Individual Data Science Project

In this project, you will follow the format of a short scientific manuscript without extended literature following the guidelines below. The guidelines are similar to those you might have to follow to submit a paper to a scientific journal. In these cases, if you don't follow the guidelines, your paper can be rejected. In our case, if you don't follow the guidelines, credit will be reduced.

Instructions

Your paper must contain no less than 3 and no more than 5 text pages in Times New Roman font, 12 point, single spaced. Page margins may not exceed 1.25in. on all sides. Your list of references will appear after the discussion section, but *does not* count toward your page count. Likewise, figures and tables *do not* count toward the page count.

You should perform **at least two** different kinds of analysis. These can involve any of the analysis techniques we have discussed (or plan to discuss). For example, you might be trying to predict an effect based on variables in your dataset—you might create a baseline model, then create a more sophisticated model and compare the two. Or, you might perform a clustering analysis to establish groups, then do some further downstream analysis on those groups. Remember that some models make assumptions that you need to test for—you must include tests to validate these assumptions as well, but those tests *do not* count as a separate analysis.

You must provide appropriate figures to communicate your analysis and results. **At least one** figure must be provided *per analysis* (but no more than three). Remember that figures don't add to the page count of your paper. You may also provide tables to aggregate results; these also do not count in the page length. If you have different options, choose the most relevant/informative and legible one—the one that communicates your results best.

On the following pages there is a description for each section that must appear in your paper. You must order your sections exactly as shown. Note that tables and figures come *after* the references. (This is common when publishing in a journal—the journal editors will choose where to put your figures/tables in the final publication.)

Title: Give a relevant title to reflect the questions you'll be addressing in your project.

The author name (yours) goes here.

Abstract

1. background and research question (≈ 2 sentences), 2. statistical analyses used ($\approx 1-2$ sentences), 3. main results ($\approx 1-2$ sentences), 4. conclusions and implications ($\approx 1-2$ sentences). NOTE: When summarizing your main results in this abstract, do not report your p-values!

Introduction

Give a little background (e.g., theory, scientific or social issue) so the reader will understand your choice of analysis techniques. State and explain your objectives and predictions (be specific: is it a difference [in mean, variance, or distribution?], a correlation, are you predicting an outcome, are you showing group relationships, etc.?).

Methods

Briefly describe the origin of your study dataset. Be sure to provide citations for the relevant references.

Write a paragraph or two about your analyses, explaining your choices. Indicate your significance level if you use a null hypothesis approach (i.e., p-values). If your analyses do not involve p-values but some other metric, explain the metric and what a "successful" value would be, etc. Also specify what you will be using in the results: confidence intervals or standard errors.

Mention the language (R / Python) that you will be using run your analyses AND give the names (italicized or in parentheses) of the R or Python functions (and associated packages if downloaded) you'll be using for the statistical tests and/or models. You do not need to indicate the R functions such as *mean* or *sqrt*. Your code should be made available either in supplemental files (you will submit these with your project paper) or in an online repository (i.e. Github, Gitlab, BitBucket, etc.). Your methods section should mention how you are making the code available (refer the reader to supplemental material, or give a link if it is online).

Do not forget to talk about the possible assumptions that need to be met to run your analyses. Explain how you handled assumptions that were not met if any. Talk about how you handled things like missing values in this section as well, or if you performed a data transformation, give details here.

Results

Report estimates (e.g., mean, slope) with its uncertainty (e.g., SE, 95% CI), the results of statistical analyses (the test statistic, df and p-value). For predictive models, report your prediction metrics (i.e. accuracy, sensitivity, specificity, Area Under the ROC-curve, or whatever metric best describes your model's performance) along with a baseline model so the reader can determine how effective your model was. For cluster analyses, just state the facts (i.e. 300 individuals were clustered into 5 groups using K-means with 5 variables (which you would describe)).

Describe your graphs. This is a section that is only descriptive (e.g., higher-smaller, increase-decrease). You can say what a graph or the result of a statistical analysis suggests but do not give interpretations.

Remember when reporting numeric values: Use a reasonable and *consistent* number of decimal places (2 or 3 is usually enough) and *always* report your units if the values have a unit.

Discussion

Give the *interpretation* of your results and graphs (no p-value, etc. here!). Did you obtain what you predicted? Why or why not? Make a *conclusion* and explain what the *implications* are for the discipline into which your data belong.

If you had problems with your data, you can discuss them too. Give some perspectives for future research/further analyses. For example, if you set out to predict an outcome from your dataset, but it turned out that your best model was no better than "guessing", discuss why you believe that happened. Sometimes the discussion section is the most interesting part — this tells other researchers what to be on the lookout for, and sometimes points out new questions that can lead to new research.

Give at least a sentence or two (no more than a paragraph) to discuss future work you would like to do now that you have done this analysis. What ideas did it give you, or what new questions did it raise?

References

This section should start on a new page.

List your references here. Use the IEEE reference format. Information and examples can be found here:
https://pitt.libguides.com/citationhelp/ieee

Tables

This section should start on a new page.

If you have any… Place a caption for the table above each. Tables may only contain horizontal lines at the top and (optionally) bottom, and separating the headings from the data. No vertical lines allowed (i.e. no grid). If you don't have any tables, omit this section completely.

Figures

This section should start on a new page.

This is where you will place your graphs, or heatmaps, or other figures. If a legend is needed to understand the figure, do not forget to add one on the graph or in the caption of the figure. The caption should be below the figure (unlike tables). Do not use a title on the graph/figure. It is the caption's job to explain what the reader is looking at. The caption should be descriptive but concise (i.e. use as few words as you can to describe what the figure shows).

Validation plots (used, for example, to validate assumptions) will not count toward the number of required figures (nor do they count against the maximum figure count) as they offer no insight into your scientific question.

Supplemental material

This is not actually a section in your paper, but additional files that you submit along with it. If you have created additional spreadsheets, or if you have more figures than you are allowed to put in the manuscript, these can be published as supplemental material. *You can reference these in your writing.* The supplemental material is delivered as additional files that are submitted with your project.

All the code that you used to complete your analysis should be included in the supplementary material or made available in an online repository (i.e. Github, Bitbucket, Gitlab, etc.). Be sure to comment your code enough for an interested reader to understand what it does. Remove any sections of commented-out or inactive code.