END-TO-END E-COMMERCE DATA SCIENCE ANALYSIS AND MACHINE LEARNING ON FLIPKART PRODUCT DATA

Ashmithaa Pradeep
August 2025

PROJECT OVERVIEW

This capstone project involves analyzing real-world e-commerce product data scraped from the Flipkart website. The aim is to identify product trends, customer preferences, and performance patterns using data science techniques. The project follows the complete data science lifecycle, including data collection, data cleaning, data storage, exploratory data analysis, unsupervised learning, supervised learning, and hyperparameter tuning, ultimately generating actionable business insights.

PROBLEM STATEMENT

Flipkart provides extensive product data, but it is available in raw and unstructured form, making analysis challenging. This project scrapes and structures Flipkart product data and applies machine learning to uncover trends, classify products, and support data-driven product and marketing decisions.

WEB SCRAPING

Source:
Flipkart E-Commerce Website

Tools & Technologies:
Python
BeautifulSoup / Requests
Pandas

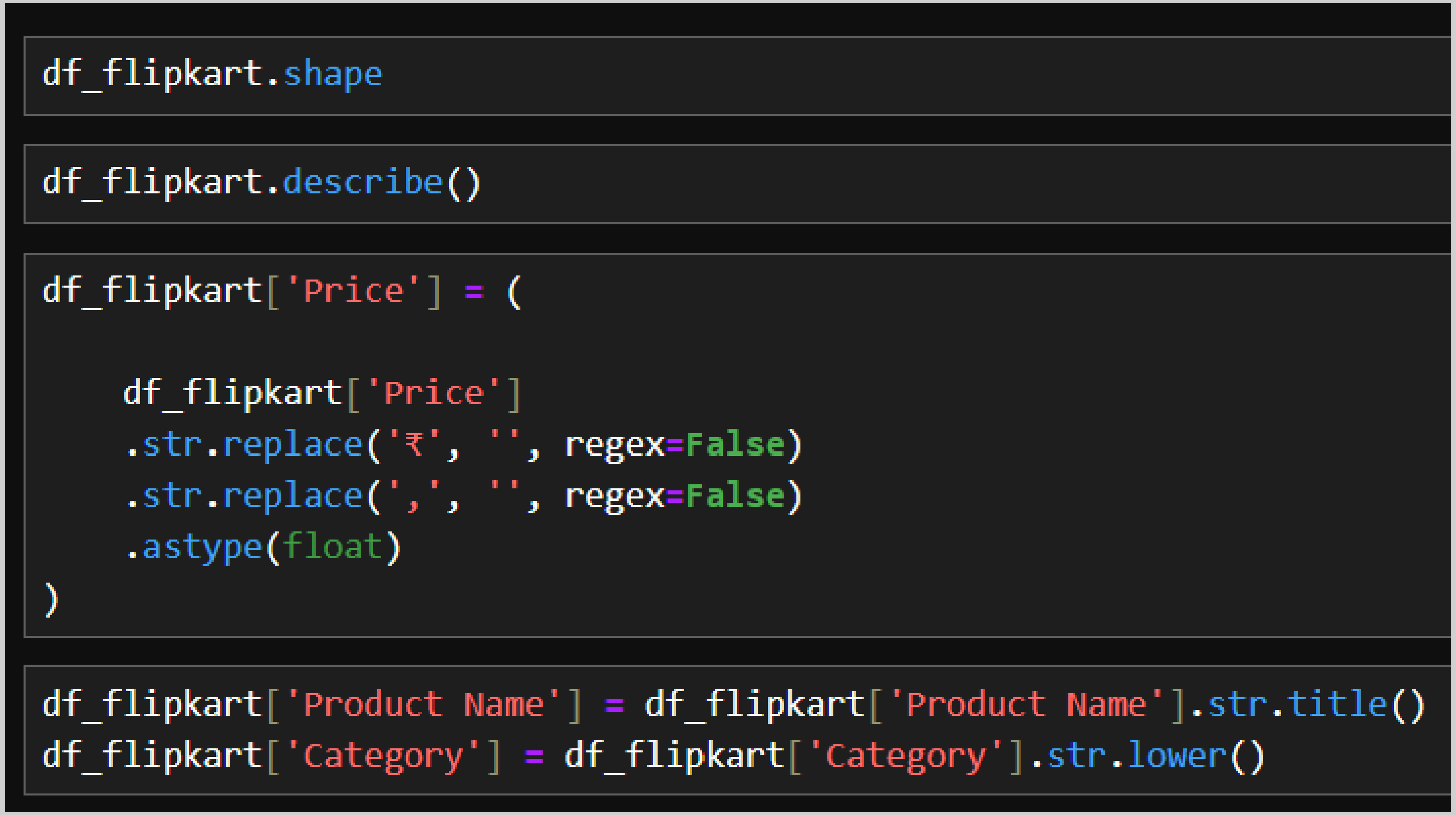
Data Scraped:
Product Name
Price
Category
Ratings
Number of Reviews

