**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

🙠🙟🕮🙝🙢



**MÔN HỌC: TƯƠNG TÁC DỮ LIỆU TRỰC QUAN**

**ĐỒ ÁN CUỐI KÌ**

**ĐỀ TÀI: TRỰC QUAN HOÁ VÀ PHÂN TÍCH DỮ LIỆU ĐIỆN THOẠI**

**GVHD**:

**Nhóm sinh viên thực hiện:** Nhóm 2

1. Nguyễn Trí Dũng 20133029
2. Nguyễn Khoa Quang Thắng 20133090
3. Lê Minh Đăng 20133

**Mã môn học**: RPAN233577\_22\_1\_02

Thành phố Hồ Chí Minh, Tháng 12 năm 2022

# PHẦN 1 – TÓM TẮT

Điện thoại - vật dụng không thể thiếu trong cuộc sống hiện nay khi mà toàn cầu đang tiến tới công nghệ hóa mọi thứ trong cuộc sống. Với những chức năng đa dạng và ngày càng hiện đại (như soạn văn bản Word, xem phim với độ phân giải cao,…), cộng thêm những chiếc điện thoại ngày nay có khả năng được những thiết bị khác (như máy ảnh, máy nhắn tin, máy ghi âm,…) thì việc càng nhiều người chọn điện thoại là vật bất ly thân là không thể chối cãi. Cấu hình điện thoại là một trong những yếu tố đóng vai trò không thể thiếu để giúp người dùng có được những trải nghiệm tuyệt vời nhất trên thiết bị di động thông minh. Tuy nhiên, cấu hình điện thoại không phải là một khái niệm hoàn toàn đơn giản như bạn nghĩ. Bởi vậy nên có rất nhiều dòng điện thoại thông minh hiện nay được nhà sản xuất trang bị chip xử lý lên tới 8 nhân nhưng vẫn hoạt động kém nhanh nhạy và trơn tru.

Nhận thấy điều đó, nhóm 2 nhận thức được rằng việc người mua hàng lựa chọn cấu hình và giá cả khi mua điện thoại luôn được người tiêu dùng đặt lên hàng đầu. Do vậy, để hiểu rõ hơn về sự ảnh hưởng của những chức năng, cấu hình điện thoại tới giá cả, nhóm 2 chọn đề tài là: “Phân tích mô hình dự đoán giá điện thoại qua cấu hình máy”.

Các phương pháp nhóm phân tích sử dụng:

* 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.

# PHẦN 2 – GIỚI THIỆU

Khi giải quyết được vấn đề “Phân tích mô hình dự đoán giá điện thoại qua cấu hình máy” giúp chúng ta có thể lựa chọn một chiếc máy điện thoại vừa có cấu hình mình lại có giá cả vừa với số tiền bỏ ra. Ngày nay, điện thoại thông minh hầu như là vật bất li thân đôi với mỗi người trong xã hội hiện đại bây giờ. Với sự phát triển công nghệ mạnh mẽ như hiện nay, các thương hiện sản xuất điện thoại đang cạnh tranh khốc liệt, không ngừng nâng cao chất lượng sản phẩm để đưa đến tay người dùng một sản phẩm hoàn hảo nhất, phù hợp nhất. Thí dụ như một số thương hiệu lớn như Apple, SamSung, Oppo, OnePlus, Xiaomi,… Với những sản phẩm với chất lượng của mình đã đưa đến cho người dùng cảm nhận về sự hiện đại của công nghệ đối với cuộc sống. Input của bài toán là những thông số máy chi tiết của từng chiếc điện thoại thông minh thuộc những hãng điện thoại hiện có trên thị trường hiện nay. Nhóm đã sử dụng thuật toán linear regression, k-NNđể dự đoán giá của điện thoại thông minh dựa trên cấu hình của điện thoại đó.

Các công cụ được nhóm sử dụng trong dự án:

* *Ngôn ngữ R:*R là một ngôn ngữ lập trình và môi trường phần mềm dành cho tính toán và đồ họa thống kê. Đây là một bản hiện thực ngôn ngữ lập trình S với ngữ nghĩa khối từ vựng lấy cảm hứng từ Scheme. Xuất hiện lần đầu vào năm 1993 và được tích hợp với nhiều hệ điều hành. R cũng có tính mở rộng cao bằng cách sử dụng các gói cho người dùng đưa lên cho một số chức năng và lĩnh vực nghiên cứu cụ thể. Do được thừa hưởng từ Schema, R có nền tảng lập trình hướng đối tượng mạnh hơn đa số các ngôn ngữ tính toán thống kê khác. Việc mở rộng R cũng dễ dàng nhờ các luật đóng khối từ vựng.
* *RStudio:* R-Studio là một nhóm các tiện ích phân tích dữ liệu đầy đủ chức năng. Chương trình có phiên bản cho cả hệ điều hành Windows, Mac và Linux. Phần mềm RStudio là một công cụ hỗ trợ IDE rất được mọi người ưa thích, nhất là trong mảng giáo dục khi làm việc với R. RStudio cho phép truy cập trực tiếp vào phần mềm R và còn hỗ trợ rất nhiều tính năng nâng cao, giúp cho việc quản lý và phân tích dữ liệu với R được hiệu quả hơn.

# PHẦN 3 – DỮ LIỆU

**3.1) Thông tin về dữ liệu:**

Và trong đề tài này, dataset được nhóm 2 thống nhất sử dụng có tên là Mobile Phone Specifications and Prices được tải xuống từ Kaggle và tác giả của dataset là PRATIK GARAI. Theo tác giả, dataset được lấy từ website Gadgets360 - là một trang báo của Ấn Độ được cập nhật các thông tin trên toàn thế giới. 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 (họ từng làm việc cho Amazon) thành lập vào tháng 10 năm 2007.

