Flipkart Mobile Sales Analysis

Table of Contents

Abstract

Mobile phones' lightning-fast revolution has changed every aspect of our lives, including the way we communicate with one another, conduct business, and connect with the outside world. Mobile phones have developed from basic communication tools to cutting-edge smartphones with a wide range of features. Online retail stores have emerged as major industry contributors thanks to consumers' increasing demand for mobile phones and their ability to choose from a wide variety of products.

One of India's biggest and most well-known online shopping destinations, Flipkart, has established itself as the country's preferred location for buying mobile phones. With a wide range of mobile phone models and thorough product information, Flipkart offers tremendous insight into consumer preferences and market trends.

With a diverse product selection, Flipkart has developed into one of the most well-known and significant e-commerce platforms. It has provided its customers with the best deals, prices, and payment options in addition to an innovative user-friendly interface. Regressions, data visualisations, and other techniques are used in this project to emphasise the significance of e-commerce data analysis.

Mobile phone sales play a significant role in overall sales as a result of their rising popularity and usage. The main goal is to examine critical factors that could have an impact on mobile phone ratings and analyse those factors. By delving into customer preferences, hoping to gain a more thorough understanding of their expectations and inclinations.

Used Kaggle dataset of Flipkart sales which included 3114 entries with parameters listed such as brand, model, color, memory, rating, selling price, original price, and so on.

There are certain steps that are followed to perform the analysis on the dataset. They are:-
Data preprocessing
Exploratory data analysis
Correlation analysis
Model Building
Evaluation

Project Motivation

One of the most powerful e-commerce sites in India, Flipkart controls a sizable portion of the online mobile phone sales market. Its enormous potential customer base and wide range of products contribute to its hegemonic position in the mobile phone sales category. In addition to well-known brands like Apple, Samsung, Xiaomi, OnePlus, Realme, and others, Flipkart promotes a sizable selection of mobile phone models from a variety of manufacturers.

A wide range of consumer tastes and budgets will find something appealing in this extensive selection of goods. Users who purchase items from Flipkart, including mobile phones, can leave reviews and ratings for those items. These user-generated reviews play a significant role in influencing the purchasing Decisions of other potential customers and in building customer trust.

The main driving force behind this project is to identify consumer preferences and propensities for mobile phones. We plan to use the "Flipkart Mobiles Dataset," which contains a wealth of information on a variety of handset models easily accessible through Flipkart, to dig deeper into each of the factors that affect customer ratings and how each of these factors may reveal recurring patterns and trends.

The study aims to arm Flipkart with more in-depth knowledge about the market, causing the platform to keep up with anticipated market developments and anticipate variability in customer demand. Flipkart could increase its competitiveness in the smartphone market by investigating emerging characteristics, well-known brands, and price inclinations. We are seeking to hunt down conclusions to glitches like:
What are the most pertinent factors influencing customer ratings on Flipkart for mobile phones?
How big of an impact does price have on the ratings as well as the recognition among different mobile phone models?

Business Requirements

The major goal of this project is to perform the analysis on the dataset that could bring out the hidden factors that drive the sales of the mobile phones. These insights could be beneficial in many ways.
The analysis desires to delegate Flipkart and other e-commerce companies insightful information so that they ought to enhance their offerings, item Decision-making, advertising strategies, and branding efforts by deeper comprehending mobile phone ratings as well as recognizing the vital factors driving these ratings and sales. They can enhance customer satisfaction and retention and boost sales and income by emphasizing on the components that matter most to potential clients.

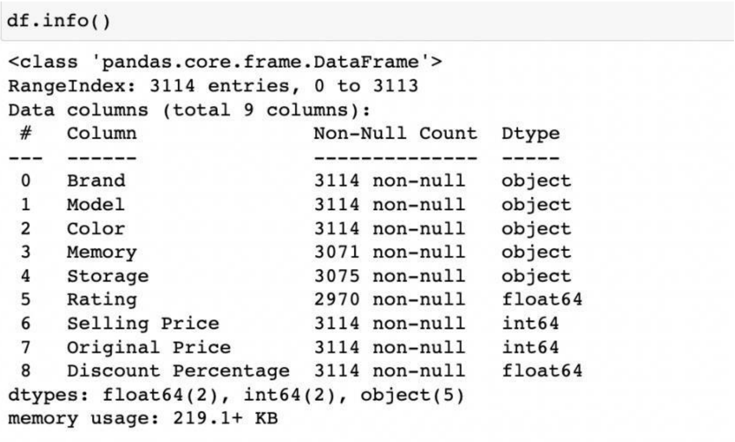
Improvements:

According to feedback review given during the presentation, we have done the modifications listed below

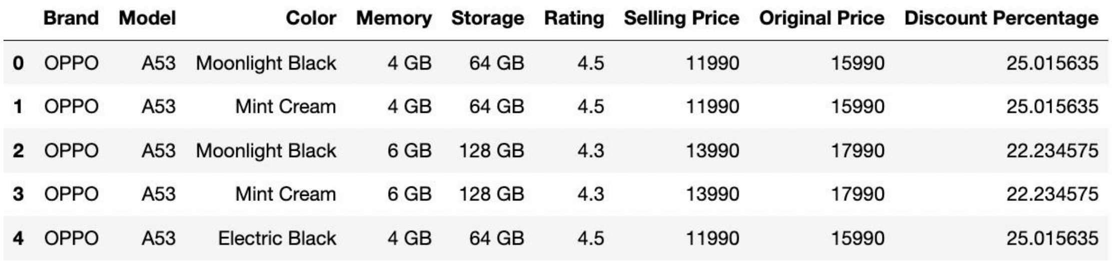
1. We have built Decision Regression Tree models using Selling Price as target variable.
2. Added extra visualization to find interesting relationship in the data using box plots and 3-d graph.
3. Evaluated the model performance as a result of using ‘selling price’ as target variable.

Overview of dataset

The selected dataset contains 3114 rows and it contains following attributes.



```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3114 entries, 0 to 3113
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Brand                3114 non-null   object 
 1   Model                3114 non-null   object 
 2   Color                3114 non-null   object 
 3   Memory               3071 non-null   object 
 4   Storage              3075 non-null   object 
 5   Rating               2970 non-null   float64
 6   Selling Price        3114 non-null   int64  
 7   Original Price       3114 non-null   int64  
 8   Discount Percentage  3114 non-null   float64
dtypes: float64(2), int64(2), object(5)
memory usage: 219.1+ KB
```



	Brand	Model	Color	Memory	Storage	Rating	Selling Price	Original Price	Discount Percentage
0	OPPO	A53	Moonlight Black	4 GB	64 GB	4.5	11990	15990	25.015635
1	OPPO	A53	Mint Cream	4 GB	64 GB	4.5	11990	15990	25.015635
2	OPPO	A53	Moonlight Black	6 GB	128 GB	4.3	13990	17990	22.234575
3	OPPO	A53	Mint Cream	6 GB	128 GB	4.3	13990	17990	22.234575
4	OPPO	A53	Electric Black	4 GB	64 GB	4.5	11990	15990	25.015635

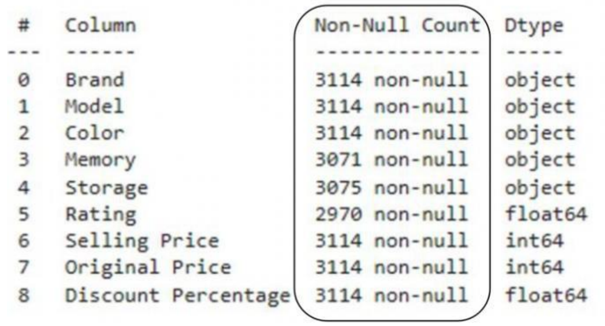
There are total 9 variables and 3114 records. These attributes are very important in order to carry out the in depth analysis and to check which are influencing more with respect to the sales.

The columns in the selected dataset are Brand, Model, Color, Memory, Storage, Rating, Selling price, Original price, Discount percentage. These attributes fall in either of the category i.e. float, int, object.

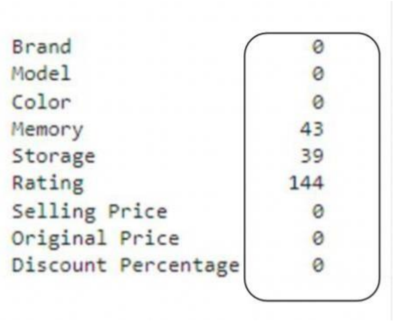
When we see the count of each column, we could observe that there are chances of having null values because the count differed in each of the column. These could be handled in further steps. The analysis is performed on this dataset, and it includes various steps. All the steps and methods applied are explained in detail in further sections.

Data Preprocessing:-

Since the data obtained from datasets or real-world data is never prepared, data preparation is essential and is regarded as one of the primary prerequisites. It is always viewed as raw information. In order to be suitable for analysis, this raw data needs to be organised and consistent. Methods for data preprocessing can help with issues like missing values, outliers, and data anomalies. By improving data quality, analysis and modelling processes become more dependable, precise, and accurate.

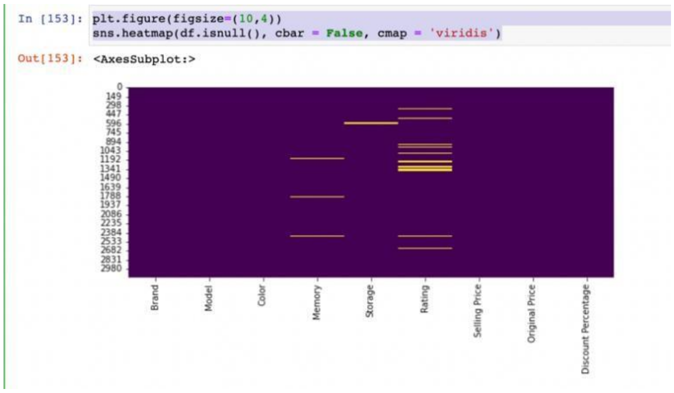


```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Brand                3114 non-null   object
 1   Model                3114 non-null   object
 2   Color                3114 non-null   object
 3   Memory               3071 non-null   object
 4   Storage              3075 non-null   object
 5   Rating               2970 non-null   float64
 6   Selling Price        3114 non-null   int64
 7   Original Price       3114 non-null   int64
 8   Discount Percentage  3114 non-null   float64
```



```
Brand                  0
Model                  0
Color                  0
Memory                43
Storage               39
Rating               144
Selling Price          0
Original Price         0
Discount Percentage    0
```

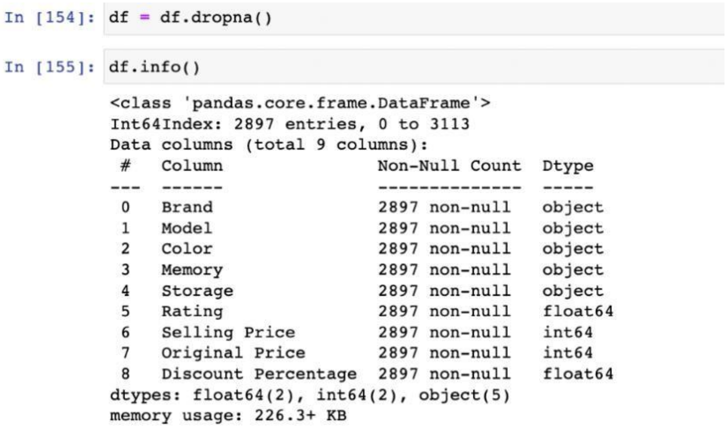
Now the first and most common approach in data preprocessing is to get rid of null values. From the above figures it is very evident that there are null values in the dataset. We can see the count of null values with respect to each column in the above diagram.



```
In [153]: plt.figure(figsize=(10,4))
          sns.heatmap(df.isnull(), cbar = False, cmap = 'viridis')

Out[153]: <AxesSubplot:>
```

Now we have visualized the null values using heatmap. The heatmap of the data is shown in the above figure.

