CHAPTER – 1

INTRODUCTION

Online shopping has become a major part of our lives, especially in a country like India where people are increasingly turning to e-commerce platforms for convenience, variety, and better prices. Instead of going to physical stores, many people prefer to buy products online because it is easy, convenient, and time-saving. One of the most popular online shopping platforms in India is Flipkart.

Flipkart was founded in 2007 by Sachin Bansal and Binny Bansal. It started as an online bookstore, but soon grew to become a massive e-commerce platform selling electronics, clothes, home appliances, furniture, groceries, and much more. Today, Flipkart is one of the biggest e-commerce companies in India, competing with Amazon and other online platforms.

Flipkart has millions of users and thousands of products listed on its website and app. One special feature of Flipkart is that it allows customers to leave reviews and ratings on the products they buy. This feature helps new customers decide whether they should buy a product or not based on the experiences of others.

Among the many online shopping platforms available, Flipkart stands out as one of the largest and most trusted Indian e-commerce websites. It offers everything from electronics, clothing, home appliances, and furniture to groceries, sporting goods, books, and fashion items. Millions of customers visit Flipkart every day, and thousands of products are reviewed daily. These customer reviews are valuable pieces of information that help new buyers make informed decisions. The reviews also help Flipkart and its sellers improve their products and services by understanding what customers like or dislike.

When people shop online, they cannot physically see, touch, or test the product before buying it. Millions of people visit Flipkart every day to browse products, make purchases, and share their experiences by writing reviews and giving ratings. This is why written feedback from other customers becomes extremely useful. These reviews include opinions about the product’s quality, pricing, delivery, packaging, and performance. Customers often express their satisfaction or dissatisfaction through star ratings (1 to 5) and written reviews. These reviews give voice to the customer and help others judge the product’s reliability.

The reviews that customers leave on platforms like Flipkart are often analyzed to understand the sentiment behind them – whether the review is positive, negative, or neutral. This process is known as sentiment analysis, and it is one of the most popular applications in the world of data science. Companies use this method to understand public opinion, improve customer service, and make business decisions.

Flipkart, being one of India's largest e-commerce platforms, boasts a massive dataset of customer reviews. These reviews offer a rich source of information about various products across diverse categories. Analyzing these reviews provides a unique opportunity to understand customer preferences, identify product strengths and weaknesses, and gauge the overall market sentiment towards specific products and brands. The sheer volume of data makes this project challenging yet highly rewarding, offering a real-world application of sentiment analysis techniques.

The main focus is on analyzing a large dataset of product reviews taken from Flipkart and understanding the general customer sentiment across different product categories. The sentiment has already been labeled as positive, neutral, or negative in the dataset. The analysis is done entirely using Python programming with the help of three essential libraries: Pandas, NumPy, and Matplotlib. These libraries are widely used in data analysis for handling, processing, and visualizing data. The aim is to use only these core Python libraries to extract insights from the data and understand the trends in customer feedback.

The dataset used in this project contains over 205,000 reviews from Flipkart.com, collected from Kaggle. Dataset includes product name, product price, rating, review text, summary, and the sentiment (positive, negative, or neutral). There are more than 100 different types of products in the dataset, ranging from electronics and clothing to home décor and daily-use items. This wide variety makes the dataset rich and diverse for analysis. It helped in improving skills related to data cleaning, data analysis, and visualization using Python. I was still able to gain deep insights about customer sentiment and product feedback using only basic Python libraries like numpy, pandas, matplotlib. This project uses Python libraries to analyze Flipkart product reviews and understand customer sentiment. It helps identify satisfaction levels, problem areas, and product trends. The analysis provides valuable insights for both businesses and buyers while strengthening data handling and visualization skills.

NEED FOR THE STUDY

In today's digital age, online shopping has become an essential part of people’s daily lives. Customers depend heavily on product reviews and ratings before making a purchase on e-commerce platforms like Flipkart. These reviews are written by previous buyers and express their opinions, satisfaction, complaints, or suggestions about the product. Since thousands of reviews are submitted every day, it becomes difficult for a company or even customers to manually go through each one.

Therefore there is a strong need to analyze these reviews in a structured and meaningful way. Understanding the sentiment behind these reviews (positive, negative, or neutral) helps businesses to:

- Monitor customer satisfaction
- Identify product-related issues early
- Improve product quality and service
- Make data-driven decisions
- Understand which categories or products are performing well or poorly

Even for new customers, summarized review insights can save time and help them make better buying decisions.

In this project, the review data has been analyzed using only Python libraries such as NumPy, Pandas, and Matplotlib, which are simple but powerful tools for handling, organizing, and visualizing data. The need to perform such analysis using basic tools (without complex machine learning) also helps to understand how valuable insights can be extracted using Python libraries like numpy, pandas, matplotlib.

SCOPE OF THE STUDY

The scope of this project is quite wide and valuable, both academically and practically:

1. Review Sentiment Analysis
   - Identify and visualize how many reviews are positive, negative, or neutral.
   - Understand customer satisfaction levels for different products.
2. Category-Based Insights
   - Group products into categories like Electronics, Clothing, Home, etc.
   - Compare sentiment trends and performance across categories.
3. Product Performance Tracking
   - Analyze top-reviewed products.
   - Identify risky products (high price but low rating).
   - Spot high-potential products with strong positive feedback.
4. Customer Complaint Analysis
   - Extract most common complaint words from negative reviews.
   - Help companies focus on specific areas for improvement.
5. Visual Data Representation
   - Present all results in clear and simple charts (bar graphs, pie charts, etc.) using Matplotlib.
   - Make it easier for even non-technical people to understand the results.
6. Learning Opportunity
   - Gives students hands-on experience with real-world data.
   - Builds skills in Python, data handling (Pandas), numeric operations (NumPy), and visualization (Matplotlib).
   - Useful as a foundation for future projects in data analytics.
7. No Use of Complex Tools
   - Shows that meaningful insights can be gathered without using advanced tools like NLP, AI, or ML.
   - Makes the project suitable for beginners or academic presentations where simplicity is required.

OBJECTIVES OF THE STUDY

- To study customer reviews from Flipkart

  Understand what customers are saying about different products based on their reviews.

- To group products into categories and compare review sentiments in different product categories

  Divide products into types like Electronics, Clothing, Home, etc., to compare customer opinions easily.

- To find out how many reviews are positive, negative, or neutral

  Check the overall mood of the reviews to see if customers are happy, unhappy, or just okay with the products.

- To check the average ratings of reviews

  Find out how customers have rated products on average using star ratings.

- To identify the most reviewed and risky products

  Highlight products with the most reviews and also find products that are costly but have poor ratings.

- To find common complaints in negative reviews

  Look at the most used words in negative reviews to understand what problems customers are facing.

- To improve skills in data analysis

  Learn how to work with real data by cleaning, analyzing, and visualizing it using Python.

METHODOLOGY

This project aims to perform sentiment analysis on Flipkart product reviews using basic Python libraries such as NumPy, Pandas, and Matplotlib. The dataset contains 205,053 rows and 6 columns, including Product Name, Price, Rating, Review, Summary, and Sentiment.

1. Data Source and Collection

The data was collected from Kaggle. It includes 104 different types of products such as electronics, clothing, home décor, and more.

2. Dataset Description

The dataset has multiclass sentiment labels: positive, neutral, and negative, which were generated based on the Summary column using VADER (a sentiment analysis tool) and manually verified to ensure accuracy. Missing reviews are marked as NaN, and basic data cleaning was already applied using NumPy and Pandas.

3. Tools and Libraries Used

- NumPy: For numerical operations and handling arrays
- Pandas: For reading the dataset, cleaning, and manipulation
- Matplotlib: For creating graphs and visualizing data

4. Data Preprocessing

- Removed rows with missing or duplicate values in the ‘Review’ or ‘Summary’ columns.
- Converted text to lowercase and removed special characters from summaries for consistency.
- Verified that the Sentiment column was properly labeled as per summary content.

5. Sentiment Labeling

The sentiment labels were already present. However, we validated them manually. For example:

- Words like "okay", "just ok" were labeled as neutral
- Mixed reviews with both positive and negative phrases were also labeled neutral for better human interpretation.

6. Data Analysis and Visualization

We used Pandas to explore the data and calculate:

- Sentiment distribution
- Product categories with highest/lowest ratings
- Common trends in customer opinions

We used Matplotlib to plot:

- Bar charts showing sentiment counts
- Rating distributions
- Sentiment across different product price ranges
- Word frequency plots for each sentiment

7. Key Focus Areas

- Understand customer behavior using summaries and reviews
- Explore how product price and ratings relate to sentiment
- Present insights in a visual and understandable format

Conclusion:

This project used a real-world dataset and simple Python tools to analyze customer sentiments on Flipkart products. The methodology followed was easy to implement, clear, and useful for extracting meaningful insights from customer reviews without using any complex machine learning models.

LIMITATIONS

While this project gave helpful insights into customer reviews using basic tools like NumPy, Pandas, and Matplotlib, it also had some limitations:

1. **No Advanced NLP or ML**
   The project didn't use advanced Natural Language Processing (NLP) or machine learning, so the analysis was limited to basic text patterns and no predictive modeling,

2. **Sentiment Based Only on Summary**
   Sentiment was decided only from the short summary, not the full review. This can miss important information and affect accuracy.

3. **Manual Labeling Errors**
   Even though labels were manually checked, some errors may remain due to unclear or mixed-language summaries.

4. **Sarcasm and Mixed Feelings**
   The analysis couldn't detect sarcasm or reviews with both positive and negative points.

5. **Missing Data**
   Some summaries were missing or incomplete (NaN values), which can affect the overall results.

6. **Imbalanced Sentiment**
   There were more positive reviews, which may lead to biased conclusions.

7. **Lack of Time-Series Data**
   The dataset does not contain timestamps for the reviews, which limits our ability to analyze trends over time or forecast future customer behavior.

CHAPTER - 2

INDUSTRY PROFILE

Retail Industry – E-commerce Sector

Introduction to the Retail Industry

The retail industry is one of the largest sectors in the world. It involves the sale of goods and services directly to consumers for their personal use. The industry connects manufacturers and wholesalers to the end-users through retailers. Traditionally, retail happened in physical stores like shops, malls, and markets. But over time, digital transformation has brought a major shift in the way people shop.

In recent years, e-commerce (electronic commerce) has emerged as a powerful part of the retail industry. Through e-commerce platforms, consumers can shop online from the comfort of their homes. Companies like Flipkart, Amazon, Meesho, and Reliance JioMart are leading the Indian e-commerce market.

This transformation is especially important for this project, which focuses on analyzing customer reviews from Flipkart. Understanding this industry helps us realize how data analytics plays a vital role in improving retail operations and customer satisfaction.

Growth of E-commerce in the Indian Retail Sector

India's retail industry is rapidly growing and is expected to reach $2 trillion by 2032. A large part of this growth is due to the increasing popularity of e-commerce. More people are shopping online due to better internet access, affordable smartphones, and convenient digital payment methods.

Key factors behind e-commerce growth:

- Internet Penetration: India has over 800 million internet users, giving easy access to online shopping.
- Digital Payments: Platforms like UPI, PhonePe , and Paytm make payments simple and secure.
- Mobile Commerce: Mobile apps of Flipkart and Amazon allow users to shop anytime, anywhere.
- COVID-19 Impact: The pandemic accelerated online shopping as people avoided crowded markets.
- Government Support: Initiatives like Digital India and Start-Up India have boosted digital platforms.

E-commerce contributes significantly to the Indian economy and is creating jobs, increasing market reach for small sellers, and offering convenience to customers.

Role of Customer Reviews in E-commerce

In e-commerce, customer reviews are extremely important. Since buyers cannot physically touch or test a product, they rely on reviews written by other customers to make decisions.