Nhóm chia tập dữ liệu ra làm 10 phần, 8 phần sẽ sử dụng làm training và 2 phần sẽ sử dụng để test.

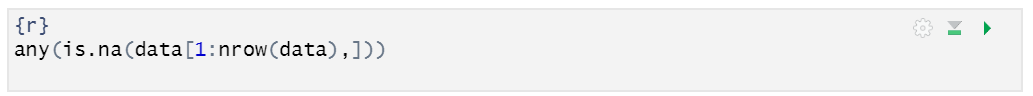
****3.2) Tiền xử lý:****

Nhóm đã chuyển đổi Rupee(đơn vị tiền tệ Ấn Độ) sang Dollar



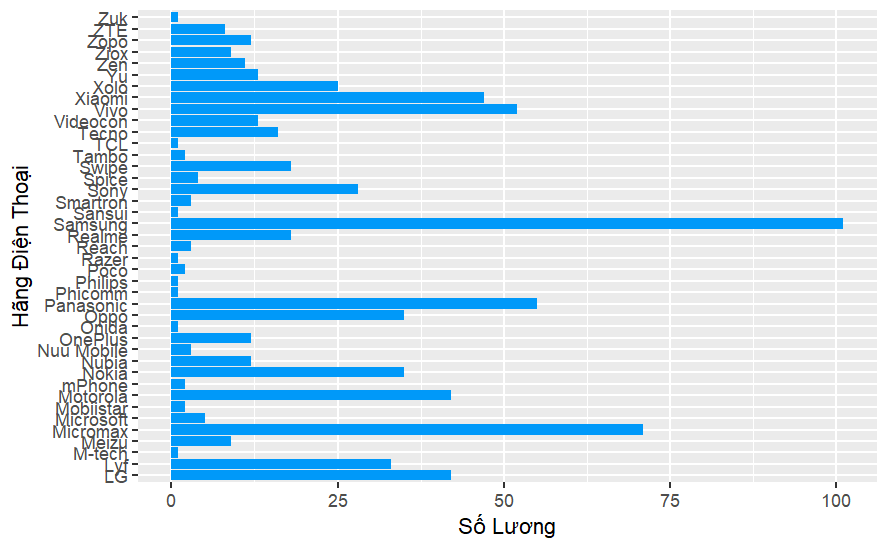
Nhóm đã chuyển đổi lại thuộc tính để thuận tiện cho việc xử lí:

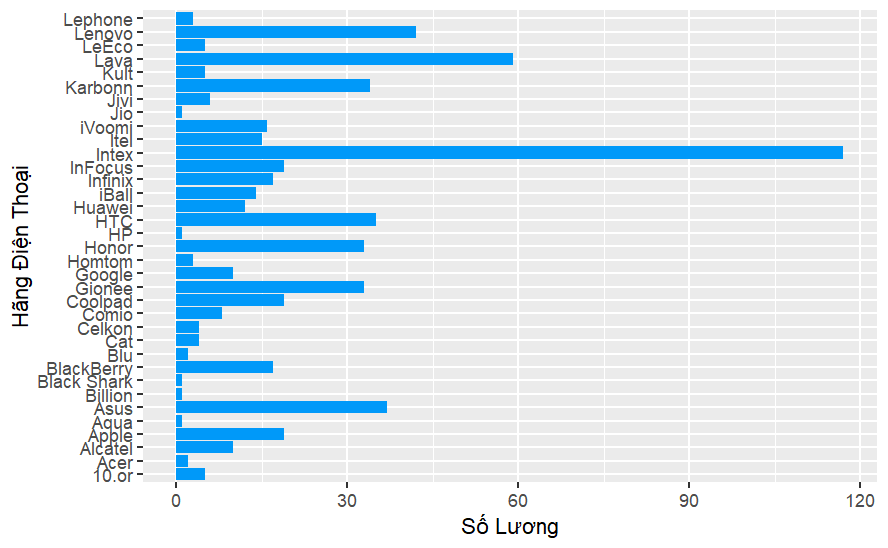


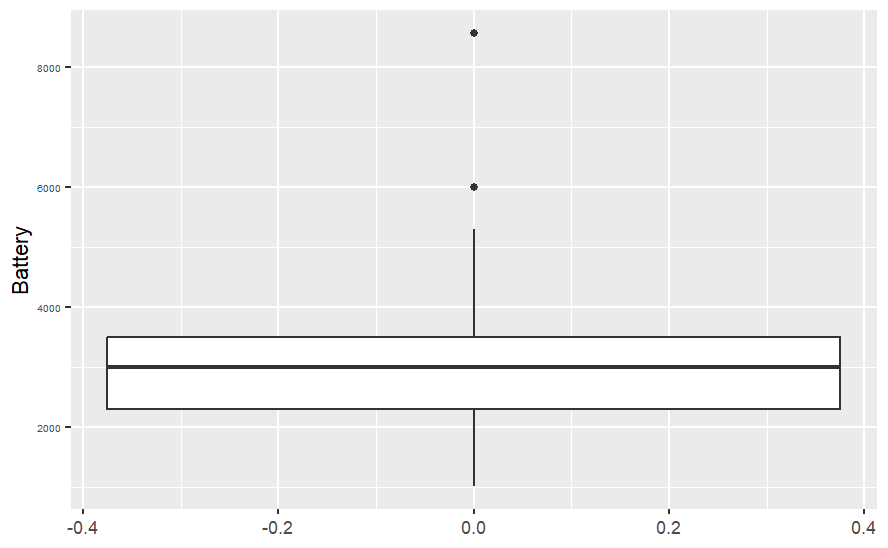
Kiểm tra xem có bất kì ô dữ liệu nào bị rỗng:

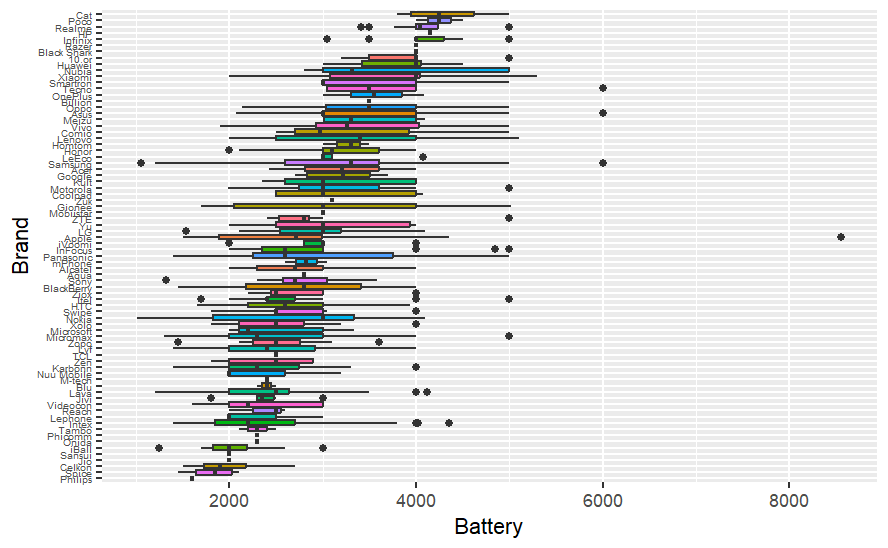
# PHẦN 4 – TRỰC QUAN HOÁ DỮ LIỆU

Thống kê số lượng điện thoại:

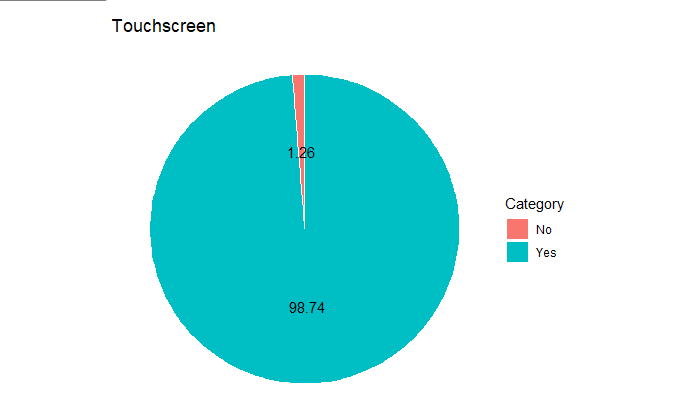


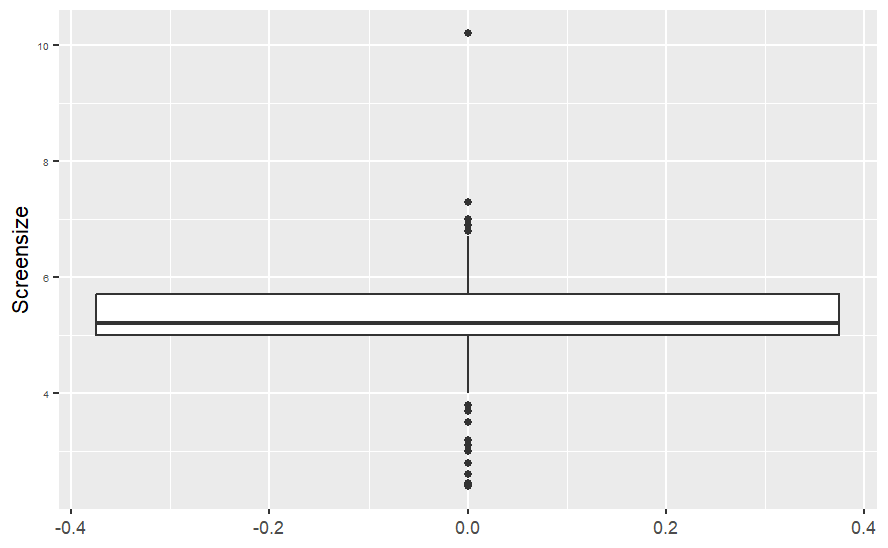


Thống kê số lượng pin:

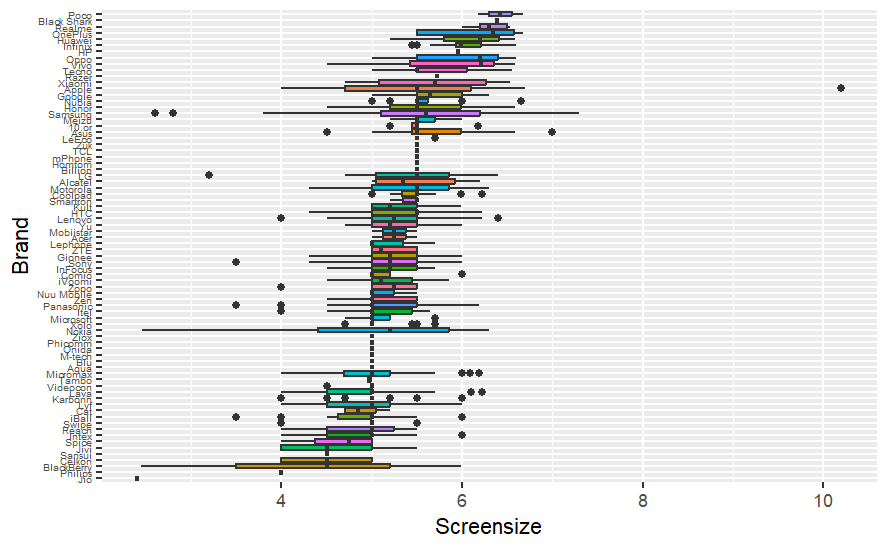


Thống kê số lượng sản phẩm có màn hình cảm ứng:

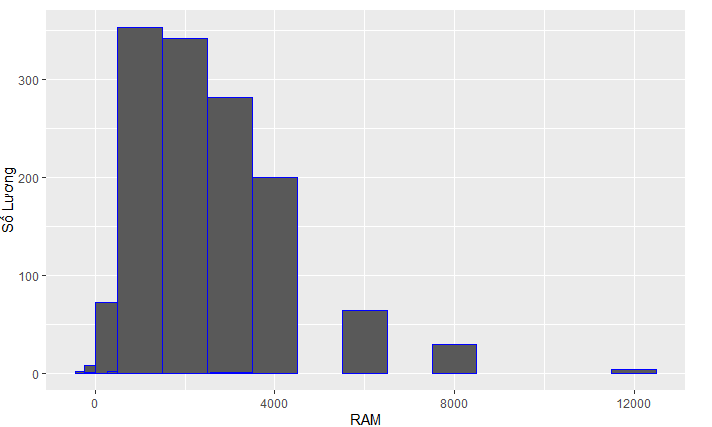


Thống kê kích cỡ màn hình của tất cả điện thoại:

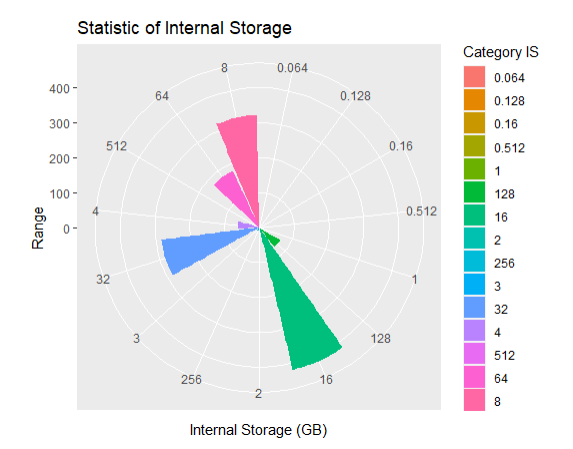
Thống kê kích cỡ màn hình của từng hãng điện thoại:



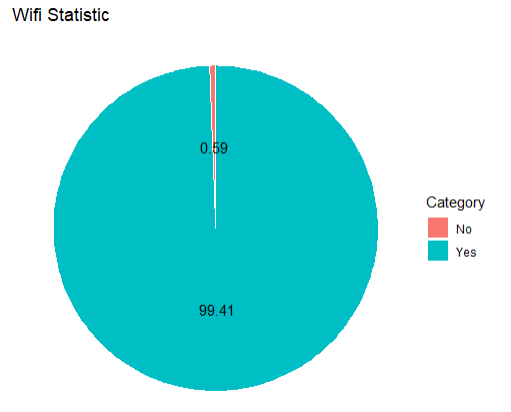
Thống kê RAM của điện thoại trên tập data phần nhiều là 1GB, 2GB, 3GB, 4GB.



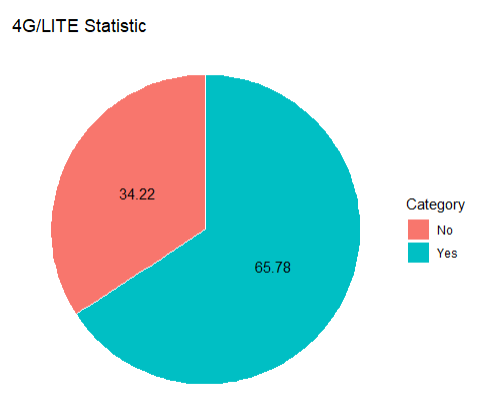
Thống kê Bộ nhớ trong của điện thoại trên tập data phần nhiều là 16GB.



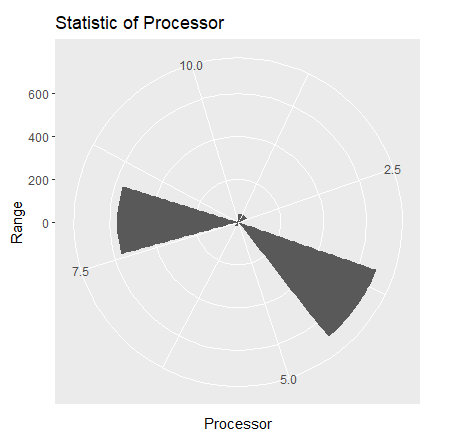
Thống kê số lượng sản phẩm có hỗ trợ wifi trên tập training là 99.41% và không có hỗ trợ wifi là 0.59%.

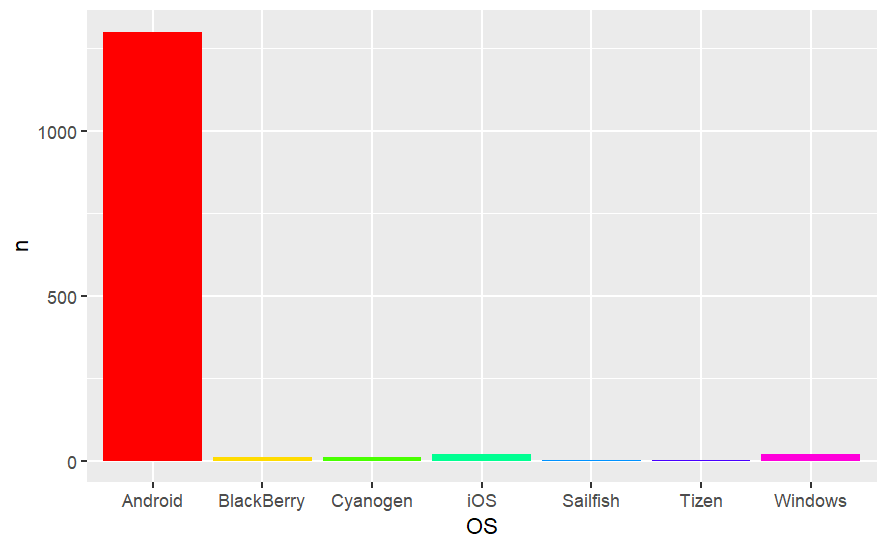


Thống kê số lượng sản phẩm có hỗ trợ 4G trên tập training là 65.78% và không có hỗ trợ 4G là 34.22%.



