1. Introduction

The dataset is about smartphone prices of various smartphones across the 5 more popular brands in the mobile phone domain. It gathered data from different sources to compile a list of 1200+ smartphones to allow a more holistic analysis, recording important specifications such as the front and rear camera megapixels, display size and refresh rate, RAM, and storage among a few others.

This analysis aims to explore the different factors that influence smartphone prices. The focus is on determining price variation and predicting prices using machine learning based on smartphone specifications such as display size, RAM, internal storage, front and rear cameras, battery capacity, and brand, employing different regression models to choose the best-performing one from popular models such as linear, lasso, and ridge regression as well as decision tree regressor.

2. Data Description

The shape of the data showed that it contained 1256 rows and 11 columns. The dependent variable is price, which means this is a regression problem. Likewise, the remaining variables are the independent variables as all of them are essentially used in the analysis. The dataset is from Kaggle, which was gathered through web scraping different websites such as OLX.

3. Data Cleaning and Descriptive Analysis

The data cleaning process involved several key steps. Null values were checked, and unique values in categorical explanatory variables were identified to facilitate the creation of dummy variables. The “battery_capacity” column was split to extract numerical values, such as "5000" from "5000mAh," and only the new “battery_mAh” column was retained. One missing value in this column was filled with appropriate information sourced from GSMArena.

Similarly, the display column was split into “display_size” and “refresh_rate”, as some rows contained both size and refresh rate (e.g., 6.5" (90hz)), while others included only the size. Due to excessive null values in the “refresh_rate” column, it was dropped from the dataset. The processor column had approximately 13 missing values, which were filled using domain-specific knowledge based on information on the smartphone model taken from GSMArena.

The “rear_camera” column was split into six columns to represent individual camera megapixels, and a seventh column, “rear_cam_count”, was created to indicate the number of rear cameras in each phone. The “brand_name” column included five unique brands, for which dummy variables were created, leaving out one column ("APPLE") to avoid the dummy variable trap.

A new column, “price_log”, was added by taking the base-10 logarithm of the price values to mitigate the effect of outliers. Outliers in the “display_size” column were reduced to a certain range to align with common smartphone display sizes. Although there was one outlier each in the RAM and “internal_storage” columns, these were not modified, as they reflected high-spec smartphones compared to the competition.

	count	mean	std	min	25%	50%	75%	max
ram	1181.0	6.236664	2.892070	1.000000	4.000000	6.000000	8.00000	16.000000
internal_storage	1181.0	146.079594	106.097032	8.000000	64.000000	128.000000	256.00000	1000.000000
price	1181.0	84065.089754	63942.996724	399.000000	39999.000000	67999.000000	93999.00000	465999.000000
display_size	1181.0	6.340440	0.503870	4.500000	6.300000	6.500000	6.67000	7.500000
rear_cam_count	1181.0	2.803556	1.003564	1.000000	2.000000	3.000000	4.00000	6.000000
rear_cam_1	1181.0	44.111770	29.353359	5.000000	13.000000	48.000000	64.00000	200.000000
rear_cam_2	1181.0	9.001693	31.535587	0.000000	2.000000	8.000000	12.00000	1048.000000
rear_cam_3	1181.0	162.261643	386.424848	0.000000	0.000000	2.000000	12.00000	1125.000000
rear_cam_4	1181.0	7.290432	102.038250	0.000000	0.000000	0.000000	2.00000	1752.000000
rear_cam_5	1181.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
rear_cam_6	1181.0	0.004234	0.145494	0.000000	0.000000	0.000000	0.00000	5.000000
front_cam_MP	1181.0	15.533446	9.236989	1.000000	8.000000	16.000000	20.00000	48.000000
battery_mAh	1181.0	4346.216765	934.126246	1715.000000	4000.000000	4500.000000	5000.00000	7000.000000
price_log	1181.0	11.098507	0.698969	5.991465	10.596635	11.127263	11.45105	13.051941

