A black background with a black square

AI-generated content may be incorrect.**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue button with white text

AI-generated content may be incorrect.

**PHÂN TÍCH DỮ LIỆU**

**BÁO CÁO**

**[Bài tập lớn 1] Phân tích khám phá dữ liệu**

Sinh viên thực hiện:

**Đỗ Hữu Lộc - 3123410201**

**Nguyễn Hoàng Phúc - 3123410277**

**Hồ Đắc Khả - 31234110143**

Giảng viên: **Đỗ Như Tài**

Thành phố Hồ Chí Minh, 10 tháng 10 năm 2025

**Mục lục**

1.3 **Thống kê mô tả. 3**

[**1.2. XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU 8**](#_audgheqvu93r)

[1.2.1. Ôn tập lý thuyết 8](#_mghrl8uavd7h)

[1. Vai trò và Tầm quan trọng của Trực quan hóa dữ liệu trong EDA 8](#_mjatjvb2c3yt)

[2. Các loại biểu đồ phổ biến và Trường hợp sử dụng 8](#_fmxu0f45gld9)

[3. Cách chọn loại biểu đồ phù hợp 9](#_a6n946ijdl8w)

[4. Sự khác biệt giữa các thư viện trực quan hóa trong Python 9](#_p3mst3upzhoe)

[5. Các nguyên tắc thiết kế biểu đồ hiệu quả 10](#_w4ssp8p1dg2d)

[**1.3. Phân tích Dữ liệu Chất lượng Rượu Vang Đỏ (Red Wine Quality Analysis Report) 35**](#_g6dubfwhg9qd)

[1. Tóm tắt Điều hành (Executive Summary) 35](#_aktnuybf6byc)

[2. Khám phá Dữ liệu Chuyên sâu (Exploratory Data Analysis - EDA) 35](#_fc85x2hx9bqd)

[2.1. Tổng quan Dữ liệu 35](#_8ny37ov734ia)

[2.2. Phân tích Phân phối Biến Mục tiêu 36](#_cjm4lsud8maq)

[2.3. Phân tích Đơn biến (Univariate Analysis) và Tiền xử lý 36](#_potju9doiezh)

[2.5. Phân tích Tương quan và Mối quan hệ Chi tiết 38](#_xo5sy6af693v)

[Nhận xét: 40](#_sjm0z6xshgcq)

[3. Phương pháp luận và Tiền xử lý 43](#_fjezmpes5n9x)

[3.1. Kỹ thuật Tiền xử lý 43](#_47bhxvh8nxxm)

[3.2. Giới thiệu Mô hình 44](#_k74t6qamg6dj)

[4. Kết quả và Đánh giá Mô hình 44](#_vqh39bwlyuvk)

[5. Tính toán điểm tầm quan trọng của các đặc trưng hàng đầu giúp dự đoán chất lượng rượu vang và trực quan hóa chúng 45](#_if7voh1a4s4f)

[6. Kết luận và Hướng Phát triển Tiếp theo 46](#_6corifwe7ghp)

# 

# 

| **MSSV** | **Họ tên** | **Phân công công việc** | **Mức độ hoàn thành** |
| --- | --- | --- | --- |
| Hồ Đắc Khả | 31234110143 | 1.1 Thống kê mô tả | 100% |
| Đỗ Hữu Lộc | 3123410201 | 1.2 Xử lí và trực quan hóa dữ liệu | 100% |
| Nguyễn Hoàng Phúc | 3123410277 | 1.3 Phân tích dữ liệu chất lượng rượu vang đỏ | 100% |

**THỰC HÀNH : PHÂN TÍCH KHÁM PHÁ DỮ LIỆU**

**1.1 Thống kê mô tả.**

**1.1.1 Ôn tập lý thuyết.**

- Thống kê mô tả là gì? Khác gì với thống kê suy luận?

Thống kê mô tả (Descriptive Statistics):  
 Là việc tóm tắt, mô tả dữ liệu hiện có bằng các con số hoặc biểu đồ (như trung bình, trung vị, độ lệch chuẩn, histogram, boxplot...).

=> Mục tiêu : Hiểu dữ liệu đang có.

-Thống kê suy luận (Inferential Statistics):

Là việc rút ra kết luận hoặc dự đoán cho toàn bộ quần thể (population) dựa trên mẫu dữ liệu (sample).

=> Dùng các kỹ thuật như: ước lượng, kiểm định giả thuyết, hồi quy...

Ví dụ:

Thống kê mô tả: “Tuổi trung bình của 100 sinh viên là 20.4 tuổi.”

Thống kê suy luận: “Tuổi trung bình của toàn bộ sinh viên trường ước tính là 20.4 tuổi ± 0.5.”

**-Các thước đo thống kê mô tả chính**

| **Nhóm** | **Thước đo** | **Ý nghĩa** |
| --- | --- | --- |
| **Trung tâm (Central tendency)** | **Trung bình (Mean)** | **Giá trị trung tâm của dữ liệu.** |
|  | **Trung vị (Median)** | **Giá trị nằm giữa khi sắp xếp dữ liệu.** |
|  | **Mode (Giá trị xuất hiện nhiều nhất)** | **Phản ánh xu hướng phổ biến.** |
| **Phân tán (Dispersion)** | **Phương sai (Variance)** | **Độ chênh lệch bình phương giữa các giá trị và trung bình.** |
|  | **Độ lệch chuẩn (Standard Deviation)** | **Mức độ dao động quanh trung bình.** |
|  | **Phạm vi (Range)** | **Khoảng cách giữa giá trị lớn nhất và nhỏ nhất.** |

-Khi nào dùng trung vị thay vì trung bình?

