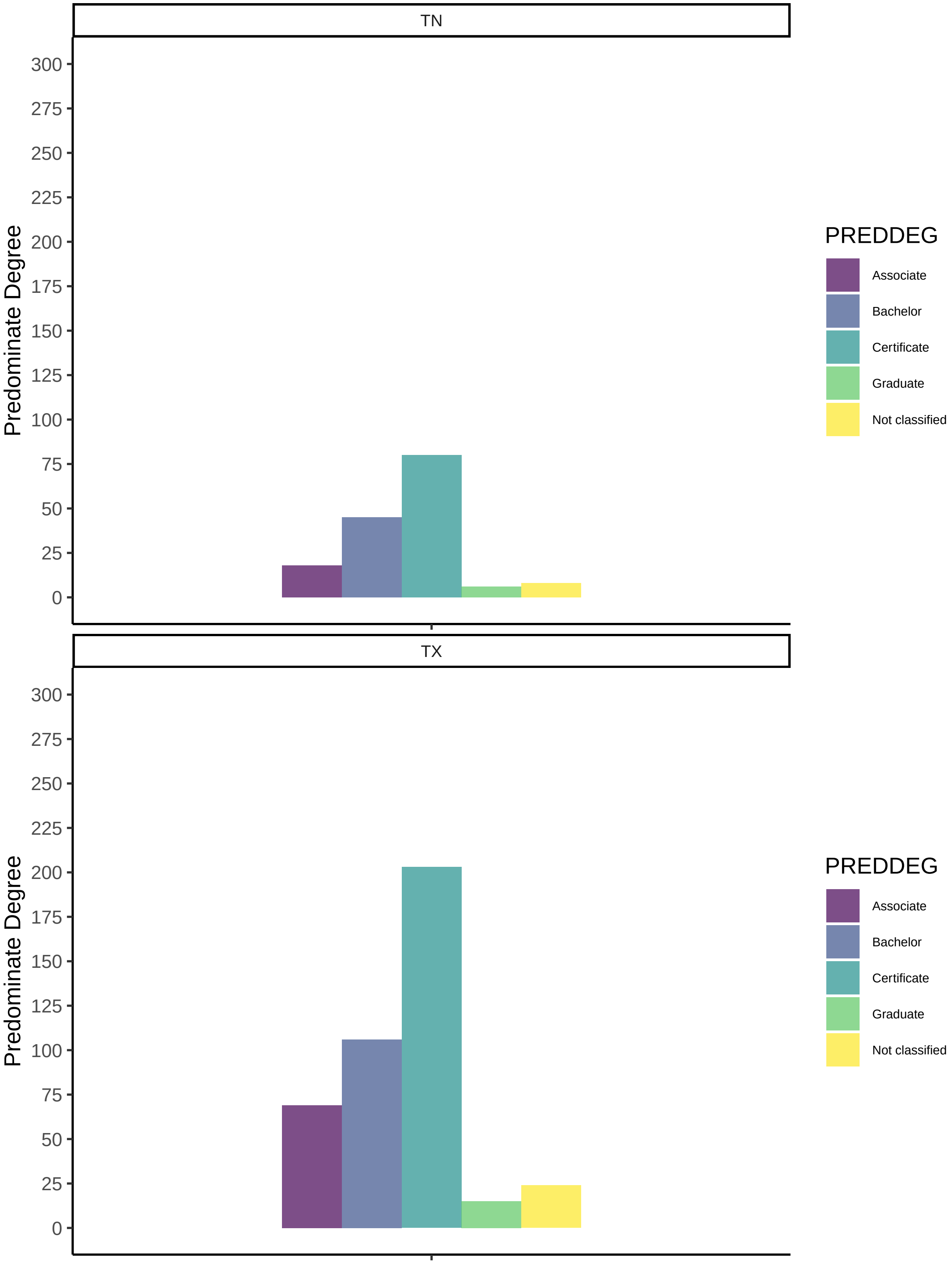

OR
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate

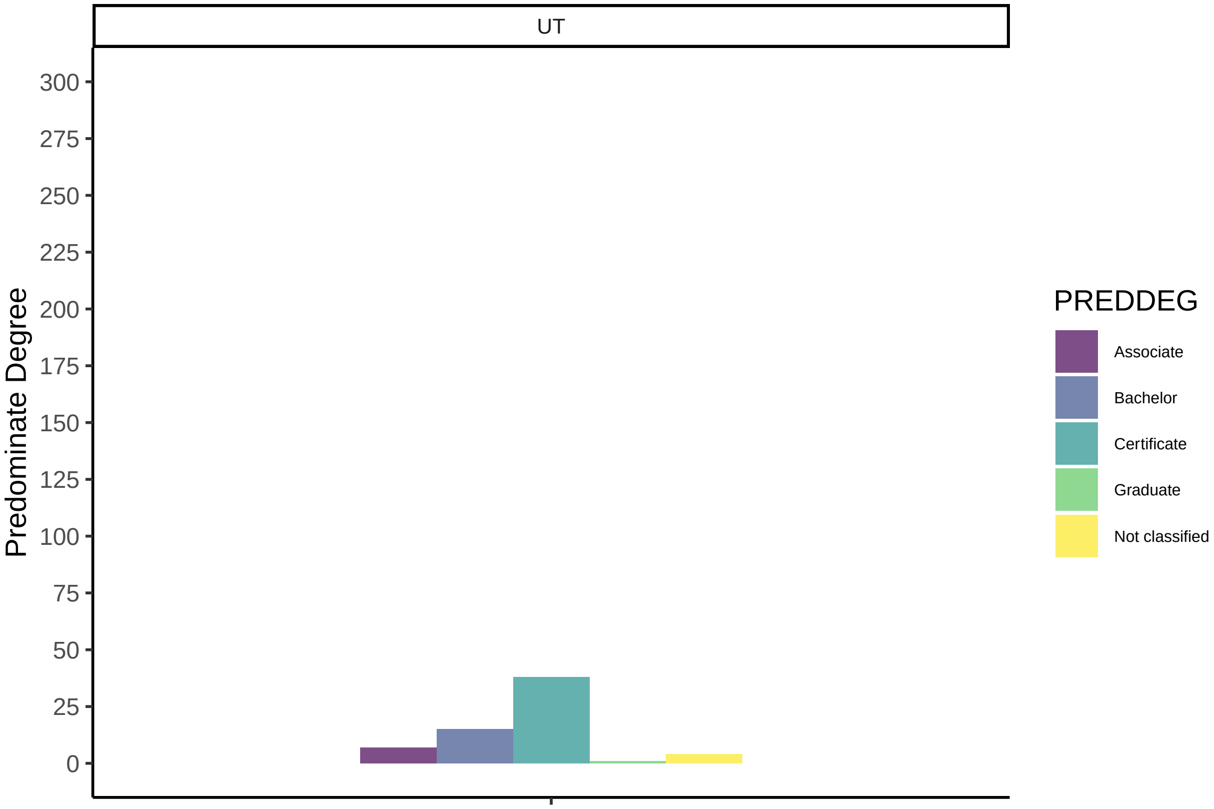

PA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

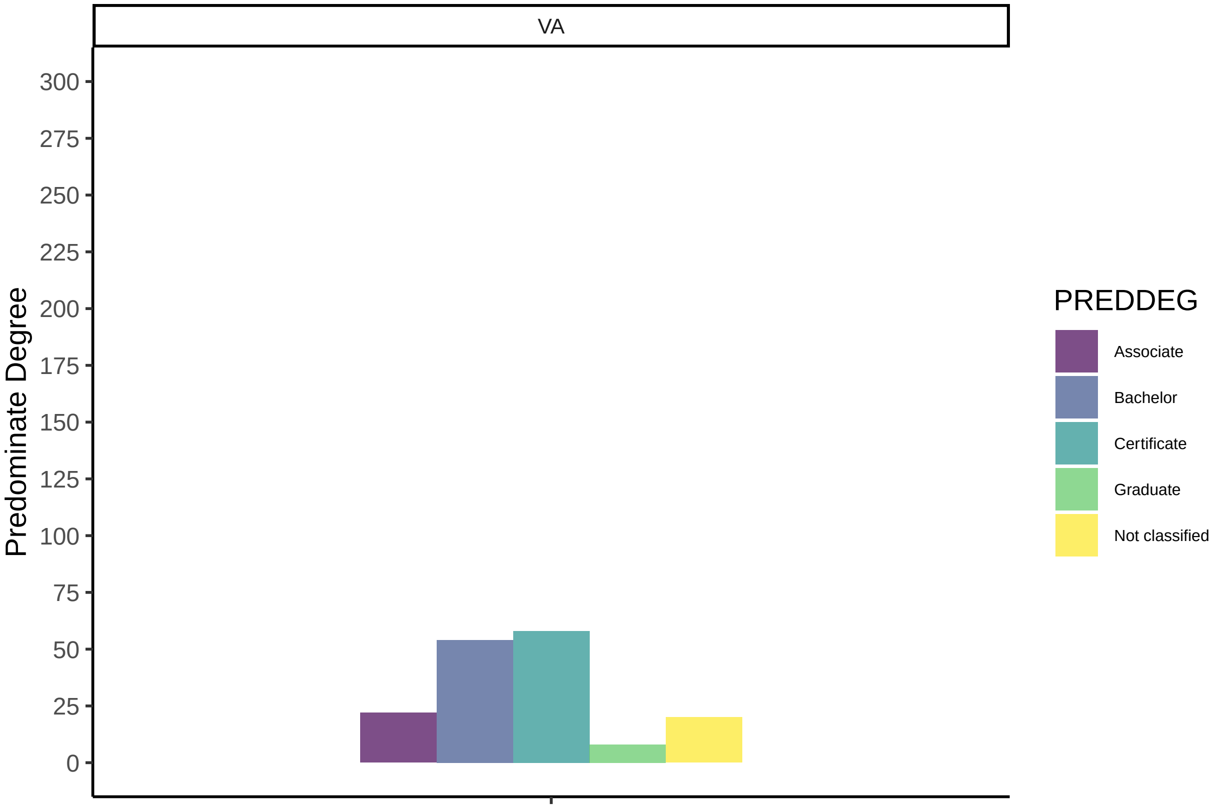

RI
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

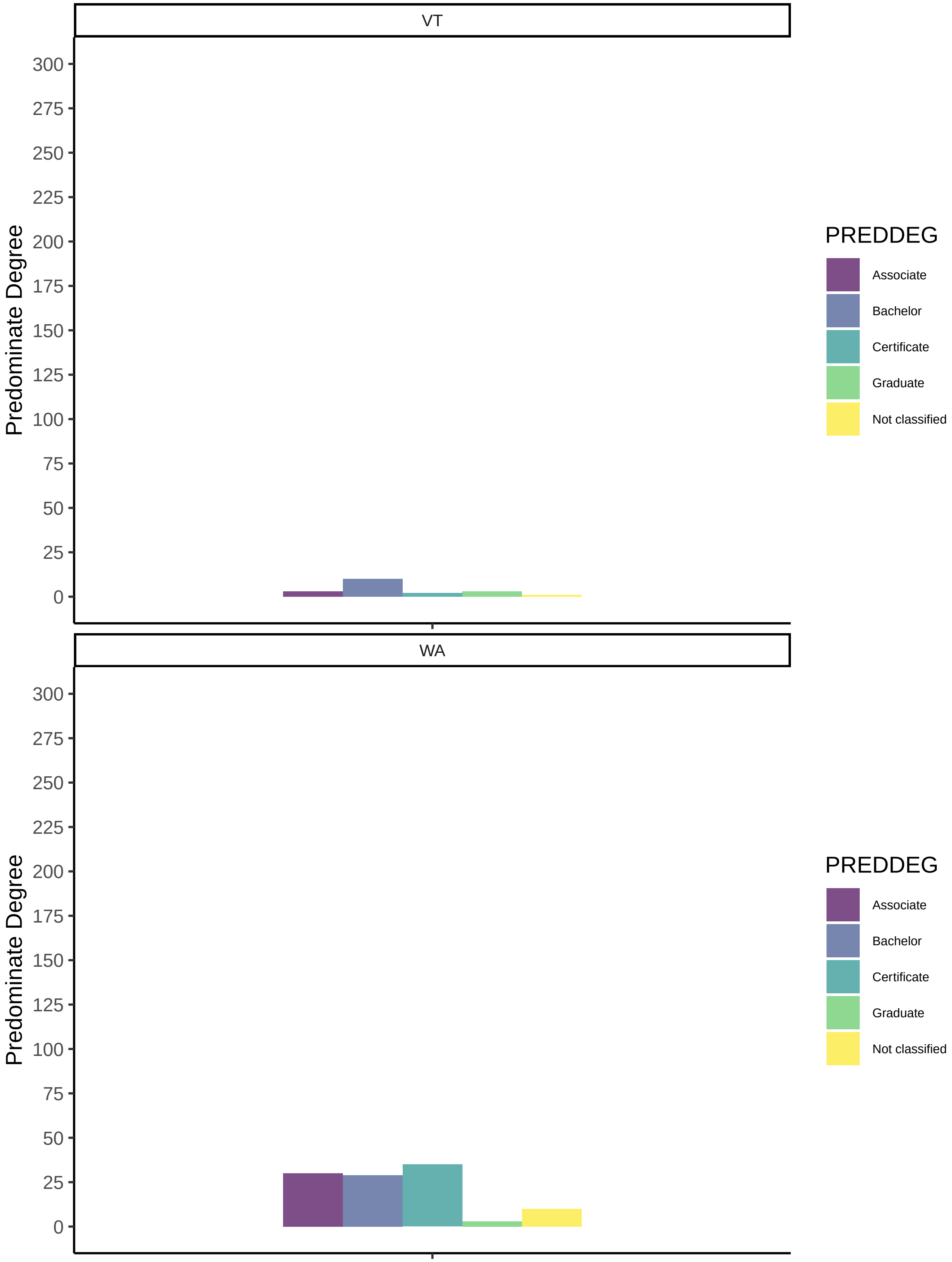

SC
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

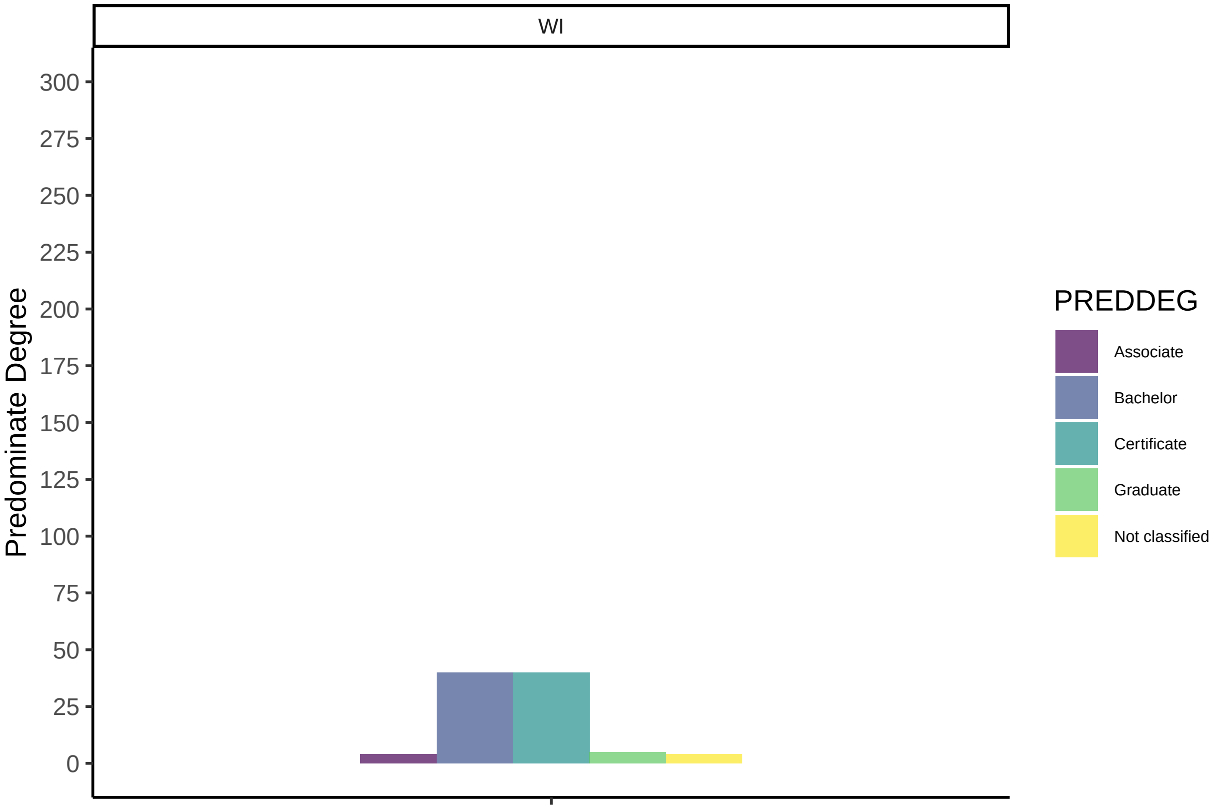

SD
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate

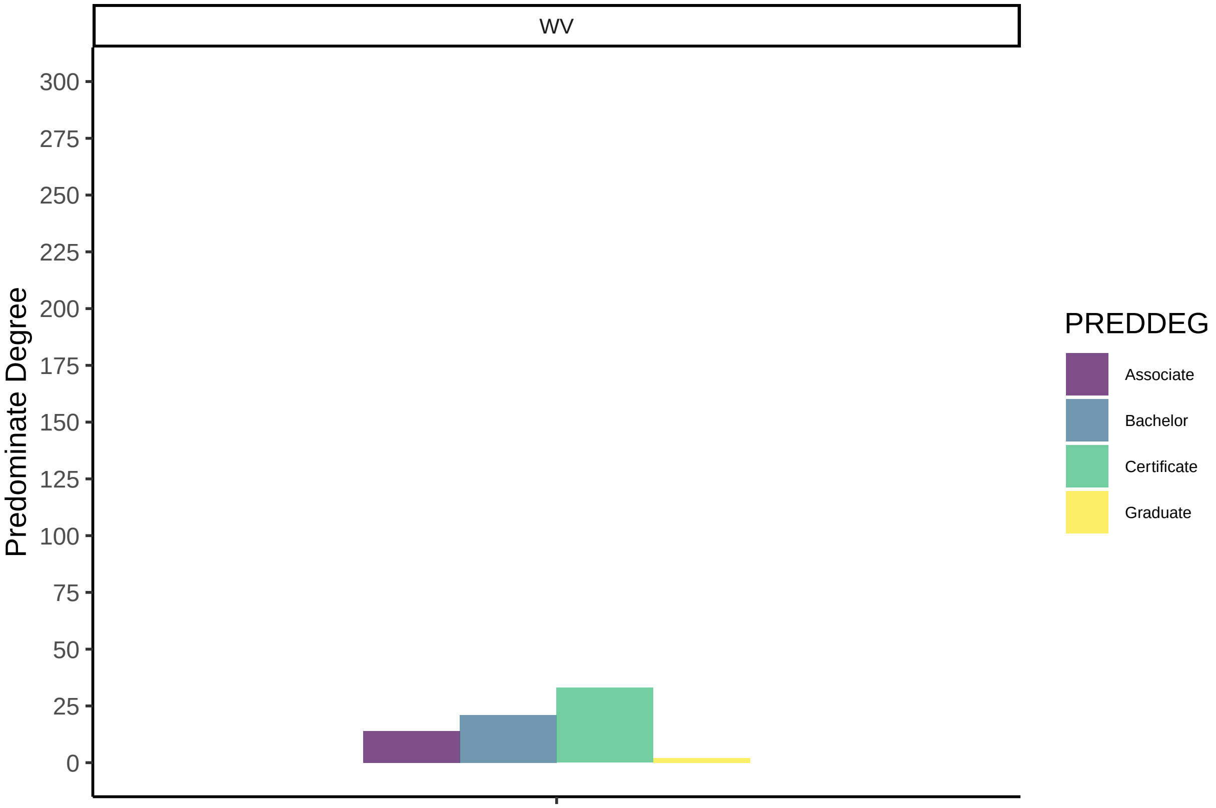

