**StudentID: 2023-PHDCS-09**

**Name: Khaqan Khawer**

**Project Report: E-Commerce Data Analysis & Predictive Modeling**

**1. Problem Definition and Objectives:**

**Problem Definition:**

The primary objective of this project is to analyze the large e-commerce dataset to uncover insights into sales patterns, order statuses, and customer behavior. The goal is to use exploratory data analysis (EDA) to understand factors that affect sales and customer actions, with the ultimate aim of **predicting key metrics** such as **order status**, **revenue generation**, **discount patterns**, and **sales volume**.

**Key Objectives:**

* **Explore customer purchase behavior**: Understand the relationship between product categories, payment methods, order statuses, and sales.
* **Identify trends**: Investigate seasonal variations and product demand patterns.
* **Sales and revenue analysis**: Examine total sales, quantities, and discounts across different months and product categories.
* **Predictive Modeling**: Build a predictive model to forecast future sales or order statuses using machine learning techniques like linear regression or logistic regression.

**2. Questions Posed and Their Relevance:**

The following research questions guided the exploratory analysis:

1. **What is the best-selling product category?**
   * **Relevance**: Identifying top-selling categories can help allocate resources efficiently, plan inventory, and target marketing efforts.
2. **What payment methods are most commonly used?**
   * **Relevance**: Payment method preferences provide insights into customer preferences and can inform payment strategy.
3. **What is the relationship between order status and sales metrics?**
   * **Relevance**: Understanding this relationship can help in operational planning, such as inventory management, shipping, and customer support.
4. **What is the impact of discounts on sales?**
   * **Relevance**: Understanding the discount-to-sales ratio helps assess the effectiveness of promotions and pricing strategies.
5. **What are the trends in sales across different months or seasons?**
   * **Relevance**: Seasonal trends help businesses plan for demand fluctuations, promotions, and inventory levels.
6. **Can we predict future sales or order statuses using existing data?**
   * **Relevance**: Predictive models help forecast future sales, enabling better planning and decision-making.

**3. Data Source and Description:**

**Data Source:**

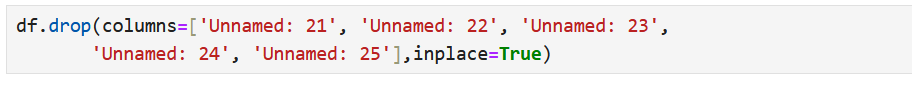
The dataset is taken from Kaggle: **Pakistan's Largest E-Commerce Dataset** (link: [Kaggle Dataset](https://www.kaggle.com/datasets/zusmani/pakistans-largest-ecommerce-dataset)).

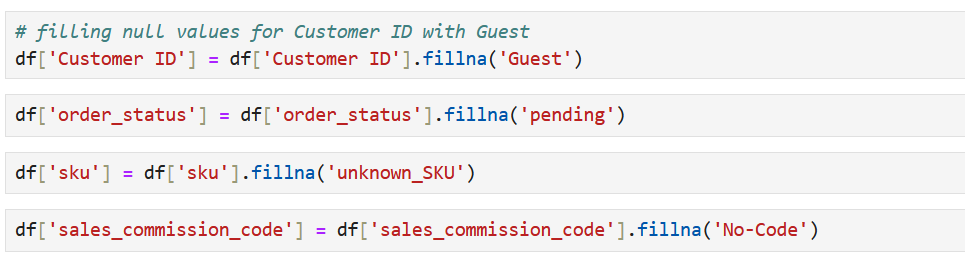
**Dataset Description:**

The dataset contains transaction data from an e-commerce platform and includes the following columns:

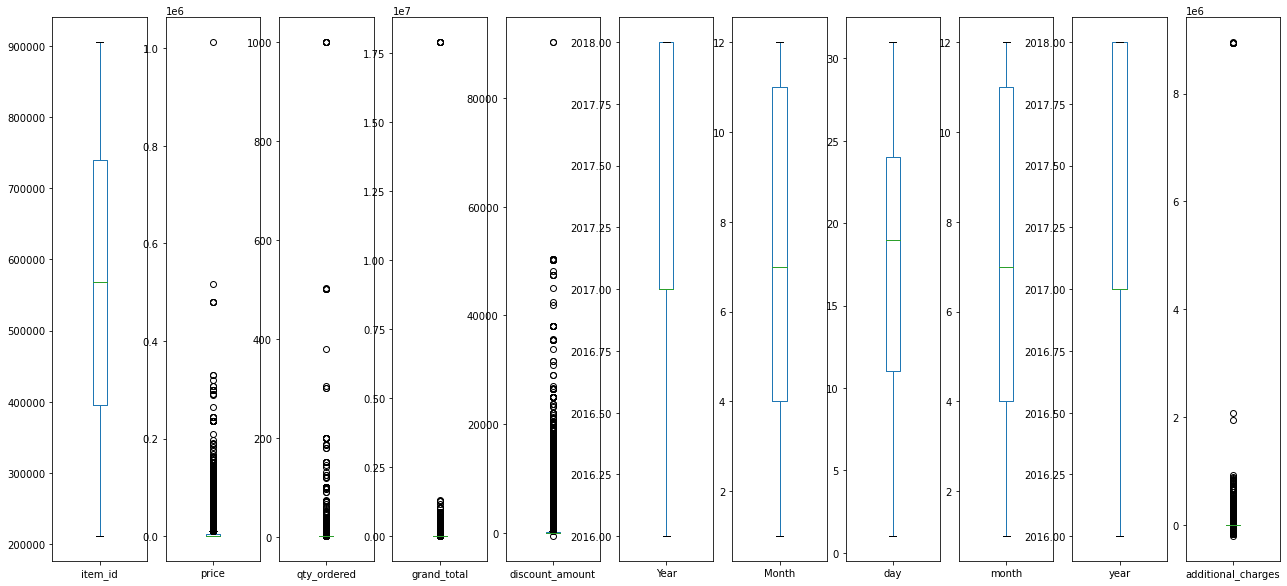
1. **item\_id**: Unique identifier for each product.
2. **status**: Current status of the order (e.g., completed, pending).
3. **created\_at**: Timestamp when the order was placed.
4. **sku**: Stock-keeping unit of the product.
5. **price**: Price per unit of the product.
6. **qty\_ordered**: Quantity of the product ordered.
7. **grand\_total**: Total price of the order after all calculations.
8. **increment\_id**: Unique identifier for each order.
9. **category\_name\_1**: Product category.
10. **sales\_commission\_code**: Sales commission information.
11. **discount\_amount**: Discount applied to the order.
12. **payment\_method**: Payment method used for the order.
13. **Working Date**: Date the order was processed.
14. **BI Status**: Business intelligence status, such as shipped or pending.
15. **MV**: Value associated with the product or order.
16. **Year**: Year of the order.
17. **Month**: Month of the order.
18. **Customer Since**: Date the customer first placed an order.
19. **M-Y**: Month-Year formatted as M-Y.
20. **FY**: Fiscal year.
21. **Customer ID**: Unique identifier for customers.

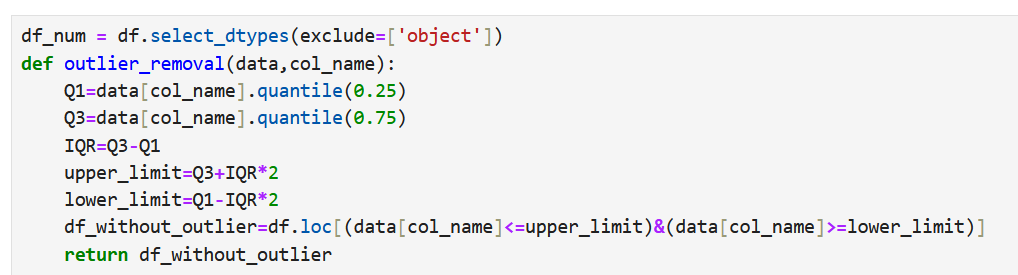
**4. Data Wrangling and Cleaning Steps:**

