Final Report: Mobile Phone Data Analysis

Objective

The goal of this project was to perform Exploratory Data Analysis (EDA) on a dataset of mobile phones, conduct predictive modeling to estimate user ratings based on various features, and perform clustering to group similar phones.

The dataset contains 984 entries and 10 columns. The columns are as follows:

1. **battery**: Describes the battery capacity.
2. **camera**: Details the camera specifications.
3. **display**: Describes the display size and type.
4. **memory**: Specifies the RAM and ROM capacities.
5. **name**: The name of the mobile phone.
6. **price**: The price of the mobile phone.
7. **processor**: Details the processor specifications.

8. **rating**: The user rating of the mobile phone.
9. **reviews**: The number of reviews.
10. **warranty**: Warranty information.

Some observations after performing data cleaning :

- The `processor`, `rating`, and `reviews` columns have missing values.
- The `warranty` column has a significant number of missing values (about 15%).
- Most columns are in the object (string) format, which will require parsing for meaningful analysis.

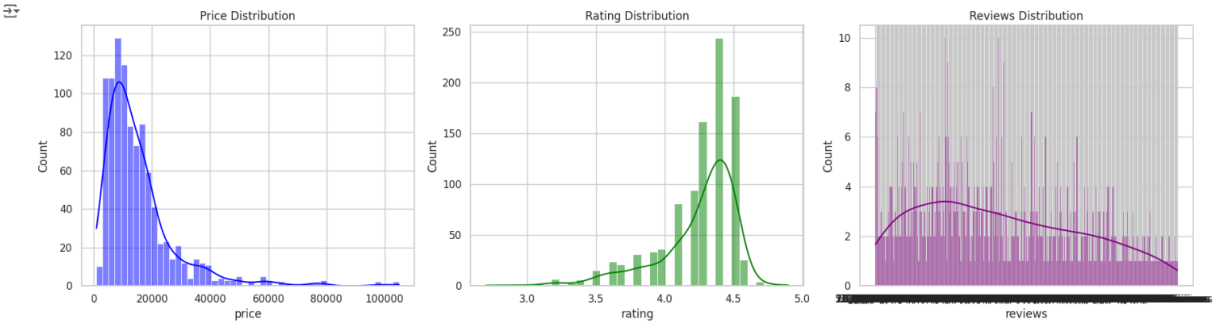
univariate analysis to understand the distribution and characteristics of each column.

The univariate analysis shows the following:

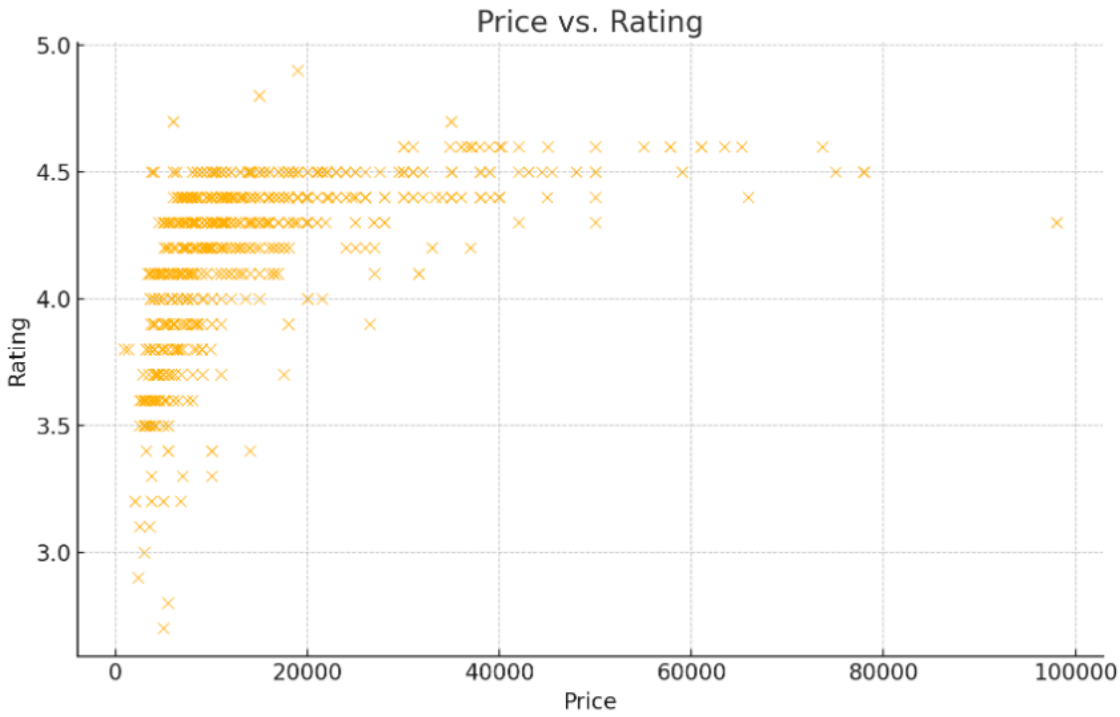
1. **Price Distribution: The prices are right-skewed, with most mobile phones priced between 5,000 and 15,000 units.**
2. **Rating Distribution: Ratings are generally high, clustering around 4.0 to 4.5.**
3. **Reviews Distribution: The number of reviews is also right-skewed, with most phones having fewer than 50,000 reviews.**

