

ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ



BÁO CÁO
ĐỒ ÁN PTDL R
NĂM HỌC 2021-2022

Giáo viên hướng dẫn: HỒ QUỐC DŨNG

Lớp: KHDL & TTNT

Sô phách
(Do hội đồng chấm thi ghi)

Thừa Thiên Huế, ngày ... tháng ... năm 2022

ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ



(MẪU BÌA PHỤ)

BÁO CÁO
ĐỒ ÁN PTDL R

NĂM HỌC 2021-2022

Giảng viên hướng dẫn: HỒ QUỐC DŨNG

Lớp: KHDL & TTNT

Sinh viên thực hiện: SỬ THÀNH CÔNG

Sô phách

(Do hội đồng chấm thi ghi)

Thừa Thiên Huế, ngày ... tháng ... năm 2022

MỤC LỤC

MỤC TIÊU CHUNG	1
CHƯƠNG I. THU THẬP DỮ LIỆU	2
1. Các nguồn dữ liệu	2
2. Sử dụng Selenium để thu thập dữ liệu	5
3. Làm sạch dữ liệu bằng Python.....	14
CHƯƠNG II. PHÂN TÍCH DỮ LIỆU.....	20
1. Tổng quan về dữ liệu:	20
2. Thống kê mô tả (EDA & Trực quan hóa dữ liệu)	21
2.1 Phân tích biến định tính (Categorical).....	22
2.2 Phân tích biến định lượng (Numerical).....	30
3 Thống kê suy diễn	35
3.1 Sử dụng ANOVA để so sánh giá giữa các cửa hàng phân phối	35
3.2 Phân tích mối quan hệ giữa các biến.....	36
KIỂM TRA ĐẠO VĂN	40
TÀI LIỆU THAM KHẢO	41

MỤC TIÊU CHUNG

Hẳn ai cũng muốn mua thiết bị tốt nhất với giá thấp nhất có thể và Giá của 1 chiếc laptop sẽ phụ thuộc vào những thành phần cấu tạo như: RAM, Core, ... **Nhưng thành phần nào có ảnh hưởng lớn nhất đến giá cả?**



Bằng việc sử dụng dữ liệu về **Máy tính xách tay** (từ những web kinh doanh laptop), sử dụng ngôn ngữ lập trình R làm công cụ chính và thực hiện việc phân tích dữ liệu để trả lời cho câu hỏi trên.

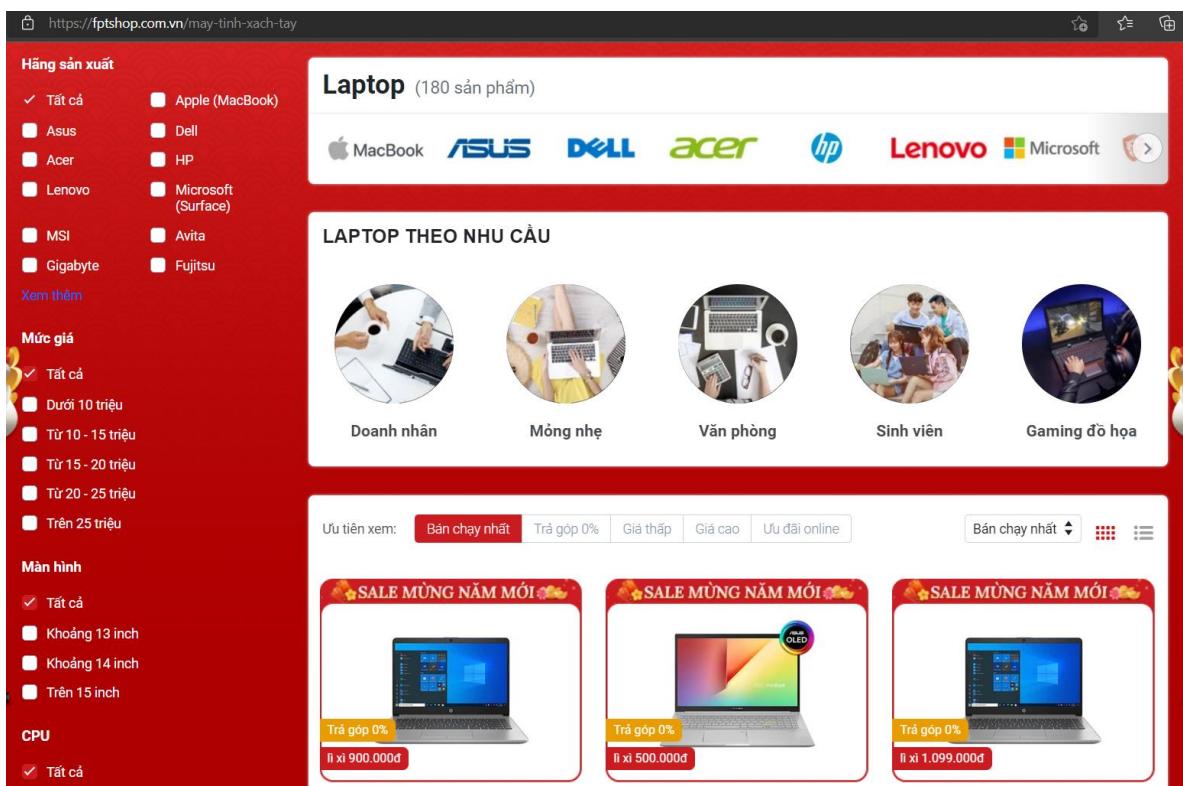
CHƯƠNG I. THU THẬP DỮ LIỆU

1. Các nguồn dữ liệu

Ngày nay có rất nhiều trang web cung cấp và phân phối Laptop online, ngay cả những trang Thương mại điện tử (như Tiki, Lazada, Shopee,...) cũng là một nền tảng trực tuyến để người mua có thêm nhiều lựa chọn.

Để đảm bảo chất lượng và giá thành của sản phẩm, trong bài viết này sẽ chỉ lấy dữ liệu sản phẩm Laptop từ các trang uy tín. Cụ thể là:

- [Laptop chính hãng, giá rẻ, trả góp 0% \(fptshop.com.vn\)](https://fptshop.com.vn/may-tinh-xach-tay)



Hình 1. Giao diện Danh mục sản phẩm laptop của FPTShop

- [Laptop | Máy tính xách tay chính hãng Giá rẻ, Trả góp 0% \(dienmayxanh.com\)](https://www.dienmayxanh.com/laptop)

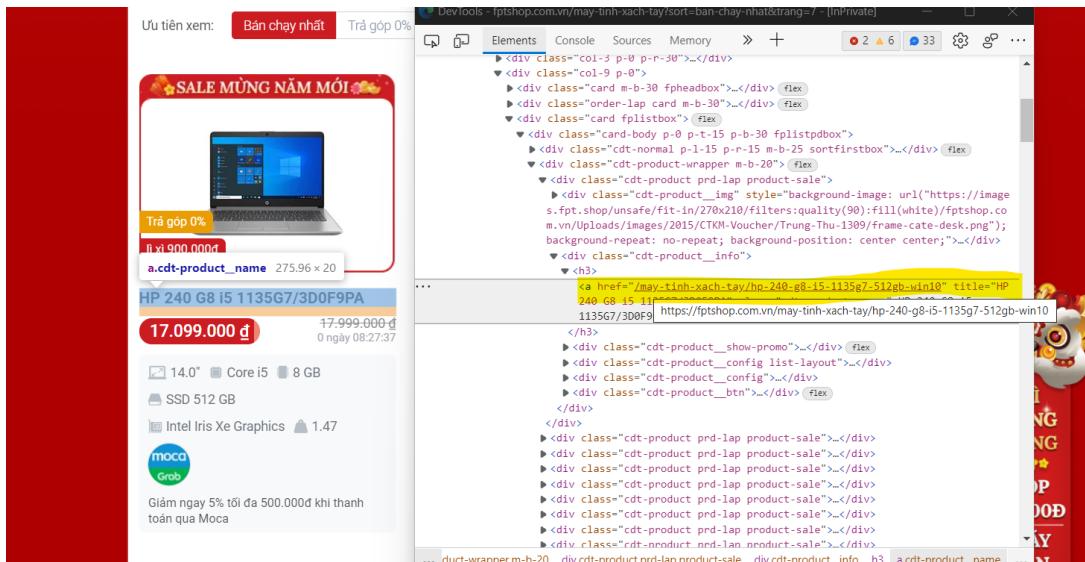
Hình 2. Giao diện Danh mục sản phẩm laptop của Điện máy XANH

- [Phong Vũ \(phongvu.vn\)](https://phongvu.vn/laptop-macbook-scat.01-N001)

Hình 3. Giao diện Danh mục sản phẩm laptop của Phong Vũ

Sau khi xem qua danh mục sản phẩm Laptop của 3 trang web trên, ta nhận thấy cả 3 đều có cấu trúc giống nhau. Vì vậy tôi sẽ hướng dẫn crawl sản phẩm trên FPTshop và thực hiện tương tự đối với 2 trang còn lại:

- **Trang danh mục sẽ chứa vùng hiển thị sản phẩm và đường dẫn đến sản phẩm**



Hình 4. Inspect trang danh mục để tìm đường dẫn

- Click vào 1 sản phẩm sẽ truy cập đến Trang thông tin của sản phẩm



Hình 5. Trang thông tin của sản phẩm - Laptop HP 240 G8 i5 1135G7/8GB/512GB/14.0''HD/Win 10 / Fptshop.com.vn

2. Sử dụng Selenium để thu thập dữ liệuⁱ

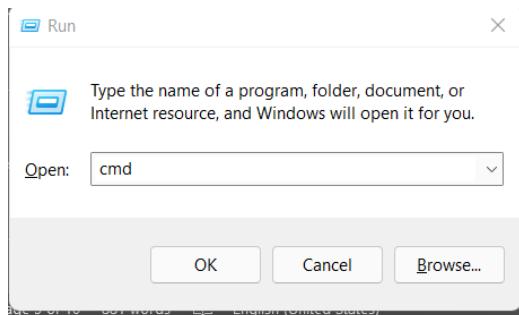
Hiện nay với sự phổ biến của Python, có rất nhiều thư viện hỗ trợ duyệt web và lấy dữ liệu, phổ biến như là:

- Thư viện Scrapy: Cung cấp cho bạn khả năng phân tích và truy cập vào các trang web là lọc dữ liệu
- Thư viện Selenium: Cung cấp một giao diện, tương tác như một web browser.
- Ngoài ra còn có: BeautifulSoup, Urllib, Requests, ...

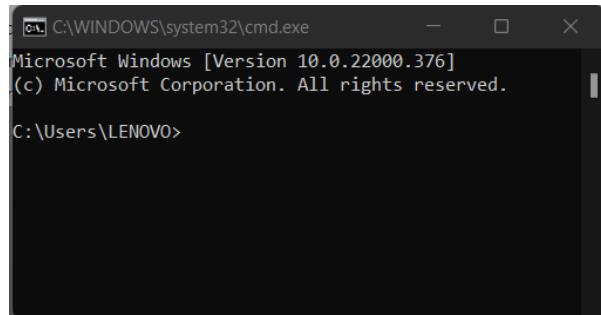
Trong phần này chúng ta sẽ chủ yếu sử dụng **Selenium 3.141.0** để tiến hành thu thập dữ liệu.

Cài đặt Selenium

Mở Command Prompt (Windows + R) và gõ cmd:



Hộp thoại Run



Cửa sổ Command Prompt

Tại cửa sổ **Command Prompt**, gõ lệnh:

```
pip install selenium
```

Hoặc nếu muốn cài đặt chính xác phiên bản **Selenium 3.141.0**:

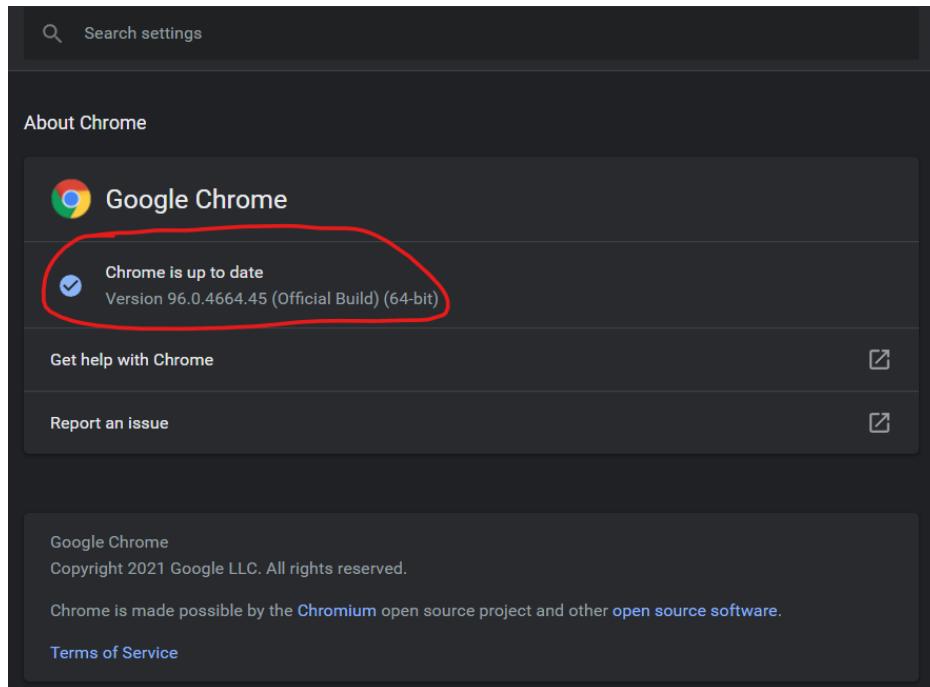
```
pip install selenium==3.141.0
```

Đối với các hệ điều hành khác xem thêm tại [đây](#).

