**1.2.2. Thực hiện trực quan hóa dữ liệu trên tập dữ liệu về phân loại chất lượng rượu đỏ.**

Mô tả: Bộ dữ liệu: chứa thông tin về các loại rượu vang đỏ, mỗi mẫu rượu được mô tả bởi nhiều thuộc tính hóa học (features). Mục tiêu: Xây dựng mô hình học máy để dự đoán chất lượng rượu vang (quality) dựa trên các thông số hóa học.

Dữ liệu vào: fixed acidity – độ axit cố định, volatile acidity – độ axit bay hơi, citric acid hàm lượng axit citric, residual sugar – đường dư, chlorides – nồng độ muối (clorua), free sulfur dioxide – SO₂ tự do, total sulfur dioxide – tổng lượng SO₂, density – mật độ rượu, pH – độ pH, sulphates – hàm lượng sunphat, alcohol – nồng độ cồn. Kết quả: Quality – điểm số chất lượng rượu (thường từ 0 đến 10, nhưng thực tế dữ liệu chủ yếu từ 3 đến 8).

**Phân tích dữ liệu (Analyze Data)**

Hiển thị các thông tin về:

+ Số dòng, số cột của dữ liệu

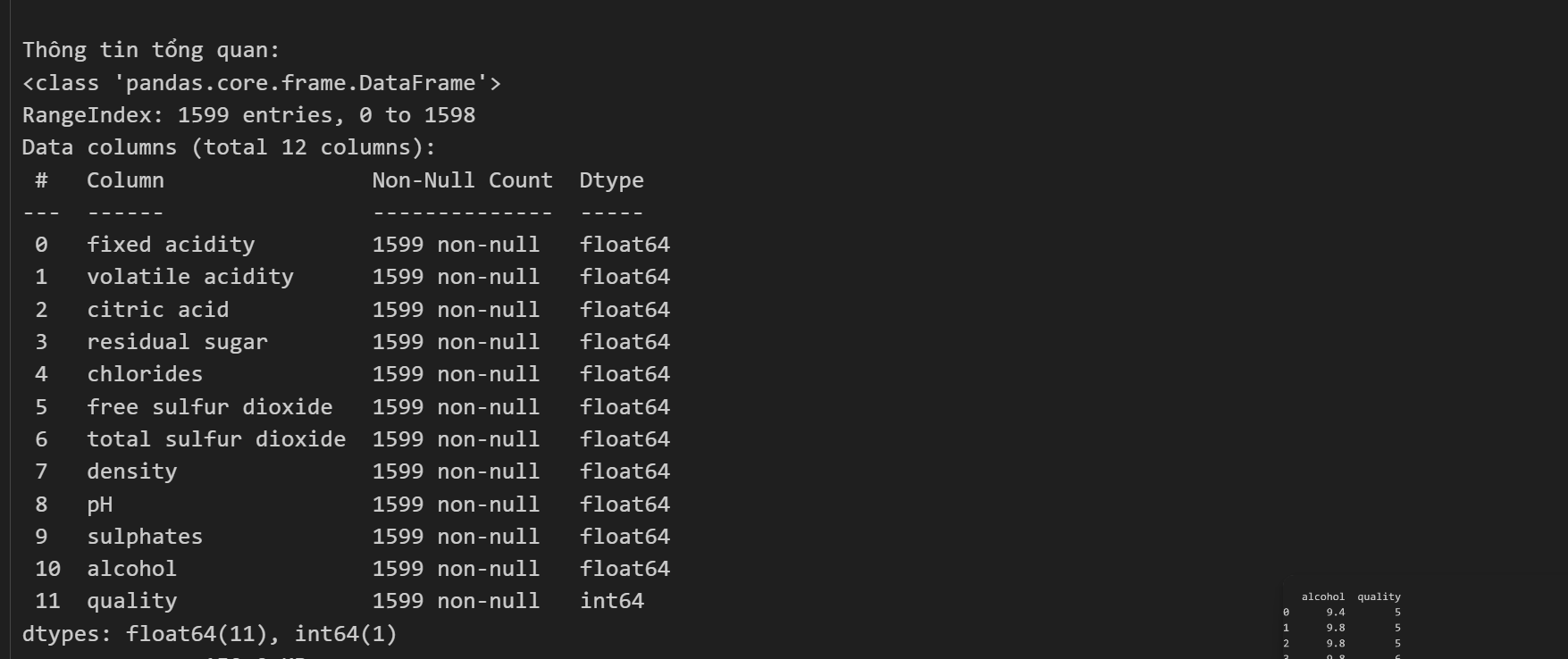
+ Kiểu dữ liệu của từng cột

+ 5 dòng đầu và 5 dòng cuối của bảng dữ liệu

+ Thông tin chung về dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Nhận xét:

Dữ liệu sạch, không thiếu giá trị, tất cả đều ở dạng số. Chất lượng rượu (quality) là biến mục tiêu (target), còn lại là biến đầu vào (features).Có thể tiến hành bước tiếp theo: phân tích thống kê mô tả + trực quan hóa để tìm mối quan hệ giữa các thuộc tính và chất lượng.

**Kiểm tra dữ liệu có bị thiếu và các thống kê cơ bản**

A screen shot of a computer

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