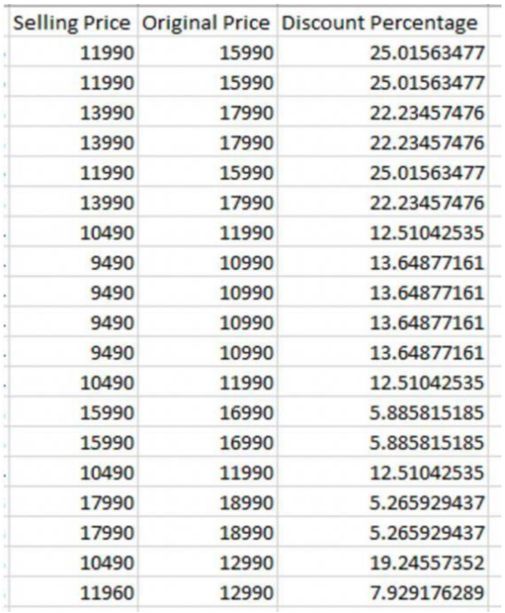
In order to handle the null values, we have chosen to drop those records that are having the null values by using the below code.



```
In [154]: df = df.dropna()

In [155]: df.info()

          <class 'pandas.core.frame.DataFrame'>
          Int64Index: 2897 entries, 0 to 3113
          Data columns (total 9 columns):
           #   Column               Non-Null Count  Dtype
          ---  ------               --------------  -----
           0   Brand                2897 non-null   object
           1   Model                2897 non-null   object
           2   Color                2897 non-null   object
           3   Memory               2897 non-null   object
           4   Storage              2897 non-null   object
           5   Rating               2897 non-null   float64
           6   Selling Price        2897 non-null   int64
           7   Original Price       2897 non-null   int64
           8   Discount Percentage  2897 non-null   float64
          dtypes: float64(2), int64(2), object(5)
          memory usage: 226.3+ KB
```

From the above figure it is very evident that after dropping the records having the null values, now the count of the records is reduced to 2897. Now the dataset contains 2897 records.

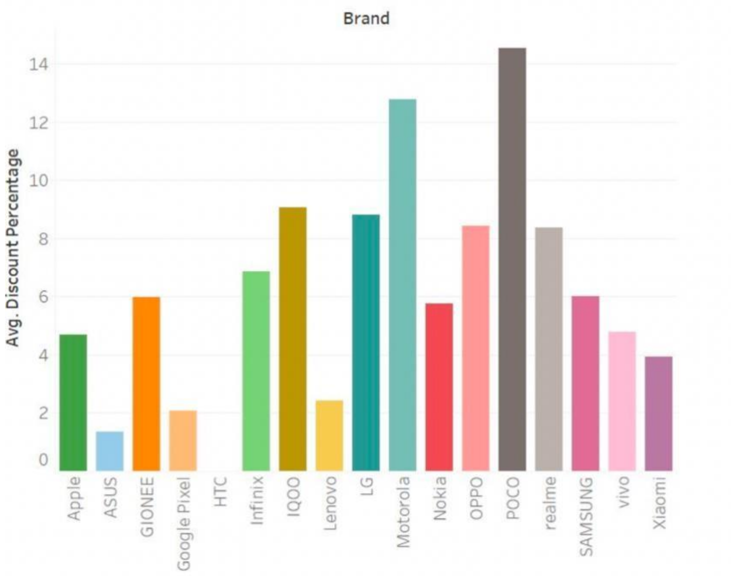
As a part of data modification we have added a new column discount percentage that gives us the percentage that is applied on the product’s original price.



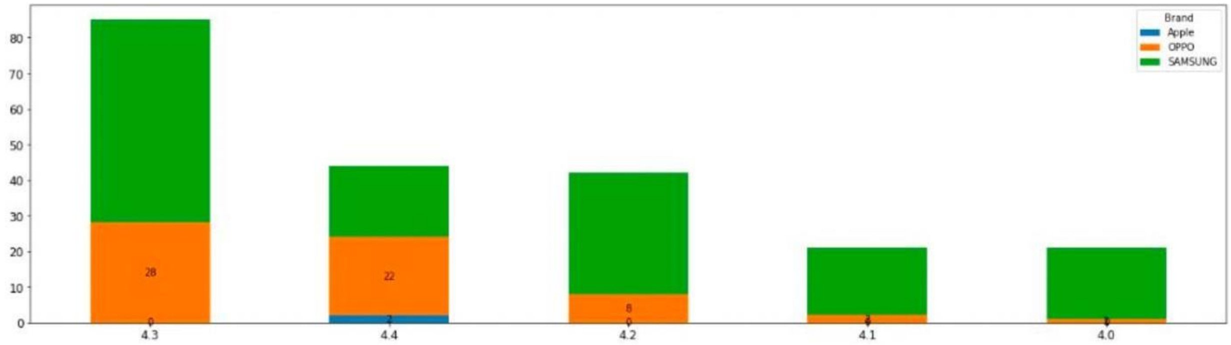
Selling Price	Original Price	Discount Percentage
11990	15990	25.01563477
11990	15990	25.01563477
13990	17990	22.23457476
13990	17990	22.23457476
11990	15990	25.01563477
13990	17990	22.23457476
10490	11990	12.51042535
9490	10990	13.64877161
9490	10990	13.64877161
9490	10990	13.64877161
9490	10990	13.64877161
10490	11990	12.51042535
15990	16990	5.885815185
15990	16990	5.885815185
10490	11990	12.51042535
17990	18990	5.265929437
17990	18990	5.265929437
10490	12990	19.24557352
11960	12990	7.929176289

Exploratory Data analysis:-

Now we have visualized the data by plotting against the variables with the target variable. This gives us the relationship or the trend in terms of numbers or the respective values.

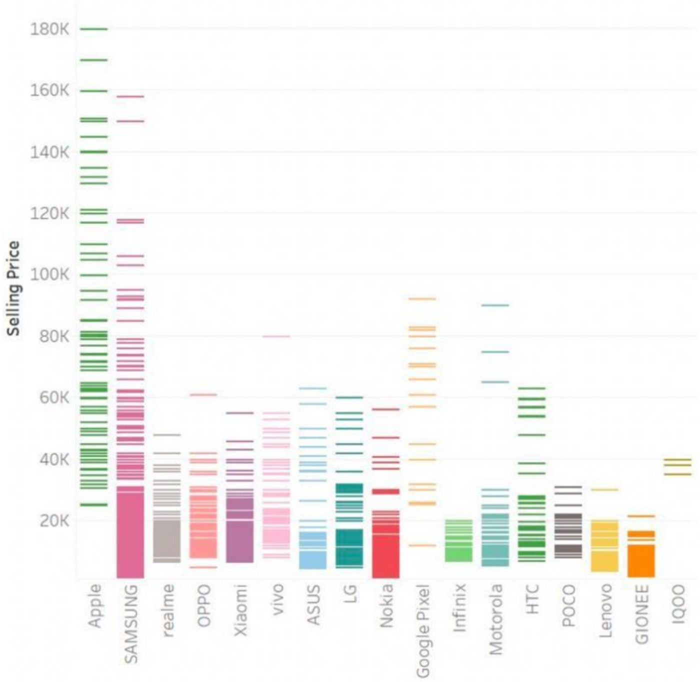


We have plotted a graph or a histogram which gives us the discount percentage offered by the various companies. The above is the resultant graph. From that it is very evident that POCO and Motorola offers a greater deals or greater discounts.



Now in order to understand the ratings of the company the above visualization is obtained. Here we have chosen three brands named as Apple, Oppo, Samsung. The legend shows the companies with its corresponding colors. By analyzing the obtained result, we have understood that Samsung has best average ratings from the customers.

It is also critical to understand which brand offers the widest selection of items for the Indian market, catering to various categories such as low, mid, and premium.

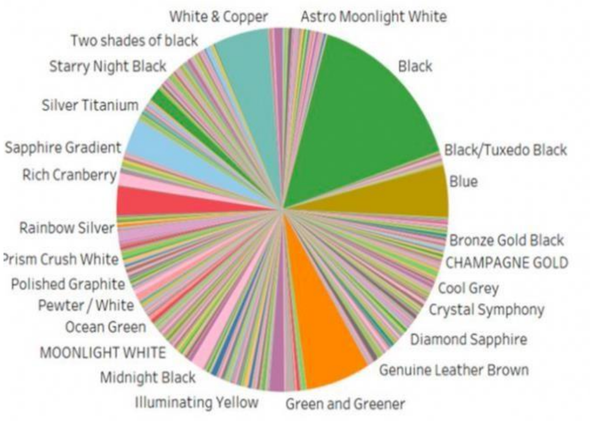


We can infer from the graph above that Apple and Samsung offer a variety of products. IQOO offers the fewest options. The reaming businesses sell products in the middle range. The premium-range price determination varied by 77.7% from those of low-range items, as shown by Apple and Samsung.

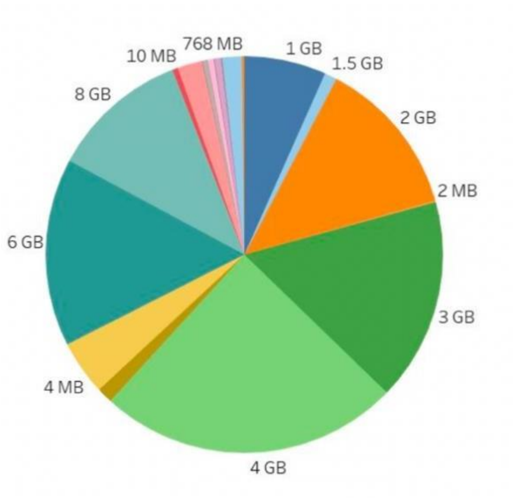
Additionally, it shows that the premium-range items cost about 77.7% more than their low-range counterparts.

The above figure also provides another intriguing insight. It is evident that the majority of the businesses create products in the middle market.

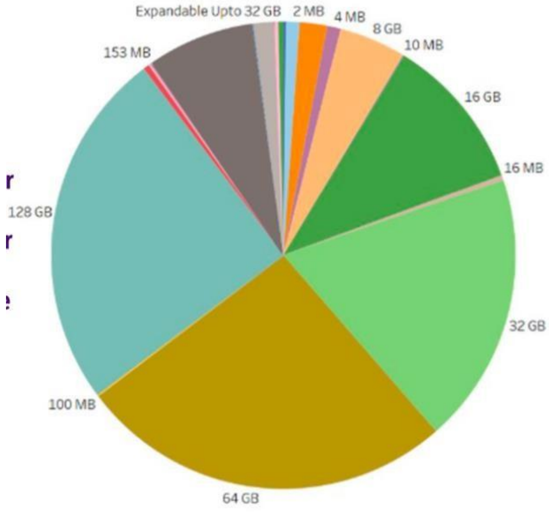
The majority of the general information about production values, widely used specifications, etc., can be found if we analyse the products that companies are offering in terms of their specifications.



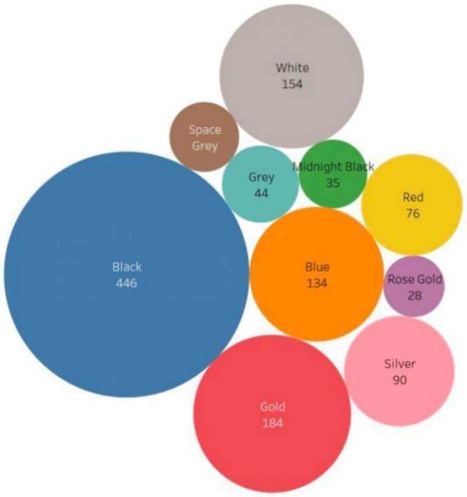
When we analyze the product with respect to the color, then from the above chart we can conclude that the most produced or utilized color for the production is black, green and blue. Colors like Amber red, Arctic blue are utilized very less.



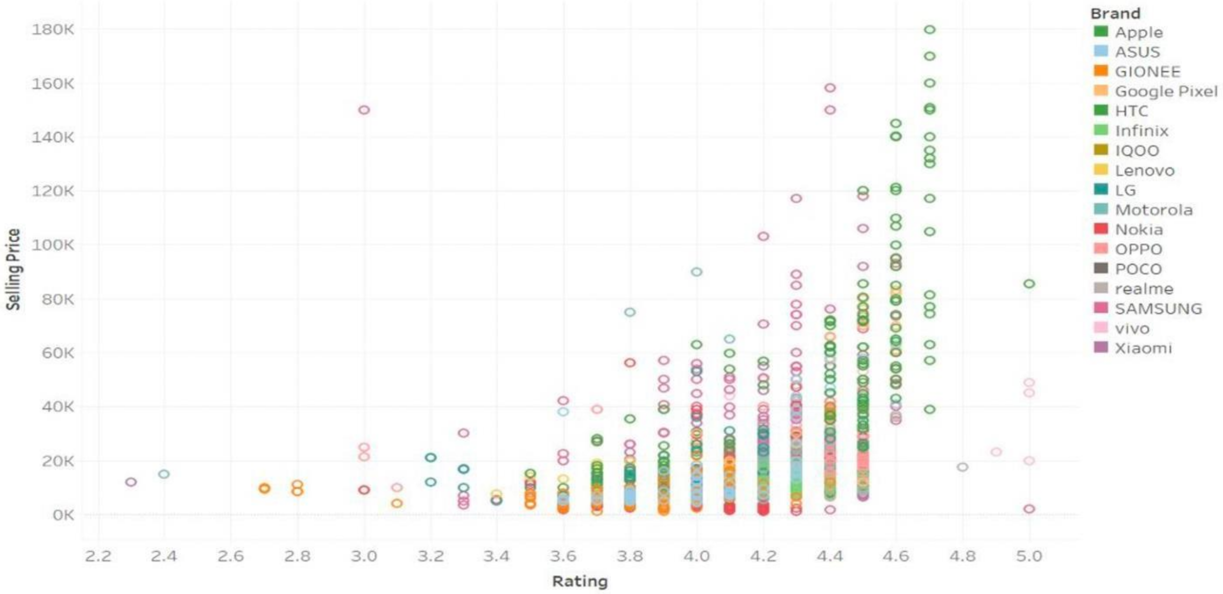
The next important specification is memory. The above chart depicts the representation of the memory with respect to majority of the phones. From the visualization we can easily say that most commonly offered Memory are 4GB, 3GB and 6GB.



