Analyze The Effect Of Laptop Predictors On Customers' Interest

Pham Thanh Hung
20194437
hung.pt194437@sis.hust.edu.vn

Hoang Nguyen Minh Nhat
20194445
nhat.hnm194445@sis.hust.edu.vn

Tran Quoc Lap
20194443
lap.tq194443@sis.hust.edu.vn

Abstract

This report is a brief summary about the laptop market data exploratory analysis. The purpose is to identify the best predictors of customer interest on laptops. Data is collected and cleaned before exploratory analysis. Further insight of customers' habits and which information draws more interest will be revealed. From marketing insight, we conducted some marketing suggestions and sale strategies.

1. Introduction

Nowadays, laptops are increasingly concerned as a demand for the work, especially in work-from-home situations caused by COVID-19. The purpose of this study was to identify the best predictors of customer interest on laptops - represented by the number of customers' comments, number of ratings and average rating, from multiple related factors such as manufacturer brand, price, release year, RAM size, storage size, etc.

Having a better understanding of factors that are most likely to affect customers' interest will allow us to identify customer buying habits and which factors to focus on in order to draw interest from customers.

Answering this question might help manufacturers to better focus on the right components so that they can optimize their revenue. Besides, findings from this research might provide customers with reliable information when they want to buy a new laptop.

2. Methods

2.1. Scraping Data

Our scraping source of data are 4 online laptop markets:

- FPT shop: fptshop.com.vn/may-tinh-xach-tay
- CellphoneS: cellphones.com.vn/laptop.html
- The gioi di dong: thegioididong.com/laptop
- Dien may xanh: dienmayxanh.com/laptop

We used the python library BeautifulSoup and Selenium Web Driver as the main tools to access the web browser and collect data. Data to collect include main information of the laptop, price, rating, number of comments, technical features and its identity code.

However, during the actual scraping process, there are some products were missing information, so we decided to gather additional details from 1 other sources:

- Laptop arena: https://www.laptoparena.net/

The main data still come from 4 main sources, while the other sources provide the missing values

2.2. Cleaning and integrating data

Cleaning and integrating are probably the two most time-consuming processes in our project. As the information in each website is displayed differently, our raw data at 5 sources have a certain dissimilarity with each other. On top of that, the information at each data source is filled out in a messy way, some attributes do not follow a common pattern and some special values must be manually handled. In this section, we will give a brief description of each step we did to clean and integrate the data. For full implementation, visit `*Data Cleaning and Integrating.ipynb*`.

Number of records in each raw data sources:

- TgddDmx data source: 181 records
- FPT store data source: 146 records
- CellphoneS store data source: 416 records
- LaptopArena data source: 77464 records

The first step is to deal with inconsistent data sources, each data source has different ways to store the information (see in

scraped_data folder). So firstly, we defined a common data form which is our expectation of the final dataset. This form will define variable names as well as the data type of each variable. Vist Laptop_EDA_data_form for more details.

In the second step, we start the data cleaning process of each source to match the expected data form. The main tool we use is **Pandas library with regular expression** to extract strings that have a common pattern from each variable. For more messy dataset like CellphoneS, we use **Open Refine** to clean the dataset before passing it to Pandas.

After cleaning all dataset to the same format, we start the data integration process: merge all data sources into a single data frame. To avoid data redundancy, we first select some variables in data form as mapping attributes to match similar laptops in different data sources). About variable selection criteria, the values of all mapping attributes must be reliable, non-nullable and representative for a laptop. To not mismatch the data, a set of mapping attributes must be an identifier for each laptop in the market. Our chosen mapping attributes are: *Series*, *Brand* , *cpu brand* , *cpu code* , *ram size*, *storage size*, *displ size* and *displ rate*.

In this part, our most challenging problem is comparing between *Series* of similar laptops. Unlike other variables, which are all laptop specs and fixed among data sources, *Series* in different data sets have different values. For example, a Series of laptop in data source A is

"Gaming Stealth15m a11sdk 061vn"

but in data source B, it may called

"Gaming Stealth 15M A11UEK 254V"

To tackle this problem, we use some string similarity algorithm, like Levenshtein distance, or an algorithm that we customized from Levenshtein distance but give higher weight to Series keywords extracted from the LaptopArena dataset.

Integrating process is still not completed after joining the data source, we have to merge the attributes together. For instance, if a laptop is match in both 3 data source, the integrated variables is computed as follow:

- *Comment count* = Sum of 3 sources
- *Rating count* = Sum of 3 sources
- *Rating* = Average of *Rating* in 3 sources with the weight is *Rating count*
- Price = Average of *Price* in 3 sources
- Other variables are equivalent among sources, so we just simply choose from those

The last step is to reclean the merged data, we once again look at all the attributes in the result to check whether some attributes may contain noisy and redundant values. Unfortunately, the GPU is such a case, so we have to recleaned it based on the GPU values extracted from LaptopArena data sources

After cleaning and integrating, the final data has a nearly the same structure with the expected data form, with 4 more attributes:

- *Availabe_in*: Number of existence in all 3 sources
- Index_tgdd: Index to original TgddDmx store data source
- Index_fpt: Index to original Fpt store data source
- Index_cellphones: Index to original CellphoneS store data source

Number of records in final dataset: 552 records and 34 columns.

Each rows consists of 3 response attributes:

- *Rating count*: the number of ratings for a laptop.
- *Rating*: average rating of a laptop, ranging from 1 to 5 and representing the level of customer satisfaction.
- *Comment count*: the number of comments for a laptop. This attribute represents the crowd attention or popularity of a laptop and is the most reliable main response attribute because the 2 above attributes miss a lot of data.

Others are exploratory attributes, divided into 2 types: numerical attributes and categorical attributes:

- *Price*: the sale price of each laptop in Vietnamese currency (VND).
- *Brand* and *Series*: the name of the laptop's brand and its particular model name.

