Customer Shopping Behavior Analysis

1. Project Overview

This project analyses customer shopping behavior using transactional data from 9,800 purchases across various product categories. The goal is to uncover insights into spending patterns, customer segments, product preferences, and subscription behavior to guide strategic business decisions**.**

2. Dataset Summary

- Rows: 9,800

- Columns: 18

- Key Features:

  Customer demographics(country, city, state, region)

  Purchase details(category ,sub-category , product name

  Shopping behavior (ship mode)

- Missing Data: 11 values in postal code

3. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

- Data Loading: Imported the dataset using pandas.
- Initial Exploration:

  Used df.info() to check structure and
  .describe() for summary statistics.

	Row_ID	Postal_Code	Sales
count	9800.000000	9789.000000	9800.000000
mean	4900.500000	55273.322403	230.769059
std	2829.160653	32041.223413	626.651875
min	1.000000	1040.000000	0.444000
25%	2450.750000	23223.000000	17.248000
50%	4900.500000	58103.000000	54.490000
75%	7350.250000	90008.000000	210.605000
max	9800.000000	99301.000000	22638.480000

4.Missing Data Handling: Checked for null values and imputed missing values in the Postal code and removed the unwanted column.

5.Check for White Space: Checked for white space using the command df.columns.

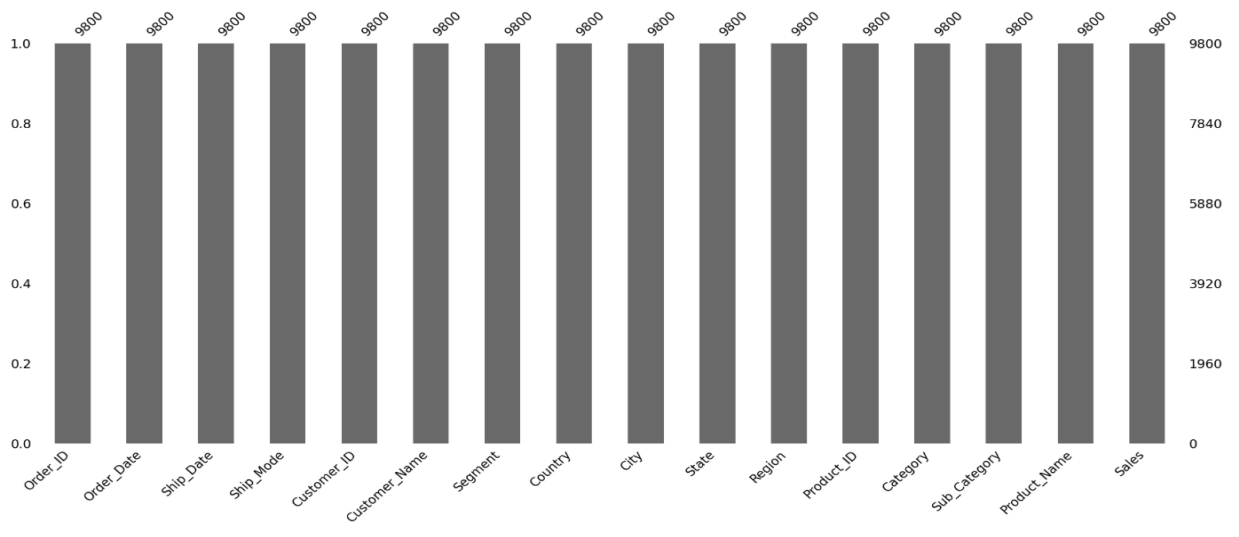
6.Data Cleaning: Removed unwanted columns that conveys the same meaning using

df = df.drop(columns=['Row_ID','Postal_Code'], axis = 1)

7.Missing values: Checked missing values using df.isnull().sum() command

```
Order_ID          0
Order_Date        0
Ship_Date         0
Ship_Mode         0
Customer_ID       0
Customer_Name     0
Segment           0
Country           0
City              0
State             0
Region            0
Product_ID        0
Category          0
Sub_Category      0
Product_Name      0
Sales             0
dtype: int64
```

8.Visualise the missingness: Checked the missingness using the command



9.Change date column to appropriate data type: Checked for the data type and changed its type accordingly .

10.Data Validation:

1 Understand Business Requirements

- Know what each column represents.
- Identify expected values and rules.
- Example: AQI must be between 0–500, Clicks ≤ Impressions.

---

2 Check Data Structure

- Verify number of rows and columns.
- Check column names and formats.
- Ensure dataset matches expected schema.

---

3 Validate Data Types

- Ensure correct datatype for each column.

- **Example:**
    - Numbers → Integer/Float
    - Dates → Datetime
    - Text → String

4️⃣ Check Missing Values

- Identify null or empty values.
- Decide whether to fill, remove, or keep them.
- Ensure critical fields are not missing.

5️⃣ Check Duplicate Records

- Detect repeated rows.
- Remove duplicates if not required.
- Important for primary keys in MySQL tables.

6️⃣ Validate Data Range

- Ensure values fall within valid limits.
- **Example:**
    - Age cannot be negative.
    - AQI cannot exceed 500.
    - Ratings between 1–5.

