Problem Set #8 Instructions

November 16, 2018

Reading for next week

- [REQUIRED] GW chapter 11 (Data import)

General instructions for R script

In this homework, you will submit an `R script`. Please make sure that your R script runs without any errors before submitting.

- General Instructions for Problem Sets Here
- Open R script with RStudio (may open with R or other application)
- Can create "comments" by using "#"
- Shortcuts for executing commands
  - Cmd/Ctrl + Enter: execute highlighted line(s)
  - Cmd/Ctrl + Shift + Enter (without highlighting any lines): run entire script
- Output from commands executed from R script
- Output will appear in the "console"

Broad overview

The goal of this problem set is to have you all get some practice with reading in data from the web, labelling data, tidying data, and merging data.

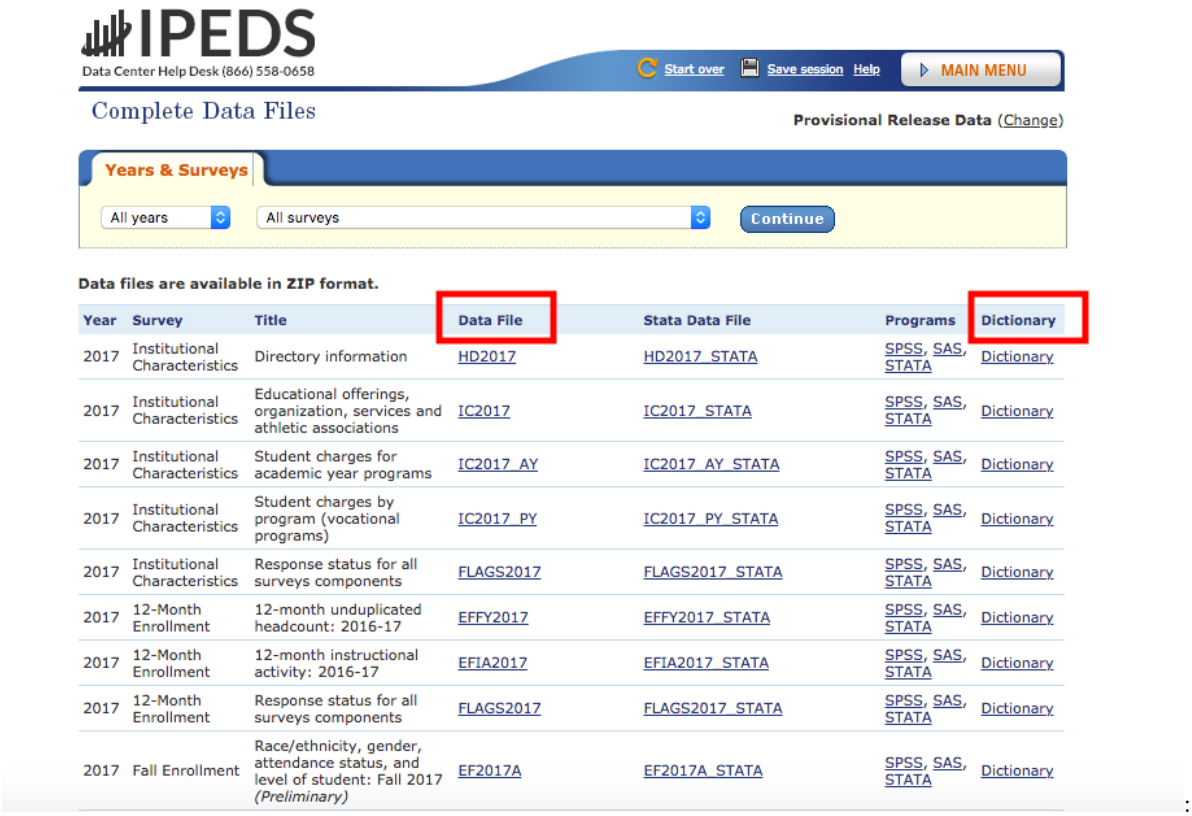
You will:

- Import three IPEDS data sets into R
  - **HD2017** (from Institutional Characteristics survey)
  - **FLAGS2017** (from Institutional Characteristics survey)
  - **EFFY2017** (from 12-month enrollment survey)
- For each data set, subset data to a few columns
- For each data set, add variable and value labels
- Tidy one of these data sources (EFFY2017), so that it has one observation per unitid
- Merge these IPEDS data sets together
  - First, merge HD2017 to FLAGS2017 using an inner_join creating the data frame `hd_enroll`
  - Second, merge the data frame `hd_enroll` as the `x` table to the EFFY2017 data frame as the `y` table using a `left_join()`

Specific instructions and steps to follow and order to complete steps are given below in the section called **Specific instructions (steps to complete the problem set)**

IPEDS Data sources you will work with in this problem set

- IPEDS consists of about a dozen "survey components" (e.g., Institutional Characteristics, Completions, Fall Enrollment, etc.).
- Each of these survey components may contain several different data sets. For example, within Institutional Characteristics there are separate data sets for: "Directory information"; "Educational offerings, organization, services and athletic association"; "Student charges for academic year program"; etc.
- In this problem set, you are required to get data directly from the IPEDS website here.
  - Click on the link above
  - Select continue for all years and all surveys
  - Hover over the corresponding data file under the "data file" column and right click to save the url
  - [see section 6 "Reading in data from the web" of lecture 8]
- You will download the following data from the ipeds website under the "data file" column


IPEDS
Data Center Help Desk (866) 558-0658
Start over | Save session | Help | MAIN MENU
Complete Data Files
Provisional Release Data (Change)
Years & Surveys
All years | All surveys | Continue
Data files are available in ZIP format.

Year	Survey	Title	Data File	Stata Data File	Programs	Dictionary
2017	Institutional Characteristics	Directory information	HD2017	HD2017_STATA	SPSS, SAS, STATA	Dictionary
2017	Institutional Characteristics	Educational offerings, organization, services and athletic associations	IC2017	IC2017_STATA	SPSS, SAS, STATA	Dictionary
2017	Institutional Characteristics	Student charges for academic year programs	IC2017_AY	IC2017_AY_STATA	SPSS, SAS, STATA	Dictionary
2017	Institutional Characteristics	Student charges by program (vocational programs)	IC2017_PY	IC2017_PY_STATA	SPSS, SAS, STATA	Dictionary
2017	Institutional Characteristics	Response status for all surveys components	FLAGS2017	FLAGS2017_STATA	SPSS, SAS, STATA	Dictionary
2017	12-Month Enrollment	12-month unduplicated headcount: 2016-17	EFFY2017	EFFY2017_STATA	SPSS, SAS, STATA	Dictionary
2017	12-Month Enrollment	12-month instructional activity: 2016-17	EFIA2017	EFIA2017_STATA	SPSS, SAS, STATA	Dictionary
2017	12-Month Enrollment	Response status for all surveys components	FLAGS2017	FLAGS2017_STATA	SPSS, SAS, STATA	Dictionary
2017	Fall Enrollment	Race/ethnicity, gender, attendance status, and level of student: Fall 2017 *(Preliminary)*	EF2017A	EF2017A_STATA	SPSS, SAS, STATA	Dictionary

:
  - **HD2017**
  - **FLAGS2017**
  - **EFFY2017**
- Download the dictionary associated with each data file
  - Each dictionary (codebook) will contain general information about the data, a list of variable

names, variable descriptions, and some frequencies.

- You can also download the SAS, SPSS, or STATA scripts to assist you with labelling variables
  - This is not absolutely necessary, but downloading one of these scripts will be helpful for questions that require you to assign variable labels and value labels