TN
TX
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

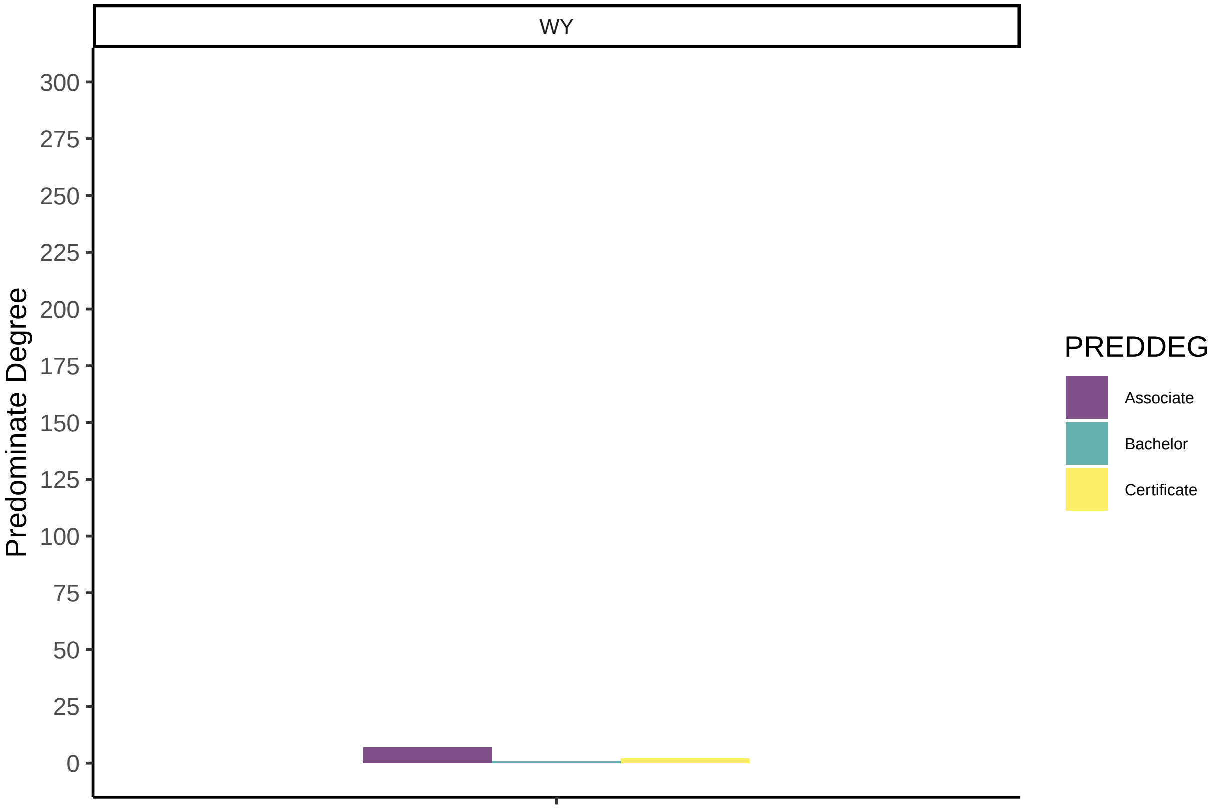

UT
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified


VA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified


VT
WA
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified


WI
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate
Not classified

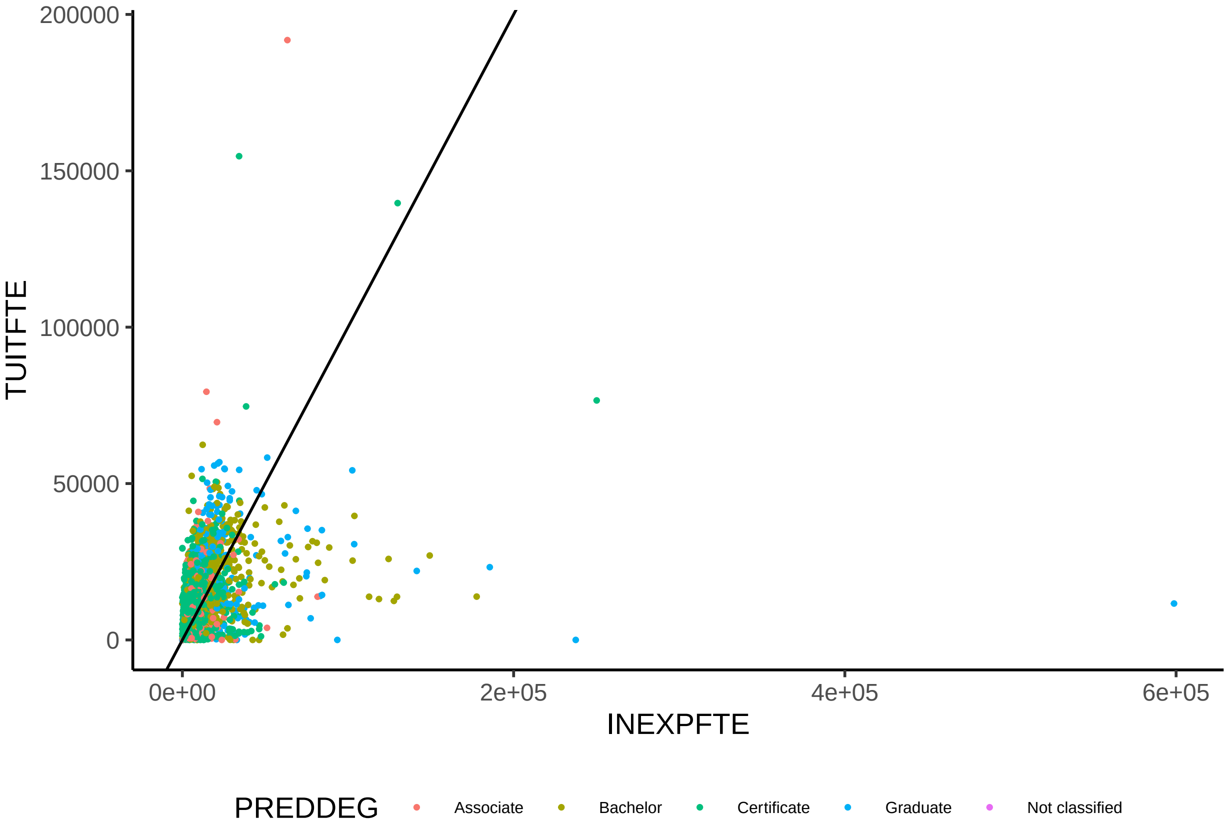

WV
Predominate Degree
300
275
250
225
200
175
150
125
100
75
50
25
0
PREDDEG
Associate
Bachelor
Certificate
Graduate