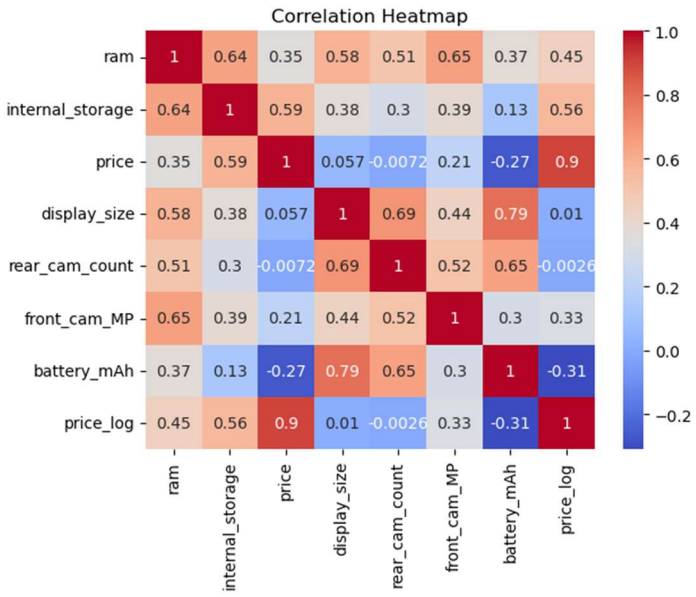
The first table provides a detailed summary of numerical columns in the dataset. The RAM variable has an average of 6.23 GB, with a maximum of 16 GB, reflecting a mix of mid-range and premium devices. Internal storage averages around 146 GB, with values ranging from 8 GB to 1,000 GB, accounting for both budget and flagship models. The price column shows significant variability, with a mean of PKR 84,065 and a maximum of PKR 465,999, indicating a wide range of device prices. Display sizes are concentrated around an average of 6.34 inches, which is as per expectations for modern smartphones. The rear camera count averages 2.8, suggesting most devices have dual or triple rear cameras. Smartphones have an average battery capacity of 4,346 mAh, which aligns with the fact that modern smartphones allow for longer use, essentially the new standard for smartphones. Lastly, log-transformed prices display a tighter range, confirming its use for reducing skewness.

The second table provides categorical insights. Five unique smartphone brands are included, with Samsung being the most frequent (387 devices). The model column features 1,173 unique entries, reflecting a diverse dataset. The rear camera column has 207 unique configurations, with 13 MP being the most common, appearing 58 times. The processor column includes 215 unique values, with Snapdragon 888 5G being the most frequent, present in 28 devices. Camera megapixels vary significantly across devices, as seen in the breakdown of rear_cam_1 to rear_cam_6, with the first two cameras being most prevalent. These tables highlight the richness of the dataset and its ability to capture both categorical and numerical diversity in smartphone specifications.

	brand_name	model	rear_camera	processor	rear_cam_1	rear_cam_2	rear_cam_3	rear_cam_4	rear_cam_6
count	1181	1181	1181	1181	1181	1181	1181	1181	1181
unique	5	1173	207	215	13	15	15	11	2
top	SAMSUNG	Xiaomi Black Shark 4 256GB	13 MP	Snapdragon 888 5G	48	8	0	0	0
freq	387	2	58	28	247	435	363	869	1180

