**Shape, square

Description automatically generated**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

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

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**A picture containing text

Description automatically generated**

**Logo

Description automatically generated**

**TP. Hồ Chí Minh, tháng 5 năm 2023**

**Báo cáo: Tương tác dữ liệu trực quan**

**TRỰC QUAN, PHÂN TÍCH VÀ DỰ ĐOÁN DOANH SỐ BÁN HÀNG CỦA MỘT SIÊU THỊ TẠI THỊ TRƯỜNG MỸ**

|  |  |
| --- | --- |
| ***Nhóm sinh viên thực hiện :*** | |
| Văn Mai Thanh Nhật | 20133076 |
| Huỳnh Minh Phước | 20133082 |
| Trần Nguyên Hạnh | 20133013 |
| Trần Đông | 20133035 |
|  |  |
|  |  |

**Giảng viên : Ths. Lê Quang Thái**

**Mục lục**

[**phần 1 – Tóm tắt 3**](#_Toc106217175)

[**phần 2 – Giới thiệu 4**](#_Toc106217176)

[**phần 3 – Dữ liệu 6**](#_Toc106217177)

[**3.1 Tiền xử lý 6**](#_Toc106217178)

[**phần 4 – trực quan hoá dữ liệu 8**](#_Toc106217179)

[**4.1 Thống kê Thương hiệu: 8**](#_Toc106217180)

[**4.2 Thống kê dòng máy: 8**](#_Toc106217181)

[**4.3 Thống kê vi xử lý: 9**](#_Toc106217182)

[**4.4 Thống kê về RAM: 9**](#_Toc106217183)

[**4.5 Thống kê về ổ đĩa: 10**](#_Toc106217184)

[**Phần 5 - Mô hình hóa dữ liệu 11**](#_Toc106217185)

[**5.1 Kiểm định giả thuyết 11**](#_Toc106217186)

[**5.2 Mô hình dự đoán giá laptop 12**](#_Toc106217187)

[**Phần 6 – THỰC NGHIỆM, KẾT QUẢ, THẢO LUẬN 13**](#_Toc106217188)

[**6.1 Kiểm định giả thuyết: 13**](#_Toc106217189)

[6.1.1 Khoảng tin cậy của giá cho từng biến không được liệt kê thành nhóm: 13](#_Toc106217190)

[6.1.2 Khoảng tin cậy cho từng nhóm mẫu máy trong từng thương hiệu 16](#_Toc106217191)

[6.1.3 Khoảng tin cậy cho từng nhóm mẫu vi xử lý trong từng hãng vi xử lý 17](#_Toc106217192)

[**6.2 Mô hình dự đoán giá laptop: 18**](#_Toc106217193)

[6.2.1 Định nghĩa: 18](#_Toc106217194)

[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. 19](#_Toc106217195)

[6.2.3 Vẽ đồ thị hiển thị giá trị dự đoán và sai số hồi quy: 22](#_Toc106217196)

[**Phần 7 – kết luận 23**](#_Toc106217197)

[**Phần 8 – Phụ lục 24**](#_Toc106217198)

[**Phần 9 – đóng góp 24**](#_Toc106217199)

[**Phần 10 – tham khảo 25**](#_Toc106217200)

# **phần 1 – Tóm tắt**

Đề tài "Trực quan, phân tích và dự đoán dữ liệu bán hàng của một siêu thị tại thị trường Mỹ" nhằm mục đích nghiên cứu và phân tích dữ liệu bán hàng của một siêu thị tại Mỹ trong vòng 4 năm để đưa ra những kết luận hữu ích giúp cải thiện hiệu quả kinh doanh. Nghiên cứu sử dụng các phương pháp trực quan hóa dữ liệu, phân tích đa biến, và dự đoán thống kê để khai thác thông tin từ dữ liệu bán hàng và đưa ra các dự đoán về xu hướng tiêu dùng và nhu cầu của khách hàng.

Trong đề tài này , các phương pháp phân tích và trực quan hóa dữ liệu được nhóm sử dụng gồm:

1. Phân tích đa biến: Dùng để phân tích mối quan hệ giữa các biến độc lập với biến phụ thuộc và đưa ra những kết luận về mức độ ảnh hưởng của các yếu tố đến doanh số bán hàng.
2. Trực quan hóa dữ liệu: Sử dụng các công cụ như biểu đồ, đồ thị để hiển thị dữ liệu bán hàng dưới dạng hình ảnh rõ ràng, giúp nhà quản lý có cái nhìn tổng quan về tình hình bán hàng và dễ dàng phát hiện ra các xu hướng và mô hình tiêu thụ của khách hàng.
3. Dự đoán thống kê: Sử dụng các phương pháp dự đoán như hồi quy tuyến tính, mô hình dự đoán chuỗi thời gian để đưa ra dự đoán về doanh số bán hàng trong tương lai dựa trên các mẫu và xu hướng trong dữ liệu quá khứ.

