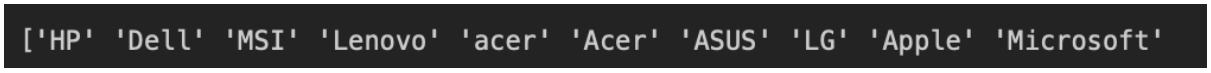
Problem 1

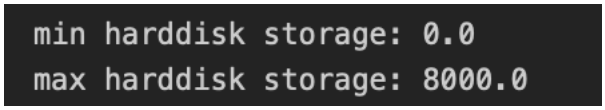
Column name	Situation of column	Cleaning actions/steps	Justification/explanation
price	Empty columns	If empty, remove the whole row	The rest of the information is useless if price is missing
brand[1]	Spellings of common brands inconsistent	I used (import) fuzzywuzzy to correct spellings	Spelling should be consistent.
price, screen_size, ram	columns were strings	Removed units and placed them in the column name. Then, converted into floats.	If converted into floats, the data is easier to manipulate for analysis.
ram	Some cells units in MB	Removed all rows with cells that have units in MB	Generally, RAM measured in GB, so if the units are MB, it must me an anomaly.
model	Some cells empty	Remove the whole row if model is unknown	The buyer can't purchase a laptop if model unknown.
harddisk[2]	Data is strings	Remove the units and add GB to the title. Then convert cells in TB to GB.	If converted into floats, the data is easier to manipulate for numerical analysis later.
harddisk[2]	Anomalies are present	Remove all rows when hard disk storage <50GB.	Anomalous results are likely a mistake.
cpu_speed	Data is strings	Remove units and add GHz to the title and convert cells in MHZ to GHz.	If converted into floats, the data is easier to manipulate for numerical analysis later.
colour	Inconsistent column capitalisation	Made whole column lower case	Makes the data less inconsistent

```
Shape before removing duplicates: (4441, 14)
Shape after removing duplicates: (2623, 14)
```

Img0.1- a screenshot of shape of data, before and after removing duplicates. So, I dropped duplicate rows to avoid repetition and redundancy.



Img1 – a screenshot of the first 10 brand names.

