**DATA PREPARATION AND CLEANING**

**Team Members: Laura Walker, Himanshi Manglunia, Kalpita Raut**

**About the Dataset:**

IPEDS is the Integrated Postsecondary Education Data System. It is a system of interrelated surveys conducted annually by the U.S. Department of Education’s National Center for Education Statistics (NCES). IPEDS gathers information from every college, university, and technical and vocational institution that participates in the federal student financial aid programs. ("About IPEDS," 2018). Our analysis will focus on comparing the University of Maryland College Park to the rest of the institutions in the Big Ten Conference.

The datasets can be found here:

<https://nces.ed.gov/ipeds/use-the-data/download-access-database>

The ***IPEDS201516TablesDoc.xlsx*** spreadsheet contains the output for the IPEDS fields chosen for our analysis from five years of data starting in 2011 and ending in 2015. . The IPEDS data collection consists of 3 collection modules: (All modules for years 2011-15)

* Institutional Characteristics--basic information about the schools.
* Admissions Fall, Graduation Rates, and Student financial Aid.
* Fall enrollment, Finance Human Resources components and the Academic Libraries.

The data in this database are final released data. During the 2016-17 collection, institutions were allowed to submit revisions to their 2015-16 submission through the 2015-16 Prior Year Data Revision system. The 2015-16 provisional data were then updated with these revisions for the final release. There are some data tables and variables that institutions are not allowed to revise. Revised data tables and variables are identified in the spreadsheet in this workbook.

**License for the Dataset**

The license for the above dataset is available here:

<https://ies.ed.gov/funding/researchaccess.asp>

**Description of the metadata**

The following data files are included in 2015-16 IPEDS Database (IPEDS\_2015-16\_Final.zip):

IPEDS201516.MDB (519mb) – This is the Microsoft Access database that contains the data tables and metadata tables.

IPEDS201516TablesDoc.xlsx – Excel workbook containing the following spreadsheets.

Metadata Tables:

* **Tables15** – This table lists each data table in the database by survey component. It includes table name, table title, brief description, and data year covered. It also identifies the data release (i.e., final data that has revised the provisional data or provisional, which has yet to be revised) and the date the table was released in the IPEDS “Use the data” website.
* **vartable15** – This variable table lists each variable in the data table by survey component. It includes variable names, descriptions, source, data types and formats. Also, vartable15 identifies variables that were subject to revisions.
* **valuesets15** – This table lists the value labels of each categorical variable in the database.
* **MetaDataDefinitions** – This table describes each variable found in the metadata tables (i.e., Tables15, vartable15, and valuesets15).

The investigators have files for each year of our analysis and the descriptions above hold true for the years previous.

**Issues with the Dataset:**

The dataset contains missing values. These missing values were replaced with “N/A” values using the package called RODBC. When null values are present, the package changes the value to “N/A.”

We wanted to do an analysis of students test scores, financial aid data in terms of grants and scholarships by income level and aggregate counts of the student population, known as cohorts. We had to get all sixty-two variables for each of the five years and then combine them all into one data frame. (Please see our attached data diction for all variables and their descriptions. )

Towards the end of our R-code, we survey our final dataset, IPEDS, and count how many values are “NA.” Unfortunately, twenty-one out of sixty-four variables came up with at least thirty null values in the final IPEDS data frame. When we do our final data visualization, we will include code that will exclude null values.

Unfortunately, the IPEDS data is stored in an access database for each year. Table names changes with the years too, so queries had to be adjusted as we extracted data from each year. Per the RODBC R package that connects RStudio to access databases, we had to open a connection with RStudio for each year. From there, we ran an SQL query from each year for which we needed data. We were sure to get all of the variables in the same order per year so that when we combined each year into a single data frame, they would all line up with each other. After we extracted a query for a given year, we added a column to the yearly data, year, so that we can distinguish each row with an institution and year. There are over 7,600 institutions of higher education who participate in this survey. We extracted all the institutions that are in the Big Ten Conference per our design. Big Ten Institutions include: (Name, UNITID used in IPEDS)

* Indiana University,151351
* Michigan State University,171100
* Northwestern University,147767
* The Ohio State University,204796
* Penn State University,214777
* Purdue University,243780
* Rutgers University,186380
* University of Illinois at Urbana-Champaign,145637
* University of Iowa,153658
* **University of Maryland,163286**
* University of Michigan - Ann Arbor,170976
* University of Minnesota,174066
* University of Nebraska,181464
* University of Wisconsin,240444

The attached R Markdown comments each step that went into getting each year’s data into R as a single data frame and the process of combining all five years of data into one frame.

**References**

**About IPEDS,** National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education