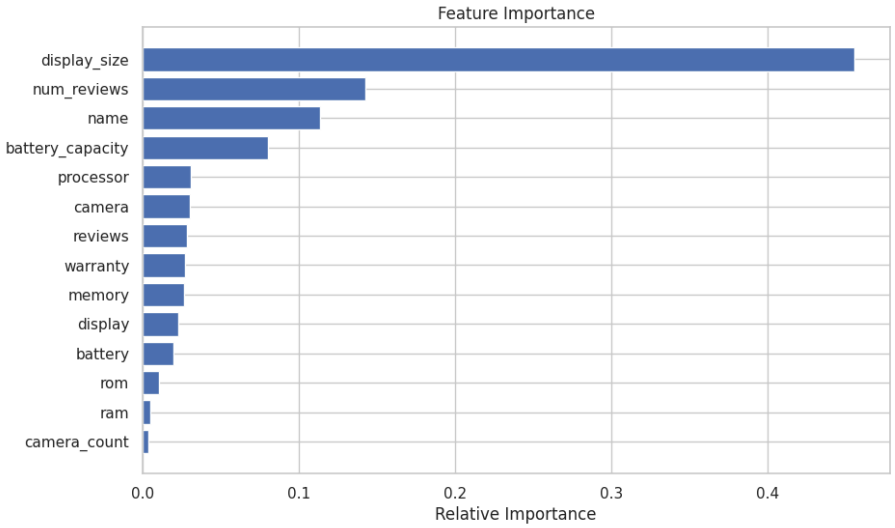
```
[28]
# Reviews Distribution
sns.histplot(data['reviews'].dropna(), kde=True, ax=axes[2], color='purple')
axes[2].set_title('Reviews Distribution')


plt.tight_layout()
plt.show()
```



