Problem Set 8

Overview:

In this problem set, you will be working on a shared repository with your homework group. We encourage you all to communicate with your group by creating issues on your shared repository.

We will not be asking for your git commands and output – you can simply use your command line interface. You are required to do your work in branches. You will only need to submit (via pushing to your Github repository) your R script and any data files you create.

The purpose of this problem set will be to give you more practice writing for loops. This problem set is based off of Ben Skinner's script to batch download Integrated Postsecondary Education Data System (IPEDS) files. The goal is to download HD data files for particular years (e.g., 2015-2018) from the IPEDS website, unzip those files, and load the .csv files of the data into R.

Part I: Setting up

Like the previous week, create a new private GitHub repository here for your group called `<team_name>_ps8` and initialize it with a `.gitignore` file.

1. Download and save the ipeds_file_list.txt file in your repository. Make sure it is in the main repository folder. Only one person in your group needs to add this file to the repository using the **master branch**. Once this file is added, everyone on the team should clone the repository to their local machines.
2. Create a new RStudio Project for this repository and complete all your work on a branch called `dev_<last_name>`.
3. Download and save the `lastname_script.R` file under "Weekly readings and assignments" >> Week 8.
   - Rename this file to your lastname and save it in your repository `<team_name>_ps8`.
4. Open the R script in your R Studio and add your header information on lines 1-8.
   - On lines 19 and 23 change the name of `data_lastname` to your lastname (e.g., `data_martin`). These lines will create the directory path and a sub-folder in your shared respository where you will save all your data files.
5. Run lines 1-101 and make sure you do not get any errors. Go through each line and make sure you understand what is going on.

Part II: Creating loops

Loop 1: Create a loop that prints the URL associated with each data file (e.g., `HD2015.zip`, `HD2015_Dict.zip`, `HD2015_Stata.zip`). Use the skeleton of the for loop on lines 107-120 to guide you through each step.

1. The URL associated with each data file should look something like this "https://nces.ed.gov/ipeds/datacenter/data/HD2015_Dict.zip". You can paste this URL in a browser and it should download a zip file for the HD2015 data dictionary.
2. You may want to use the functions `str_c()` and `writeLines()` to create and print out each URL.

Loop 2: Download the HD datasets using a loop.

1. You will use the `download.file()` function. The first argument `url` is a character string of the url where our data will be downloaded from. (**Hint: these are the URLs we created in the previous question**) The second argument `destfile` is the file path to where we want to save our downloaded data. (**Hint: this is the data directory we created on line 19**)
2. Follow the code and hints in the `<last_name>_script.R` script. You should end up with three zip files for every HD dataset in the `hd` vector. For example, for HD2018, you should have these files downloaded in your data folder: `HD2018.zip`, `HD2018_Dict.zip`, and `HD2018_Stata.zip`. And so on for the rest of the HD data.

Loop 3: Unzip data files.

1. All the files we downloaded are zip files. Instead of manually unzipping each one, we are going to create a for loop to unzip the files for us.

2. We use the `unzip` function to unzip all the files. The `zipfile` argument takes the file path where our zip file(s) are located plus the name of the zip file. The `unzip` argument takes the value `"unzip"`. The `exdir` takes the file path to where we want out zip files to be unzipped.

3. We will write a nested loop where the first for loop loops through the `hd` vector and the second for loop loops through the file suffixes (i.e., `""`, `"Dict"`, `"_Stata"`). Follow the code and hints in `<last_name>_script.R`. You should end up with unzipped files in your data folder.

Loop 4: Read in data and make all column names lowercase.

1. Now that we have all the unzipped data, let's create a loop to read in the .csv files for each HD dataset.

2. In this loop we will change the dataframe name to lowercase instead of uppercase, read in the csv files, and change the column names to lowercase.

3. Follow the code and hints in `<last_name>_script.R`. You should end up with dataframes in your environment for all HD file (`hd2015-hd2018`).

Part III: I got issues

1. Navigate to the issues tab for the **rclass2** repository here, then:
   - Create a new issue posting either a question you have or something new you learned about from reading Ch.2-5 of Advanced R
   - Respond to an issue that another student had posted

   Paste the links to the 2 issues you contributed to as a comment in `<last_name>_script.R`.

   Please make sure to close your issue within 2 weeks.

Part IV: Wrapping up

1. How much time did you spend on this problem set? Write your response as a comment in `<last_name>_script.R`.
2. Add your `<last_name>_script.R` file to the `dev_<last_name>` branch and make a commit. Switch back to the `master` branch and merge in your `dev_<last_name>` branch. Push your work to the remote.