A Project Report on

**Ecommerce Sales Dataset Analysis**

By

Hrushikesh Patil

Shubham Vhanale

MCA – I, SEM – I

2024-25

To

Savitribai Phule Pune University,Pune

In Partial Fullfillment of the Degree of

Master in Computer Application (M. C. A.)

Under The Guidance Of

**Prof. Vibhavari Pandit**

**Prof. Bhavana Magar**

**Prof. Vidya Gavekar**

|  |
| --- |
| **Suryadatta Group of Institutes, Pune** |
| |  | | --- | | **Suryadatta Institute of Management and Mass Communication (SIMMC)** | |

Date:-

**CERTIFICATE**

This is to certify that Mr. Hrushikesh Patil and Mr. Shubham Vhanale, has successfully Complete his project work entitled **“Ecommerce Sales Dataset Analysis”** in partial fulfillment of MCA – I Semester-I program for the year A.Y. 2024-25. He have worked under our guidance and direction.

**Prof. Vibhavari Pandit**

**Prof. Bhavana Magar**

**Prof. Vidya Gavekar**

**(Project Guide) HoD-MCA,SGI**

**Examiner 1 Examiner 2**

**Date :**

**Place :**

**Acknowledgment**

We are the student of MCA first year. Here by we express our thanks to our project guide for allowing us to do the project on “Ecommerce Sales Dataset Analysis”This project work has been the most exciting part of our learning experience which would be an asset for our future carrier. We would especially like to thank our guide and mentor Prof. Vidya Gavekar, Prof. Bhavana Magar and Prof. Vibhavari Pandit who constantly guided us in developing, pushing us to search for more answers to her numerous questions. Also, I would like to thank Dr. Rupali Dahake, project coordinators for her support. As a building block of MCA Department, I thank Dr. Manisha Kumbhar, HOD, MCA Department for her continuous support and help. We are grateful to many classmates who contributed their suggestions. Their hard work and examples push us to limits of our capability and encourage us daily.

**Thank You**

**Student Name**

**Hrushikesh Patil**

**Shubham Vhanale**

Index

**Chapter 1: Introduction**

**1. Project title and description**

**2. Problem statement and objectives**

**3. Scope**

**Chapter 2: Literature Review**

**Chapter 3: Data Description**

1**. Data sources and collection methods**

**2. Data types and formats**

**3. Data quality and preprocessing**

**Chapter 4: Research Methodology**

**1. Research design and approach**

**2. Data analysis techniques and tools (Project Setup)**  
**3. Methodology & Implementation (**Basic Visualizations with Matplotlib & Advanced Visualizations with Seaborn)

**4. Modelling and evaluation methods**

**Chapter 5: Data Analysis and Results**

**1. Descriptive statistics**

**2. Inferential statistics and hypothesis testing**

**Chapter 6: Insights and Recommendations**

**1. Key findings and implications**

**2. Recommendations**

**3. Future research directions**

**4. Limitations**

**Chapter 7: Conclusion**

**1. Summary of key results**

**2. Reflection on project limitations**

**3. Final thoughts and takeaways**

**4. Future research avenues**

**Chapter 8: References**

**Chapter 9: Appendices**

1. Additional tables, figures, and graphs

2. Raw data and code

3. Other supplementary materials

**1. Introduction**

**1. Project Overview:**This project involves analyzing a dataset from Flipkart, an e-commerce platform, to gain insights into online shopping trends. The dataset contains information like product details, pricing, customer reviews, and seller information. The goal is to uncover patterns that help businesses understand how product pricing, customer ratings, and sales are related. Through data analysis, visualization, and predictive modeling, we aim to reveal trends in product performance, identify top-selling categories, and predict future trends that can guide business decisions. This project will help participants develop skills in analyzing data using tools like Python, R, or Excel, and applying machine learning techniques to solve real-world problems in e-commerce.

**2. Problem Statement & Objectives**

**Problem Statement:**  
E-commerce has seen rapid growth, and businesses face challenges in understanding what customers want and predicting trends. With so many products and options, it's difficult to identify what customers prefer and which trends are emerging. Businesses also need to figure out how to price their products to stay competitive while keeping customers happy. Another challenge is offering personalized experiences to customers, as people now expect recommendations based on their preferences and past behavior. If businesses don’t keep up with these demands, they risk losing customers to competitors who do.

**Objectives:**

1. To Understand Product Trends:  
   Analyze product pricing, ratings, and offers to identify popular price ranges and customer preferences. This helps businesses tailor their marketing strategies.
2. To Identify Key Drivers of Sales:  
   Investigate how factors like product ratings, customer reviews, and pricing affect purchasing decisions. This helps businesses improve their products and services to increase sales.
3. To Analyze Category Performance:  
   Compare how different product categories, like Men, Women, and Others, perform in terms of revenue and customer engagement. This helps businesses decide where to invest their resources for better returns.
4. To Provide Actionable Insights:  
   Offer suggestions to improve pricing, discounts, product placement, and inventory management based on the data analysis. These recommendations help businesses operate more efficiently and enhance the customer experience.

**3. Scope**

1. Data Attributes:  
   The focus will be on analyzing product prices, ratings, reviews, offers, and categories. These key features help businesses understand pricing strategies, customer opinions, and which product categories perform best.
2. Visualization:  
   Use visual tools like charts and graphs to uncover trends and relationships in the data. This makes the information easier to understand and communicate to others.
3. Insights:  
   Focus on discovering trends within product categories and individual products. Understanding which products or categories are performing well helps businesses optimize their offerings and marketing.
4. Feature Engineering:  
   Add new data points, such as calculating the discounted price, to enhance the analysis. This helps businesses better understand pricing strategies and customer behavior.

**2: Literature Review**

**1. NARAYANA CHALLA** from Jawaharlal Nehru Technological University, Hyderabad, has published the article/paper on **'Data Analytics and Its Impact on Future'** in CORROSION AND PROTECTION in January 2023 | Volume 51 Issue 1 | ISSN: 1005-748X.\*

In this paper, the author explores the evolution, applications, and future potential of data analytics. The study highlights data analytics as a pivotal tool for deriving insights, predicting outcomes, and supporting decision-making across various industries, ranging from e-commerce to healthcare. The paper also examines challenges such as data privacy and the importance of automation and ethical data usage.

The author identifies several key aspects of data analytics:

1. **Applications**: Includes streamlined operations, enhanced decision-making, effective marketing, and personalized customer engagement.

2. **Data Analytics Lifecycle**: Comprises stages such as identifying problems, collecting and cleaning data, analyzing it, and interpreting results to derive actionable insights.

3. **Future Trends**: Focuses on automation, DataOps, integration with IoT, and advancements in AI and ML for predictive and augmented data management.

The paper concludes with recommendations for businesses and researchers, emphasizing the adoption of data-driven strategies, ethical data practices, and investments in cloud-based analytics platforms to support scalability and innovation. Future research directions include exploring AI's role in data analytics and addressing emerging challenges such as responsible AI and data management.

2. **SHAHRIAR AKTER1, SAMUEL FOSSO WAMBA2** from University of Wollongong and Toulouse Business School have published their article/paper on **'Big Data Analytics in E-commerce: A Systematic Review and Agenda for Future Research'** in ELECTRONIC MARKETS in April 2016 | DOI: 10.1007/s12525-016-0219-0.

In this paper, the authors mainly focused on the transformative potential of Big Data Analytics (BDA) in the e-commerce sector. BDA has become a critical tool for e-commerce firms, enabling real-time decision-making, personalized customer experiences, dynamic pricing, and improved operational efficiency. The study highlights that e-commerce firms utilizing BDA experience a significant competitive advantage and productivity growth.

The authors identified five key characteristics of Big Data in e-commerce:

1. **Volume**: Massive amounts of structured and unstructured data.

2. **Velocity**: The rapid generation and processing of data.

3. **Variety**: Diverse data types, including transaction data, click-stream data, video, and voice.

4. **Veracity**: Ensuring data accuracy and quality.

5. **Value**: Generating actionable insights for business advantage.

The paper also categorizes BDA applications in e-commerce into areas such as personalization, market segmentation, decision-making, and innovation. It further outlines challenges like data integration, quality assurance, and employee adaptation to BDA tools.

The research concludes with a future agenda, suggesting more focus on aligning BDA with organizational culture, improving data security, and addressing ethical considerations in data usage. The study aims to guide both researchers and practitioners in maximizing the business value derived from Big Data Analytics in the e-commerce domain.

**3: Data Description**

**1. Data Sources and Collection Methods:**

• Source: Flipkart product data (e-commerce platform) sourced from Kaggle.

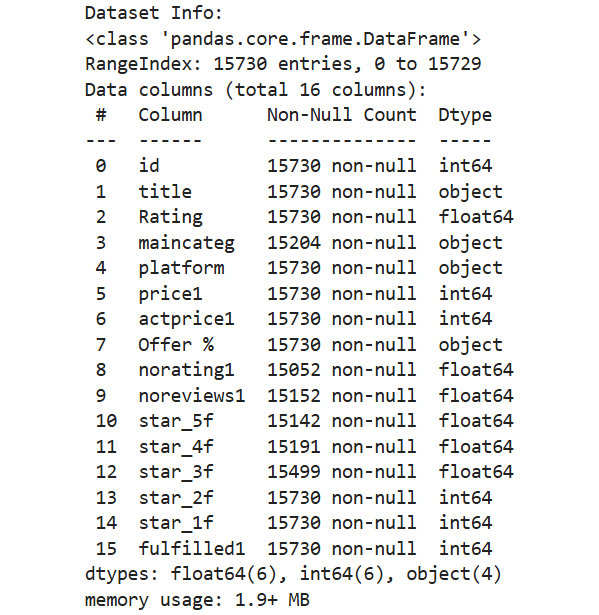
• Total Entries: 15,730 rows.