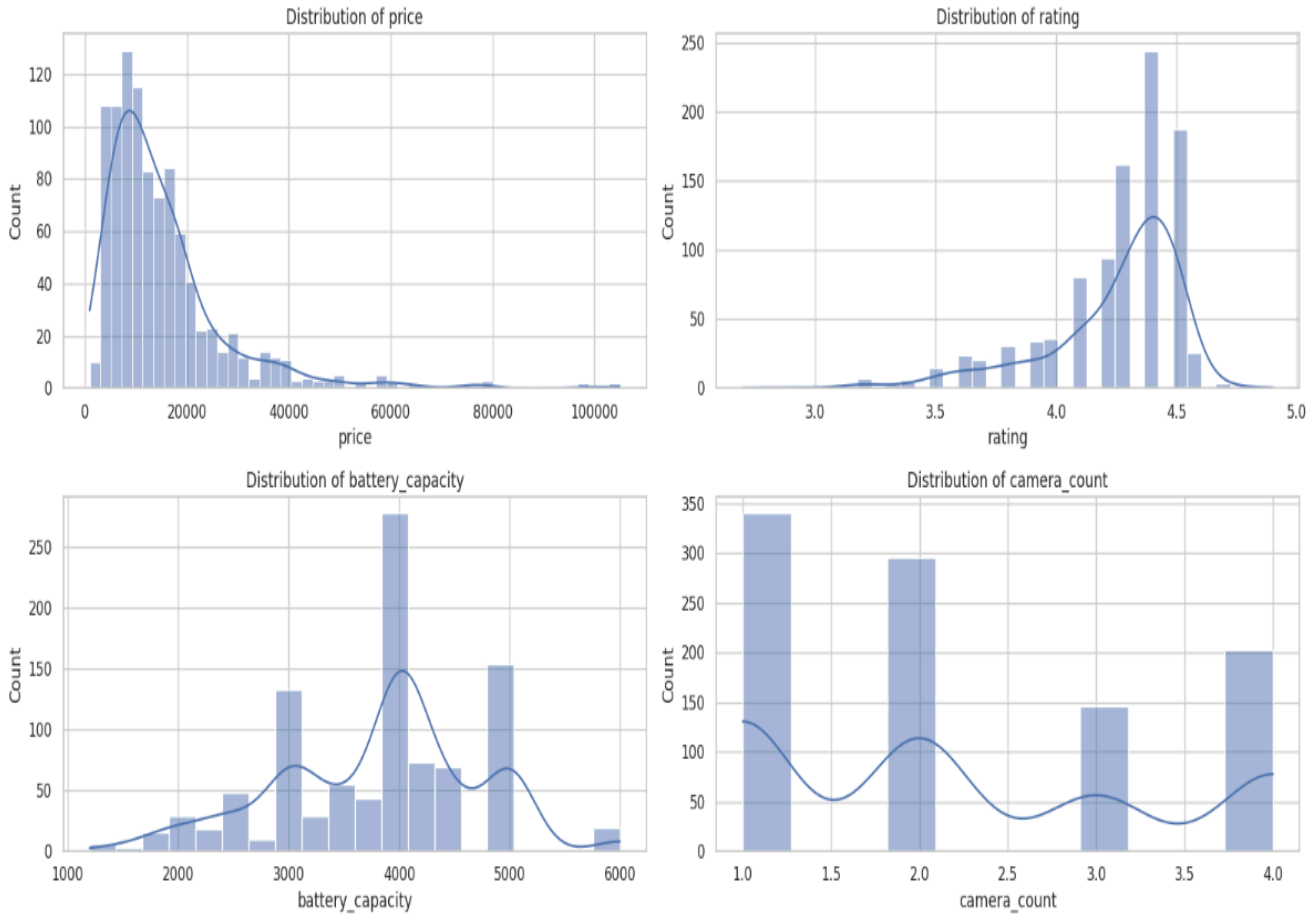
bivariate analysis to identify the relationships between price, rating, and other parameters