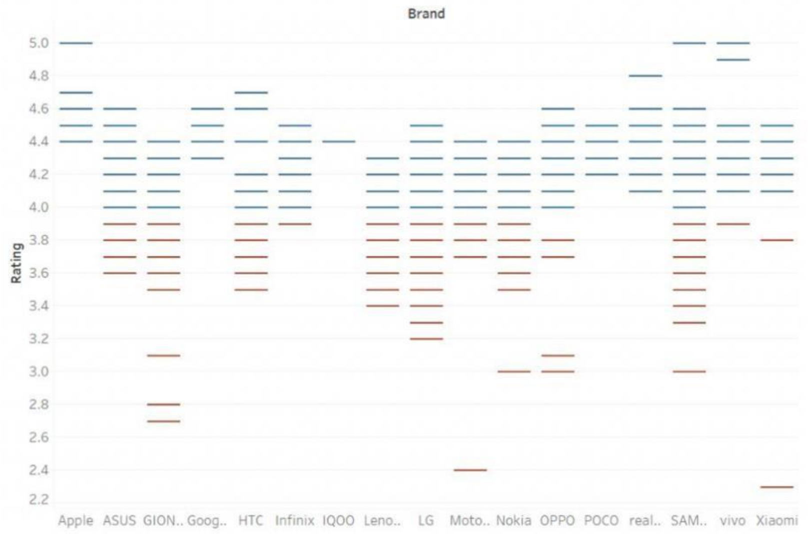
The next visualized specification is storage. From the above chart we can draw few conclusions. 64 GB, 128 GB, and 32GB were the most frequently acknowledged storage options, with 64 GB given as a choice 757 times and 16 MB being considered only 10 times.



Sales data show that there is a high demand for mobile phones with 4GB of memory and 64GB of storage in the colours black, green, and blue. On the other hand, less common colour combinations like Amber Red and Arctic Blue, as well as smaller memory or storage options, are less widely accessible on the market.

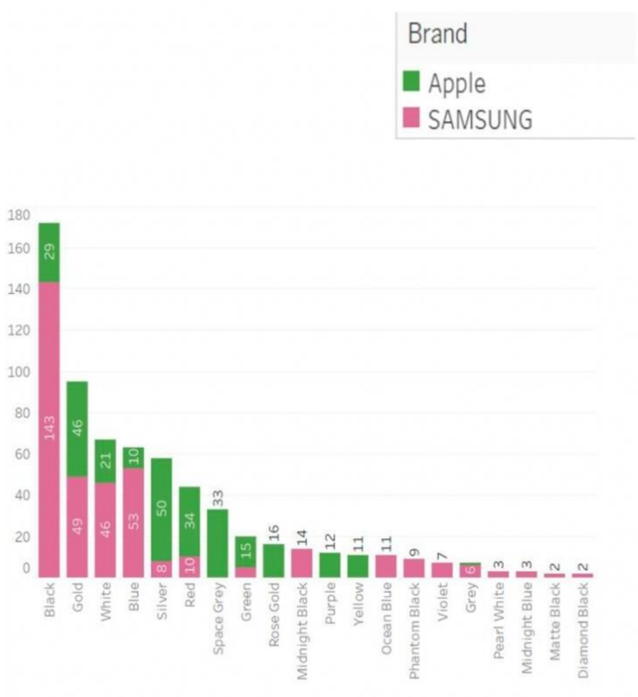


When we plot a graph against rating and selling price, we got the above results where it clearly demonstrates that even high-end mobile phones are not considered to be that pricey.



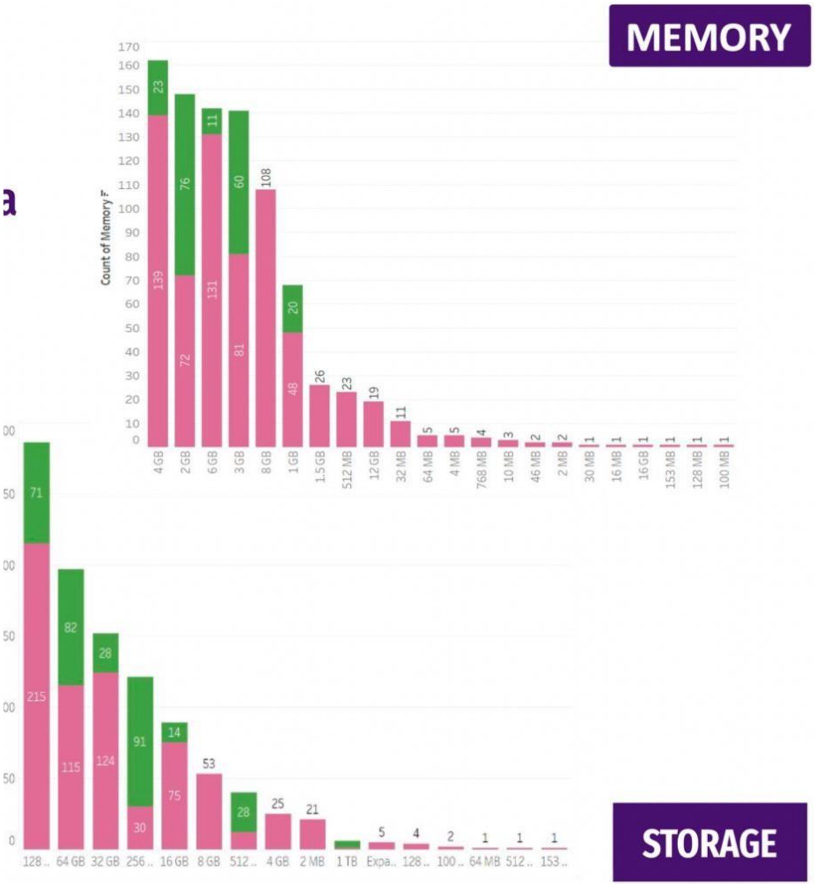
Finding out a brand's average rating after taking customer ratings into account would be interesting. It is very obvious from the above figure that companies like Apple, IQOO, Google, Poco, and realme have all inescapably amassed ratings over 4 for every product they sell.

Now let's look at an intriguing comparison between two brands that were regarded as the best. We can infer from the results above that Apple and Samsung are the top two businesses in terms of ratings or overall product scope.



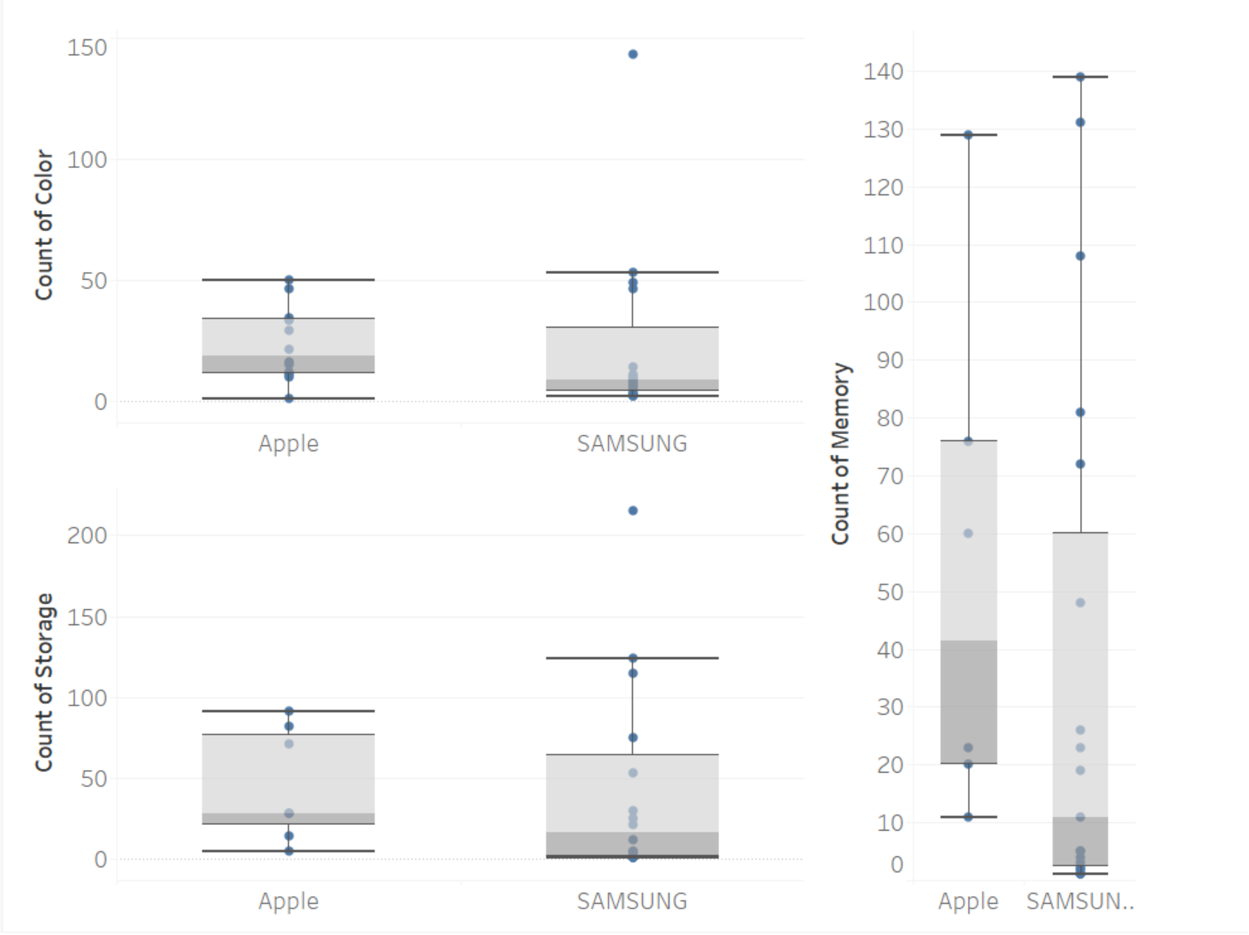
The visualisation shows that the Samsung brand offers a wider variety of colours than the Apple brand.
Apple-Space Grey, Rose Gold, and Purple are examples of distinctive colours that are thought to be extremely picky or specific with regard to both brands.
Samsung's darker tones

These brands can also be contrasted in terms of additional features like memory and storage. With the help of all these analyses, we can show which attributes are the most and least used. We can even infer from these that wider use indicates that the product is well-liked or frequently bought by customers.
These could also lead to the reasons for the fluctuations in the sales that make them unique or different from the other brands and makes them stay little ahead from rest of them.

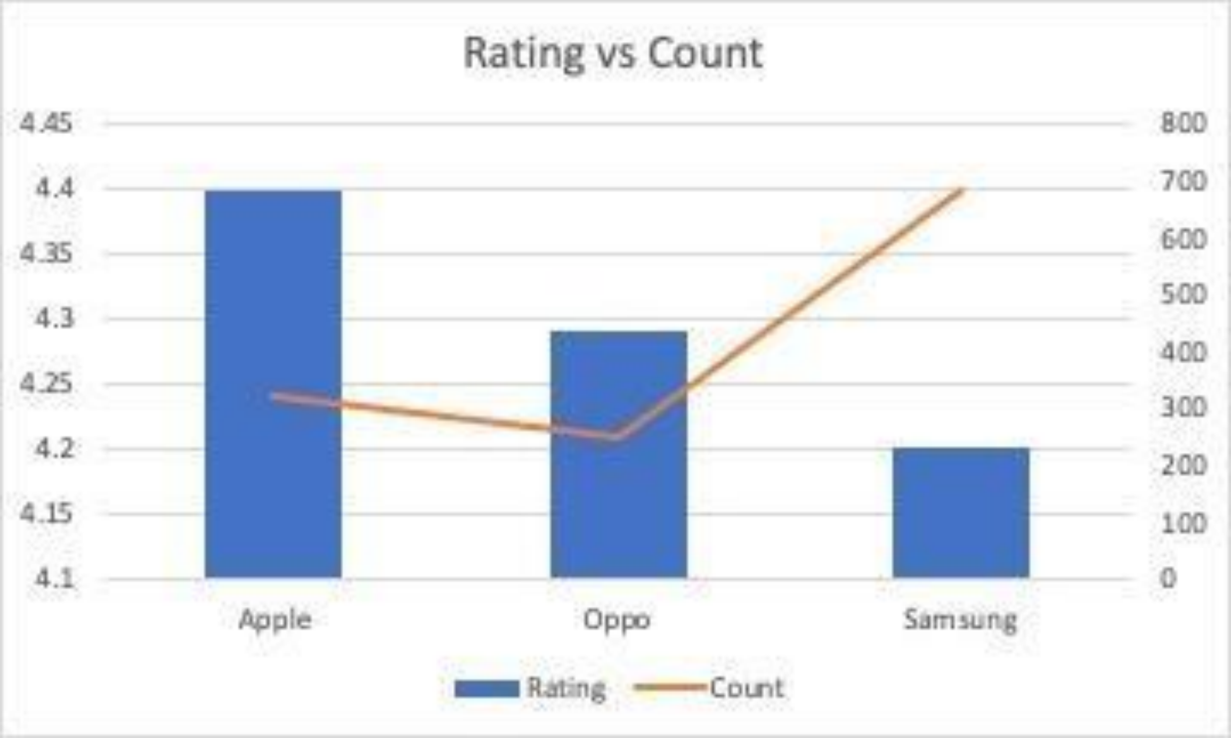


In accordance with memory, Samsung provides a broader selection of products all throughout several different categories. Samsung offers simultaneously smaller and larger range items in the market, unlike Apple, which only offers higher range storage. As shown in the graph, customers prefer high storage and memory options.