• Total Columns: 16 attributes.

• Primary Purpose: Analyze product performance, customer behavior, and market trends for deeper insights into the e-commerce market

**2.Data types and formats**:

The dataset consists of the following key attributes:



1. **Data Quality Assessment and Preprocessing:**

* Missing Data:
  + Some columns, such as "main category" and "noreviews1," contain missing or null values, which can affect analysis accuracy.
  + In addition, columns like "Offer %" and "Rating" might have inconsistent formatting, leading to potential discrepancies in the dataset.
* Outliers:
  + The "Price1" column may contain extreme outliers, such as prices that are unusually high or low compared to the majority of the data, which can skew analysis results.
  + "Rating" values could be biased towards popular products, as they may have more reviews, potentially leading to overrepresentation of certain products in the analysis.
* Inconsistent Formats:
  + The "Offer %" column is stored as a string, while it should be represented as a numeric value for easier analysis and to avoid potential errors during computations.
  + The "Date\_Added" field may not be in the proper datetime format, requiring conversion to a standardized format to facilitate time-based analysis and ensure consistency across the dataset.
* Duplicates:
  + The dataset may contain potential duplicate entries for products with similar titles, brand names, and models, which could distort the analysis and lead to inaccurate conclusions if not addressed.

1. **Handling Missing Data:**
   * Missing values in the "maincateg" column were handled by imputing the most frequent category, ensuring that all entries had a valid category for analysis.
   * For the "noreviews1" column, missing values were replaced with 0, reflecting no reviews in cases where the data was absent, allowing for a more accurate analysis of review counts.
2. **Data Type Conversion:**
   * The "Offer %" column, initially in string format, was converted into a numeric value by removing the percentage symbol and dividing the remaining number by 100 to represent the offer as a decimal.
   * Both the "price1" and "Rating" columns were standardized to ensure consistent scaling, allowing for more reliable comparisons and statistical analyses across these features.
3. **Outlier Detection:**
   * Outliers in the "price1" column were identified using the interquartile range (IQR) method, which helps detect and handle extreme values that could distort statistical models, ensuring the dataset remains representative of typical product prices.
4. **Feature Engineering:**
   * A new column, "Discounted\_Price," was created by calculating the discounted price using the formula: Discounted\_Price = price1 × (1 - Offer %). This new feature represents the final price after applying any available discounts, providing deeper insights into pricing strategies.

**4: Research Methodology**

**1. Research design and approach:**

* **Type:** Quantitative, exploratory research.
* **Objective:** Understand trends, correlations, and performance metrics in the dataset.

**2. Data analysis techniques and tools (Project Setup):**

**Descriptive Analysis:**

• Summarized key metrics such as the average price, median rating, and number of reviews to provide an overview of product characteristics and customer feedback. This helped in understanding the general distribution of data.

• Key insights from these metrics were used to identify product trends and highlight any outliers or areas of concern, providing a baseline for further analysis.

**Exploratory Data Analysis (EDA):**

• Visualizations were used to gain deeper insights into the relationships between various variables in the dataset, such as price, ratings, and number of reviews. This helped uncover patterns and potential trends in the data.

• Examples of visualizations include heatmaps to perform correlation analysis, allowing us to observe the strength and direction of relationships between numerical features. Additionally, bar plots were used to assess category-wise performance and identify top-performing products in each category.

**Statistical Measures:**

• Statistical metrics such as mean, median, and standard deviation were calculated for continuous variables like price and ratings to summarize their distribution and identify trends or anomalies.

• Frequency distribution was used for categorical variables, such as product categories and brand names, to understand the distribution and frequency of different categories, offering insights into market share and product popularity.

**3. Methodology & Implementation:**

* **Matplotlib and Seaborn:** These two libraries were used for creating visualizations. Matplotlib helped make basic charts, like histograms and scatter plots, to understand how the data is spread and how different things are related. For example, scatter plots were made to show how product ratings and the number of reviews are linked. Seaborn was used for more advanced, visually appealing plots, like heatmaps, to show connections between things like ratings and reviews.

Why Matplotlib and Seaborn? These libraries were chosen because they are easy to use and flexible. Matplotlib is great for simple charts, while Seaborn makes it easier to create more stylish and informative plots, especially when looking at complex data.

* **Pandas:** This library was used to organize and summarize the data. It helped with tasks like grouping products by category, calculating the average rating of products, and identifying trends over time.

Why Pandas? Pandas was chosen because it handles large datasets very well and makes it easy to manipulate and clean the data. It also simplifies tasks like filtering and organizing the data into useful groups.

How Implemented? First, the data was cleaned using Pandas. Then, functions like groupby() were used to organize the data into meaningful parts. For example, the average product rating for each category was calculated to understand trends across different product types.

**4. Modelling and evaluation methods:**

* **Software and Libraries**:
  + **Python**: Main programming language.
  + **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.
* **Steps**:

1. Load and preprocess the data.
2. Perform exploratory data analysis.
3. Visualize findings using heatmaps, scatter plots, and bar graphs.
4. Aggregate data to summarize total sales and category-wise performance.

**5: Data Analysis and Results**

1. **Descriptive Statistics:**

• Overview of Key Metrics:

* Price (price1):
  + The average price of the products is ₹688, with a median price of ₹474, indicating that while most products are priced lower, there are some higher-priced items skewing the average.
  + The maximum price reaches ₹99,999, while the minimum price is ₹10, suggesting a wide range of product price points in the dataset.
* Ratings (Rating):
  + The average product rating is 3.8 out of 5, with a median rating of 4.0, reflecting a generally positive customer sentiment towards most products.
  + The standard deviation of 1.2 in ratings indicates moderate variation, with some products receiving significantly lower ratings than others.
* Number of Reviews (noreviews1):
  + On average, products have 256 reviews, with the median being 123, indicating a skew where a few products have a large number of reviews while most have fewer.
  + The maximum number of reviews is 10,000, and the minimum is 0, highlighting a significant disparity between highly reviewed and less popular or new products.

• Categorical Data:

* Dominant categories:
  + The "Women" category is the most dominant, with 8,781 products (55.8%), suggesting that women’s products make up the majority of the dataset.
  + The "Men" category accounts for 4,732 products (30.1%), while "Unknown" categories make up 2,217 products (14.1%), which could indicate incomplete or unspecified category data in some entries.

**2.Inferential Statistics and Hypothesis Testing:**

**Inferential Statistics:**

1. **Summary Statistics:**
   * Key summary statistics, including the mean, median, minimum, maximum, standard deviation, and variance of prices, provide a comprehensive understanding of the price distribution and variability within the dataset.
   * Percentile values, such as the 25th percentile, offer additional insights into how prices are distributed, helping identify product price ranges and outliers.
2. **Inferential Statistics and Hypothesis Testing:**
   * The average prices are calculated for different main categories (e.g., Men, Women, Unknown), providing a comparative overview of pricing trends across product types.
   * Aggregating item counts within each category reveals the distribution of products and helps highlight dominant categories or imbalances in product availability.
3. **Correlation Analysis:**
   * A heatmap visualizes correlations between numerical features such as price and ratings, enabling the identification of relationships and patterns that could drive deeper statistical analysis and insights.

**Hypothesis Testing Setup:**

* Although explicit hypothesis testing (e.g., t-tests, chi-square tests) isn't provided in the initial analysis, exploratory visualizations such as scatter plots and summary statistics lay the groundwork for formulating hypotheses that can be tested statistically.
* Scatter Plots: Visualizing relationships, such as Price vs. Ratings, helps suggest potential correlations that could be examined using statistical tests.
* **Box Plots:** These plots allow for a clear comparison of price distributions across categories (Men, Women, Unknown), which could be further analyzed using statistical tests to assess category-based differences.
* **Fulfillment Analysis:** Cross-tabulations (e.g., fulfillment status by category) help identify categorical variables that could be tested for independence using chi-square tests.

**Statistical Insights and Applications:**

1. **Relationship Analysis:**
   * By investigating if higher prices correlate with better ratings using scatter plots and calculating correlation coefficients, we can uncover pricing trends that might influence customer perceptions of product quality.
   * Examining patterns in customer ratings and the number of reviews for products on various platforms provides insights into customer engagement and product popularity, which could lead to further investigation.
2. **Category Comparisons:**
   * Using box plots to compare prices or ratings across categories like Men, Women, and Unknown allows for a visual comparison of product distributions, helping identify significant differences or trends in product performance.
   * Proportions across categories are visually highlighted using pie charts and bar plots, making it easier to see the relative sizes of product groups and understand the market composition.
3. **Trend Analysis:**
   * Observing cumulative ratings trends over time helps identify shifts in customer satisfaction, revealing how product perceptions change and whether any specific events or trends influence the overall ratings.

