AJEENKYA
D Y PATIL UNIVERSITY | THE INNOVATION UNIVERSITY

Artificial Intelligence And Machine Learning

ASSIGNMENT

MINI PROJECT REPORT ON

Analysing Amazon Sales Trends with Python: A Machine Learning Approach

UNDER THE GUIDANCE OF

PROF.VIVEK MORE

Name: Ruchika Shevale

URN NO: 2023-B-26012005A

Year: 2nd Year

Divison: B

Semester: 4th

Course: Artificial Intelligence and Machine Learning

Course code: BD201E

Faculty: Prof. Vivek More

College: Ajeenkya D Y Patil University

Submission Date: 14th April 2025

TOPIC: Analysing Amazon Sales Trends with Python: A Machine Learning Approach

This is to certify that ________, a student of BCA(AIML) Sem-4 URN No. _____________, has successfully completed the Dashboard Report on:

“Analysing Amazon Sales Trends with Python”

As per the requirement of Ajeenkya DY Patil University, Pune was carried out under my supervision. I hereby certify that; he has satisfactorily completed his/herTerm-Work Project work.

Place: - Pune

Index page

Introduction

In the modern era of digitalization, online stores such as Amazon have revolutionized the manner in which customers purchase and companies’ function. With an astronomical number of transactions occurring every day, data becomes the key driver of strategic decisions. The project involves examining a dataset that includes Amazon sales data for the year 2025. The dataset has 250 records and features order dates, product names, categories, quantities, prices, total sales, customer data, payment modes, and order status.

The objective of the project is to extract valuable information from this data with Python and the required libraries. By using data visualization and analysis skills, we are able to know how various products are performing, determine customer demand by location, and learn sales patterns. The information can aid stakeholders in making informed decisions in marketing, stock management, and targeting customers.

Objective:

The main aim of this project is to study Amazon's 2025 sales figures to identify concealed patterns and business trends. This entails:

- Determining best-selling products and categories generating most of the revenue.
- Examining geographic location of customers to identify major market areas.
- Measuring effectiveness of payment types and customer choices.
- Understanding order statuses to determine cancellations or pending orders rates and their possible reasons.
- Investigating trends in sales over time, which would be possible to utilize in demand planning or seasonal forecasting.

At the completion of this project, we should have constructed an exhaustive understanding of the sales data that can be used as a starting point for data-driven decision-making.

Methodology:

Dataset Understanding—

The dataset used in this project contains sales transaction data from Amazon for the year 2025. It includes attributes such as Order Date, Region, Product Category, Sub-Category, Sales, Quantity, Discount, Profit, and Shipping Cost. The dataset was initially loaded into a Pandas DataFrame to begin the analysis. A preliminary inspection was done to understand the structure of the data, including column names, data types, and basic statistics.

STEP—1

Data Cleaning and Preprocessing:

To get the dataset ready for analysis, some data cleaning and preprocessing operations were used. Missing values were handled by replacing non-critical columns like Shipping Cost with the mean, and rows lacking crucial data like Sales or Order Date were deleted to maintain integrity. Duplicate records were identified through the `duplicated () ` function and removed to prevent biased results. The `Order Date` column was converted to a correct datetime column for time-based analysis, and rows containing invalid dates were dropped. Feature engineering was also carried out—new columns for `Year`, `Month`, and `Day` were derived to view the seasonal pattern, and a `Profit Margin` column was created by dividing Sales by Profit. Infinite or undefined values in this new column were substituted with zeros to ensure consistency and prevent errors during analysis.

Implementing code:

1] Cleaning Data:

- Dropped Rows with Missing Key Data

We eliminated all the rows that contained missing values in key columns such as Price, Quantity, Total Sales, and Date. These are key fields to ensure accurate analysis, and having incomplete records here may produce incorrect results.

- Replaced Missing Non-Critical Fields with "Unknown"

For columns like Customer Name, Customer Location, Payment Method, and Status, we filled missing values with the placeholder "Unknown." This method maintains the records in their original form while distinctly indicating the lack of data.

