Homework 05: Instructions

Overview

For you part of the team project (e.g., symmetry span, go/no-go, demographic measures, etc), you will be responsible for managing all files for a project sub-task, for managing specific steps or stages of multiple task, or form managing some some combination of files as determined by your team members. Whether you manage all steps associated with preparing data for a sub-task or you manage only specific files, all team members should understand how the code files are structured from the initial script and data file that serves as the foundation to the final step of the R Markdown report, which depends upon all of the previous data-preparation scripts, data visualization scripts, and modeling scripts. You can think about the R Markdown report as the outside wrapper of an onion and each dependent script file as respective inner layers.

This homework assignment requires you to **create the script files necessary for the project base layer and the data preparation layers (not the data interpretation layers)** as outlined in the steps below. You should create these files in your personal version `<liaison-yourname>`.

A Hierarchy of Script Files

The concept of reproducible research and code is new for many students and is somewhat difficult to mentally represent. Let's start with conceptualizing a hierarchy of scripts that will result in reproducing all steps of the process from cleaning to model testing automatically.

Report layer

- report: `/report/PSYC166_team_<liaison>_report.Rmd`

Data Interpretation

- layer 5a: `/r/<subtask>/<subtask>_analyze.R` (1 or more, named accordingly)
- layer 5b: `/r/<subtask>/<subtask>_visualize.R` (1 or more, named accordingly)
- layer 5c: `/r/<subtask>/<subtask>_summarize.R` (1 or more named accordingly)

Data Processing

- layer 4: `/r/<subtask>/<subtask>_long_clean.R`
- layer 3: `/r/<subtask>/<subtask>_wide_to_long.R`
- layer 2: `/r/<subtask>/champ_all_waves_create_<subtask>_subset.R`

Base Script

- layer 1: `/r/champ_all_waves_initial_clean.R`

Outlining the Script Goals

The Base Script

(1) You have so far worked on an exercise to **read and initially clean the full data file** (a smaller practice subset anyway) and save it. You also worked on integrating these steps into a workflow model of scripts that perform simple tasks and write out files that you used by other scripts.

- That script named `/r/champ_all_waves_initial_clean.R`, was created for the purpose of producing a data file of the same name: `/data/champ_all_waves_initial_clean.Rds`
- In order to produce that data file, the script read `/data/raw/champ_all_waves_practice_subset.Rds` (which will later be the full file named `/r/champ_all_waves.Rds`).

Data Preparation

After loading libraries, one of the first lines of code in each script is to source an .R script file associate with a previous step. When R sources that script file from the previous step, one of the first lines of code in that file is to source a script file associated with the previous step, which also first sources a script from the previous step and so forth all the way down to the base script. After reaching that deepest layer, each script will execute code line-by-line until there is no code and will continue executing from base script all the way back up to the initial source. The scripts listed below help with that process.

(2) The next step is to **subset the full data** based on key variables (e.g., `id_subject`, etc.) that you will use to for merging subset files in addition to variable relevant to a specific data component (e.g., go/no-go, symmetry span, demographics, etc.) of the larger project.

- The sub-setting script is appropriately named:
  - `/r/champ_all_waves_create_<subtask>_subset.R`
- The script produces a data file named:
  - `/data/<subtask/<subtask>_subset_wide.Rds`
- In order to produce that data file, the script:
  - sources the previous step: `/r/champ_all_waves_initial_clean.R`
  - and reads data produced: `/data/champ_all_waves_initial_clean.Rds`

(3) The next step is to **transform/pivot the data from wide to long** in order to prepare the long version of the data file.

- That script for transforming from wide to long is appropriately named:
  - `/r/<subtask>/<subtask>_wide_to_long.R`
- The script produces a data file named:
  - `/data/<subtask/<subtask>_long.Rds`
- In order to transform the data, the script:
  - sources the previous step: `/r/champ_all_waves_create_<subtask>_subset.R`
  - and reads data to pivot: `/data/<subtask/<subtask>_subset_wide.Rds`

(4) The next step is to **read, clean, and prepare the long data** for processing by cleaning relevant variables, adding variables, dealing with `NA`s, etc. so that you can later produce data summaries.