Scatter Plot

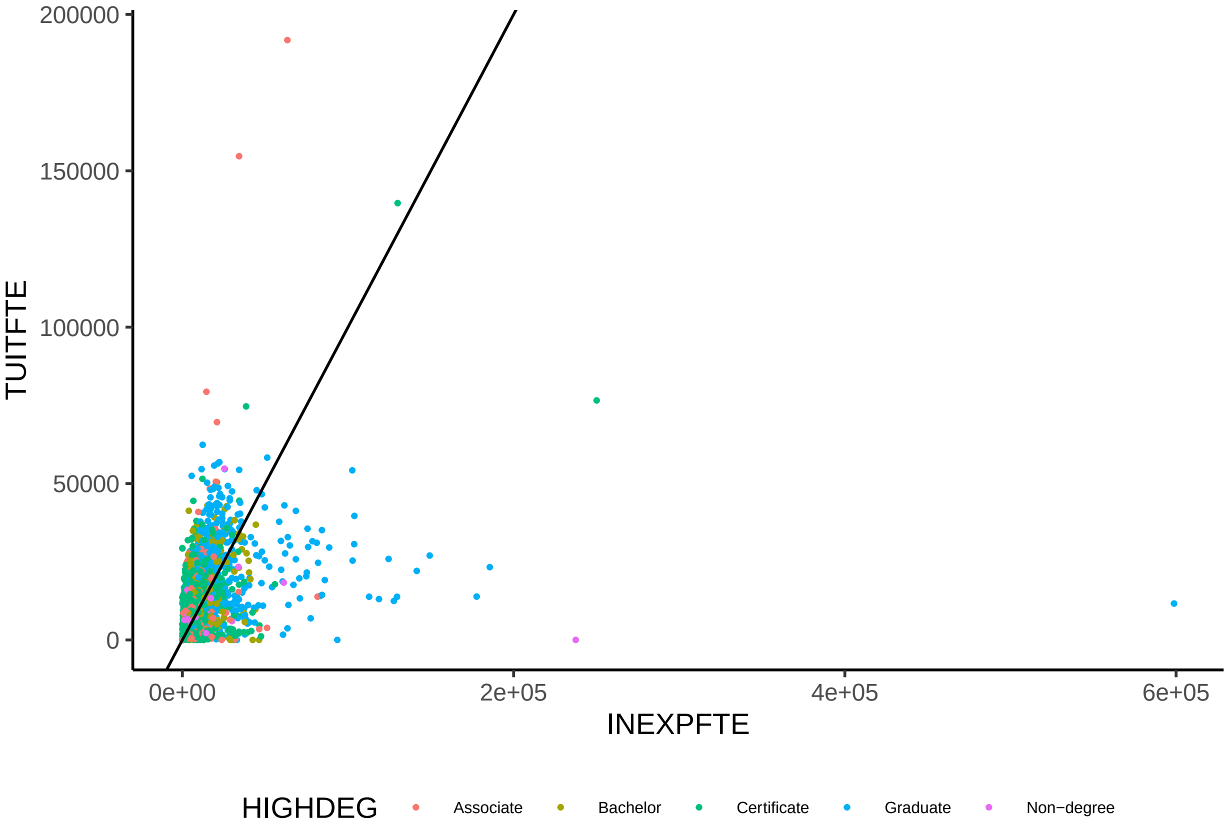
```
finance_df <- cbind(STABBR = filtered_df[,6], filtered_df[,14:15], CCBASIC = filtered_df[,22], TUITFTE =

finance_df %>%
  ggplot(aes(x=INEXPFTE, y=TUITFTE)) +
  geom_point(aes(color = PREDDEG),size = 0.5) +
  scale_fill_viridis(discrete = TRUE, alpha=0.6, option= 'viridis')+
  theme_classic() +
  theme(legend.text = element_text(size = 6), # font text of variables in legend
        legend.position = 'bottom')+
  geom_abline(method= 'lm', color = 'black')
```

```
## Warning in geom_abline(method = "lm", color = "black"): Ignoring unknown
## parameters: `method`

## Warning: Removed 519 rows containing missing values (`geom_point()`).
```



```
finance_df %>%
  ggplot(aes(x=INEXPFTE, y=TUITFTE)) +
  geom_point(aes(color = HIGHDEG),size = 0.5) +
  scale_fill_viridis(discrete = TRUE, alpha=0.6, option= 'viridis')+
  theme_classic() +
  theme(legend.text = element_text(size = 6), # font text of variables in legend
        legend.position = 'bottom')+
  geom_abline(method= 'lm', color = 'black')
```

```
## Warning in geom_abline(method = "lm", color = "black"): Ignoring unknown parameters: `method`
## Removed 519 rows containing missing values (`geom_point()`).
```

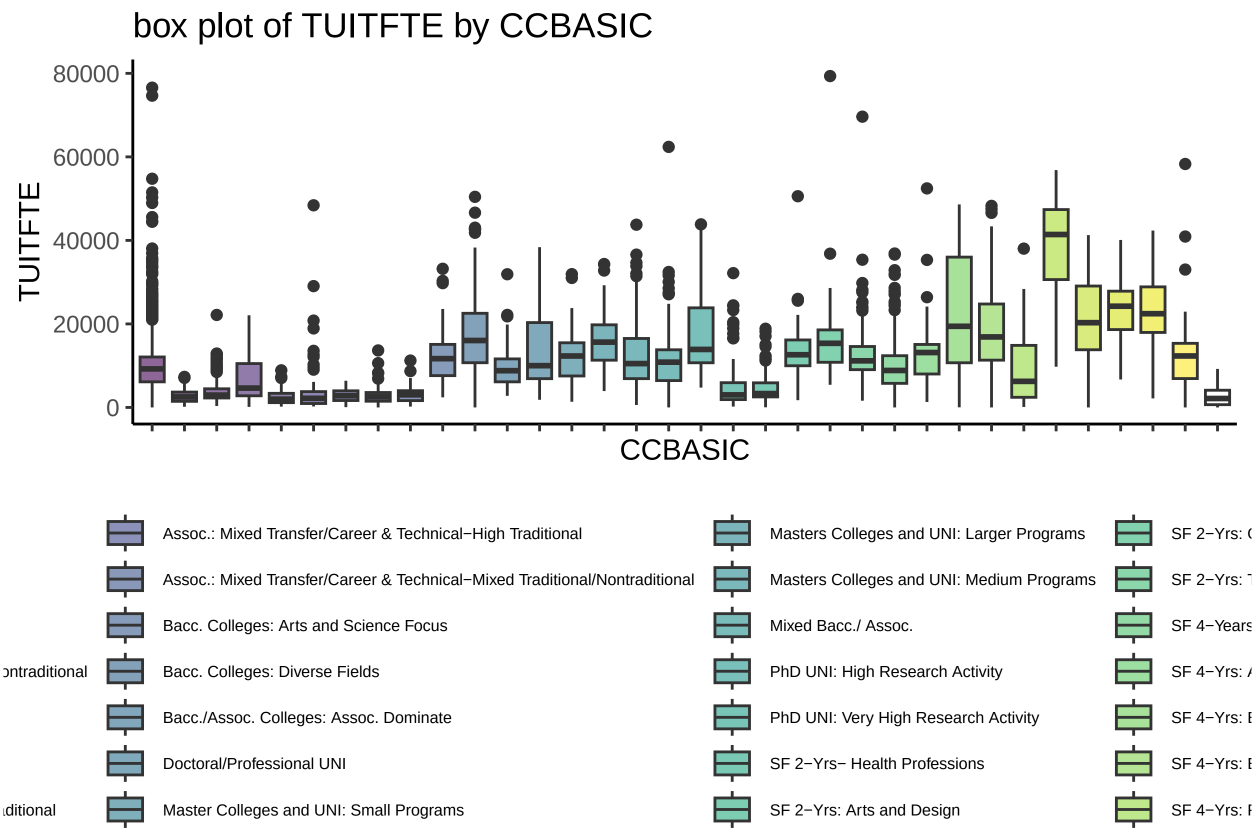


ANALYSIS

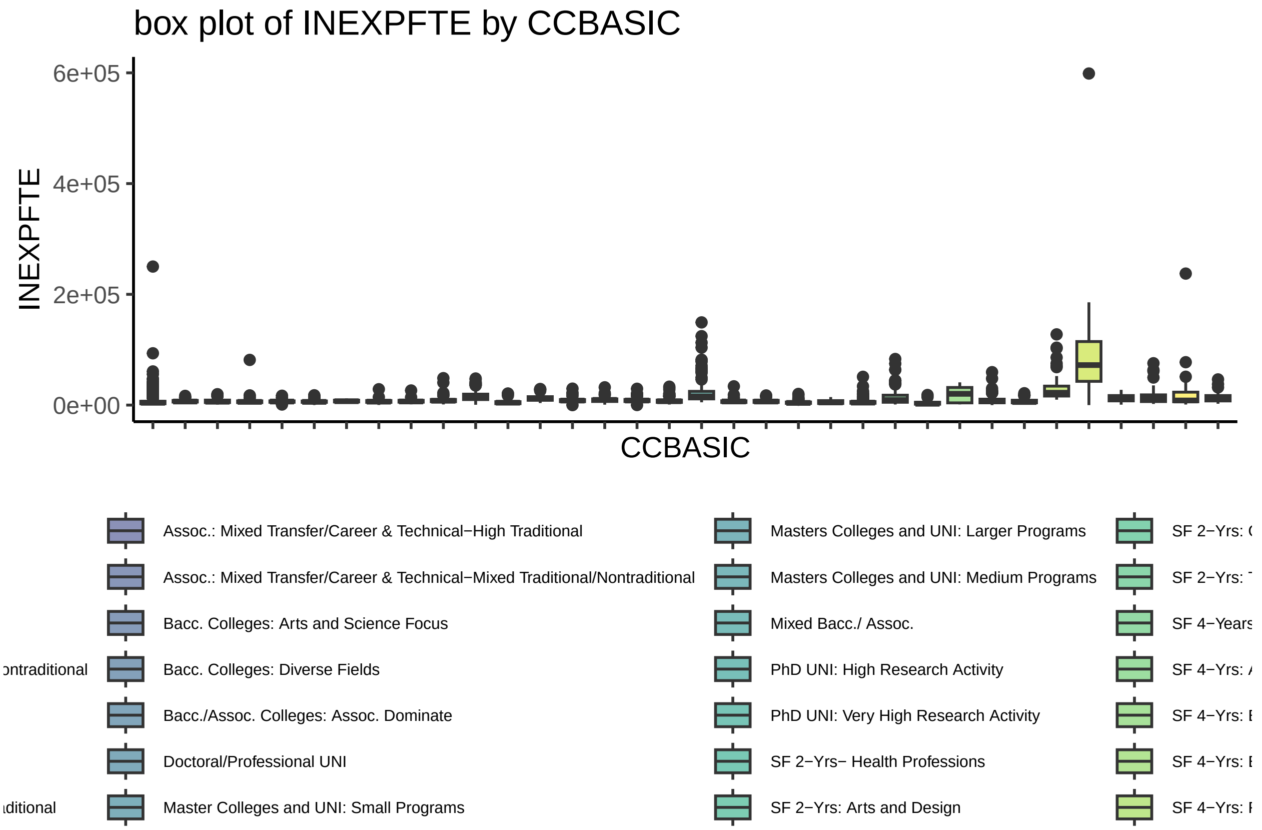
A majority of institution's instructional spending in relation to its tuition revenue is cluster around the tuition revenue < 50000 and instructional expenditures < 200000. The range of the associate data points is highest on the y-axis and the range of graduate insitutions is highest on the x-axis. This could suggest, that graduate institutions are more likely to spend more in instructional expenses in comparison to their tuition revenue. While associate insitutions are more likely to spend less in instructional expenses in comparsion to their tuition revenue.

```
box_plot <- function(df, money, category){
  summary_data <- summary(df$TUITFTE)

  max_height_graph <- (summary_data['Max.']+summary_data['3rd Qu.'])/2
  df <- df %>%
    filter(TUITFTE <= max_height_graph)

  plot <- df %>%
    ggplot(aes(x=!!sym(category), y= !!sym(money), fill = !!sym(category))) +
    geom_boxplot(position = 'dodge') +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option= 'viridis')+
    theme_classic() +
    theme(legend.text = element_text(size = 6), # font text of variables in legend
        legend.position = 'bottom')+
    theme(axis.text.x = element_blank())+
    labs(title = sprintf('box plot of %s by %s', money, category))

  print(plot)
}
```