- Removed Duplicate Records

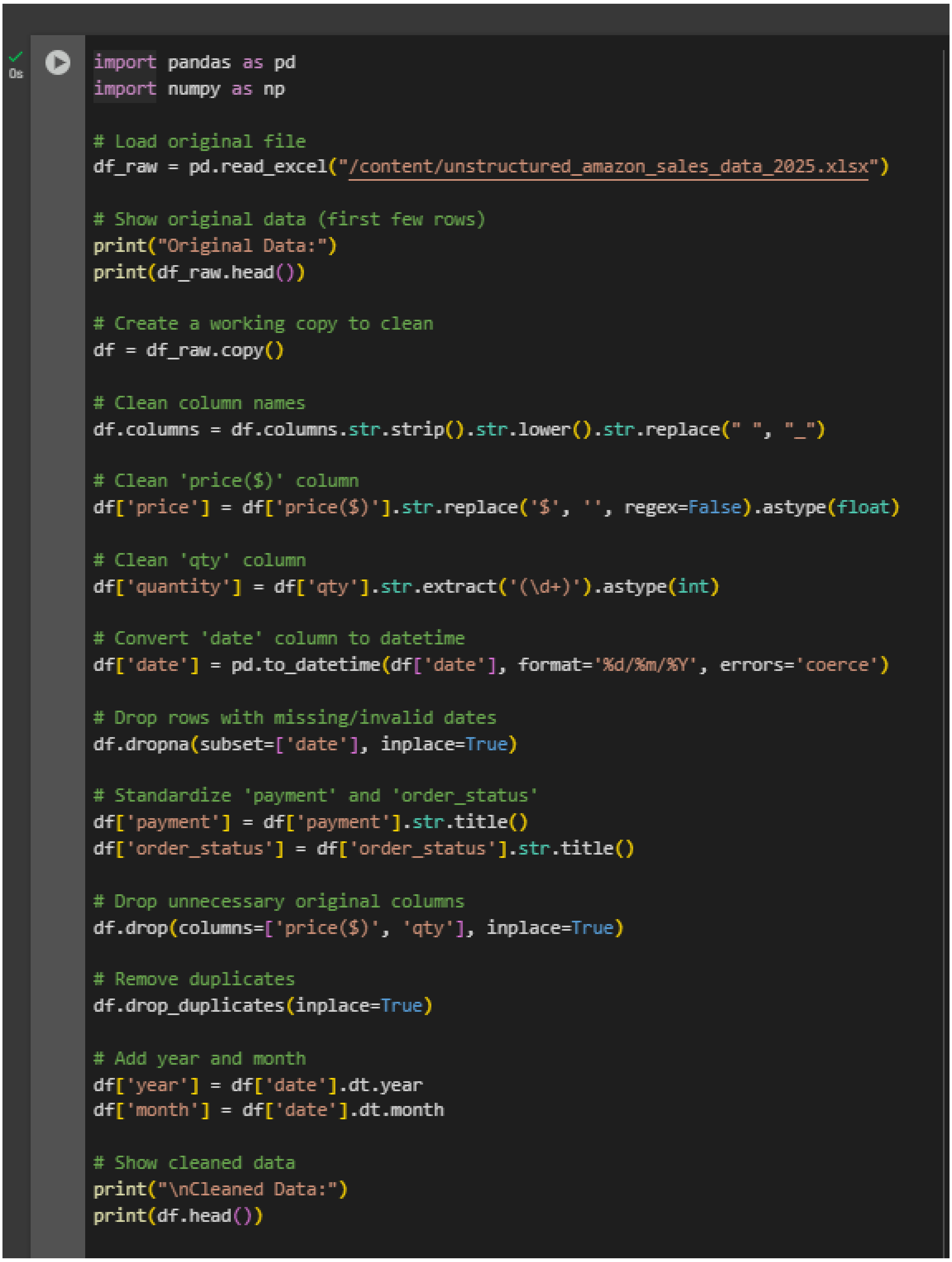
We removed all duplicate rows from the dataset to avoid counting each transaction more than once. This avoids inflated numbers or duplicated analysis of the same data.

- Handled Invalid or Corrupted Date Entries

The Date column was changed to the correct datetime format. Rows that contained invalid or unrecognized dates were deleted to prevent problems in time-based reporting or trend analysis.

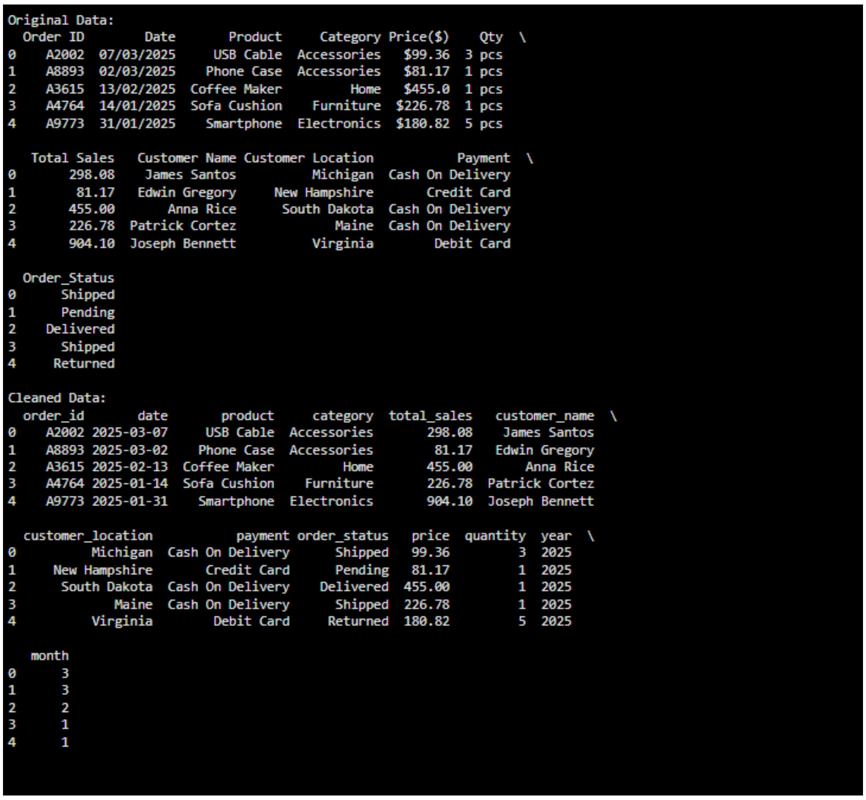
- Cleaned Dataset Structure and Validity

We utilized functions such as info () and isnull(). sum() before and after cleaning to check for the structure of the dataset so that the steps of cleaning have been successful and the data are now ready to be analyzed.

