1. Introduction

This dataset contains information about various smartphone models from different brands, capturing key specifications and features such as price, brand, model, number of ratings, display size, RAM, battery capacity, and internal memory. The data provides insights into the diversity of smartphones, catering to a wide range of users based on performance, design, and affordability.

The objective of this project is to predict the price of a smartphone based on its specifications. By analyzing this dataset, we aim to understand how features like display size, RAM, battery capacity, and internal memory influence the pricing of smartphones. This prediction model can help manufacturers, retailers, and consumers make informed decisions regarding smartphone pricing and purchases.

The expected outcome is a machine learning model that accurately estimates the price of a smartphone based on its attributes, which can then be used for pricing strategies and market analysis.

2. Data Preparation

Data Loading and Exploration

The dataset was loaded using the pandas library with the read_csv() function.

Below is a summary of the initial dataset:

```
[32] data.head()
```

	Brand	Model	Price	Number of Ratings	Display Size	RAM	Battery	Internal Memory
0	Infinix	Zero 40 4G	Rs.70,000	23	6.78 inches	8GB	500mAh	256GB
1	Samsung	Galaxy Z Flip 6	Rs.385,000	39	6.7 inches	12GB	4000mAh	512GB
2	Samsung	Galaxy Z Fold 6	Rs.605,000	45	7.6 inches	12GB	4400mAh	512GB
3	Samsung	Galaxy A05	Rs.25,000	56	6.7 inches	4GB	5000mAh	64GB
4	Tecno	Phantom V Fold 2 5G	Rs.370,000	37	7.85 inches	12GB	5750mAh	512GB

```
[33] data.tail()
```

	Brand	Model	Price	Number of Ratings	Display Size	RAM	Battery	Internal Memory
1340	gfive	GFive Disco	Rs 3,199	59	2.4 InchesDisplay	32 MBRAM	3000 mAhBattery	32 MB
1341	gfive	GFive Spark	Rs 2,325	3 Ratings	1.8 inchesDisplay	32 MBRAM	3000 mAhBattery	32 MB
1342	e-tachi	E-Tachi E888	Rs 3,749	38	2.8 InchesDisplay	32 MBRAM	3000 mAhBattery	32 MB
1343	sparx	SparX Edge 20	Rs 5,000	24	6.67 inchesDisplay	8GB+8GB RAMRAM	5000 mAhBattery	256 GB
1344	gfive	GFive 4G Style	Rs 6,999	39	2.8 InchDisplay	2GBRAM	4000 mAhBattery	16gb

```
data.describe()
```

	Brand	Model	Price	Number of Ratings	Display Size	RAM	Battery	Internal Memory
count	1345	1345	1345	1345	1345	1345	1345	1345
unique	63	689	457	204	142	112	98	70
top	Samsung	Galaxy S24 Ultra	Rs.25,000	2	6.6 inches	8GB	5000mAh	128GB
freq	182	14	34	92	135	339	712	340

```
[34] data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1345 entries, 0 to 1344
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Brand               1345 non-null   object
 1   Model               1345 non-null   object
 2   Price               1345 non-null   object
 3   Number of Ratings   1345 non-null   object
 4   Display Size        1345 non-null   object
 5   RAM                 1345 non-null   object
 6   Battery             1345 non-null   object
 7   Internal Memory     1345 non-null   object
dtypes: object(8)
memory usage: 84.2+ KB
```

Data Cleaning

- **Handling Missing Data**

Some columns including price, RAM and battery had some missing values. The threshold was set to 5%. The count of all the rows with missing values was below the threshold, so they were removed from the dataset.

```
[15] data.isna().sum()
```

	0
Brand	0
Model	0
Price	16
Number of Ratings	0
Display Size	0
RAM	6
Battery	2
Internal Memory	0

dtype: int64

	0
Brand	0
Model	0
Price	0
Number of Ratings	0
Display Size	0
RAM	0
Battery	0
Internal Memory	0

dtype: int64

Duplicate Removal

The dataset was cleaned using the drop_duplicates() function. Duplicates were successfully removed.

```
print("Before Removing:", data.shape)
data.drop_duplicates(inplace=True)
print("After Removing:", data.shape)
```

```
Before Removing: (1321, 8)
After Removing: (1263, 8)
```

Data Type Conversion

- Price column was converted from strings (Rs.70,000) to numeric values using the clean_price() function. Non-numeric characters such as currency symbols and commas were removed, and invalid entries like 'Price Not Available' were replaced with NaN.
- Number of Ratings column contained both numeric and non-numeric values (3 Ratings). The clean_number_of_ratings() function was used to extract numeric parts, converting them to integers.
- Display Size column used extract_screen_size() function to convert screen size strings (6.78 inches) to float values representing the size in inches.
- The clean_ram() function extracted and standardized RAM values from a mix of GB/MB units. The conversion normalized RAM values, converting MB to GB where necessary.
- Battery capacities in both mAh and Wh were cleaned using the convert_battery_values() function. Wh values were converted to mAh for standardization.
- Internal Memory was cleaned using the convert_memory_values() function, converting values from MB to GB where necessary and handling multiple values by averaging them.

