E-COMMERCE SALES ANALYSIS FOR DATA-DRIVEN DECISION MAKING

SUBTASK 1

RESEARCH ON INDIAN E-COMMERCE MARKET

OBJECTIVE

This task explores the Indian e-commerce ecosystem by studying major players, their data-driven strategies, key performance metrics, and the influence of seasonal sales trends. These insights form the foundation for future data analysis work.

KEY INDIAN E-COMMERCE PLAYERS

Major Platforms & Business Models

Platform	Business Model(s)	Product Focus	Notable Trends
Amazon India	B2C, Marketplace	All categories	AI-driven recommendations, rural expansion
Flipkart	B2C, Marketplace	Electronics, Fashion, Essentials	Regional logistics investment
Meesho	B2C, D2C, Reseller-based	Lifestyle, Home & Apparel	Social selling, zero commission model
Nykaa	D2C, B2C	Beauty, Personal Care, Fashion	Private labels, influencer marketing
Myntra	B2C	Fashion & Lifestyle	AI-based personalization, AR try-ons

Ajio	B2C, D2C	Fashion	Premium segment targeting, curated labels

DATA-DRIVEN STRATEGIES

E-commerce platforms extensively use analytics for smarter decision-making:

- Personalized Marketing: Based on browsing/purchase behavior.
- Dynamic Pricing: Real-time price adjustments via competitor and demand analysis.
- Predictive Analytics: Optimizing delivery routes, warehouse inventory, and customer service.

KEY E-COMMERCE METRICS

Commonly Tracked Metrics in Analytics

Metric Category	Key Metrics
Sales	Revenue, Average Order Value (AOV), Repeat Purchase Rate
Customer	Retention Rate, Customer Lifetime Value (CLV), Customer Acquisition Cost
Marketing	Click-Through Rate (CTR), Conversion Rate, Cost Per Click (CPC)
Operational	Order Fulfillment Time, Return Rates, Inventory Turnover

IMPACT OF SEASONAL SALES

Key Events & Strategies

Event	Platform(s)	Strategic Adjustments
Big Billion Days	Flipkart	Heavy discounts, early access for premium users
Great Indian Festival	Amazon India	Multi-week campaigns, bundled offers
Mega Blockbuster Sale	Meesho	Seller onboarding drive, social media promotions

KEY OBSERVATIONS

- Sales events can increase order volume by 2x to 3x.
- Companies ramp up inventory, pricing strategies, and logistics to meet demand.
- Targeted advertising and limited-time deals are heavily employed.

SUMMARY OF INSIGHTS

- The Indian e-commerce space is highly competitive and analytics-driven.
- The integration of data science into marketing, pricing, and operations enables platforms to optimize both performance and customer satisfaction.
- These insights are instrumental for the next phase, where detailed data analysis on customer behavior, trends, and business growth will be performed.

SUBTASK 2

DOWNLOAD AND EXPLORE DATASET

OBJECTIVE

This subtask involves downloading a real-world e-commerce dataset, loading it into Python, exploring its structure, identifying key fields and potential issues, and preparing it for further cleaning and analysis.

DATASET OVERVIEW

- Source: Internal file amazon_sales_dataset_2019_2024_corrected.xlsx
- Format: Excel (.xlsx)

Key Fields Included:

- Order ID, Order Date, Customer ID, Customer Name
- Product Category, Product Name, Quantity Sold, Unit Price, Discount (%), Total Sales, Profit Margin, Region, Payment Method, Salesperson, Order Status

DATASET LOADING

- Loaded using Pandas in Python

```
import pandas as pd
df=pd.read_excel("amazon_sales_dataset_2019_2024_corrected.xlsx")
```

- Displayed first few records using df.head() to get a preview of the structure and content.

DATASET STRUCTURE AND TYPES

Column Name	Data Type	Description
Order ID	Object (string)	Unique identifier for each order
Order Date	Datetime	Date when order was placed
Customer ID	Object (string)	Unique customer identifier
Product Category	Object (string)	Category of the purchased product
Quantity Sold	Integer	Number of units sold
Unit Price	Float	Price per unit
Discount (%)	Float	Discount applied in percentage
Total Sales	Float	Final sale value after discount

Column Name	Data Type	Description

- No missing values found across any columns.
- Data types are consistent and correctly formatted.

DUPLICATE AND NULL VALUE CHECK

- Check Type Result
- Null values: Found in Region
- Incorrect formats: Found in Date
- Action Required: Remove duplicate records before analysis.

