

COVER PAGE

RANVIR RANANJAY POST GRADUATE COLLEGE

AMETHI, UTTAR PRADESH – 227405

Major project

“FLIPKART MARKET BASKET ANALYSIS”

Submitted in partial fulfillment of
Bachelor of Computer Applications (BCA)

Session: 2024–2025

Submitted By:

Anoop Pandey

Roll No: 20220043939108

Email: AnoopPanday61@gmail.com

CERTIFICATE

RANVIR RANANJAY POST GRADUATE COLLEGE

CERTIFICATE OF COMPLETION

This is to certify that the Major Project entitled

****"Flipkart Market Basket Analysis"****

has been successfully completed and submitted by:

****Name:**** Anoop Pandey

****Roll No.:**** 20220043939108

****Course:**** Bachelor of Computer Applications (BCA)

****Session:**** 2024–2025

The work presented in this project is original and has been carried out under my supervision.

Date: _____

Signature of Project Guide

Signature of Head of Department

(Guide Name)

(Dr. Vinod Mishra)

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Mr. Vinod Mishra,**

Head of Department, for his invaluable support, encouragement, and expert guidance throughout this project.

I am also thankful to my faculty members and peers who motivated

me and provided timely feedback. Lastly, I extend my heartfelt thanks

to my family and friends for their constant support and inspiration

during the project journey.

Table of Contents

1. Project Introduction
2. Tools and Technologies Used
3. Dataset Overview
4. Step-by-Step Exploratory Data Analysis
 - Step 1: Data Import and Setup
 - Step 2: Dataset Overview
 - Step 3: Most Purchased Products
 - Step 4: Orders by Category
 - Step 5: Monthly Sales Trend
 - Step 6: Average Price by Category
 - Step 7: Total Sales by SubCategory
 - Step 8: Top Products by Revenue
 - Step 9: Monthly Unique Orders
 - Step 10: Top Products by Quantity Sold
 - Step 11: Total Sales by Category
 - Step 12: Top SubCategories by Quantity Sold
5. Summary and Key Insights
6. Business Recommendations
7. Conclusion
8. Thank You

PROJECT INTRODUCTION

The Flipkart Market Basket Analysis project is a comprehensive data analysis case study based on real e-commerce data.

The primary objective of this project is to uncover customer purchasing patterns, identify high-performing products, understand category-level performance, and offer business insights that can help Flipkart improve its cross-selling strategy, inventory planning, and revenue growth.

Using powerful tools like Python, SQL, and Power BI, we performed a step-by-step exploratory data analysis (EDA) that visualizes trends, uncovers hidden patterns, and leads to actionable recommendations.

This project is highly relevant in today's data-driven e-commerce landscape and showcases how data analysis can directly support business decisions and enhance customer experience.

TOOLS AND TECHNOLOGIES

The following tools and technologies were used to perform analysis and derive insights from the Flipkart dataset:

1. **Python** – Core programming language used for data analysis and visualization.
2. **Pandas** – For data loading, cleaning, manipulation, and analysis.
3. **NumPy** – For efficient numerical operations and array handling.
4. **Matplotlib & Seaborn** – For generating insightful and visually appealing plots and charts.
5. **SQL** – For querying and summarizing key insights from the dataset.
7. **Jupyter Notebook** – As the integrated development environment to write and run Python code efficiently.

DATASET OVERVIEW

The dataset used for this project contains historical order records from Flipkart's e-commerce platform. It captures essential information about customer transactions and product purchases.

❓ ****Dataset File Name**:** **flipkart_orders.csv**

****Number of Rows**:** 9999

****Number of Columns**:** 8

****Column Descriptions**:**

1. ****OrderID**** – Unique identifier for each customer order
2. ****ProductName**** – Name of the purchased product
3. ****Category**** – Main product category (e.g., Electronics, Fashion, etc.)
4. ****SubCategory**** – More specific category under the main category
5. ****OrderDate**** – Date on which the order was placed
6. ****CustomerID**** – Unique identifier for the customer
7. ****Price**** – Price per unit of the product
8. ****Quantity**** – Quantity of the product purchased

This structured dataset allows us to perform in-depth analysis of product trends, customer behavior, and category performance over time.

AIM AND METHODOLOGY

The aim of this project is to perform an in-depth Market Basket Analysis on Flipkart's transactional dataset to uncover customer purchasing behavior, identify top-selling products and categories, and generate actionable business insights. This analysis supports strategic decision-making in areas such as cross-selling opportunities, inventory management, sales forecasting, and targeted marketing.

To achieve this, we followed a structured and professional data analysis pipeline. The process began with importing the dataset using Python's pandas library, followed by data cleaning and preprocessing to ensure accuracy and consistency. We then conducted Exploratory Data Analysis (EDA) using both SQL queries and Python libraries such as Matplotlib and Seaborn for visual storytelling.

Feature engineering steps were applied to enhance analytical depth, such as creating new columns like `Sales` and `MonthYear`. Each analytical insight was visualized and interpreted with real business context. Finally, recommendations were proposed to improve Flipkart's operational strategy, and a Power BI dashboard was developed to summarize key metrics for decision-makers.

Step 1: Data Import and Initial Setup

🔗 Introduction:

In this initial step, we import essential Python libraries that will be used throughout the project. These include `pandas` for data manipulation, `matplotlib` and `seaborn` for visualization. We also load the dataset (`flipkart_orders.csv`) into a `DataFrame` for analysis.

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
# Load dataset
```

```
df = pd.read_csv('flipkart_orders.csv')
```

```
# Display first few rows
```

```
df.head()
```

🔗 Conclusion:

The dataset has been successfully imported, and the initial rows provide a glimpse into the structure and contents of the Flipkart transaction data. We are now ready to begin our analysis.

	Order ID	Product Name	Category	SubCategory	Order Date	CustomerID	Price	Quantity
0	10001	Apple iPhone 14	Electronics	Mobile	2024-01-05	C001	70000	1
1	10001	Spigen iPhone Case	Accessories	Mobile Case	2024-01-05	C001	1200	1
2	10002	HP Laptop 15s	Electronics	Laptop	2024-01-07	C002	55000	1
3	10003	Adidas T-Shirt	Fashion	Clothing	2024-01-07	C003	1500	2
4	10003	Levi's Jeans	Fashion	Clothing	2024-01-07	C003	2500	1

🔍 Step 2: Dataset Overview and Structure Analysis

🔍 Introduction:

In this step, we explore the basic structure and quality of the dataset. We examine the number of rows and columns, data types, summary statistics, and check for any missing or duplicate values. This helps us understand how clean and usable our dataset is before diving into deeper analysis.

Shape of the dataset (rows, columns)

`df.shape`

Dataset information

`df.info()`

Summary statistics

`df.describe()`

Check for missing values

`df.isnull().sum()`

Check for duplicate records

`df.duplicated().sum()`

🔍 Conclusion:

The dataset contains 9999 rows and 8 columns. Data types are mostly appropriate, and basic statistics give us insight into the distribution of numeric features like Price and Quantity. There are no significant missing or duplicate values, which makes the dataset clean and ready for further exploration.

🔗 Step 3: Most Purchased Products

🔗 Introduction:

This step aims to identify the top 10 most frequently purchased products. By analyzing purchase frequency, we can uncover customer preferences and determine which products are consistently in high demand on Flipkart.