	Brand	Model	Price	Number of Ratings	Display Size	RAM	Battery	Internal Memory
0	Infinix	Zero 40 4G	Rs.70,000	23	6.78 inches	8GB	500mAh	256GB
1	Samsung	Galaxy Z Flip 6	Rs.385,000	39	6.7 inches	12GB	4000mAh	512GB
2	Samsung	Galaxy Z Fold 6	Rs.605,000	45	7.6 inches	12GB	4400mAh	512GB
3	Samsung	Galaxy A05	Rs.25,000	56	6.7 inches	4GB	5000mAh	64GB
4	Tecno	Phantom V Fold 2 5G	Rs.370,000	37	7.85 inches	12GB	5750mAh	512GB

Outlier Detection

Outliers were detected using the Interquartile Range (IQR) method. The columns analyzed for outliers included Number of Ratings, RAM, and Internal Memory. Boxplots were generated to visualize outliers before and after their removal. Rows containing outliers were removed to ensure a clean dataset for further analysis. The outliers in Price and Display Size were not removed as these columns have such information which should not be removed.

Data Transformation

Standardization was performed on the following columns (Price, Number of Ratings, Battery, RAM, Display Size, Internal Memory) using StandardScaler(). Scaling transformed the dataset into a form that is more suitable for machine learning models, normalizing the values to a standard scale (mean of 0 and variance of 1).

	Brand	Model	Price	Number of Ratings	Display Size	RAM	Battery	Internal Memory
1	Samsung	Galaxy Z Flip 6	2.033061	0.792020	0.376645	1.838843	-0.805408	2.562082
2	Samsung	Galaxy Z Fold 6	3.607899	1.094605	1.152898	1.838843	-0.337167	2.562082
3	Samsung	Galaxy A05	-0.543947	1.649345	0.376645	-0.623003	0.365194	-0.787905
4	Tecno	Phantom V Fold 2 5G	1.925686	0.691158	1.368524	1.838843	1.243146	2.562082
5	Tecno	Phantom V Flip 2 5G	0.708765	-0.166167	0.549145	0.607920	0.037426	0.647804

3. Data Analysis

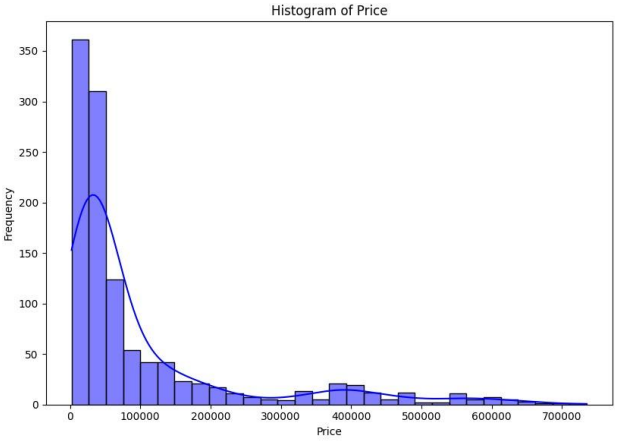
Univariate Analysis

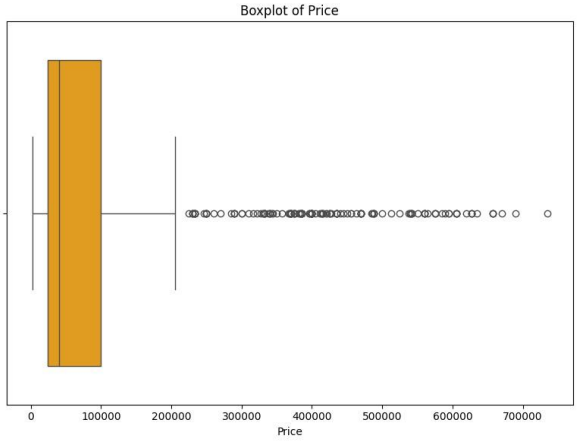
Key numerical variables analyzed include Price, Number of Ratings, Display Size, RAM, Battery, and Internal Memory. Below are the findings:

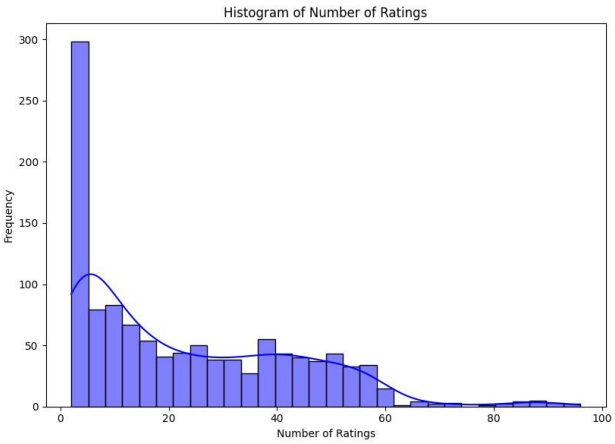
Histograms:
Most variables exhibit a skewed distribution.

