Project Submission Document- Sai Jyothsna Pulivarthi

Project Portfolio & Case Studies

1. E-commerce Data Analysis

GitHub Repository: E-commerce Data Analysis

I have Worked on an E-commerce project dealing with large data sets. As I can not present those work samples I have recreated an E-commerce project that can demonstrate similar Insights and visualization. I have collected this dataset from https://www.kaggle.com/

Objective:

Analyzing customer behavior and transaction patterns is crucial for e-commerce businesses. My goal in this project was to extract meaningful insights from customer purchase data, identify key trends, and develop predictive models to assist in decision-making. The analysis covers customer segmentation, revenue impact of discounts, and forecasting customer spending behavior.

Tools Used:

- **Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn)**
- **Jupyter Notebooks**
- **SQL for data querying**

Methodologies Applied:

1. **Data Cleaning**

    - **The dataset contained inconsistencies, missing values, and redundant information, which I addressed through preprocessing.**
    - **Date formats and categorical variables were standardized to ensure uniformity across the dataset.**
    - **Duplicates were identified and removed to maintain data accuracy and consistency.**
    - **Reference:** 01_data_cleaning.ipynb

2. **Exploratory Data Analysis (EDA)**

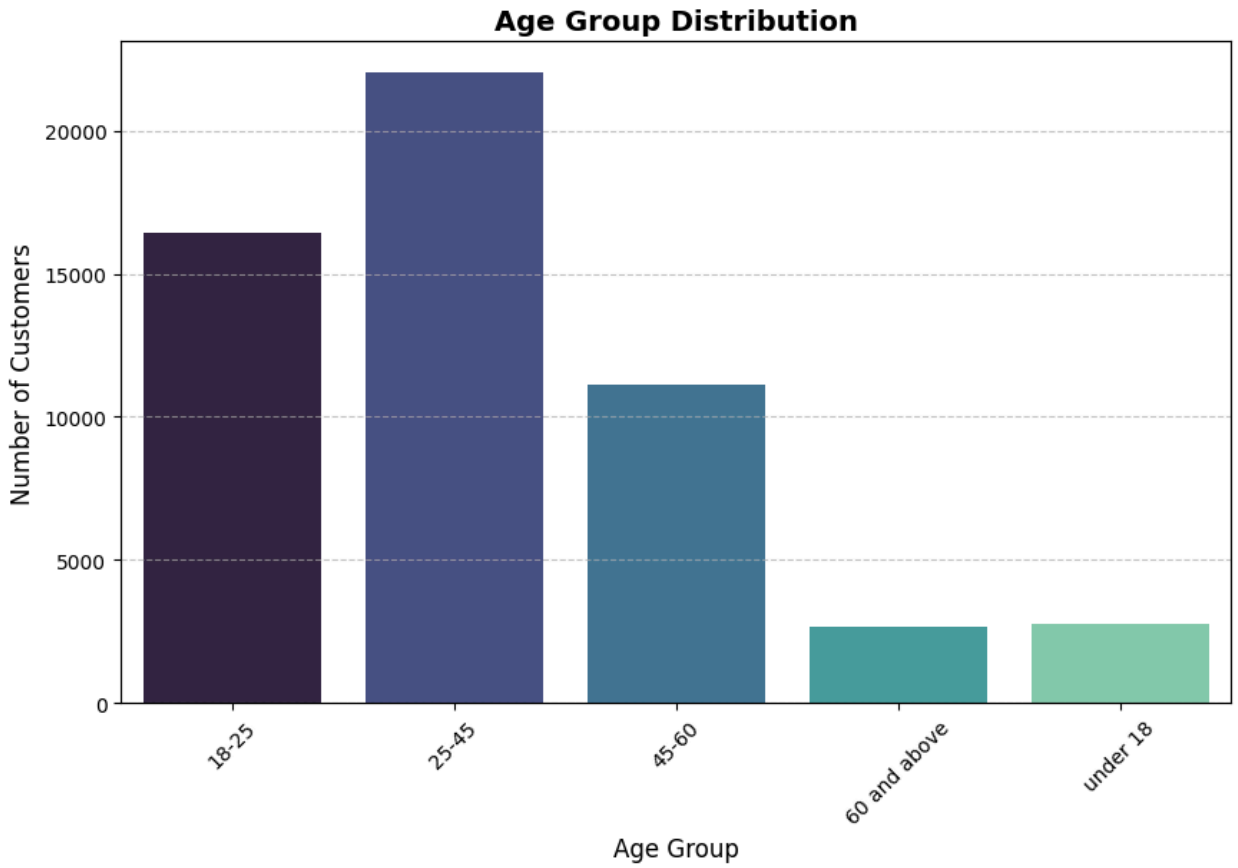
    - **I conducted an in-depth analysis of customer demographics, purchasing trends, and product preferences.**
    - **Visualizations helped me uncover trends in sales, discount effectiveness, and customer retention.**
    - **I examined how discount strategies influenced customer purchases to optimize future marketing campaigns.**
    - **Reference: 02_eda.ipynb**
3. **Feature Engineering**