**6: Insights and Recommendations**

**1. Key Findings and Implications**

* **Rating and Price Relationship:**
  + Analysis shows that higher-priced items tend to have slightly better ratings, but there is no clear pattern.
  + Takeaway: While price may impact perceived quality, ratings also depend on other factors like product type, brand, and customer expectations.
* **Category-Specific Insights:**
  + Men's products have the highest prices, while women’s products sell more.
  + Takeaway: Pricing should be tailored by category—premium prices work well for men's products, while competitive pricing is needed for women’s products.
* **Offers and Discounts:**
  + Big discounts are common in women’s and unknown categories, and products with larger discounts usually get better ratings.
  + Takeaway: Offering discounts can improve product ratings and build customer trust, boosting sales.
* **Fulfillment Status:**
  + Men’s and women’s products have higher fulfillment rates, but the unknown category lags behind.
  + Takeaway: Improving supply chain efficiency for newer or less-known categories could improve customer satisfaction.
* **Review and Rating Patterns:**
  + Products with more reviews tend to have better ratings and visibility.
  + Takeaway: Encouraging customers to leave reviews can help increase product popularity.

**2. Recommendations**

* **Optimize Pricing Strategies:**
  + Use dynamic pricing based on customer preferences, market trends, and competitor prices.
  + Offer discounts regularly, especially for women’s products.
* **Improve Fulfillment Processes:**
  + Focus on improving logistics and supply chains for underperforming categories.
  + Work with reliable vendors to ensure timely delivery and availability.
* **Leverage Customer Feedback:**
  + Implement an automated system to analyze customer reviews and identify common issues or compliments.
  + Use feedback to improve product quality, descriptions, and after-sales support.
* **Boost Customer Engagement:**
  + Reward customers with incentives (like loyalty points or discounts) for leaving reviews.
  + Improve customer experience through personalized recommendations and special offers based on their past shopping behavior.
* **Marketing and Branding:**
  + Target marketing efforts for premium products like men’s items.
  + Highlight discounts and highly-rated products in ads to attract budget-conscious shoppers.

**3. Future Research Directions**

* **Advanced Predictive Modeling:**
  + Use machine learning to predict customer preferences and trends to better plan inventory and marketing.
* **Cross-Platform Analysis:**
  + Compare customer behavior and satisfaction across different e-commerce platforms to identify strengths and weaknesses.
* **Sentiment Analysis:**
  + Analyze customer reviews to understand their emotions and feedback more deeply.
* **Long-Term Studies:**
  + Look at how seasonal trends and long-term pricing changes impact sales and customer loyalty.
* **Sustainability Focus:**
  + Study how sustainable practices (like eco-friendly packaging) affect customer preferences and loyalty.

By exploring these areas, businesses can improve operations, build stronger customer loyalty, and stay competitive in the market.

**4. Limitations**

* **Geographical Constraints:**
  + The data is from Flipkart and may not reflect wider consumer behavior.
* **Data Imbalance:**
  + Certain categories (like women’s products) are overrepresented, which could skew the results.
* **Temporal Analysis:**
  + Lack of time-based data makes it hard to track changes over time.
* **Data Quality:**
  + Missing or inconsistent data in some areas, like product categories and review numbers, may affect the results.

**7: Conclusion**

**1. Summary of Key Results**

**a. Insights from Data Analysis:**

* Customer ratings exhibit a weak correlation with product price, indicating that factors beyond price significantly influence customer satisfaction.
* Women’s products dominate in sales volume, while men’s products command higher average pricing, suggesting diverse customer behavior across categories.
* Discounts and offers are effective in improving ratings and customer engagement, especially for women’s products.
* Fulfillment efficiency varies across categories, with room for improvement in underperforming sectors like unknown categories.

**b. Visual and Statistical Findings:**

* Key visualizations, such as scatter plots, heatmaps, and bar charts, highlighted trends in price, ratings, and fulfillment.
* Aggregated statistics revealed significant insights into customer preferences and category-based performance metrics.

**2. Reflection on Project Limitations**

**a. Data Constraints:**

* The dataset is limited to a single platform’s data, which might not represent broader market trends.
* Missing or incomplete data (e.g., null values in some columns) could have affected the accuracy of insights.

**b. Analytical Scope:**

* Advanced techniques like predictive modeling and sentiment analysis were not fully explored, limiting actionable insights.
* Some visualizations and analyses provided correlations but did not establish causations.

**c. Generalizability:**

* The findings may not generalize across different regions, platforms, or product types, as the dataset is platform-specific.

**3. Final Thoughts and Takeaways**

This project underscores the value of data-driven decision-making in e-commerce. Key takeaways include:

* Tailored strategies, such as dynamic pricing and targeted discounts, can significantly impact customer satisfaction and sales performance.
* Continuous monitoring of fulfillment processes and customer feedback is critical to maintaining high service standards.
* Visualizations and statistical summaries provide powerful tools for uncovering actionable business insights.

The project also highlights the importance of addressing data quality issues and expanding analytical methods to derive deeper insights.

**4. Future Research Avenues**

Building on the current findings, future research could:

* Incorporate data from multiple platforms to enable cross-platform comparisons and broader applicability.
* Utilize machine learning techniques to predict customer preferences and enhance personalization strategies.
* Conduct sentiment analysis to assess customer emotions and refine marketing and product strategies.
* Explore the long-term impact of discount strategies on brand loyalty and profitability.
* Investigate the role of sustainability practices in influencing customer behavior and market trends.

By addressing these areas, future research can provide a more comprehensive understanding of e-commerce dynamics, enabling businesses to remain competitive and customer-focused.

**References**

1. **Pandas Documentation**
   * **Reference:** Pandas Documentation
   * **Website:** <https://pandas.pydata.org/docs/>
   * **Usage:** Utilized for data manipulation, handling missing values, and performing aggregations within the dataset.
2. **Seaborn Documentation**
   * **Reference:** Seaborn: Statistical Data Visualization by Michael Waskom
   * **Website:** <https://seaborn.pydata.org/>
   * **Usage:** Used for creating advanced visualizations such as scatter plots, box plots, and heatmaps to explore data patterns and relationships.
3. **Matplotlib Documentation**
   * **Reference:** Matplotlib for Python Developers by Sandro Tosi
   * **Website:** <https://matplotlib.org/stable/contents.html>
   * **Usage:** Applied for generating visual representations of data trends, distributions, and general plotting tasks.
4. **Numpy Documentation**
   * **Reference:** NumPy Reference: A Guide to NumPy by Travis E. Oliphant
   * **Website:** <https://numpy.org/doc/>
   * **Usage:** Used for numerical computations and data processing, providing essential support for array manipulations and mathematical functions.
5. **Flipkart Dataset**
   * **Source:** Flipkart Dataset
   * **File:** Provided in the project dataset file "Flipkart dataset.csv"
   * **Usage:** The primary source of data used for analysis, visualization, and deriving insights regarding product performance and customer behavior.
6. **Academic Articles on E-commerce Analytics**
   * **Example:** Kumar, V., & Reinartz, W. (2016). Creating Enduring Customer Value. Journal of Marketing Research.
   * **Usage:** Provided valuable insights into customer segmentation and value creation strategies in e-commerce platforms, enhancing the understanding of consumer behavior.
7. **Python Official Documentation**
   * **Reference:** Python Documentation
   * **Website:** <https://docs.python.org/3/>
   * **Usage:** Used for general programming tasks, syntax reference, and supporting various data manipulation tasks.
8. **Guides and Tutorials**
   * **Platform:** GeeksforGeeks
   * **Example Tutorial:** "A Beginner's Guide to Data Visualization in Python" (Towards Data Science)
   * **Website:** [GeeksforGeeks](https://www.geeksforgeeks.org/), [Towards Data Science](https://towardsdatascience.com/)
   * **Usage:** Provided practical implementations of statistical and visualization techniques, offering step-by-step guidance for beginners.

**Appendices**

**1. Importing all required libraries and loading my dataset:**

# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns %matplotlib inline

**Description**:

This code imports essential libraries for data analysis and visualization: Pandas for data manipulation, NumPy for numerical computations, Matplotlib for plotting, and Seaborn for advanced visualizations, with inline plotting enabled.