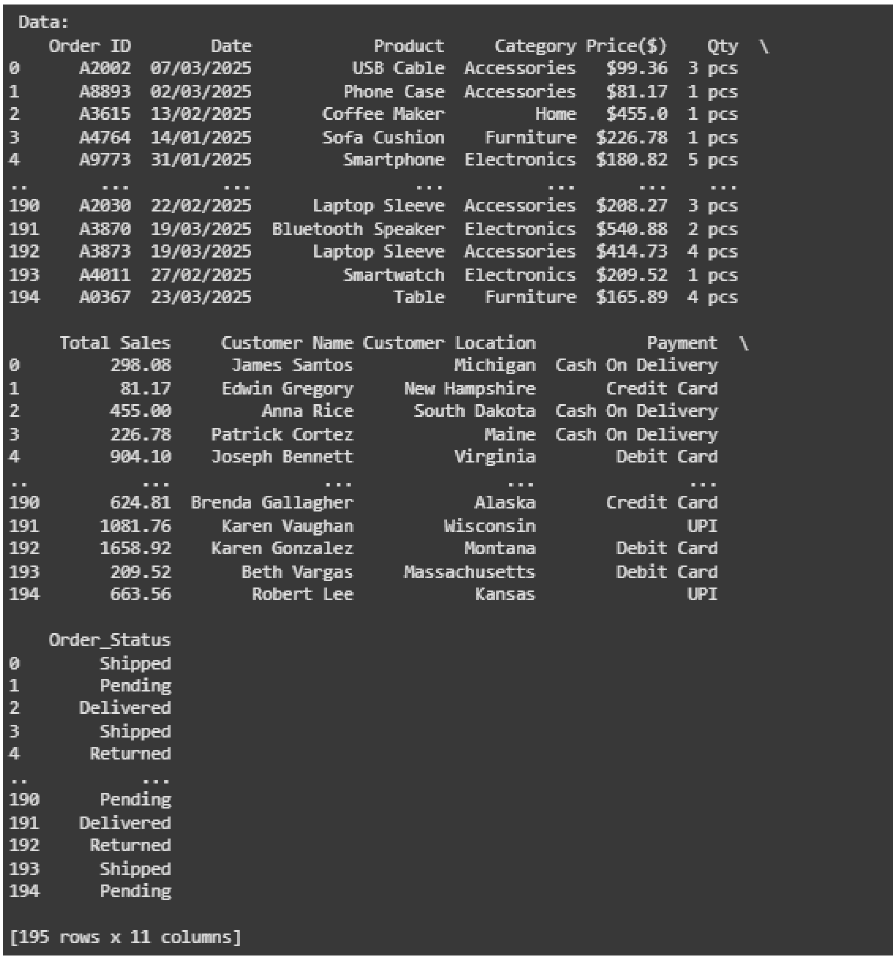


```python
import pandas as pd
import numpy as np

# Load original file
df_raw = pd.read_excel("/content/unstructured_amazon_sales_data_2025.xlsx")

# Show original data (first few rows)
print("Original Data:")
print(df_raw.head())

# Create a working copy to clean
df = df_raw.copy()

# Clean column names
df.columns = df.columns.str.strip().str.lower().str.replace(" ", "_")

# Clean 'price($)' column
df['price'] = df['price($)'].str.replace('$', '', regex=False).astype(float)

# Clean 'qty' column
df['quantity'] = df['qty'].str.extract('(\d+)').astype(int)

# Convert 'date' column to datetime
df['date'] = pd.to_datetime(df['date'], format='%d/%m/%Y', errors='coerce')

# Drop rows with missing/invalid dates
df.dropna(subset=['date'], inplace=True)

# Standardize 'payment' and 'order_status'
df['payment'] = df['payment'].str.title()
df['order_status'] = df['order_status'].str.title()

# Drop unnecessary original columns
df.drop(columns=['price($)', 'qty'], inplace=True)

# Remove duplicates
df.drop_duplicates(inplace=True)

# Add year and month
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month

# Show cleaned data
print("\nCleaned Data:")
print(df.head())
```

