CAPSTONE PROJECT DATA SCIENCE

Latha Christiyana J

1.Introduction:

The mobile phone market is highly competitive, with numerous brands offering products across various price ranges and specifications. Understanding customer preferences, ratings, and pricing trends is crucial for market analysis and strategic decision-making. This project focuses on analyzing Flipkart mobile phone data to extract insights using statistical analysis, clustering, and supervised learning techniques. Data cleaning and preprocessing were performed to ensure accuracy, followed by hypothesis testing, K-Means clustering, and classification models including Logistic Regression, SVM, KNN, Random Forest, and XGBoost. The goal is to identify patterns, segment the market, and predict mobile phone performance based on key features such as price, ratings, and number of reviews**.**

2.Objective of the Project:

The main objective of this project is to analyzed mobile phone data from Flipkart to gain insights into pricing, customer ratings, and product performance. The project aims to:

- Scraped the real world data from flipkart using beautiful soup.
- Perform data cleaning and preprocessing to ensure accuracy and consistency.
- Analyze the relationship between price and customer ratings using statistical hypothesis testing.
- Segment mobiles into clusters based on price, ratings, and number of reviews using K-Means clustering.
- Build and evaluate supervised learning models (Logistic Regression, SVM, KNN, Random Forest, XGBoost) to classify mobiles and predict trends.
- Identify key patterns and actionable insights for market segmentation and customer preference analysis.

3.Web Scraping:

performed web scraping using Python libraries, including Beautiful Soup, to extract mobile phone data from Flipkart.

A total of 10 pages were scraped, capturing essential product information.

The dataset contains 5 columns—Price, Rating, Number of Reviews, Product Name, and Category—and 668 rows of data.

Finally, the dataset was saved as a CSV file for further analysis and modeling.

```
# Set Chrome options for the web driver
options = Options()
# options.add_argument("--headless")
options.add_argument("--disable-gpu")
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')

service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service, options=options)

# URL of the Flipkart mobile search results page
url = 'https://www.flipkart.com/search?q=mobiles+5g'
driver.get(url)
time.sleep(5)

# Initialize empty lists to store scraped data
titles = []
prices = []
categories = []
ratings = []
number_of_reviews = []
page_counter = 0
```

```
# Define the maximum number of pages to scrap
max_pages = 10

while page_counter < max_pages:
    content = driver.page_source
    soup = BeautifulSoup(content, "html.parser")
# Find product containers
    products = soup.find_all("div", class_="cPHDOP col-12-12")

    for product in products:
```

	Title	Price	Category	Rating	Number of Reviews
0	Unknown	0	mobiles 5g	Unknown	0
1	Unknown	0	mobiles 5g	Unknown	0
2	OPPO K13x 5G 6000mAh and 45W SUPERVOOC Charger...	₹11,999	mobiles 5g	4.5	13,953 Ratings & 1,243 Reviews
3	Nothing Phone (3) (Black, 256 GB)	₹79,999	mobiles 5g	4.4	1,735 Ratings & 158 Reviews
4	vivo T4x 5G (Pronto Purple, 128 GB)	₹14,499	mobiles 5g	4.4	1,18,524 Ratings & 5,772 Reviews
...	...	...	...	...	...
663	Tecno Pop 9 5G (Aurora Cloud, 64 GB)	₹8,699	mobiles 5g	4.1	127 Ratings & 7 Reviews
664	MOTOROLA Edge 60 Pro (Pantone Walnut, 256 GB)	₹29,999	mobiles 5g	4.3	10,271 Ratings & 889 Reviews
665	Unknown	0	mobiles 5g	Unknown	0
666	Unknown	0	mobiles 5g	4.4	0
667	Unknown	0	mobiles 5g	Unknown	0

668 rows × 5 columns

4.Data Cleaning Process:

Before performing the analysis, Performed a thorough data cleaning process to ensure the dataset was accurate and consistent.

All unknown and zero values were removed to avoid misleading results.