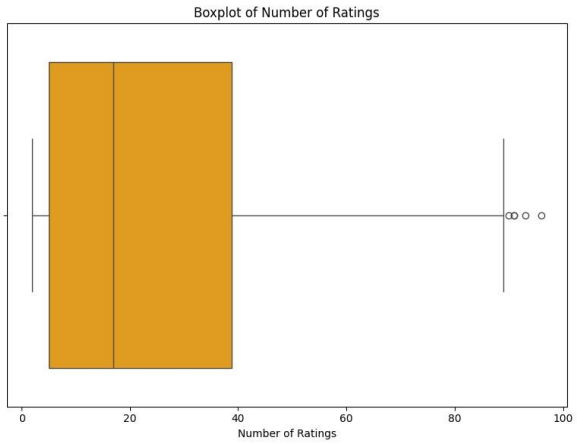
1. Price is heavily right-skewed, indicating a concentration of lower-priced smartphones, with fewer high-end devices.
2. Battery and RAM distributions are more uniform, showing a wide range of options across different brands.
3. Internal memory also indicate that its normal distributed.

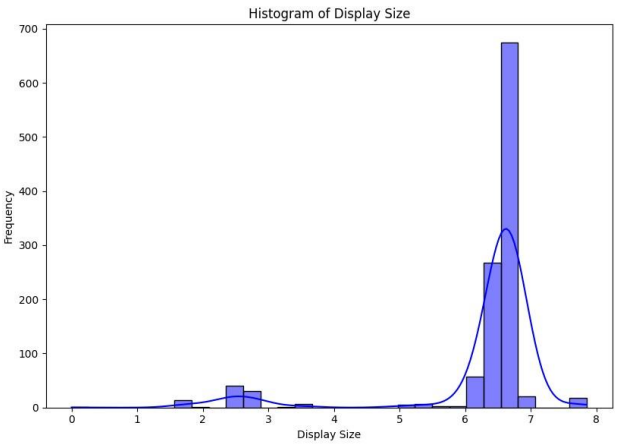
Box Plots:
Significant outliers were observed in the Price variable, likely representing premium or flagship devices. The Battery capacity has some outliers but is largely consistent across products, with typical capacities ranging between 4,000– 5,000mAh.

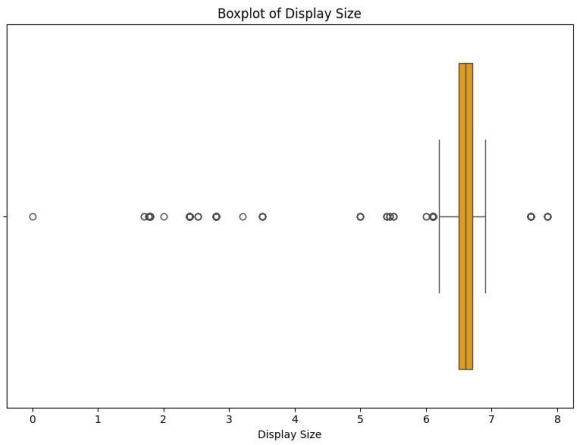

Histogram of Price
Frequency
Price
0
50
100
150
200
250
300
350
0
100000
200000
300000
400000
500000
600000
700000

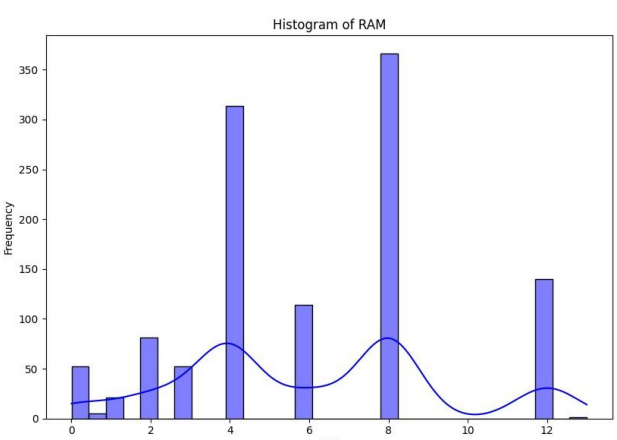

Boxplot of Price
Price
0
100000
200000
300000
400000
500000
600000
700000

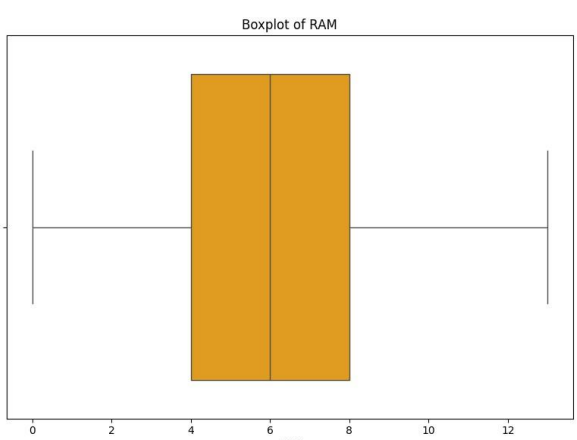

Histogram of Number of Ratings
Frequency
Number of Ratings
0
50
100
150
200
250
300
0
20
40
60
80
100