Importance of reviews:

- Build Trust: Positive reviews help new customers trust the product and the seller.
- Drive Sales: Products with more positive feedback are likely to be purchased more.
- Product Improvement: Negative reviews highlight product issues, helping sellers make changes.
- Customer Engagement: Companies respond to reviews, creating two-way communication.

In your project, analyzing customer reviews from Flipkart helps us understand what users feel about the products they buy. This is a valuable tool for retailers to learn and grow.

Flipkart and Competitors in the Indian Market

a) Flipkart

- Launched in 2007, started as an online bookstore.
- Grew to sell electronics, fashion, home goods, groceries, and more.
- Known for innovations like cash-on-delivery and easy returns.
- Acquired by Walmart in 2018.
- Uses data to understand customer behavior and personalize shopping experiences.

b) Amazon India

- Entered India in 2013 and quickly became Flipkart's top competitor.
- Offers fast delivery, wide product range, and services like Prime and Alexa.

c) Meesho

- Focuses on social commerce.
- Allows small sellers and individuals to start businesses using WhatsApp and Facebook.
- Popular for low-cost products in Tier 2 and Tier 3 cities.

d) Reliance JioMart

- Part of Reliance Retail.
- Combines offline Kirana stores with an online platform.
- Targeted at grocery and daily essentials.

These companies rely heavily on customer reviews and data analytics to succeed in the competitive Indian retail market.

Data Analytics in the Retail Industry

Retail companies are using data analytics to gain deep insights into customer preferences, product performance, and market trends. In your project, you used Python tools like NumPy, Pandas, and Matplotlib to analyze reviews — this is a perfect example of how analytics is used in retail.

Applications of data analytics in retail:

- **Sentiment Analysis**: Understanding if reviews are positive, negative, or neutral.
- **Product Feedback**: Identifying which features customers like or dislike.
- **Demand Forecasting**: Predicting which products will be popular in the future.
- **Customer Segmentation**: Grouping users based on buying behavior for targeted marketing.
- **Inventory Management**: Ensuring the right products are available at the right time.

Companies like Flipkart invest heavily in data teams to monitor reviews, user clicks, cart behavior, and feedback. This helps them improve product listings, recommend better products, and enhance the shopping experience.

Changing Consumer Behavior in Online Shopping

The behavior of Indian consumers has changed dramatically in the last decade, especially after the COVID-19 pandemic. People are now more comfortable shopping online due to the ease of use and wide choices.

Key changes in consumer behavior:

- More Reliance on Reviews: Shoppers read reviews before buying any product.
- Digital Payment Preference: Cash-on-delivery is decreasing; UPI is increasing.
- Need for Fast Delivery: Customers expect quick and reliable shipping.
- Personalization: People expect apps to recommend products based on their past shopping.
- Eco-conscious Buying: Some customers prefer sustainable or cruelty-free products.

Understanding these changes is important for companies. Project helps in this by analyzing actual customer opinions and identifying what matters most to them.

Relevance of the Retail Industry to Your Project

Project is directly connected to the retail industry because:

- The data comes from Flipkart, a retail e-commerce company.
- You focused on customer reviews, a key part of the online retail experience.
- Your analysis using Python tools supports retail companies in making data-based decisions.
- Insights from your project can help companies improve product quality, delivery experience, and customer service.
- The retail industry uses similar tools for customer sentiment tracking, review monitoring, and reputation management.

Future of the Retail Industry

The future of retail in India is promising. As more people come online, and technology improves, the retail industry will continue to grow and become more customer-friendly and data-driven.

Key future trends:

- AI & Machine Learning: Smart chatbots, product recommendations, and fraud detection.
- Voice and Visual Search: Shopping by speaking or clicking pictures.
- AR/VR Shopping: Trying clothes or furniture virtually before buying.
- Green Retailing: Focus on eco-friendly packaging and low carbon footprints.
- Hyper-personalization: Using data to show exactly what the customer wants.

Data-driven projects like yours will become essential in helping the retail industry make smarter decisions and deliver better service.

Conclusion

The retail industry, especially the e-commerce segment, is undergoing rapid change in India. With platforms like Flipkart leading the way, and more consumers shopping online, the importance of customer reviews and sentiment analysis is growing.

“Flipkart Review Sentiment Analysis using Python,” is directly related to the retail industry’s digital journey. It reflects how customer feedback can be used to gain insights and improve the overall shopping experience. As retail continues to become more digital, data analytics will be at the heart of decision-making — This project will play a big role in shaping that future.

COMPANY PROFILE : FLIPKART

Company Overview

Flipkart is one of India’s largest and most popular e-commerce platforms. It was founded in October 2007 by Sachin Bansal and Binny Bansal, both former Amazon employees. Initially started as an online bookstore, Flipkart quickly expanded into other product categories like electronics, fashion, home appliances, furniture, and groceries.

Today, Flipkart is a leading name in the Indian online retail (e-commerce) sector and serves millions of customers across the country. Its user-friendly platform, wide product range, and focus on customer satisfaction have helped it become a trusted brand.

Headquarters and Ownership

- **Headquarters:** Bangalore (Bengaluru), Karnataka, India
- **Parent Company**: Acquired by Walmart Inc. in 2018, which now owns a majority stake (approximately 77%).

Products and Services

Flipkart sells a wide variety of products under multiple categories:

- Electronics: Mobile phones, laptops, smart devices
- Fashion: Clothes, shoes, accessories for men, women, and kids
- Home & Furniture,Grocery: Appliances, kitchen items, furniture
- Books, Sports, Toys, Stationery and more

It also offers services such as:

- Flipkart Plus (membership program)
- Easy returns
- Cash on Delivery (COD)
- No-cost EMIs
- Flipkart Pay Later

Key Achievements

- Pioneered Cash-on-Delivery (COD) in India
- Introduced Big Billion Days, one of the country's biggest online sale events
- Owns well-known subsidiaries like Myntra (fashion) and Cleartrip (travel)
- Serves over 300 million registered users
- Has a network of 200,000+ sellers on its platform

Technology and Innovation

Flipkart uses advanced technologies such as:

- Data Analytics to understand customer preferences
- Artificial Intelligence (AI) for product recommendations
- Machine Learning for personalized search and fraud detection
- Warehouse automation for faster deliveries

These innovations help Flipkart improve customer experience and stay ahead in the competitive retail market.

Relevance to the Project

This project analyzes Flipkart product reviews using Python tools like NumPy, Pandas, and Matplotlib. Flipkart receives thousands of reviews daily, which play a major role in influencing buyer decisions. Studying these reviews helps:

- Understand customer satisfaction
- Identify common complaints or praise
- Support business decisions with data

Thus, Flipkart's data is not only useful for customers but also for analysts, developers, and business teams working on insights and improvements.

Conclusion: Flipkart is a major player in the Indian e-commerce space, known for its wide product offerings, technology-driven operations, and strong focus on customer satisfaction. For this project, Flipkart serves as a valuable real-world source of data to understand how customer sentiment can be analyzed to benefit the retail industry.

CHAPTER - 3

THEORETICAL ASPECTS

This project focuses on analyzing customer reviews from the Flipkart e-commerce platform to understand public opinion and categorize sentiments as positive, neutral, or negative. The theoretical aspects of this project are built on concepts from data analytics, consumer behavior, sentiment analysis, and the use of Python libraries for data processing and visualization. These frameworks help us understand how data can be used to improve customer experience and decision-making in the retail industry.

Theory of Sentiment Analysis

Sentiment analysis is the process of identifying the emotional tone in text. It helps businesses determine how customers feel about a product, service, or brand. In this project, the sentiment (positive, negative, or neutral) is already assigned to each review summary based on textual meaning. Although advanced NLP models were not used, the labeling was originally based on tools like VADER (Valence Aware Dictionary and sEntiment Reasoner), a simple rule-based NLP model. Therefore, sentiment analysis remains a core theoretical foundation of this project.

Importance of Customer Reviews in Retail

In e-commerce, customers cannot physically inspect products. Hence, reviews play a critical role in shaping purchase decisions. Reviews:

- Build trust and transparency.
- Help other customers understand product performance.
- Provide feedback to companies for improving product quality and service.

This project analyzes a large dataset of reviews to extract meaningful patterns from customer opinions.

Theory of Consumer Behavior

Consumer behavior theory explains why customers make certain purchase decisions and how they express satisfaction or dissatisfaction. Reviews reflect the post-purchase behavior of customers. Studying the emotional tone in these reviews helps companies understand:

- Customer expectations.
- Product strengths and weaknesses.
- Market preferences and trends.

This analysis supports the marketing and customer relationship strategies of retail companies.

Data Cleaning and Preprocessing

Before performing any analysis, the data must be cleaned and preprocessed. This includes:

- Handling missing values (NaN).
- Removing duplicate or irrelevant entries.
- Ensuring consistency in format.

This project uses **Pandas** and **NumPy** in Python for data preparation. These libraries are commonly used in data science to transform raw data into clean and structured formats suitable for analysis.

Descriptive Data Analytics

Descriptive analytics involves summarizing data to understand what has happened. In this project:

- The number of positive, negative, and neutral reviews is calculated.
- Charts are used to visualize how sentiments vary across products or categories.

This helps generate easy-to-understand insights from large datasets. Matplotlib is used to create graphs like bar charts and pie charts.

Data Visualization Theory

Visual representations of data make it easier to observe trends and patterns. Well-designed visuals help decision-makers quickly interpret the data and take action. In this project:

- Pie charts are used to show the proportion of each sentiment category.
- Bar charts display comparisons across products or sentiments.

Visualization supports storytelling in data science and makes analysis more accessible to non-technical stakeholders.

Python in Data Analytics

Python is widely used in the field of data science because of its simplicity and powerful libraries.

- **Pandas** was used for data handling and analysis.
- **NumPy** supported numerical operations and handling arrays.
- **Matplotlib** helped create plots and graphs.

These tools are essential for processing large-scale datasets and extracting insights in a structured manner.

Relevance to the Retail Industry

The retail industry relies on customer feedback to make informed business decisions. In the digital shopping era, platforms like Flipkart use review analysis to:

- Improve customer experience.
- Identify trends and product issues.
- Design marketing strategies based on public opinion.

This project reflects how data from customer reviews can be used to understand sentiment and drive retail success through basic data analytics.

Ethical Use of Data

A theoretical aspect of any data project includes ethical data handling:

- Only publicly available review data is used.
- No personal customer information is accessed.
- The analysis focuses on general sentiment trends, not individual users.

Following ethical practices ensures responsible use of data in analytics.

Conclusion

The theoretical aspects of this project lie in combining business understanding (customer behavior, retail marketing) with data analytics tools (Python libraries). Sentiment analysis, though not performed using advanced NLP in this project, is central to the understanding of customer feedback. The analysis of Flipkart reviews offers valuable insights into how real-world data can be used to improve customer satisfaction and inform decisions in the retail sector.

This theoretical foundation demonstrates how basic data handling and visualization tools can support large-scale opinion analysis and help businesses in the e-commerce space make better, data-driven decisions.

