Session ID : ASSI2079081020

Subject/Topic name is BUSINESS INSIGHT THROUGH DATA VISUALIZATION

Section 1: Dataset selection

- **Process of Dataset Selection:**
    1. **Visit Kaggle:**
        - Navigated to Kaggle's website and logged in.
    2. **Explore Datasets:**
        - Used search terms like "Amazon Seller" and "Order Status Prediction" to find relevant datasets.
    3. **Filter and Evaluate:**
        - Filtered results based on relevance, popularity, and recent updates.
        - Evaluated datasets considering descriptions, sizes, and download counts.
    4. **Check Dataset Details:**
        - Clicked on the "Amazon Seller - Order Status Prediction" dataset to view details.
        - Examined the dataset description, column details, and any accompanying documentation.
    5. **Read Reviews and Discussions:**
        - Explored user reviews and discussions to understand potential challenges or insights shared by others.
    6. **Verify Data Quality:**
        - Ensured the dataset met quality standards, checking for completeness, absence of significant missing values, and clarity in variable definitions.
    7. **Download the Dataset:**

    - Downloaded the "Amazon Seller - Order Status Prediction" dataset to the local machine.

- **Criteria for Choosing the Dataset:**

  The criteria for choosing this dataset included:

  - **Relevance:** The dataset is directly related to Amazon sellers and order status prediction, aligning with the objective of deriving business insights.
  - **Completeness:** The dataset appeared to have comprehensive information needed for analysis, including order details and status.
  - **Quality:** Initial examination indicated good data quality, with no apparent issues like excessive missing values or unclear variable definitions.
  - **Popularity:** The dataset's popularity, as indicated by download counts and positive reviews, suggested it was well-received by the Kaggle community.

- **Brief Overview of the Selected Dataset:**

  - The "Amazon Seller - Order Status Prediction" dataset comprises columns such as `order_no`, `order_date`, `buyer`, `ship_city`, `ship_state`, `sku`, `description`, `quantity`, `item_total`, `shipping_fee`, `cod`, and `order_status`.
  - It provides detailed information about individual orders, including customer details, shipping information, product details, and order status.
  - This dataset will be instrumental in predicting and understanding order statuses in the context of Amazon sellers.

Section 2: Business Context and Background

- **Business Context:**

  The dataset is relevant to the e-commerce sector, specifically Amazon sellers. Understanding and predicting order statuses are critical for managing logistics, inventory, and customer expectations. Accurate predictions can optimize operations, reduce delivery times, and enhance customer satisfaction.

- **Background Information:**

Amazon, being a global e-commerce giant, operates in a highly competitive environment. Efficient order management is crucial for sustaining customer loyalty and outperforming competitors. Predicting order statuses helps sellers streamline their supply chain, reduce costs, and improve overall service quality.

Section 3: Data Cleaning and Transformation

1. **Handling Missing Data:**

   - There are 3 features with missing values: `item_total` , `shipping_fee` and `cod`
   - Imputing with **mode** as `shipping_fee` is fixed based on package size and weight and we are considering the item that is sold the most

   ```
   df['shipping_fee'].fillna(df['shipping_fee'].mode()[0], inpla
   df['item_total'].fillna(df['item_total'].mode()[0], inplace=
   df['cod'].fillna('online', inplace=True)
   ```

2. **Remove rupee symbol from amount feature**

   ```
   amounts = ['item_total', 'shipping_fee']
   for i in amounts:
       df[i] = df[i].apply(lambda x: x.replace(',', ''))
       df[i] = df[i].apply(lambda x: x[1:])
   ```

   ```
   # change data types
   i = 'int64'
   f = 'float64'
   df = df.astype({'item_total': f, 'shipping_fee': f, 'quantit
   ```

3. **To drop the features from a dataframe, convert city and states into uppercase and removing comma from city values**

```
defdrop(df, *features):
    for i in features:
        df.drop(i, axis=1, inplace=True)
```

```
places = ['ship_city', 'ship_state']
for i in places:
        df[i] =df[i].apply(lambdax:x.upper())
```

```
df['ship_city'] = df['ship_city'].apply(lambdax:x.replace(',',
df['ship_state'] = df['ship_state'].apply(lambdax:x.replace(','
```

Section 4: Problem Statement or Research Question

We will try to answer few questions:

- Does order success depend on the mode of payment?
- Sales Trend (Sales Pattern over the period)
- Which days of the week draw more sales?
- People from which states are ordering the most
- Which are the top sold products?