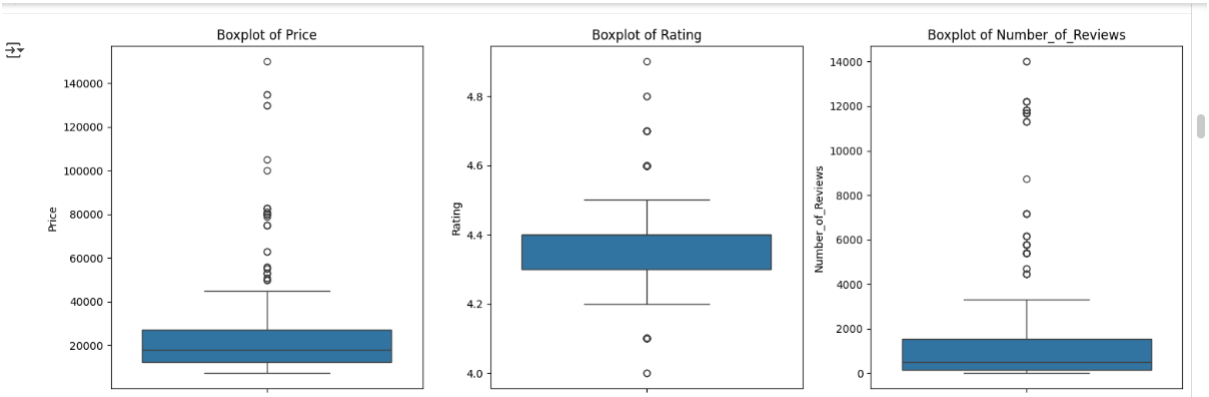
The currency symbols were stripped and the data formats were standardized for uniformity.

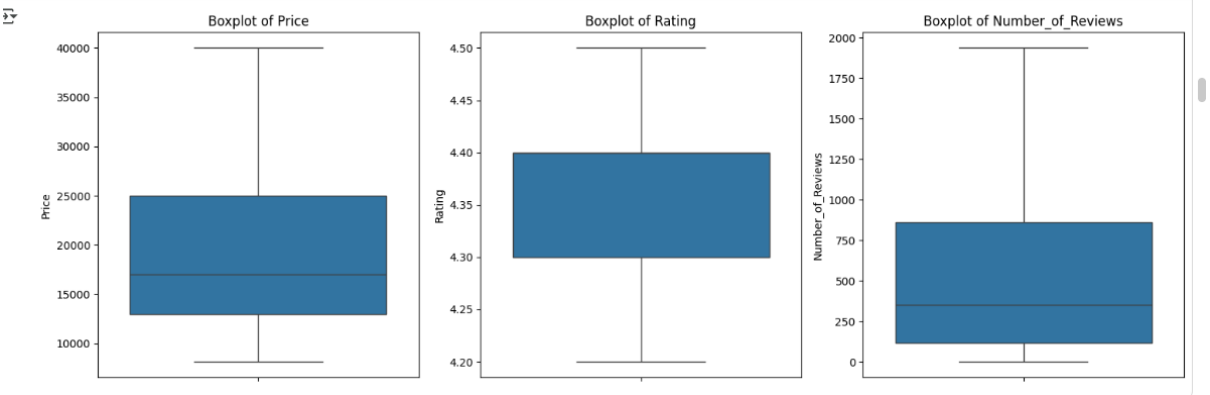
Duplicate entries were eliminated to maintain data integrity.

Additionally, outliers were identified and removed by IQR method to improve the quality of the analysis.

	Title	Price	Category	Rating	Number of Reviews
0	OPPO K13x 5G 6000mAh and 45W SUPERVOOC Charger...	₹11,999	mobiles 5g	4.5	13,953 Ratings & 1,243 Reviews
1	Nothing Phone (3) (Black, 256 GB)	₹79,999	mobiles 5g	4.4	1,735 Ratings & 158 Reviews
2	vivo T4x 5G (Pronto Purple, 128 GB)	₹14,499	mobiles 5g	4.4	1,18,524 Ratings & 5,772 Reviews
3	vivo T4x 5G (Glacial Teal, 128 GB)	₹13,499	mobiles 5g	4.4	1,57,603 Ratings & 7,149 Reviews
4	vivo T4 5G (Phantom Grey, 128 GB)	₹20,999	mobiles 5g	4.5	70,882 Ratings & 3,232 Reviews

	Title	Price	Category	Rating	Number_of_Reviews
0	OPPO K13x 5G 6000mAh and 45W SUPERVOOC Charger...	11999.0	mobiles 5g	4.5	1243
1	Nothing Phone (3) (Black, 256 GB)	79999.0	mobiles 5g	4.4	158
2	vivo T4x 5G (Pronto Purple, 128 GB)	14499.0	mobiles 5g	4.4	5772
3	vivo T4x 5G (Glacial Teal, 128 GB)	13499.0	mobiles 5g	4.4	7149
4	vivo T4 5G (Phantom Grey, 128 GB)	20999.0	mobiles 5g	4.5	3232





5.Statistical Analysis:

Performed basic statistical Analysis mean,median,mode.

performed hypothesis tesing to explore the relationship between mobile phone price and customer ratings.