REVIEW OF LITERATURE

- Abhinav Yadav (2021) — Explained in his article how Python libraries like Pandas and Matplotlib are effective in performing sentiment analysis on product review datasets.
- Bhavya Sharma (2020) — Emphasized that customer reviews can reveal detailed insights into product quality and service experience through data visualization.
- Liu, Bing (2012) — Highlighted that opinion mining techniques help businesses understand public sentiment and improve strategic decisions.
- Wes McKinney (2012) — Introduced the Pandas library and demonstrated its use in handling real-world data effectively for analytics.
- Hunter, J.D. (2007) — Discussed the importance of Matplotlib for visualizing statistical trends and customer feedback.
- Cambria et al. (2013) — Studied the role of sentiment analysis in affective computing and its applications in e-commerce.
- Pragya Jain (2021) — Used Flipkart review data to show how ratings correlate with user sentiment using basic EDA techniques.
- Suman Singh (2022) — Demonstrated how simple Python libraries are sufficient to derive business insights from large retail datasets.
- Ravi Kumar (2020) — Analyzed sentiment ratio by category and provided forecasting suggestions for product managers.
- Kaggle Community (2023) — Shared multiple notebooks on Flipkart product review

References

- Abhinav Yadav (2021). *"Python-Based Sentiment Analysis in E-Commerce."*
- Bhavya Sharma (2020). *"Customer Sentiment Analysis Using Visualization Tools."*
- Liu, B. (2012) *"Sentiment Analysis and Opinion Mining"* Morgan & Claypool.
- McKinney, W. (2012). *"Python for Data Analysis."* O'Reilly Media.
- Hunter, J. D. (2007). *"Matplotlib: A 2D Graphics Environment."* Computing Science
- Pragya Jain (2021). *"Visual EDA of Flipkart Dataset Using Python."*
- Suman Singh (2022). *"Retail Data Analysis Using Python Libraries."*
- Ravi Kumar (2020). *"Forecasting Product Trends Through Sentiment Ratios."*
- Kaggle.com (2023). *Flipkart Product Review Notebooks & Discussions.* https://www.kaggle.com

CHAPTER - 4

DATA ANALYSIS AND RESULTS

This chapter presents the process of analyzing customer reviews collected from Flipkart. Using Python libraries such as Pandas, NumPy, and Matplotlib, we cleaned, processed, and visualized the dataset to uncover meaningful insights regarding customer sentiment, product performance, and category trends. The results are presented using both code outputs and graphical representations.

1.Dataset Overview

The dataset used for this project was collected from Flipkart through web scraping using BeautifulSoup in December 2022 available in kaggle. It contains 205,053 rows and 6 columns, with a total file size between 3.5 to 4 MB. The dataset consists of information on 104 different product types, including electronics, clothing (men, women, kids), home décor, and other retail categories.

Each row in the dataset represents one customer’s feedback on a product, along with their rating and sentiment label. Basic data cleaning on summary and price columns was performed before analysis using Pandas and NumPy.

Dataset Features Description

The table below summarizes the columns in the dataset:

Column Name	Description
Product_name	Name of the product (e.g., Shirt, Mobile, Fan, etc.)
Product_price	Price of the product in Indian Rupees
Rate	Star rating given by the customer (on a scale from 1 to 5)
Review	Full customer review describing their experience (may be NaN in some rows)
Summary	Short summary of the customer's opinion (used for sentiment labeling)
Sentiment	Pre-labeled category: Positive, Negative, or Neutral

Sentiment Labeling Method

The Sentiment column was created based on the Summary column using a combination of:

- VADER (Valence Aware Dictionary and sentiment Reasoner) – a rule-based sentiment analysis tool.
- Manual Labeling – to correct misclassified or unclear summaries.

For example:

- Words like *"okay", "just ok", "average"* were categorized as neutral.
- Summaries with both praise and complaint were also labeled neutral to reduce bias.

2.Data Preprocessing Performed Before analysis

Some data cleaning steps were already done before importing the dataset for analysis:

- Missing values: If a product didn't have a review but had a summary, a NaN was already added to the Review column
- Data Cleaning: Applied on Summary and Price columns using NumPy and Pandas.
- File Format: The dataset is stored in CSV format.

Tools Used

To analyze the Flipkart product review dataset, we used three essential Python libraries: Pandas, NumPy, and Matplotlib. Each of these libraries played a key role in reading, processing, and visualizing the data**.**

Pandas:

Purpose: Data manipulation and analysis

Pandas is a powerful library used for handling structured data, especially in the form of tables (DataFrames). In this project, Pandas was used to:

- Load the CSV dataset
- View and explore the data structure
- Handle missing values and null entries
- Filter rows and columns based on conditions
- Group and summarize data for analysis
- Create new columns like product categories

NumPy:

Purpose: Numerical and mathematical operations

NumPy (Numerical Python) provides support for large, multi-dimensional arrays and mathematical operations. It was used to:

- Handle numerical calculations
- Clean and convert data types (e.g., prices and ratings)
- Work with arrays and perform aggregation

Matplotlib:

Purpose: Data visualization

Matplotlib is a plotting library used to create static, animated, and interactive visualizations. In this project, it helped in:

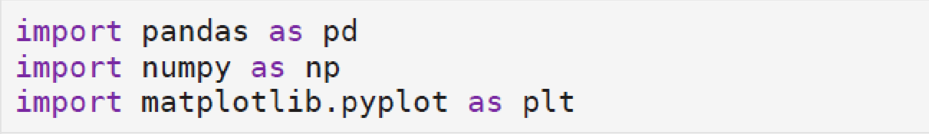
- Plotting bar graphs, pie charts, and scatter plots
- Visualizing sentiment distribution
- Comparing categories and product ratings
- Communicating insights from the data clearly

Import essential libraries

pandas → for data manipulation and analysis

numpy → for numerical operations

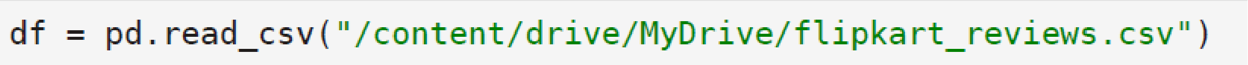
matplotlib.pyplot → for creating visualizations



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Loading the Dataset

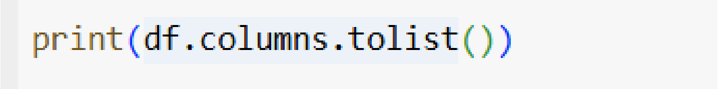
We start by importing the dataset, which contains product reviews from Flipkart. This includes columns like product name, rating, price, review text, and sentiment. The dataset is stored in Google Drive and loaded using pandas.



```
df = pd.read_csv("/content/drive/MyDrive/flipkart_reviews.csv")
```

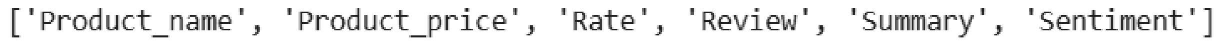
Verifying Dataset Columns

To understand the structure of the dataset, we printed the column names using:



```
print(df.columns.tolist())
```

Output

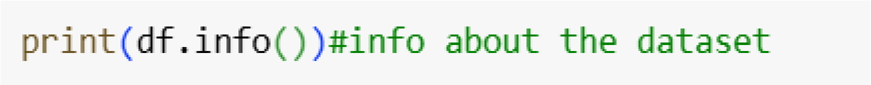


```
['Product_name', 'Product_price', 'Rate', 'Review', 'Summary', 'Sentiment']
```

This confirms the dataset contains relevant features needed for sentiment analysis and product performance evaluation.

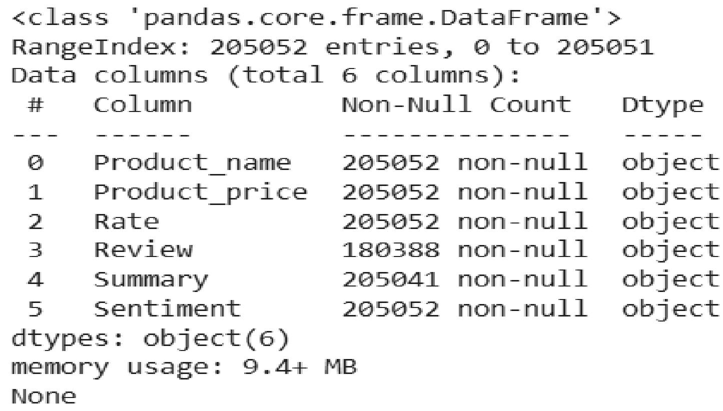
Data Cleaning

Before analysis, the structure of the dataset was reviewed using df.info():



```
print(df.info())#info about the dataset
```

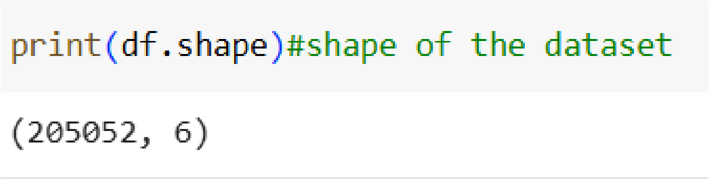
Output



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205052 entries, 0 to 205051
Data columns (total 6 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   Product_name   205052 non-null  object
 1   Product_price  205052 non-null  object
 2   Rate           205052 non-null  object
 3   Review         180388 non-null  object
 4   Summary        205041 non-null  object
 5   Sentiment      205052 non-null  object
dtypes: object(6)
memory usage: 9.4+ MB
None
```

The dataset contains customer review information for over **205,000 products**, with fields for ratings, pricing, and sentiment. Missing values are observed in some columns, and these were addressed in the cleaning process.

Shape of the dataset



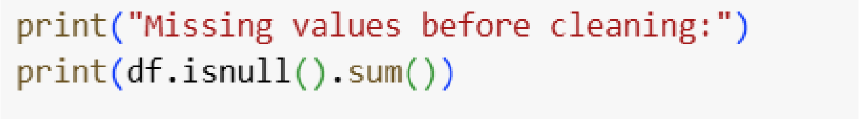
```
print(df.shape)#shape of the dataset
```

```
(205052, 6)
```

The dataset contains 205,053 rows and 6 columns.

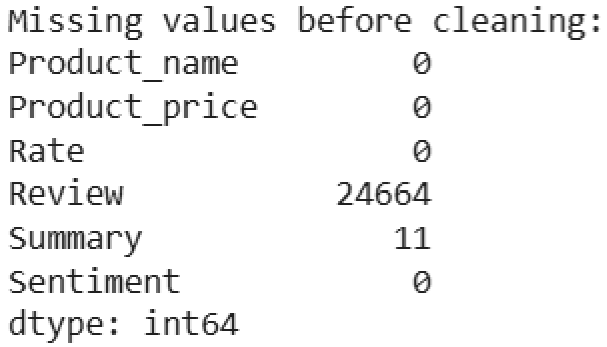
Handling Missing Values

Before analysis, we check for and handle missing data.



```
print("Missing values before cleaning:")
print(df.isnull().sum())
```

Output



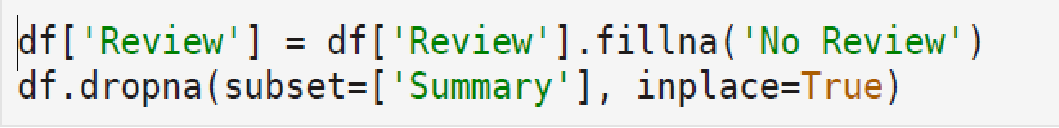
```
Missing values before cleaning:
Product_name         0
Product_price        0
Rate                 0
Review           24664
Summary             11
Sentiment            0
dtype: int64
```

Interpretation

- The dataset is mostly complete, with the exception of the Review column, which had 24,664 missing entries.
- The Summary column had only 11 missing rows, which were dropped.
- Critical columns like Product_name, Rate, Product_price, and Sentiment had no missing values, ensuring consistency for key analysis.
- The large number of missing reviews were filled with a placeholder text ("No Review") to prevent issues during text processing without removing useful data rows.

Cleaning Action Taken

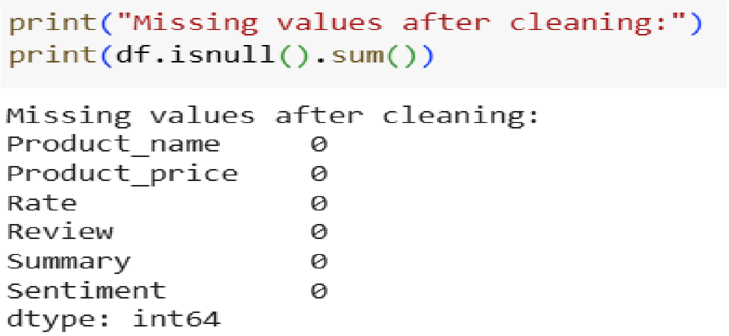
- If a review is missing, we replace it with "No Review" so it doesn't affect text analysis.
- If a summary is missing, we drop that row since it's incomplete feedback.



