Mobiles Discount Data Analysis & Estimation

Team Members :
Srujana (Tr)
Venkatesh
Vamsi

Project Guide :
Mr. Mahendra

Business study & Factors !

Business Study:

- Predict the Discount_Price for mobiles based on product features. Help business teams optimize pricing and discounts by understanding which features influence discounting.
- Factors affecting Discount :
    - Brand & Platform
    - MRP & Selling Price
    - Technical specifications (RAM, ROM, Processor, Battery, Cameras, Display Size)
    - Customer engagement (Ratings, Review_Count, Rating_Count)

Data Collection

- **Source**: Data taken from Amazon and Flipkart websites.
- **Format:** CSV file with Amazon 456 rows and 8 columns, Flipkart 689 rows & 7 columns columns
- **Details Collected:** Product Name, MRP, Selling Price, Discount %, Brand, Rating, Review Count, RAM, ROM, Display Size, Camera, Processor, Battery.
- **Method:** Collected using Python web scraping (BeautifulSoup).
- **Time:** Collected in August 2025.
- **Purpose:** To predict mobile phone discount Price based on product features.

Brand	Brand_Model	Color	Platform	MRP	Selling_Price	Discount_Price	Discount	RAM	ROM	Display_Size	Battery	Front_Cam(MP)	Back_Cam(MP)	Proce
POCO	POCO C75 5G	Enchanted Green	Flipkart	10999	7699	3300	30	4	64	6.8800	5160	5	50	4s Gen Proce
POCO	POCO M6 Plus	Graphite Black	Flipkart	15999	10080	5919	36	6	128	6.7900	5030	13	50	Snapdra 4 Gen. Proce
CMF	CMF by Nothing	Black	Flipkart	22999	18999	4000	17	8	128	6.7700	5000	16	108	Dimer 7300 Proce
Motorola	Motorola G85 5G	Viva Magenta	Flipkart	20999	15999	5000	23	8	128	6.6700	5000	32	108	6s G Proce
OPPO	OPPO K13 5G	Prism Black	Flipkart	22999	17999	5000	21	8	128	6.6700	7000	16	50	Snapdra 6 G Proce
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Samsung	Samsung Galaxy M36	Orange Haze	Amazon	22999	17499	5500	24	6	128	6.6000	5000	13	108	Medi
Redmi	Redmi Note 14	Ivy Green	Amazon	22999	17999	5000	22	8	128	6.8000	5000	16	50	Snapdra 4 Gen

Data Validation

- Checking Column Wise data (Unique, nunique, dtype)
- Changing dtypes as per the column details
- Verified duplicates and there are no duplicates
- Verified present positive or negative discounts in discount price & Rating ranges also
- Verified missing values
- Dataset consistency checked

EDA
Uni-Variate

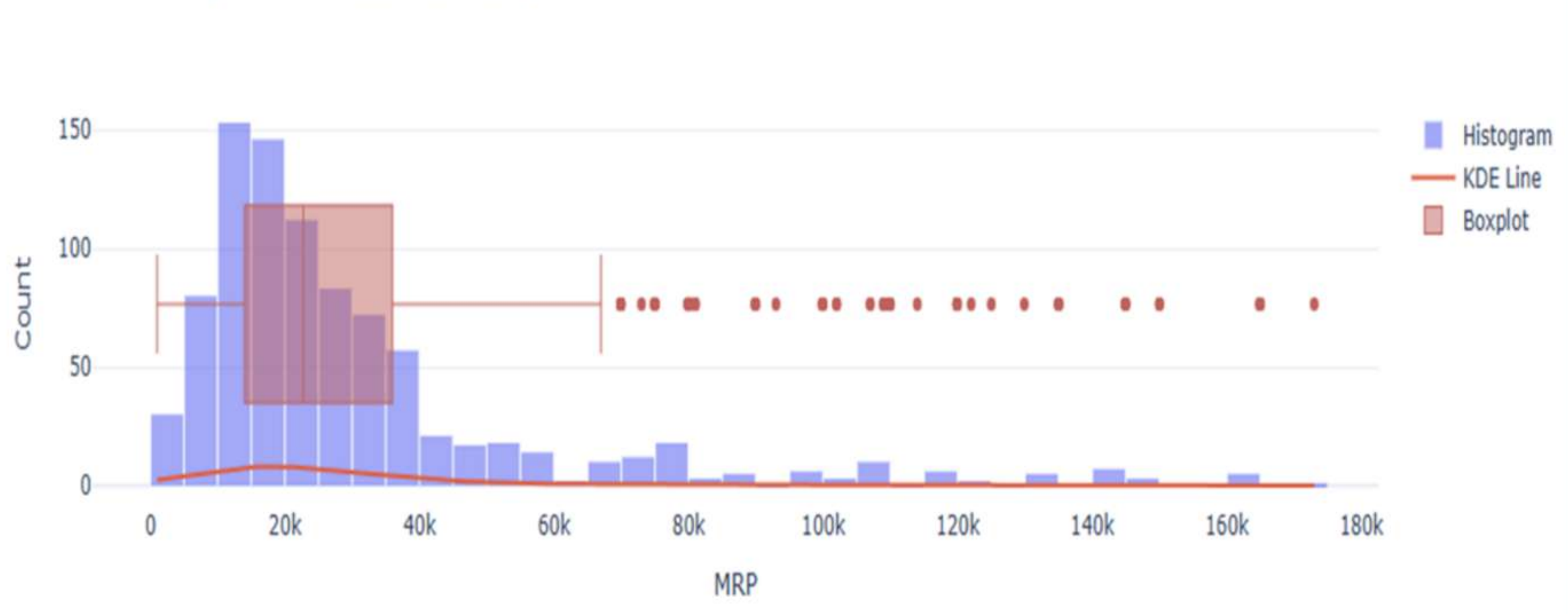
Univariate Statistics:

Feature	Count	Mean	Std Dev	Min	Q1	Median	Q3	Max
MRP	904	31660.16	29276.22	999.00	13999.00	22749.00	35999.00	172999.00
Selling_Price	904	24561.91	24483.21	267.00	10072.25	16998.00	27999.00	154900.00
Discount_Price	904	7098.24	7698.98	250.00	3500.00	5000.00	7925.75	72000.00
Discount	904	24.61	10.83	1.00	17.00	23.00	31.00	73.00
RAM	904	7.36	2.99	2.00	6.00	8.00	8.00	16.00
ROM	904	166.16	102.21	32.00	128.00	128.00	256.00	512.00
Display_Size	904	6.57	0.76	1.50	6.60	6.70	6.77	7.80
Battery	904	5107.03	837.40	800.00	5000.00	5000.00	5160.00	7550.00
Front_Cam(MP)	904	14.77	8.20	2.00	8.00	13.00	16.00	50.00
Back_Cam(MP)	904	57.05	30.32	0.00	50.00	50.00	50.00	200.00
Ratings	904	4.30	0.23	3.00	4.20	4.30	4.40	4.80
Rating_Count	904	15476.64	40685.76	21.00	2066.00	4090.50	7325.00	313551.00
Review_Count	904	1099.39	2011.32	0.00	262.00	560.00	970.00	16878.00

Uni-Variate EDA

Histograms/boxplots used to see spread ,skewness and outliers.

