Data Preparation & Documentation Group Assignment:

The dataset that our group has chosen to use for our team project is the US Department of Education’s College Scorecard dataset, which is freely accessible online from the department’s Scorecard website (https://collegescorecard.ed.gov/data/). Specifically, this data contains records on all Title IV aid receiving higher education institutions in the US, as indicated by the Integrated Postsecondary Education Data System (IPEDS) federal reporting surveys. The purpose of the US Department of Education providing these data freely online is so that prospective high school students and their parents could have a way to compare US higher education institutions across a wide range of data points dealing with college costs and student success outcomes, in order to allow them to make the most data driven and informed opinion as possible on where potential institutions to attend. The data are compiled by the US Department of Education using federal reporting from institutions, data on federal financial aid, and tax information.

There are several data files made available for consumers, but we have chosen to use the “Most recent data” dataset provided on the College Scorecards main website that was linked in the previous paragraph, but the exact URL to download our chosen data is (<https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv>). The data is in the CSV file format and is approximately 145 MB in storage size, but our group saved the data in the XLSX file format.

APA Citation:

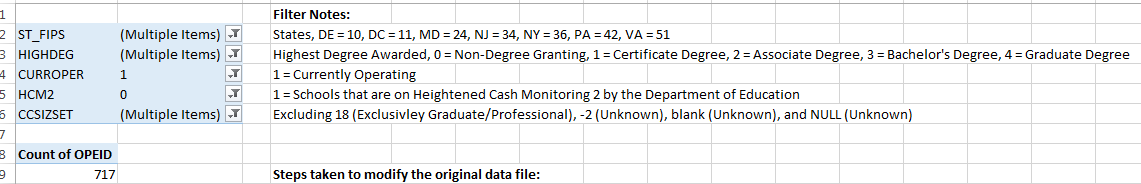
US Department of Education. (2017). College Scorecard Data. Retrieved September 11, 2017, from

<https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv>

According to a notice (https://www2.ed.gov/notices/copyright/index.html) by the U.S. Department of Education, “Unless specifically stated otherwise, all information on the U.S. Department of Education's (ED's) website at www.ed.gov is in the public domain, and may be reproduced, published or otherwise used without ED's permission”. This notice details that the data downloaded from the College Scorecard is publically available and is free to use or reuse.

There is a great deal of metadata provided by the US Department of Education for this data set available from the main webpage mentioned in the first paragraph. The “Data Dictionary” (<https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx>) and the “Documentation Report” (<https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf>) were the primary pieces of metadata provided that we made use of. Both the Data Dictionary and the Documentation Report provide information regarding data element variable names, long description names, exact sources of the data, data types, and potential values (as well as translations in case certain fields were codified). Our group also added noted on to the column headers in our XLSX data file in order to make analysis easier, since several variable names were not very descriptive of the data that they represented. Also, since several nominal fields were codified to numbers in the data set to reduce overall file size and therefore required the separate data dictionary in order to be understood, we added columns with the translated variable descriptions.

We’ve encountered other various issues that were able to be resolved as well. One issue is that this file contains data on over 7,000 instructions, which was more than we found to be necessary, so we filtered the list to a more relevant and manageable total of 717. The majority of the filtering done was limiting our data set to schools located in states in or near the Mid East region. We also wanted to only include instructions that catered to the average undergraduate student attending a 2-year or 4-year institution. A complete set of steps of our entire data scanning and cleaning process is below.

1. After downloading the “Most Recent Data” CSV file, we saved to an XLSX file format.
2. Using the Data Dictionary, we identified ~150 variables/fields that we would be interested in using for our analysis, so between the 4 group members, we equally browsed all values contained within those variables/fields. This was done in order to make sure no values were found that were not documented in the Data Dictionary or the Documentation Report.
3. We specifically deleted the following XLSX columns that we did not plan on using for analysis: BPK, BMX-BPG, BKJ-BMT, BJG-BJZ, BJB, BIX-BIY, QB-BIT, PS-PW, ON-OW, OD-OH, MQ-NN, LC-LD, KH-KU, BJ-KE, AM, V-X, O, I-K, F-G, A.
4. We deleted any rows (instructional records) that we did not want to include in our relevant pool for analysis. The exact filters (along with notes) used are below, which left us with 717 institutions:  
   
5. For any variables/fields that were codified (as opposed to using their English-worded text value), we inserted a column with the code description taken from the Data Dictionary.
6. We removed any fields/variables from the data dictionary that we did not keep in the data set, in order to make the data dictionary easier to navigate and use.

*Word Count: 784*