Cài đặt WebBrowser

Lưu ý cần chọn phiên bản webdrive phù hợp với trình duyệt hiện tại:

Để check phiên bản Chrome hiện tại: chrome://settings/help



Hình 6. Phiên bản Chrome hiện tại

Vì trình duyệt hiện tại đang sử dụng là **Chrome 96.0.4664.45** nên mình chọn webdrive tương ứng là ChromeDrive Version 96.0.4664.45

Tìm phiên bản Chrome tương ứng và tải về tại trang ChromeDriver:
<https://chromedriver.chromium.org/downloads>

Current Releases

- If you are using Chrome version 97, please download [ChromeDriver 97.0.4692.71](#)
- If you are using Chrome version 96, please download [ChromeDriver 96.0.4664.45](#)
- If you are using Chrome version 95, please download [ChromeDriver 95.0.4638.69](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

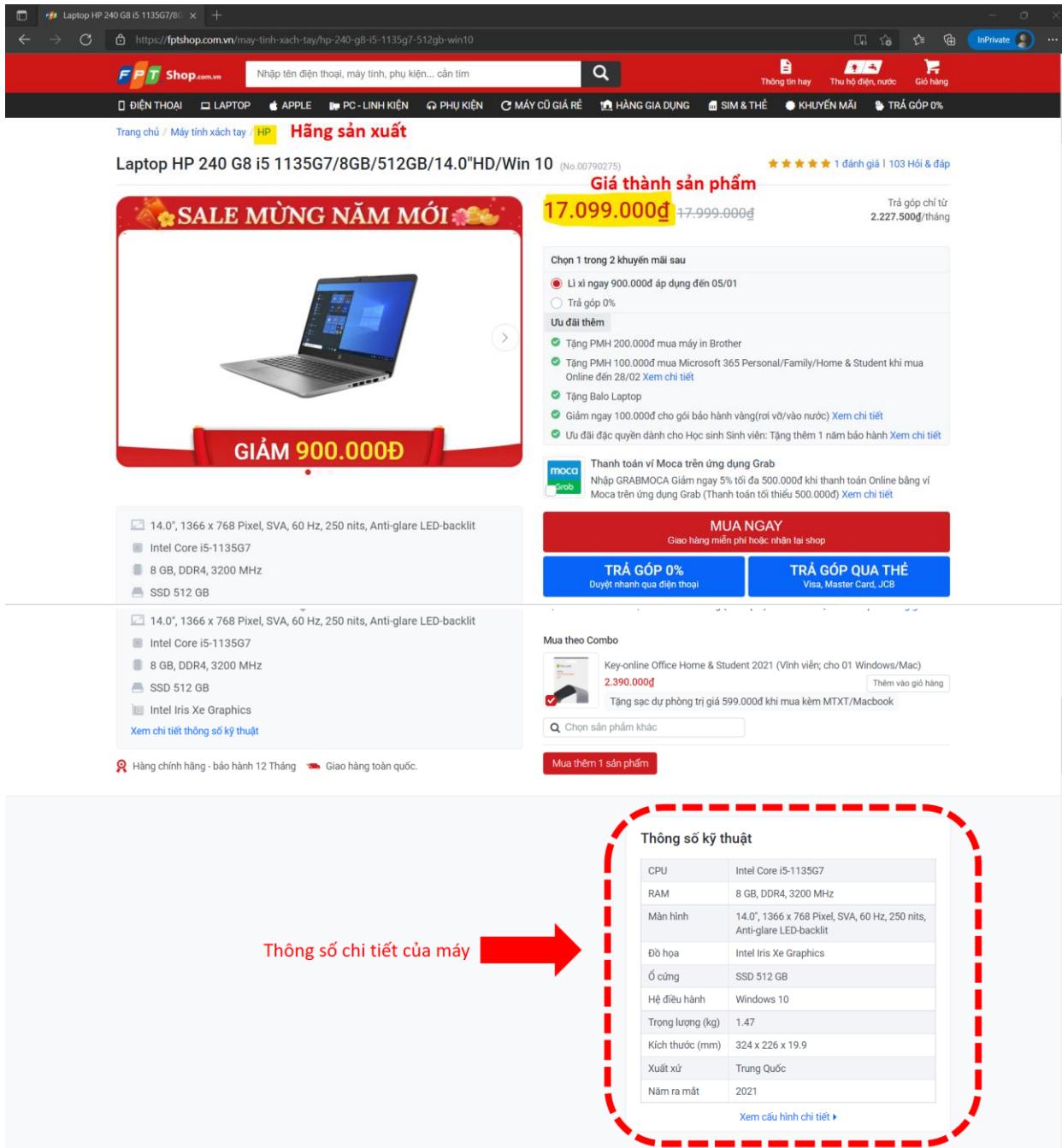
If you are using Chrome from Dev or Canary channel, please following instructions on the [ChromeDriver Canary](#) page.

For more information on selecting the right version of ChromeDriver, please see the [Version Selection](#) page.

Hình 7. Giao diện trang ChromeDriver

Xác định những thông tin cần lấy

Khi truy cập vào trang thông tin sản phẩm, ta sẽ xác định được những thông tin cụ thể như: Tên sản phẩm, Giá sản phẩm, Hãng sản xuất, CPU, RAM, Kích thước màn hình (inch), Card Đồ họa, Ổ cứng, Hệ điều hành, Trọng lượng, Xuất xứ và Năm ra mắt.



The screenshot shows a product page for a Laptop HP 240 G8 i5 1135G7/8GB/512GB/14.0"HD/Win 10. The page includes a banner for a New Year sale, a price of 17,099,000đ, and a detailed technical specification table.

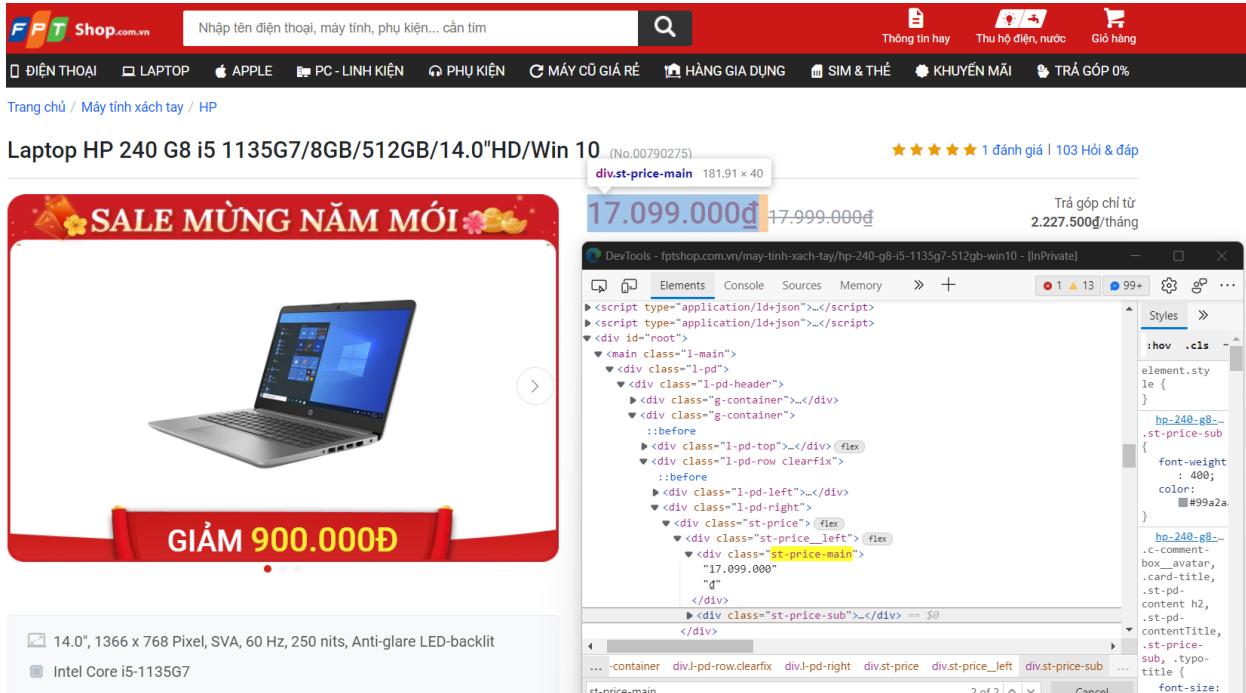
Thông số kỹ thuật

CPU	Intel Core i5-1135G7
RAM	8 GB, DDR4, 3200 MHz
Màn hình	14.0", 1366 x 768 Pixel, SVA, 60 Hz, 250 nits, Anti-glare LED-backlit
Đồ họa	Intel Iris Xe Graphics
Ổ cứng	SSD 512 GB
Hệ điều hành	Windows 10
Trọng lượng (kg)	1.47
Kích thước (mm)	324 x 226 x 19.9
Xuất xứ	Trung Quốc
Năm ra mắt	2021

Hình 8. Giao diện trang thông tin sản phẩm

Phân tích trang web

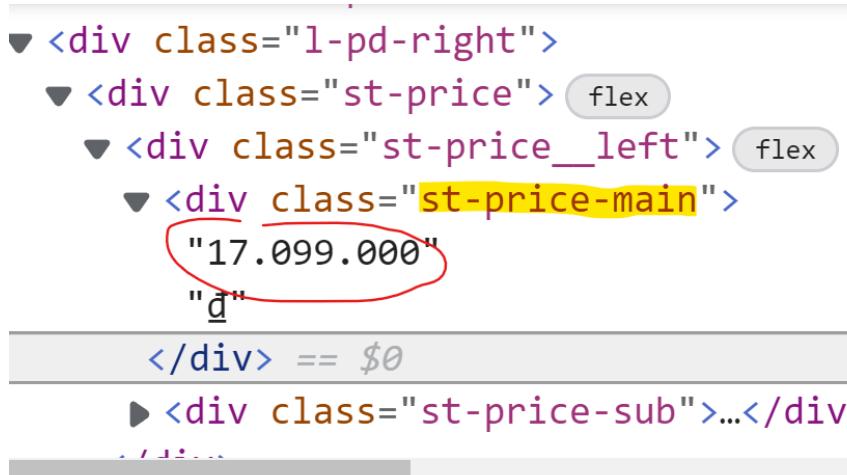
Chương trình sẽ không thể nào tự xác định được vị trí những thông tin cần lấy, vì vậy chúng ta cần phân tích trang HTML để trở đến vị trí chính xác chứa thông tin rồi click trái chọn **Inspect**, cửa sổ **DevTools** sẽ bật lên (*Hình 9*)



Hình 9. Cửa sổ DevTools bật lên khi Inspect

Như vậy **Giá của sản phẩm** sẽ nằm trong `<div class = "st-price-main">`

XPath tương ứng: `//*[@class="st-price-main"]`

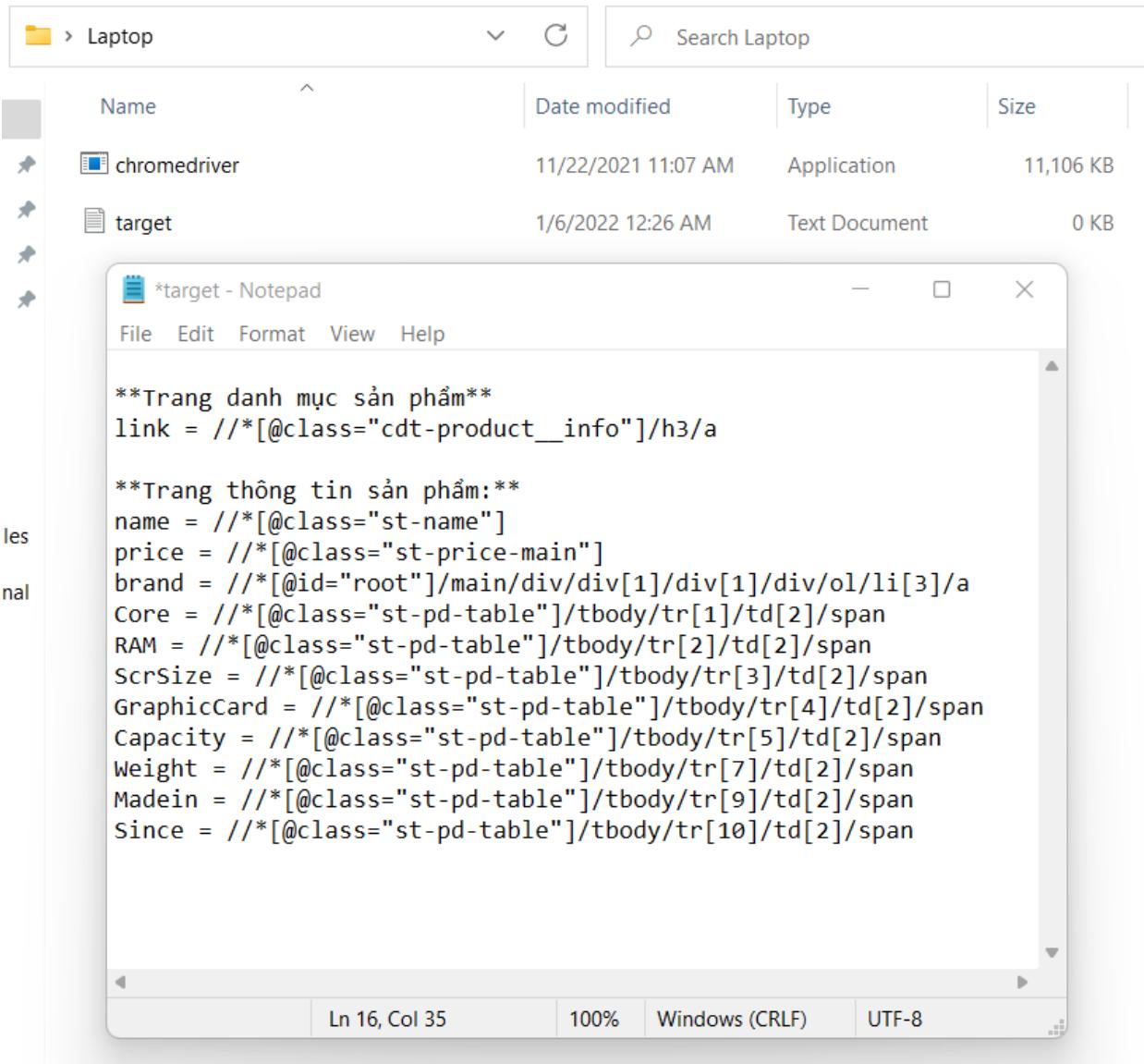


Hình 10. Khi zoom lên

Tạo thư mục làm việc (**Laptop**) và chuyển file **ChromeDrive đã tải** vào thư mục vừa tạo.

Tạo file **target.txt** để ghi lại XPath của **Giá của sản phẩm** và thực hiện tương tự với những thông số khác.