Section 5: Data Analysis

- Techniques:
    1. **Mode of payment:** Grouped data by `cod` and `order_status`, then visualized using a countplot.
    2. **Sales Trend Analysis:** Extracted year, month, day, and time from the `order_date` feature. Extracted a unique identifier from the `sku` column. Grouped sales data by date, calculating the total sales for each day. Created a line plot using Plotly Express to visualize the sales trend over time.

3. **Day-wise Sales Analysis:** Grouped sales data by year, month, and day of the week, calculating the total sales for each combination. Created a bar plot using Plotly Express to visualize month-wise sales by day of the week.
4. **Top Ordering States:** Grouped sales data by `ship_city`, `ship_state`, calculating the total sales for each combination. Sorted the DataFrame by total sales in descending order. Extracted the top states with the highest total sales and visualized the results using a bar plot.
5. **Top Sold Products:** Grouped the data by `sku`, calculating the total number of orders for each product. Calculated the percentage of total orders for each product. Visualized the top 5 sold products as a bar chart.

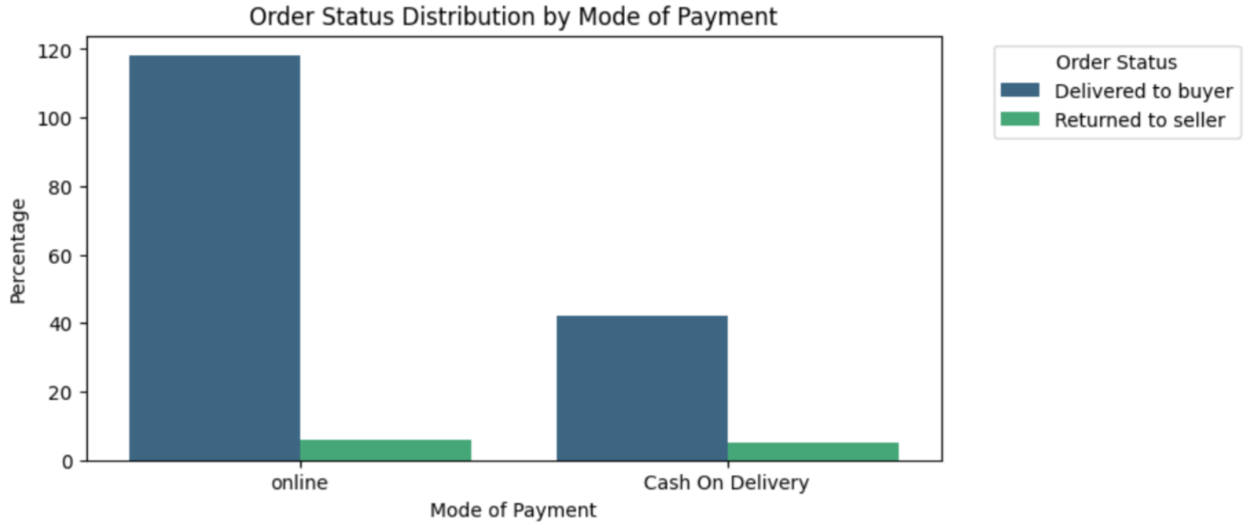
- Results:
  - We see that most of the orders were prepaid and the percentage of return orders is more in cod mode.
  - No significant growth observed over the period.
  - December exhibited the highest sales, marked by a couple of breakthroughs.
  - We visualized the sales across the months grouped by day of week to find that Sundays and Wednesdays emerged as the highest contributors to sales. Saturdays were identified as the least impactful day for sales.
  - The bar plot shows the total sales amounts for the top ordering states.
  - The bar chart displays the percentage of total orders for the top 5 sold products.