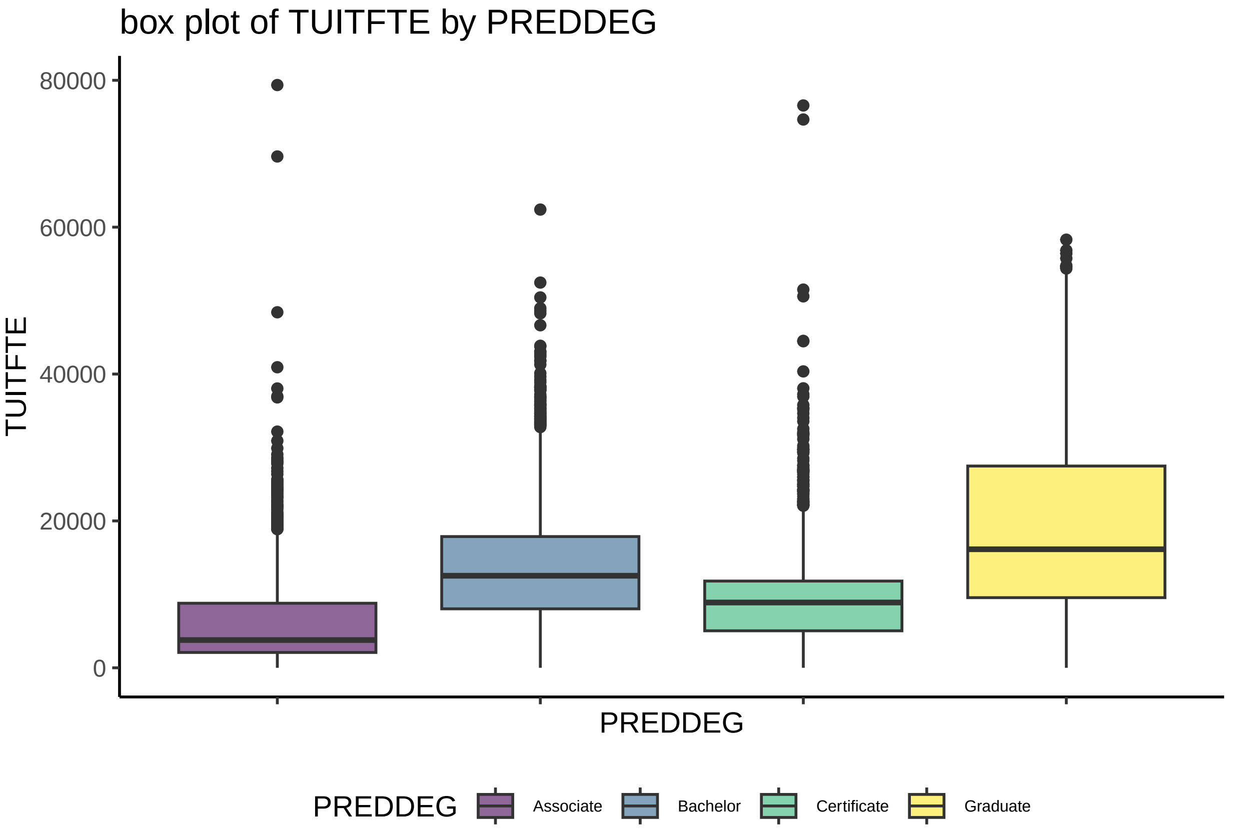
```
violin_plot <- function(df, money, category){
  summary_data <- summary(df$TUITFTE)

  max_height_graph <- (summary_data['Max.']+summary_data['3rd Qu.'])/2
  df <- df %>%
    filter(TUITFTE <= max_height_graph)
  plot <- df %>%
    ggplot(aes(x=!!sym(category), y= !!sym(money), fill = !!sym(category))) +
    geom_violin(position = 'dodge') +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option= 'viridis')+
    theme_classic() +
    theme(legend.text = element_text(size = 6), # font text of variables in legend
        legend.position = 'bottom')+
    theme(axis.text.x = element_blank())+
    labs(title = sprintf('box plot of %s by %s', money, category))

  print(plot)
}

cc_rev_box <- box_plot(finance_df, 'TUITFTE', 'CCBASIC')
```



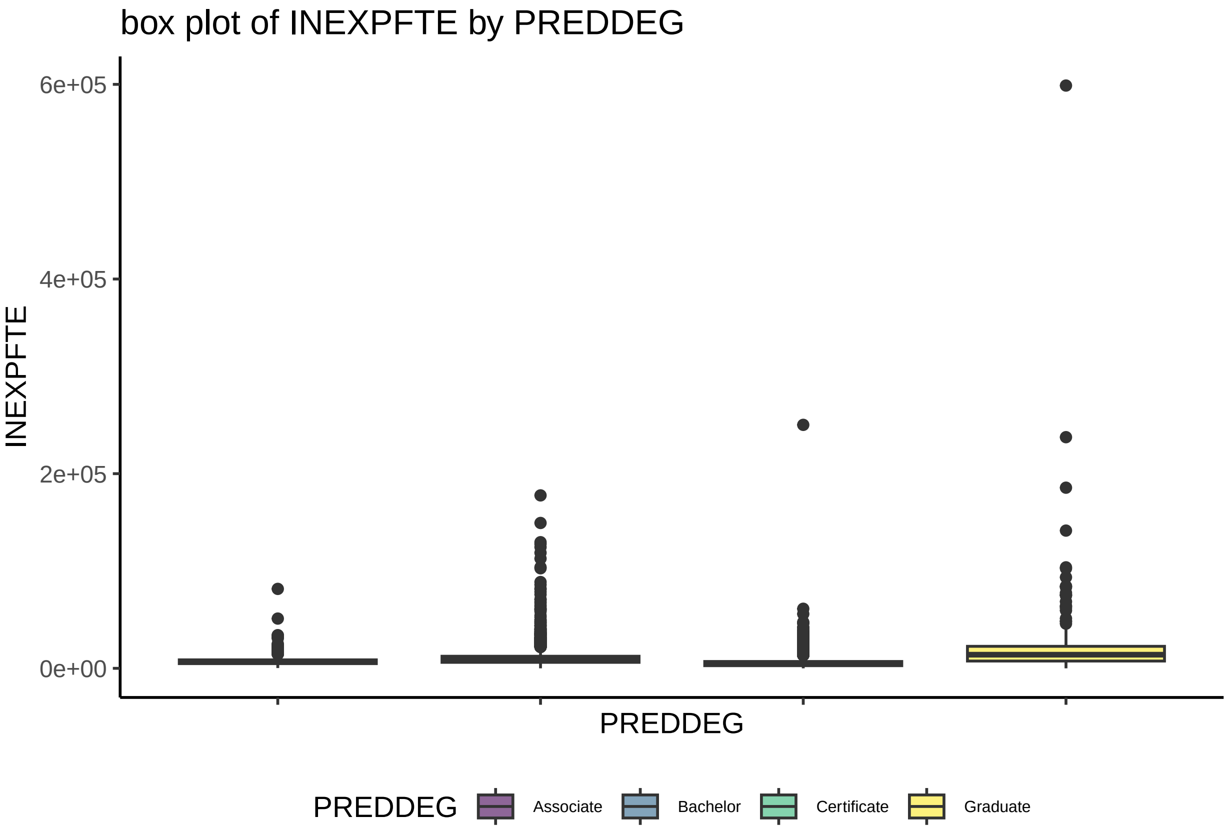
```
cc_exp_box <- box_plot(finance_df, 'INEXPFTE', 'CCBASIC')
```



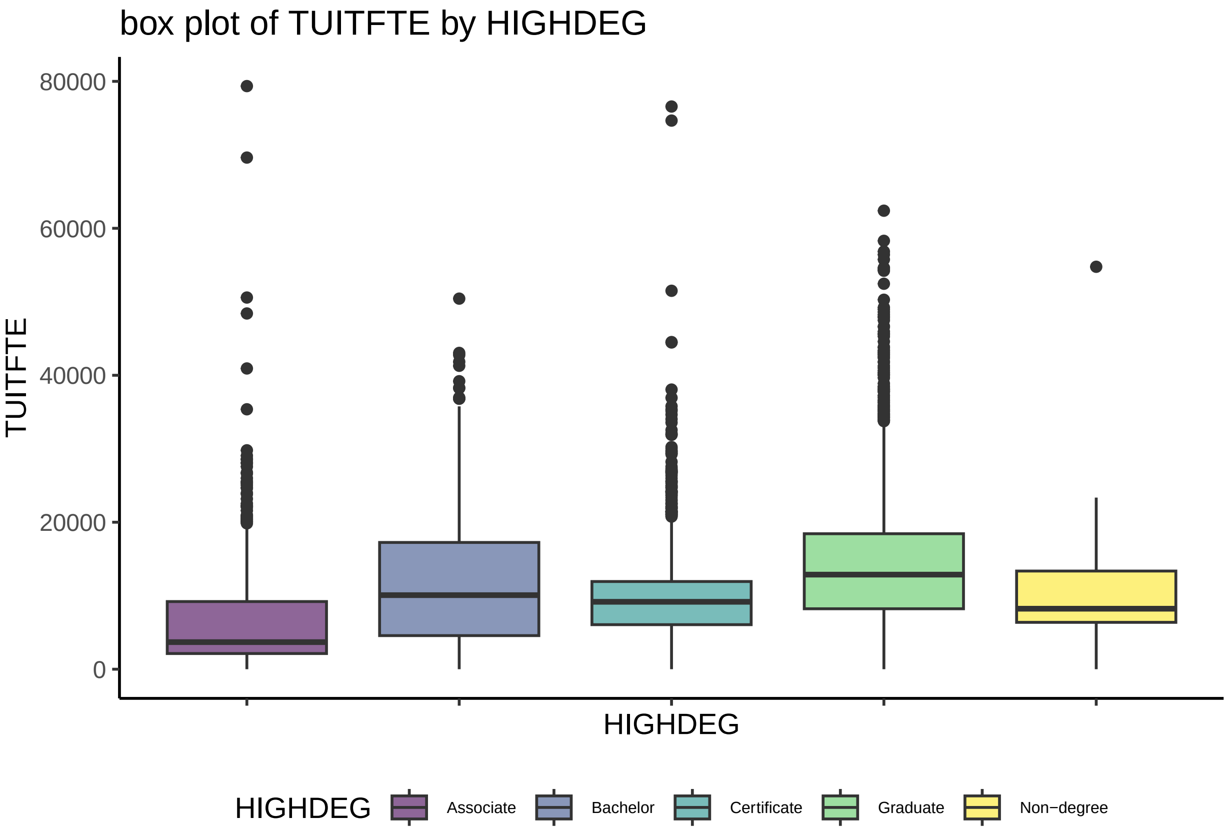
```
pred_rev_box <- box_plot(finance_df, 'TUITFTE', 'PREDDEG')
```



```
pred_exp_box <- box_plot(finance_df, 'INEXPFTE', 'PREDDEG')
```

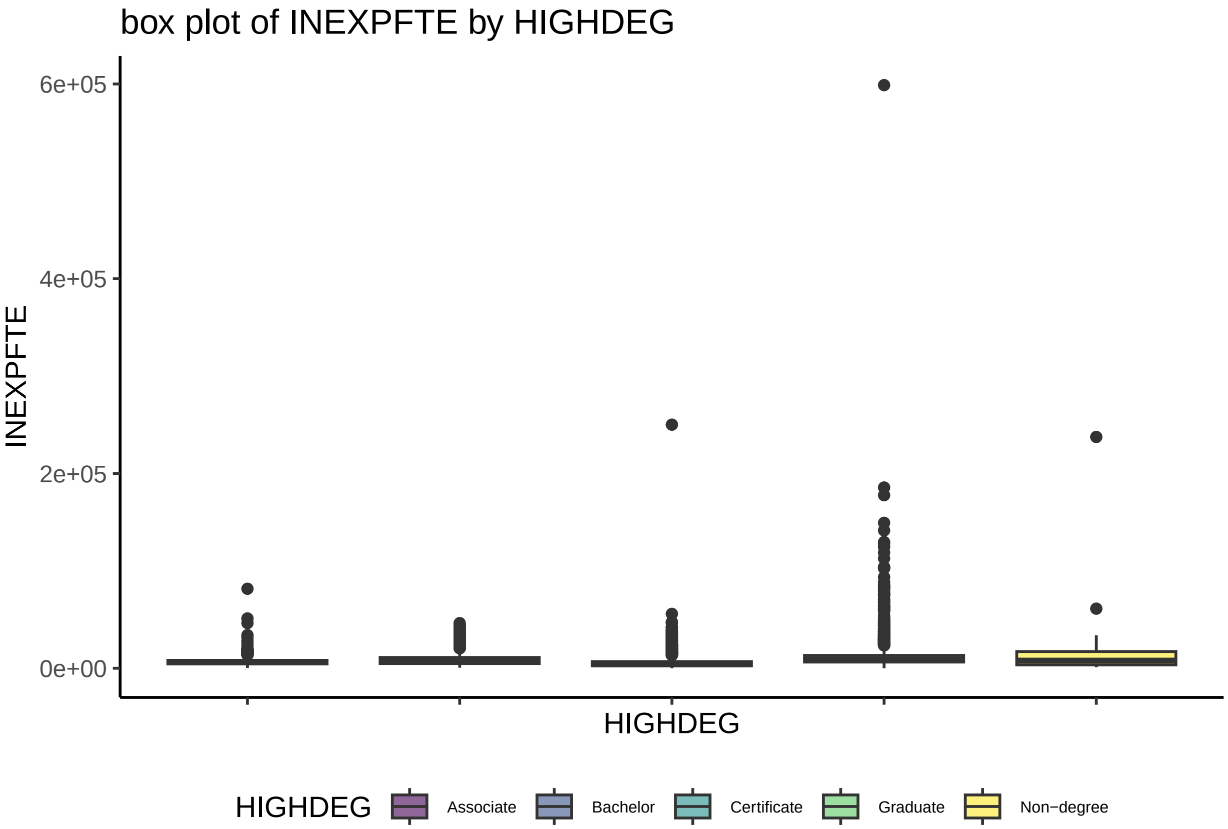


```
high_rev_box <- box_plot(finance_df, 'TUITFTE', 'HIGHDEG')
```



