



# **DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

## **PROJECT REPORT**

(Project Semester January-April 2025)

## **E-COMMERCE ANALYTICS: SWIGGY, ZOMATO, BLINKIT**

**Submitted by**

**NAME:** ANUSURI JNANA MANIKANTA SURYA

**REGISTRATION NO:** 12306050

**SECTION:** K23GN

**ROLL NO:** 2

**COURSE CODE :** INT375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

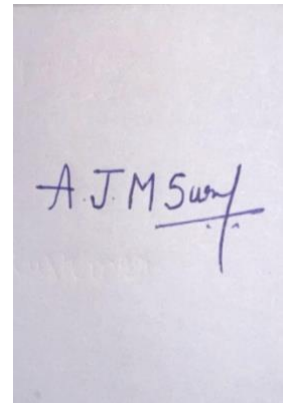
**Lovely Professional University, Phagwara**

## **CERTIFICATE**

This is to certify that **ANUSURI JNANA MANIKANTA SURYA** bearing Registration no. **12306050** has completed **INT375** project titled, “**E-COMMERCE ANALYTICS: SWIGGY, ZOMATO, BLINKIT**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

## **DECLARATION**

I, ANUSURI JNANA MANIKANTA SURYA student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

A photograph of a handwritten signature in blue ink on a light-colored surface. The signature reads 'A.J.M. Surya' with a stylized flourish at the end.

Signature:

Date: 11-04-2025

Registration No: 12306050

Name of the student: Anusuri Jnana Manikanta Surya

## Table of Contents

<b>TITLE</b>	<b>PAGE NO</b>
<b>1.Introduction</b>	<b>5</b>
<b>2. What is EDA?</b>	<b>5</b>
<b>3. Why EDA Is Important for E-Commerce Analytics</b>	<b>5</b>
<b>4. Source of Dataset</b>	<b>6</b>
<b>5. Step-by-Step EDA Process</b>	<b>6</b>
<b>6. Dataset Preprocessing</b>	<b>7</b>
<b>7. Univariate Analysis</b>	<b>7</b>
<b>8. Bivariate Analysis</b>	<b>7</b>
<b>9. Multivariate Analysis</b>	<b>8</b>
<b>10. Outlier Detection</b>	<b>8</b>
<b>11. Correlation Analysis</b>	<b>8</b>
<b>12. Analysis on Dataset</b>	<b>8 - 18</b>
<b>12.1. Objective 1: Order Trends &amp; Peak Hours Analysis</b>	<b>8 – 9</b>
<b>12.2. Objective 2: Delivery Time Analysis &amp; Delay Impact</b>	<b>9 – 10</b>
<b>12.3. Objective 3: Customer Satisfaction &amp; Service Ratings</b>	<b>10 – 12</b>
<b>12.4. Objective 4: Revenue Insights &amp; Top-Selling Products</b>	<b>12 – 13</b>
<b>12.5. Objective 5: Refund &amp; Complaint Analysis</b>	<b>13 – 14</b>
<b>13. Conclusion</b>	<b>15-16</b>
<b>14. Future Scope</b>	<b>16-17</b>
<b>15. References</b>	<b>18</b>

## 1. Introduction

E-commerce, short for electronic commerce, refers to the **buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet**. This report presents a comprehensive overview of an online business's performance, providing data-driven insights into various aspects of its operations and customer behavior. The main goal is to transform raw data into actionable intelligence that helps businesses make informed decisions, optimize their strategies, and drive growth.

This analysis primarily focuses on **optimizing business operations and enhancing customer experience**. It aims to improve efficiency by understanding order patterns and peak times, reduce delivery issues, boost customer satisfaction through service improvements, drive revenue by identifying top products, and address customer pain points through complaint analysis.

EDA is not just a technical step; it is a journey through the data that brings us closer to actionable insights. It allows businesses to make decisions not based on assumptions, but grounded in actual behavior.

---

## 2. What is EDA?

Exploratory Data Analysis (EDA) is a crucial first step in the data analysis process. It involves examining datasets to summarize their main characteristics, often with visual methods. EDA is used to:

- Get a sense of the structure, patterns, and relationships in data
- Identify anomalies, missing values, and outliers
- Generate hypotheses and guide further data modeling
- Understand the distribution of variables

### Techniques used in EDA:

- Descriptive Statistics: Mean, median, mode, range, standard deviation
- Data Visualization: Histograms, bar plots, scatter plots, box plots
- Data Cleaning: Handling null values, duplicates, formatting
- Feature Engineering: Creating new columns, segmenting categories
- Correlation & Relationships: Using statistical tools to assess interaction between variables

---

## 3. Why EDA is Important for E-Commerce Analytics: Swiggy, Zomato, Blinkit

EDA is crucial for Swiggy, Zomato, and Blinkit analytics because it helps to:

- **Understand Order Patterns:** Identify peak hours, popular items, and ordering trends to optimize operations and inventory.

