What features influence the price of the laptop?

Joram Mutenge and Noel Thomas
IS 517

Introduction

There was a time when new technology meant introducing new features on gadgets. Today, technological innovation is not so much about introducing new features as it is about improving on already existing features. For example, the features of laptops have remained the same for years. The focus of change is usually on the size or capacity of those features. What is interesting is that despite this consistency in the types of features present in laptops, the prices of laptops vary immensely. So why are some laptops more expensive than others? We decided to investigate this topic by focusing on the research question: What features have a significant impact on the price of a laptop? To answer this question, we performed multiple linear regression and analyzed the coefficients of variables using the laptop specifications and price dataset obtained from Kaggle. Additionally, we performed cluster analysis and investigated the segments created.

Literature Review

We hypothesized that Random-access memory (RAM), Solid State Drive (SSD), and the Operating System (OS) would have a significant influence on the price of the laptop. This hypothesis is not unfounded. Grindstaff (2022) also mentions RAM (memory) as one of the features that affect computer price along with brand and processor speed. She adds that hard drive space is another essential component, and that pricing is typically influenced by hard drive capacity (Grindstaff, 2022). Quality—which is related to brand—is another factor that impacts the price of a laptop. Analysis of product quality as it relates to the buying decision for Toshiba laptops by Prasetyo and Purwantini (2017) concludes "that the quality of a product has a positive and significant influence on buying decision. The better the quality of a product, the higher the likelihood of buying decision." Therefore, manufacturers who believe that the public perceives their brand as having higher quality are likely to charge more for the laptops.

While the two resources consulted give us an idea of what could influence the price of laptops, they have some limitations. For instance, in her article, Grindstaff did not support her claims about the features that influence computer prices with data. By contrast, the findings in our research are supported by data. Likewise, the started by Prasetyo and Purwantini was limited in that analysis was performed on a single brand, Toshiba. In other words, the research was not as extensive as it could have been. Thus, our analysis will expand on this work by incorporating multiple brands.

Data Cleaning and Preprocessing

The dataset used in our research was obtained from Kaggle, but the original data was scraped from Flipkart, which is an Indian e-commerce store. The dataset consisted of 896 rows and 23 columns. The columns represented features and the rows represented various laptops.

Before analyzing the data, we had to clean and preprocess it. This process involved four steps. To begin with, we selected the relevant columns based on domain knowledge. Out of 23 columns, we selected 12. Secondly, we transformed string data into numerical data. This was extremely important because we were analyzing the data with multiple linear regression, which performs better with numerical data. The third stage involved one-hot encoding categorical values. For example, the column *Touchscreen* had values "yes" and "no." To convert the values to numerical data two columns were created: *touch_Yes* and *touch_No*. If a row represented a touchscreen laptop, 1 was inserted in *touch_Yes* column and 0 in *touch_No* column. The original column *Touchscreen* was deleted from the dataframe. The final step involved dropping columns with more than 25% missing data. Out of the 12 selected columns, *processor_gen* was removed.

To get a deeper understanding of the data, we performed some exploratory data analysis. Thus, we implemented a Pearson correlation coefficient test to see how various columns correlated with price (in Indian Rupees). Fig 1 below shows correlation values for each column.

	processor_gen	ram	ssd	hdd	os_bit	graphic_card_gb	warranty	price	star_rating
processor_gen	1.000000	0.110624	0.226558	-0.143252	0.018421	-0.004779	0.123125	0.108406	-0.017618
ram	0.110624	1.000000	0.396407	-0.180229	0.122495	0.275676	0.087017	0.413835	-0.044368
ssd	0.226558	0.396407	1.000000	-0.579884	0.229685	0.285320	0.202927	0.488710	-0.119142
hdd	-0.143252	-0.180229	-0.579884	1.000000	-0.174198	-0.032831	-0.143631	-0.246925	0.060426
os_bit	0.018421	0.122495	0.229685	-0.174198	1.000000	-0.139860	0.290472	-0.007944	0.043147
graphic_card_gb	-0.004779	0.275676	0.285320	-0.032831	-0.139860	1.000000	-0.042251	0.462580	-0.013381
warranty	0.123125	0.087017	0.202927	-0.143631	0.290472	-0.042251	1.000000	0.051954	0.093033
price	0.108406	0.413835	0.488710	-0.246925	-0.007944	0.462580	0.051954	1.000000	-0.075707
star_rating	-0.017618	-0.044368	-0.119142	0.060426	0.043147	-0.013381	0.093033	-0.075707	1.000000

Fig 1. Pearson correlation plot.

