Customer Order & Revenue Analysis

1. Project Overview

This project analyzes transactional order data generated from a custom-built Next.js e-commerce platform. The dataset captures customer demographics, product details, pricing, discounts, payment methods, and order status.

The objective of the project is to prepare raw application data using Python (Pandas) and perform structured analysis using PostgreSQL to derive insights into revenue, product performance, and customer behavior.

2. Dataset Summary

Rows: 200

Columns: 16

Key Attributes:

- Customer demographics: Age, Gender, City, State
- Order details: Order Date, Order Status, Payment Method
- Product information: Product Name, Product Category
- Pricing fields: Quantity, Price, Discount, Revenue

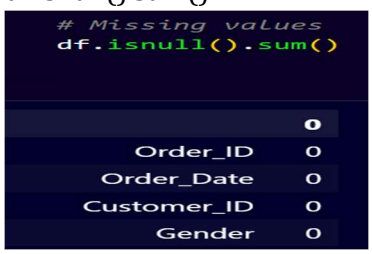
Initial Data Observations:

- Missing values present in the Age column
- Revenue not provided in raw data and required calculation
- Inconsistent text formatting across categorical columns

3. Exploratory Data Analysis using Python

Data exploration and cleaning were performed using Python (Pandas) across two Jupyter notebooks.

Data Loading & Initial Inspection:

- The dataset was imported using pandas.read_csv(). Dataset structure and data types were examined using df.head(), df.shape, and df.info(). Missing values were identified using df.isnull().sum().
- 

Age Group Feature Engineering:

- A custom function was used to group customer ages into categories: 0–18, 19–30, 31–45, 46–60, and 60+. A new column age_group was created.

```
def age_group(age):
    if pd.isna(age):
        return "Unknown"
    elif age <= 25:
        return "18-25"
    elif age <= 35:
        return "26-35"
    elif age <= 45:
        return "36-45"
    elif age <= 59:
        return "46-59"
    else:
        return "60+"

df["age_group"] = df["age"].apply(age_group)
```

Categorical Data Standardization:

- Text-based columns were cleaned by stripping extra whitespace and standardizing capitalization.

```
text_cols = ["gender","city","state","product_category","payment_method","order_status"]

for col in text_cols:
    df[col] = df[col].str.strip().str.title()
```

Revenue Calculation:

- Revenue = Quantity × Price × (1 − Discount)

```
# Revenue column
df["revenue"] = df["quantity"] * df["price"] * (1 - df["discount"])
```

Final Output:

- The cleaned dataset was exported as cleaned_orders.csv.

4. Data Analysis using SQL

Revenue analysis, demographic analysis, product performance, discount impact, order status distribution, payment methods, city-level orders, high-value orders, and customer ordering behavior were analyzed using PostgreSQL.

- **Sales & Revenue Analysis**

Q1. What is the total revenue generated by the company?

```
Query  Query History
1  SELECT
2      ROUND(SUM(revenue), 2) AS total_revenue
3  FROM orders_cleaned
4  WHERE order_status = 'Delivered';
5
```

Data Output | Messages | Notifications

	total_revenue numeric
1	266887.10

Q2. Which products are generating the lowest revenue (low-performing products)?

```
Query  Query History
1  SELECT
2      product_name,
3      ROUND(SUM(revenue), 2) AS total_revenue
4  FROM orders_cleaned
5  WHERE order_status = 'Delivered'
6  GROUP BY product_name
7  ORDER BY total_revenue ASC
8  LIMIT 10;
9
```

Data Output | Messages | Notifications

	product_name character varying (100)	total_revenue numeric
1	Kitchen Set	2396.10
2	Wireless Mouse	2464.80
3	Bluetooth Speaker	4711.50
4	Power Bank	8429.70
5	Jeans	12679.20
6	Smart Watch	16759.00
7	T-Shirt	19826.80
8	Lipstick	24263.65
9	Skincare Kit	24850.05
10	Headphones	26092.35