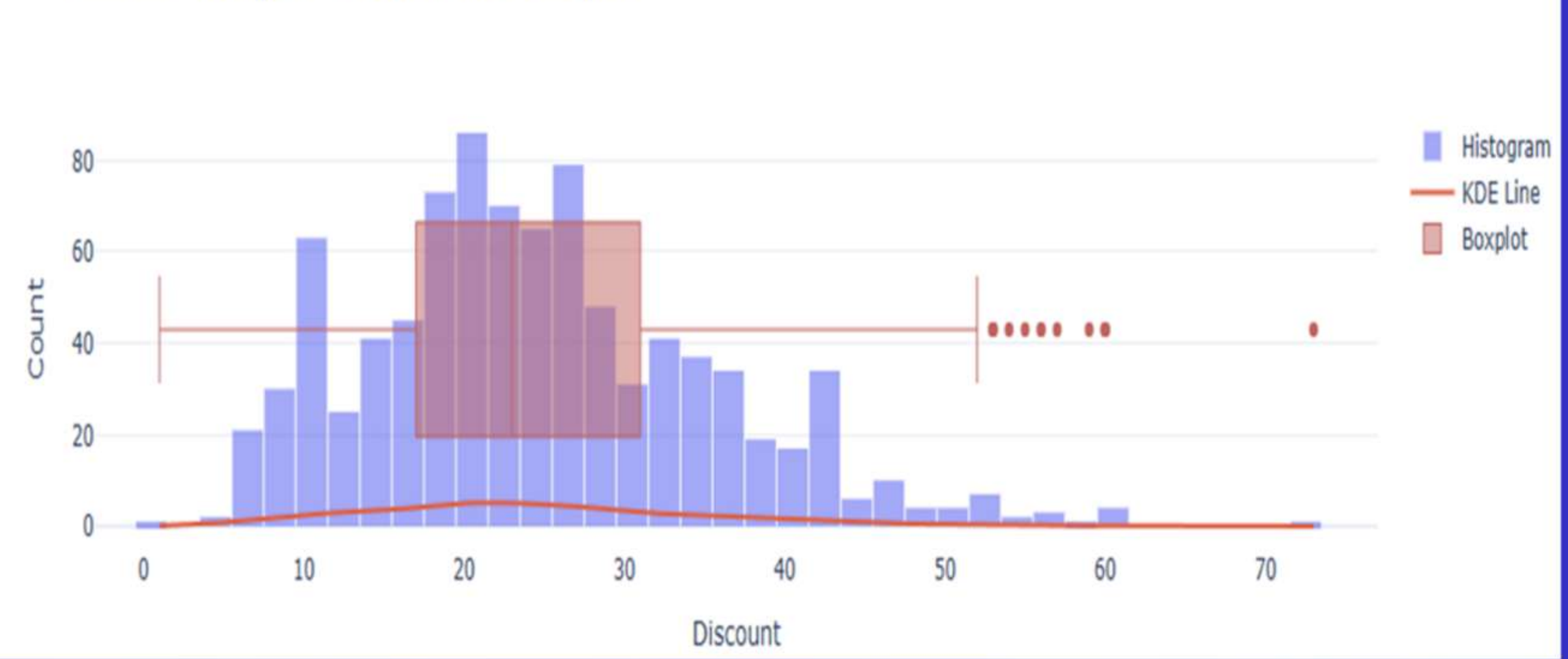
MRP - Histogram with KDE Line & Boxplot



Selling_Price - Histogram with KDE Line & Boxplot

Discount_Price - Histogram with KDE Line & Boxplot

Discount - Histogram with KDE Line & Boxplot



Bi-Variate(Stats & Visual)

Scatter plots help you see relationships between target and independent variables.

Na & Out Handling

- There are no present null values.
- Using isolationforest found outliers and removed those 20 rows

----- Isolation Forest Outliers -----

	Brand_Model	ROM	RAM	Battery	MRP	Display_Size
9	samsung guru music	64	6	800	2349.0000	1.8000
10	samsung guru 1200	128	8	800	1699.0000	1.5200
34	samsung guru music	128	8	800	1999.0000	1.8000
42	samsung sm 310e	64	4	800	1999.0000	1.9000
63	samsung guru1200	64	6	800	1799.0000	1.8000
132	samsung guru 1200	64	4	800	1699.0000	1.5200
142	samsung guru music	64	4	1100	1999.0000	2.0000
204	motorola a200	64	4	800	1549.0000	1.7700
211	lava a3 torch	128	6	1750	1649.0000	1.8000
213	samsung 1200	64	4	800	1699.0000	1.5000
235	samsung guru music	64	6	800	1999.0000	1.8000
285	samsung guru music	64	6	800	1999.0000	1.8000
288	motorola a50v dual	64	4	1750	1849.0000	1.8000
300	samsung b310ed	128	6	850	1599.0000	2.4000
338	karbonn kx29 ds	32	6	2700	1790.0000	2.4000
348	motorola a10v ds	64	4	800	1499.0000	1.8000
353	samsung sm-b310ezddins	64	4	800	1999.0000	1.8000
383	samsung sm 1207	64	6	800	1799.0000	1.5000
712	jiobharat v4 4g	512	12	5000	1999.0000	2.4000

Predictive modeling

X @ Y SELECTION

- Selected Discount_price target variable.
- This is the value to predict (dependent variable).
- Dropped columns(discount_price,discount,MRP,selling_price)that would cause data leakage or are the same as the target.

Column	Use in X	Reason
Brand	Yes	Brand affects pricing. Encode as category.
Brand_Model	Maybe	Too granular, almost unique per row — can overfit.
Color	Yes	Color sometimes affects price, but small effect.
Platform	Yes	Flipkart/Amazon might affect price.
MRP	Careful	Strong correlation with selling price — could leak if predicting price, but fine if predicting
Selling_Price	No	Usually target or leaks into discount calculation.
Discount_Price	No	This is basically target if predicting discounts.
Discount	No	Calculated from Selling_Price & MRP → direct leakage.
RAM	Yes	Important spec.
ROM	Yes	Important spec.
Display_Size	Yes	Can influence price.
Battery	Yes	Can influence price.
Front_Cam(MP)	Yes	Useful feature.
Back_Cam(MP)	Yes	Useful feature.
Processor	Yes	Important categorical feature.
Ratings	Maybe	Could leak if target is popularity-related.
Rating_Count	Maybe	Highly correlated with popularity and pricing.
Review_Count	Maybe	Same as above.