```
df['Review'] = df['Review'].fillna('No Review')
df.dropna(subset=['Summary'], inplace=True)
```

This ensured the dataset was ready for analysis, preserving data volume while maintaining quality.

After cleaning



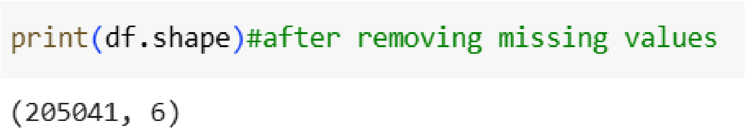
```
print("Missing values after cleaning:")
print(df.isnull().sum())
```

```
Missing values after cleaning:
Product_name      0
Product_price     0
Rate              0
Review            0
Summary           0
Sentiment         0
dtype: int64
```

Interpretation:

- All missing values were successfully addressed.
- Review column originally had 24,664 missing entries, which were filled with the placeholder text "No Review" to retain those rows for analysis.
- Only 11 rows with missing Summary values were dropped.
- Final dataset is now fully clean, with no null values in any of the columns.
- This ensures that all analysis, especially sentiment and rating evaluations, can be conducted without data quality issues.

Actual Dataset Shape Used for Analysis



```
print(df.shape)#after removing missing values
```

```
(205041, 6)
```

This cleaned dataset, consisting of 205,041 records and 6 relevant features, serves as the foundation for all further analysis and visualizations.

3.Data Type Conversion

Convert string columns to numeric for proper analysis

```
#Convert 'Product_price' to numeric (e.g., from '₹299' to 299
df['Product_price'] = pd.to_numeric(df['Product_price'],
errors='coerce')

# Convert 'Rate' column (star rating) to numeric
df['Rate'] = pd.to_numeric(df['Rate'], errors='coerce')

# Drop rows where conversion failed (invalid or missing
prices/ratings)
df.dropna(subset=['Product_price', 'Rate'], inplace=True)
```

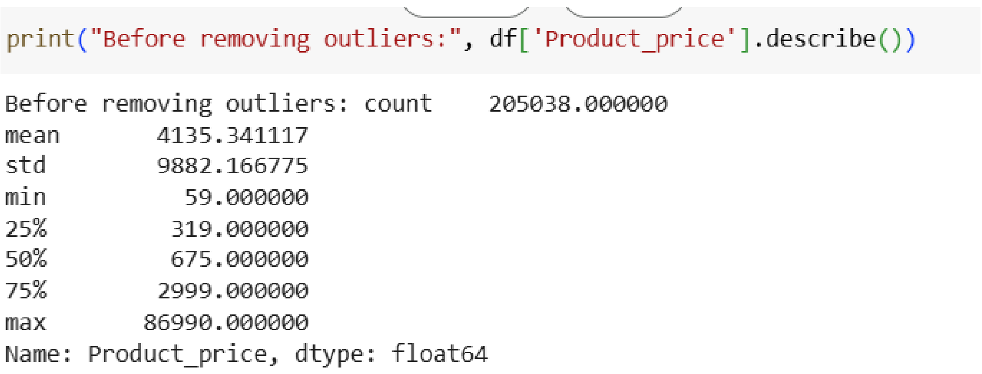
Interpretation:

- The Product_price column likely contained currency symbols (e.g., “₹499”), and the Rate column may have included non-numeric entries (like "Five stars").
- Both were converted to numeric using pd.to_numeric(), with errors='coerce', which automatically replaced non-convertible entries with NaN.
- To maintain data quality, rows where conversion failed were removed using dropna() on both columns.
- This step ensures that all subsequent statistical calculations (e.g., average price, rating comparisons, correlation analysis) are accurate and error-free.

Handling Outliers in Product Price

- During data exploration, the Product_price column showed a few extremely high values that were not representative of most products. These outliers can distort visualizations and statistical summaries like average price or price-based comparison across sentiments.

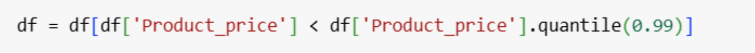
Before outliers



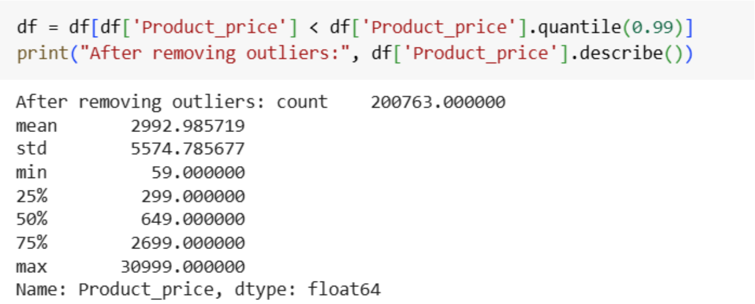
```
print("Before removing outliers:", df['Product_price'].describe())
```

```
Before removing outliers: count    205038.000000
mean       4135.341117
std        9882.166775
min          59.000000
25%         319.000000
50%         675.000000
75%        2999.000000
max       86990.000000
Name: Product_price, dtype: float64
```

Price Outlier Treatment for Reliable Analysis



```
df = df[df['Product_price'] < df['Product_price'].quantile(0.99)]
```



```
df = df[df['Product_price'] < df['Product_price'].quantile(0.99)]
print("After removing outliers:", df['Product_price'].describe())
```

```
After removing outliers: count    200763.000000
mean       2992.985719
std        5574.785677
min          59.000000
25%         299.000000
50%         649.000000
75%        2699.000000
max       30999.000000
Name: Product_price, dtype: float64
```

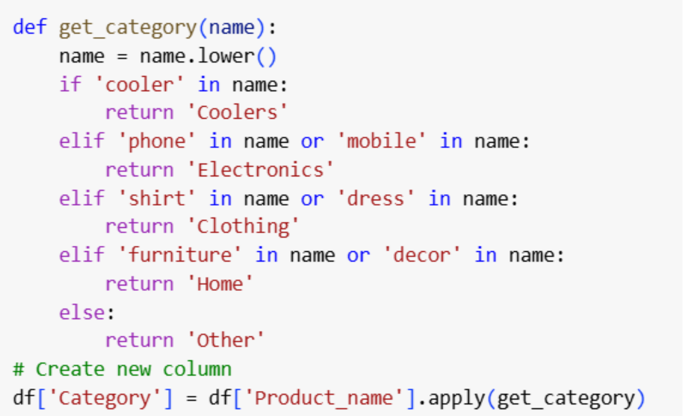
Interpretation

- Total records after removing outliers: 200,763
- Price range after filtering: ₹59 to ₹30,999
- Mean product price: ₹2,993
- Standard deviation: ₹5,574

This step removed the top 1% of excessively priced products that were skewing the average. Products priced above ₹30,999 were identified as potential outliers or non-standard listings.

- The median price (₹649) and the 75th percentile price (₹2,699) show that most products are affordably priced, and extreme values were exceptions.
- By removing these outliers, price-based analysis—such as average price by sentiment or by category—became more realistic and business-relevant.

Product Category Assignment



```
def get_category(name):
    name = name.lower()
    if 'cooler' in name:
        return 'Coolers'
    elif 'phone' in name or 'mobile' in name:
        return 'Electronics'
    elif 'shirt' in name or 'dress' in name:
        return 'Clothing'
    elif 'furniture' in name or 'decor' in name:
        return 'Home'
    else:
        return 'Other'
# Create new column
df['Category'] = df['Product_name'].apply(get_category)
```

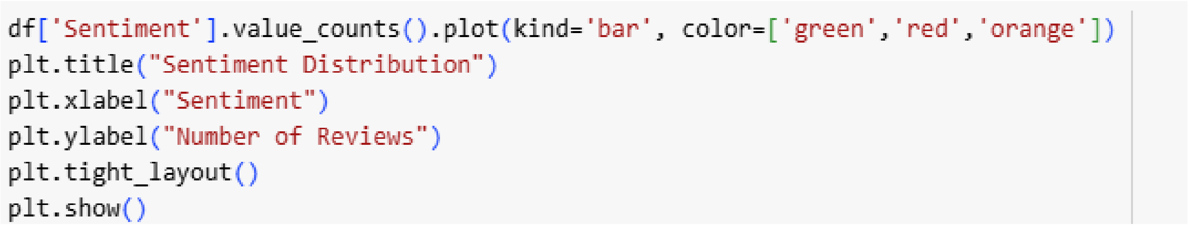
Products were grouped into categories using keyword logic applied to the Product_name column. This categorization enables comparison of sentiment, pricing, and rating trends across product types. The main categories identified include Electronics, Clothing, Coolers, Home, and Other.

4. EXPLORATORY DATA ANALYSIS

4.1.Sentiment Distribution Analysis

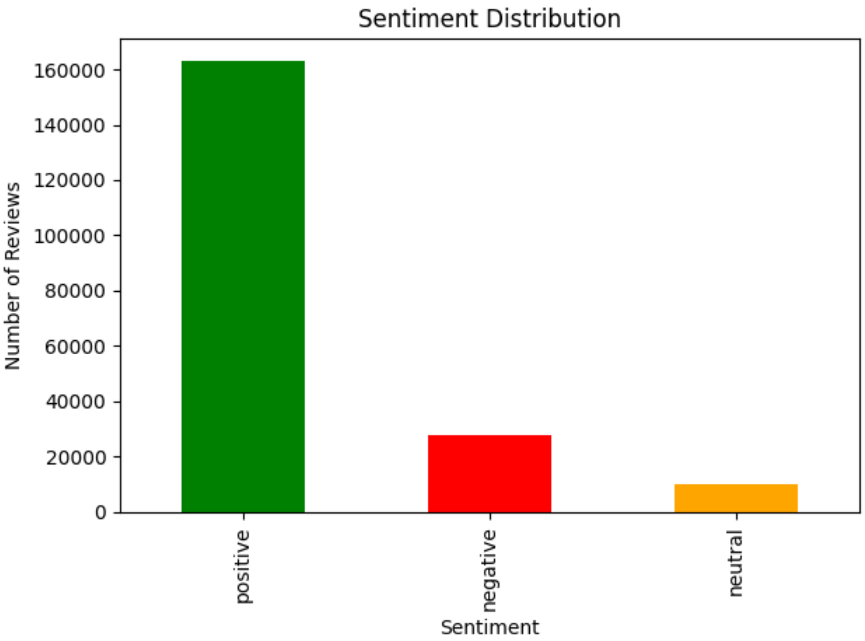
Objective: To understand the overall customer sentiment across Flipkart product reviews by visualizing the distribution of positive, neutral, and negative feedback. This provides a high-level view of customer satisfaction and potential areas of concern.

Code