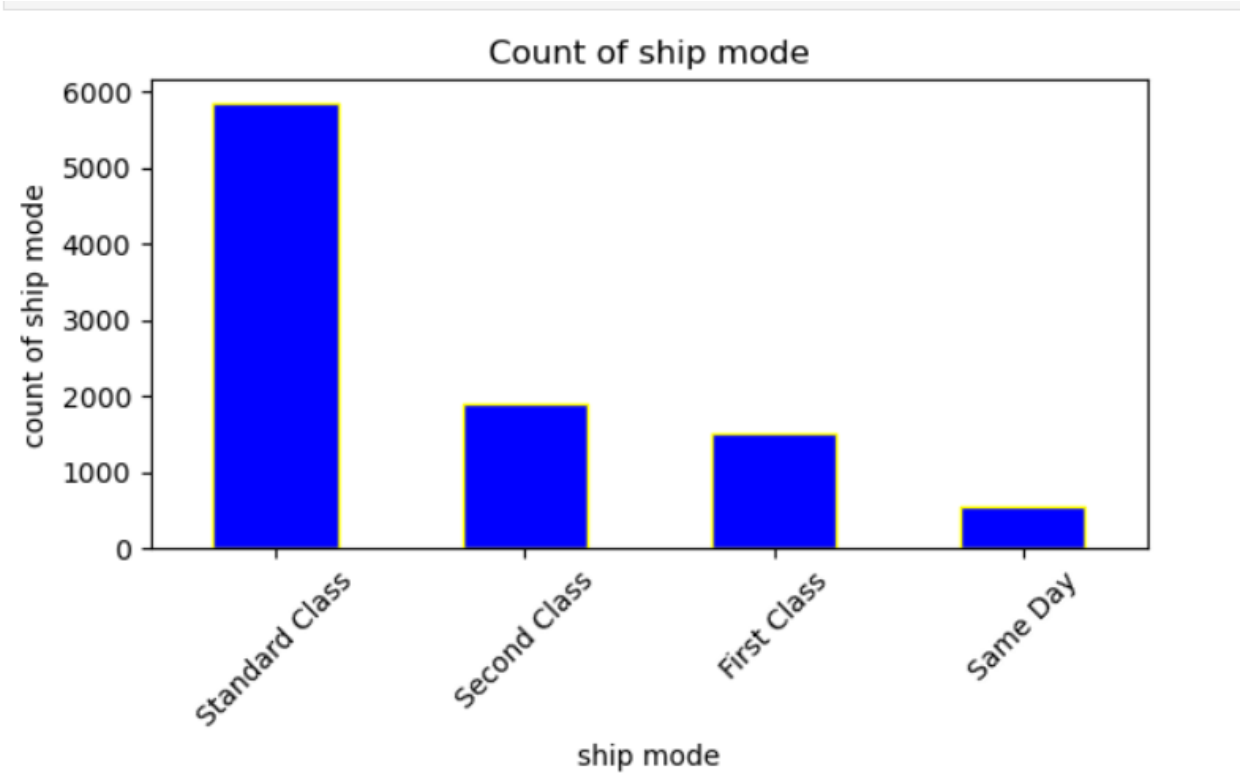
7️⃣ Validate Categorical Values

- Check unique values in category columns.
- Detect spelling mistakes or invalid categories.
- Example: "Indai" → "India".

11.Exploratory data analysis:

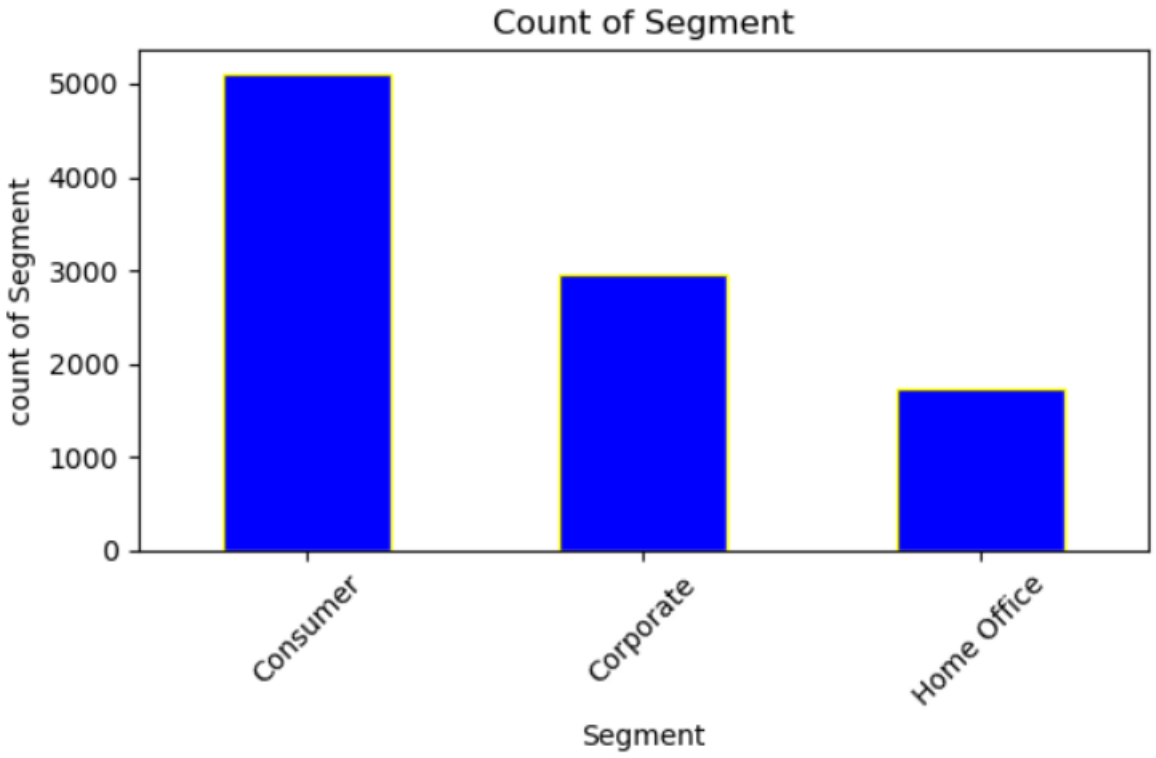
Univariate Analysis: From the analysis, the company operates in three different shipping modes: standard class, second class, first class, and same day. The standard mode of shipping has the highest count of ship mode followed by seccond class with the same day has the lowest . From this analysis we can understand customer prefer standard shipping process compared to the othe shipping mode. The reason is that standard shipping mode offer cheaper rate compared to other shipping modes.

--Business recomendation -The price of other shipping modes should be revised to attract more customer -Promo programs can be used to entice customers that make use of shipping modes aside the standard class -Further analysis should be made to understand the difference between the delivery for all shipping modes , that way we can determine that customer actually get the value for their money they opt for other shipping modes aside from the standard shipping.



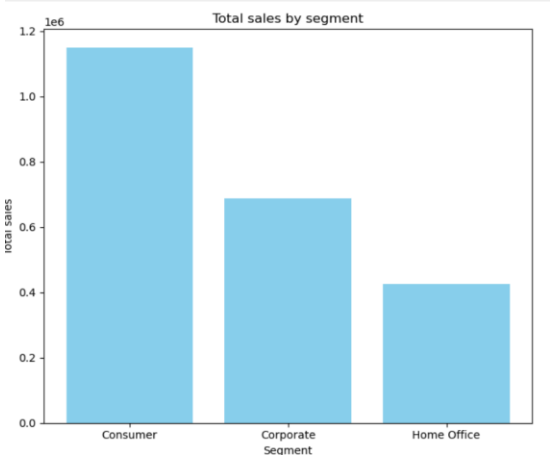
This analysis shows the consumer segment has the highest number of sales orders, followed by the corporate, and then the home office. This indicates that the business serves a predominantly consumer base with most of their transaction coming from business clients.

Business Recommendation --Since most of the orders come from consumers, we should ensure the buying experience is seamless. Product availability should be optimised, including the pricing and consumer support. --The lower volume of the other segmants suggest a room for growth. We should consider targeted outreach or tailored offerings to increase their engagement. --Segmant based marketing should also be considered.

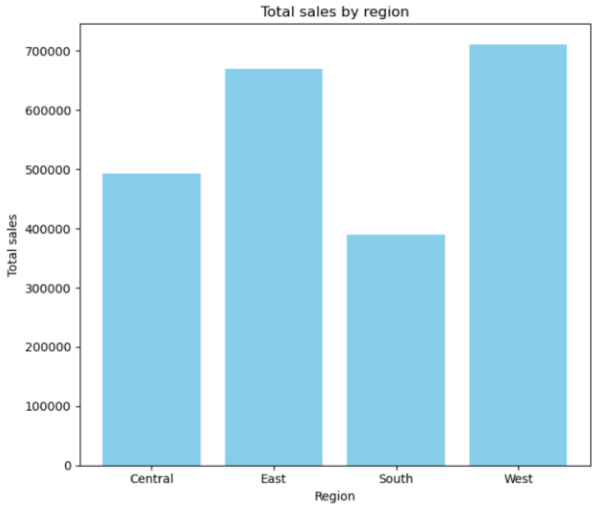


12.Bivariate Exploratory Analysis:

The Consumer segment generates the highest sales and is the primary revenue driver for the business. The company should continue investing in targeted marketing and loyalty programs to maximize profits from this segment. The Corporate segment shows good potential and can be improved through bulk discounts and long-term contracts. The Home Office segment has the lowest sales, so focused promotions and tailored product bundles can help increase its performance.



From the analysis we can see central region has highest sales followed by east,south and west region.



13.Multivariant analysis: The west region leads all other region exceeding 700k, driven mainly by consumers and corporations using the standard shipping class. The

follows South Falls slowly

Across all standard shipping mode dominates. This shows consumers' preference for affordability over speed