ĐẠI HỌC HUẾ



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

-----🙞🙜🕮🙞🙜-----

**BÁO CÁO ĐỒ ÁN**

**Học kỳ I năm học 2023-2024**

**Học phần:**

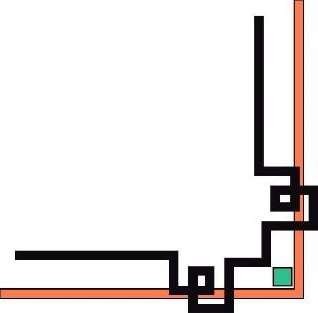
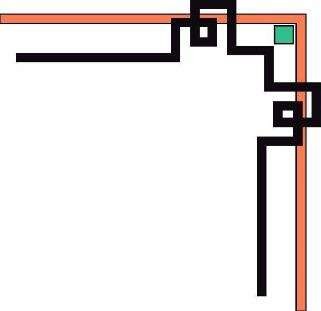
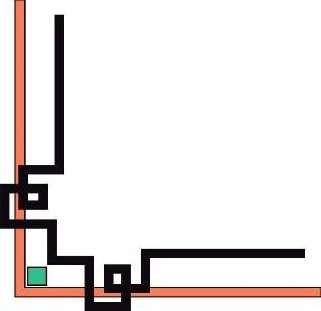
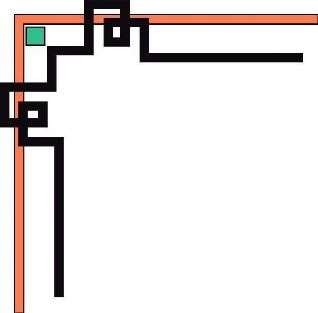
**Học máy 1**

**Số phách**

(*do hội đồng chấm thi ghi*)

Thừa Thiên Huế, tháng … năm 2023

**ĐẠI HỌC HUẾ**



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

**🙢**



**(Bìa phụ 2)**

**BÁO CÁO ĐỒ ÁN**

**Học kỳ I năm học 2023 - 2024**

**Học phần:**

**Học máy 1**

**Giảng viên hướng dẫn: Hoàng Hữu Trung**

**Lớp: Khoa học dữ liệu & Trí tuệ nhân tạo Khóa 3**

**Sinh viên thực hiện: Nguyễn Thanh Tuấn - 22E1020033**

**Số phách**

*(Do hội đồng chấm thi ghi)*

**Thừa Thiên Huế, tháng….năm 2023**

# LỜI CÁM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành đến Khoa Kỹ Thuật Và Công Nghệ đã đưa môn học Học Máy 1 chương trình giảng dạy. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – Thầy Hoàng Hữu Trung đã dạy dỗ, truyền đạt những kiến thức quý giá cho em trong suốt quá trình học tập vừa qua. Trong thời gian tham gia lớp học Học Máy 1 của Thầy, em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý giá, là hành trang để em có thể vững bước sau này.

Bộ môn Học Máy 1 là một môn học thú vị, vô cùng bổ ích. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu của sinh viên. Tuy nhiên do vốn kiến còn nhiều hạn chế và khả năng tiếp thu còn nhiều bỡ ngỡ. Mặc dù em đã cố gắng hết sức nhưng chắc chắn bài đồ án khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong Thầy xem xét và góp ý để bài đồ án của em được hoàn thiện hơn.

Em xin chân thành cảm ơn.

# DANH MỤC HÌNH ẢNH

[Hình 1: Dữ liệu (Supervised Learning) 7](#_Toc155433171)

[Hình 2: Kết quả tiền xử lí dữ liệu (SuperVised Learning) 27](#_Toc155433172)

[Hình 3: Kết quả dự đoán bằng Hồi Quy Logistic (LogisticRegression) 33](#_Toc155433173)

[Hình 4: Kết quả dự đoán bằng thuật toán Support Vector Machine (SVM) 34](#_Toc155433174)

[Hình 5: kết quả thuật toán K-Nearest-NeighBors (KNN) 35](#_Toc155433175)

[Hình 6: Kết quả dự đoán bằng thuật toán LGBMClassifer (LGBM) 36](#_Toc155433176)

[Hình 7: Kết quả so sánh của các thuật toán học có giám sát 39](#_Toc155433177)

[Hình 8: Kết quả dự đoán bằng thuật toán hồi quy tuyến tính (LinearRegression) 40](#_Toc155433178)

[Hình 9: Dữ liệu (Unsupervised Learning) 42](#_Toc155433179)

[Hình 10: Kết quả tiền xử lí dữ liệu (Unsupervised Learning) 52](#_Toc155433180)

[Hình 11: Kết quả phân cụm bằng thuật toán K-means 58](#_Toc155433181)

[Hình 12: Kết quả phân cụm bằng thuật toán DBSCAN 62](#_Toc155433182)

[Hình 13: Kết quả phân cụm bằng thuật toán MeanShift 64](#_Toc155433183)

[Hình 14: Kết quả so sánh các thuật toán học không có giám sát 65](#_Toc155433184)

# DANH MỤC BẢNG BIỂU

# MỤC LỤC

[LỜI CÁM ƠN i](#_Toc155433063)

[DANH MỤC HÌNH ẢNH ii](#_Toc155433064)

[DANH MỤC BẢNG BIỂU iii](#_Toc155433065)

[MỤC LỤC iv](#_Toc155433066)

[PHẦN 1: HỌC CÓ GIÁM SÁT (SUPERVISED LEARNING) 7](#_Toc155433067)

[1 Giới thiệu dữ liệu. 7](#_Toc155433068)

[2 Tiền xử lí dữ liệu. 8](#_Toc155433069)

[2.1 Gọi thư viện và đọc dữ liệu. 8](#_Toc155433070)

[2.2 Trực quan hóa dữ liệu. 9](#_Toc155433071)

[2.3 Xử lí dữ liệu. 21](#_Toc155433072)

[2.3.1 Kiểm tra thông tin dữ liệu. 21](#_Toc155433073)

[2.3.2 Trực quan các dữ liệu trống trong dữ liệu. 22](#_Toc155433074)

[2.3.3 Loại bỏ những cột chứa nhiều giá trị trống và những cột không cần thiết trong dữ liệu. 23](#_Toc155433075)

[2.3.4 Xử lí những giá trị trống. 24](#_Toc155433076)

[2.3.5 Kiểm tra lại dữ liệu trống sau khi xử lí dữ liệu. 24](#_Toc155433077)

[2.4 Chuẩn hóa dữ liệu. 25](#_Toc155433078)

[2.5 Mã hóa dữ liệu. 25](#_Toc155433079)

[2.6 Xử lí giá trị nhiễu. 26](#_Toc155433080)

[2.6.1 Xử lí giá trị nhiễu. 26](#_Toc155433081)

[2.6.2 Trực quan dữ liệu trước và sau khi xử lí nhiễu. 27](#_Toc155433082)

[3 Ứng dụng thuật toán học có giám sát (Supervised Learning). 28](#_Toc155433083)

[3.1 Chia mô hình. 28](#_Toc155433084)

[3.2 Phân tích tương quan dữ liệu. 28](#_Toc155433085)

[3.2.1 Vẽ ma trận tương quan. 28](#_Toc155433086)

[3.2.2 Phân tích tương quan. 30](#_Toc155433087)

[3.2.3 Loại bỏ các biến tương quan cao. 30](#_Toc155433088)

[3.3 Ứng dụng LazyPredict để tìm thuật toán phù hợp cho mô hình. 31](#_Toc155433089)

[3.4 Ứng dụng các thuật toán học có giám sát (Supervised Learning). 32](#_Toc155433090)

[3.4.1 Dự đoán bằng hồi Quy Logistic (LogisticRegression). 33](#_Toc155433091)

[3.4.2 Dự đoán bằng Máy Vector hỗ trợ (Support Vector Machine). 33](#_Toc155433092)

[3.4.3 Dự đoán bằng K-Nearest-NeighBors. 34](#_Toc155433093)

[3.4.4 Dự đoán bằng LightGBM (LGBMClassifier). 35](#_Toc155433094)

[3.4.5 Trực quan dữ liệu thực tế với dữ liệu dự đoán của 4 thuật toán sau 37](#_Toc155433095)

[3.4.6 So sánh độ chính xác của các thuật toán. 38](#_Toc155433096)

[3.4.7 Dự đoán bằng hồi quy tuyến tính. 39](#_Toc155433097)

[PHẦN 2: HỌC KHÔNG CÓ GIÁM SÁT (UNSUPERVISED LEARNING) 41](#_Toc155433098)

[1 Giới thiệu dữ liệu. 41](#_Toc155433099)

[2 Tiền xử lí dữ liệu. 42](#_Toc155433100)

[2.1 Đọc dữ liệu và gọi thư viện. 42](#_Toc155433101)

[2.2 Trực quan hóa dữ liệu. 43](#_Toc155433102)

[2.3 Xử lí dữ liệu. 49](#_Toc155433103)

[2.4 Chuẩn hóa dữ liệu. 51](#_Toc155433104)

[2.5 Mã hóa dữ liệu. 51](#_Toc155433105)

[3 Phân tích tương quan dữ liệu. 52](#_Toc155433106)

[4 Ứng dụng các thuật toán học không có giám sát (Unsupervised Learning). 54](#_Toc155433107)

[4.1 Trực quan dữ liệu trước khi phân cụm. 55](#_Toc155433108)

[4.2 Phân cụm bằng thuật toán K-means. 55](#_Toc155433109)

[4.2.1 Sử dụng phương pháp khuỷu tay (Elbow Method). 56](#_Toc155433110)

[4.2.2 Thực hiện phân cụm. 57](#_Toc155433111)

[4.2.3 Trực quan hóa sau khi phân cụm. 57](#_Toc155433112)

[4.2.4 Phân phối cụm k-means. 58](#_Toc155433113)

[4.3 Phân cụm bằng thuật toán DBSCAN (Density-Based Spatial Clustering of Applications with Noise). 60](#_Toc155433114)

[4.3.1 Thực hiện phân cụm và trực quan hóa DBSCAN 61](#_Toc155433115)

[4.3.2 Tính tỉ lệ nhiễu và phân phối cụm. 62](#_Toc155433116)

[4.4 Phân cụm bằng thuật toán MeanShift 63](#_Toc155433117)

[4.4.1 Thực hiện phân cụm MeanShift. 63](#_Toc155433118)

[4.4.2 Trực quan hóa sau khi phân cụm MeanShift 64](#_Toc155433119)

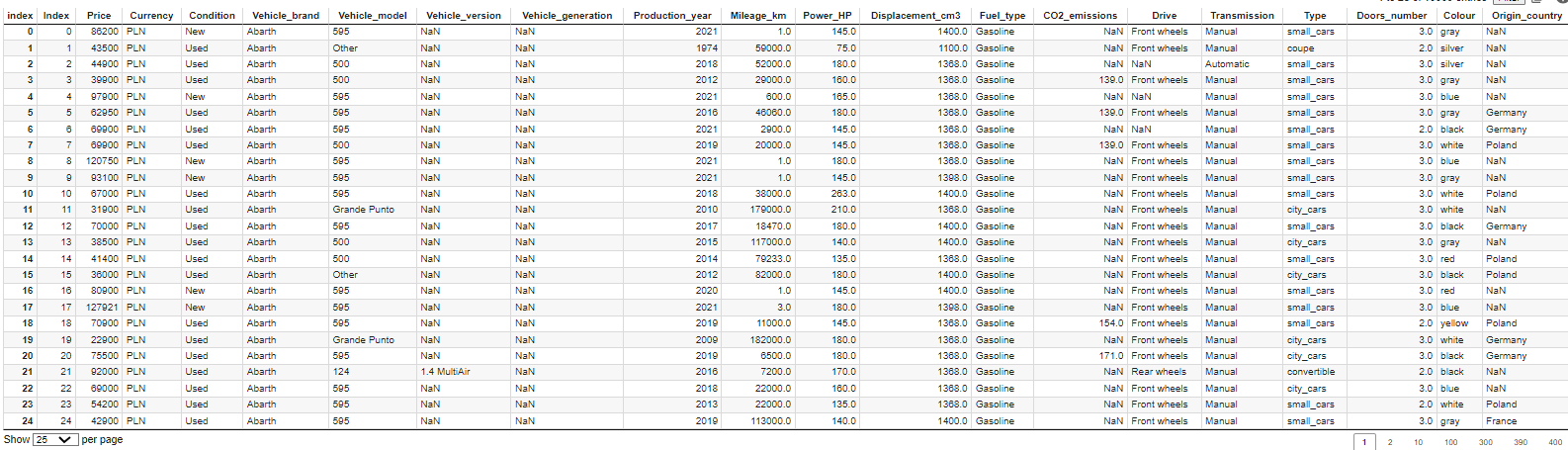
[4.4.3 Phân phối cụm trên dữ liệu với MeanShift. 65](#_Toc155433120)

[4.5 Đánh giá hiệu suất mô hình và so sánh các thuật toán phân cụm trên. 66](#_Toc155433121)

# PHẦN 1: HỌC CÓ GIÁM SÁT (SUPERVISED LEARNING)

# 1 Giới thiệu dữ liệu.

Mã nguồn dữ liệu: <https://www.kaggle.com/datasets/bartoszpieniak/poland-cars-for-sale-dataset>



Hình 1: Dữ liệu (Supervised Learning)

Tập dữ liệu “Poland cars for sale dataset” chứa nhiều thông tin quan trọng về đặc điểm của xe ô tô, được thể hiện qua 25 cột dữ liệu sau:

* ID: Mã định danh duy nhất của đề xuất.
* Price: Giá trị của mục đề xuất.
* Currency: Loại tiền tệ của giá
* Condition: Tình trạng của xe ô tô (mới hoặc đã sử dụng).
* Vehicle\_brand: Thương hiệu của xe ô tô trong đề xuất.
* Vehicle\_model: Mô hình của xe ô tô trong đề xuất.
* Vehicle\_generation: Thế hệ của xe ô tô trong đề xuất.
* Vehicle\_version: Phiên bản của xe ô tô trong đề xuất.
* Production\_year: Năm sản xuất của xe ô tô.
* Mileage\_km: Tổng quãng đường mà chiếc xe ô tô đã đi được tính bằng đơn vị kilômét.
* Power\_HP: Công suất động cơ của xe ô tô tính bằng mã lực.
* Displacement\_cm3: Dung tích động cơ của xe ô tô tính bằng đơn vị centimet khối.
* Fuel\_type: Loại nhiên liệu của xe ô tô.
* CO2\_emissions: Lượng khí CO2 phát ra từ xe ô tô tính bằng đơn vị gram/kilômét.
* Drive: Loại hệ thống dẫn động của xe ô tô.
* Transmission: Loại hộp số của xe ô tô.
* Type: Kiểu dáng của xe ô tô.
* Doors\_number: Số lượng cửa của xe ô tô.
* Colour: Màu sắc của bề mặt xe ô tô.
* Origin\_country: Quốc gia xuất xứ của xe ô tô.
* First\_owner: Cho biết liệu chủ sở hữu có phải là chủ sở hữu đầu tiên không.
* First\_registration\_date: Ngày đăng ký đầu tiên của xe ô tô.
* Offer\_publication\_date: Ngày công bố đề xuất bán xe ô tô.
* Offer\_location: Địa chỉ được cung cấp bởi người đề xuất.
* Features: Danh sách các tính năng của xe ô tô được liệt kê (ví dụ: ABS, túi khí, cảm biến đỗ xe, v.v.).

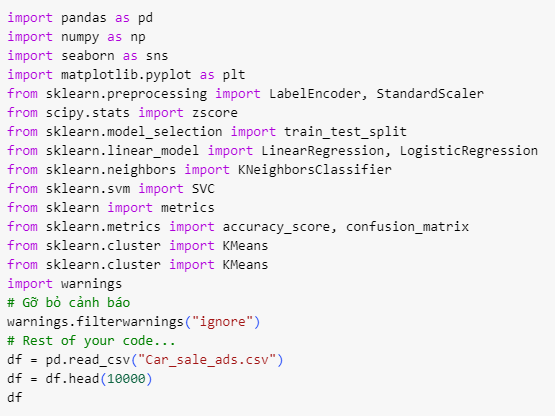
Dữ liệu này có thể hỗ trợ nhiều ứng dụng và phân tích trong lĩnh vực mua bán xe ô tô:

* Phân tích thị trường xe ô tô
* Đánh giá giá trị xe ô tô
* Dự đoán giá xe ô tô
* Phân loại và lọc xe ô tô
* Phân tích tình trạng xe ô tô đã sử dụng
* Phân tích ảnh hưởng của các yếu tố môi trường
* Dự đoán nhu cầu thị trường

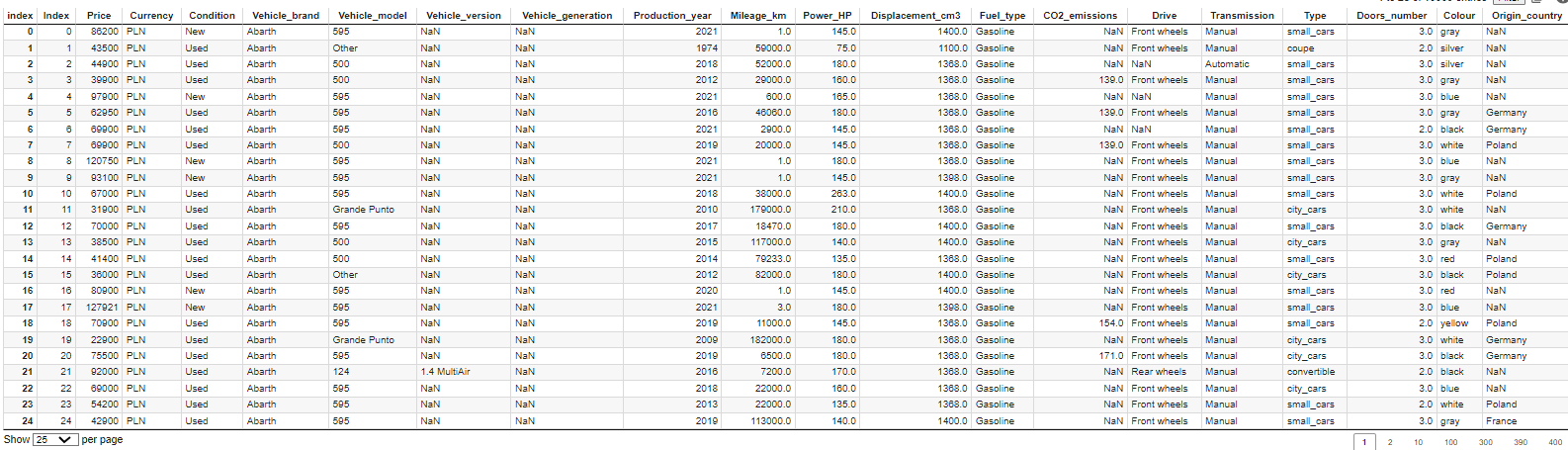
# 2 Tiền xử lí dữ liệu.

## 2.1 Gọi thư viện và đọc dữ liệu.

Vì dữ liệu quá lớn nên trong bài đồ án này chúng ta chỉ phân tích 10000 dòng đầu tiên của tập dữ liệu.