The data was scraped in an ethical manner, following the website’s structure and usage policies, and stored in CSV format for subsequent processing.

DATA CLEANING & PREPROCESSING

A relational database was designed using SQLAlchemy, where the cleaned data was stored in structured tables. This setup enabled efficient querying and seamless reuse of data for modeling and analysis.

DATA CLEANING



```
df_flipkart.shape
```

```
df_flipkart.describe()
```

```
df_flipkart['Price'] = (

    df_flipkart['Price']
    .str.replace('₹', '', regex=False)
    .str.replace(',', '', regex=False)
    .astype(float)
)
```

```
df_flipkart['Product Name'] = df_flipkart['Product Name'].str.title()
df_flipkart['Category'] = df_flipkart['Category'].str.lower()
```

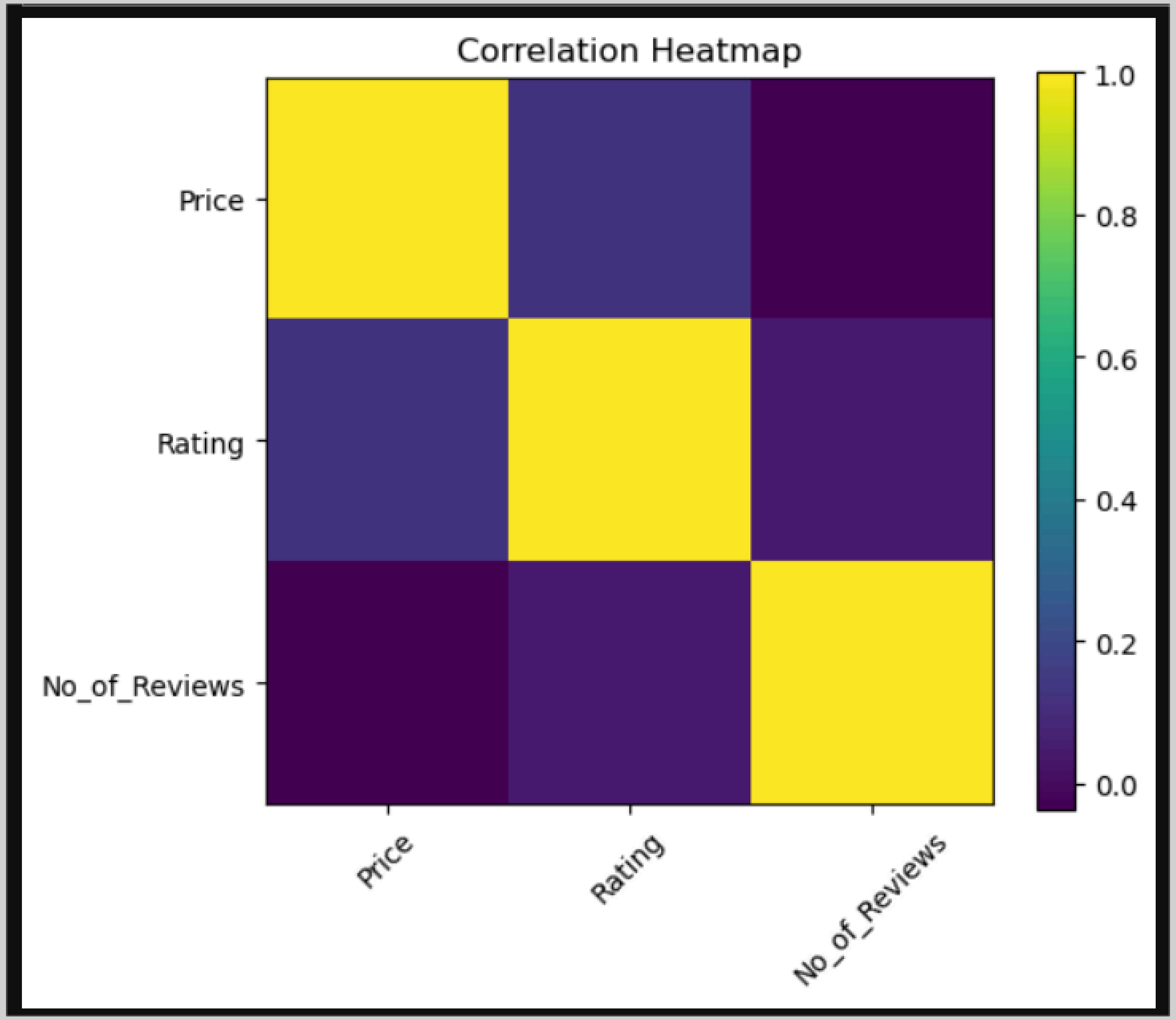
EXPLORATORY DATA ANALYSIS(EDA)