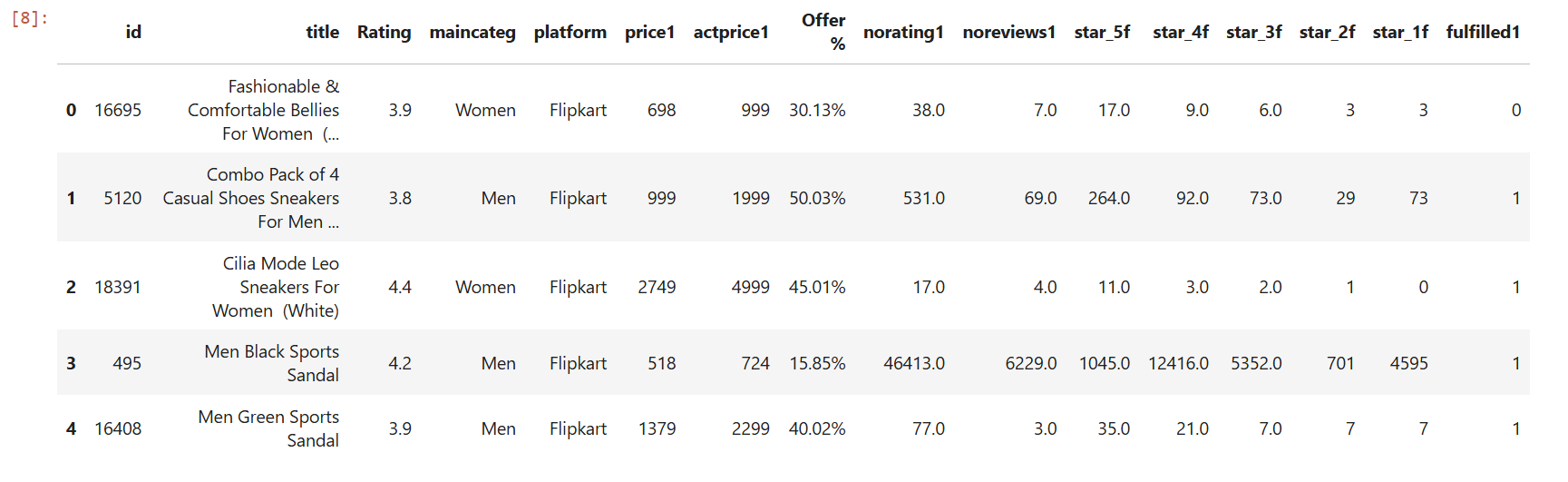
2. **Loading Data:**

**# Load the dataset**

data = pd.read\_csv('train.csv')

# Preview the dataset

data.head()



**Description**:

The code loads the dataset from 'train.csv' using Pandas and displays the first five rows with data.head(), providing a quick preview of the dataset's structure and contents.

**3.Data Columns Details:**

**# Data Overview**

print("Dataset Info:")

data.info()

**Dataset Info:**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 15730 entries, 0 to 15729

Data columns (total 16 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 id 15730 non-null int64

1 title 15730 non-null object

2 Rating 15730 non-null float64

3 maincateg 15204 non-null object

4 platform 15730 non-null object

5 price1 15730 non-null int64

6 actprice1 15730 non-null int64

7 Offer % 15730 non-null object

8 norating1 15052 non-null float64

9 noreviews1 15152 non-null float64

10 star\_5f 15142 non-null float64

11 star\_4f 15191 non-null float64

12 star\_3f 15499 non-null float64

13 star\_2f 15730 non-null int64

14 star\_1f 15730 non-null int64

15 fulfilled1 15730 non-null int64

dtypes: float64(6), int64(6), object(4)

memory usage: 1.9+ MB

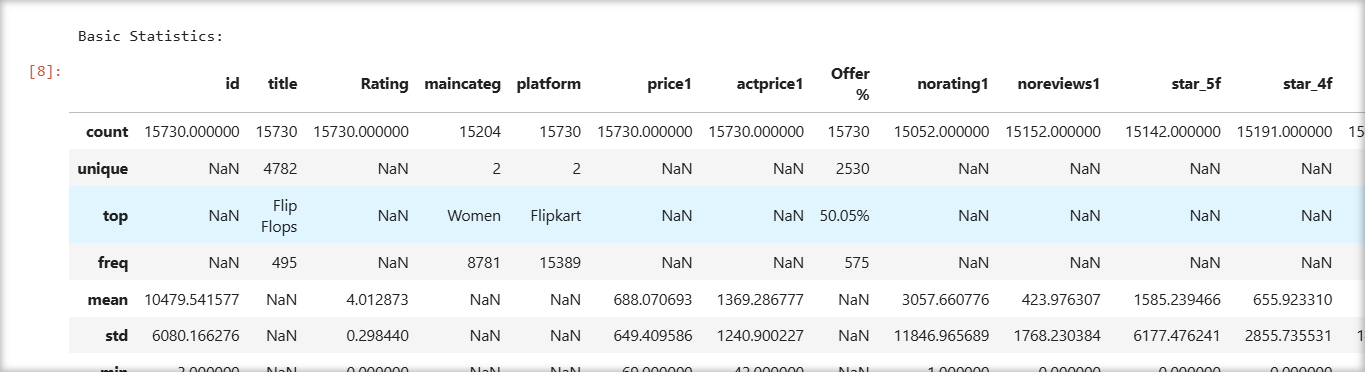
**Description**:

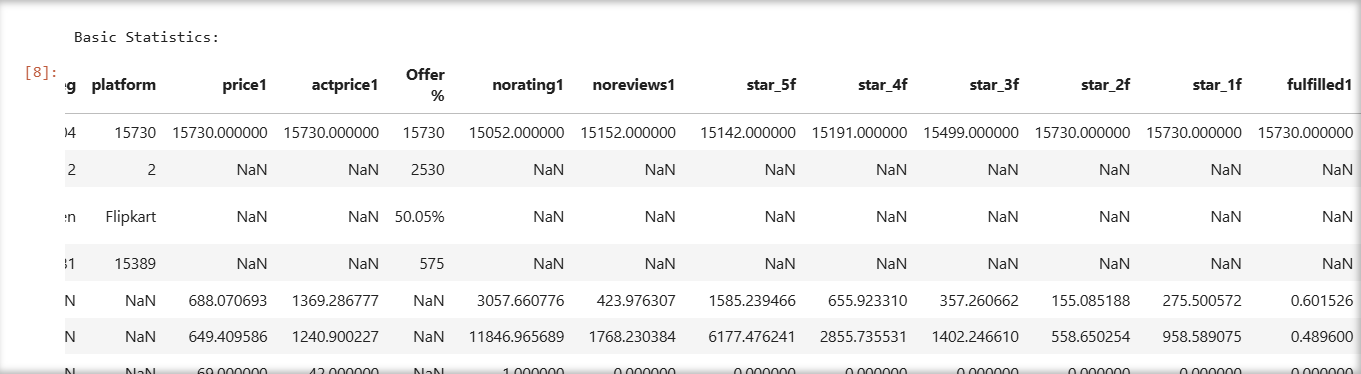
The data.info() method provides a concise summary of the dataset, including the number of entries, data types, non-null counts, and memory usage, aiding in understanding its structure and completeness.

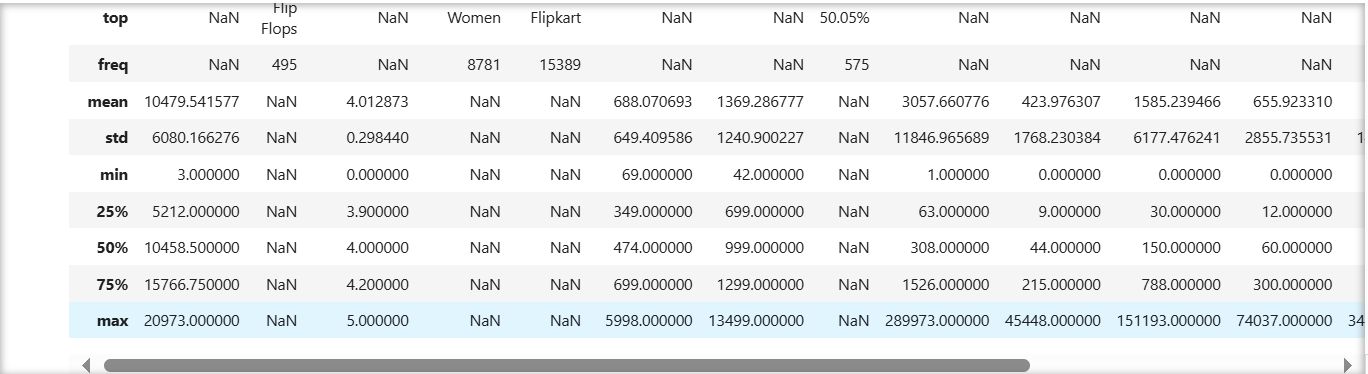
**4.Basic Statistics:**

print("\nBasic Statistics:")

data.describe(include='all')







**Description**:

The data.describe(include='all') function generates a summary of key statistics for the dataset, including counts, means, standard deviations, minimums, maximums, and percentiles for numerical and categorical columns.

**5.Check for any Null Values:**

**# Handling Missing Values**

missing\_summary = data.isnull().sum()

print("Missing Values per Column:\n", missing\_summary)

# Example: Filling missing values in `maincateg`

data['maincateg'].fillna('Unknown', inplace=True)

# Converting `Offer %` to numeric

data['Offer %'] = data['Offer %'].str.replace('%', '').astype(float)

# Drop or fill other missing values as necessary

data.fillna(0, inplace=True)

**Output**:

Missing Values per Column:

id 0

title 0

Rating 0

maincateg 526

platform 0

price1 0

actprice1 0

Offer % 0

norating1 678

noreviews1 578

star\_5f 588

star\_4f 539

star\_3f 231

star\_2f 0

star\_1f 0

fulfilled1 0

dtype: int64

**Description**:

Missing values were identified and handled by filling maincateg with "Unknown" and replacing others with 0. The Offer % column was cleaned and converted to a numeric format for analysis.

