BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo: Nhập môn lập trình Python cho phân tích

YẾU TỐ NÀO ẢNH HƯỞNG ĐẾN GIÁ LAPTOP HIỆN NAY

Giảng viên : Ths. Quách Đình Hoàng

Nhóm sinh viên thực hiện :

Văn Mai Thanh Nhật 20133076

Huỳnh Minh Phước 20133082

Nguyễn Trí Dũng 20133029

Lương Gia Huy 20133047

TP. Hồ Chí Minh, tháng 6 năm 2022

Mục lục

PHẦN 1 – TÓM TẮT

Trong những năm gần đây, ta có thể nhận thấy giá, cấu hình của một chiếc laptop đã có nhiều sự thay đổi đáng kể. Còn trong thị trường hiện nay, ta cũng có thể bắt gặp nhiều loại, mẫu mã laptop khác nhau dẫn đến giá cũng khác nhau.

Vậy yếu tố nào ảnh hưởng đã ảnh hưởng đến giá của một chiếc laptop? Cũng như làm sao để chọn được một chiếc laptop có hiệu năng trên giá thành tốt tuỳ theo nhu cầu sử dụng của mỗi cá nhân? Đó là mục đích của bài báo cáo này.

Ở đây, nhóm sử dụng 4 phương pháp phân tích, bao gồm:

- **Phân tích miêu tả (Descriptive Snalysis)** sắp xếp, thao tác và diễn giải dữ liệu thô thành những góc nhìn sâu sắc có giá trị cho bài báo cáo.

- **Phân tích khám phá (Exploratory Snalysis)** giúp nhóm tìm ra các kết nối và đưa ra các giả thuyết về mối quan hệ giữa những biến giải thích với nhau.

- **Phân tích chuẩn đoán (Diagnostic Analysis)** tìm hiểu sâu vào các bộ dữ liệu để tìm kiếm thông tin chi tiết có giá trị.

- **Phân tích dự đoán (Predictive Analysis)** dựa vào kết quả của phân tích mô tả, khám phá và chuẩn đoán ở trên, bên cạnh công cụ học máy (Machine learning) và Trí tuệ nhân tạo (AI) để phát hiện những xu hướng trong tương lai, cũng như các vấn đề tiềm ẩn cần khai thác trong dữ liệu.

Qua góc nhìn tổng quan thì ta có thể thấy hiện hữu khá nhiều yếu tố ảnh hưởng đến giá của một chiếc laptop, những yếu tố này bao gồm thương hiệu sản xuất và những chính sách khuyến mãi của hãng, hay là những tiện ích đi kèm. Ngoài ra, ở phần cứng, quan trọng chính là dung lượng bộ nhớ để có thể lưu trữ dữ liệu, tốc độ của vi xử lý cũng không kém phần ảnh hưởng. Và đóng góp một phần nhỏ có thể kể đến thiết kế bên ngoài và hệ điều hành mà nó sử dụng, cũng đã tác động đến giá của một chiếc laptop hiện nay.

PHẦN 2 – GIỚI THIỆU

Trong những tác vụ xử lí công việc hay học tập hằng ngày, sự tiện lợi và linh hoạt của một chiếc laptop đã giúp chúng dần dần thay thế những chiếc máy bàn cồng kềnh và nặng nề, không thể di chuyển. Mọi thứ của một chiếc PC giờ đây đã gói gọn chỉ bằng một cuốn sách và ta có thể đem nó đến mọi nơi từ nhà riêng cho đến văn phòng hay thậm chí là quán cà phê để có thể làm việc, do đó trong thị trường thiết bị công nghệ hiện nay, nó chiếm thị phần khá cao, chỉ xếp sau điện thoại thông minh và cũng đang trở thành xu hướng hiện đại thay cho những chiếc máy bàn vì lợi ích của nó khó có thể bỏ qua.

Ngay trong việc chọn mua 1 chiếc lap cũng đã dễ dàng hơn so với phải mua từng linh kiện để lắp ráp thành máy bàn. Ngoài ra, laptop còn được tích hợp nhiều công nghệ mới, hay sự đồng bộ trong phần cứng để tối ưu hoá hiệu năng do chính nhà sản xuất cung cấp tạo ra những lợi ích vượt trội mà máy bàn không thể có được.

Những lí do này cùng với đề cập ở phần trên, có hàng tá những yếu tố có thể ảnh hưởng đến giá của một chiếc laptop, dẫn đến cho ta những câu hỏi: vậy thì nó ảnh hưởng như thế nào? Tác động của nó là nhỏ hay lớn? Những mối liên hệ này có tuân theo quy tắc nào hay không? Và những mối quan hệ này sẽ giúp ích được gì trong dự đoán tương lai? Đó là những câu hỏi mà ta sẽ đi giải thích trong bài báo cáo này. Đồng thời cũng cung cấp cho người đọc một góc nhìn sâu sắc hơn về thị trường laptop hiện nay để có thể giúp cho bản thân chọn được một chiếc laptop phù hợp với nhu cầu trong thời đại công nghệ 4.0 đang dần chiếm lĩnh xã hội.

Trong báo cáo này, nhóm sử dụng Input của bài toán là tập các :

- Thương hiệu
- Tên vi xử lí
- Thương hiệu vi xử lí
- Thế hệ vi xử lí
- Dung lượng ram
- Loại ram
- Dung lượng ổ cứng ssd
- Dung lượng ổ cứng hdd
- Hệ điều hành
- Loại hệ điều hành
- Dung lượng card đồ hoạ
- Cân nặng
- Kích thước màn hình
- Số năm bảo hành của hãng
- Cảm ứng màn hình
- Ứng dụng MSOffice

Nhóm sử dụng những thuật toán:

- ANOVA F test: Nhiều nhóm độc lập
- T-test : 2 nhóm độc lập
- Multicollinearity(đa cộng tuyến): dùng VIF
- Backward elimination: Loại bỏ các biến có p-values > 0,05

Để:

- Tìm ra mối quan hệ của từng biến thông số kĩ thuật với giá (USD) của laptop
- Dự đoán giá của một chiếc laptop dựa trên những thuộc tính được người dùng lựa chọn

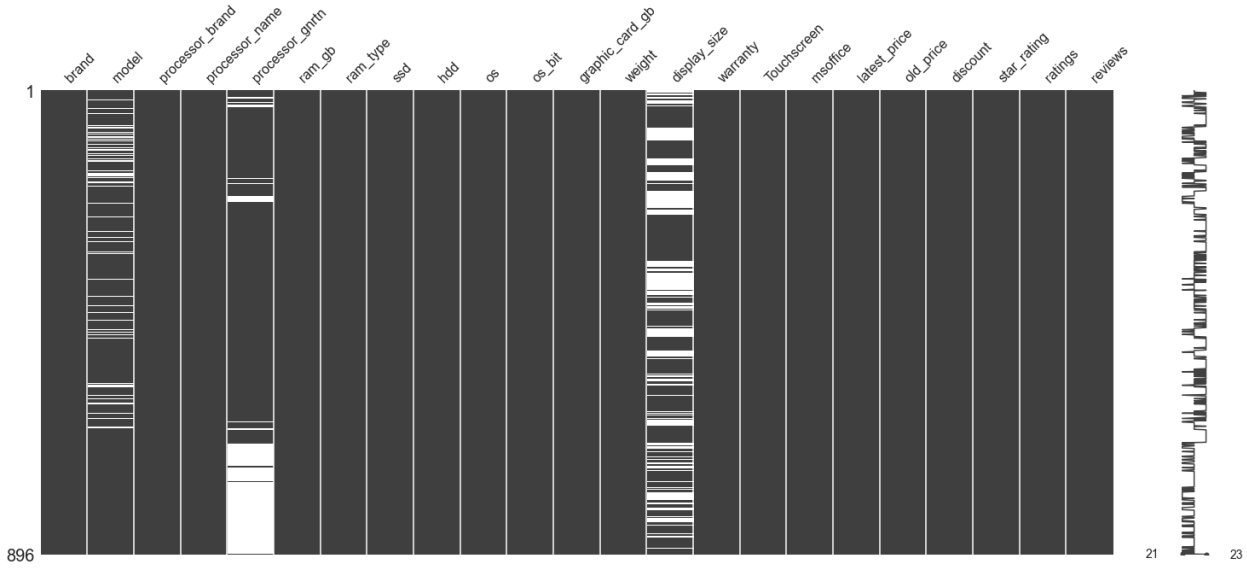
PHẦN 3 – DỮ LIỆU

Trong đề tài này, nhóm xin dùng bộ dữ liệu về *thông số kỹ thuật và giá thành của laptop hiện nay* để trực quan hoá mối liên hệ và tác động kể trên để phần nào hiểu rõ hơn về việc việc phân tích dữ liệu.