Thống kê Processor của điện thoại trên tập training phần nhiều là 4 và 8.



Thống kê OS:

# PHẦN 5 – MÔ HÌNH HOÁ DỮ LIỆU

****5.1) Kiểm định giả thuyết:****

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

****5.1.1) Point Biserial:****

a) Hệ số tương quan điểm

là giá trị trung bình trên biến liên tục X đối với tất cả các điểm

dữ liệu trong nhóm 1

là giá trị trung bình trên biến liên tục X đối với tất cả các điểm dữ liệu trong nhóm 0

là số điểm dữ liệu trong nhóm 1

là số điểm dữ liệu trong nhóm 0

b) Có công thức sử dụng

****5.1.2) ANOVA F test:****

D

* Tổng bình phương giữa các điều trị/điều kiên

N: Tổng số đối tượng

* Tổng bình phương cho các đối tượng trong các điều trị/điều kiện

* Tổng bình phương tổng số sửa đổi
* Tổng bình phương sai số

-

**b) Tính bậc tự do:**

-

c) Đánh giá bình phương trung bình:

**d) Tính F-ratio:**

Kiểm định giả thiết F-test: /

5.2) Phương pháp thực hiện:

**Đối với ANOVA F test:**

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

**Đối với Correlation test:**

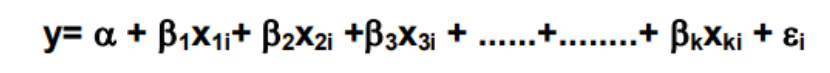
* H0: Biến độc lập và biến kết quả không có sự tương quan với nhau
* H1: Biến độc lập và biến kết quả có sự tương quan với nhau

**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.

5.3) Xây dựng mô hình dự đoán giá điện thoại:

5.3.1) Mô hình tuyến tính multi linear regression:

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

5.3.2) Mô hình k-NN:

K-nearest neighbors là thuật toán học máy có giám sát, đơn giản và dễ triển khai. Thường được dùng trong các bài toán phân loại và hồi quy.

Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm củng có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp.

# PHẦN 6 - THỰC NGHIỆM, KẾT QUẢ, VÀ THẢO LUẬN

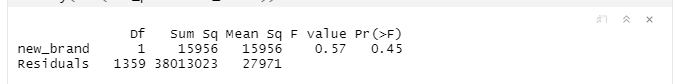
6.1) Kiểm định giả thuyết:

6.1.1) Anova F test:

**a) Brand:**

H0: Không có sự khác biệt về giá trung bình giữa các hãng điện thoại.

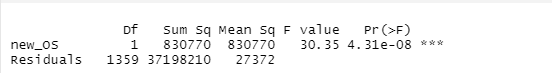
H1: Có sự khác biệt về giá trung bình giữa các hãng điện thoại.



Do p = 0.369 > 0.05 => Không có sự khác biệt về giá trung bình giữa các hãng điện thoại.

**b) Operating system:**

H0: Không có sự khác biệt về giá trung bình giữa điện thoại có hệ điều hành khác nhau.

H1: Có sự khác biệt về giá trung bình giữa điện thoại có hệ điều hành khác nhau.

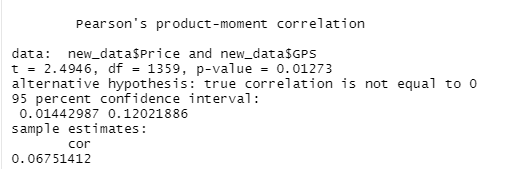
Do p = 0.369 > 0.05 => Không có sự khác biệt về giá trung bình giữa các hãng điện thoại.

6.1.2) Point Biserial:

**a) GPS:**

H0: Không có mối quan hệ tuyến tính giữa GPS và giá.

H1: Có mối quan hệ tuyến tính nào giữa GPS và giá.

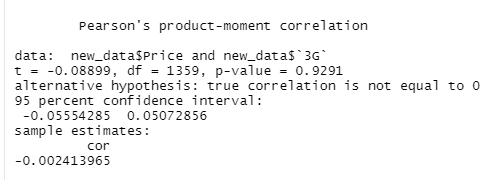


Do cor khác 0=> Có mối quan hệ tuyến tính giữa GPS và giá.

**b) 3G:**

H0: Không có mối quan hệ tuyến tính giữa 3G và giá.

H1: Có mối quan hệ tuyến tính nào giữa 3G và giá.

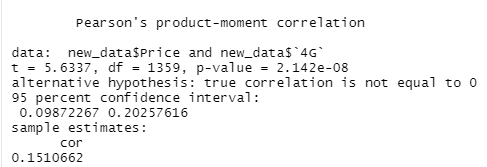


Do cor xấp xỉ 0=> Không mối quan hệ tuyến tính giữa 3G và giá.

**c) 4G:**

H0: Không có mối quan hệ tuyến tính giữa 4G và giá.

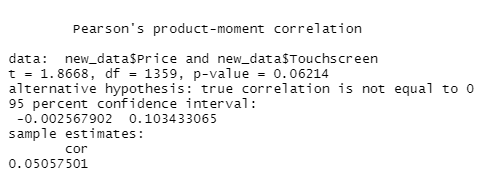
H1: Có mối quan hệ tuyến tính nào giữa 4G và giá.

****

**d) Touchscreen:**

H0: Không có mối quan hệ tuyến tính giữa điện thoại có cảm ứng và không có cảm ứng.

