Final Project - Step 2 (15 Points)

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

- STUDENT 1 (NetID 1)
- STUDENT 2 (NetID 2)
- STUDENT 3 (NetID 3)
- STUDENT 4 (NetID 4)
- STUDENT 5 (NetID 5)

Due Date

The deadline for this step is **Friday, May 9, 2025**.

Before Proceeding to Step 2

Please carefully review any comments or feedback on Step 1. If you have any questions or need clarification, reach out to the TA for assistance. Ensuring a clear understanding of Step 1 will help you complete the next step successfully.

Instructions

In this step, you will develop clear research questions and hypotheses based on your selected dataset, and conduct a thorough Exploratory Data Analysis (EDA). This foundational work is crucial for guiding your analysis in the following steps.

1 Step 2: Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

1.1 Research Questions

- **Formulate Insightful Questions**: Develop 2–3 well-defined research questions that align closely with your dataset. These should investigate interesting patterns, trends, or relationships and help guide your data analysis—whether your project involves regression, classification, clustering, or exploratory insights.

  Examples:

- What demographic factors are associated with increased likelihood of purchasing insurance?
- Which user behavior patterns predict customer churn in a subscription model?
- How does the availability of public transportation influence property values in urban areas?
- Are there noticeable regional patterns in access to healthcare facilities?
- **Ensure Relevance**: Your questions should be closely tied to your dataset and aim to uncover specific insights.

1.2 Hypotheses

Develop Clear, Testable Hypotheses: For each research question, craft specific, measurable hypotheses that can be tested with the available data. These can reflect expected relationships, differences between groups, or distributional characteristics.

Examples:

- Users aged 60 and above are more likely to choose comprehensive insurance plans than younger users.
- Higher daily app engagement is associated with lower churn rates among users.
- Urban neighborhoods with nearby transit hubs have higher average property values.
- Regions with higher population density have greater access to hospitals per capita.

Link to Research Questions: Make sure your hypotheses flow directly from your research questions and form a logical foundation for your analytical methods.

1.3 Exploratory Data Analysis (EDA)

- **Conduct EDA**: Perform a comprehensive exploratory data analysis on your dataset. This should include:

1.3.1 Data Cleaning: Identify and handle missing values and outliers.

Example: Use functions like `na.omit()` or `mutate()` from the `dplyr` package to manage missing data.

1.3.2 Descriptive Statistics: Provide summary statistics for your variables.

Example: Use the `summary()` function to calculate mean, median, and quartiles.

1.3.3 Data Visualization: Create various visualizations to explore relationships and distributions:

Histograms: To visualize the distribution of continuous variables. *Example*: `ggplot(data, aes(x=variable)) + geom_histogram(bins=30)`

Box Plots: To compare distributions across categorical variables.

Example: `ggplot(data, aes(x=categorical_variable, y=continuous_variable)) + geom_boxplot()`

Scatter Plots: To examine relationships between two continuous variables. *Example*: `ggplot(data, aes(x=var1, y=var2)) + geom_point()`

Heatmaps: To visualize the correlation matrix between variables. *Example*: `corrplot::corrplot(cor(data), method="circle")`

Word Clouds: For text data, use word clouds to identify common terms. *Example*: Use the `wordcloud2` package to create a word cloud based on text data.

Maps: If your dataset includes geographical information, consider using maps to visualize data spatially. *Example*: Use `ggmap` or `leaflet` to create interactive maps displaying relevant data points.

Feature Relationships: Analyze how independent variables relate to your response variable. *Example*: Create visualizations showing the impact of different features on the outcome of interest.

Distribution Comparisons: Compare distributions across different groups. - *Example*: Use faceted histograms to visualize how a continuous variable is distributed across categories of a categorical variable.

Skewness and Kurtosis: Assess the normality of distributions. *Example*: Calculate skewness and kurtosis using the `e1071` package.

1.4 Documentation

- **Clearly Document Your Process**: Ensure your `R` markdown file is organized, documenting each step of your research questions, hypotheses formulation, and EDA findings.
- **Initial Insights**: Include a brief discussion of what makes your dataset interesting or relevant for analysis, and highlight any preliminary insights derived from your EDA.

1.4.1 Do:

- Include clear, informative captions for every plot (add captions within the chunk options).
- Write in complete correct sentences.
- Annotate all visual elements so the reader can easily interpret them.
- Limit the report to **7** pages or fewer, focusing primarily on graphics.
- Submit the `.pdf` file on **Gradescope**.
- Develop clear, focused, and insightful research questions and testable hypotheses to guide your analysis.
- Perform a thorough exploratory data analysis (EDA) and clearly document your insights.
- Use a variety of visualizations (e.g., heat maps, word clouds, geographic maps) to effectively convey your findings.
- **Conduct all data processing and analysis in R—this includes any changes to the dataset.**
- **Ensure that all data transformations and modifications are documented in your R script so your work is fully reproducible.**

1.4.2 Do Not:

- Include warning or error messages. Display only relevant and meaningful code (use echo = FALSE when appropriate).
- Print raw lists of data.
- Pose irrelevant or unrelated research questions.
- Formulate vague or non-testable hypotheses.
- Submit incomplete, disorganized, or poorly structured analyses.
- **Manually edit the spreadsheet or make any data changes outside of `R`.**

1.5 Deliverables and Submission Requirements

Submit only a `.pdf` file rendered from 'Final Project 100 Step 2 Template.qmd' on **Gradescope** by the specified due date.

1.6 Additional Notes

- This step is foundational for your final analysis, so invest time in developing thoughtful questions and conducting thorough EDA.
- Remember that your analysis may involve various methods beyond regression, such as clustering, classification, or forecasting, depending on your dataset, research questions and group skills.