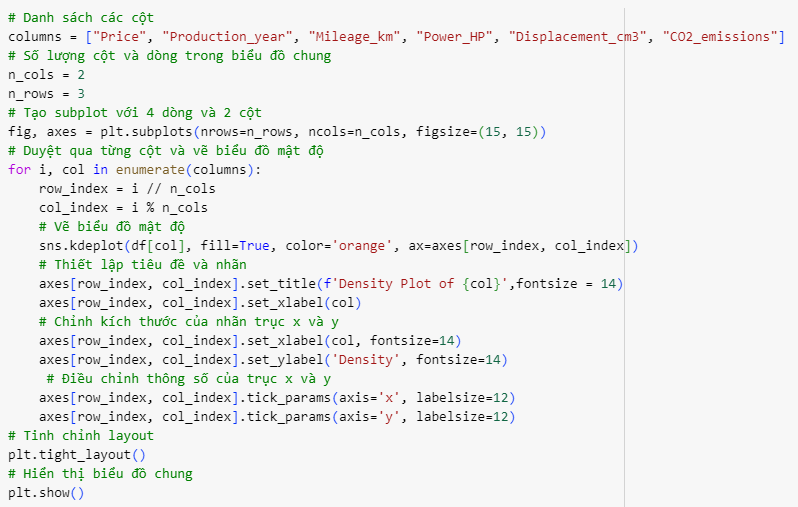


Kết quả:

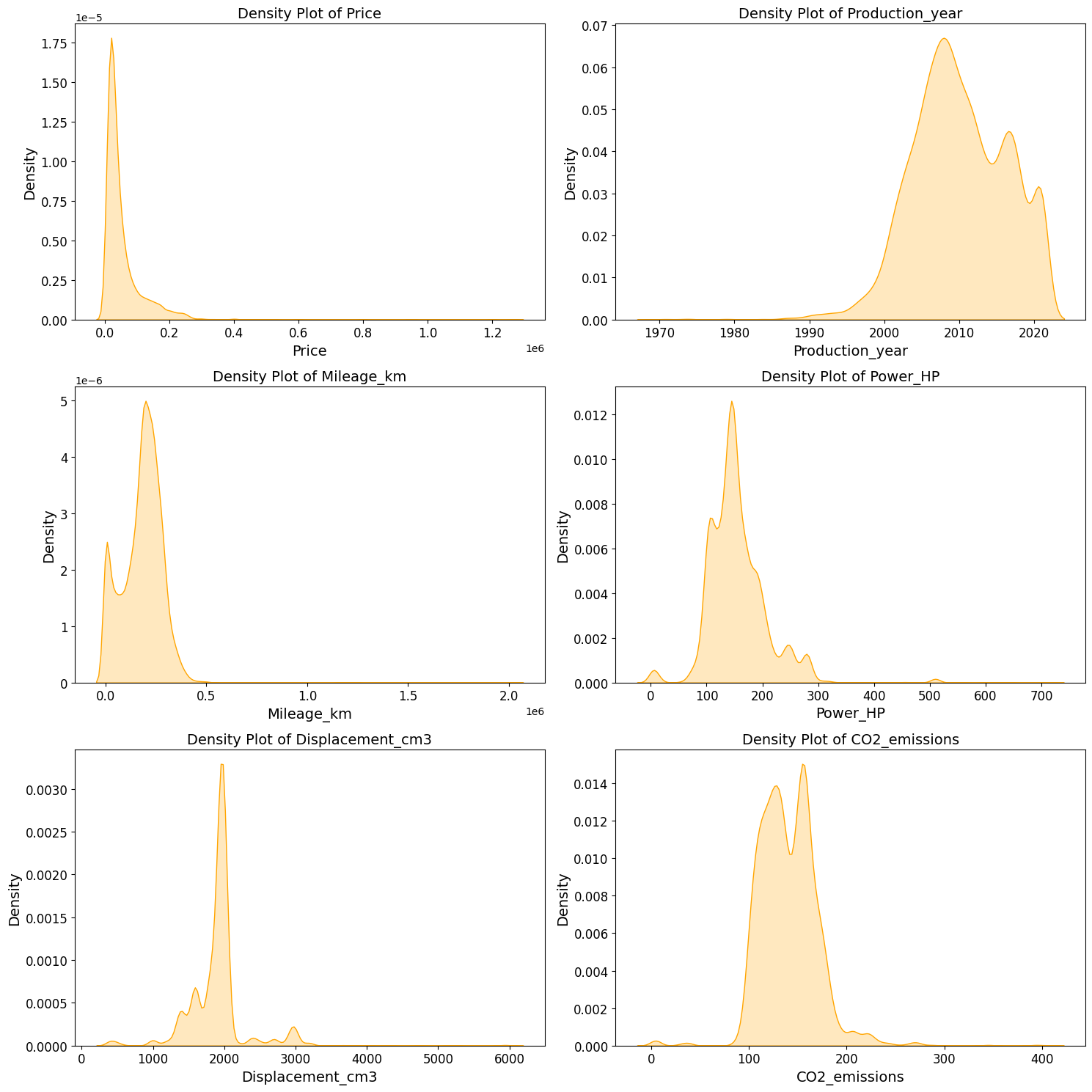


## 2.2 Trực quan hóa dữ liệu.

Mã nguồn:



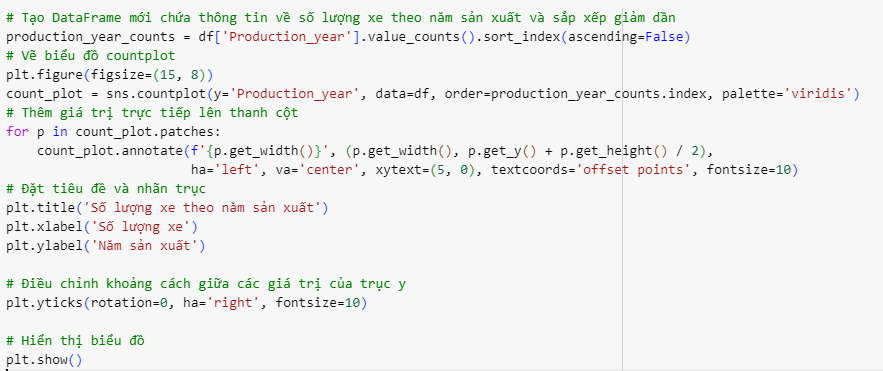
Kết quả:



Nhận xét:

* Biểu đồ 1 (Density plot of Price): thể hiện mật độ giá của một sản phẩm, được đo bằng số lượng sản phẩm có giá nằm trong một khoảng nhất định. Từ biểu đồ, ta có thể thấy:
* Giá của sản phẩm phân bố khá rộng, từ khoảng 0 đến hơn 1.
* Giá của sản phẩm tập trung nhiều nhất trong khoảng 0 đến 0,2.
* Số lượng sản phẩm có giá cao hơn 0,2 là khá ít.
* Biểu đồ 2 (Density Plot of Production\_year): Biểu đồ này thể hiện mật độ năm sản xuất của sản phẩm, được đo bằng số lượng sản phẩm được sản xuất trong một năm nhất định. Từ biểu đồ, ta có thể thấy:
* Sản phẩm được sản xuất nhiều nhất trong giai đoạn từ năm 2000 đến năm 2020.
* Sản phẩm được sản xuất ít nhất trong giai đoạn từ năm 1970 đến năm 1990
* Biểu đồ 3 (Density Plot of Mileage\_km): Biểu đồ này thể hiện mật độ số km đã đi của một chiếc xe, được đo bằng số lượng xe có số km đã đi nằm trong một khoảng nhất định. Từ biểu đồ, ta có thể thấy:
* Số km đã đi của xe phân bố khá rộng, từ khoảng 0 đến hơn 200.000 km.
* Số km đã đi của xe tập trung nhiều nhất trong khoảng 0 đến 50.000 km.
* Số lượng xe có số km đã đi cao hơn 100.000 km là khá ít.
* Biểu đồ 4 (Density Plot of Power\_HP): Biểu đồ này thể hiện mật độ công suất động cơ của một chiếc xe, được đo bằng số lượng xe có công suất động cơ nằm trong một khoảng nhất định. Từ biểu đồ, ta có thể thấy:
* Công suất động cơ của xe phân bố khá rộng, từ khoảng 100 đến 700 mã lực.
* Công suất động cơ của xe tập trung nhiều nhất trong khoảng 200 đến 400 mã lực.
* Số lượng xe có công suất động cơ cao hơn 500 mã lực là khá ít.
* Biểu đồ 5 (Density Plot of Displacement\_cm3): Biểu đồ này thể hiện mật độ dung tích xy-lanh của một chiếc xe, được đo bằng số lượng xe có dung tích xy-lanh nằm trong một khoảng nhất định. Từ biểu đồ, ta có thể thấy:
* Dung tích xy-lanh của xe phân bố khá rộng, từ khoảng 1.000 đến 6.000 cm3.
* Dung tích xy-lanh của xe tập trung nhiều nhất trong khoảng 1.500 đến 3.000 cm3.
* Số lượng xe có dung tích xy-lanh cao hơn 4.000 cm3 là khá ít.
* Biểu đồ 6 (Density Plot of CO2\_emissions): Biểu đồ này thể hiện mật độ khí thải CO2 của một chiếc xe, được đo bằng số lượng xe có khí thải CO2 nằm trong một khoảng nhất định. Từ biểu đồ, ta có thể thấy:
* Khí thải CO2 của xe phân bố khá rộng, từ khoảng 100 đến 400 g/km.
* Khí thải CO2 của xe tập trung nhiều nhất trong khoảng 150 đến 300 g/km.
* Số lượng xe có khí thải CO2 cao hơn 350 g/km là khá ít.

Mã nguồn:



Kết quả:

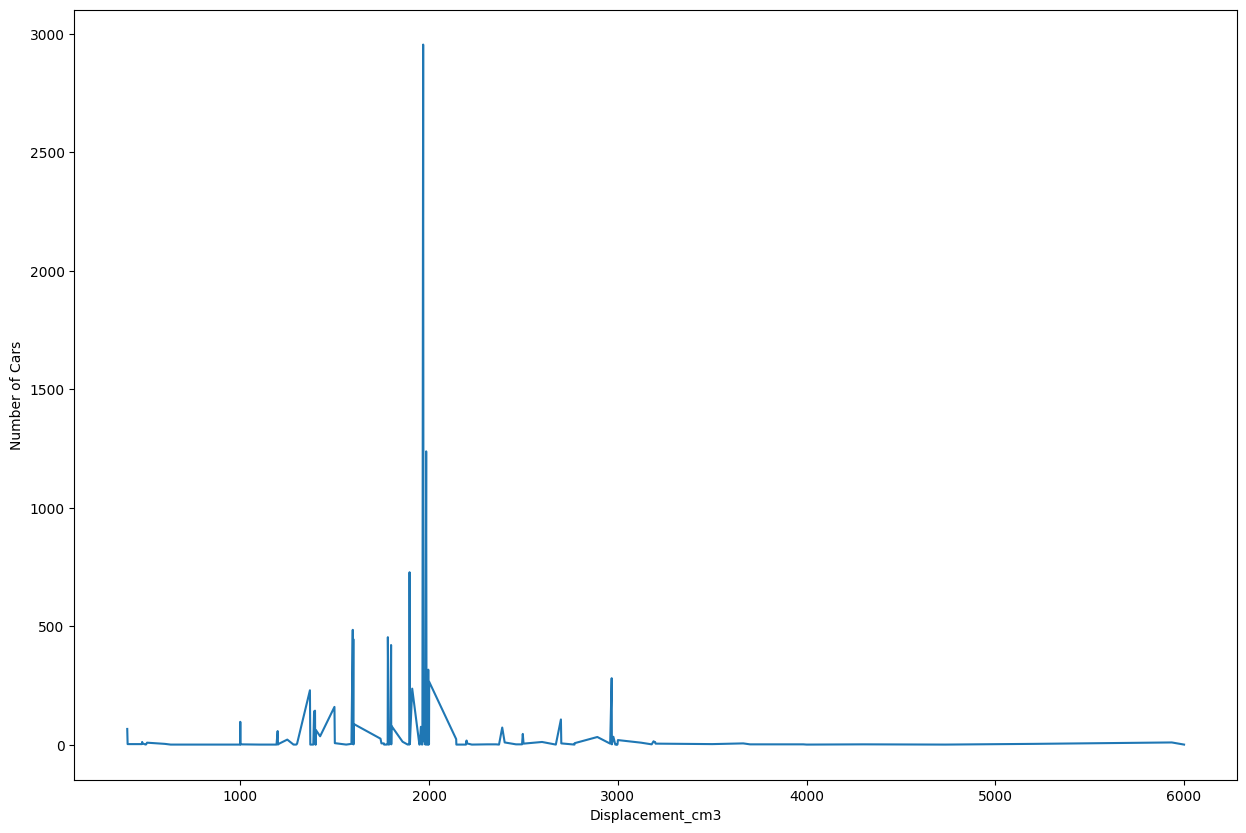


Nhận xét: Biểu đồ này thể hiện số lượng ô tô được sản xuất theo năm sản xuất từ năm 1970 đến năm 2021. Biểu đồ sử dụng thang đo tuyến tính với đơn vị là số xe.

Mã nguồn:



Kết quả:

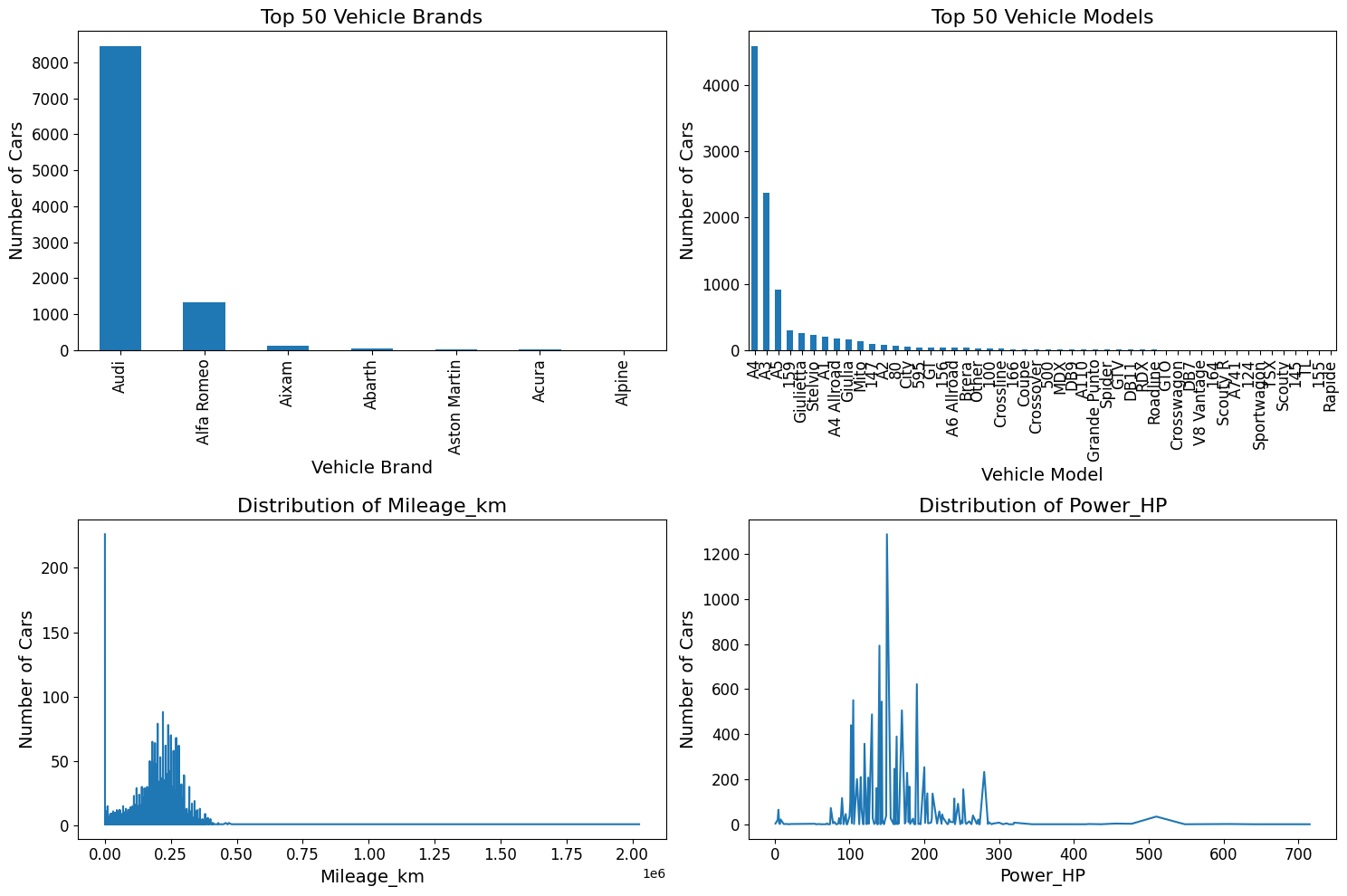


Nhận xét: Biểu đồ này thể hiện phân phối của số lượng xe ô tô dựa trên dung tích xi lanh (Displacement\_cm3).

Mã nguồn:



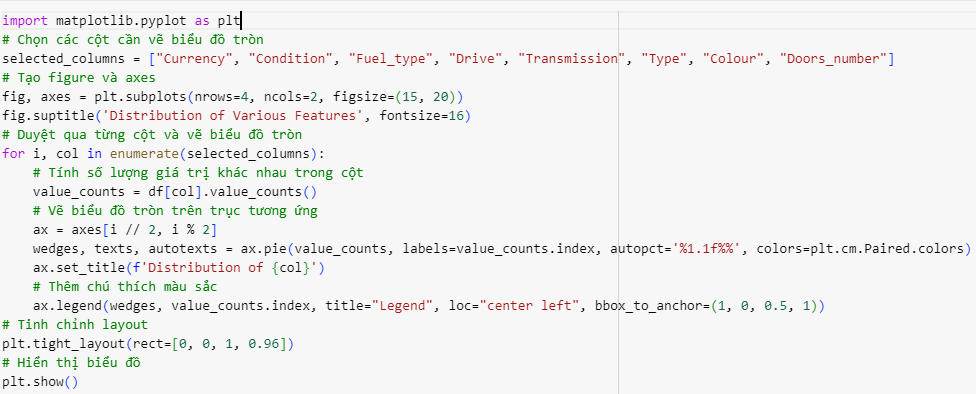
Kết quả:



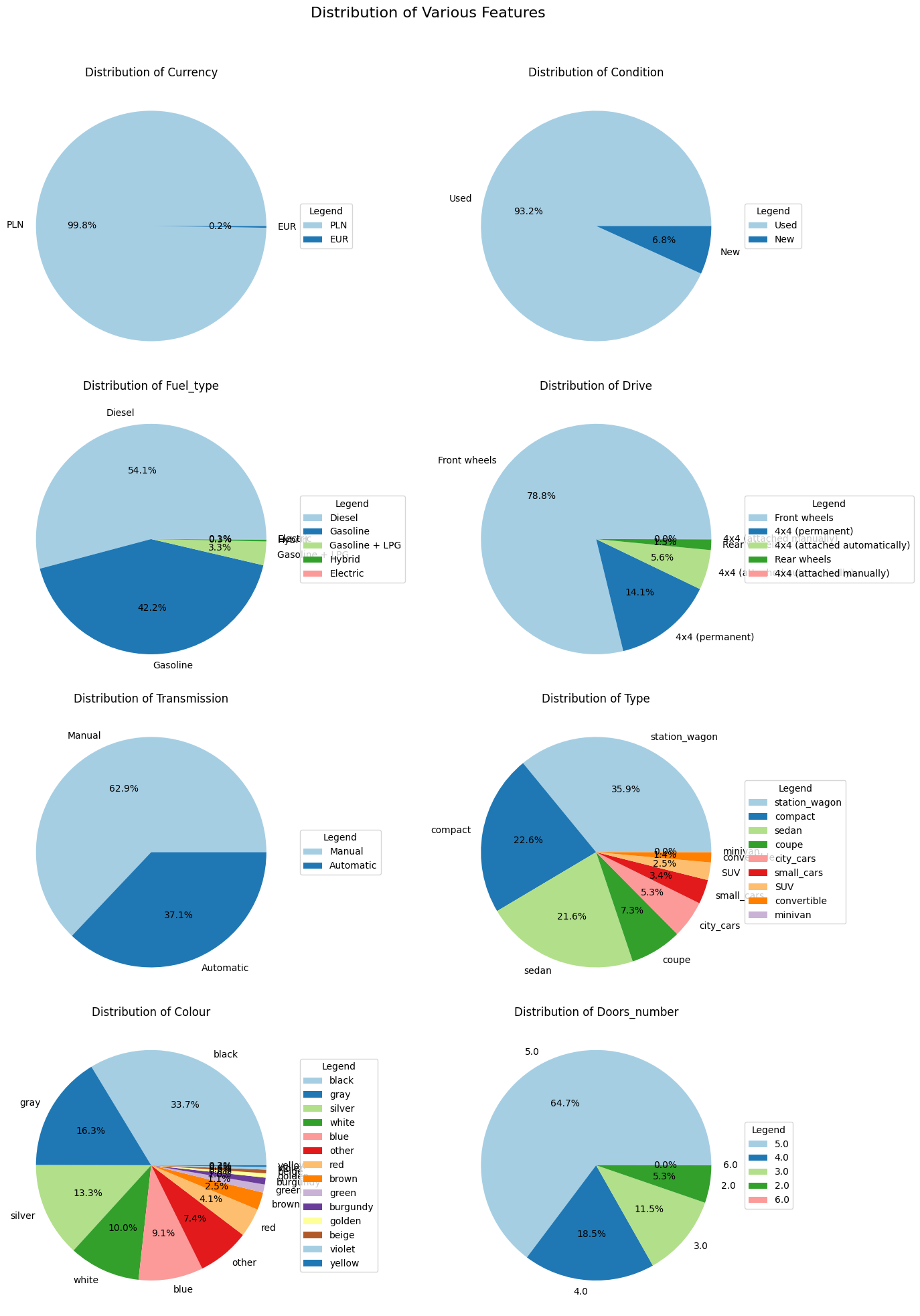
Nhận xét:

* Biểu đồ 1 (Top 50 Vehicle Brands): Biểu đồ thể hiện các thương hiệu xe trong 50 hàng đầu tiên của dữ liệu. Từ biểu đồ, ta thấy nhãn hiệu xe Audi có số lượng vượt trội hơn hẳn so với các thương hiệu xe còn lại.
* Biểu đồ 2 (Top 50 Vehicle Brands): Biểu đồ thể hiện các mô hình xe trong 50 hàng đầu tiên của dữ liệu. Từ biểu đồ, ta thấy mô hình A4 có số lượng vượt trội hơn hẳn so với các mô hình xe còn lại.
* Biểu đồ 3 (Distribution of Mileage\_km): Biểu đồ thể hiện phân phối của số km đã đi. Từ biểu đồ ta thấy số km đã đi của các xe chủ yếu từ 0-50km.
* Biểu đồ 4 (Distribution of Power\_HP): Biểu đồ đường thể hiện phân phối của công suất động cơ.