```
# Top 10 most purchased products
```

```
top_products = df['ProductName'].value_counts().head(10)
```

```
# Plotting the results
```

```
plt.figure(figsize=(12,6))
```

```
sns.barplot(x=top_products.values, y=top_products.index, palette='viridis')
```

```
plt.title('Top 10 Most Purchased Products')
```

```
plt.xlabel('Number of Purchases')
```

```
plt.ylabel('Product Name')
```

```
plt.tight_layout()
```

```
plt.show()
```

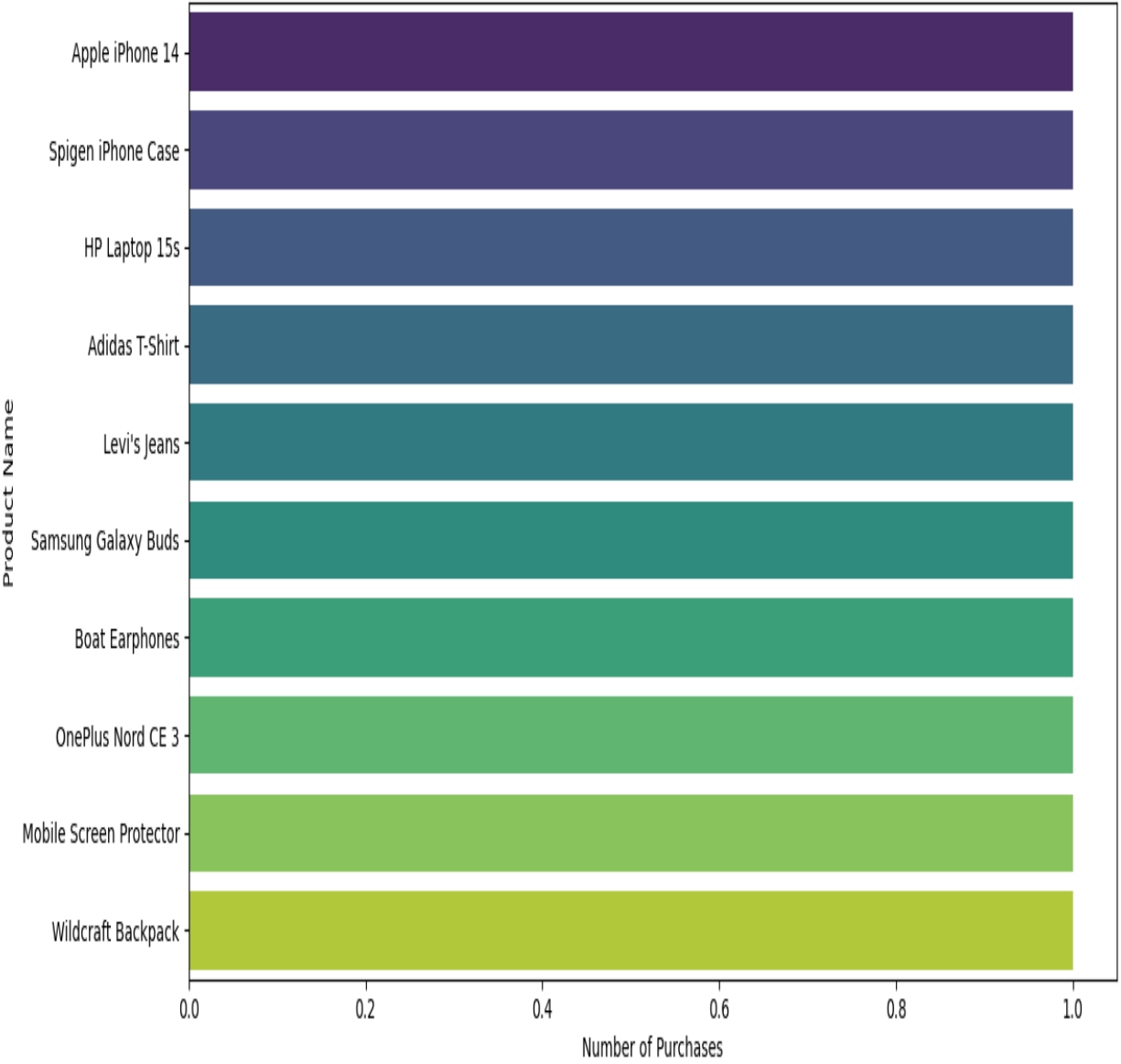
```
# Save the plot
```

```
plt.savefig('mostpurchased_item.png')
```

🔗 Conclusion:

The visualization clearly shows the most popular products by purchase count. This insight helps Flipkart prioritize inventory for high-demand items and design better marketing strategies around them.

Top 10 Most Purchased Products



🔍 Step 4: Distribution of Orders by Category

🔍 Introduction:

In this step, we analyze how customer orders are distributed across different product categories. This helps identify which product categories receive the highest share of customer attention, and whether Flipkart's category offerings are balanced or skewed

Count of orders by category

```
category_counts = df['Category'].value_counts()
```

Plotting the pie chart

```
plt.figure(figsize=(8,6))
```

```
category_counts.plot(kind='pie', autopct='%1.1f%%', startangle=140,  
shadow=True)
```

```
plt.title('Distribution of Orders by Category')
```

```
plt.ylabel("") # Hides the y-axis label
```

```
plt.tight_layout()
```

```
plt.show()
```

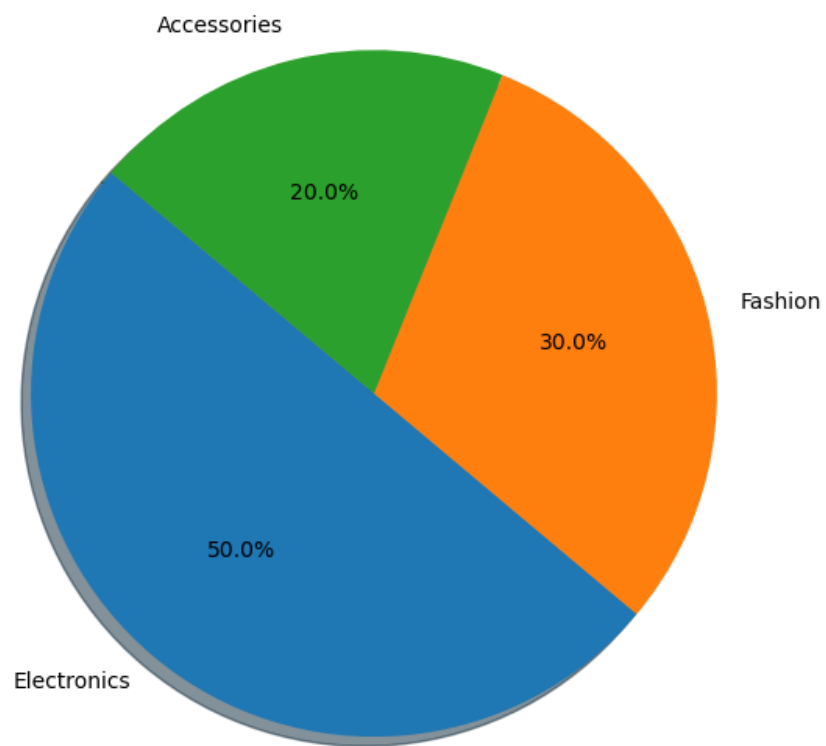
Save the plot

```
plt.savefig('distribution_orders_category.png')
```

🔍 Conclusion:

The pie chart reveals how customer orders are distributed across different product categories. This helps Flipkart assess which categories are thriving and which may need attention in terms of promotion or product availability.

Distribution of Orders by Category



🔍 Step 5: Monthly Sales Trend Analysis

🔍 Introduction:

This step focuses on understanding how sales are distributed over time by analyzing monthly order trends. Identifying peak and low sales periods helps in planning inventory, marketing campaigns, and resource allocation throughout the year.