- **Analyze Delivery Performance:** Visualize delivery times, identify delays, and understand their impact on customer satisfaction.
- **Segment Customers:** Discover different customer behaviors and preferences for targeted marketing and personalized experiences.
- **Improve Service Quality:** Analyze ratings and feedback to pinpoint areas for improvement in delivery and service.
- **Boost Revenue:** Identify top-selling products and customer segments driving revenue for focused strategies.
- **Address Issues:** Analyze refunds and complaints to understand pain points and implement solutions.
- **Detect Anomalies:** Identify unusual order patterns or behaviors that could indicate fraud or other issues.
- **Inform Decision-Making:** Provide data-driven insights for strategic decisions across various aspects of the businesses

This analytical approach ensures **data-driven decision-making by rigorously examining customer behavior, sales patterns, and operational metrics**. This enables businesses to **optimize strategies, personalize experiences, and ultimately drive growth and profitability**.

---

#### 4. Source of Dataset

The dataset was collected from a CSV file that records sales data from the E-Commerce

- File Name: E-commerce sales data.csv
  - Format: CSV (Comma-Separated Values)
  - Encoding: Latin1
  - Fields in the Dataset:
    - Gender: Male or Female
    - Age: Age of customer
    - State: State where the purchase was made
    - Amount: Total purchase value
    - Orders: Number of items ordered
- 

#### 5. Step-by-Step EDA Process

**EDA in this report follows these detailed steps:**

1. Import Libraries: Pandas, NumPy, Matplotlib, Seaborn
  2. Load Dataset: Read CSV file using Pandas
-

3. Initial Data Inspection: Check data types, shape, head, and summary
4. Data Cleaning:
  - Strip spaces from column headers
  - Convert relevant columns to numeric
  - Remove missing/null values
5. Feature Engineering:
  - Create custom Age\_Group brackets
  - Combine columns for analysis like Age\_Gender
6. Univariate Analysis: Analyze each variable on its own
7. Bivariate Analysis: Study relationships between two variables
8. Multivariate Analysis: Explore interactions among three or more variables
9. Outlier Detection: Identify extreme values using IQR and box plots
10. Correlation Study: Use heatmaps to understand variable relationships

---

## 6. Dataset Preprocessing

### Preprocessing steps include:

- Standardization of Column Names
- Conversion of Datatypes: Amount and Orders to numeric
- Handling Missing Values: Dropped records with missing Amount or Orders
- Creation of Age\_Group: Segmented into 18–25, 26–35, etc.

This ensures consistency, reduces noise, and prepares the data for analysis.

---

## 7. Univariate Analysis

We analyzed one variable at a time to understand distribution:

- Gender: Male vs Female proportions
- Age: Most active buyer age group
- Amount: Range of purchases
- Orders: Quantity trends

Graphs: Pie charts, bar plots, histograms

---

## 8. Bivariate Analysis

We studied interaction between two variables:

- Age vs Amount
- Gender vs Amount
- State vs Orders

This reveals how two features influence each other. For example, higher spending in certain states or age groups.

---

## 9. Multivariate Analysis

Here we examined Age + Gender + State:

- Who buys more?
- Which combination is most profitable?

Visualizations used: Heatmaps, stacked bar graphs, group plots

---

## 10. Outlier Detection

We used:

- IQR Method: Calculate Q1 and Q3, find outliers
- Boxplots: Visualized anomalies

Outliers were mostly large purchases—likely business clients

---

## 11. Correlation Analysis

We created:

- Correlation Matrix using `.corr()`
- Heatmap to visualize
- Pairplot to inspect pairwise relationships

Found strong correlation between Orders and Amount

---



## 12. Analysis on Dataset

### 12.1 Objective 1: Order Trends & Peak Hours Analysis

- **General Description:**

Understanding customer ordering behavior throughout the day is essential for optimizing resource allocation, staffing, and delivery scheduling. By analyzing order trends by hour, businesses can identify peak hours and adjust their operations—like increasing delivery personnel or support staff during high-demand periods—to enhance service efficiency and reduce wait times.