Theo như tác giả của bộ dữ liệu, nguồn dữ liệu ở đây được trích xuất từ Flipkart.com. Flipkart Pvt Ltd. là một công ty thương mại điện tử ở Ấn Độ có trụ sở tại Bengaluru, chuyên cung cấp danh sách các sản phẩm như phụ kiện điện tử, đồ gia dụng, quần áo, thiết bị tập thể dục và nhiều danh sách lựa chọn khác. Công ty được Sachin Bansal và Binny Bansal (không có họ hàng, họ từng làm việc cho Amazon) thành lập vào tháng 10 năm 2007. Sau quãng thời gian dài phát triển, hiện nay Flipkart shopping hiện là công ty Thương mại điện tử lớn nhất và mạnh nhất ở Ấn Độ, trị giá hơn 30 tỷ USD.

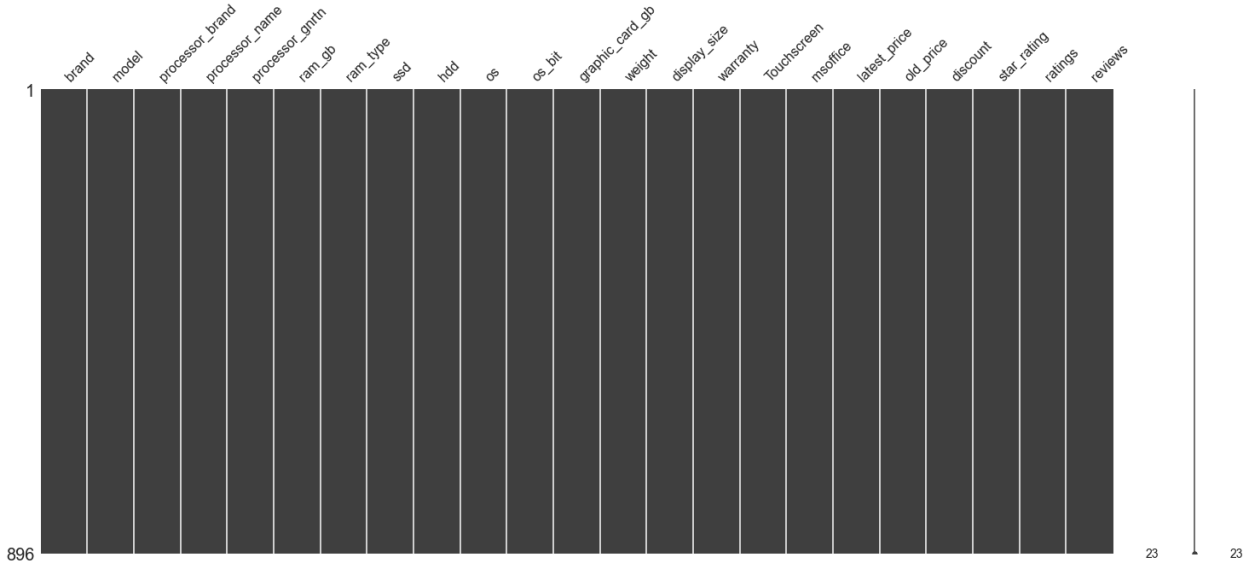
Phương pháp thu thập dữ liệu đến từ một tiện ích mã nguồn mở tự động trên Chrome có tên là Instant Data Scrapper, được phát triển bởi webrobots.io, là một công cụ trích xuất dữ liệu tự động từ bất kì trang web nào bằng cách sử dụng AI để dự đoán dữ liệu nào là liên quan, thích hợp nhất từ trang HTML và cho phép lưu dữ liệu đó thành file xlsx hay csv.

3.1 Tiền xử lý



Ta có thể thấy ở dữ liệu có khá nhiều dữ liệu bị thiếu (missing value) ở những cột model (dòng máy), processor_gnrtn (thế hệ vi xử lý), display_size (kích thước màn hình).

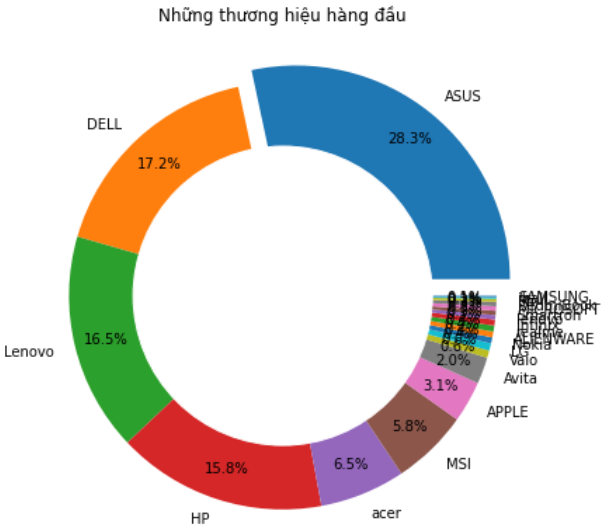
Nhóm thực hiện tiền xử lí bằng cách điền vào những giá trị bị thiếu ở thuộc tính Model bằng giá trị "Unknow". Còn ở thuộc tính processor_gnrtn và display_size thì dùng giá trị phổ biến nhất, xuất hiện nhiều nhất trong cột để thay thế cho những giá trị bị thiếu.



Có thể thấy dữ liệu sau khi được xử lí đã không còn dữ liệu trống và trở thành một bộ dữ liệu hoàn chỉnh.

PHẦN 4 – TRỰC QUAN HOÁ DỮ LIỆU

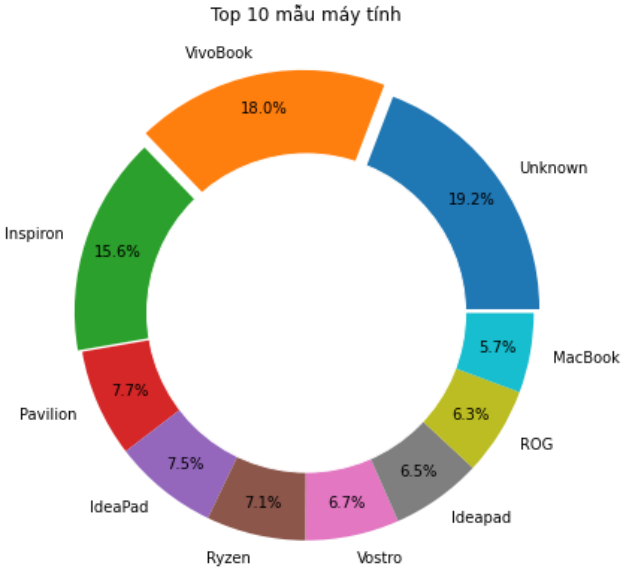
4.1 Thống kê Thương hiệu:



Kết luận:

- Asus đứng đầu với 28.3%
- Dell, Lenovo, HP có thị phần tương đương nhau xấp xỉ 16%
- Apple với hệ điều hành OS riêng nhưng đừng ở vị trí thứ 7 với 3.1%
- Những thương hiệu cần tăng gia sản xuất (sản phẩm bán ra ít hơn 5 trong dữ liệu):
  - 'LG','Vaio', 'realme', 'Nokia', 'Infinix', 'ALIENWARE', 'Smartron', 'lenovo', 'MICROSOFT', 'RedmiBook', 'Mi', 'SAMSUNG', 'iball'

