**END TO END DATA ANALYSIS PROJECT REPORT**

**OBJECTIVE:** To analyze Amazon sales data and derive useful insights from it.

**SKILLS AND TECHNOLOGY USED:**

1. Programming language used: PYTHON being a simple easy to use and learn programming language which is highly portable, having extensive libraries supported and can be run on cross platform makes it a perfect language to be used for data analytics.
2. Primary steps involved:
3. Data loading and inspection
4. Cleaning and processing the data
5. Exploratory data analysis
6. Visualization for insights
7. Data visualization with power BI: creating an interactive dashboard with power BI to have a single screen format which enables users to understand the data insights visually.

**DATA INFORMATION:**

The data used for the implementation of this project is taken from a widely used website: Kaggle.com.

The amazon sales report data has several columns giving necessary information which can be used to analyze the sales.

**OVERVIEW OF DATA ANALYTICS DONE:**

The project focuses on the descriptive analysis of data:

Descriptive analysis: this type of analysis answers the question – **“what happened”**. It mainly summarizes past performance and sales trend from the given data.

**DATA ANALYSIS BY USING PYTHON**

1. Data loading and inspection:

* Data loaded into dataframe
* First few rows are shown which helps us to get a overview of what actually the dataset contains.
* A basic information about the column names and data types is known
* A summary statistic for numerical columns is printed
* Missing values/ null values are checked

1. Data cleaning and processing

 Clean the dataset by handling missing values.

 Convert the date format for easier filtering.

 Remove duplicates based on a unique identifier.

 Clean up and standardize text-based fields.

 Prepare columns for further analysis or model training.

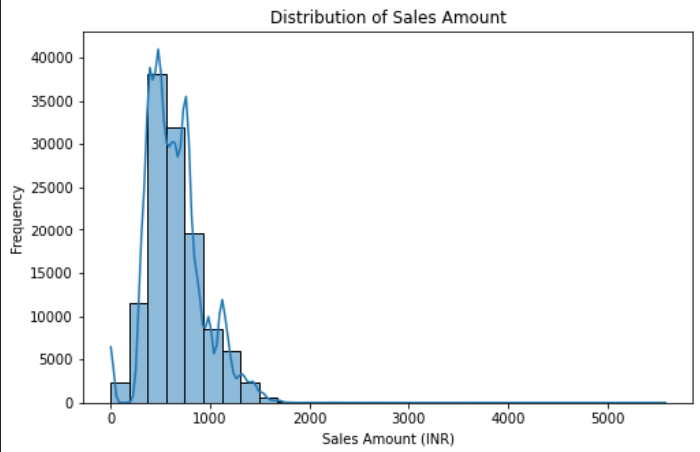
1. Exploratory data analysis:

**Descriptive analysis:**

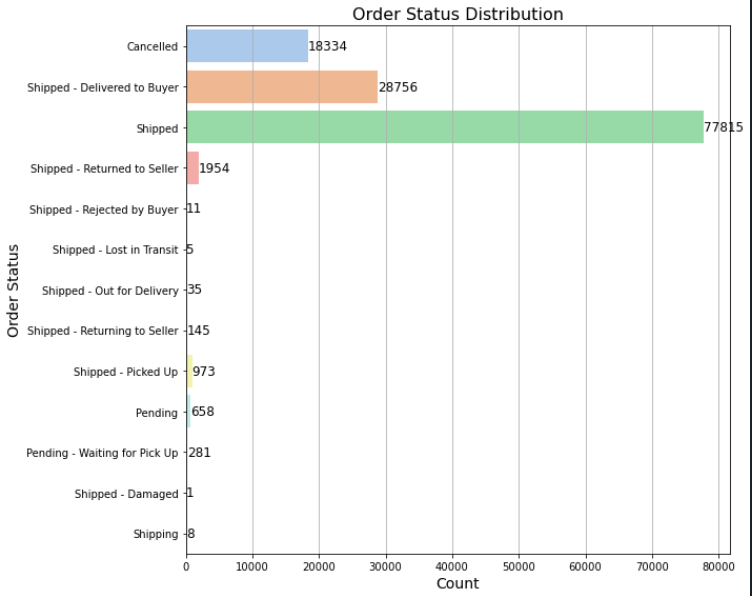
1. **Data summary**

* Summary statistics: which consists of mean, standard deviation, max min etc
* Value counts for status of delivery: counts the number of each type of delivery made by the company. It helps to study the delivery condition and performance.
* Value counts for category: depending upon the data, several category of items are listed. This value count helps to analyse how many items of each category were traded.