The columns that are more positively correlated with price are *ram*, *ssd*, and *graphic_card_gb*. We also investigated how data was distributed in these three columns. Fig 2 shows bar charts for each of the three columns.

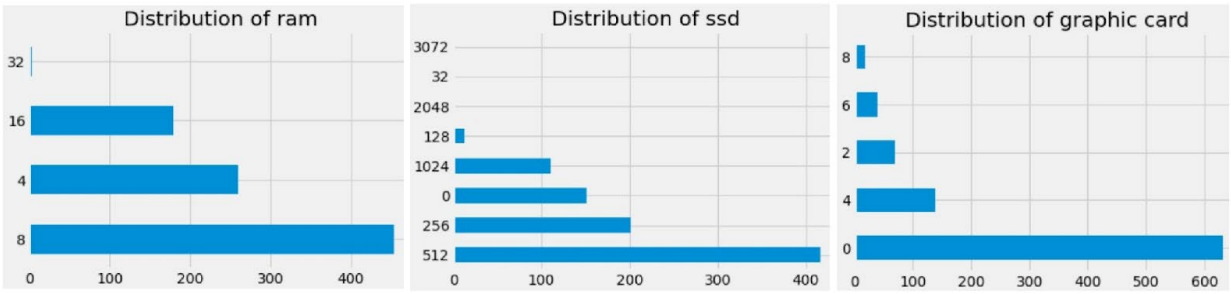


Fig 2. Distribution of data in ram, ssd, and graphic card.

From the bar charts, we can see that more laptops had 8GB memory, and many had 512GB storage space. As for graphic cards, most laptops in the dataset did not have them.

Methods Applied

i) Multiple Linear Regression

Before implementing regression, we split the data using the train_test_split() function from the scikitlearn library in Python. The training set comprised 70% of the data and while the testing set comprised 30% of the data. We used the statsmodels package to execute Multiple Linear Regression. The first implementation produced an adjusted R-squared value of 0.502. This means that only 50% of the variance for the dependent variable (price) is explained by the independent variables used in our model such as *ram*, *ssd*, *warranty*, etc. This is equivalent to a coin toss and therefore not a good model. Fig 3 below shows the results produced by the first execution of the regression model.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.511
Model:	OLS	Adj. R-squared:	0.502
Method:	Least Squares	F-statistic:	58.37
Date:	Fri, 25 Nov 2022	Prob (F-statistic):	5.15e-88
Time:	00:18:12	Log-Likelihood:	-7536.3
No. Observations:	627	AIC:	1.510e+04
Df Residuals:	615	BIC:	1.515e+04
Df Model:	11		
Covariance Type:	nonrobust		