Output:



```
Original Data:
  Order ID        Date       Product     Category Price($)    Qty  \
0    A2002  07/03/2025     USB Cable  Accessories   $99.36  3 pcs
1    A8893  02/03/2025    Phone Case  Accessories   $81.17  1 pcs
2    A3615  13/02/2025  Coffee Maker         Home   $455.0  1 pcs
3    A4764  14/01/2025  Sofa Cushion    Furniture  $226.78  1 pcs
4    A9773  31/01/2025    Smartphone  Electronics  $180.82  5 pcs

   Total Sales   Customer Name Customer Location          Payment  \
0       298.08    James Santos          Michigan  Cash On Delivery
1        81.17   Edwin Gregory     New Hampshire       Credit Card
2       455.00       Anna Rice      South Dakota  Cash On Delivery
3       226.78  Patrick Cortez             Maine  Cash On Delivery
4       904.10  Joseph Bennett          Virginia        Debit Card

  Order_Status
0      Shipped
1      Pending
2    Delivered
3      Shipped
4     Returned

Cleaned Data:
  order_id       date       product     category  total_sales   customer_name  \
0    A2002 2025-03-07     USB Cable  Accessories       298.08    James Santos
1    A8893 2025-03-02    Phone Case  Accessories        81.17   Edwin Gregory
2    A3615 2025-02-13  Coffee Maker         Home       455.00       Anna Rice
3    A4764 2025-01-14  Sofa Cushion    Furniture       226.78  Patrick Cortez
4    A9773 2025-01-31    Smartphone  Electronics       904.10  Joseph Bennett

  customer_location           payment order_status   price  quantity  year  \
0          Michigan  Cash On Delivery      Shipped   99.36         3  2025
1     New Hampshire       Credit Card      Pending   81.17         1  2025
2      South Dakota  Cash On Delivery    Delivered  455.00         1  2025
3             Maine  Cash On Delivery      Shipped  226.78         1  2025
4          Virginia        Debit Card     Returned  180.82         5  2025

   month
0      3
1      3
2      2
3      1
4      1
```



```
 Data:
    Order ID        Date           Product     Category Price($)    Qty  \
0      A2002  07/03/2025         USB Cable  Accessories   $99.36  3 pcs
1      A8893  02/03/2025        Phone Case  Accessories   $81.17  1 pcs
2      A3615  13/02/2025      Coffee Maker         Home   $455.0  1 pcs
3      A4764  14/01/2025      Sofa Cushion    Furniture  $226.78  1 pcs
4      A9773  31/01/2025        Smartphone  Electronics  $180.82  5 pcs
..       ...         ...               ...          ...      ...    ...
190    A2030  22/02/2025      Laptop Sleeve  Accessories  $208.27  3 pcs
191    A3870  19/03/2025  Bluetooth Speaker  Electronics  $540.88  2 pcs
192    A3873  19/03/2025      Laptop Sleeve  Accessories  $414.73  4 pcs
193    A4011  27/02/2025        Smartwatch  Electronics  $209.52  1 pcs
194    A0367  23/03/2025             Table    Furniture  $165.89  4 pcs

     Total Sales     Customer Name Customer Location          Payment  \
0         298.08      James Santos          Michigan  Cash On Delivery
1          81.17     Edwin Gregory     New Hampshire       Credit Card
2         455.00         Anna Rice      South Dakota  Cash On Delivery
3         226.78    Patrick Cortez             Maine  Cash On Delivery
4         904.10    Joseph Bennett          Virginia        Debit Card
..           ...               ...               ...              ...
190       624.81  Brenda Gallagher            Alaska       Credit Card
191      1081.76     Karen Vaughan         Wisconsin               UPI
192      1658.92    Karen Gonzalez           Montana        Debit Card
193       209.52       Beth Vargas     Massachusetts        Debit Card
194       663.56        Robert Lee            Kansas               UPI

    Order_Status
0        Shipped
1        Pending
2      Delivered
3        Shipped
4       Returned
..           ...
190      Pending
191    Delivered
192     Returned
193      Shipped
194      Pending

[195 rows x 11 columns]
```

Step—2 Exploratory Data Analysis (EDA) Report: Amazon Sales Data

The category breakdown of sales indicated that the Home category dominated overall revenue, with more than $61,000 in total sales, followed by Accessories and Furniture, each with more than $51,000. Electronics and Stationery comprised the rest of the top five, each with consistent but relatively lower performance. This indicates that there is high consumer demand for home and personal use items, which tend to be higher-value items. The Home category dominance might also be the result of lifestyle changes like spending more time at home or cyclical home remodelling seasons.

One review of the distribution of order status in a pie chart revealed nearly equal distribution between Pending (27%), Shipped (26%), Delivered (24%), and Returned (24%) orders. The fairly high rate of return orders suggests problems in product satisfaction, delivery expectations, or post-purchase experience. This is an opportunity to examine product quality, packaging, or customer support procedures to lower return rates and enhance satisfaction.