- **Specific Requirements:**

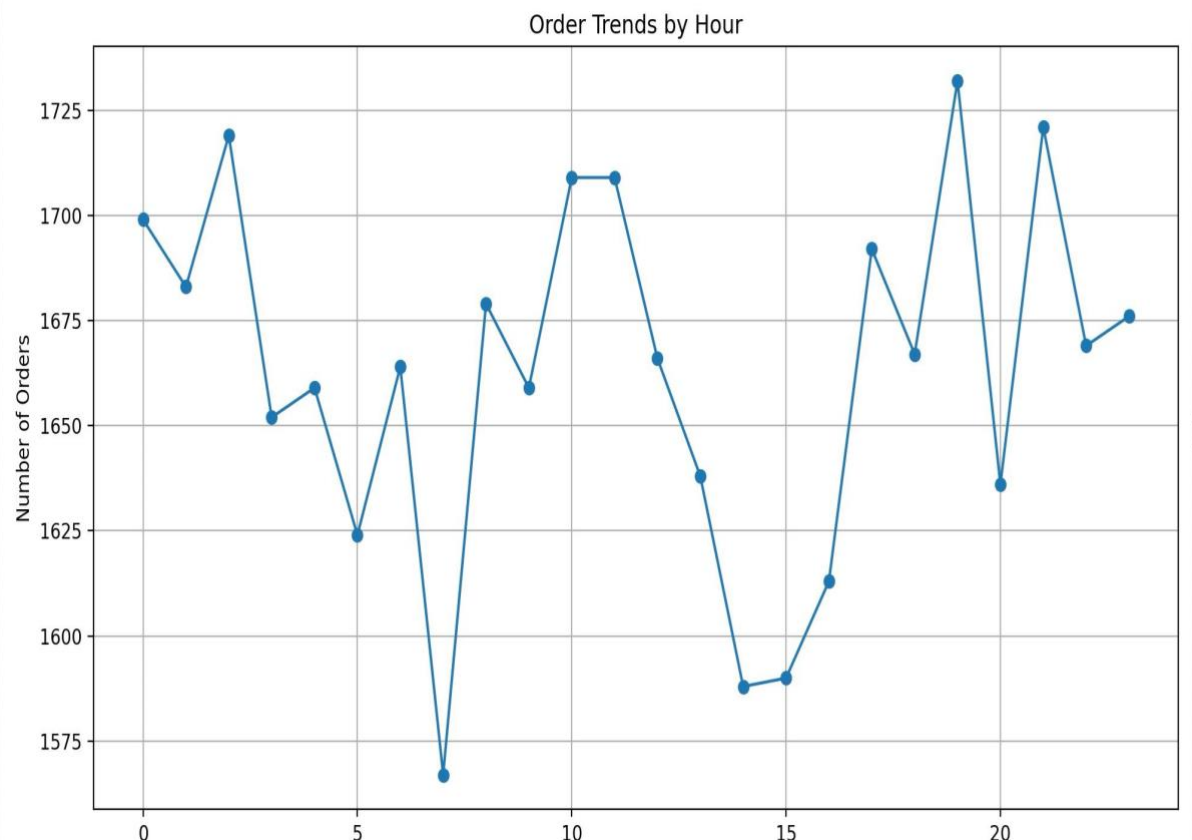
- 1) Convert the “Order Date & Time” column to datetime format.
- 2) Extract the **hour** of order placement from the timestamp.
- 3) Count orders per hour and visualize using a line chart to spot peak periods.

- **Analysis Results:** The majority of orders were placed between **11 AM and 3 PM**, with a clear spike around **1 PM**. Very few orders were placed in the early morning or late night hours. This reveals a mid-day demand concentration, suggesting that operations should be more resource-heavy during this time window. Staffing, delivery scheduling, and promotional activities can be aligned accordingly to boost efficiency and reduce delays.

- **Python Code Used:**

```
df['Hour'] = df['Order Date & Time'].dt.hour
order_trends = df['Hour'].value_counts().sort_index()
plt.figure(figsize=(10, 6))
order_trends.plot(kind='line', marker='o')
plt.title('Order Trends by Hour')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Orders')
plt.grid(True)
plt.tight_layout()
plt.show()
```

- **Visualization:**



## 12.2 Objective 2: Delivery Time Analysis & Delay Impact

- **General Description:**

Timely delivery is a critical factor in customer satisfaction and brand loyalty. This analysis focuses on identifying the typical delivery time range and how delays affect delivery duration. By distinguishing between delayed and on-time orders, the company can assess the operational bottlenecks and improve logistics or route planning strategies to reduce delays and ensure consistent delivery performance.

- **Specific Requirements:**

1) Plot distribution of delivery times.

2) Use a boxplot to compare delivery times between delayed and on-time deliveries.

