Final Project – Data Analysis (Round 3)

Background

You are working as a **Data Analyst** for an **E-commerce company** that sells products across multiple countries (Egypt, Saudi Arabia, USA, Germany).
The company has collected a large amount of transactional and customer data. They now want to:

- Understand customer behavior.
- Identify high-value and at-risk customers.
- Predict customer churn to improve retention strategies.

Objective

- Perform **data cleaning and preprocessing** to ensure data quality.
- Build a **data model** (star/snowflake schema) for analysis.
- Conduct **customer segmentation** using RFM and clustering.
- Develop a **churn prediction model** to identify customers likely to leave.
- Provide **business insights and recommendations** through analysis and visualization.

Dataset

The dataset follows a **Snowflake Schema**:

- **Fact Table:**
    - Fact_Sales → Sales transactions (InvoiceNo, InvoiceDate, CustomerID, ProductID, Quantity, UnitPrice, TotalAmount).
- **Dimension Tables:**
    - Dim_Customer → Customer details (Name, Gender, Age, Location, RegistrationDate).
    - Dim_Product → Product details (ProductName, SubCategory, Price, Cost).
    - Dim_Location → City, State, Country.
    - Dim_Country → Country & Region.
    - Dim_Category → Product categories.
    - Dim_SubCategory → Product subcategories.

- Dim_Date → Date dimension (Day, Month, Quarter, Year).

Tasks

1. **Data Preparation & Cleaning**
    - Handle missing values, duplicates, incorrect data types.
    - Build Data Modeling.
2. **Exploratory Data Analysis (EDA)**
    - Analyze sales trends (monthly/quarterly/yearly).
    - Identify top products, categories, and regions.
    - Study customer demographics and purchasing patterns.
3. **Customer Segmentation**
    - Perform **RFM Analysis** (Recency, Frequency, Monetary).
    - Interpret the customer segments (e.g., loyal, at-risk, new customers).
4. **(Supervised & Unsupervised ML)**: you have some use cases like:
    - Make churn prediction: Define churn (inactive and active customers) and then start building your classification model
    - Monthly trend predictions using regression (time series)
    - Apply **Clustering (K-Means)** to group customers unlike RFM
5. **Visualization & Business Insights**
    - Build dashboards with KPIs:
    - Provide actionable recommendations to improve customer retention and marketing.

Team Structure & Responsibilities

Team 1 – BI & Reporting (Power BI + Excel)

- Data Cleaning (Excel / Power Query).
- Build **data model** in Power BI.
- Create **dashboards** for:
    - Sales trends by category & country.

    - Top customers & top products.
    - Customer demographics analysis.
- Provide **business insights** from visualizations.

Team 2 – Data & ML (Python + SQL)

- Structure and query data using SQL.
- Perform **EDA** (Exploratory Data Analysis).
- Customer Segmentation:
- Build Machine Learning Models

Deliverables

- **Team 1 (BI & Excel):**
    - Cleaned dataset (Excel/Power Query).
    - Power BI Dashboard with KPIs & insights.
- **Team 2 (Python & SQL):**
    - Jupyter Notebook / Python scripts with EDA & ML models.
    - SQL queries for data exploration.
    - ML results with evaluation metrics.
- **Final Presentation (all team):**
    - Business problem & dataset.
    - Analysis & models.
    - Insights & recommendations.

Duration & Final Presentation

Project Duration: 6 days.

- **Final Presentation & Discussion:**
    - Date: **24 / 9 / 2025**
    - Format: **15 minutes per team** (presentation + discussion).