Khi dữ liệu có ngoại lai (outlier) hoặc lệch mạnh (ví dụ: thu nhập, giá nhà, điểm thi).

Ví dụ:  
 Thu nhập = [10, 12, 11, 9, 300]  
 → Trung bình = 68.4 (bị lệch do 300)  
 → Trung vị = 11 (phản ánh thực tế hơn)

-Xác định phân bố của dữ liệu

Cách xác định:

Dùng biểu đồ histogram hoặc QQ-plot.

Kiểm tra hệ số lệch (skewness) và độ nhọn (kurtosis).

**Các loại phân bố phổ biến:**

| **Loại phân bố** | **Mô tả** | **Ví dụ** |
| --- | --- | --- |
| **Chuẩn (Normal)** | **Dạng chuông, đối xứng** | **Chiều cao, cân nặng** |
| **Lệch phải (Right-skewed)** | **Đuôi dài bên phải** | **Thu nhập, giá nhà** |
| **Lệch trái (Left-skewed)** | **Đuôi dài bên trái** | **Tuổi nghỉ hưu** |

-Độ lệch chuẩn và phạm vi

Phạm vi (Range) = Max – Min → Cho biết khoảng dao động tổng thể.

Độ lệch chuẩn (Standard Deviation) → Cho biết mức độ biến động trung bình quanh giá trị trung bình.

ð Độ lệch chuẩn nhỏ → dữ liệu tập trung, ổn định.  
 Độ lệch chuẩn lớn → dữ liệu phân tán mạnh.

**-Q1, Q2, Q3 trong biểu đồ hộp (Boxplot)**

| **Ký hiệu** | **Tên** | **Ý nghĩa** |
| --- | --- | --- |
| **Q1** | **Quartile 1** | **25% dữ liệu nhỏ hơn Q1** |
| **Q2 (Median)** | **Quartile 2** | **50% dữ liệu nhỏ hơn Q2** |
| **Q3** | **Quartile 3** | **75% dữ liệu nhỏ hơn Q3** |

Khoảng tứ phân vị (IQR) = Q3 - Q1 → đo độ phân tán trung bình của 50% dữ liệu giữa.

-Xử lý giá trị thiếu (Missing Values)

Một số cách phổ biến:

1. Loại bỏ (nếu số lượng nhỏ và không ảnh hưởng lớn).

2. Điền giá trị thay thế:

o Trung bình (mean)

o Trung vị (median)

o Mode (giá trị phổ biến nhất)

3. Sử dụng kỹ thuật nội suy (interpolation) hoặc mô hình dự đoán.

-Cách đọc và diễn giải biểu đồ

Histogram:

Trục X: giá trị dữ liệu

Trục Y: tần suất xuất hiện  
 → Dạng chuông = phân bố chuẩn  
 → Đuôi lệch phải/trái = dữ liệu không đối xứng

Boxplot:

Hộp = khoảng [Q1, Q3]

Đường giữa = trung vị (Q2)

“Whiskers” = phạm vi dữ liệu thông thường

Dấu chấm ngoài = outliers (ngoại lai)

-Xử lý giá trị ngoại lai (Outliers)

Một số cách:

1. Phát hiện bằng boxplot hoặc Z-score.

2. Kiểm tra nguyên nhân:

o Lỗi nhập liệu? → sửa hoặc loại bỏ

o Dữ liệu thực nhưng hiếm? → cân nhắc giữ lại

3. Cách xử lý:

o Loại bỏ nếu gây sai lệch nghiêm trọng

o Dùng trung vị thay cho trung bình

o Áp dụng IQR trimming (giới hạn giá trị trong [Q1–1.5×IQR, Q3+1.5×IQR])

## 1.2. XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU

### 1.2.1. Ôn tập lý thuyết

Phần này trình bày các khái niệm lý thuyết cốt lõi về **trực quan hóa dữ liệu**, một công cụ không thể thiếu trong quá trình **Phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA)**.

#### Vai trò và Tầm quan trọng của Trực quan hóa dữ liệu trong EDA

Trực quan hóa dữ liệu là **nghệ thuật và khoa học biểu diễn dữ liệu dưới dạng đồ họa** như biểu đồ, đồ thị, bản đồ. Vai trò chính của nó là **chuyển đổi các tập dữ liệu phức tạp** thành hình ảnh trực quan, dễ hiểu.

Trong EDA, trực quan hóa đặc biệt quan trọng vì:

* **Hiểu nhanh về dữ liệu**: Biểu đồ như histogram, boxplot cung cấp cái nhìn tổng quan tức thì về phân bố dữ liệu.
* **Phát hiện quy luật và xu hướng**: Scatter plot giúp nhận diện mối tương quan; line chart thể hiện xu hướng theo thời gian.
* **Xác định giá trị ngoại lai (Outliers)**: Boxplot, scatter plot hỗ trợ phát hiện dữ liệu bất thường.
* **Truyền đạt thông tin hiệu quả**: Hình ảnh trực quan giúp chia sẻ kết quả EDA dễ hiểu hơn so với bảng số liệu thô.

#### Các loại biểu đồ phổ biến và Trường hợp sử dụng

* **Histogram (Biểu đồ tần suất):**

*Mục đích sử dụng*: Được dùng để xem xét sự phân bố của một biến số liên tục. Nó cho biết tần suất xuất hiện của các giá trị trong từng khoảng dữ liệu.