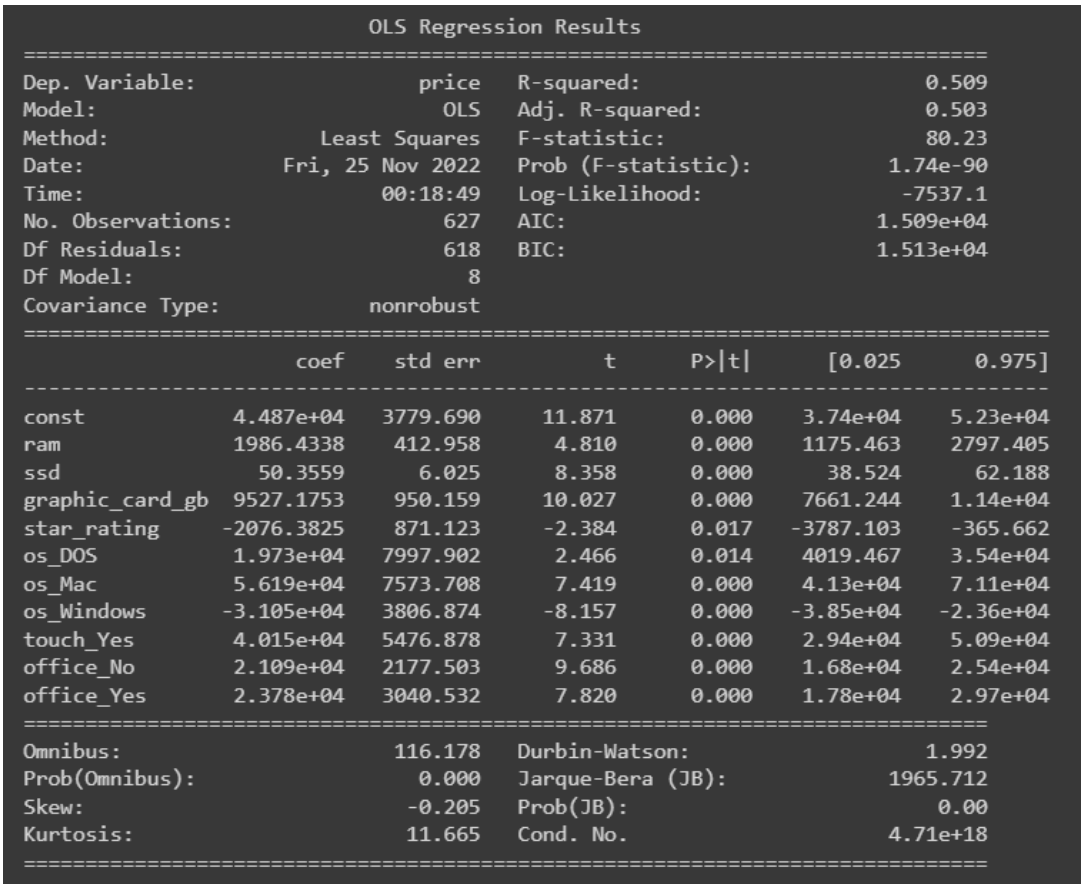
	coef	std err	t	P>|t|	[0.025	0.975]
const	4.527e+04	4772.493	9.486	0.000	3.59e+04	5.46e+04
ram	1962.2219	415.669	4.721	0.000	1145.919	2778.524
ssd	49.5054	7.184	6.891	0.000	35.398	63.613
hdd	-1.9109	4.979	-0.384	0.701	-11.689	7.867
os_bit	11.7902	159.597	0.074	0.941	-301.632	325.212
graphic_card_gb	9690.2915	971.232	9.977	0.000	7782.958	1.16e+04
warranty	-3848.4297	3159.568	-1.218	0.224	-1.01e+04	2356.422
star_rating	-2016.3093	874.752	-2.305	0.021	-3734.172	-298.447
os_DOS	1.791e+04	8125.279	2.205	0.028	1955.713	3.39e+04
os_Mac	5.815e+04	8140.767	7.143	0.000	4.22e+04	7.41e+04
os_Windows	-3.079e+04	4638.948	-6.638	0.000	-3.99e+04	-2.17e+04
touch_No	2127.4673	3163.380	0.673	0.501	-4084.870	8339.804
touch_Yes	4.314e+04	4102.728	10.516	0.000	3.51e+04	5.12e+04
office_No	2.02e+04	2729.383	7.401	0.000	1.48e+04	2.56e+04
office_Yes	2.507e+04	3543.869	7.075	0.000	1.81e+04	3.2e+04

Omnibus:	114.131	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1857.743
Skew:	-0.202	Prob(JB):	0.00
Kurtosis:	11.423	Cond. No.	6.71e+18

Fig 3. Regression results for the first execution.