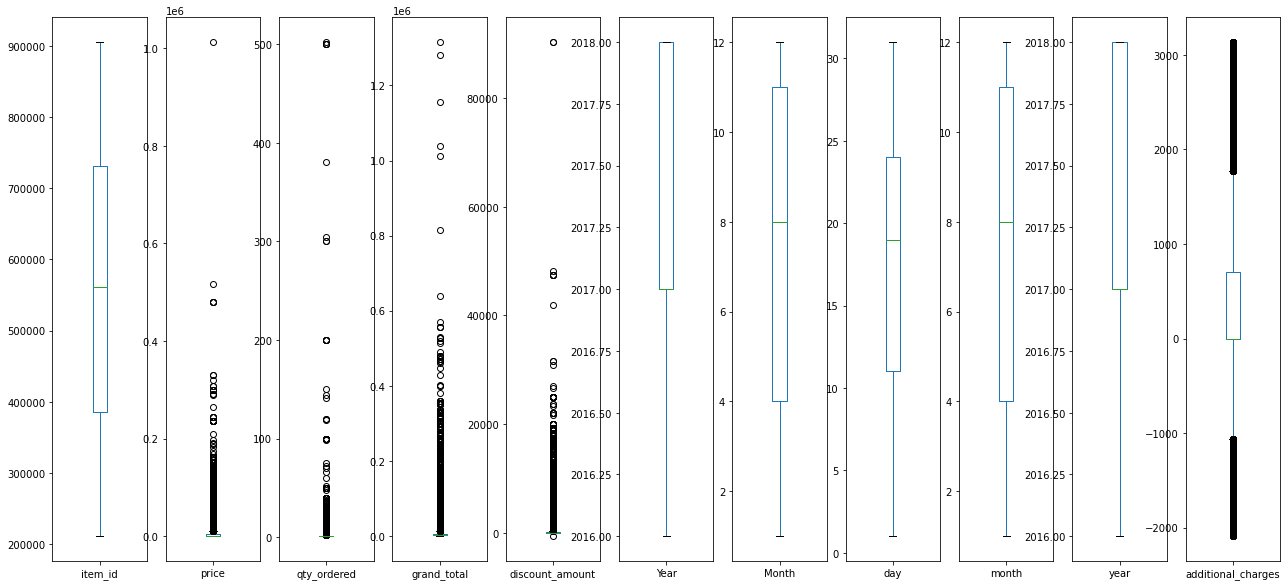
1. **Handling Missing Values**:
   * Identified missing values in certain columns like Unnamed: 21 to Unnamed: 25, which were irrelevant, and removed th
   * Filled or removed missing values in key columns such as status, customer\_ID, order\_status etc.



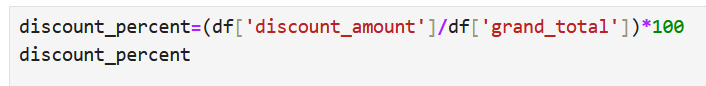
1. **Correcting Data Types**:
   * The created\_at and Working Date columns were in string format and were converted to datetime objects to enable temporal analysis.
   * 
   * The year, month, and Customer Since columns were correctly typed as integers and dates.
2. **Handling Outliers**:
   * Removed extreme outliers in numerical columns (e.g., price, grand\_total) based on domain knowledge to ensure a more accurate analysis.



* + Applied **Z-score** or **IQR** methods to detect and handle outliers for specific columns like qty\_ordered and grand\_total.

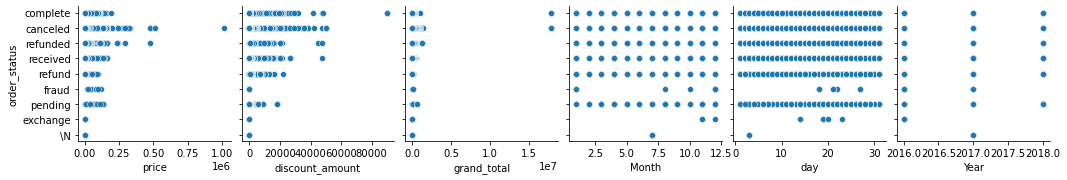
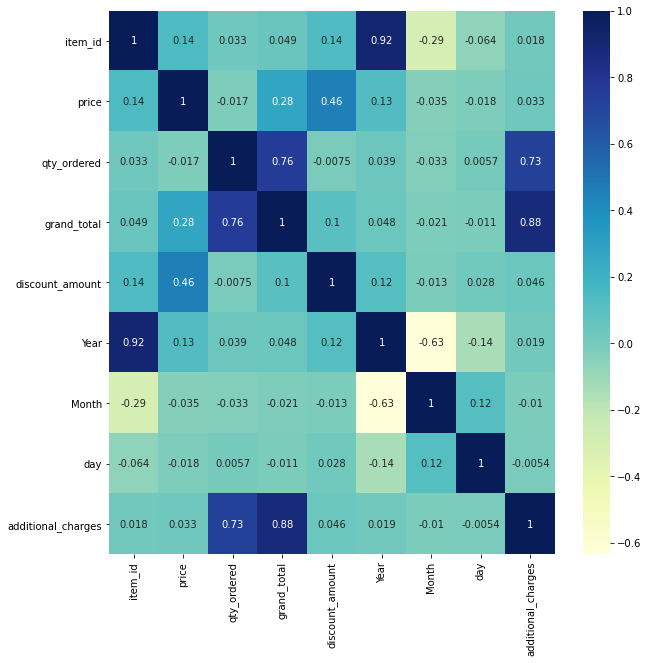