Nhận xét:

Bộ dữ liệu đầy đủ, không missing. Đa số rượu vang đỏ có chất lượng trung bình (5–6 điểm), dữ liệu hướng chất lượng rượu trung bình. Có một số thuộc tính chứa outliers (đặc biệt sulphates, chlorides, residual sugar, total sulfur dioxide) có thể ảnh hưởng xấu đến thống kê mô tả (mean, std).

**Kiểm tra sự tương quan của từng thuộc tính**

A colorful squares with black text

AI-generated content may be incorrect.

Nhận xét:

alcohol có tương quan dương khá rõ (0.48), rượu có nồng độ cồn cao thường được đánh giá chất lượng tốt hơn. volatile acidity có tương quan âm (-0.39), nhiều acid bay hơi thì rượu bị chua, làm giảm chất lượng. citric acid (0.23) và sulphates (0.25) cũng có mối liên hệ dương nhẹ với chất lượng. Các yếu tố khác (density, pH, chlorides, residual sugar, SO₂) có tương quan rất yếu với chất lượng (< |0.2|).

**Sử dụng Univariate Feature Selection để chọn các đặc trưng quan trong**

A computer screen shot of a program code

AI-generated content may be incorrect.

**Trực quan hóa dữ liệu:**

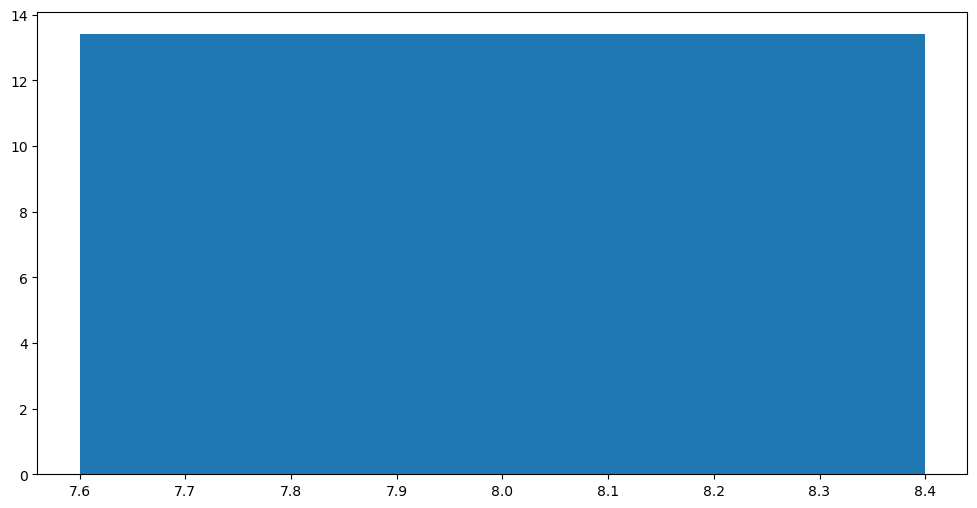
**A group of colored bars

AI-generated content may be incorrect.**

Nhận xét:

Alcohol và Volatile acidity là 2 đặc trưng quan trọng nhất (tác động rõ ràng, độ chênh lệch lớn). Citric acid và Sulphates có tác động bổ sung, tuy xu hướng tăng nhưng mức ảnh hưởng nhỏ hơn. Rượu vang đỏ càng có độ cồn cao, ít axit bay hơi, chứa citric acid và sulphates hợp lý thì chất lượng càng cao.