- *CPU brand* and *CPU code*: Name of CPU's manufacturer and the CPU model number.
- *RAM size* and *Upgradable capability*.
- *Storage size*: storage capacity of laptop's hard drive (GB).
- *Storage type*: hard drive type (SSD, HDD, eMMC)
- *Extra storage slot available*: has extra slot to expand storage capacity or not (Yes, No).
- *Display size* (inch)
- *Display resolution* (for example: 1920x1080)
- *Display refresh rate* (Hz)
- *Screen technology*: panel type (IPS, VA, OLED,..)
- *Weight* (kg).
- *Length, Width, Thickness* (mm)
- *GPU*: main graphics card (integrated or discrete)
- *Battery capacity* (Wh).
- *OS*: Operating system pre-installed on the device.
- Surface material of the case (plastic, aluminum, carbon,...).
- *Release year*: Release year of the laptop.
- *HDMI slot available* (Yes, No).
- *Headphone 3.5mm slot available* (Yes, No).
- *LAN slot available* (Yes, No)
- *Type-C slot available* (Yes, No)
- *USB-A slot available* (Yes, No)
- *SD-card slot available* (Yes, No)
- *Backlit keyboard available* (Yes, No)

2.2 Visualization and statistical tools

In order to conduct experiment and to give an illustrative explanation of our analysis, we use some visual aids such as:

- Barplot and pie plot to compare various numeric variables.
- Boxplot, density and histogram is used to analyze the distribution of numerical attributes and categorical attributes value count.
- Scatter plot to illustrate the distribution and intuitively determine the dependency between pairs of attributes.
- ANOVA test and T-test to compare whether two (or more than two) samples means are significantly different or not.

3. Result of EDA

3.2 Univariate analysis

In this section, we will look at the distribution, mean and variation of some important attributes…..

First, let's look at our 3 most important numerical attributes, Comment count, Rating count and Rating

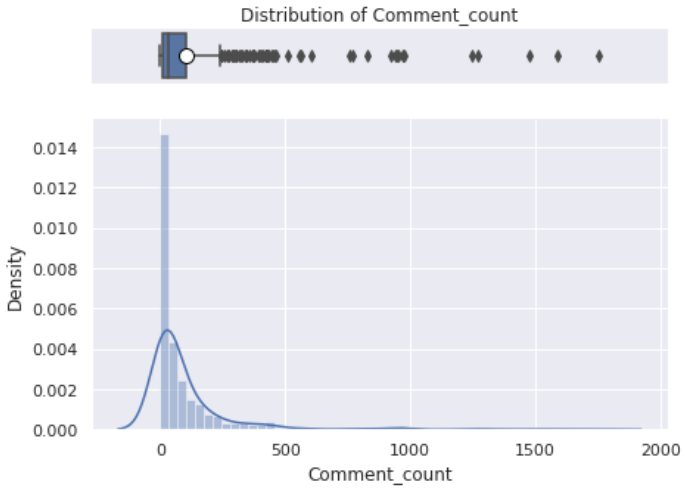

Figure 1: Distribution of Comment count

The distribution has the shape of an exponential distribution, with most of the values between 0 and 104 comments (75% quantile). Although this range is quite small, there are some laptops that have an extremely high number of comments (up to 1756), which is very interesting for further analysis.

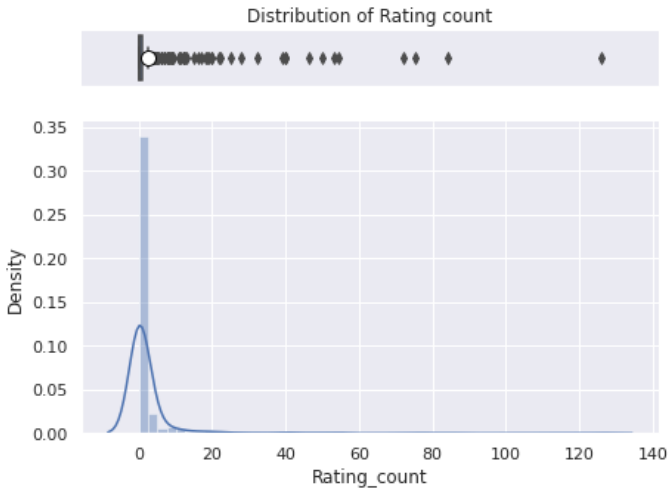

Figure 2: Distribution of Rating Count

The Rating count distribution makes us really confused. The zero value dominates the distribution, with more than 75% interval of 0 to 1. In other words, more than half of the products have not been rated. This will be a challenge for us to use it as a response variable and find its link with other attributes. Besides, as the distribution shape of this variable is very similar to Comment_count, we will examine their relationship in the next section.

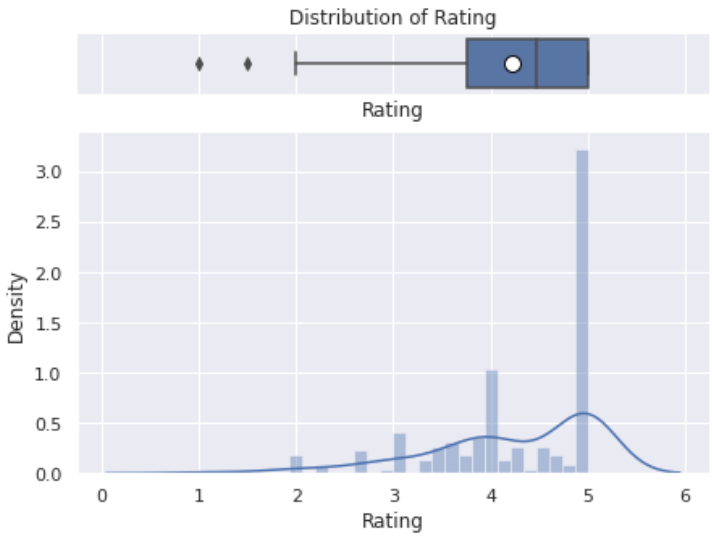

Figure 3: Distribution of Rating

The distribution of Rating has 2 peaks, with a clear mode 5 and another smaller mode 4. Most of the values are ranging between 3.7 and 5 stars, pointing out that the overall Rating in our dataset is very positive. However, the number of null values in Rating is nearly 70%, so we should notice that this distribution of Rating is just based on 30% of products that are rated in the dataset.