Mã nguồn:



Kết quả:



Nhận xét: Đây là các biểu đồ tròn thể hiện phân phối của các giá trị trong mỗi cột dữ liệu

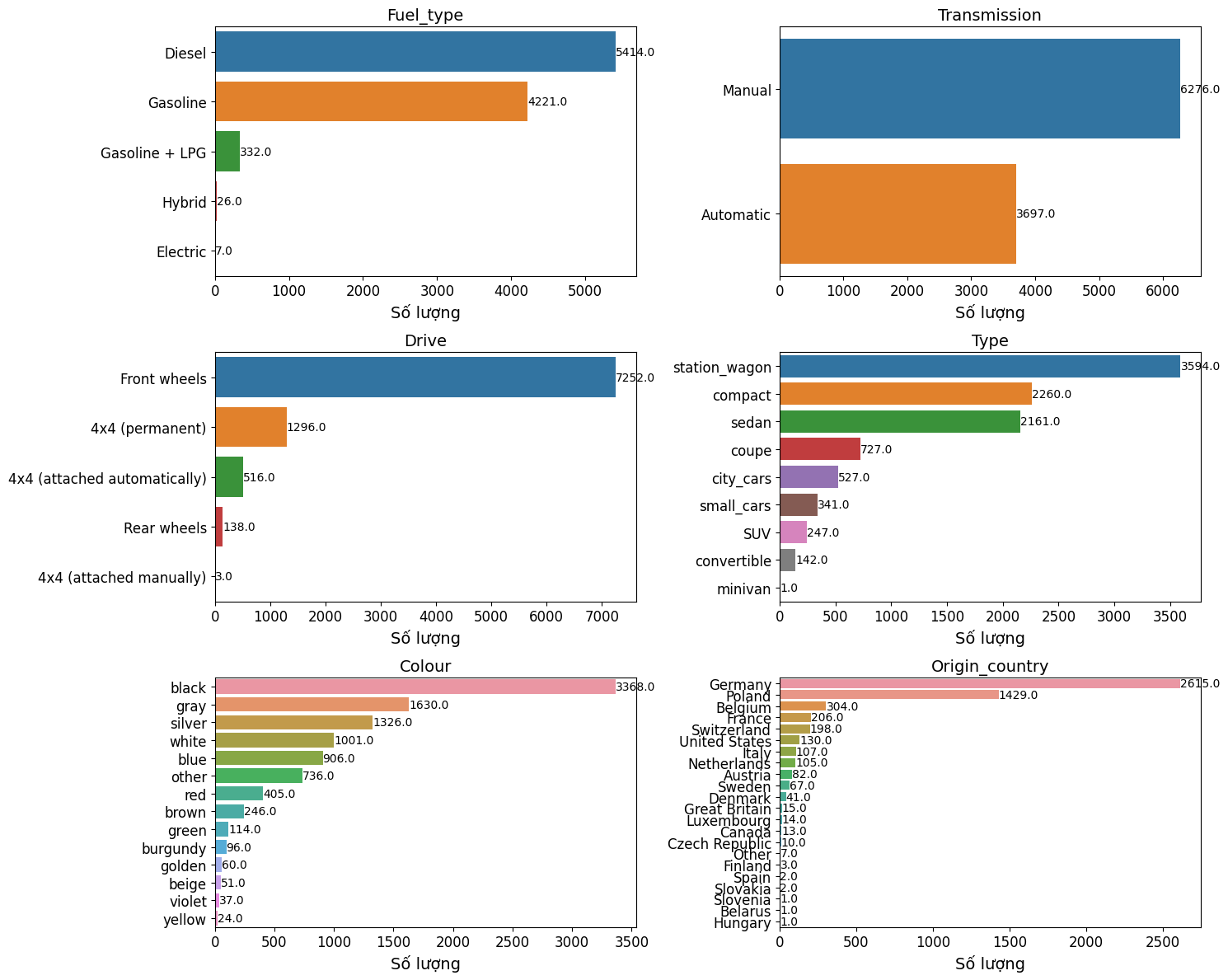
Mã nguồn:





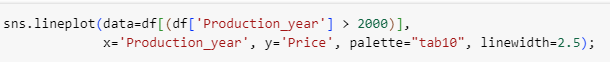
Kết quả:



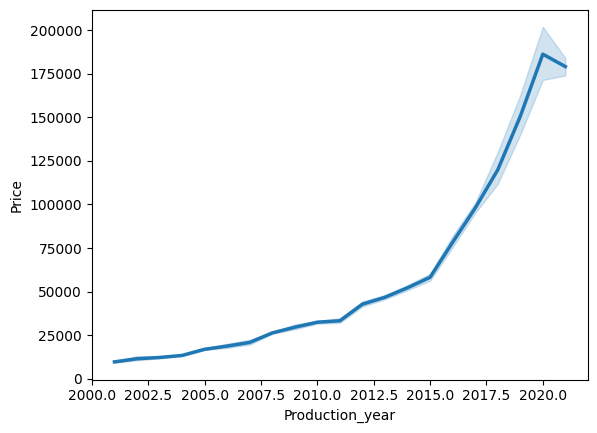


Nhận xét: Các biểu đồ này thể hiện số lượng giá trị của từng cột trong dữ liệu

Mã nguồn:



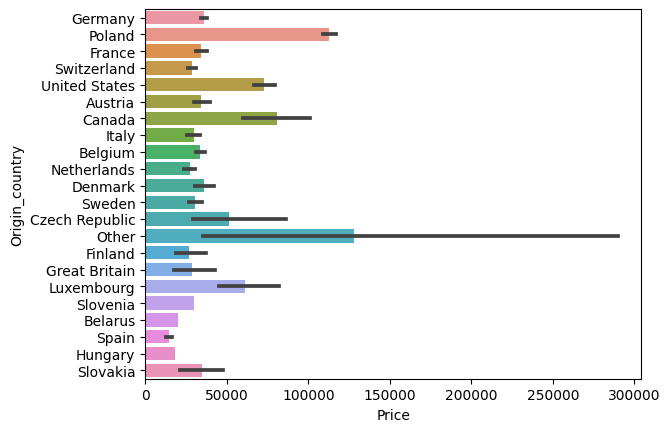
Kết quả:



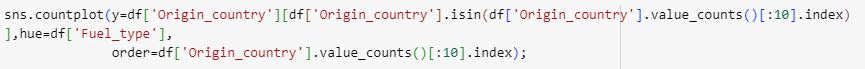
Nhận xét: Đây là biểu đồ đường, nhìn vào biểu đồ cho thấy giá xe tăng từ năm 2000 trở đi

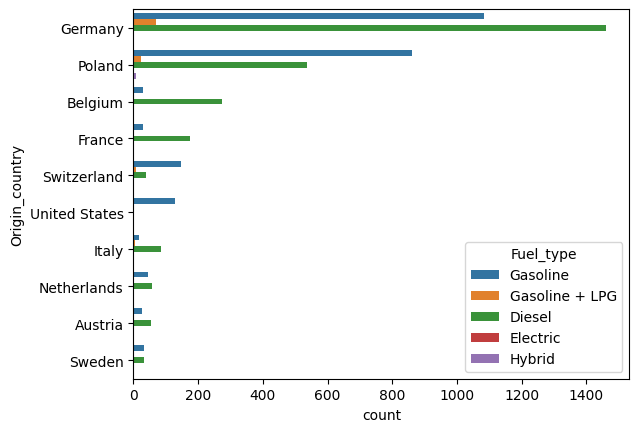
Mã nguồn:



Kết quả:

Nhận xét: Biểu đồ thể hiện giá xe của các quốc gia, nhìn vào biểu đồ ta thấy giá xe của các quốc gia này đa số là trên 25000 USD

Mã nguồn: 

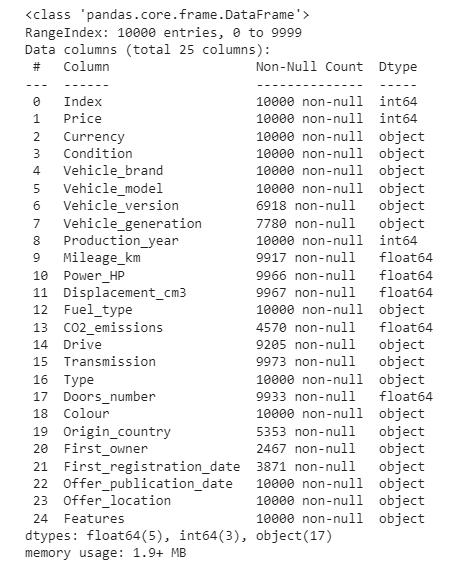
Kết quả: 

Nhận xét: Biểu đồ cho thấy rõ sự phân phối của các loại dầu của 10 quốc gia phổ biến nhất, nhìn vào biểu đồ ta thấy chỉ có 2 loại dầu phổ biến ở các quốc gia này đó là Gasoline và Diesel

## 2.3 Xử lí dữ liệu.

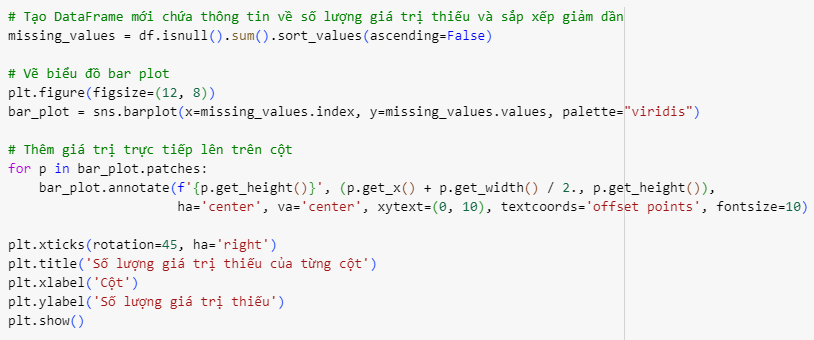
### 2.3.1 Kiểm tra thông tin dữ liệu.

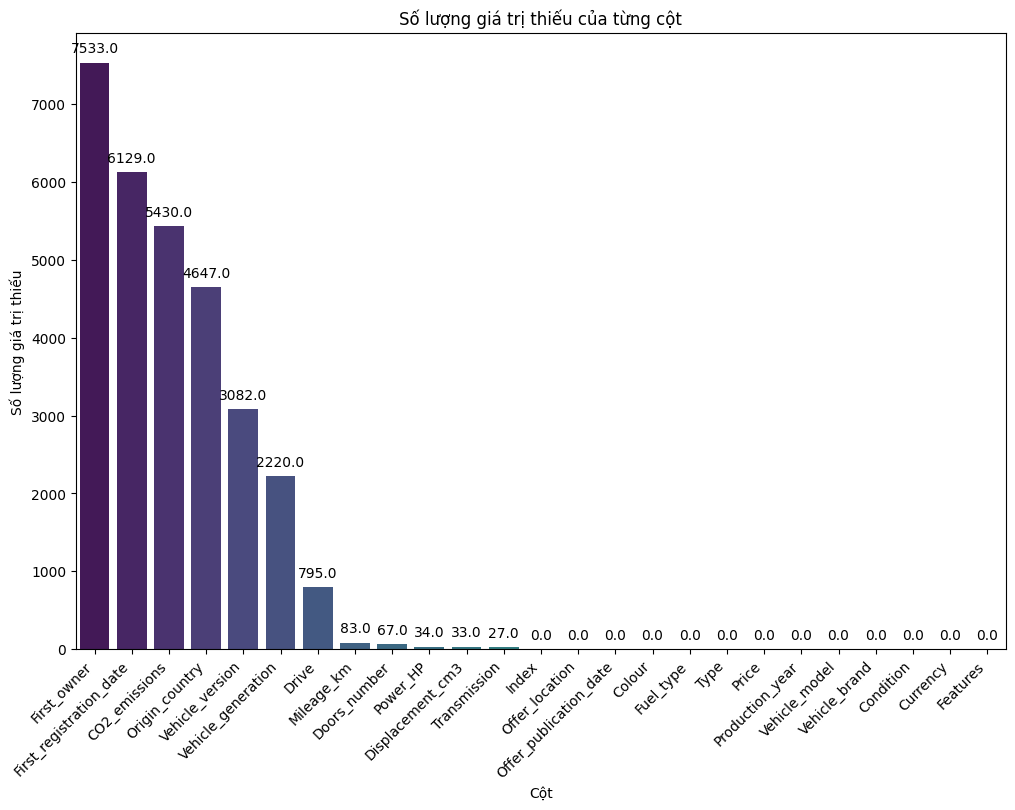




* Nhìn vào dữ liệu ta có thể thấy được các dữ liệu nào là số, dữ liệu nào là chữ và các dữ liệu trống trong dữ liệu

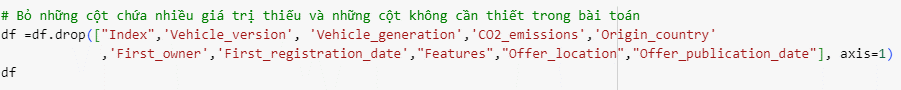
### 2.3.2 Trực quan các dữ liệu trống trong dữ liệu.

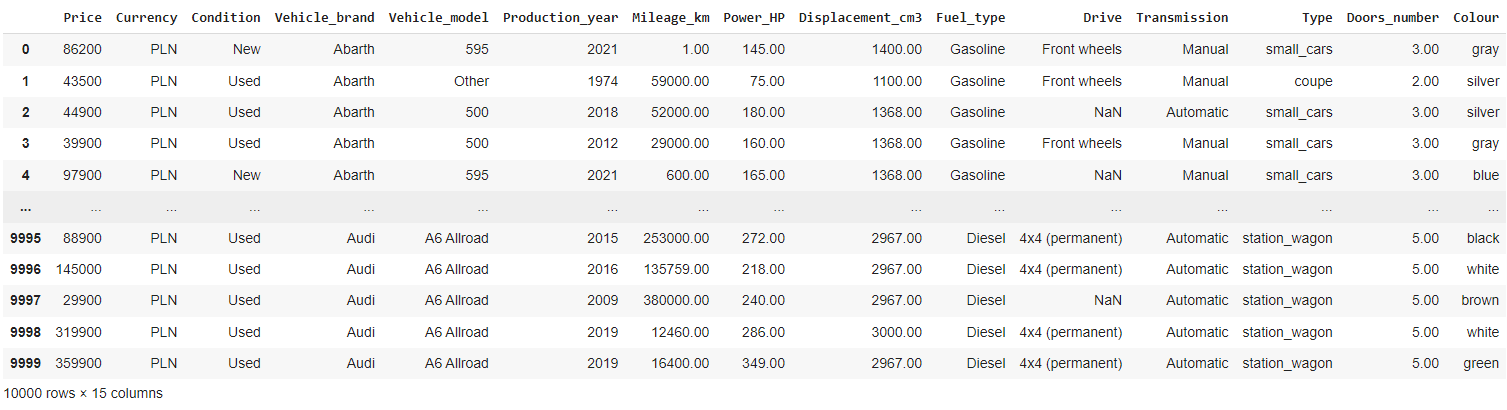




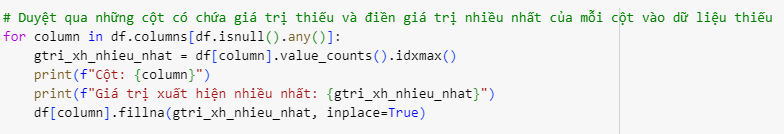
* Nhìn vào biểu đồ ta có thể thấy rõ các dữ liệu trống của từng cột.

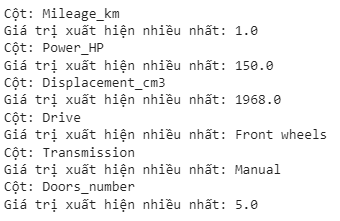
### 2.3.3 Loại bỏ những cột chứa nhiều giá trị trống và những cột không cần thiết trong dữ liệu.





### 2.3.4 Xử lí những giá trị trống.

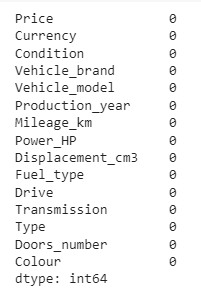




* Các giá trị trống đã được xử lí bằng cách điền các giá trị xuất hiện nhiều nhất trong cột của giá trị đó.

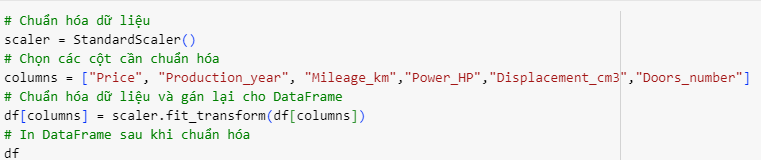
### 2.3.5 Kiểm tra lại dữ liệu trống sau khi xử lí dữ liệu.

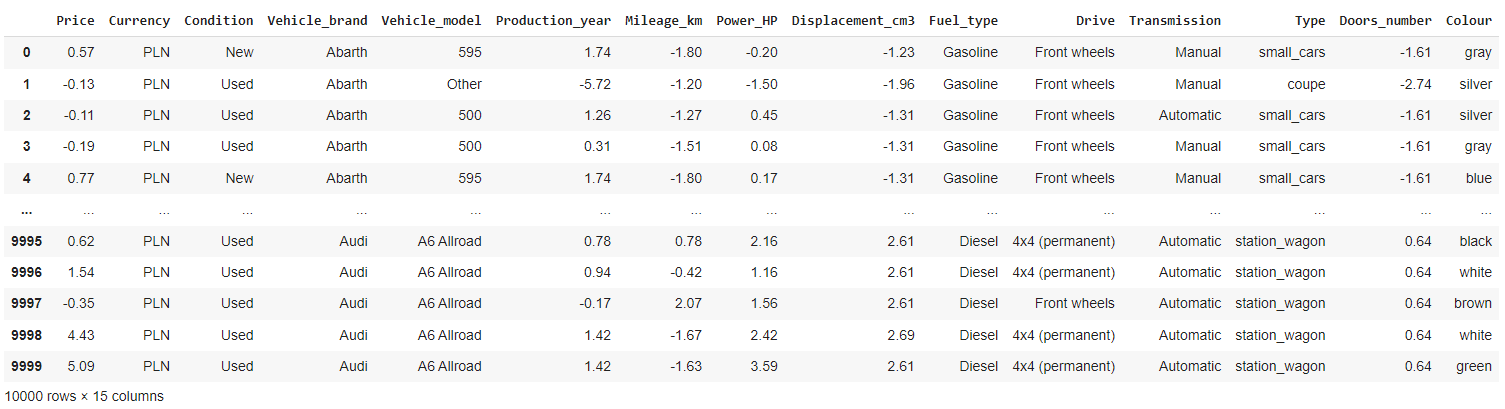




* Các giá trị trống đã được xử lí xong.