1. **Distribution analysis:**
2. Visualize the amount by frequency distribution using the pyplot library.

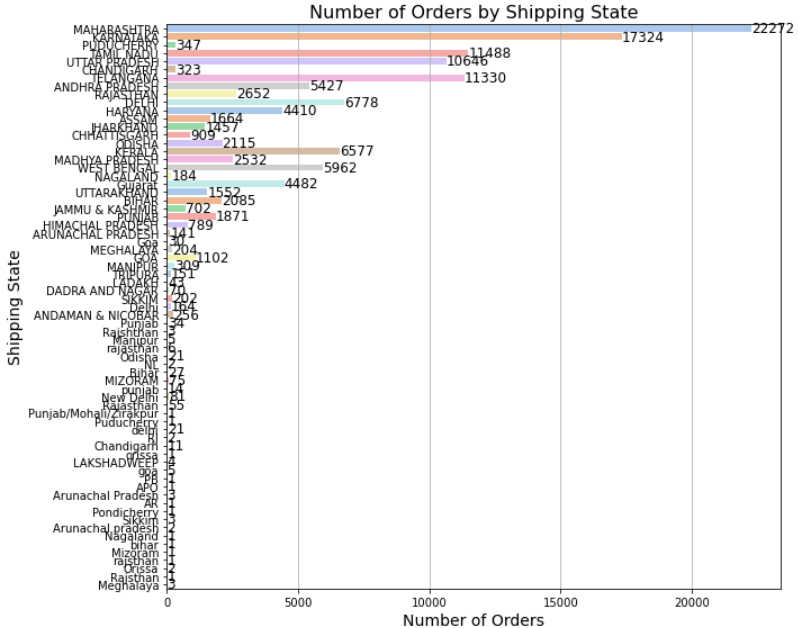


1. Frequency distribution: visually study the order status distribution to analyse the order status condition.

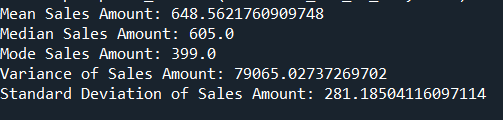


1. City wise sales: comparing which city has most sales

Interpretation: Maharashtra has most number of orders followed by Karnataka and Uttar Pradesh.



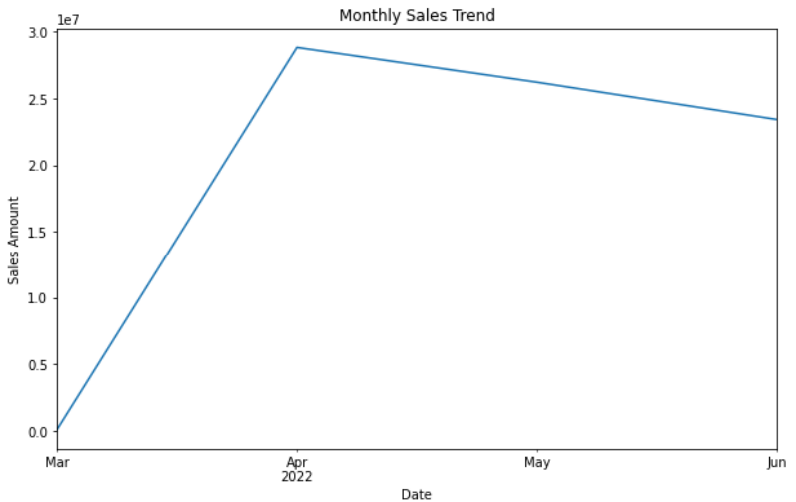
1. Central tendencies and dispersion: calculating mean, mode, median, standard deviation of numerical data (sales amount here)

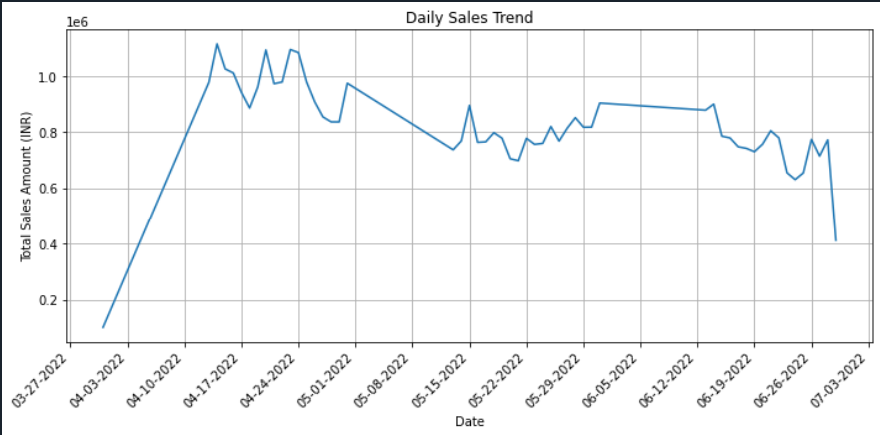


* Mean: the average value of amount, indicating central tendencies.
* Mode: the most frequent occurring amount. Most common transaction value
* Median: middle value when all amounts are sorted in ascending order. Helps to understand the distribution and potential skewness of data
* Variance: the measure of how much the amount vary from mean.