1. **Feature Engineering**:
   * Created new features, such as discount\_percentage (discount\_amount / grand\_total), price\_per\_item (grand\_total / qty\_ordered)



* + Used **One-Hot Encoding** for categorical features like category\_name\_1 and payment\_method.



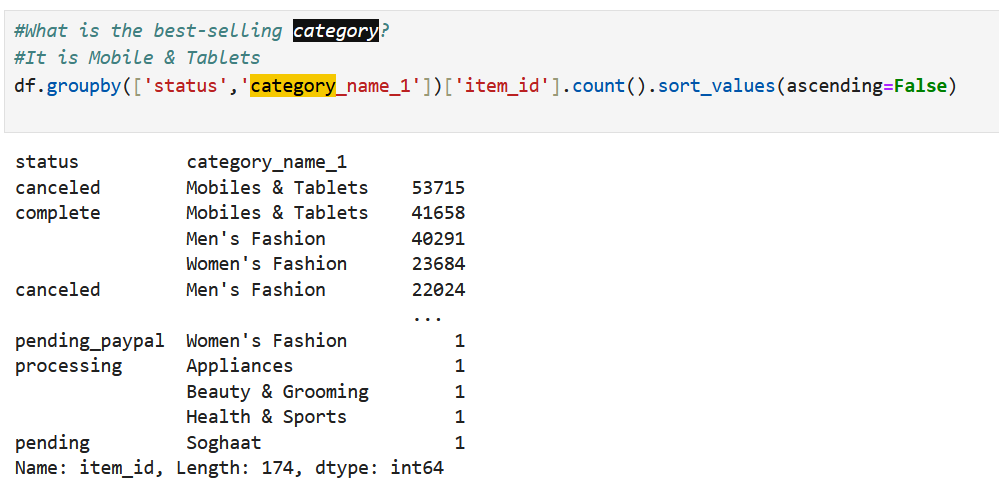


1. **Aggregating Data**:
   * Aggregated data on a **monthly** level to analyze sales trends across time and calculated metrics such as total sales, quantity sold, and average discounts per month.
   * Analyzed customer-level data by aggregating purchases per Customer ID to determine frequent buyers.

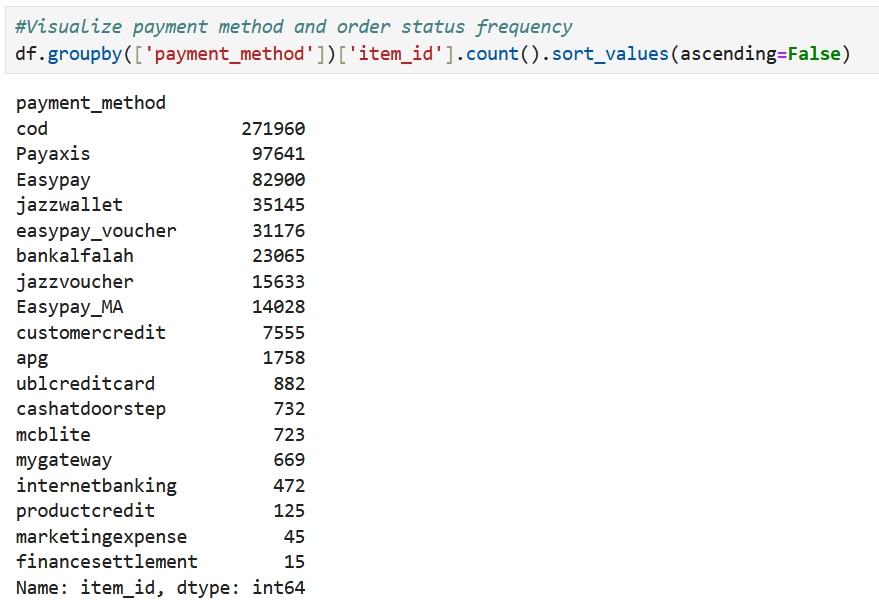
**5. Insights from Exploratory Data Analysis (EDA):**