H1: Có mối quan hệ tuyến tính giữa điện thoại có cảm ứng và không có cảm ứng.



cor != 0=> Có mối quan hệ tuyến tính giữa Touchscreen và giá.

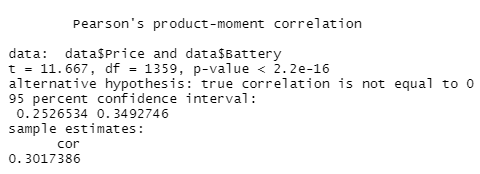
6.1.3) Phân tích tương quan Pearson:

**a) Battery Capacity:**

H0: Không có mối quan hệ tuyến tính giữa battery capacity và giá.

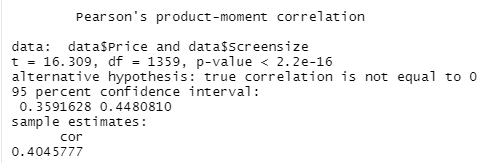
H1: Có mối quan hệ tuyến tính giữa battery capacity và giá.

cor != 0=> Có mối quan hệ tuyến tính giữa Battery Capacity và giá.

**b) Screensize:**

H0: Không có mối quan hệ tuyến tính giữa screensize và giá.

H1: Có mối quan hệ tuyến tính giữa screensize và giá.

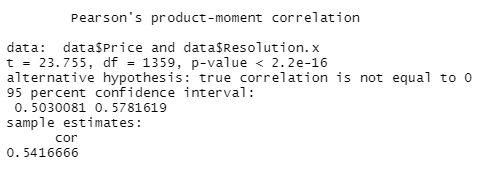


Cor !=0=> Có mối quan hệ tuyến tính giữa Screensize và giá

**c) Resolution.x:**

H0: Không có mối quan hệ tuyến tính giữa resolution.x và giá

H1: Có mối quan hệ tuyến tính giữa resolution.x và giá

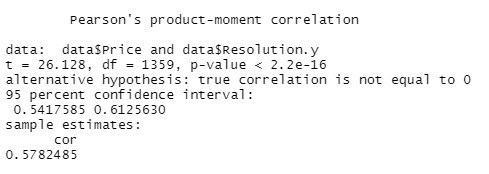
****

Cor !=0=> Có mối quan hệ tuyến tính giữa Resolution.x và giá

**d) Resolution.y:**

H0: Không có mối quan hệ tuyến tính giữa resolution.y và giá

H1: Có mối quan hệ tuyến tính giữa resolution.y và giá

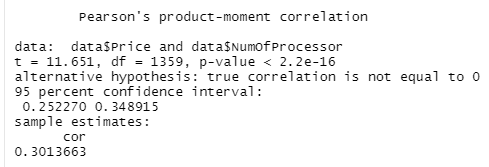
****

cor != 0=> Có mối quan hệ tuyến tính giữa Resolution.y và giá.

**e) Processor:**

H0: Không có mối quan hệ tuyến tính giữa NumOfProcessor và giá

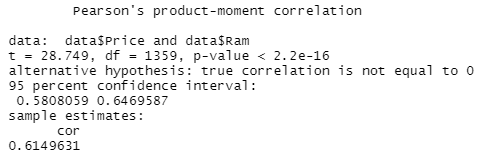
H1: Có mối quan hệ tuyến tính giữa NumOfProcessor và giá

****

cor != 0=> Có mối quan hệ tuyến tính giữa NumOfProcessor và giá

**f) RAM:**

H0: Không có mối quan hệ tuyến tính giữa ram và giá

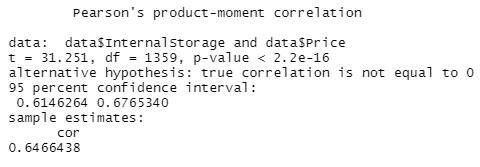
H1: Có mối quan hệ tuyến tính giữa ram và giá****

cor != 0=> Có mối quan hệ tuyến tính giữa RAM và giá

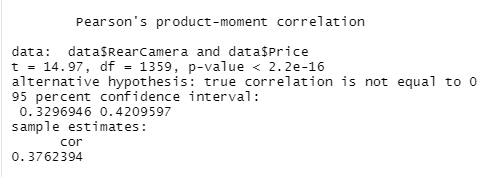
**g) Internal storage:**

H0: Không có mối quan hệ tuyến tính giữa Internal storage và giá

H1: Có mối quan hệ tuyến tính giữa Internal storage và giá

****cor != 0=> Có mối quan hệ tuyến tính giữa Internal Storage và giá

**h) Rear camera:**

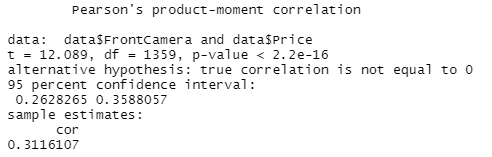
****H0: Không có mối quan hệ tuyến tính giữa Rear camera và giá

H1: Có mối quan hệ tuyến tính giữa Rear camera và giá

cor != 0=> Có mối quan hệ tuyến tính giữa Rear Camera và giá.

**i) Front camera:**

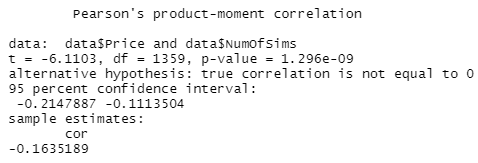
H0: Không có mối quan hệ tuyến tính giữa Front camera và giá

****H1: Có mối quan hệ tuyến tính giữa Front camera và giá

cor != 0=> Có mối quan hệ tuyến tính giữa Front camera và giá.

**j) SIM:**

H0: Không có mối quan hệ tuyến tính giữa NumOfSims và giá

****H1: Có mối quan hệ tuyến tính giữa NumOfSims và giá

cor != 0=> Có mối quan hệ tuyến tính giữa NumOfSims và giá.