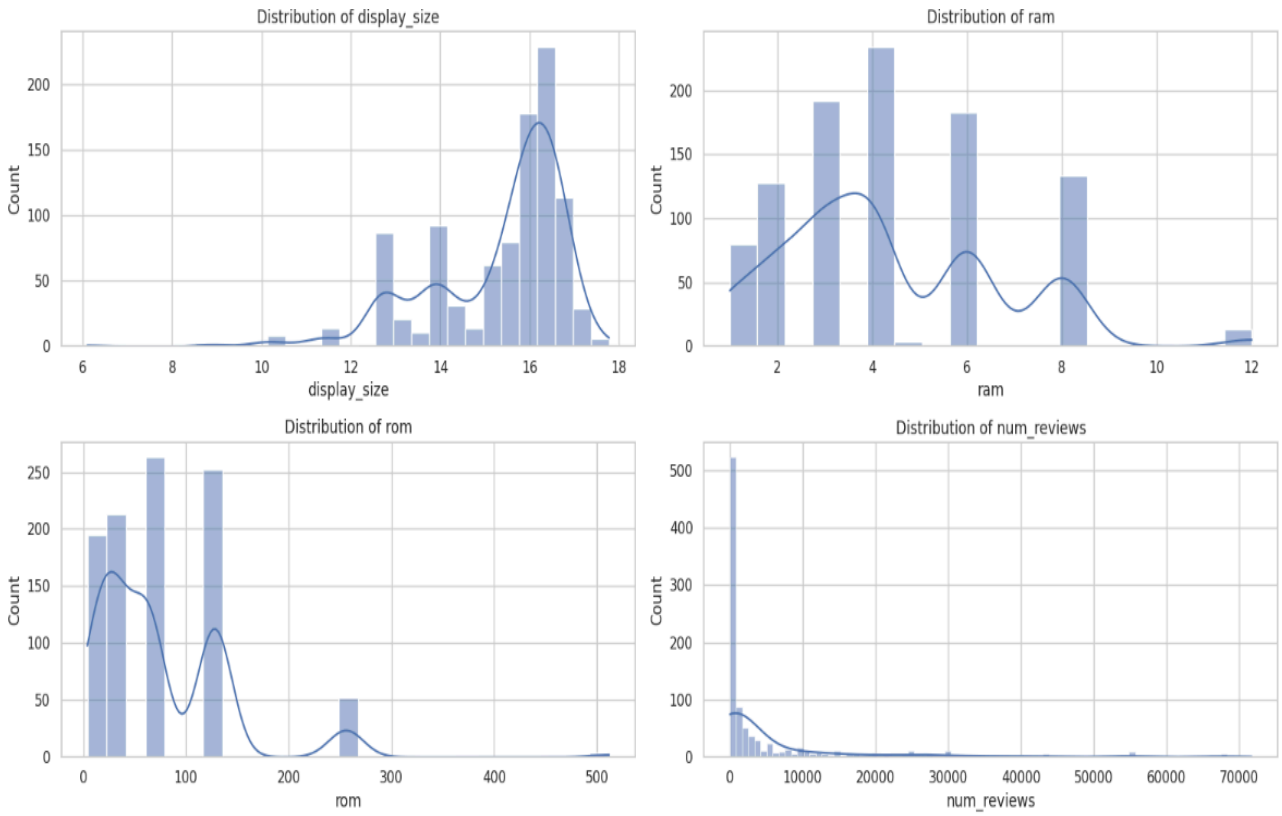





```
# Scatter plot of Actual vs. Predicted Ratings
plt.figure(figsize=(8, 6))
```

[18]


Distribution of price
Count
price
Distribution of rating
Count
rating
Distribution of battery_capacity
Count
battery_capacity
Distribution of camera_count
Count
camera_count



The univariate analysis reveals the following:

1. **Price: The distribution of phone prices is right-skewed, with most phones priced under 15,000 INR.**