Exploratory Data Analysis (EDA) was conducted to examine:

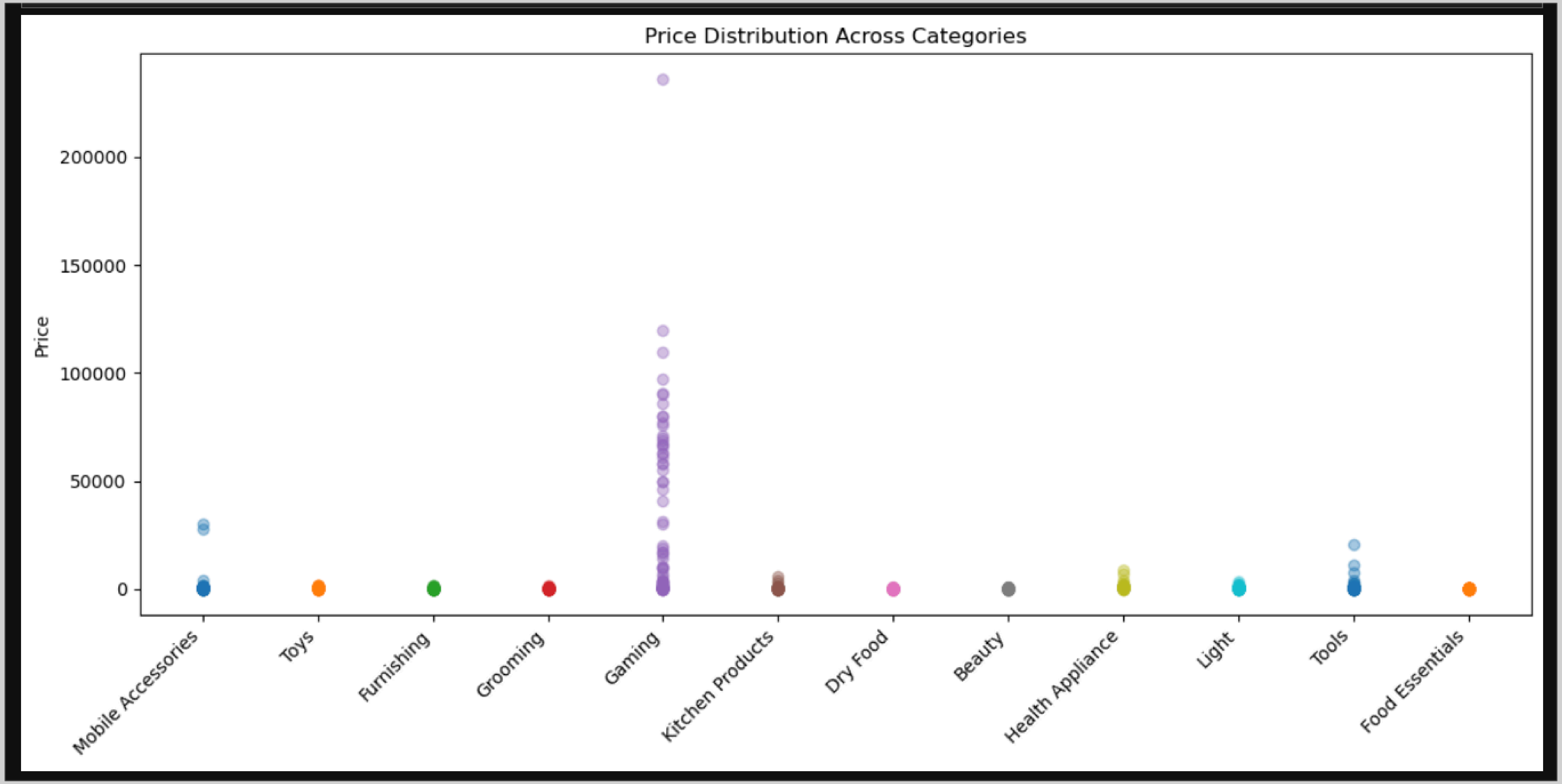
Price distributions
Rating trends
Category-wise product distribution
Review count patterns.

