Term Project–Guidelines

STAT 4690–Applied Multivariate Analysis

The main objective of the term project is to **apply the multivariate methods** we discussed in class on a real data analysis. Secondary objectives include:

- Improve your proficiency in `R`.
- Practice your communication skills (both oral and written).
- Foster curiosity.

This project is to be completed individually. You will be assessed on both a short in-class presentation of your results, as well as a written report.

Finding a dataset

The first step is to find an interesting dataset for the analysis. The dataset should be amenable to multivariate statistical methods, but ideally it should be of interest *to you*. Here are some repositories where you can find datasets, but you are definitely not required to use one of these websites:

- ICPSR: https://www.icpsr.umich.edu/icpsrweb/ICPSR/
- DASL: https://dasl.datadescription.com/
- TCGA: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php
- Kaggle: https://www.kaggle.com/datasets

It is important to validate your dataset with me before starting your analysis. You may send me a short description of the data, or you can generate a short description through `R` (e.g. see the functions `utils::str` or `tibble::glimpse`). The deadline for finding a dataset is **October 31st at 8PM**.

Data analysis

For the data analysis, you should start by exploring the dataset through the use of summary statistics and data visualizations. This exploration should logically lead you to a (scientific) hypothesis about the data and a choice of statistical methods to investigate this hypothesis.

The analysis does not need to be restricted to methods presented in class; however, your analysis should be centered around the use of multivariate methods that we discussed in class.

You should provide justification for the selected methods. In particular, you should pay special attention to the assumptions of each method and explore the extent to which they are met in your data.

The analysis should be carried out using the statistical software `R`, if possible.

In-class presentation

The **last two lectures of the semester** (i.e. December 4$^{\text{th}}$ and 6$^{\text{th}}$) will be devoted to in-class presentations. Each student will have 10 minutes (8 minutes for the presentation and 2 minutes for questions), and the order of presenting will be determined randomly.

You are required to prepare slides as a visual aid for your presentation. The goal of the presentation is to share your exploratory results, your main research question, and any preliminary results you may want to

discuss. In particular, your analysis does not need to be complete at this point. Indeed, you should view the oral presentation as an opportunity to receive feedback from me and your colleagues that you can then incorporate into your analysis.

You should send your slides to me at least one hour before the lecture during which you are scheduled to give your presentation. The only file formats that will be accepted are PDF and HTML (no Powerpoint or Keynote file).

Written report

Finally, you should submit a final written report **no later than December 20^th at 8PM**. This report should be between no more than 10 pages long (excluding tables and graphs), and should contain:

- Your exploratory results (e.g. summary statistics and data visualizations), including a description of your dataset and where it comes from.
- Your research question.
- A description of the statistical methods used to address the question, including the necessary justification.
- Results and discussion.
- Bibliography (if needed).

The report needs to be prepared electronically (i.e. no handwritten report) and submitted in PDF form, but you are free to use whatever tool to create the report (e.g. Word, Latex, RMarkdown, Lyx).

Evaluation criteria

Oral presentation (10% of your final grade)

- Provided sufficient background information
- Clearly described study objectives and methods.
- Presented results in appropriate detail.
- Presentation was interesting and engaging.
- Presentation respected the alloted time.

Written report (30% of your final grade)

Your mark will reflect the level of completeness and effort put into each of the following:

1. *Introduction*: summarized the data; provided relevant literature review; stated purpose of the research clearly
2. *Methods*: role of each method clearly stated; methods and assumptions described accurately
3. *Results*: results accurately stated; research question adequately answered
4. *Discussion/Conclusion*: results clearly and completely summarized; limitations and/or concerns stated
5. *General Considerations*: ideas presented in logical order; sections well-organized; no grammatical, spelling and punctuation errors; appropriate level of details results clearly and completely summarized; limitations and/or concerns stated

Bonus marks will be awarded for any challenging feature of the data that is appropriately accounted for in your analysis (e.g. missing data, correlated data, skewed and/or heavy tailed distributions).