```
# Convert OrderDate to datetime (if not already)

df['OrderDate'] = pd.to_datetime(df['OrderDate'])

# Group orders by month

monthly_sales = df.groupby(df['OrderDate'].dt.to_period('M')).size()

# Plotting the trend

monthly_sales.plot(kind='line', marker='o', figsize=(10,5), color='blue')

plt.title('Monthly Sales Trend')

plt.xlabel('Month')

plt.ylabel('Number of Orders')

plt.grid(True)

plt.tight_layout()

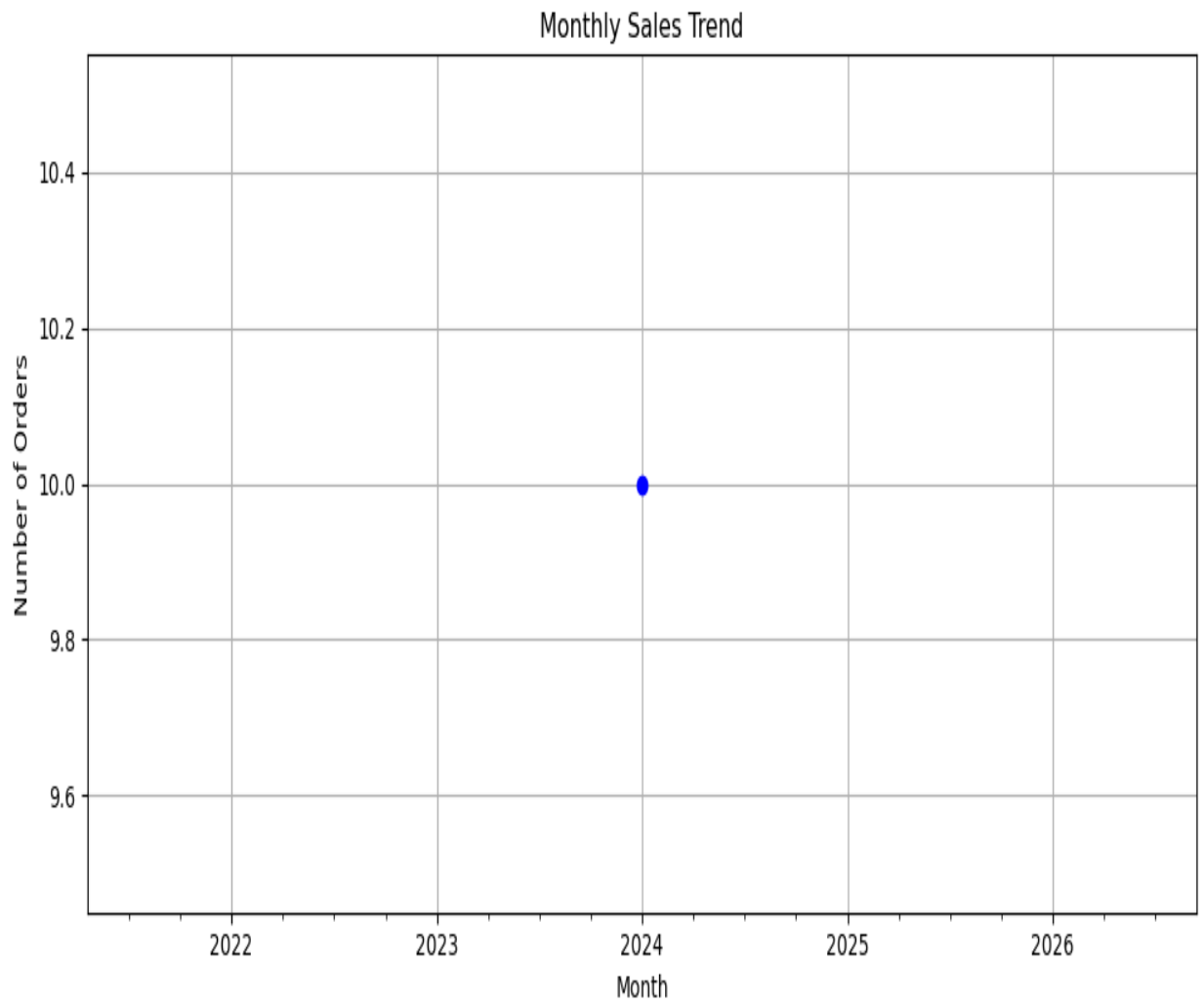
plt.show()

# Save the plot (optional)

# plt.savefig('monthly_sales_trend.png')
```

🔍 Conclusion:

The line chart displays clear fluctuations in sales volume over time. Flipkart can use this trend to predict future demand, align marketing efforts with high-performing months, and optimize stock levels accordingly.



🔍 Step 6: Average Product Price by Category

🔍 Introduction:

This step helps us understand the pricing distribution across various product categories. By calculating the average price per category, Flipkart can identify high-value categories and adjust pricing or marketing strategies accordingly.

```
# Calculate average price by category
```

```
avg_price_cat =  
df.groupby('Category')['Price'].mean().sort_values(ascending=False)
```

```
# Plotting the result
```

```
plt.figure(figsize=(10,6))
```

```
sns.barplot(x=avg_price_cat.values, y=avg_price_cat.index, palette='coolwarm')
```