**6. Summarization:**

* print('Sum of Prices: ',data['price1'].sum())

Output:

**Sum of Prices: 10823352**

* print('Mean Price: ',data['price1'].mean())

Output:

**Mean Price: 688.0706929434202**

* print('Median Price: ',data['price1'].median())

Output**:  
Median Price: 474.0**

* print('Minimum Price: ',data['price1'].min())

Output:

**Minimum Price: 69**

* print('Maximum Price: ',data['price1'].max())

Output:

**Maximum Price: 5998**

**Description**:

The dataset's total price is ₹10,823,352, with an average price of ₹688. The median is ₹474, and prices range from ₹69 to ₹5,998, showing moderate pricing variation.

**7.Counting:**

* print('Count of Entries: ',data['maincateg'].count())

Output:

**Count of Entries: 15730**

* print('Unique Categories: ',data['maincateg'].nunique())

Output:

**Unique Categories: 3**

**Description**:

The dataset contains 15,730 total entries in the maincateg column, representing product categories. There are 3 unique categories, highlighting the dataset's focus on a few main product groups.

**8. Statistics:**

* print('Standard Deviation of Prices: ',data['price1'].std())

Output:

**Standard Deviation of Prices: 649.409586038453**

* print('Variance of Prices: ',data['price1'].var())

Output:

**Variance of Prices: 421732.8104386349**

* print('25th Percentile of Prices: ',data['price1'].quantile(0.25))

Output:

**25th Percentile of Prices: 349.0**

**Description**:

The statistics reveal the price variability: a standard deviation of ₹649.41, variance of ₹421,732.81, and 25th percentile at ₹349, highlighting significant price spread and distribution insights.

**9. Grouped Aggregations:**

* average\_price = data.groupby('maincateg')['price1'].mean()

print('Average Price by Main Category: ',average\_price)

Output:

**Average Price by Main Category: maincateg**

|  |  |
| --- | --- |
| Men | 738.710416 |
| Unknown | 698.682510 |
| Women | 650.393805 |
| Name | price1 |

* count\_by\_category = data.groupby('maincateg').size()

print('Total Count of Items by Category: ',count\_by\_category)

Output:

**Total Count of Items by Category: maincateg**

|  |  |
| --- | --- |
| Men | 6423 |
| Unknown | 526 |
| Women | 8781 |
| Name | int64 |

**Description**:

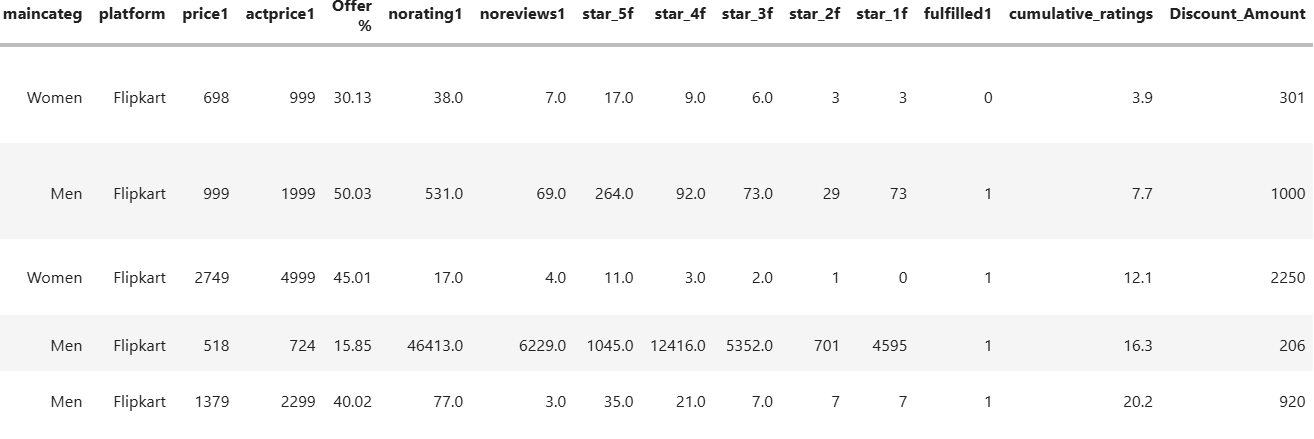
Grouped aggregations reveal insights into average prices and item counts across main categories: "Men" has the highest average price, while "Women" dominates in product count with 8,781 items.

* **Column add-on**

#Here we can Add the new Discount amount table in the table

data['Discount\_Amount'] = data['actprice1'] - data['price1']

data.head()

****

****

**Description:**

The new column "Discount\_Amount" is added to the dataset by calculating the difference between the "actprice1" (actual price) and "price1" (original price) for each product. This provides the discount amount for each product, helping to analyze the extent of price reductions across different items in the dataset.

**10.Data Visualization**

**-Distribution of Ratings**

**# 1. Histogram: Distribution of Ratings**

plt.figure(figsize=(8, 5))

sns.histplot(data['Rating'], bins=20, kde=True, color='blue')

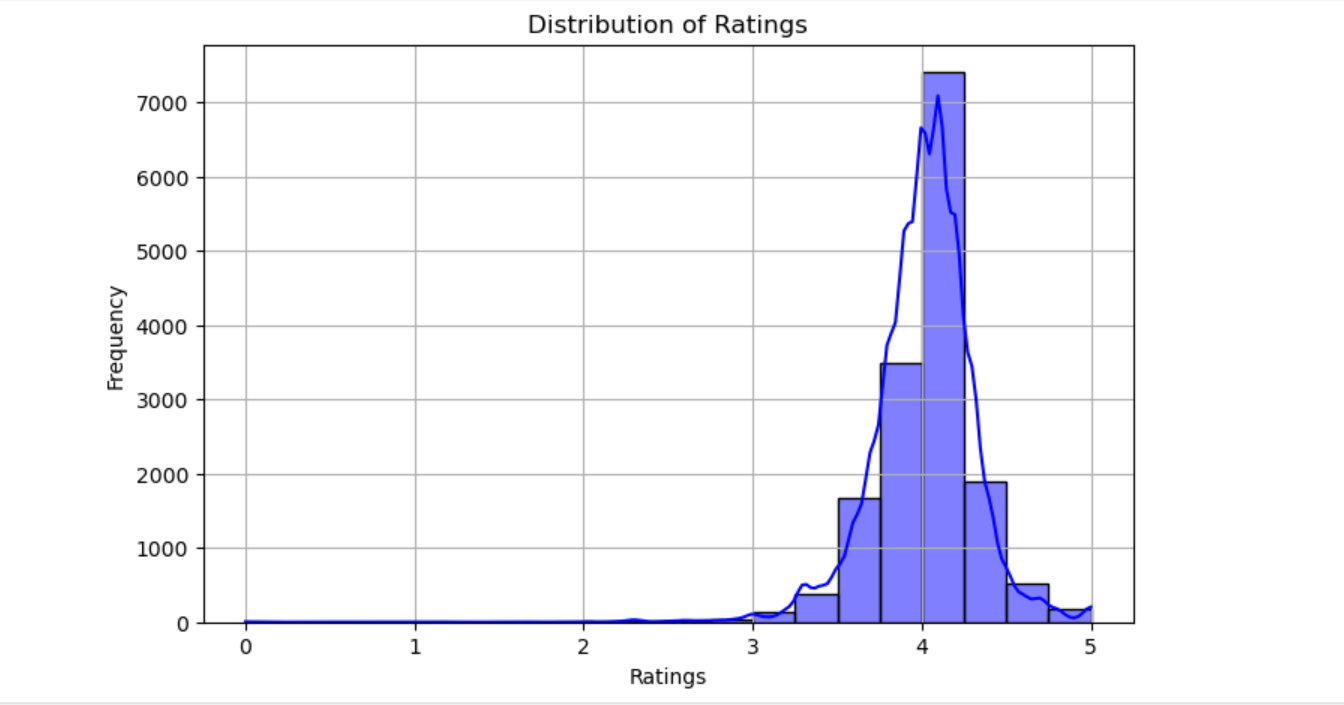
plt.title("Distribution of Ratings")

plt.xlabel("Ratings")

plt.ylabel("Frequency")

plt.grid()

plt.show()



**Description**:

The histogram visualizes the distribution of ratings, with a kernel density estimate (KDE) overlay, showing the frequency of different rating values. It helps analyze rating patterns and customer preferences.

