Problem Statement 1:

You are given a dataset of movie reviews. Each review contains text and a sentiment label (positive or negative). Your task is to build a sentiment analysis model that can classify new reviews into one of these two categories. You can use the validation set provided in dataset link.

Dataset:

- https://huggingface.co/datasets/stanfordnlp/imdb

Task:

- Preprocess the text data (remove stop words, punctuation, etc.).
- Train a simple machine learning model (e.g., Logistic Regression, Naive Bayes) or a deep learning model (e.g., LSTM or BERT-based model) to classify the sentiment of the reviews.
- Evaluate the model using accuracy, precision, recall, and F1-score.

Problem Statement 2:

Outlier Impact and Seasonal Revenue Analysis in Retail Data on UCI Online Retail Dataset. Your task is to perform EDA and get the following results.

Dataset:

- **UCI Online Retail Dataset:**
  https://archive.ics.uci.edu/ml/datasets/Online+Retail

Task:

- Remove duplicate records and handle missing values appropriately.
- Standardize data types (e.g., converting date fields) to ensure consistency.
- Use statistical methods (e.g., z-score or IQR) to identify outliers in invoice revenue or unit sales.
- Quantify how these outliers affect summary statistics like the mean, median, and overall revenue distribution.
- Visualize the revenue distribution before and after outlier treatment (e.g., using box plots or histograms).
- Aggregate the data to a monthly level and plot the revenue trends.
- Evaluate whether outliers are concentrated in specific time periods, potentially distorting seasonal patterns.
- Discuss how outlier-induced distortions could mislead business decisions such as forecasting or performance evaluations.
- Propose strategies for managing or mitigating the impact of outliers in future analyses.