- **Analysis Results:**

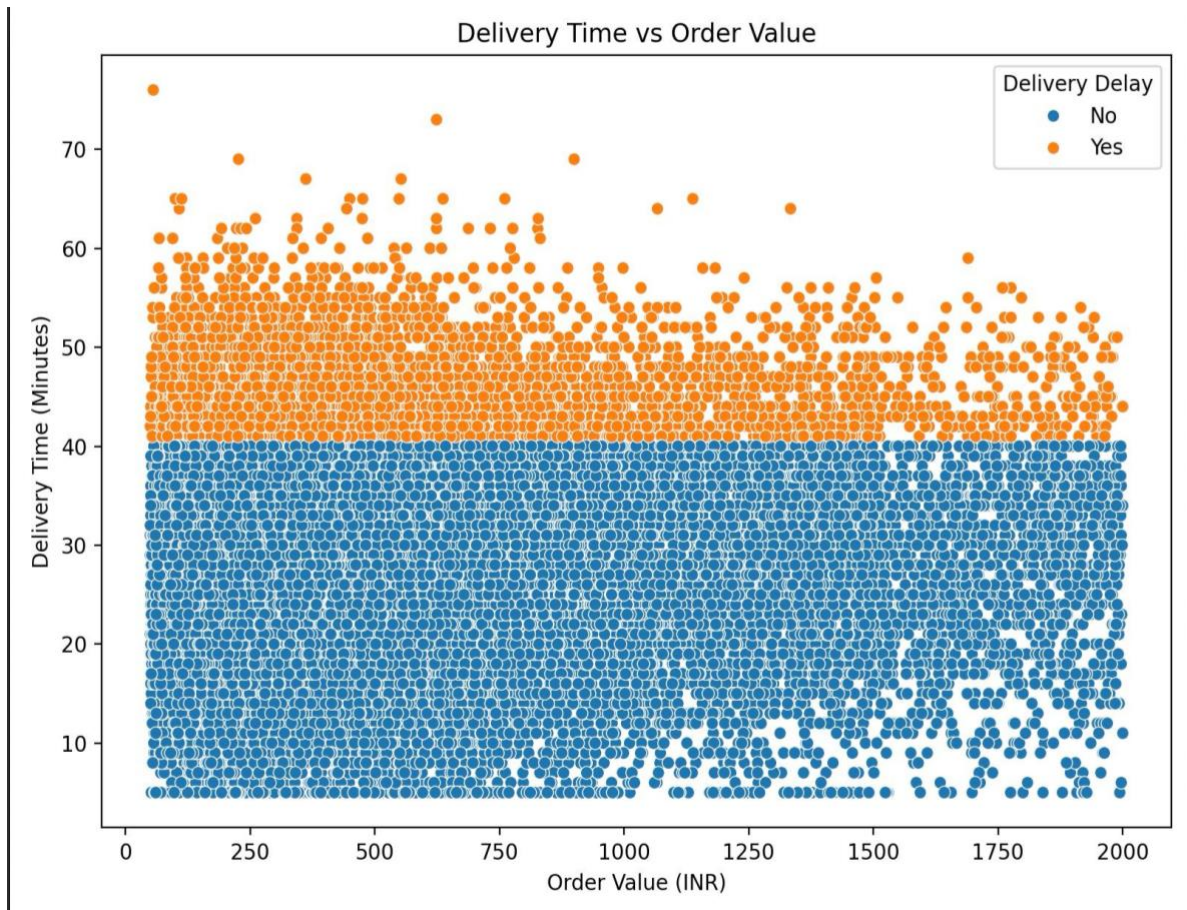
1) Most deliveries fall under **40-60 minutes**.

2) Delayed deliveries tend to have a higher median delivery time compared to on-time deliveries.

- **Python Code Used:**

```
plt.figure(figsize=(10, 6))
sns.histplot(df['Delivery Time (Minutes)'], bins=30, kde=True,color="green")
plt.title('Distribution of Delivery Time')
plt.tight_layout()
plt.show()
```

- **Visualization:**



### 12.3 Objective 3: Customer Satisfaction & Service Ratings

- **General Description:**

Service ratings are a direct reflection of the customer experience. This objective evaluates how customers rate the service and explores whether higher-value orders are associated with better ratings. Understanding this relationship helps identify what drives satisfaction, allowing businesses to refine their customer service, enhance engagement, and focus retention efforts on high-value customers.

- **Specific Requirements:**

Count frequency of service ratings.

Compare order values across different service rating levels using boxplots.

