**InnoByte Data Analysis Internship**

**Problem Statement:**

Analyze and Provide actionable Insights on Amazon Sales Reports to support business decision-making.

**Problem Description:**

The provided dataset contains information about sales transactions on Amazon, including details such as order ID, date, status, fulfilment method, sales channel, product category, size, quantity, amount, shipping details, and more. The objective is to conduct a comprehensive analysis of the data and extract actionable insights to support business decision-making.

**Datasets:** [**Datasets**](https://drive.google.com/file/d/1YrjYKtS1WHmINL6eafRsrDzrZaw2_WvX/view?usp=sharing)

**Key Objectives:**

**1. Sales Overview:** Understand the overall sales performance, trends, and patterns overcome.

**2. Product Analysis:** Analyze the distribution of product categories, sizes, and quantities sold to identify popular products.

**3. Fulfilment Analysis:** Investigate the fulfilment methods used and their effectiveness in delivering orders.

**4. Customer Segmentation:** Segment customers based on their buying behaviour, location, and other relevant factors.

**5. Geographical Analysis:** Explore the geographical distribution of sales, focusing on states and cities.

**6. Business Insights:** Provide actionable insights and recommendations based on the analysis to optimize sales strategies, improve customer satisfaction, and enhance overall business performance.

**Deliverables:**

1. Comprehensive analysis report summarizing key findings, insights, and recommendations.

2. Visualizations (charts, graphs) illustrating various aspects of the data analysis.

3. Insights on product preferences, customer behaviour, and geographical sales distribution.

4. Recommendations for improving sales strategies, inventory management, and customer service.

**Expected Outcome:**

By conducting a thorough analysis of the Amazon sales report, the goal is to gain valuable insights that can be leveraged to optimize business operations, enhance customer experience, and drive revenue growth. The analysis should provide actionable recommendations tailored to the specific needs and challenges of the business.

## **Data Overview:**

**1.** **Dataset name:** Amazon Sales Report.

**2.** **Total number of rows before data cleaning:** 128976.

**3. Total number of columns before data cleaning:** 21.

**Tech Stacks:**

**1. Programming Language:** Python

**2. Libraries:**

**Numpy -** Numerical Calculations.

**Pandas -** Data Manipulations.

**Matplotlib -** Data Visualizations.

**Seaborn -** Advanced Data Visualizations.

**Scipy -** Statistical Analysis.

**3. IDE/Environment:**

**Jupyter Notebook -** For running python scripts and interactive analysis.

**Power BI -** For building interactive dashboards and visualizations.

**Dataset Description:**

**Total\_column names:**

['index', 'Order ID', 'Date', 'Status', 'Fulfilment',' Sales Channel', 'ship-service-level', 'Category', 'Size',

'Courier Status', 'Qty', 'currency', 'Amount', 'ship-city', 'ship-state', 'ship-postal-code', 'ship-country', 'B2B', 'fulfilled-by', 'New', 'PendingS']

|  |  |
| --- | --- |
| **VARIABLES** | **DESCRIPTION** |
| Index | Unique identifier for each row. |
| Order ID | Unique identifier for each order. |
| Date | Order date ranging from January to December. |
| Status | Current order status (e.g., Shipped, cancelled). |
| Fulfilment | Whether the order was fulfilled by Amazon or a Merchant. |
| Sales Channel | The platform where the sales occurred (e.g., Amazon.in). |
| Ship-Service-Level | Shipping method used (e.g., Standard, Expedited). |
| Category | Product type (e.g., T shirt, Shirt, Blazzer). |
| Size | Product size (e.g., S, XL, 3XL). |
| Courier Status | Status of the shipment (e.g., On the way, Shipped, Cancelled). |
| Qty | Quantity of products in the orders. |
| Currency | The currency in which the transaction occurred (e.g., INR). |
| Amount | Total order values. |
| Ship-City | City where the order was shipped. |
| Ship-State | State where the order was shipped. |
| Ship-Postal-code | Postal code of the shipping address. |
| Ship-Country | Country where the order was shipped. |
| B2B | Indicates if the order is a Business-to-Business transactions. |
| Fulfilled-By | Indicates who fulfilled the orders (e.g.,Easy Ship). |
| New, PendingS | Unused columns like full of null values only. |

In this dataset date datatype is in object so I change that to date datatype.

## **Data cleaning**

Clean the data to handle Missing values, Duplicates, Outliers, and inconsistencies. This step involves tasks like Imputation, Removal and etc.

**Removing a Duplicates**

**Duplicate Rows** = 165

**Duplicate columns =** 0

Remove duplicate rows using drop\_duplicates ()

**Handling Missing Values**

Identify missing values using isnull() or info()

**currency 7789**

**Amount 134**

**ship-city 33**

**ship-state 33**

**ship-postal-code 33**

**ship-country 33**

**fulfilled-by 89597**

**New 128811**

**PendingS 128811**

“New” and “PendingS” columns contain only null values and were dropped.

“Currency”, “Ship Country” and “Fulfilled-by” these three columns are zero variance. Currency is only INR, Ship Country is only IN, and Fulfilled-by is only Easy Ship, so I drop those columns also.

For Amount column, missing values were filled using median imputation.

Missing values for ship-city and ship-state were filled with “Unknown”.

Ship-postal-code missing values were filled with 0.

**Standardizing Date format**

Dates were inconsistent, with different formats (**mm/dd/yyyy** and **dd/mm/yyyy)**.

Convert all dates into Consistent format **dd/mm/yyyy**.

**Outlier Analysis**

Outliers were identified in numerical columns (Amount and Qty).

However, outliers were kept as removing them would distort the sales analysis, particularly revenue calculations.

**After cleaning Dataset contains**

Rows = 128811

Columns = 16

**Business Decisions**

|  |  |  |
| --- | --- | --- |
|  | **Amount** | **Qty** |
| Mean | 610.0214375325089 | 0.9044258642507239 |
| Median | 587.0 | 1.0 |
| Variance | 97819.88117629236 | 0.09819413721607577 |
| Range | 5584.0 | 15 |
| Standard Deviation | 312.7617003027902 | 0.31335943773257535 |
| Skewness | 0.44711202524751154 | -0.658544510965909 |
| Kurtosis | 1.917389848001982 | 60.428680941783455 |

**Visualizations**

1. Total Orders and Total Sales by Month, Year, and Quarter.

2. Sales by Category, Size, and Fulfilment Method.

3. Total Orders by Ship State and Ship City.

4. Sales Distribution by Amount, Category, Size, and Status.

5. Outlier Flag Analysis.

6. Product Preference Insights (Top Categories, Sizes).

7. Geographical Distribution of Sales.

**Conclusion**

**1. Sales Overview:**

Visualizations of total orders and total sales across different time periods (monthly, quarterly, yearly) address the objective of understanding overall sales performance and trends.

**2. Product Analysis:**

Analysis of sales by product category, size, and quantity helps identify high-demand products and popular categories, fulfilling the product analysis requirement.

**3. Fulfilment Analysis:**

The analysis of fulfilment methods and their effectiveness, including sales by fulfilment type, meets the requirement to investigate how different fulfilment methods impact sales.

**4. Customer Segmentation:**

While direct segmentation isn’t performed, your analysis of sales by geographic data (ship state, ship city) provides valuable insight into customer locations. Additional segmentation can be done if needed based on order behaviour.

**5. Geographical Analysis:**

Detailed geographical analysis (total orders by ship state, ship city, and sales distribution across locations) fully meets the geographical analysis objective.

**6. Business Insights:**

My insights, including outlier analysis and trends in product and sales data, provide actionable business recommendations, helping optimize operations and improve customer service.