Capstone Project: COLLEGE SCORECARD DATA
Jeevarani Radhakrishnan

Data Source

The college scorecard data at the link below is designed to increase transparency, putting the power in the hands of students and families to compare how well individual postsecondary institutions are preparing their students to be successful. This provides data to help students and families to compare college costs and outcomes as they weigh the tradeoffs of different colleges. These data are provided through federal reporting from institutions.

https://collegescorecard.ed.gov/data/

The college scorecard provides the following documentation:

Data Dictionary: The Scorecard data dictionary provides information on each metric in the API and downloadable data files, including the variable name from the data files, a longer descriptive name, the API location and developer-friendly metric name, values and value labels for the metrics, data source, and high-level notes for each metric. In addition, the data dictionary contains a cohort map that identifies the data elements that are present in each academic year.
Data: This contains the data that appear on the College Scorecard, as well as supporting data on student completion, debt and repayment, earnings, and more. The files include data from 1996 through 2017 for all undergraduate degree-granting institutions of higher education.

Data Cleaning Notebook

Notebook

Data Cleaning Steps

1. As a first step, data dictionary and cohort map are imported. The cohort map was analyzed to examine which academic year has most non null values and has the data required for analysis. Based on that, 2014_2015 data was chosen as that had comparably low non null values.
2. Then the all the category columns were examined again to see if the relevant fields have reasonable non null values
3. The 2014_2015 academic year file is imported into the dataframe.
4. Since the study is focused on bachelors and graduate degree, all the institutions whose highest degree is academic, or associate is removed
5. The column data types are updated as Categorical and numeric by looking up the datatype column in the data dictionary.
6. The columns with all non-null values are removed.
7. Since UNIT ID is unique, it was set as an Index.
8. Then for each category, the columns are examined to check the number of non-null values. All columns with less non null value is removed. The columns which are not required for the study are also removed.
9. The variable name in dataframe are changed to developer friendly names. The categorical column numbers are changed to values.
10. The shape of the dataframe was changed from 7703x1977 to 3426x264
11. The cleaned-up data frame and the column details are written into excel sheet.