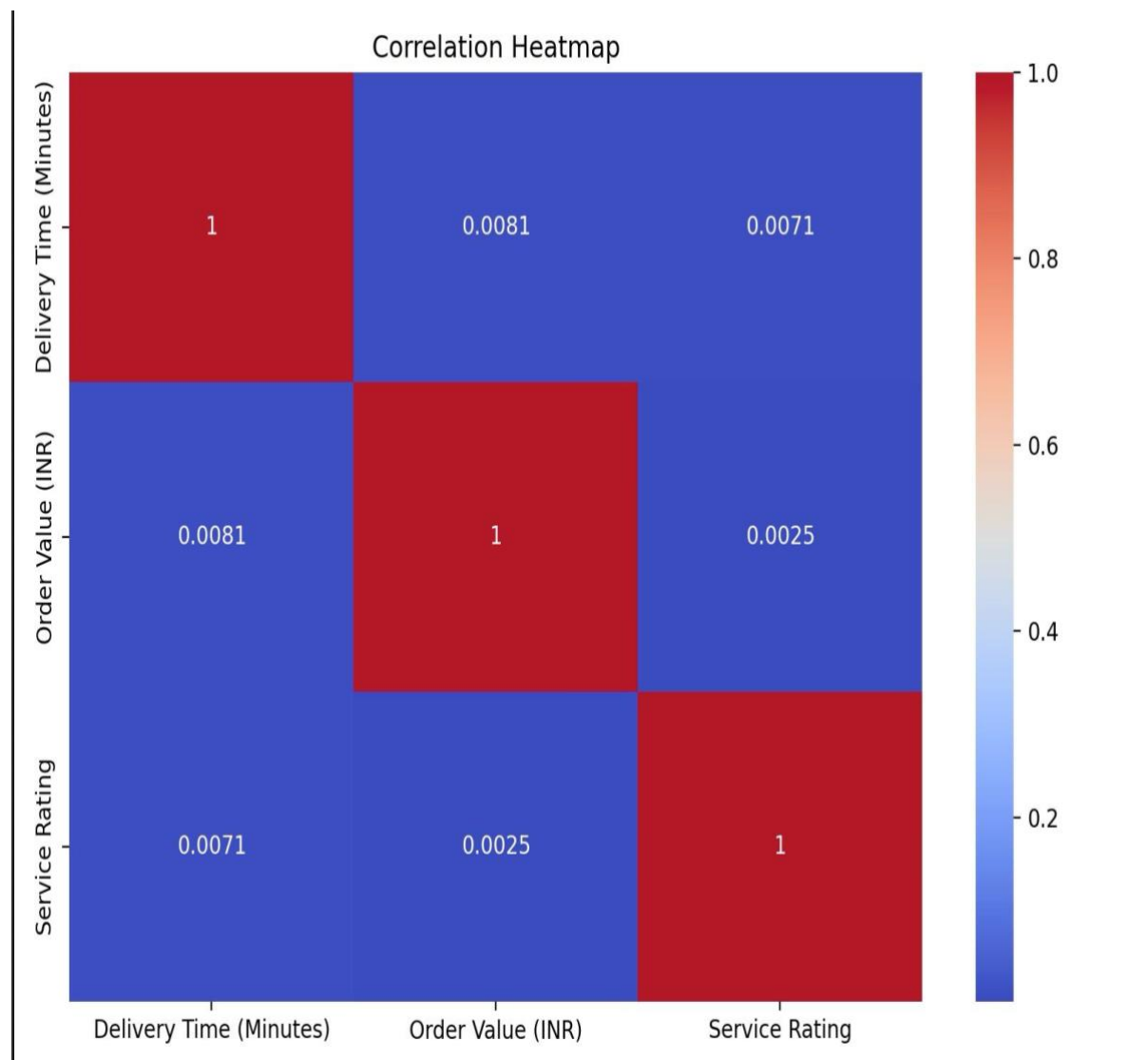
- **Analysis Results:**

Most customers gave **4 or 5-star** ratings, indicating strong satisfaction levels. A positive trend was observed where higher order values received better ratings, suggesting that premium customers are more satisfied. Lower ratings, while fewer, may indicate issues with delivery timing or product quality. Monitoring and responding to low ratings is key to maintaining strong customer relationships.