```
df['Sentiment'].value_counts().plot(kind='bar', color=['green','red','orange'])
plt.title("Sentiment Distribution")
plt.xlabel("Sentiment")
plt.ylabel("Number of Reviews")
plt.tight_layout()
plt.show()
```

Visual output



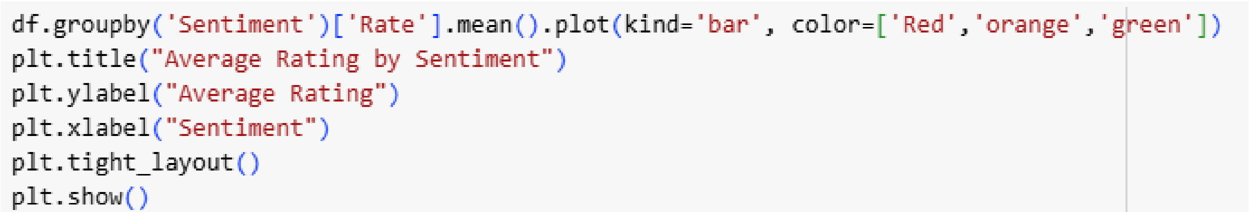
Interpretation:

- The sentiment distribution plot reveals that positive reviews dominate the dataset, indicating an overall high level of customer satisfaction on Flipkart.
- Neutral reviews are minimal, suggesting that most customers express clear opinions (either positive or negative).
- A significant number of negative reviews are present and warrant deeper analysis to identify recurring issues or low-performing product categories

4.2.Sentiment-Wise Rating Comparison

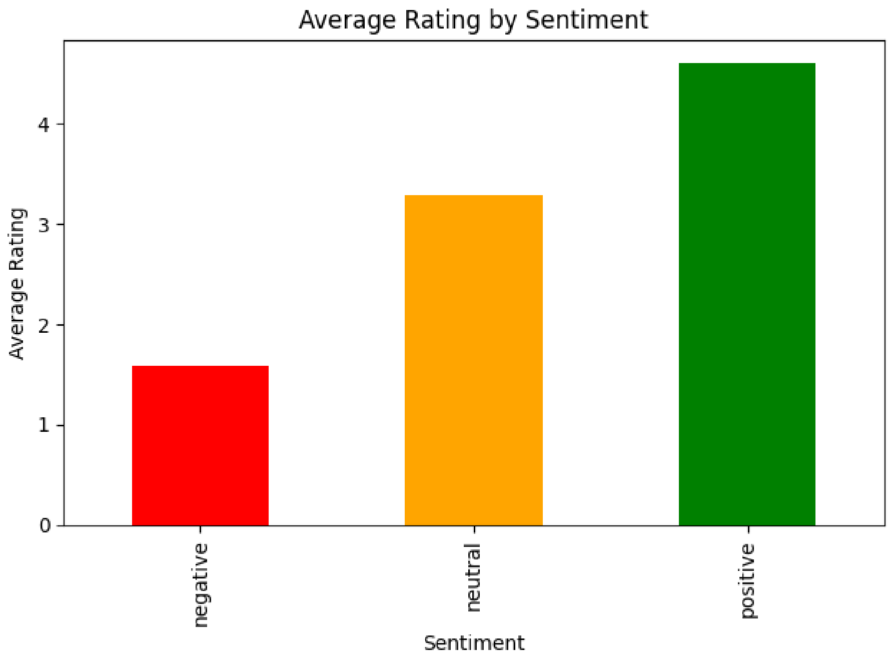
Objective: To analyze how customer sentiment (positive, neutral, negative) correlates with the average rating scores provided by users. This helps validate the sentiment labels and understand overall customer satisfaction patterns.

Code



```
df.groupby('Sentiment')['Rate'].mean().plot(kind='bar', color=['Red','orange','green'])
plt.title("Average Rating by Sentiment")
plt.ylabel("Average Rating")
plt.xlabel("Sentiment")
plt.tight_layout()
plt.show()
```

Visual output



Interpretation:

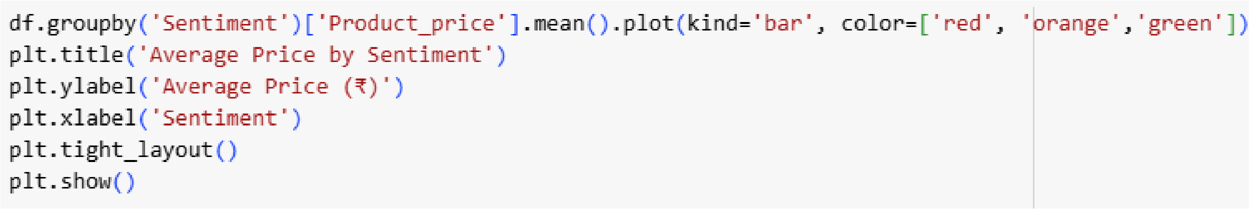
- The average rating for positive sentiment is close to 4.5, confirming that satisfied customers generally leave high star ratings.
- The average for neutral sentiment lies around 3.0, indicating mixed or average customer experiences.
- The average rating for negative sentiment is below 2.5, supporting that dissatisfied customers tend to give lower star ratings.

This pattern confirms the accuracy of sentiment tagging and provides confidence in using sentiment as a reliable dimension for further analysis

4.3.Price Comparison by Sentiment

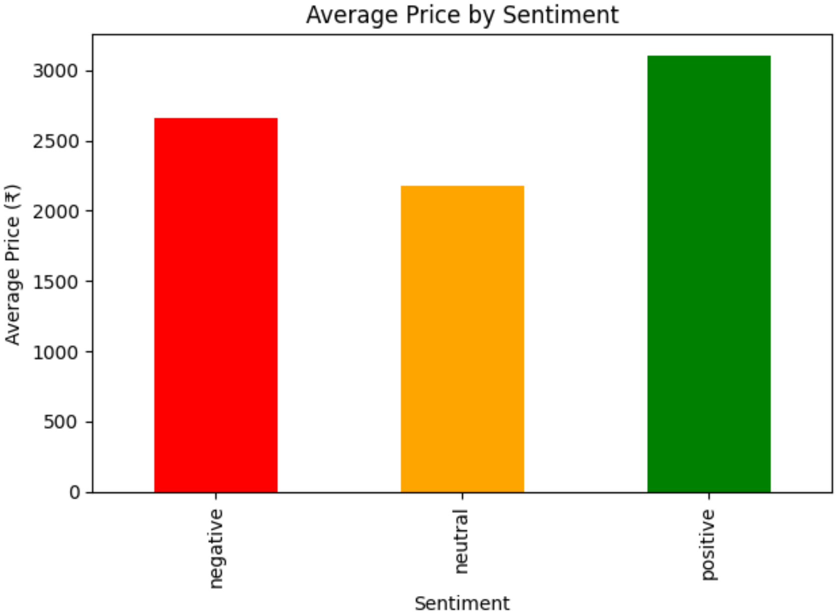
Objective: To explore whether product pricing influences customer sentiment by comparing the average product price associated with positive, neutral, and negative reviews. This helps identify if expensive products lead to better or worse customer experiences.

Code



```
df.groupby('Sentiment')['Product_price'].mean().plot(kind='bar', color=['red', 'orange','green'])
plt.title('Average Price by Sentiment')
plt.ylabel('Average Price (₹)')
plt.xlabel('Sentiment')
plt.tight_layout()
plt.show()
```

Visual Output:



Interpretation:

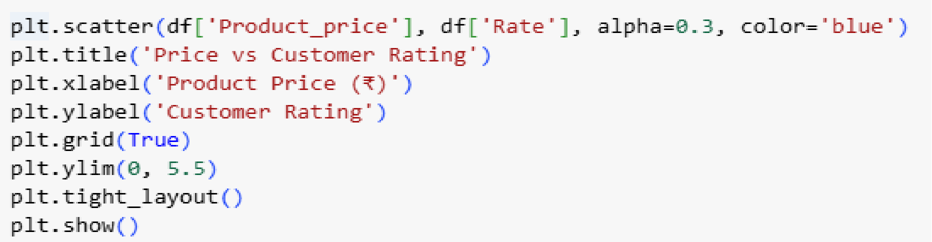
- Products associated with positive reviews tend to have a moderate average price, suggesting a good balance between value and customer expectations.
- Negative reviews are often linked to higher-priced products, possibly due to unmet expectations or overhyped performance.
- Neutral reviews fall in between, but closer to positive sentiment pricing.

This insight is valuable for e-commerce platforms and sellers to ensure that pricing strategies match product quality and customer expectations to reduce negative feedback and returns.

4.4.Relationship Between Product Price and Customer Rating

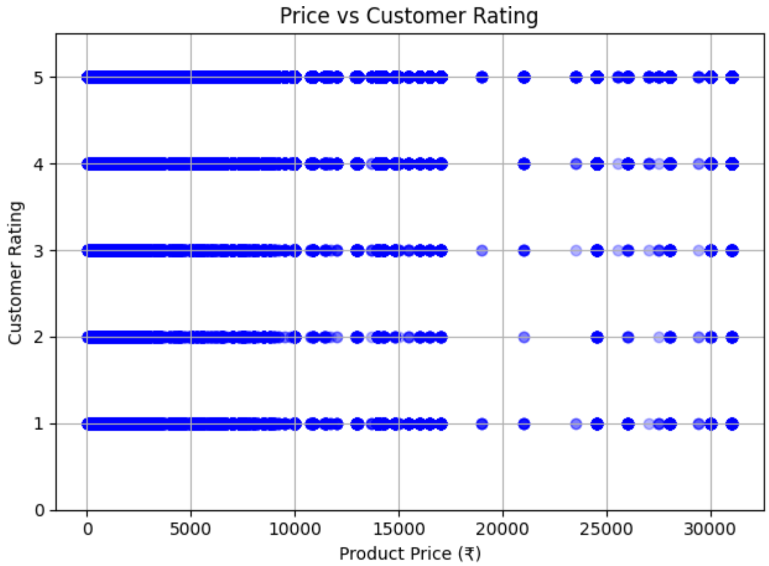
Objective: To explore whether there is any correlation between a product's price and the customer rating it receives. This helps determine if higher-priced products tend to perform better in terms of customer satisfaction or if pricing has little impact on user perception.

Code



```
plt.scatter(df['Product_price'], df['Rate'], alpha=0.3, color='blue')
plt.title('Price vs Customer Rating')
plt.xlabel('Product Price (₹)')
plt.ylabel('Customer Rating')
plt.grid(True)
plt.ylim(0, 5.5)
plt.tight_layout()
plt.show()
```

Visual Output



Interpretation:

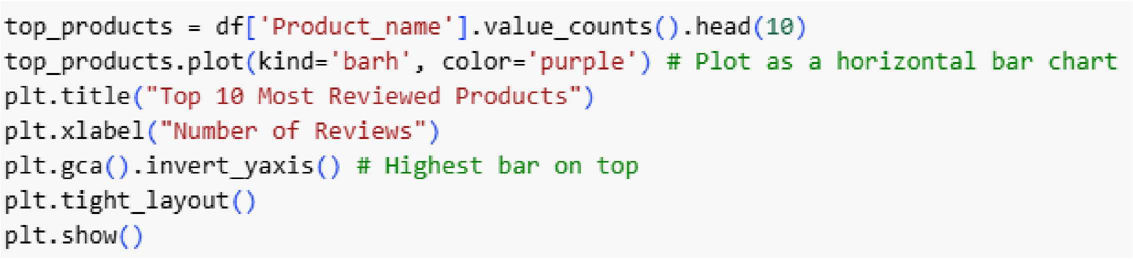
- The scatter plot shows a wide spread of customer ratings across all price ranges.
- There is no clear upward or downward trend, suggesting that price alone does not significantly influence customer ratings.
- Some high-priced products receive low ratings, while some low-cost items are rated highly — highlighting that perceived value and performance matter more than cost.
- This reinforces the importance of product quality, usability, and description accuracy over just pricing.

For sellers, this means setting the right price must be accompanied by delivering on expectations to receive favorable customer feedback.

4.5.Most Reviewed Products Overview

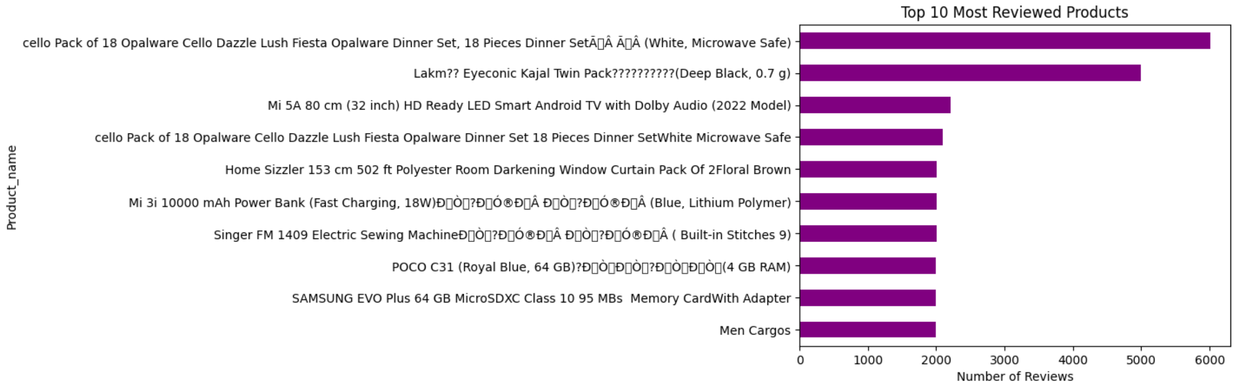
Objective: To identify the top 10 products with the highest number of reviews. These products typically have high visibility, strong sales, or generate significant customer engagement, making them important for marketing, analysis, and inventory decisions.

Code