Kết quả của bài nghiên cứu này sẽ giúp định hướng chiến lược kinh doanh cho siêu thị thời gian đó, đồng thời cung cấp thông tin hữu ích cho các nhà quản lý và nhà đầu tư trong lĩnh vực bán lẻ tại Mỹ.

# **phần 2 – Giới thiệu**

Đề tài sẽ sử dụng các phương pháp phân tích và trực quan hóa dữ liệu để hiểu rõ hơn về sự phân bố, phân tích và tương quan giữa các yếu tố ảnh hưởng đến doanh số bán hàng của siêu thị, bao gồm: ngày, tháng, mùa, các ngày lễ, giá cả, phương thức vận chuyển, phân khúc khách hàng, loại sản phẩm, và các yếu tố kinh tế khác.

Ngoài ra, đề tài cũng sử dụng các thuật toán dự đoán chuỗi thời gian như ARIMA, Prophet để dự đoán doanh số bán hàng tương lai của siêu thị, giúp các nhà quản lý có thể đưa ra các quyết định kinh doanh phù hợp để tăng trưởng doanh số và cải thiện kinh doanh.

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

* Mã đơn hàng
* Ngày đặt hàng
* Ngày nhận hàng
* Phương thức vận chuyển
* Mã khách hàng
* Tên khách hàng
* Phân khúc khách hàng
* Quốc gia
* Thành phố
* Tiểu bang
* Mã bưu điện
* Vùng miền
* Mã sản phẩm
* Loại sản phẩm
* Loại sản phẩm phụ
* Tên sản phẩm
* Giá bán

Nhóm sử dụng những thuật toán:

* Linear Regression (Hồi quy tuyến tính): Sử dụng để xác định mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc, ví dụ như mối quan hệ giữa số lượng sản phẩm được bán và giá bán.
* K-Means Clustering (Phân cụm K-means): Sử dụng để phân cụm các dữ liệu bán hàng thành các nhóm tương đồng với nhau, giúp nhìn rõ hơn các đặc tính của khách hàng và thị trường bán hàng.
* ARIMA (Auto-Regressive Integrated Moving Average): Sử dụng để mô hình hóa và dự đoán chuỗi thời gian, có thể giúp đưa ra dự đoán về doanh số bán hàng trong tương lai.
* Prophet: là một mô hình dự đoán chuỗi thời gian được phát triển bởi Facebook. Mô hình sử dụng một phương pháp phân tích chuỗi thời gian mới, gọi là mô hình tuyến tính theo ngày, để đưa ra dự đoán. Sử dụng Prophet ta có thể phân tích chuỗi thời gian về doanh số bán hàng của siêu thị, từ đó đưa ra các thông tin về xu hướng tăng giảm, định hướng và dự báo cho tương lai.

# **phần 3 – Dữ liệu**

Trong đề tài này, nhóm xin dùng bộ dữ liệu *Superstore Sales Dataset* 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.