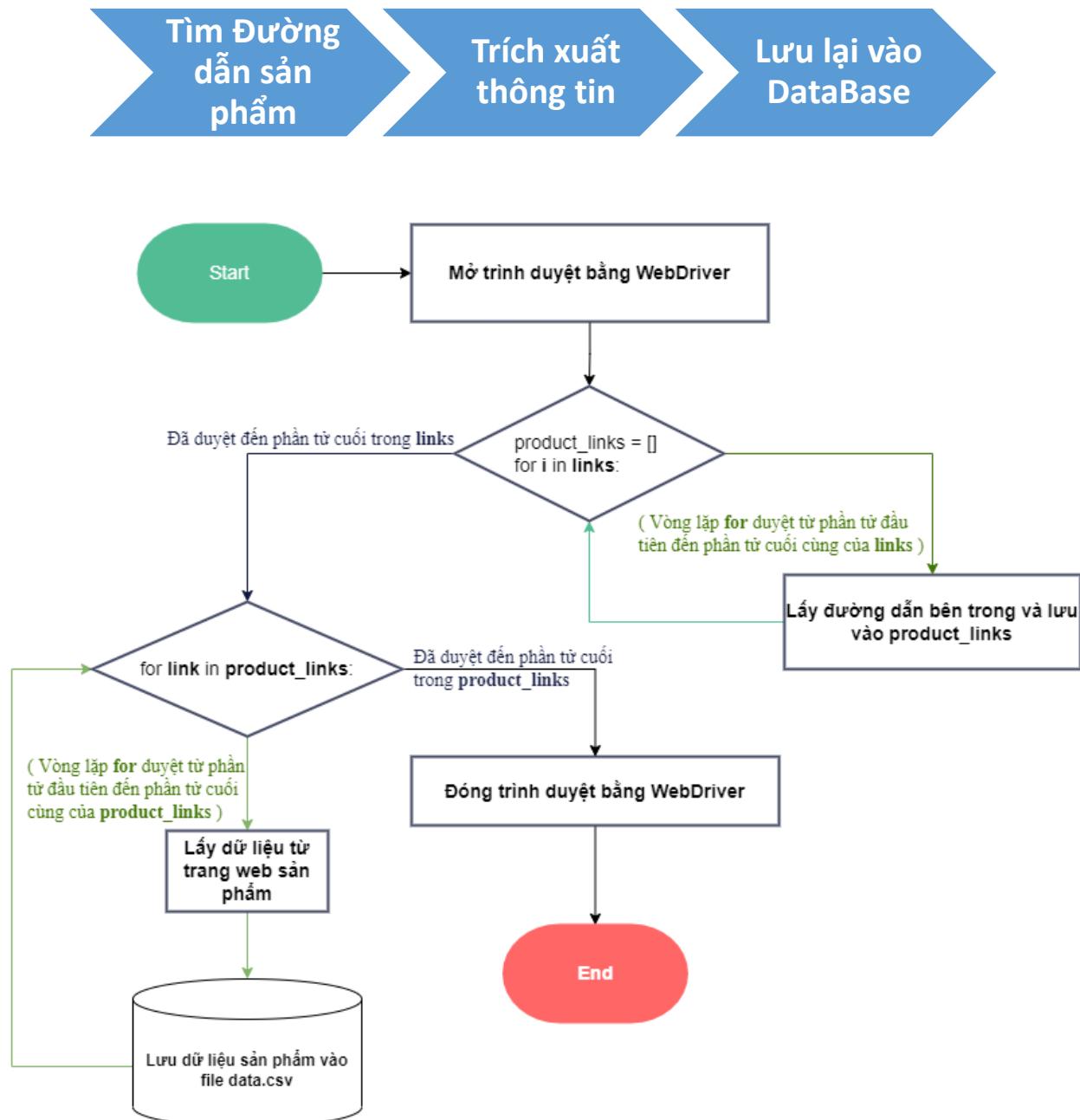
Kết quả thu được:



Hình 11. Thư mục làm việc vừa tạo (tên thư mục: Laptop)

Cấu trúc chương trình

Ý tưởng chính:



Hình 12. Sơ đồ thuật toán

Xây dựng chương trình

Tại thư mục làm việc, tạo file fpt.py sau đó tiến hành khai báo các thư viện sử dụng:

```
from selenium import webdriver
from time import sleep
from selenium.webdriver.chrome.options import Options
import csv
```

Cài đặt thông số cho WebDriver (ẩn danh và chạy ngầm)

```
# Options
chrome_options = Options()
chrome_options.add_argument("--incognito")
chrome_options.add_argument("--headless")
```

Truyền đường dẫn ban đầu (Trang danh mục) <https://ftpshop.com.vn/may-tinh-xach-tay?sort=ban-chay-nhat&trang=10> và biến **url**. Tạo biến **product_links** để lưu trữ những đường dẫn tìm được.

Sau đó set đường dẫn **ChromeDrive** bằng `webdriver.Chrome(executable_path="chromedriver.exe")`

Truy cập vào trang danh mục bằng `browser.get(url)`

```
product_links = []
url = 'https://ftpshop.com.vn/may-tinh-xach-tay?sort=ban-chay-nhat&trang=10'
browser = webdriver.Chrome(executable_path="chromedriver.exe")
browser.get(url)
```

Sau khi truy cập vào trang danh mục, tìm các đường link nằm trong các **class="cdt-product_info"**, chúng ta sử dụng hàm **find_elements_by_xpath** để tìm đến các class, bạn có thể sử dụng nhiều phương thức truy tìm khác (**by_CSS Selector**, **by_class_name**, **by_id_name**,...) sau đó sử dụng **get_attribute('href')** để lấy đường sản phẩm bên trong và lưu đường dẫn vào **product_links**

```
# Vị trí chứa đường dẫn sản phẩm
links = browser.find_elements_by_xpath('//*[@[@class="cdt-product_info"]]/h3/a')
for i in links:
    link=i.get_attribute("href") # Lấy đường dẫn
    product_links.append(link) # và thêm vào product_links
```

Tương tự, bạn có thể lấy được thông số khác (feature và Xpath tương ứng đã được xác định trong file **target.txt**)

```

for link in product_links:
    browser.get(link)
    try:
        data = {
            "Product" : browser.find_element_by_xpath('//*[@[@class="st-name"]]').text,
            "Price" : browser.find_element_by_xpath('//*[@[@class="st-price-main"]]').text,
            "Brand" : browser.find_element_by_xpath('//*[@[@id="root"]/main/div/div[1]/div[1]/div[ol/li[3]/a')).text,
            "Core" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[1]/td[2]/span').text,
            "RAM" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[2]/td[2]/span').text,
            "ScrSize" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[3]/td[2]/span').text,
            "GraphicCard" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[4]/td[2]/span').text,
            "Capacity" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[5]/td[2]/span').text,
            "Weight" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[7]/td[2]/span').text,
            "Madein" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[9]/td[2]/span').text,
            "Since" : browser.find_element_by_xpath('//*[@[@class="st-pd-table"]]/tbody/tr[10]/td[2]/span').text,
            "Shop": 'FPTShop',
            "URL":link,
        }
    except:
        pass

```

Tạo file **data.csv** bằng python

```

csv_columns = [
    'Product','Price','Brand','Core','RAM','ScrSize',
    'GraphicCard','Drive_Type','Capacity','OperSystem',
    'Weight','Madein', 'Since','Shop','URL']
with open('data.csv', "a", encoding="utf8") as f:
    writer = csv.DictWriter(f, fieldnames=csv_columns)
    writer.writeheader()

```

Thao tác với file CSV bằng Python: Mở và ghi lại dữ liệu thu thập được

```

with open('data.csv', "a", encoding="utf8") as f: # Mở file data.csv
    writer = csv.DictWriter(f, fieldnames=csv_columns)
    writer.writerow(data) # Ghi dữ liệu theo cột

```

Xử lý dữ liệu (xóa ký tự thừa, giữ lại số) đối với các feature như **Kích thước màn hình, Dung lượng, RAM, Giá:**

```

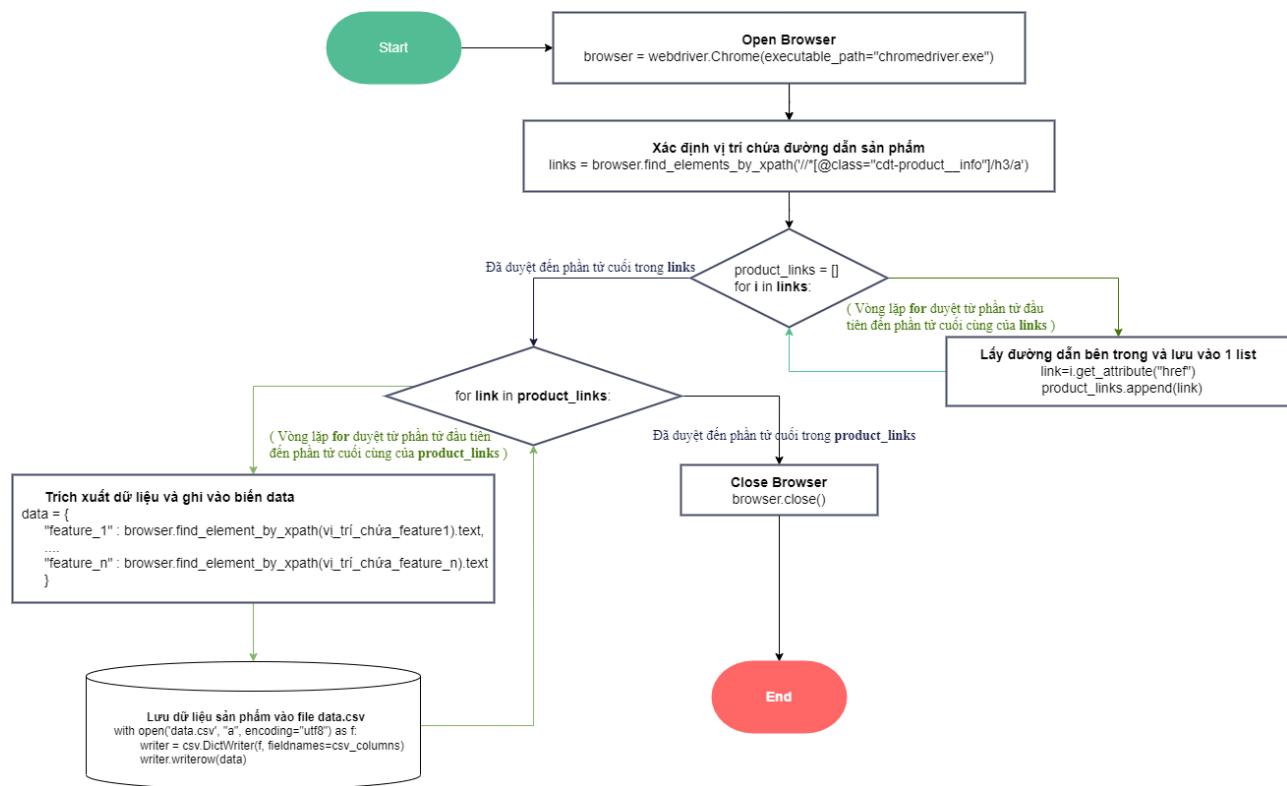
import re
#text_after = re.sub(regex_search_term, regex_replacement, text_before)
def convert(value):
    value = re.sub(r'D', "", value)
    return int(value)

```

sleep(10) dùng để tạm dừng chương trình là 10s để selenium có đủ thời gian để hiện thị cũng như tránh bị chặn IP (“detected unusual traffic”)

```
sleep(10)
```

Lưu đồ thuật toán chi tiết



Hình 13.

Thực hiện tương tự với **ĐiệnMáyXanh (dmx.py)** và **PhongVu (phongvu.py)**

Full SourceCode

Data thu thập được từ 3 nguồn:

	Product	Price	Brand	Core	RAM	ScrSize	GraphicCard	Drive_Type	Capacity	OperSystem	Weight	Madein	Since	Shop	URL
0	Laptop Acer Nitro Gaming AN515 57 74NU i7 1180...	29999000	Acer	Intel Core i7-1180H 4.60 GHz	8	15.6	NVIDIA GeForce RTX 3050Ti 4 GB & Intel UHD Gra...	SDD	512	Window	2.2	Trung Quốc	2021	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/acer-...
1	Laptop MSI Modern 14 B11MOU 852VN i5 1155G7/8G...	18799000	MSI	Intel Core i5-1155G7 4.50 GHz	8	14.0	Intel Iris Xe Graphics	SDD	512	Window	1.3	Trung Quốc	2021	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/msi-m...
2	Laptop Dell Inspiron N3511 i3 1115G4/4GB/256GB...	15299000	Dell	Intel Core i3-1115G4 4.10 GHz	4	15.6	Intel UHD Graphics	SDD	256	Window	1.7	Trung Quốc	2021	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-...
3	Laptop Dell Vostro V3500 i3 1115G4/8GB/256GB/1...	15599000	Dell	Intel Core i3-1115G4 4.10 GHz	8	15.6	Intel UHD Graphics	SDD	256	Window	2.0	Trung Quốc	2021	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-...
4	Laptop Dell Inspiron N3511 i5 1135G7/4GB/512GB...	19999000	Dell	Intel Core i5-1135G7	4	15.6	Intel Iris Xe Graphics	SDD	512	Window	1.7	Trung Quốc	2021	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-...

Hình 14. Dataset

3. Làm sạch dữ liệu bằng Pythonⁱⁱ

- Dataset gồm **273** hàng, **15** cột:

```
[5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 273 entries, 0 to 272
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Product     273 non-null    object 
 1   Price       273 non-null    int64  
 2   Brand       273 non-null    object 
 3   Core        273 non-null    object 
 4   RAM         273 non-null    int64  
 5   ScrSize     273 non-null    float64
 6   GraphicCard 273 non-null   object 
 7   Drive_Type  273 non-null    object 
 8   Capacity    273 non-null    int64  
 9   OperSystem  273 non-null    object 
 10  Weight      273 non-null    float64
 11  Madein      133 non-null    object 
 12  Since       273 non-null    int64  
 13  Shop        273 non-null    object 
 14  URL         273 non-null    object 
dtypes: float64(2), int64(4), object(9)
memory usage: 32.1+ KB
```

- Kiểm tra cột chứa giá trị khuyết (missing value):

• • •

```
[4]: missing_values_table(df)

Your selected dataframe has 15 columns.
There are 1 columns that have missing values.

[4]:      Missing Values  % of Total Values
Madein           140           51.3
```

Chỉ duy nhất cột **Madein** bị thiếu giá trị, và lượng giá trị khuyết chiếm **51.3%**, có thể trong quá trình crawl dữ liệu, cửa hàng không hiển thị nơi sản xuất sản phẩm vì vậy trường dữ liệu này cần bị loại bỏ.