General guidelines merging data and tidying data

- Guidelines to follow for tidying data [this will only be applicable for the data set effy]
  - So that it follows the three interrelated rules of a tidy data set:
    - Each variable must have its own column
    - Each observation must have its own row
    - Each value must have its own cell
  - After tidying the data, check that there is one observation per row and one variable per column
    - Run some quick checks (frequencies, check missing, etc.)
    - [see 4.3.1 "Student exercise: real-world example of spreading" of Lecture 6]
- Prior to merging, investigate the two data frames
  - Try to identify variable(s) that uniquely identify object or dataframe
  - Based on this investigation, decide which variable (or variables) is the "key" variable for the merge
  - You may find it helpful to print a few observations of the data frame (or view with the `View()` function) to help you understand data structure
- Join the two data frames
  - Note: for some joins, you may have to rename a "key" variable; you can do this prior to the join or within the join code [see section 3.2.2 of lecture 7]
- After joining, spend some time investigating the quality of the join
  - You **must** use an `anti_join()` to identify observations that did not merge

Specific instructions (steps to complete the problem set)

1. Set your working directory and run the code we created for you
   - This code: reads in IPEDS `hd` data set; changes variable names from upper-case to lower-case; keeps selected variables; adds value labels and variable labels; investigate variables and data patterns
   - It may be helpful to run one line at a time, just to see what we are doing to read-in and clean data

2. Read in FLAGS2017 data and assign it to object `flags`
   - Change variable names from upper-case to lower-case
     - Can follow the example we give above
   - Subset data to columns `unitid`, `stat_e12`, `lock_e12`, `prch_e12`, `idx_e12`, `pce12_f`, `imp_e12`
     - Essentially, we are keeping `unitid` and any variable that contains the text "e12"
     - The variable `stat_e12` gives us the response status for 12-month enrollment data (EFFY2017)
   - Add variable labels to all variables and add value labels to categorical variables
     - Note: This is where the codebook you download from IPEDS Data Center and the SAS/SPSS/Stata script (which is optional to download) will be useful
   - Conduct any investigations of the data you see fit [not more than 10-15 minutes]
3. Using `hd` as the `x` table and `flags` as your `y` table, perform an `inner_join()`, assigning the resulting object the name `hd_flags`. Basically, the resulting data frame will have one observation per unitid and

will have a variable `stat_e12` for the response status of 12-month enrollment data.

4. Read in 12-month enrollment data for the 2016-2017 academic year (headcount rather than institutional activity) and assign it to object `enroll`
   - Subset data to columns `unitid`, `effylev`, `efytotlt`

   - Label the variables and add value labels
   - Investigate the data frame (e.g., investigate specific variables, investigate which combination of variables uniquely identify observations)
   - Don't spend more than 10-15 minutes on this
   - Tidy data, assign tidy data to object `enroll_v2`
     - HINT: prior to tidying (but in the same line of code that tidies the data) you may want to create a new version of the variable effylev that is a character variable rather than a numeric variable.
5. Using `hd_flags` as the `x` table and `enroll_v2` as your `y` table, perform a `left_join()`, assigning the resulting object the name `hd_enroll`. Basically, the resulting data frame will have one observation per unitid and 12-month enrollment data with additional characteristics (sector, carnegie, city, etc).
   - Follow guidelines above for investigating data prior to merge, and investigating quality of merge (e.g., must use anti_join to investigate obs that don't merge); but don't spend more than 10-15 minutes investigating quality of merge [HINT: variable stat_e12 important for investigating obs that don't merge]

6. **Bonus** Question:
   - Use Ben Skinner's script to label variables and variable values to both the IC directory data (HD2016, FLAGS2017) and the 12-month enrollment headcount data (EFFY2017)
     - LINK

     - Note: this script assumes the data are already downloaded. If you have not downloaded the data, you can use his package to download data.
       * **This is not required**, you only need to use the add-ipeds-labels-in-r program for the bonus question.

Prior to submitting your homework, please do the following:

- Make sure that your R script runs without any errors before submitting
- Make sure to include your name on line three of the R script
- Use this naming convention for your R script "lastname_firstname_ps7"
- Comment out or delete code that you don't want Patricia to see when grading
  - e.g., you might choose to comment out (but not delete) lines of code that provided helpful insights about the data to avoid having many pages of output.
  - e.g., you might decide to delete (as opposed to comment out) lines of code that were part of your investigations but did not provide big insights about the data
  - These decisions are subjective; don't spend much time worrying about which lines to delete or comment out.