**Top 10 rượu có chất lượng cao nhất**

****

Nhận xét:

Top 10 rượu có chất lượng tốt nhất có giá trị alcohol tập trung chủ yếu trong khoảng 13-13.5

**Trực quan hóa với Seaborn**

A graph of different colored squares

AI-generated content may be incorrect.

Nhận xét:

Phần lớn rượu có chất lượng trung bình: Chất lượng 5 và 6 chiếm tỷ lệ lớn nhất (gần như toàn bộ dataset, mỗi mức có trên 600 mẫu). Đây là mức "average quality", cho thấy đa số rượu trong tập dữ liệu có chất lượng trung bình. Rượu chất lượng cao hiếm: Chỉ có một số ít mẫu đạt chất lượng 7 hoặc 8 (dưới 250 mẫu). Chất lượng 8 đặc biệt hiếm (< 20 mẫu). Rượu chất lượng thấp cũng rất ít: Chất lượng 3 và 4 có số lượng rất nhỏ (< 70 mẫu).

**Biểu đồ so sánh từng thuộc tính với chất lượng**

**A group of colorful bars

AI-generated content may be incorrect.**

Nhận xét:

Cồn có xu hướng tăng dần theo chất lượng, Alcohol liên quan tích cực đến chất lượng. Volatile acidity (độ axit bay hơi) Giảm rõ rệt khi chất lượng tăng. Volatile acidity càng thấp, rượu càng ngon. Citric acid (axit xitric) Có xu hướng tăng dần theo chất lượng, Citric acid cao là đặc trưng của rượu ngon.

Sulphates (sunphat). Có xu hướng tăng nhẹ theo chất lượng. Sulphates cũng đóng vai trò tích cực trong chất lượng rượu, nhưng tác động yếu hơn alcohol và volatile acidity.

**1.2.3.1. Thực hiện trực quan hóa dữ liệu trên tập dữ liệu về bệnh tiểu đường.**

Mô tả:

Bộ dữ liệu Pima Indians Diabetes được thu thập bởi National Institute of Diabetes and Digestive and Kidney Diseases, Tập dữ liệu bao gồm thông tin y tế của 768 bệnh nhân nữ từ 21 tuổi trở lên, đều có nguồn gốc Pima Indian. Mục tiêu của bài toán là dự đoán khả năng mắc bệnh tiểu đường của bệnh nhân dựa trên các thuộc tính y tế. Dữ liệu vào (Input Attributes): Number of times pregnant (Số lần mang thai) Plasma glucose concentration (Nồng độ glucose trong huyết tương sau 2 giờ) Diastolic blood pressure (Huyết áp tâm trương - mm Hg) Triceps skin fold thickness (Độ dày nếp gấp da cánh tay - mm)2-Hour serum insulin (Insulin huyết thanh sau 2 giờ - mu U/ml) Body mass index (Chỉ số BMI = cân nặng/chiều cao²). Diabetes pedigree function (Chỉ số phả hệ tiểu đường). Age (Tuổi - năm)

 Kết quả (Output / Class variable): 0 = Không mắc tiểu đường, 1 = Mắc tiểu đường

**Phân tích dữ liệu (Analyze Data)**

Hiển thị một số thông tin về dữ liệu:

+ Số dòng, số cột của dữ liệu

+ Kiểu dữ liệu của từng cột

+ 5 dòng đầu và 5 dòng cuối của bảng dữ liệu

+ Thông tin chung về dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Nhận xét:

Kích thước dữ liệu: tổng cộng 768 mẫu (rows), số cột 9 thuộc tính (features).Kiểu dữ liệu: gồm 7 cột số nguyên (int64) và 2 cột số thực (float64), toàn bộ dữ liệu đều là dạng số học. Dữ liệu thiếu: một số cột như Glucose, BloodPressure, SkinThickness, Insulin, BMI có nhiều giá trị bằng 0, thực chất đây là dữ liệu thiếu chứ không phải giá trị hợp lệ.