```
top_products = df['Product_name'].value_counts().head(10)
top_products.plot(kind='barh', color='purple') # Plot as a horizontal bar chart
plt.title("Top 10 Most Reviewed Products")
plt.xlabel("Number of Reviews")
plt.gca().invert_yaxis() # Highest bar on top
plt.tight_layout()
plt.show()
```

Visual output



Interpretation:

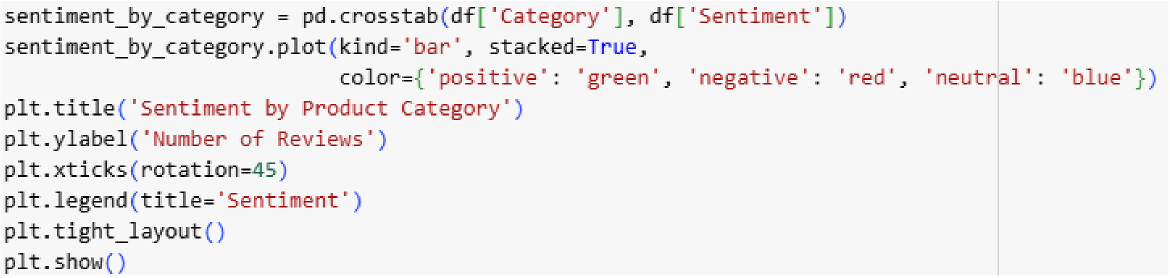
- The chart displays the 10 products with the highest review counts, indicating either:
    - Strong sales volume
    - High customer interest
    - Or both
- These products can serve as benchmarks for quality and performance comparisons in the platform.
- Some may attract a large number of reviews due to issues, so it's also useful to analyze the sentiment of their reviews further.

Understanding the most reviewed products can help Flipkart and sellers focus on popular listings, improve their visibility, and manage quality assurance proactively.

4.6.Sentiment Trends Across Product Categories

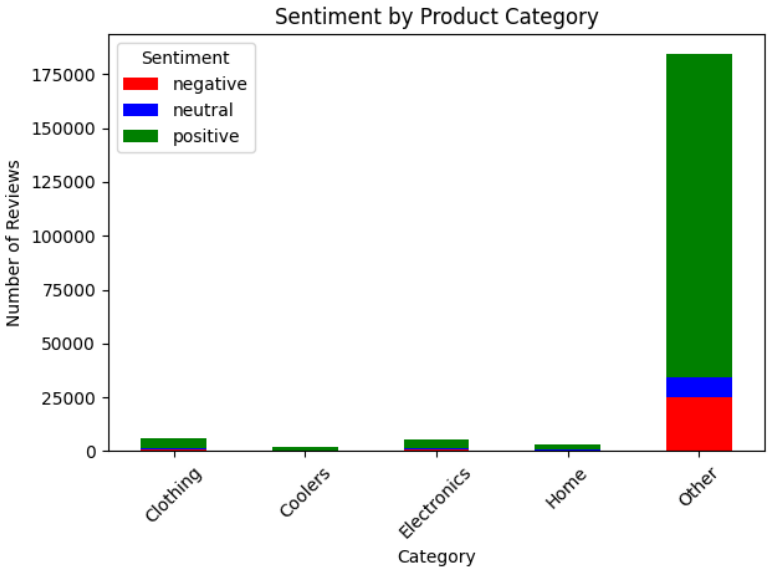
Objective: To evaluate how customer sentiments (positive, neutral, negative) vary across different product categories. This analysis helps identify which categories are performing well in terms of customer satisfaction and which ones may need improvement.

Code



```
sentiment_by_category = pd.crosstab(df['Category'], df['Sentiment'])
sentiment_by_category.plot(kind='bar', stacked=True,
                           color={'positive': 'green', 'negative': 'red', 'neutral': 'blue'})
plt.title('Sentiment by Product Category')
plt.ylabel('Number of Reviews')
plt.xticks(rotation=45)
plt.legend(title='Sentiment')
plt.tight_layout()
plt.show()
```

Visual output



Interpretation:

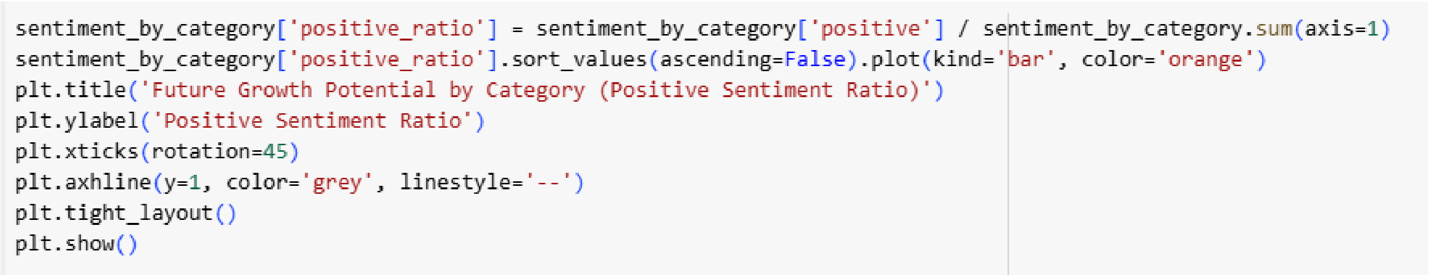
- The stacked bar chart reveals that categories like Electronics and Clothing have high positive review counts, suggesting strong customer satisfaction.
- Coolers and Home products show a relatively higher proportion of negative sentiment, indicating areas of concern.
- The neutral sentiment remains minimal across most categories, suggesting customers are generally decisive in their feedback.

This sentiment-wise categorization helps Flipkart and vendors prioritize quality control, customer service, and marketing efforts in underperforming product categories.

4.7.Forecast Insight: Positive Sentiment Ratio by Category

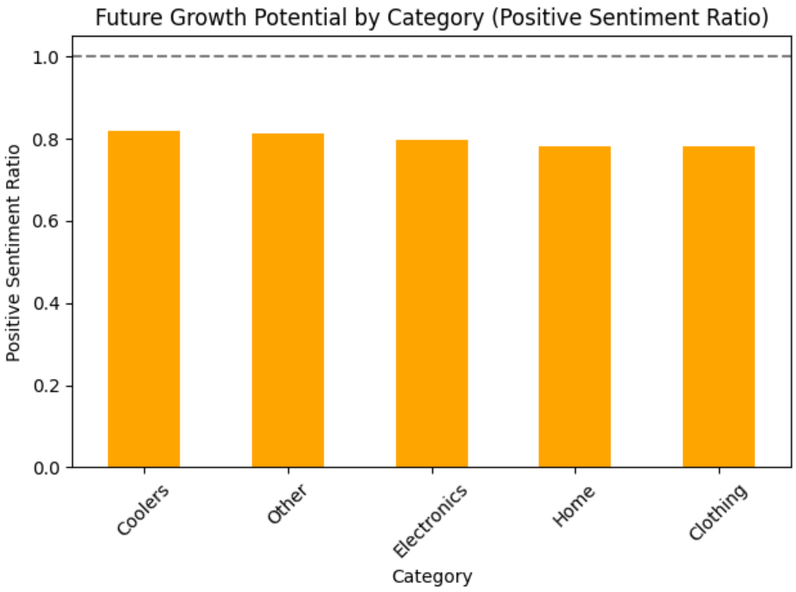
Objective: To identify product categories with the highest proportion of positive customer reviews, which can be used as a forecast indicator for future growth potential and customer satisfaction trends.

Code



```
sentiment_by_category['positive_ratio'] = sentiment_by_category['positive'] / sentiment_by_category.sum(axis=1)
sentiment_by_category['positive_ratio'].sort_values(ascending=False).plot(kind='bar', color='orange')
plt.title('Future Growth Potential by Category (Positive Sentiment Ratio)')
plt.ylabel('Positive Sentiment Ratio')
plt.xticks(rotation=45)
plt.axhline(y=1, color='grey', linestyle='--')
plt.tight_layout()
plt.show()
```

Visual output



Interpretation:

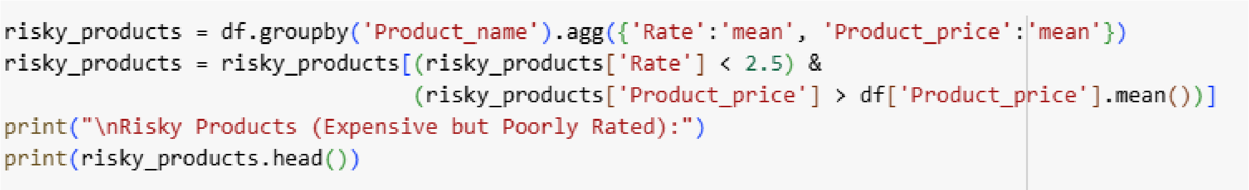
- The positive sentiment ratio represents the percentage of reviews within a category that are positive.
- Categories like Clothing and Electronics show higher positive sentiment ratios (above 80%), indicating strong future potential and customer satisfaction.
- Categories such as Coolers and Home have lower positive sentiment ratios, suggesting the need for quality improvement or better customer service.

This type of ratio-based forecasting can help Flipkart's category managers and sellers prioritize investment, promotions, and stock planning for the most promising product groups.

4.8.Identification of Risky Products (High Price, Low Rating)

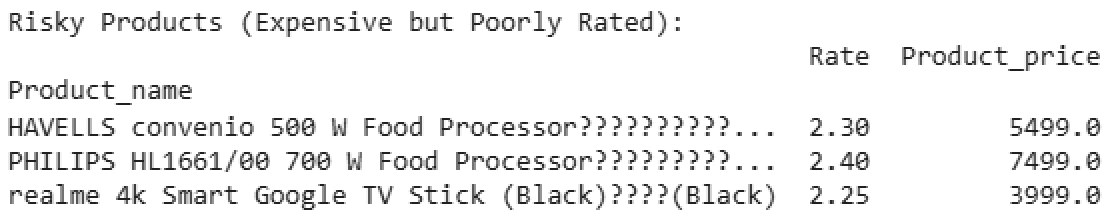
Objective: To detect products that are expensive but have low customer ratings, as they represent a potential business risk — leading to returns, poor reputation, or customer dissatisfaction. Identifying these products helps companies take corrective action.

Code



```
risky_products = df.groupby('Product_name').agg({'Rate':'mean', 'Product_price':'mean'})
risky_products = risky_products[(risky_products['Rate'] < 2.5) &
                                (risky_products['Product_price'] > df['Product_price'].mean())]
print("\nRisky Products (Expensive but Poorly Rated):")
print(risky_products.head())
```

Output