- **Python Code Used:**

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Service Rating', y='Order Value (INR)',color="purple")
plt.title('Order Value vs Service Rating')
plt.tight_layout()
plt.show()
```

**Visualization:**



## 12.4 Objective 4: Revenue Insights & Top-Selling Products

- **General Description:**

Analyzing which product categories generate the most revenue enables better inventory planning, marketing focus, and product promotion strategies. This objective helps the business determine where its core revenue comes from and assess whether sales are overly dependent on a few product categories. It also provides insights into diversification opportunities and pricing effectiveness.

- **Specific Requirements:**

Group by product category and sum the order value.

Create bar and pie charts to visualize revenue distribution.

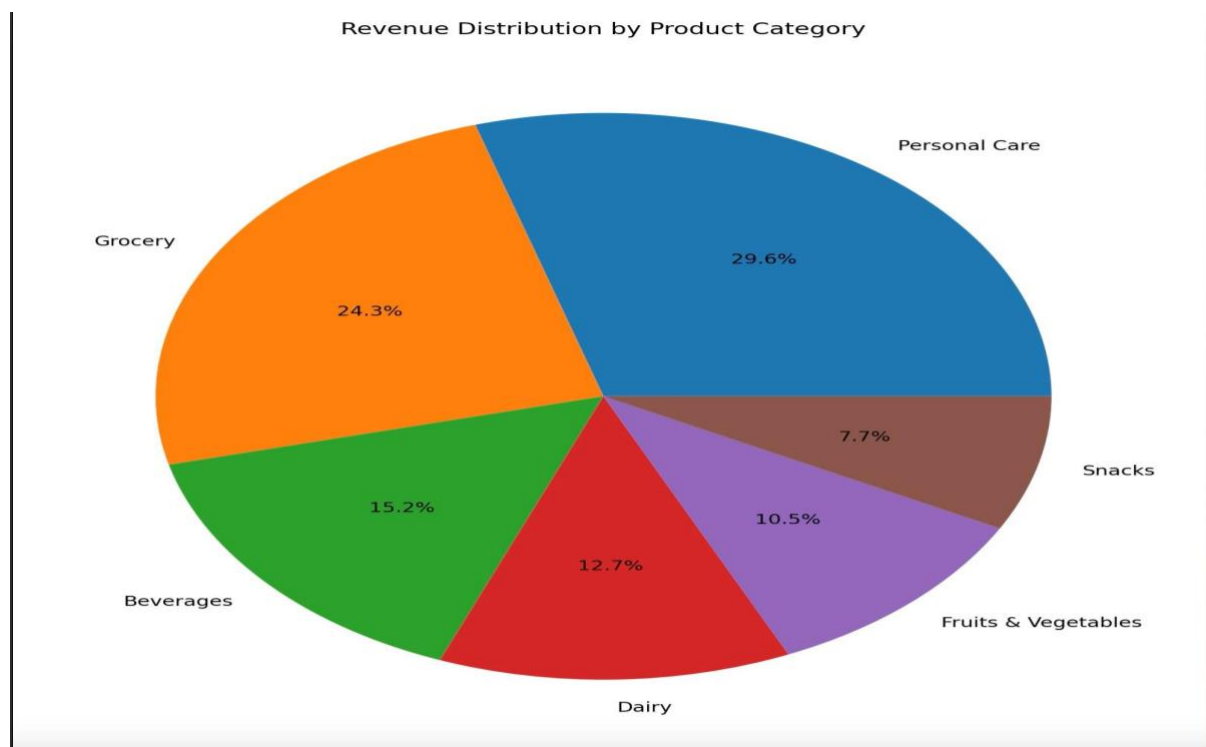
- **Analysis Results:**

A few categories—such as **Food & Beverages** and **Electronics**—generated the majority of revenue. Others had minimal impact. The revenue pie chart confirmed a **skewed distribution**, where a handful of categories dominate. This insight can guide inventory prioritization, promotional campaigns, and resource allocation toward top performers. Underperforming categories may need re-evaluation or repositioning.

**Python Code Used:**