Ý nghĩa cột: Pregnancies (số lần mang thai), Glucose (nồng độ glucose), BloodPressure (huyết áp tâm trương), SkinThickness (độ dày da), Insulin (nồng độ insulin), BMI (chỉ số khối cơ thể), DiabetesPedigreeFunction (chỉ số di truyền), Age (tuổi), Outcome (0 = không mắc, 1 = mắc tiểu đường).Đặc điểm: Outcome là biến mục tiêu dạng nhị phân, dữ liệu phân loại có thể mất cân bằng nên cần kiểm tra thêm.

**Kiểm tra dữ liệu có bị thiếu và các thống kê cơ bản**

A screenshot of a computer

AI-generated content may be incorrect.

Nhận xét:

Dữ liệu không có missing values theo thống kê nhưng nhiều cột có giá trị bằng 0 bất thường (Glucose, BloodPressure, SkinThickness, Insulin, BMI) cho thấy thực chất đây là giá trị thiếu cần xử lý; giá trị trung bình Glucose ~121 và BMI ~32 phản ánh tình trạng thừa cân/tiền tiểu đường khá rõ; phân phối Outcome mất cân bằng (chỉ ~35% mẫu dương tính tiểu đường); một số thuộc tính có min=0 nhưng max rất cao (ví dụ Insulin max=846) chứng tỏ có nhiều outliers cần xem xét

**Kiểm tra sự tương quan của từng thuộc tính**

A screenshot of a computer screen

AI-generated content may be incorrect.

Nhận xét:

Ma trận tương quan cho thấy Outcome có mối quan hệ mạnh nhất với Glucose (0.47), tiếp theo là BMI (0.29), Age (0.24) và Pregnancies (0.22), trong khi các thuộc tính khác có tương quan rất thấp; Pregnancies và Age có tương quan cao (0.54), Insulin và SkinThickness có tương quan đáng kể (0.44), cho thấy có khả năng dư thừa thông tin; nhìn chung các đặc trưng không quá đa cộng tuyến, nhưng nên ưu tiên các đặc trưng Glucose, BMI, Age, Pregnancies cho việc dự đoán Outcome.

**Sử dụng Univariate Feature Selection để chọn các đặc trưng quan trong**

A screen shot of a computer code

AI-generated content may be incorrect.

**Chuẩn bị dữ liệu cho trực quan hóa**

A screenshot of a computer program

AI-generated content may be incorrect.

**Trực quan hóa với Matplotlib**

A graph with orange lines

AI-generated content may be incorrect.Nhận xét:

Biểu đồ cho thấy 10 bệnh nhân có chỉ số Glucose cao nhất đều đạt mức gần 200, phân bố chủ yếu trong độ tuổi từ 20 đến hơn 60, trong đó tập trung nhiều ở nhóm tuổi 28–32 và 53–62, điều này cho thấy chỉ số Glucose cao không chỉ xảy ra ở một độ tuổi nhất định mà xuất hiện ở nhiều lứa tuổi khác nhau.

**So sánh song song Glucose và BMI của 10 bệnh nhân**

**A red and blue lines

AI-generated content may be incorrect.**

Nhận xét:

Biểu đồ bên trái: top 10 bệnh nhân có Glucose cao nhất tập trung chủ yếu ở mức gần 200 mg/dL, độ tuổi trải đều từ khoảng 21 đến 62 tuổi, không có xu hướng rõ ràng theo tuổi; Biểu đồ bên phải: top 10 bệnh nhân có BMI cao nhất dao động từ khoảng 26 đến 43, độ tuổi trải rộng từ khoảng 22 đến 62 tuổi, có một vài bệnh nhân trẻ nhưng BMI vẫn cao.

**Trực quan hóa với Seaborn**

A graph of different colored bars

AI-generated content may be incorrect.