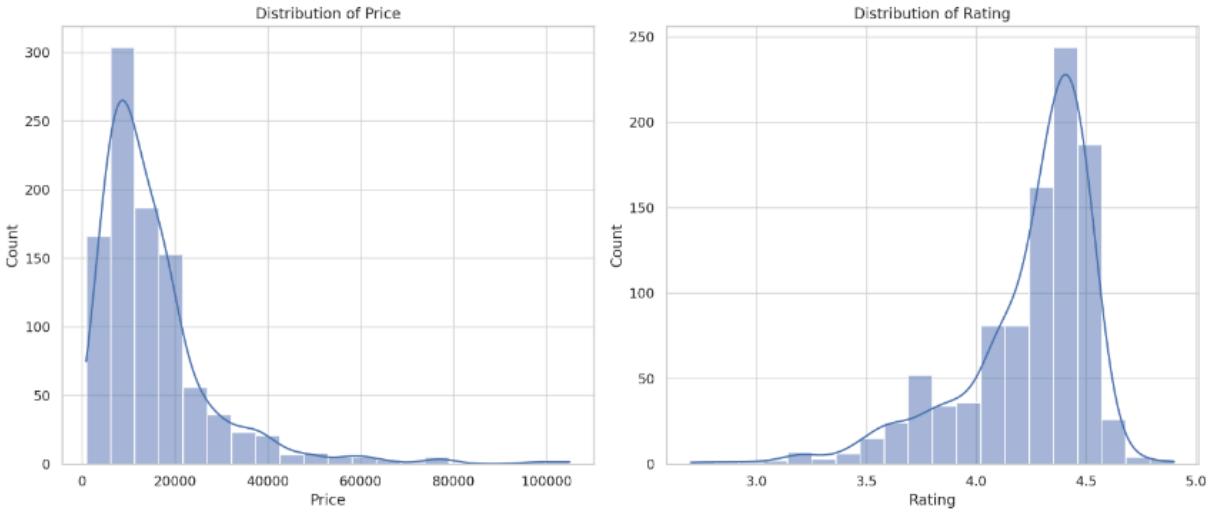
2. **Rating: Ratings are mostly clustered between 4 and 4.5, indicating generally favorable reviews.**
3. **Battery Capacity: The majority of phones have a battery capacity around 4000-5000 mAh.**
4. **Camera Count: Most phones have 2 or 4 cameras, reflecting the popularity of multi-camera setups.**
5. **Display Size: Phone displays are predominantly in the range of 15 to 16.5 inches.**
6. **RAM: The most common RAM sizes are 3 GB and 4 GB.**
7. **ROM (Storage): The most common storage sizes are 32 GB and 64 GB.**
8. **Number of Reviews: The number of reviews varies significantly, with a few phones receiving an extremely large number of reviews.**

Bivariate Analysis Insights:

1. **Price vs. Rating:**
    - **There is a slight positive correlation between price and rating. Higher-priced phones tend to have slightly better ratings, but the relationship is not very strong.**
2. **Price vs. Reviews:**
    - **The relationship between price and the number of reviews shows some variability, with a few high-priced phones receiving a large number of reviews. Most phones with moderate prices have a wide range of reviews.**
3. **Price vs. Battery:**
    - **Phones with higher battery capacities (like 4300 mAh Lithium-ion Battery) are typically more expensive. However, there is considerable variation within the mid-range battery categories.**
4. **Price vs. Camera:**
    - **More sophisticated camera setups, particularly those with higher megapixel counts and multiple cameras (e.g., 108MP cameras), are associated with higher prices.**

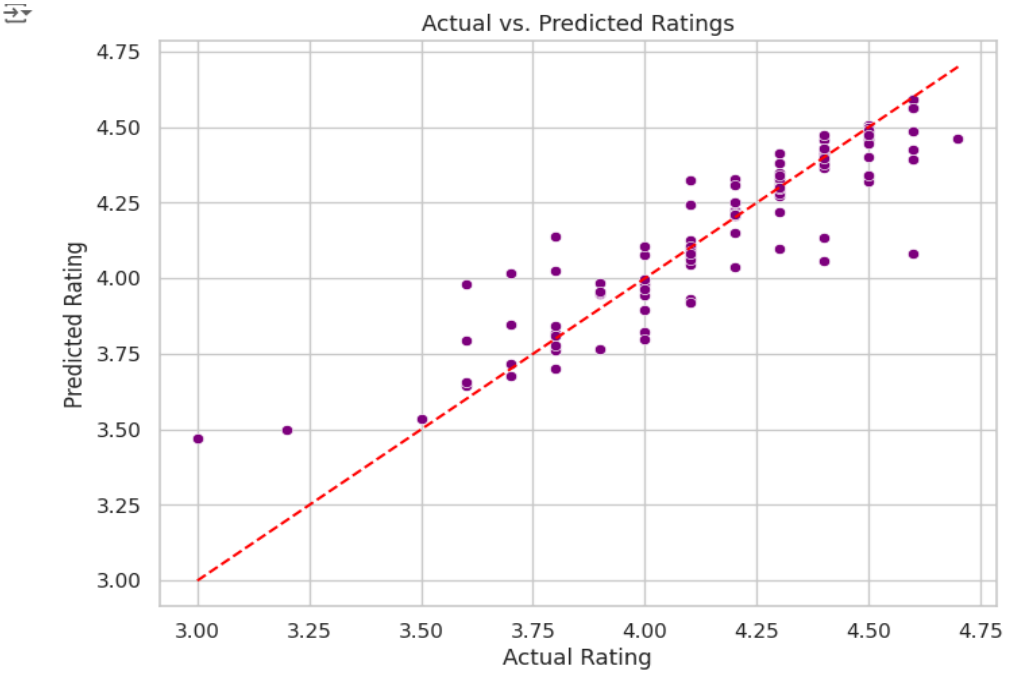
5. **Price vs. Memory:**
    - **Higher RAM and ROM configurations (e.g., 12 GB RAM and 256 GB ROM) correspond to higher prices, as expected.**