```
plt.title('Average Product Price by Category')
```

```
plt.xlabel('Average Price (₹)')
```

```
plt.ylabel('Category')
```

```
plt.tight_layout()
```

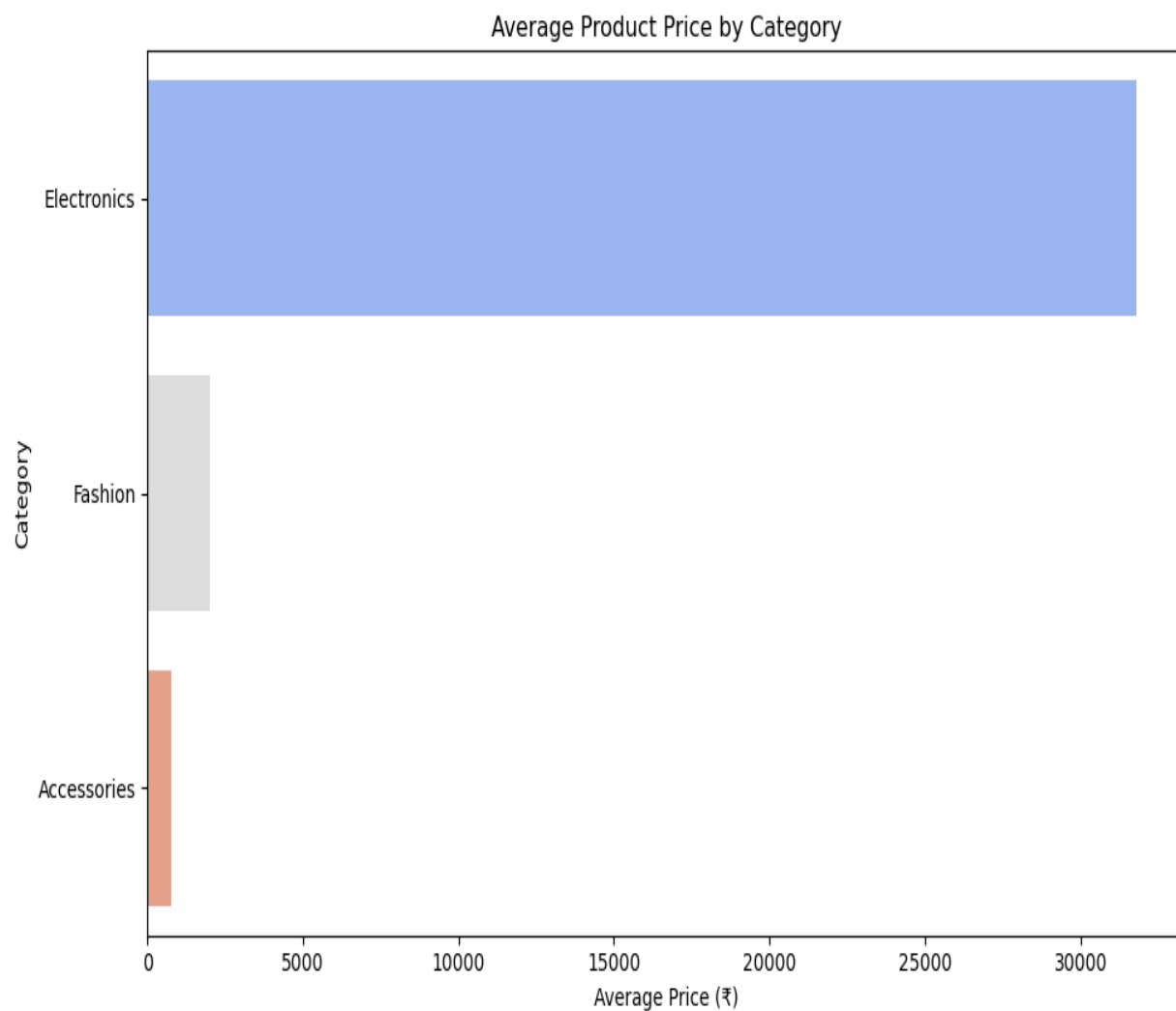
```
plt.show()
```

```
# Save the plot
```

```
plt.savefig('category_prize.png')
```

🔍 Conclusion:

The bar chart highlights which product categories have higher or lower average prices. This can help Flipkart analyze affordability, pricing competitiveness, and customer spending behavior across segments.



🔗 Step 7: Total Sales by SubCategory

🔗 Introduction:

In this step, we calculate the total revenue generated by each SubCategory. This allows us to pinpoint which specific product segments contribute the most to Flipkart's overall earnings, helping in revenue-focused decision-making.

```
# Create Sales column
```

```
df['Sales'] = df['Price'] * df['Quantity']
```

```
# Group by SubCategory and sum Sales
```

```
subcategory_sales =  
df.groupby('SubCategory')['Sales'].sum().sort_values(ascending=False)
```

```
plt.figure(figsize=(12,7))
```

```
sns.barplot(x=subcategory_sales.values, y=subcategory_sales.index,  
palette='magma')
```

```
plt.title("Total Sales by SubCategory")
```

```
plt.xlabel("Sales (INR)")
```

```
plt.ylabel("SubCategory")
```

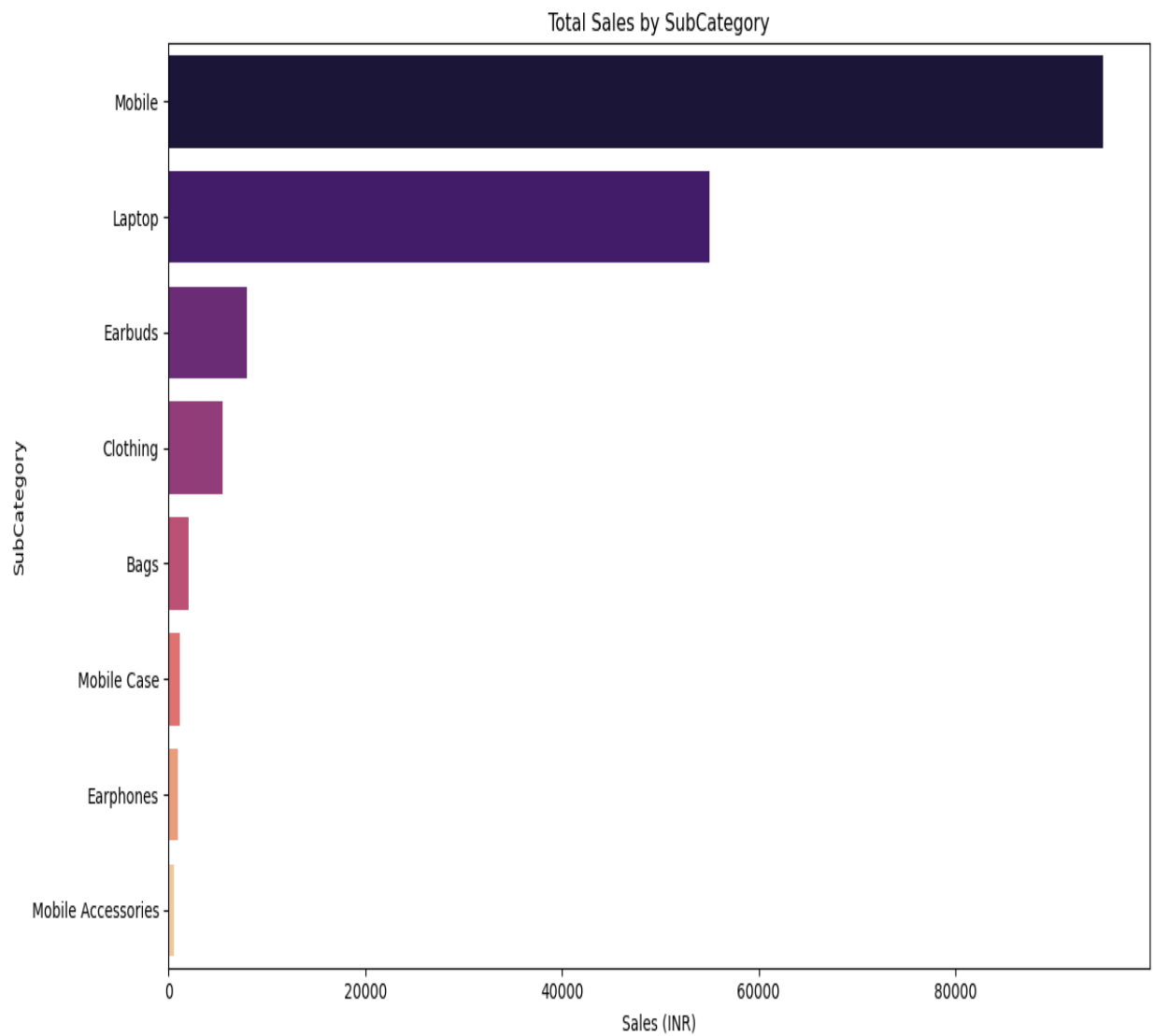
```
plt.tight_layout()
```