This analysis helped identify outliers, popular product categories, and highly rated items.

CORRELATION HEAT MAP

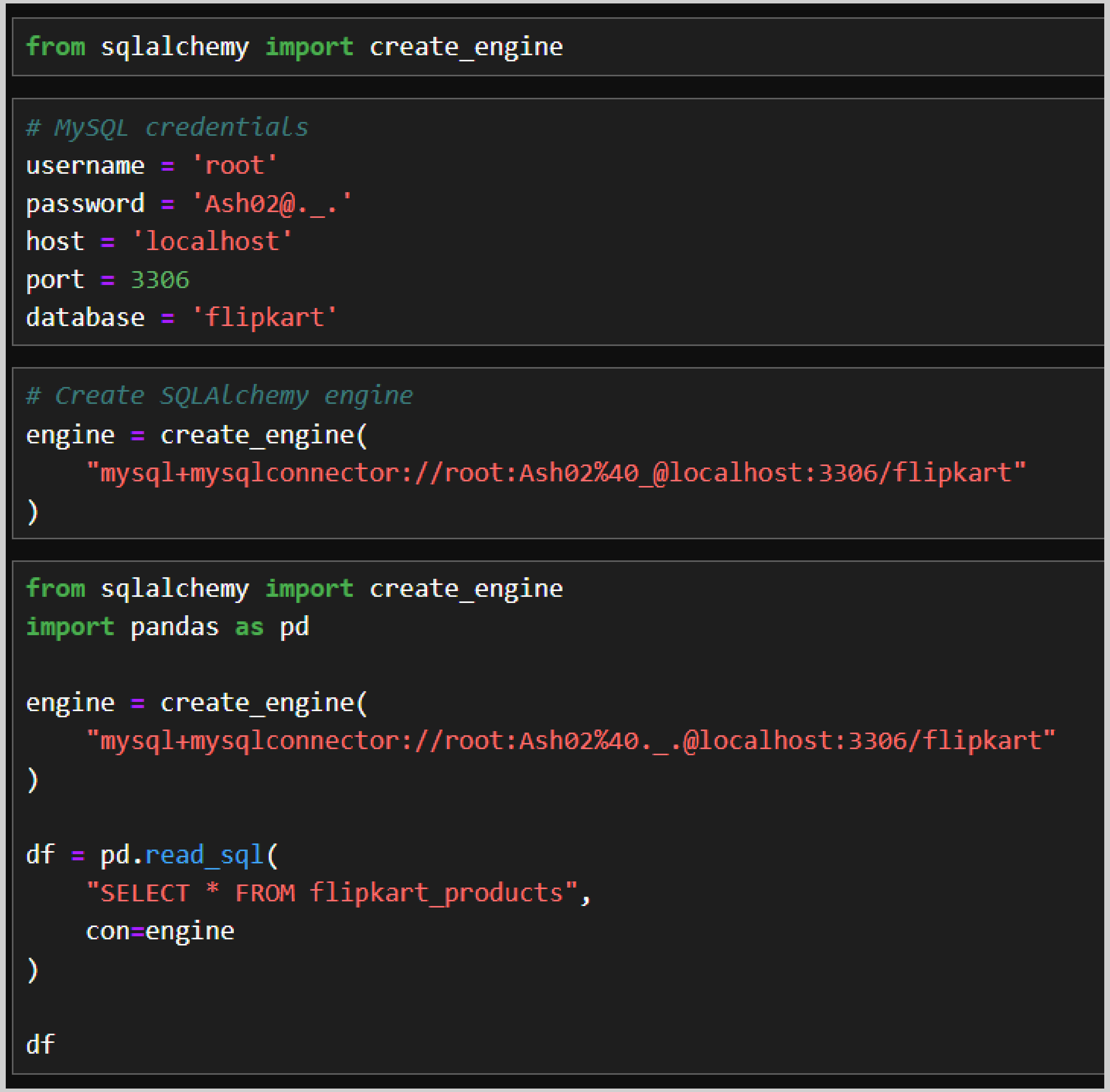


PRICE DISTRIBUTION



DATA STORAGE

A relational database was designed using SQLAlchemy, where the cleaned data was stored in structured tables.
This setup enabled efficient querying and seamless reuse of data for modeling and analysis.



```python
from sqlalchemy import create_engine
```

```python
# MySQL credentials
username = 'root'
password = 'Ash02@._.'
host = 'localhost'
port = 3306
database = 'flipkart'
```

```python
# Create SQLAlchemy engine
engine = create_engine(
    "mysql+mysqlconnector://root:Ash02%40_@localhost:3306/flipkart"
)
```