Figure 4:.Distribution of Price

The distribution of Price is a bell curve, skewed right, centered around 26.000.000 VND and has standard variation of 12.000.000 VND. The majority of our data focus on mid range laptops, with most common values lies between 18.000.000 to 30.000.000 VND. (~ 1000 - 1500 USD).

The lowest price to afford a laptop from those sites is 7.000.000 VND (~ 350 USD) and the most luxury laptop costs ~ 93.000.000 VND (~ 4500 USD)

For further exploration, we will divide the price to different segments by the prior knowledge about the market:

- Low price range: smaller than 10.000.000 VND
- Low-middle price range: between 10.000.000 and 20.000.000
- Middle price range: between 20.000.000 and 30.000.000
- High-middle price range: between 30.000.000 and 40.000.000
- High price range: Between 40.000.000 and 50.000.000
- Ultra high price range: Higher than 50.000.000

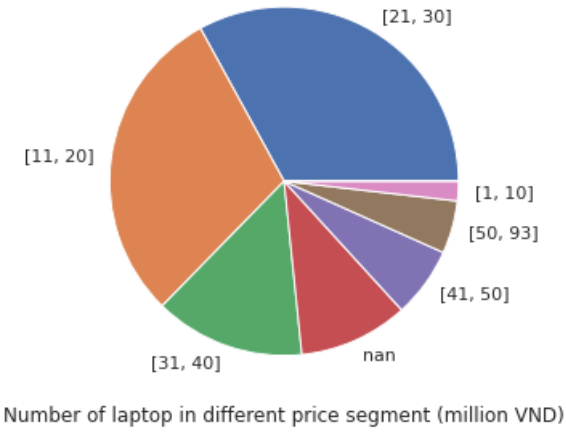

Figure 5: Number of laptop in different price segment (million VND)

Some other numerical attributes (of which distribution are visualized in the `EDA.ipynb`)

- Battery: almost like the distribution of Price (bell curved, skewed right), common range from 42 to 65
- Display rate: discrete numeric, 8 distinct values with highest mode is 60 (Hz)
- Display size: discrete numeric, 19 distinct value with 3 highest mode equal to 13.3, 14 and 15.6
- Ram size: discrete numeric, 5 distinct values with highest mode is 8 (GB)
- Storage size: discrete numeric, 7 distinct values with highest mode is 512 (GB)

For categorical attributes, what is most concerned is Brand, which contains laptops from 13 different manufacturers. According to Figure 6, ASUS occupies the majority of the share, other dominant brands are MSI, Lenovo, HP, Acer. Besides, the percentage of laptops not commented on by each brand is relatively small, so further analysis upon brand is seemingly reliable. The popularity of these manufacturers in reality and in the Figure 6 is easy to raise an initial expectation: ASUS is the most interesting brand and, therefore, there are statistical significant differences in the level of interest between brands.

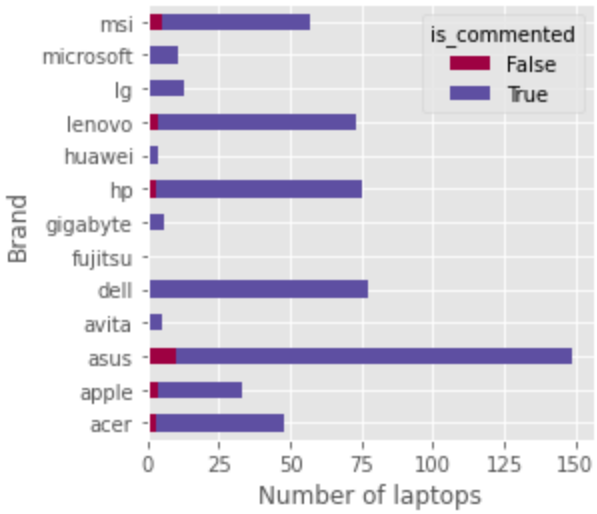

Figure 6: Stacked bar plot representing the number of laptops by brand. Red bar denotes laptops without comments, purple bar denotes laptops with comments.

3.3 Bivariate analysis

Our first concern focuses on numeric attributes. Because the number of numerical attributes is quite large, we would first look at the Pearson and Spearman correlation matrix to get an overview of the correlation between these variables.

