Exploratory Data Analysis

Assignment 1: Basic - Understanding Data Structure and Univariate Analysis

Context: As a data analyst at One Hour AI Solution platform, you need to understand the basic characteristics of your session data to help improve service quality.

Task: Perform basic exploratory data analysis focusing on understanding data structure and univariate analysis.

Dataset: sessions_basic.csv

Requirements:

1. Perform data structure analysis:
    - Check the shape of the dataset (number of rows and columns)
    - Examine data types of each column
    - Check for missing values using appropriate commands
2. Conduct univariate analysis for numerical variables:
    - Calculate central tendency (mean, median, mode) for session_duration, engineer_experience, and client_satisfaction
    - Calculate dispersion measures (range, variance, standard deviation)
    - Identify potential outliers using the IQR method
3. Conduct univariate analysis for categorical variables:
    - Create frequency tables for problem_type and solution_provided
    - Calculate proportions for each category
4. Write a brief summary (3-5 sentences) of your key findings from the univariate analysis.

Assignment 2: Intermediate - Bivariate Analysis

Context: The management team at One Hour AI Solution wants to understand the relationships between different aspects of their consulting sessions.

Task: Perform bivariate analysis to explore relationships between pairs of variables in the session data.

Dataset: sessions_bivariate.csv

Requirements:

1. Analyze relationships between numerical variables:
    - Calculate Pearson correlation coefficient between session_duration and client_satisfaction
    - Calculate Pearson correlation coefficient between engineer_experience and client_satisfaction
    - Interpret the correlation results (strength and direction)
2. Analyze relationships between categorical and numerical variables:
    - Calculate average client_satisfaction for each problem_type
    - Calculate average client_satisfaction for each client_experience level
    - Calculate standard deviation of client_satisfaction for each group
    - Use group comparison methods to determine if differences are meaningful
3. Analyze relationships between categorical variables:
    - Create a contingency table between problem_type and solution_provided
    - Create a contingency table between client_experience and solution_provided
    - Calculate percentages to understand the relationships
4. Based on your findings, write a brief summary (maximum 150 words) explaining the key relationships discovered in the data and what they might suggest for the One Hour AI Solution platform.

Assignment 3: Advanced - Data Issues and Feature Engineering

Context: One Hour AI Solution is launching a new feature that will match clients with the most appropriate engineers. Your task is to examine session data, handle data issues, and create new features that might help predict successful sessions.

Task: Handle data issues (missing values, outliers, skewed distributions) and perform feature engineering based on your EDA findings.

Dataset: sessions_advanced.csv

Requirements:

1. Handle missing values:
    - Identify columns with missing values and calculate the percentage of missing values in each
    - Decide on an appropriate strategy for each column (deletion, imputation, etc.)
    - Implement your chosen approach and explain your reasoning
2. Detect and handle outliers:
    - Use the IQR method or Z-score to identify outliers in numerical variables
    - Analyze whether the outliers represent actual data issues or meaningful extreme values
    - Implement an appropriate strategy (keep, remove, transform) with explanation
3. Perform feature engineering to create at least 3 new features:
    - Create a feature that represents the experience gap between engineer and client
    - Convert session_date to a day of the week feature
    - Create a binary feature indicating whether the problem complexity is high (4-5) or low (1-3)
    - Create any other feature you think might be useful
4. Analyze the engineered features:
    - Calculate descriptive statistics for the new features
    - Perform correlation analysis between new features and key outcome variables (client_satisfaction, solution_provided)
    - Determine which engineered features show the strongest relationships with session outcomes