Retail Sales Analysis and Prediction Report

1. Objective:

The primary objective of this analysis was to explore a retail sales dataset to uncover patterns and trends in customer behavior and sales performance. A secondary objective was to build and evaluate regression models to predict the Total Amount of a transaction based on available customer and product features.

2. Methodology:

The analysis was conducted using Python with libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-learn/XGBoost for modeling. The key steps were:

- **Data Loading:** The dataset (retail_sales_dataset.csv) was loaded into a Pandas DataFrame.
- **Data Cleaning & Preparation:**
    - Initial inspection revealed 1000 entries and 9 columns.
    - Checked for missing values (none found).
    - Checked for duplicate entries (none found).
    - Converted the Date column from object to datetime format for time-based analysis.
- **Feature Engineering:**
    - Extracted Month and Year from the Date column.
    - Created an Age Group categorical feature ('Young', 'Adult', 'Senior') based on the Age column.
    - Calculated Revenue per Item (Total Amount / Quantity). *Note: This wasn't used in the provided modeling section but was created.*
    - For modeling, categorical features (Gender, Product Category) were label encoded into numerical representations.
- **Exploratory Data Analysis (EDA) & Visualization:**
    - **Sales by Product Category:** A bar plot visualized the total sales amount for each product category (Clothing, Beauty, Electronics).

- **Monthly Revenue Trends:** A line plot showed the total revenue generated per month over the period covered by the data.
- **Customer Age Distribution:** A histogram displayed the frequency distribution of customer ages.

- **Modeling (Predicting Total Amount):**
  - **Features (X):** Gender (encoded), Age, Quantity, Price per Unit.
  - **Target (y):** Total Amount.
  - **Data Split:** The data was split into training (80%) and testing (20%) sets.
  - **Models Trained:**
    1. Linear Regression
    2. Random Forest Regressor
    3. XGBoost Regressor
    4. Support Vector Regressor (SVR)
  - **Evaluation Metrics:** Mean Absolute Error (MAE) and R-squared (R^2) were calculated on the test set.
  - **Prediction Visualization:** A line plot compared actual vs. predicted Total Amount for a sample of 100 test data points using the Linear Regression model.

3. Results:

- **EDA Insights:**
  - *Product Sales:* The bar plot indicated that 'Clothing' and 'Electronics' generated the highest total revenue, while 'Beauty' generated significantly less.
  - *Monthly Trends:* The line plot showed fluctuations in monthly revenue, suggesting potential seasonality or specific sales events driving revenue changes (exact trends depend on the specific data period).
  - *Age Distribution:* The histogram revealed the concentration of customer ages, likely showing a peak in the 'Adult' age group.

- **Model Performance:**
    - **Linear Regression:** MAE = 172.95, R^2 = 0.857
    - **Random Forest:** MAE = 0.0, R^2 = 1.0
    - **XGBoost:** MAE = 0.015, R^2 = ~1.0 (0.99999...)
    - **SVR:** MAE = 259.45, R^2 = 0.282
- **Discussion of Model Results:**
    - The Linear Regression model showed a reasonable fit ($R^2 \approx 0.86$), capturing a significant portion of the variance, although with a noticeable error (MAE ≈ 173). The actual vs. predicted plot visually confirms this.
    - Random Forest and XGBoost achieved perfect or near-perfect scores ($R^2 \approx 1.0$, MAE ≈ 0). This is because the target variable, Total Amount, is a direct calculation from two of the input features (Quantity * Price per Unit). Tree-based models like RF and XGBoost are highly effective at learning such exact deterministic relationships from the training data, leading to perfect predictions on the test set.
    - SVR performed poorly compared to the other models on this specific task.
    - The perfect scores highlight that predicting Total Amount using Quantity and Price per Unit as features is trivial for capable models. A more challenging and potentially useful prediction task might involve predicting Product Category, Quantity, or forecasting future sales.

4. Conclusion:

The analysis provided insights into sales distribution across product categories, monthly revenue patterns, and customer demographics. The modeling exercise demonstrated the predictability of the Total Amount when Quantity and Price per Unit are included as features, with Random Forest and XGBoost perfectly capturing the underlying relationship. Linear Regression offered a good approximation, while SVR struggled. For future work, focusing on predicting less deterministic variables or forecasting would be more indicative of real-world predictive power.