Women/Men Focused Training

Ghana R-Users Community

Group Project 1 – Cleaning and Visualizing a Dataset

Learning Objectives:

- Apply Week 1 and Week 2 concepts to clean and explore a messy dataset
- Create polished visualizations using `ggplot2`
- Collaborate as a team to present insights from the dataset

Session Outline

1. Introduction to Group Project

- **Objective:** Participants will work in groups to clean, explore, and visualize a messy dataset provided by the trainer.
- **Steps to Follow:**
  1. Import and inspect the dataset.
  2. Clean the dataset (handle missing values, filter, and transform as necessary).
  3. Perform basic exploratory data analysis (EDA).
  4. Create at least three visualizations showcasing insights from the dataset.
- **Deliverable:** Each group will present their cleaned dataset, findings, and visualizations.

2. Dataset Details and Instructions

- **Dataset:** A simulated dataset `student_performance.csv` containing information about students' performance, including:
  - `StudentID`: Unique identifier for each student
  - `Age`: Age of the student
  - `Gender`: Male or Female
  - `Math_Score`, `Reading_Score`, `Writing_Score`: Test scores (0-100)
  - `Study_Hours`: Weekly hours spent studying
  - `Parental_Education`: Highest education level of the parents
  - `Lunch`: Type of lunch received (Standard or Free/Reduced)
  - **Issues in the Dataset:** Missing values, inconsistent capitalization, and outliers

Instructions:

1. **Cleaning**
   - Handle missing values in `Math_Score`, `Reading_Score`, and `Writing_Score`.
   - Normalize the `Study_Hours` column to scale between 0 and 1.
2. **EDA**
   - Calculate summary statistics for each score.
   - Group by `Gender` and calculate the average scores.
3. **Visualizations**
   - Create a bar chart comparing average scores by gender.
   - Create a scatter plot of `Study_Hours` vs. `Math_Score`, coloring by `Gender`.
   - Create a histogram of `Math_Score` distribution.

3. Group Work

- **Breakout Groups:** Participants will work in teams of 3-5 people.
- **Trainer Support:** Trainers will assist each group with challenges during the breakout session.

4. Presentations

- Each group will have 5 minutes to present:

    1. The cleaned dataset and summary of issues they addressed.
    2. Insights from their EDA.
    3. Their visualizations and what they reveal about the data.