**Milestone 1 & 2 Report**

**Dataset Description:**

| | |
|---|---|
| **Transaction Data** | **Transaction_ID:** Unique identifier for each transaction.<br>**Customer_ID:** Unique identifier for the customer associated with the transaction.<br>**Product_ID:** Unique identifier for the product purchased in the transaction.<br>**Transaction_Date:** The date when the transaction occurred. |
| **Product Information** | **Category:** The category to which the product belongs (i.e., Electronics, Clothing).<br>**Units_Sold:** The quantity of the product sold in the transaction.<br>**Discount_Applied:** The discount percentage applied to the product during the transaction.<br>**Revenue:** Total revenue generated from the transaction, calculated as Price x Units Sold x (1 - Discount). |
| **Customer Demographics** | **Customer_ID:** Unique identifier for each customer (repeated for easier reference). |
| **Advertising Metrics** | **Clicks:** Number of ad clicks associated with the product during the time of the transaction.<br>**Impressions:** Number of ad impressions served during the campaign.<br>**Conversion_Rate:** Calculated as Clicks / Impressions, representing the percentage of impressions that resulted in clicks.<br>**Ad_CTR:** Click-through rate (CTR) for the advertisement, representing the effectiveness of the ad campaign.<br>**Ad_CPC:** Cost-per-click for the advertisement.<br>**Ad_Spend:** Total advertising spend for the product, calculated as Ad_CTR x Ad_CPC x 1000. |

| | |
|---|---|
| **Seasonal and Regional Information** | **Region:** The geographical region where the transaction occurred (e.g., North America, Europe, Asia).<br>**Seasonality Effects:** Implied through patterns in transaction dates and revenue, reflecting holiday promotions and season-based purchasing trends. |

**Data Cleaning:**

We have checked the missing values, duplicates in addition to removing the trivial columns ('Transaction_ID', 'Customer_ID', 'Product_ID') and converting ‘Transaction_Date’ into a time column. Furthermore, we have used the interquartile range method to remove the outliers, remove the highly correlated features to avoid overfitting occurrence, and round the float columns to avoid noise effects which may sometimes lead to overfitting. The categorical (nominal) columns are dealt with through one-hot encoding then checking the categories balance, finding that almost all the categories are divided equally. We also have checked each column distribution to do the scaling based on their distribution: for the ('Ad_Spend', 'Revenue', 'Conversion_Rate'), they are exponentially distributed so a logarithmic scaling is done while for the ('Ad_CPC', 'Ad_CTR', 'Clicks', 'Impressions', 'Discount_Applied'), they are uniformly distributed so a minmax scaling is done.

**Data visualization:**

**Univariate analysis:** 1. We did pie chart for the ‘Region’ and ‘Category’ columns with finding that almost the distribution are balanced in both categorical features

2. We also did a histogram for the numerical variables to check the data distribution in order to decide which scaling technique will be used

**Bivariate Analysis:** 1. We have checked the Revenue over time (aggregated Monthly), finding that November is the month with the highest revenue which is reassured through checking the units sold among months to get November also scores the highest all over the year.

2. We have checked the revenue per category, finding almost equal revenues with the electronics, the highest one but a very slight increase among the other categories with checking the revenue per region, which appears to be similar across them.

3. We have looked for the Units Sold by Region and Season, getting that the highest number of units sold is in North America in November and also checked the advertising effects on the revenue which is observed to have a real impact.

4. We have also checked the holidays versus the units sold because most of the time, the businesses earn more at holidays, so we have found that the median number of units

sold is slightly higher on holidays than on non-holidays which indicates a potential boost in sales during holidays. However, there are significantly more extreme high outliers on non-holidays than holidays, indicating occasional promotional campaigns or irregular sales spikes on non-holidays to attract customers.

5. We have done a time series plot which shows as the previous plots that November has the highest number of ‘units_sold’.

**Multivariate Analysis:** we have done a pairplot between the columns and the target feature to check how they are related to each other.

**Feature engineering:**

We have checked the correlation between the numerical features to remove the highly correlated ones with the target column to prevent or lessen the probability of overfitting occurrence, then checked the heatmap of revenue by day of week and month. We found that October and November have highest correlations as demonstrated from the other analysis in addition to having the highest correlation on Friday (which does make sense).