- Thay đổi tên các trường dữ liệu cho phù hợp hơn

```
[6]: del df['Madein']
df = df.rename(columns=str.lower)
df.columns

[6]: Index(['product', 'price', 'brand', 'core', 'ram', 'scrsizes', 'graphiccard',
       'drive_type', 'capacity', 'opersystem', 'weight', 'since', 'shop',
       'url'],
       dtype='object')

[7]: df = df.rename(
      columns={'product':'id',
                'graphiccard':'gpu',
                'core':'cpu',
                'capacity':'memory'})
df.columns

[7]: Index(['id', 'price', 'brand', 'cpu', 'ram', 'scrsizes', 'gpu', 'drive_type',
       'memory', 'opersystem', 'weight', 'since', 'shop', 'url'],
       dtype='object')
```

- Mỗi cửa hàng sẽ có 1 cách đặt tên sản phẩm khác nhau, vì vậy để đồng bộ cho việc phân tích, ta chuyển tên sản phẩm **product** thành **id** (dạng số):

```
[17]: df['id'] = np.arange(0, len(df))
df.head(5)

[17]:   id  price  brand      cpu  ram  scrsizes      gpu  drive_type  memory  opersystem  weight  since  shop                                     url
0   0  29999000  Acer  Intel Core i7-11800H  8  15.6  NVIDIA GeForce RTX 3050Ti 4 GB & Intel UHD Gra...  SDD  512  Window  2.2  2021  FPTShop  https://ftpshop.com.vn/may-tinh-xach-tay/acer-...
1   1  18799000  MSI  Intel Core i5-1155G7  8  14.0  Intel Iris Xe Graphics  SDD  512  Window  1.3  2021  FPTShop  https://ftpshop.com.vn/may-tinh-xach-tay/msi-m...
2   2  15299000  Dell  Intel Core i3-1115G4  4  15.6  Intel UHD Graphics  SDD  256  Window  1.7  2021  FPTShop  https://ftpshop.com.vn/may-tinh-xach-tay/dell-...
3   3  15599000  Dell  Intel Core i3-1115G4  8  15.6  Intel UHD Graphics  SDD  256  Window  2.0  2021  FPTShop  https://ftpshop.com.vn/may-tinh-xach-tay/dell-...
4   4  19999000  Dell  Intel Core i5-1135G7  4  15.6  Intel Iris Xe Graphics  SDD  512  Window  1.7  2021  FPTShop  https://ftpshop.com.vn/may-tinh-xach-tay/dell-...
```

- Ở cột **brand** xuất hiện giá trị trùng lặp (chữ hoa chữ thường)

```
[9]: df['brand'].value_counts()
```

```
[9]: Asus      56
Dell      52
MSI       45
Acer      34
Lenovo    24
HP        14
GIGABYTE  8
LG        8
ACER      8
ASUS      8
MacBook   7
Macbook   5
Microsoft 3
Fujitsu   1
Name: brand, dtype: int64
```

Vì vậy cần in hoa để đồng bộ các giá trị lặp

```
[10]: df['brand'] = df['brand'].str.upper()
df['brand'].value_counts()
```

```
[10]: ASUS      64
DELL      52
MSI       45
ACER      42
LENOVO    24
HP        14
MACBOOK   12
GIGABYTE  8
LG        8
MICROSOFT 3
FUJITSU   1
Name: brand, dtype: int64
```

- Kiểm tra biến **cpu**:

```
[9]: print("Before clean 'cpu':")
print('*****')
print(df['cpu'].value_counts())

Before clean 'cpu':
*****
i7 1165G7 2.8GHz           15
Intel Core i5-1135G7 4.2GHz 15
Intel Core i7-11800H 4.60 GHz 13
i5 1135G7 2.4GHz           13
AMD Ryzen 5-5500U 4 GHz     12
.
.
.
AMD Ryzen 5-3450U 3.5 GHz    1
Intel Core i5-10th-gen 4.30 GHz 1
i5 11320H 3.2GHz           1
Ryzen 7 5700U 1.8GHz        1
AMD Ryzen 5 5500U 4.0 GHz    1
Name: cpu, Length: 87, dtype: int64
***
```

Nhận thấy các giá trị hiển thị bao gồm:

- **Loại CPU:** Intel Core i5, AMD Ryzen 5, ...
- **Xung nhịp:** 2.8GHz, 4.2GHz, ...

Vì vậy tiến hành trích xuất thêm cột dữ liệu **cpu_GHz** và đồng bộ các giá trị ở **cpu** và đồng thời thêm biến **cpu_brand**:

```
[13]: df = clean_cpu(df)
df[['cpu', 'cpu_GHz']].sample(5)
```

	cpu	cpu_GHz
70	Intel i7	4.6
155	Apple M1	3.2
33	Ryzen 7	4.4
112	Intel i5	4.2
141	Intel i7	3.3

```
[14]: df['cpu_brand'] = df['cpu'].str.extract(r'^(\w+)')
df['cpu_brand'].value_counts()
```

```
[14]: Intel      209
Ryzen      54
Apple      10
Name: cpu_brand, dtype: int64
```

- Kiểm tra biến **gpu**:

```
[32]: df['gpu'].value_counts()

[32]: Intel Iris Xe Graphics           48
      Intel Iris Xe                  32
      AMD Radeon Graphics           22
      Intel UHD Graphics             20
      RTX 3050 4GB                  19
      RTX 3050Ti 4GB                8
      NVIDIA GeForce RTX 3050 4GB & Intel UHD Graphics 7
      NVIDIA GeForce MX330 2 GB & Intel Iris Xe Graphics 6
      RTX 3060 6GB                  6
      M1 8-Core GPU                 6
      NVIDIA GeForce GTX 1650 4 GB & AMD Radeon Graphics 5
      M1 7-Core GPU                 4
      GTX 1650 4GB                  4
      NVIDIA GeForce RTX 3050Ti 4 GB & Intel UHD Graphics 4
      NVIDIA GeForce RTX 3050 4GB & Intel Iris Xe Graphics 3
      RTX 3060 Max-Q 6GB             3
      AMD Radeon RX 5500M, 4GB       3
      NVIDIA GeForce RTX 3050Ti 4 GB 3
      NVIDIA GeForce RTX 3050 4GB & AMD Radeon Graphics 3
      Intel UHD Graphics 600           3
      Intel Iris Plus Graphics       3
      Intel UHD Graphics 605           3
      NVIDIA GeForce GTX 1650 4GB GDDR6 / AMD Radeon Graphics 3
      RTX 3070 8GB                  3
      NVIDIA GeForce RTX 3050 4GB GDDR6 / AMD Radeon Graphics 2
      NVIDIA GeForce RTX 3050Ti 4GB GDDR6 / AMD Radeon Graphics 2
      Radeon                         2
      NVMe SSD                      1
```

Nhận thấy có sự trùng lặp vì mỗi shop sẽ có 1 cách hiển thị khác nhau

VD: NVIDIA GeForce RTX 3060 giống với RTX 3060

Vì vậy cần đồng bộ các giá trị trong **gpu** và đồng thời thêm biến **gpu_brand**:

```
[19]: df['gpu'].value_counts()
```

```
[19]: Intel Iris           93
      NVIDIA GeForce RTX 78
      AMD Radeon          34
      Intel UHD            32
      NVIDIA GeForce GTX 22
      M1 GPU              10
      NVIDIA GeForce MX   4
      Name: gpu, dtype: int64
```

```
[20]: df['gpu_brand'] = df['gpu'].str.extract(r'^(\w+)')
      df['gpu_brand'].value_counts()
```

```
[20]: Intel      125
      NVIDIA    104
      AMD       34
      M1        10
      Name: gpu_brand, dtype: int64
```

- Thay đổi thứ tự các cột và hiển thị Dataset sau khi đã được “làm sạch”

```
[22]: column_names = ['id', 'brand', 'cpu', 'cpu_GHz', 'cpu_brand', 'ram', 'scrsizes', 'gpu', 'gpu_brand', 'memory', 'drive_type', 'opersystem', 'weight', 'since', 'shop', 'price', 'url']
df = df.reindex(columns=column_names)
df
```

		id	brand	cpu	cpu_GHz	cpu_brand	ram	scrsizes	gpu	gpu_brand	memory	drive_type	opersystem	weight	since	shop	price	url
0	0	ACER	Intel i7	4.6	Intel	8	15.6	NVIDIA GeForce RTX	NVIDIA	512	SDD	Window	2.2	2021	FPTShop	29999000	https://ftpshop.com.vn/may-tinh-xach-tay/acer-...	
1	1	MSI	Intel i5	4.5	Intel	8	14.0	Intel Iris	Intel	512	SDD	Window	1.3	2021	FPTShop	18799000	https://ftpshop.com.vn/may-tinh-xach-tay/msi-m...	
2	2	DELL	Intel i3	4.1	Intel	4	15.6	Intel UHD	Intel	256	SDD	Window	1.7	2021	FPTShop	15299000	https://ftpshop.com.vn/may-tinh-xach-tay/dell-...	
3	3	DELL	Intel i3	4.1	Intel	8	15.6	Intel UHD	Intel	256	SDD	Window	2.0	2021	FPTShop	15599000	https://ftpshop.com.vn/may-tinh-xach-tay/dell-...	
4	4	DELL	Intel i5	4.2	Intel	4	15.6	Intel Iris	Intel	512	SDD	Window	1.7	2021	FPTShop	19999000	https://ftpshop.com.vn/may-tinh-xach-tay/dell-...	
...	
268	268	DELL	Ryzen 5	4.0	Ryzen	8	14.0	AMD Radeon	AMD	256	SDD	Window	1.4	2021	Phongvu	19090000	https://phongvu.vn/may-tinh-xach-tay-laptop-de...	
269	269	DELL	Intel i5	4.5	Intel	8	14.0	Intel Iris	Intel	512	SDD	Window	1.4	2021	Phongvu	23990000	https://phongvu.vn/may-tinh-xach-tay-laptop-de...	
270	270	DELL	Ryzen 5	4.0	Ryzen	8	15.6	AMD Radeon	AMD	512	SDD	Window	1.7	2020	Phongvu	21790000	https://phongvu.vn/may-tinh-xach-tay-laptop-de...	
271	271	DELL	Ryzen 5	4.0	Ryzen	8	14.0	AMD Radeon	AMD	512	SDD	Window	1.4	2020	Phongvu	20890000	https://phongvu.vn/may-tinh-xach-tay-laptop-de...	
272	272	DELL	Intel i5	4.5	Intel	8	14.0	Intel Iris	Intel	512	SDD	Window	1.4	2020	Phongvu	23390000	https://phongvu.vn/may-tinh-xach-tay-laptop-de...	

273 rows × 17 columns

Sourse **SourceCode CleanData:** [clickhere](#)

- Lưu lại vào file **laptop_clean.csv** để chuẩn bị tiến hành phân tích:

```
[23]: df.to_csv('laptop_clean.csv', index=False)
```

CHƯƠNG II. PHÂN TÍCH DỮ LIỆUⁱⁱⁱ

1. Tổng quan về dữ liệu:

- Nguồn dữ liệu về laptop thu thập được từ 3 trang web:

<https://ftpshop.com.vn/>
<https://www.dienmayxanh.com/>
<https://phongvu.vn/>

- Có 17 trường dữ liệu:

id	ID laptop
brand	Thương hiệu của laptop
cpu	Bộ xử lý trung tâm.
cpu_GHz:	Tốc độ xử lý CPU hay tần số tính toán và làm việc của nó được đo bằng đơn vị GHz
cpu_brand	Hãng sản xuất CPU
ram	Bộ nhớ tạm (đơn vị GB)
scrsize	Kích thước màn hình laptop (đơn vị inch)
gpu	Đơn vị xử lý đồ họa chuyên dụng, nhiệm vụ chính là tăng tốc, xử lý đồ họa
gpu_brand	Hãng sản xuất GPU
memory	Dung lượng lưu trữ (đơn vị GB)
drive_type	Loại ổ đĩa
opersystem	Hệ điều hành
since	Năm sản xuất
shop	Cửa hàng phân phối
price	Giá sản phẩm
url	Đường dẫn đến sản phẩm

- Gồm có 273 hàng và 17 cột

A data.frame: 5 x 17																
id	brand	cpu	cpu_GHz	cpu_brand	ram	scrsize	gpu	gpu_brand	memory	drive_type	opersystem	weight	since	shop	price	url
164	ACER	Intel i7	2.3	Intel	16	15.6	NVIDIA GeForce RTX	NVIDIA	1000	SDD	Window	2.3	2021	Dienmayxanh	36990000	https://www.dienmayxanh.com/laptop/acer-predator-helios-ph315-54-75yd-i7-nhqq2sv002?src=osp
50	MACBOOK	Apple M1	3.2	Apple	16	13.3	M1 GPU	M1	256	SDD	Mac OS	1.4	2020	FPTShop	39999000	https://ftpshop.com.vn/may-tinh-xach-tay/macbook-pro-13-2020-touch-bar-m1-ram-16gb7dung-luong=256gb
181	GIGABYTE	Intel i5	2.5	Intel	16	15.6	NVIDIA GeForce RTX	NVIDIA	1000	SDD	Window	2.2	2021	Dienmayxanh	27990000	https://www.dienmayxanh.com/laptop/gigabyte-gaming-g5-i5-5s11130sh?src=osp
29	DELL	Intel i5	4.4	Intel	8	15.6	Intel Iris	Intel	256	SDD	Window	1.7	2021	FPTShop	22999000	https://ftpshop.com.vn/may-tinh-xach-tay/dell-inspiron-n5510-i5-11300h
231	MSI	Intel i3	3.0	Intel	8	14.0	Intel UHD	Intel	1000	SDD	Window	1.3	2021	Dienmayxanh	13950000	https://www.dienmayxanh.com/laptop/msi-modern-14-b11mol-i3-813vn?src=osp

2. Thống kê mô tả (EDA & Trực quan hóa dữ liệu)^{iv}

- Top10 sản phẩm có giá cao nhất:

A data.frame: 10 × 3

	price	shop	url
	<int>	<chr>	<chr>
1	76990000	Phongvu	https://phongvu.vn/may-tinh-xach-tay-laptop-asus-rog-flow-gv301qc-k6029t-amd-ryzen-9-5980hs-den-s210902962.html?sku=210902962
2	75190000	Dienmayxanh	https://www.dienmayxanh.com/laptop/msi-gaming-ge66-raider-11uh-i7-259vn?src=osp
3	62490000	Dienmayxanh	https://www.dienmayxanh.com/laptop/msi-gaming-gs66-stealth-11ug-i7-219vn?src=osp
4	58190000	Dienmayxanh	https://www.dienmayxanh.com/laptop/msi-gaming-ge66-raider-11ug-i7-258vn?src=osp
5	56990000	Dienmayxanh	https://www.dienmayxanh.com/laptop/dell-xps-13-9310-i7-jgnh62?src=osp
6	52900000	Phongvu	https://phongvu.vn/may-tinh-xach-tay-laptop-lg-gram-2021-17z90p-g-ah76a5-i7-1165g7-bac-s210500992.html?sku=210500992
7	52140000	Dienmayxanh	https://www.dienmayxanh.com/laptop/lg-gram-17-i7-17z90pgah78a5?src=osp
8	51490000	Dienmayxanh	https://www.dienmayxanh.com/laptop/msi-gp76-11ug-i7-435vn?src=osp
9	50999000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/acer-triton-gaming-pt315-53-71dj-i7-11800
10	50900000	Phongvu	https://phongvu.vn/may-tinh-xach-tay-laptop-lg-gram-2021-16z90p-g-ah75a5-i7-1165g7-den-s210500990.html?sku=210500990