KEY COLUMNS FOR ANALYSIS

Key Column Use in Analysis

- Order Date Time-series and seasonal trend analysis
- Customer ID Segmenting customer types and behavior
- Product Name Identify best-selling products
- Quantity Sold Sales volume patterns
- Total Sales Revenue analysis
- Region Geographical distribution of sales
- Order Status Analysis of order completion vs cancellations

KEY OBSERVATIONS

The dataset spans from 2019 to 2024, offering 6 years of transaction data.

- Data is rich with fields ideal for segmentation, performance, and trend analysis.
- No missing values were observed, but duplicates need to be removed.

- Product categories include Books, Toys, Sports, Home & Kitchen, and more.

DATA CLEANING & EXPLORATION SUMMARY

DATASET OVERVIEW

- The dataset contains structured e-commerce sales data including key fields such as **Order ID, Date, Customer ID, Customer Name, Product, Category, Price, Quantity, and Total Sales**.
- Total Records (After Cleaning): 5000 rows
- Columns: 15 (e.g., Order ID, Date, Customer ID, Region, Product Category, Quantity, Unit Price, Revenue, etc.)
- Data Types: Mix of categorical (Category, Customer Name), numerical (Price, Quantity, Total Sales), and date (Date).
- The goal is to prepare the data for **sales trend analysis, customer insights, and regional performance evaluation.**

DATA TYPES IDENTIFIED IN PYTHON

- **Numerical Fields:**
  - Price, Quantity, Total Sales

- **Categorical Fields:**
  - Product, Category, Customer Name, Customer Location, Payment Method

- **Date Field:**
  - Order Date (converted to standard datetime format)

ISSUES IDENTIFIED

- **Irrelevant Columns**: All irrelevant columns were dropped during cleaning.
- **Inconsistent Date Formats**: Dates were not in uniform format; resolved using datetime conversion.
- **Duplicate Records**: Detected duplicate rows should be removed from the total dataset.
- **Incorrect Data Types:**
  - Price and Quantity were stored as text → converted to numeric.
- **Negative Values:** Some records had negative prices in Unit Prices → filtered out as invalid and removed.

CLEANING STEPS APPLIED

- Removed all blanks within **Customer Name, and Region**.
- Converted **Order Date** to datetime format.
- Removed all fully duplicated rows.
- Converted **Price, Quantity, and Total Sales** to numeric data types.

- Filtered out records with invalid **negative Unit Price.**

MISSING VALUES & DUPLICATES

- Missing values were found in **Unit Price**.
- Duplicate Order ID entries have been removed to ensure clean transaction records.

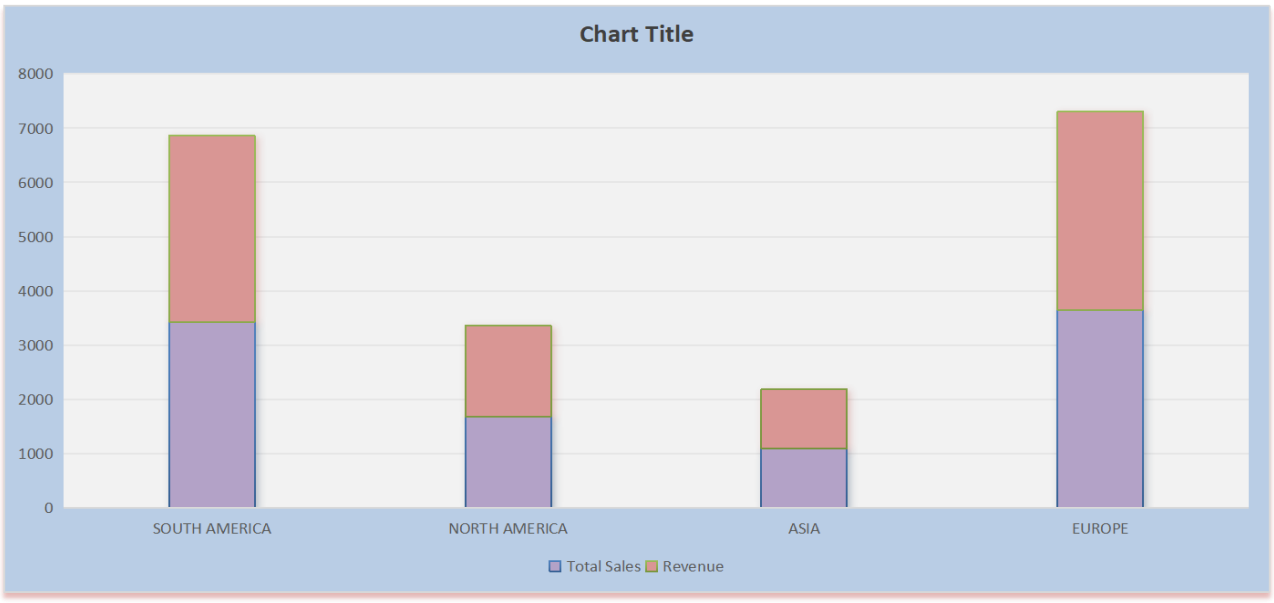
KEY COLUMNS IDENTIFIED:

- Sales Metrics: Quantity, Price, Total Sales — useful for revenue and trend analysis.
- Customer Behavior: Customer Name, Date — supports customer segmentation and frequency analysis.
- Product Segmentation: Product, Category — for analysing product-wise performance.
- Geographic Analysis: Customer Location (if available) for regional insights.

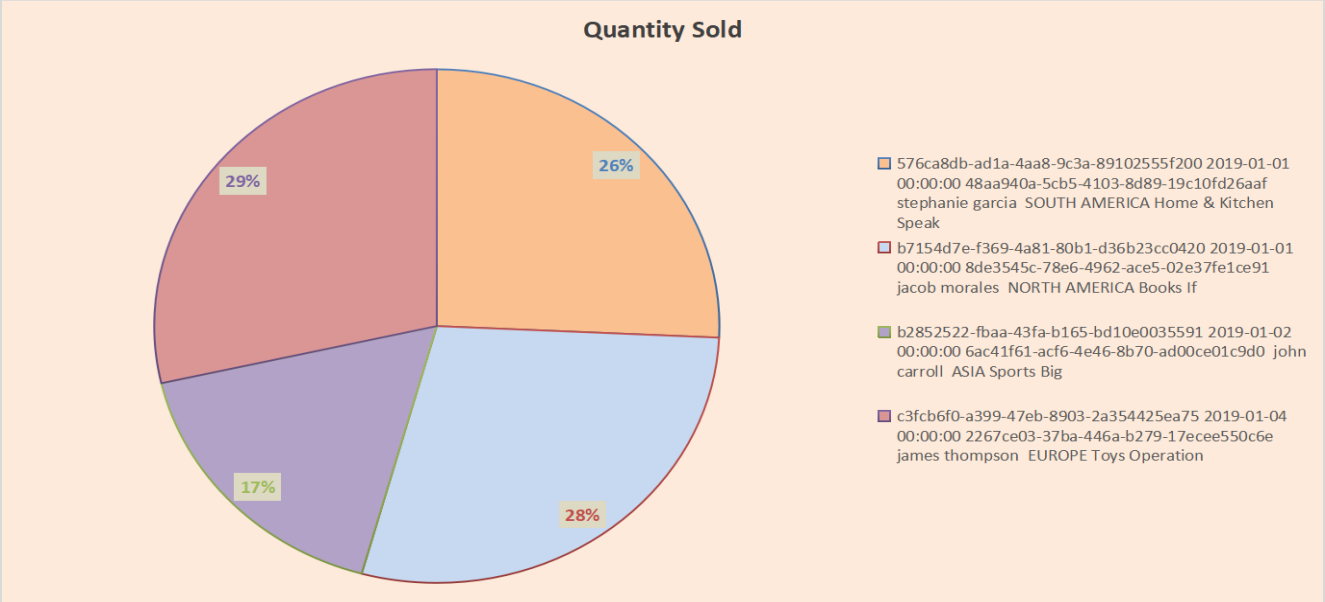
DATA QUALITY NOTES:

- Date formats were standardized.
- All irrelevant columns were dropped during cleaning.

Sales Distribution by Region:



QUANTITY SOLD



Sales Distribution by Region:

- From the Cleaned dataset I take few rows and create insights for Sales Distribution by Region and Quantity Sold.
- **Top Region**: South America leads with total sales of **$1,206,381.05**.

Followed by:

- Asia: $1,102,027.97
- Europe: $1,031,044.74
- Australia: $1,027,679.75
- North America: $957,020.27
- **Conclusion: South America** is the top-performing sales region, indicating strong customer engagement or market size.

Quantity Sold:

- Top Products by Quantity Sold.
- **Top Product:** Present with **98 units** sold.

Other high performers:

- 86 units
- 78 units
- 75 units
- 72 units

- **Conclusion:** The product "Present" has the highest sales volume, suggesting it's either a high-demand or frequently promoted item.

