Escel

ETL using Advanced Excel

STEP-BY-STEP: Data Analysis & Visualization using Advanced Excel

STEP 1: Open Your Dataset

• Open sales_data.xlsx in Microsoft Excel.

STEP 2: Clean the Data (Basic ETL)

A. Remove Blank Rows
• Select the entire dataset.
• Go to Data > Filter → Apply filter to check for blanks.
• Select blanks in a column (e.g., Sales) and delete the rows.

B. Convert to Excel Table

• Select the dataset.
• Press Ctrl + T → This helps in dynamic referencing.
• Table Name: Rename it to SalesTable under Table Design > Table Name.

STEP 3: Add Calculated Columns

A. Extract Month from Order Date

• Add a new column: =TEXT([@[Order Date]], "mmmm")
→ This will show month names like January, February.

B. Extract Year from Order Date

• Add a new column: =YEAR([@[Order Date]])

C. Calculate Profit Margin (if applicable)

• =([@[Profit]] / [@[Sales]])*100

STEP 4: Use Pivot Tables for Data Analysis

A. Insert a Pivot Table

• Select your SalesTable.
• Go to Insert > PivotTable > From Table/Range.
• Place PivotTable in a new sheet.

B. Monthly Sales Analysis

• Rows: Month
• Values: Sales → Summarize by Sum
• Sort Months:
o Add a helper column with =MONTH([@[Order Date]]) for correct sorting if needed

C. Sales by Region or Segment

• Rows: Region
• Columns: Segment
• Values: Sales

STEP 5: Create Visualizations

A. Insert a Bar Chart

• Select your PivotTable showing monthly sales.

• Go to Insert > Column or Bar Chart > Clustered Column

B. Insert a Pie Chart

• Create a PivotTable:

o Rows: Month, Values: Sales

• Select PivotTable > Insert > Pie Chart

C. Create a Heatmap (Conditional Formatting)

• Use PivotTable: Rows = Product Category, Columns = Month

• Highlight the values.

• Go to Home > Conditional Formatting > Color Scales.

STEP 6: Apply Filters and Slicers

A. Add Slicers

• Click inside PivotTable.

• Go to PivotTable Analyze > Insert Slicer.

• Choose fields like Region, Segment, or Month.

STEP 7: Create a Dashboard (Optional)

• Create a new sheet.

• Copy charts and slicers.

Align them neatly.

• Use shapes/textboxes for headings and KPIs.

Useful Excel Commands Summary:

Task Formula / Command

Convert to table Ctrl + T

Extract Month =TEXT([@[Order Date]], "mmmm")

Extract Year =YEAR([@[Order Date]])

Profit Margin % =([@[Profit]] / [@[Sales]])*100

Insert PivotTable Insert > PivotTable

Conditional Formatting Home > Conditional Formatting > Color Scales

Insert Slicer PivotTable Analyze > Insert Slicer

…

 pandas: Used for data manipulation and analysis (dataframes, etc.).

 numpy: For numerical computations, arrays, etc.

 matplotlib.pyplot: For plotting graphs.

 seaborn: For advanced statistical plotting (built on matplotlib).

 train_test_split: Splits the dataset into training and testing sets.

 LinearRegression: Used to train a linear regression model.

- **mean_squared_error, r2_score**: Evaluation metrics for regression.
- **metrics**: General module for additional metrics like MAE, etc.

- **imdb**: Loads the IMDB dataset (preprocessed for sentiment analysis).
- **pad_sequences**: Ensures input sequences have equal lengths.
- **Sequential**: A linear stack of Keras layers (for neural network models).
- **Embedding**: Turns positive integers into dense vectors.
- **LSTM**: Long Short-Term Memory, a type of RNN used for sequence data.
- **Dense**: Fully connected neural network layer (used for classification output).

- **classification_report**: Precision, recall, f1-score, and accuracy.
- **confusion_matrix**: Matrix to visualize correct vs. incorrect predictions.

Your script sets up for **two major tasks**:

1. **Linear Regression Analysis** – likely for numerical prediction tasks (like house prices).
2. **IMDB Review Classification** using an **LSTM-based deep learning model** – for sentiment analysis.