- Top10 sản phẩm có giá rẻ nhất:

A data.frame: 10 × 3

	price	shop	url
	<int>	<chr>	<chr>
1	9490000	Phongvu	https://phongvu.vn/may-tinh-xach-tay-laptop-lenovo-ideapad-1-11igl05-81vt006fvn-n5030-xam-s211000648.html?sku=211000648
2	9699000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/asus-flip-br1100fka-bp0531t-n4500
3	10499000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/asus-flip-br1100fka-bp0660t-n6000
4	10999000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-inspiron-n3502-celeron-n4020-win-10-nk
5	10999000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-inspiron-n3510-celeron-n4020-win-10-nk
6	10999000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-inspiron-n3510-n4020-nk
7	11699000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-inspiron-n3510-pentium-n5030-win-10-nk
8	11699000	FPTShop	https://fptshop.com.vn/may-tinh-xach-tay/dell-inspiron-n3502-pentium-n5030-win-10-nk
9	11990000	Phongvu	https://phongvu.vn/may-tinh-xach-tay-laptop-acer-aspire-3-a3155637dv-nxhs5sv001-i31005g1-den-s200400313.html?sku=200400313
10	12990000	Phongvu	https://phongvu.vn/may-tinh-xach-tay-laptop-acer-aspire-3-a315-58-3939-nx-addsv-001-i3-1115g4-den-s210801044.html?sku=210801044

2.1 Phân tích biến định tính (Categorical)

- Các biến định tính và giá trị của chúng:

```
$brand
[1] "ACER"      "MSI"       "DELL"      "LENOVO"    "ASUS"      "HP"
[7] "MACBOOK"   "MICROSOFT" "GIGABYTE" "FUJITSU"   "LG"

$cpu
[1] "Intel i7"   "Intel i5"   "Intel i3"   "AMD Ryzen 5"
[5] "AMD Ryzen 7" "Intel Celeron" "Intel Pentium" "AMD Ryzen 9"
[9] "Apple M1"    "AMD Ryzen 3"

$cpu_brand
[1] "Intel" "AMD"  "Apple"

$gpu
[1] "NVIDIA GeForce RTX" "Intel Iris"      "Intel UHD"
[4] "NVIDIA GeForce GTX"  "AMD Radeon"   "M1 GPU"
[7] "NVIDIA GeForce MX"

$gpu_brand
[1] "NVIDIA" "Intel"  "AMD"    "M1"

$drive_type
[1] "SDD"

$opersystem
[1] "Window" "Mac OS"

$shop
[1] "FPTShop" "Dienmayxanh" "Phongvu"
```

Nhận xét:

- Đối với biến **drive_type** chỉ có duy nhất 1 giá trị: SDD
- Xét biến **opersystem** (hệ điều hành) gồm 2 giá trị:
 - Mac OS: Hệ điều hành dành riêng cho máy tính của Apple (macbook)
 - Còn laptop của các hãng còn lại sử dụng hệ điều hành Window

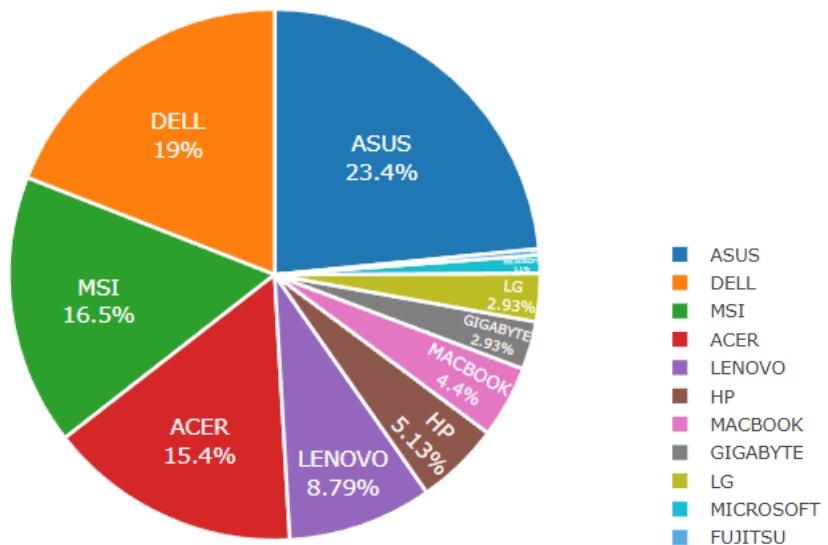
Kết luận: 2 biến này hoàn toàn không ảnh hưởng đến giá cả, vì toàn bộ ổ đĩa của các sản phẩm trong dataset đều là SDD và hệ điều hành phụ thuộc vào thương hiệu của laptop (brand).

2.1.1 Thương hiệu máy tính (brand)

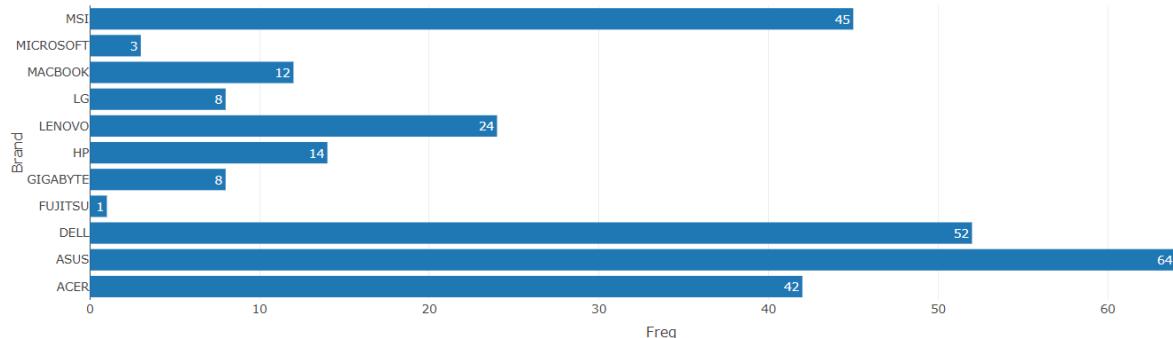
A data.frame: 11 × 3

Var1	Freq	Percent
<fct>	<int>	<dbl>
ASUS	64	23.4432234
DELL	52	19.0476190
MSI	45	16.4835165
ACER	42	15.3846154
LENOVO	24	8.7912088
HP	14	5.1282051
MACBOOK	12	4.3956044
GIGABYTE	8	2.9304029
LG	8	2.9304029
MICROSOFT	3	1.0989011
FUJITSU	1	0.3663004

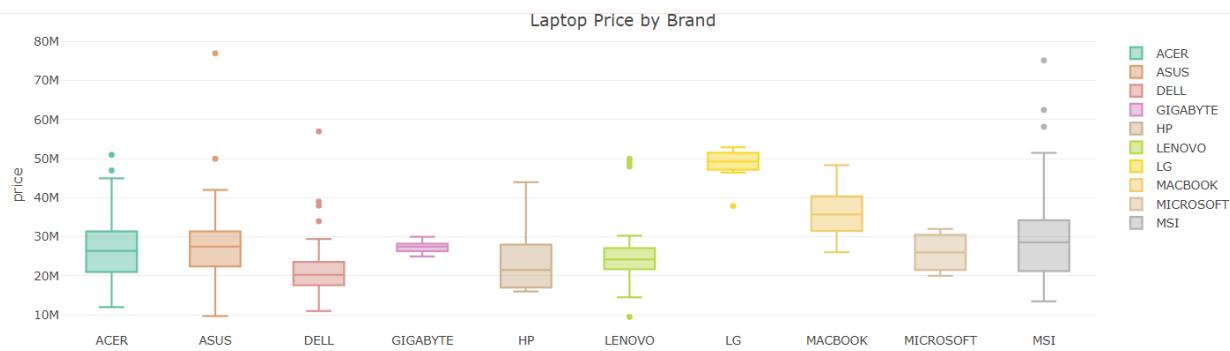
What's the percentage each brand?



Which brand is the most frequent?



- Trong bộ dữ liệu thu thập được, có 10 thương hiệu Laptop:
 - ASUS: 23.4% (Phổ biến nhất)
 - DELL: 19.0%
 - MSI: 16.5%
 - ACER: 15.4%
 - LENOVO: 8.8%
- Hãng ít phổ biến:
 - FUJITSU (1 sản phẩm)
 - MICROSOFT (3 sản phẩm)



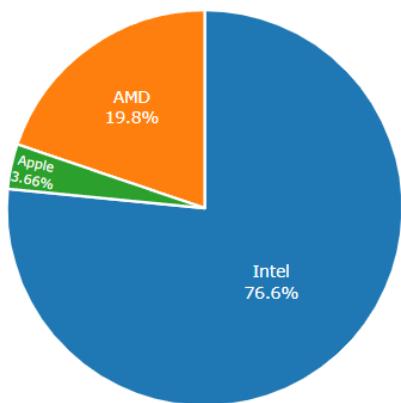
- Đa số các hãng sẽ có khoảng giá trung bình 20-30tr đồng.
- Tuy các thương hiệu đều có khoảng giá tương đối gần nhau, nhưng một số trong chúng có giá trị cao bất thường (sản phẩm cao cấp so với các sản phẩm còn lại), chẳng hạn như ASUS hoặc MSI:
 - ASUS: khoảng 77tr đồng
 - MSI: khoảng 75.2tr đồng
- Sản phẩm của LG có giá trung bình cao nhất - Khoảng 49 triệu đồng, nhưng số lượng máy tính của công ty này (chỉ 8 chiếc).
- Sản phẩm MACBOOK của APPLE có giá trung bình cao hơn mặt bằng chung (khoảng 35tr đồng) đồng thời khoảng giá sản phẩm tương đối hẹp (từ 25-50tr) bởi vì đây là dòng sản phẩm cao cấp, hướng đến tệp khách hàng đắt tiền.
- DELL là một trong những hãng phổ biến nhưng sản phẩm của DELL có giá trung bình tương đối thấp hơn so với các hãng khác (khoảng 20tr đồng) vì các máy tính của Dell chủ yếu phục vụ cho học tập và công việc.
- MSI là hãng chuyên cung cấp dòng sản phẩm laptop GAMING nên giá của hầu hết sản phẩm sẽ tương đối cao hơn so với mặt bằng chung

Kết luận: Thương hiệu dường như có thể ảnh hưởng đến giá thành sản phẩm. Đồng thời thương hiệu cũng là điều mà nhiều người quan tâm gửi gắm niềm tin khi lựa chọn laptop.

2.1.2 Bộ xử lý (cpu, cpu_brand)

CPU brand

What's the percentage of each CPU brand?

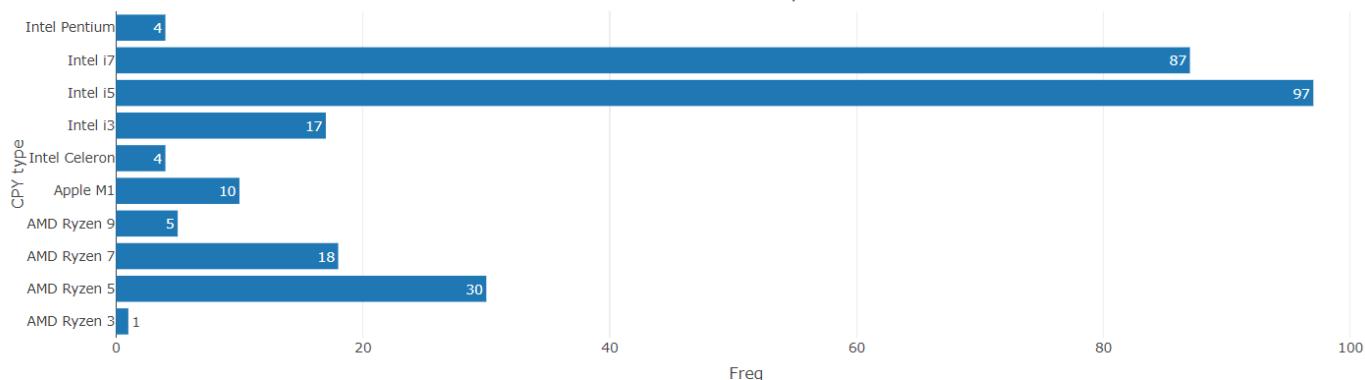


Trong bộ dữ liệu thu thập được, có 3 thương hiệu CPU

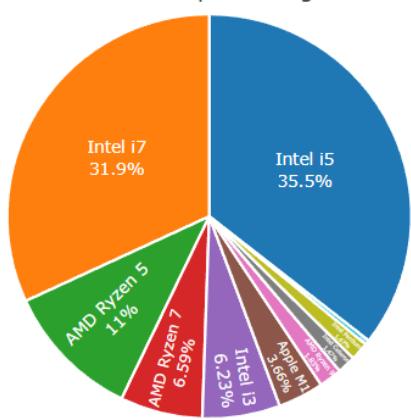
- Intel chiếm 76.6% là hàng phổ biến nhất
- AMD: chiếm 19.8%
- Apple: chiếm 3.7%

CPU core

Which CPU is the most frequent?



What's the percentage of each CPU?



Trong bộ dữ liệu thu thập được, có 10 loại CPU

- CPU của hãng Intel:
 - Intel i7: chiếm 31.9% (Top Phổ biến)
 - Intel i5: chiếm 35.5% (Top Phổ biến)
 - Intel i3: chiếm 6.23%
 - Intel Celeron: chiếm 1.46%
 - Intel Pentium: chiếm 1.46%
- CPU của hãng AMD Ryzen:
 - AMD Ryzen 9: chiếm 1.83%
 - AMD Ryzen 7: chiếm 6.6%
 - AMD Ryzen 5: chiếm 11.0% (phổ biến nhất trong dòng này)
 - AMD Ryzen 3: chiếm 0.36% (1 sản phẩm)
- CPU của hãng Apple:
 - Apple M1: chiếm 3.66%



Đánh giá sản phẩm theo hãng CPU

- Những laptop có CPU của AMD và Intel có khoảng giá chủ yếu từ 20-30tr. Các dòng CPU của 2 hãng này đều phổ biến trên thị trường laptop hiện nay.
- Laptop có CPU của Apple có mức giá cao hơn so với 2 hãng CPU còn lại (30-40tr). Vì Apple là thương hiệu cao cấp nên điều đó bảo chứng cho mức giá sản phẩm cao trong thị trường để phục vụ tệp khách hàng cao cấp.
- Tuy nhiên sản phẩm mức sử dụng CPU của 2 hãng AMD và Intel lại có giá cao vượt trội: Chúng tỏ 2 hãng này có phân khúc thị trường mạnh, phục vụ nhiều tệp khách hàng.