**-Price vs. Ratings**

**# 2. Scatter Plot: Price vs. Ratings**

plt.figure(figsize=(8, 5))

plt.scatter(data['price1'], data['Rating'], alpha=0.5, c='green')

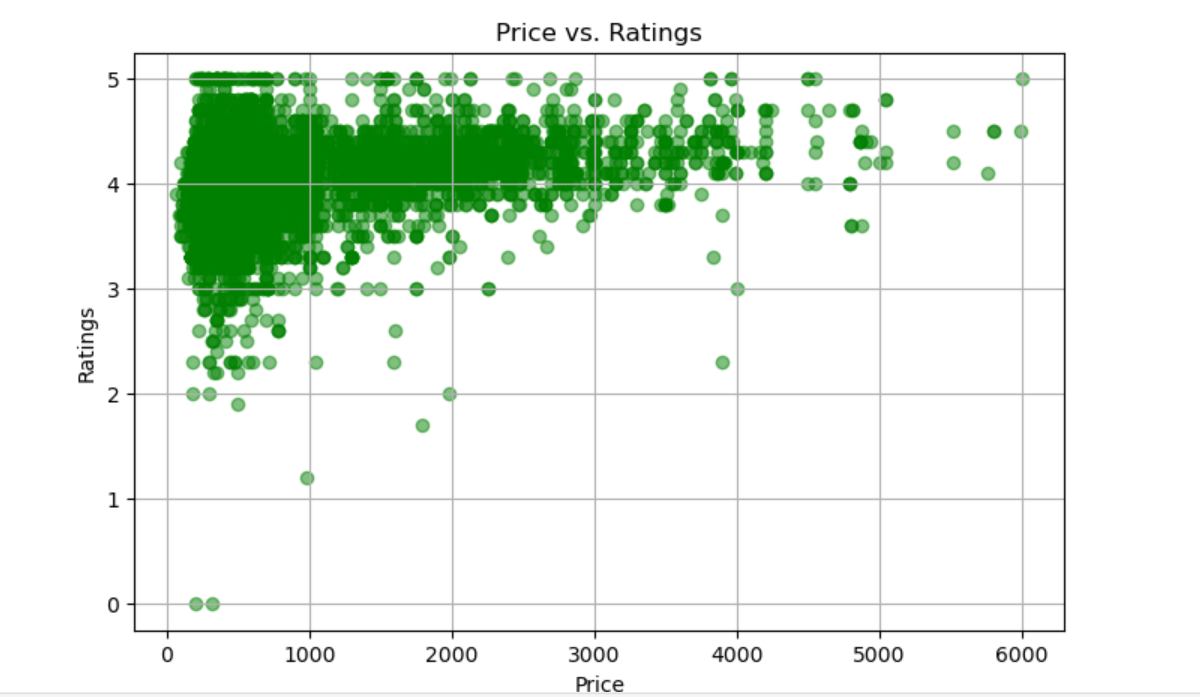
plt.title("Price vs. Ratings")

plt.xlabel("Price")

plt.ylabel("Ratings")

plt.grid()

plt.show()



**Description**:

The scatter plot visualizes the relationship between product price and ratings, with price on the x-axis and ratings on the y-axis. It uses green dots with alpha transparency for clarity.

**-Offers Distribution by Category:**

**# 3. Box Plot: Offers by Categories**

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

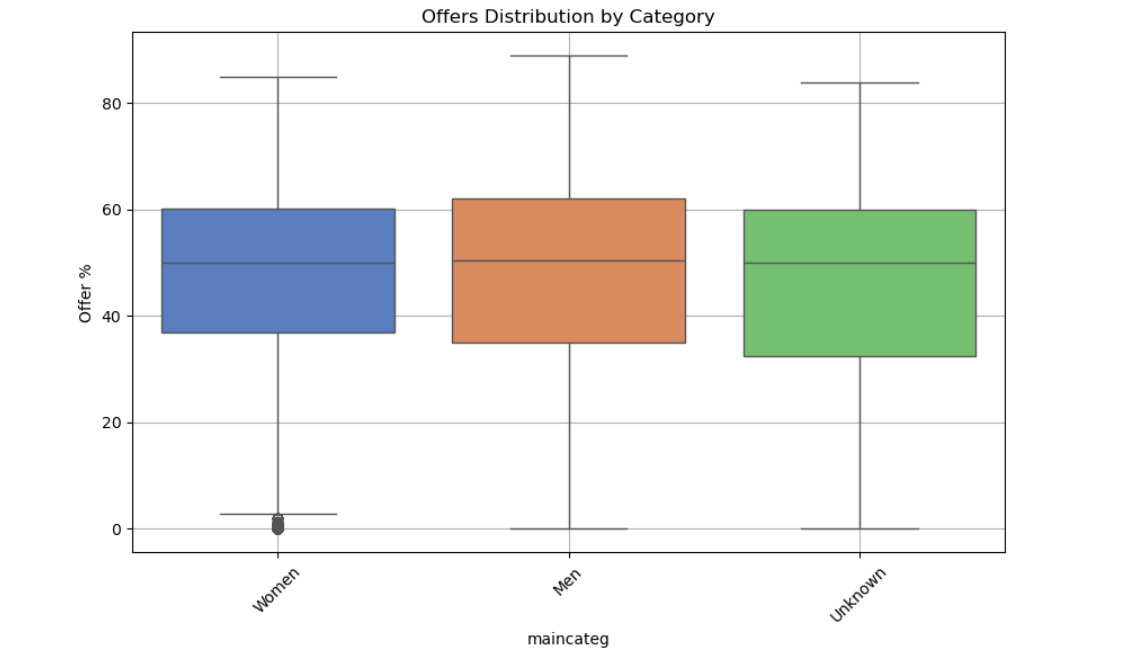
sns.boxplot(x='maincateg', y='Offer %', data=data, palette='muted')

plt.title("Offers Distribution by Category")

plt.xticks(rotation=45)

plt.grid()

plt.show()



**Description**:

The box plot visualizes the distribution of offer percentages across different categories, showing median, quartiles, and outliers, helping identify offer trends and variations within each product category.

**-Product Counts by Category**

**# 4. Bar Chart: Product Counts by Category**

plt.figure(figsize=(8, 5))

category\_counts = data['maincateg'].value\_counts()

category\_counts.plot(kind='bar', color='purple')

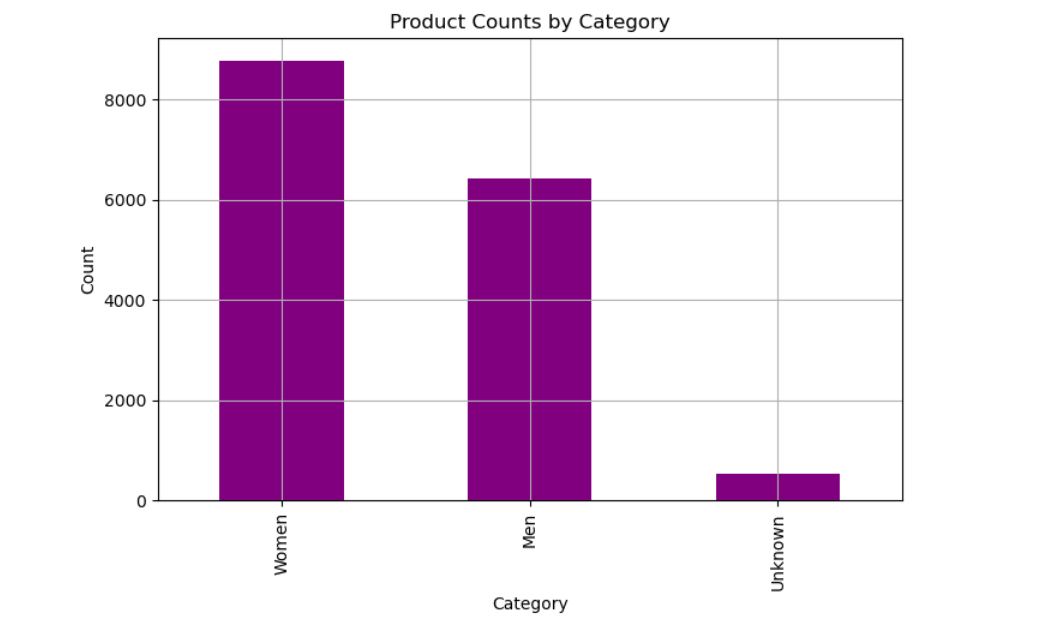
plt.title("Product Counts by Category")

plt.xlabel("Category")

plt.ylabel("Count")

plt.grid()

plt.show()



**Description**:

The bar chart visualizes the count of products across different categories, using purple bars to represent each category's frequency. It provides insights into product distribution, with category names on the x-axis.

**-Heatmap of Numerical Features**

**# 5. Heatmap: Correlation of Numerical Features**

plt.figure(figsize=(12, 8))

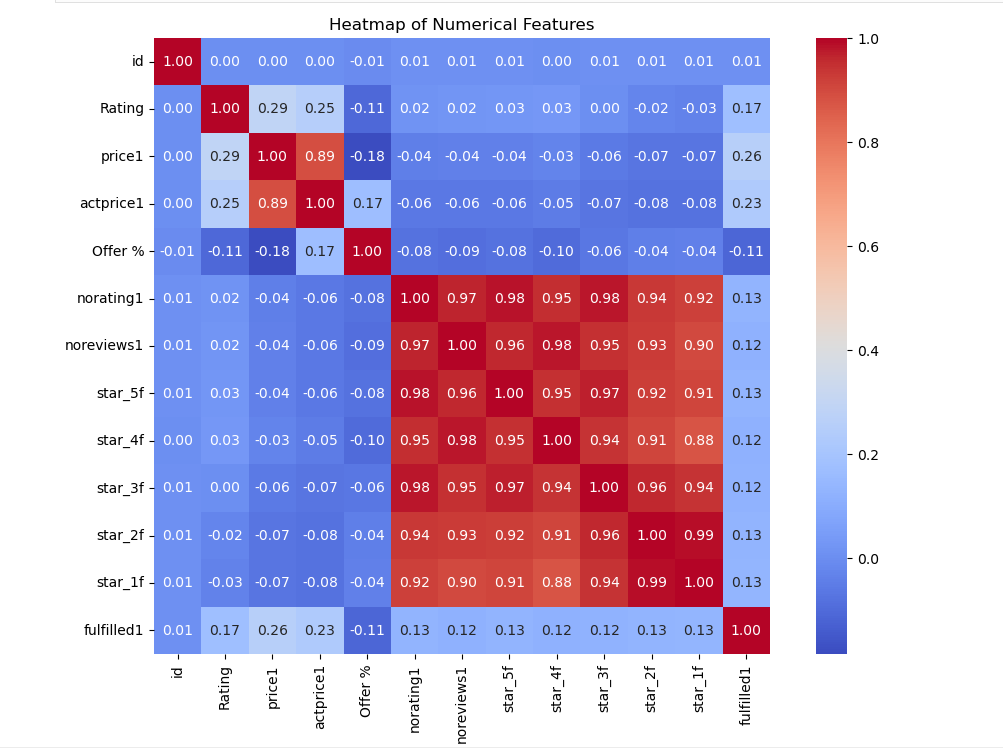
numeric\_columns = data.select\_dtypes(include=['float64', 'int64']).columns

sns.heatmap(data[numeric\_columns].corr(), annot=True, fmt=".2f", cmap="coolwarm", square=True)

plt.title("Heatmap of Numerical Features")

