INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Title: Web Scraping and Exploratory Data Analysis of Flipkart - Mobile

INNOMATICS
RESEARCH LABS

Introduction:

- **Flipkart** is one of India's leading **e-commerce platforms**, offering a wide range of products such as **mobiles, electronics, fashion, home essentials**, and more. Among all categories, **mobiles** are one of the **top-selling and most-searched products** on Flipkart.
- Flipkart's **mobile data** refers to all the **information collected and displayed** about smartphones listed on the Flipkart platform.
- It includes both **structured** (numbers, categories) and **unstructured** (text, reviews, images) data.
- Analyzing mobile data helps both **businesses** and **data analysts** derive meaningful insights.

INNOMATICS RESEARCH LABS

Business Problem:

- **Business teams** struggle to **benchmark products** based on price, ratings, and features.
- **Marketing and pricing analysts** lack **clear visibility into discount effectiveness**, customer satisfaction trends, and brand competitiveness.
- **Customers** face information overload when comparing devices, leading to decision fatigue and suboptimal purchase choices.
- Flipkart offers discounts on almost every product, but the platform lacks a clear metric to **measure how discounts impact sales and customer ratings**.

INNOMATICS RESEARCH LABS

Objectives:

- To analyze the relationship between **mobile prices, discounts, and customer ratings** to identify optimal pricing strategies.
- To examine **customer reviews and ratings** to identify the key factors influencing product satisfaction.
- To evaluate how **discount percentages** affect **customer interest, ratings, and purchase likelihood**.
- To compare **different mobile brands** on key performance metrics such as **average price, rating, and discount.**
- To create a benchmarking framework that evaluates **feature-level performance** across mobile models.

INNOMATICS RESEARCH LABS

Web Scraping:

- Flipkart official website was selected as the main data source.
- Mobiles is selected in flipkart website and dragged the information.
- Used browser developer tools (Inspect Element) to identify relevant HTML elements in match data pages
- Used Beautiful Soup and Requests libraries in Python to extract structured data from the web pages
- Sent HTTP requests to automate the retrieval of multiple match records across seasons
- Cleaned and consolidated scraped data for further analysis and visualization.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import re
import requests
from bs4 import BeautifulSoup
import warnings
warnings.filterwarnings("ignore")

url='https://www.flipkart.com/search?q=all+4g+mobile&sid=tyy%2C4io&as=on&as-show=on&otracker=AS_QueryStore_OrganicAutoSuggest_0_6&otracker1=AS_QuerySto

response= requests.get(url)
response
```

```
<Response [200]>
```

```
    for x in soup.find_all('div', attrs={'class':'yKfJKb row'}):
        pname = x.find('div', attrs={'class':'KzDlHZ'})
        cost = x.find('div', attrs={'class':'Nx9bqj _4b5DiR'})   # discounted price
        rat = x.find('div', attrs={'class':'XQDdHH'})
        specs = x.find('div', attrs={'class':'_6NESgJ'})
        op = x.find('div', attrs={'class':'yRaY8j ZYYwLA'})
        discount = x.find('div', attrs={'class':'UkUFwK'})
        reviews = x.find('span', attrs={'class':'Wphh3N'})


        productname.append(pname.text if pname else np.nan)
        price.append(cost.text if cost else np.nan)
        rating.append(rat.text if rat else np.nan)
        features.append(specs.text if specs else np.nan)
        original_price.append(op.text if op else np.nan)
        Discount.append(discount.text if discount else np.nan)
        Review.append(reviews.text if reviews else np.nan)

        pagenum.append(i)

    print(f'Page {i} completed in {time.time()-start_time:.2f} seconds')