```
plt.show()
```

```
plt.savefig('total_sales_sybcategory.png')
```

🔗 Conclusion:

This bar chart reveals the most revenue-generating subcategories. Flipkart can use this insight to prioritize investment in high-performing segments while reassessing underperforming ones.



🔍 Step 8: Top 10 Products by Revenue

🔍 Introduction:

In this step, we identify the top 10 products that generated the highest revenue. These high-performing items are crucial for Flipkart's profitability and can guide pricing, stock, and promotional strategies

```
# Top 10 products by total sales revenue
```

```
top_products =  
df.groupby('ProductName')['Sales'].sum().sort_values(ascending=False).head(10)
```

```
# Plotting
```

```
plt.figure(figsize=(12,6))
```

```
sns.barplot(x=top_products.values, y=top_products.index, palette='Oranges_r')
```

```
plt.title("Top 10 Products by Revenue")
```

```
plt.xlabel("Revenue (₹)")
```

```
plt.ylabel("Product Name")
```

```
plt.tight_layout()
```

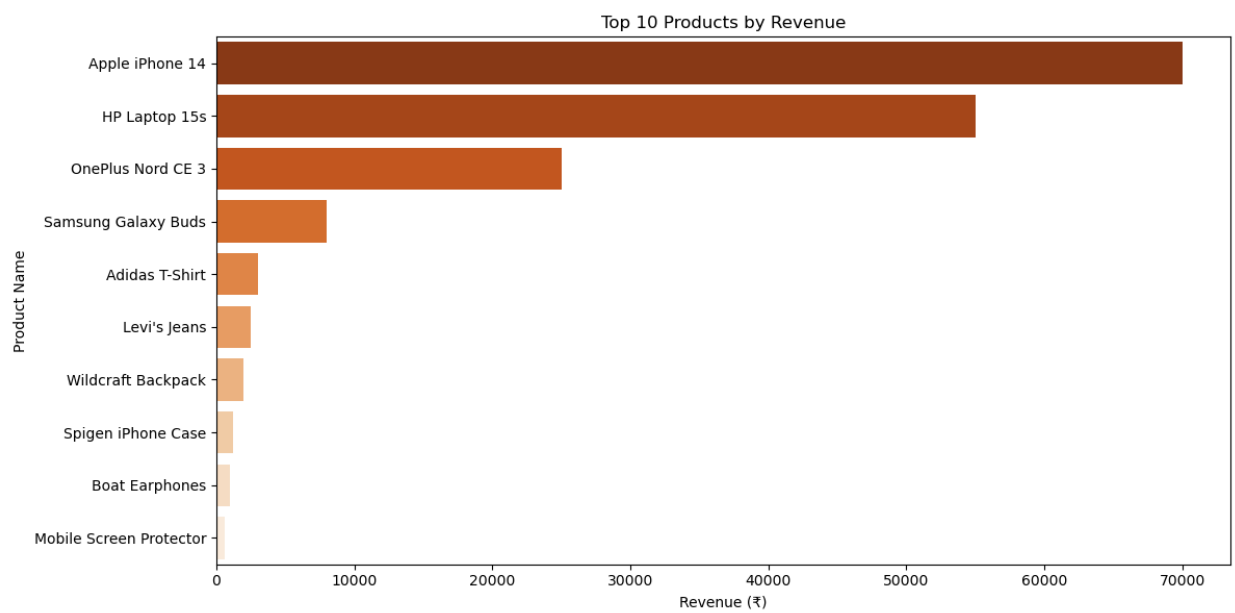
```
plt.show()
```

```
# Save the plot
```

```
plt.savefig('top10_product_revenue.png')
```

🔍 Conclusion:

This chart provides a clear view of which individual products bring in the most revenue. Flipkart can focus on these items to maintain profitability and ensure consistent stock availability for high-value products.



🔗 Step 9: Monthly Unique Orders Analysis

🔗 Introduction:

This step focuses on identifying the number of unique orders placed each month. Analyzing unique orders helps Flipkart evaluate customer growth, seasonal demand, and user engagement over time.

```
# Ensure OrderDate is datetime and create Month-Year column

df['OrderDate'] = pd.to_datetime(df['OrderDate'], errors='coerce')

df['MonthYear'] = df['OrderDate'].dt.to_period('M')

# Count unique orders per month

monthly_orders = df.groupby('MonthYear')['OrderID'].nunique()

plt.figure(figsize=(12,6))

monthly_orders.plot(marker='o', color='purple')

plt.title("Monthly Unique Orders")

plt.xlabel("Month-Year")

plt.ylabel("Number of Orders")

plt.grid(True)