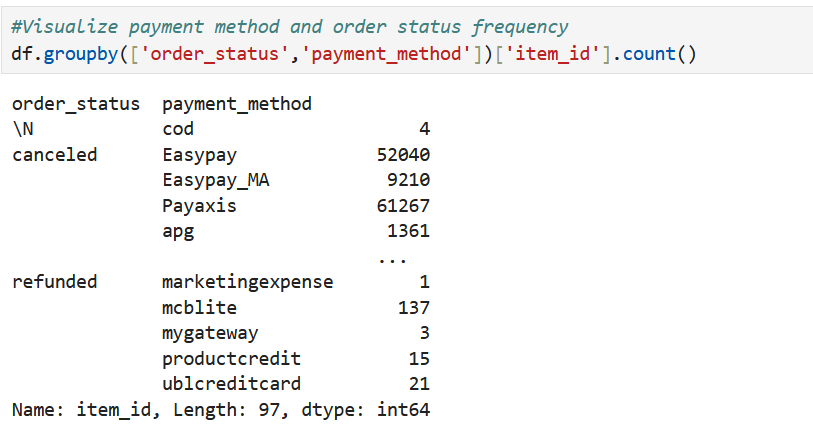
**Key Insights:**

1. **Best-Selling Category**:
   * The **Mobile & Tablets** category emerged as the best-selling category based on the number of units sold.

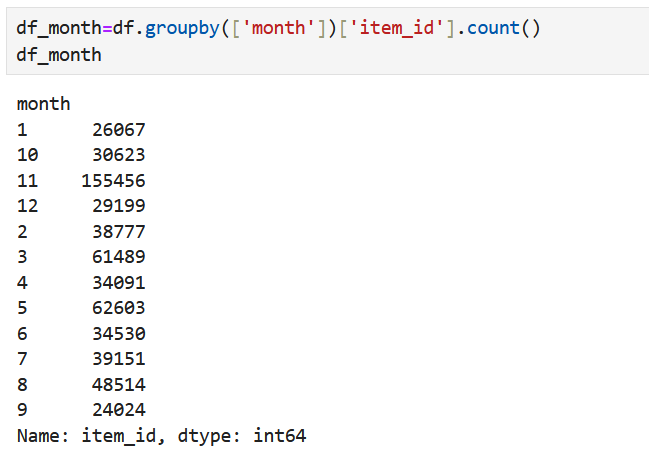


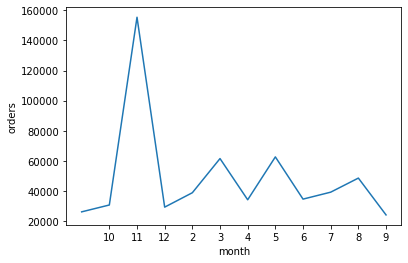
1. **Payment Method and Order Status**:
   * **Credit Card** was the most frequently used payment method, followed by **Cash on Delivery**.
   * The majority of orders were **completed**, with a small percentage marked as **cancelled** or **pending**.





1. **Sales Over Time**:
   * Total **grand total** sales increased significantly during certain months (e.g., November), showing the impact of **11.11 sale** in e-commerce.
   * There is a clear upward trend in **discount amounts** in months leading up to festive seasons.