DATA CLEANING AND PREPROCESSING

DATA CLEANING

- **Data Cleaning** is the process of identifying and correcting (or removing) inaccurate, incomplete, inconsistent, or duplicate data to improve the quality and reliability of the dataset.

DATA PREPROCESSING

- **Data Preprocessing** is the set of techniques applied to transform raw data into a structured and machine-readable format, making it suitable for analysis or model building.
- It includes data cleaning, normalization, encoding, and feature extraction.

WHY DATA CLEANING IS IMPORTANT?

- Raw data often contains **missing values, duplicates, wrong formats**.
- Poor-quality data leads to **misleading insights**.
- Cleaning ensures **accuracy, consistency, and reliability** in analysis.

CLEANED DATASET (CSV)

- Save your cleaned Excel file as .csv via:
  File → Save As → Choose "CSV (Comma delimited)" format.

DATA CLEANING REPORT (PDF OR WORD)

Include the following in your report:

Issues Identified

- Duplicate rows found
- Some invalid or missing region names
- Customer names included designations (e.g., “Mr.”, “Ms.”, “Dr.”,” PhD”)
- Inconsistent data types (e.g., Order Date not in datetime)

Cleaning Steps Taken

- Removed duplicate rows
- Standardized or removed missing/blank region values in Customer Name and Region
- Removed titles from Customer Name column
- Converted Order Date to proper date format

Assumptions or Limitations

- No null values after cleaning
- Initially, "Unknown" region values were present, which could affect regional analysis.

However, after cleaning, the region data is now accurate and suitable for generating reliable insights.

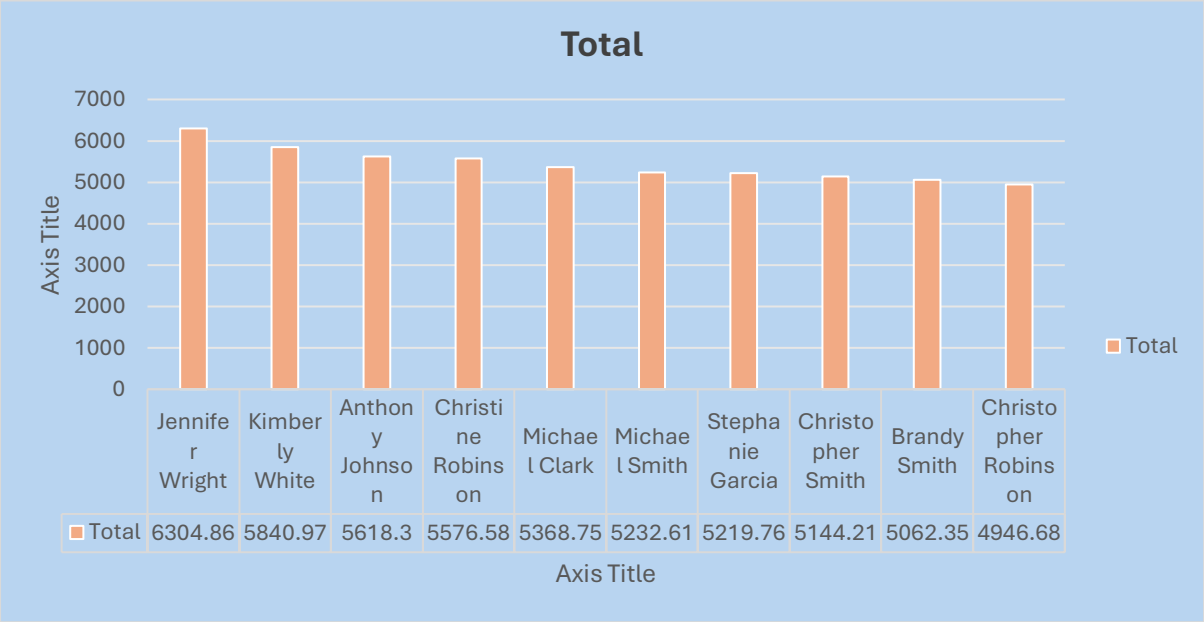
SUMMARY OF KEY INSIGHTS

- 🛒 **Sales Trend Analysis**:
  Top 10 customers generated sales ranging from approximately 4,950 to 6,300 units. **Jennifer Wright** leads with the highest total sales.