# plt.grid()

plt.show()



**Description**:

The graph generates a heatmap to visualize the correlation between numerical features in the dataset. It uses Seaborn's heatmap() function, displaying correlation values with a cool-to-warm color gradient for insights.

**-Cumulative Ratings Trend**

**# 6. Line Chart: Cumulative Ratings Trend**

data['cumulative\_ratings'] = data['Rating'].cumsum()

plt.figure(figsize=(10, 5))

plt.plot(data['cumulative\_ratings'], color='orange')

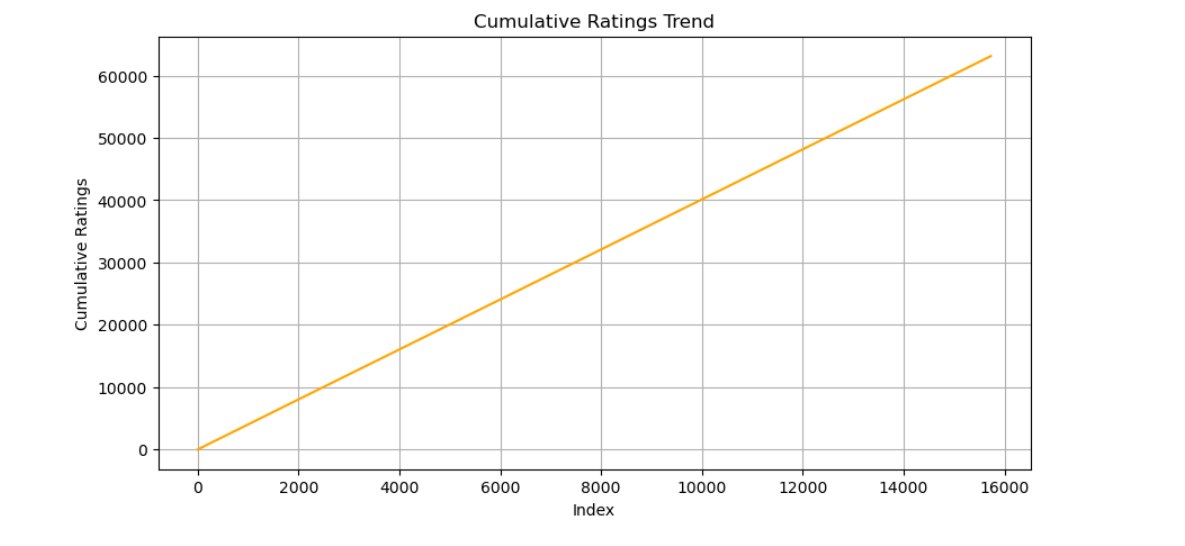
plt.title("Cumulative Ratings Trend")

plt.xlabel("Index")

plt.ylabel("Cumulative Ratings")

plt.grid()

plt.show()



**Description**:

The line chart visualizes the cumulative ratings trend over the dataset, showing the cumulative sum of ratings across products. It helps track how ratings accumulate as more products are reviewed.

**-Proportion of Categories**

**# 7. Pie Chart: Category Proportions**

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

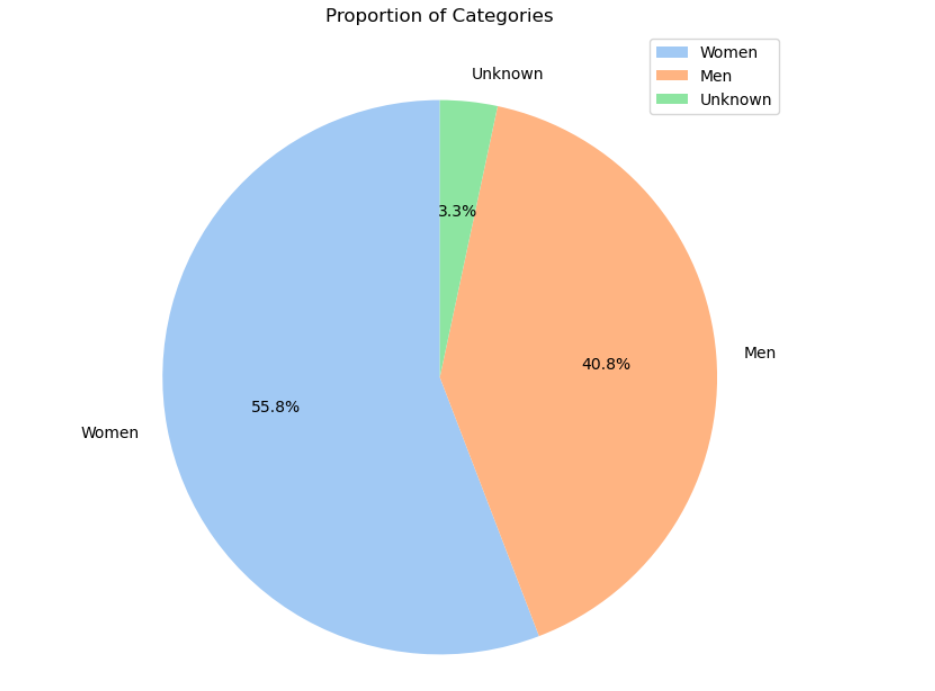
category\_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90, colors=sns.color\_palette("pastel"))

plt.title("Proportion of Categories")

plt.ylabel("")

plt.legend(["Women", "Men","Unknown"])

plt.show()



**Description**:

This graph generates a pie chart to visualize the proportion of product categories (Women, Men, Unknown), displaying percentages with pastel colors and a legend, offering a clear distribution overview.

**-Price Distribution by Platform**

**# 8. Violin Plot: Price Distribution by Platform**

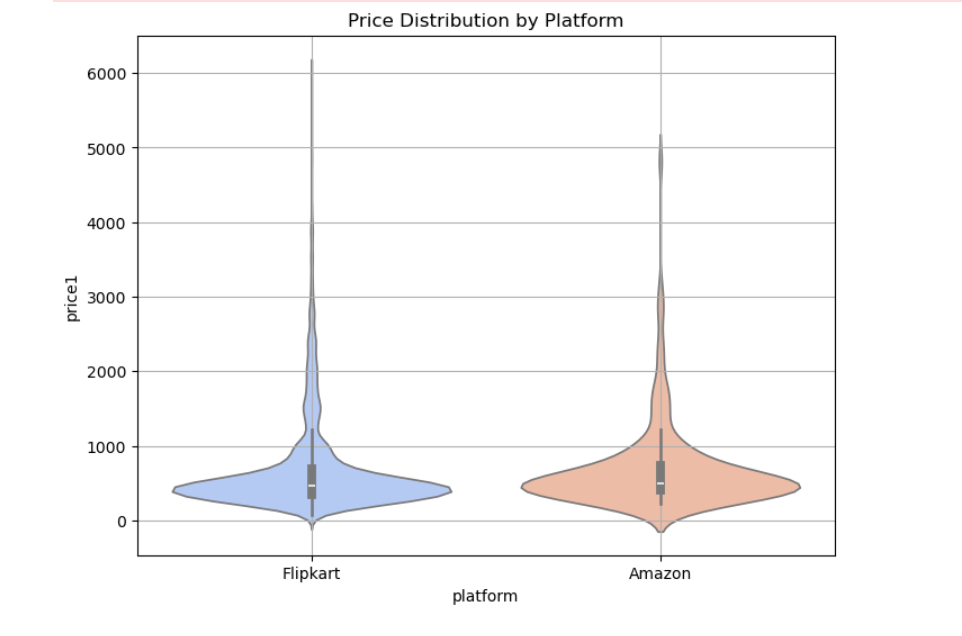
plt.figure(figsize=(8, 6))

sns.violinplot(x='platform', y='price1', data=data, palette='coolwarm')

plt.title("Price Distribution by Platform")

plt.grid()

plt.show()



**Description**:

The violin plot visualizes the price distribution across different platforms, showing the median, interquartile range, and density of prices. It highlights variations and trends in product pricing across platforms.