## 2.4 Chuẩn hóa dữ liệu.

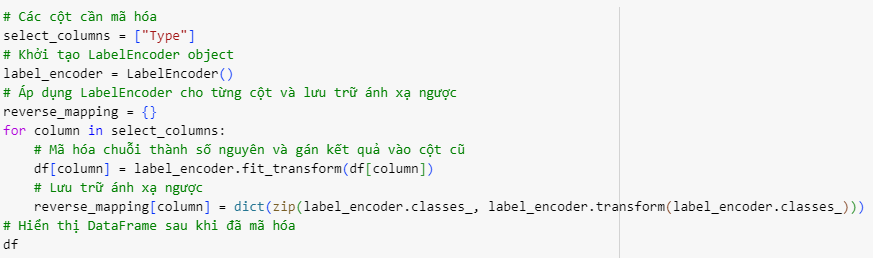


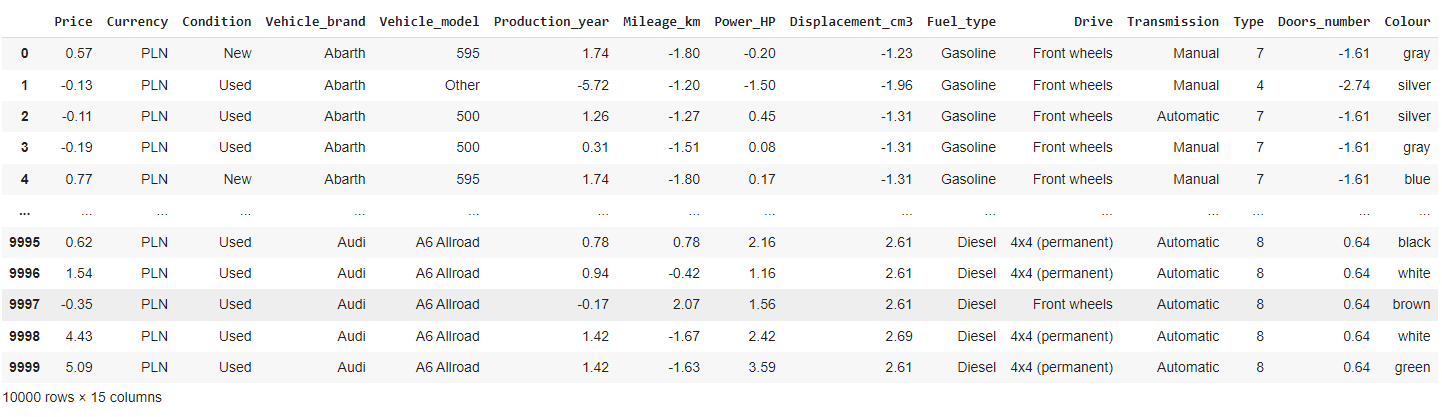


* Dữ liệu đã được chuẩn hóa

## 2.5 Mã hóa dữ liệu.

* Mã hóa cột dữ liệu cần dự đoán trong mô hình (Cột “Type”) và hiển thị các giá trị ảnh xạ:



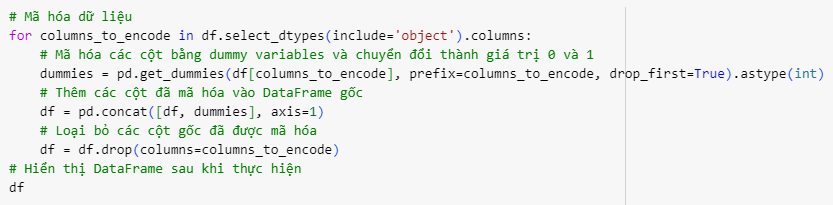


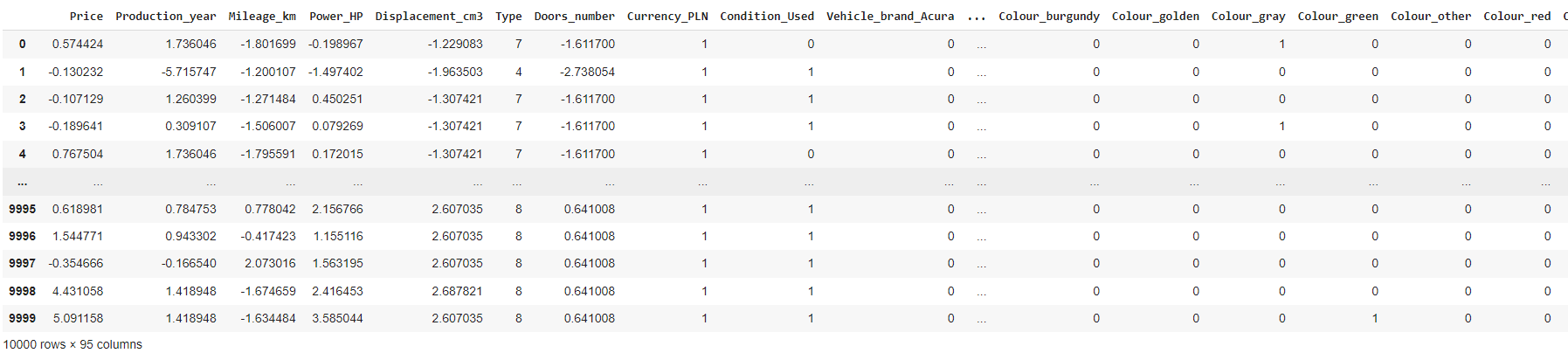


* Cột “Type” đã được mã hóa và các giá trị trong cột “Type” được ánh xạ như sau:



* Mã hóa những cột dữ liệu dùng để dự đoán (Các đặc trưng):

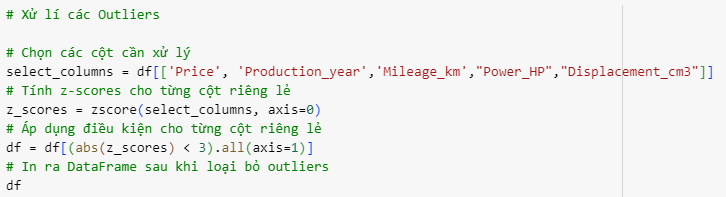


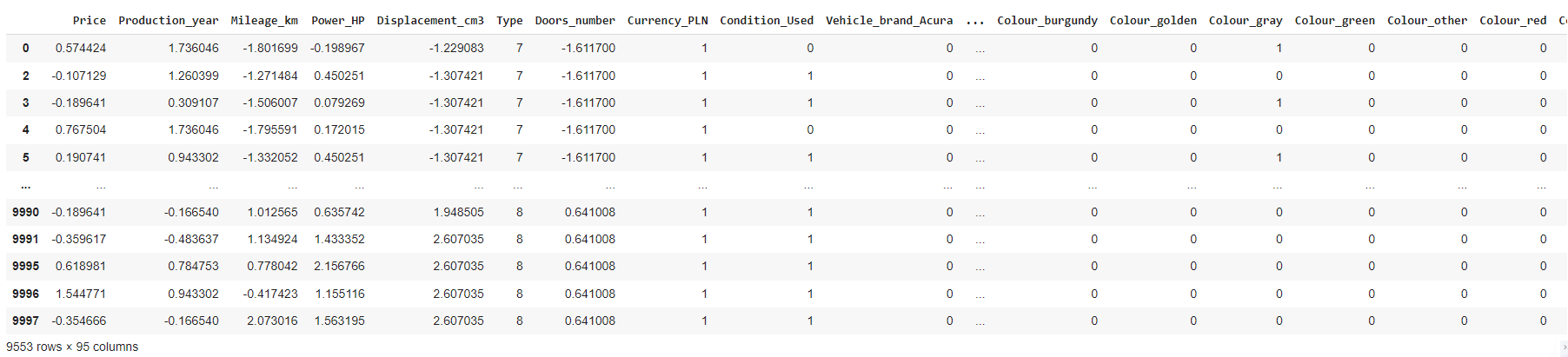


* Dữ liệu đã được mã hóa và không còn những cột dữ liệu có giá trị là chữ.

## 2.6 Xử lí giá trị nhiễu.

### 2.6.1 Xử lí giá trị nhiễu.

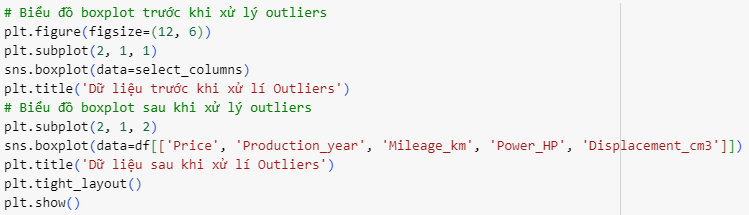


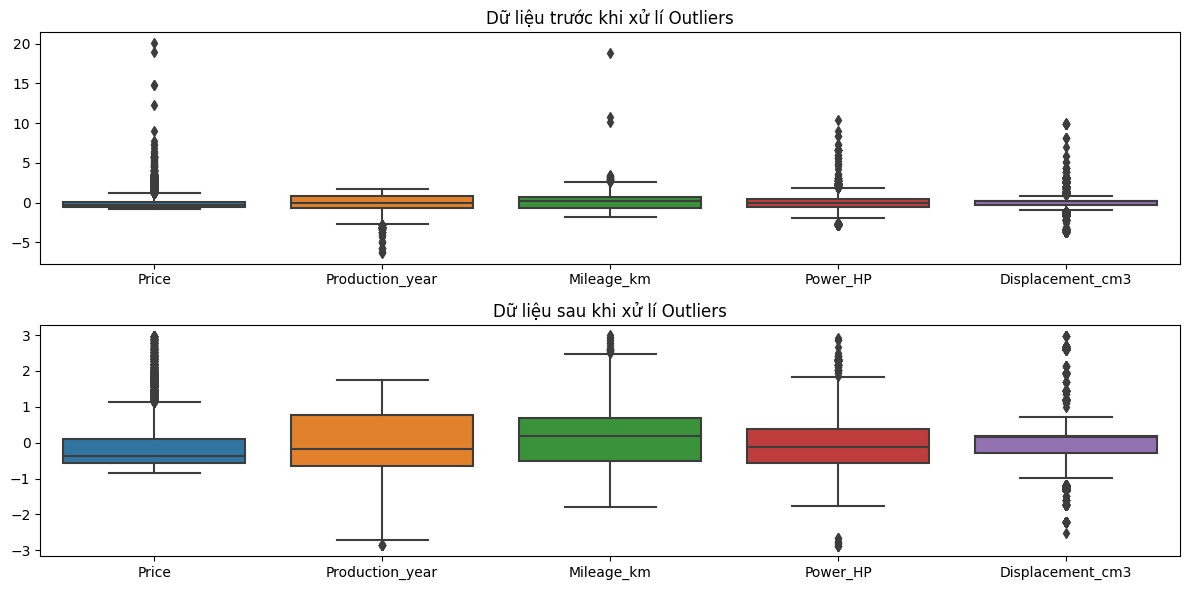


Hình 2: Kết quả tiền xử lí dữ liệu (SuperVised Learning)

* Dữ liệu ban đầu gồm 10000 phần tử, sau khi xử lí Outliers chúng ta còn lại 9553 phần tử và đây cũng là dữ liệu cuối cùng sau tiền xử lí.

### 2.6.2 Trực quan dữ liệu trước và sau khi xử lí nhiễu.

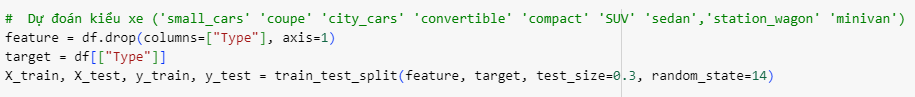




* Sau khi xử lí nhiễu các giá trị chỉ nằm trong ngưỡng từ -3 đến 3.

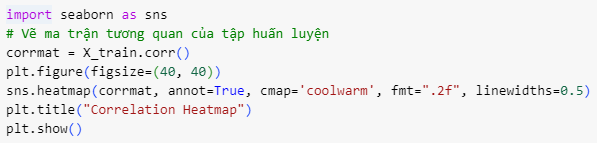
# 3 Ứng dụng thuật toán học có giám sát (Supervised Learning).

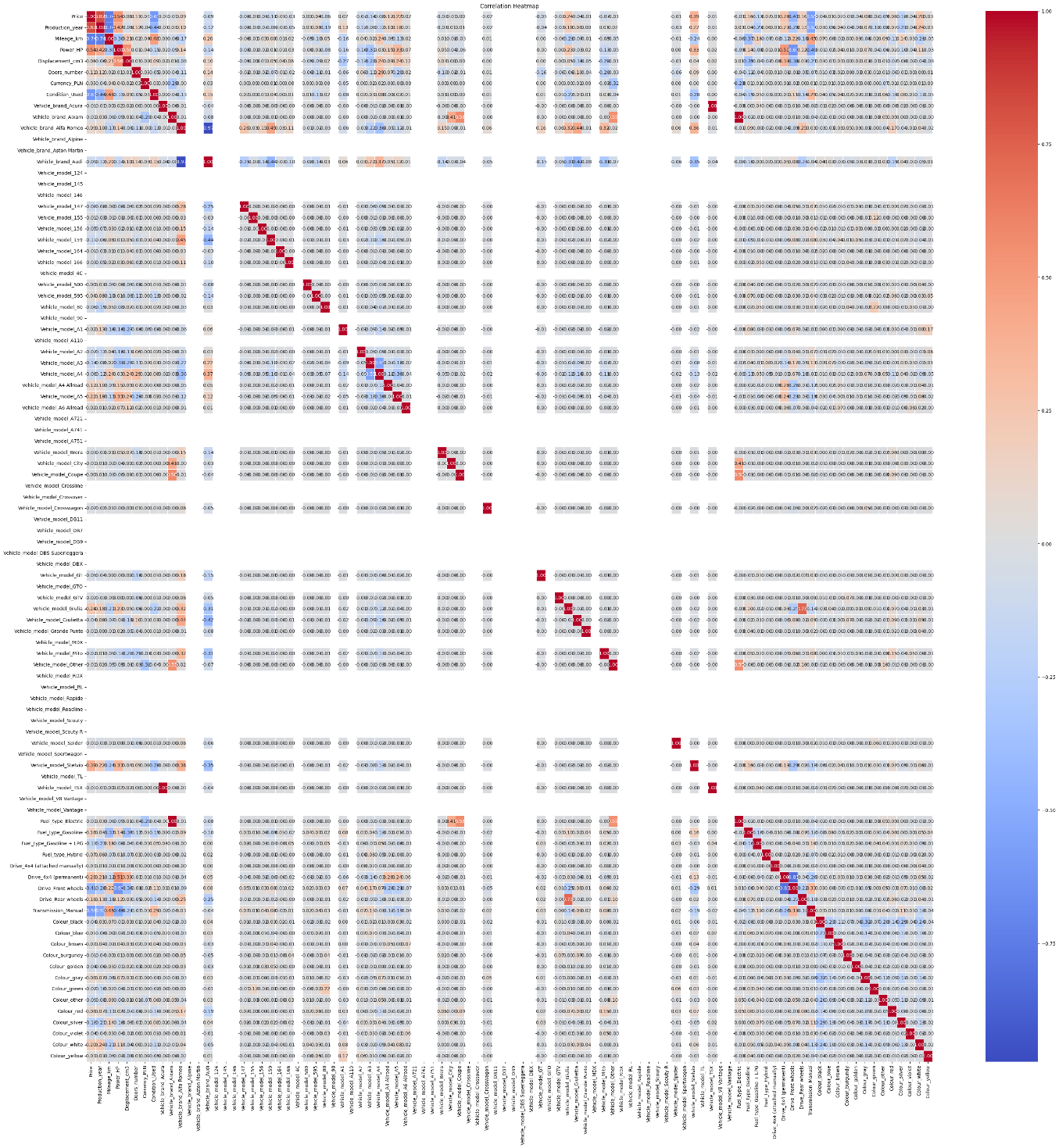
## 3.1 Chia mô hình.



## 3.2 Phân tích tương quan dữ liệu.

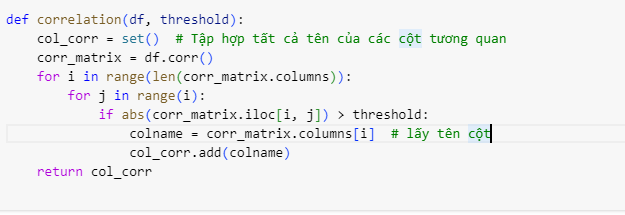
### 3.2.1 Vẽ ma trận tương quan.

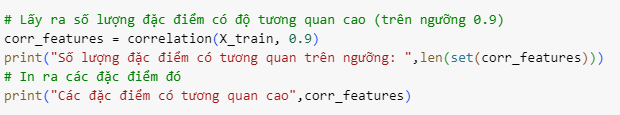




* Vì dữ liệu sau tiền xử lí quá nhiều đặc trưng nên khi trực quan ra ta sẽ không thể hình dung được.

### 3.2.2 Phân tích tương quan.

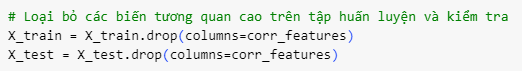






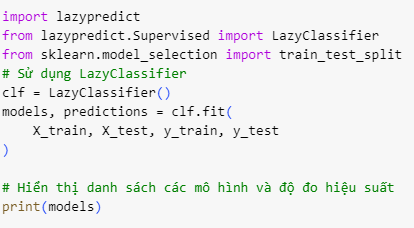
* Sau khi phân tích tương quan ta có thể tìm ra được 3 biến có độ tương quan cao trong ma trận tương quan.

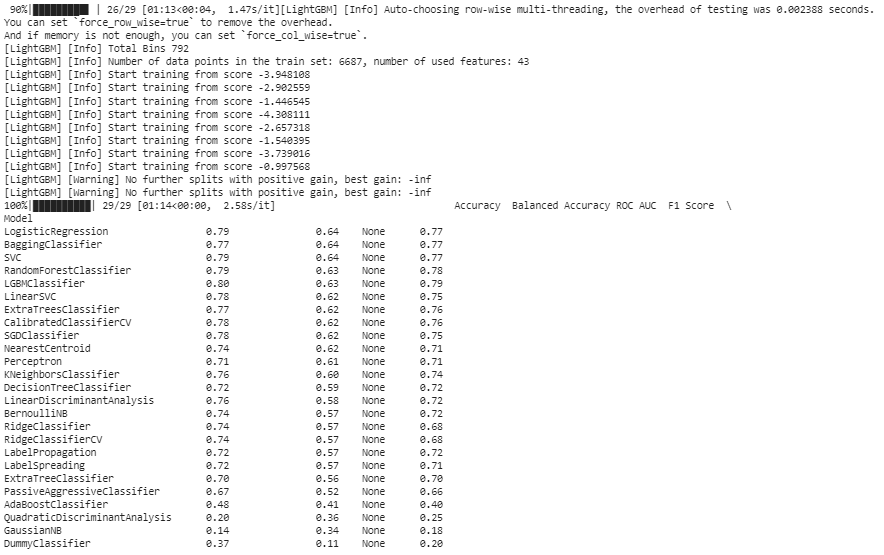
### 3.2.3 Loại bỏ các biến tương quan cao.

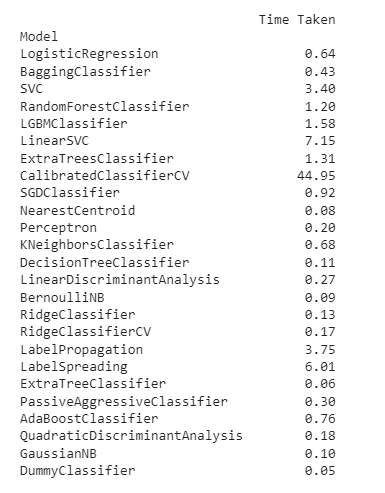


* Loại bỏ các biến tương quan cao là một chiến lược phổ biến để cải thiện hiệu suất và giảm độ phức tạp của mô hình, đặc biệt trong các mô hình yêu cầu tính toán hiệu quả và dự đoán chính xác.

## 3.3 Ứng dụng LazyPredict để tìm thuật toán phù hợp cho mô hình.





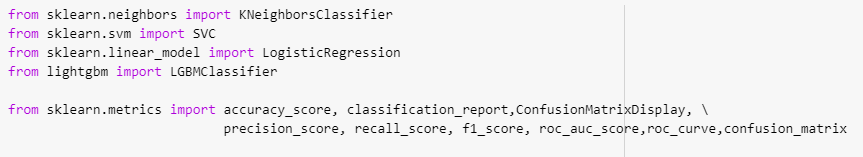


* Sau khi thực hiện LaZyPredict ta có thể tìm ra được thuật toán có độ chính xác cao nhất là LGBMClassifer với độ chính xác là 0,80 (80%).

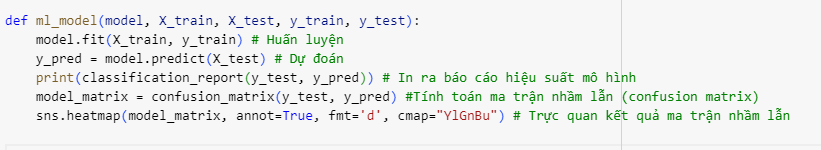
## 3.4 Ứng dụng các thuật toán học có giám sát (Supervised Learning).