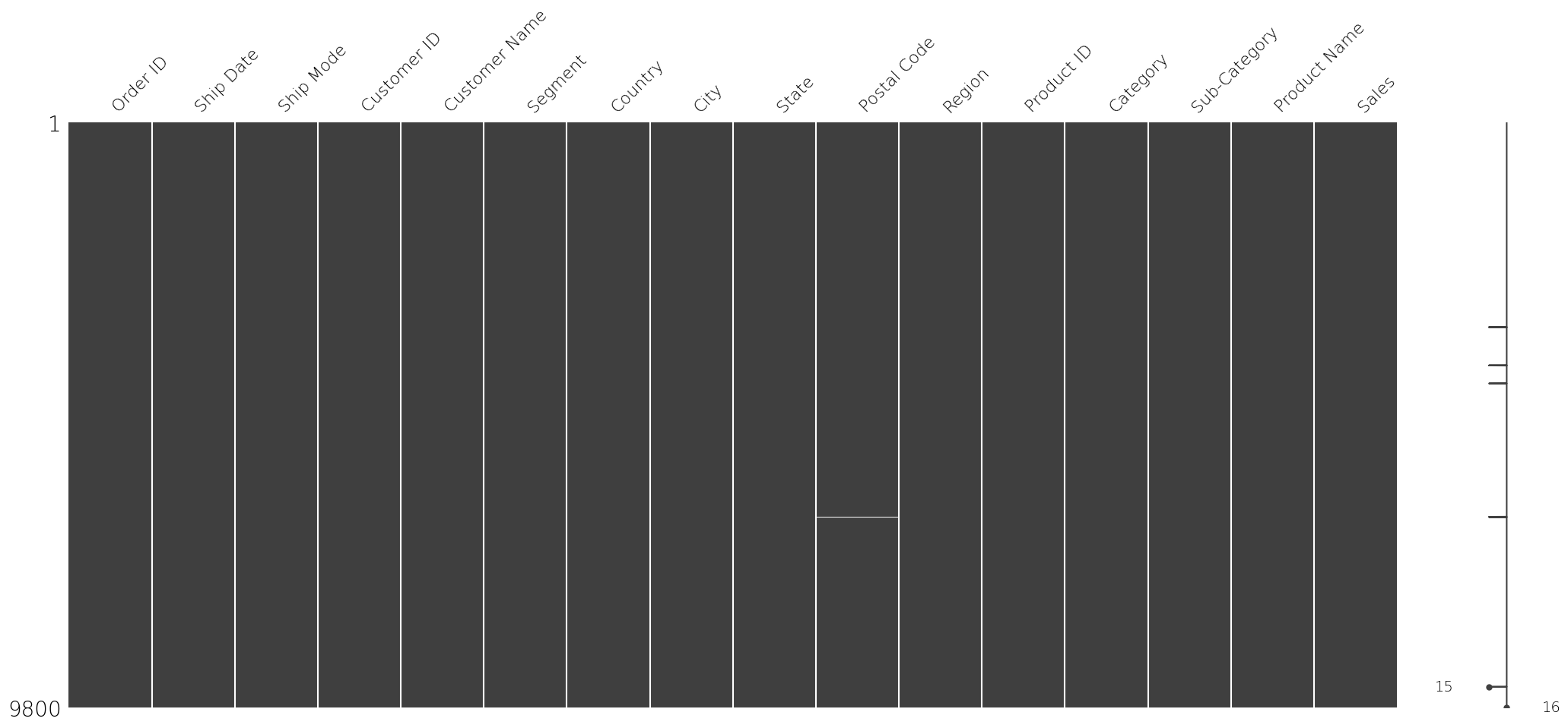
Theo như tác giả trên kaggle, tập dữ liệu xuất hiện trong nhóm người dùng tháng 12 của cộng đồng tableau. Hiện tại thì nhóm này đã không còn hoạt động và vẫn không tìm được tác giả của tập dữ liệu này là ai, nó xuất phát từ công ty hay doanh nghiệp nào và cũng không biết cách mà tập dữ liệu được thu thập.

Bộ dữ liệu chứa hơn 9800 hàng, đã được thực hiện tiền xử lý, mỗi hàng đại diện cho một giao dịch bán hàng.

* Bộ dữ liệu gồm 18 cột, :
  + Row ID: Số thứ tự
  + Order ID: Mã đơn hàng
  + **Product ID:** Mã định danh duy nhất cho mỗi sản phẩm.
  + **Order Date:** Ngày đặt hàng.
  + Ship Date: Ngày nhận hàng
  + Ship Mode: Phương thức vận chuyển
  + **Sales:** Giá bán cho mỗi sản phẩm.
  + **Customer ID:** Mã định danh duy nhất cho mỗi khách hàng.
  + Customer Name: Tên khách hàng
  + **Segment:** Phân khúc của khách hàng, chẳng hạn như bán lẻ hoặc bán buôn.
  + **Country:** Quốc gia nơi khách hàng đặt hàng.
  + **City:** Thành phố nơi khách hàng đặt hàng.
  + **State:** Tiểu bang hoặc tỉnh nơi khách hàng đặt hàng.
  + **Postal Code:** Mã bưu chính nơi khách hàng đặt hàng.
  + **Region:** Khu vực nơi khách hàng đặt hàng.
  + Product ID: Mã sản phẩm
  + **Category:** Loại sản phẩm, chẳng hạn như đồ nội thất hoặc công nghệ.
  + **Sub-Category:** Danh mục con của sản phẩm, chẳng hạn như bàn hoặc laptop.
  + Product name: Tên sản phẩm