print("Total Time Completed in seconds", str(time.time()-total_time))
```

```
Page 1 completed in 0.57 seconds
Page 2 completed in 0.59 seconds
Page 3 completed in 0.89 seconds
Page 4 completed in 0.72 seconds
Page 5 completed in 0.61 seconds
```

INNOMATICS RESEARCH LABS

Tools Used:


BeautifulSoup


pandas


.[RegEx]*


matplotlib


NumPy


seaborn

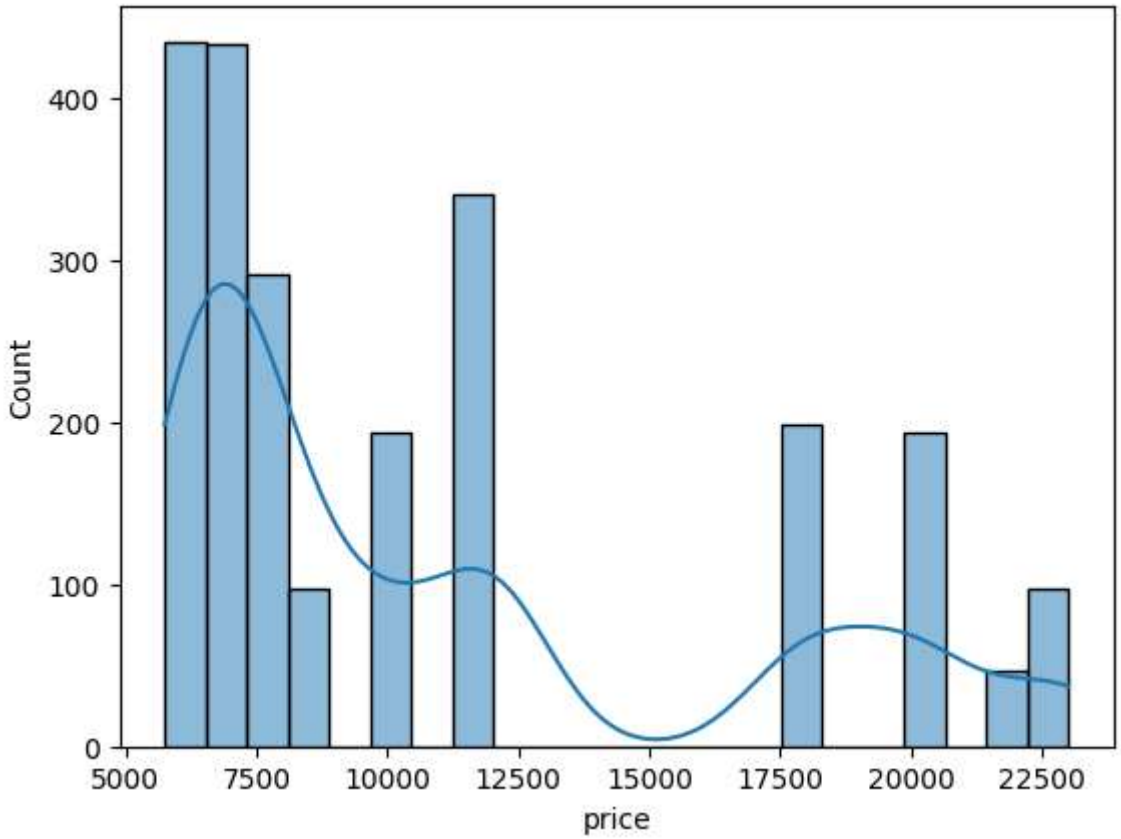

INNOMATICS
RESEARCH LABS

Data Cleaning Steps:

- After **web scraping Flipkart mobile data**, the next key step is **data cleaning** — turning messy, inconsistent data into a structured and analysis-ready format.
- To clean, preprocess, and prepare the scraped Flipkart mobile data for analysis — ensuring it is accurate, consistent, and ready for visualization or modeling.
- Identify and treat missing data in columns such as price, rating, or reviews.
- Scraped prices often include currency symbols and commas.
- Ratings may be in string form — convert them to numeric.

Data Visualization:

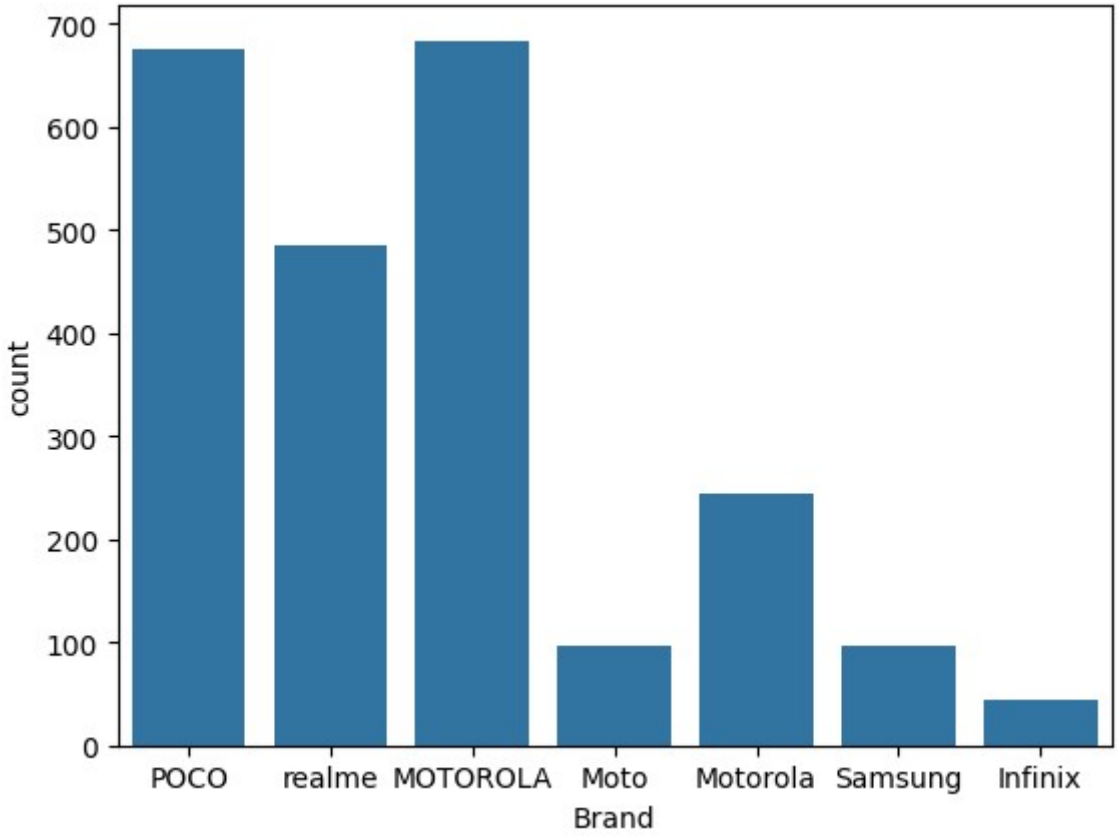
- The histogram shows how product prices are distributed across different ranges.
- Most prices are concentrated between ₹5,000 and ₹10,000, indicating that the majority of products fall in this affordable range.