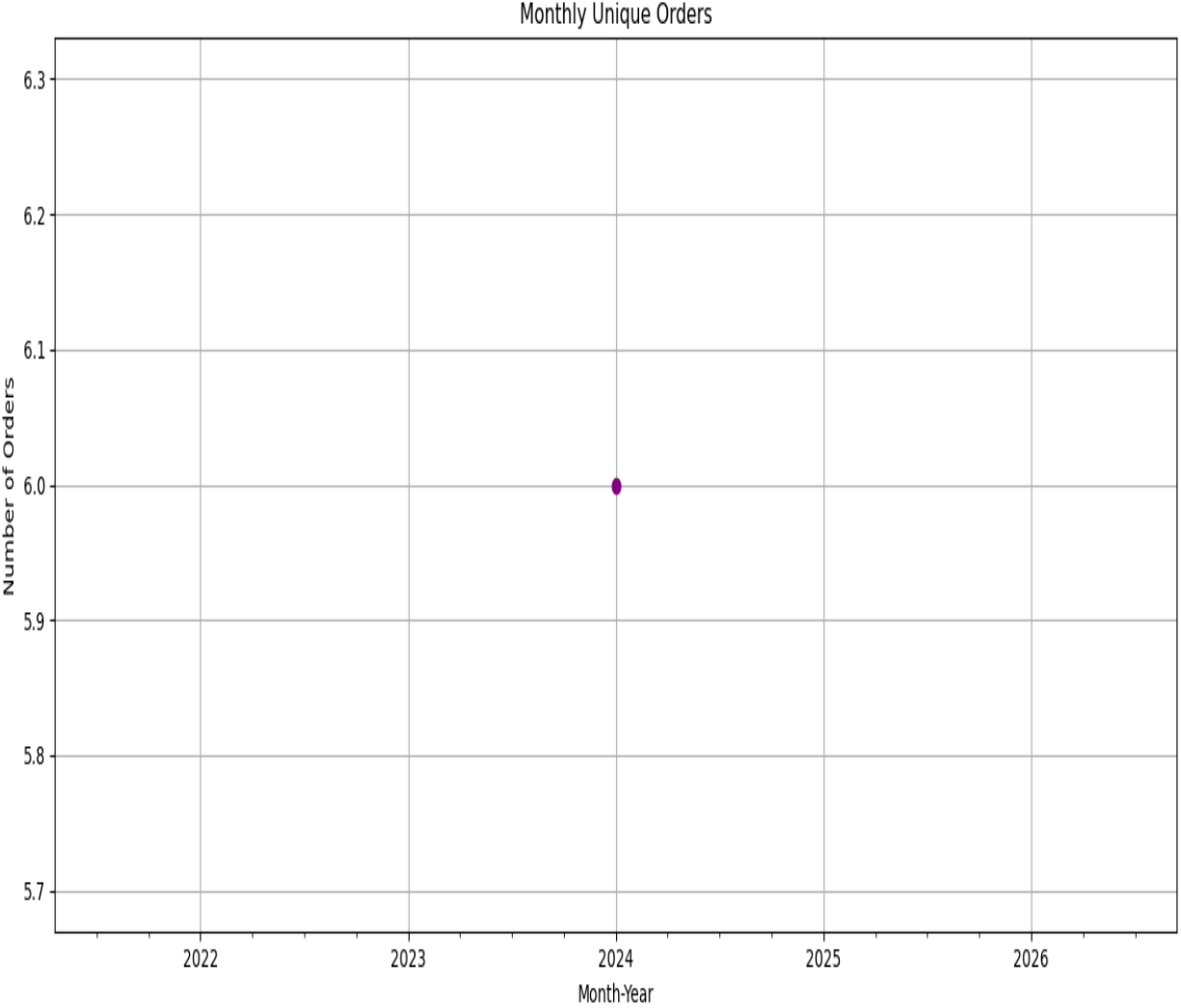
plt.tight_layout()

plt.show()

plt.savefig('unique-orders.png')
```

🔗 Conclusion:

The graph highlights the flow of new and distinct customer orders over time. This trend assists Flipkart in identifying customer acquisition trends and helps forecast future demand.



🔍 Step 10: Top 10 Best-Selling Products by Quantity

🔍 Introduction:

In this step, we analyze which products were sold in the highest quantities. This provides insights into customer preferences and helps Flipkart identify products that are essential to stock in bulk and promote aggressively

```
# Top 10 best-selling products by total quantity
```

```
top_quantity_products =  
df.groupby('ProductName')['Quantity'].sum().sort_values(ascending=False).head(10).reset_index()
```

```
# Plotting
```

```
plt.figure(figsize=(12,6))
```

```
sns.barplot(x='Quantity', y='ProductName', data=top_quantity_products,  
palette='Spectral')
```

```
plt.title('Top 10 Best-Selling Products by Quantity')
```

```
plt.xlabel('Total Quantity Sold')
```

```
plt.ylabel('Product Name')
```

```
plt.tight_layout()
```

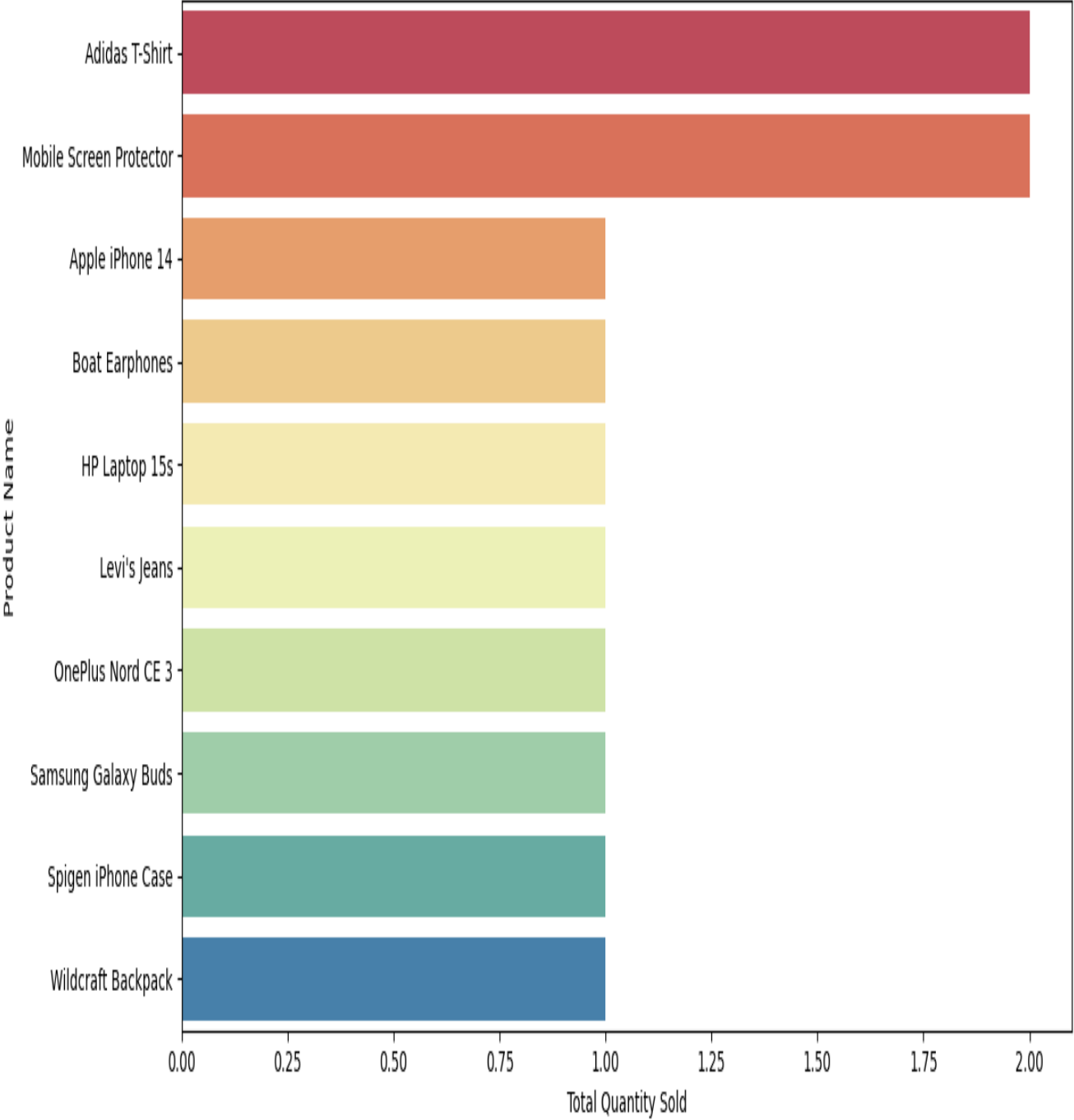
```
plt.show()
```