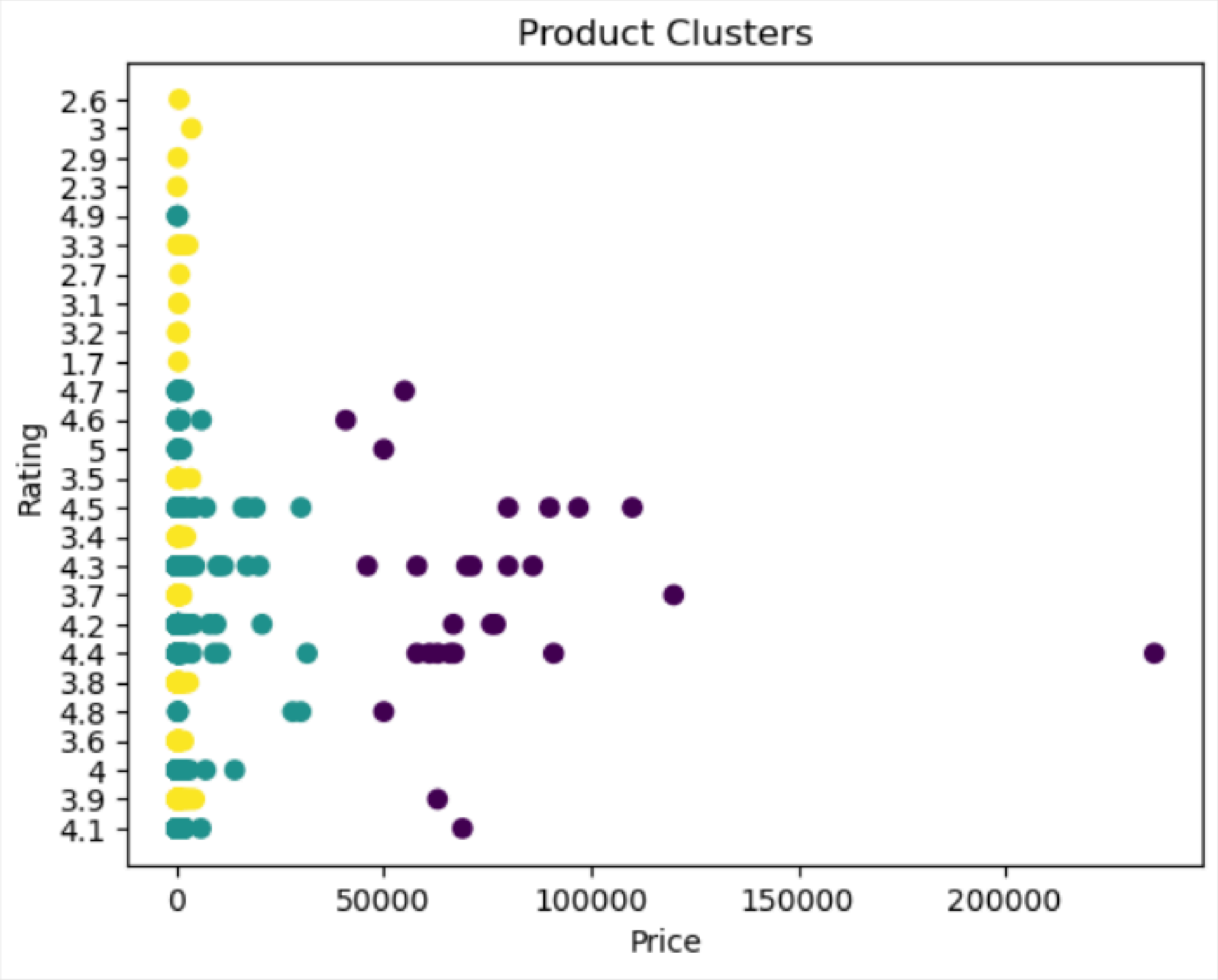
```python
from sqlalchemy import create_engine
import pandas as pd

engine = create_engine(
    "mysql+mysqlconnector://root:Ash02%40._.@localhost:3306/flipkart"
)

df = pd.read_sql(
    "SELECT * FROM flipkart_products",
    con=engine
)

df
```

