Exploratory Data Analysis Project

The primary goal of this project is to gain comprehensive insights into the dataset through exploration and analysis. By applying visual techniques and statistical methods, your group aims to uncover patterns or relationships within the data. The insights obtained will be used to inform decision-making.

Instruction

- This is **a group project**. A group can have two or three members.
- One of you submit your report onto Gradescope by indicating your partner(s) on the Gradescope. Your project will be graded as a group effort. This means that you are responsible for your own work and your partner's work. I will not assign different grades to one project.
- You are not allowed to discuss your projects with anyone other than the instructor or TA before submitting your report.
- Any outside help beyond that from the instructor or TA is considered plagiarism. This includes copying someone's works, asking a tutor and your classmates for direct help, posting the questions to homework help sites, etc.
- You can use or modify your previous code or the instructor's code posted online.
- The maximum length of the report is **4 pages** excluding title, header, tables, plots, and R appendix. Tables, plots, and R appendix should be attached at the end of your report.
- Formatting will be a significant portion of your grade for this project (take-home exam). There should be an appendix of code, including all your codes. Codes or raw R results (directly copied and pasted from R with no additional formatting) should not be in the body of the report.
- Your report should be in full paragraph form. You are allowed to have tables, and/or use R Markdown, but it should have clearly labeled sections. You can also use Word or Google Docs (or Latex) if you are more comfortable with those.

The Report Format

The Goal: The goal is to answer your questions by Exploratory Data Analysis (EDA - Visualization) and Confidence Interval/Hypothesis Test. You should write up a full, paragraph form report on your findings, which should include the following sections:

I. Introduction: Introduce your dataset. State your three questions and why these are interesting to you.

II. Data Preparation: Explain your data source. Explain your data cleaning/variable selection procedures. You may choose your explanatory and response variables with reasonings.

III. Descriptive Statistics: Present any numerical summaries you find interesting facts and connect to your questions.

IV. Visualization: Present your visualizations to answer your two of three questions.

V. Statistical Inference: Construct Confidence intervals for parameters to answer your questions. Conduct Hypothesis testing on your claim/questions. Interpret the estimates, any confidence intervals, or p-values that you calculated.

VI. Conclusion: Summarize your findings and suggest potential next steps to enhance your analysis.

Report Format

Your report should be in the following format:

- Typed
- A title page or a header, including the title of your project, your names, the name of the class, and the name of your instructor (me)
- Tables and Graphs are clearly labeled and attached at the end of the report.
- R appendix at the end of the report (after the tables/plots). The R appendix contains your R code used to produce the results. Do not include R code in the body of your report. Write your code clearly with explanations so we can understand your code easily.

For example, your project report should be put together in the following order:

Title page/Header
Part I-VI.
Table/Plots
R appendix.

Presentation

Each team will present their projects either in class or upload their recorded presentations. The assigned presentation time is between 5-10 minutes for each group. I will provide details after the Midterm.

Peer Evaluation

There will be two types of peer evaluation

- Within groups: regarding Team contribution, collaboration, communication skills, and responsibility.
- Between groups: Clarity of Objectives, Depth of Analysis, Data Visualization, and Presentation Quality.

I will provide evaluation sheets (or via Canvas).

Project Timeline

Week 3: Data Selection
Week 4: Data Cleaning
Week 5: Data Visualization (Project Check-in due at 2/9)
Week 8: Statistical Inference
Week 9: Project Report, Slides, Presentation (if recorded) at 3/8
Week 10: Peer Evaluation due at 3/15

Grading Distribution

Project Report	80%
Project Presentation	10%
Peer Evaluation	10%

Schedules or details might be changed depending on course circumstances.