- A few higher-priced products (₹15,000–₹22,000) are present but less frequent, suggesting they may be premium models.
- The KDE curve indicates a slightly right-skewed distribution, meaning there are more low-priced products compared to high-priced ones.



INNOMATICS RESEARCH LABS

Data Visualization:

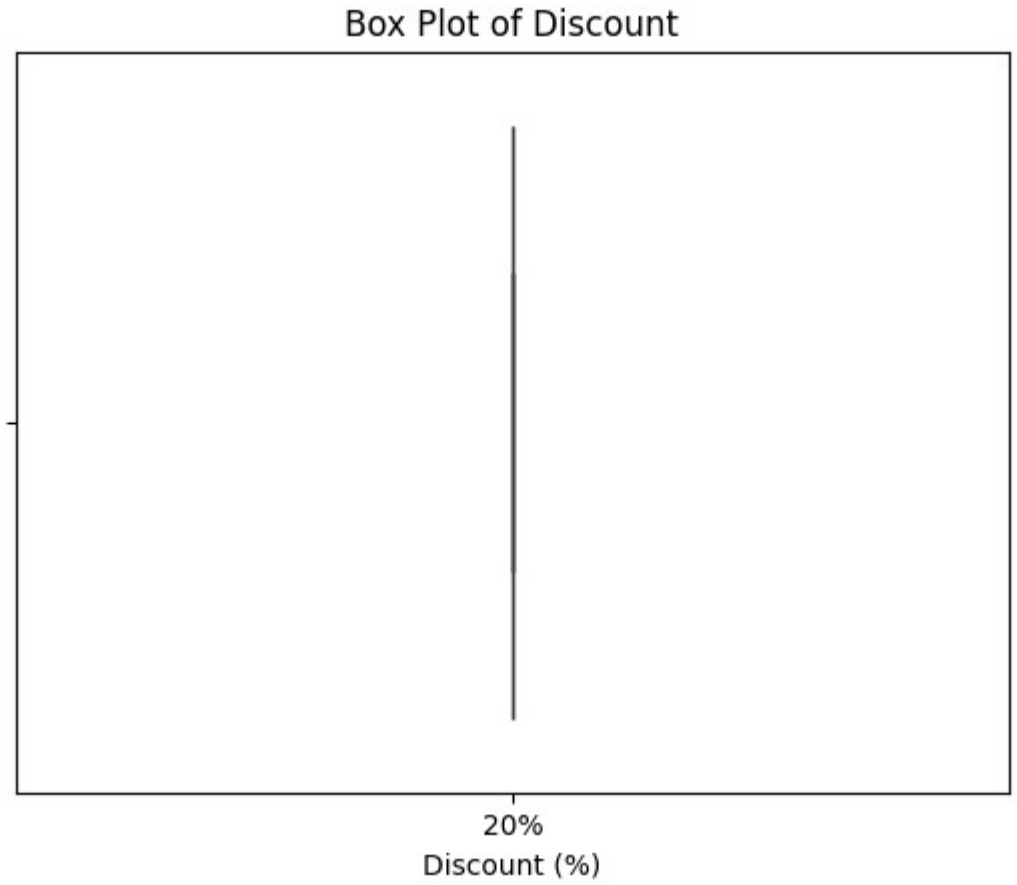
- POCO and MOTOROLA have the highest number of listings, each with nearly 700 products, indicating strong market availability or higher product variety.
- realme follows with around 480 listings, showing a significant presence.
- Motorola (alternate name form) and Moto have moderate listings, which may indicate data inconsistency due to different naming formats.
- Samsung and Infinix have relatively fewer listings, suggesting a smaller share in this dataset.



INNOMATICS RESEARCH LABS

Data Visualization:

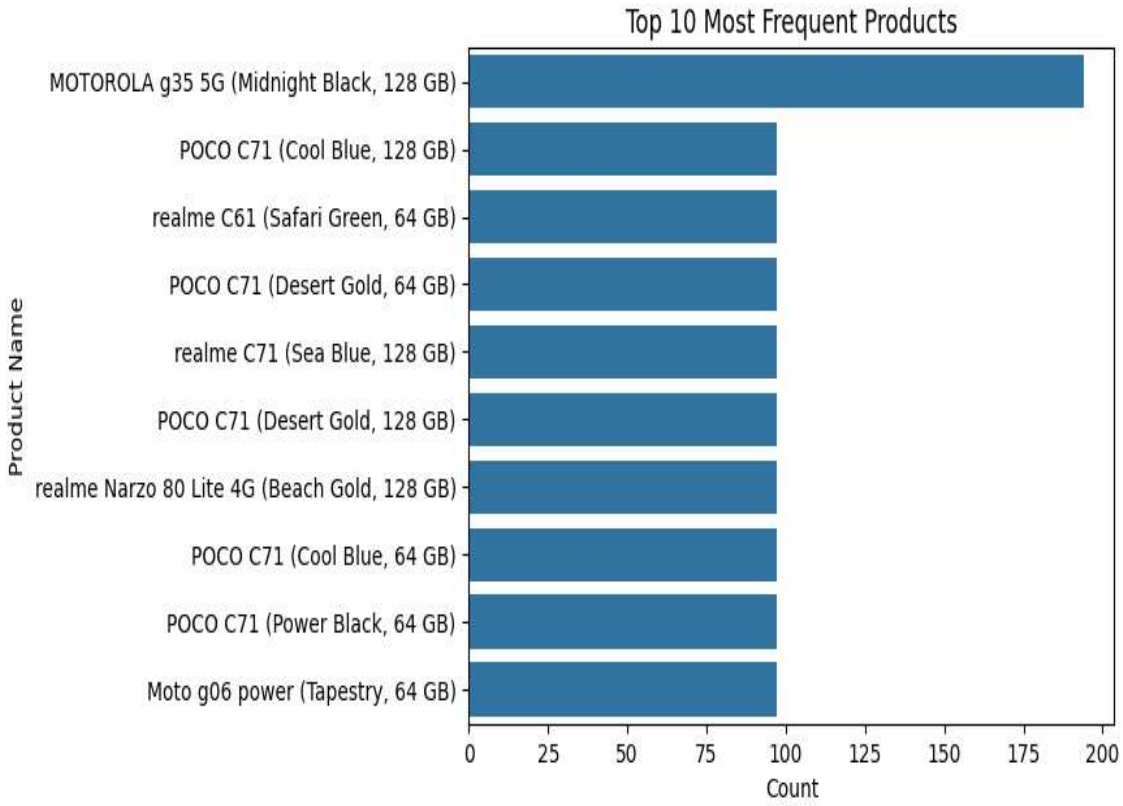
- The median discount is approximately 20%, indicating that most laptops have moderate discounts.
- The box plot is narrow, suggesting that discounts do not vary significantly across products.
- There are no significant outliers, meaning that extreme discounts are rare in this dataset.
- Most of the laptops fall within a similar discount range, indicating consistent promotional pricing strategy by the sellers.



INNOMATICS RESEARCH LABS