*Trường hợp áp dụng:* Phân tích sự phân bố của độ tuổi khách hàng, điểm thi của sinh viên, hoặc giá nhà trong một khu vực.

* **Scatter Plot (Biểu đồ phân tán):**

*Mục đích sử dụng:* Được dùng để khám phá mối quan hệ giữa hai biến số liên tục.

*Trường hợp áp dụng:* Tìm hiểu mối tương quan giữa chi tiêu quảng cáo và doanh thu; mối quan hệ giữa diện tích nhà và giá bán.

* **Boxplot (Biểu đồ hộp):**

*Mục đích sử dụng:* Tóm tắt sự phân bố của một biến số qua các đại lượng thống kê (trung vị, tứ phân vị), phát hiện các giá trị ngoại lai, và đặc biệt hữu ích khi so sánh sự phân bố của một biến số qua nhiều nhóm khác nhau.

*Trường hợp áp dụng:* So sánh phân bố mức lương giữa các phòng ban trong một công ty; kiểm tra sự khác biệt về giá phòng giữa khách sạn thành phố và khách sạn nghỉ dưỡng.

* **Bar Chart (Biểu đồ cột):**

*Mục đích sử dụng:* Dùng để so sánh giá trị (ví dụ: trung bình, tổng) hoặc tần suất của các nhóm (biến phân loại) riêng biệt.

*Trường hợp áp dụng:* So sánh số lượng sản phẩm bán được theo từng danh mục; thống kê số lượng khách hàng đến từ các quốc gia khác nhau.

#### Cách chọn loại biểu đồ phù hợp

* **Phân tích một biến (Univariate Analysis)**

Dữ liệu số: Histogram, Boxplot

Dữ liệu phân loại: Bar Chart, Pie Chart

* **Phân tích hai biến (Bivariate Analysis)**

Số ↔ Số: **Scatter Plot**

Số ↔ Phân loại: **Boxplot, Violin Plot**

Phân loại ↔ Phân loại: **Grouped Bar Chart, Heatmap**

* **Phân tích dữ liệu chuỗi thời gian (Time-series Data)**

Sử dụng **Line Chart** để thể hiện sự thay đổi liên tục theo thời gian.

#### Sự khác biệt giữa các thư viện trực quan hóa trong Python

| **Thư viện** | **Đặc điểm chính** | **Ưu điểm** |
| --- | --- | --- |
| **Matplotlib** | Thư viện nền tảng, kiểm soát chi tiết ở mức low-level | Linh hoạt, tùy biến cao, nền tảng cho nhiều thư viện khác |
| **Seaborn** | Xây dựng trên Matplotlib, tối ưu cho biểu đồ thống kê | Cú pháp đơn giản, biểu đồ đẹp mặc định, tích hợp tốt với Pandas |
| **Plotly** | Tập trung vào biểu đồ tương tác (interactive) | Hỗ trợ zoom, hover thông tin; phù hợp cho dashboard và báo cáo web |

#### Các nguyên tắc thiết kế biểu đồ hiệu quả

* **Rõ ràng và Tối giản**: Tập trung vào thông điệp chính, loại bỏ yếu tố rườm rà.
* **Chọn đúng loại biểu đồ**: Phù hợp với loại dữ liệu và mục tiêu phân tích.
* **Ghi nhãn đầy đủ**: Có tiêu đề, nhãn trục, đơn vị đo, và chú thích khi cần.
* **Tỷ lệ trục hợp lý**: Nên bắt đầu từ 0 để tránh sai lệch khi so sánh.
* **Sử dụng màu sắc có chủ đích**: Dùng màu để phân biệt nhóm hoặc làm nổi bật thông tin quan trọng; tránh quá nhiều màu gây rối mắt.

1. **Code mẫu tạo biểu đồ đơn giản bằng Matplotlib**

**Khái niệm cơ bản về Histogram:**

**Histogram** (biểu đồ tần suất) là một công cụ trực quan hóa mạnh mẽ dùng để thể hiện sự phân bố của một tập dữ liệu số liên tục. Nó hoạt động bằng cách chia toàn bộ phạm vi giá trị của dữ liệu thành một loạt các khoảng (gọi là bins) có độ rộng bằng nhau, sau đó đếm xem có bao nhiêu điểm dữ liệu rơi vào mỗi khoảng đó. Chiều cao của mỗi cột trong histogram tương ứng với số lượng (tần suất) dữ liệu trong khoảng đó.

**Khi nào nên sử dụng Histogram?**

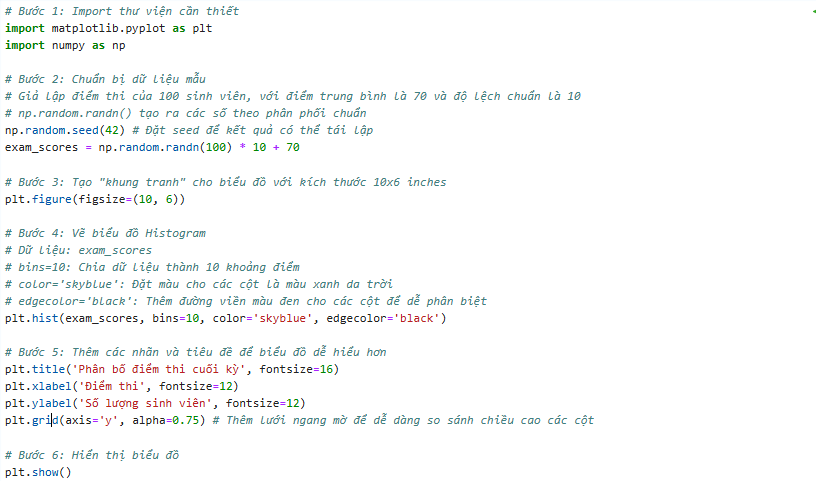
Bạn nên sử dụng histogram khi muốn trả lời các câu hỏi như:

* Dữ liệu của tôi trông như thế nào? Nó có đối xứng không?
* Đa số các giá trị tập trung ở đâu?
* Dữ liệu có bị lệch về phía giá trị cao hay thấp không?
* Có các đỉnh (peaks) nào trong dữ liệu không?