Train & Test split

- Split the into train 80% ,test 20%.
- Split the data before encoding to prevent leakage .
- After done the encoding.
- One-hot-encoding (model,platform).
- Target encoding (brand_model,processor,colour).
- Numerical columns done scaling.

Model Selection & Training of Different Models

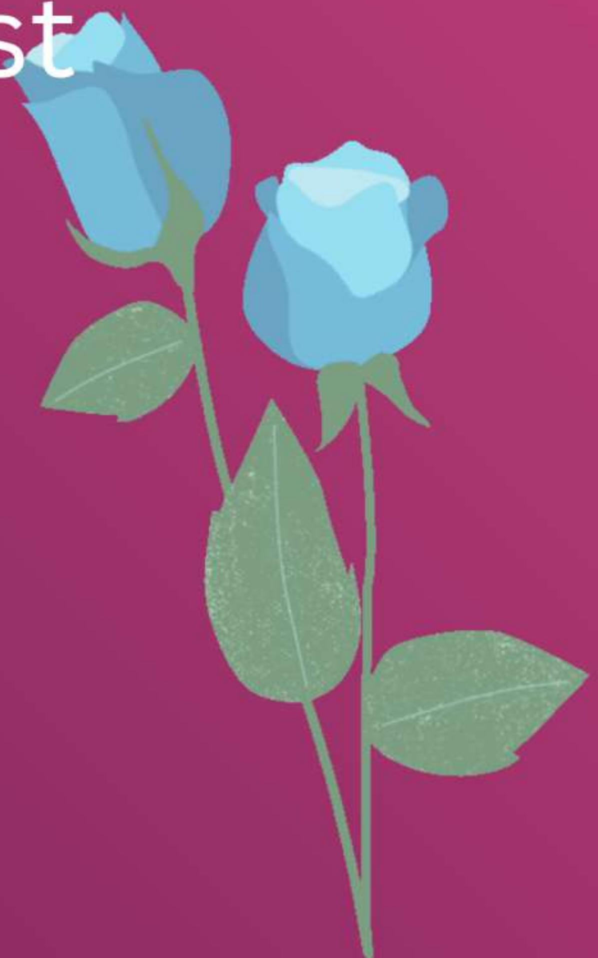
- Trained and compared multiple regression models:
  - o Linear, Ridge, Lasso, ElasticNet
  - o Decision Tree, Random Forest
  - o Gradient Boosting, XGBoost, LightGBM
  - o Support Vector Regression, KNN
- Process followed:
  - o Train-Test Split (80–20)
  - o Fit model → Predict → Evaluate
- Evaluation metrics: R^2 Score, RMSE, MSE
- Best model selected based on highest R^2 & lowest error

Model Performance Comparison:

	Model	Train R^2	Test R^2	Train MSE	Test MSE	Train RMSE	Test RMSE	Fit Status
0	Ridge (L2)	0.737	0.593	1.249399e+07	2.242996e+07	3534.684	4736.028	Good fit
1	ElasticNet	0.740	0.584	1.234891e+07	2.294787e+07	3514.102	4790.394	Overfit
2	Linear Regression	0.740	0.581	1.234409e+07	2.307842e+07	3513.416	4804.001	Overfit
3	Lasso (L1)	0.740	0.580	1.234409e+07	2.312745e+07	3513.416	4809.101	Overfit
4	XGBoost	0.993	0.577	3.225588e+05	2.329058e+07	567.943	4826.032	Overfit
5	Gradient Boosting	0.999	0.531	6.542587e+04	2.582835e+07	255.785	5082.160	Overfit
6	Random Forest	0.617	0.516	1.819623e+07	2.665885e+07	4265.704	5163.220	Good fit
7	SVR	-0.006	0.004	4.781702e+07	5.488536e+07	6914.986	7408.466	Good fit

Model Evaluation

- Used R^2, MSE, RMSE to evaluate performance
- R^2 Score: Explains variance in target
- MSE / RMSE: Show prediction error
- Compared models on test set → selected best performing model
- Example:
- High R^2 (≈ 1) + Low RMSE = Better model



Best Model From Evaluation & hyper parameter Tuning

- Finalize the best model based low mse and high r2
- Hyper parameter tuning through the gridsearchcv and selected randomforest regression model

Saving Best Model & Real Time Prediction

- Saved the Random Forest
- https://deployment-2eneyeckebhp4qquofuomp.streamlit.app/

Thank You