    - **I created new features such as Customer Lifetime Value (CLV), Recency, Frequency, and Monetary (RFM) metrics to strengthen predictive modeling.**
    - **Customer Lifetime Value (CLV): This metric estimates the total revenue a business can expect from a customer over their relationship duration. It helps identify high-value customers and optimize marketing efforts.**
    - **Recency: Represents the time since a customer's last purchase, useful for predicting future purchase likelihood.**
    - **Frequency: Measures how often a customer makes purchases, indicating their engagement level.**
    - **Monetary (RFM) Analysis: Determines spending behavior, classifying customers into different value groups.**
    - **Reference: 03_feature_engineering.ipynb**
4. **Model Training**

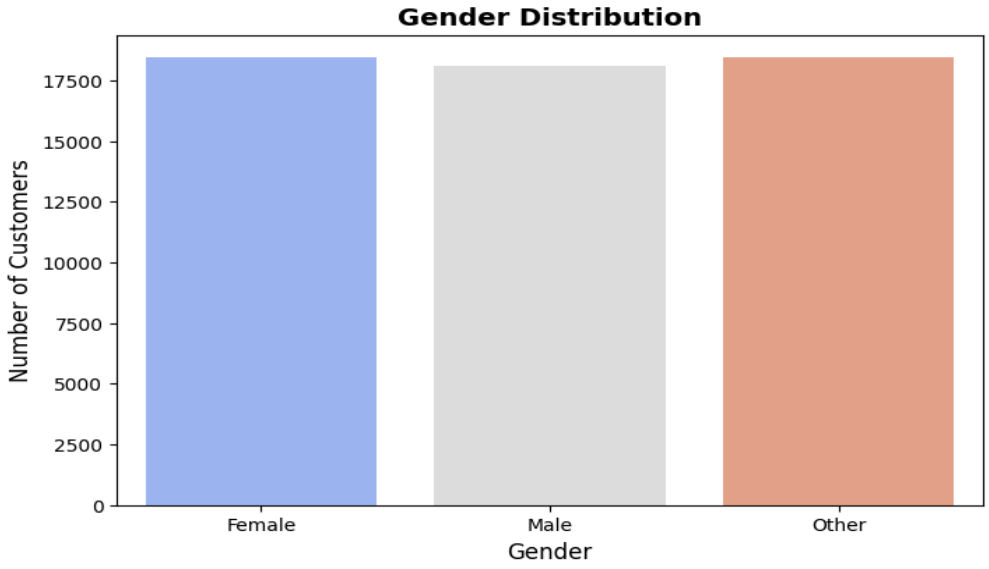
    - **I trained a Random Forest Classifier to predict customer churn based on CLV, Recency, Frequency, and Engagement Level.**
    - **The model was trained using a well-balanced dataset after performing label encoding on categorical variables like Engagement Level.**
    - **Feature Scaling: I applied StandardScaler to normalize numerical features for improved model performance.**
    - **Performance Evaluation: The model was assessed using key metrics like Classification Report, Confusion Matrix, and ROC-AUC Score to ensure accuracy and reliability.**
    - **Visualizations: Histograms were generated to compare training and testing data distributions for features like CLV, Recency, and Frequency.**
    - **Reference: 04_model_training.ipynb**

Findings from Visualizations:

➔ **Customer Demographics Analysis:**



➔ **The majority of customers fall into the 25-45 age group, with a significant portion being repeat buyers.**



