MSDS 6306: Doing Data Science – Tidy Data

Live session Unit 05 assignment

Due: 1 hour before your 6th live session (February 12th, 2018)

Submission

ALL MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT. Formatting can be basic, but it should be easily human-readable. Unless otherwise stated, please enable {r, echo=TRUE} so your code is visible.

Questions

Backstory: Your client is expecting a baby soon. However, he is not sure what to name the child. Being out of the loop, he hires you to help him figure out popular names. He provides for you raw data in order to help you make a decision.

1. **Data Munging (30 points):** Utilize **yob2016.txt** for this question. This file is a series of popular children's names born in the year 2016 in the United States. It consists of three columns with a *first name*, a *gender*, and the *amount of children* given that name. However, the data is raw and will need cleaning to make it tidy and usable.

    a. First, import the .txt file into R so you can process it. Keep in mind this is not a CSV file. You might have to open the file to see what you're dealing with. Assign the resulting data frame to an object, **df**, that consists of three columns with human-readable column names for each.

    b. Display the summary and structure of **df**

    c. Your client tells you that there is a problem with the raw file. One name was entered twice and misspelled. The client cannot remember which name it is; there are thousands he saw! But he did mention he accidentally put three y's at the end of the name. Write an R command to figure out which name it is and display it.

    d. Upon finding the misspelled name, please remove this particular observation, as the client says it's redundant. Save the remaining dataset as an object: **y2016**

2. **Data Merging (30 points):** Utilize **yob2015.txt** for this question. This file is similar to yob2016, but contains names, gender, and total children given that name for the year 2015.

    a. Like 1a, please import the .txt file into R. Look at the file before you do. You might have to change some options to import it properly. Again, please give the dataframe human-readable column names. Assign the dataframe to **y2015**.

    b. Display the last ten rows in the dataframe. Describe something you find interesting about these 10 rows.

c. Merge **y2016** and **y2015** by your Name column; assign it to **final**. The client only cares about names that have data for both 2016 and 2015; there should be no NA values in either of your *number of children* rows after merging.

3. **Data Summary (30 points)**: Utilize your data frame object **final** for this part.

    a. Create a new column called "Total" in **final** that adds the *number of children* in 2015 and 2016 together. In those two years combined, how many people were given popular names?

    b. Sort the data by Total. What are the top 10 most popular names?

    c. The client is expecting a girl! Omit boys and give the top 10 most popular girl's names.

    d. Write these top 10 girl names and their Totals to a CSV file. Leave out the other columns entirely.

4. **Upload to GitHub (10 points)**: Push at minimum your RMarkdown for this homework assignment and a Codebook to one of your GitHub repositories (you might place this in a Homework repo like last week). The Codebook should contain a short definition of each object you create, and if creating multiple files, which file it is contained in. You are welcome and encouraged to add other files—just make sure you have a description and directions that are helpful for the grader.

Reminder

To complete this assignment, please submit **one** RMarkdown and matching HTML file at least one hour before your live session on February 12, 2018. You do not need to submit a link to your GitHub: just note where the assignment is (URL) in your RMarkdown file. Please make sure it is public and submit all files at the same time; only one submission is granted.