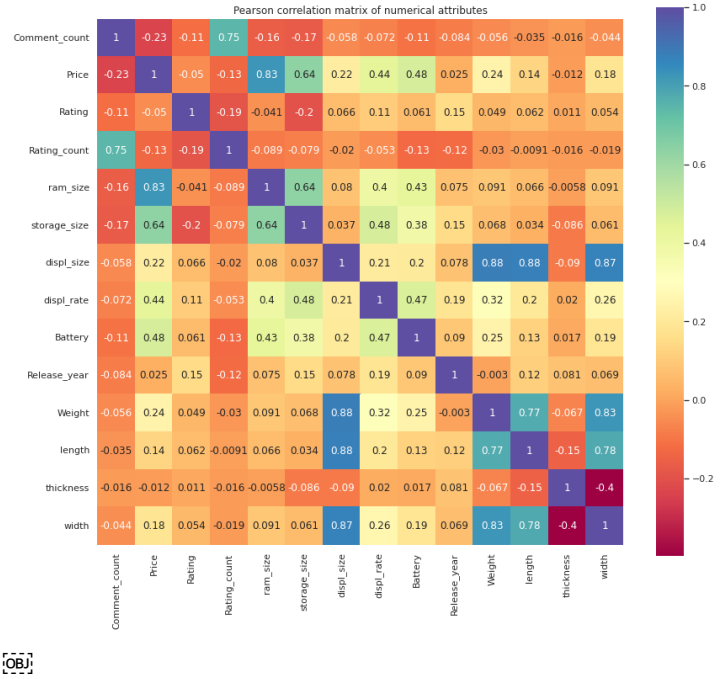

Figure 7: Pearson correlation matrix of numerical attributes

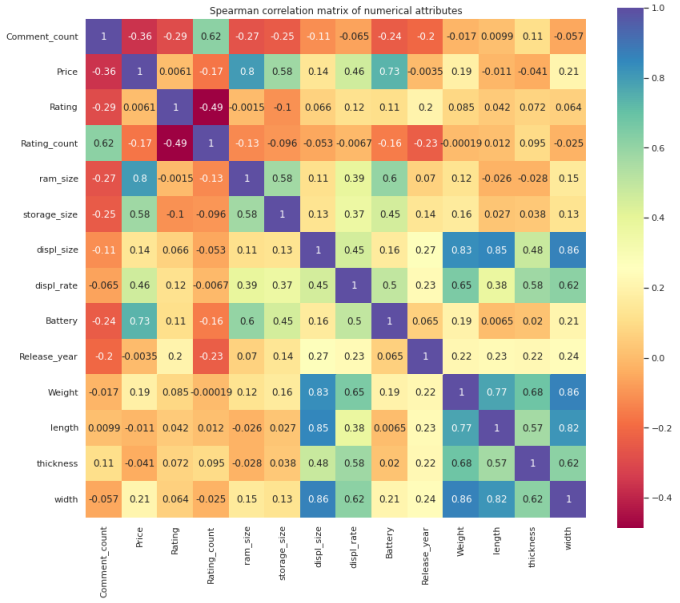

Figure 8: Spearman correlation matrix of numerical attributes

By observing the pearson coefficient and spearman coefficient in 2 matrix, we can conclude some relationships as follow:

- The rating count and the comment count are highly depended with each other, with the pearson coefficient = 0.75
- Display size influence a lots the size and weight of laptop, which is easy to understand
- Price is dependent on ram size, storage size and battery capacity
- Ram size is dependent on storage size

Except for Rating count, the relationship of all numerical attributes with Comment count is quite dubious. Maybe the correlation exists in the form of 2 by 1 variables, which we will examine further later on.

In the next section, we will look at the relationship of Price and Comment count.

Figure 9: Scatter plot of Price with Comment count in different price segment

As we discussed in the previous section, prices are divided into different price ranges, and the range of Comment count in different price ranges is different. This can be explained by the number of quantities supplied and demanded at each price range. The important question is “In the same price range, what factors make a product really standout?”. To answer this question for numeric attributes, we have plotted out the Pearson correlation matrix, the results obtained at each price range are quite similar to Figure 8, which does not tell much. We will leave this question open here and leave it for the next section of categorical bivariate analysis.

In the next part, we have a question about the correlation of ram size and storage size to the number of Comment counts. Is higher ram size and higher storage size equivalent to higher Comment count?

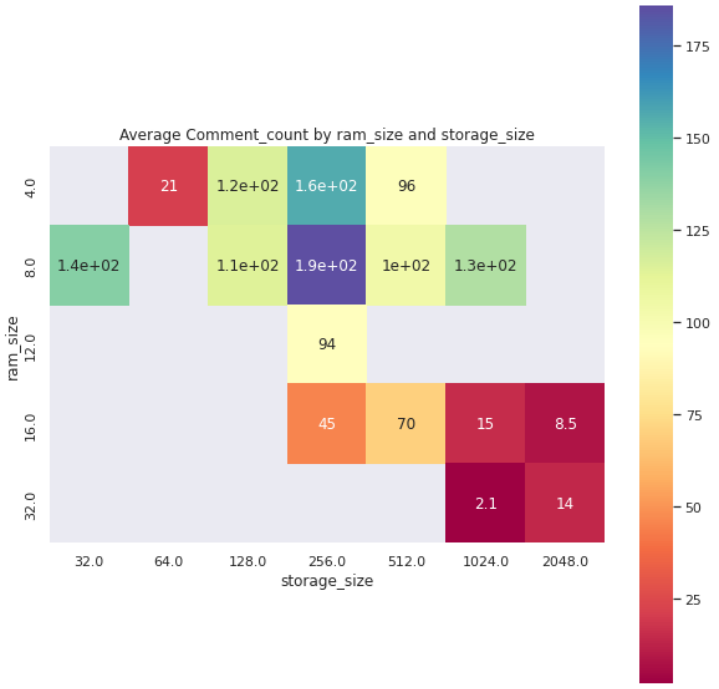

Figure 10: .Average comment count by ram size and storage size

The highest comment count does not lie on the highest ram size nor the highest storage size. The interest focuses on the middle ram (8 GB) and middle storage (256 GB), as this is common configuration for mid-range laptops.

3.3 Categorical analysis

In this section, we analyze the effect of categorical attributes on the level of interest on a laptop - represented by the number of comments.

Because there are many categorical attributes, before analyzing each attribute, we do a quick one-way ANOVA test of independence between each of these attributes with the number of comments. For one categorical attribute, a one-way ANOVA test is used to compare whether two (or more than two) samples means are significantly different or not. In our problem, each sample is the collection of the number of comments belonging to a particular group of a categorical attribute. The result of the ANOVA test is described in Table 1, where the number of comments is the measure of interest.

