Midterm project

Independence

In contrast to Homework assignments, you must work completely independently on this project – do not discuss your approach, your code, or your results with any other students, and do not use the discussion board for questions related to this project. If questions do arise, please email the instructor and lead TAs.

Context

At this point, we've covered Building Blocks (topic_building_blocks.html), Data Wrangling I (topic_data_wrangling_i.html), Visualization and EDA (topic_visualization_and_eda.html), and Data Wrangling II (topic_data_wrangling_ii.html). These three topics give a broad introduction into the commonly-used tools of data science, and are the main focus of this project.

Due date

Due: October 25 at 4:00pm.

Reproducibility

The course's emphasis on workflow – especially the use git and GitHub for reproducibility, R Projects to organize your work, R Markdown to write reproducible reports, relative paths to load data from local files, and reasonable naming structures for your files – will be reflected in your Midterm Project submission.

To that end:

- create a **private** GitHub repo + local R Project; we suggest naming this repo / directory `p8105_mtp_YOURUNI` (e.g. `p8105_mtp_ajg2202` for Jeff), but that's not required
    - non-private repos will be treated as inconsistent with the independent work requirement and as violations of the academic integrity policy
- add the GitHub user "bst-p8105" as a collaborator on the project, which will give us (and only us) access to your repo
- create a single .Rmd file named `p8105_mtp_YOURUNI.Rmd` that renders to `github_document`
- submit a link to your repo via Courseworks

We will assess adherence to the instructions above and whether we are able to knit your .Rmd in the grading of this project. Adherence to appropriate styling and clarity of code will be assessed. This project includes figures; the readability of your embedded plots (e.g. font sizes, axis labels, titles) will be assessed.

Deliverable

For this project, you should write a report describing your work in a way that targets a reasonably sophisticated collaborator – not an expert data scientist, but an interested observer. Structure your report to include an introduction and four sections corresponding to the problems below. Write in a reproducible way (e.g. using inline R code where necessary) and include relevant code chunks and their output. Include only relevant information, and adhere to a strict-500 word limit (this excludes figures and tables, code chunks, inline code, YAML, and other non-text elements). You can check your word count using `wordcountaddin::text_stats("p8105_mtp_YOURUNI.Rmd")` ; installation instructions can be found on the `wordcountaddin` package website (https://github.com/benmarwick/wordcountaddin). We'll use the "koRpus" count.

Data

In 2018, Nature Scientific Reports published an article (https://www.nature.com/articles/s41598-018-21625-1) by Australian researchers published describing the relationship between posture and "enlarged protuberances", especially among younger subjects. This was later reported in the popular press (https://www.washingtonpost.com/nation/2019/06/20/horns-are-growing-young-peoples-skulls-phone-use-is-blame-research-suggests/) in articles drawing the conclusion that phone usage was leading to "horns" growing in the back of millenials' heads. Later articles were more skeptical (https://www.nytimes.com/2019/06/20/health/horns-cellphones-bones.html), and a limited author correction (https://www.nature.com/articles/s41598-019-49153-6) was published in Nature.

The data for this project (data/p8105_mtp_data.xlsx) are those that accompany the author correction.

Problems

Problem 1 – Data.

Import and clean the data. Format the data to use appropriate variable names; fill in missing values with data where appropriate (as indicated in the header information); create character and ordered factors for categorical variables.

Briefly describe the data cleaning process and the resulting dataset, identifying key variables based on your understanding of the original scientific report. How many participants are included? What is the age and gender distribution (a human-readable table may help here)?

Note (but don't correct) issues in the available data – in particular, whether categorical variables in the dataset correctly implement the definitions based on underlying continuous variables. Use tables, figures, or specific examples (i.e. data for particular subjects) as needed to illustrate these issues.

Problem 2 – Visualization.

In the original scientific report, Figures 3 and 4 show data or derived quantities. Both are flawed. Figure 3 shows only the mean and standard deviation for FHP, but does not show the distribution of the underlying data. Figure 4 shows the number of participants in each age and sex group who have an enlarged EOP (based on categorical EOP Size – groups 0 and 1 vs groups 2, 3, 4, and 5). However, the number of participants in each age and sex group was controlled by the researchers, so the *number* with enlarged EOP in each group is not as informative as the *rate* of enlarged EOP in each group. Create a two-panel figure that contains improved versions of both of these.

Although the authors are interested in how FHP size, age, and sex affect EOP size, no figure contains each of these. Create a 2 x 5 collection of panels, which show the association between FHP size and EOP size in each age and sex group.

Comment on your plots with respect to the scientific question of interest.

Problem 3 – Reproducing reported results.

Are the authors' stated sample sizes in each age group consistent with the data you have available?

Are the reported mean and standard deviations for FHP size consistent with the data you have available?

The authors find "the prevalence of EEOP to be 33% of the study population". What is the definition of EEOP, and what variables can you use to evaluate this claim? Is the finding consistent with the data available to you?

FHP is noted to be more common in older subjects, with "FHP >40 mm observed frequently (34.5%) in the over 60s cases". Are the broad trends and specific values consistent with your data?

Problem 4 – Discussion.

Summarize your results, the quality of the data analysis / presentation of results in the original report, and comment on the conclusions of the reports' authors. Do you think the data provide evidence that cell phones are causing horn growth? What other data would you like to have to address this hypothesis?

Some past examples

Last year's assignment was more open-ended, and approaches varied widely. However, some things we look for are clearly-written code, conducting necessary exploratory analyses, and documenting results in a concise report. These submissions from last year were strong in each of those areas.

- Project 1 (https://github.com/Raarbiarsan1899/p8105_mtp_zf2213)
- Project 2 (https://github.com/sal2222/p8105_mtp_sal2222)
- Project 3 (https://github.com/tiffanyxi/p8105_mtp_hx2263)