Instructions: Data portfolio

Near the end of the course, you will submit a data portfolio demonstrating your implementation of statistics in R across the topics we have covered. You will be provided a dataset which includes a set of (simulated) variables pertaining to a certain subject. Use this dataset to formulate a research question, which you will seek to answer by carrying out the various types of statistical analyses we explore in this course. This is meant to be a cumulative project, which you will continuously build upon with the new set of statistical procedures that you add to your scientist toolkit each week.

The final portfolio will count for 30% of your grade and will consist of two components:

1) A **poster**, as suitable for presentation at a scientific conference
2) The **R script** used to explore your dataset and generate results shown in the poster

Both components will be submitted via Canvas (Assignments >> Data portfolio), **due October 16th, 2020 before 22:00.**

Datasets

Simulated datasets and an accompanying data dictionary with a description of the study design and variables can be found on Canvas (Modules >> Data portfolio). All members of a student working group (SWG) will be assigned the same dataset and multiple groups may have the same dataset. Your dataset assignment will be announced via Canvas/email. ***Important:*** although multiple people will work with the same dataset, and you may consult with your fellow students on statistical questions and R code, **this is an individual – NOT group – assignment!** That means you need to come up with your own research question and turn in your own poster and R script. Although multiple people may come up with similar research questions, copying work from other students will be treated as plagiarism.

Choosing your research question

In principle, your research question can be about any relationship you find interesting between two (or more) variables in your dataset. However, it will be most helpful to formulate a broad question of interest so that you can conduct multiple (sub-)analyses under the umbrella of your research project. You should be able to generate one or more testable, falsifiable hypotheses.

Conducting analyses

Each week, you will learn how to carry out a new set of statistical analyses in R during the practicals and then apply these analyses in a structured way in the assignment. The data portfolio provides an unstructured challenge to apply these same tools to a dataset like you would experience in your own research. Each week (*of weeks 1-5; no multilevel analysis*), choose a statistical technique from the current unit (i.e., week) that would be suitable to apply to a set of variables in your dataset (ideally an analysis that will directly connect to your research question, but given the number of different analyses, some will be more relevant than others). Use the R code from the practicals, assignments, and R cheat sheet to help you carry out this analysis, ***making sure to save all your R code in the script file you will submit.*** Although there is only one deadline for the final data portfolio, it will serve you well to work on these analyses progressively each week while the information is still fresh!

Poster requirements

The submitted poster file should be an A0-sized PowerPoint or PDF document created using the template on Canvas (Modules >> Data portfolio >> SiN 2020 – data portfolio template). Only submit the PPT/PDF file – **do not actually print your poster** (a very big, very expensive piece of paper)!

Your poster should very clearly convey the design of your study, your research question, the methods you used to answer it, and the results (and interpretations) of your analyses in a visually appealing way. Large graphics and concise text (bullet points) are best - try to boil each section down into a couple of key points. Your poster does not need to (and should not) include every single analysis you ever ran, but should highlight the most important analyses; the ones needed to answer the research question.

The poster should be suitable for presentation at a scientific conference (though we're using fake data so don't submit it to one!), including the sections:

1) Background
2) Methods
3) Results, including
    a. 1-2 tables (usually one for descriptive statistics)
    b. 1-3 figures (highlighting your main results)
4) Discussion

You can move around the different elements of the template and change the formatting, but make sure all of these sections are present in your poster. For tips on designing good scientific posters, see https://guides.nyu.edu/posters and for some examples of existing posters, see https://phdposters.com/gallery (or search many available online).

R script requirements

Unlike the poster, your R script **should** include every analysis that you ran throughout this project: data exploration, descriptive statistics, assumption testing, subgroup- or within-time comparisons, etc. At minimum, your R script should include **at least one** type of analysis from each of weeks 1-5, e.g.:

1) Week 1 – Descriptive statistics (mean, SD, skew, kurtosis, normality tests)
2) Week 2 – Correlation (Pearson's, Spearman's) or simple regression
3) Week 3 – Between-subject comparisons (independent t-test, Wilcoxon, ANOVA, Welch, Kruskall-Wallis)
4) Week 4 – Within-subject comparisons (paired t-test, Wilcoxon, repeated measures ANOVA, factorial ANOVA, Friedman's)
5) Week 5 – Multiple regression, regression with dummy coding, regression with interaction

Each dataset contains an array of variables suited for carrying out these different types of analyses, although you may need to run some analyses that are not directly relevant to your research question/poster (but informative as descriptive statistics on your data sample!). *Regardless, include all analyses you conduct in the R script you submit, not only the ones pertaining to your main research question!*

Your R script should also be **well-documented**, with clear and sufficient commentary so that an outside reader (i.e., us: the people grading your R-script!) with no knowledge of your project could follow everything that is happening in the script.

Grading rubric

A grading rubric can be found on Canvas (Modules >> Data portfolio).