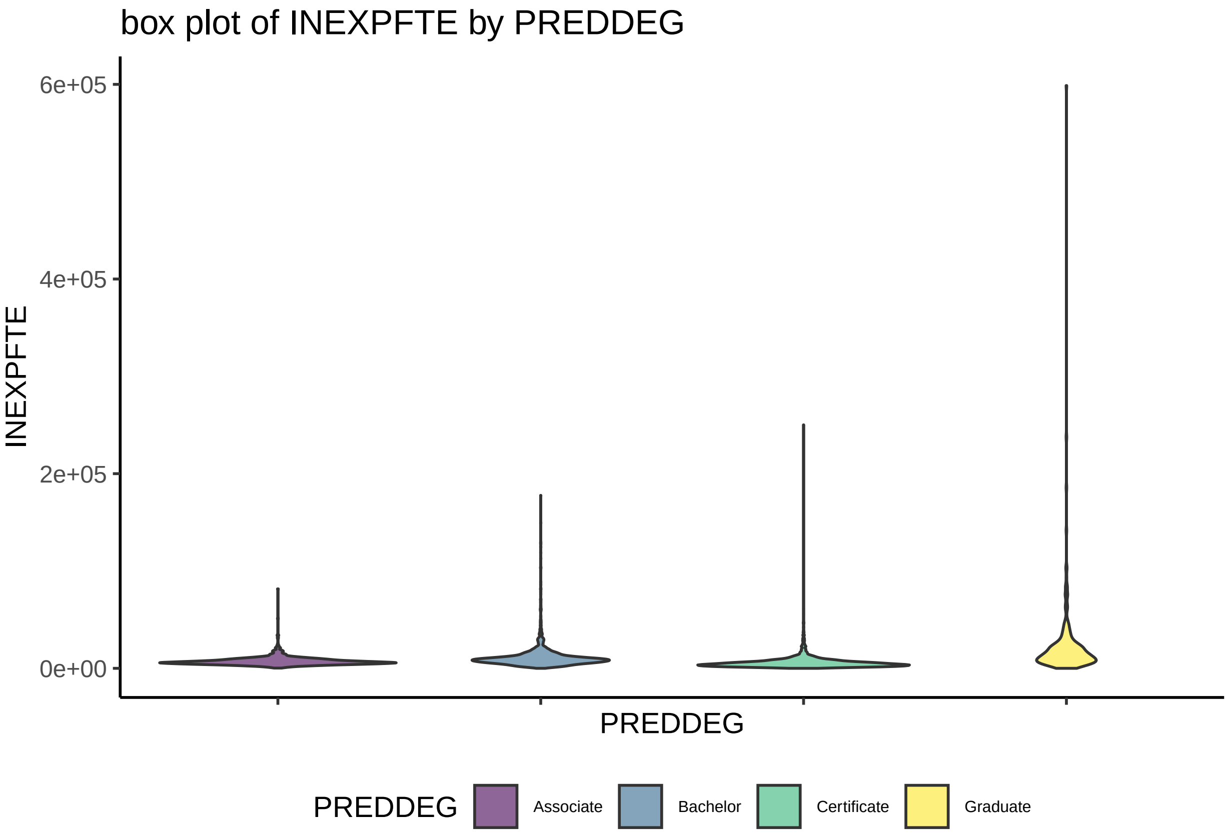
box plot of TUITFTE by HIGHDEG

```r
high_exp_box <- box_plot(finance_df, 'INEXPFTE', 'HIGHDEG')
```



box plot of INEXPFTE by HIGHDEG

```
pred_exp_violin <- violin_plot(finance_df, 'INEXPFTE', 'PREDDEG')
```



ANALYSIS

We can see from the graphs that the revenue seems to be significantly less than the

In relation to box plot of TUITFTE

The highest box plot is related to SF 4-Yrs Other Health Professionals, the cost for such a specialized degree may be greater. Medical schools may have a more limited admissions rate which could lead to an increase in price.

Z-Score Finance Plot Function

```
z_finance_plot <- function(df, fill, colname){
  plot <- ggplot(df, aes(x = STABBR, group = !!sym(colname)))+
    geom_bar(aes(y = !!sym(colname), # z-score value
                 fill = !!sym(fill)), # coloring the bars based on PREDDEG, HIGHDEG, CCBASIC
             stat = 'identity',
             position = 'dodge', # positions the bars next to each other instead of on top of each othe
             width = 0.5)+
    scale_fill_viridis(discrete = TRUE,
                       option = 'viridis',
                       direction = -1)+
    labs(
      title = 'Mean Tuition Revenue vs Mean Instruct. Expense Per Full-Time Equivalent Student',
      x = 'State',
      y = 'z-score of difference between tuition revenue and instruction expenses',
      fill = 'Category'
    )+
    theme_classic()

  # plot <- plot + theme(legend.position = 'none')
  print(plot)
  return(plot)
}
```

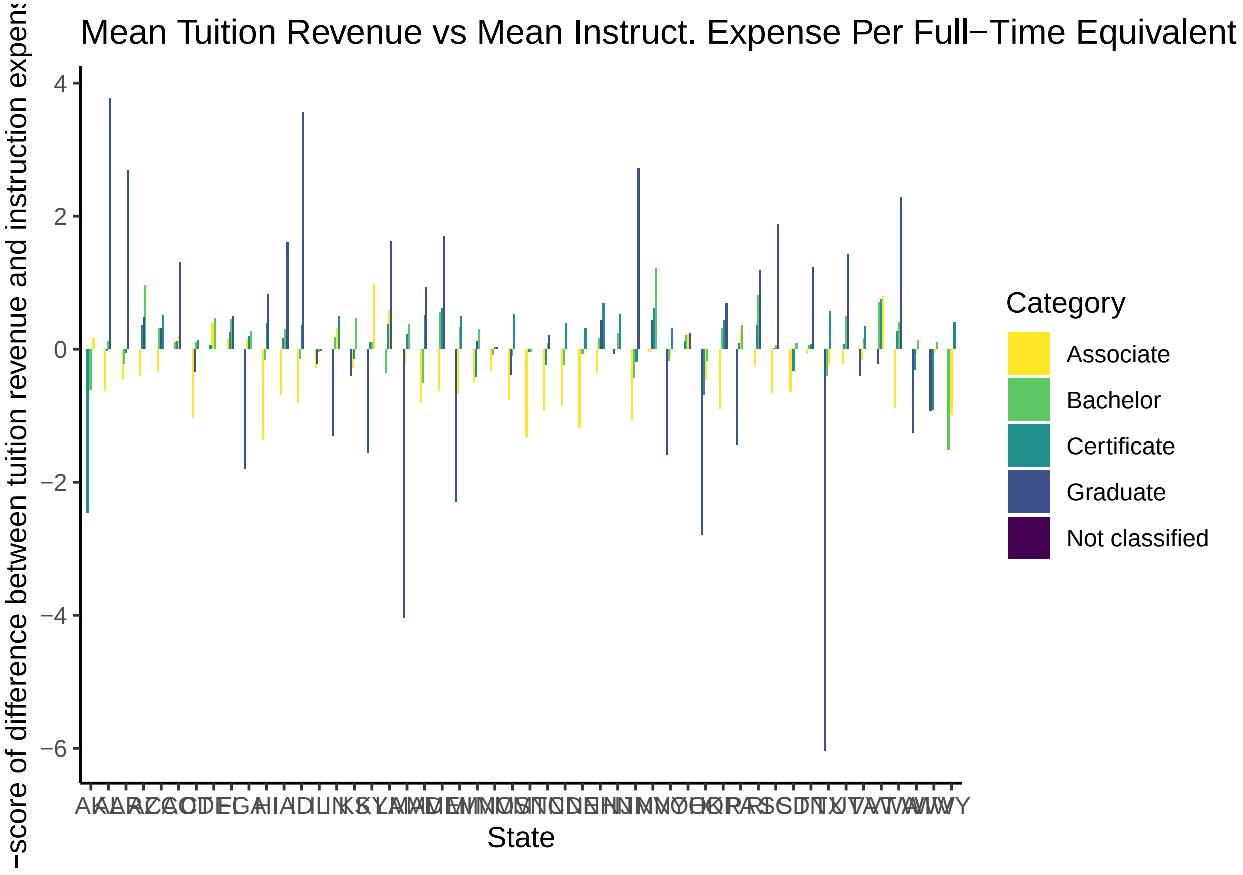
Creating Z-Score Finance Plot for Predominant Degree getting difference based on grouping of Predominate degree

```
# Filtering the data frame based on 'PREDDEG'
preddeg_exp_vs_profit_df <- exp_vs_profit_filter(df_trimmed, 'PREDDEG')
```

```
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
```

```
# Calculating the Z-score for the 'difference' column and adding it as a new column 'zscore'
preddeg_exp_vs_profit_df$zscore <- scale(preddeg_exp_vs_profit_df$difference)[,1] # normalizing calcula

# Creating a Z-Score Finance Plot for Predominant Degree
preddeg_exp_vs_profit_plot <- z_finance_plot(preddeg_exp_vs_profit_df, 'PREDDEG', 'zscore')
```

```
## Warning: Removed 40 rows containing missing values (`geom_bar()`).
```



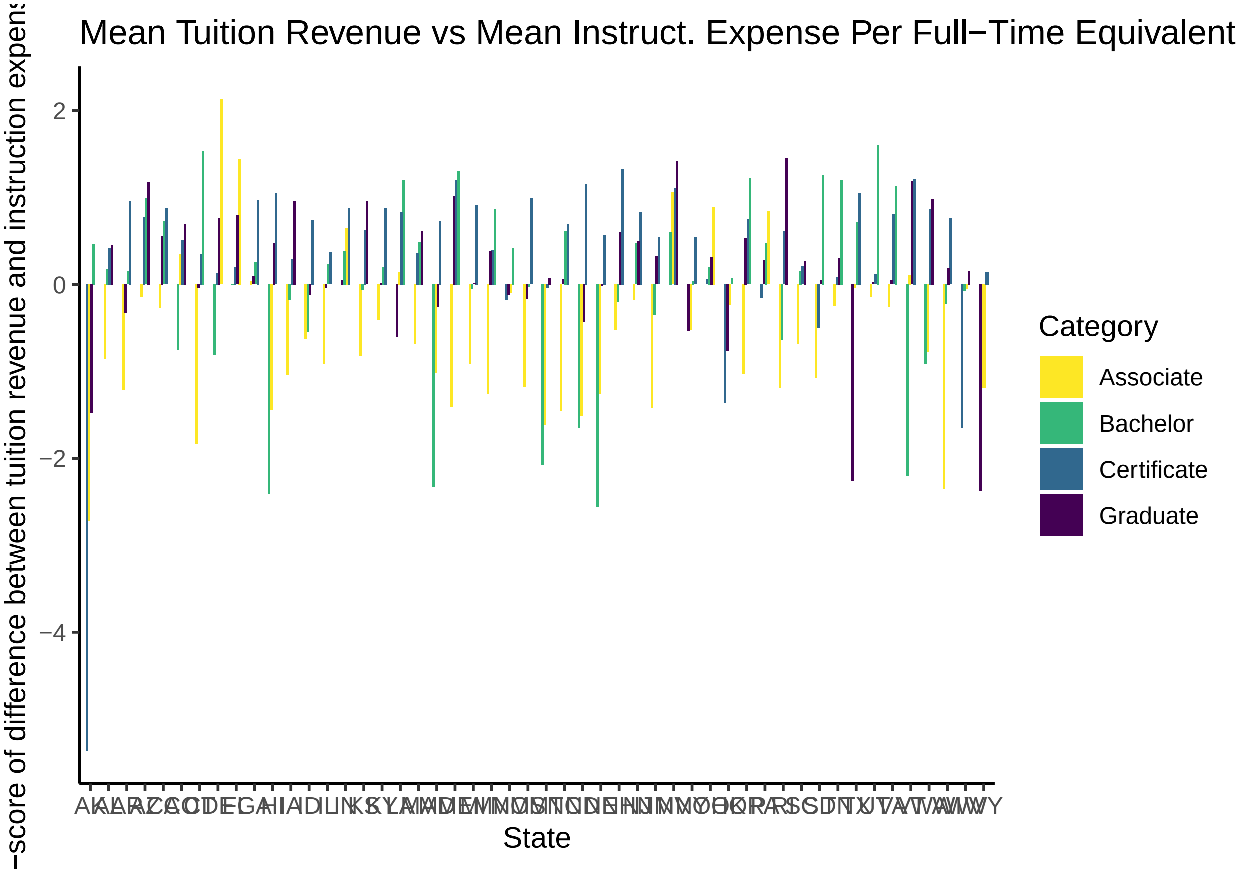
```
#
```

Creating Z-Score Finance Plot for Highest degree

```
# Filtering the data frame based on 'HIGHDEG' and excluding 'Non-degree'
highdeg_exp_vs_profit_df <- exp_vs_profit_filter(df_trimmed, 'HIGHDEG') %>%
  filter(HIGHDEG != 'Non-degree')
```

```
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
```

```
# Calculating the Z-score for the 'difference' column and adding it as a new column 'zscore'
highdeg_exp_vs_profit_df$zscore <- scale(highdeg_exp_vs_profit_df$difference)[,1] # normalizing calcula

# Creating a Z-Score Finance Plot for Highest Degree
highdeg_exp_vs_profit_plot <- z_finance_plot(highdeg_exp_vs_profit_df, 'HIGHDEG', 'zscore')
```



Creating Z-Score Finance Plot for Carnegie Classification, getting difference based on grouping of CCBASIC

```
# Creating Z-Score Finance Plot for Carnegie Classification (There's some sort of error here on 351)
# getting difference based on grouping of CCBASIC

# Filtering the data frame based on 'CCBASIC'
ccbasic_exp_vs_profit_df <- exp_vs_profit_filter(df_trimmed, 'CCBASIC')
```

```
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'STABBR'. You can override using the
## `.groups` argument.
```

```
# Calculating the Z-score for the 'difference' column and adding it as a new column 'zscore'
ccbasic_exp_vs_profit_df$zscore <- scale(ccbasic_exp_vs_profit_df$difference)[,1] # normalizing calcula
```

ANALYSIS

The bars in the negative means the instructional expenses are greater than the tuition revenue,this could indicate that they are wanting to spend more towards delivering educational services.

