What features influence laptop price?

Joram Mutenge, Noel Thomas

objectives

- Our objective:
  - Use Multiple Linear regression to determine which features greatly influence laptop price.
  - Use Clustering to learn about the various segments of laptops in the dataset.

brand	model	processor_brand	processor_name	processor_gnrtn	ram_gb	ram_type	ssd	hdd	os
Lenovo	A6-9225	AMD	A6-9225 Processor	10th	4 GB GB	DDR4	0 GB	1024 GB	Windows
Lenovo	Ideapad	AMD	APU Dual	10th	4 GB GB	DDR4	0 GB	512 GB	Windows
Avita	PURA	AMD	APU Dual	10th	4 GB GB	DDR4	128 GB	0 GB	Windows
Avita	PURA	AMD	APU Dual	10th	4 GB GB	DDR4	128 GB	0 GB	Windows

data cleaning and preprocessing

- Step 1. Selecting columns based on domain knowledge.

  Columns = [processor_gen, ram, ssd, hdd, os, os_bit, graphic_card_gb, warranty, Touchscreen, msoffice, price, star_rating]

- Step 2. Transforming object columns to numeric.
  - processor_gen : original value 10th. Changed to 10
  - ram , ssd, and hdd : original value 8GB. Changed to 8

- Step 3. One- hot encoding.
  - Columns: os, Touchscreen, and msoffice

- Step 4. Dropping columns with > 25% missing data.
  - processor_gen.

investigating features

- How do features correlate with price?

	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

multiple linear regression

- Data Split: Train 70%, Test 30%
- Performed MLR using the statsmodels package
- Received an Adjusted R-squared value of 0.502
- Analyzed P- values of the independent variables
- Variables with P - values larger than 0.05 were removed, and the model was run again

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		

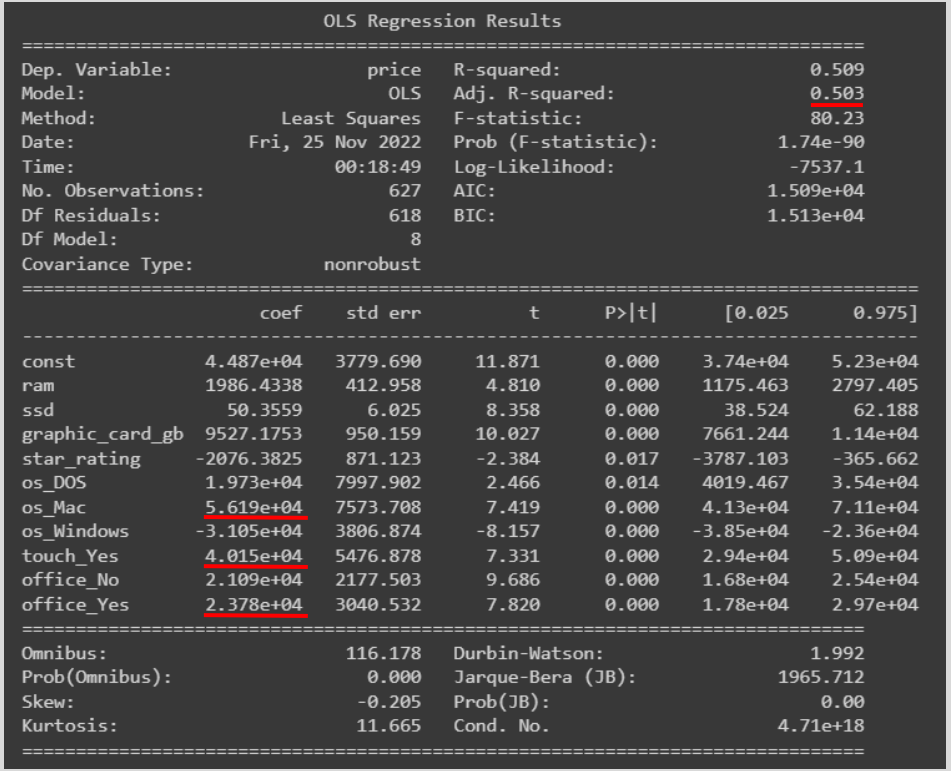
	coef	std err	t	P>|t|	[0.025	0.975]
const	4.527e+04	4772.493	9.486	0.000	3.59e+04	5.46e+04
ram	1962.2219	415.669	4.721	0.000	1145.919	2778.524
ssd	49.5054	7.184	6.891	0.000	35.398	63.613
hdd	-1.9109	4.979	-0.384	0.701	-11.689	7.867
os_bit	11.7902	159.597	0.074	0.941	-301.632	325.212
graphic_card_gb	9690.2915	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

multiple linear regression

- The new model received a similar Adjusted R - squared value of 0.503

- The independent variables with the highest coefficients were the:
  - touch screen capabilities of the laptop
  - whether it had Microsoft Office installed prior to purchase, and
  - Operating system