Above three graphs showing comparisons between Apple and Samsung is being depicted in the box plot below. The outliers are there in plot which represents the values which are very far from the frequently occurring data.



The graph below depicts the link between three prominent smartphone brands' ratings and counts: Apple, Oppo, and Samsung. The different brands are represented by the X-axis, while the Y-axis displays the respective counts and average ratings. Apple gets the highest average rating of 4.4 among the brands, based on a total of 319 reviews. Oppo comes in second with an average rating of 4.29 and a total of 259 reviews. Samsung, on the other hand, has a better average rating of 4.2 based on 685 reviews. We can compare the popularity and satisfaction levels of different brands based on their respective counts and ratings, which provides useful insights into customer preferences and brand success.

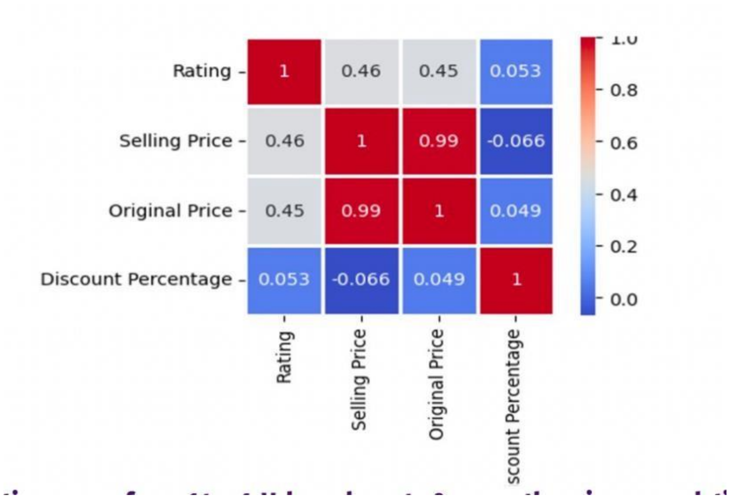


Correlation analysis

A technique called correlation analysis is used to investigate the statistically significant relationship between two or more variables in a dataset. It makes it simpler to determine the relationship between changes in one variable and changes in another.

Feature selection is a crucial step in data mining and machine learning to find the most important and insightful variables to create predictive models. Data miners can identify highly related features and give priority to those that contain unique information to avoid reoccurring words by using correlation analysis.

In correlation analysis, heat maps are frequently used to visually represent the strength and evolution of correlations among numerous variables in a dataset. In correlation analysis, a correlation matrix has been created that contains two correlation coefficients that specify the strength of the relationship between two pairs of variables.
Data analysts and researchers can easily navigate and spot associations and trends in the data thanks to heat maps, which provide a clear and understandable way to represent these correlation coefficients.



Correlation may fluctuate from -1 to +1.

Near 0 values indicate that there is no correlation between the two variables. If the correlation value is close to 1, the variables are more positively correlated; if it is close to -1, the variables are more negatively correlated.

According to the heatmap Since the original price and the selling price are positively correlated, both variables rise as the first does. Since the selling price and the discount percentage are inversely correlated, they both rise as one variable falls.

Data Modeling:-

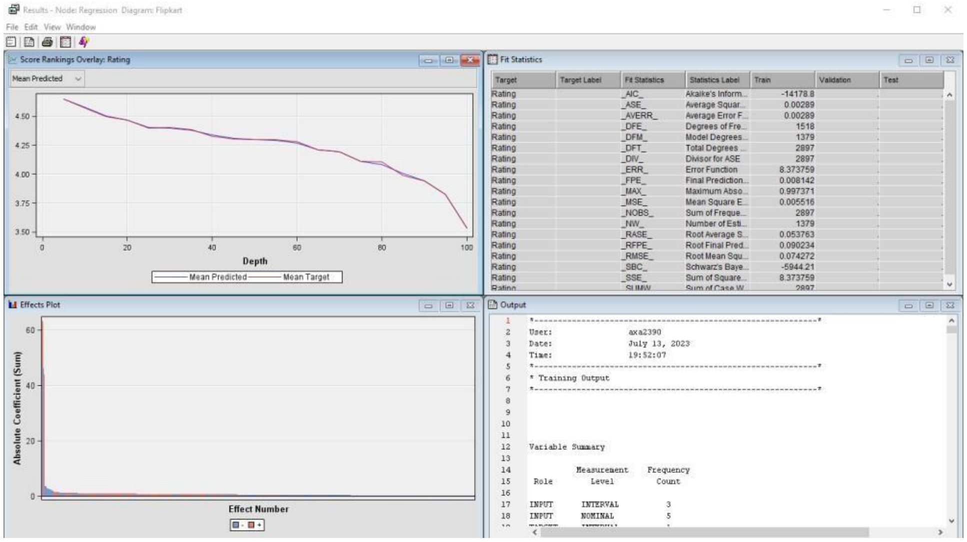
By using models built on historical data and characteristics, we can predict and anticipate events. Models help us to understand the connections and interactions between numerous different variables in a complex network. By using models, we can gain understanding of the underlying processes that underlie prevalent patterns and behaviour. Models for maximising operations can be used to achieve data-driven alternatives.

Model 1 :- Multiple linear regression

To investigate the relationship between a dependent variable and two or more independent variables, multiple linear regression is used. It is a modified form of fundamental linear regression that considers only one independent variable.

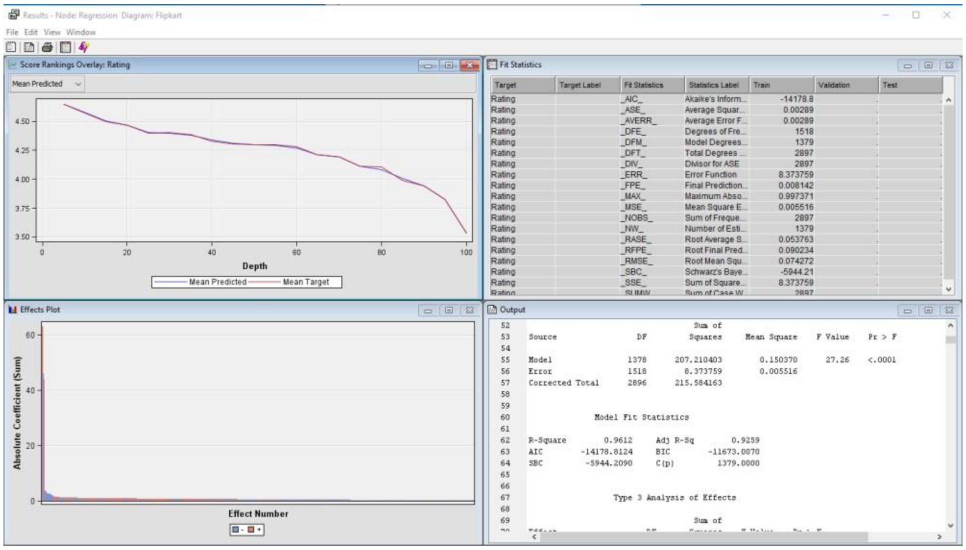
The goal of multiple linear regression is to identify the best-fitting linear equation that demonstrates how the dependent variable (target) changes when the independent variables are combined linearly.

Rating is the dependent variable in this specific case. By using multiple linear regression, we hope to identify a linear relationship among a wide range of independent variables. Brand, Colour, Model, Storage, Memory, Selling Price, Original Price, and Discount Percentage are the factors that predict characteristics.



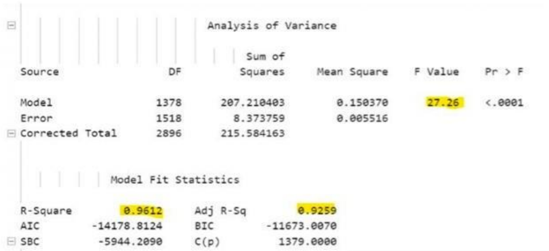
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating		_AIC_	Akaike's Inform...	-14178.8	.	.
Rating		_ASE_	Average Squar...	0.00289	.	.
Rating		_AVERR_	Average Error F...	0.00289	.	.
Rating		_DFE_	Degrees of Fre...	1518	.	.
Rating		_DFM_	Model Degrees...	1379	.	.
Rating		_DFT_	Total Degrees ...	2897	.	.
Rating		_DIV_	Divisor for ASE	2897	.	.
Rating		_ERR_	Error Function	8.373759	.	.
Rating		_FPE_	Final Prediction...	0.008142	.	.
Rating		_MAX_	Maximum Abso...	0.997371	.	.
Rating		_MSE_	Mean Square E...	0.005516	.	.
Rating		_NOBS_	Sum of Freque...	2897	.	.
Rating		_NW_	Number of Esti...	1379	.	.
Rating		_RASE_	Root Average S...	0.053763	.	.
Rating		_RFPE_	Root Final Pred...	0.090234	.	.
Rating		_RMSE_	Root Mean Squ...	0.074272	.	.
Rating		_SBC_	Schwarz's Baye...	-5944.21	.	.
Rating		_SSE_	Sum of Square...	8.373759	.	.
Rating		_SUMW_	Sum of Case W...	2897		

```
*------------------------------------------------------------*
User:                   axa2390
Date:                   July 13, 2023
Time:                   19:52:07
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          3
INPUT      NOMINAL           5
```

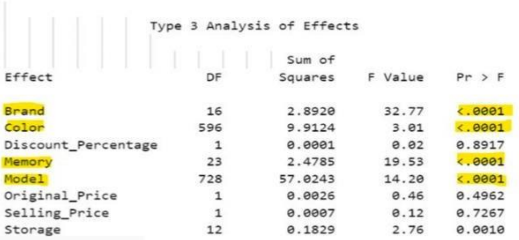


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating		_AIC_	Akaike's Inform...	-14178.8	.	.
Rating		_ASE_	Average Squar...	0.00289	.	.
Rating		_AVERR_	Average Error F...	0.00289	.	.
Rating		_DFE_	Degrees of Fre...	1518	.	.
Rating		_DFM_	Model Degrees...	1379	.	.
Rating		_DFT_	Total Degrees ...	2897	.	.
Rating		_DIV_	Divisor for ASE	2897	.	.
Rating		_ERR_	Error Function	8.373759	.	.
Rating		_FPE_	Final Prediction...	0.008142	.	.
Rating		_MAX_	Maximum Abso...	0.997371	.	.
Rating		_MSE_	Mean Square E...	0.005516	.	.
Rating		_NOBS_	Sum of Freque...	2897	.	.
Rating		_NW_	Number of Esti...	1379	.	.
Rating		_RASE_	Root Average S...	0.053763	.	.
Rating		_RFPE_	Root Final Pred...	0.090234	.	.
Rating		_RMSE_	Root Mean Squ...	0.074272	.	.
Rating		_SBC_	Schwarz's Baye...	-5944.21	.	.
Rating		_SSE_	Sum of Square...	8.373759	.	.
Rating		_SUMW_	Sum of Case W...	2897		