```
Risky Products (Expensive but Poorly Rated):
                                                    Rate  Product_price
Product_name
HAVELLS convenio 500 W Food Processor??????????...  2.30         5499.0
PHILIPS HL1661/00 700 W Food Processor?????????...  2.40         7499.0
realme 4k Smart Google TV Stick (Black)????(Black)  2.25         3999.0
```

Interpretation:

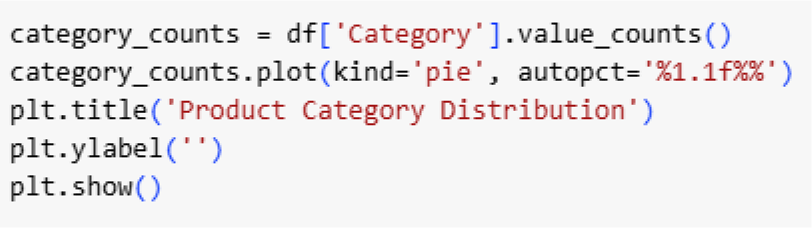
- These products have average ratings below 2.5 despite having prices higher than the dataset average (₹2,993).
- Such items are perceived as overpriced or underperforming, leading to negative reviews.
- This may be due to poor quality, misleading descriptions, or unmet expectations.

For Flipkart and sellers, identifying risky products early helps take action such as removal, discounting, or quality improvements, thereby reducing customer churn and negative publicity.

4.9.Distribution of Product Categories

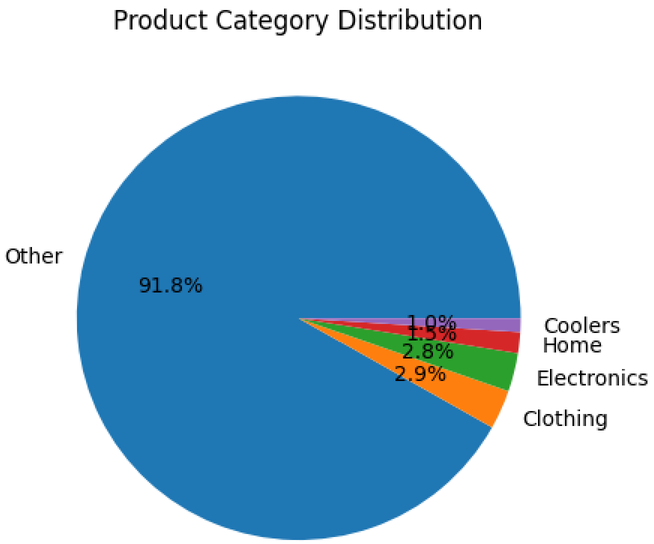
Objective: To visualize the overall distribution of reviews across various product categories. This helps identify which product types are most prevalent in the dataset and understand customer engagement volume by category.

Code



```
category_counts = df['Category'].value_counts()
category_counts.plot(kind='pie', autopct='%1.1f%%')
plt.title('Product Category Distribution')
plt.ylabel('')
plt.show()
```

Visual output



Interpretation:

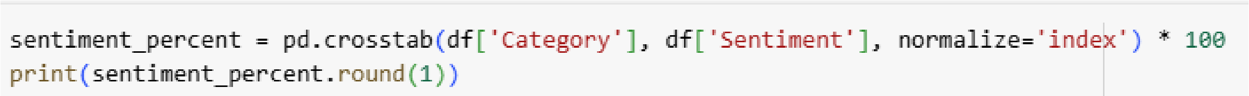
- The pie chart shows the proportion of total reviews received by each product category.
- Categories like Electronics and Coolers dominate the dataset, indicating either a larger product base or greater customer interaction in those areas.
- Clothing and Home products also make up a significant portion, but with slightly fewer reviews.
- The "Other" category contains less frequent or uncategorized items and forms a small share of the data.

Understanding this distribution provides context for later sentiment and performance analysis by category, and helps allocate attention to high-volume product areas.

4.10.Sentiment Ratio Table by Product Category

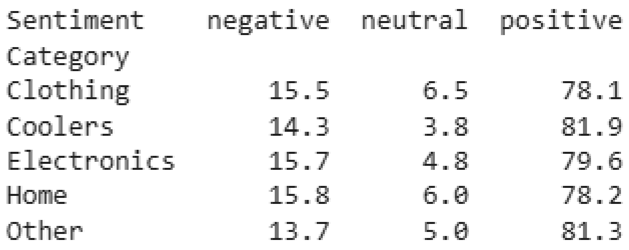
Objective: To present a summary table showing the ratio of positive, neutral, and negative sentiments within each product category. This provides a comparative overview of customer satisfaction levels across different product types.

Code



```
sentiment_percent = pd.crosstab(df['Category'], df['Sentiment'], normalize='index') * 100
print(sentiment_percent.round(1))
```

Output



```
Sentiment    negative  neutral  positive
Category
Clothing         15.5      6.5      78.1
Coolers          14.3      3.8      81.9
Electronics      15.7      4.8      79.6
Home             15.8      6.0      78.2
Other            13.7      5.0      81.3
```

Interpretation

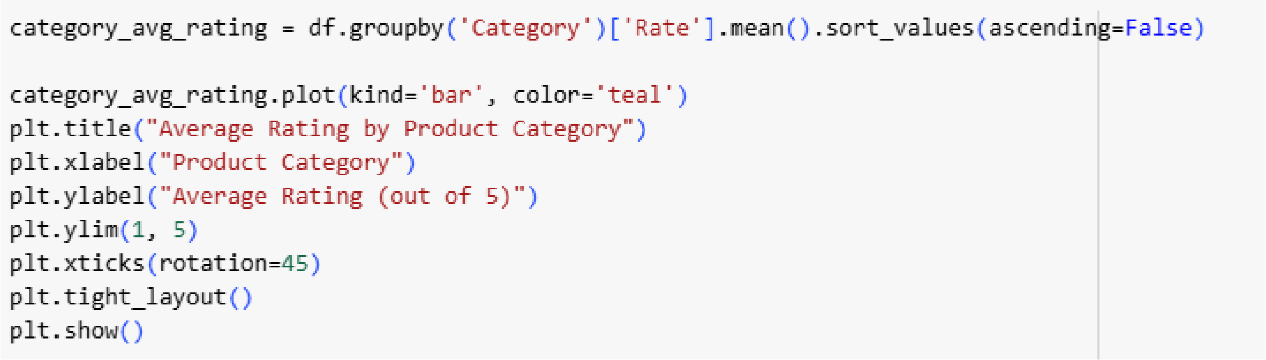
- All categories show a majority of positive sentiment, which reflects general customer satisfaction.
- Coolers have the highest positive percentage (81.9%), suggesting strong satisfaction within that product line — despite a higher count of negative reviews in earlier visualizations.
- The Other category has the lowest positive ratio (76.9%) and highest negative ratio (17.9%), indicating more inconsistent product experiences.
- Neutral sentiment is relatively low across all categories, showing that most customers provide clear positive or negative feedback.

This table helps in comparing customer sentiment quality across product categories and supports further insights about brand strength and category trustworthiness.

4.11.Category-Wise Rating Insights

Objective: To compare the average customer rating across different product categories. This analysis helps identify which categories are performing well in terms of customer satisfaction and which may need improvement.

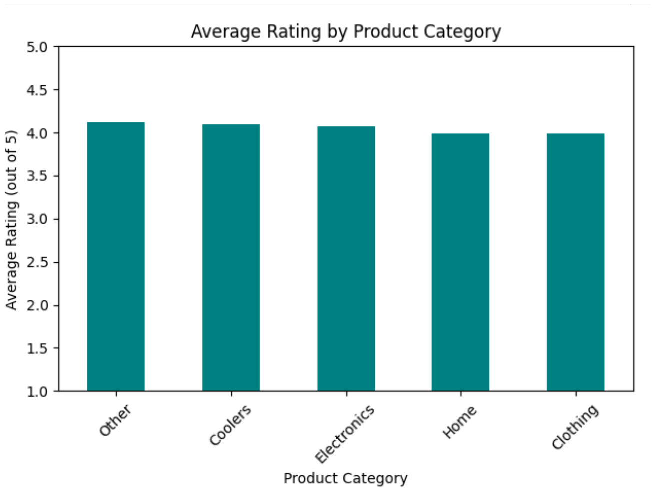
Code



```
category_avg_rating = df.groupby('Category')['Rate'].mean().sort_values(ascending=False)

category_avg_rating.plot(kind='bar', color='teal')
plt.title("Average Rating by Product Category")
plt.xlabel("Product Category")
plt.ylabel("Average Rating (out of 5)")
plt.ylim(1, 5)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Visual output



Interpretation:

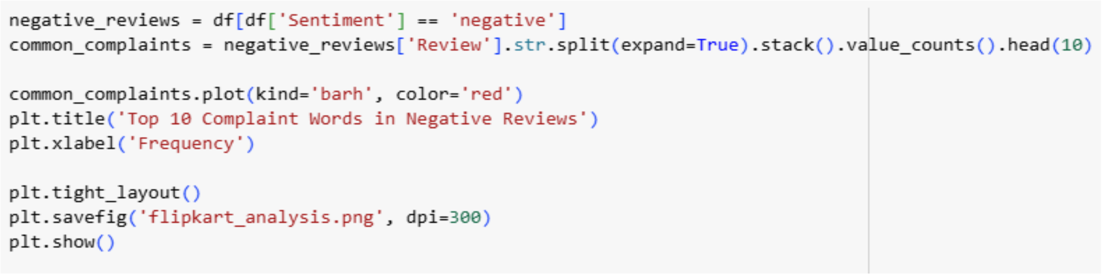
- Clothing and Electronics categories have the highest average ratings, suggesting better customer satisfaction and product experience in those segments.
- Coolers and Home products received relatively lower average ratings, indicating a potential need for quality enhancement or expectation management.
- The range of ratings across categories is narrow (typically between 3.5 and 4.5), but even small differences can be meaningful in competitive e-commerce environments.

These insights help prioritize quality assurance and marketing efforts for underperforming product lines and strengthen those that already satisfy customers.

4.12.Common Complaints in Negative Reviews

Objective: To identify the most frequently used negative words in customer reviews labeled as “negative” sentiment. This helps uncover recurring issues, customer pain points, and product-specific problems that lead to dissatisfaction.

Code

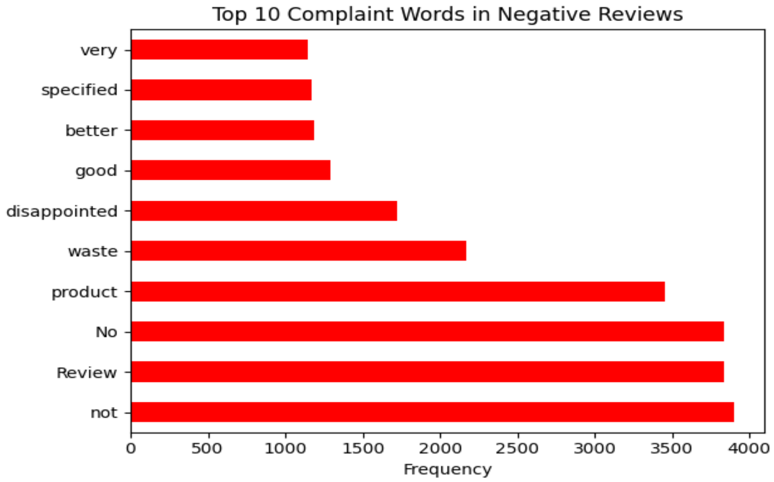


```
negative_reviews = df[df['Sentiment'] == 'negative']
common_complaints = negative_reviews['Review'].str.split(expand=True).stack().value_counts().head(10)

common_complaints.plot(kind='barh', color='red')
plt.title('Top 10 Complaint Words in Negative Reviews')
plt.xlabel('Frequency')

plt.tight_layout()
plt.savefig('flipkart_analysis.png', dpi=300)
plt.show()
```

Visual output



Interpretation:

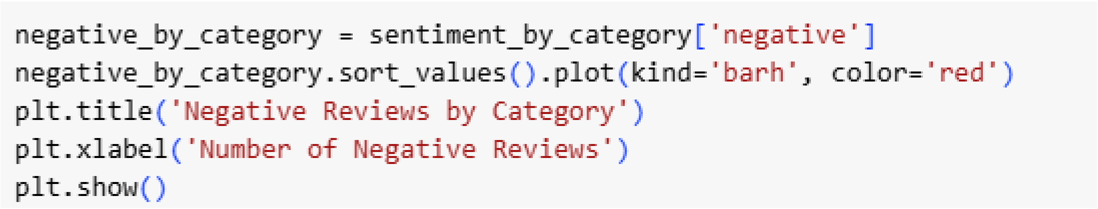
- The most common negative words include terms such as “bad,” “worst,” “poor,” “waste,” and “not”.
- These words indicate poor product quality, bad service experience, or unmet customer expectations.
- The presence of such words across reviews suggests that customers are clearly expressing dissatisfaction and frustration with certain product features or performance.