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

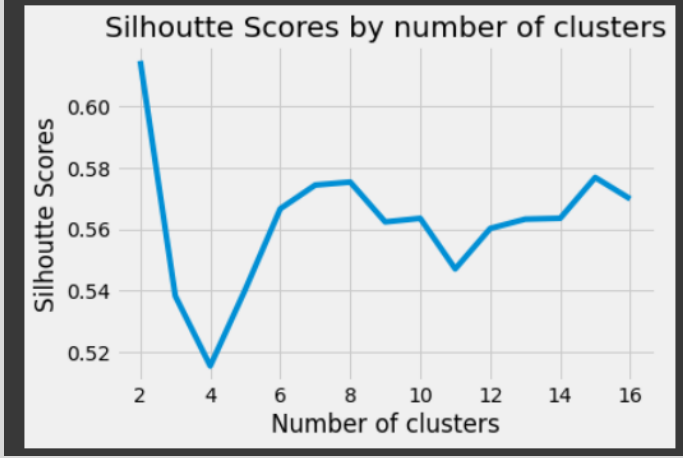
Omnibus:	116.178	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1965.712
Skew:	-0.205	Prob(JB):	0.00
Kurtosis:	11.665	Cond. No.	4.71e+18

clustering

- To calculate the right number of clusters, the silhouette scores were plotted.

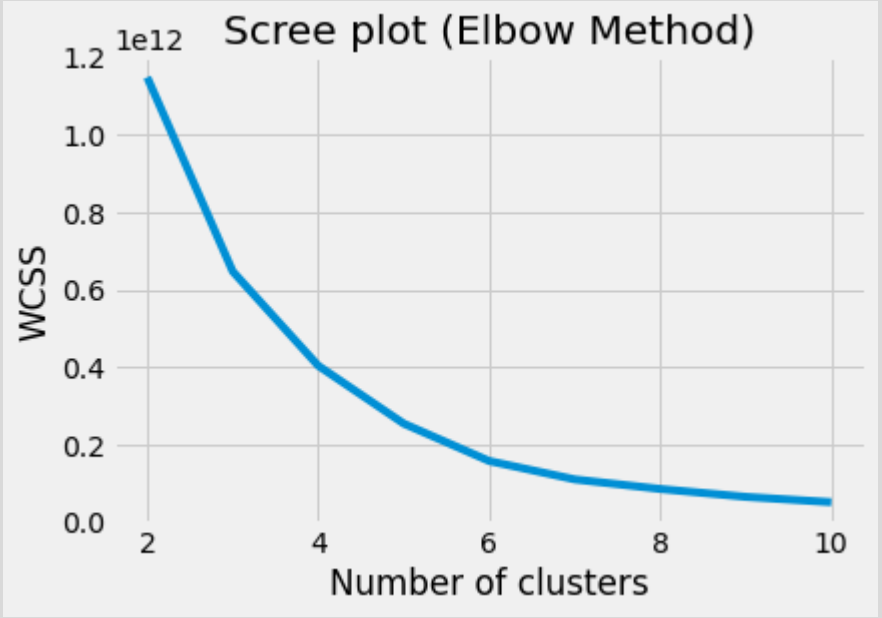
- The highest scores were achieved by 2,8,7 and 6 clusters

```
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
```



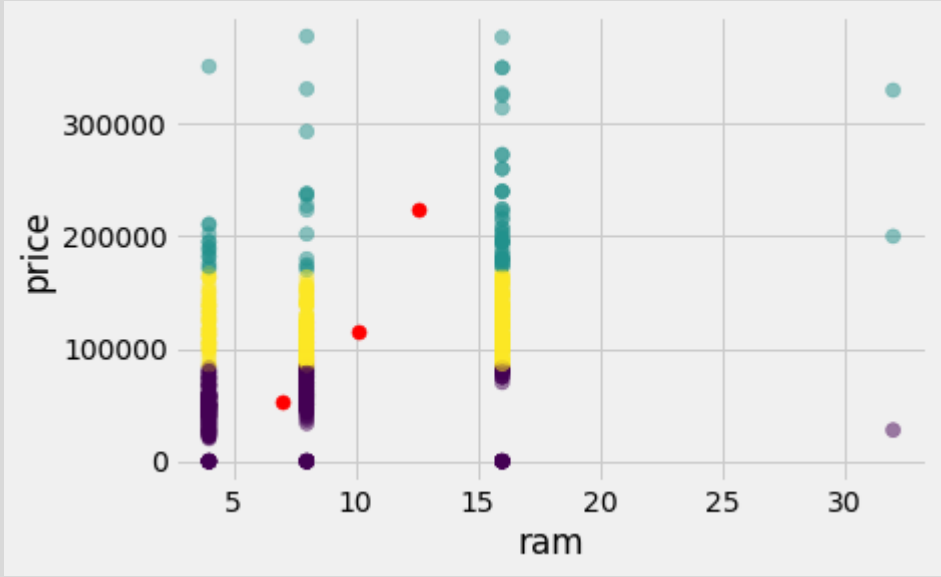
clustering

- The Scree plot was also plotted to help us choose the right number of clusters
- We can see that values between 2 and 6 are ideal
- Using what we learned from both these plots, 2,3 or 6 would be the ideal number of clusters
- We decided to select 3 clusters



clustering

- The three clusters could represent cheap, mid - priced and expensive laptops



limitations and future scope

- Limitations
  - The quality of data in the dataset: too many missing values.
  - Most columns were string, which is not ideal for regression.

- Future Scope
  - Perform classification on the dataset into the clusters created using clustering.
  - Grabbing data from multiple e - commerce websites.

conclusion

- The touch screen capabilities of the laptop, presence of Microsoft Office prior to purchase, and the operating system are the variables that influence the price of a laptop the most.

- The laptops in the dataset can be grouped into three clusters depending on the price.