- **Product Performance Analysis**

Q3. Which products are sold the most by quantity?

```
Query  Query History
1  SELECT
2      product_name,
3      SUM(quantity) AS total_units_sold
4  FROM orders_cleaned
5  WHERE order_status = 'Delivered'
6  GROUP BY product_name
7  ORDER BY total_units_sold DESC;
8
```

Data Output | Messages | Notifications

	product_name character varying (100)	total_units_sold bigint
1	Protein Powder	22
2	Curtains	20
3	Skincare Kit	12
4	Dress	12
5	Bedsheet	11
6	Smart Watch	10
7	Lipstick	10
8	Headphones	8
9	Power Bank	7
10	T-Shirt	5
11	Kitchen Set	5

Q4. Are certain products associated with higher return rates?

Query | Query History

```
SELECT
    product_name,
    COUNT(*) FILTER (WHERE order_status = 'Returned') * 100.0 / COUNT(*)
        AS return_rate_percentage
FROM orders_cleaned
GROUP BY product_name
HAVING COUNT(*) > 3
ORDER BY return_rate_percentage DESC;
```

Data Output | Messages | Notifications

	product_name character varying (100)	return_rate_percentage numeric
1	T-Shirt	50.0000000000000000
2	Jeans	50.0000000000000000
3	Wireless Mouse	42.8571428571428571
4	Smart Watch	41.6666666666666667
5	Kitchen Set	36.3636363636363636
6	Dress	33.3333333333333333
7	Headphones	33.3333333333333333
8	Curtains	31.5789473684210526
9	Lipstick	29.4117647058823529
10	Skincare Kit	26.9230769230769231
11	Protein Powder	25.0000000000000000
12	Power Bank	23.0769230769230769
13	Bedsheet	23.0769230769230769

- **Customer Demographics Analysis**

Q5. How are orders distributed across age groups?

Query | Query History

```
SELECT
    age_group,
    COUNT(*) AS total_orders
FROM orders_cleaned
GROUP BY age_group
ORDER BY total_orders DESC;
```

Data Output | Messages | Notifications

	age_group character varying (10)	total_orders bigint
1	Unknown	99
2	46-59	28
3	36-45	20
4	60+	19
5	18-25	18
6	26-35	16

Q6. Are there differences in purchasing behavior by gender in terms of revenue?

Query | Query History

```
1  SELECT
2      gender,
3      COUNT(*) AS total_orders,
4      ROUND(SUM(revenue), 2) AS total_revenue
5  FROM orders_cleaned
6  WHERE order_status = 'Delivered'
7  GROUP BY gender;
8
```

Data Output | Messages | Notifications

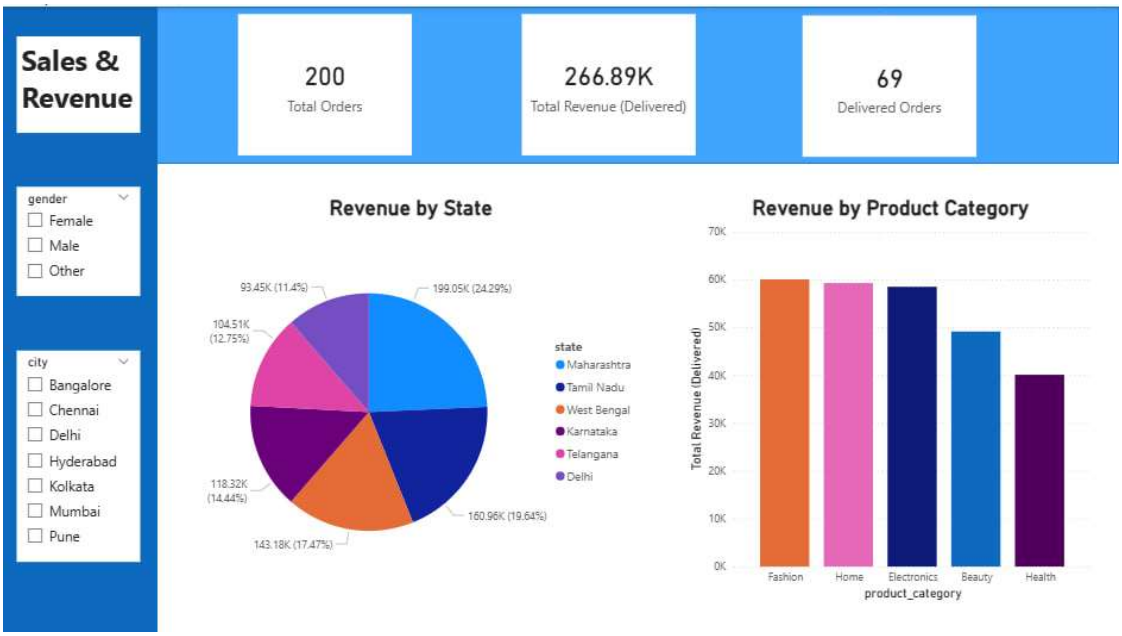
	gender character varying (10)	total_orders bigint	total_revenue numeric
1	Other	28	105054.45
2	Male	19	63981.20
3	Female	22	97851.45