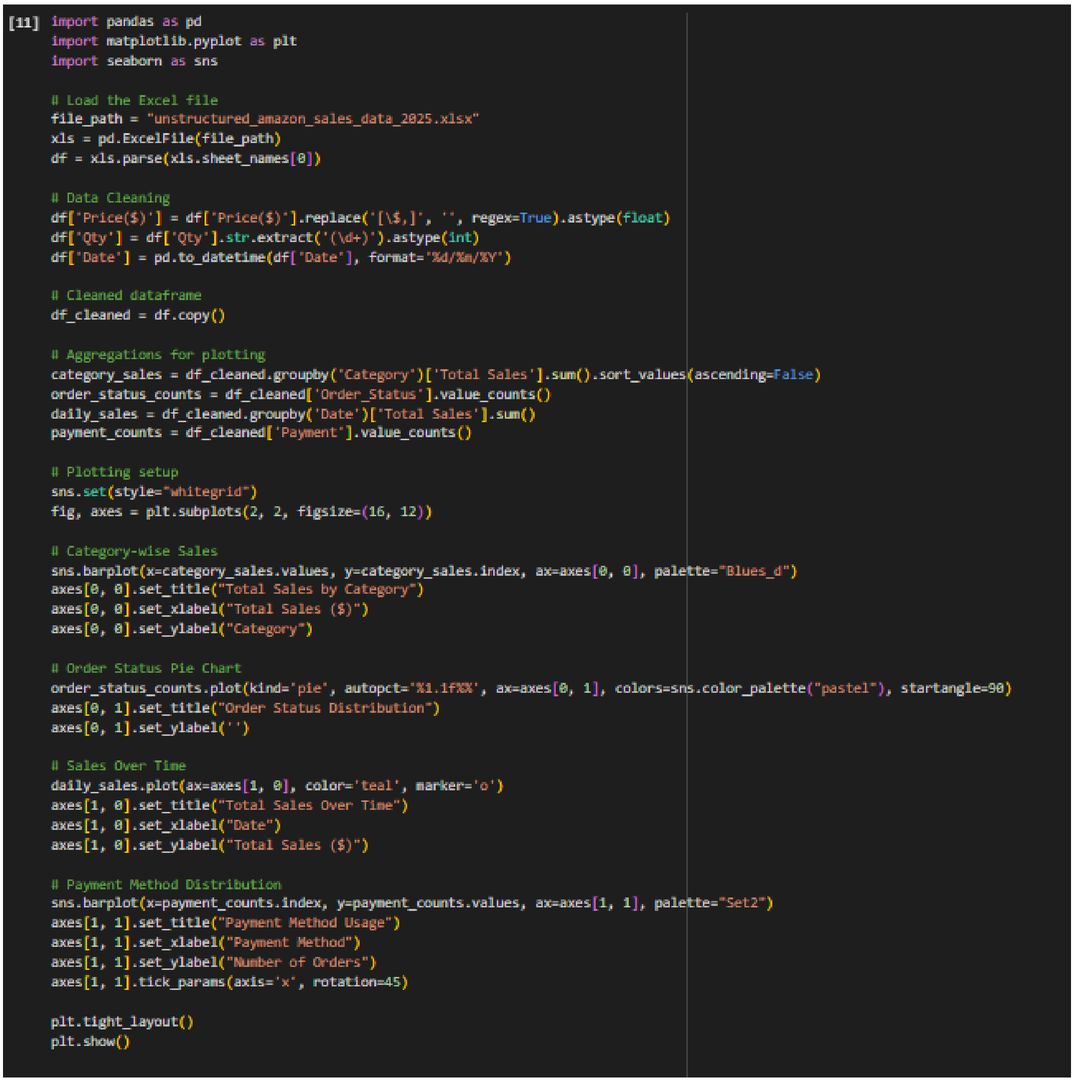
The trend of sales over time, presented using a time-series line chart, showed some spikes in the daily revenue, most probably representing special promotional events or weekends. Generally, the sales were steady during January to March 2025, with a slight increase towards the end of February. Identification of these peak demand periods is crucial in deciding inventory planning, marketing initiatives, and order fulfilment strategies.

The payment method analysis revealed that the most favoured customer choice is Cash on Delivery (COD), and then Credit and Debit Cards. Other types like UPI and Net Banking were used lesser but significantly. This suggests maintaining the prominence of providing COD as a convenient and reliable payment means, particularly for markets where acceptance of digital payment can be diversified.

Summary statistics showed a mean product price of around $463.12, a mean number of 2.81 items per order, and a mean total sale value per order of $1,290.14. The Coffee Maker was the most often ordered product, and the state of Maryland led the list as the most often visited customer location. These observations indicate that although customers are making fairly high-value transactions, they are purchasing in small to medium volumes, which is in line with the nature of products in leading categories.