Nhận xét: biểu đồ thể hiện top 10 bệnh nhân có chỉ số Glucose cao nhất, được sắp xếp và gắn nhãn theo độ tuổi. Nhóm tuổi từ 22 đến 62 đều có chỉ số Glucose rất cao, vượt ngưỡng 125 mg/dL. Đặc biệt, bệnh nhân 62 tuổi có chỉ số cao nhất (khoảng 200 mg/dL), tiếp theo là bệnh nhân 57 tuổi (khoảng 175 mg/dL). Điều này cho thấy chỉ số Glucose có xu hướng tăng cao ở nhóm bệnh nhân lớn tuổi hơn trong top 10 này, nhưng vẫn có những trường hợp trẻ tuổi (22, 28 tuổi) cũng nằm trong nhóm có mức Glucose cao bất thường.

A screenshot of a graph

AI-generated content may be incorrect.

Nhận xét biểu đồ boxplot so sánh phân phối chỉ số Glucose giữa hai nhóm Outcome. Nhóm Outcome = 1 (có bệnh) có chỉ số Glucose trung bình cao hơn rõ rệt so với nhóm Outcome = 0 (không có bệnh). Phân vị thấp nhất (đáy hộp) của nhóm có bệnh cũng cao hơn đáng kể so với phân vị cao nhất (đỉnh hộp) của nhóm không bệnh, cho thấy Glucose là một chỉ số phân biệt rất mạnh và có giá trị dự đoán cao giữa hai nhóm đối tượng.

A screenshot of a graph

AI-generated content may be incorrect.

Nhận xét biểu đồ boxplot so sánh phân phối chỉ số BMI giữa hai nhóm Outcome. Nhóm Outcome = 1 (có bệnh) có chỉ số BMI trung vị cao hơn nhóm Outcome = 0 (không bệnh). Phạm vi phân bố BMI của nhóm có bệnh cũng rộng hơn và nằm ở vị trí cao hơn, cho thấy xu hướng những người có BMI cao sẽ có nguy cơ mắc bệnh cao hơn. Tuy nhiên, sự khác biệt giữa hai nhóm không quá lớn như chỉ số Glucose.

**1.2.3.2. Thực hiện EDA trên tập dữ liệu mua sắm tại siêu thị.**

Online Retail.xlsx là một tập dữ liệu bán lẻ trực tuyến thường được dùng trong học tập và nghiên cứu về khai phá dữ liệu, phân tích khách hàng và xây dựng mô hình dự đoán.Phân tích hành vi khách hàng: tìm khách hàng tiềm năng, phân khúc khách hàng (RFM analysis). Phân tích giỏ hàng: tìm các sản phẩm hay mua kèm nhau (market basket analysis – luật kết hợp). Phân tích kinh doanh: doanh thu theo thời gian, quốc gia, sản phẩm. Dự báo: dự đoán doanh số bán hàng hoặc nhu cầu theo mùa.

    Input: InvoiceNo: Mã hóa đơn (mỗi hóa đơn có thể chứa nhiều sản phẩm). StockCode: Mã sản phẩm. Description: Mô tả sản phẩm. Quantity: Số lượng sản phẩm được mua/bán. InvoiceDate: Ngày giờ phát sinh giao dịch. UnitPrice: Giá đơn vị (giá của một sản phẩm). CustomerID: Mã khách hàng (ẩn danh). Country: Quốc gia của khách hàng. Output Là thông tin rút ra để hỗ trợ quyết định kinh doanh

**Chuẩn bị vấn đề (Prepare Problem)**

**2.1. Khai báo thư viện (Load Libraries)**

**A computer screen shot of a program code

AI-generated content may be incorrect.**

**2.2. Nạp dữ liệu (Load Dataset)**

A screen shot of a computer

AI-generated content may be incorrect.

**3. Phân tích dữ liệu (Analyze Data)**

 Hiển thị một số thông tin về dữ liệu:

+ Số dòng, số cột của dữ liệu

+ Kiểu dữ liệu của từng cột

+ 5 dòng đầu và 5 dòng cuối của bảng dữ liệu

+ Thông tin chung về dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.