5. Dashboard / Visualization Layer

The cleaned and analyzed dataset was visualized using an interactive dashboard to present insights in an intuitive and business-friendly manner. The dashboard is divided into three analytical sections: Sales & Revenue, Product Performance, and Customer Demographics.

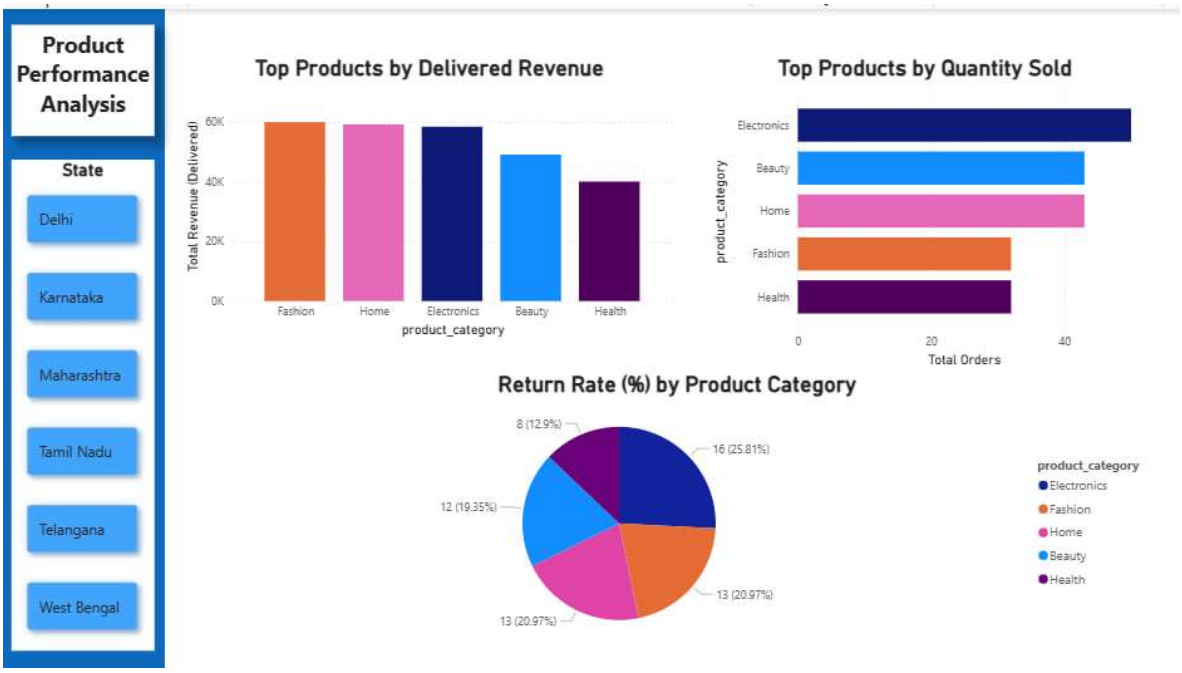
1. Sales & Revenue Analysis

This section analyzes overall business performance by measuring total orders, delivered revenue, and successful deliveries. It identifies key revenue-contributing states and product categories, helping understand geographical demand and high-earning segments.