6.2) Mô hình dự đoán giá điện thoại:

6.2.1) Mô hình tuyến tính:

a) 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 𝑋1, 𝑋2, ..., 𝑋𝑛 và sai số ngẫu nhiên ε.

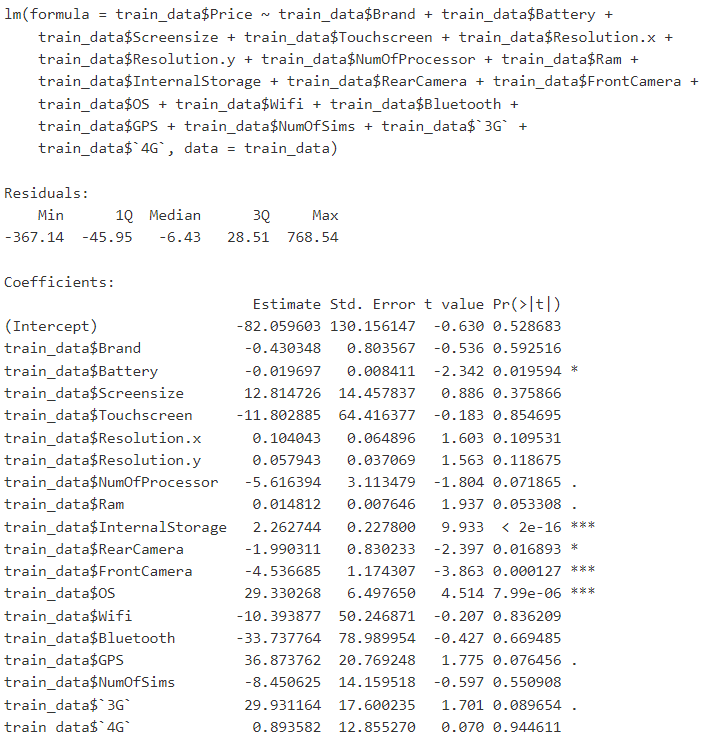
𝑌 = 𝛼 + 𝛽1𝑋1 + 𝛽2𝑋2 + ⋯ + 𝛽𝑛𝑋𝑛 + ε

(với 𝛽1, 𝛽2,…, 𝛽𝑛 là các tham số)

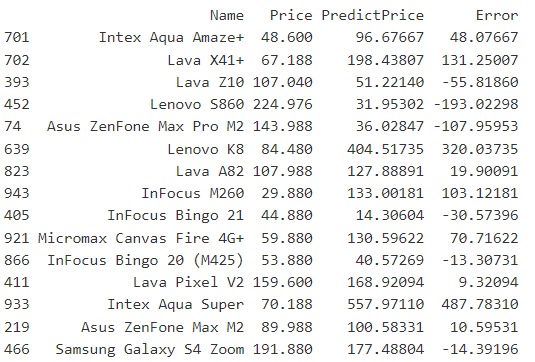
Hệ số xác định 𝑟2: dùng để đo mức độ ảnh hưởng của yếu tố được xem xét trong mô hình đối với sự biến động của các giá trị của các biến ngẫu nhiên quanh giá trị trung bình của nó, 𝑟2 càng lớn thì mô hình càng ý nghĩa.

b) Thực nghiệm:

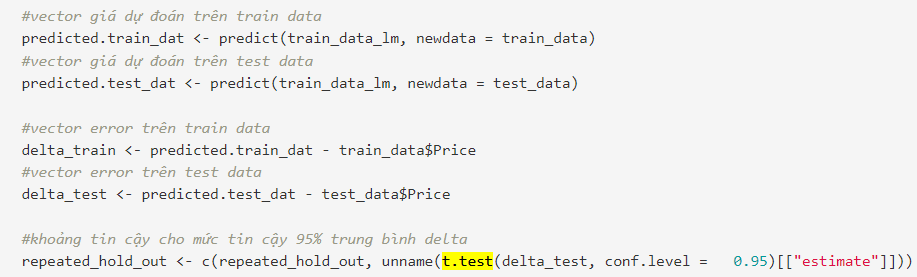
**Mô hình tuyến tính cho tất cả hãng:**

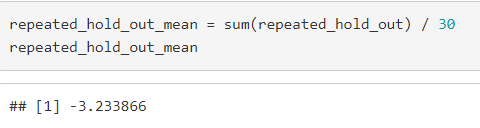
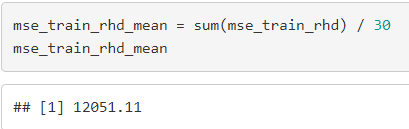


* Kết quả dự đoán:

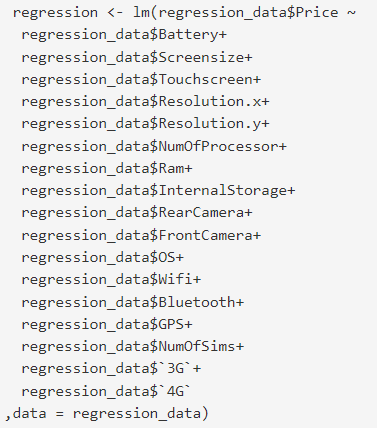


* Trung bình khoảng tin cậy của error:

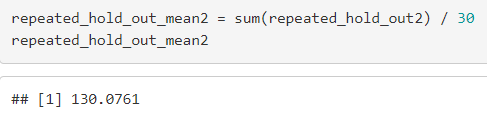
****

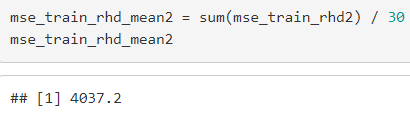
* ****Trung bình squared error train data:
* Trung bình squared error train data:

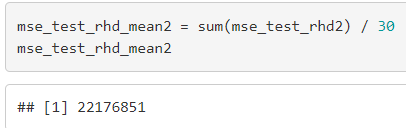
**Mô hình tuyến tính cho mỗi hãng:**

****

* Trung bình khoảng tin cậy error:



* Trung bình squared error train data:
* Trung bình squared error test data:



6.2.2) Mô hình k-NN:

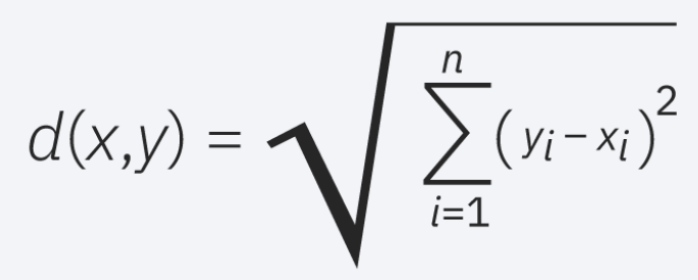
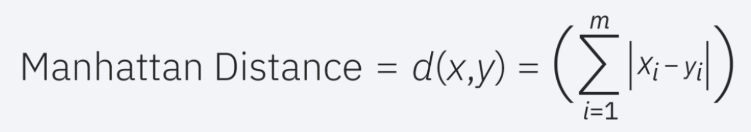
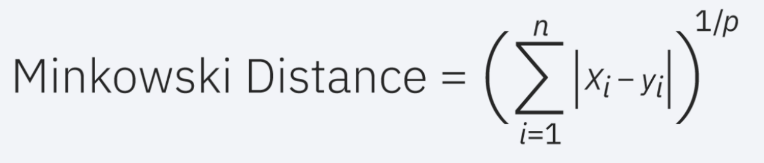
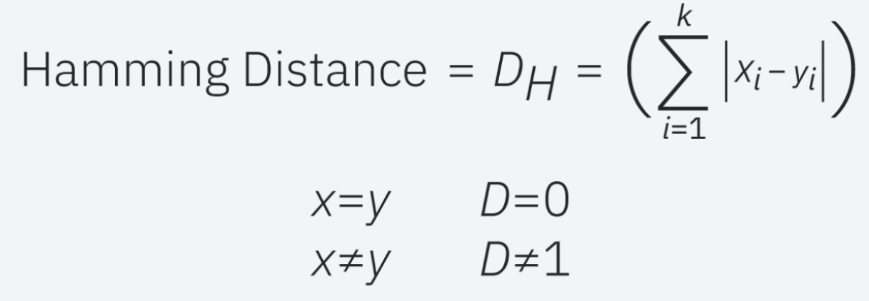
Định nghĩa:

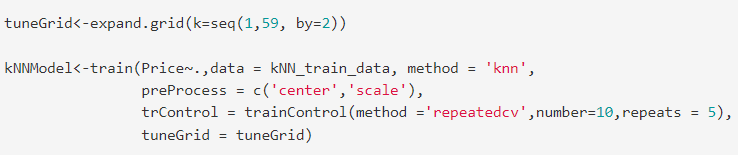
K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression. KNN còn được gọi là một thuật toán Instance-based hay Memory-based learning.

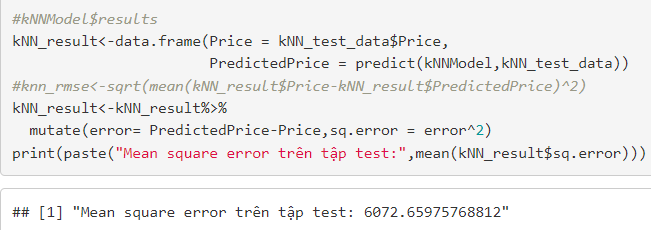
Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới (hay kết quả của câu hỏi trong bài thi) được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label.

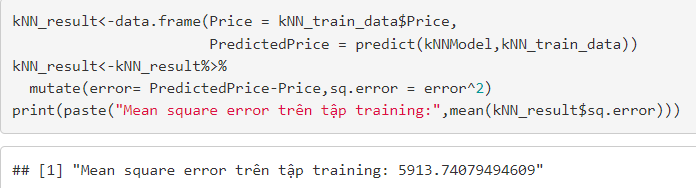
Trong bài toán Regresssion, đầu ra của một điểm dữ liệu sẽ bằng chính đầu ra của điểm dữ liệu đã biết gần nhất (trong trường hợp K=1), hoặc là trung bình có trọng số của đầu ra của những điểm gần nhất, hoặc bằng một mối quan hệ dựa trên khoảng cách tới các điểm gần nhất đó.

**Các công thức để ước lượng khoảng cách:**

1. Euclidean distance:
2. Manhattan distance:
3. Minkowski distance:
4. Hamming distance:

**Thực nghiệm mô hình:**

**Trung bình squared error test data:**

**Trung bình squared error test data:**

# PHẦN 7 – KẾT LUẬN:

Sau khi thực hiện xong đồ án, nhóm chúng em một số kết luận như sau:

1. Các điện thoại hiện tại đều là điện thoại cảm ứng, có RAM nằm trong khoảng 1GB tới 4GB.Bộ nhớ trong điện thoại nhiều nhất là 16GB. Và các mẫu điện thoại hầu hết hỗ trợ những chức năng cơ bản là wifi, bluetooth, 3G, 4G,…
2. Hầu hết các biến là có mối tương quan với giá, ngoại trừ một số biến do giá trị của nó quá ít nên không thể kiểm định (ví dụ không thể kiểm định liệu 3G có ảnh hưởng tới giá không vì hầu hết các điện thoại đều có 3G và ít có điện thoại nào không có 3G).
3. Các biến khi kiểm định riêng thì có thể không ảnh hưởng tới giá nhiều, nhưng khi đưa vào mô hình tuyến tính thì lại là yếu tố quan trọng trong model.