Assignment 4

<u>Data Preparation</u>

‘Mine Dime’, a stock trend forecasting company has just employed you as a Data Scientist. As a first task in your new job, your manager has provided you with a company’s stock data and asked you to check the quality of the data for the next step of analysis. Following are the additional description and information about the data which your manager has shared with you.

a) The data set contains six variables namely-
   i. Date
   ii. Open
   iii. High
   iv. Low
   v. Close
   vi. Volume

b) Typically, the stock market opens at 9:15 hours and closes at 15:30 hours. Each stock is defined by an opening price and a closing price which are the prices it opens and closes with. Within the operating hours, the stock price touches a maximum and minimum which are the highest and lowest prices achieved by the stock in the working hours of the stock market. You have access to ten years of monthly stock price data with the Open, High, Low and Close price and the number of stocks traded for each day given by the feature Volume. On some days when there is no trading, the parameters Open, High, Low and Close remain constant and Volume is zero.

Furthermore, your manager also claims that the model prediction is too bad since the data is polluted. Try to impress your new boss by preprocessing the data and by giving a proper rationale behind the steps you would follow. The two datasets should be merged before preprocessing.

To ensure data quality and consistency for model training or data analysis, kindly make sure to perform the following steps:

1. Sort by Date

- Ensure the final dataset (after merging) is sorted chronologically by Date.

2. Data Type and Format Validation

- Date: Convert to datetime format (YYYY-MM-DD). Ensure no null or malformed dates.
- If any numerical variable value is recorded in the file as a string type instead of numerical data type, convert it to numerical data type (int or float).
- Price Variables (Open, High, Low, Close): Convert to float. Remove any currency symbols or strings if present.
- Volume: Convert to int or float (as needed). Ensure it is non-negative.

3. Missing or Null Values
    - Drop rows with nulls in any of these columns: Date, Open, High, Low, Close, Volume.
    - Alternatively (if essential), perform data imputation using appropriate strategy for each variable. Example: Impute “Volume” with rolling average or median over last 5 valid trading days

4. Validate Price Relationships
    - Ensure logical integrity of price data. Example: Low ≤ Open ≤ High, Low ≤ Close ≤ High
    - Drop rows violating this unless explainable (e.g., erroneous recording)

5. Handling duplicate data
    - Check for duplicate rows. Remove exact duplicates.

6. Remove Non-Trading Days if present
    - Filter out rows where Open = High = Low = Close and Volume = 0
    - These are non-trading days and should not be used for modeling

7. Outlier Detection and Treatment
    - Identify extreme outliers in Price and Volume using appropriate method and verify once using visual inspection.
        - Investigate the outliers and take suitable actions. Example: Removal of outliers if it is due to data error