Final Project - Step 1 (8 Points)

PSTAT100: Data Science Concepts and Analysis

The purpose of this document is to provide guidance for forming the groups, dataset selection and structuring your project.

The deliverables will consist of a pdf file generated from `Final Project Step 1 Template.qmd` to be submitted on **Gradescope**.

> Due Date
>
> The deadline for this step is Friday, April 25.

1 Formation of groups:

Students will work in groups of 4 to 5 members **from the same discussion section** to collaboratively complete their final project. Each group should aim to include members with diverse skills to enhance teamwork and ensure a professional presentation of findings.

1.1 Group Formation Instructions

1. **Group Size**: Each group must consist of **4 to 5 students**. No more, no less.
2. List your group names in the Excel sheet here.
3. **Group Composition**: Form groups with students possessing **different skills** (e.g., research, writing, technical expertise) to promote collaboration and comprehensive project execution.
4. **Self-Selection Process**: Students may choose their own groups. Encourage members to communicate their skills and preferences when forming groups.
5. **Group Registration**: Each group must submit the following by the deadline:
   - **Group Name**
   - **List of Group Members** with:
     - Student Name
     - NetID
6. **Initial Meeting**: Schedule an initial meeting for all group members to discuss project ideas and agree on roles.
7. **Role Distribution**: Clearly define roles within each group to ensure accountability and efficient workflow. Flexibility is encouraged, allowing members to adjust roles as needed.
8. **Regular Check-ins**: Groups should schedule regular meetings (weekly or bi-weekly) to monitor progress and address any challenges.
9. **Conflict Resolution**: Establish a protocol for addressing conflicts. Encourage open communication and seek assistance from the instructor and TA if necessary.

10. **Feedback Mechanism**: Implement a peer feedback system to evaluate each member's contributions throughout the project.
11. **Final Submission**: Mind the **submission format** and **deadline** for the final project. Ensure all group members understand expectations.

2 Tentative Role Distribution Table

Name	Responsibilities
STUDENT 1	
STUDENT 2	
STUDENT 3	
STUDENT 4	
STUDENT 5	

3 Data Requirements

To ensure a comprehensive analysis for your project, please adhere to the following dataset specifications:

3.1 Variable Specifications

- The dataset must include a minimum of **10 variables**. Specifically:
  - At least **3 independent quantitative variables** (e.g., age, income, score).
  - At least **2 independent categorical variables** (e.g., gender, education level).
- **Note**: Labels or identifiers (e.g., ID, name) do not count as variables for analysis purposes, as they cannot be utilized in statistical modeling. Please use full variable names or provide clear descriptions for any abbreviations to enhance clarity for your readers.

3.2 Creating Categorical Variables

- If you encounter challenges in locating a dataset that meets these criteria, you may select a dataset containing only continuous variables. In such cases, you are encouraged to create categorical variables by partitioning the continuous variables based on reasonable criteria.
- **Example**: For instance, you could take the Body Mass Index (BMI) variable and categorize it into segments such as "Very Low," "Low," "Medium," "High," and "Very High" based on common percentiles. Be sure to mention this process in your final report.

3.3 Observation Requirements

- The dataset should consist of at least **100 independent cases/observations**. Ideally, aim for a sample size between **200 and 500 observations**.
- Be vigilant regarding missing data; excessive missing observations may compromise your analysis. If your dataset contains too many missing values, consider randomly selecting 500 observations to maintain a robust sample.

3.4 Population Representation

- As your project will involve hypothesis testing, it is crucial to define the population that your data represents.
- For example, if your dataset is derived from a census, discuss whether it could represent a broader population. For instance, data from a state census conducted in 2015 might reasonably be extrapolated to inform insights about the population in 2016.
- Additionally, consider and articulate the limitations of generalizing findings to a larger population based on your dataset.

By following these guidelines, you will ensure that your dataset is well-suited for thorough analysis and provides meaningful insights in your project.

4 Recommended Data Sources

If you plan to select the dataset yourself, consider utilizing the following reputable data sources:

1. **UCI Machine Learning Repository**:
   - This extensive collection of datasets is maintained by the University of California, Irvine. Many datasets are suitable for linear regression analysis.
   - Access the UCI Repository (https://archive.ics.uci.edu/ml/index.php).
2. **Kaggle**:
   - Kaggle is a well-known platform for data science competitions and hosts a diverse range of datasets. Many datasets available here are appropriate for linear regression analysis.
   - Explore Kaggle Datasets (https://www.kaggle.com/datasets).
3. **OpenML**:
   - OpenML is an open-source platform designed for sharing and organizing machine learning data and experiments. It offers a large collection of datasets suitable for linear regression analysis.
   - Visit OpenML (https://www.openml.org).
4. **R Datasets Package**:
   - The R programming language includes a built-in collection of datasets in its base installation. These datasets can be loaded directly into R and are available for linear regression analysis.

By considering these sources, you can find a suitable dataset that meets the requirements for your analysis project.

5 Format

5.1 Do:

- **Provide Comprehensive Information**:
  - List all relevant details about the selected dataset, including:
    - **Data Name/Title**: The official name of the dataset.
    - **Author/Owner**: The individual or organization responsible for the dataset.
    - **Date of Publication**: When the dataset was published.
    - **Publication Venue**: Where the dataset is hosted or published.
    - **Retrieval Date**: The date you accessed the dataset to ensure its relevance.
    - **Link**: Include the URL for the dataset if it is publicly accessible.
  - This information will facilitate future access to the dataset.

- **Initial Insights**:
  - Provide preliminary insights about the dataset, highlighting what makes it interesting or relevant for your analysis team. Consider discussing potential implications, unique features.
- **Submission Requirements**:
  - Submit your work in both `.qmd` and `.pdf` formats via Gradescope. Ensure that the `.qmd` file is properly configured to knit without errors.

5.2 Don't:

- **Do Not Print Data Lists**: Refrain from displaying entire lists of data or raw output. Focus on the narrative and insights.
- **No Plots, Modeling, or Hypothesis Testing**: At this stage, do not include any plots, modeling outputs, or hypothesis testing results. This section is for preliminary insights only.