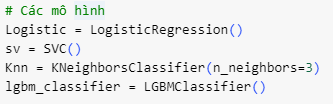
Bài toán đặt ra: Dựa vào nguồn dữ liệu để dự đoán kiểu dáng của xe (Type)

* Gọi thư viện:



* Tạo hàm để chạy mô hình:

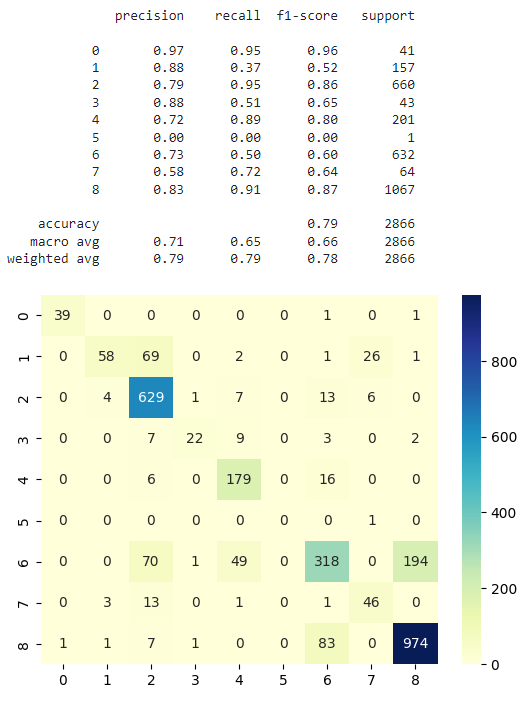




* Ma trận nhầm lẫn (Confusion Matrix) giúp chúng ta biết được số lượng dự đoán các giá trị đúng và sai

### 3.4.1 Dự đoán bằng hồi Quy Logistic (LogisticRegression).



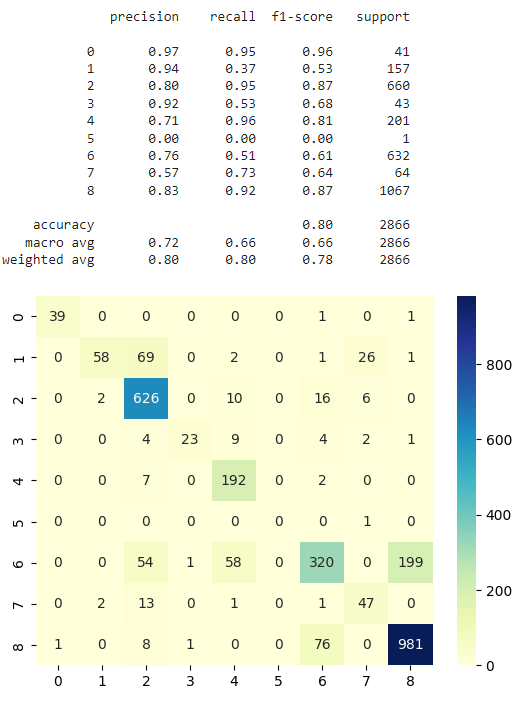


Hình 3: Kết quả dự đoán bằng Hồi Quy Logistic (LogisticRegression)

* Sử dụng hồi Quy Logistic dự đoán, ta có thể thấy được độ chính xác khá cao là 0,79 (79%).

### 3.4.2 Dự đoán bằng Máy Vector hỗ trợ (Support Vector Machine).



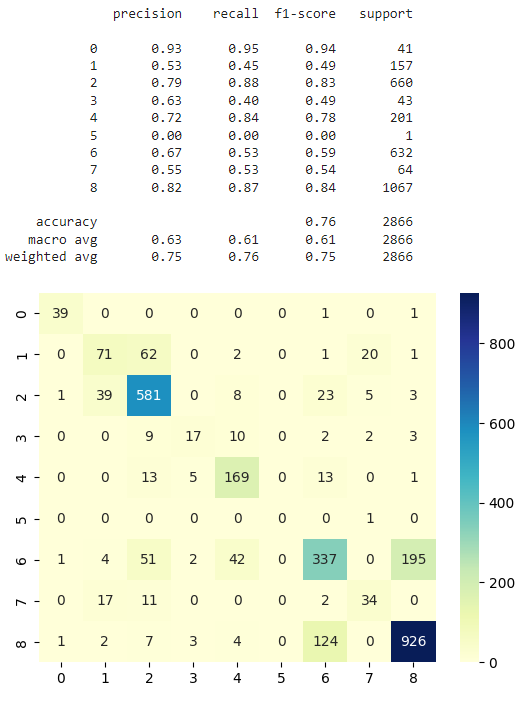


Hình 4: Kết quả dự đoán bằng thuật toán Support Vector Machine (SVM)

* Sử dụng Support Vector Machine dự đoán, ta có thể thấy được độ chính xác khá cao là 0,80 (80%).

### 3.4.3 Dự đoán bằng K-Nearest-NeighBors.



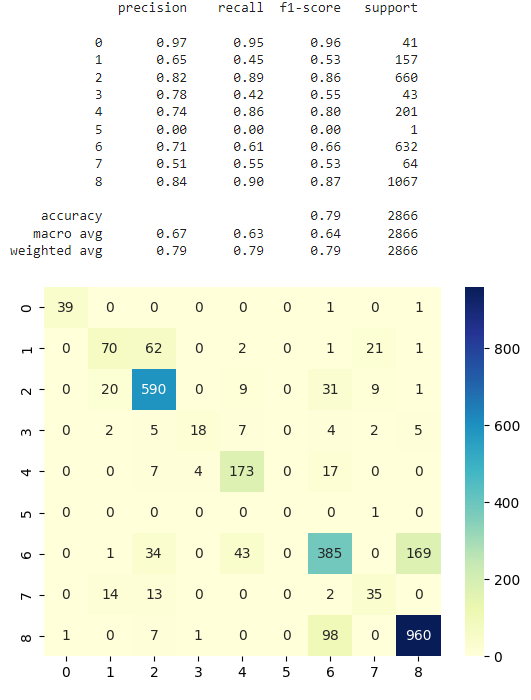


Hình 5: Kết quả thuật toán K-Nearest-NeighBors (KNN)

* Sử dụng Support Vector Machine dự đoán, ta có thể thấy được độ chính xác khá cao là 0,76 (76%).

### 3.4.4 Dự đoán bằng LightGBM (LGBMClassifier).



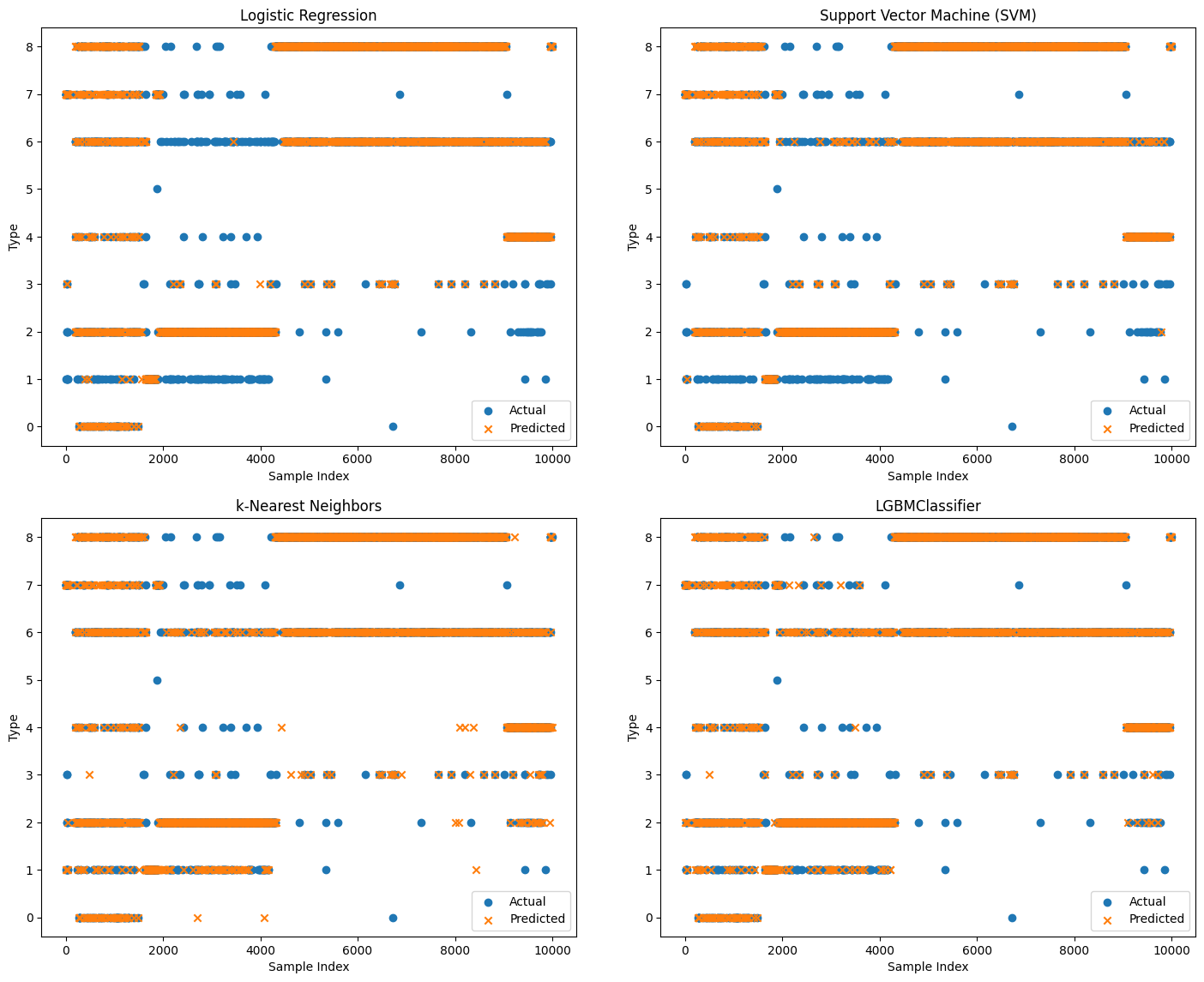


Hình 6: Kết quả dự đoán bằng thuật toán LGBMClassifer (LGBM)

* Sử dụng LightGBM dự đoán, ta có thể thấy được độ chính xác khá cao là 0,79 (79%).

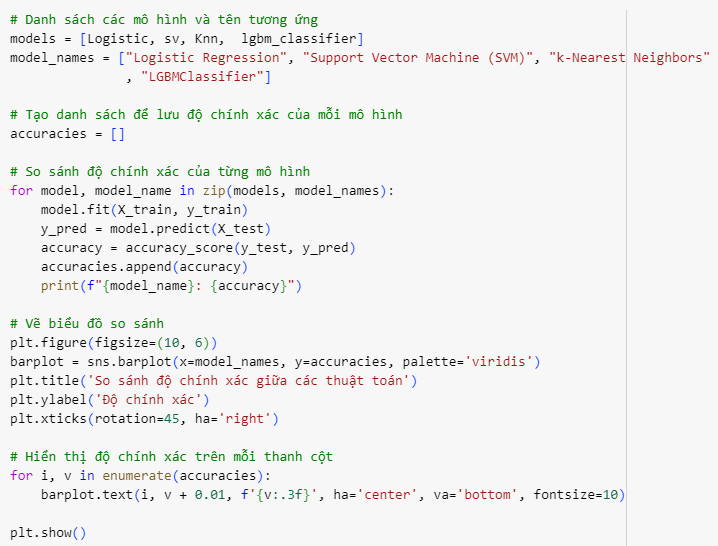
### 3.4.5 Trực quan dữ liệu thực tế với dữ liệu dự đoán của 4 thuật toán sau

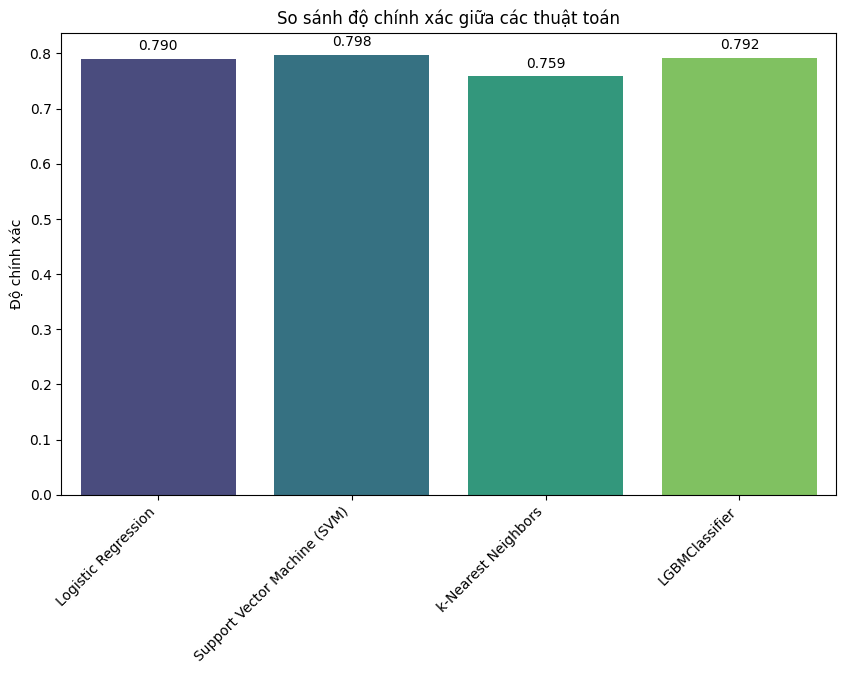




* Nhìn vào 4 biểu đồ trên ta có thể thấy rõ độ chính xác của 4 biểu đồ đều cao vì các điểm của dữ liệu dự đoán trùng với dữ liệu thực tế khá nhiều.

### 3.4.6 So sánh độ chính xác của các thuật toán.



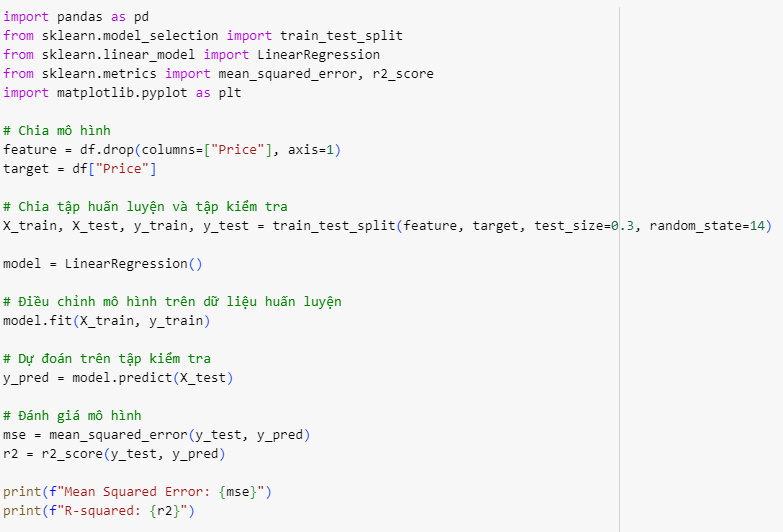


Hình 7: Kết quả so sánh của các thuật toán học có giám sát

### 3.4.7 Dự đoán bằng hồi quy tuyến tính.

Thuật toán hồi quy tuyến tính chỉ có thể dự đoán được biến liên tục nên không thể sử dụng để dự đoán biến phân loại và so sánh độ chính xác cùng với các thuật toán trên.

Vậy để ứng dụng hồi quy tuyến tính ta sẽ đặt lại bài toán như sau: Dự đoán giá xe dựa vào các đặc điểm trong dữ liệu.





Hình 8: Kết quả dự đoán bằng thuật toán hồi quy tuyến tính (LinearRegression)

* Dự đoán giá xe bằng hồi quy tuyến tính trả về độ chính xác khá cao là 0,85 (85%)

# PHẦN 2: HỌC KHÔNG CÓ GIÁM SÁT (UNSUPERVISED LEARNING)

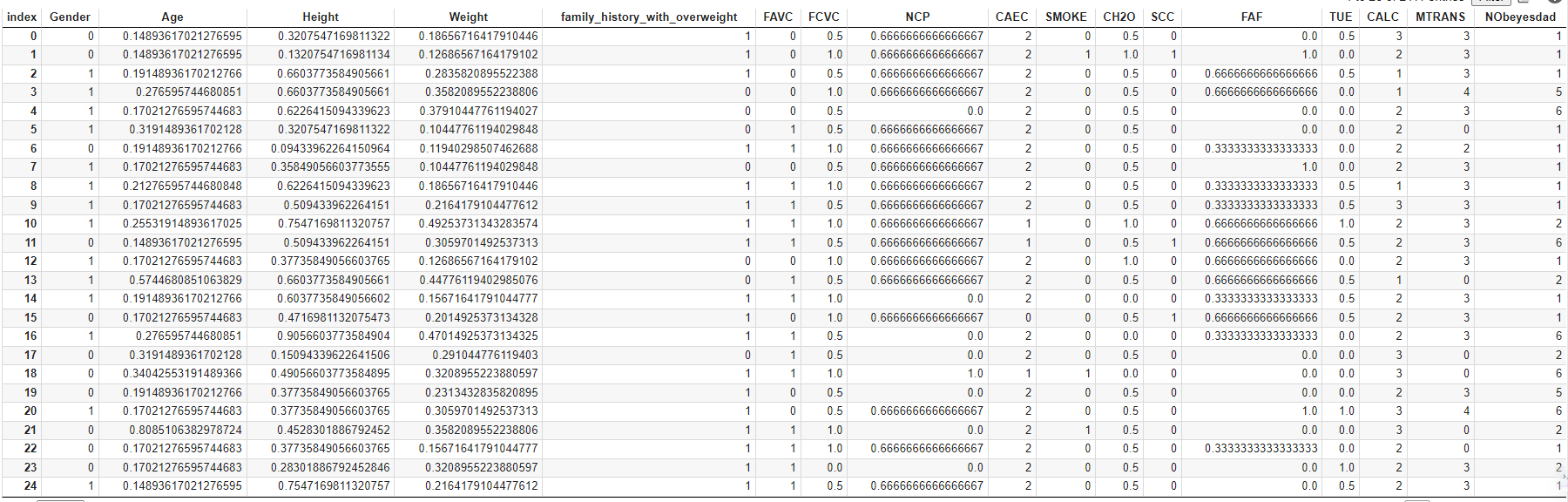
# 1 Giới thiệu dữ liệu.

Mã nguồn dữ liệu: <https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data>

Dữ liệu “Béo phì hoặc nguy cơ bệnh tim mạch” bao gồm thông tin về 2111 người từ các quốc gia Mexico, Peru và Colombia. Dữ liệu được thu thập thông qua một nền tảng web sử dụng cuộc khảo sát, trong đó người dùng ẩn danh trả lời từng câu hỏi. Sau đó, thông tin được xử lý để tạo ra một tập dữ liệu với 17 thuộc tính. Dưới đây là mô tả của các thuộc tính:

* Gender (Giới tính): Thuộc tính này mô tả giới tính của người tham gia khảo sát, có thể là "Male" hoặc "Female".
* Age (Tuổi): Đây là thuộc tính biểu thị độ tuổi của người tham gia khảo sát, có giá trị từ 14 đến 61.
* Height (Chiều cao): Chiều cao của người tham gia, có thể được đo bằng các đơn vị tiêu chuẩn như cm hoặc inch.
* Weight (Cân nặng): Cân nặng của người tham gia, có thể được đo bằng các đơn vị tiêu chuẩn như kg hoặc pound.
* family\_history\_with\_overweight (Tiền sử gia đình với béo phì): Thuộc tính mô tả xem người tham gia có tiền sử gia đình với tình trạng béo phì hay không.
* FAVC (Frequent consumption of high caloric food): Mức độ tiêu thụ thường xuyên thức ăn có nhiều calo.
* FCVC (Frequency of consumption of vegetables): Mức độ tiêu thụ rau củ.
* NCP (Number of main meals): Số bữa ăn chính trong một ngày.
* CAEC (Consumption of food between meals): Mức độ tiêu thụ thức ăn giữa các bữa ăn chính.
* SMOKE (Hút thuốc): Thuộc tính này có giá trị "yes" hoặc "no", mô tả xem người tham gia có hút thuốc hay không.
* CH2O (Consumption of water daily): Mức độ tiêu thụ nước hàng ngày.
* SCC (Calories consumption monitoring): Mức độ theo dõi lượng calo tiêu thụ.
* FAF (Physical activity frequency): Tần suất hoạt động thể chất.
* TUE (Time using technology devices): Thời gian sử dụng thiết bị công nghệ.
* CALC (Consumption of alcohol): Mức độ tiêu thụ rượu.
* MTRANS (Transportation used): Phương tiện giao thông sử dụng.
* NObeyesdad: Thuộc tính mô tả mức độ béo phì của người tham gia dựa trên các thông tin thu thập được. Có thể chia thành các nhóm như "Insufficient Weight", "Normal Weight", "Overweight Level I", "Overweight Level II", "Obesity Type I", "Obesity Type II", "Obesity Type III".