An independent two-sample t-test was conducted to compare the average ratings of expensive (> ₹20,000) and cheap (≤ ₹20,000) phones.

Got p-value< 0.05(reject null hypothesis)

The test revealed a significant difference in ratings, indicating that expensive phones generally receive higher customer ratings.

This analysis helped in understanding the impact of price on customer perception.

	Price	Rating	Number_of_Reviews
count	376.000000	376.000000	376.000000
mean	19385.571809	4.346011	539.430851
std	7825.977532	0.098712	528.450440
min	8150.000000	4.200000	2.000000
25%	12999.000000	4.300000	115.000000
50%	16999.000000	4.400000	351.000000
75%	24999.000000	4.400000	862.000000
max	39999.000000	4.500000	1936.000000

```
import scipy.stats as stats
import numpy as np

# Create groups
expensive = df_cleaned[df_cleaned['Price'] > 20000]['Rating'].dropna()
cheap = df_cleaned[df_cleaned['Price'] <= 20000]['Rating'].dropna()

# Perform independent t-test
t_stat, p_value = stats.ttest_ind(expensive, cheap, equal_var=False)

print(f"T-statistic: {t_stat:.3f}")
print(f"P-value: {p_value:.4f}")

if p_value < 0.05:
    print(" Significant difference in average ratings between expensive and cheap phones.")
else:
    print(" No significant difference in average ratings between expensive and cheap phones.")
```

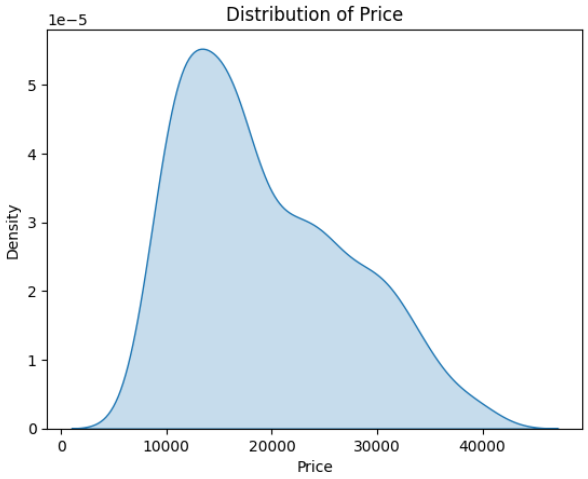
```
T-statistic: 10.723
P-value: 0.0000
 Significant difference in average ratings between expensive and cheap phones.
```

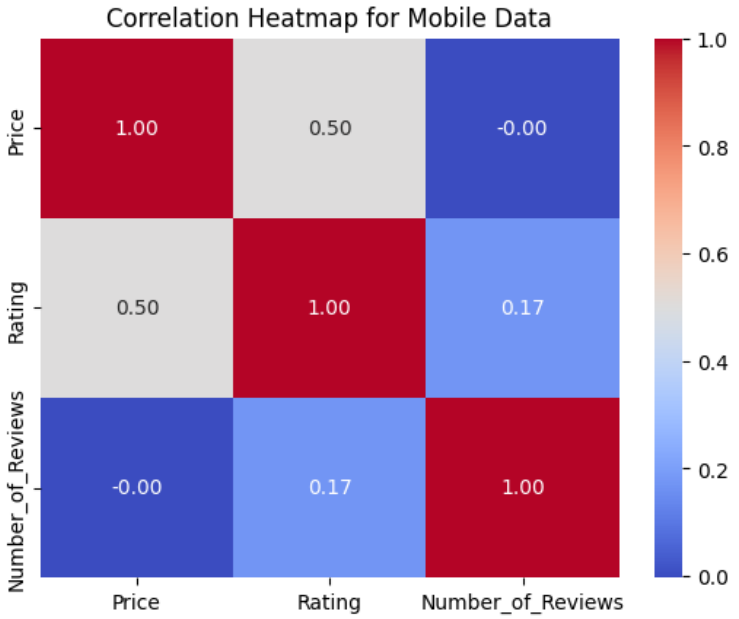
6.Exploratory Data Analysis:

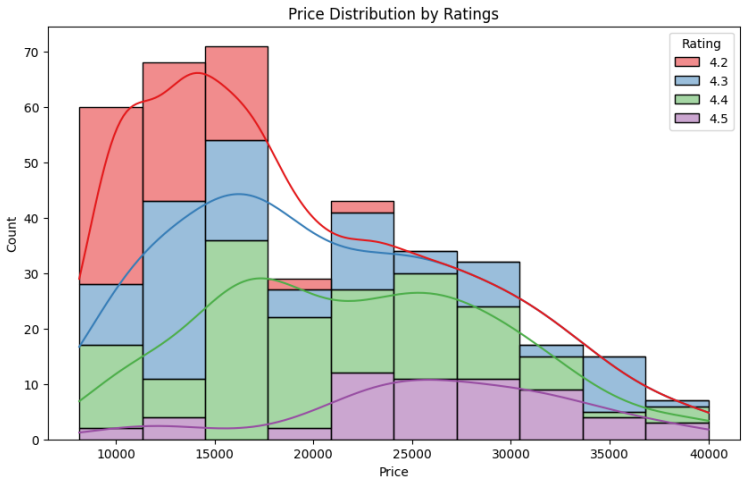
Exploratory Data Analysis (EDA) , examined the price distribution to understand the range and spread of mobile phone prices.

A correlation heatmap was generated to identify relationships between key features such as price, ratings, and number of reviews.

Additionally, performed the price distribution by ratings to explore how customer ratings vary across different price points.

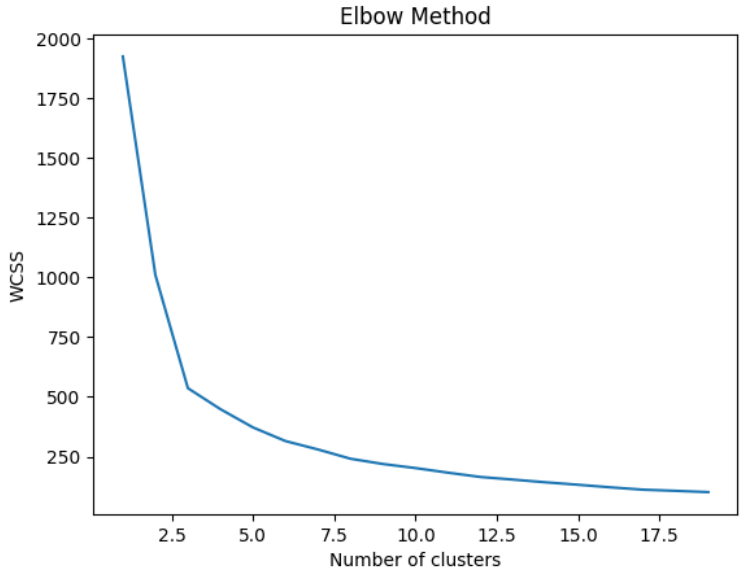

Distribution of Price
1e−5
Density
Price
0
10000
20000
30000
40000
0
1
2
3
4
5

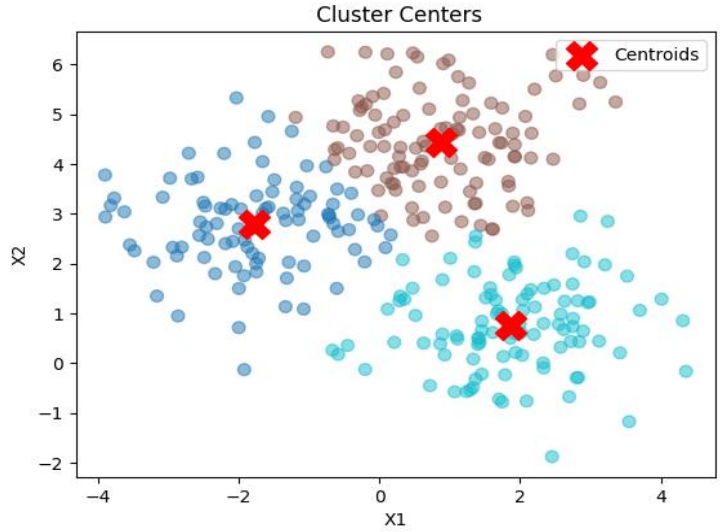

Correlation Heatmap for Mobile Data
Price
Rating
Number_of_Reviews
1.00
0.50
-0.00
0.50
1.00
0.17
-0.00
0.17
1.00
1.0
0.8
0.6
0.4
0.2
0.0


Price Distribution by Ratings
Rating
4.2
4.3
4.4
4.5
Count
Price
0
10
20
30
40
50
60
70
10000
15000
20000
25000
30000
35000
40000

7.Unsupervised Learning:

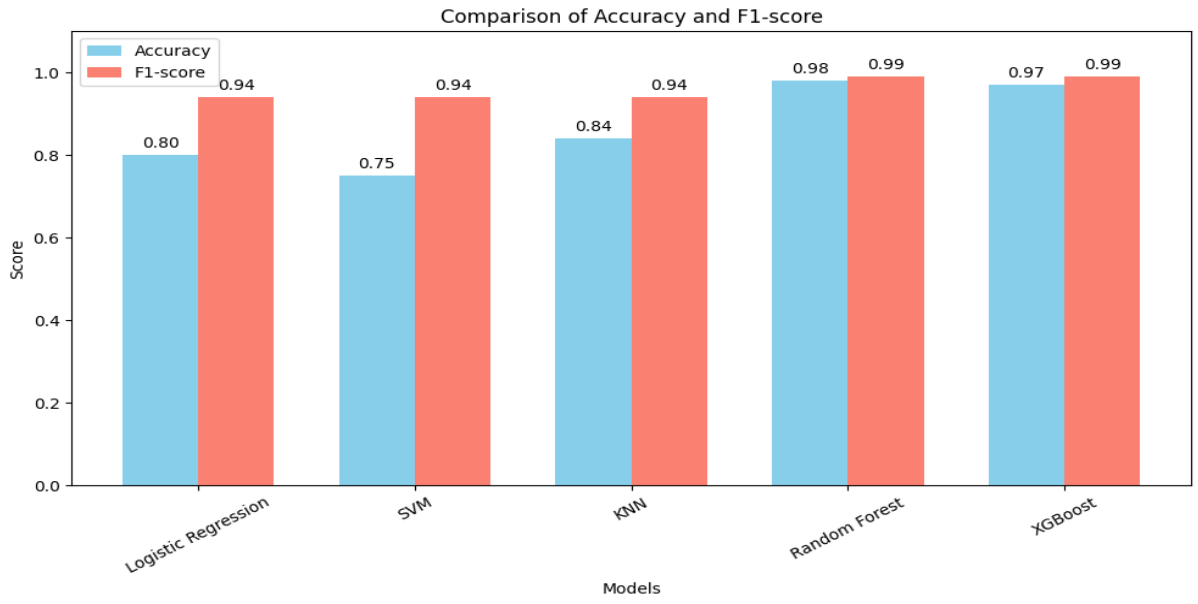
Performed K-Means clustering to group mobile phones based on Price, Rating, and Number of Reviews.
The Elbow Method was used to determine the optimal number of clusters for meaningful segmentation.
A new column was created in the dataset to label each product with its cluster group.
This clustering helped in identifying patterns and similarities among mobiles in different segments.
The analysis provided insights into customer preferences and market trends based on pricing and ratings.





8.Supervised Learning:

In the supervised learning phase, I implemented various classification algorithms to predict mobile phone categories based on features.
The models used included Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost.
The dataset was divided into training and testing sets to evaluate model performance on unseen data.
Each model was assessed using accuracy, precision, recall, and F1-score, and hyperparameter tuning was performed to improve results.
The evaluation helped identify the best-performing model for predicting mobile phone categories.Random Forest has got the highest accuracy rate and F1-score of 0.98 and 0.99.



8.Hyperparameter Tuning

In logistic Regression:

- C = 10 → The model allows less regularization, giving more flexibility to fit the data.
- penalty = 'l2' → Uses Ridge regularization to prevent overfitting.
- solver = 'lbfgs' → Optimization algorithm suitable for small-to-medium datasets.

- cross-validation score = 0.99 → The model was highly accurate on unseen data.

In KNN:

- neighbors = 9 → The model considers the 9 closest neighbors for classification.
- weights = 'distance' → Closer neighbors have more influence on the prediction.
- algorithm = 'auto' → Automatically selects the best algorithm for computing nearest neighbors.
- metric = 'euclidean' → Uses Euclidean distance to measure closeness between points.
- Best cross-validation score = 0.9 → KNN performing well

9.Key insights:

- Expensive mobiles (> ₹20,000) tend to have significantly higher ratings than cheaper ones.
- K-Means clustering identified three distinct groups of mobiles based on price, ratings, and number of reviews.
- Random Forest algorithm got both accuracy and Fl-score highly.
- Hyperparameter tuning improved model performance, and features like price, ratings, and reviews were key predictors.

10.Conclusion:

Scraped the data from Flipkart using beautiful soup. Performed data cleaning process , statistical Analysis, unsupervised learning , supervised learning and performed hyperparameter tuning identify the trends of real-world dataset. There is a strong relationship between price and ratings, with clustering revealing clear market segments. Random forest best for predicting mobile dataset.