Đánh giá sản phẩm theo loại CPU

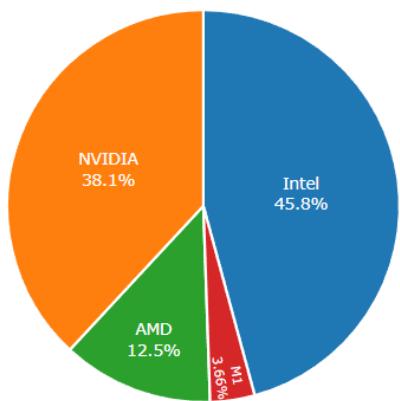
- Đời CPU càng cao, mức giá càng tăng:
 - Ryzen5 < Ryzen7 < Ryzen9
 - Celeron < Pentium < i3 < i5 < i7
- Những laptop có CPU là Intel Celeron và Intel Pentium có mức giá rẻ hơn rất nhiều so với những loại CPU khác, điều đó chứng tỏ 2 dòng này đều là dòng CPU dành riêng cho laptop giá rẻ.
- Ryzen 9 và Intel i7 đều có mức giá cao -> 2 loại CPU hiện đại nhất

Kết luận: Chất lượng CPU (thương hiệu và loại cpu) có ảnh hưởng đến giá laptop (tỷ lệ thuận)

2.1.3 GPU (gpu, gpu_brand)

GPU brand

What's the percentage of each GPU brand?



Trong bộ dữ liệu thu thập được, có 4 thương hiệu GPU trong đó 2 hãng phổ biến nhất là Intel và NVIDIA

- Intel: chiếm 45.8%
- NVIDIA: chiếm 38.1%
- AMD: chiếm 12.5%
- Apple: chiếm 3.7%

GPU core

Trong bộ dữ liệu thu thập được, có 7 loại GPU (4 hãng)

- GPU của hãng Intel:
 - Intel Iris: chiếm 34.1% (Phổ biến nhất trong dòng này)
 - Intel UHD: chiếm 11.7%
- GPU của hãng NVIDIA:
 - NVIDIA GeForce RTX: chiếm 28.6% (Phổ biến nhất trong dòng này)
 - NVIDIA GeForce GTX: chiếm 8.06%
 - NVIDIA GeForce MX: chiếm 1.46%
- GPU của hãng AMD:
 - AMD Radeon: chiếm 12.5%
- GPU của hãng Apple (M1):
 - Apple M1: chiếm 3.66%

A data.frame: 7 × 3

Var1	Freq	Percent
	<fct>	<int>
Intel Iris	93	34.065934
NVIDIA GeForce RTX	78	28.571429
AMD Radeon	34	12.454212
Intel UHD	32	11.721612
NVIDIA GeForce GTX	22	8.058608
M1 GPU	10	3.663004
NVIDIA GeForce MX	4	1.465201



Đánh giá sản phẩm theo hãng GPU

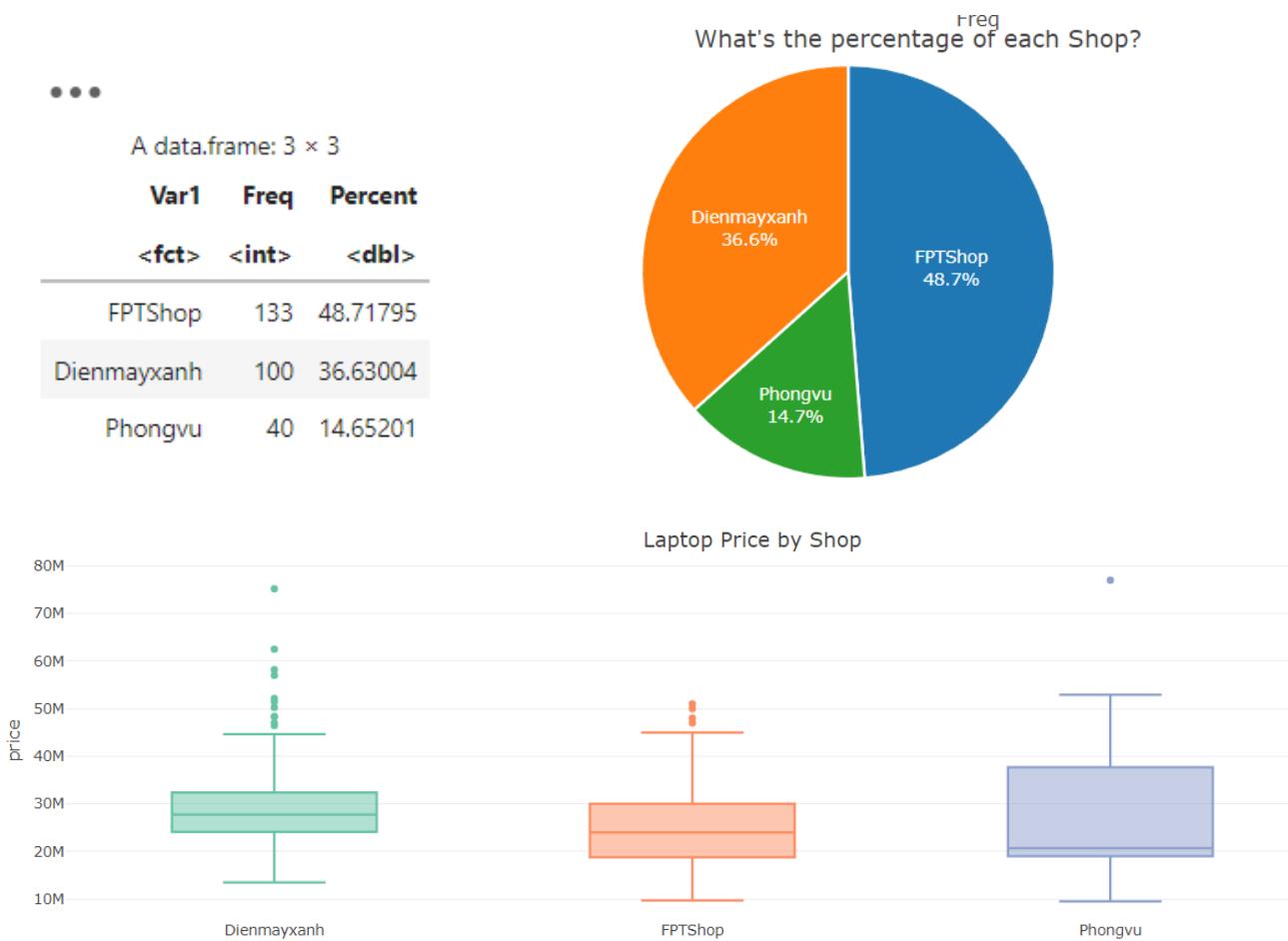
- Những laptop có GPU của AMD và Intel có khoảng giá chủ yếu từ 20-30tr. Các dòng GPU của 2 hãng này đều phổ biến trên thị trường laptop hiện nay.
- Laptop có GPU của Apple có mức giá cao hơn so với 2 hãng GPU còn lại (30-40tr).
- Sản phẩm mức sử dụng GPU của hãng NVIDIA có giá cao vượt trội -> Chuyên biệt cho những dòng máy cần cấu hình đồ họa cao như dòng GAMING hay WORKSTATION

Đánh giá sản phẩm theo loại GPU

- Giữa các dòng GPU không có sự chênh lệch giá quá đáng kể, giá dao động trong khoảng 25-35tr.
- NVIDIA GeForce RTX là dòng GPU có mức giá cao nhất còn Intel UHD có mức giá thấp nhất

Kết luận: Nhận thấy 1 số hãng CPU đồng thời cung cấp cả GPU cho sản phẩm laptop vì vậy chúng ta cần kiểm định mối quan hệ giữa gpu và cpu sẽ ảnh hưởng như thế nào đến giá sản phẩm.

2.1.4 Cửa hàng phân phối (shop)



Trong bộ dữ liệu, các sản phẩm lấy từ 3 Shop:

- **Điện máy xanh:** có 100 sản phẩm chiếm 36.6%
- **FPTshop:** có 133 sản phẩm chiếm 48.7%
- **Phong vũ:** có 40 sản phẩm chiếm 14.7%

Phân phối giá sản phẩm laptop theo shop:

- Tuy FPTShop có số lượng sản phẩm chiếm nhiều hơn so với số lượng sản phẩm phân phối bởi Dienmayxanh và Phongvu, nhưng khoảng giá của Fpt lại hẹp hơn (tầm giá chủ yếu 20-30tr)
- ĐMX và PhongVu lại có bán sản phẩm với giá cao vượt trội (trên 50tr) -> Có bán những sản phẩm cao cấp đắt tiền có cấu hình khủng (như các dòng laptop GAMING hay WORKSTATION)

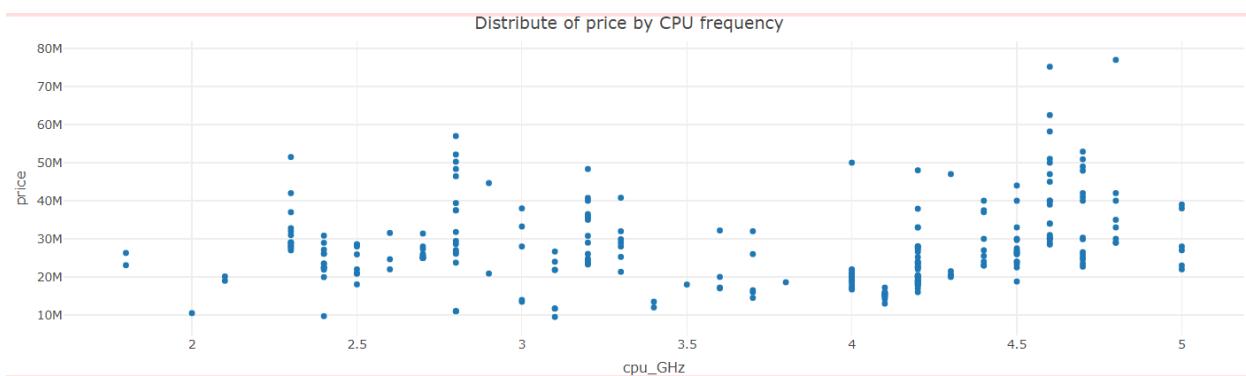
Kết luận: Nhận thấy phân phối giá giữa các cửa hàng trong có vẻ tương đối giống nhau vì vậy chúng ta cần kiểm định giả thiết này kỹ càng hơn.

2.2 Phân tích biến định lượng (Numerical)

- Sử dụng hàm **summary()** để tính nhanh các giá trị thống kê:

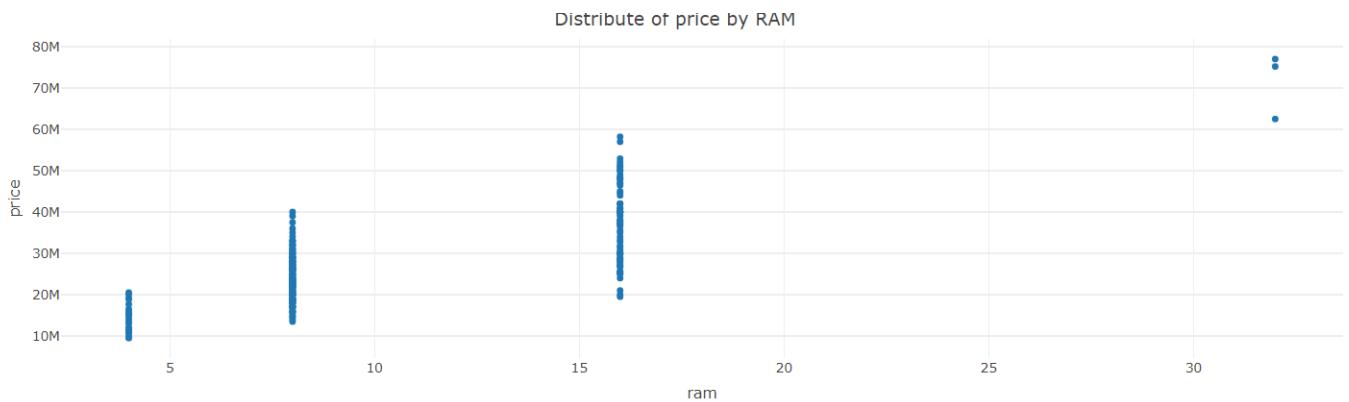
```
cpu_GHz      ram      scrsizes      memory      weight
Min. 1.800  Min. 4.0  Min. 11.60  Min. 64.0  Min. 0.80
1st Qu.:2.800 1st Qu.: 8.0  1st Qu.:14.00  1st Qu.: 512.0  1st Qu.:1.40
Median :4.000  Median : 8.0  Median :15.60  Median : 512.0  Median :1.70
Mean   :3.688  Mean   :10.2  Mean   :14.87  Mean   : 595.5  Mean   :1.74
3rd Qu.:4.400 3rd Qu.:16.0  3rd Qu.:15.60  3rd Qu.:1000.0  3rd Qu.:2.10
Max.   :5.000  Max.   :32.0  Max.   :17.30  Max.   :1000.0  Max.   :2.90
since      price
Min. :2018  Min. : 9490000
1st Qu.:2021 1st Qu.:20290000
Median :2021  Median :25990000
Mean   :2021  Mean   :27556059
3rd Qu.:2021 3rd Qu.:31390000
Max.   :2021  Max.   :76990000
                           ...
[1] "Mode of cpu_GHz: 4.2"
[1] "Mode of RAM: 8"
[1] "Mode of scrsizes: 15.6"
[1] "Mode of memory: 512"
[1] "Mode of weight: 1.4"
[1] "Mode of since: 2021"
[1] "Mode of price: 39999000"
```

2.2.1 CPU frequency (GHz)

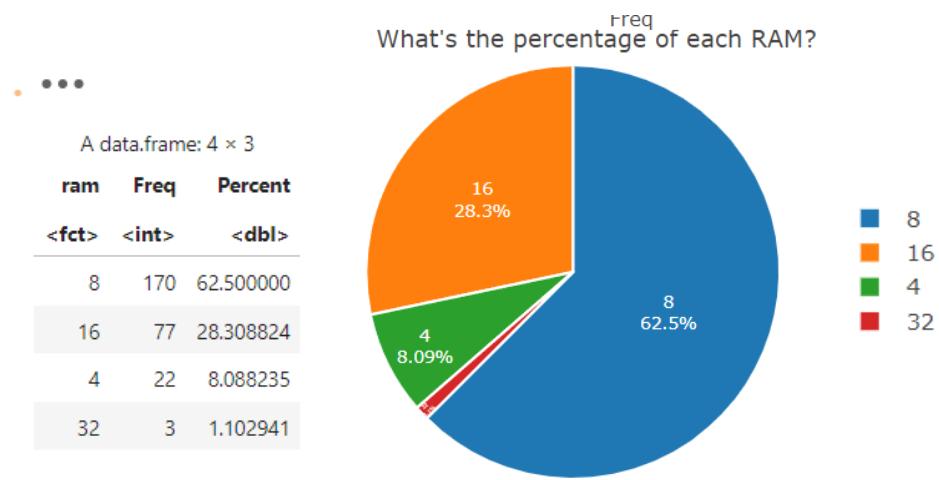


Phân phối giá sản phẩm theo tần số CPU (GHz) tương đối dàn đều, điều này chứng tỏ giá sản phẩm phụ thuộc nhiều hơn vào những thông số khác

2.2.2 RAM



Dung lượng RAM là một biến định lượng nhưng chỉ có 4 loại dung lượng cố định: 4GB, 8GB, 16GB và 32GB. Vì vậy đây là biến rời rạc.