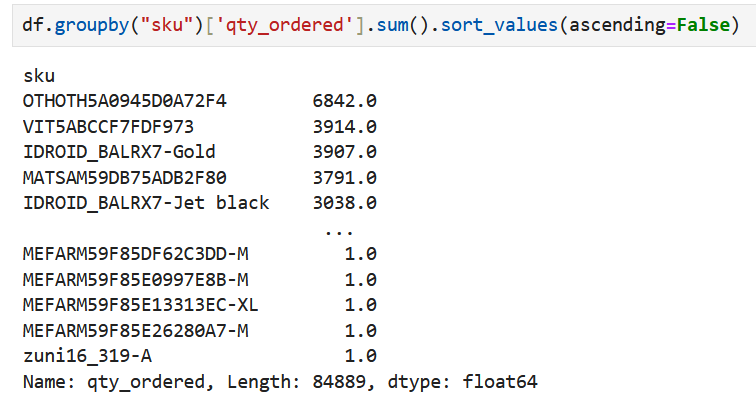




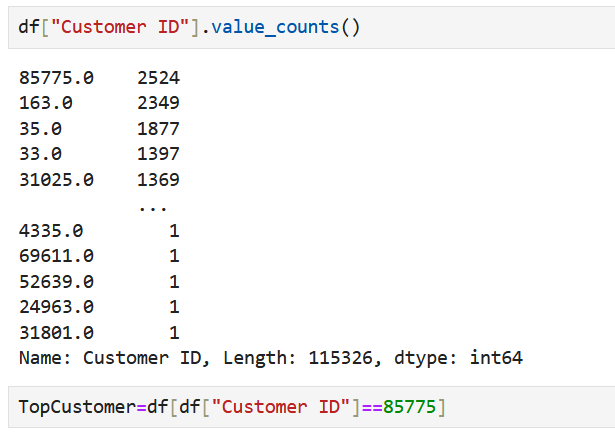
1. **Discount Analysis**:
   * **Discounts** had a significant impact on sales, but the **discount-to-revenue ratio** varied seasonally. Higher discounts often led to a higher sales volume in certain months.



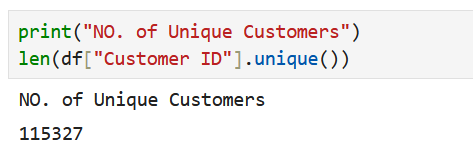
1. **Product and SKU Performance**:
   * Certain products had significantly higher **quantity ordered**, indicating strong demand in specific SKUs, often related to mobile phones and accessories.



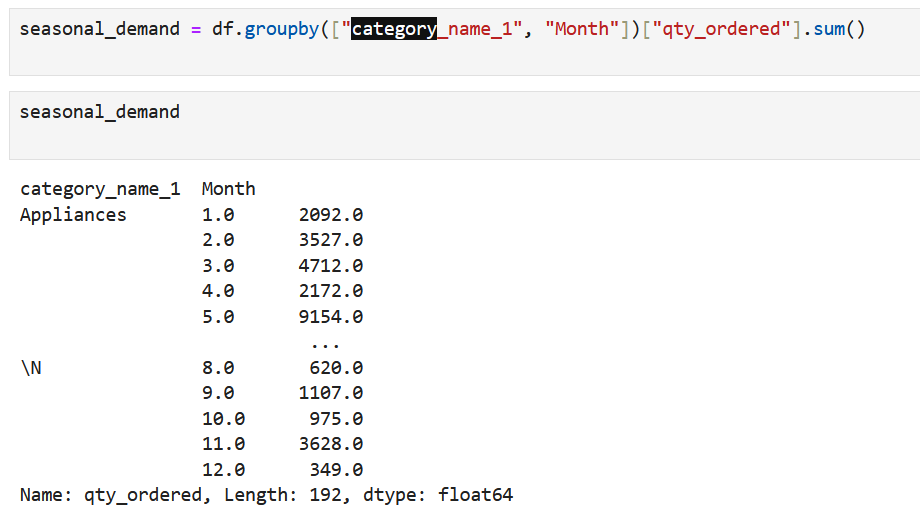
1. **Customer Insights**:
   * Customers who have been on the platform longer tend to make larger purchases and buy more frequently.



* + A considerable portion of the customers were **new** (based on the Customer Since field), indicating the growing customer base.



1. **Seasonal Demand**:
   * There was a noticeable **seasonal pattern** in product demand across categories, with certain months showing higher sales, likely due to holidays or promotional events.



***6*. Model Choice, Performance Metrics, and Interpretation:**

**Model Choice:**

Given the dataset's characteristics, the following models were considered for predicting continuous outcomes (e.g., revenue or order quantity) or categorical outcomes (e.g., order status):

1. **Linear Regression**: For predicting continuous variables like **grand\_total**, **qty\_ordered**, or **discount\_amount**.

**Example Model (Linear Regression for Predicting Revenue):**

* **Target Variable**: grand\_total (Total order value)
* **Features**: price, qty\_ordered, discount\_amount, category\_name\_1, year, month, days\_since\_first\_purchase, etc.

**Steps:**

1. Preprocess data (One-hot encoding, scaling, handling missing values).
2. Split data into training and test sets.
3. Train the model on the training data.
4. Evaluate using **R-squared** and **Mean Squared Error**.



**Evaluation Metrics:**

* **R-squared (R²)**: Measures how well the model explains the variance in the target variable.
* **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted values.

**Example Evaluation:**

