**Process for Analyzing and Visualizing Willamette’s Accreditation**

Elise Fashimpaur, Fall 2024

Office of Institutional Research and the Department of Psychology

**Background on Project**

Every seven years, Willamette University collects data on the state of the institution for the benefit of the Northwest Commission on Colleges and Universities (NWCCU). This information is used to draw conclusions on the trends that have been taking place at an institution over time. The accreditation process is an important function of assuring the quality of the institutions that students choose to attend. Additionally, it assures that a university has a culture of self-assessment. Thus, this project is important to judge how Willamette University is meeting its goals and how it understands those goal trends.

**Steps in the Process**

1. The data used comes from College Scorecard (https://collegescorecard.ed.gov/). College Scorecard holds years of information, going back to the 1996-1997 academic year. From this website, one can download a zipped folder of all the CSV files for each year, as well as a spreadsheet with the data dictionary.
2. With all the spreadsheets downloaded, the next step taken was to merge all the files into the same spreadsheet, including an extra variable for the year of the file from which it was taken. In this instance, as will be explained in the data dictionary for this project, there will both need to be *integer*-type year information as well as *text*-type year information to allow for the most amount of calculation and visualization options. This can happen as the data is being used or can be done in the creation process of the *year* variable.

At this time, all variables should be checked and edited to be in the correct form as would make sense for their calculations. For example, variables relating to rates and proportion should be classified as *numeric*.

1. At this same time, use the CollegeComparators.csv and VariableComparators.csv tables to adjust which schools are of interest to the school during that round of accreditation. This removes the necessity of working this all in code, as these sheets can be edited separately from any code changes to keep the analyses updated.
2. Once all the above steps are complete, filter the completed set of College Scorecard information (with all the years present) to only include the colleges and variables of interest, based on the spreadsheet modified in Step #3. Then, use this filtered set to conduct the desired analyses and visualizations that are needed for this project.

**Explanation for R Code**

For this initial project, R Studio was used primarily, while Jamovi was utilized as an intermediary step in fixing the data types. In this section, we will step through specifically how each portion of the R Markdown (rmd) file in the Github is accomplishing its goal. This information may also be included in the comments of the rmd file but will be expounded upon in further detail here.

**Accreditation Visuals**

Set-Up Phases

The following packages were utilized in R for this process: ***dplyr***, ***readr***, ***tidyverse***, and ***ggplot2***. The beginning of the R file is installing each package and using the library() function to load them into this session.

Next, the csv file that is to be used for analysis is read into the filtered\_path variable, which will then be read into the name jamovi\_filtered, which will be used throughout the rest of the code. The only part here to be adjusted per use is the quoted file path, which should include the full path including the file extension.

There are additional variables defined at the start to make later visualization simpler. This includes defining a regional\_comparators variable, a national\_comparators variable and an all\_comparators variable. In addition, this section also defines which colors should be used for graphics that involve the all\_comparators variable, specifically in the order that is used and defined in said variable. If anything is wished to be changed, it should be done here with the colors listed or the schools included, rather than throughout the code.

Example: PCTPELL and C100\_4

This first graphic uses the previously defined jamovi\_filtered variable and the ***ggplot*** feature to create a scatterplot with an overlaid line graph. The x variable is PCTPELL\_DCS, which is the proportion of first-time, first-year students receiving Pell, multiplied by 100 to be a percentage rather than a decimal. The y variable is C100\_4, which is the graduation rate for all students within 4 years of matriculation. The geom\_jitter with an alpha of less than 1 (in this case, set at 0.8) allows for dots that are overlapping to be pushed slightly apart for better visualization with a slight translucency to see how much they overlap. The labs and theme additions create labels for the graph and text positioning respectively. Geom\_smooth is what overlays a regression line over the scatterplot. In the first example, it is a regression line as calculated, but the second graph (otherwise with all the same features) uses method= “lm” for a linear model of the regression, if that is preferred. The final graph of this section subsets the jamovi\_filtered data by the all\_comparators variable to see the trends within each comparator group.

**What Could Be Improved Going Forward**

While the current state of this project is a significant leap from where it began, there is still room for adding more information that could detail the place Willamette University stands at present. For example, the R code is structured in a way that takes the same variables from all CSV files and only collects common variable names. These column names were from the most recently updated data dictionary provided by College Scorecard. Because of this, there are numerous variables that had been renamed and recalculated over the years. It may be worthwhile to do a transformation of this data to ensure that no matching information is being left out due to the structure of the code. This would have taken a considerable amount of manual effort, which was not available during the semester when the beginning of this project took place.