**3.1 Tiền xử lý**

Qua một vài bước chuẩn hóa dữ liệu ngày tháng, loại bỏ những cột không cần thiết, thì ta thực hiện trực quan hóa dữ liệu bị thiếu như hình

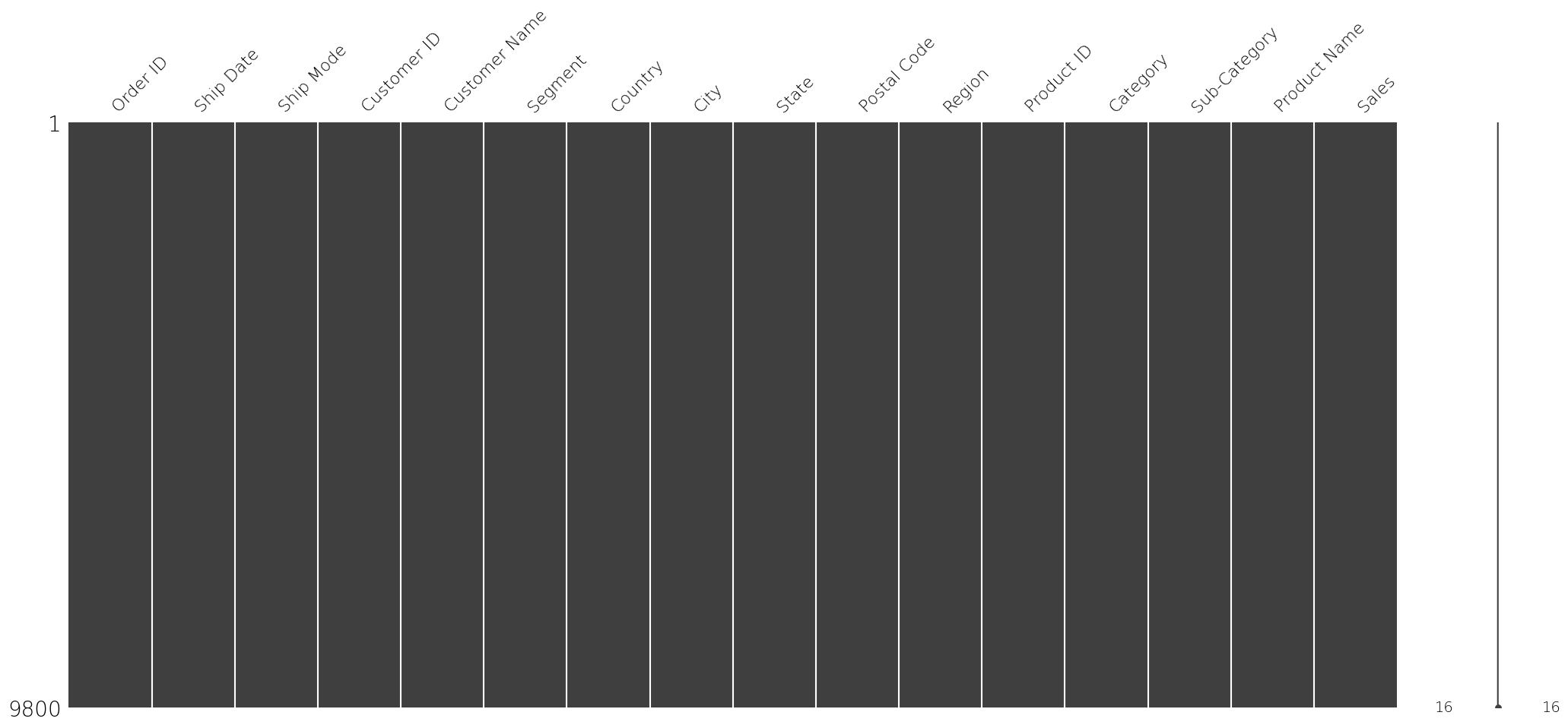
Ta có thể thấy ở dữ liệu một ít dữ liệu bị thiếu ( missing value) ở những cột Postal Code.