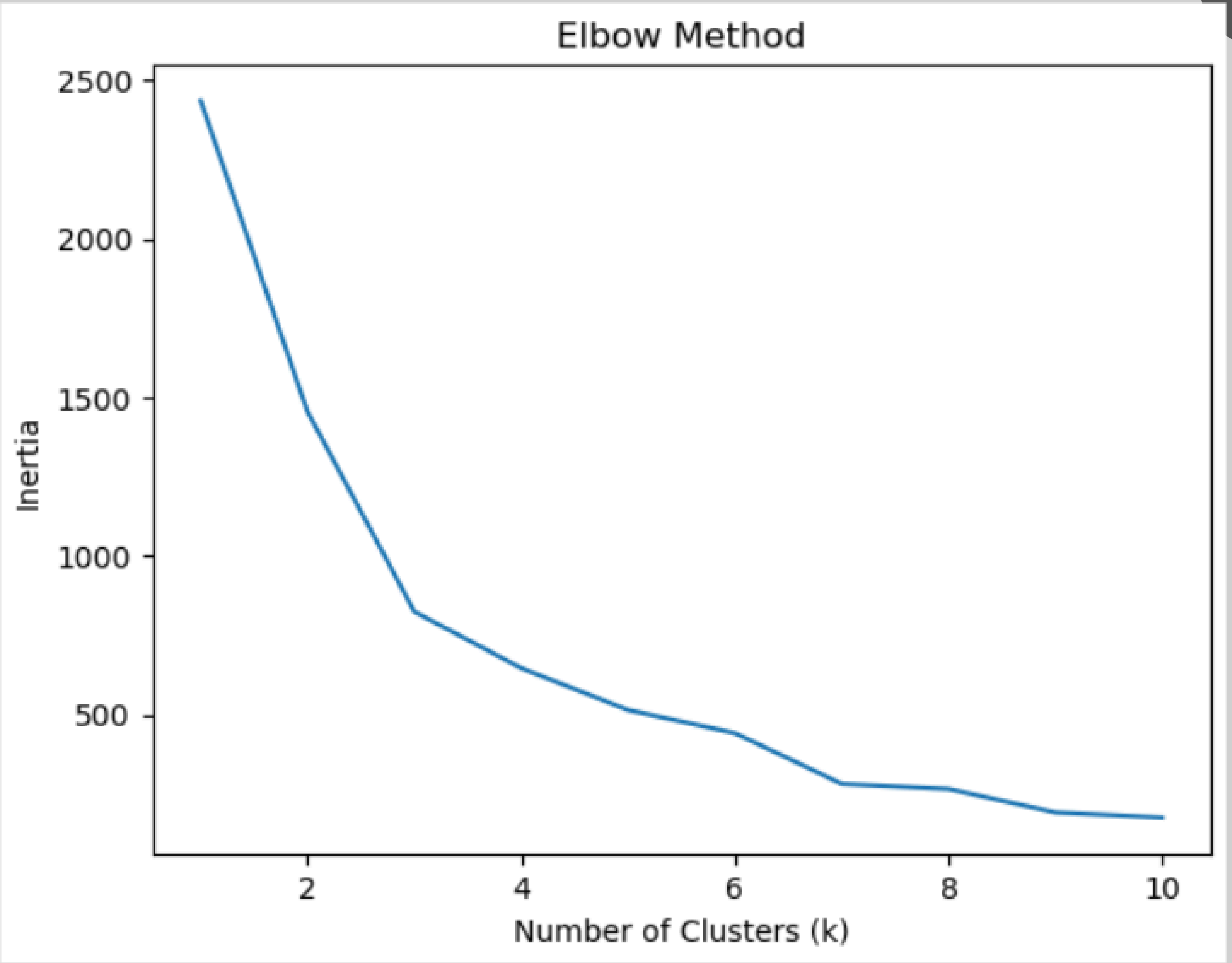
UNSUPERVISED LEARNING (CLUSTERING)

Technique Used:
K-Means Clustering

Process:
Relevant numerical features were selected, and multiple cluster values were evaluated using the Elbow Method. Cluster labels were then assigned to each product.

Outcome:
Products were grouped based on similarities in price, ratings, and review counts, enabling the identification of premium, mid-range, and budget product clusters.


Product Clusters
Rating
2.6
3
2.9
2.3
4.9
3.3
2.7
3.1
3.2
1.7
4.7
4.6
5
3.5
4.5
3.4
4.3
3.7
4.2
4.4
3.8
4.8
3.6
4
3.9
4.1
0
50000
100000
150000
200000
Price