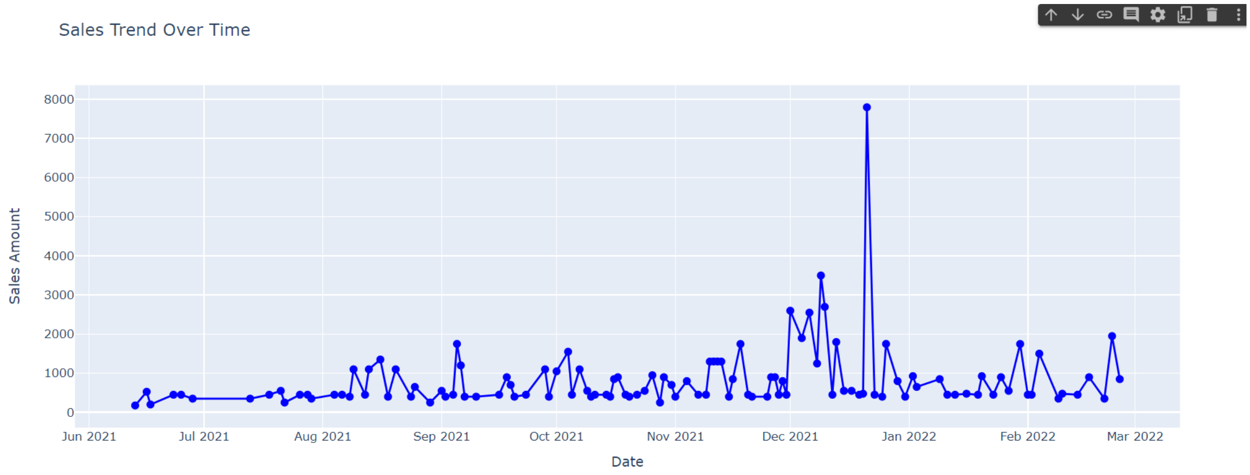
Section 6: Business Insights

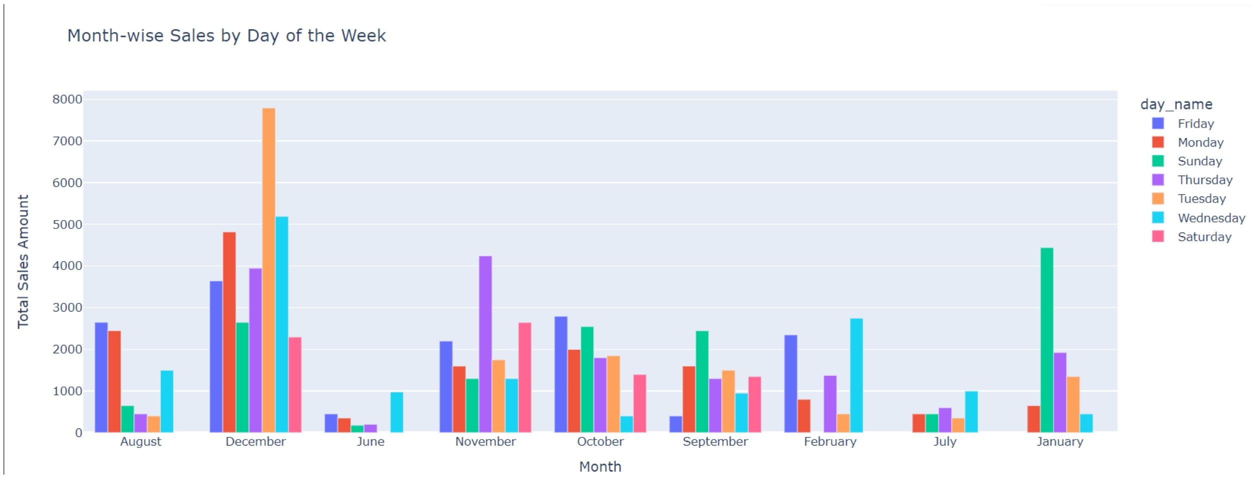
- Insights:
  - **Mode of payment:** Prepaid dominance; COD mode associated with higher return rates.
  - **Sales Trend Analysis:** Identifying periods of peak sales (e.g., December) can aid in planning marketing strategies and inventory management.
  - **Dominant Sales Days:** Sundays and Wednesdays emerged as the highest contributors to sales.