Data Visualization:

- The MOTOROLA g35 5G (Midnight Black, 128 GB) is the most frequently listed product, indicating it's among the top-selling or most featured models.
- The POCO C71 series (available in multiple color and storage variants such as Cool Blue, Desert Gold, and Power Black) appears several times in the top 10 list — showing its strong popularity and product variety.
- realme also has multiple entries, including realme C61 and realme Narzo 80 Lite 4G, highlighting its consistent presence in the competitive mid-range segment.

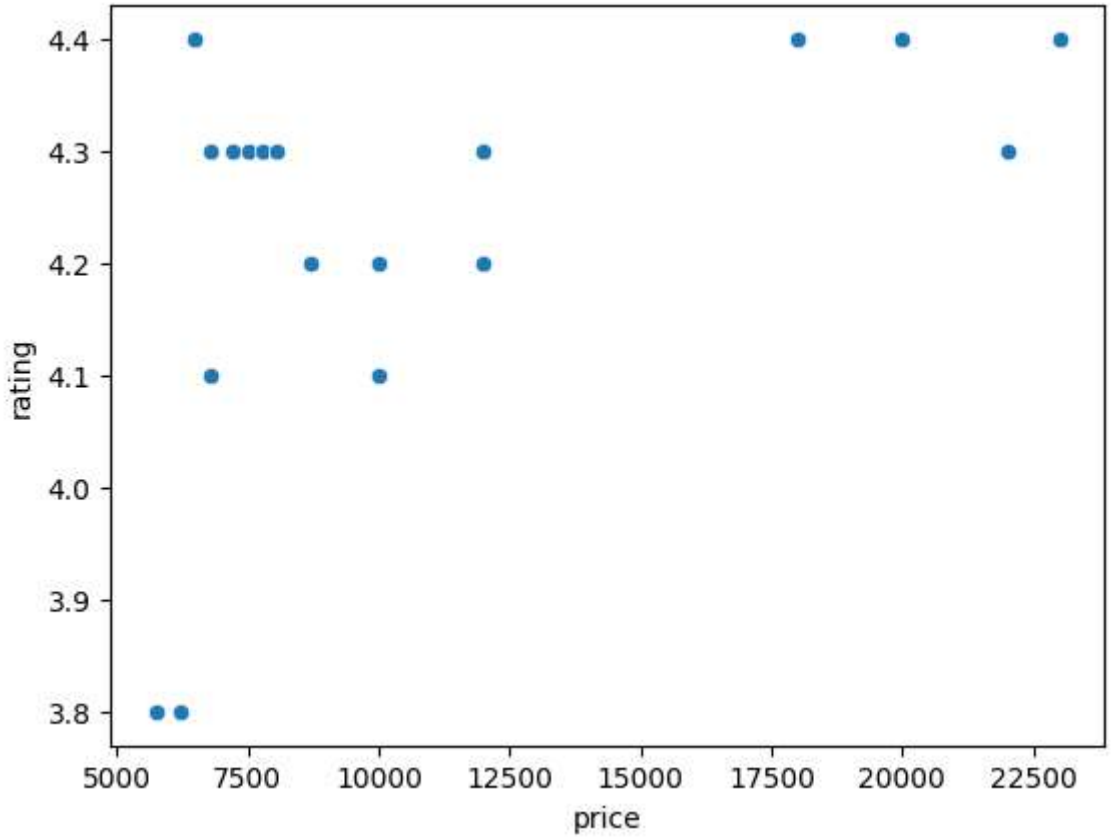


INNOMATICS RESEARCH LABS

Data Visualization:

2. Bivariate Analysis

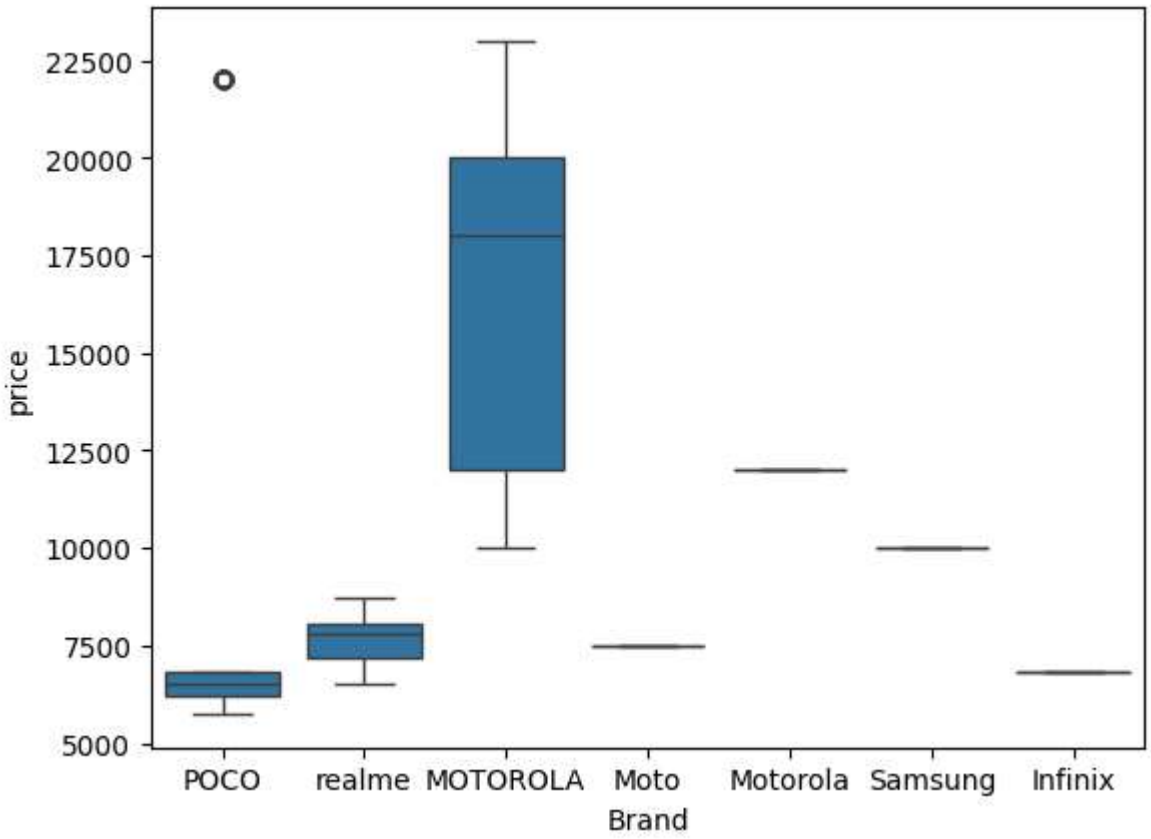
- The scatter plot shows that ratings remain consistently high (around 4.1–4.4) across all price ranges.
- There is no strong correlation between price and rating — indicating that higher-priced laptops do not necessarily receive better ratings.
- Customers seem satisfied across different price categories, suggesting that both budget and premium laptops meet consumer expectations in terms of performance and quality.
- The dense clustering of points at certain rating levels indicates that most laptops maintain similar



INNOMATICS RESEARCH LABS

Data Visualization:

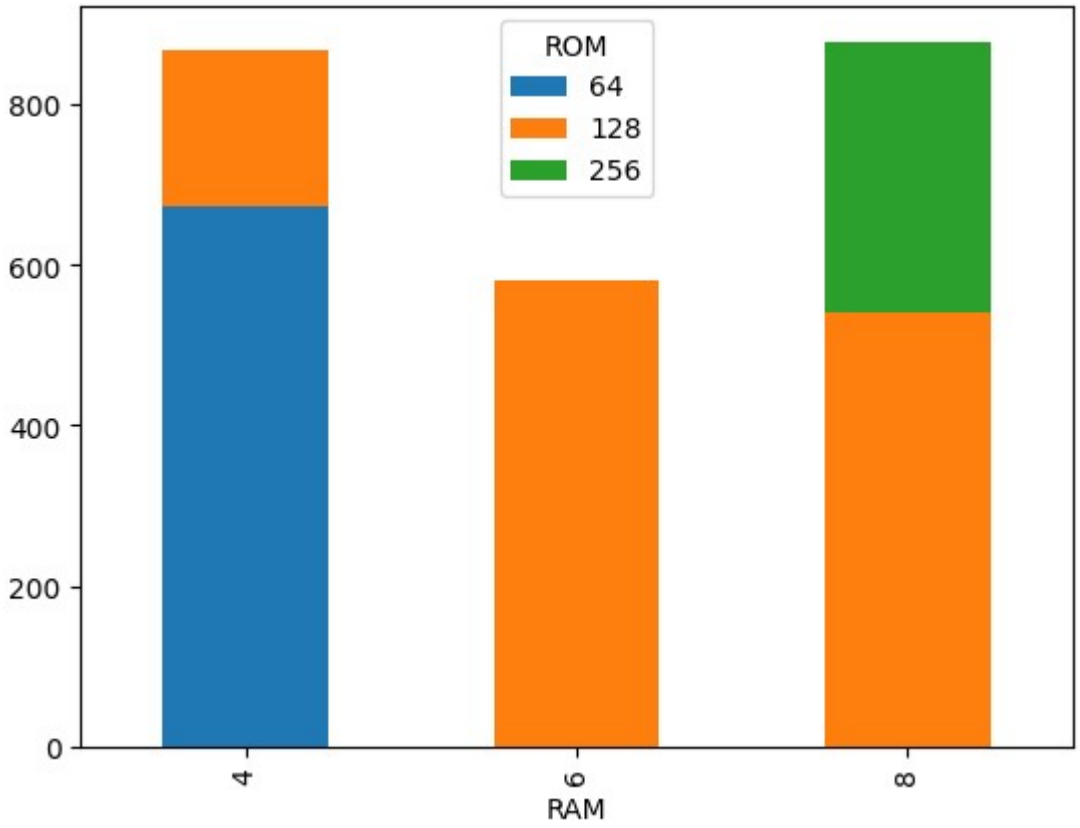
MOTOROLA laptops exhibit the highest median price range (₹12,000–₹20,000) with a wide interquartile range, suggesting the brand offers both mid-range and premium models.
realme laptops are mostly in the mid-price segment (₹6,000–₹9,000), showing moderate variation.
POCO and Infinix are in the budget-friendly category, priced mostly belov ₹7,000.
Samsung and Motorola (alternate name) have prices clustered in the ₹10,000–₹12,000 range, indicating stable and consistent pricing.
A few outliers in POCO and MOTOROLA suggest the presence of high-end variants or special editions.



INNOMATICS RESEARCH LABS

Data Visualization:

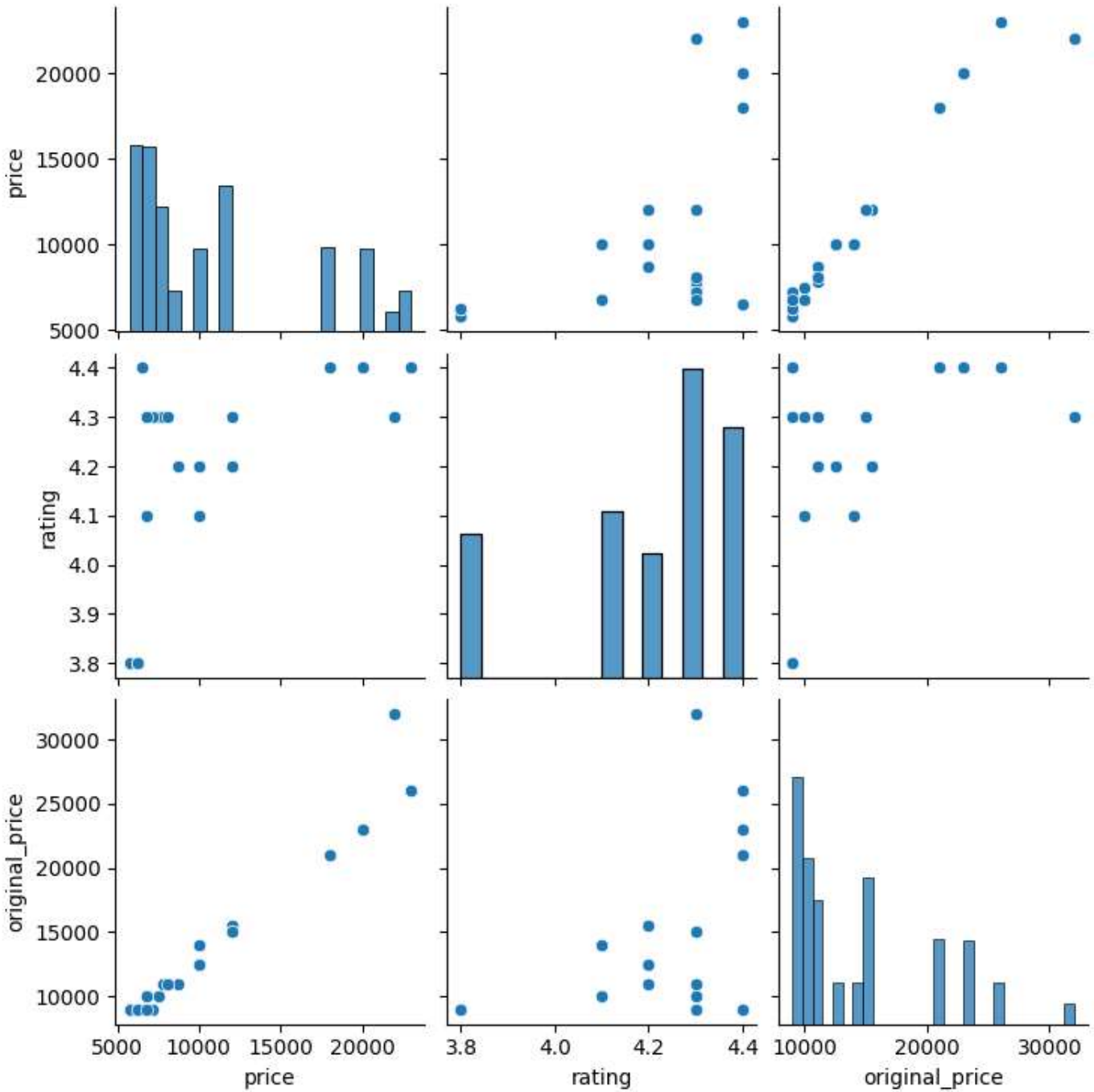
Devices with 4 GB RAM mostly come with 64 GB ROM, making them entry-level configurations aimed at budget users.
Laptops with 6 GB RAM are mainly paired with 128 GB ROM, reflecting a balance between performance and storage.
Systems with 8 GB RAM show a mix of 128 GB and 256 GB ROM, indicating that higher RAM configurations are often coupled with larger storage capacities — suitable for high-performance or professional use.
The stacked distribution suggests a positive association between RAM and ROM, where higher RAM models tend to come with more internal storage.



INNOMATICS RESEARCH LABS