Analysis variable	#categories	ANOVA p-value	Reject hypothesis
Support extra storage slot	2	9.534354e-10	True
Price range	6	1.136593e-06	True
Support backlit keyboard	2	2.817380e-06	True
Body material	9	1.981282e-05	True
Screen technology	6	7.527494e-05	True
RAM upgradable	2	6.555863e-04	True
Storage type	3	5.946819e-03	True
Support headphone jack	2	6.033891e-02	False
Support LAN	2	1.022081e-01	False
Support HDMI	2	1.872486e-01	False
Support SD card	2	2.267507e-01	False
GPU brand	4	2.913708e-01	False
Brand	13	3.284112e-01	False
OS	4	4.403031e-01	False
CPU brand	3	5.159711e-01	False
GPU	52	5.879526e-01	False
CPU code	67	5.920354e-01	False
Support USB-A	2	6.499917e-01	False
Support Type-C	2	6.543112e-01	False
Display resolution	23	7.967517e-01	False

Table 1: ANOVA test for each categorical attribute. *#categories* denotes the number of groups in one attribute, for example *CPU brand* has 3 categories: *Intel*, *AMD*, *Apple*. Value in column *Reject hypothesis* is *False* if the corresponding value of column "ANOVA p-value" is less than *0.05*.

According to Table 1, there are significant differences in the interest between laptops of different price ranges (p-value = 1e-6), and likewise with attributes *Body material*, *Screen technology*, *RAM upgradable*, *Storage type*, *Support extra storage slot*, *Support backlit keyboard* as they all have p-value smaller than 0.05. Besides, there is no evidence for significant differences in the interest between laptops of different manufacturer brands, or CPU brands, or display resolutions.

The outcome of ANOVA test on price range does match our initial expectations where laptops with low and medium prices are intuitively more accessible than laptops with high prices. However the outcome of the test for brand, or display resolution does not match our expectations. For example with

the *Brand* attribute, the remarkable impression when investigate the Figure 11 on the first sight is that Acer or Lenovo are likely to have significant difference in the interest as compared to other brands, such as Microsoft, due to higher medians and IQR range, plus a considerable number of laptops with extremely high comment count (more than 500 comments).

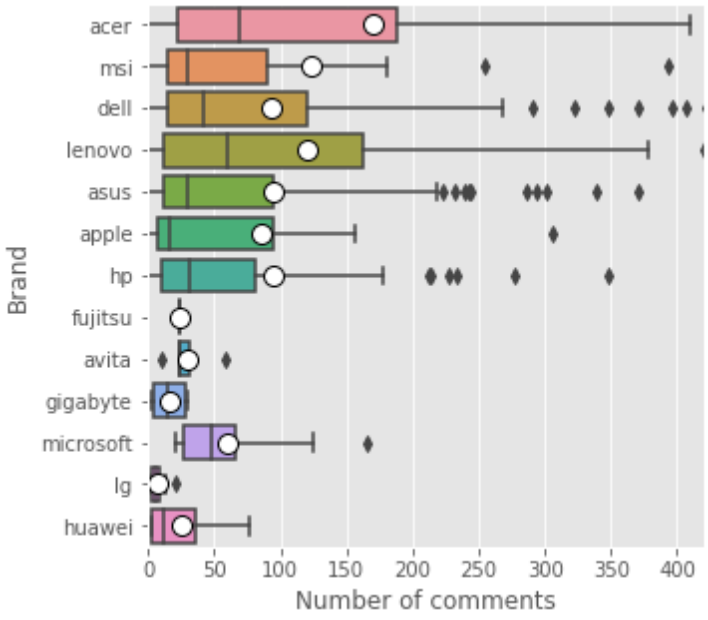

Figure 11: Box plot of number of comments against each brand. **Note**: the lim of *x*-axis is cut up to 420 to ensure visibility. The white dot represents the average number of comments.

As the ANOVA test does not match some of our expectations, we analyze each attribute in more detail. Remarkable findings are described below:

Pairs (million dongs)	T-t pvalue	Reject hs
[31, 40]-[11, 20]	0.000174	True
[41, 50]-[11, 20]	0.000910	True
[50, 93]-[11, 20]	0.003611	True

Table 2: T-test for the significant difference between price ranges. For the *Price range* attribute, Table 2 just shows pairs with null-hypothesis rejected, and all pairs with null-hypothesis not rejected are discarded due to page size.

According to Table 2, there are only 3 pairs of price range that are statistically significantly different in level of interest. In particular, the level of interest of laptops in the price range (11, 20) million dongs is significantly different from laptops that cost more than 30 million dongs. Table 3 represents the number of laptops and the average number of comments for each price range. Comparing the average values, we can infer: *laptops of (10, 20) million dongs are more interesting than laptops that cost more than 30 million dongs*. By the way, the number of laptops in range (11, 20) million dongs is way more than the number of laptops which cost more than 30 million dongs. This raises a question: *Whether manufacturers focus more on laptops of range (11, 20) million dongs and if in that case, whether this strategy optimizes their revenue*? This is an interesting question because the topic around sale price is highly concerned, especially for Apple - the manufacturer well-known for selling products with high prices. However, answering this question is out of the scope of this capstone project and the available data is simply not sufficient due to the lack of sales numbers.

Price range	Number of laptops	Average #comments
[1, 10]	10	157.400000
[11, 20]	164	171.618902
[21, 30]	182	109.256410
[31, 40]	77	47.259740
[41, 50]	36	20.402778
[50, 93]	27	15.796296

Table 3: Number of laptops and average number of comments for each price range.