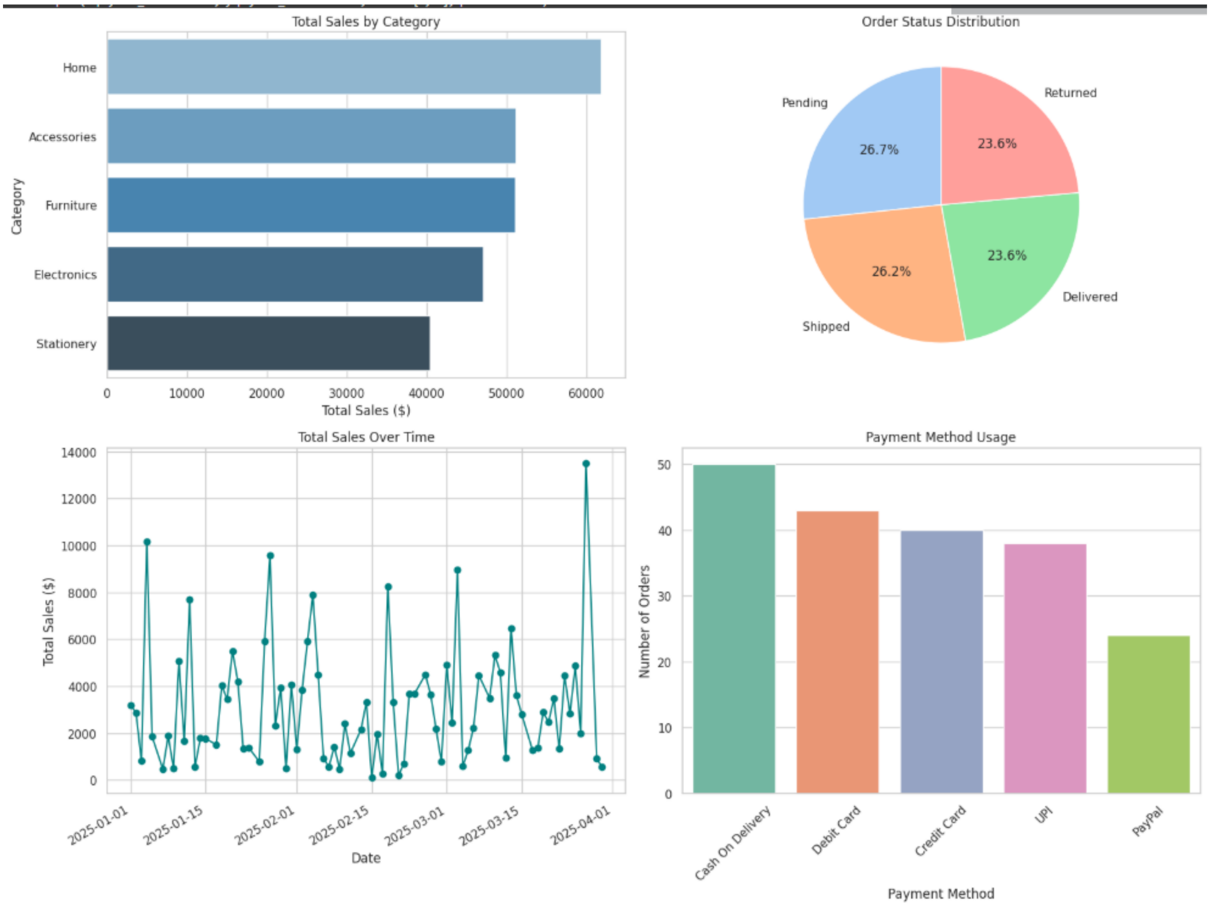
In summary, this EDA gives obvious and actionable conclusions regarding sales behavior in the first quarter of 2025. The robust performance of Home and Accessory categories indicates a scope for targeted promotion and inventory control. The even split of order statuses, specifically the high returns, necessitates further examination of post-purchase issues. COD dominance reiterates the necessity of ongoing support for varied payment methods. The determination of sales peaks over time can assist in scheduling marketing events and supply chain optimization. Further analysis in the future can include customer segmentation, regional demand mapping, and product-specific profitability to further support strategic decision-making.

code:



```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the Excel file
file_path = "unstructured_amazon_sales_data_2025.xlsx"
xls = pd.ExcelFile(file_path)
df = xls.parse(xls.sheet_names[0])

# Data Cleaning
df['Price($)'] = df['Price($)'].replace('[\$,]', '', regex=True).astype(float)
df['Qty'] = df['Qty'].str.extract('(\d+)').astype(int)
df['Date'] = pd.to_datetime(df['Date'], format='%d/%m/%Y')

# Cleaned dataframe
df_cleaned = df.copy()

# Aggregations for plotting
category_sales = df_cleaned.groupby('Category')['Total Sales'].sum().sort_values(ascending=False)
order_status_counts = df_cleaned['Order_Status'].value_counts()
daily_sales = df_cleaned.groupby('Date')['Total Sales'].sum()
payment_counts = df_cleaned['Payment'].value_counts()

# Plotting setup
sns.set(style="whitegrid")
fig, axes = plt.subplots(2, 2, figsize=(16, 12))

# Category-wise Sales
sns.barplot(x=category_sales.values, y=category_sales.index, ax=axes[0, 0], palette="Blues_d")
axes[0, 0].set_title("Total Sales by Category")
axes[0, 0].set_xlabel("Total Sales ($)")
axes[0, 0].set_ylabel("Category")

# Order Status Pie Chart
order_status_counts.plot(kind='pie', autopct='%1.1f%%', ax=axes[0, 1], colors=sns.color_palette("pastel"), startangle=90)
axes[0, 1].set_title("Order Status Distribution")
axes[0, 1].set_ylabel('')

# Sales Over Time
daily_sales.plot(ax=axes[1, 0], color='teal', marker='o')
axes[1, 0].set_title("Total Sales Over Time")
axes[1, 0].set_xlabel("Date")
axes[1, 0].set_ylabel("Total Sales ($)")

# Payment Method Distribution
sns.barplot(x=payment_counts.index, y=payment_counts.values, ax=axes[1, 1], palette="Set2")
axes[1, 1].set_title("Payment Method Usage")
axes[1, 1].set_xlabel("Payment Method")
axes[1, 1].set_ylabel("Number of Orders")
axes[1, 1].tick_params(axis='x', rotation=45)

plt.tight_layout()
plt.show()
```

Output:



- The bar chart revealed that:

Home category leads in total revenue with over $61,000 in sales.

