AMS 597 Statistical Computing

Group Project Requirements

Instructor: Silvia Sharna

Group Selection:

- Please form groups with 4-5 members.
- For those who find it difficult to form a group, **please email me latest Friday, March 14, 2025 by 5:00PM**. Late request(s) will be declined.
  - Then I will randomly assign those individuals to different groups.
- Once you decide your group (or you are assigned to a group by the instructor), one member from each group will be required to email the instructor and give cc to the rest of the group members along with both the TAs.
  - The group member who will be sending this email, will be considered the group representative and will be responsible to submit the required materials in the Brightspace upon completion of the project. Hence, choose your representative wisely!
  - Once you send the email, it becomes your FINAL group. No group modification is possible after that.
  - You will then be assigned with a group number.

Dataset Selection:

- Find a real dataset (preferably messy and dirty; so that you have the chance to show your data cleaning proficiency) which has $n \geq 100$ samples and $p \geq 20$ variables.
- Although your dataset contains $p \geq 20$ variables, not all variables need to be included in the analysis.
- The variables should include both categorical and continuous type.
- The dataset can be from any domain (e.g., healthcare, finance, social sciences, etc.).
- Provide a brief description of the dataset, including its source and relevance to the research question.

Data Preprocessing:

- Perform necessary data cleaning (e.g., handling missing values, outliers, etc.).
- Conduct exploratory data analysis (EDA) using advanced visualization techniques (e.g., ggplot2, plotly, etc.).
- Transform variables if needed (e.g., scaling, encoding categorical variables, etc.).

Statistical Analysis:

- The research should be addressing THREE different questions (but can be related) of interest. That is, it SHOULD NOT be using two different methods to answer the same scientific question.

- The first research question can be answered using any of the methods that we have learned in class or more advanced methods.
- The rest of the two research questions should use **two different advanced statistical methods** such as:
    - Principal Component Analysis (PCA) or Factor Analysis.
    - Clustering (e.g., k-means, hierarchical clustering).
    - Resampling method.
    - GLM
    - Analysis of multi factor experiments
    - Mixed effect model
    - Regression modeling (e.g., linear regression, logistic regression, Lasso , Ridge).
    - Time series analysis (if applicable).
    - Machine learning techniques (e.g., decision trees, random forests, etc.).
    - Justify the choice of methods based on the dataset and research question.
- You will analyze your data using R only.

Model Evaluation:

- Evaluate the performance of your models using appropriate metrics (e.g., RMSE, AUC, accuracy, etc.).
- Interpret the results and discuss their implications.

Reproducibility and Code Quality:

- Write clean, well-documented R code using functions, loops, and vectorized operations where appropriate.
- Use only **R Markdown** to create a reproducible report that includes:
    - Code chunks.
    - Visualizations.
    - Interpretations of results.
- Ensure the report is well-organized and easy to follow.

Report Writing:

- Write a report (**maximum 25 pages** excluding appendices (shorter is fine), font size ≥ 11).
- The report should be self-contained. Submit your report as a pdf file. Be sure to include:
    - Introduction and research question.
    - Description of the dataset.
    - Methodology and implementation details.
    - Results and discussion.
    - Conclusion and limitations.
- Include your R source code, together with the data. If the data is large, use SBU google drive to upload the data. Your R source code can be saved as .R file that the instructor can open with RStudio. If you use R markdown, submit the markdown source file (i.e., the .Rmd file) together with the html, pdf etc.

- Make sure your code include comments describing the purpose of each chunk of code (i.e., what each chunk of the code is trying to analyze and should be clear from the comments included).
- Each group will also prepare and submit presentation slides (**maximum 15 slides**) summarizing your project.
- The slides will be shared with the rest of the class.

Submission Deadline (Read this section carefully!): April 16, 2025 by 5:00 PM

- **Your group representative** is required to submit both the project report (pdf) and presentation slides on 04/16/2025 (Wednesday) by 5:00 PM EST in Brightspace.
- **The group representative is also responsible to submit all the project materials** (report, slides, R codes and outputs) to the instructor (silvia.sharna@stonybrook.edu) and cc the TAs and the other group members.
- **Finally, each student is required to submit the report (pdf) on Brightspace**. If you do not submit the report on Brightspace, there will be a penalty on your project score.
- In the Subject line of the email, type “AMS 597 Spring 2025 Group XX Project", where XX will be replaced by your assigned group number.

Presentation:

- Presentations will start on 4/18/2025 with a hope that we will be able to wrap up everything by 5/9/2025.
  - But in case we are not able to, then we will be meeting during the Final Exam day which is Wednesday, May 21 from 11:15am 1:45 pm.
- Each group should prepare a 20-minute presentation.
- We have 86 students in total. Total number of groups will be in between (18 to 23).
- **We plan to do 6 presentations each day**, but the **7th group should always be ready** in case we can accommodate them within the duration of the assigned class time.
- Be prepared to answer questions from the instructor, TAs and peers.

Some Potentially Useful Data Archives/Repositories:

- You are encouraged to find your own real dataset. If you have trouble finding your own dataset, you can check the publicly available datasets from the following repositories:
  - Data and Story Library (DASL): http://lib.stat.cmu.edu/DASL/
  - NIST Statistical Reference Datasets: http://www.itl.nist.gov/div898/strd/
  - UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
  - UMASS Datasets: http://www.umass.edu/statdata/statdata/data/
  - DataSF: https://datasf.org/
  - Sports Statistics: https://sports-statistics.com/sports-data/
  - Other Data Repositories: https://www.kdnuggets.com/datasets/index.html

Grading Criteria:

- Are the report and presentation slides well-organized?
- Is supporting computer output provided (in edited form, that is, edit out all the extraneous information in the report)? R must be used for model fitting and plotting.
- Is the model appropriate for the design and questions of interest? Have you checked the assumptions?
- Are correct interpretations given for the parameters in the model?
- Are conclusions drawn from the model correct and do they answer the question of interest?
- Are the research questions interesting and non-trivial?
- Can the instructor reproduce the results reported using the code the group provided if the model is correct?

Group Synergy:

- If you have concerns with non-contributing members and are not able to resolve within the group, please speak to the instructor immediately (Please **DO NOT** wait till project due date).
- There will be an optional peer evaluation for groups with potential group synergistic issue:
    - For the **groups with non-contributing members**, for each member of the group, fill in the peer evaluation in the scale of 0-100%, how much each of the other group member should get from the group project score (e.g., if the group gets 24/30 and member A is given 60% by member B, 50% by member C, 80% by member D and 70% by member E and member A evaluates own-self 100%, the final score for member A will be 17.28/30). The instructor will arrange for zoom meeting with these groups to ensure fair evaluation.

*****Note:** Plagiarism or the use of pre-existing analyses without proper attribution will result in a failing grade. Be creative and demonstrate your understanding of statistical computing in R!