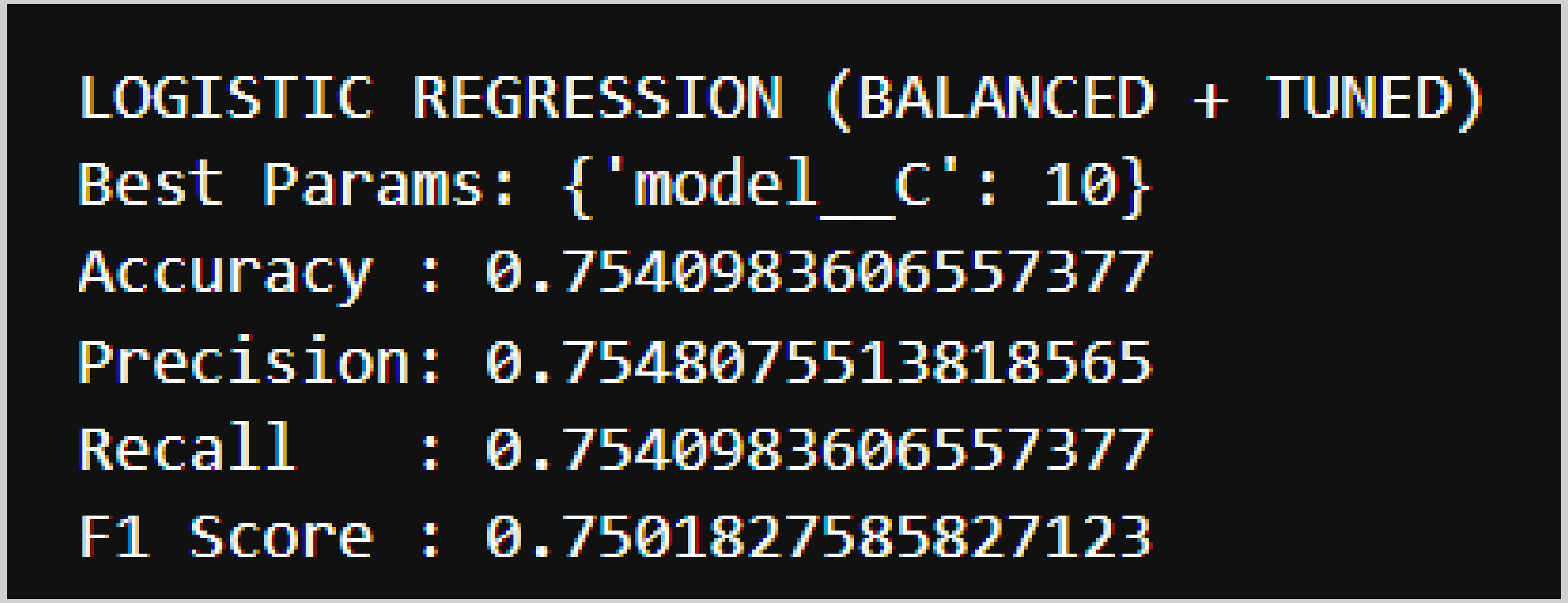

Elbow Method
Inertia
2500
2000
1500
1000
500
2
4
6
8
10
Number of Clusters (k)

SUPERVISED LEARNING

Evaluation Metrics:

Accuracy
Precision
Recall
F1 Score were used to assess model performance.

HYPERPARAMETER TUNING



```
LOGISTIC REGRESSION (BALANCED + TUNED)
Best Params: {'model__C': 10}
Accuracy : 0.7540983606557377
Precision: 0.7548075513818565
Recall   : 0.7540983606557377
F1 Score : 0.7501827585827123
```

Precision and recall were well balanced, and the tuned Logistic Regression model outperformed more complex models while remaining interpretable and computationally efficient.

INSIGHTS

Proper data balancing and hyperparameter tuning contributed more to performance improvement than increasing model complexity. Logistic Regression proved to be an ideal choice due to its high accuracy, interpretability, and faster training and deployment. Additionally, the Flipkart product data exhibited strong linear patterns, with ratings and review counts playing a significant role in product classification.

CONCLUSION

This project successfully demonstrates an end-to-end data science workflow using real-world data scraped from Flipkart. By integrating web scraping, data preprocessing, exploratory analysis, clustering, supervised learning, and hyperparameter tuning, meaningful insights were derived to support data-driven business decisions. The final tuned Logistic Regression model delivered the best performance and was selected for deployment.

THANK YOU