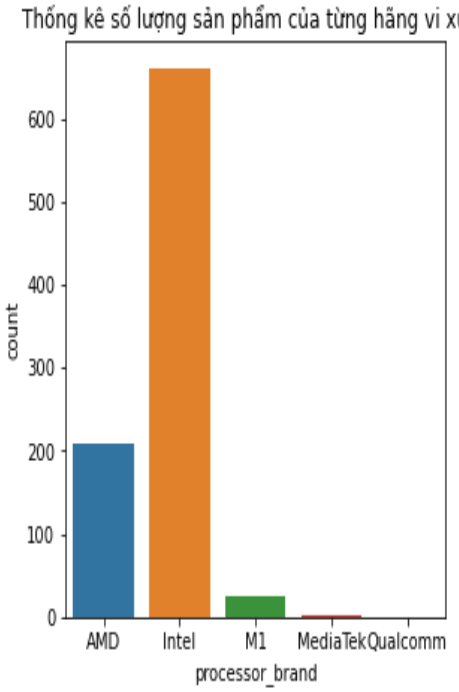
4.2 Thống kê dòng máy:

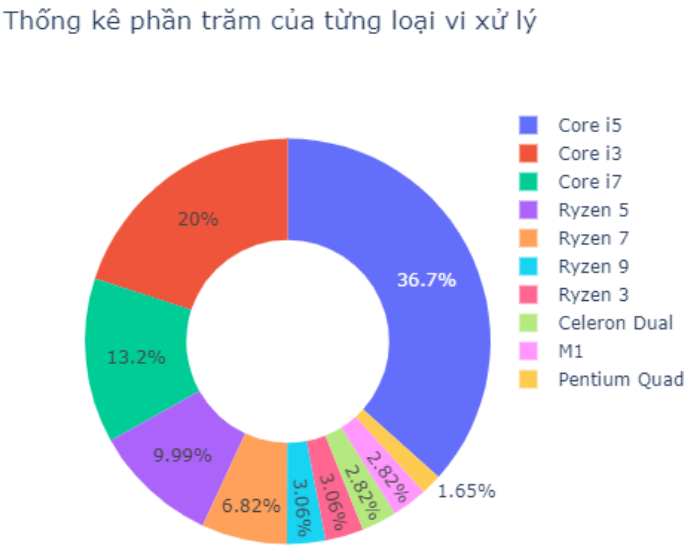


Kết luận:

- Vivobook là dòng laptop được ưa chuộng nhiều nhất
- Trong khi đó ta có: Những dòng Inspiron, Pivillion của DELL, IdeaPad của Lenovo, Ryzen có thể thuộc HP hoặc ASUS thì nằm trong top 5
- Gần 19.3% là những giá trị bị thiếu, với số lượng lớn như vậy thì có thể là gia tăng 1 hay vài giá trị của những dòng máy tính khác

4.3 Thống kê vi xử lý:

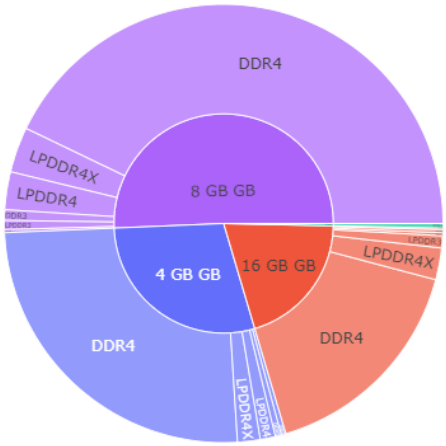




Kết luận

- Intel là hãng có sản lượng cao nhất, cứ 4 laptop được bán ra thì sẽ có 3 cái sử dụng vi xử lý của Intel
- Ta có thể nhận thấy top 3 đều thuộc về Intel đó là i5, i3, i7
- Core i5 được ưa chuộng nhiều nhất có thể là do sự linh hoạt trong loại công việc của nó có thể đảm nhiệm, không quá yếu để xử lý những tác vụ văn phòng như i3, và không quá dư thừa để xử lý những tác vụ nặng về đồ hoạ như i7

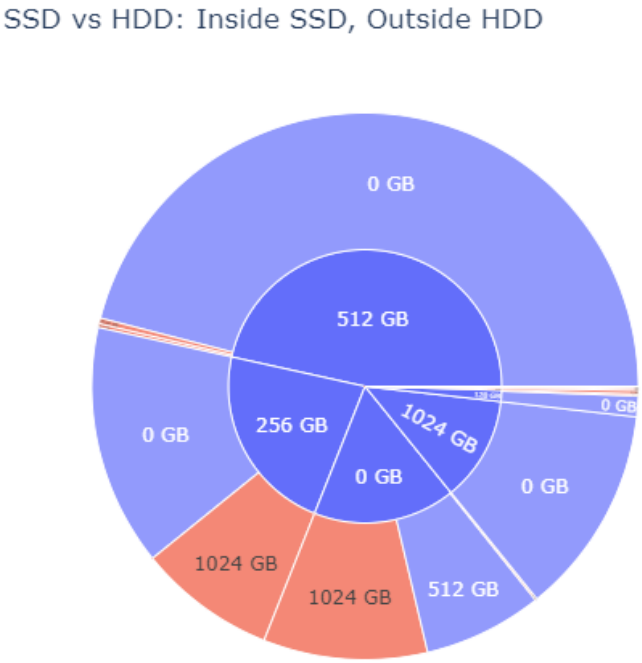
4.4 Thống kê về RAM:



Kết luận:

- 50% laptop hiện giờ đang được trang bị 8GB ram
- Trong đó có 85% đang sử dụng loại DDR4, chiếm phần trăm cao nhất trong bảng xếp hạng loại ram

4.5 Thống kê về ổ đĩa:



Kết luận:

- Gần một nửa số máy trên thị trường sở hữu bộ nhớ SSD 512GB
- Với các máy có bộ nhớ SSD lớn sẽ có xu hướng không sử dụng SSD, và với các SSD mức thấp hơn có xu hướng sử dụng bộ nhớ HDD 1024GB

PHẦN 5 - MÔ HÌNH HÓA DỮ LIỆU

5.1 Kiểm định giả thuyết

Kiểm định chứng minh các yếu tố phần cứng có ảnh hưởng tới giá thành laptop.

Thuật toán sử dụng:

- ANOVA F-test đối với các biến có nhiều nhóm độc lập

ANOVA F-test statistic:

$$F = \frac{\sum_{i=1}^{k} n_i \frac{(\overline{x}_i - \overline{x})^2}{k-1}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(x_{ij} - \overline{x}_i)^2}{n-k}} \sim F(k-1, n-k)$$

- t-test đối với các biến có hai nhóm độc lập

▶ t-test statistic:

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \sim t(n-2)$$

$$SE(\hat{\beta}) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- Công thức tính khoảng tin cậy:

▶ Khoảng tin cậy $(1-\alpha)$ ($(1-\alpha)$ CI) cho β:

$$\hat{\beta} \pm t_{1-\alpha/2} \times SE(\hat{\beta})$$

Phương pháp thực hiện:

- Đặt giả thuyết liệu có sự khác nhau về giá giữa các nhóm độc lập
  - H0: Giá bằng nhau giữa các nhóm độc lập
  - H1: Giá khác nhau giữa các nhóm độc lập
- Thực hiện kiểm định giả thuyết với độ tin cậy 95%.
  - Với các biến có giá trị p-value < mức ý nghĩa (0.05), phủ nhận H0 và kết luận biến có ảnh hưởng tới giá cuối cùng.
  - Với các biến có giá trị p-value > mức ý nghĩa (0.05), chấp nhận H0 và kết luận biến không ảnh hưởng tới mức giá cuối cùng.

5.2 Mô hình dự đoán giá laptop

Xây dựng mô hình hồi quy tuyến tính. Chọn 80% dữ liệu để huấn luyện cho mô hình và 20% test mô hình.