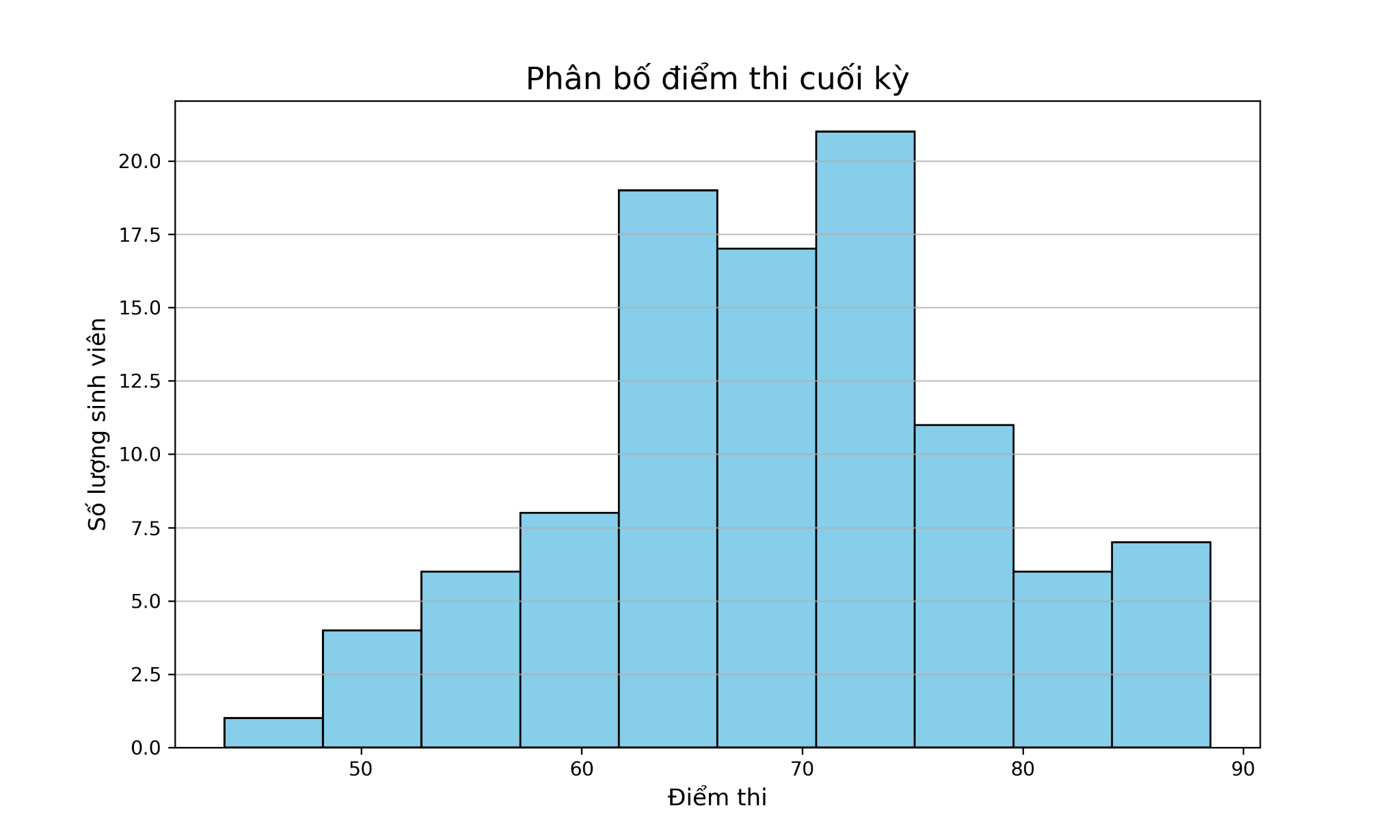
**Các bước tạo Histogram với Matplotlib:**

* Import thư viện: Đầu tiên, bạn cần import module pyplot từ thư viện matplotlib.
* Chuẩn bị dữ liệu: Tạo một danh sách (list) hoặc một Series trong Pandas chứa các giá trị số mà bạn muốn phân tích.
* Tạo hình (Figure): (Tùy chọn nhưng được khuyến khích) Sử dụng plt.figure() để tạo một "khung tranh" cho biểu đồ của bạn. Điều này cho phép bạn tùy chỉnh kích thước và các thuộc tính khác.
* Vẽ Histogram: Sử dụng hàm plt.hist(). Tham số quan trọng nhất là dữ liệu đầu vào và số lượng bins.
* Thêm nhãn và tiêu đề: Sử dụng các hàm plt.title(), plt.xlabel(), plt.ylabel() để làm cho biểu đồ trở nên rõ ràng và dễ hiểu.
* Hiển thị biểu đồ: Cuối cùng, gọi plt.show() để hiển thị biểu đồ.