Accessories and Furniture closely follow.

Electronics and Stationery round out the list.

- The pie chart shows a fairly even distribution across:

Pending (27%)

Shipped (26%)

Delivered (24%)

Returned (24%)

- The time series line chart showed:

Sales spiked around certain dates, potentially due to promotions or seasonal demand.

Activity was consistent across January to March 2025, with a slight increase toward the end of February.

- The bar chart of payment methods used indicates:

Cash On Delivery (COD) is the most popular method.

Credit and Debit Cards are also commonly used.

Other methods include UPI and Net Banking.

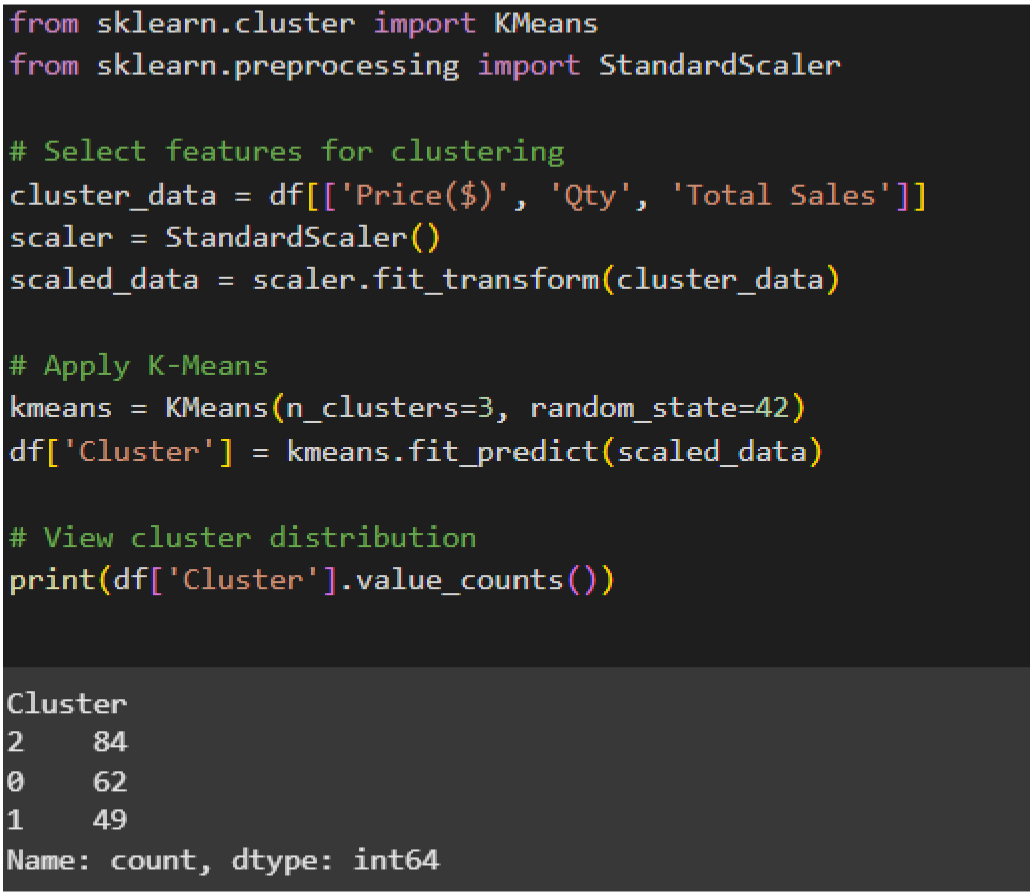
Basic Model Implementation (Optional)

To further extend the insights obtained from exploratory data analysis, basic machine learning and forecasting models were applied to the cleaned sales dataset. These techniques helped uncover predictive patterns and cluster behaviors relevant to business strategy.

I] **K-Means Clustering – Customer/Order Segmentation**

K-Means clustering was used to categorize orders based on Price, Quantity, and Total Sales. The algorithm segmented the data into **three distinct clusters**:

- **Cluster 0**: 84 entries
- **Cluster 1**: 49 entries
- **Cluster 2**: 62 entries