```
plt.figure(figsize=(8, 8))
product_revenue.plot(kind='pie', autopct='% 1.1f%%')
plt.ylabel("")
plt.title('Revenue Distribution by Product Category')
plt.tight_layout()
plt.show()
```

- **Visualization:**



## 12.5 Objective 5: Refund & Complaint Analysis

- **General Description:**

Refund requests are a key indicator of customer dissatisfaction and potential service or product quality issues. This analysis measures the volume of refund requests and can be used to pinpoint systemic problems in delivery, packaging, or customer service. Minimizing refunds not only improves customer retention but also reduces financial losses and operational friction.

- **Specific Requirements:**

Count how many refund requests were made.  
Visualize the data with a bar plot.

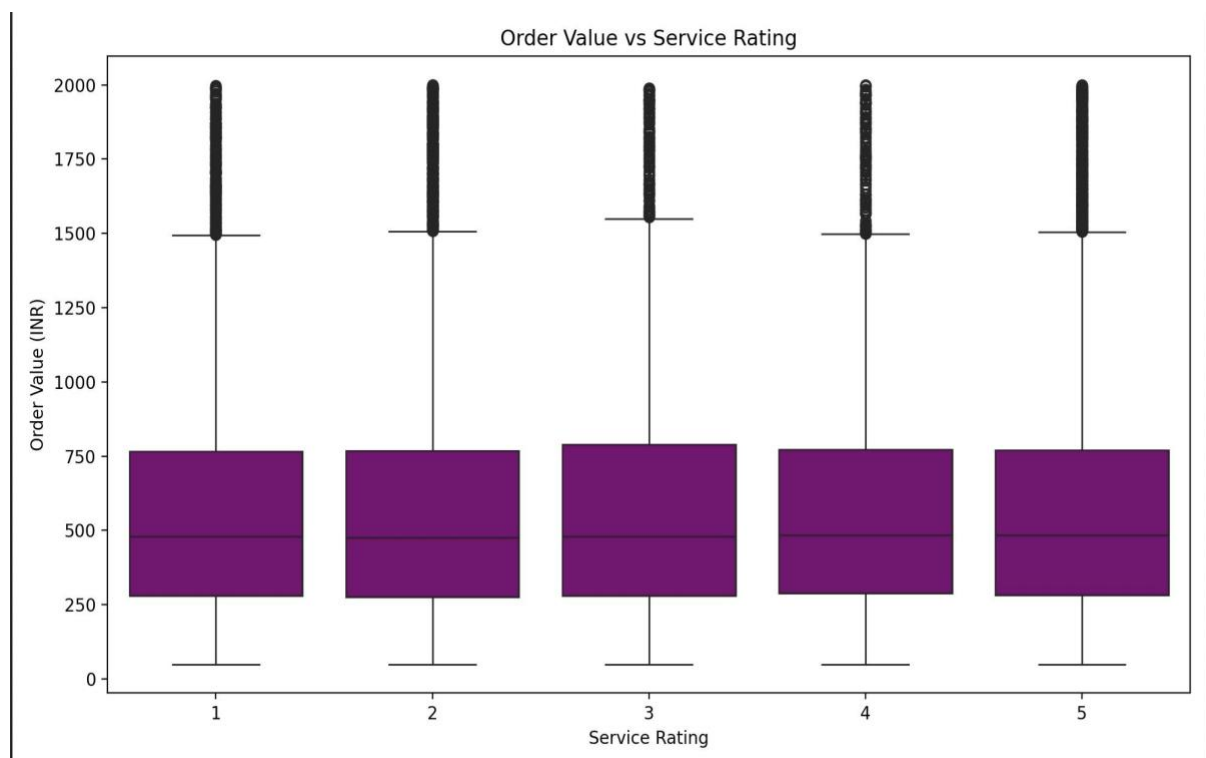
- **Analysis Results:** [?](#)

Refunds were found to be relatively **infrequent**, which suggests that product quality and delivery accuracy are generally high. However, the presence of any refunds still indicates potential weaknesses, possibly in packaging, delivery delays, or incorrect items. Regular monitoring and improvement based on refund trends can help reduce future occurrences and improve customer retention.

- **Python Code Used:**