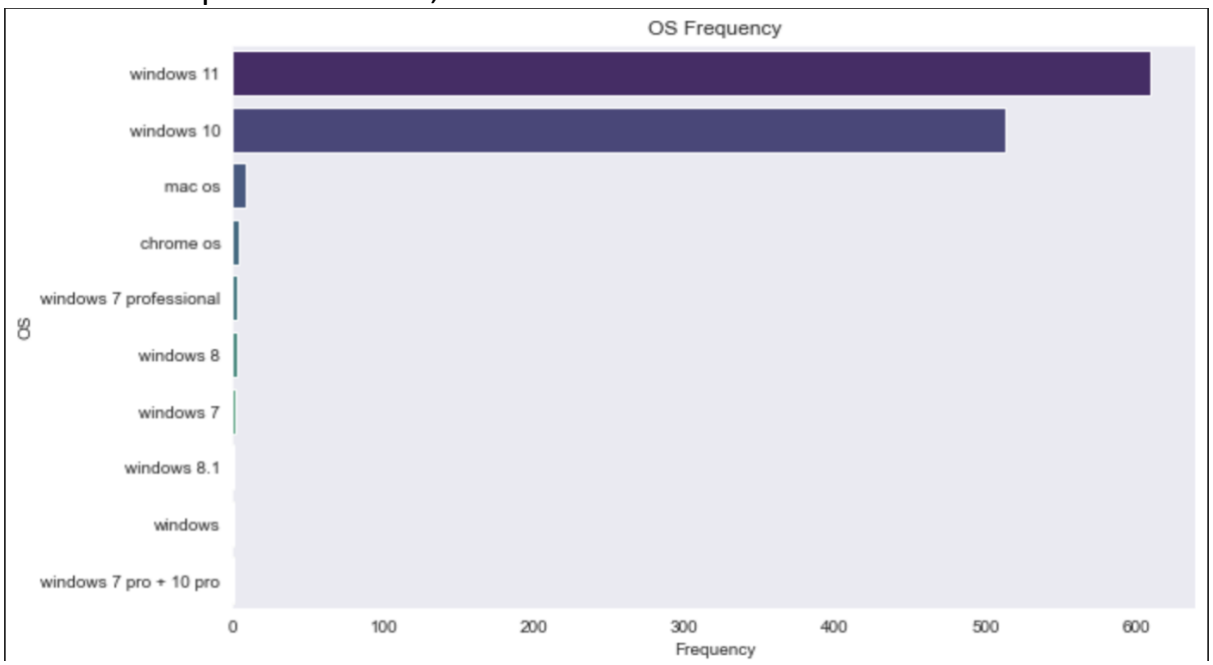


Img2- min/max storage

Note: I have changed the names of the column attribute names, mostly to include units.

Problem 2

The two types of customers I have chosen are: student and gamer. For both, they won't need the laptops with low rating, so I will filter out rows when *rating<3.* Furthermore, I have removed all rows when the OS system is an old version of windows (below 10), since for both customers that will be the minimum requirement. Also, I have allowed Mac OS and Chrome OS.



Img3: screenshot of seaborn graph, shows which OS systems are present, so I was able to remove all old versions of windows without removing a different OS system that was valid.

Student

Students generally have a low budget and require moderate hard disk storage and RAM. Also, I don't want to consider laptops with a rating below 3 but I will allow laptops without ratings. I have assigned the following requirements:

- RAM(GB)>=8,
- harddisk(GB)>=200,
- rating>=3|rating=na.

As, for price, the recommendations will aim to have price low, however, the minimum specs will be the priority in the algorithm.

Gamer

Gamers will generally want a high hard disk storage, RAM, and CPU speed, while budget is often less important (so long as it is under $1500), so the weighting for price will be lower. The same reasoning for students rating applies here too. Furthermore, screen_size will be important too, but the other

parameters will be more important. So, I will have the following minimum requirements:

- RAM(GB)>=8
- harddisk(GB)>=500
- cpu_speed(GHz)>=3
- rating>=3 | rating=na.
- screen_size(inches)>=14