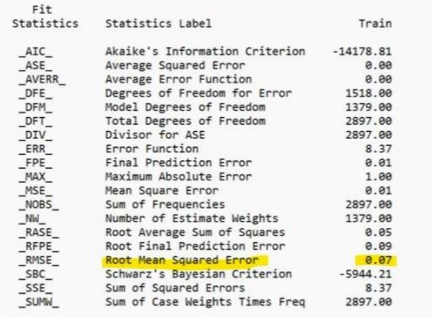
```
                                   Sum of
Source                DF          Squares     Mean Square    F Value    Pr > F

Model               1378       207.210403        0.150370      27.26    <.0001
Error               1518         8.373759        0.005516
Corrected Total     2896       215.584163


              Model Fit Statistics

R-Square        0.9612     Adj R-Sq        0.9259
AIC        -14178.8124     BIC        -11673.0070
SBC         -5944.2090     C(p)         1379.0000


                Type 3 Analysis of Effects

                                   Sum of
```



```
                        Analysis of Variance

                                  Sum of
Source                DF         Squares    Mean Square   F Value   Pr > F

Model               1378      207.210403       0.150370     27.26   <.0001
Error               1518        8.373759       0.005516
Corrected Total     2896      215.584163


              Model Fit Statistics

R-Square        0.9612     Adj R-Sq        0.9259
AIC        -14178.8124     BIC        -11673.0070
SBC         -5944.2090     C(p)         1379.0000
```



```
              Type 3 Analysis of Effects

                                   Sum of
Effect                  DF        Squares    F Value    Pr > F

Brand                   16         2.8920      32.77    <.0001
Color                  596         9.9124       3.01    <.0001
Discount_Percentage      1         0.0001       0.02    0.8917
Memory                  23         2.4785      19.53    <.0001
Model                  728        57.0243      14.20    <.0001
Original_Price           1         0.0026       0.46    0.4962
Selling_Price            1         0.0007       0.12    0.7267
Storage                 12         0.1829       2.76    0.0010
```

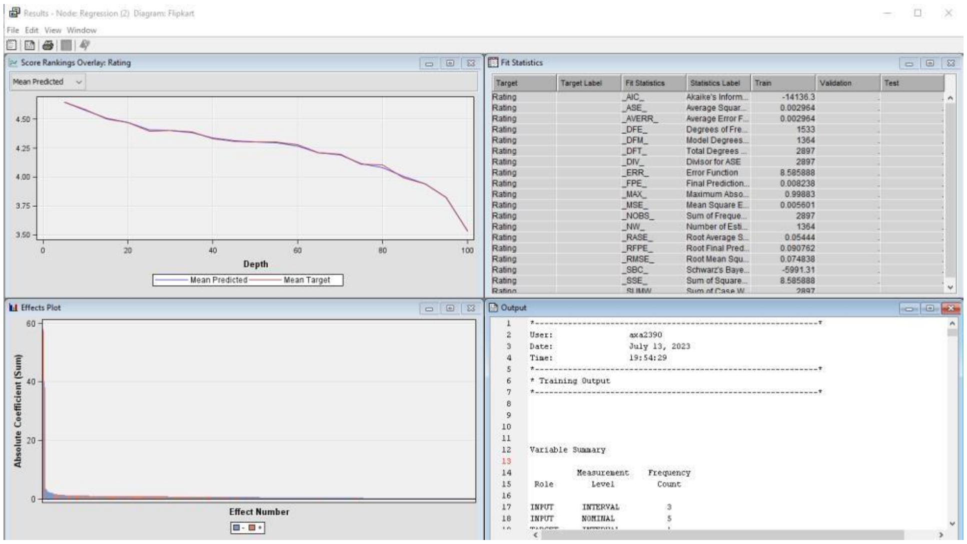


```
   Fit
Statistics    Statistics Label                         Train

 _AIC_        Akaike's Information Criterion       -14178.81
 _ASE_        Average Squared Error                     0.00
 _AVERR_      Average Error Function                    0.00
 _DFE_        Degrees of Freedom for Error           1518.00
 _DFM_        Model Degrees of Freedom               1379.00
 _DFT_        Total Degrees of Freedom               2897.00
 _DIV_        Divisor for ASE                        2897.00
 _ERR_        Error Function                            8.37
 _FPE_        Final Prediction Error                    0.01
 _MAX_        Maximum Absolute Error                    1.00
 _MSE_        Mean Square Error                         0.01
 _NOBS_       Sum of Frequencies                     2897.00
 _NW_         Number of Estimate Weights             1379.00
 _RASE_       Root Average Sum of Squares               0.05
 _RFPE_       Root Final Prediction Error               0.09
 _RMSE_       Root Mean Squared Error                   0.07
 _SBC_        Schwarz's Bayesian Criterion          -5944.21
 _SSE_        Sum of Squared Errors                     8.37
 _SUMW_       Sum of Case Weights Times Freq         2897.00
```

The F-test score of 27.26, as well as the low AIC and SBC values, indicate that the model is reasonably suited. According to the model, R-square explains 96.12% of the variation in the dependent variable. The modified R-square of 92.59% suggests that the model fits the data well.

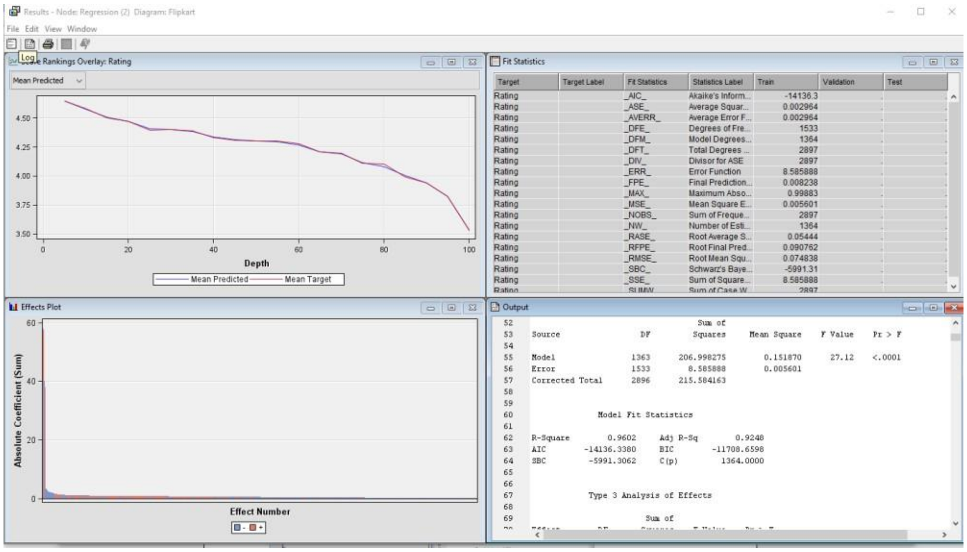
Model 2: Running Linear regression model

Running a linear regression model utilizing the only significant variables observed. Brand, Memory, Model, and Storage are the predictive factors in this case.



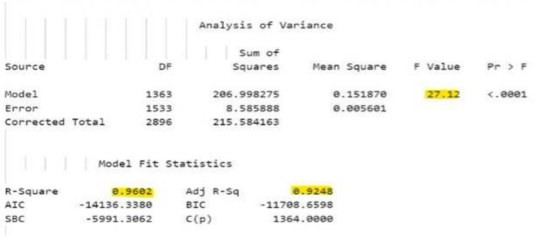
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating		_AIC_	Akaike's Inform...	-14136.3	.	.
Rating		_ASE_	Average Squar...	0.002964	.	.
Rating		_AVERR_	Average Error F...	0.002964	.	.
Rating		_DFE_	Degrees of Fre...	1533	.	.
Rating		_DFM_	Model Degrees...	1364	.	.
Rating		_DFT_	Total Degrees ...	2897	.	.
Rating		_DIV_	Divisor for ASE	2897	.	.
Rating		_ERR_	Error Function	8.585888	.	.
Rating		_FPE_	Final Prediction...	0.008238	.	.
Rating		_MAX_	Maximum Abso...	0.99883	.	.
Rating		_MSE_	Mean Square E...	0.005601	.	.
Rating		_NOBS_	Sum of Freque...	2897	.	.
Rating		_NW_	Number of Esti...	1364	.	.
Rating		_RASE_	Root Average S...	0.05444	.	.
Rating		_RFPE_	Root Final Pred...	0.090762	.	.
Rating		_RMSE_	Root Mean Squ...	0.074838	.	.
Rating		_SBC_	Schwarz's Baye...	-5991.31	.	.
Rating		_SSE_	Sum of Square...	8.585888	.	.
Rating		_SUMW_	Sum of Case W...	2897		

```
*------------------------------------------------------------*
User:                    axa2390
Date:                    July 13, 2023
Time:                    19:54:29
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          3
INPUT      NOMINAL           5
```

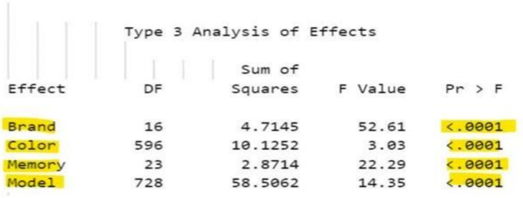


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating		_AIC_	Akaike's Inform...	-14136.3	.	.
Rating		_ASE_	Average Squar...	0.002964	.	.
Rating		_AVERR_	Average Error F...	0.002964	.	.
Rating		_DFE_	Degrees of Fre...	1533	.	.
Rating		_DFM_	Model Degrees...	1364	.	.
Rating		_DFT_	Total Degrees ...	2897	.	.
Rating		_DIV_	Divisor for ASE	2897	.	.
Rating		_ERR_	Error Function	8.585888	.	.
Rating		_FPE_	Final Prediction...	0.008238	.	.
Rating		_MAX_	Maximum Abso...	0.99883	.	.
Rating		_MSE_	Mean Square E...	0.005601	.	.
Rating		_NOBS_	Sum of Freque...	2897	.	.
Rating		_NW_	Number of Esti...	1364	.	.
Rating		_RASE_	Root Average S...	0.05444	.	.
Rating		_RFPE_	Root Final Pred...	0.090762	.	.
Rating		_RMSE_	Root Mean Squ...	0.074838	.	.
Rating		_SBC_	Schwarz's Baye...	-5991.31	.	.
Rating		_SSE_	Sum of Square...	8.585888	.	.
Rating		_SUMW_	Sum of Case W...	2897		

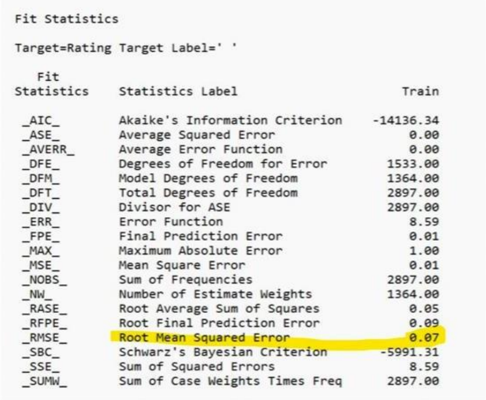
```
                                  Sum of
Source                DF         Squares      Mean Square    F Value    Pr > F

Model               1363      206.998275         0.151870      27.12    <.0001
Error               1533        8.585888         0.005601
Corrected Total     2896      215.584163


            Model Fit Statistics

R-Square        0.9602     Adj R-Sq        0.9248
AIC        -14136.3380     BIC        -11708.6598
SBC         -5991.3062     C(p)         1364.0000


           Type 3 Analysis of Effects

                          Sum of
```



```
                         Analysis of Variance

                                   Sum of
Source                DF          Squares    Mean Square    F Value    Pr > F

Model               1363       206.998275       0.151870      27.12    <.0001
Error               1533         8.585888       0.005601
Corrected Total     2896       215.584163

                Model Fit Statistics

R-Square          0.9602    Adj R-Sq          0.9248
AIC          -14136.3380    BIC          -11708.6598
SBC           -5991.3062    C(p)           1364.0000
```



```
              Type 3 Analysis of Effects

                         Sum of
Effect        DF        Squares    F Value    Pr > F

Brand         16         4.7145      52.61    <.0001
Color        596        10.1252       3.03    <.0001
Memory        23         2.8714      22.29    <.0001
Model        728        58.5062      14.35    <.0001
```



```
Fit Statistics

Target=Rating Target Label=' '

   Fit
Statistics    Statistics Label                    Train

 _AIC_        Akaike's Information Criterion   -14136.34
 _ASE_        Average Squared Error                 0.00
 _AVERR_      Average Error Function                0.00
 _DFE_        Degrees of Freedom for Error       1533.00
 _DFM_        Model Degrees of Freedom           1364.00
 _DFT_        Total Degrees of Freedom           2897.00
 _DIV_        Divisor for ASE                    2897.00
 _ERR_        Error Function                        8.59
 _FPE_        Final Prediction Error                0.01
 _MAX_        Maximum Absolute Error                1.00
 _MSE_        Mean Square Error                     0.01
 _NOBS_       Sum of Frequencies                 2897.00
 _NW_         Number of Estimate Weights         1364.00
 _RASE_       Root Average Sum of Squares           0.05
 _RFPE_       Root Final Prediction Error           0.09
 _RMSE_       Root Mean Squared Error               0.07
 _SBC_        Schwarz's Bayesian Criterion      -5991.31
 _SSE_        Sum of Squared Errors                 8.59
 _SUMW_       Sum of Case Weights Times Freq     2897.00
```

The output discloses that a result for the F-test of 27.12 and a lower value than that of the previous model in AIC and SBC, indicating that the model is appropriately fit. R-square explains 96.02% of the variation in the dependent variable as described by the model. The modified R-square of 92.48% suggests that the model is a strong match in predicting the analysis.
The Type 3 analysis reveals that the Brand, Memory, model, and storage endure low p-values and possess a statistically substantial influence on affecting the rating target variable.