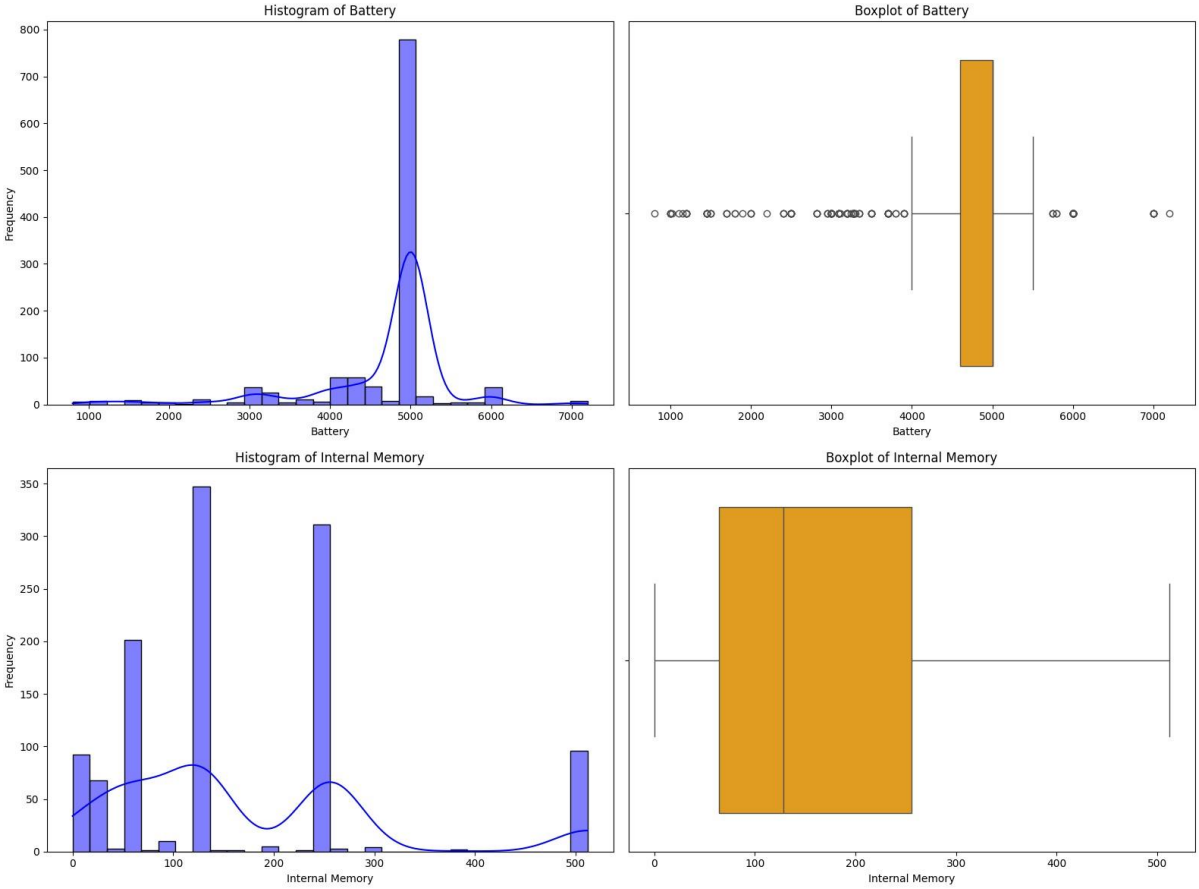

Boxplot of Number of Ratings
Number of Ratings
0
20
40
60
80
100


Histogram of Display Size
Frequency
Display Size
0
100
200
300
400
500
600
700
0
1
2
3
4
5
6
7
8