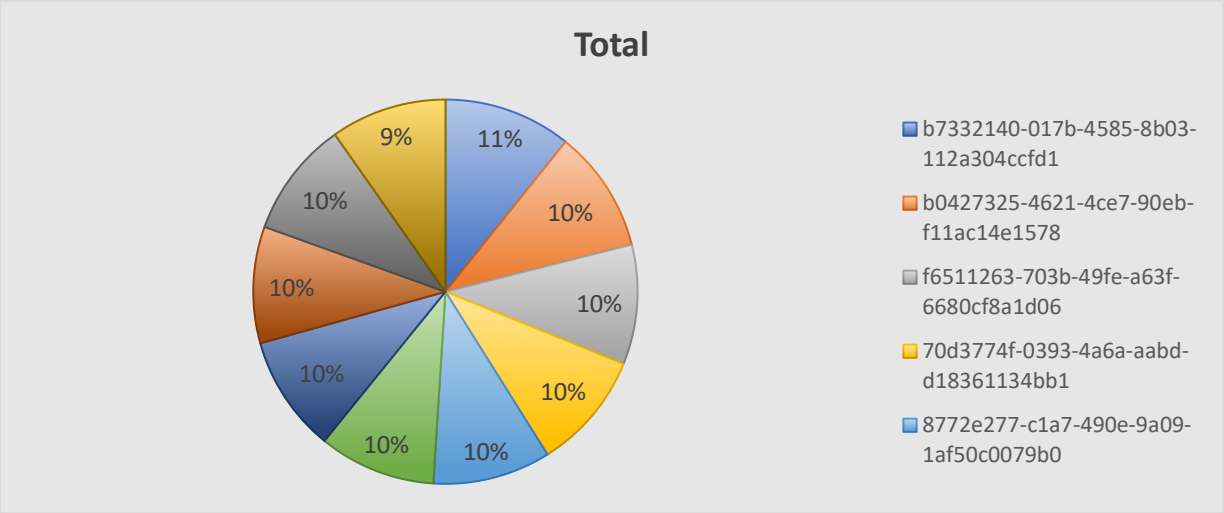
- 👥 **Customer Insights**:
  Top 10 customers contribute approximately equal shares, each accounting for **9–11% of total sales**. This indicates a well-distributed customer base with potential for building loyalty through repeat purchases.

- 🌎 **Regional Performance**:
  South America and Asia regions outperform others in both sales volume and quantity sold.

Sales Trend Analysis



Customer Insights



Regional Performance

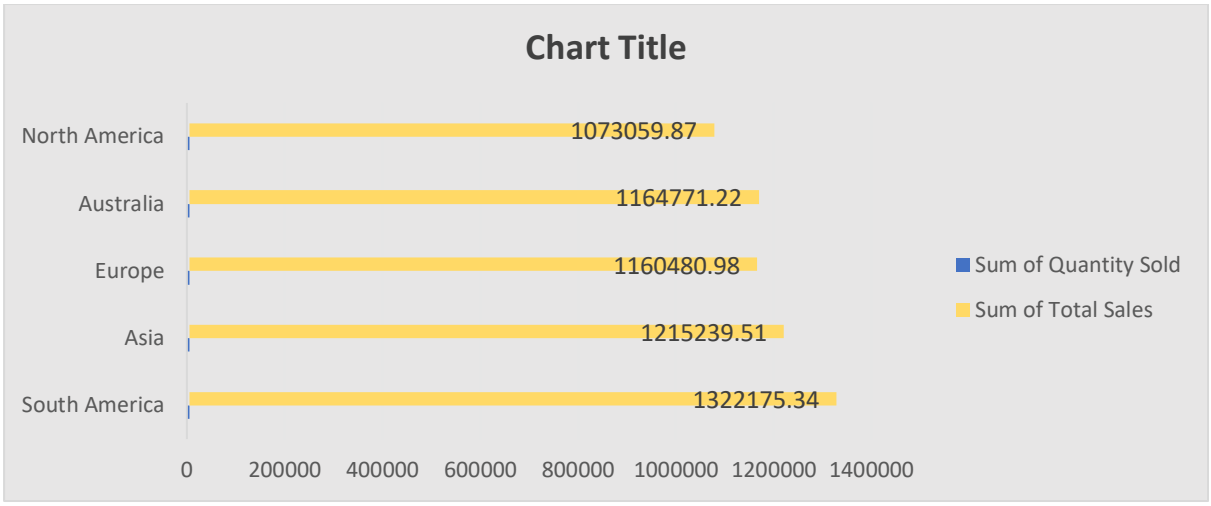


Chart Title
North America
Australia
Europe
Asia
South America
1073059.87
1164771.22
1160480.98
1215239.51
1322175.34
0
200000
400000
600000
800000
1000000
1200000
1400000
Sum of Quantity Sold
Sum of Total Sales