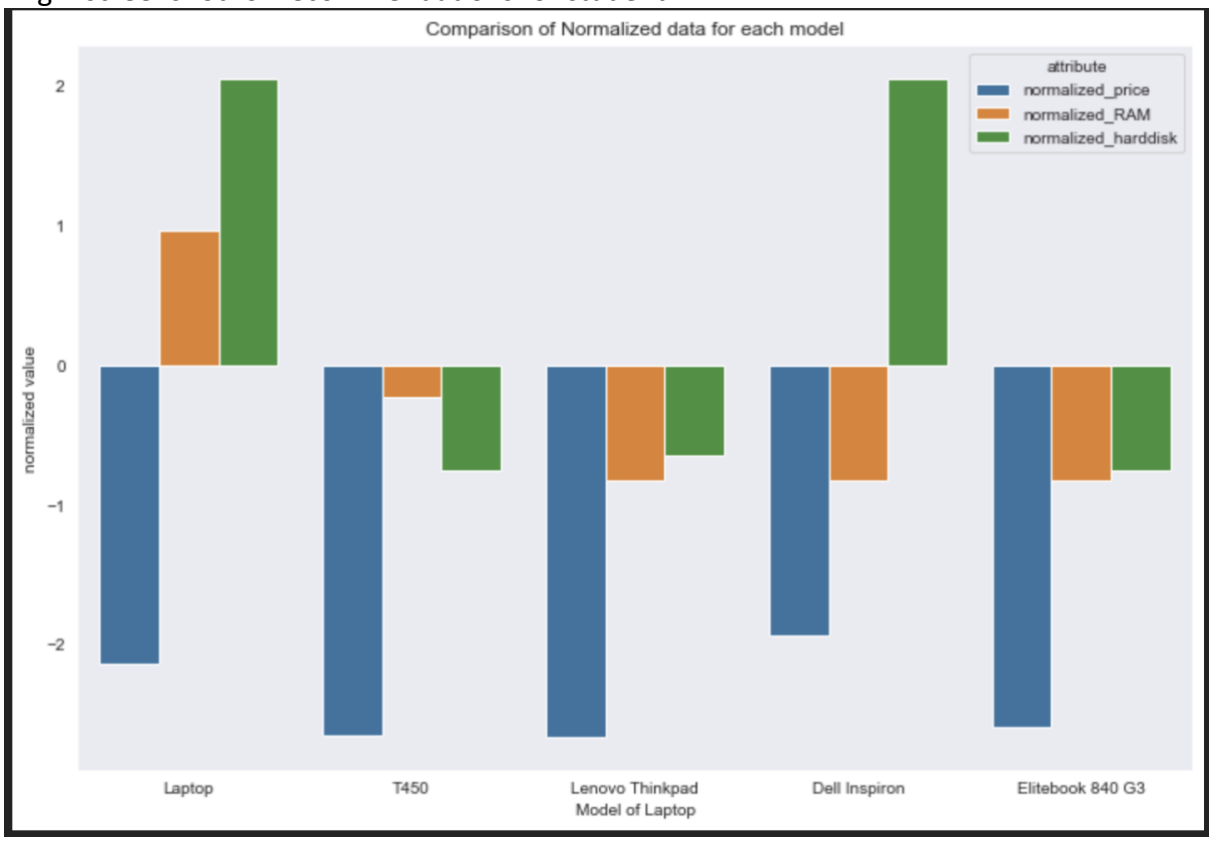
Recommendation Algorithm

The recommendation algorithm first normalises the numerical data and works out the z-score of the data. Then it calculates a ranking_score using this data and multiplying by the weights (which I assigned).

Recommendations for student:

brand	model	price($)
HP	Laptop	298.70
Lenovo	T450	139.98
Lenovo	Lenovo Thinkpad	138.00
Dell	Dell Inspiron	356.99
HP	Elitebook 840 G3	159.99