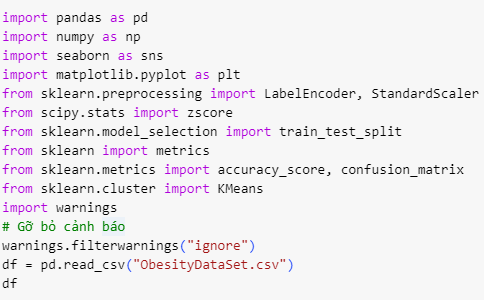
Dữ liệu này có thể được sử dụng để thực hiện các phân tích về mối quan hệ giữa các thuộc tính và mức độ béo phì, cũng như để xây dựng mô hình dự đoán mức độ béo phì dựa trên các thông tin về ăn uống và tình trạng vận động.

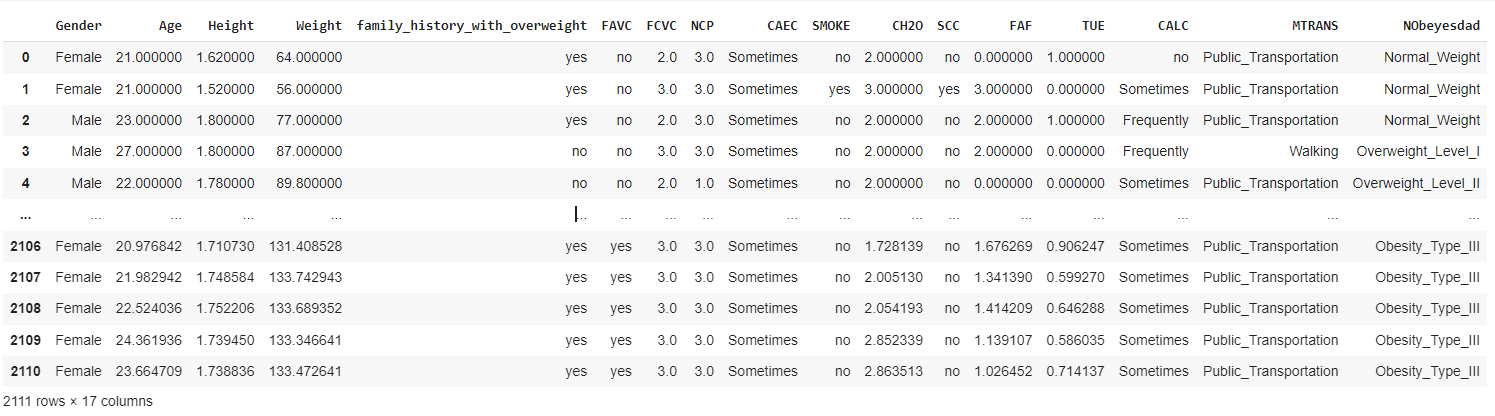


Hình 9: Dữ liệu (Unsupervised Learning)

# 2 Tiền xử lí dữ liệu.

## 2.1 Đọc dữ liệu và gọi thư viện.

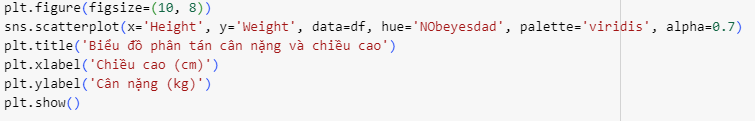




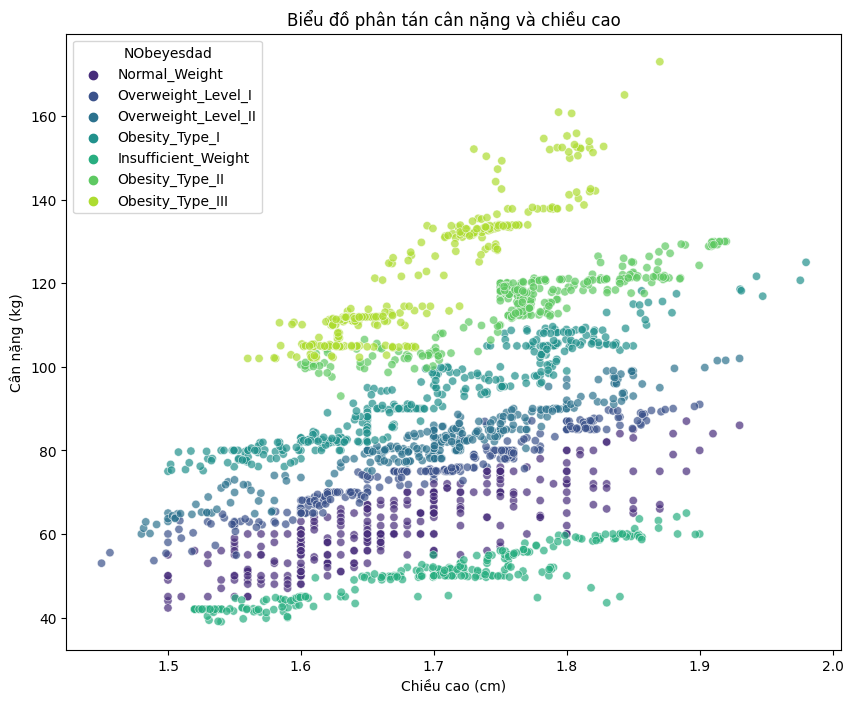
* Dữ liệu gồm 2111 hàng và 17 cột

## 2.2 Trực quan hóa dữ liệu.

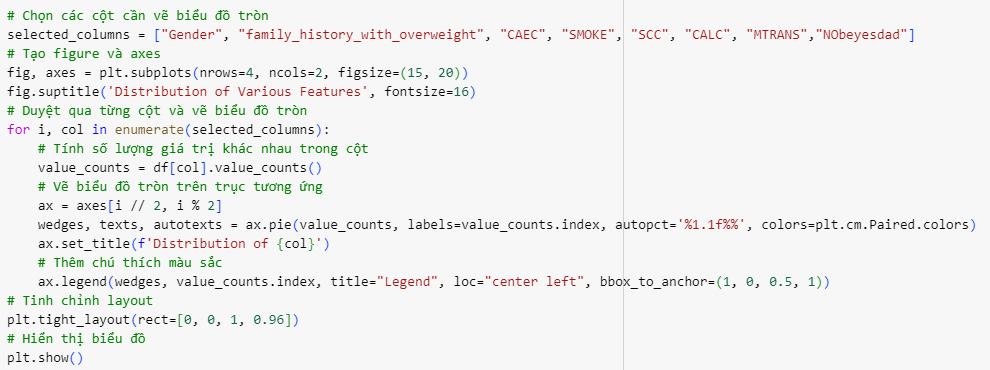
Mã nguồn:



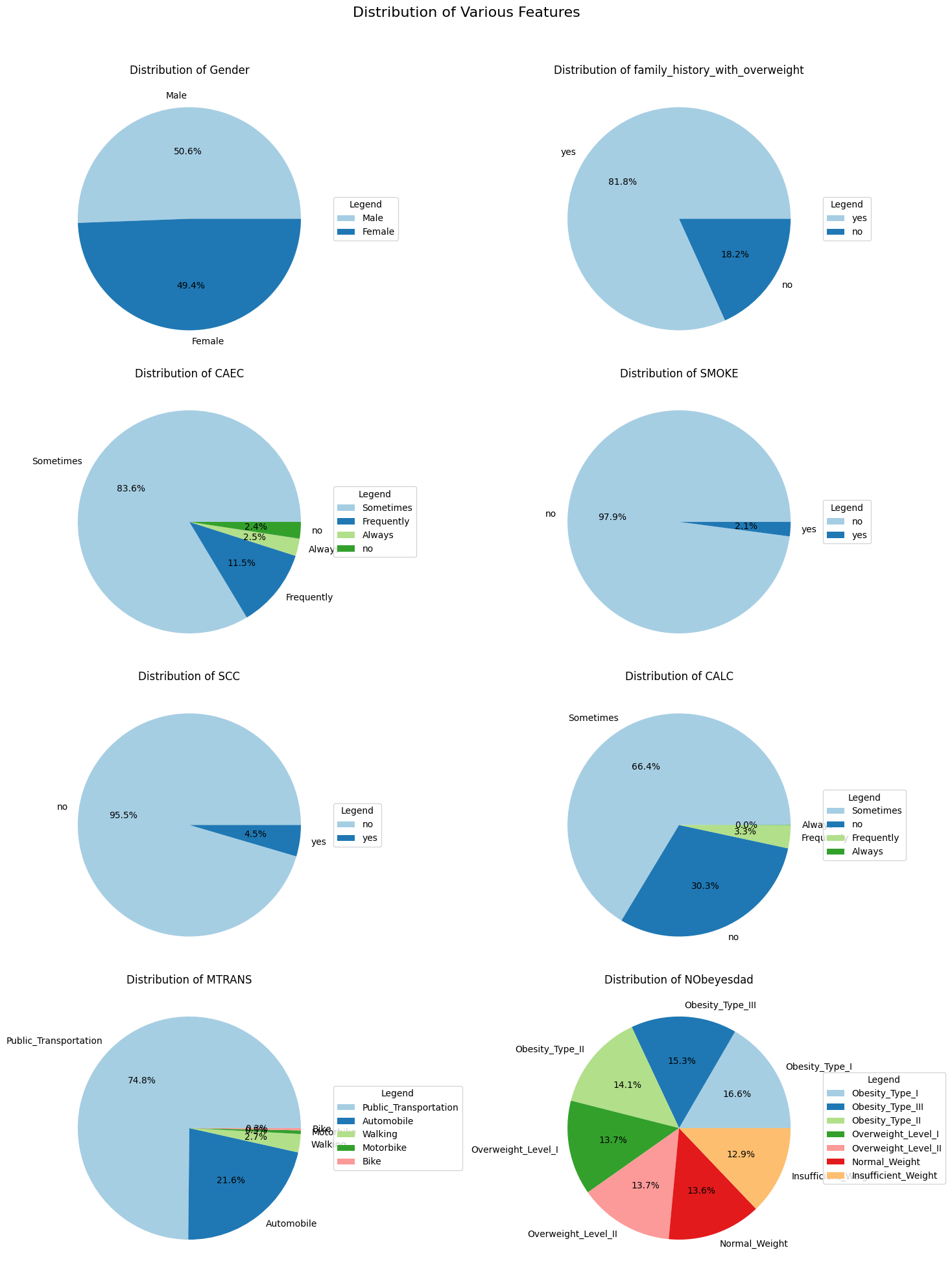
Kết quả:



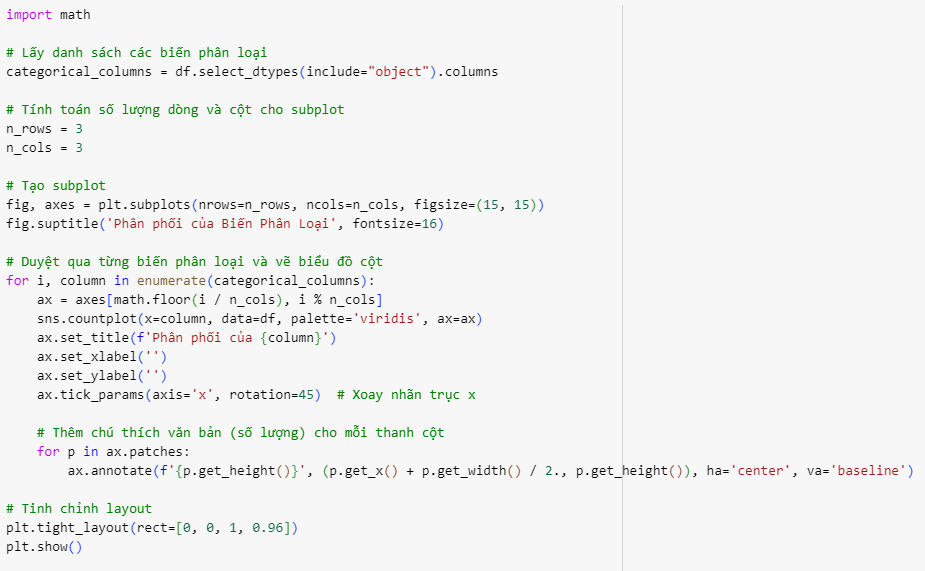
Mã nguồn:



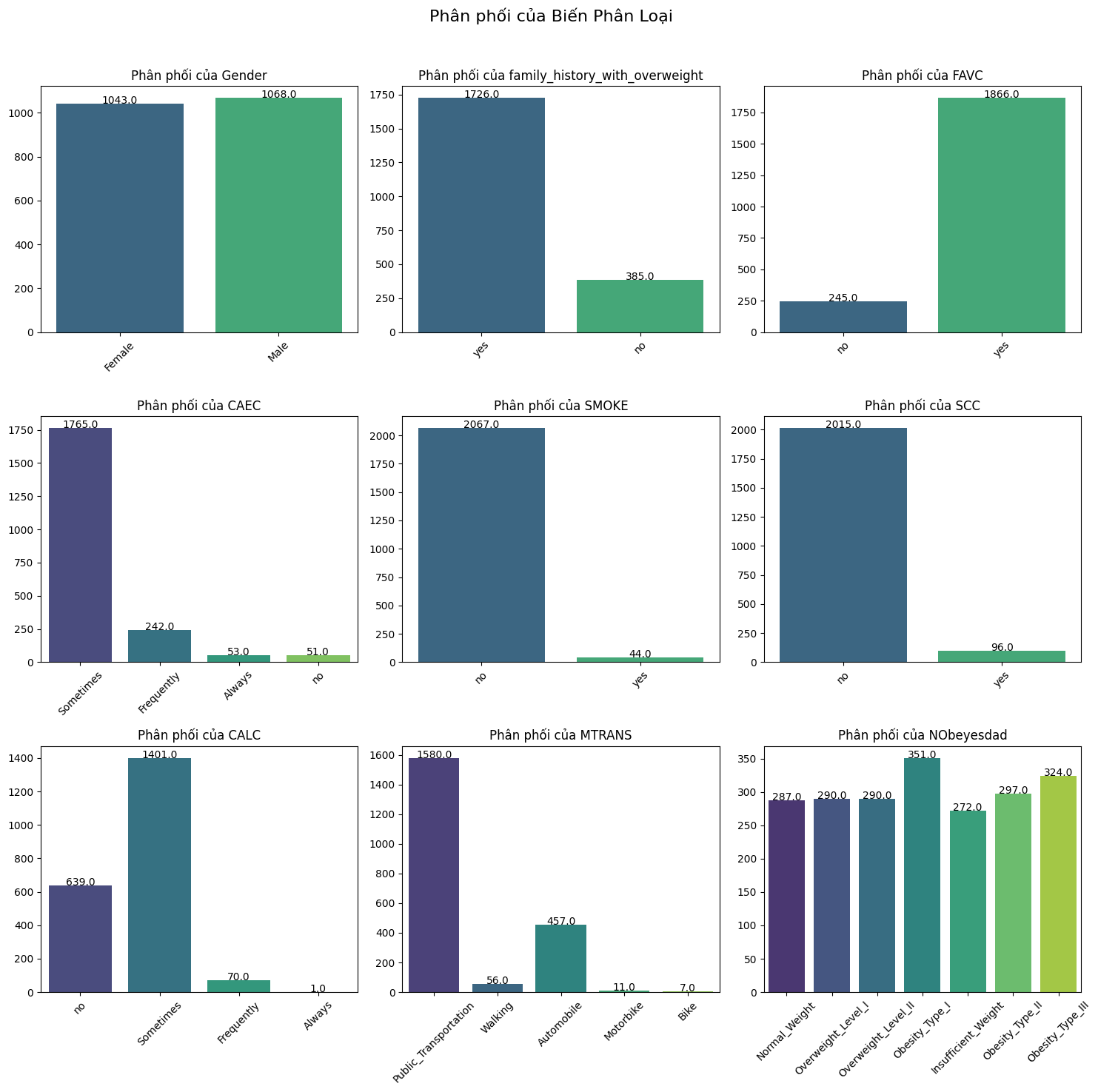
Kết quả:



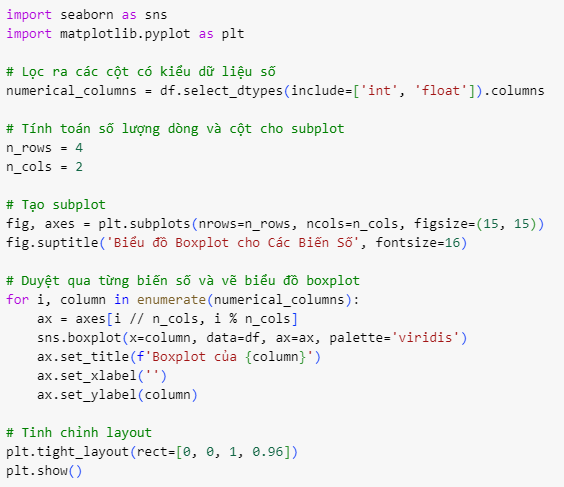
Mã nguồn:



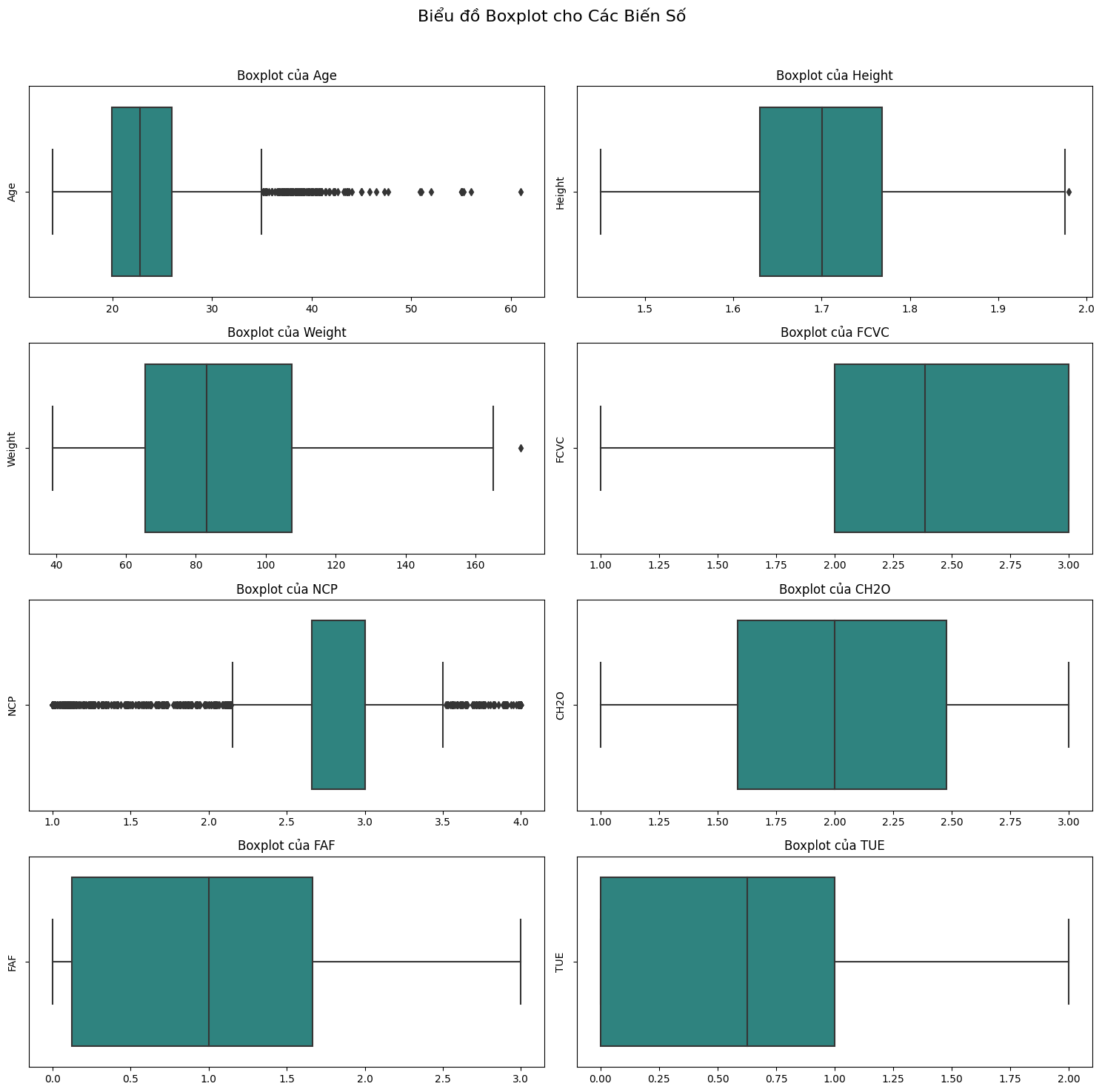
Kết quả:



Mã nguồn:

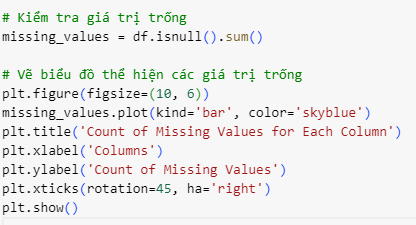


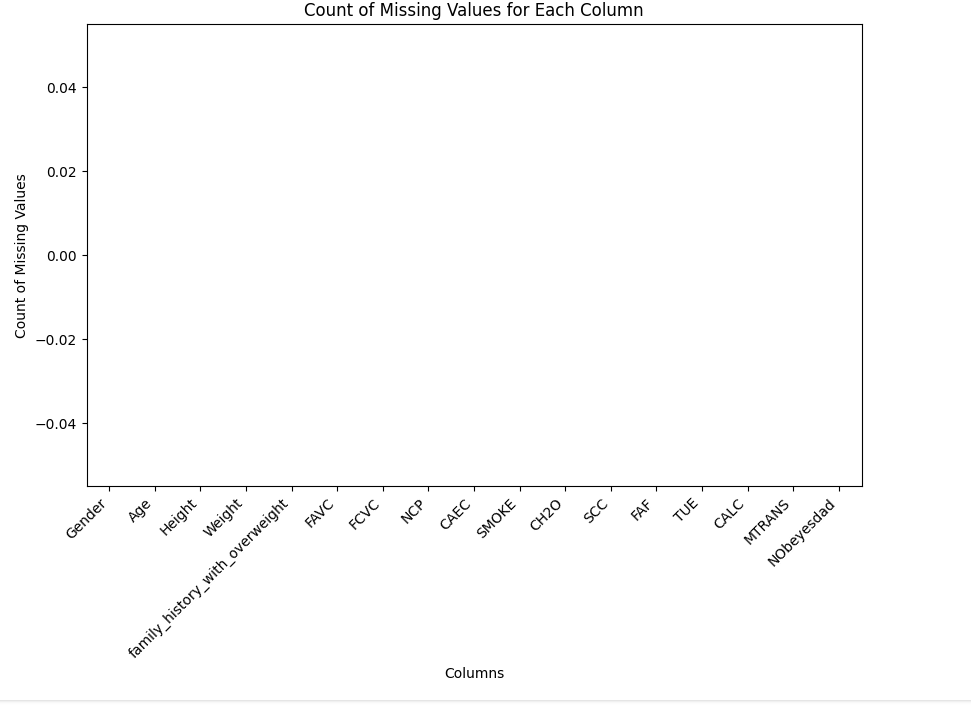
Kết quả:



## 2.3 Xử lí dữ liệu.

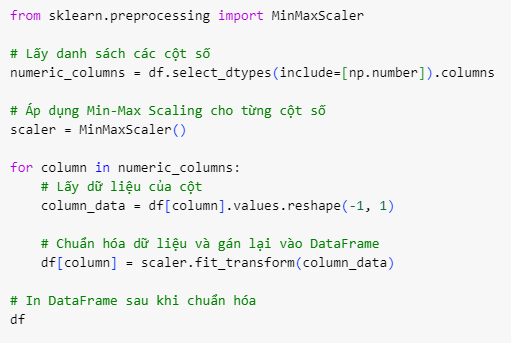
* Trực quan hóa dữ liệu trống:

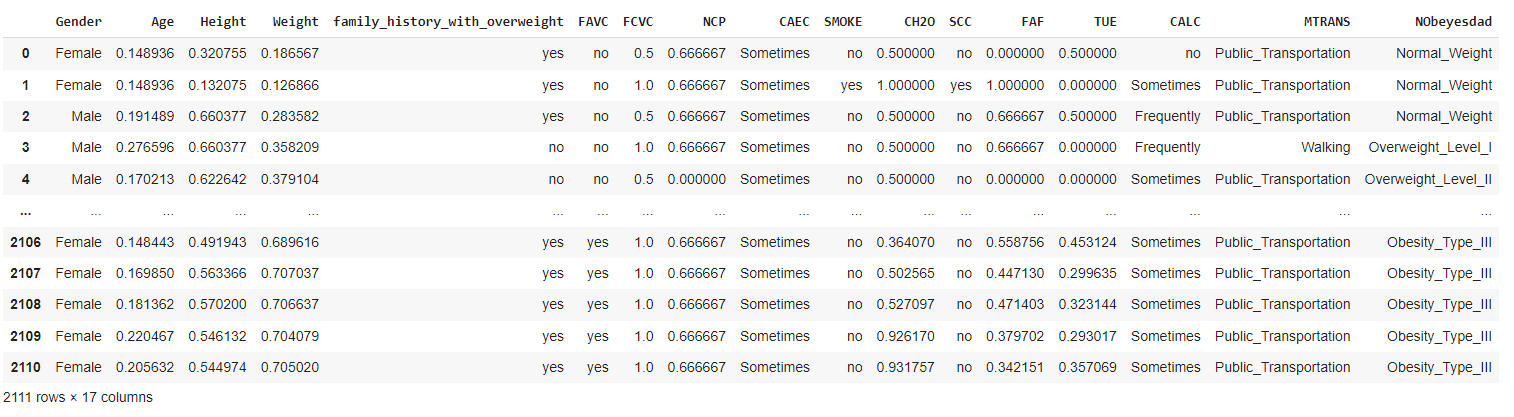




* + Nhìn vào biểu đồ ta có thể thấy không có giá trị trống nào cả nên không cần xử lí giá trị trống.

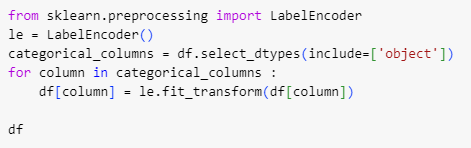
## 2.4 Chuẩn hóa dữ liệu.

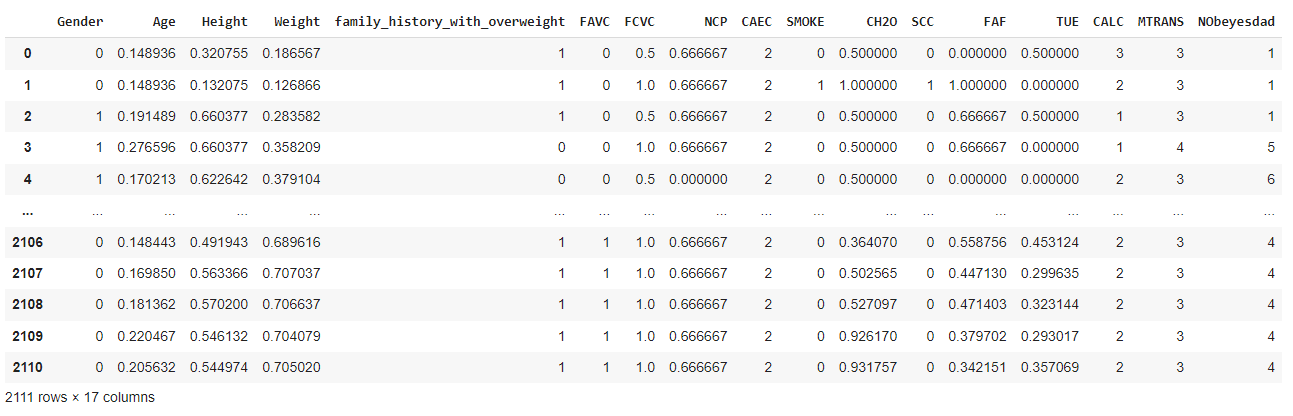




* + Dữ liệu đã được chuẩn hóa.

## 2.5 Mã hóa dữ liệu.



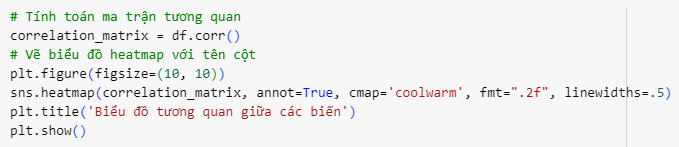


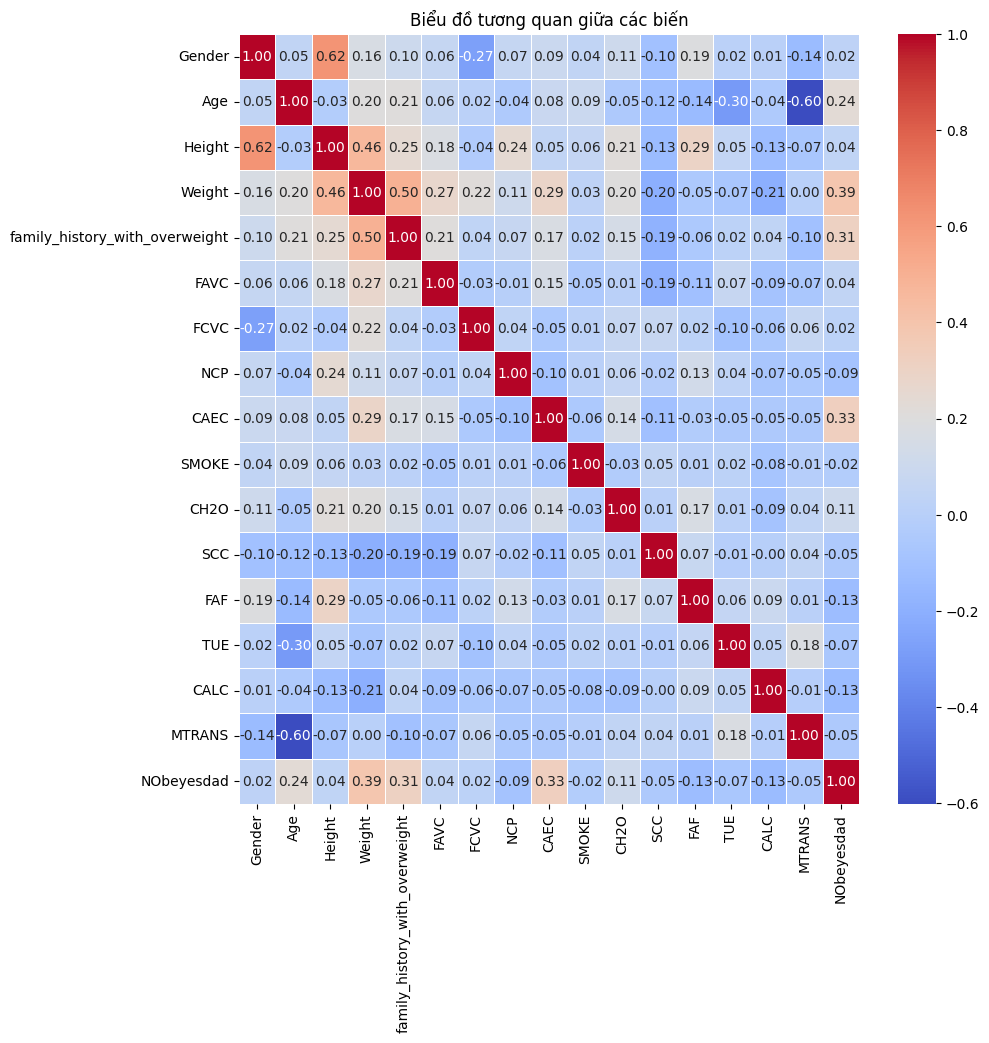
Hình 10: Kết quả tiền xử lí dữ liệu (Unsupervised Learning)

* + Dữ liệu đã được mã hóa về dạng số và đây cũng là dữ liệu cuối cùng sau tiền xử lí dữ liệu.

# 3 Phân tích tương quan dữ liệu.

* Vẽ ma trận tương quan:





Dựa trên ma trận tương quan, có thể thấy rằng:

* + Các biến có tương quan mạnh nhất với nhau là:
* Tuổi và chiều cao (r = 0.62)
* Cân nặng và chiều cao (r = 0.50)
* Tuổi và cân nặng (r = 0.46)
  + Các biến có tương quan yếu nhất với nhau là:
* Tuổi và tiền sử gia đình bị thừa cân (r = 0.25)
* Tuổi và tần suất ăn vặt (r = 0.20)
* Chiều cao và tiền sử gia đình bị thừa cân (r = 0.21)

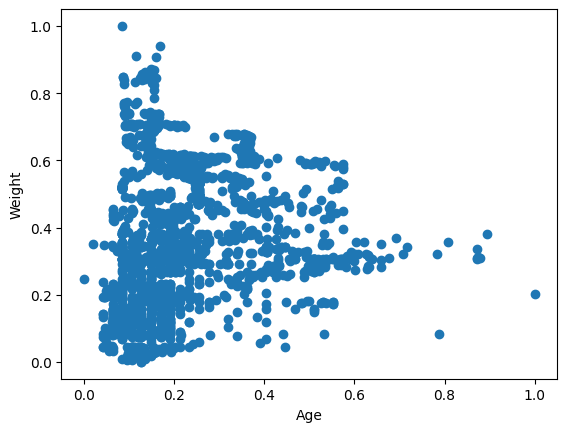
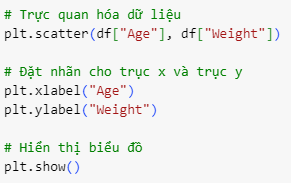
Mối quan hệ giữa các biến:

* Tuổi và chiều cao có mối quan hệ tuyến tính dương mạnh, nghĩa là tuổi càng cao thì chiều cao càng cao. Điều này là dễ hiểu vì con người thường cao lên theo tuổi tác.
* Cân nặng và chiều cao cũng có mối quan hệ tuyến tính dương mạnh, nghĩa là cân nặng càng cao thì chiều cao càng cao. Tuy nhiên, mối quan hệ này không mạnh bằng mối quan hệ giữa tuổi và chiều cao.
* Tuổi và tiền sử gia đình bị thừa cân có mối quan hệ tuyến tính dương yếu, nghĩa là người lớn tuổi có nhiều khả năng có tiền sử gia đình bị thừa cân hơn người trẻ tuổi. Điều này có thể là do tuổi tác làm tăng nguy cơ mắc các bệnh mãn tính, chẳng hạn như tiểu đường, là những bệnh có liên quan đến thừa cân.
* Tuổi và tần suất ăn vặt có mối quan hệ tuyến tính dương yếu, nghĩa là người lớn tuổi có nhiều khả năng ăn vặt nhiều hơn người trẻ tuổi. Điều này có thể là do người lớn tuổi có nhiều thời gian rảnh hơn và có nhiều khả năng bị căng thẳng hơn.
* Chiều cao và tiền sử gia đình bị thừa cân có mối quan hệ tuyến tính dương yếu, nghĩa là người cao có nhiều khả năng có tiền sử gia đình bị thừa cân hơn người thấp. Điều này có thể là do người cao có nhiều diện tích bề mặt hơn để lưu trữ chất béo.

# 4 Ứng dụng các thuật toán học không có giám sát (Unsupervised Learning).

Bài toán: Phân cụm tuổi và cân nặng

## 4.1 Trực quan dữ liệu trước khi phân cụm.

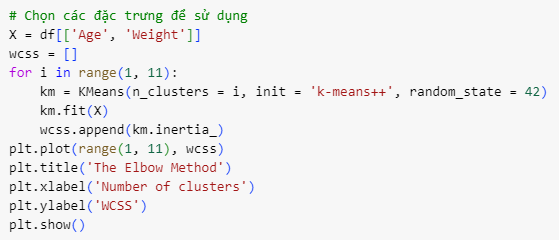


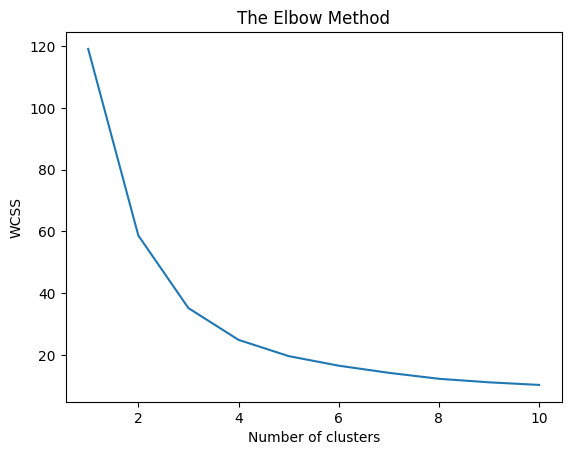
## 4.2 Phân cụm bằng thuật toán K-means.

* Cách Hoạt Động: K-means là một thuật toán phân cụm tuyến tính và yêu cầu xác định trước số lượng cụm cần tìm.
* Quy Trình:
* Chọn k trung tâm cụm ban đầu ngẫu nhiên.
* Gán mỗi điểm dữ liệu vào cụm có trung tâm gần nhất.
* Cập nhật trung tâm cụm bằng cách tính trung bình của tất cả các điểm thuộc cụm đó.
* Lặp lại quá trình gán và cập nhật cho đến khi sự thay đổi giữa các trung tâm là nhỏ.

### 4.2.1 Sử dụng phương pháp khuỷu tay (Elbow Method).

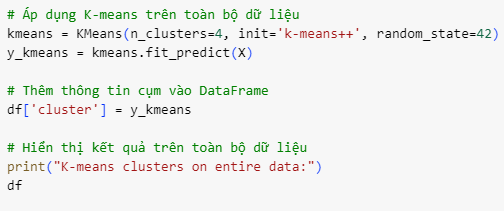
Chúng ta sử dụng phương pháp này để chọn ra số lượng cụm phù hợp cho bài toán.

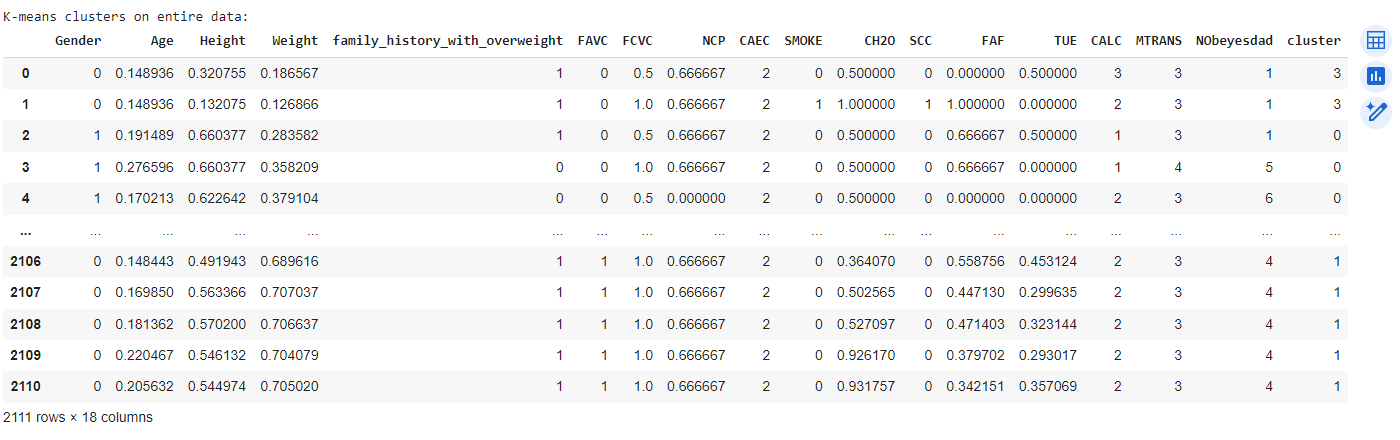




* + Nhìn vào biểu đồ ta có thể chọn số lượng cụm bằng 4.