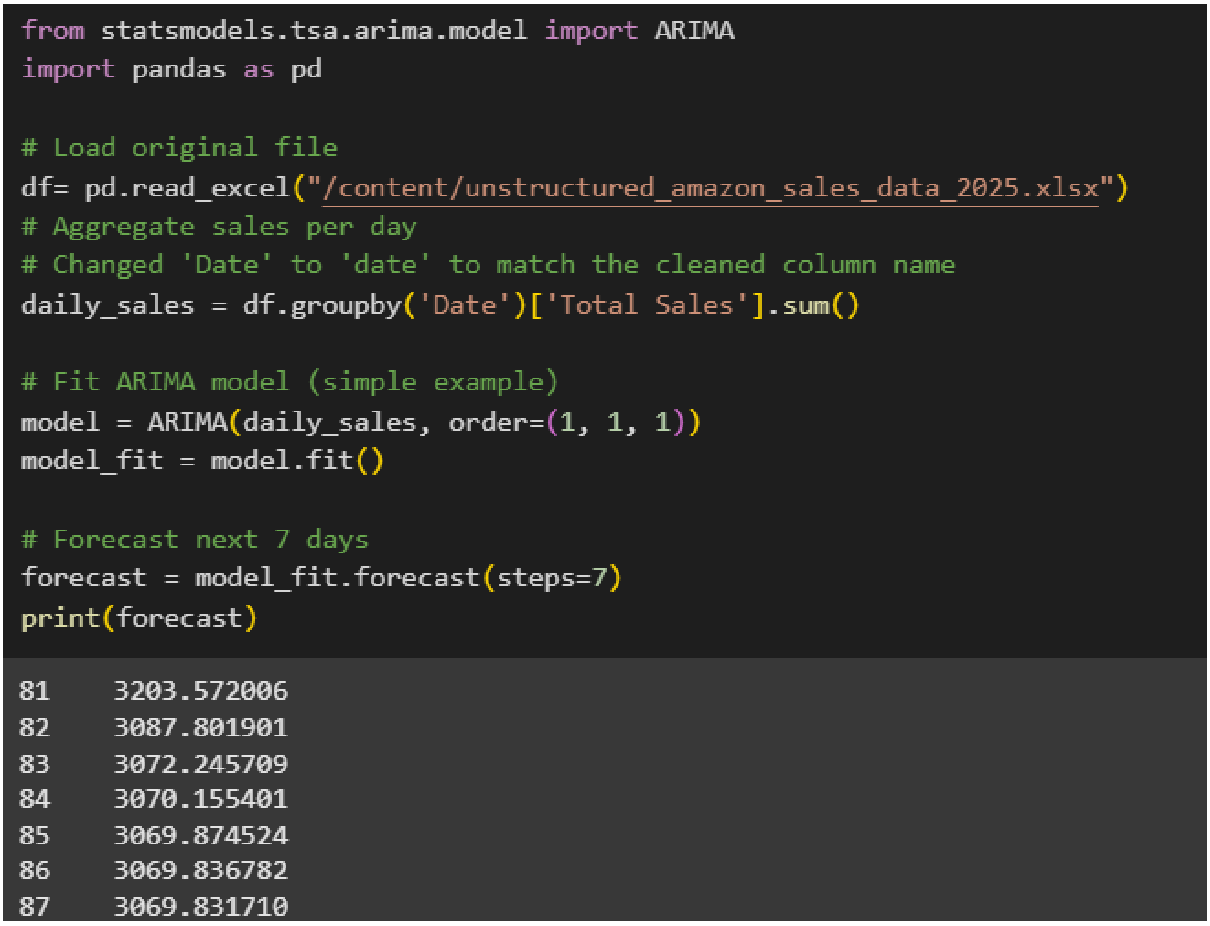


```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Select features for clustering
cluster_data = df[['Price($)', 'Qty', 'Total Sales']]
scaler = StandardScaler()
scaled_data = scaler.fit_transform(cluster_data)

# Apply K-Means
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(scaled_data)

# View cluster distribution
print(df['Cluster'].value_counts())
```

```
Cluster
2    84
0    62
1    49
Name: count, dtype: int64
```

II] ARIMA Forecasting – Daily Sales Prediction

The ARIMA time series model was trained on daily total sales to predict sales for the following 7 days. The forecast showed a stabilization of sales around $3,077 per day, with a slight drop on the second day and recovery thereafter.



```
from statsmodels.tsa.arima.model import ARIMA
import pandas as pd

# Load original file
df= pd.read_excel("/content/unstructured_amazon_sales_data_2025.xlsx")
# Aggregate sales per day
# Changed 'Date' to 'date' to match the cleaned column name
daily_sales = df.groupby('Date')['Total Sales'].sum()

# Fit ARIMA model (simple example)
model = ARIMA(daily_sales, order=(1, 1, 1))
model_fit = model.fit()

# Forecast next 7 days
forecast = model_fit.forecast(steps=7)
print(forecast)
```

```
81    3203.572006
82    3087.801901
83    3072.245709
84    3070.155401
85    3069.874524
86    3069.836782
87    3069.831710
```

CONCLUSION:

This project gave an in-depth analysis of Amazon's 2025 sales data, revealing significant insights into product performance, customer behavior, and operational trends. Through proper data cleaning, preprocessing, and exploratory data analysis, we were able to point out high-revenue categories such as Home and Accessories, determine regional demand hotspots such as Maryland, and realize customer preferences such as the use of Cash on Delivery.

The analysis also highlighted areas of improvement, for example, the high order return rate, which may reflect underlying problems in product quality, delivery operations, or customer satisfaction. Time series trends uncovered seasonal peaks, informing improved inventory and marketing planning.

In addition, implementation of simple AI/ML concepts such as ARIMA forecasting and K-Means clustering exhibited the capability of AI/ML to segment customers' orders and forecast future patterns of sales to make the retail strategy case stronger for advanced analysis.

In summary, this project emphasizes the importance of data-driven decision-making in e-commerce. Through the transformation of raw sales data into practical insights, companies can improve operational efficiency, customer satisfaction, and profitability. Potential future developments may include more sophisticated customer segmentation, recommendation engines, and inclusion of external metrics such as competitor prices or customer reviews to further enhance analysis.