Boxplot of Display Size
Display Size
0
1
2
3
4
5
6
7
8

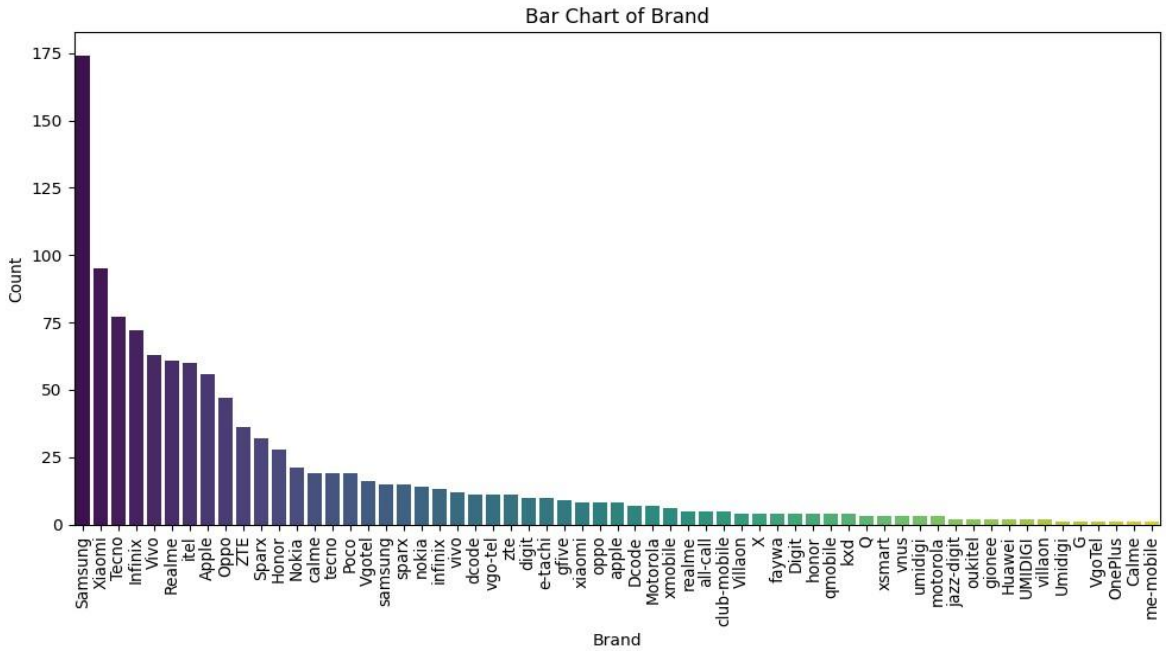

Histogram of RAM
Frequency
RAM
0
50
100
150
200
250
300
350
0
2
4
6
8
10
12


Boxplot of RAM
RAM
0
2
4
6
8
10
12



Categorical Variables

- The Brand column showed that Samsung(almost 15.2%) is the most dominant brand.
- Xiaomi(8.3%) and tecno(6.7%) are 2nd and 3rd respectively.
- The Model column has a wide variety of unique entries, indicating a diverse product lineup.



Bivariate Analysis

Scatter plots and correlation analysis were used to explore relationships between pairs of numerical variables:

Price vs Display Size:

A weak positive correlation was observed. As the price increase there should be minor change observe which in not very significant because most of the values are in the mid ranges.

Price vs Battery:

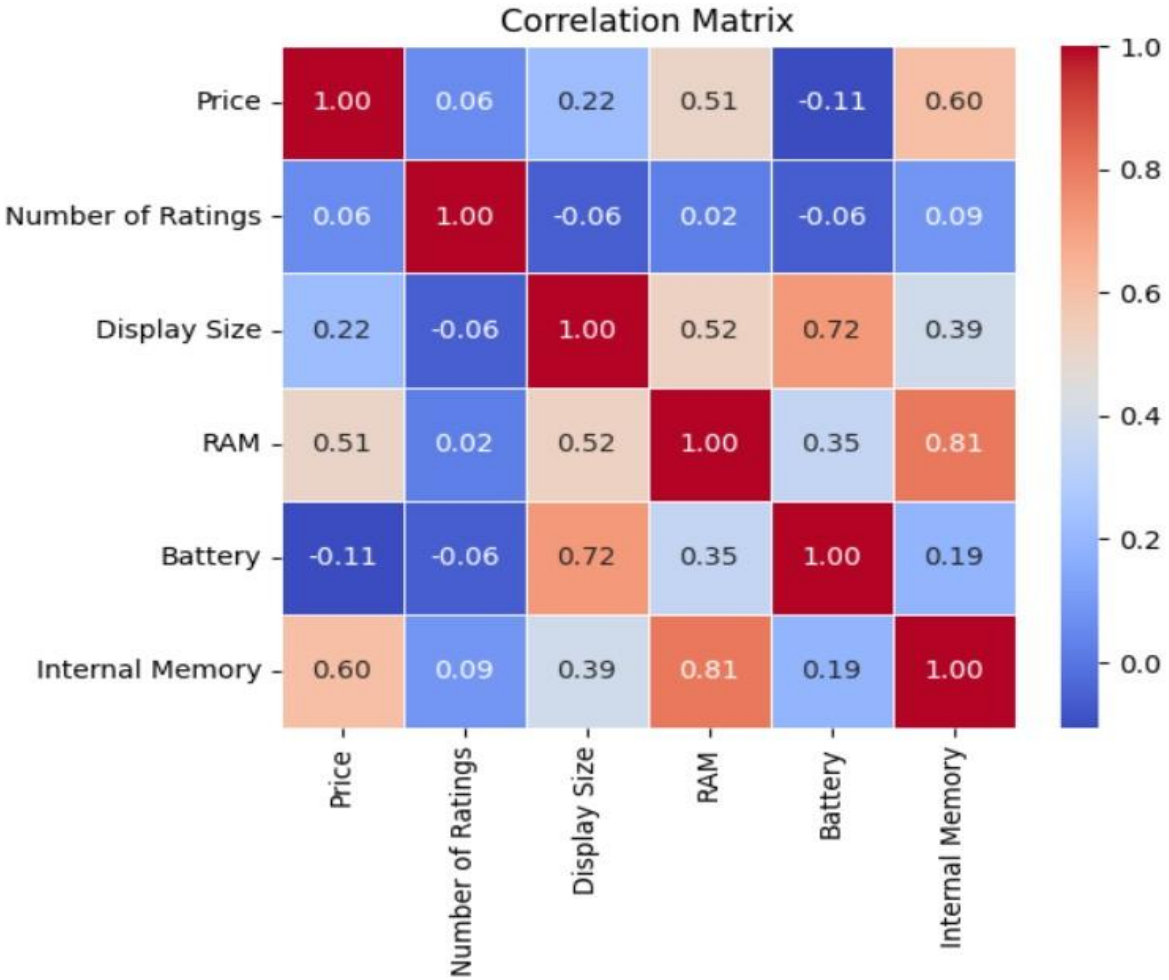
A weak correlation was observed, indicating that battery capacity alone does not significantly influence the price.