Phương pháp sử dụng:

Backward elimination: Ta loại bỏ những biến có chỉ số p-value (P>|t|) cao hơn 0,05. Rồi dựng lại mô hình cho đến khi đạt được mô hình tốt nhất gồm:

- R-square, F-statistic cao
- AIC, BIC thấp

VIF (Variance inflation factor): Kiểm tra coi có Đa cộng tuyến (Multicollinearity) không bằng cách check VIF. Nếu VIF của biến nào lớn hơn 10 thì loại ra khỏi mô hình.

Mô hình hồi qui tuyến tính đa biến có dạng:

$$y= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots\ldots + \ldots\ldots + \beta_k x_{ki} + \varepsilon_i$$

PHẦN 6 – THỰC NGHIỆM, KẾT QUẢ, THẢO LUẬN

6.1 Kiểm định giả thuyết:

Tất cả những biến kiểm định dưới đều là biến phân loại bao gồm: brand, model, processor_brand, processor_name, processor_gnrtn, ram_gb, ram_type, ssd, hdd, os, os_bit, graphic_card_gb, weight, display_size, warranty, Touchscreen, msoffice.

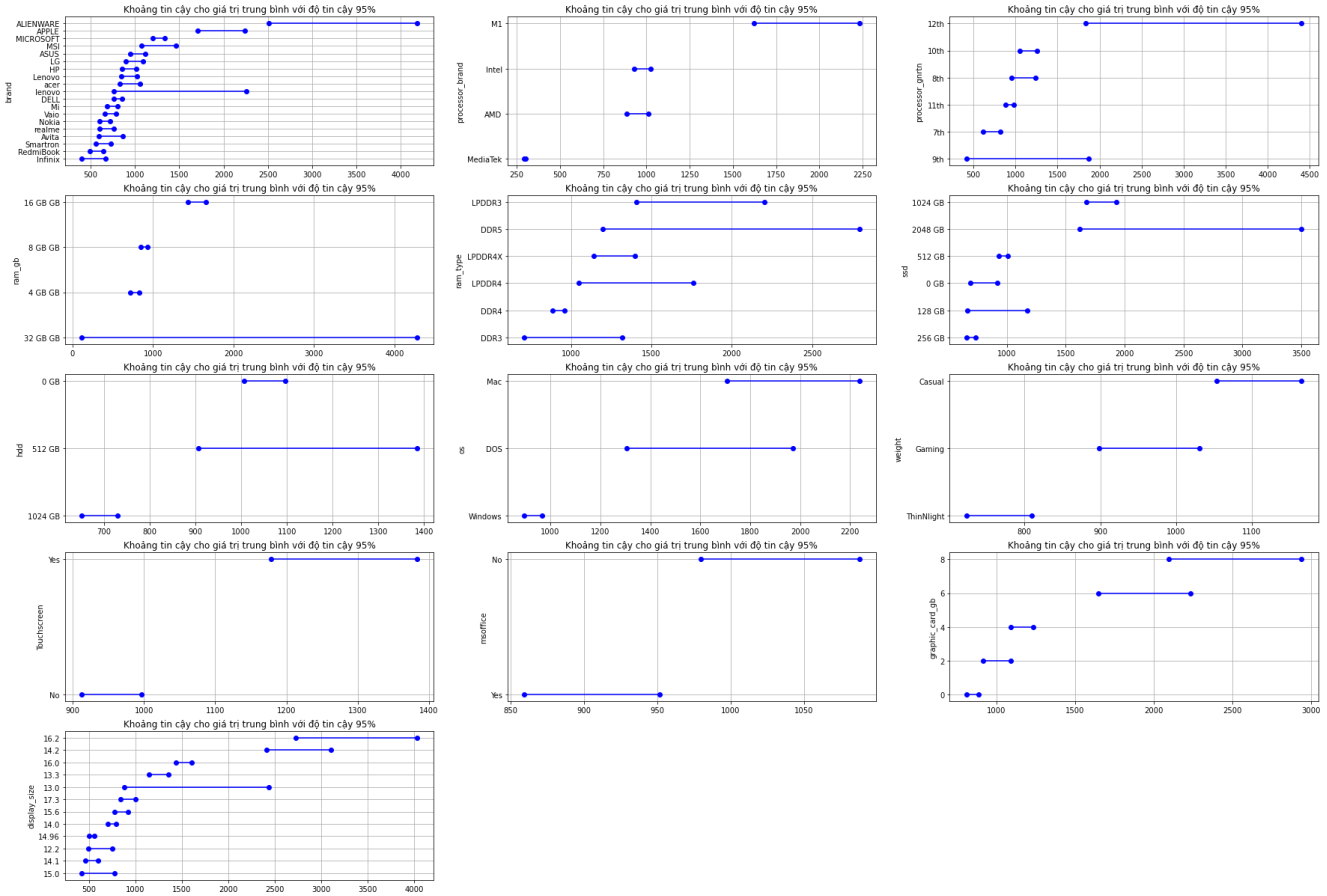
Trong phần này, ta sử dụng t-test cho những biến có hai nhóm độc lập gồm: Touchscreen, msoffice. Còn lại là những biến có nhiều nhóm độc lập nên ta sử dụng ANOVA F test. Mục đích để kiểm tra p-value < 0,05.

Sau khi thực hiện kiểm định ta loại được các biến os_bit(số bit hệ điều hành) và warranty(sô năm bảo hành) do p-value > 0,05 nên có thể kết luận chúng không ảnh hưởng tới giá thành sản phẩm.

Cuối cùng, ta tính và vẽ biểu đồ khoảng tin cậy cho:

- Những biến không được liệt kê thành nhóm
- Những biến được liệt kê thành nhóm: Model được liệt kê theo từng brand, processor_name được liệt kê theo từng processor_brand.

6.1.1 Khoảng tin cậy của giá cho từng biến không được liệt kê thành nhóm:



Kết luận:

- Thương hiệu:
  - Một số những hãng máy tính chỉ có 1 máy trong dữ liệu sẽ không xuất hiện trong biểu đồ
  - Đa số các hãng đều có khoảng tin cậy nhỏ, riêng Lenovo và ALIENWARE thì có khoảng tin cậy lớn hơn. Do Alienware có kích thướng mẫu nhỏ, còn Lenovo có nhiều phân khúc máy khác nhau dẫn đến sự chênh lệch giá khá nhiều.
  - Riêng Alienware có Giá cao hơn hầu như tất cả các máy còn lại.
  - Đa số sẽ có giá từ 50000 - 100000 (15tr - 30tr), những giá trị cao hơn sẽ là những phân khúc cao cấp

- Thương hiệu vi xử lý:
  - M1 luôn nằm ở phân khúc khác so với các đối thủ còn lại, Intel và AMD có giá khá tương đượng mặc dù Intel được ưa chuộng nhiều hơn

- Thế hệ vi xử lý:
  - Giá của gen 10th vs 11th khá ổn định, Giá của gen 12h bị lớn là do vừa ra mắt

- Dung lượng RAM:
  - 4gb vs 8gb đc ưa chuộng nên giá khá ổn đinh từ > 50000, 32gb lớn là do có kích thước mẫu nhỏ
  - Với sự chênh lệch giá của 4gb vs 8gb là không quá khác biệt, 16gb cũng không thực sự cao so với 8gb

- Loại RAM:
  - DDR4 được ưa chuộng nhất

- Dung lượng SSD:

  - Đa số laptop được trang bị ssd 256Gb và 512Gb
  - Có 1 số lượng nhiều máy không sử dụng SSD
  - 1024Gb khá cao so với 512Gb

- Dung lượng HDD:
  - Hiện nay laptop đã được trang bị SSD sử dụng công nghệ mới nên không còn sử dụng HDD
  - Nếu có thì sử dụng 1TB để lưu trữ

- Hệ điều hành:
  - Máy có sử dụng hệ điều hành MACOS có giá cao hơn cả, tiếp đó là DOS và cuối cùng là Window với vị trí thấp nhất

- Máy sử dụng hdh Windows có khoảng tin cậy nhỏ do có số lượng mẫu lớn, từ đó có thể khẳng định có độ tin cậy cao nhất
- MAC và DOS có khoảng tin cậy dài do số lượng mẫu nhỏ, độ tin cậy thấp hơn so với Windows