4. Exploratory Data Analysis

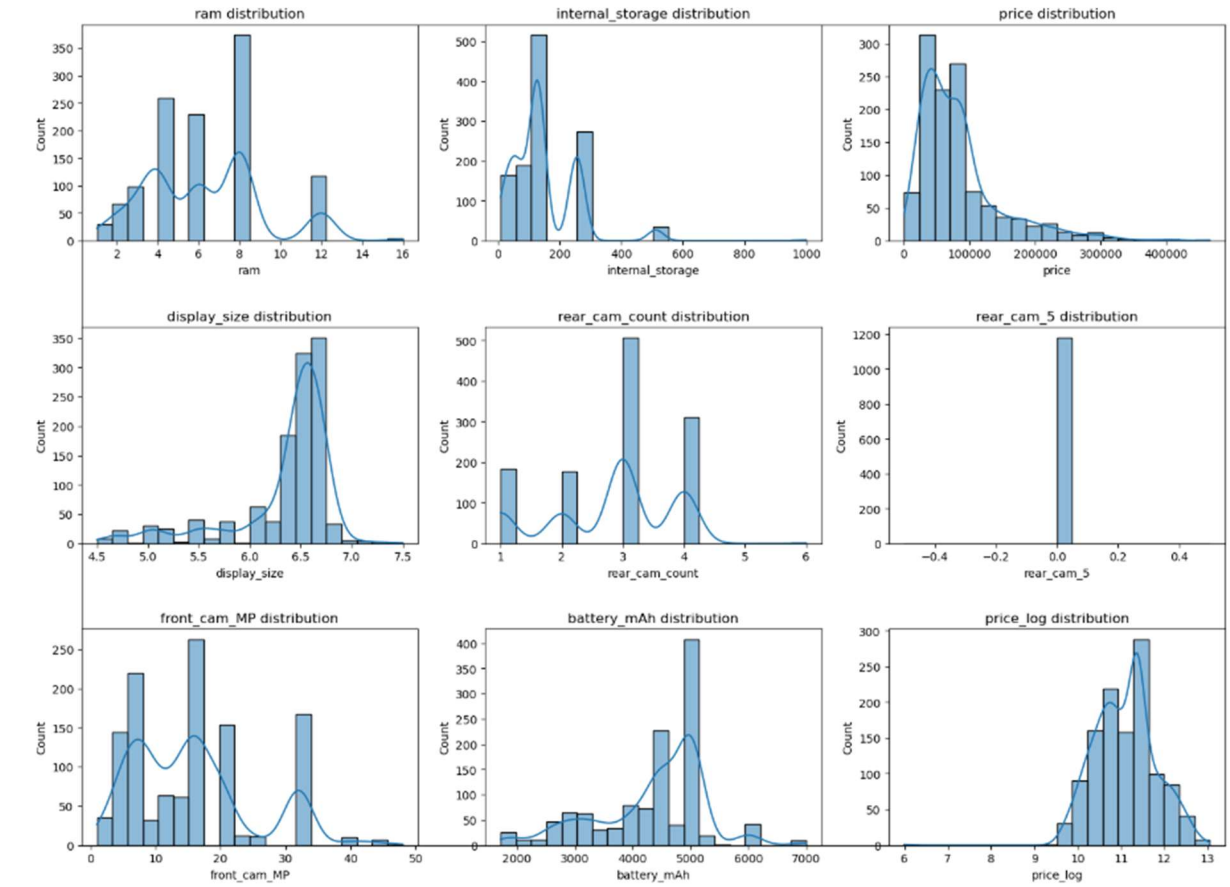
The exploratory data analysis includes visualizations to better understand the relationships between variables. The correlation matrix is useful in determining which specification has a stronger impact on smartphone prices. Count plots are used to examine the distribution of categorical variables such as “brand_name”.



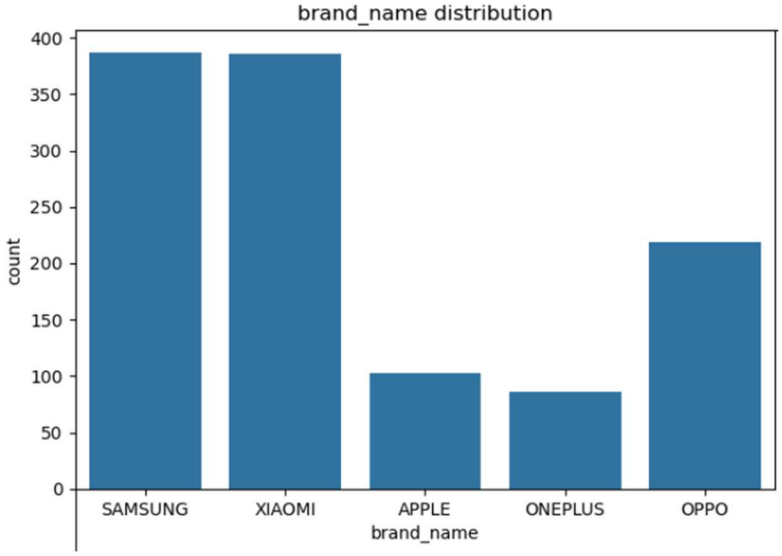
Similarly, KDE plots were used to analyze the distribution of numerical data such as price, RAM, and storage among a few others.