Perform T-test similarly with *Body material*, *Screen technology*, *RAM upgradable*, *Storage type*, *Support extra storage slot*, *Support backlit keyboard*, we achieve the following conclusions:

- Laptops without backlit keyboards receive more interest than laptops with backlit keyboards.
- Laptops that are upgradable of RAM receives more interest than laptops that are not upgradable

- Laptops with LED displays receive more interest than laptops with IPS, VA, or TN displays, though the number of LED displays is relatively small as compared to IPS (31/331).

As for other attributes, no conclusion can be produced due to insufficient amount of data.

The greatest question is now pointed to *Brand* which is expected to affect level of interest but failed in the ANOVA test. We've done another pairwise T-test for *Brand*, and as demonstrated in Table 4: none of these brand pairs produce p-value less than 0.05. So, with the scraped data there is no evidence to prove the significant differences in level in interest between different manufacturer brands.

Pairs of brand	T-t pvalue	Reject hs
apple-acer	0.989765	False
asus-acer	0.843240	False
avita-acer	0.999933	False
dell-acer	0.944806	False
fujitsu-acer	1.000000	False
...	...	...
microsoft-lenovo	1.000000	False
msi-lenovo	1.000000	False
microsoft-lg	1.000000	False
msi-lg	0.985958	False
msi-microsoft	1.000000	False

Table 4: T-test for the significant difference between manufacturer brands. The table only shows several pairs due to limited page size. The complete result is that all null-hypotheses are false to be rejected.

3.4. Top 10% analysis

As the aforementioned, the distribution of *Comment counts* and *Rating counts* witness spectacular numbers in these stats of some product. This phenomenon leads to the question why these products are more concerned than the rest and what features make them outperformed.

To solve the quest, we will extract out some of the products with highest *Comment counts* and *Rating counts* - two attributes that, again, represent *popularity* of laptops.

3.4.1. Statistical description:

Stats	Highest Comment count	Highest Rating count
Min - Max	277 - 1756	5 - 126
Mean	590.8	23.67
Average Price	20.124.000 vnd (~ 1000USD)	21.261.000 vnd (~ 1050USD)
Max Price	33.949.000 vnd (~ 1500 USD)	46.999.000 vnd (~ 2000 USD)

Table 5: Comment count and Rating count and Price value range of top 10% highest.

At first glance, we can deduce that most of the concerned laptops based on the number of comments and ratings stay in the low-to-medium price range. No high-end product appears in the top 10%.

To systematically examine the factors that made up the top 50 hottest laptops, we walked through numerical and categorical attributes.

3.4.2. Numerical Analysis

Similar to the general numerical attributes analysis, continuous and discrete numerical attributes are separately examined since they have different types of impact on the top 50.

a. Continuous Numerical Attributes

Firstly, the correlation of all continuous attributes are calculated to basically determine which one plays a role in another's value, specially in Comment count and Rating count.

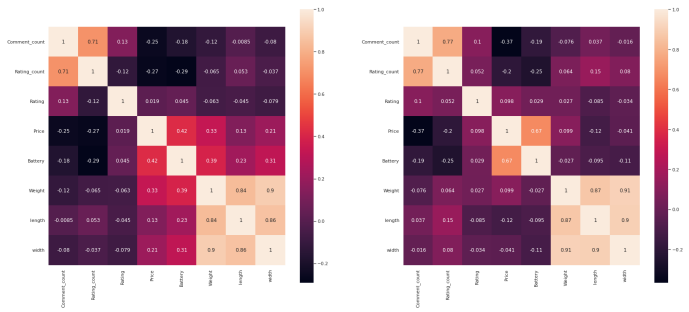

Figure 12: Correlation heatmap of numerical attributes in the top 10% highest Comment count (left) and Rating count (right).

As inferred from the heatmap, Price and battery are the 2 most impacted. The scatterplot of Comment count, rating count with these 2 attributes tells the same thing:

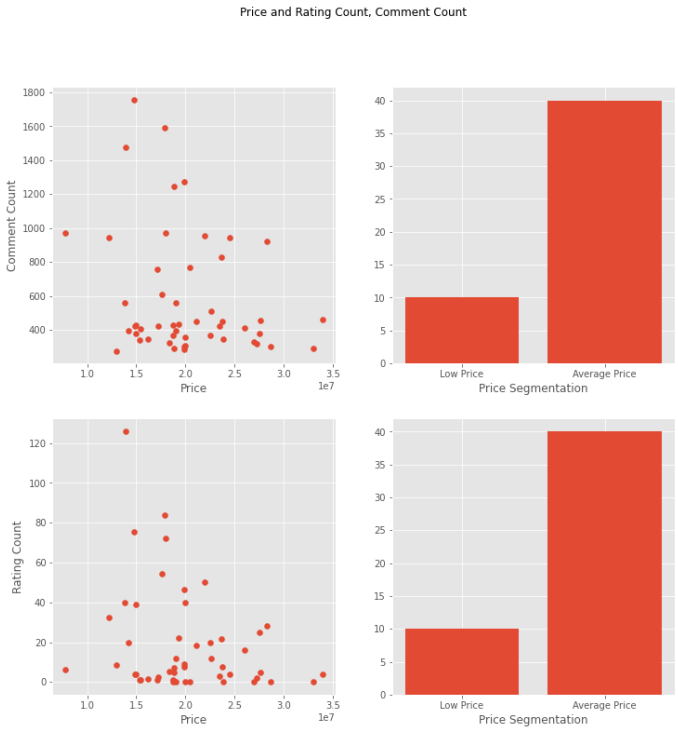

Figure 13: Scatter plot of price (left) corresponding to Comment count (upper) and Rating count (lower) and bar plot by price segmentation (right).

The majority of the most concerned laptops, additionally the most controversial products, are in the Average Price Segmentation.