Display Size vs Battery:

A strong positive correlation was observed, as display size increase the battery size also increase.

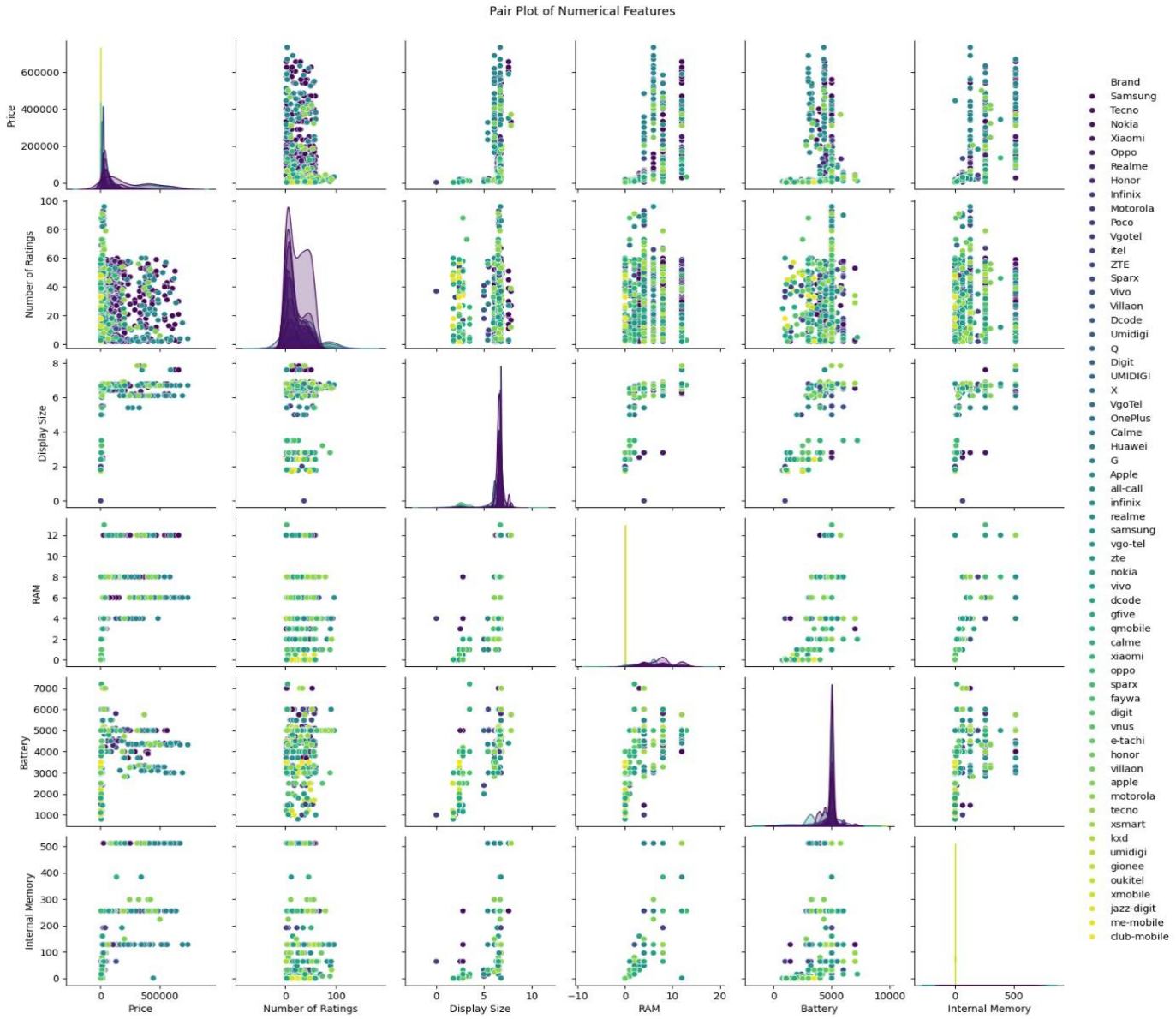
Correlation Matrix:

The result of heat map is given in below also key insights:



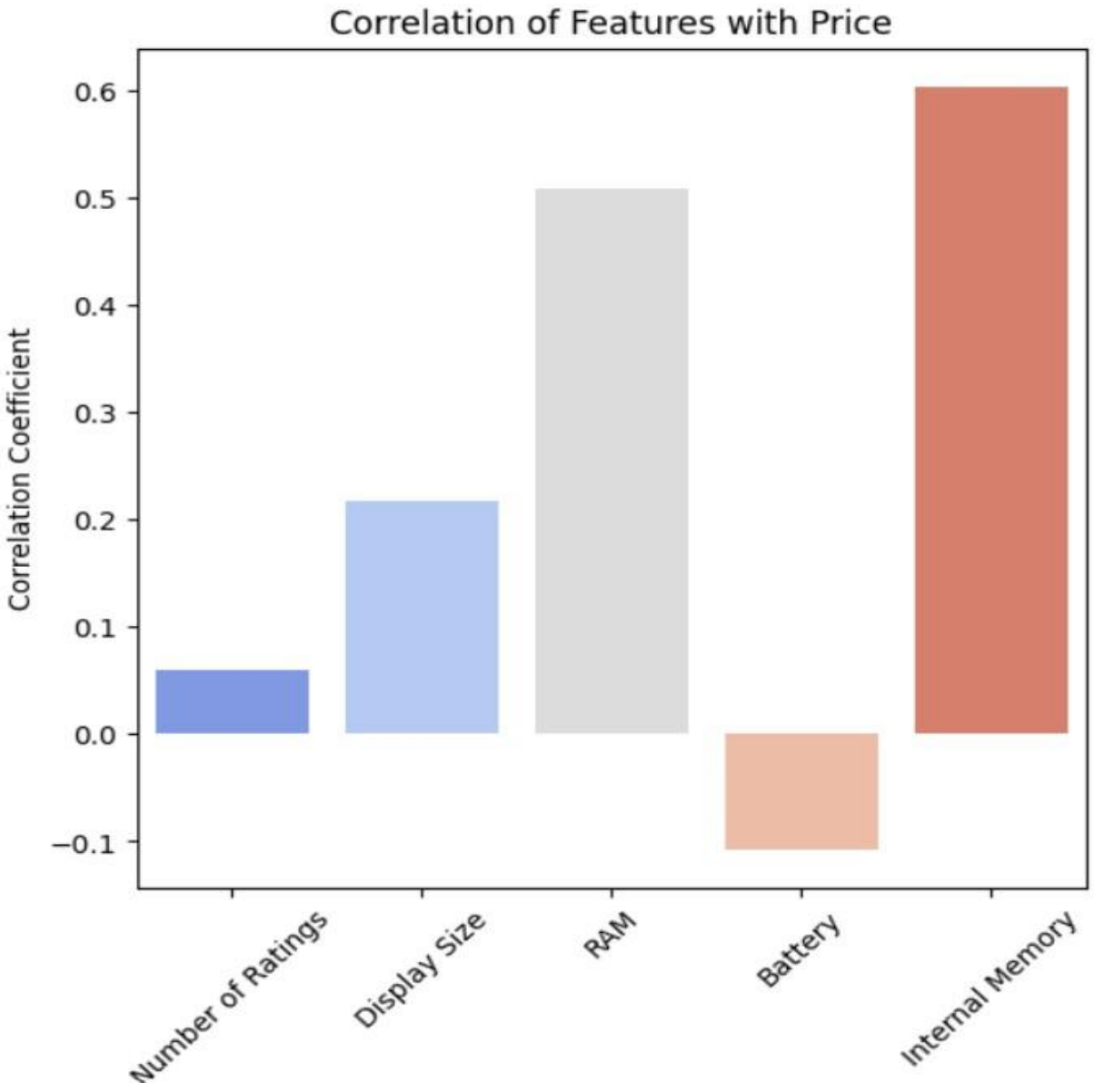
Multivariate Analysis

Using pair plots, relationships across multiple variables were visualized. Devices with higher RAM and Internal Memory cluster in the higher-price range. Clusters for low-end, mid-range, and high-end devices were clearly distinguishable. Battery and Display Size showed less discernible clustering, indicating weaker interaction effects with other features.



Feature Analysis

Correlation analysis with the target variable (Price) revealed the following:



4. Model Training

Feature Selection

In our analysis, we used the covariance matrix to evaluate feature dependencies. The covariance between the Rating and Price was observed to be the lowest, indicating minimal correlation. Therefore, we removed the Rating feature from our model. Additionally, variance was computed for all features, and Rating exhibited the lowest variance. This reinforced our decision to exclude it.

We identified a high correlation between RAM and Internal Memory. To evaluate multicollinearity, we computed the Variance Inflation Factor (VIF). Since both features had VIF values below 5, indicating moderate multicollinearity, we retained both in the model.

Lastly, when comparing the categorical features Brand and Model, we observed that Model had high deviations, increasing the risk of overfitting. Consequently, the Model feature was removed. Brand, however, showed significant value in predicting Price, so it was retained.

Model Selection

For this project, **Random Forest** and **Neural Networks** were chosen as the primary models for addressing the problem of price prediction.

Random Forest:

- A machine learning algorithm well-suited for regression tasks.
- Capable of handling non-linear trends in data due to its ensemble nature, combining multiple decision trees to enhance predictive accuracy and reduce overfitting.
- Chosen for its robustness and ability to handle complex relationships between features and the target variable.

Neural Networks:

- A deep learning model that excels in capturing intricate patterns and non-linear relationships in data.
- Particularly suitable for this regression problem where the underlying trends are non-linear, making traditional linear models insufficient.
- Provides flexibility in learning from large datasets, adjusting to hidden patterns that may not be evident in simpler models.
- The selection of these models reflects the non-linear nature of the data and the objective to achieve accurate price predictions while leveraging the strengths of both machine learning and deep learning approaches.

