Project 1: EDA

36-600

The goal of this project is to present an exploratory analysis of a dataset, utilizing the basic summarization and visualization tools at your disposal in `R`.

Your `HTML` file (generated by knitting an `R Markdown` source file) should be uploaded to Canvas by Wednesday, September 25th, at 11:59 PM. There is no limit on the length here: the report should try to find that sweet spot between being so short that it leaves important information out and so long that it becomes as boring as a laundry list. Think of this as summarizing the data for a distracted advisor: hit the highlights, but at the same time try to be complete.

Data

You will examine the dataset `cmu-sleep.csv`, which is linked to from the `Canvas` Project 1 module page. A description of the data are given here:

https://cmustatistics.github.io/data-repository/psychology/cmu-sleep.html (https://cmustatistics.github.io/data-repository/psychology/cmu-sleep.html)

In this project, *you can assume your audience knows what each variable represents*; don't add text explaining defining the variables. Here, you just have to craft a story about how the data are (empirically) distributed, and how variables are associated, etc., while referring to each variable by just its name.

Note: assume that these data might be used in the future in supervised learning exercises, and that the response variable is `term_gpa`. This may affect some of your visualization choices!

Pointers

Exploratory data analysis is an open-ended exercise, and no two people working with the same data will produce the same results. This is a feature, not a bug. However, there are some things that everyone will want to keep in mind:

- What questions would you ask if someone was to present a dataset to you? Assume your audience would like to have that information. Don't overthink this…what is basic information about a dataset that everyone should know? (Perhaps start by determining the number of rows and columns, and doing a summary.)
- EDA is not a (linear) recording of all that you type when examining data. You might go in particular directions and find that they are not, in your mind, informative about the data (or perhaps better put, they do not yield information that would necessarily be interesting to your audience).
- Keep in mind any tips about plotting given in the class notes! (If they are indeed applicable here. They might not be…)
- Don't try to go (too far) beyond what we have done in class (if you choose to go beyond what we've done in class at all). EDA is not meant to be "fancy" and you won't win prizes for pushing the envelope on methods and algorithms. In the end, your audience needs to walk away with an understanding of the data, not marveling over that javascript-based web plugin you (shouldn't have) created.

- Faceted plots generally look better than a vertical column of plots defined one after the other. Use faceting when you can.
- You are allowed to remove uninformative columns, if they exist, but you should state clearly that you did.
- You are allowed to remove outliers that affect visualization and summarization, but you should state clearly that you did.
- If data are missing, you should indicate via words or directly via code how many data are missing, and you should be prepared to remove the rows with missing data.
- We note above that `term_gpa` is the response variable. That means that if you choose to make scatter plots or grouped box plots, the most relevant ones would have `term_gpa` along the y-axis. (If you want to do scatter plots, etc., of the predictor variables against each other, just use `ggpairs` …but it might get messy…I don't know. Just be sure to only include plots that you can explain, and which are easy to interpret [hence my comment by `ggpairs` and messiness…is it just too much? maybe a `corrplot` would be just as good here].)
- Comment on whether there appears to be clear associations between `term_gpa` and the predictor variables…if there are, then that implies that statistical learning done in the native space of the data will uncover useful associations and thus be able to produce models that can predict `term_gpa` with some degree of precision better than that afforded by random guessing.
- Every bit of output you show should be reproducible by your audience. What that really means is, show all *relevant* code.
- Regarding *relevant* code: try not to show code that is not relevant to the story. For instance, in theory your audience does not need to see the specifics of inputting data to `R`. If you choose to "hide" code, do it by using code chunks with `echo=FALSE` in the chunk definition, e.g., ```` ```{r echo=FALSE} ````. The chunk will run during knitting but the code will not be echoed to the output `HTML` file.
- If in doubt, email me. If you bring up an important point that I've forgotten to make here, I will bring that point to the attention of the class.
- As a very last point: I've made this project somewhat free-form, because I'd prefer that you explore options rather than simply go through a checklist of plots. That said, many of you (most of you…all of you?) will make sub-optimal choices here and there…and we will provide feedback on those. Don't worry about making them…just go forth and build intuition about the dataset!