Lab: Data Preprocessing

Objective:
By the end of this lab, you should understand the process of collecting, cleaning, and transforming data, handling missing values, dealing with outliers, normalizing data, and performing feature engineering and feature selection.

Tools Required:

- Python
- Jupyter Notebook
- Libraries: `pandas`, `numpy`, `scikit-learn`, `matplotlib`, `seaborn`

Step 1: Data Collection

Instructions:

1. Download a dataset from Kaggle or use an inbuilt dataset from `seaborn` or `sklearn`.
    - Example: Titanic dataset (available via `seaborn`)

Step 2: Data Cleaning

Instructions:

1. Inspect for missing values.
2. Handle missing values by either:
    - Dropping rows or columns with missing data.
    - Imputing missing values with mean, median, or mode.

Step 3: Handling Outliers

Instructions:

1. Identify outliers in the **`fare`** and **`age`** columns using box plots.
2. Use techniques like capping or removing outliers.

Step 4: Data Normalization

Instructions:

1. Normalize the numerical features such as `age` and `fare` using Min-Max scaling or Z-score normalization.

Step 5: Feature Engineering

Instructions:

1. Create new features such as:
    - A `family_size` column by summing `sibsp` (siblings/spouses aboard) and `parch` (parents/children aboard).
    - A `title` column extracted from the `name` column (if available).

Step 6: Feature Selection

Instructions:

1. Select the most important features using techniques like correlation analysis or feature importance from a machine learning model.

Step 7: Model Building

Instructions:

- the data into train/test sets and build a simple classifier like Logistic Regression or Random Forest using the preprocessed data.