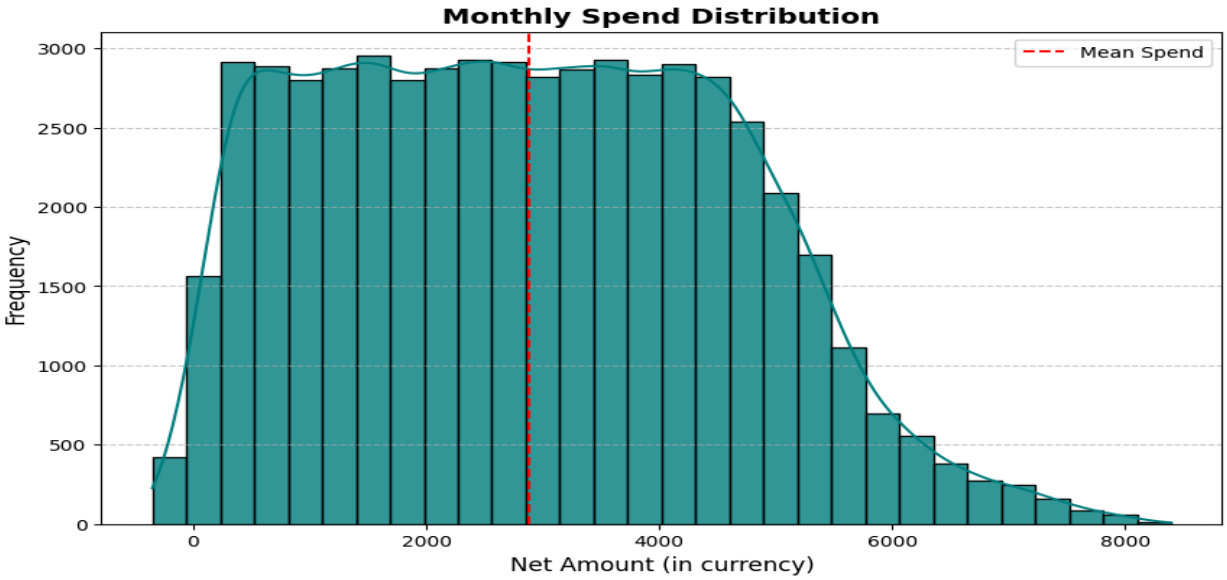
- ➔ **Gender distribution shows a slight skew toward female customers.**

➔ **Discount Impact Analysis:**

- ◆ **While discounts increase sales volume, they do not always lead to higher profit margins.**
- ◆ **Returning customers tend to purchase at regular prices rather than relying on discounts.**

➔ **Revenue Contribution by Customer Segments:**

- ◆ **High-value customers contribute disproportionately to total revenue.**



- **Identifying these customers allows businesses to tailor personalized marketing efforts.**

Outcomes:

- **I successfully identified high-value customers using CLV analysis, aiding in targeted marketing.**
- **Discount strategies were evaluated for their effectiveness in boosting sales while maintaining profitability.**
- **The predictive model provides insights into expected future sales based on past purchase behaviors.**

2. Cricket Analysis - SQL-Based Data Processing

GitHub Repository: Cricket Analysis

Objective:

This project is centered on structuring cricket match data using SQL, focusing on designing efficient queries, stored procedures, and views to streamline data retrieval and analysis.

Tools Used:

- **SQL (MySQL, PostgreSQL)**
- **Stored Procedures & Views**
- **Data Aggregation Functions**

Methodologies Applied:

1. **SQL Query Development**
    - **I wrote SQL queries to extract match statistics, such as player averages, team performance, and match outcomes.**
    - **SQL views were created to simplify structured data retrieval for further analysis.**
2. **Stored Procedures Implementation**
    - **Automated stored procedures were developed to calculate batting and bowling statistics efficiently.**
    - **I ensured modularity so that the procedures could be reused for future cricket datasets.**
3. **Optimized Data Retrieval**
    - **Indexing and query optimization techniques were implemented to enhance database performance.**
    - **The structured approach allows for handling large volumes of match data efficiently.**

Outcomes:

- **I developed efficient SQL queries and stored procedures for structured cricket match data analysis.**
- **Database performance was optimized, ensuring faster query execution.**
- **The SQL-based solution provides reusable components for similar future projects.**

Code Samples & Notebooks

- **E-commerce Data Analysis: Jupyter notebooks document the entire workflow, from data cleaning to model building.**
- **Cricket Analysis: SQL scripts demonstrate structured query design and stored procedure implementation.**

Data Visualizations & Dashboards

E-commerce Project:

- **Customer segmentation and purchase trends visualized using bar charts and scatter plots.**
- **The effectiveness of discounts displayed through comparative revenue charts**
- **Sales predictions plotted against historical data for validation.**

Documentation & Reports

- **E-commerce Analysis: Markdown explanations are included in Jupyter notebooks to provide clear context on each step.**
- **Cricket Analysis: SQL documentation outlines query logic, stored procedures, and optimization strategies.**

Data Samples & Management

➢ **Datasets Used:**

dataset.csv: Original e-commerce transaction data.

cleaned_dataset.csv: Processed dataset after data cleaning.

engineered_dataset.csv: Feature-enriched dataset for modeling.

➢ **Data Quality Management:**

- **I ensured data integrity through validation checks and preprocessing steps.**
- **Outliers and missing values were carefully handled to maintain data consistency.**
- **Query performance was optimized for handling large datasets efficiently.**

GitHub Repositories:

1. **E-commerce Data Analysis**
2. **Cricket Analysis**