- Cân nặng:
  - Đa số các hãng sản xuất laptop chủ yếu sản xuất những máy casual hướng đến đối tượng người dùng phổ thông
  - Casual cao hơn Gaming vì lí do trên nên mức giá của casual
  - Laptop thông thường lại có giá cao hơn các loại máy tính khác, trong khi đó laptop mỏng và nhẹ có giá trị thấp nhất do cắt giảm phần cứng

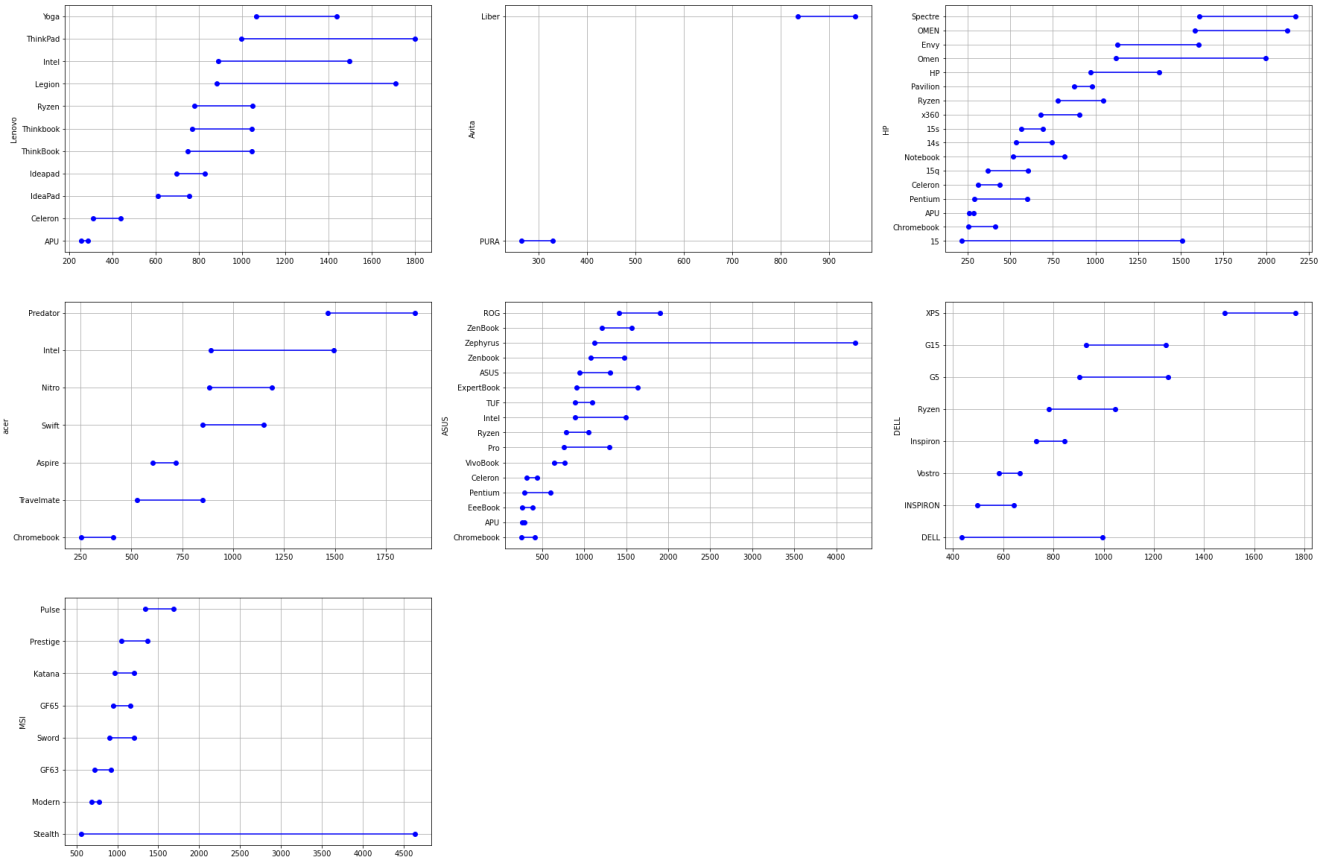
- Màn hình cảm ứng:
  - Hầu như máy tính hiện nay đều không xài đến màn hình cảm ứng, có lẽ vì sự hiện diện của table nên tính năng cảm ứng trên màn hình laptop không được ưa chuộng
  - Khi máy tính có màn hình cảm ứng đồng nghĩa với giá sẽ tăng theo, giá sửa chữa khi hư hỏng màn hình cũng không hề rẻ

- MSOffice:
  - Ta có thể thấy 1 điều khá thú vị, khi giá laptop càng tăng thì nó không được hỗ trợ MSOffice
  - Có lẽ là vì giá đã không hề rẻ nên nhà sản xuất muốn loại bỏ 1 số khuyến mãi đi kèm để giá bán ra tốt nhất có thể
  - Những máy có tích hợp sẵn msoffice có giá rẻ hơn vì đó là tích hợp của hãng, còn những máy còn lại phải mua kèm theo nhà phân phối nên giá sẽ cao hơn, sự chênh lệch khoảng 10000 rupee ~ 2.900.000vnd = 1 năm sử dụng msoffice

- Dung lượng card đồ hoạ:
  - Đa số các máy đều không có card đồ hoạ
  - Máy có dung lượng card đồ hoạ càng lớn thì giá càng cao

- Kích thước màn hình:
  - Dựa vào bảng và biểu đồ, ta có thể nhận ra giá trung bình của 1 chiếc laptop có kích thước màn hình 15,6 (phổ biến nhất) inch nằm khoảng > 50000
  - Với những loại có số lượng bản ghi thấp như thì 13.0 và 16.2 sẽ có khoảng tin cậy rộng, dẫn đến độ tin tưởng khá thấp
  - Nhìn chung, ta có thể thấy với mọi kích thước màn hình thì giá sẽ vẫn nằm ở khoảng từ 50000 đến 125000

6.1.2 Khoảng tin cậy cho từng nhóm mẫu máy trong từng thương hiệu



Kết luận:

Lenovo:

- Có thể thấy rằng APU là dòng máy phổ thông nhất vì mức độ phổ biến cũng như giá thành rẻ.
- Trong khi đó Yoga và ThinkPad lại nhắm vào phân khúc giá cao hơn với số lượng máy không nhiều.

Avita:

- Có thể thấy rõ hai phân khúc mà hãng nhắm tới là giá rẻ và tầm trung.
- Số lượng máy sản xuất tương đối đồng đều cũng phản ánh nhu cầu của thị trường.

HP:

- Sở hữu số lượng máy đa dạng theo từng phân khúc.

Acer:

- Chromebook sở hữu phân khúc giá thấp nhất.
- Phổ biến nhất là Aspire.
- Cao cấp nhất là dòng Predator.

ASUS:

- Sở hữu nhiều mẫu máy với nhiều phân khúc thị trường.
- Zephyrus sở hữu ít máy và có mức giá cao hơn nhiều so với đa số các mẫu khác.