Data Visualization:

- **Multi-Variate Analysis**.



INNOMATICS RESEARCH LABS

A positive relationship is visible between price and original_price, which is expected since discounts are applied on the original price.
Discount and price show an inverse relationship, where higher discounts are generally associated with lower prices.
Rating appears to have no strong correlation with price or discount, indicating that customer satisfaction is not solely influenced by pricing strategies.
The scatter patterns and diagonal histograms reveal that most products cluster within the mid-price range, with moderate discounts and consistently high ratings.

Data Visualization:

- There is a very strong positive correlation (0.98) between price and original_price, indicating that as the original price increases, the selling price also increases proportionally.

- A moderate positive correlation (0.61) exists between price and rating, suggesting that higher-priced laptops tend to receive slightly better ratings, though the relationship is not very strong.

- The correlation between reviews and other variables (price, rating, and original_price) is weak (0.18–0.31), meaning the number of reviews does not strongly depend on the price or rating.

- Rating and original_price also show a moderate relationship (0.57), implying that moderately priced products may maintain consistent quality and ratings.

-

INNOMATICS RESEARCH LABS

Data Visualization:

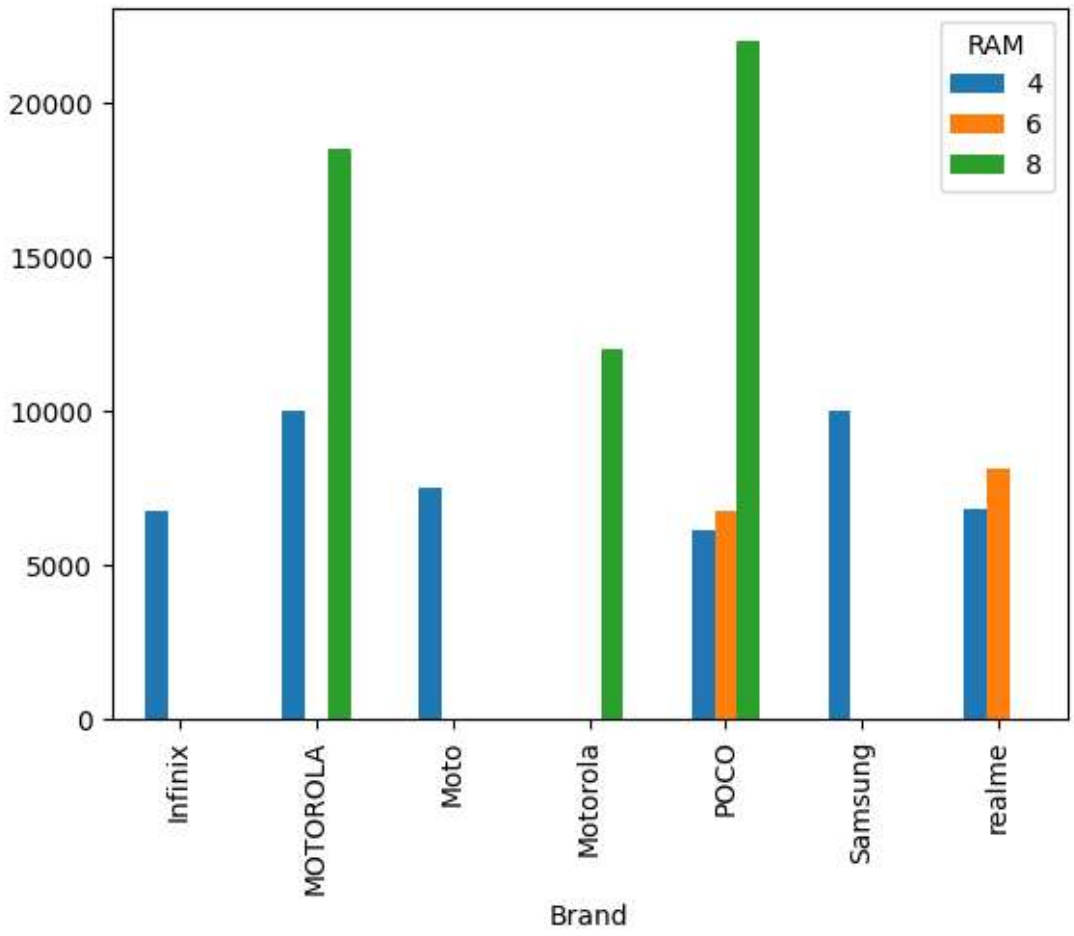
- POCO and MOTOROLA have the highest average prices for models with 8 GB RAM, crossing ₹20,000, indicating their focus on high-performance devices in the premium category.

- realme and Samsung also show moderate price growth with higher RAM configurations, suggesting balanced pricing across performance levels.

- Infinix remains the most budget-friendly brand, maintaining lower prices even for higher RAM models.

- Brands like Moto and Motorola (duplicate entry) indicate some inconsistency in labeling, which may require data cleaning to merge identical brands.

- The chart clearly shows that price increases consistently with higher RAM, confirming the expected relationship between device performance and cost.



INNOMATICS RESEARCH LABS

Conclusion:

1.The **Flipkart Mobile Dataset Analysis** provided valuable insights into the **pricing patterns, brand performance, customer preferences, and rating trends** across various mobile products listed on Flipkart.