**Nhận xét:**

Bộ dữ liệu có kích thước (541909, 8), gồm 541909 dòng và 8 cột, các cột bao gồm cả dữ liệu dạng object, số nguyên, số thực và thời gian, trong đó cột Description có 1454 giá trị thiếu và cột CustomerID có khoảng 135080 giá trị thiếu, dữ liệu chủ yếu mô tả các giao dịch bán hàng với thông tin mã hóa đơn, mã sản phẩm, mô tả sản phẩm, số lượng, thời gian mua, giá đơn vị, mã khách hàng và quốc gia, dung lượng bộ nhớ khoảng 33.1 MB.

**Kiểm tra tính toàn vẹn của dữ liệu**

+ Dữ liệu có bị trùng lặp không? Hiển thị dòng bị vi phạm.

+ Dữ liệu có tồn tại giá trị Null không? Hiển thị dòng bị vi phạm.

+ Dữ liệu có tồn tại giá trị NaN không? Hiển thị dòng bị vi phạm.

A screenshot of a computer

AI-generated content may be incorrect.

Nhận xét

Kết quả kiểm tra cho thấy dữ liệu có tồn tại giá trị Null/NaN chủ yếu ở cột Description và CustomerID, ngoài ra còn có các dòng trùng lặp với số lượng đáng kể, điều này ảnh hưởng đến tính toàn vẹn dữ liệu, do đó cần thực hiện xử lý như loại bỏ hoặc điền giá trị thay thế để đảm bảo chất lượng dữ liệu trước khi phân tích.

Các tính chất thống kê trên dữ liệu số

+ Count, Mean, Standard Deviation, Minimum Value

+ 25th Percentile, 50th Percentile (Median), 75th Percentile, Maximum Value

A screenshot of a computer screen

AI-generated content may be incorrect.

Nhận xét

Cột Quantity có giá trị trung bình khoảng 9.55 nhưng tồn tại giá trị âm rất lớn (-80995) và dương cực lớn (80995) cho thấy dữ liệu có ngoại lệ, phần lớn phân phối tập trung ở khoảng từ 1 đến 10; Cột InvoiceDate trải dài từ 01/12/2010 đến 09/12/2011 cho thấy dữ liệu ghi nhận trong 1 năm; Cột UnitPrice có giá trị trung bình khoảng 4.61 nhưng cũng tồn tại ngoại lệ âm (-11062.06) và dương rất cao (38970) so với khoảng giá phổ biến từ 1.25 đến 4.13; Cột CustomerID có 406,829 giá trị hợp lệ, trung bình khoảng 15287, nằm trong khoảng từ 12346 đến 18287, thể hiện sự phân bổ tương đối đồng đều nhưng vẫn có khoảng thiếu dữ liệu đáng kể.

**Mối tương quan giữa các tính chất (Correlations)**

A black screen with white numbers

AI-generated content may be incorrect.

Nhận xét:

Ma trận tương quan cho thấy các biến trong df\_retail gần như không có mối quan hệ tuyến tính với nhau, cụ thể Quantity và UnitPrice có tương quan rất yếu và âm (-0.0012), Quantity và CustomerID cũng gần như không liên hệ (-0.0036), tương tự UnitPrice và CustomerID (-0.0046), do đó các biến số trong dữ liệu này gần như độc lập, không ảnh hưởng tuyến tính lẫn nhau.

**Hiển thị dữ liệu (Visualize Data)**

**Hiển thị trên từng tính chất đơn (Univariate Plots)**

Box and whisker plots

+ So sánh các trung vị (median) tương ứng của mỗi ô hộp (box plot). Nếu đường trung vị của một ô hộp nằm bên ngoài ô của một ô hộp so sánh, thì có thể có sự khác biệt giữa hai nhóm.

+ So sánh chiều dài hộp để kiểm tra cách dữ liệu được phân tán giữa mỗi mẫu. Hộp càng dài thì dữ liệu càng phân tán. Dữ liệu càng nhỏ càng ít bị phân tán.

+ Một ngoại lệ (outlier) được định nghĩa là một điểm dữ liệu nằm bên ngoài phần rìa (whiskers) của ô hộp.