**Đoạn code mẫu**

****

**Kết quả**

****

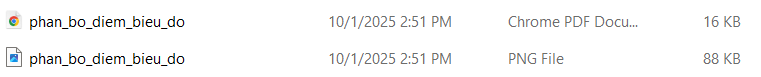
1. **Cách xuất biểu đồ ra các định dạng file**

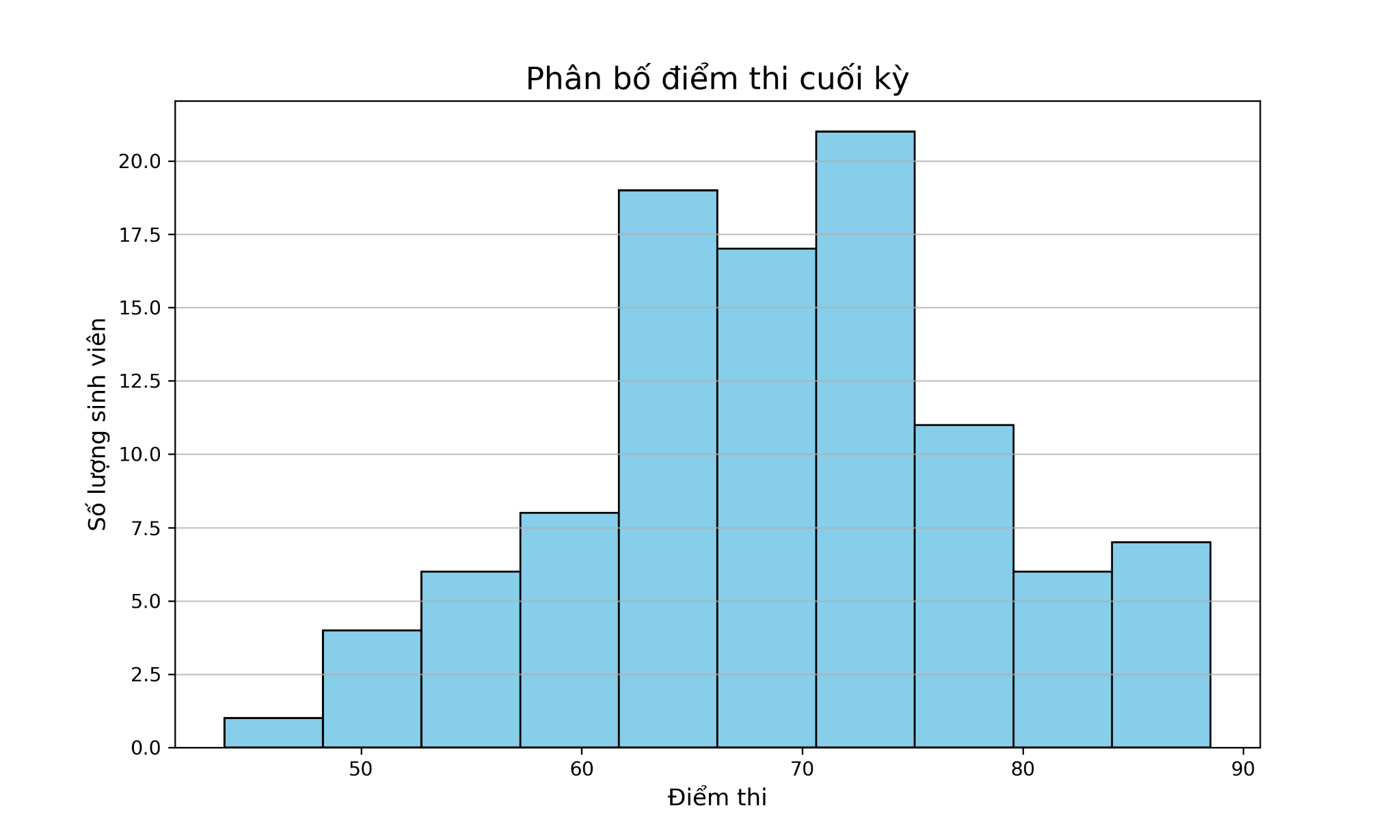
Trong Python, thư viện Matplotlib cho phép lưu biểu đồ thành nhiều định dạng file khác nhau như PNG, PDF, SVG bằng cách sử dụng hàm plt.savefig(). Lệnh này cần được gọi trước khi hiển thị biểu đồ bằng plt.show().

Ví dụ code lưu biểu đồ ở mục 6:



Kết quả:





1.2.1 Bài mẫu.

1.2.2 Bài thực hành 1

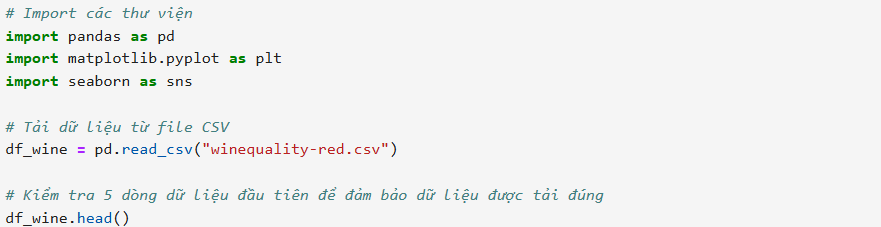
***Thực hiện trực quan hóa dữ liệu trên tập dữ liệu về phân loại chất lượng rượu đỏ. Dữ liệu lấy tại***

[***https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification***](https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification)