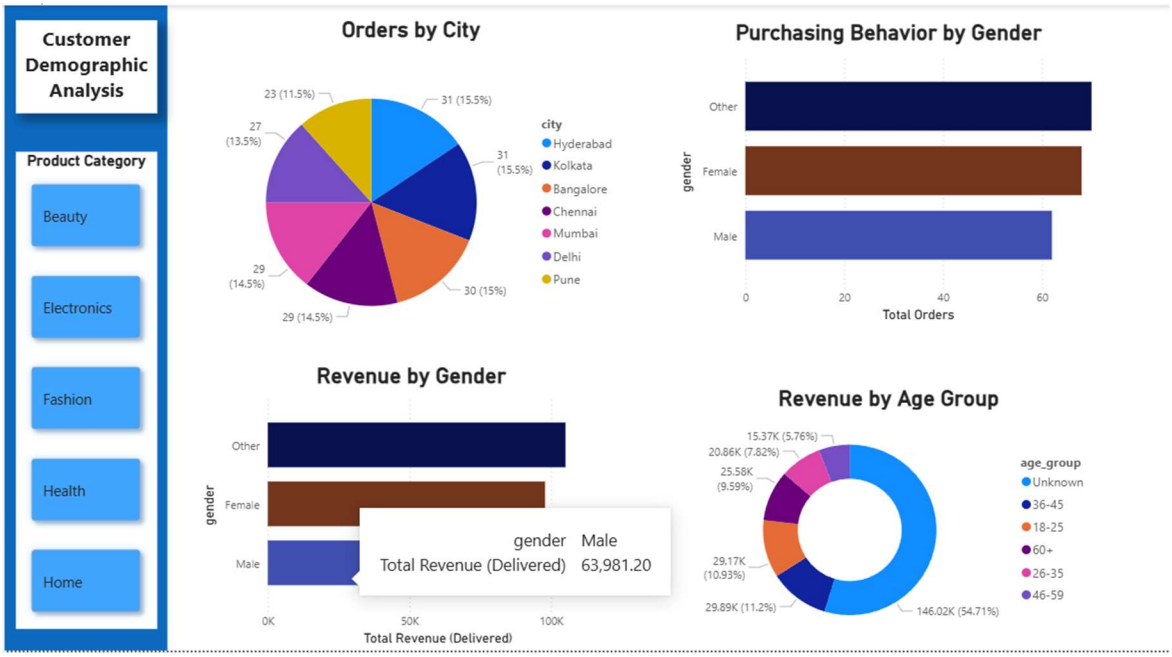
2. Product Performance Analysis

This analysis evaluates product categories based on revenue, quantity sold, and return rates. It helps identify top-performing products as well as low-performing categories that generate sales volume but contribute less to revenue or have higher returns.



3. Customer Demographic Analysis

This section studies customer purchasing behavior across cities, gender, and age groups. It highlights major customer locations and demographic segments contributing to revenue, supporting targeted marketing and customer segmentation strategies.



6. Business Recommendations

- **Revenue is concentrated, not widespread**
  Out of **200 total orders**, only **69 delivered orders** generate the full **₹266.9K revenue**.
  👉 This indicates **significant leakage due to cancellations and returns**, which directly impacts profitability.

- **Top revenue-driving categories**
  - **Fashion, Home, and Electronics** together contribute the **major share of revenue**.
  - **Health** generates the **lowest revenue**, making it a potential **low-performing category**.

- **Geographic concentration of orders**
  Cities like **Hyderabad, Kolkata, and Bangalore** show the **highest order volumes**, indicating:
  - Strong demand in metro and tier-1 cities
  - Opportunity for **regional marketing and faster delivery optimization**

- **Gender-based purchasing differences**

  Female category contribute slightly **more revenue than male category**.
  👉 Suggests potential for:

  Gender-targeted promotions

  Category-specific campaigns

- **City-level demand is evenly distributed**
  No single city dominates orders completely.
  👉 Indicates a **stable multi-city customer base**, reducing overdependence on one region.