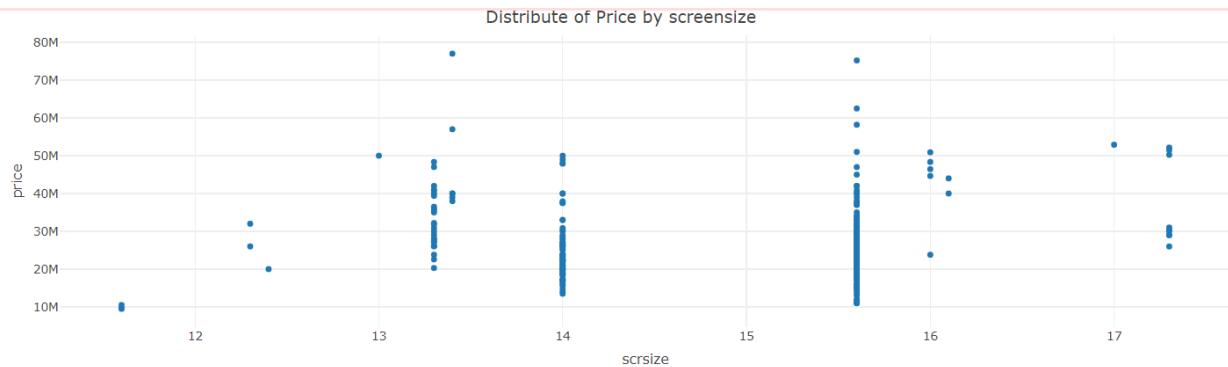


Trong bộ dữ liệu:

- Có 4 loại dung lượng RAM(GB):4GB, 8GB, 16GB và 32GB
- Trong đó phổ biến nhất là loại 8GB (chiếm 62.5%) và 16GB (chiếm 28.3%)
- Dung lượng RAM càng lớn giá sản phẩm càng đắt: các laptop có dung lượng RAM trên 16GB có giá trên 20 triệu, và ngược lại laptop dung lượng RAM 4GB có giá dưới 20 triệu

Kết luận:RAM là 1 thông số quan trọng vì càng nhiều RAM, máy càng "mạnh" và đồng thời giá của laptop cũng sẽ tỷ lệ thuận theo dung lượng RAM

2.2.3 Kích thước màn hình (scrsizes)

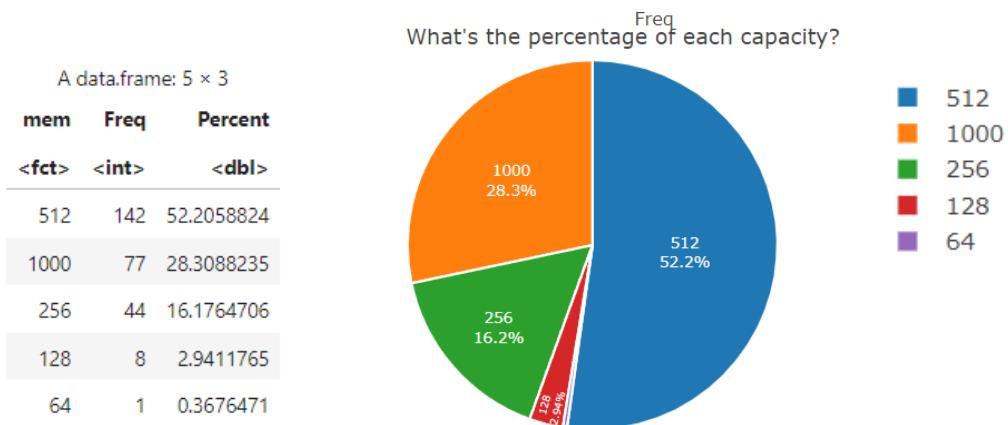


Nhận xét:

- Tuy kích thước màn hình (scrsizes) là một biến định lượng, nhưng hầu hết các máy tính xách tay đều có kích thước cố định (ví dụ: 13.3inch, 14inch,...) vì vậy đây là biến rời rạc.
- Kích thước phổ biến thường là 13.3inch, 14inch và 15.6inch
- Những dòng laptop kích thước màn hình bé (bé hơn 14inch) có giá thấp hơn so với những laptop kích thước màn hình lớn. [Ngoại lệ có 1 sản phẩm 13.3inch nhưng có giá 77tr](#).

Kết luận: Giá thành giữa các dòng laptop có kích thước khác nhau có sự chênh lệch nhẹ, cần kiểm định để làm rõ hơn.

2.2.4 Dung lượng bộ nhớ (memory)



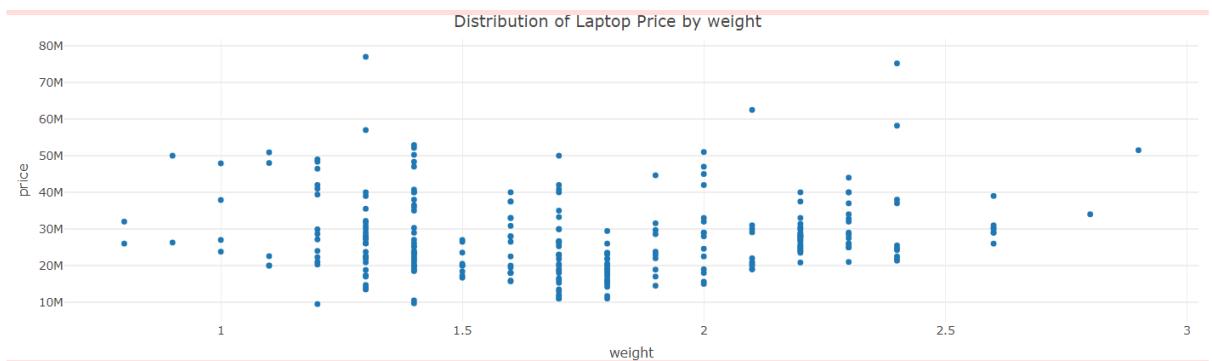
- Tương tự như kích thước màn hình, dung lượng bộ nhớ cũng là một biến định lượng nhưng chỉ có 5 loại dung lượng bộ nhớ cố định: 64GB, 128GB, 256GB, 512GB và 1000GB (1TB). Vì vậy đây là biến rời rạc.
- Trong đó phổ biến nhất là loại **512GB (chiếm 52.2%)** kế đến là 1TB(28.3%) và 256GB(16.2%)



- Giá laptop có xu hướng tăng theo dung lượng bộ nhớ của laptop.

Kết luận: Dung lượng bộ nhớ là 1 thông số quan trọng vì dung lượng càng lớn sẽ giúp máy lưu trữ được càng nhiều dữ liệu, vì vậy giá của laptop cũng sẽ tỷ lệ thuận theo dung lượng bộ nhớ

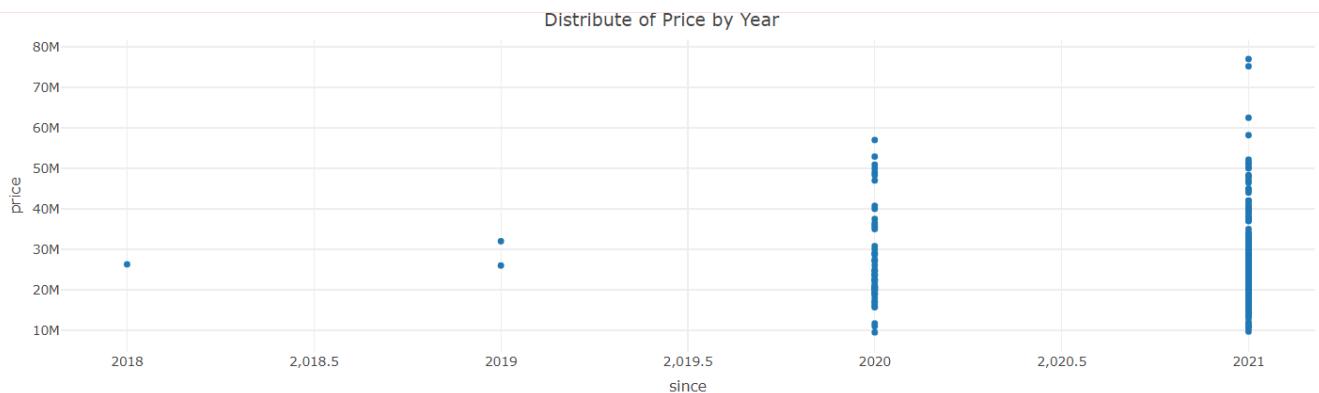
2.2.5 Trọng lượng (weight)



Nhận xét:

- Hầu hết ai cũng muốn máy tính xách tay nhẹ nhất có thể, nhưng các thành phần khác như card đồ họa và ổ đĩa,... sẽ khiến laptop tay nặng hơn.
- Trọng lượng trung bình của laptop sẽ vào khoảng 1.5-2Kg và đồng thời những máy tính có trọng lượng trong khoảng này thường có giá rẻ hơn (mức độ phổ biến, bình dân)
- Những laptop trọng lượng bé hơn mức trung bình có giá cao hơn điều này chứng tỏ những laptop này thuộc dòng máy cao cấp như dòng Macbook của Apple, với thiết kế hướng đến tinh gọn, nhẹ nhàng nhưng vẫn đảm bảo độ sang trọng và hiệu năng tốt.
- Những laptop có trọng lượng lớn hơn mức trung bình cũng đồng thời có giá cao hơn vì những laptop có trọng lượng nặng thường là các dòng GAMING và WORKSTATION, có cấu hình cao, đồ sộ.

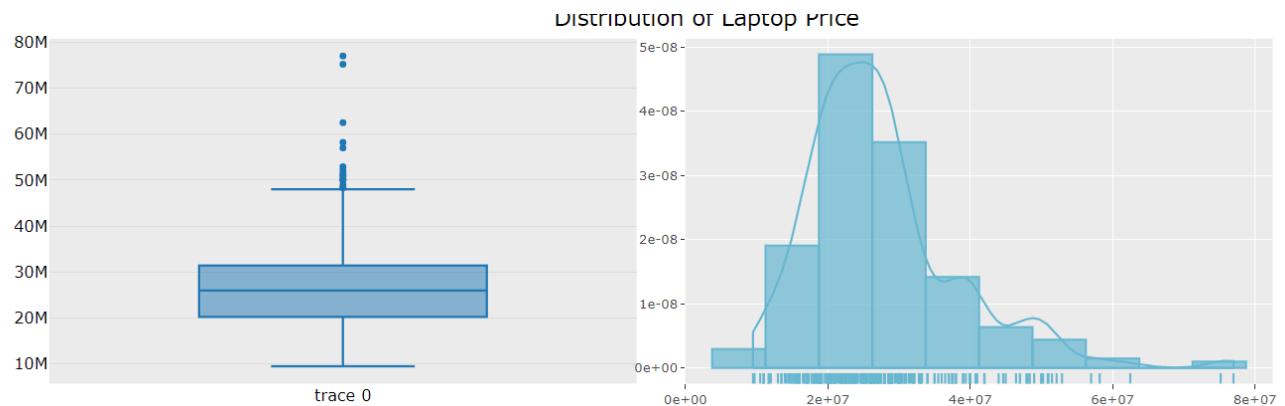
2.2.6 Năm sản xuất (since)



Nhận xét:

- Hầu hết các sản phẩm đều sản xuất trong vòng 2 năm gần đây (2020 và 2021)
- **Giá của sản phẩm không phụ thuộc quá nhiều vào năm sản xuất**

2.2.7 Giá thành sản phẩm (price)



Nhận xét:

- Có thể thấy giá laptop chủ yếu trong khoảng 20-30tr
- Thấp nhất là 9.5tr và cao nhất là 77tr

SourceCode Thông kê mô tả: [clickhere](#)

3 Thống kê suy diễn

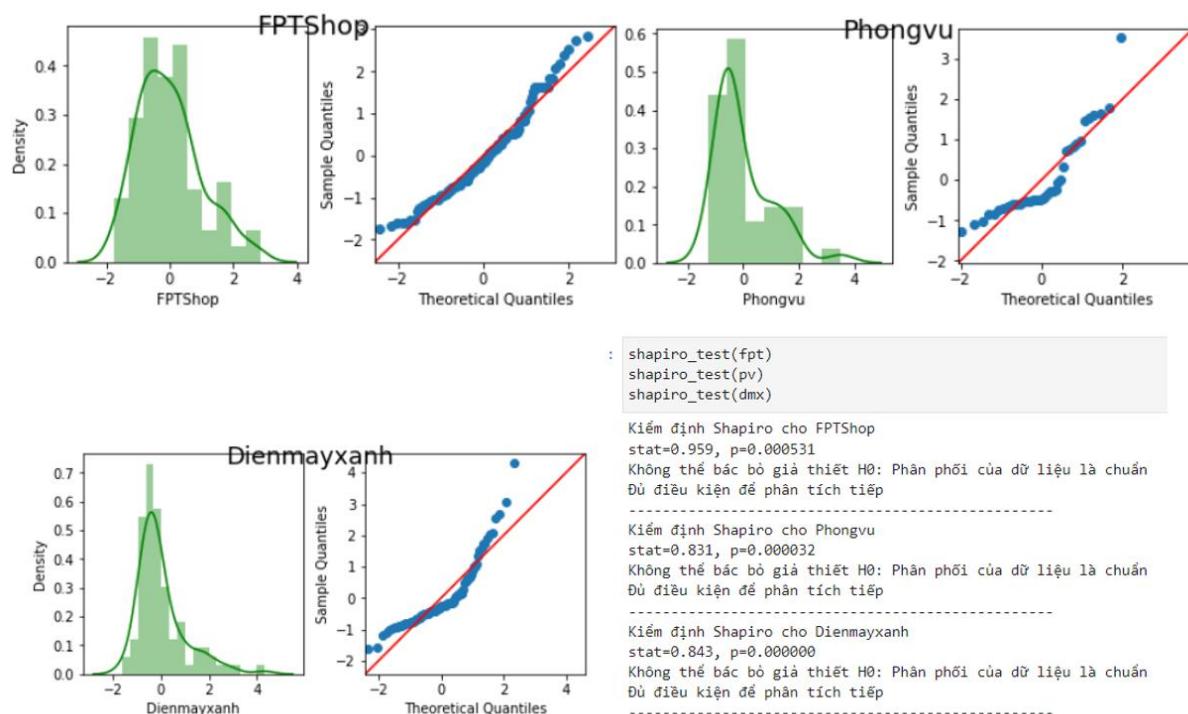
3.1 Sử dụng ANOVA để so sánh giá giữa các cửa hàng phân phối