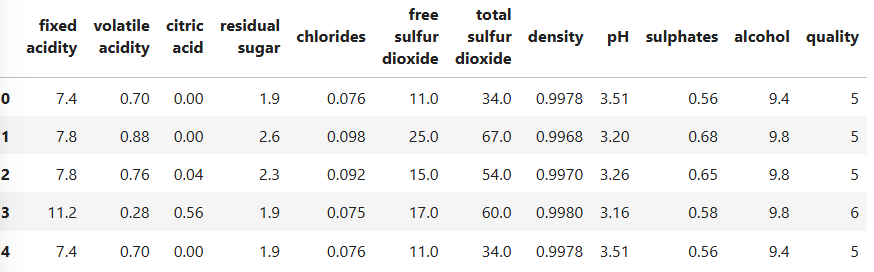
***Bài làm***

### 1. Chuẩn bị dữ liệu cho trực quan hóa dữ liệu

Bước đầu tiên là tải dữ liệu và các thư viện cần thiết vào môi trường làm việc



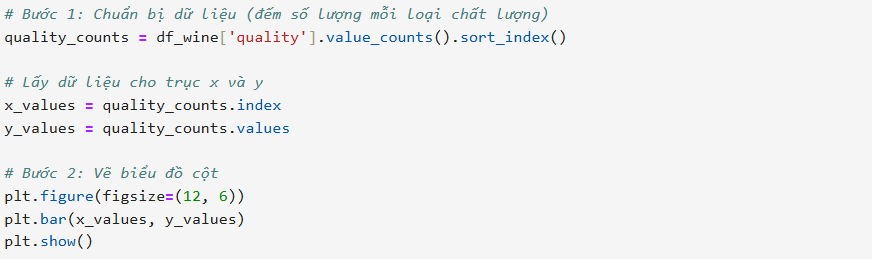
.

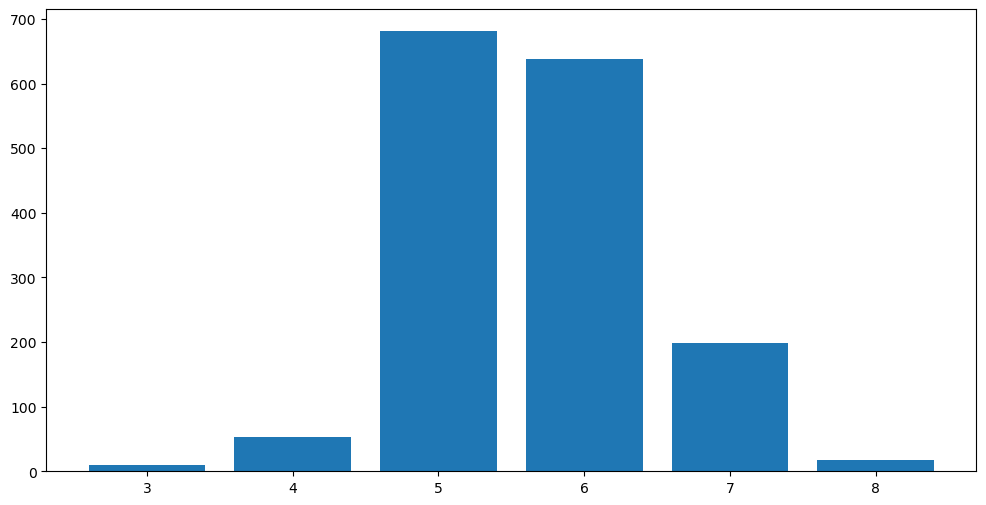
u

### 2.Trực quan hóa dữ liệu với thư viện Matplotlib

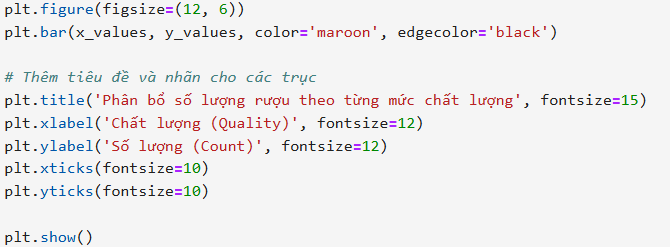
Chúng ta sẽ sử dụng Matplotlib để vẽ biểu đồ thể hiện số lượng rượu vang cho mỗi mức chất lượng.

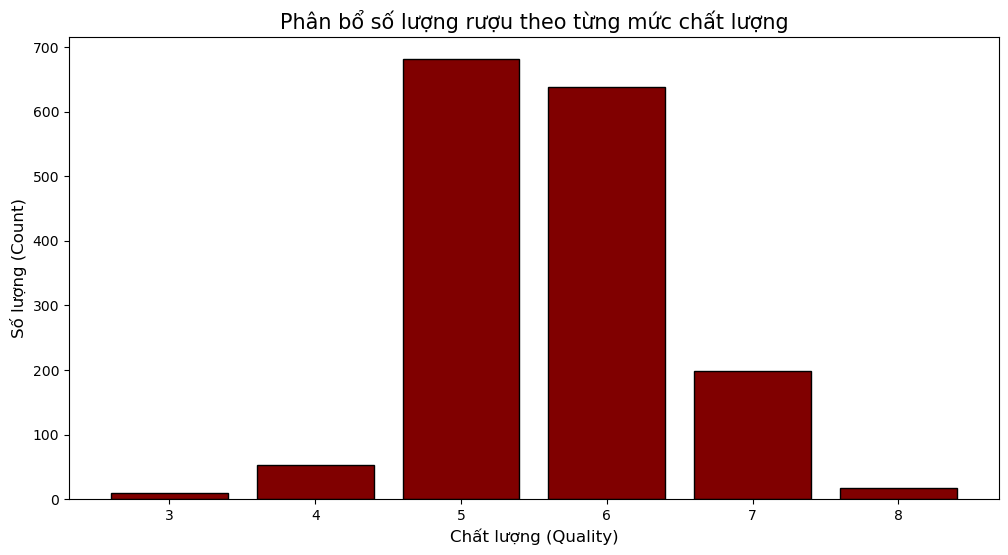
Case 1: Basic (Vẽ biểu đồ cơ bản)