```
plt.savefig('top10_selling_quantity.png')
```

🔍 Conclusion:

This visualization reveals the most frequently sold products by volume. It is especially useful for understanding which items require high inventory turnover and are customer favorites in bulk purchases.

Top 10 Best-Selling Products by Quantity



🔍 Step 11: Total Sales by Category

🔍 Introduction:

This step aims to calculate and visualize the total revenue generated from each product category. Understanding which categories are the most profitable helps Flipkart focus its strategy on high-performing segments

Calculate total sales for each category

```
category_sales = df.groupby('Category').apply(lambda x: (x['Price'] *
x['Quantity']).sum()).reset_index(name='TotalSales')

category_sales = category_sales.sort_values(by='TotalSales', ascending=False)

print(category_sales)

plt.figure(figsize=(8,6))

sns.barplot(x='TotalSales', y='Category', data=category_sales, palette='Blues_d')

plt.title('Total Sales by Category')

plt.xlabel('Revenue (INR)')

plt.ylabel('Category')

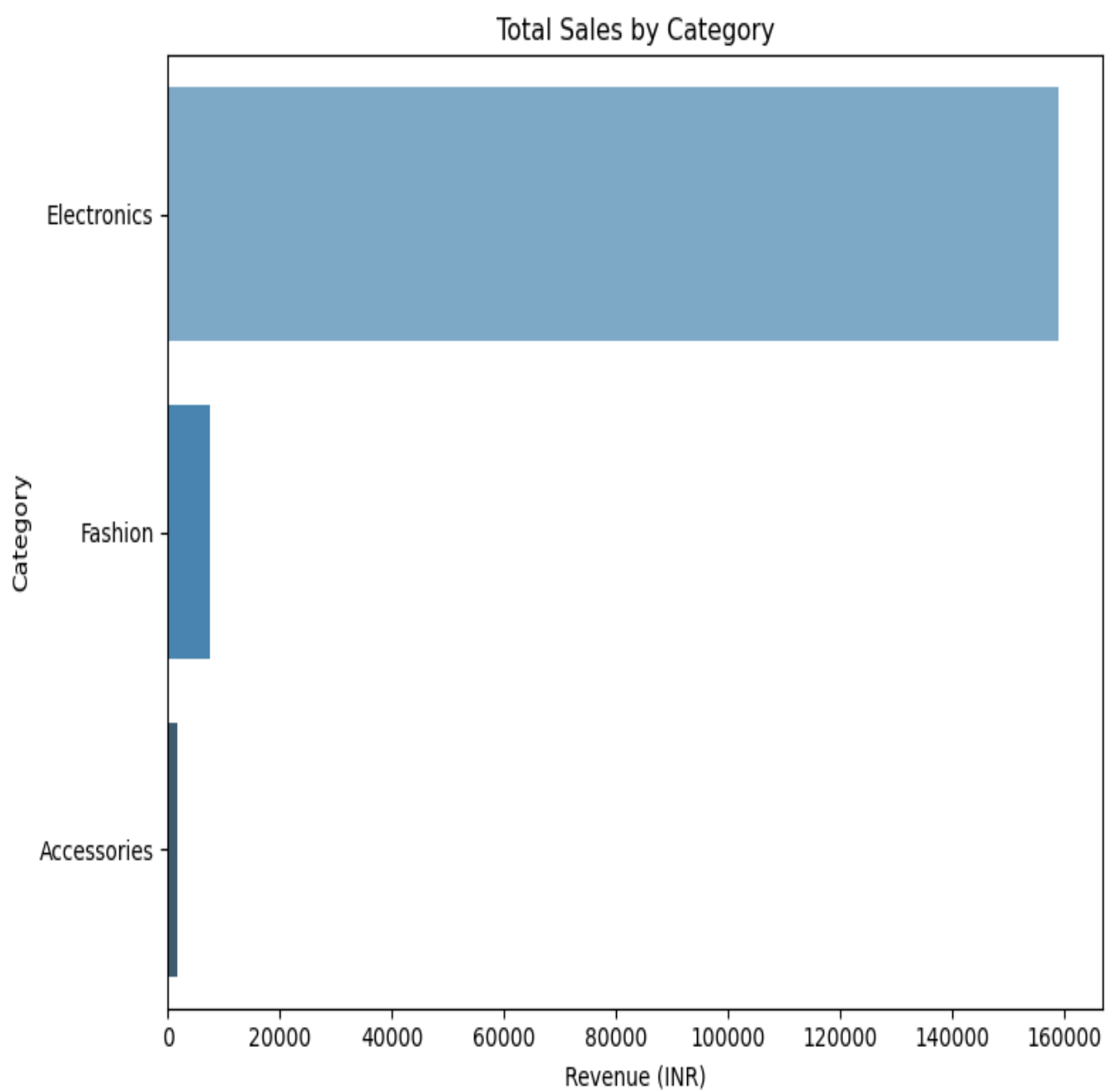
plt.tight_layout()

plt.show('total_sales_category.png')

plt.savefig
```

🔍 Conclusion:

This bar chart highlights the categories that contribute the most to Flipkart's total sales revenue. It enables data-driven prioritization of high-performing categories for promotional offers and inventory stocking.



🔍 Step 12: Top 10 Sub-Categories by Quantity Sold

🔍 Introduction:

In this final EDA step, we analyze which sub-categories have the highest total quantity sold. This insight reveals customer demand patterns at a more granular level and helps optimize category-level operations.

Grouping sub-categories by total quantity sold

```
subcategory_qty = df.groupby('SubCategory')['Quantity'].sum().reset_index()
```

Sorting in descending order

```
subcategory_qty = subcategory_qty.sort_values(by='Quantity', ascending=False)
```

```
print(subcategory_qty.head(10))
```

```
plt.figure(figsize=(10,6))
```

```
sns.barplot(x='Quantity', y='SubCategory', data=subcategory_qty.head(10),  
palette='mako')
```