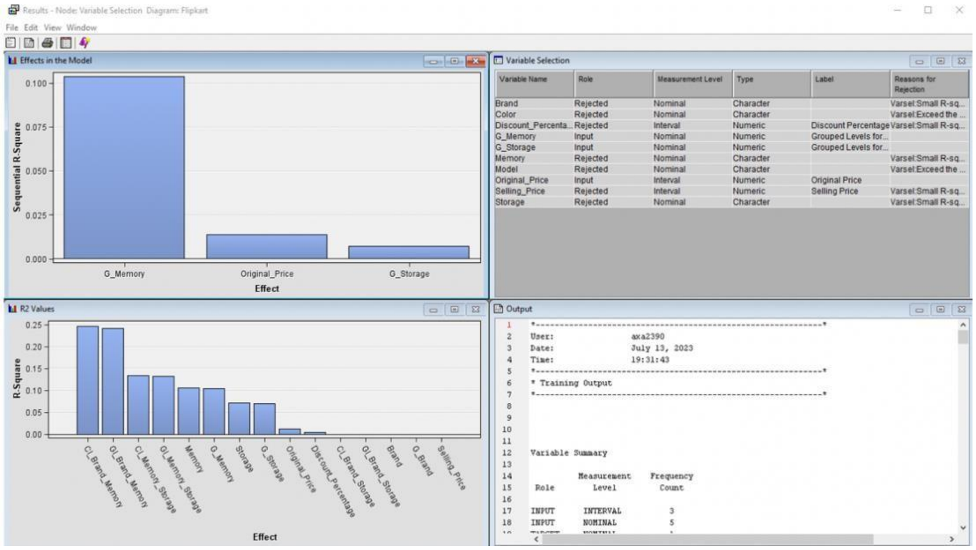
Model 2: Decision Regression Tree regression model

Non-linear relationships between the dependent and independent variables can be handled by Decision Regression Tree regression. It is especially useful when linear models fall short of accurately capturing the interrelationship between the variables.Decision Regression Trees are useful for gaining insights into the factors influencing the goal variable because they are easy to recognize and analyse.

People can explore the Decision pathways and comprehend how the algorithm arrives at its predictions using the Decision Regression Tree structure.Decision Regression Tree regression does not require regularly distributed data, unlike some other regression methods, and is robust to outliers. It functions well with data that deviates from the assumptions of linear regression.

A non-parametric supervised learning method for continuous regression is Decision Regression Tree regression. To put it another way, it learns by making choices based on training the model with a pair of inputs and their associated outputs.

Therefore, we divided the dataset into 40% for validation and 60% for training. A Decision model that calculates the value of a target variable—often a nominal class value—by acquiring essential Decision rules from data features is the desired outcome.

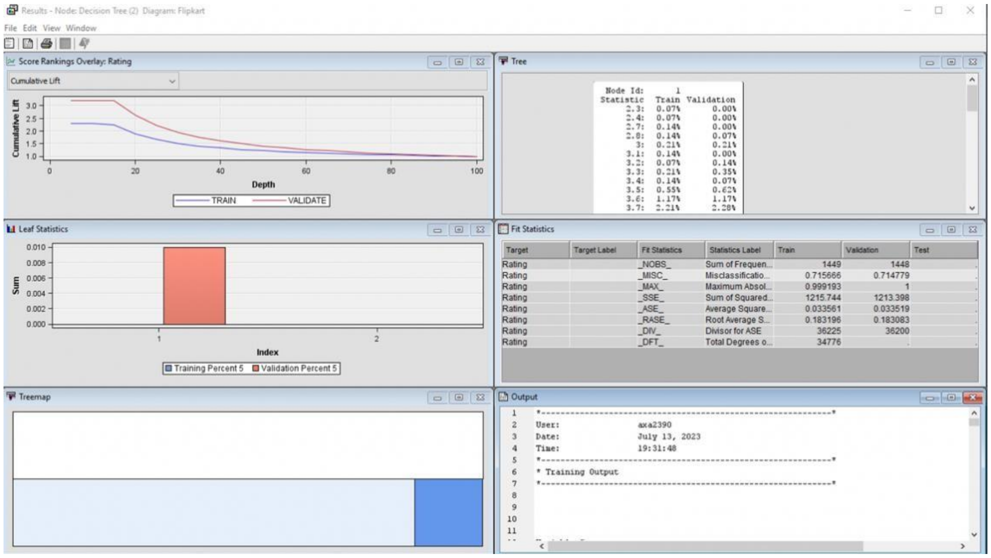


Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
Brand	Rejected	Nominal	Character		Varsel:Small R-sq...
Color	Rejected	Nominal	Character		Varsel:Exceed the ...
Discount_Percenta...	Rejected	Interval	Numeric	Discount Percentage	Varsel:Small R-sq...
G_Memory	Input	Nominal	Numeric	Grouped Levels for...	
G_Storage	Input	Nominal	Numeric	Grouped Levels for...	
Memory	Rejected	Nominal	Character		Varsel:Small R-sq...
Model	Rejected	Nominal	Character		Varsel:Exceed the ...
Original_Price	Input	Interval	Numeric	Original Price	
Selling_Price	Rejected	Interval	Numeric	Selling Price	Varsel:Small R-sq...
Storage	Rejected	Nominal	Character		Varsel:Small R-sq...

Output

```
*------------------------------------------------------------*
User:               axa2390
Date:               July 13, 2023
Time:               19:31:43
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          3
INPUT      NOMINAL           5
```



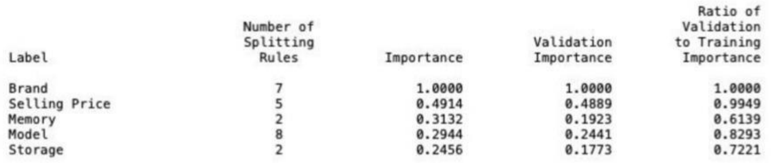
```
 Node Id:        1
Statistic   Train  Validation
     2.3:   0.07%       0.00%
     2.4:   0.07%       0.00%
     2.7:   0.14%       0.00%
     2.8:   0.14%       0.07%
       3:   0.21%       0.21%
     3.1:   0.14%       0.00%
     3.2:   0.07%       0.14%
     3.3:   0.21%       0.35%
     3.4:   0.14%       0.07%
     3.5:   0.55%       0.62%
     3.6:   1.17%       1.17%
     3.7:   2.21%       2.28%
```

Fit Statistics

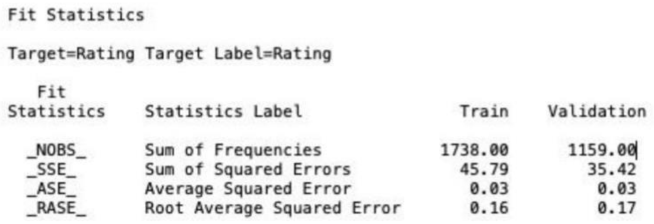
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating		_NOBS_	Sum of Frequen...	1449	1448	.
Rating		_MISC_	Misclassificatio...	0.715666	0.714779	.
Rating		_MAX_	Maximum Absol...	0.999193	1	.
Rating		_SSE_	Sum of Squared...	1215.744	1213.398	.
Rating		_ASE_	Average Square...	0.033561	0.033519	.
Rating		_RASE_	Root Average S...	0.183196	0.183083	.
Rating		_DIV_	Divisor for ASE	36225	36200	.
Rating		_DFT_	Total Degrees o...	34776	.	.

Output

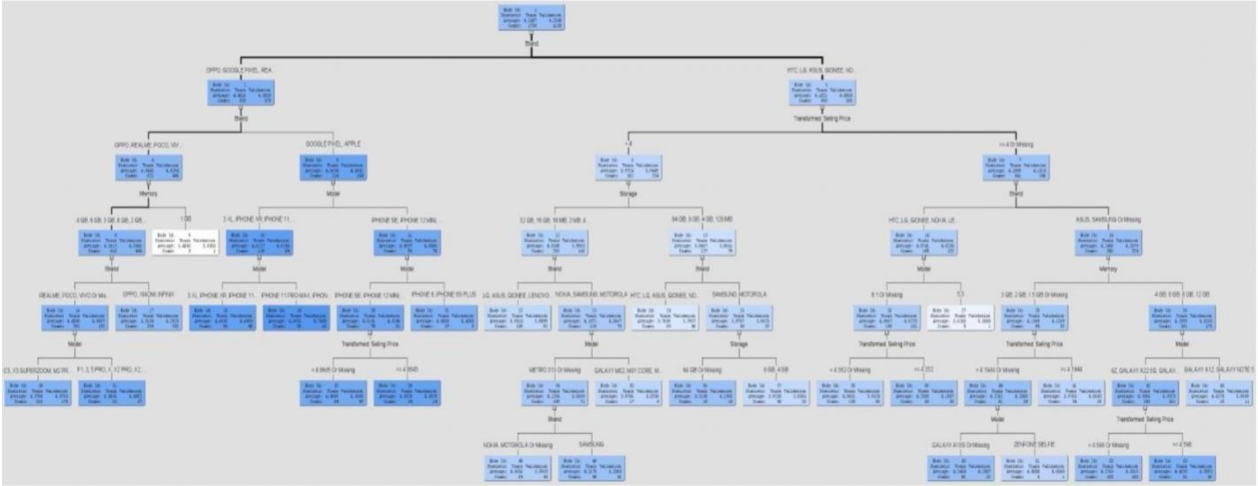
```
*------------------------------------------------------------*
User:               axa2390
Date:               July 13, 2023
Time:               19:31:48
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```



```
                        Number of                                        Ratio of
                        Splitting                       Validation    to Training
Label                     Rules        Importance       Importance     Importance

Brand                       7            1.0000           1.0000         1.0000
Selling Price               5            0.4914           0.4889         0.9949
Memory                      2            0.3132           0.1923         0.6139
Model                       8            0.2944           0.2441         0.8293
Storage                     2            0.2456           0.1773         0.7221
```