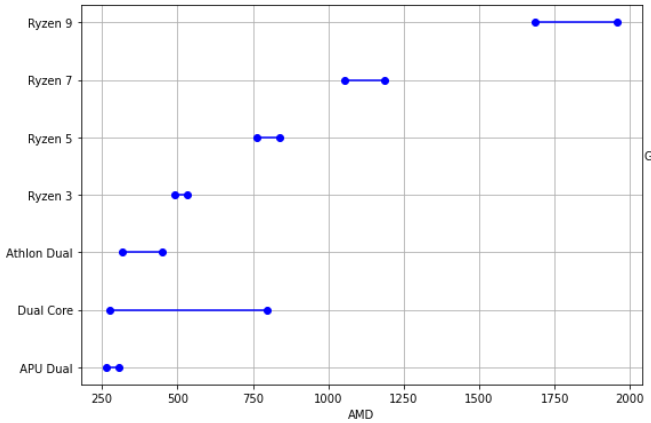
DELL:

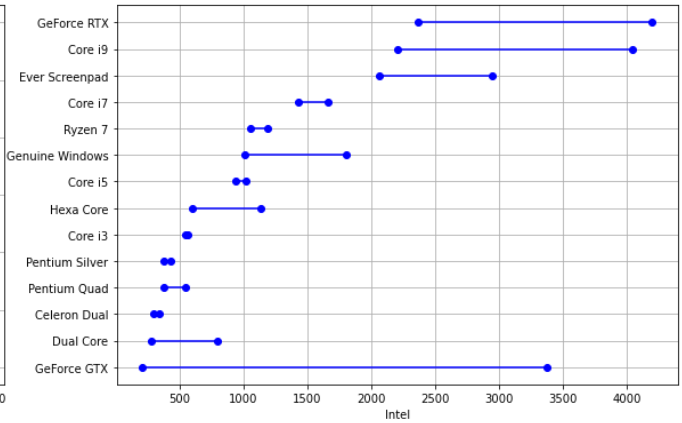
- Có thể thấy rõ rằng dòng Inspiron là dòng máy phổ biến nhất.
- XPS sở hữu phân khúc giá cao hơn hẳn so với các dòng máy khác.

MSI:

- Đa số các máy nằm ở phân khúc tầm trung.
- Có số ít máy ở phân khúc cao cấp nằm trong dòng Stealth.

6.1.3 Khoảng tin cậy cho từng nhóm mẫu vi xử lý trong từng hãng vi xử lý





Kết luận:

AMD:

- Đa số các máy sử dụng AMD sẽ dùng APU Dual hoặc Ryzen 3.
- Số ít máy sử dụng Dual Core.
- APU Dual dành cho phân khúc giá rẻ khá được ưa chuộng.
- Ryzen 9 dành cho phân khúc cao cấp.

Intel:

- Sở hữu nhiều dòng chip khác nhau ở nhiều phân khúc giá.
- Phổ biến nhất là Core i5 và Pentium Silver nằm ở phân khúc tầm trung.
- GeForce GTX và GeForce RTX không quá phổ biến trên thị trường.

6.2 Mô hình dự đoán giá laptop:

6.2.1 Định nghĩa:

Mô hình hồi quy tuyến tính đa biến là phương trình mô tả quan hệ giữa biến phụ thuộc Y với các biến độc lập X_1, X_2, ..., X_n và sai số ngẫu nhiên ε.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

(với $\beta_1, \beta_2, \ldots, \beta_n$ là các tham số)

- Hệ số xác định r^2: dùng để đo mức độ ảnh hưởng của yếu tố được xem xét trong mô hình đối với sự biến động của các giá trị của các biến ngẫu nhiên quanh giá trị trung bình của nó, r^2 càng lớn thì mô hình càng ý nghĩa.

$$r^2 = 1 - \frac{SSE}{SST}$$

- Tiêu chí thông tinh Akaike (AIC)
- Tiêu chí thông tin Bayesian (BIC)

$$AIC = -2(loglikelihood) + 2k$$

$$BIC = -2(loglikelihood) + klogn$$

(với k là tham số của mô hình)

Đối với hồi quy tuyến tính, với giả định Guass thì $-2(loglikelihood)$ tỉ lệ với $nlog(\frac{SSE}{n})$

Vì tổng bình phương sai số được giải thích bởi mô hình (SSE) càng nhỏ thì AIC và BIC càng lớn. Một mô hình đơn giản và đầy đủ phải là mô hình có trị số AIC hoặc BIC càng thấp càng tốt và các biến độc lập phải có ý nghĩa thống kê.

6.2.2 Xây dựng mô hình hồi quy tuyến tính để đánh giá các nhân tố có thể ảnh hưởng đến giá thành của laptop.

OLS Regression Results

Dep. Variable:	latest_price	**R-squared (uncentered):**	0.872
Model:	OLS	**Adj. R-squared (uncentered):**	0.870
Method:	Least Squares	**F-statistic:**	436.3
Date:	Wed, 15 Jun 2022	**Prob (F-statistic):**	7.46e-306
Time:	17:04:09	**Log-Likelihood:**	-5336.3
No. Observations:	716	**AIC:**	1.069e+04
Df Residuals:	705	**BIC:**	1.074e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ram_gb	34.0602	4.012	8.491	0.000	26.184	41.936
ssd	0.7758	0.071	10.861	0.000	0.636	0.916
hdd	0.1278	0.050	2.572	0.010	0.030	0.225
processor_gnrtn	-93.0450	17.212	-5.406	0.000	-126.838	-59.252
graphic_card_gb	90.7803	8.448	10.746	0.000	74.194	107.367
display_size	82.1740	11.864	6.926	0.000	58.881	105.467
warranty	-37.6434	28.124	-1.338	0.181	-92.860	17.573
Touchscreen	301.3702	51.040	5.905	0.000	201.162	401.578
star_rating	-19.2139	8.339	-2.304	0.022	-35.587	-2.841
ratings	0.0937	0.073	1.285	0.199	-0.049	0.237
reviews	-0.8240	0.602	-1.368	0.172	-2.007	0.359

Omnibus:	435.033	**Durbin-Watson:**	2.078
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4558.745
Skew:	2.593	**Prob(JB):**	0.00
Kurtosis:	14.222	**Cond. No.**	4.08e+03

Mô hình hồi quy tuyến tính bao gồm biến Target là một biến dự đoán và các biến còn lại đều là biến độc lập.

Biến Target là một biến dự đoán dựa vào các biến độc lập "ram_gb", "ssd", "hdd", "processor_gnrtn", "graphic_card_gb", "display_size", "warranty", "Touchscreen", "star_rating", "ratings" và "reviews". Dựa trên thông tin của mô hình model_new_DF_all:

− Ta có thể xem được biết được các tiêu chí thông tin AIC và BIC, chỉ số F-statistic, chỉ số R-squared giải thích mức độ biến động giá laptop.

– Giá trị P> |t| cung cấp thông tin mức độ của những yếu tố dự báo(biến độc lập) có ảnh hưởng đến Target, nếu giá trị dưới 0,05 (alpha) thì biến độc lập đó có ảnh hưởng lớn đến mô hình hồi quy được xây dựng. Giá trị P>|t| càng nhỏ thì càng có ý nghĩa.

Bằng phương pháp Backward elimination, ta loại bỏ những biến có chỉ số p-value(P>|t|) cao hơn 0.05. Rồi dựng lại mô hình cho đến khi đạt được mô hình tốt nhất (R-squared cao, F-statistic cao, AIC thấp, BIC thấp.)

Mô hình 2: Sau khi đã loại bỏ biến ratings

OLS Regression Results

Dep. Variable:	latest_price	**R-squared (uncentered):**	0.872
Model:	OLS	**Adj. R-squared (uncentered):**	0.870
Method:	Least Squares	**F-statistic:**	479.4
Date:	Wed, 15 Jun 2022	**Prob (F-statistic):**	7.56e-307
Time:	17:10:29	**Log-Likelihood:**	-5337.1
No. Observations:	716	**AIC:**	1.069e+04
Df Residuals:	706	**BIC:**	1.074e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ram_gb	34.1603	4.013	8.513	0.000	26.282	42.038
ssd	0.7772	0.071	10.876	0.000	0.637	0.917
hdd	0.1290	0.050	2.596	0.010	0.031	0.227
processor_gnrtn	-92.6252	17.217	-5.380	0.000	-126.428	-58.822
graphic_card_gb	90.3991	8.447	10.702	0.000	73.815	106.983
display_size	81.6796	11.863	6.885	0.000	58.388	104.971
warranty	-34.9024	28.056	-1.244	0.214	-89.985	20.181
Touchscreen	300.1217	51.054	5.879	0.000	199.886	400.357
star_rating	-19.3328	8.343	-2.317	0.021	-35.712	-2.953
reviews	-0.0647	0.117	-0.553	0.581	-0.294	0.165