The other attributes - weight, length and width - seems to not cause so much attention from customers according to their relatively low correlations.

b. Discrete Numerical Attributes

With discrete numerical Attributes, the display size, RAM size and storage size attributes show their distribution reasonably fit with the price segmentation that dominates the top 10% most popular products:

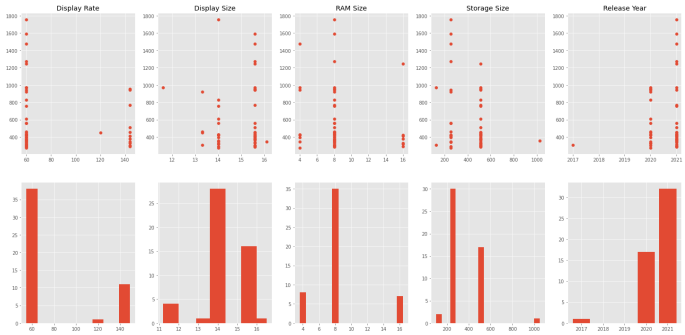

Figure 14: Scatter plot (upper) and barplot (lower) of number of comment counts according to some important categorical attributes.

The most popular RAM size in our range is 8GB, storage size is crowded from 256GB to 512GB, and display size ranges from 14 to 16 inches. These properties are nearly close to the majority of average-price laptops, of which segmentation is also being the most interested.

This detail, together with the price segmentation scatter plot, informed us that vast Vietnamese buyers are fond of middle-quality laptops with price lies between 15.000.000 to 30.000.000, with the evidence of highest comment count, rating count shown above.

3.4.3. Categorical Analysis

With numerical analysis, we have had an inference about price segmentation that dominate the top popular laptops. In the categorical analysis ahead, information about laptop brand, CPU and GPU brand and how the categorical attributes relate to the popularity of top concerned laptops will be explored.

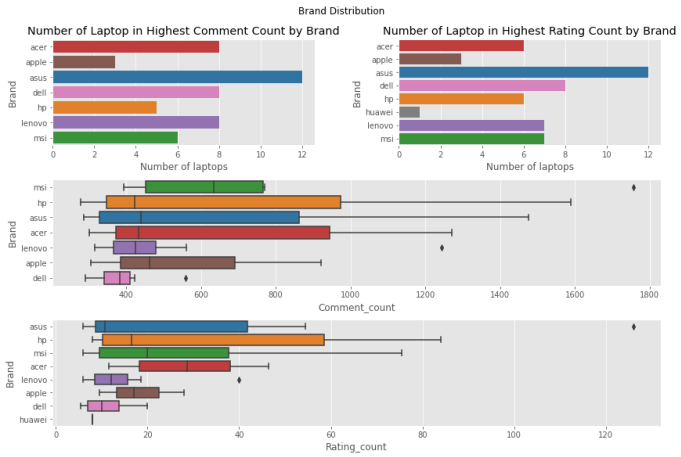

Figure 15: Number of laptops by brands (Upper) with highest Comment counts (left) anf Rating counts (right) and Distributions (Lower boxplots) by Comment counts and Rating counts respectively.

The brand distribution gives us the evidence of a competitive environment in the laptop market: Asus is leading the race but no brand persuasively conquers the market. The number of top interested laptops of Acer, Asus, Dell, HP, Lenovo are just nose to tail with each other. However, Acer, Asus and HP seem to draw more attention with a wide range of Comment count distribution.

Opposite to the racing in laptop brands environment, most interested laptops distribution in GPU brands is the duel of NVIDIA and Intel where the latter is taking advantage.

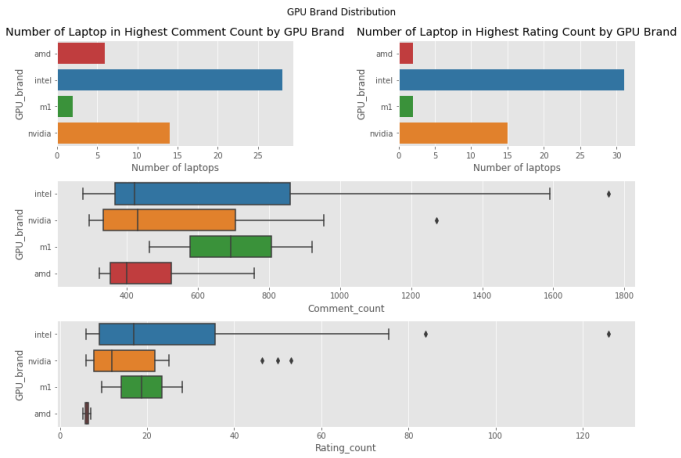

Figure 16: Number of laptops by GPU brands (Upper) with highest Comment counts (left) anf Rating counts (right) and Distributions (Lower boxplots) by Comment counts and Rating counts respectively.

Similarly, in the CPU brands field, Intel is dominating the highly popular laptop domain in both numbers and attention.

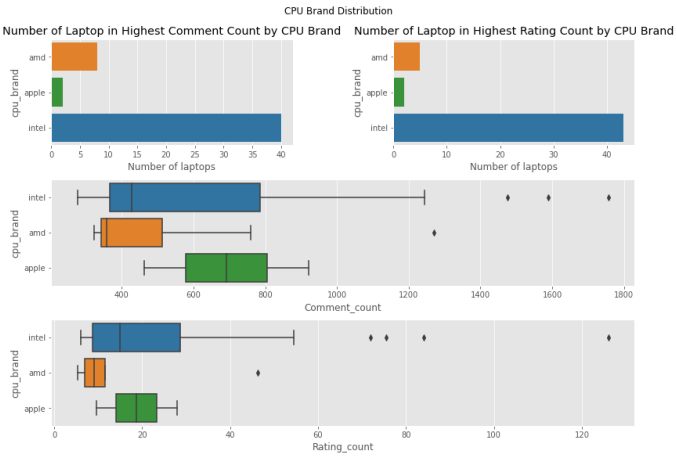

Figure 17: Number of laptops by CPU brands (Upper) with highest Comment counts (left) anf Rating counts (right) and Distributions (Lower boxplots) by Comment counts and Rating counts respectively.

