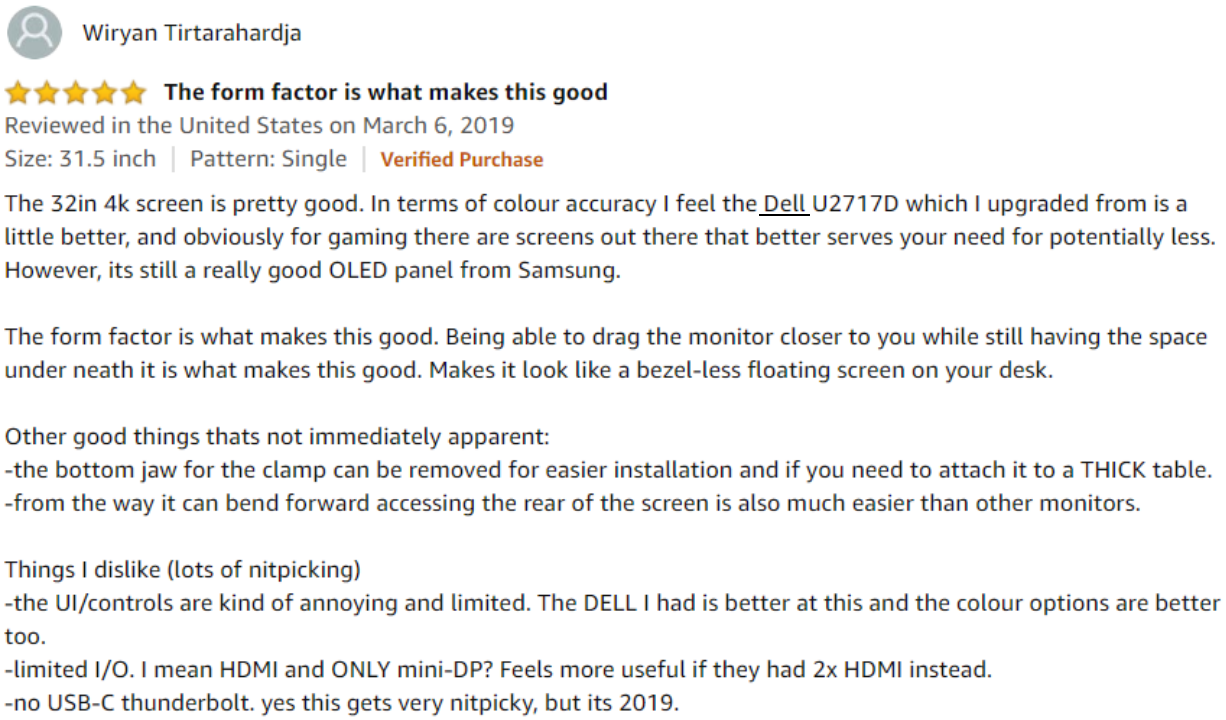
Group Assignment 2: Building a perceptual map based on Amazon.com reviews

Select an Amazon.com product category (e.g., Appliance). Collect 20 ~ 30 brands in the category. You can do it manually. If there are more than 30 brands in the category, either work in a sub-category (e.g., TVs in Appliance), or use the top 30 brands (in terms of average of sales rank), or applying both criteria to further reduce the number of brands to 20 ~ 30.

Analyze the product reviews in the category. The purpose of this text mining is to identify other brand name(s) that is mentioned in a focal brand's review text (co-occurrence). Below is a co-occurrence example of a Samsung monitor review:



Wiryan Tirtarahardja

★★★★★ **The form factor is what makes this good**

Reviewed in the United States on March 6, 2019

Size: 31.5 inch | Pattern: Single | **Verified Purchase**

The 32in 4k screen is pretty good. In terms of colour accuracy I feel the Dell U2717D which I upgraded from is a little better, and obviously for gaming there are screens out there that better serves your need for potentially less. However, its still a really good OLED panel from Samsung.

The form factor is what makes this good. Being able to drag the monitor closer to you while still having the space under neath it is what makes this good. Makes it look like a bezel-less floating screen on your desk.

Other good things thats not immediately apparent:
-the bottom jaw for the clamp can be removed for easier installation and if you need to attach it to a THICK table.
-from the way it can bend forward accessing the rear of the screen is also much easier than other monitors.

Things I dislike (lots of nitpicking)
-the UI/controls are kind of annoying and limited. The DELL I had is better at this and the colour options are better too.
-limited I/O. I mean HDMI and ONLY mini-DP? Feels more useful if they had 2x HDMI instead.
-no USB-C thunderbolt. yes this gets very nitpicky, but its 2019.

Note in this example, the occurrence of Samsung is implicit in the sense that the name is not mentioned. But it is a review on a Samsung product in that product's page.

Define *lift* as the ratio of the actual co-occurrence of two terms to the frequency with which we would expect to see them together. The *lift* between terms *A* and *B* can be calculated as:

$$Lift(A,B) = \frac{P(A,B)}{P(A) \times P(B)}$$

where P(X) is the probability of occurrence of term X in a given review, and P(X, Y) is the probability that both X and Y appear in a given review (one of them is the focal brand so its appearance could be implicit).

Calculate and use *lift* as a proxy for similarity between two brands. Construct a similarity matrix of the brands based on the resulting lift values. Submit your similarity matrix in a separate excel file (1 pts.) Why using *lift*, instead of the simple co-occurrence, to proxy brand similarity? (2 pts.)

Employ Multidimensional Scaling (MDS) to build a perceptual map of the selected brands. Put your resulting plot on a separate word document. Identify the two underling dimensions and give an interpretation. Do they make sense to you? (2 pts.)

All the codes have been done using R. Submit all your source codes. Code will be graded on correctness and cleanness. (5pts.) Codes that cannot be run will not be graded and get zero; codes that do not include sufficient AND MEANINGFUL comments will not be graded and get zero.

This is an open-ended question so no certain answer. It may require your team to explore more than one product category in the case that your initial product category is not good (e.g., all of the brands in a category are very close to each other. The MDS program will bomb in the case.)

There are many opportunities to earn bonus points in this assignment. Below are just a few examples:

- Instead of manually collecting brand names, your team can apply a supervised learning approach to obtain the list of brands in a category semi-automatically.
- Work on the model level instead of the brand level.
- While co-occurrence is a pretty good proxy of similarity, the similarity measure could be improved by adding sentimatic analysis, e.g., A review such as “Do not buy Brand A. Go for Brand B” certainly suggests dissimilarity rather than similarity.

As such, I would like to put up to 5 pts. bonus points to this assignment.