Omnibus:	434.290	**Durbin-Watson:**	2.081
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4531.312
Skew:	2.588	**Prob(JB):**	0.00
Kurtosis:	14.184	**Cond. No.**	1.80e+03

Mô hình 3: Sau khi đã loại bỏ biến review

OLS Regression Results

Dep. Variable:	latest_price	**R-squared (uncentered):**	0.872
Model:	OLS	**Adj. R-squared (uncentered):**	0.870
Method:	Least Squares	**F-statistic:**	533.1
Date:	Wed, 15 Jun 2022	**Prob (F-statistic):**	3.71e-308
Time:	17:10:29	**Log-Likelihood:**	-5337.3
No. Observations:	716	**AIC:**	1.069e+04
Df Residuals:	707	**BIC:**	1.073e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ram_gb	34.3129	4.001	8.576	0.000	26.457	42.168
ssd	0.7818	0.071	11.021	0.000	0.643	0.921
hdd	0.1286	0.050	2.590	0.010	0.031	0.226
processor_gnrtn	-92.9467	17.199	-5.404	0.000	-126.714	-59.180
graphic_card_gb	90.2542	8.439	10.695	0.000	73.687	106.822
display_size	81.7148	11.857	6.892	0.000	58.435	104.994
warranty	-36.1941	27.945	-1.295	0.196	-91.059	18.670
Touchscreen	302.5016	50.847	5.949	0.000	202.672	402.331
star_rating	-20.2502	8.172	-2.478	0.013	-36.294	-4.206

Omnibus:	435.117	**Durbin-Watson:**	2.081
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4573.414
Skew:	2.592	**Prob(JB):**	0.00
Kurtosis:	14.244	**Cond. No.**	1.79e+03

OLS Regression Results

Dep. Variable:	latest_price	**R-squared (uncentered):**	0.871
Model:	OLS	**Adj. R-squared (uncentered):**	0.870
Method:	Least Squares	**F-statistic:**	599.0
Date:	Wed, 15 Jun 2022	**Prob (F-statistic):**	3.39e-309
Time:	17:10:29	**Log-Likelihood:**	-5338.1
No. Observations:	716	**AIC:**	1.069e+04
Df Residuals:	708	**BIC:**	1.073e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ram_gb	34.0976	4.000	8.525	0.000	26.245	41.950
ssd	0.7666	0.070	10.952	0.000	0.629	0.904
hdd	0.1281	0.050	2.578	0.010	0.031	0.226
processor_gnrtn	-94.2342	17.178	-5.486	0.000	-127.961	-60.508
graphic_card_gb	91.3830	8.397	10.882	0.000	74.896	107.870
display_size	81.8189	11.863	6.897	0.000	58.529	105.109
Touchscreen	294.1301	50.459	5.829	0.000	195.064	393.197
star_rating	-21.7716	8.091	-2.691	0.007	-37.657	-5.887

Omnibus:	438.493	**Durbin-Watson:**	2.074
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4700.689
Skew:	2.611	**Prob(JB):**	0.00
Kurtosis:	14.415	**Cond. No.**	1.76e+03

Mô hình 4: Sau khi đã loại bỏ biến warranty

Nhận xét:

- Ta dễ dàng thấy mô hình 4 là mô hình tốt nhất, thích hợp nhất để làm mô hình hồi quy tuyến tính. Từ đó, ta thấy được sự tác động của các biến lên biến dự đoán.
- Hệ số hồi quy của một biến dự báo khác 0 thì có ý nghĩa thống kê. Vì vậy, nhóm có thể biết được ảnh hưởng đến biến dự đoán.
- Xét đường hồi quy tuyến tính mẫu:

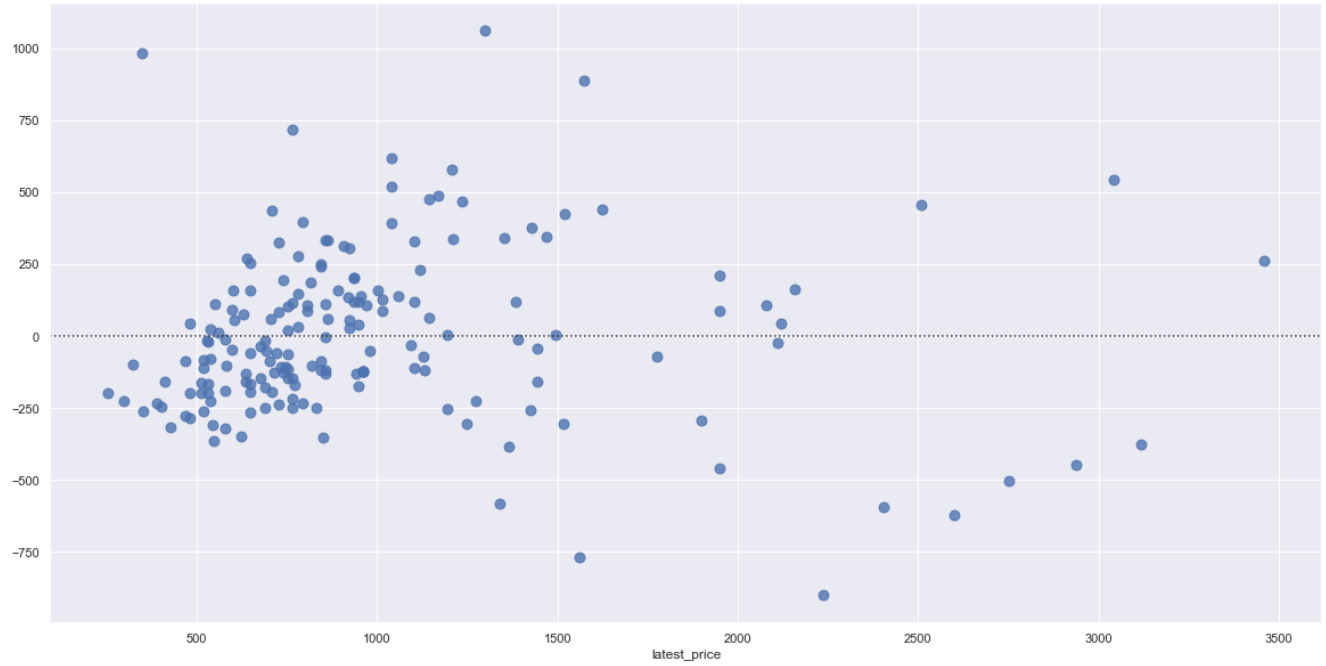
$$y = 2622.8929 * ramgb + 58.9719 * ssd + 9.8531 * hdd - 7248.7835 * processorgnrtn + 7029.4591 * graphiccardgb + 6294.7592 * displaysize + (2.263e + 04) * touchscreen - 1674.7416 * starrating$$