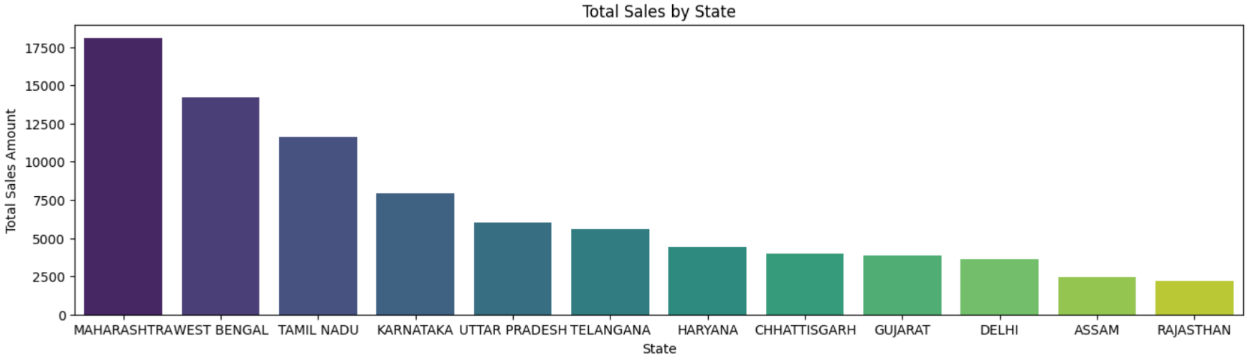
- **Least Contributor:** Saturdays were identified as the least impactful day for sales.
- **Top Ordering States:** Maharashtra, West Bengal, Tamil Nadu, Karnataka, Uttar Pradesh, and Telangana emerged as the leading states in terms of order amounts.
- **Top Sold Products:** DN-0WDX-VYOT, 0M-RFE6-443C, SB-WDQN-SDN9, CR-6E69-UXFW, and 2X-3C0F-KNJE are the top 5 products.

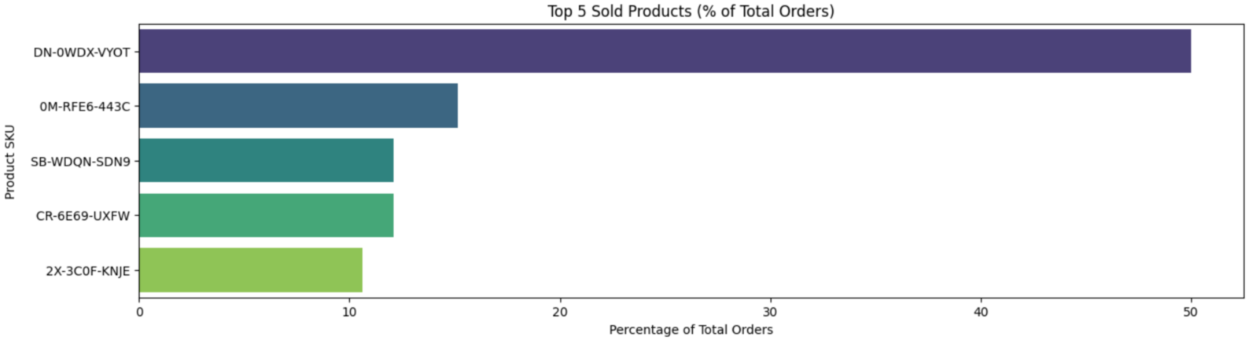
Section 7: Dashboard of Appropriate Visualization











Section 8: Recommendations

- **Promote Prepaid Options:**
  - **Benefit:** Streamline payment process, reduce cancellations, and incentivize efficient order fulfillment.

- **Investigate COD Issues:**
    - **Benefit:** Identify and address Cash on Delivery (COD) challenges, enhancing customer trust and loyalty.
- **Improve COD Experience:**
    - **Benefit:** Provide clearer communication, accurate tracking, and efficient payment processing, leading to increased customer satisfaction and loyalty.
- **Promotional Strategies:**
    - **Benefit:** Understanding sales trends helps in optimizing inventory, marketing efforts, and overall business strategies. Capitalizing on peak sales months can result in increased revenue and customer engagement.
- **Maximized Sales Potential:**
    - **Benefit:** By strategically targeting high-sales days and adjusting strategies for lower-performing days, the business can optimize its overall sales performance, leading to increased revenue and customer engagement.
- **Regional Targeting:**
    - **Benefit:** Targeting regions with higher order amounts allows for more efficient allocation of marketing resources, leading to increased sales and customer engagement in key geographic areas.
- **Inventory Focus:**
    - **Benefit:** By focusing on the top-selling products, the business can streamline inventory management, reduce stockouts, and ensure customer demand is consistently met, resulting in increased customer satisfaction and potentially higher revenue.

Section 9: Conclusion and Report

- The dataset came with some missing values in 2 of its features
- Order return rates are higher in COD type orders
- The sales has been stagnant throughout with few breakthroughs in December

- We visualized the sales across the months grouped by day of week to find that Sundays and Wednesdays dominate over the sales whereas the least contributor turned out to be the Saturdays
- We saw the region wise order amounts
- Top ordering cities are: Mumbai, Kolkata, Bangalore, Chennai and Pune

Appendix:

- Here is the link of google colab:
    - **google colab**

References:

- Reference