The graphs below provide valuable insights into the dataset. The RAM distribution indicates that most smartphones have RAM between 4 to 8 GB, reflecting mid-range and premium devices, while few exceed 12 GB. Internal storage is concentrated around 128 GB, a standard capacity, with fewer devices offering higher storage for premium models. The price distribution is heavily skewed to the right, with most devices priced below PKR 100,000, highlighting the need for log transformation, which resulted in the price_log column showing a more normal distribution. Display sizes mostly remain around 6 to 6.5 inches, while smaller and larger sizes are rare.

Rear camera counts are typically 2 to 3, with multi-camera setups becoming increasingly popular. Front cameras predominantly feature resolutions between 20 and 30 MP. Battery capacities peak around 4,000 to 5,000 mAh, which is standard for modern devices, while some outliers suggest some niche or premium designs. Overall, these distributions highlight common specifications and outliers within the smartphone market.



Moreover, the count plots for categorical variables like brand name and rear camera megapixels for the 6 cameras show the distribution of smartphones among brands and the number of cameras on those smartphones respectively. Below is the graph for brand names which shows that the data contained almost the same amount of Samsung and Xiaomi phones followed by Oppo in terms of the number of phones in the dataset. This helps in concluding that these 3 companies have put out a large lineup of phones overtime, indicated by the high number of models for each.



5. Modeling

The analysis employed several machine learning models to predict smartphone prices, including Linear Regression, Lasso Regression, and Ridge Regression along with a Decision Tree Regression. Each model was trained using the cleaned dataset, and evaluation metrics such as Root Mean Square Error (RMSE), R-squared (R^2), and the Mean Absolute Error (MAE) were recorded for all 4 models were calculated to compare their performance. The summary for these models is provided in the table below.

Model	R-squared (R^2)	MSE	RMSE	MAE
Linear regression	0.646	0.157	0.3960	0.3159
Lasso regression	0.526	0.210	0.4580	0.3619
Ridge regression	0.638	0.160	0.4002	0.3215
Decision tree	0.7722	0.1008	0.3176	0.1998

Linear Regression demonstrates a moderate R-squared (0.646), indicating that around 64.6% of the variance in the dependent variable is explained by the model. It has a Mean Squared Error (MSE) of 0.157 and a Root Mean Squared Error (RMSE) of 0.3960, showing moderate prediction accuracy, with a Mean Absolute Error (MAE) of 0.3159. **Lasso Regression** performs slightly worse, with a lower R-squared (0.526) and higher errors (MSE 0.210, RMSE 0.4580, and MAE 0.3619), likely due to its regularization effects penalizing coefficients. **Ridge Regression** achieves performance close to Linear Regression, with an R-squared of 0.638 and slightly higher RMSE (0.4002), but slightly better MAE (0.3215), making it comparable but not superior.

The **Decision Tree** model outperforms the others with the highest R-squared (0.7722), and the lowest MSE (0.1008), RMSE (0.3176), and MAE (0.1998). This indicates that the Decision Tree model captures non-linear relationships effectively and provides the most accurate predictions among the models evaluated.

6. Conclusion

This analysis explored the key factors influencing smartphone prices using a comprehensive dataset of smartphone specifications. After data cleaning and transformation, including handling missing values, creating dummy variables, and addressing outliers, the dataset was prepared for analysis.

The Exploratory Data Analysis revealed significant relationships between price and features like RAM, internal storage, and display size, while visualizations provided insights into the distributions and trends of key variables.

In the modeling phase, four regression models were evaluated: Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree. The Decision Tree model emerged as the best performer, achieving the highest R-squared value (0.7722) and the lowest error metrics (MSE, RMSE, and MAE), indicating its ability to effectively capture complex, non-linear relationships between smartphone features and price. In contrast, while Linear Regression and Ridge Regression performed moderately well, Lasso Regression's penalization of coefficients led to reduced accuracy.