**Milestone:1 Exploratory Data Analysis**

**Retail Sales Analysis**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sno** | **Use Case** | **Skill Covered** | **Expected Outcome** | **Complexity** | **Industry** | **Total Marks** |
| 1 | **Milestone :1**  Exploratory Data Analysis | Data Manipulation  Statistics  Visualization | Data Understanding & Cleaning  Descriptive Statistics  Data Manipulation & Transformation  Visualization | Medium-Complex | Retail | 50 |

A US-based retail store has provided historical sales data, capturing transactions across different product categories, customer segments, and regions. The company aims to analyse store performance, customer purchasing behaviour, and the impact of discounts on profitability.

**The goal is to:**

* Segment customers based on purchasing behaviour and loyalty.
* Identify sales trends across different time periods and regions.
* Evaluate discount strategies to optimize profitability.
* Provide AI-driven business recommendations to improve sales, marketing effectiveness, and customer retention strategies.

**Dataset**: store.csv

# **Data Description:**

|  |  |
| --- | --- |
| **Order ID** | A unique identifier for each order placed by a customer. |
| **Order Date** | The date when the order was placed, useful for time-series analysis and sales trend predictions. |
| **Ship Date** | The date when the order was shipped, useful for analyzing shipping times and operational efficiency. |
| **Customer ID** | A unique identifier for each customer, allowing for customer segmentation and lifetime value analysis. |
| **Customer Name** | The name of the customer, often used for customer-centric reporting. |
| **Segment** | The business segment of the customer, such as Consumer, Corporate, or Home Office. |
| **Region** | The geographical region where the sale occurred (e.g., East, West, Central, South). |
| **Product Category** | The category of the purchased product, such as Furniture, Office Supplies, or Technology. |
| **Product Sub-Category** | A more specific classification of the product (e.g., Chairs, Phones, Binders). |
| **Product ID** | A unique identifier for each product in the inventory. |
| **Product Name** | The name or description of the product. |
| **Quantity** | The number of units sold for a particular product in a single transaction. |
| **Discount** | The discount applied to the product during the transaction, expressed as a percentage. |
| **Sales** | The revenue generated from the transaction (after discounts). |
| **Profit** | The monetary profit earned from the transaction (Sales - Cost). |
| **Shipping Mode** | The method of shipping used (e.g., Standard Class, Second Class, Same Day). |

**Initial Guidelines:**

1. Ensure to follow to Use Id’s provided by UNext for naming file as conventions.
2. Create GitHub account and submit the GitHub link.

**Task 1: Load the dataset and perform preliminary EDA (Exploratory Data Analysis) with key observations and insights- (weightage - 40 marks)**

|  |  |  |  |
| --- | --- | --- | --- |
| **T1.1** | Load the store dataset. | (weightage - 2marks) | **AE** |
| **T1.2** | (i)Count duplicate rows and remove them.  (ii) Find unique values in "Customer Segment", "Ship Mode", "Product Category" | (weightage - 2marks) | **AE** |
| **T1.3** | **Overall Sales Performance:** (i) What is the total sales volume (Total Amount) over the entire dataset period?  (ii) Are there any missing values in the data? If yes, handle the missing values by using appropriate technique.  (iii) What is the distribution of Sales? | (weightage - 3marks) | **AE** |
| **T1.4** | **Customer Segmentation:**  (i)Which customer segment (Consumer, Corporate, Home Office) contributes the most to total sales and profits?  (ii) How does customer purchasing behaviour vary across different regions?  (iii) Which regions have the highest number of high-spending customers? | (weightage –3marks) | **AE** |
| **T1.5** | **Product Category Analysis**:  (i)Which Product Categories are most popular in terms of quantity sold and total sales revenue?  (ii) What is the average quantity purchased per transaction for each product category? | (weightage –2marks) | **AE** |
| **T1.6** | **Product Preferences & Purchasing Behavior:**  **(i)**What is the relationship between Quantity and Sales?  (ii) Are there any common trends or patterns in Quantity and Sales for specific Product Categories?  (iii) Which customer segments generate the highest profits? | (weightage –3marks) | **AE** |
| **T1.7** | **Calculate new metrics**  Shipping time: Ship Date - Order Date to get delivery time.  Discounted Price**:** Sales - (Sales \* Discount) to determine the actual price after applying the discount.  Revenue per Day**:** Sales / Shipping Time to estimate how much revenue is generated per day of shipping. | (weightage –3marks) | **AE** |
| **T1.8** | **Identify Most Discounted Products**  Find the top 10 products with the highest discount applied. | (weightage –2marks) | **AE** |
| **T1.9** | **Analyze Revenue Efficiency**  Determine the **top 5 orders** with the highest "Revenue per Day" | (weightage –2marks) | **AE** |
| **T1.10** | **Visualization:**  (i)Plot a sales distribution histogram.  (ii) Create a bar chart showing sales by customer segment.  (iii)Plot a line chart of total sales over time (monthly or yearly trends). (iv) Compare order count trends over time. | (weightage - 4marks) | **ME** |
| **T1.11** | **Validate Hypotheses**  (i)Is there a significant difference in the average sales across product categories? Apply ANOVA.  (ii) Is there a significant difference in sales performance between Consumer and Corporate Segments? Apply an independent samples t-test.  (iii) Is there a relationship between product category and Region? Apply the chi-square test. | (weightage - 6marks) | **AE** |
| **T1.12** | Detect Outliers in the data and do outlier treatment.  Plot a boxplot to visualize outliers in the data. | (weightage –3marks) | **AE  ME** |
| **T1.13** | **Analyse the relationships between different variables:** Create a correlation matrix and visualize it using a heatmap. | (weightage - 1mark) | **AE** |
| **T1.14** | **Data Transformation**: remove the unnecessary variable for model building process [“Order\_ID", “CustomerID”]. Apply normalization technique (standard scaler). | (weightage - 2marks) | **AE** |
| **T1.15** | **Handling categorical features**: Apply encoding technique to convert categorical variable into numerical. | (weightage - 2marks) | **AE** |

**Task 2: Summarize the findings of the analysis and draw conclusions with PPT / PDF.**  **(weightage - 10 marks). ME**

**AE Score: 35 ME Score: 15**

**Final Submission guidelines:**

1. Download the Jupyter notebook in the format of html.
2. Upload it in the lumen (UNext LMS)
3. Summarized PPT/ PDF prepared in Task 4 to be uploaded in the lumen (UNext LMS)

-----------------------------------------------------------------------------------------------