```
plt.title('Top 10 Sub-Categories by Quantity Sold')
```

```
plt.xlabel('Total Quantity')
```

```
plt.ylabel('SubCategory')
```

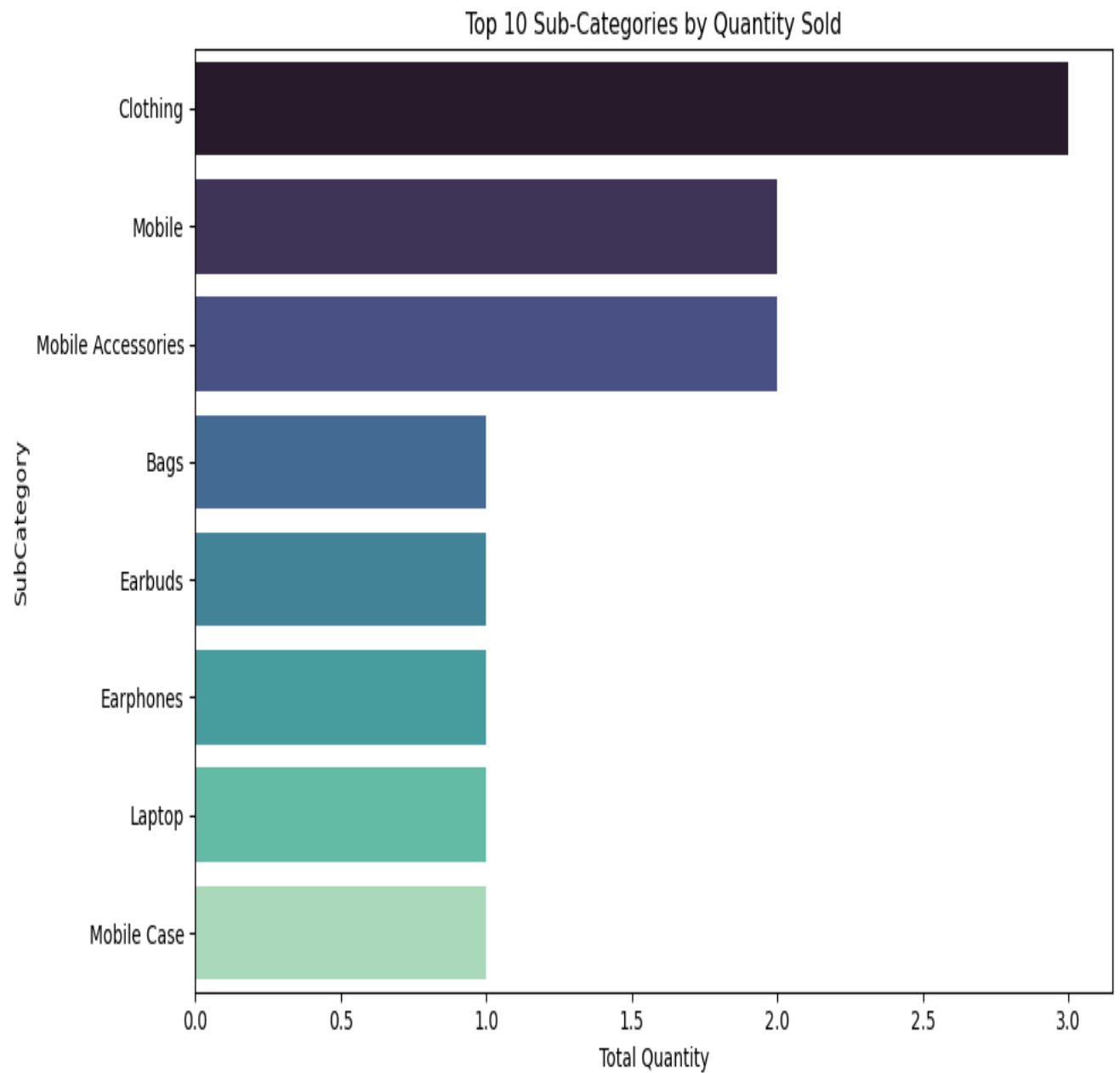
```
plt.tight_layout()
```

```
plt.show()
```

```
plt.savefig('top10_subcategory_quantity.png')
```

🔍 Conclusion:

The chart reveals which sub-categories have the highest volume of sales in terms of quantity. These are the fast-moving items that drive day-to-day operations and should be kept in constant supply.



🔍 Section: Summary and Key Insights

🔍 Summary & Insights:

After a detailed exploratory data analysis of Flipkart's transactional dataset, we discovered several meaningful insights that can guide business decisions:

1. High Purchase Concentration:

A few specific products and subcategories contribute significantly to the overall sales volume and revenue.

2. Category Imbalance:

Some categories, such as Electronics and Apparel, dominate the order volume and revenue share, while others remain underutilized.

3. Seasonal Trends:

Monthly sales trends show fluctuations, with visible peaks during festival seasons. This suggests the importance of seasonal promotions.

4. Revenue vs. Quantity Gap:

Products with the highest revenue are not always the highest-selling by quantity — indicating high-margin products vs. high-volume products.

5. SubCategory Revenue Focus:

Certain subcategories generate higher sales despite moderate quantities, suggesting premium pricing potential in those areas.

6. Clean & Complete Data:

The dataset had minimal missing values and duplicates, making it ideal for direct analysis and business planning.

❓ Section: Business Recommendations

❓ Recommendations for Flipkart:

Based on the analysis and observed trends, the following strategic recommendations can be made for Flipkart:

- 1. Focus on High-Revenue Subcategories:**
Allocate more inventory and marketing budget to subcategories that consistently generate higher revenue (e.g., Electronics Accessories, Footwear).
- 2. Boost Low-Performing Categories:**
Investigate underperforming categories and consider removing, replacing, or promoting them through bundled offers or discounts.
- 3. Optimize for Seasonal Demand:**
Increase stock levels and ramp up marketing campaigns during high-order months (e.g., festival seasons) to maximize conversions.
- 4. Bundle Best-Sellers:**
Create combo offers with top-selling products and medium performers to push volume and increase Average Order Value (AOV).
- 5. Promote High-Margin Products:**
Identify products that generate high revenue with low quantity and target them through premium placements and influencer marketing.
- 6. Inventory Strategy:**
For high-quantity low-revenue items, ensure low procurement costs and high availability to avoid stockouts or overstocking.
- 7. Leverage Data for Personalization:**
Use insights to recommend products based on previous purchase trends by category and subcategory.

Conclusion

📌 Conclusion:

This project provided a comprehensive exploratory analysis of Flipkart's product order dataset using Python, SQL, and visualization tools. By examining key metrics like product performance, category-wise revenue, monthly trends, and customer preferences, we uncovered actionable insights that can significantly impact Flipkart's business strategy.

Through clear visualizations and structured analysis, the project successfully:

- Identified high-performing products and subcategories.
- Highlighted category-wise revenue contributions.
- Analyzed seasonal sales fluctuations.
- Differentiated between high-quantity and high-revenue items.

The results not only inform Flipkart's inventory and marketing decisions but also establish a strong analytical foundation for future forecasting and machine learning applications.

This project demonstrates how data-driven decisions can optimize operations, enhance customer experience, and ultimately increase profitability.

Section: Thank You Note

❓ Thank You

I would like to extend my sincere gratitude to **Dr. Vinod Mishra**, Head of Department, for his continuous support, guidance, and encouragement throughout the duration of this project.

I am also thankful to all the faculty members of **Ranvir Rananjay Post Graduate College, Amethi**, whose valuable inputs and academic resources have significantly contributed to the successful completion of this project.

Special thanks to my friends and mentors who inspired and supported me at every step of the journey. Their feedback and motivation were crucial in shaping the direction and quality of this work.

This project has been an enriching learning experience, and I am grateful to everyone who played a role in making it successful.

Warm Regards,
Anoop Pandey
Roll No. 20220043939108