The distribution of laptop brands, GPU brands and CPU brands tell us that the popularity of laptops reliably depends on the brand of hardware: In the top 50 most popular laptops, the division of brands becomes vividly different.

To verify the effects of laptop brands, GPU brands and CPU brands on the Comment count and Rating count i.e. the population of the most concerned laptops, we apply Chi square test on 2 contingency tables for each attribute. Chi square test with H0 hypothesis is “The calculating attribute does not have significant relation to Comment count in 95% confidence”. Below is the plotted result for the test with Comment count and Rating count:

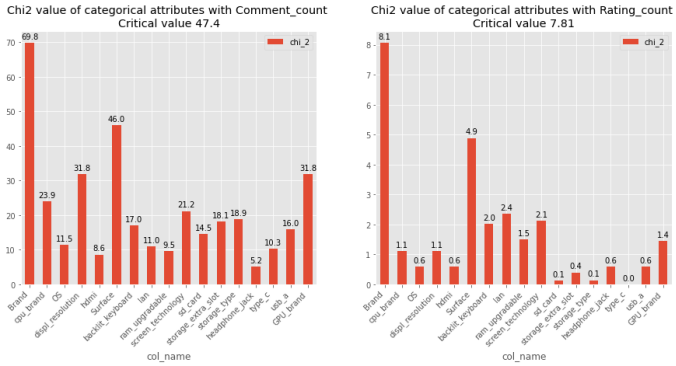

Figure 18: Chi square test results of categorical attributes according to 10% highest Comment count (left) and Rating count (right).

According to the Chi square test on both attributes, only the Brand attribute has a Chi square score bigger than the critical value and passed the test (69.8 with Comment count and 8.1 with Rating count) with H0 hypothesis "The variables have a significant relation". This test is conducted based on the number of laptops on each category within a comment count range, which can lead to biases by the number of products of each brand (which is significantly different among brands). Other attributes such as GPU brands and CPU brands, unexpectedly, do not affect the number of comment count and rating count as much as discussed earlier.

4. Conclusion

After collecting and analyzing, several truths of the laptop have been told. There are three main points will soon to be discussed:

4.1. The efficiency of response variables

In this project, we chose 3 response variables of *Comment count, Rating* and *Rating count* to represent the popularity and controversy of each product. Logically, these variables tell how much attention the product draws from customers: more comments and ratings means more buyers and responses. However, in fact, there are plenty of natural factors that reduce how close the response variables are to their meaning.

For example, the *Comment count* and *Rating count* attributes have an intensive number of 0 to 10, much less than their average. This situation comes from the fact that not every buyer will return to the site to leave a comment or rating, or the website display encourages viewers to catch more eyes on the earlier shown products.

To overcome this fog of efficiency, another analysis on the top 10% highest of response variables was conducted and gives us more detailed insights.

For further analysis and deeper marketing development, it is best to use response variables of sales (i.e. number of laptops sold) because this variable is not biased by environment and more pure to the ultimate concerns - selling laptops.

4.2. Customer habits are surprising

Since we are working with an economic objective, the habits and favorites of customers can help a lot in improving sales propagation, leading attractive products to appropriate buyers.

In general, most attention comes from customers of average-price laptops. This indicates the fact that the laptop market in Vietnam has become popular and general to people's life, and the laptop market in Vietnam seems to prefer average-price laptops, with decent laptop quality and affordable price.

From the analysis, customers are more interested in the product's brand, as shown in the top 10% analysis. This is quite expectable in common sense. However, the second closest to the brand in attraction was the *Surface material* which tells what type of laptop's physical cover (Plastic ỏ metallic). Another surprise point is that the *RAM size* property does not catch so much attention as expected. Through correlation calculating, the *RAM size* shows almost no impact on comment and rating count - or our popularity of product. This can be caused by the distractions of other attributes or because customers don't discuss much about the laptop on its ram size.

This is a surprising contrast of how customers care about a detailed appearance feature rather than technical property and power.

In the top 10% look, the battery information draws a lot of attraction. Buyers may leave comments on how long the battery lasts - a common and important question to general

people but is not explicitly shown in the battery number. Laptops with higher battery capacity also attract more comments.

4.3. Marketing suggestions

It is important to understand customers' habits and interests, however, that information is only worthy when decisive suggestions are made to improve sales and reach more customers. In our case of laptop market, through examine product's popularity and controversy, we have some advices to online laptop selling:

+ Hottest comes first: In website display, laptops with better interactive information such as comment, rating (i.e. more attractive) should be displayed in the first page and shown to customers without clicking "Next page". Similarly, brands with higher sales should occupy a larger ratio in the first page since the brand name plays an essential role to buyers' first attraction.
+ Laptops should be classified by their purposes such as gaming, office, graphic design, etc… Since different usages catch different classes of customers. From there, customers can get to their desired laptop faster - and distribute the products more evenly to reduce the bias of evaluating variables.

5. Acknowledgement and Contribution

Project proposal & Data source selection.	Hung 50%, Lap 50%
Scraping and cleaning FPT data	Nhật
Scraping and cleaning CellphoneS, LaptopArena data	Lập
Scraping and cleaning Thegioididong, Dienmayxanh data	Hưng
Reclean mapping attributes, Integrate data	Hưng
EDA - top 10% analysis	Nhật
EDA - numerical analysis	Hưng
EDA - categorical analysis	Lập
Write Conclusion, top 10% analysis, abstract of the report	Nhật
Write categorical part of univariate and bivariate analysis, descriptive analysis of the report	Lập
Write numerical part of univariate and bivariate analysis, cleaning and integrating part	Hưng