- That script for cleaning and preparing the long file is appropriately named:
  - `/r/<subtask>/<subtask>_long_clean.R`
- The script produces a data file named:
  - `/data/<subtask/<subtask>_long_clean.Rds`
- In order to produce the prepared long data file, the script:
  - sources the previous step: `/r/<subtask>/<subtask>_wide_to_long.R`
  - and reads data to clean: `/data/<subtask/<subtask>_long.Rds`

Data Interpretation

Once you have a cleaned version of your long data file, you are ready to manipulate it in order to aggregate as appropriate for a specific purpose. After aggregating the data for a specific purpose, you should be read for the next steps involving further processing of the aggregated data. For example, you would likely (a) create tabular data summaries that group by predictors and outcome variables along with measures of central tendency and dispersion, (b) create data visualizations that will help communicate patterns in data and tell the story of the data, and (c) perform statistical or predictive models.

(5a) After aggregating data at the participant level, you will want to **create tabular data summaries** based on predictor and outcome variables.

- The script for aggregating and producing data summaries is appropriately named:
  - `/r/<subtask>/<subtask>_summarize.R`
- The script produces any data files that would represent data summary subsets:
  - e.g., `/data/<subtask>/<subtask>_summarize_<descriptive and meaningful name>.Rds`
- In order to aggregate the data and produce data summaries, the script:
  - sources the cleaned long file from the previous step: `/r/<subtask>/<subtask>_long_clean.R`
  - and reads data to summarize: `/data/<subtask/<subtask>_long_clean.Rds`

(5b) After describing data with tabular data summaries, you will **create data visualizations** from the data, likely the aggregate data from the data aggregation step.

- The script for aggregating and producing data summaries is appropriately named:
  - `/r/<subtask>/<subtask>_vizualize.R`
- The script produces any plots/figures
  - (e.g., `/figs/<subtask/<subtask>_<meaningful_figure_name>.png`)
- In order to produce data visualizations, the script:
  - sources the cleaned long file from the previous step: `/r/<subtask>/<subtask>_long_clean.R`
  - and reads data to visualize: `/data/<subtask/<subtask>_long_clean.Rds`

(5c) After describing data with tabular data summaries, you will **run machine-learning or statistical models** on the data.

- The script for analyzing/modeling data is appropriately named:
  - `/r/<subtask>/<subtask>_analyze.R`

- The script produces any data files that would:
  - summarize model parameters
    - e.g., `/data/<subtask>/<subtask>_model_<meaningful name based on predictor and outocome variables>.Rds`
  - or figure files to visualize the model
    - e.g., `/figs/<subtask>/<subtask>_model_<descriptive and meaningful plot name>.png`
- In order to model the data, the script:
  - sources the cleaned long file from the previous step:
    - `/r/<subtask>/<subtask>_long_clean.R`
  - and reads data to manipulate for modeling:
    - `/data/<subtask/<subtask>_long_clean.Rds`

(6) The report layer inherits all of the necessary processing steps outlined in executed by scripts.

- sources any script files on which the report depends;
- reads in summarized data frames to reference as in-line R objects or to present in a table
- reads in relevant model parameters to reference as in-line R objects or to present in a table
- reads in figure files to include

Future Directions

At a later point, you should work on the scripts for the **data interpretation** layer. In addition, the **reporting lead** should take on the responsibility of creating the R Markdown file for the final report layer and integrate the written components contributed by the team members. Similarly, the **coding lead** should take on the responsibility of organizing the final versions of the individual files create by team members for each step of a project subset, which will later be migrated to the team's GitHub repository. The **project manager** should take on the responsibility of created a timeline for team members to complete their tasks, motivate the team to complete them, and ensure that the tasks are completed in a timely manner.