Distribution of price and rating

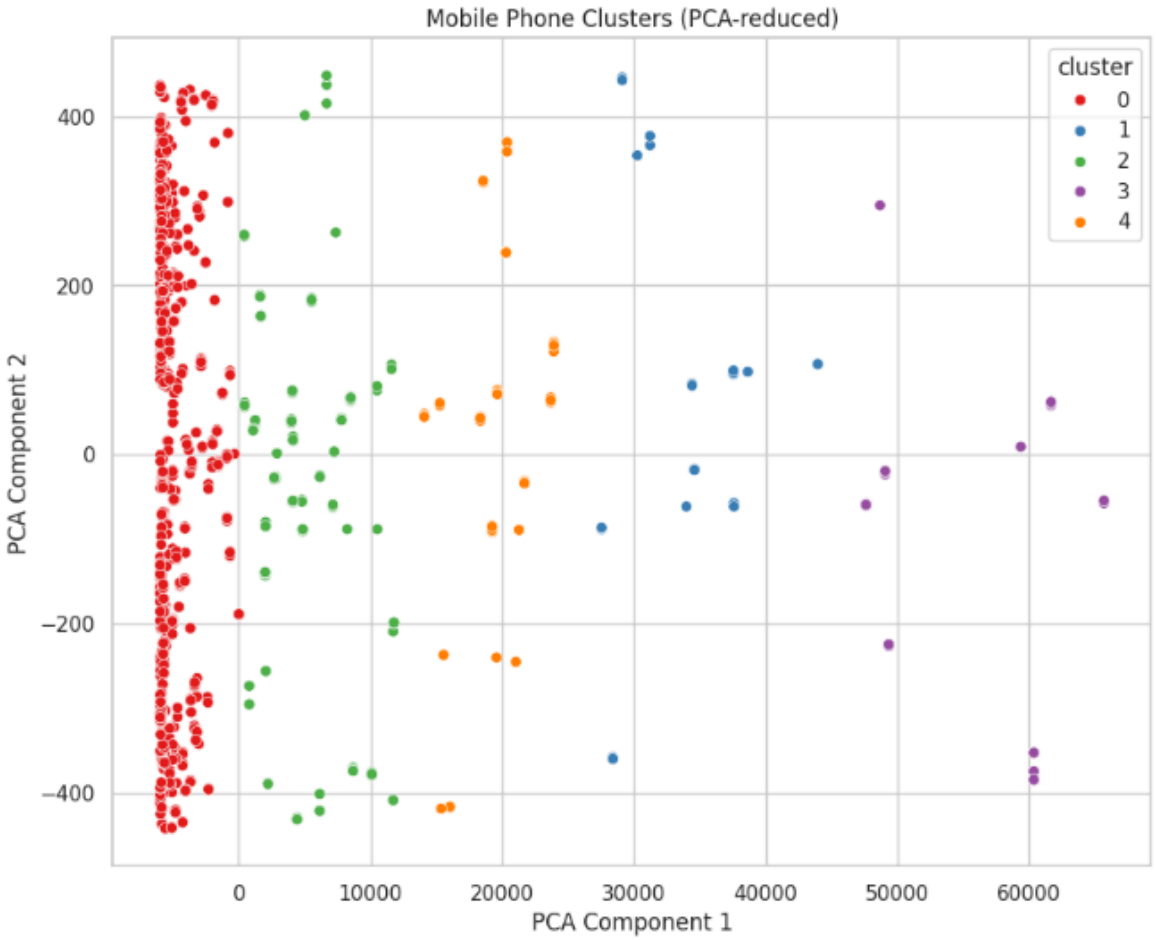


