AMS 325.20: Computing and Programming Fundamentals

Fall 2025

Final Project Requirements

Due Date: 11/23/2025 , Sunday by 11:59 PM

General Instructions:

- Each group should have 3-4 members.
- For those who find it difficult to form a group, please email me at the latest, Friday 10/24/2025 by 5:00 PM. Then I will randomly assign those individuals to different groups.
- Late requests will be declined.

Dataset Criteria:

- Find a real dataset which has n ≥ 100 samples and p ≥ 8 variables not including ID columns or irrelevant metadata.
- Although your dataset contains p ≥ 8 variables, not all variables need to be included in the analysis.
- The dataset can be from any domain (e.g., healthcare, finance, social sciences, etc.).You may use publicly available datasets (e.g., from Kaggle, UCI Machine Learning Repository, data.gov, etc.).
- It is highly recommended to avoid datasets that are too simple (e.g., classic toy datasets like Iris or Titanic) or too complex.
- One student from each group must email the instructor and copy the TAs with the dataset/link of the dataset, names and SBU IDs of your group members (along with herself/himself), cc'ing the rest of the group members by no later than Friday 10/26/2025 by 11:59 PM.
    - This student will be considered as the group representative and will be required to submit all the project materials on behalf of the group.
    - Other group members will be required to submit the report (.pdf file) only.
    - Once you send the email, it becomes your FINAL group. No group modification is possible after that.
- Once I get the dataset info along with the group info from the representative, I will assign a group number for each group (first come first serve!).

Analysis Scope:

You are required to analyze your data in R (RMarkdown) or Python (Jupyter Notebook) only. Your project should include the following components:

1. Introduction & Objective

- Briefly describe the dataset and explain the goal(s) of your analysis.
- The research should be addressing Two different questions (but can be related) of interest. That is, it SHOULD NOT be using two different methods to answer the same scientific question.

2. Data Cleaning & Preprocessing

- If your dataset is not clean, then you will have to handle - missing values, incorrect data types, and outliers if any.
- Provide a brief justification of any data cleaning process that you have performed (if any).

3. Exploratory Data Analysis (EDA)

- Summarize the data using descriptive statistics.
- Create at least 5 meaningful visualizations (e.g., histograms, box plots, scatter plots, correlation heatmaps).
- Interpret trends, patterns, and relationships between variables.

4. Modeling or Statistical Inference

- Apply at least 3 methods of basic modeling that we are learning in the class (e.g., linear regression, logistic regression, KNN, Decision Tree Classifier, Random Forest Classifier, K- Means Clustering, Principal Component Analysis, or hypothesis testing).
- Clearly explain what the model is doing, why you chose it, and what the results mean in the context of your data.

5. Model Evaluation

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

6. 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:
  - o Code chunks.
  - o Visualizations.
  - o Interpretations of results.
- Ensure the report is well-organized and easy to follow.

Report Writing:

- Write a report (maximum 12 pages (shorter is fine), font size ≥ 11).
- You can use Word/LaTeX/Markdown to write the report.
- Submit your report as a pdf file.
- Be sure to include introduction, dataset description and exploratory data analysis.
- The report should be brief and self-contained.
- Summarize the key findings of your analysis.
- Discuss any limitations and suggest future directions or questions raised by your findings.

Presentation Slides:

- Each group will also prepare and submit PPT presentation summarizing your project.
- The PPT slides might be shared with the rest of the class.

Submission Guidelines:

- You must submit the dataset along with the source code (as a .csv). If the data is large, use SBU google drive to upload the data.
- Submit your source file (i.e., ipynb file or .Rmd file) with well-commented, organized code together with the pdf.
  - Your notebook should be **readable, reproducible, and executable** from top to bottom.
  - Include markdown cells for explanations, interpretations, and transitions between sections.
- Submit the report as a pdf file
- Submit the PPT slides.

Timeline:

- The project report (pdf), code, dataset and poster slide are due 11/23/2025 (Sunday) by 11:59 PM EST. ONLY one the group representative will submit all the project materials (dataset, report, poster slide, Python code) in the Brightspace, and email to the instructor (silvia.sharna@stonybrook.edu) and cc the TAs and the other group members.
- In the "Subject" header of the email, type "AMS 325 Fall 2025 Group XX Project", where XX will be replaced by your group number.
- Although you work in a group, still each student is expected to submit the report (only the pdf) on Brightspace. If you do not submit the report on Brightspace, there will be a penalty on your project score.

Grading Criteria (not in the order of importance):

- Is the report and presentation well-organized?

- Is supporting computer output provided (in edited form, that is, edit out all the extraneous information in the report)?
- R or Python 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?

******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 (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.