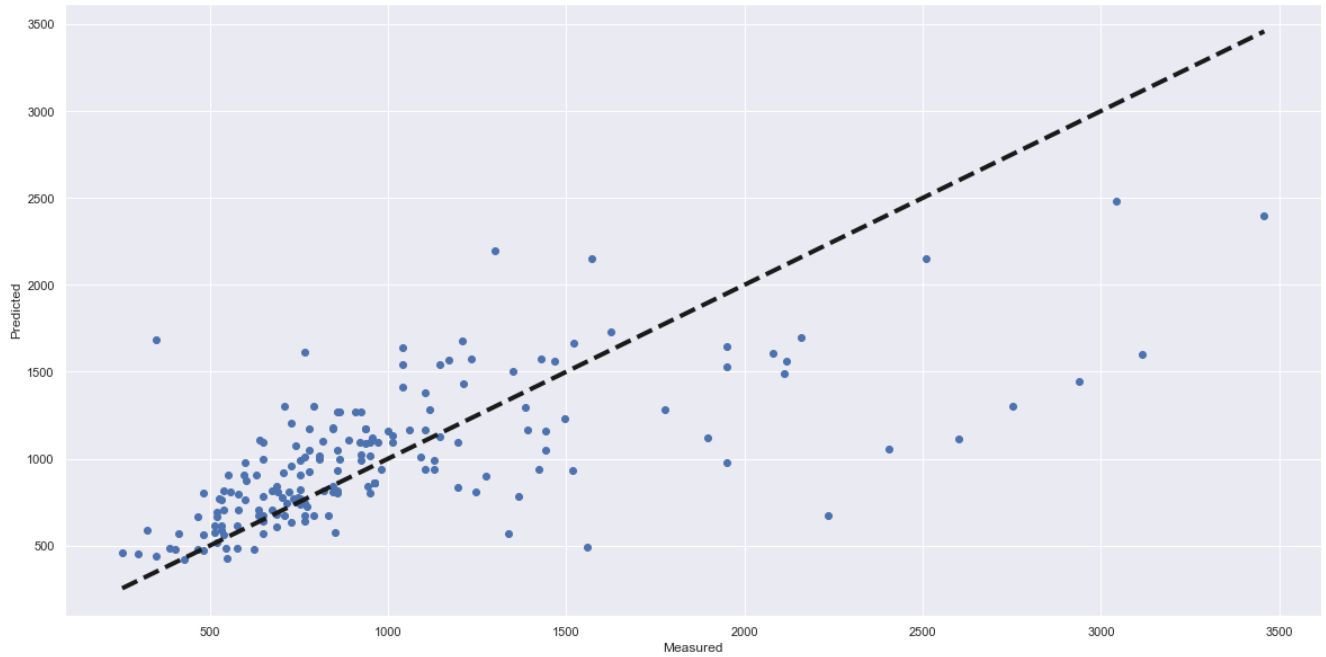
Đạo hàm từng biến độc lập để thấy mức độ hảnh hưởng khi tăng đơn vị của biến dự báo nào đó.

Sau khi tìm được mô hình tốt nhất, ta có thể dự đoán giá trị y_pred (latest_price)

	Lastest Price	Prediction
145	597.090	905.096
60	519.987	694.214
733	649.870	572.447
31	388.570	485.361
506	844.870	840.002
...	...	...
635	3457.974	2395.691
308	766.870	1611.099
384	935.870	1087.862
895	747.370	776.979
1	254.670	457.627

6.2.3 Vẽ đồ thị hiển thị giá trị dự đoán và sai số hồi quy:





− Ý nghĩa: đồ thị biểu thị sai số hồi quy và giá trị dự báo cho biến dự đoán.

− Nhận xét: giá trị phần dư (sai số) tập trung quanh đường y = 0 nên giả định các sai số có kỳ vọng bằng 0 thoả mãn. Mô hình dự đoán khá chính xác.Mô hình giải thích được 87,1% sự biến động của giá laptop.

PHẦN 7 – KẾT LUẬN

Qua những phần trên, nhóm đã cung cấp những góc nhìn tổng quan và sâu sắc về thị trường laptop hiện nay trong đó có sự hiện diện của các yếu tố phần cứng, phần mềm có ảnh hưởng đến giá laptop, qua đó ta có thể nhận thấy một vài thông tin hữu ích khi muốn mua một chiếc laptop bao gồm:

- Số lượng lớn nhất máy tính xách tay được sản xuất bởi các thương hiệu ASUS, DELL, HP, Lenovo. Những thương hiệu quá nổi tiếng này nên là sự lựa chọn hàng đầu khi muốn mua một chiếc laptop có thể sử dụng lâu dài với giá cả vừa phải.
- Ta có thể thấy sự chênh lệch về giá thành giữa các kích thước màn hình là không nhiều, nên việc chọn loại kích thước màn hình nào sẽ phụ thuộc vào nhu cầu sử dụng của mỗi cá nhân. Giá trị đa số ở đây là 15,6inch, cho thấy nhu cầu của người mua ở loại màn hình này khá cao. Điều này có lẽ do thực tế những máy tính xách tay được sử dụng linh hoạt ở nơi làm việc cũng như ở nhà. Màn hình 15,6inch là độ rộng vừa đủ có thể hiển thị một cách tốt nhất cũng như có thể bỏ vào tất cả mọi loại ba lô.
- Các bộ vi xử lý phổ biến nhất là bộ vi xử lý Intel, đặc biệt là Core i5, Core i7, Core i3. Có thể dễ hiểu bởi vì vi xử lý của intel có giá thành thấp và hiệu suất cao.
- Trong biến ram_gb, giá trị 8 Gb là xu hướng. Đồng nghĩa với việc dung lượng này đủ cho hầu hết các tác vụ. Ngay cả về giá cũng không quá chênh lệch so với 4Gb.
- Hầu hết các máy tính xách tay chỉ được trang bị một ổ SSD. Nhiều nhà sản xuất, muốn giành chiến thắng trong cuộc cạnh tranh, đã tiết kiệm các khe cắm trong máy tính xách tay bằng cách chỉ lắp đặt một loại ổ cứng. Ví dụ, chọn một laptop với chỉ một SSD, bạn có thể có nguy cơ hỏng hệ thống và mất luôn dữ liệu cá nhân. Lựa chọn một ổ cứng HDD cũng không thích hợp bởi vì loại ổ cứng này dùng nguyên lý hoạt động dựa trên đĩa cứng so với nguyên lý hoạt động của SSD tương tự như bộ nhớ RAM hay các loại thẻ nhớ, USB đó là sử dụng các chip nhớ flash. Hệ điều hành sử dụng HDD sẽ khởi động rất chậm và các ứng dụng sẽ gây khó chịu vì sự chậm chạp khi khởi chạy và xử lý dữ liệu. Do đó, khi chọn thiết bị, hãy ưu tiên hệ thống ổ đĩa kép SSD + HDD. Có thể với lựa chọn này giá sẽ chênh lệch từ 1tr-2tr nhưng là lợi ích lâu dài cũng như bảo toàn dữ liệu.
- Máy tính xách tay dành cho gaming với giá thành cao không thực sự phổ biến, bởi vì đối với nhu cầu chơi game, máy tính bàn là sự lựa chọn tốt hơn nhiều khi có giá thành rẻ, cấu hình mạnh cũng như không cần sự linh hoạt trong nhiều môi trường sử dụng khác nhau.
- Hầu như laptop hiện nay không được trang bị card đồ hoạ. Còn với tính chất công việc cần sử dụng đồ hoạ thì 4Gb dung lượng card là một sự lựa chọn tối ưu cho hiệu suất và giá thành. Đủ đáp ứng cho bạn những công việc đồ hoạ đơn giản, không quá phức tạp.

PHẦN 8 – PHỤ LỤC

PHẦN 9 – ĐÓNG GÓP

STT	MSSV	Họ và Tên	Công việc	Hoàn thành
35	20133076	Văn Mai Thanh Nhật	- Tóm tắt, giới thiệu, dữ liệu, kết luận - Kiểm định giả thuyết	100%
42	20133082	Huỳnh Minh Phước	- Tiền xử lý, Tham khảo - Kiểm định giả thuyết	100%
3	20133029	Nguyễn Trí Dũng	- Kiểm định giả thuyết - Trực quan hoá dữ liệu	100%
14	20133047	Lương Gia Huy	- Xây dựng mô hình dự đoán giá laptop - Kiểm định giả thuyết	100%

PHẦN 10 – THAM KHẢO

1. Quách Đình Hoàng, Lecture, 2021. Slide and video.
2. Mine Cetinkaya-Rundel, OpenIntro. OpenIntro Statistics, 4th Edition.
3. Santosh Kumar, Laptop Specs and latest price, Kaggle , 2022, đường dẫn: https://www.kaggle.com/datasets/kuchhbhi/latest-laptop-price-list
4. Santosh Kumar, Laptop Data Visualization, Kaggle , 2022, đường dẫn: https://www.kaggle.com/code/kuchhbhi/laptop-data-visualization
5. Georgy Zubkov, Kaggle, Laptop.EDA with recommendations, Kaggle, 2022, đường dẫn: https://www.kaggle.com/code/georgyzubkov/laptop-eda-with-recomendations
6. Thư viện python: pandas, scipy, numpy, plotly, seaborn, statsmodel, matplotlib, missingno, sklearn