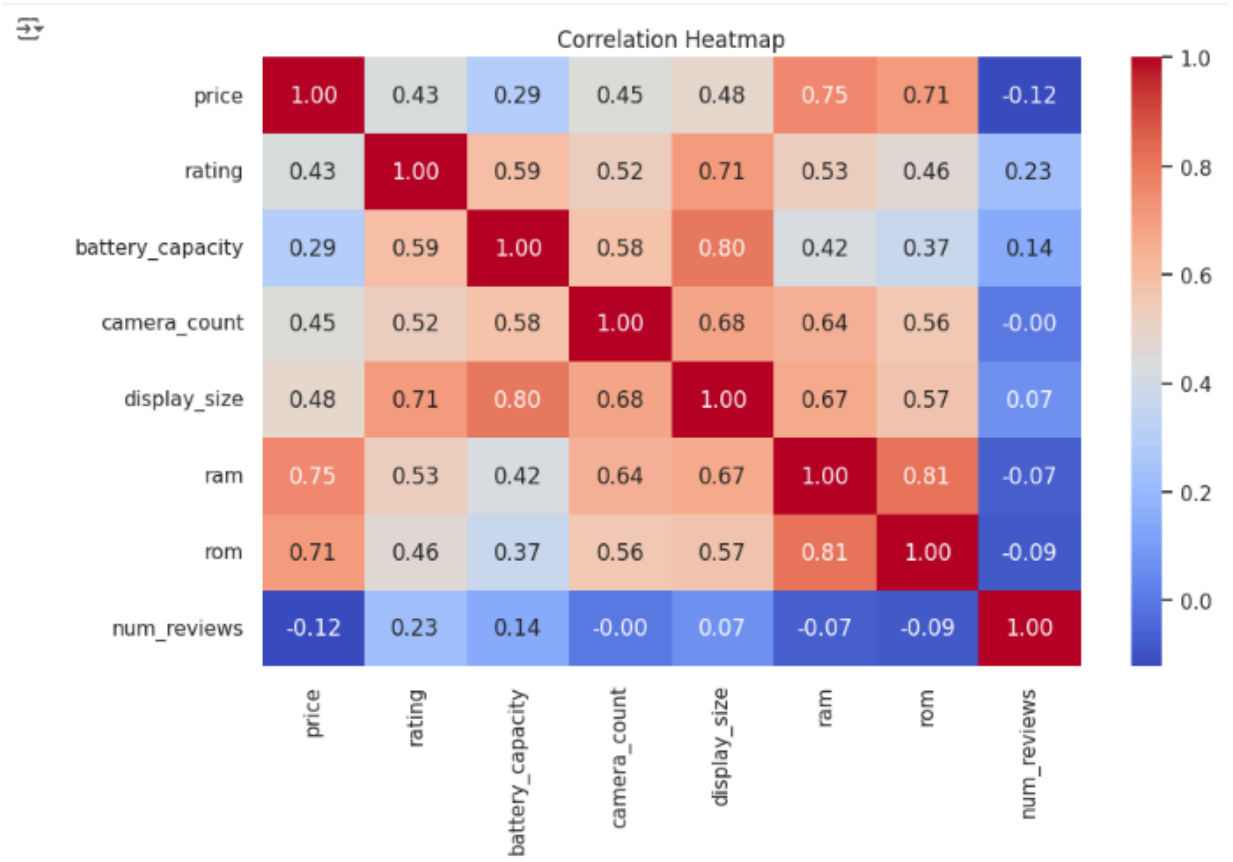
The histograms show the following:

- **Price Distribution: The prices of phones are somewhat right-skewed, indicating that more affordable phones are common, with fewer expensive options.**
- **Rating Distribution: The ratings are left-skewed, with many phones receiving high ratings (above 4.0).**



Connected to Python 3 Google Compute Engine backend





Correlation heatmap of all columns above

Blue sketch of project :-

1. Data Cleaning

During the data cleaning phase:

- **Rows with missing values in the `rating` column were removed.**
- **Categorical features (`battery_capacity`, `camera_count`, `display_size`, `ram`, `rom`) were encoded using Label Encoding to convert them into numerical values suitable for modeling.**
- **The dataset was split into `X` (features) and `y` (target: `rating`) for modeling purposes.**

---

2. Exploratory Data Analysis (EDA)

Univariate Analysis

- **Price Distribution**: Prices were right-skewed, with most mobile phones priced between 5,000 and 15,000 units.
- Rating Distribution: Ratings were generally high, clustering around 4.0 to 4.5.
- Reviews Distribution: The number of reviews was also right-skewed, with most phones having fewer than 50,000 reviews.

Bivariate Analysis

- **Price vs. Rating**: There was a slight positive correlation between price and rating, indicating that higher-priced phones tend to have better ratings.
- Price vs. Reviews: Phones with a moderate price range received a wide range of reviews, while a few high-priced phones received a large number of reviews.
- Price vs. Other Features: Higher battery capacity, camera count, and RAM/ROM configurations were associated with higher prices.

---

3. Predictive Modeling

A RandomForestRegressor model was trained to predict the `rating` based on other features.

- Feature Importance: RAM, ROM, and battery capacity were among the most influential features in predicting ratings.
- Model Performance:

- **Mean Squared Error (MSE): The model's MSE was evaluated to understand the average squared difference between the actual and predicted ratings.**
- **R-Squared (R^2): The model's R^2 score was calculated to determine how well the model explains the variance in ratings.**

Key Visualizations:

- **Feature Importance: The most important features for predicting ratings were RAM, ROM, and battery capacity.**
- **Actual vs. Predicted Ratings: The scatter plot indicated a reasonable correlation between actual and predicted ratings, with some variance.**

---

4. Clustering Analysis

Using KMeans clustering, the dataset was grouped into 5 clusters based on the features:

- **PCA for Dimensionality Reduction: Principal Component Analysis (PCA) was used to reduce the feature space to two dimensions for visualization.**
- **Cluster Visualization: A scatter plot of the clusters was generated, revealing distinct groupings of mobile phones based on their specifications.**

Conclusion and Insights

- **Price and Specifications: Higher specifications such as larger battery capacity, more cameras, and higher RAM/ROM were associated with higher prices.**
- **Rating Prediction: The RandomForest model provided a reasonable prediction of ratings based on mobile specifications, with RAM, ROM, and battery capacity being key predictors.**
- **Cluster Analysis: The clustering analysis grouped phones into distinct segments, likely reflecting different market tiers (e.g., budget, mid-range, flagship).**

Recommendations

- **For Retailers: Focus on stocking mobile phones with higher RAM and ROM configurations, as these significantly impact ratings and are associated with higher customer satisfaction.**
- **For Manufacturers: Consider the balance between price and key features (like battery and camera quality) as these have a strong influence on consumer ratings.**
- **Further Analysis: Explore more advanced modeling techniques or additional features (e.g., brand reputation, after-sales service) to enhance rating prediction accuracy.**

This concludes the analysis and modeling of the mobile phone dataset. The code used for this analysis is attached separately. If you have any further questions or need additional analysis, feel free to reach out! At email amitrathore110409@gmail.com