Based on the graphs, it looks like, Associate degree institutions spend more on instructional expenditures then receiving tuition revenue.

Some factors that could play into this

1. a majority of associate programs could be public which means they will more likely be funded by the government

2. the expenses to get an AA is cheaper than a different type of degree so tuition prices may have a lower threshold

Looking at the HIGHDEG filtering on tuition revenue and instruction expenditure, Certificate programs look to have higher revenue in comparison to instruction expenses.

Some factors that could play into this

1. Certificate programs could be online

2. The time frame to complete it is not as long

3. There is not as much upfront cost to get certain materials (for example in a chemistry lab and the need to get ppe,chemicals, etc.)

Unkwn Variables

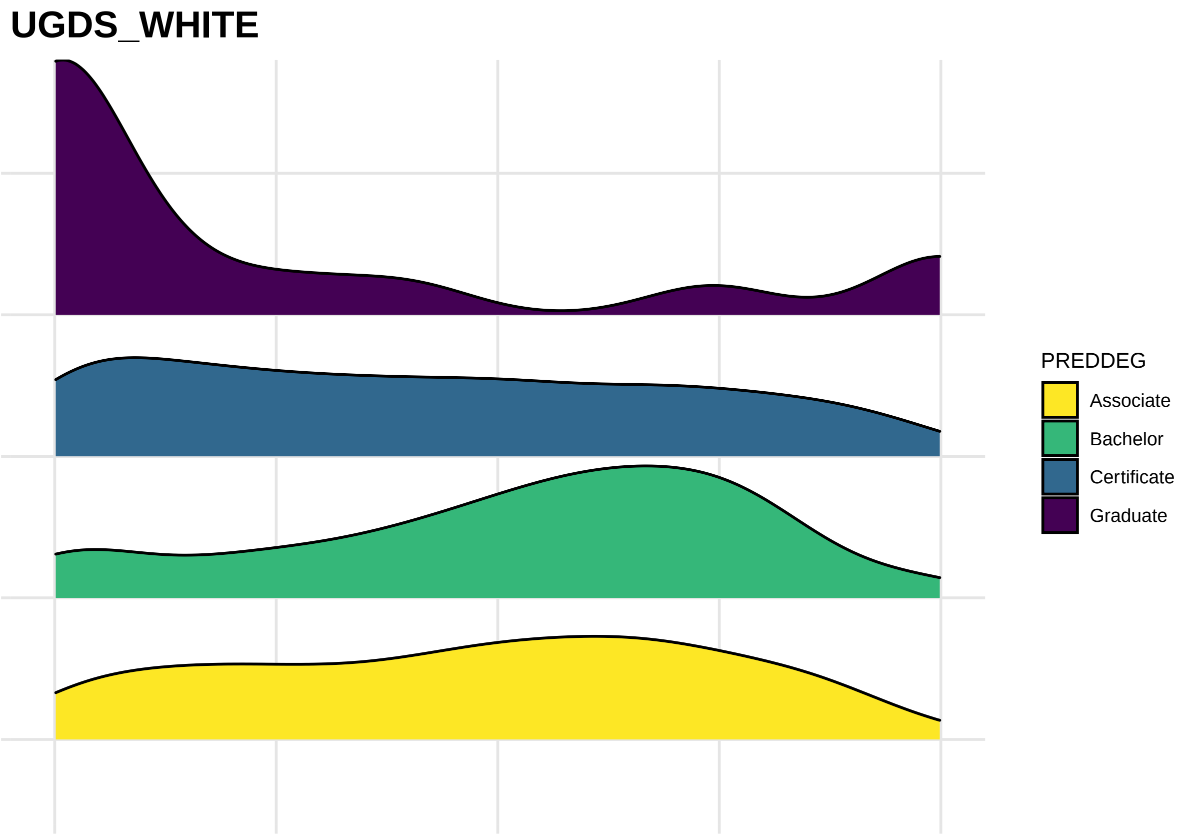
Does financial aid/ pell grant recipient alter the revenue cost? If so then more people who get these may mean the institution get more federal money from their enrollment.

Ridge Plots for Demographic Variables # Function to create ridge plots for different demographic variables # within each PREDDEG group and returns a pdf of all the graphs

```
ridge_plots <- function(df) {
  dem_group <-colnames(df[,5:13]) # getting the column names of the demographics that are listed
  plots <- lapply(dem_group, function(dem) { # looking at groupings of plots based on demographic index
    ggplot(df, aes(x = as.numeric(.data[[dem]]), y = PREDDEG, fill = PREDDEG)) +
      geom_density_ridges() +
      scale_fill_viridis(discrete = TRUE,
                         option = 'viridis',
                         direction = -1) +
      theme_ridges() +
      scale_x_continuous(limits = c(0, 1)) + # set limits to range of data
      labs(title = dem) +  # Set the title for the plot
      theme(legend.position = 'right', # position of legend
            text = element_text(size = 8)) + # size of words in legend
      theme(axis.text.y = element_blank(),  # removing any x or y axis labels
            axis.title.y = element_blank(),
            axis.text.x = element_blank(),
            axis.title.x = element_blank())

  })

  for (i in 1:length(plots)) {
      print(plots[[i]])
    }

  return(plots)
}

# Call the ridge_plots function
dem_ridge_plots <- ridge_plots(ridge_plot_df)
```

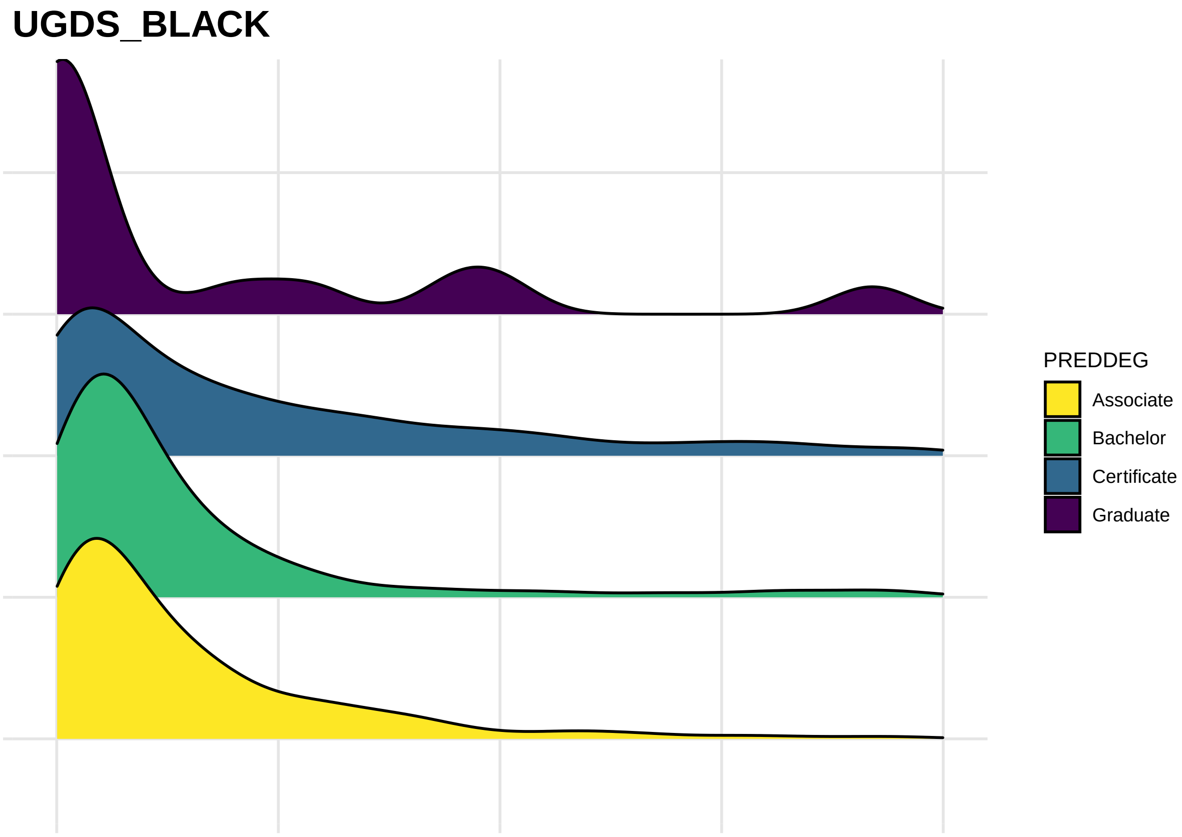
```
## Picking joint bandwidth of 0.074
```

```
## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```



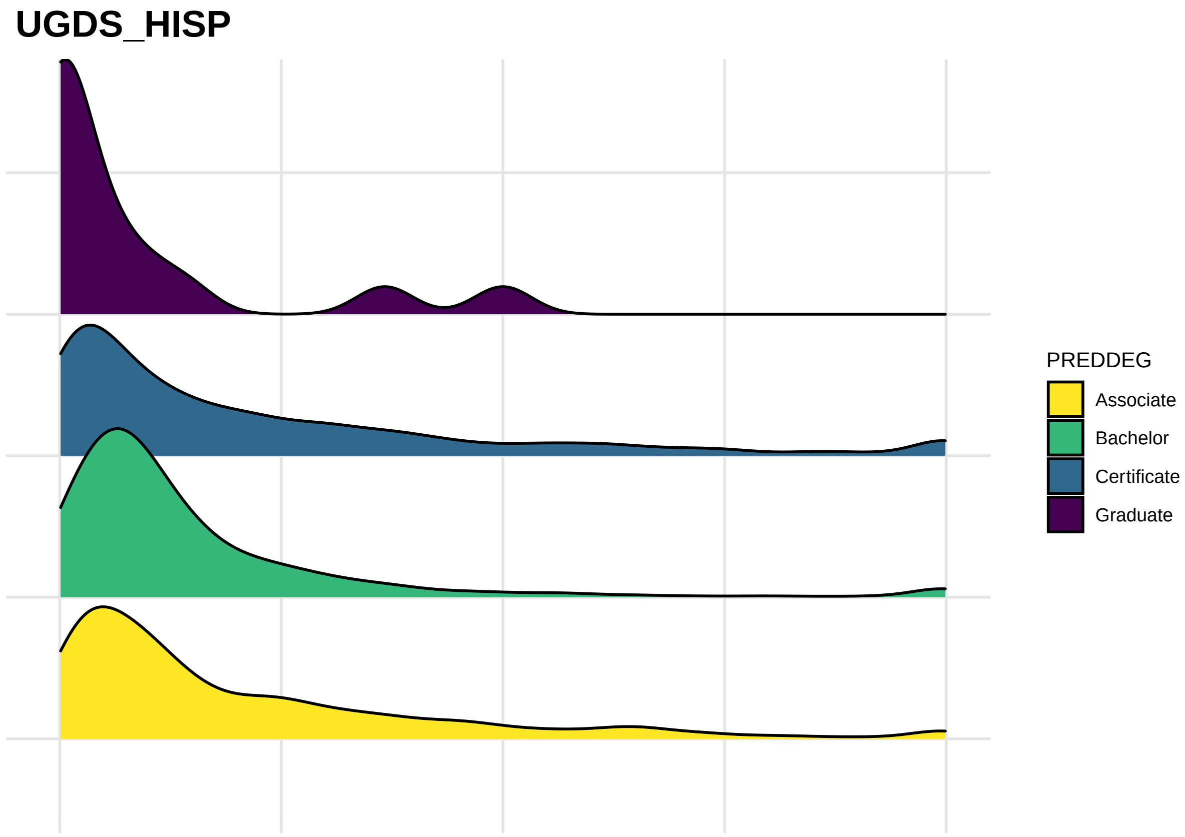
```
## Picking joint bandwidth of 0.0455
```