Thì để xử lý những giá trị bị thiếu này, ta sẽ không bỏ nguyên bản ghi này mà mà thay vào đó ta sẽ thêm Postal code cho từng thành phố tương ứng

1. Đầu tiên thì ta cần tìm những thành phố có postal code bị thiếu

2. Thực hiện điền postal code tương ứng với thành phố ở từng hàng bị thiếu giá trị

Ta có thể thấy giá trị mã bưu chính bị thiếu ở đây chỉ có thành phố Burlington ở bang Vermont. Ta có mã bưu chính của thành phố này là 5401

****

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 Khách hàng:**

4.1.1. Những khách hàng có doanh số cao nhất

4.1.2 Tỷ lệ doanh thu theo từng phân khúc khách hàng

## **4.2 Địa chỉ khách hàng:**

4.2.1. Những tiểu bang có doanh số cao nhất

4.2.2. Những thành phố có doanh số cao nhất

4.2.3. Tỷ lệ doanh thu theo từng vùng miền

4.2.4. Tỷ lệ doanh thu theo từng phương thức vận chuyển

## **4.3 Sản phẩm:**

4.3.1. Những sản phẩm có doanh số cao nhất

4.3.2. Tỷ lệ doanh thu theo từng danh mục sản phẩm, danh mục phụ sản phẩm

## **4.4 Thống kê về RAM:**

## **4.5 Thống kê về ổ đĩa:**

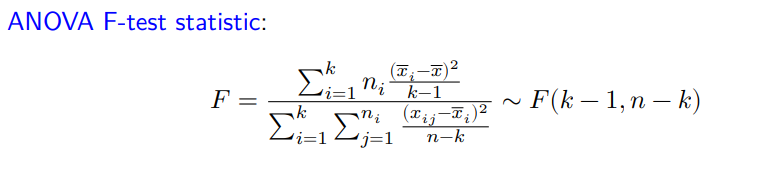
# **Phần 5 - Mô hình hóa dữ liệu**

## **5.1 Kiểm định giả thuyết**

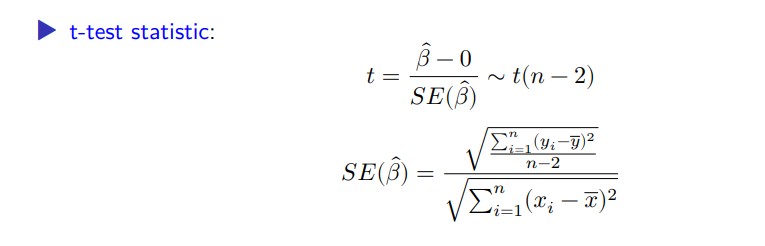
Kiểm định chứng minh các yếu tố phần cứng có ảnh hưởng tới giá thành laptop.

**Thuật toán sử dụng:**

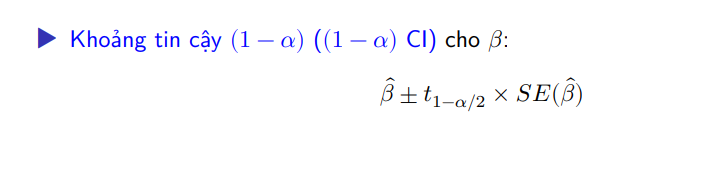
* ANOVA F-test đối với các biến có nhiều nhóm độc lập



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



* Công thức tính khoảng tin cậy:



**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**

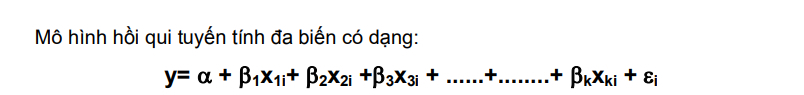
Xây dựng mô hình hồi quy tuyến tính. Chọn 80% dữ liệu để huấn luyện cho mô hình và 20% test mô hình.

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.



# **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:

A picture containing text, wall, electronics, display

Description automatically generated

**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**

A picture containing text, crossword puzzle, electronics, display

Description automatically generated

**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**

Graphical user interface, application

Description automatically generated

**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 , , ..., và sai số ngẫu nhiên ε.

(với ,,…, là các tham số)

* Hệ số xác định : 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, càng lớn thì mô hình càng ý nghĩa.
* Tiêu chí thông tinh Akaike (AIC)
* Tiêu chí thông tin Bayesian (BIC)

(với k là tham số của mô hình)

Đối với hồi quy tuyến tính, với giả định Guass thì tỉ lệ với

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.

Graphical user interface, table

Description automatically generatedMô 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 Mô hình 3: Sau khi đã loại bỏ biến review

Graphical user interface, table

Description automatically generated

Graphical user interface, table

Description automatically generated

Graphical user interface, table

Description automatically generated

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:

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)

Graphical user interface, table

Description automatically generated

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

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

− 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