Img4: screenshot for recommendations for student

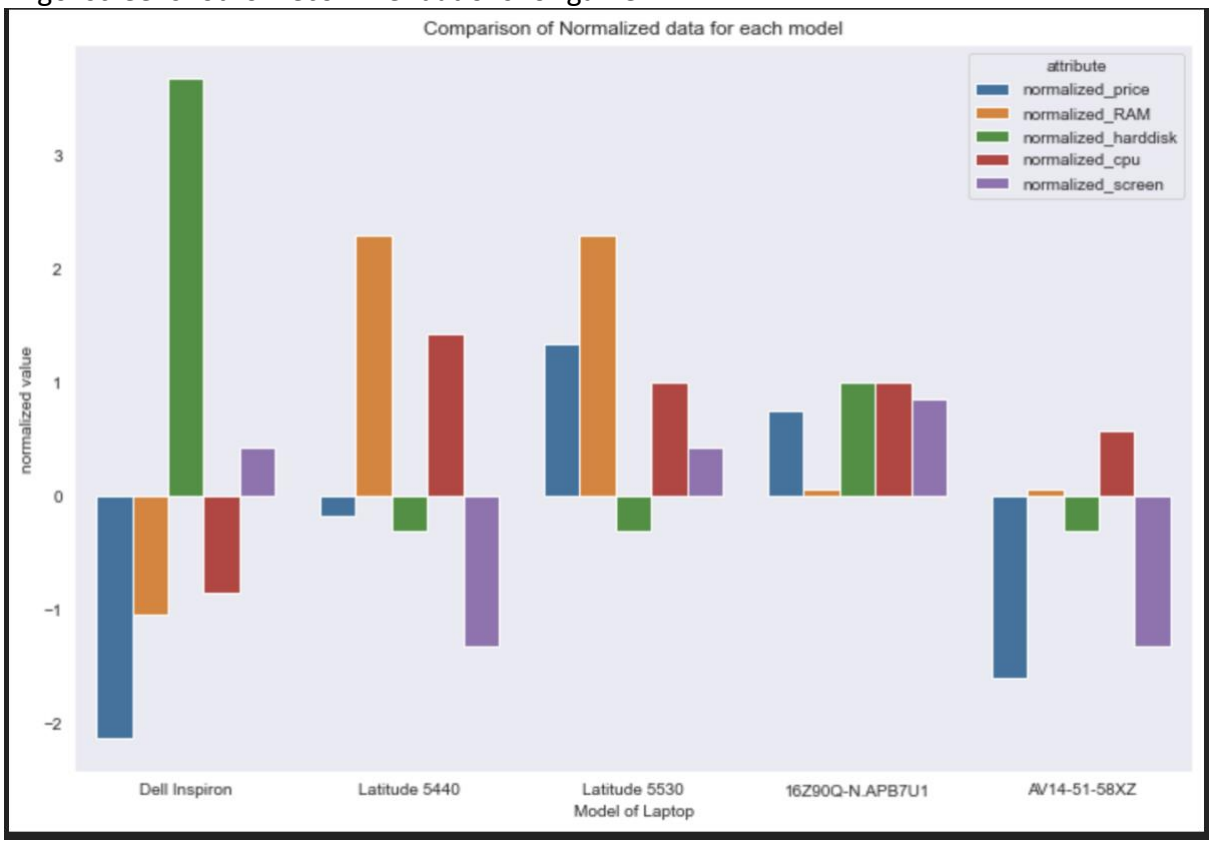


Img5: graph to compare normalized price ram and harddisk of the 5 recommended models. The price bar should ideally be a large negative number, and the ram and harddrive should be as large as possible (for ideal laptop).

Recommendations for gamer:

brand	model	price($)
Dell	Dell Inspiron	356.99
Dell	Latitude 5440	997.74
Dell	Latitude 5530	1488.95
LG	16Z90Q-N.APB7U1	1299.99
Acer	AV14-51-58XZ	532.49

Img6: screenshot for recommendations for gamer



img7: graph to compare normalized data for gamer's laptop recommendations. The price column should ideally be as negative as possible while the other columns should be as large as possible, for the ideal laptop.

References

Reference style used is the APA 7th edition.

[1] *barciewicz. (2019, March 11). Find a best fuzzy match for a string. Code Review Stack Exchange. https://codereview.stackexchange.com/questions/215174/find-a-best-fuzzy-match-for-a-string*

[2] Majumder, P. (2021, June 13). *FuzzyWuzzy Python Library*. Kaggle.com; Kaggle. https://www.kaggle.com/code/prateekmaj21/fuzzywuzzy-python-library

[3] *Saturn Cloud. (2023, June 19). How to Convert a Column in Pandas DataFrame from String to Float | Saturn Cloud Blog. Saturncloud.io. https://saturncloud.io/blog/how-to-convert-a-column-in-pandas-dataframe-from-string-to-float/*

[4] *Python isinstance(). (n.d.). Www.programiz.com. https://www.programiz.com/python-programming/methods/built-in/isinstance*

[5] Shubham. (2022, December 2). Replace column values based on conditions in Pandas - thisPointer. ThisPointer. https://thispointer.com/replace-column-values-based-on-conditions-in-pandas/

[6] Mcleod, S. (2022, December 2). Replace column values based on conditions in Pandas - thisPointer. SimplyPsychology; Olivia Guy-Evans. https://thispointer.com/replace-column-values-based-on-conditions-in-pandas/