```
## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```

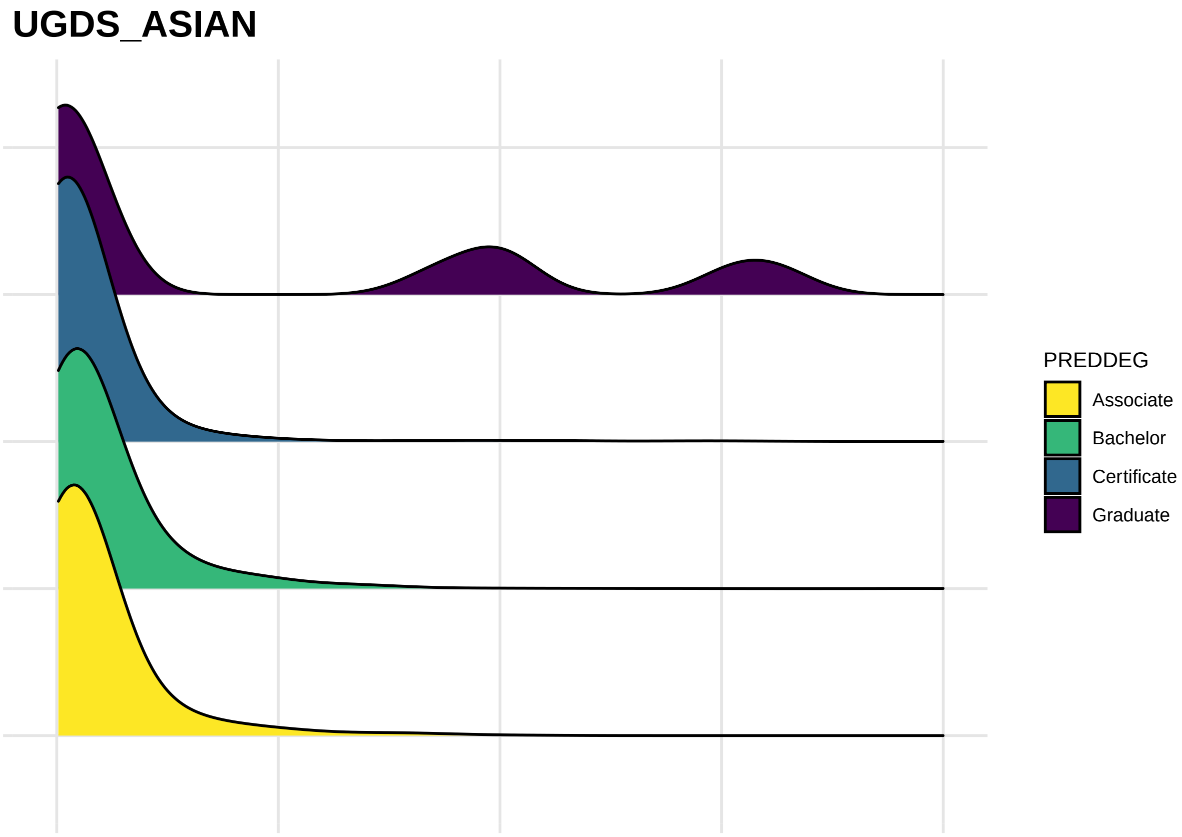


```
## Picking joint bandwidth of 0.0324
```

```
## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```

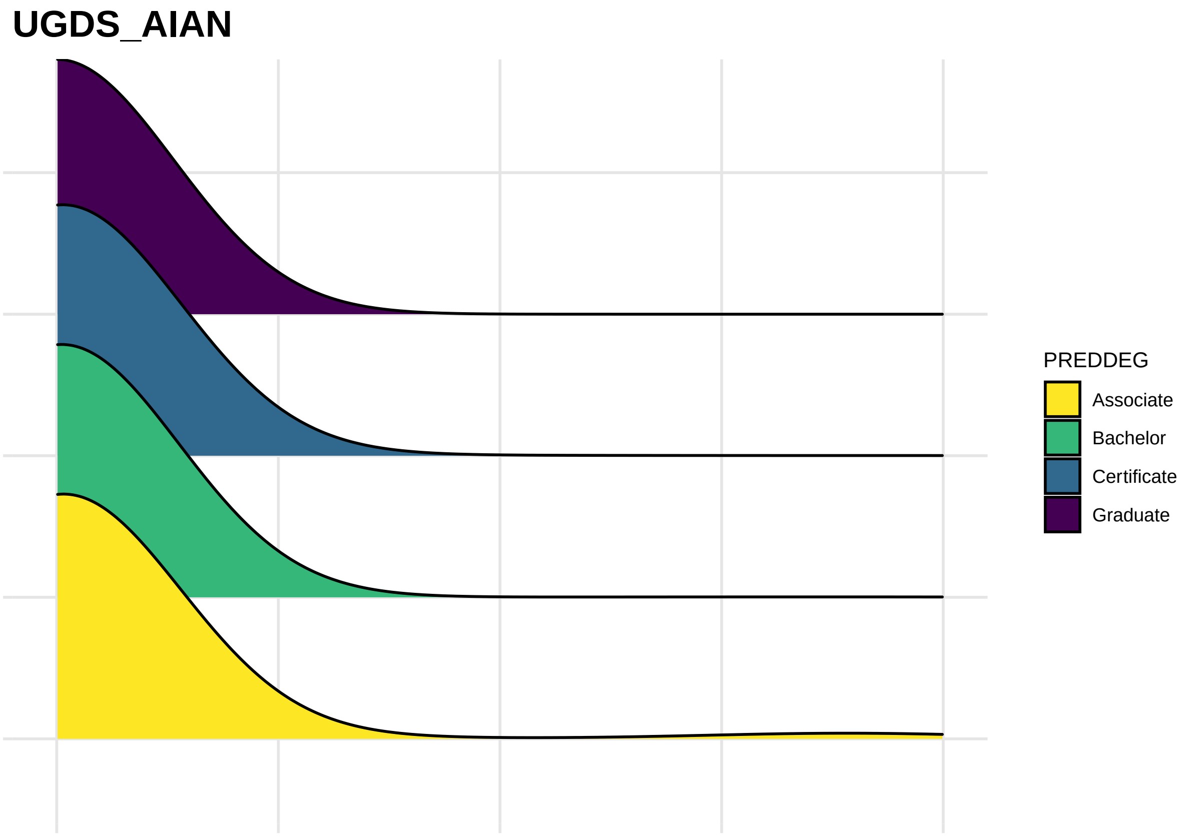


```
## Picking joint bandwidth of 0.0436

## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```

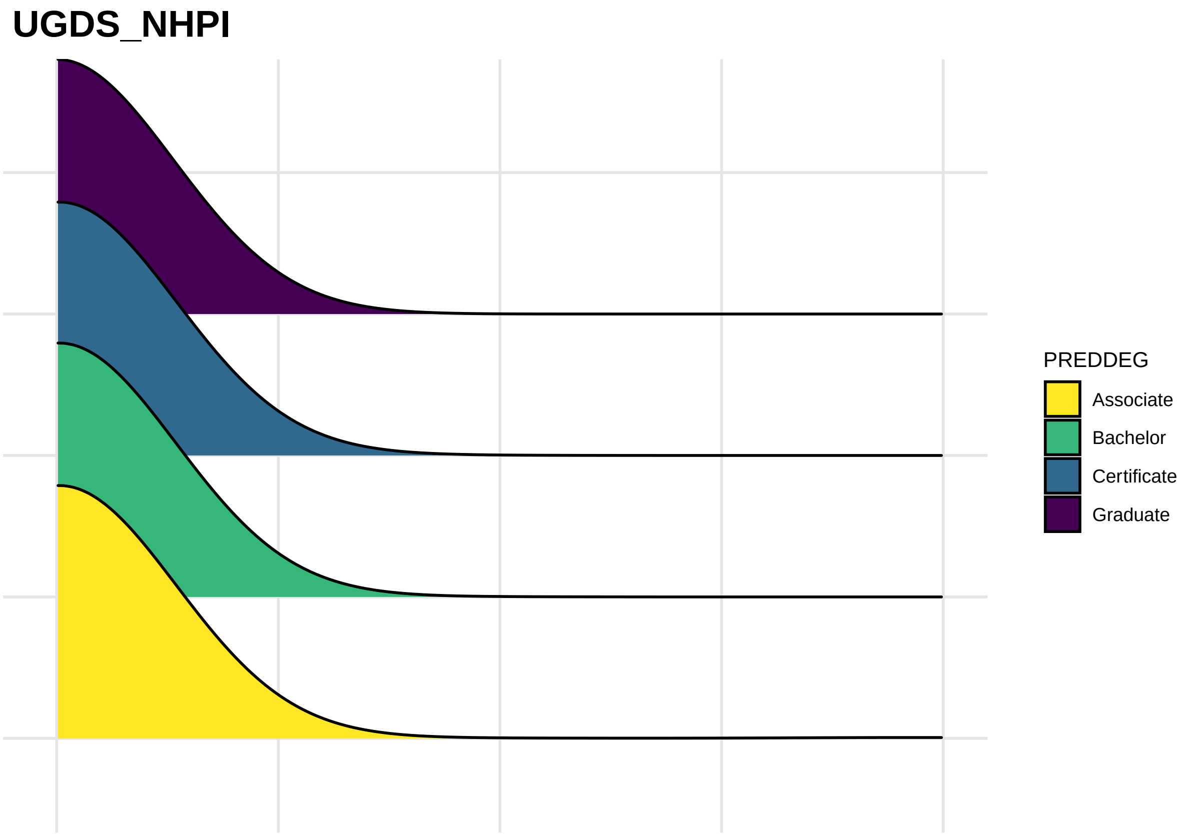


```
## Picking joint bandwidth of 0.132
```

```
## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```

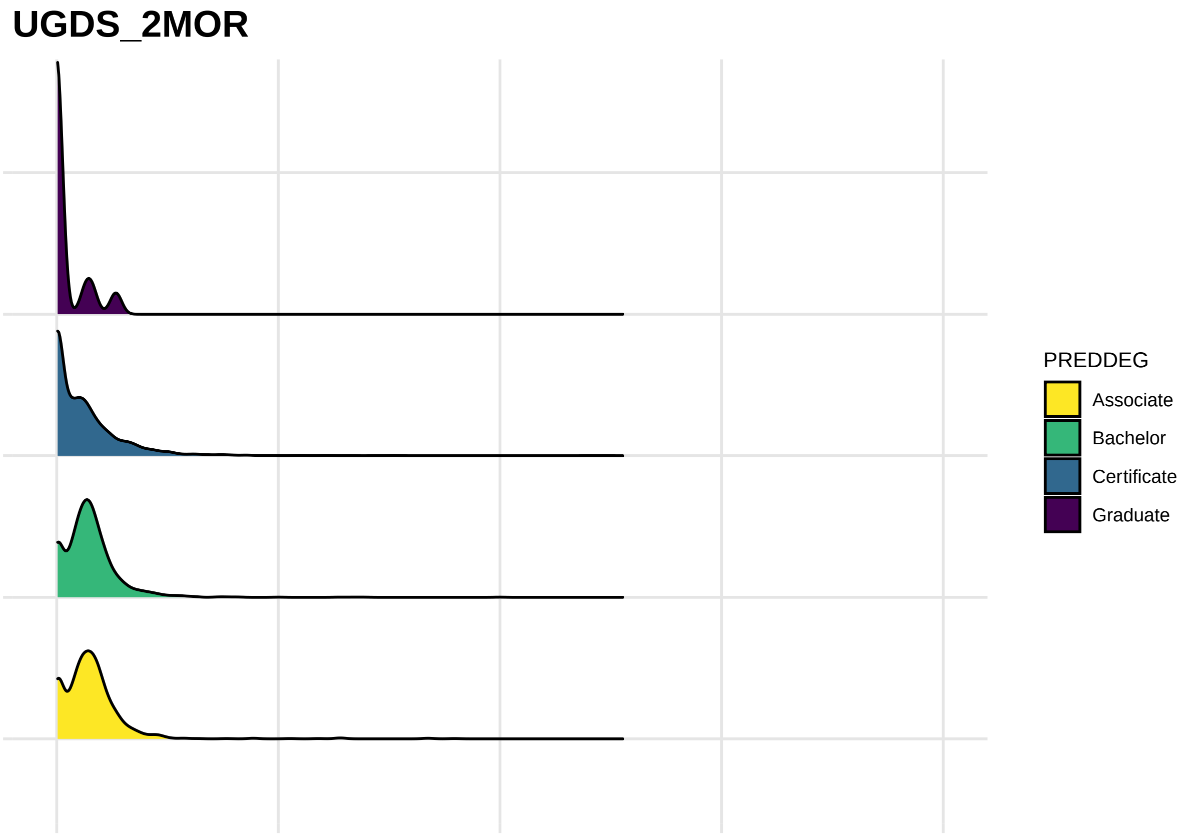


