DATA ANALYSIS

Exploratory Data Analysis (EDA) Boilerplate:

Data Summary and Understanding :

- Load the data and examine the first few rows to understand its structure.
- Check the data types of each column and the number of observations.

Missing Values Analysis :

- Identify missing values in each column and assess their impact on analysis.
- Decide on a strategy to handle missing values (imputation or deletion) based on the context of the data.

Univariate Analysis :

- For numerical features:

    - Generate summary statistics (mean, median, standard deviation, min, max).
    - Plot histograms or density plots to visualize the distribution.
    - Identify outliers using box plots.
- For categorical features:
    - Count the frequency of each category.
    - Visualize using bar plots or pie charts.

Bivariate Analysis :

- Explore relationships between pairs of variables:
    - For numerical vs. numerical: scatter plots, correlation analysis.
    - For numerical vs. categorical: box plots, violin plots.
    - For categorical vs. categorical: contingency tables, stacked bar plots.

Multivariate Analysis :

- Explore relationships between multiple variables simultaneously using techniques like:
    - Pair plots (for small datasets).
    - Heatmaps for correlation analysis.
    - Parallel coordinates plots for high-dimensional data.

Outlier Detection and Treatment :

- Identify outliers using domain knowledge or statistical methods (e.g., z-score, IQR).
- Decide whether to remove outliers or transform them based on their impact on the analysis.

Feature Engineering Boilerplate :

- **Creation of New Features**:
  - Derive new features based on domain knowledge or insights gained from EDA.
  - Combine existing features or create interaction terms to capture relationships.
- **Encoding Categorical Variables**:
  - Convert categorical variables into numerical format using techniques like one-hot encoding or label encoding.
- **Handling Date/Time Variables**:
  - Extract relevant information from date/time variables (year, month, day of the week, etc.).
  - Encode cyclical features like month or day of the week.
- **Scaling/Normalization**:
  - Scale numerical features if necessary to ensure they are on the same scale.
  - Common techniques include min-max scaling or standardization.

Correlation Analysis :

- Compute the correlation matrix to identify highly correlated features.
- Consider removing features that are highly correlated with each other to reduce multicollinearity.

Visualization for Insights :

- Use visualizations like heatmaps, pair plots, or bar plots to gain insights into relationships between features and the target variable.

Documentation :

- Document the steps taken during EDA and feature engineering for reproducibility.
- Record any assumptions made and decisions taken during the process.

Remember, EDA and feature engineering are iterative processes, and it's essential to continuously refine your analysis based on the insights gained along the way.