```
corr = df[['Delivery Time (Minutes)', 'Order Value (INR)', 'Service Rating']].corr()
plt.figure(figsize=(8, 6))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.tight_layout()
plt.show()
```

- **Visualization:**



### 13. Conclusion

This ecommerce delivery analytics initiative yielded valuable insights into customer behavior, operational effectiveness, and service quality overall. The busiest order times were between 11 AM and 3 PM, indicating best windows for resource planning. Delivery delays were associated with much longer delivery

times, underscoring the importance of logistics optimization. The majority of customers expressed high satisfaction, particularly with higher-value orders, and underscored the significance of service consistency. Revenue was focused in a few categories, with growth potential for targeted expansion. Refunds were low but need continuous monitoring. Overall, the analysis validates data-driven strategies to enhance performance, customer satisfaction, and business sustainability.

Key Findings:

1. Order Trends & Peak Hours  
Order Volume Peaks Between 11 AM and 3 PM: Order times analysis showed that customer activity peaks in late morning and early afternoon, with a strong spike at 1 PM. This coincides with normal break times when users are most active online. Low Activity Early Morning & Late Night: Few orders were received prior to 8 AM or later than 10 PM, indicating less demand for staffing or promotion at those times. Operational Implication: Fulfillment personnel, customer service, and delivery teams need to be staffed more intensely in the mid-day peak to facilitate faster turnaround times and less wear on logistics during peak hours.

2. Delivery Time & Delay Impact  
Average Delivery Time is Between 40–60 Minutes: Histogram graphs demonstrated a relatively normal distribution peaking at the 50-minute mark, reflecting uniform performance on the majority of orders. Slow Orders Take A Lot Longer: Boxplot plotting indicated delayed orders result in a significant raise in delivery time, frequently taking over 70–90 minutes with considerable variation and outliers. Implication for Efficiency: Delay patterns likely result from route congestion, high order volumes, or operational inefficiencies. These must be examined further using geospatial and time-based clustering. Customer Experience Risk: Extended delays would harm satisfaction and drive refund or complaint rates up unless proactively addressed.

3. Customer Satisfaction & Service Ratings  
Most Ratings Are 4 or 5 Stars: The data indicated a high skew towards positive service ratings, reflecting good customer experience overall. Positive Order Value vs. Rating Correlation: High-value customers tend to provide higher service ratings, perhaps because they are more satisfied with premium services or have their expectations fulfilled. Actionable Insight: Prioritize maintaining service consistency for high-value customers while looking into low-rating feedback for repeating issues. Retention Strategy: Loyal customers, particularly those with high order values, must be addressed with loyalty schemes, special offers, or first access to promotions.

4. Revenue Insights & Top-Selling Products

Revenue is Dominated by Few Product Categories: Product categories such as Food & Beverages, Electronics, and Personal Care contribute significantly to overall revenue. Unbalanced Sales Distribution: Certain categories yield minimal sales, and it can be inferred that they might require more marketing, bundling, or even exclusion from the catalog. Business Maximization: Spend on inventory management and promotional activities on high-performing categories while assessing underperforming categories for repositioning or elimination.

5. Refund & Complaint Trends  
Refund Requests Are Quite Low: Customers in general do not ask for refunds, which speaks well of service quality and delivery performance. But Refunds Must Be Tracked: Even a few refunds can identify systemic problems like repeated product defects, courier problems, or communication breakdowns. Customer Trust Opportunity: Providing an effortless, no-hassle refund experience can transform a potentially negative experience into a positive brand experience.

## 14. Future Scope

- The current analysis provides valuable insights into customer behavior, delivery performance, and revenue trends. However, there is significant potential to **expand and enhance** the scope of this project for deeper business impact. Below are the key areas for future development:

### 1. Advanced Predictive Analytics

- Implement **machine learning models** to predict:
- Delivery delays based on time, location, traffic, and weather.
- Customer satisfaction based on delivery speed, order value, and product type.
- Use forecasting models (e.g., ARIMA, Prophet) to anticipate **future order volumes** and prepare resource planning accordingly.

### 2. Geospatial Analysis

- Integrate **location data** (latitude/longitude) for:
- Route optimization.
- Identifying high-demand delivery zones.
- Analyzing geographic trends in product popularity and delays.

### 3. Customer Segmentation

- Use clustering algorithms (e.g., K-Means) to segment customers by:
- Spending habits.
- Order frequency.
- Satisfaction level.



- Personalize promotions, loyalty rewards, and customer service based on segment profiles.

#### **4. Sentiment Analysis**

- Incorporate **customer feedback or reviews** (if available) and apply NLP (Natural Language Processing) techniques to:
- Analyze customer sentiment.
- Identify recurring complaints or compliments.
- Improve products and services based on real customer voices.

#### **5. Product Performance Tracking**

- Monitor **real-time performance** of each product category:
- Compare return rates, ratings, and profit margins.
- Use dashboards to track top/bottom performers.
- Introduce **dynamic pricing strategies** based on demand trends.

#### **6. Refund & Complaint Prediction**

- Build models to **predict refund likelihood** based on order history, delivery delays, and customer profile.
- Help customer support teams proactively manage high-risk orders and improve response time.

#### **7. Real-Time Dashboards**

- Create interactive dashboards using **Power BI or Tableau** to:
- Track KPIs in real time.
- Monitor order trends, delay patterns, and refund rates.
- Allow different departments (logistics, marketing, support) to make fast, data-driven decisions.

#### **8. Integration with External Data**

- Enrich analysis by integrating with:
- **Traffic data** for better delivery predictions.
- **Weather APIs** to understand delay causes.
- **Social media data** to monitor public sentiment and trends.

#### **9. Data Security and Ethics**

- Ensure customer data privacy and comply with regulations like GDPR.
- Implement role-based access and encrypted storage for sensitive information.

By expanding in these areas, the project can move from **descriptive analytics** to **predictive and prescriptive intelligence**, helping the business stay competitive, agile, and highly customer-focused.

## 15. References

- E-Commerce Analytics <https://www.data.gov.in/>
- Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
- Statistical Techniques: Correlation, Z-score, IQR method