```
## Picking joint bandwidth of 0.131

## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```



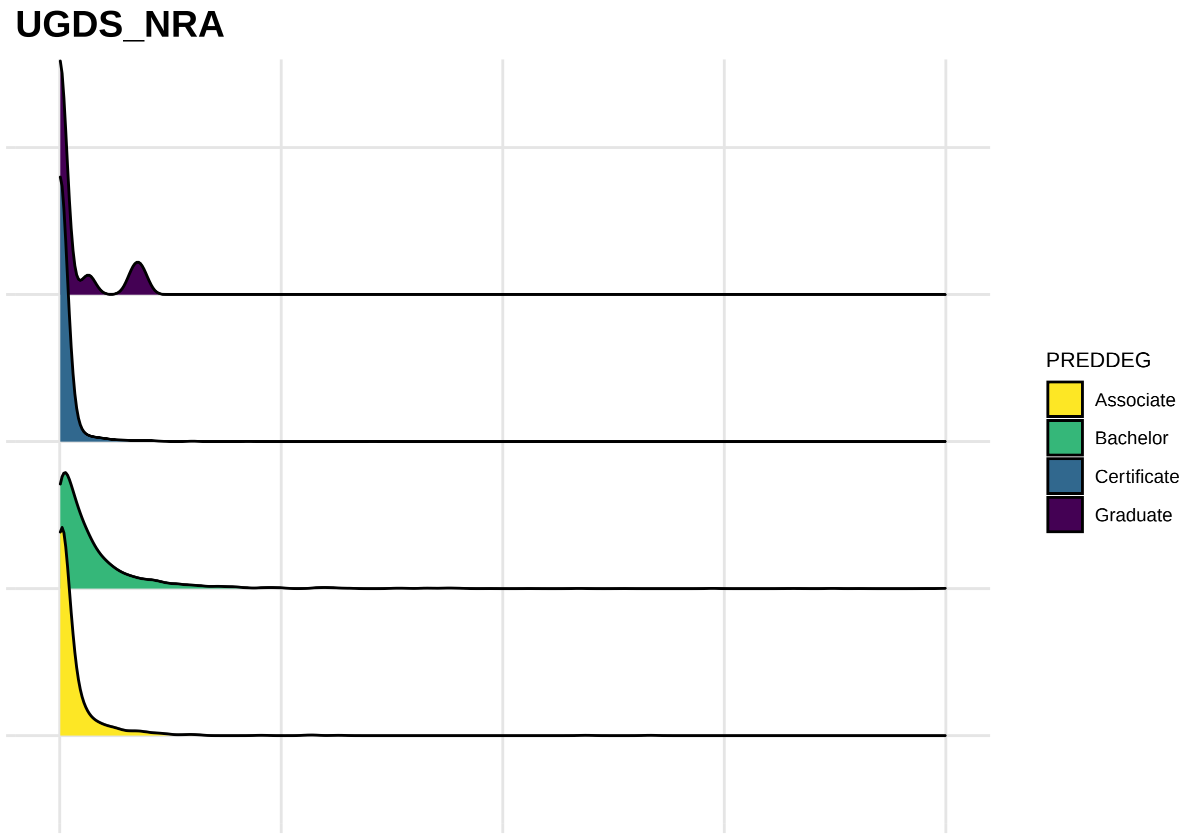
```
## Picking joint bandwidth of 0.00652
```

```
## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```

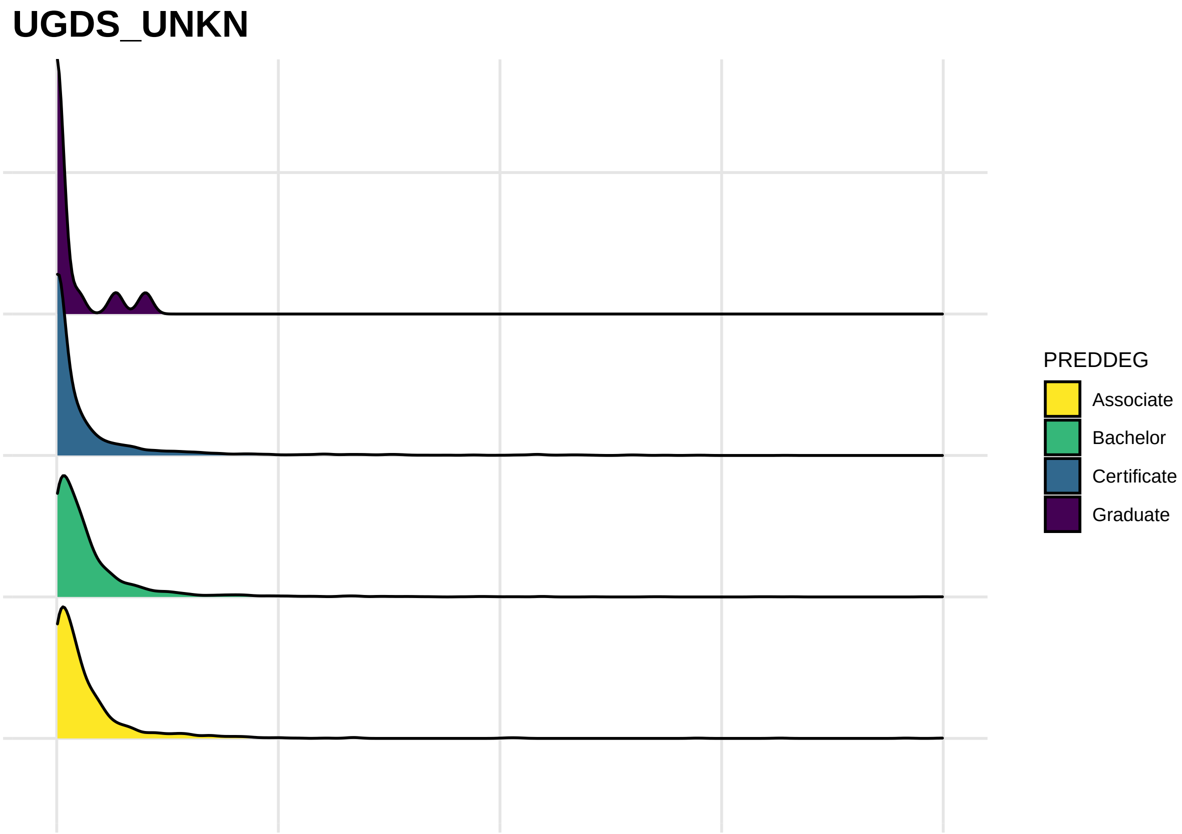


```
## Picking joint bandwidth of 0.00798
```

```
## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```



```
## Picking joint bandwidth of 0.00792

## Warning: Removed 774 rows containing non-finite values
## (`stat_density_ridges()`).
```



ANALYSIS

Those who identify as white have a more balanced distribution withing all the categories except graduate. There could be an association between access to higher education and identifying as white. There is a lower density of individuals who identity as another race who seek higher education (there is a predominate right skew to all the graphs)

Lurking variable: population size of each demographic

Heatmap for Carnegie Classification vs. Highest Degree # Function creates a heatmap that visually represents the correlation # between the degree awarded and its Carnegie classification

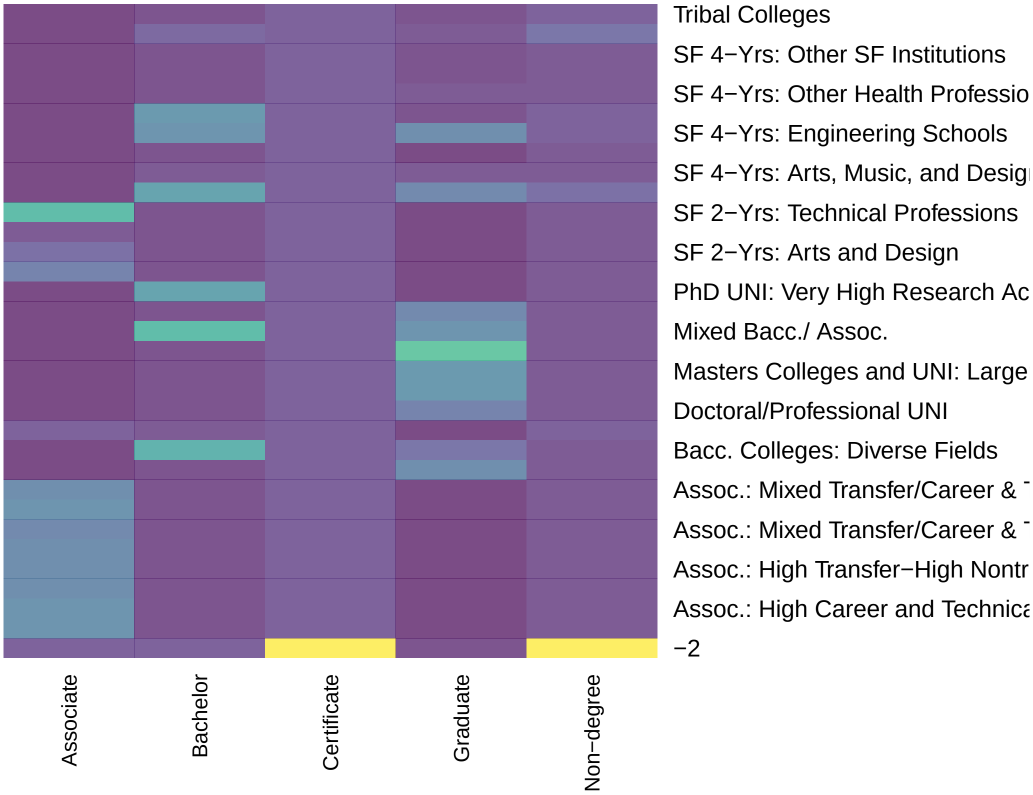
```
heat_map <- function(df){
  # color palette can be found at here:
  # https://r-graph-gallery.com/38-rcolorbrewers-palettes.html
  my_color_palette <- viridis_pal(alpha = 0.7,
                                  direction = 1,
                                  option = 'viridis')(34)# Color palette, # number = # of colors
  heatmap_data <- as.matrix(df) # making sure data is in format that will return an image
  # displaying heatmap between Carnegie classification and the highest degree offered
  map <- heatmap(heatmap_data,
          Rowv = NA, # Do not cluster rows
          Colv = NA, # Do not cluster columns
          col = my_color_palette,
          scale = "column", # Scale by column
          main = "Carnegie classification vs. Degree", # title of heatmap
          cexCol = 0.8) # changing the text size of the degrees offered

  return(map)
}
```

Data Preparation for Heatmap

```
highest_heatmap_data <- table(filtered_df$CCBASIC, filtered_df$HIGHDEG)
predominate_heatmap_data <- table(filtered_df$CCBASIC, filtered_df$PREDDEG)
```

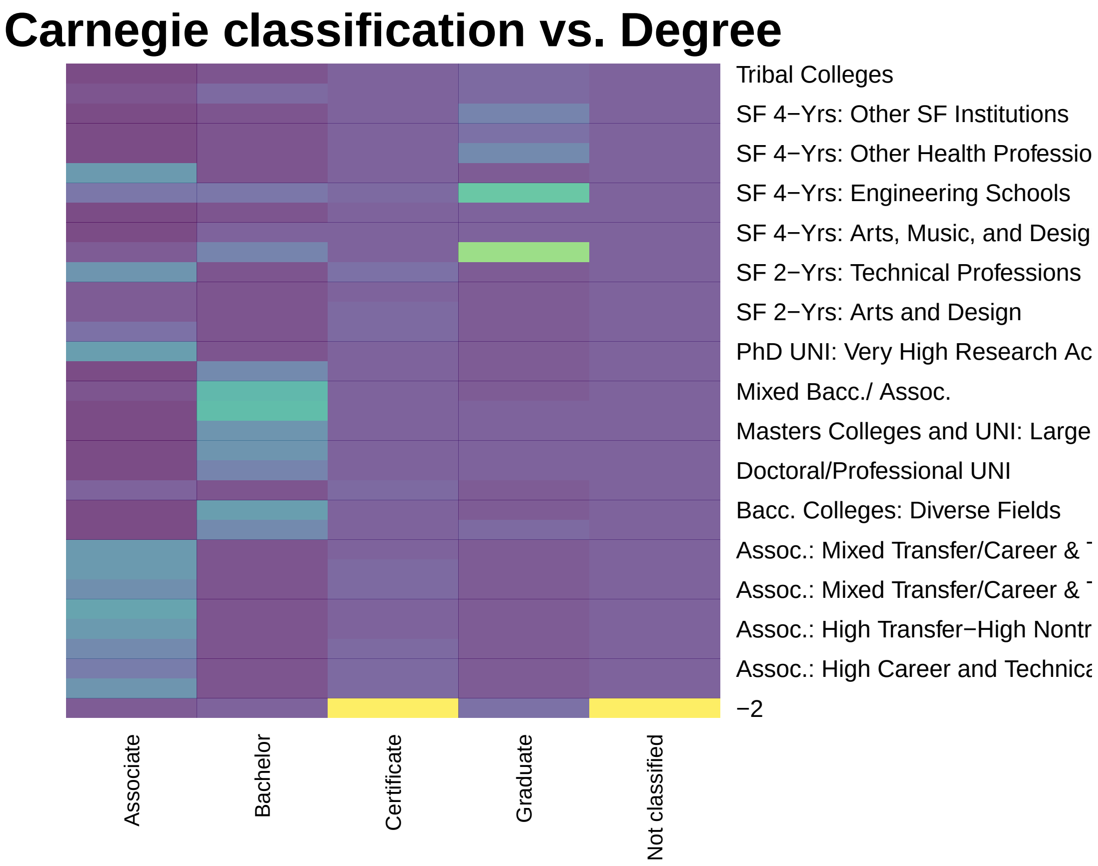
Creating Heatmaps

```
highestvsCCBASIC <- heat_map(highest_heatmap_data)
```



Carnegie classification vs. Degree

```
predominatevsCCBASIC <- heat_map(predominate_heatmap_data)
```



ANALYSIS

General: -2 looks to be referring to institutions that are certificate or non-degree grant. Both also have a similar correlation between CCBASIC scores listed as Associate or Special Focus 2-Years and associate degree awarded.

CCBASIC vs HIGHDEG

Intuitively, there is a better correlation between the CC classification and highest degree awarded. We can see this in the Graduate column where a majority of the lighter colors are associated with masters and doctoral programs. We can see a similar grouping of Special Focus 4-Year institutions and baccalaureate institutions with the Bachelor degree. (I would expect institutions that say 4-years are working towards a bachelor's degree.)

CCBASIC vs PREDDEG

In this map we can see that Bachelor degree has a higher correlation with baccalaureate, masters, and doctoral institutions. This could be because individuals are only looking at getting bachelors, but the institution that they go to allows for further studies.