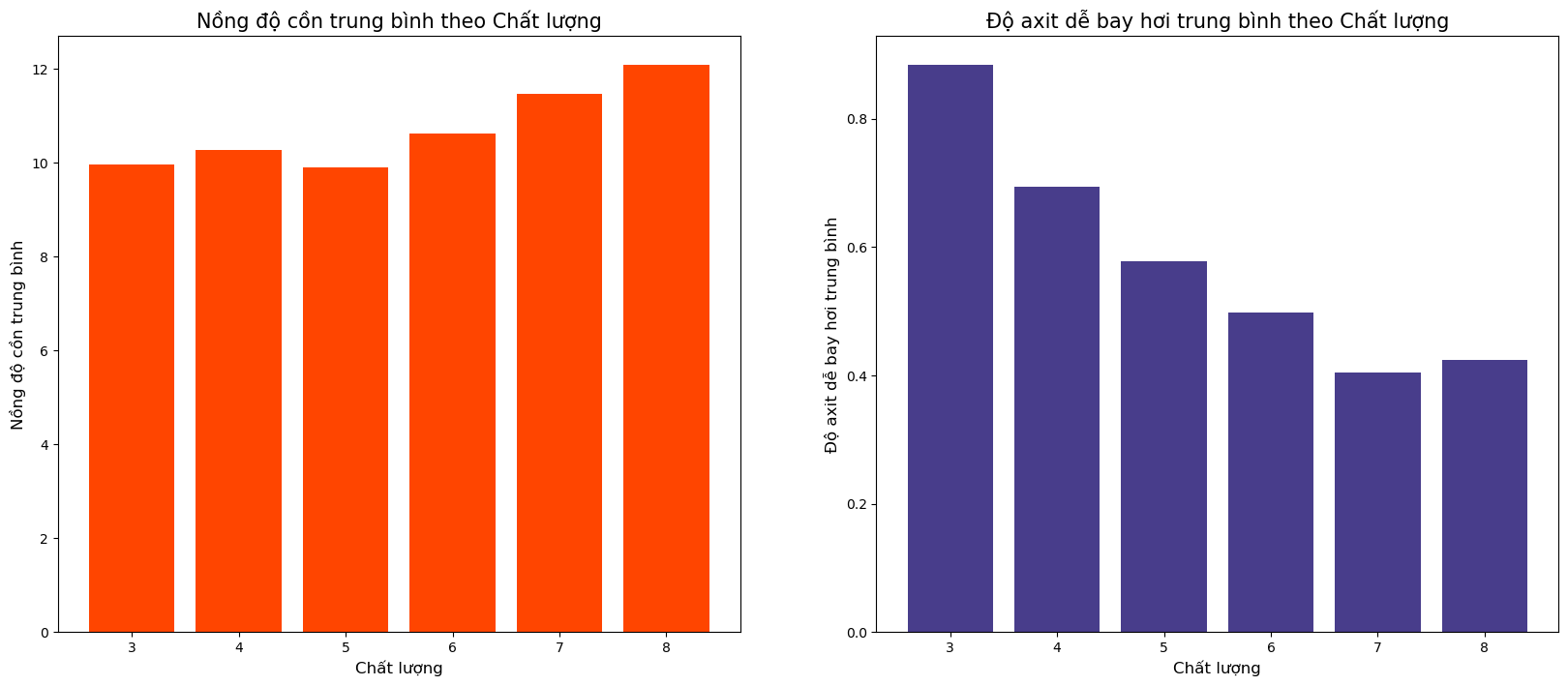
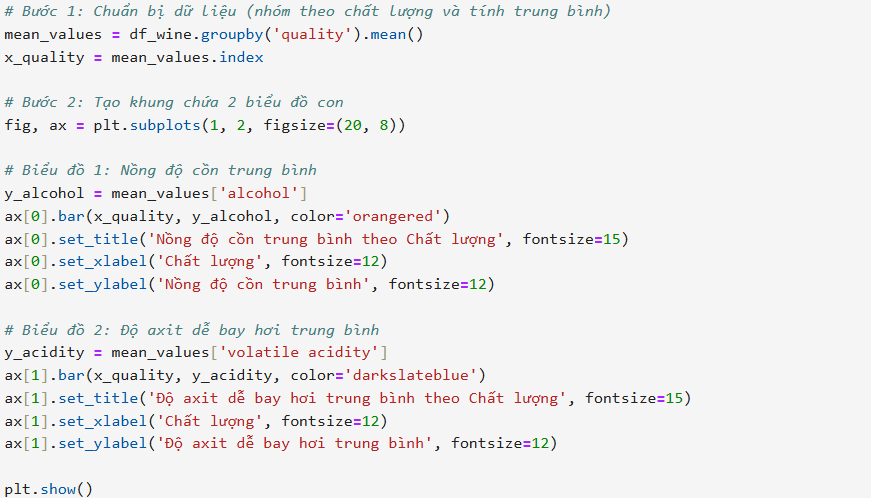


Case 2: Advanced 1 (Thêm thông tin cho biểu đồ)