To improve the performance of the model as measured by the R-squared value, we removed the features with p-values higher than 0.05. The purpose of this was to eliminate

multicollinearity among the features. The features removed were *ssd*, *os_bit*, *warranty*, and *touch_No*. The new model produced an adjusted R-squared value of 0.503, which is a very slight improvement from the original model. Fig 4 shows the results of the second implementation of the model.



OLS Regression Results

Dep. Variable:	price	R-squared:	0.509
Model:	OLS	Adj. R-squared:	0.503
Method:	Least Squares	F-statistic:	80.23
Date:	Fri, 25 Nov 2022	Prob (F-statistic):	1.74e-90
Time:	00:18:49	Log-Likelihood:	-7537.1
No. Observations:	627	AIC:	1.509e+04
Df Residuals:	618	BIC:	1.513e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.487e+04	3779.690	11.871	0.000	3.74e+04	5.23e+04
ram	1986.4338	412.958	4.810	0.000	1175.463	2797.405
ssd	50.3559	6.025	8.358	0.000	38.524	62.188
graphic_card_gb	9527.1753	950.159	10.027	0.000	7661.244	1.14e+04
star_rating	-2076.3825	871.123	-2.384	0.017	-3787.103	-365.662
os_DOS	1.973e+04	7997.902	2.466	0.014	4019.467	3.54e+04
os_Mac	5.619e+04	7573.708	7.419	0.000	4.13e+04	7.11e+04
os_Windows	-3.105e+04	3806.874	-8.157	0.000	-3.85e+04	-2.36e+04
touch_Yes	4.015e+04	5476.878	7.331	0.000	2.94e+04	5.09e+04
office_No	2.109e+04	2177.503	9.686	0.000	1.68e+04	2.54e+04
office_Yes	2.378e+04	3040.532	7.820	0.000	1.78e+04	2.97e+04

Omnibus:	116.178	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1965.712
Skew:	-0.205	Prob(JB):	0.00
Kurtosis:	11.665	Cond. No.	4.71e+18

Fig 4. Regression results for the second execution.

Focusing on the features with higher coefficients in Fig 4, we can conclude that touch screen capabilities of laptops, the presence of Microsoft Office prior to purchasing, as well as the type of Operating System have a significant influence on the price of the laptop.

ii) K-Means Clustering

The second step you know analysis was to perform cluster analysis. The purpose was to see if there were any clusters formed from the data and then investigate what those clusters entailed. To determine the right number of clusters, we used silhouette scores and the scree plot. Fig 5 below shows the plot of silhouette scores and the scree plot.

```
Overall Average Silhouette Score:0.6149 for 2 clusters
Overall Average Silhouette Score:0.5382 for 3 clusters
Overall Average Silhouette Score:0.5154 for 4 clusters
Overall Average Silhouette Score:0.5403 for 5 clusters
Overall Average Silhouette Score:0.5666 for 6 clusters
Overall Average Silhouette Score:0.5744 for 7 clusters
Overall Average Silhouette Score:0.5753 for 8 clusters
Overall Average Silhouette Score:0.5624 for 9 clusters
Overall Average Silhouette Score:0.5635 for 10 clusters
Overall Average Silhouette Score:0.5470 for 11 clusters
Overall Average Silhouette Score:0.5602 for 12 clusters
Overall Average Silhouette Score:0.5633 for 13 clusters
Overall Average Silhouette Score:0.5635 for 14 clusters
Overall Average Silhouette Score:0.5769 for 15 clusters
Overall Average Silhouette Score:0.5699 for 16 clusters
```