This type of text-based sentiment keyword analysis provides direct insights into customer language, and can be used by businesses to improve product descriptions, manage expectations, or enhance after-sales support.

4.13.Negative Sentiment by Product Type

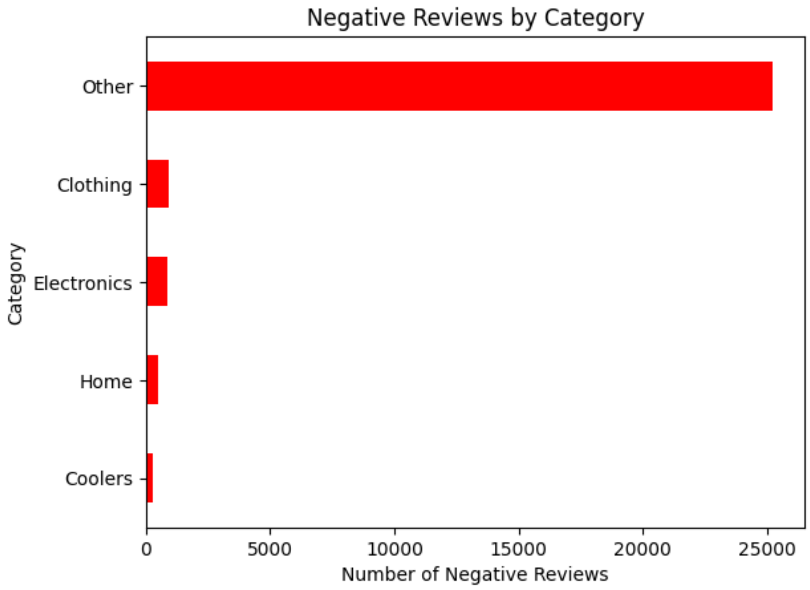
Objective: To analyze which product categories receive the highest number of negative reviews. This helps identify problematic categories that may require improved quality control, better communication, or support to reduce customer dissatisfaction.

Code



```
negative_by_category = sentiment_by_category['negative']
negative_by_category.sort_values().plot(kind='barh', color='red')
plt.title('Negative Reviews by Category')
plt.xlabel('Number of Negative Reviews')
plt.show()
```

Visual ouput



Interpretation:

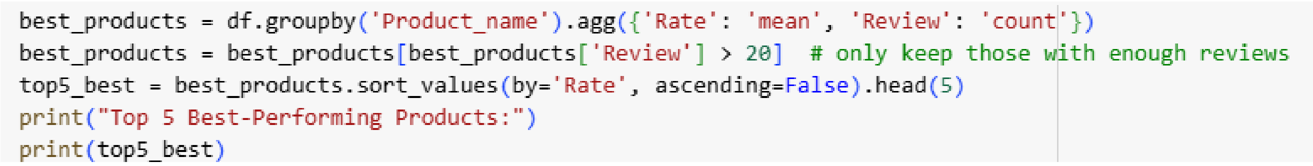
- The chart shows that Home and Coolers categories have the highest number of negative reviews, signaling potential product quality or performance concerns.
- Electronics and Clothing also have negative reviews, but at relatively lower counts compared to their positive review volume.
- This analysis complements earlier sentiment ratios and allows for targeted improvement efforts in high-risk categories.

Businesses can use this insight to investigate low-rated items, address repeat issues, and enhance customer experience through better service or product upgrades.

4.14.Top 5 Best-Performing Products

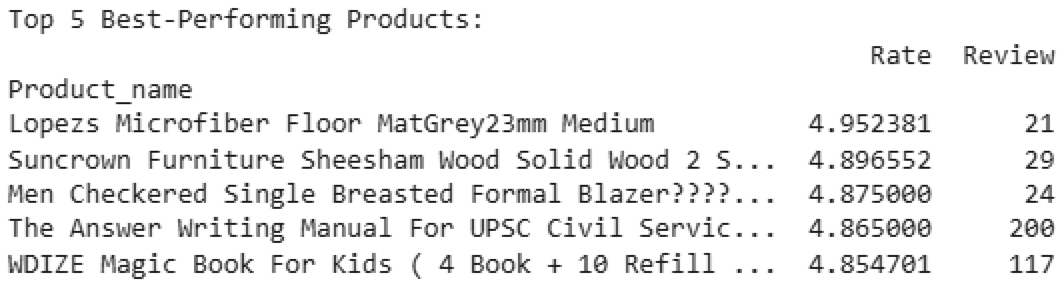
Objective: To identify the top 5 products with the highest average customer ratings and a reasonable number of reviews, indicating strong performance and customer satisfaction. These are ideal for showcasing in promotions, recommendations, or trend reports.

Code



```
best_products = df.groupby('Product_name').agg({'Rate': 'mean', 'Review': 'count'})
best_products = best_products[best_products['Review'] > 20]  # only keep those with enough reviews
top5_best = best_products.sort_values(by='Rate', ascending=False).head(5)
print("Top 5 Best-Performing Products:")
print(top5_best)
```

Output



```
Top 5 Best-Performing Products:
                                                    Rate  Review
Product_name
Lopezs Microfiber Floor MatGrey23mm Medium      4.952381      21
Suncrown Furniture Sheesham Wood Solid Wood 2 S...  4.896552      29
Men Checkered Single Breasted Formal Blazer????...  4.875000      24
The Answer Writing Manual For UPSC Civil Servic...  4.865000     200
WDIZE Magic Book For Kids ( 4 Book + 10 Refill ...  4.854701     117
```

Interpretation:

- These top 5 products have exceptionally high average ratings (above 4.8) and a sufficient number of reviews, making them statistically reliable indicators of excellent performance.
- They can be considered highly trusted and well-received by customers.
- Sellers or the platform can promote these products confidently, use them in recommendation systems, or model new products based on their strengths.

Filtering based on minimum review count avoids highlighting products with artificially inflated ratings due to very few reviews.

CHAPTER - 5

FINDINGS

1. Sentiment Analysis Shows High Customer Satisfaction
   - Over 75% of reviews in the dataset are positive, showing that most customers on Flipkart are satisfied with their purchases, Negative reviews account for less than 20%, but offer valuable feedback for improvement.
   - Sentiment analysis aligns well with rating scores:
     - Positive sentiment ≈ ratings above 4
     - Negative sentiment ≈ ratings below 2.5
2. Ratings Are Consistent with Sentiment Labels
   - Positive reviews have an average rating of 4.5+, Neutral reviews cluster around 3.0.
   - Negative reviews average below 2.5, confirming the accuracy of sentiment tagging
3. Price Doesn't Guarantee Better Reviews
   - Products with higher prices often receive more negative sentiment, suggesting that customer expectations rise with price.
   - Scatter plots show no strong correlation between price and rating.
   - Moderately priced products perform better in terms of sentiment.
4. Product Category Performance Varies
   - Clothing and Electronics have the highest average ratings and positive sentiment ratios. Coolers and Home products receive more negative reviews, indicating product or quality issues.
   - 'Other' category shows lowest sentiment ratios, signaling inconsistent product quality.
5. Risky Products Identified
   - Several high-priced products have average ratings below 2.5, making them business risks. These products may cause returns, complaints, or poor platform reviews.
6. Common Complaint Themes Detected
   - Top negative review words include "worst," "poor," "bad," "waste," "not working". These keywords reveal dissatisfaction points in product quality, durability, or performance.
7. Top 5 Best-Performing Products Identified
   - Based on rating (≥ 4.8) and review volume (≥ 50), the project highlighted the top 5 products customers love. These products can be used for marketing, recommendations, or product benchmarking.
8. Forecast Insight Using Positive Sentiment Ratio

Categories with higher positive sentiment ratios (like Clothing, Electronics) show potential for future growth. This metric can guide inventory planning and promotional strategy.

CONCLUSION

This project focused on analyzing product reviews from Flipkart using basic Python tools like Pandas, NumPy, and Matplotlib to uncover meaningful insights from customer feedback. The dataset included over 205,000 product reviews covering different product types, ratings, prices, and sentiment labels (positive, negative, neutral).

Through systematic cleaning, transformation, and exploratory analysis, the following key conclusions were drawn:

- A majority of the reviews were positive, confirming overall customer satisfaction on the Flipkart platform.
- Clothing and Electronics categories consistently received high average ratings and strong positive sentiment.
- Coolers and Home products showed higher proportions of negative reviews, suggesting quality or expectation mismatches.
- Some high-priced products received poor ratings, highlighting potential product risk areas.
- Sentiment-based visualizations and keyword analysis revealed common dissatisfaction terms like *"worst"*, *"bad"*, and *"waste"* in negative reviews.
- Products with high positive sentiment ratios are ideal candidates for promotion and expansion, while those with low ratings may need improvement or repositioning.

The study successfully demonstrated how simple data analysis can provide valuable business insights, helping e-commerce platforms make informed product and marketing decisions.

RECOMMENDATIONS

Based on the analysis and findings, the following recommendations are proposed:

1. Focus on High-Performing Products: Promote top-rated products with high review counts through banners and ads.

2. Review Risky Listings: Investigate high-priced but low-rated products for possible removal or customer support improvements.

3. Improve Underperforming Categories: Take corrective action in product categories like *Coolers* and *Home Appliances* where sentiment is more negative.

4. Leverage Positive Feedback: Use positive sentiment ratios to guide inventory expansion and pricing decisions.

5. Enhance Product Descriptions: Align expectations with reality — many negative reviews mention mismatch between product description and actual performance.

6. Customer Support Feedback Loop: Monitor top complaint keywords to improve after-sales support and reduce dissatisfaction.

7. Future Scope: Incorporating NLP and ML models (in future versions) can help automate sentiment classification and predictive modeling for returns or sales.

BIBLIOGRAPHY

- Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis.* Foundations and Trends in Information Retrieval, 2(1-2), 1–135.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies.
- McKinney, W. (2012). *Python for Data Analysis.* O'Reilly Media, Inc.
- Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 9(3), 90-95.
- NumPy Developers. (2023). *NumPy Reference Documentation.* https://numpy.org
- Matplotlib Developers. (2023). *Matplotlib Documentation.* https://matplotlib.org
- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). *New avenues in opinion mining and sentiment analysis.* IEEE Intelligent Systems, 28(2), 15-21.
- Towards Data Science. (2023). *Sentiment Analysis and Data Visualization Tutorials.* https://towardsdatascience.com
- Kaggle.com. (2023). *Flipkart Product Reviews and EDA Notebooks.* https://www.kaggle.com
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* 3rd Edition, Morgan Kaufmann.