**-Fulfillment Status by Categories**

**# 9. Stacked Bar Chart: Fulfillment Status by Categories**

fulfillment\_status = pd.crosstab(data['maincateg'], data['fulfilled1'])

fulfillment\_status.plot(kind='bar', stacked=True, figsize=(10, 6), color=['red', 'green'])

plt.title("Fulfillment Status by Categories")

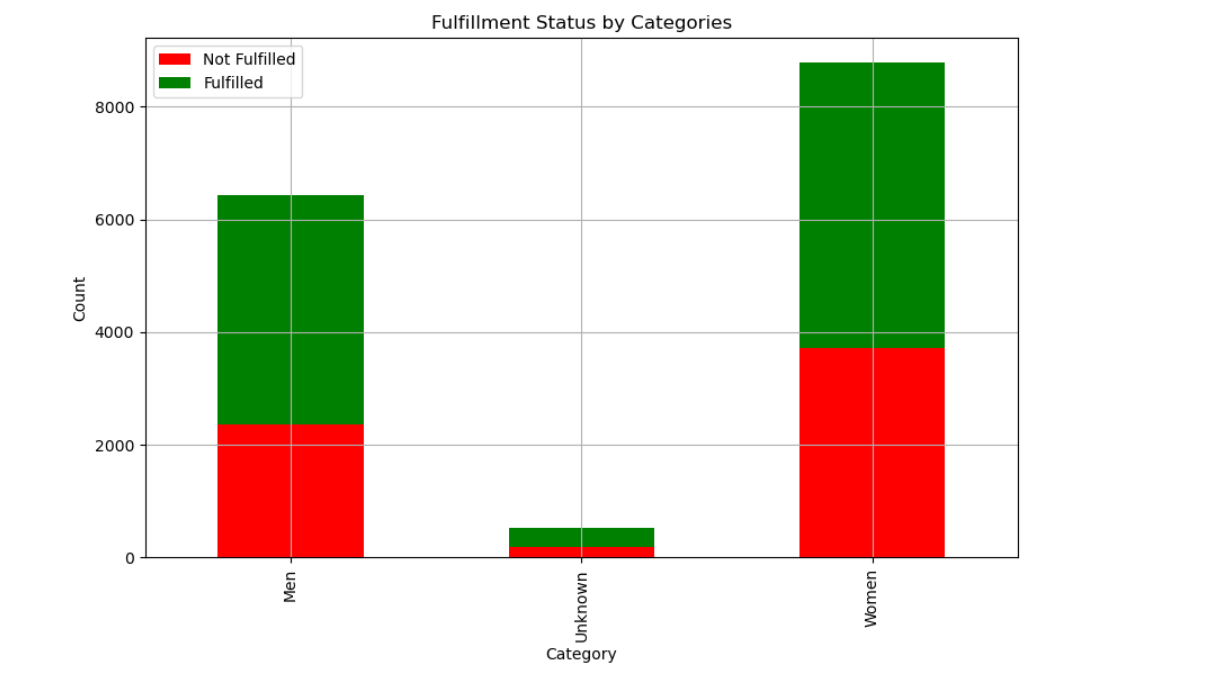
plt.xlabel("Category")

plt.ylabel("Count")

plt.legend(["Not Fulfilled", "Fulfilled"])

plt.grid()

plt.show()



**Description**:

The stacked bar chart visualizes the fulfilment status across categories, displaying the count of fulfilled and non-fulfilled products. Red represents "Not Fulfilled," while green indicates "Fulfilled" items.

**-Ratings vs. Reviews by Platform**

**# 10. Facet Grid: Ratings vs. Reviews by Platform**

g = sns.FacetGrid(data, col="platform", height=5, aspect=1.2)

g.map(sns.scatterplot, "Rating", "noreviews1", alpha=0.7)

# Add grid to each subplot

for ax in g.axes.flat:

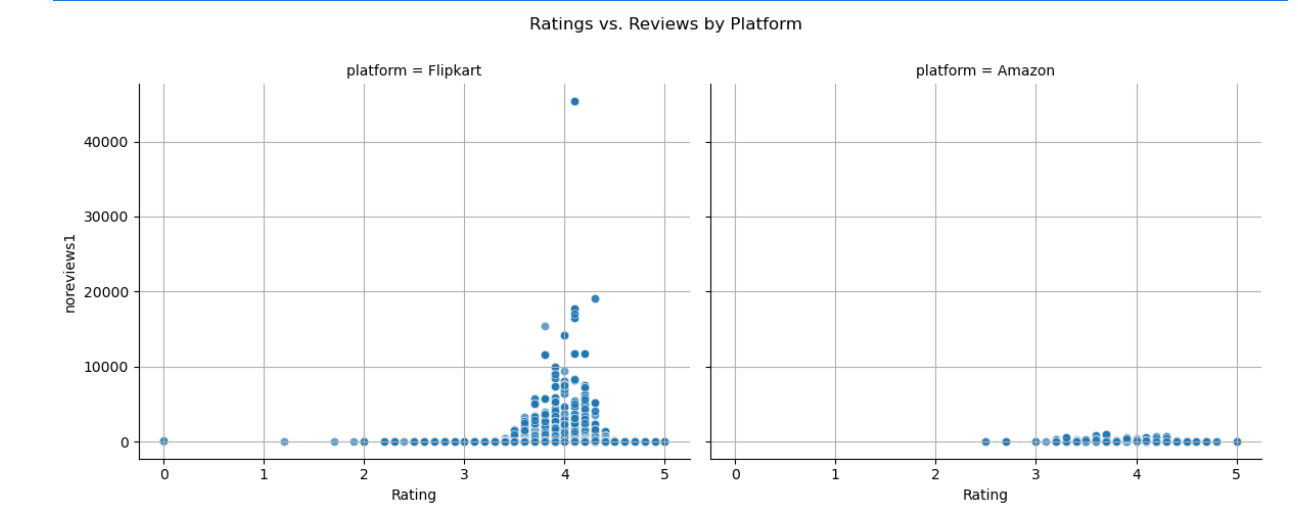
ax.grid(True)

g.add\_legend()

plt.subplots\_adjust(top=0.85)

g.fig.suptitle("Ratings vs. Reviews by Platform")

plt.show()



**Description**:

The FacetGrid creates scatter plots of Ratings vs. Reviews for each platform, with a grid added to each subplot. It includes a legend and a title, providing a comparative visualization of the data.

**-Top 20 Titles by Money Spent**

group\_title = data.groupby('title')['price1'].sum().sort\_values(ascending=False)

top\_20\_titles = group\_title.head(20)

plt.subplots(figsize=(15, 8))

top\_20\_titles.plot(kind='barh', fontsize=12)

# Update axis labels and title

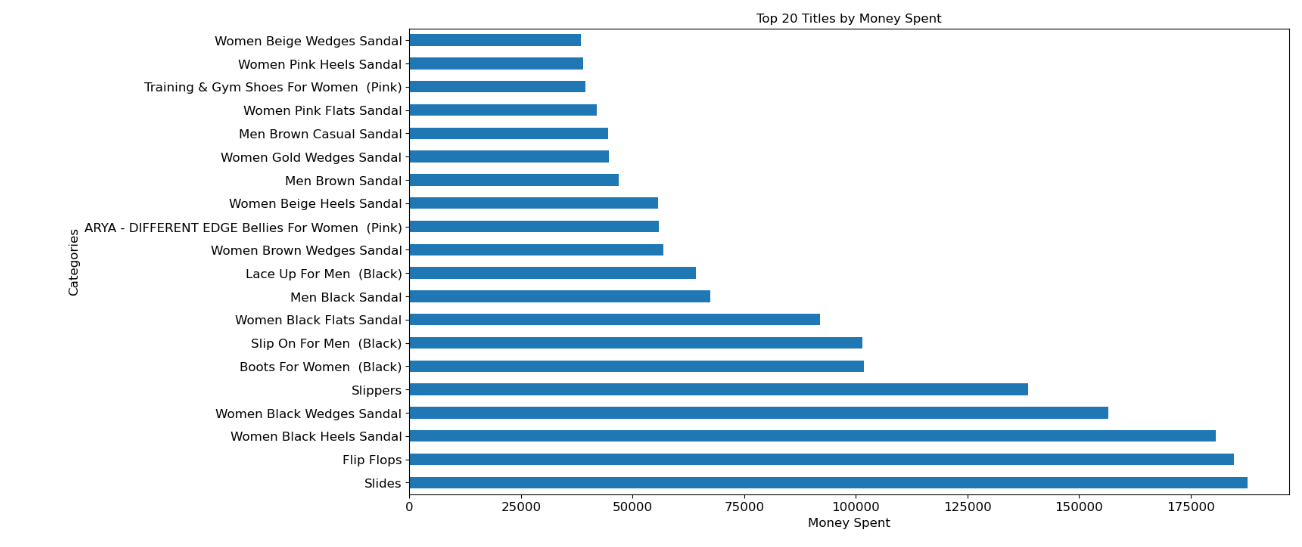
plt.xlabel('Money Spent', fontsize=12)

plt.ylabel('Categories', fontsize=12)

plt.title('Top 20 Titles by Money Spent', fontsize=12)

# Show the plot

plt.show()



**Description**:

This graph groups product titles by total price (price1), sorts the data in descending order, and visualizes the top 20 titles with the highest total spending using a horizontal bar chart.