1. **Time Series Analysis:** time series analysis allows us to analyze trends over time to see how certain metrices change over time.

* Month-wise sales:





* Interpretation:
* The maximum sales is seen between 10th to 14th April 2022.
* There has been a noticeable decline in sales between 26th june to 3rd July 2022.
* A almost constants sales has been observed between 29th may to 12th June 2022.

1. **Aggregation analysis:**

Aggregation involves summarizing data on different levels like daily, monthly or yearly or by categorial groups to uncover patterns, trends and insights. Helps to understand data and answer business questions.

1. Average Order Value (AOV): AOV: 653.6706638997244
2. Top Performing Categories:

Category sum mean count

T-shirt 39206756.65 779.582372 50292

Shirt 21297770.08 427.005836 49877

Blazzer 11215104.12 723.555105 15500

Trousers 5346286.30 503.606471 10616

Perfume 789419.66 682.298755 1157

**USEFUL INSIGHT FROM DISCRIPTIVE ANALYSIS:**

1. A significant proportion of sales is attributed to items priced below ₹1000.
2. An important observation is that there are significant number of canceled orders, which needs further investigation into potential causes like logistical issues or customer dissatisfaction.
3. Maharashtra, Karnataka, and Tamil Nadu are the leading states in terms of sales performance. ALSO, states like Uttar Pradesh, Delhi, Haryana, and West Bengal are emerging markets that have a potential to contribute significantly to overall sales.
4. April was the as the most successful month, having peak sales. However, we notice a gradual decline in sales from June to July.
5. The top-performing product categories include T-shirts, shirts, blazers, trousers, and perfumes

**Understanding the observations:**

1. High sales of low cost items: this indicates that affordability is the key driver for the customers.
2. Cancelled items: high number of cancelled items indicate the overall customer dissatisfaction from the goods or mismatched expectations
3. Regional insights: the states with higher sales record must continue with their strategies, emerging areas should be look for some improvement which will subsequently increase their overall performance and share in the sales. For the states with lower sales record should incorporate new market strategies for improving performance
4. The sales methodologies, marketing and advertising strategies of June and July month should be compared with those of may to see how effective those strategies are .
5. The company should ensure that highest selling items should be carefully continued to be sold without any unnecessary interruptions.

**Time Series Forecasting**

To move beyond simple trend visualization, forecasting was implemented using the **Prophet model**.

1. Data was resampled on a daily basis to capture fine-grained patterns.
2. Prophet was trained to forecast the next 30 days of sales.
3. The forecast plot showed the expected future trend with upper and lower confidence bounds, allowing the business to anticipate sales fluctuations.
4. This approach enables proactive inventory planning and marketing alignment.

**Key Observation:**  
The forecast indicated continuity of the general trend with small seasonal variations. Peaks and dips could be anticipated in advance, helping the sales team to prepare accordingly.

**Anomaly Detection**

An anomaly detection method was implemented using **z-scores**:

* Each sales amount was compared with the dataset mean and standard deviation.
* Transactions with |Z-Score| > 3 were flagged as anomalies.

**Findings:**

* A small number of unusually high sales orders were detected, which may indicate **bulk purchases or possible data errors**.

**Rolling Averages & Lag Features**

To smooth fluctuations and identify consistent sales patterns, rolling averages and lag features were introduced:

* A **7-day rolling average** was calculated to reduce noise and highlight long-term sales patterns.
* Lag features (e.g., previous day’s sales) were created to identify dependencies over time.

**Insights:**

* Rolling averages revealed more stable underlying trends that were hidden in daily volatility.
* These engineered features can be powerful inputs for predictive models, enabling better forecasting accuracy.

**Overall Extended Insights**

* Forecasting allows the company to **anticipate demand** and optimize supply chain operations.
* Anomaly detection helps in **monitoring unusual customer activity or data quality issues**.
* Rolling averages and lag features provide **deeper temporal understanding**, essential for predictive modeling.