# Data Cleaning Assignment

## Objective:

To clean and organize the raw financial data provided, making it ready for analysis using Excel.

## Problems to Address:

### 1. Understand the Data:

- Examine the structure of the dataset.  
 - Identify the data types of each column.  
 - Check for any obvious inconsistencies or anomalies.

### 2. Handle Missing Values:

- Identify columns with missing values:  
 - Use Excel functions to detect missing values in each column.  
 - For each column, use the formula `=COUNTBLANK(A:A)` (replace `A` with the actual column letter) to count the number of blank cells.  
 - Document which columns contain missing values and the number of missing values in each.  
  
 - Decide on appropriate strategies for handling missing values:  
 - Determine the best method to handle missing values based on the type of data and the proportion of missing values:  
 - **Numeric Columns**:  
 - Replace missing values with a statistical measure such as the mean or median.  
 - If the proportion of missing values is very high, consider removing the entire column.  
 - **Categorical Columns**:  
 - Replace missing values with the most frequent value (mode).  
 - If the column has a high proportion of missing values, consider removing it.  
 **- Date Columns**:  
 - Use a common date such as the start of the year or month, or the median date.  
 - If the date is not critical, consider removing the row or column with missing dates.  
 - **Mixed Data Types**:  
 - Use a relevant default value or category (e.g., "Unknown" for missing categorical data).  
  
 **- Implement the chosen strategies:**  
 - Use Excel's `Find & Select` -> `Go To Special` -> `Blanks` to highlight all blank cells in the dataset.  
 - Fill in the missing values based on the chosen imputation strategy:  
 - For numeric columns, input the calculated mean or median into blank cells.  
 - For categorical columns, input the mode or a default value into blank cells.  
 - For date columns, input the chosen common date or median date.  
 - Document the changes made and the rationale behind the chosen strategies.

### 3. Correct Data Types:

- Ensure each column has the correct data type.  
 - Convert columns to appropriate data types if necessary.

### 4. Remove Duplicates:

- Check for duplicate rows and remove them.

### 5. Standardize Data:

- Standardize text data to a consistent format (e.g., title case for names, consistent date formats).

### 6. Handle Outliers:

- Identify and handle outliers in numeric columns.

### 7. Validate Data:

- Ensure that each column contains valid and sensible data.  
 - Validate specific constraints (e.g., dates are in the past, transaction amounts are positive).

### 8. Feature Engineering:

- Extract Date Components:  
 - Extract the year, month, and day from the `TransactionDate` column.  
 - Categorize Transaction Amount:  
 - Create bins for `TransactionAmount` (e.g., low, medium, high).  
 - High-Risk Transaction Flag:  
 - Create a flag for high-risk transactions based on the `FraudScore`.