Model Training

The selected models are trained using a dataset that includes input features and corresponding labels. The training process involves:

- **Splitting the Data**: Dividing the dataset into training and testing sets, typically using an 80-20 or 70-30 ratio.
- **Parameter Optimization**: Fine-tuning hyperparameters such as learning rate, tree depth, or regularization strength to improve model performance.
- **Model Fitting**: Feeding the training data into the algorithm so it learns patterns and relationships to make accurate predictions.

Model Evaluation

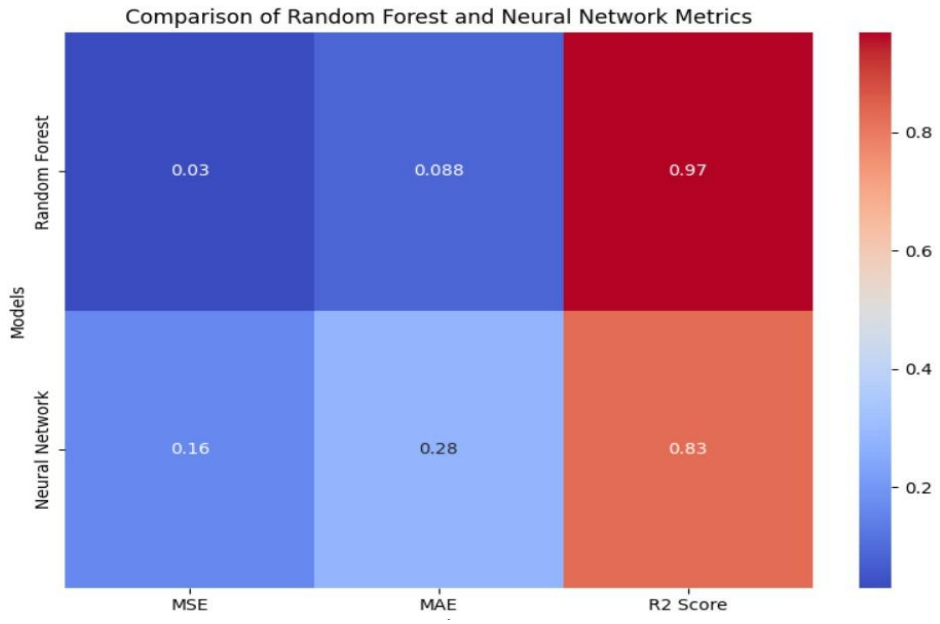
The evaluation of the selected models, Random Forest and Neural Networks, was conducted using the following performance metrics:

- **Mean Squared Error (MSE)**:
- Measures the average squared difference between the predicted and actual values.
- Lower MSE indicates better predictive accuracy.
- **Mean Absolute Error (MAE)**:
- Captures the average magnitude of errors in the predictions, regardless of direction.

- Provides a straightforward interpretation of the model's performance in terms of absolute error.
- **R^2 (Coefficient of Determination)**:
- Indicates the proportion of variance in the target variable explained by the model.
- Ranges from 0 to 1, with higher values signifying better model performance.
- Used to understand how well the model captures the variability in the data. These metrics were selected to ensure a comprehensive evaluation of the models, focusing on both error magnitude and the overall explanatory power of the models. The combination of MSE, MAE, and R^2 allows us to analyze the predictive accuracy and reliability of the models from multiple perspectives.

Comparison

- The Random Forest model has a significantly lower MSE (0.03) compared to the Neural Network model (0.16), indicating better prediction accuracy for the Random Forest model.
- Random Forest also outperforms the Neural Network in terms of MAE, with a value of 0.088 versus 0.28 for the Neural Network, reflecting lower average errors in predictions.
- While both models show strong explanatory power, the Random Forest model has a higher R^2 score of 0.97, compared to the Neural Network's 0.83, suggesting better overall fit for the Random Forest model.



5. Conclusion and Future Work

Conclusion

This project successfully developed a predictive model to estimate smartphone prices based on key specifications like display size, RAM, battery capacity, and internal memory. The analysis highlighted the following key findings:

- **Feature Importance:** Specifications like RAM and internal memory showed a significant correlation with price, whereas attributes like battery capacity and display size exhibited weaker relationships.

- **Model Performance:** Among the chosen models, Random Forest and Neural Networks, both demonstrated the ability to handle non-linear relationships effectively. Random Forest provided robust predictive performance, while Neural Networks captured complex patterns in the data.
- **Evaluation Metrics:** Metrics such as MSE, MAE, and R^2 confirmed the reliability and accuracy of the models, with Random Forest outperforming in terms of error magnitude and explanatory power.

The results underline the influence of smartphone specifications on pricing, offering valuable insights for manufacturers and retailers to optimize product offerings and pricing strategies.

Future Work

- **Feature Expansion:** Incorporating additional features, such as camera specifications, processor type, and connectivity options, may further improve the model's predictive accuracy.
- **Dynamic Pricing Analysis:** Extending the model to account for temporal trends in pricing, influenced by market dynamics, seasonality, or brand value.
- **Real-time Deployment:** Developing a real-time prediction system integrated into a web or mobile application to assist users in price comparison and decision-making.
- **Regional and Demographic Insights:** Analyzing regional variations in pricing trends and user preferences to tailor insights for different markets.