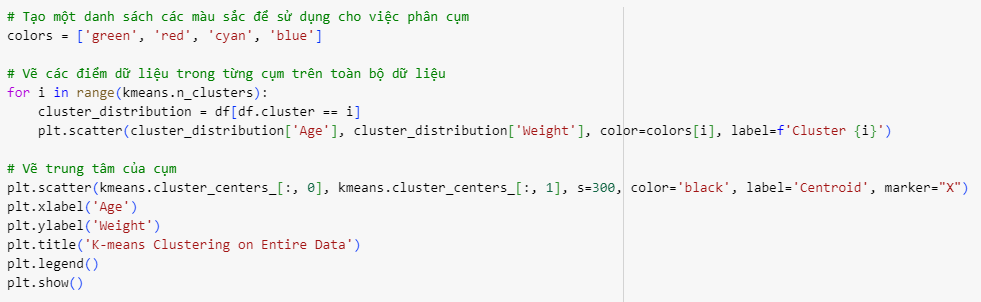
### 4.2.2 Thực hiện phân cụm.

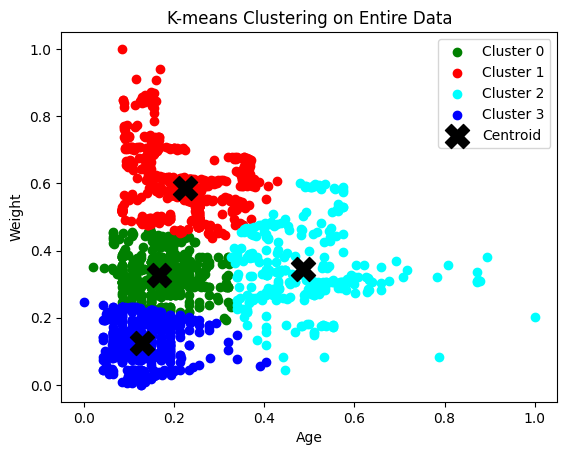




* + Cột “Cluster” đã được thêm vào dữ liệu để cho ta thấy rõ các cụm đã được phân ra.

### 4.2.3 Trực quan hóa sau khi phân cụm.



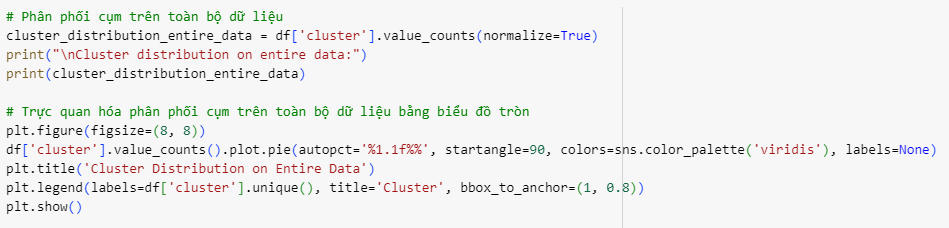


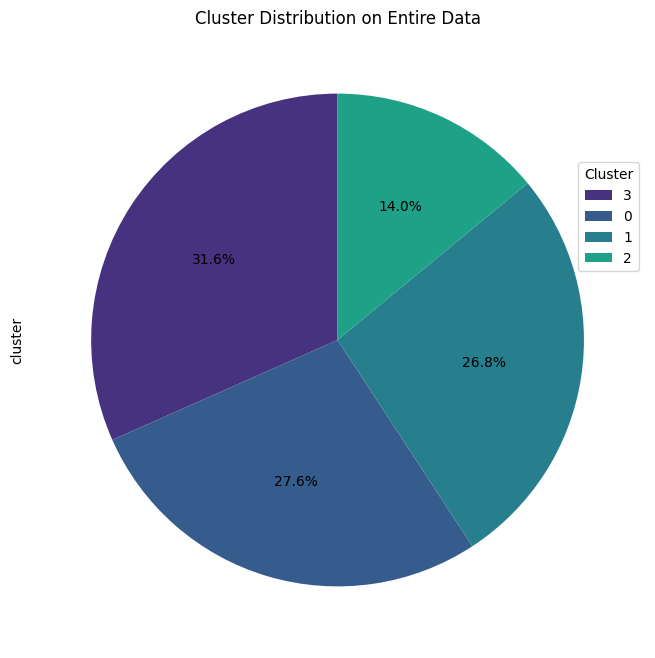
Hình 11: Kết quả phân cụm bằng thuật toán K-means

* + Nhìn vào biểu đồ, ta thấy dữ liệu đã được phân thành 4 cụm.
* Giải thích ý nghĩa:
  + Cụm 0: là cụm có tuổi trung bình và cân nặng trung bình. Cụm này bao gồm những người bình thường.
  + Cụm 1: là cụm có tuổi trung bình nhưng cân nặng rất cao. Cụm này bao gồm những người béo phì.
  + Cụm 2: là cụm có độ tuổi trung bình cao và cân nặng trung bình cao. Cụm này bao gồm những người hơi thừa cân.
  + Cụm 3: là cụm có tuổi trung bình thấp và cân nặng thấp. Cụm này bao gồm những người trẻ - gầy.

### 4.2.4 Phân phối cụm k-means.

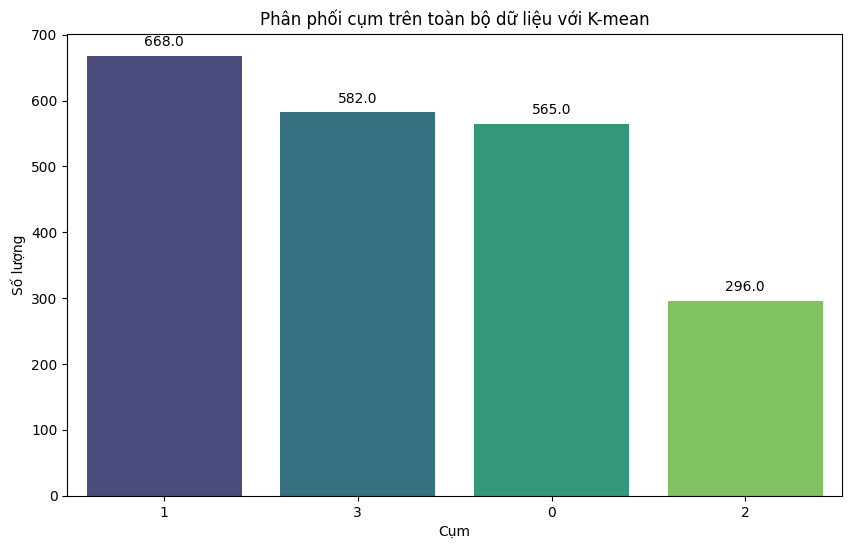
* Trực quan phân phối cụm theo tỉ lệ phần trăm (%):





* + Dựa vào biểu đồ ta có thể thấy được Cụm 0,1,3 phân bố đồng đều. Cụm 2 phân bố thấp nhất.
* Trực quan phân phối cụm biểu đồ cột:



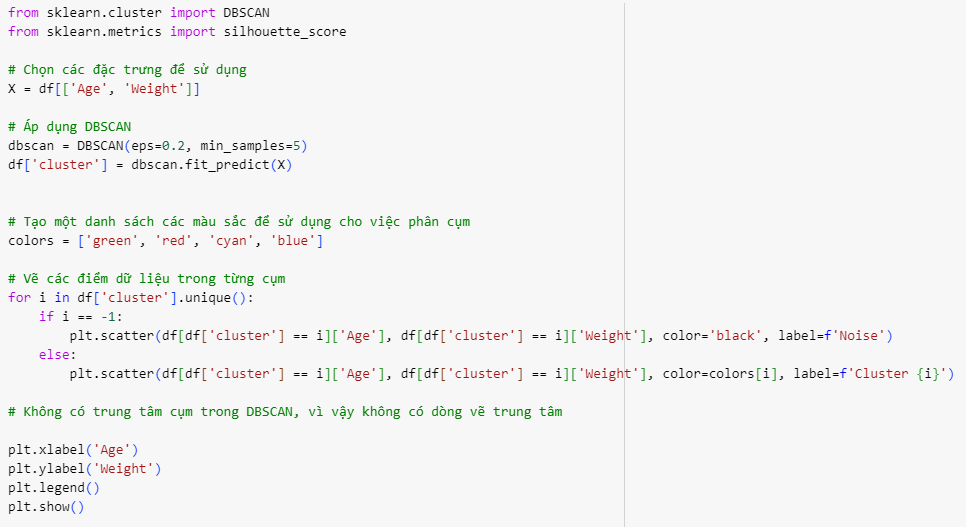


* + Dựa vào biểu đồ ta có thể thấy được số lượng của từng cụm.

## 4.3 Phân cụm bằng thuật toán DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

* Cách Hoạt Động: DBSCAN là một thuật toán dựa trên mật độ, giúp phân loại các điểm thành các cụm dựa trên mật độ của chúng.
* Quy Trình:
* Chọn một điểm dữ liệu ngẫu nhiên và xác định tất cả các điểm nằm trong khoảng cách ε từ nó.
* Nếu số lượng điểm trong khoảng cách ε là lớn hơn một ngưỡng (MinPts), tạo một cụm mới và mở rộng cụm bằng cách liên kết các điểm có thể tiếp xúc với nhau.
* Lặp lại quá trình cho tất cả các điểm dữ liệu.

### 4.3.1 Thực hiện phân cụm và trực quan hóa DBSCAN

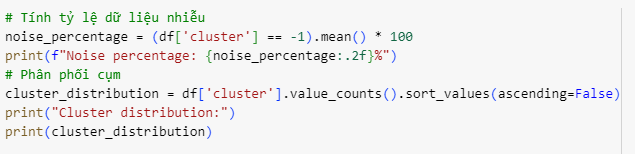




Hình 12: Kết quả phân cụm bằng thuật toán DBSCAN

* + Dựa vào biểu đồ ta thấy đa số các phần tử đều được phân vào cụm 0 và chỉ có một giá trị nhiễu duy nhất.

### 4.3.2 Tính tỉ lệ nhiễu và phân phối cụm.



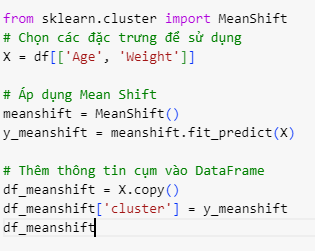


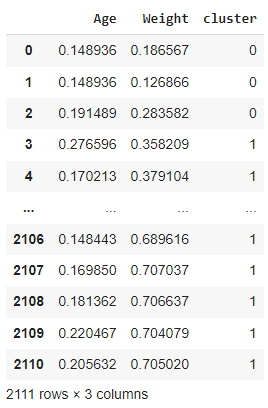
* + Nhìn vào kết quả, ta có thể thấy tỉ lệ nhiễu là 0.05% và có 2 cụm được xác định: Cụm chính được đánh số là 0 và có 2110 điểm dữ liệu thuộc về nó. Cụm khác được đánh số là -1 và chỉ có 1 điểm dữ liệu thuộc về nó.

## 4.4 Phân cụm bằng thuật toán MeanShift

* Cách Hoạt Động: Meanshift là một thuật toán không giả định hình (non-parametric) không yêu cầu đặc tả trước về số lượng cụm. Nó hoạt động bằng cách di chuyển các trung tâm cụm dựa trên gradient của hàm mật độ xác suất (probability density function) của dữ liệu.
* Quy Trình:
* Bắt đầu từ các trung tâm ước lượng ban đầu.
* Tính toán vectơ dịch chuyển dựa trên gradient của hàm mật độ xác suất.
* Di chuyển trung tâm theo vectơ dịch chuyển.
* Lặp lại quá trình cho đến khi không có sự thay đổi đáng kể trong trung tâm.

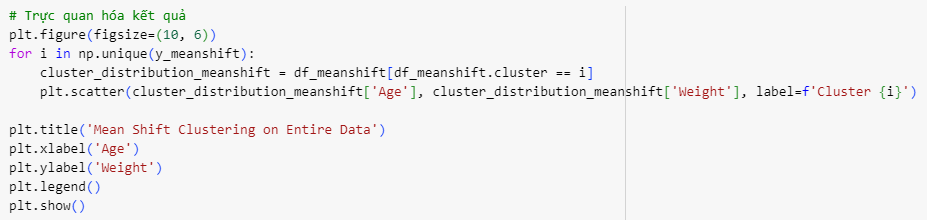
### 4.4.1 Thực hiện phân cụm MeanShift.

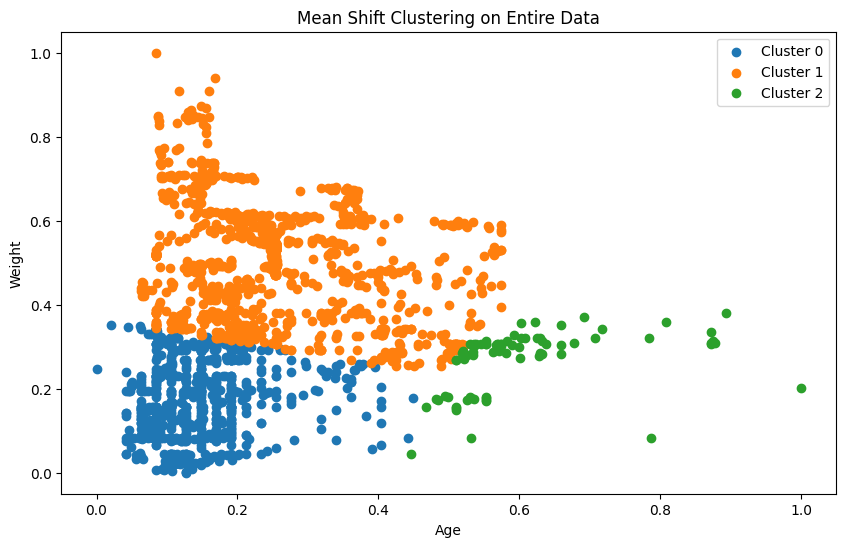




* + Dữ liệu đã được phân cụm.

### 4.4.2 Trực quan hóa sau khi phân cụm MeanShift

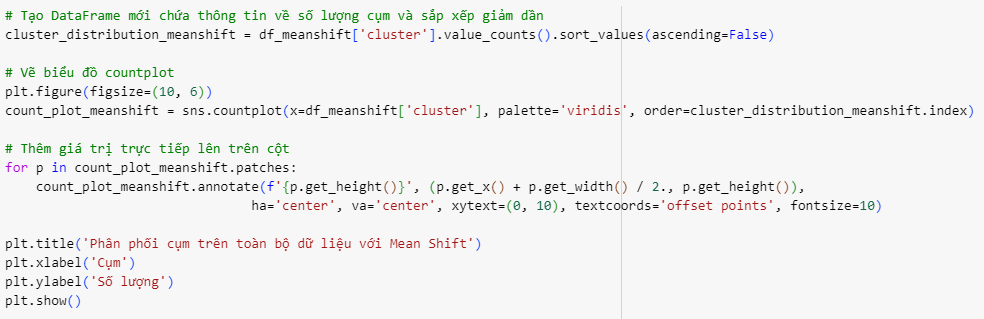


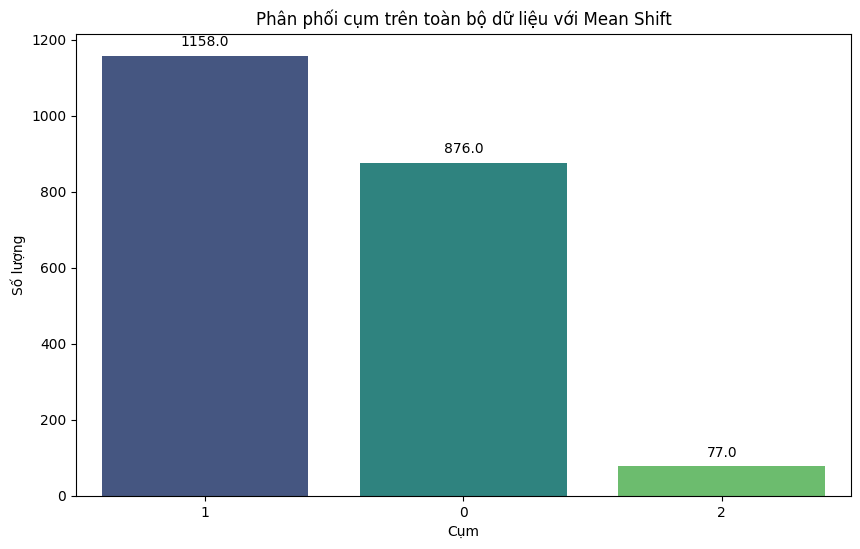


Hình 13: Kết quả phân cụm bằng thuật toán MeanShift

* + Dữ liệu đã được phân thành 3 cụm như trong hình.
* Giải thích ý nghĩa:
  + Cụm 0: là cụm có tuổi trung bình và cân nặng trung bình. Cụm này bao gồm những người có thể trạng bình thường.
  + Cụm 1: là cụm có tuổi trung bình nhưng cân nặng cao hơn. Cụm này bao gồm những người có thể trạng hơi thừa cân.
  + Cụm 2: có tuổi trung bình cao nhưng cân nặng thấp. Cụm này bao gồm những người gầy.

### 4.4.3 Phân phối cụm trên dữ liệu với MeanShift.

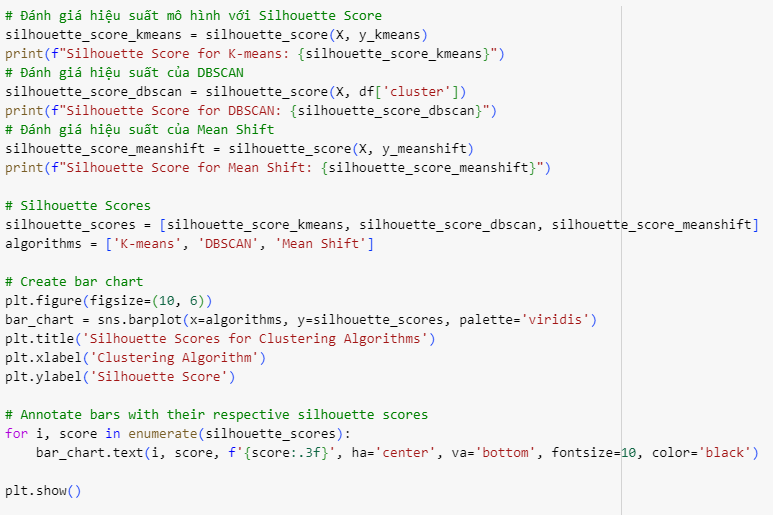


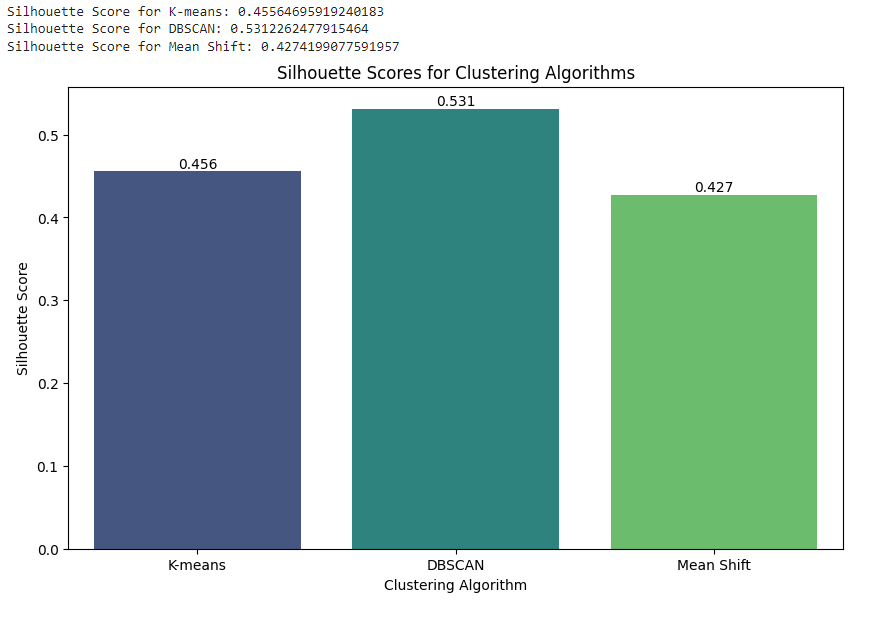


Hình 14: Kết quả so sánh các thuật toán học không có giám sát

* + Dựa vào biểu đồ ta thấy Cụm 1 và 0 được phân bố đồng đều, còn cụm 2 phân bố không đều và rất thấp

### 4.5 Đánh giá hiệu suất mô hình và so sánh các thuật toán phân cụm trên.





* + Ta thấy hiệu suất của 3 thuật toán này nằm ở mức trung bình từ 0,4 đến 0,5. Đây được xem là kết quả khá thấp.