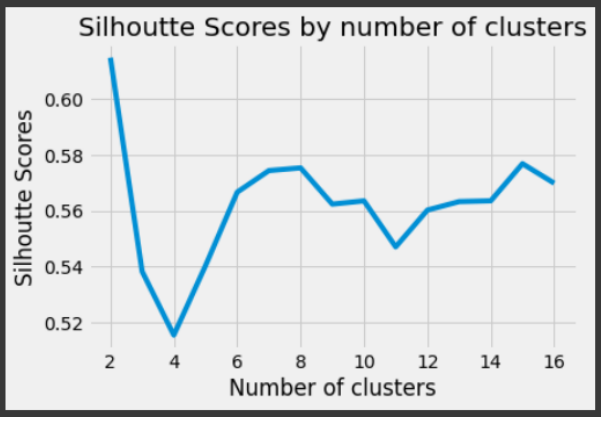
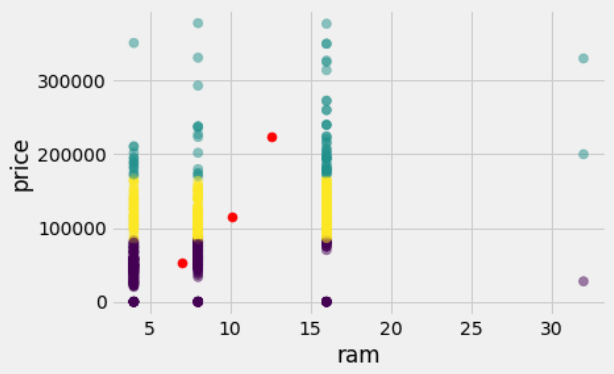


Fig 5. Silhouette scores plot and scree plot.

The highest silhouette scores were achieved by 2,7,8, and 6 clusters. In the scree plot, the elbow bend comes somewhere between 2 and 6 clusters. Using the information gained from both these plots, we decided to select three clusters. This is what the three clusters look like with respect to laptop memory (*ram*) and graphic card size (*graphic_card_gb*).



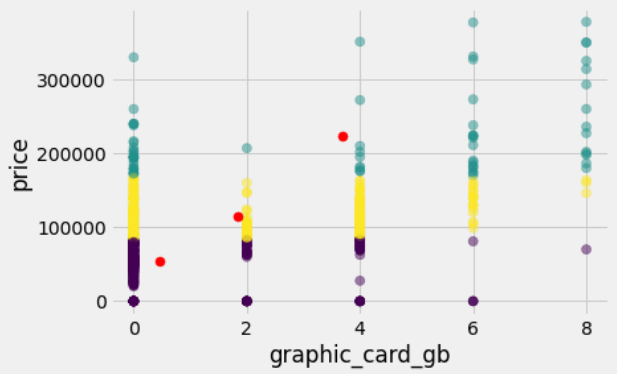


Fig 6. Congregation of clusters on plots of *price* vs *ram*, and *price* vs *graphic_card_gb*
Note: Prices are in Indian Rupees.

The three clusters are represented by red dots and looking at their locations on the two plots, they may represent the three laptop price categories namely low-priced, mid-priced and expensive.

Conclusion

There are many factors that may influence the price of laptops. Our goal was to single out features that significantly influenced the price of laptops. Based on our analysis of the laptop dataset used, our findings reveal that the features that greatly influence laptop prices are the touchscreen capabilities of the laptop, the presence of Microsoft Office prior to purchase, and the type of operating system installed on the laptop. Although the results are different from our hypothesis, some of them are in accordance with our experience. For example, when I was recently purchasing a laptop, I accepted the offer to have Microsoft Office installed. However, when I learned that the price would be increased, I declined the offer. Likewise, touchscreen laptops tend to be a little bit more expensive than those that are not.

Investigating the three clusters created through K-means clustering shows that the laptops in our dataset can be grouped into three categories based on price namely low-priced, mid-priced, and expensive laptops. Hence, these clusters may represent the budget needed to purchase laptops in those respective clusters.

References

Grindstaff, S. (2022, October 30). *What factors affect desktop computer prices? (with pictures)*. Easy Tech Junkie. https://www.easytechjunkie.com/what-factors-affect-desktop-computer-prices.htm

Kumar, S. (2022). *Laptop specs and latest price*. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/kuchhbhi/latest-laptop-price-list

Prasetyo, E. T., & Purwantini, S. (2017). An Influence Analysis of Product Quality, Brand Image, and Price on the Decision to Buy Toshiba Laptop. *Economics & Business Solutions Journal*, *1*(2).