```
Fit Statistics

Target=Rating Target Label=Rating

    Fit
Statistics    Statistics Label                Train     Validation

  _NOBS_      Sum of Frequencies            1738.00        1159.00
  _SSE_       Sum of Squared Errors           45.79          35.42
  _ASE_       Average Squared Error            0.03           0.03
  _RASE_      Root Average Squared Error       0.16           0.17
```



Brand
OPPO, GOOGLE PIXEL, REA...
HTC, LG, ASUS, GIONEE, NO...
Transformed Selling Price
OPPO, REALME, POCO, VIV...
GOOGLE PIXEL, APPLE
Memory
Model
Storage

The first node in Node is referred to as the "root node." The entire dataset is represented by the root node. The algorithm chooses the most suitable feature to divide the data into two or more subsets, making it the first action point in the tree.The root node of this project is Brand.

The distance along the path leading from the root node to a leaf node is known as the "depth" of a Decision Regression Tree. A leaf node is one that doesn't divide any further and denotes a final Decision or result. This project has a node depth of 6.

Model comparison/Evaluation:

It is now crucial to compare the models in order to assess the performance of the model and its predictions.

Some typical metrics for evaluating regression models are listed below:

The average squared difference between anticipated and actual results is known as the mean squared error (MSE). A lower MSE indicates better performance.

Mean Absolute Error (MAE): This statistic measures the average absolute difference between expected and actual values. A lower MAE indicates a higher efficiency.

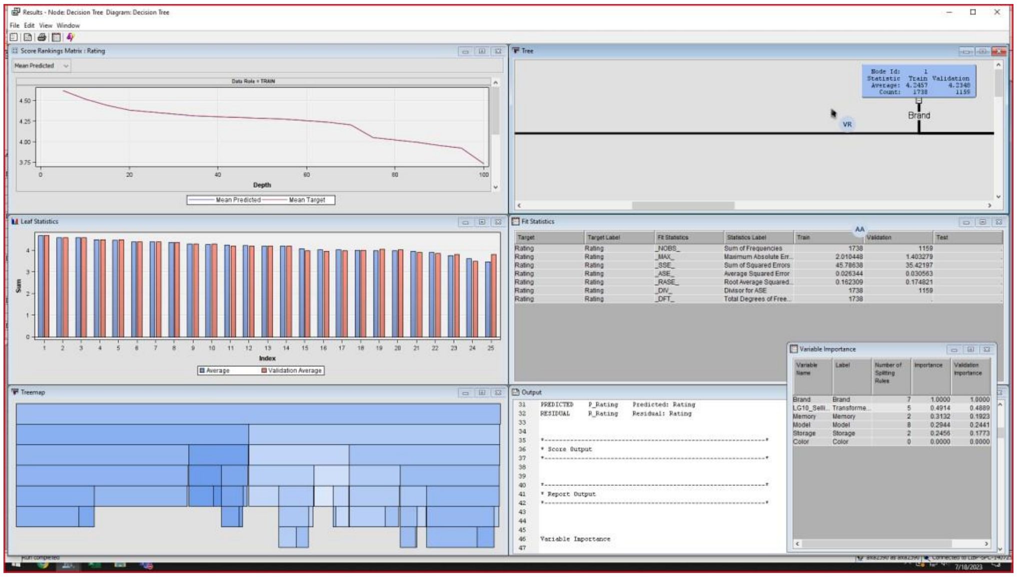
R-squared (R2): a measurement of how much variation in the dependent variable is explained by the model. The better the performance and best fit are indicated by higher R-squared values.

Root Mean Squared Error (RMSE): The square root of MSE, which provides a statistic that can be understood in the original unit of the target variable.

A model is said to be better at making predictions on unobserved data if it has a lower MSE, MAE, or RMSE and a higher R-squared value. However, it's crucial to avoid overfitting and assess the model's clarity and interpretability.

Let's compare the regression and Decision Regression Tree results now.

Decision Regression Tree:-



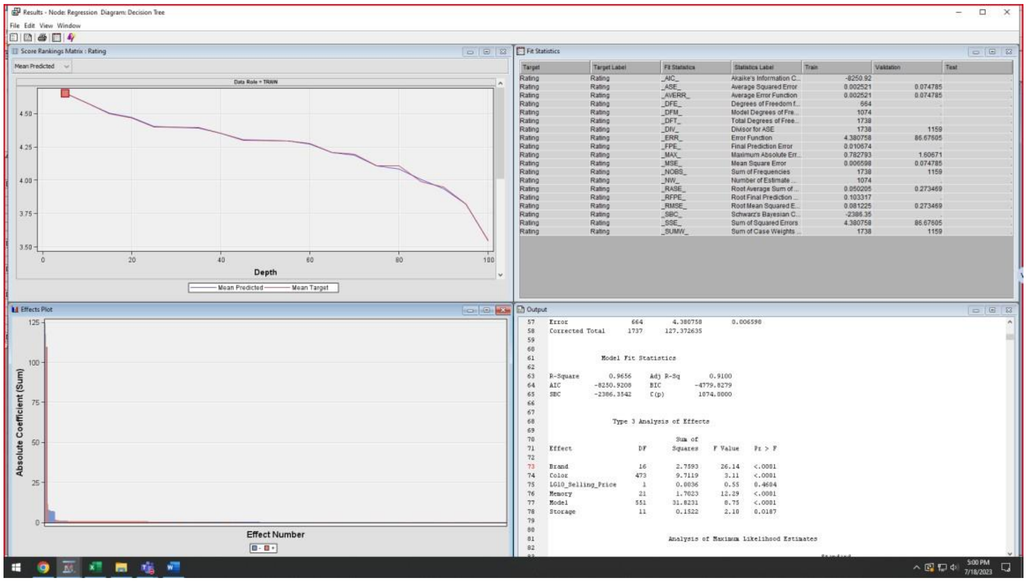
```
Node Id:        1
Statistic   Train  Validation
 Average:  4.2457      4.2340
   Count:    1738        1159
```

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating	Rating	_NOBS_	Sum of Frequencies	1738	1159	.
Rating	Rating	_MAX_	Maximum Absolute Err...	2.010448	1.403279	.
Rating	Rating	_SSE_	Sum of Squared Errors	45.78638	35.42197	.
Rating	Rating	_ASE_	Average Squared Error	0.026344	0.030563	.
Rating	Rating	_RASE_	Root Average Squared...	0.162309	0.174821	.
Rating	Rating	_DIV_	Divisor for ASE	1738	1159	.
Rating	Rating	_DFT_	Total Degrees of Free...	1738	.	.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
Brand	Brand	7	1.0000	1.0000
LG10_Selli...	Transforme...	5	0.4914	0.4889
Memory	Memory	2	0.3132	0.1923
Model	Model	8	0.2944	0.2441
Storage	Storage	2	0.2456	0.1773
Color	Color	0	0.0000	0.0000

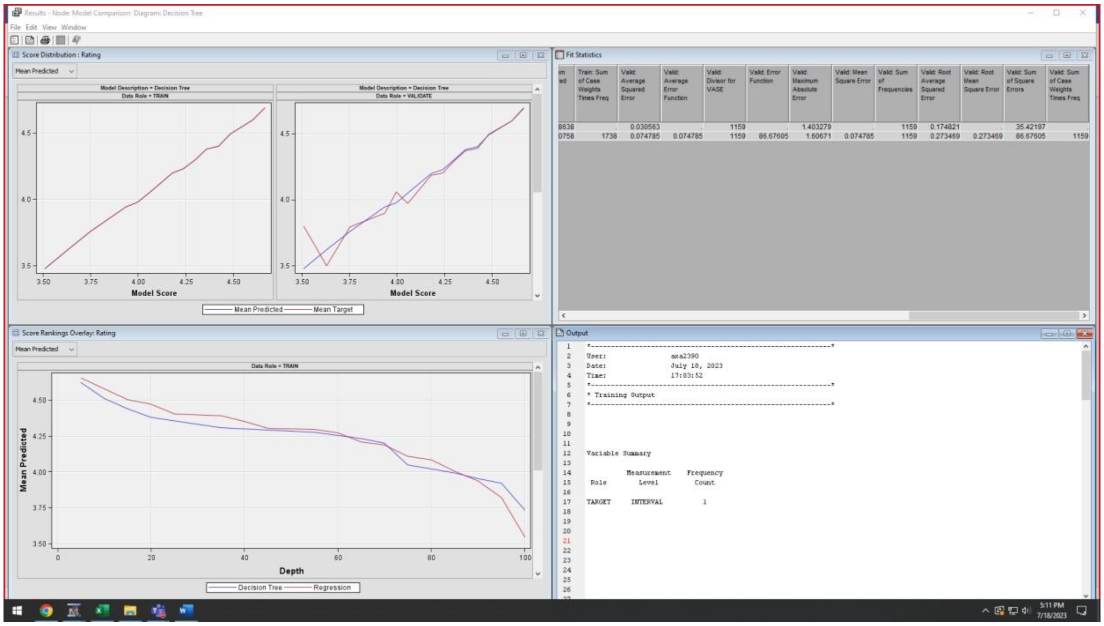
```
PREDICTED     P_Rating     Predicted: Rating
RESIDUAL      R_Rating     Residual: Rating


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*



Variable Importance

```

Regression:-



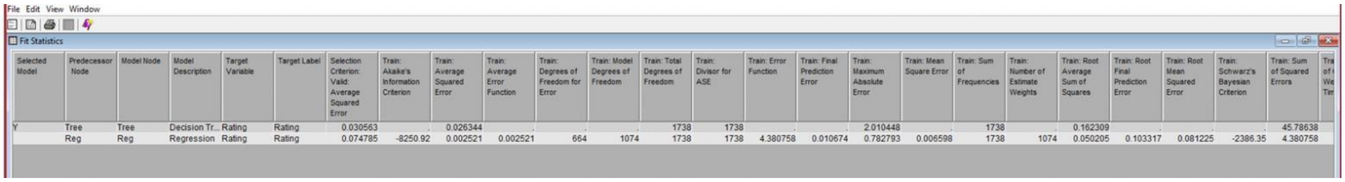
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating	Rating	_AIC_	Akaike's Information C...	-8250.92	.	.
Rating	Rating	_ASE_	Average Squared Error	0.002521	0.074785	.
Rating	Rating	_AVERR_	Average Error Function	0.002521	0.074785	.
Rating	Rating	_DFE_	Degrees of Freedom f...	664	.	.
Rating	Rating	_DFM_	Model Degrees of Fre...	1074	.	.
Rating	Rating	_DFT_	Total Degrees of Free...	1738	.	.
Rating	Rating	_DIV_	Divisor for ASE	1738	1159	.
Rating	Rating	_ERR_	Error Function	4.380758	86.67605	.
Rating	Rating	_FPE_	Final Prediction Error	0.010674	.	.
Rating	Rating	_MAX_	Maximum Absolute Err...	0.782793	1.60671	.
Rating	Rating	_MSE_	Mean Square Error	0.006598	0.074785	.
Rating	Rating	_NOBS_	Sum of Frequencies	1738	1159	.
Rating	Rating	_NW_	Number of Estimate ...	1074	.	.
Rating	Rating	_RASE_	Root Average Sum of ...	0.050205	0.273469	.
Rating	Rating	_RFPE_	Root Final Prediction ...	0.103317	.	.
Rating	Rating	_RMSE_	Root Mean Squared E...	0.081225	0.273469	.
Rating	Rating	_SBC_	Schwarz's Bayesian C...	-2386.35	.	.
Rating	Rating	_SSE_	Sum of Squared Errors	4.380758	86.67605	.
Rating	Rating	_SUMW_	Sum of Case Weights ...	1738	1159	.

```
Error                  664       4.380758        0.006598
Corrected Total       1737     127.372635


              Model Fit Statistics

R-Square         0.9656     Adj R-Sq        0.9100
AIC          -8250.9208     BIC         -4779.8279
SBC          -2386.3542     C(p)         1074.0000


                 Type 3 Analysis of Effects

                                   Sum of
Effect                   DF       Squares    F Value    Pr > F

Brand                    16        2.7593      26.14    <.0001
Color                   473        9.7119       3.11    <.0001
LG10_Selling_Price        1        0.0036       0.55    0.4604
Memory                   21        1.7023      12.29    <.0001
Model                   551       31.8231       8.75    <.0001
Storage                  11        0.1522       2.10    0.0187


                          Analysis of Maximum Likelihood Estimates
```

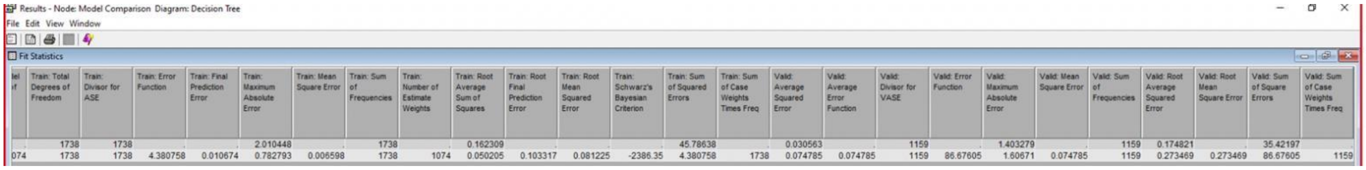


im ed	Train: Sum of Case Weights Times Freq	Valid: Average Squared Error	Valid: Average Error Function	Valid: Divisor for VASE	Valid: Error Function	Valid: Maximum Absolute Error	Valid: Mean Square Error	Valid: Sum of Frequencies	Valid: Root Average Squared Error	Valid: Root Mean Square Error	Valid: Sum of Square Errors	Valid: Sum of Case Weights Times Freq
8638	.	0.030563	.	1159	.	1.403279	.	1159	0.174821	.	35.42197	.
0758	1738	0.074785	0.074785	1159	86.67605	1.60671	0.074785	1159	0.273469	0.273469	86.67605	1159

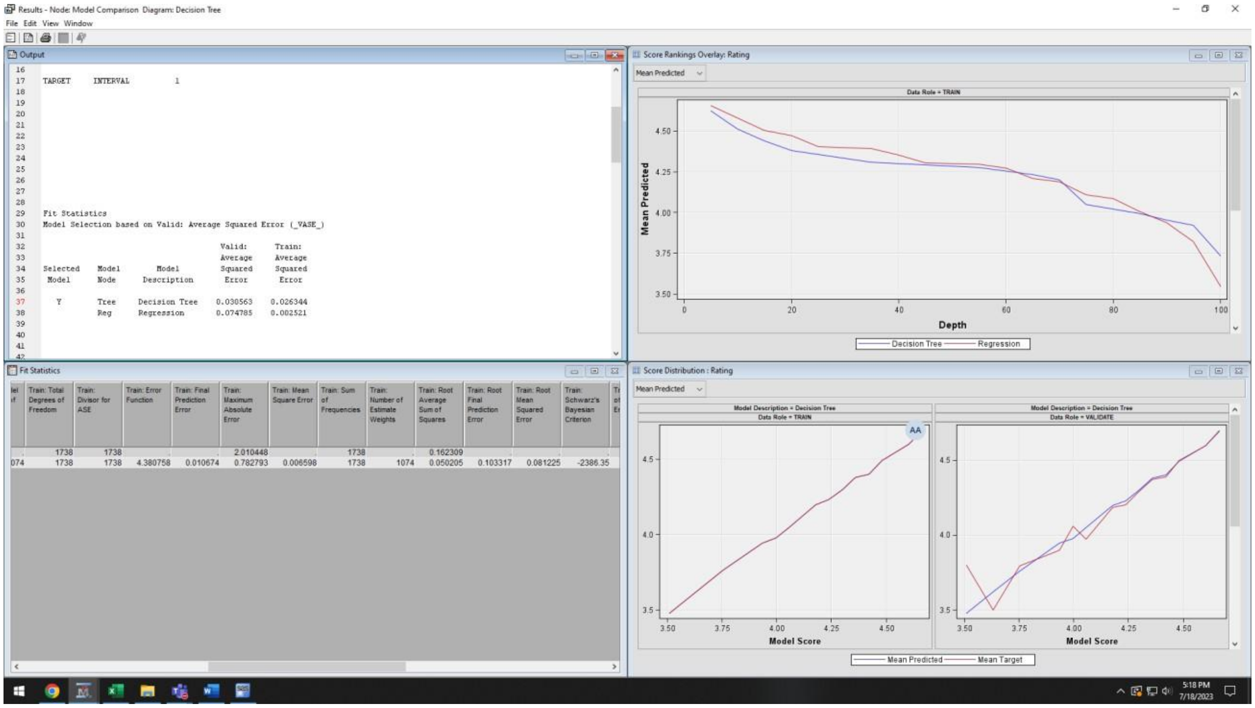
```
*------------------------------------------------------------*
User:                    axa2390
Date:                    July 18, 2023
Time:                    17:03:52
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary


          Measurement    Frequency
 Role        Level         Count

TARGET     INTERVAL          1
```



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Train: Sum of Squared Errors	Tra of We Tim
Y	Tree	Tree	Decision Tr...	Rating	Rating	0.030563	.	0.026344	.	.	.	1738	1738	.	.	2.010448	.	1738	.	0.162309	.	.	.	45.78638	
	Reg	Reg	Regression	Rating	Rating	0.074785	-8250.92	0.002521	0.002521	664	1074	1738	1738	4.380758	0.010674	0.782793	0.006598	1738	1074	0.050205	0.103317	0.081225	-2386.35	4.380758	



lel if	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Train: Sum of Squared Errors	Train: Sum of Case Weights Times Freq	Valid: Average Squared Error	Valid: Average Error Function	Valid: Divisor for VASE	Valid: Error Function	Valid: Maximum Absolute Error	Valid: Mean Square Error	Valid: Sum of Frequencies	Valid: Root Average Squared Error	Valid: Root Mean Square Error	Valid: Sum of Square Errors	Valid: Sum of Case Weights Times Freq
.	1738	1738	.	.	2.010448	.	1738	.	0.162309	.	.	.	45.78638	.	0.030563	.	1159	.	1.403279	.	1159	0.174821	.	35.42197	.
074	1738	1738	4.380758	0.010674	0.782793	0.006598	1738	1074	0.050205	0.103317	0.081225	-2386.35	4.380758	1738	0.074785	0.074785	1159	86.67605	1.60671	0.074785	1159	0.273469	0.273469	86.67605	1159



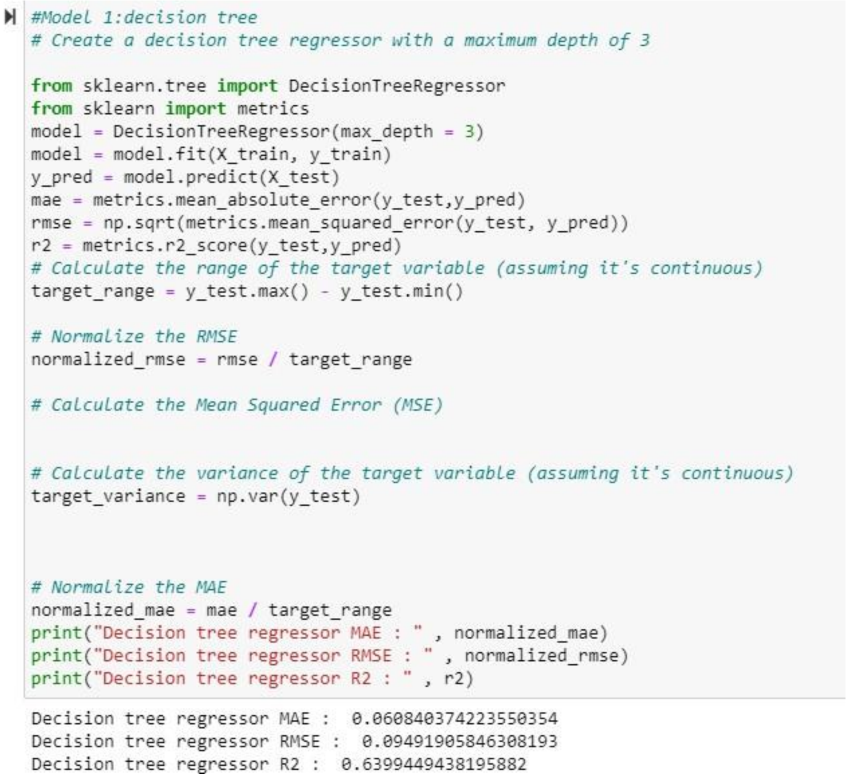
```
TARGET     INTERVAL         1









Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

                                    Valid:      Train:
                                    Average     Average
Selected    Model        Model      Squared     Squared
 Model      Node      Description    Error       Error

   Y        Tree     Decision Tree  0.030563    0.026344
            Reg      Regression     0.074785    0.002521
```

lel if	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Tr of Er
.	1738	1738	.	.	2.010448	.	1738	.	0.162309	.	.	.	
074	1738	1738	4.380758	0.010674	0.782793	0.006598	1738	1074	0.050205	0.103317	0.081225	-2386.35	

Based on the above results we can indicate that **Decision Regression Tree is better choice and have slightly better performance than the regression**. It is due to its R-Square values. The higher the value , the better is the performance. And also, the error value is lower in Decision Regression Tree when compared to the regression model.

Implementation:

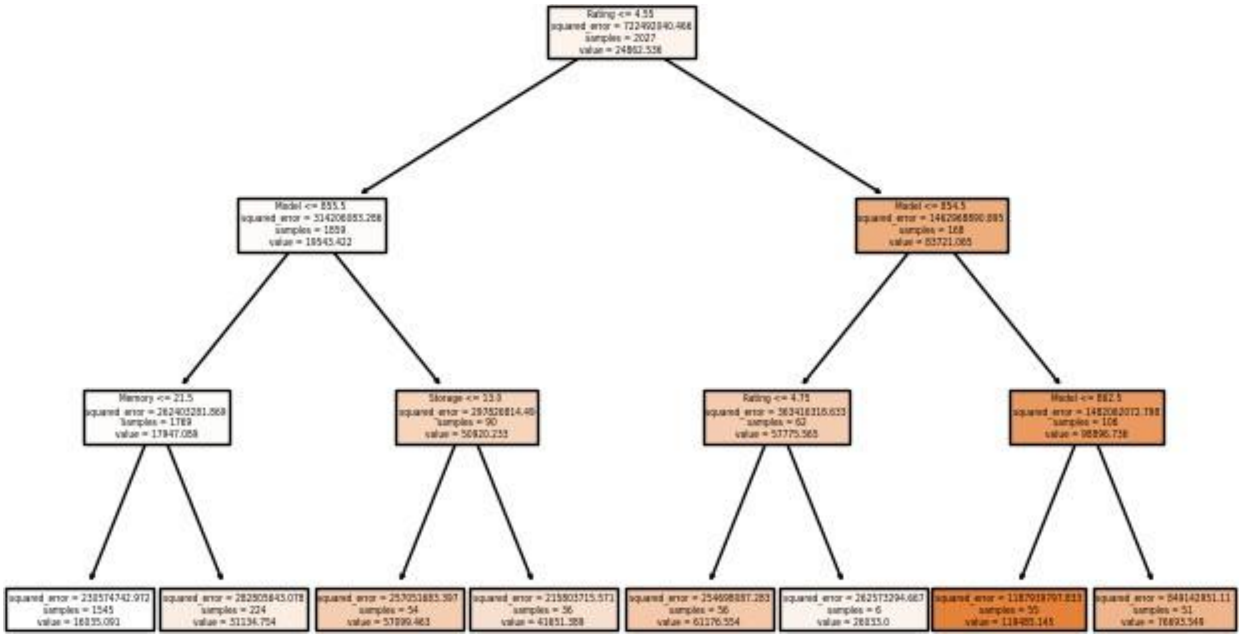
Decision Regression Tree using Selling Price as target variable



```
#Model 1:decision tree
# Create a decision tree regressor with a maximum depth of 3

from sklearn.tree import DecisionTreeRegressor
from sklearn import metrics
model = DecisionTreeRegressor(max_depth = 3)
model = model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mae = metrics.mean_absolute_error(y_test,y_pred)
rmse = np.sqrt(metrics.mean_squared_error(y_test, y_pred))
r2 = metrics.r2_score(y_test,y_pred)
# Calculate the range of the target variable (assuming it's continuous)
target_range = y_test.max() - y_test.min()

# Normalize the RMSE
normalized_rmse = rmse / target_range

# Calculate the Mean Squared Error (MSE)


# Calculate the variance of the target variable (assuming it's continuous)
target_variance = np.var(y_test)



# Normalize the MAE
normalized_mae = mae / target_range
print("Decision tree regressor MAE : " , normalized_mae)
print("Decision tree regressor RMSE : " , normalized_rmse)
print("Decision tree regressor R2 : " , r2)
```

```
Decision tree regressor MAE :  0.060840374223550354
Decision tree regressor RMSE :  0.09491905846308193
Decision tree regressor R2 :  0.6399449438195882
```

According to the regression results, the performance was satisfactory or exceptional, with a Mean Absolute Error (MAE) score of 0.060840374223550354. The model significance is reflected in the Root Mean Squared Error (RMSE) value of 0.09491905846308193 and the R-squared (Coefficient of Determination) value of 0.6399449438195882, indicating that the model is a good fit.



Decision regression trees first split is rating, then model and storage.

As far as the significance of the variables are concerned, from the results of both above models it has been proved that ‘Rating’,‘Model’, ‘storage’ and ‘memory’ are having highest significance.That means these factors are having a significant impact on the target variable Selling Price.

Inferences:
By using Selling price as target variable, we are getting getting lower R-square compared to rating as target variable

Conclusion:

The model evaluation and data visualization revealed insightful understandings into the variables affecting Flipkart sales as well as numerous significant discoveries regarding customer preferences.

Let's compile some important conclusions:

Important variables affecting sales while taking rating as target variable: According to models, the mobile phone BRAND, MODEL, MEMORY, and COLOUR are the most important variables affecting sales at Flipkart. These factors have a significant impact on the Decisions that customers make when making purchases, demonstrating that when buying mobile phones on the platform, customers consider brand reputation, memory capacity, and colour options.

Important variables affecting sales while taking selling price as target variable: According to models, the mobile phone ‘RATING’,‘MODEL’, ‘STORAGE’ AND ‘MEMORY’ are the most important variables affecting sales at Flipkart.

Astonishingly, the study's findings indicated that the selling price had very little bearing on customer ratings. This suggests that customers may favour features, brand loyalty, or customer service over price when evaluating mobile phones. This study highlights the need for businesses to pay attention to factors other than price in order to increase customer satisfaction and ratings.

Top-Selling Variants: According to statistics, customers are most likely to choose the 64GB memory model and the black colour variant. Planning for production and inventory management can both benefit from this information. Flipkart can successfully meet demand and reduce stockouts by offering a greater number of 64GB memory and black colour options.

Demand forecasting and just-in-time inventory can be planned.

This can assist e-commerce companies in managing inventory in accordance with demand and plan.

Planning for target marketing can be done using exploratory data analysis.

References:

- Shaik, M. A., & Verma, D. (2022, May). Predicting present day mobile phone sales using time series based hybrid prediction model. In *AIP Conference Proceedings* (Vol. 2418, No. 1). AIP Publishing.

- Nagaraj, P., Reddy, H. N., Srinivas, S. L., Jaipal, M., & Nagendra, S. V. (2023, March). Machine Learning-based Mobile Application for Store-to-Door using Sentimental Analysis. In *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 138-142). IEEE.

- Kumar, K. P., Dharshini, C. P., Nivedha, V., Santhiya, R., Logeswaran, K., & Ponselvakumar, A. P. FORESEE THE MOBILE PHONE SALES USING HYBRID DATA BASED PREDICTION MODEL.

- Khanna, P., & Sampat, B. (2015). Factors influencing online shopping during Diwali Festival 2014: case study of Flipkart and Amazon. In. *Journal of International Technology and Information Management*, *24*(2), 5.