+ Kiểm tra hướng lệch của dữ liệu (cân đối, các phần tử tập trung trái, phải).

 + Median ở giữa hộp và râu (whiskers) ở hai bên như nhau thì phân bố là đối xứng.

+ Median ở gần đáy hộp hơn và nếu râu ngắn hơn ở đầu dưới của hộp, thì phân phối là lệch dương (lệch phải).

+ Median ở gần đầu hộp hơn và nếu râu ngắn hơn ở đầu trên của hộp, thì phân bố bị lệch âm (lệch trái). A graph with lines and dots

AI-generated content may be incorrect.

**Nhận xét**:

Biểu đồ này hiển thị hai boxplot cho hai cột 'Quantity' (Số lượng) và 'UnitPrice' (Đơn giá) từ dataset Bán lẻ trực tuyến. Cả hai biến đều có phân phối lệch phải mạnh, với rất nhiều điểm outliers ở phía trên, đặc biệt là ở biến 'UnitPrice'. Điều này cho thấy đa phần các giao dịch có số lượng và đơn giá thấp, chỉ có một số ít giao dịch có giá trị hoặc số lượng rất cao

**Biểu đồ Histogram**

A graph of a graph

AI-generated content may be incorrect.

 Nhận xét :

Biểu đồ histogram thể hiện phân phối của 4 thuộc tính trong dataset Bán lẻ. Phân phối của 'Quantity' và 'UnitPrice' cực kỳ lệch phải, với phần lớn giá trị tập trung gần 0 và một số ít giá trị rất lớn ở xa. 'InvoiceDate' cho thấy dữ liệu được thu thập chủ yếu trong năm 2011, với tần suất giao dịch tăng dần và đạt đỉnh vào cuối năm. 'CustomerID' có phân phối tương đối đồng đều, cho thấy dữ liệu khách hàng được phân bố khá cân bằng trong khoảng từ 13000 đến 18000.

**Hiển thị nhiều tính chất (Multivariate Plots)**

A screenshot of a graph

AI-generated content may be incorrect.

**Nhận xét:**

Phân phối: Quantity và UnitPrice tập trung mạnh ở giá trị nhỏ, đuôi rất dài; xuất hiện nhiều giá trị âm và ngoại lệ cực lớn. CustomerID gần như đều vì chỉ là mã định danh. Quan hệ giữa biến: Không thấy tương quan tuyến tính rõ giữa Quantity và UnitPrice. Các ô liên quan CustomerID tạo các “vệt” dọc/ngang, không mang ý nghĩa phân tích. Vấn đề dữ liệu: Giá/SL âm và ngoại lệ rất lớn gợi ý lỗi nhập liệu/hoàn trả. Thang trục quá rộng làm loãng cấu trúc chính. Tiêu đề “Pima Indians Diabetes Dataset” có vẻ không khớp với các biến hiển thị.

A screenshot of a graph

AI-generated content may be incorrect.**Nhận xét:**

Mối quan hệ giữa các biến: biểu đồ cho thấy sự phân bố và mối quan hệ giữa Quantity và UnitPrice. Dữ liệu ngoại lai: có nhiều giá trị bất thường trong cả hai biến, đặc biệt là Quantity với một số điểm rất cao. Phân phối biến: phần lớn giá trị của Quantity tập trung ở mức thấp, trong khi UnitPrice có nhiều điểm giá trị âm, điều này cần được kiểm tra. Sự tương quan: có thể có một mối quan hệ nhất định giữa Quantity và UnitPrice, nhưng cần xem xét kỹ lưỡng do sự hiện diện của dữ liệu ngoại lai. Kích thước biểu đồ: lớn giúp dễ dàng quan sát, nhưng sự hiện diện của nhiều giá trị bất thường có thể làm khó khăn cho việc phân tích. Tóm lại, biểu đồ cung cấp cái nhìn tổng quan về mối quan hệ giữa các biến nhưng cần thêm phân tích để hiểu rõ hơn về dữ liệu.

**Xóa dữ liệu trùng nhau**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Xử lý giá trị rỗng, không hợp lệ**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Xử lý giá trị nan và null**

**A screen shot of a computer program

AI-generated content may be incorrect.**