Final Project Guidelines: Introduction to R Programming STSCI 2120/5120

Instructor: Nayel Bettache

Project Overview

The goal of this project is for students to apply the R programming skills learned throughout the course to solve a real-world data problem. Students will choose a dataset, analyze it, and present meaningful insights using R's data manipulation, analysis, and visualization capabilities.

Project Requirements

1. **Dataset Selection**:
   - You must choose a real-world dataset for analysis from https://www.kaggle.com/datasets.
2. **Problem Definition**:
   - Define a clear research question or objective based on the dataset. For example:
     - "What factors influence housing prices?"
     - "Can we predict the survival rate of passengers on the Titanic?"
   - Clearly state the **hypothesis** or **goals** of your analysis in your project report.
3. **Data Cleaning**:
   - Perform necessary data cleaning operations:
     - Handle missing data.
     - Transform or standardize variables (e.g., converting date formats).
     - Deal with outliers or irrelevant data points.
4. **Exploratory Data Analysis (EDA)**:
   - Conduct an in-depth exploratory analysis of the data:

- Use **summary statistics** (mean, median, standard deviation, etc.).
- Visualize data distributions using **histograms**, **boxplots**, and **density plots**.
- Investigate relationships between variables using **scatterplots**, **correlation matrices**, etc.

5. **Data Visualization**:
   - Provide at least **three meaningful visualizations** that illustrate important aspects of the data. These can include:
     - Bar charts, scatter plots, line plots, or other advanced visualizations (e.g., `ggplot2`).
   - Ensure the visualizations are properly labeled and easy to interpret.
6. **Statistical Analysis**:
   - Apply at least one **statistical test** (t-test, chi-squared test, etc.) or **linear regression model** to derive insights from the data.
   - Provide the results, interpretations, and a summary of the findings.
7. **Conclusion and Insights**:
   - Summarize your key findings in the conclusion section:
     - What did the data reveal? Were your hypotheses correct?
     - How do your results answer the initial research question?
8. **R Code**:
   - All your analysis should be done in R, and your project should include **well-commented code** that is easy to follow.
   - Make sure the code runs smoothly from start to finish, with all necessary libraries and functions loaded at the beginning.
9. **Report Structure**: Your final project should be presented as a report with the following sections:
   - **Introduction**: Describe the dataset, the research question, and your hypothesis.
   - **Data Preparation**: Explain the data cleaning and transformation steps.
   - **Exploratory Data Analysis (EDA)**: Include your summary statistics and visualizations.
   - **Statistical Analysis**: Detail the statistical tests or models you used, along with their results.
   - **Conclusion**: Summarize your key findings and insights.

- **References**: Cite any sources for data or external libraries used.

10. **Submission**:
    - Submit a **well-formatted report** (PDF) along with your **R script** (Rmd file with Html output).
    - The report should be between **5-10 pages**, not including code.
    - Ensure that all plots and tables are properly integrated into the report (as well as in the notebook).

Grading Criteria

- **Clarity of Research Question (10%)**: How well is the problem or question defined?
- **Data Preparation (15%)**: Is the data properly cleaned and prepared for analysis?
- **Exploratory Data Analysis (20%)**: Quality and depth of EDA, including insights from summary statistics and visualizations.
- **Statistical Analysis (20%)**: Correct application of statistical methods and interpretation of results.
- **R Code Quality (15%)**: Well-commented, functional code that is easy to follow.
- **Presentation and Report (20%)**: Professional presentation of results, including well-labeled plots and clear conclusions.

Timeline

- **Groups Proposal** (Due Friday 10/11): Send me the list of students forming your group. One email per group.
- **Project Proposal** (Due Tuesday 10/15): Submit a brief description of your dataset, research question, and analysis plan.
- **Final Project Submission** (Due 10/26): Submit the full report and code.

Additional Notes

- You should form groups of 3-5 students. Each group will submit a single project and share the same grade.
- Use R functions and packages taught in the course. You may also explore new packages, but make sure to explain them in your report.