2.By using **web scraping**, the dataset was successfully collected from Flipkart's mobile listings and then carefully **cleaned and preprocessed** to remove inconsistencies, duplicates, and missing values.

3.Price vs. Rating Relationship: Helped identify how pricing influences customer satisfaction.

4.Discount Impact: Showed how discount offers correlate with customer ratings and product popularity.

5.Brand Performance: Enabled comparison of leading brands (Samsung, Redmi, Realme, etc.) based on price, average rating, and feature offerings.

6.Customer Preference Trends: Highlighted which specifications (RAM, storage, battery life) contribute most to higher customer ratings.

INNOMATICS RESEARCH LABS

Q&A

Experience– Web Scraping & Data Analysis:

- Scraped *Flipkart- mobile* data from multiple websites using Python (Beautiful Soup, pandas).
- Cleaned and standardized data.
- Created visualizations showing mobile performance and prices.
- Learned end-to-end data workflow: **scraping → cleaning → analysis → visualization.**

INNOMATICS RESEARCH LABS

Challenges – Web Scraping & Data Analysis:

- Inconsistent webpage structures across seasons.
- Missing or mismatched match details.
- Handling duplicate and unclean data.
- Overcoming request blocks and slow scraping.
- Choosing the most meaningful visuals for insights.

INNOMATICS RESEARCH LABS

INNOMATICS
RESEARCH LABS