Sử dụng Shapiro-Wilk test để kiểm tra Phân phối của dữ liệu

Phát biểu giả thiết:

H_0 : Phân phối của nhóm là chuẩn

H_1 : Phân phối của nhóm là không chuẩn



Kiểm định tính thuần nhất phương sai (Bartlett's Test)

Phát biểu các giả thiết:

H_0 : Các nhóm đều có phương sai đồng nhất

H_1 : Có ít nhất 2 nhóm có phương sai khác nhau (lớn)

```
2]: from scipy.stats import bartlett
bartlett = bartlett(fpt, dmx, pv)
print(bartlett)
if bartlett.pvalue > 0.05:
    print('Không thể bác bỏ giả thiết  $H_0$ : Các nhóm đồng nhất về phương sai')
else:
    print('Các nhóm không đồng nhất về phương sai')
```

BartlettResult(statistic=0.004815074822490079, pvalue=0.9975953583825714)
Không thể bác bỏ giả thiết H_0 : Các nhóm đồng nhất về phương sai

Kiểm định ANOVA

Phát biểu giả thiết:

- H_0 : Giá sản phẩm (price) giữa các Cửa hàng (shop) không có sự khác biệt
- H_1 : Giá sản phẩm (price) giữa các Cửa hàng (shop) có sự khác biệt

F_onewayResult(statistic=1.0285388739512266e-30, pvalue=1.0)

Không đủ bằng chứng bác bỏ giả thuyết H_0 . Điều này ngụ ý rằng:

Giá sản phẩm (price) giữa các Cửa hàng (shop) không có sự khác biệt.

Kết luận: Về mặt thống kê, Giá sản phẩm (price) giữa các Cửa hàng (shop) không có sự khác biệt với mức ý nghĩa 5%.

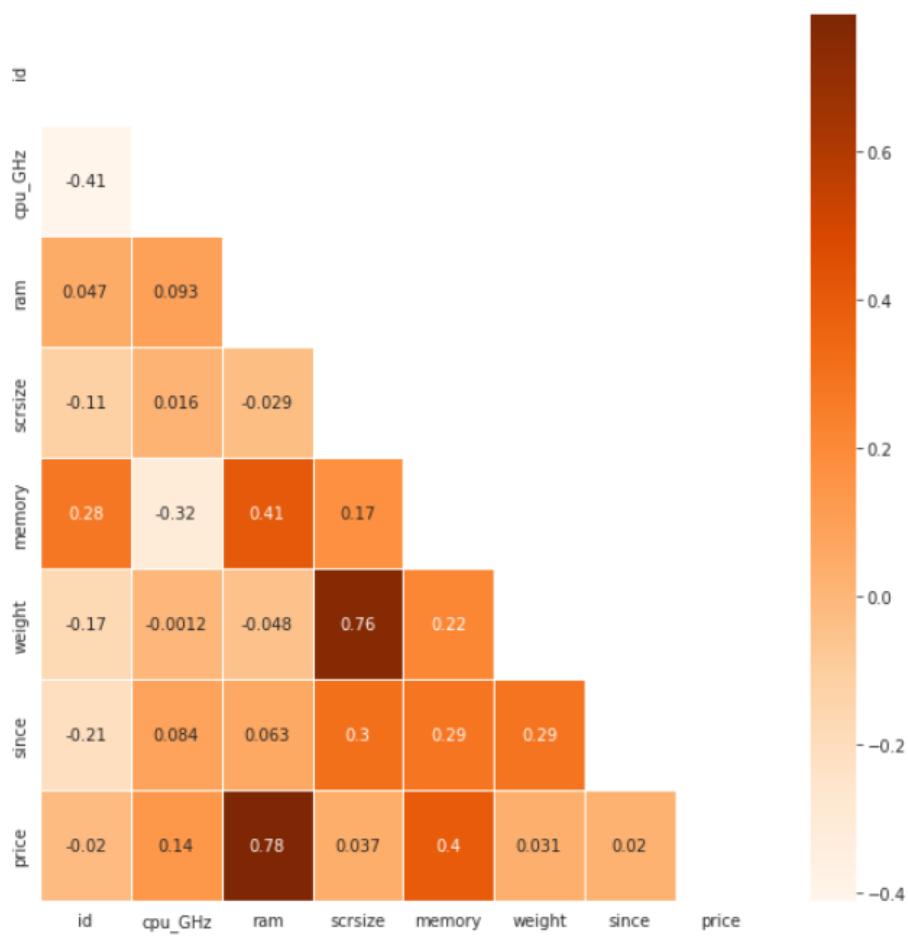
- Giữa các cửa hàng phân phối không có sự khác biệt quá lớn về giá cả.
- Không có sự độc quyền về giá ở thị trường laptop
- Nếu sản phẩm laptop mà bạn mong muốn bị hết hàng hoặc không có ở shop này
bạn có thể tìm kiếm ở shop khác mà vẫn yên tâm về giá cả

3.2 Phân tích mối quan hệ giữa các biến

Phân tích mối quan hệ tương quan:

```
df = df.sort_values(by='price')
corr = df[numerical].corr(method = "pearson")
corr
```

	id	cpu_GHz	ram	scrsize	memory	weight	since	price
id	1.000000	-0.411241	0.047324	-0.113216	0.275464	-0.170441	-0.212669	-0.020470
cpu_GHz	-0.411241	1.000000	0.093140	0.016351	-0.315072	-0.001243	0.084166	0.135903
ram	0.047324	0.093140	1.000000	-0.029363	0.409429	-0.048001	0.063276	0.783801
scrsize	-0.113216	0.016351	-0.029363	1.000000	0.170599	0.755016	0.303132	0.037228
memory	0.275464	-0.315072	0.409429	0.170599	1.000000	0.216019	0.285916	0.399240
weight	-0.170441	-0.001243	-0.048001	0.755016	0.216019	1.000000	0.291478	0.030639
since	-0.212669	0.084166	0.063276	0.303132	0.285916	0.291478	1.000000	0.020238
price	-0.020470	0.135903	0.783801	0.037228	0.399240	0.030639	0.020238	1.000000



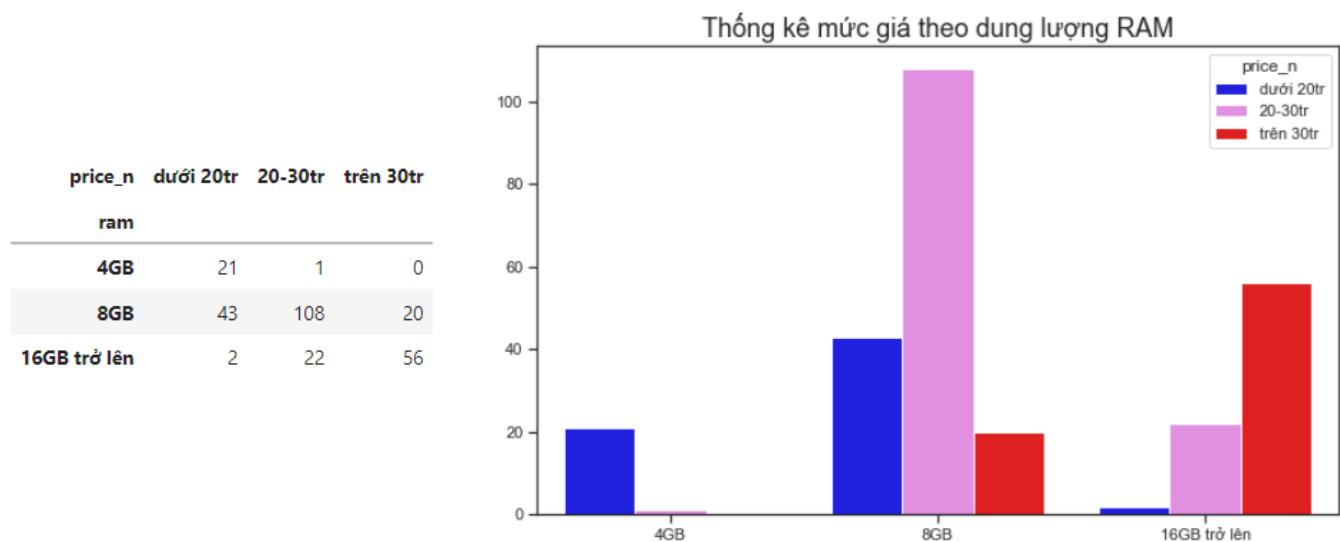
Ta có 2 giả thuyết

- **RAM** là yếu tố ảnh hưởng rất lớn đến **Giá sản phẩm (price)**
- Giữa **Kích thước màn hình (scrsizes)** và **Trọng lượng (weight)** có mối liên hệ với nhau

Vì vậy chúng ta cần sử dụng kiểm định Chisquare để kiểm định 2 giả thuyết trên

3.2.1 Sử dụng kiểm định Chi2 để nghiên cứu sự ảnh hưởng của RAM (ram) đến Giá sản phẩm (price)

Vì chỉ có 3 sản phẩm laptop có RAM 32GB nên chúng ta sẽ label chung với những laptop có ram 16GB thành **16GB trở lên**



Dùng kiểm định Chi2 để nghiên cứu sự ảnh hưởng của **RAM (ram)** đến **Giá sản phẩm (price)**.

Phát biểu giả thiết:

- H_0 : biến **RAM (ram)** biến **Giá sản phẩm (price)** là 2 biến độc lập.
- H_1 : biến **RAM (ram)** ảnh hưởng đến biến **Giá sản phẩm (price)**.

```
from scipy import stats as st # Kiểm định thống kê
score, p_value, dof, expected = st.chi2_contingency(ram)
p = {}
p['score'] = score
p['p_value'] = p_value
p['dof'] = dof
results(p)
```

score	p_value	dof	Kết luận
158.912558	2.501131e-33	4	Chấp nhận H_1 với mức ý nghĩa 0.05

Kết luận: Về mặt thống kê, **Dung lượng RAM (ram)** ảnh hưởng đến biến **Giá sản phẩm (price)** với mức ý nghĩa 5%.

Điều này chứng tỏ: Laptop càng nhiều dung lượng RAM sẽ có giá thành càng cao.

3.2.2 Dùng kiểm định Chi2 để nghiên cứu sự ảnh hưởng của Kích thước (scrsizes) đến Cân nặng (weight).

weight_n	light	normal	heavy
scrsizes_n			
<14inch	40	0	0
14inch	46	27	2
>14inch	6	54	98



Dùng kiểm định Chi2 để nghiên cứu sự ảnh hưởng của Kích thước (scrsizes) đến Cân nặng (weight).

Phát biểu giả thiết:

- H_0 : biến Kích thước màn hình (scrsizes) biến Cân nặng (weight) là 2 biến độc lập.
- H_1 : biến Kích thước màn hình (scrsizes) ảnh hưởng đến biến Cân nặng (weight) .

```
] : from scipy import stats as st # Kiểm định thống kê
score, p_value, dof, expected = st.chi2_contingency(kg_inch)
p = {}
p['score'] = score
p['p_value'] = p_value
p['dof'] = dof
results(p)
```

	score	p_value	dof	KetLuan
	191.14243	3.011594e-40	4	Chấp nhận H1 với mức ý nghĩa 0.05

Kết luận: Về mặt thống kê, biến Kích thước màn hình (scrsizes) ảnh hưởng đến biến Cân nặng (weight) . với mức ý nghĩa 5%.

Điều này chứng tỏ Kích thước màn hình càng lớn, thì laptop càng nặng

SourceCode Thống kê suy diễn: [clickhere](#)

KIỂM TRA ĐẠO VĂN



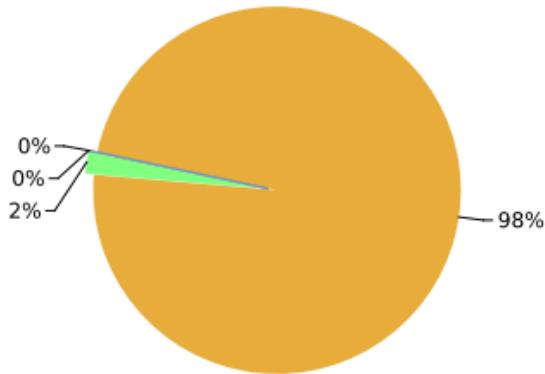
Hệ thống hỗ trợ nâng cao chất lượng tài liệu

KẾT QUẢ KIỂM TRA TRÙNG LẶP TÀI LIỆU

THÔNG TIN TÀI LIỆU

Tác giả	Sử Thành Công
Tên tài liệu	Phân tích dữ liệu về laptop bằng ngôn ngữ lập trình R
Thời gian kiểm tra	09-01-2022, 14:31:21
Thời gian tạo báo cáo	09-01-2022, 14:41:20

KẾT QUẢ KIỂM TRA TRÙNG LẶP



- Câu (đoạn) người dùng phản hồi
- Phần trăm câu (đoạn) hệ thống không kiểm tra
- Phần trăm câu (đoạn) không trùng lặp
- Phần trăm câu (đoạn) trùng lặp

[Chi tiết](#)

TÀI LIỆU THAM KHẢO

i [1] Thu thập dữ liệu bằng Selenium:

- [Hướng dẫn lấy dữ liệu web, Web Crawling với Selenium - Python - WebDrive \(trinhtuantai.com\)](#)
- [Selenium with Python — Selenium Python Bindings 2 documentation \(selenium-python.readthedocs.io\)](#)

ii [2] Xử lý dữ liệu bằng Python (pandas)

- [pandas - Python Data Analysis Library \(pydata.org\)](#)

iii [3] Thao tác với dữ liệu bằng ngôn ngữ R:

- [Manipulating data with R](#)
- [Phân tích số liệu và biểu đồ bằng R – Nguyễn Văn Tuấn](#)

iv [4] Trực quan hóa dữ liệu bằng thư viện Plotly:

- [Plotly R Graphing Library | R | Plotly](#)