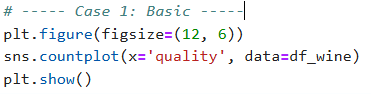


Case 3: Xem nhiều góc nhìn cùng lúc

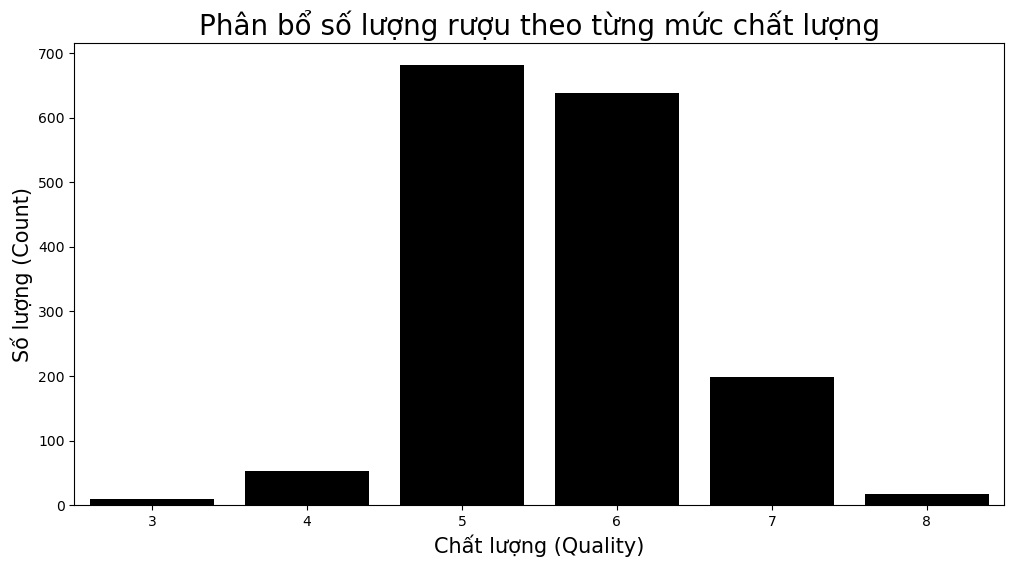


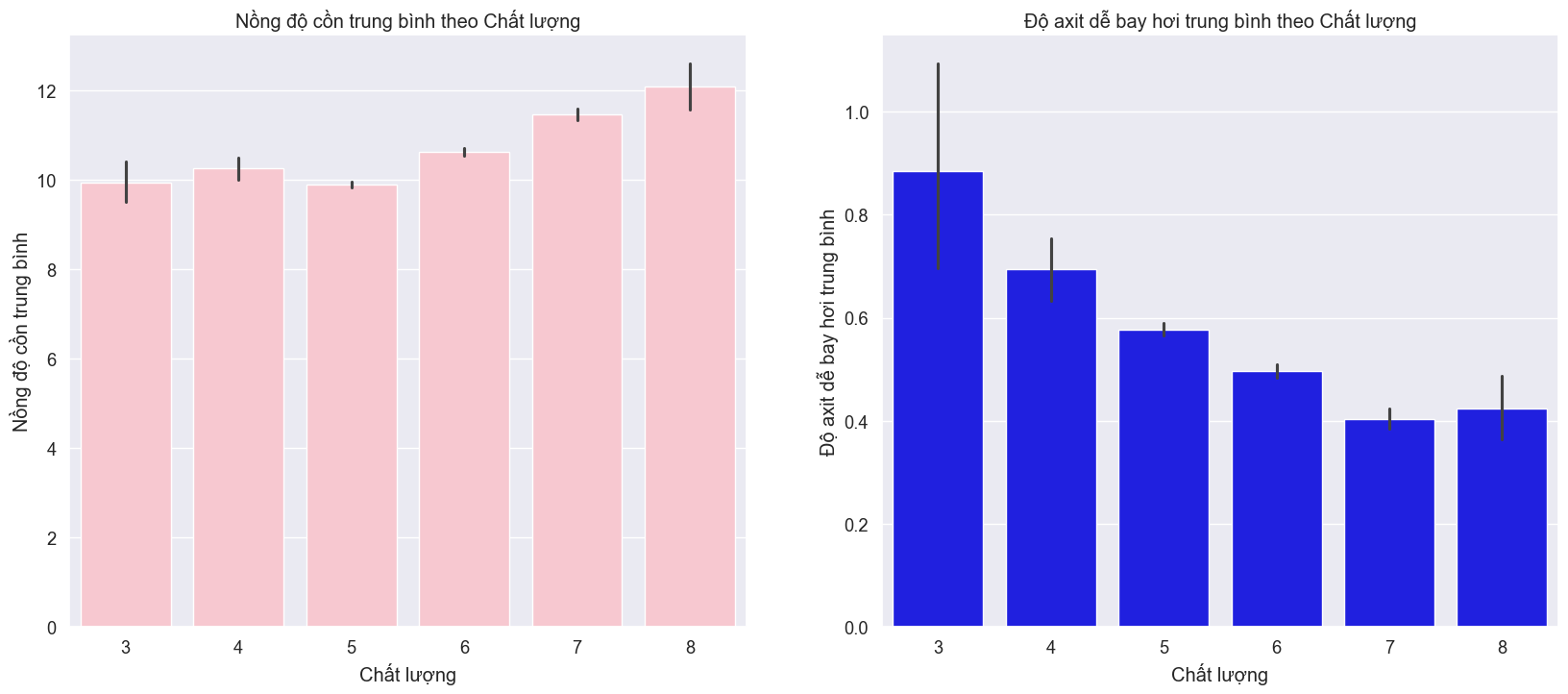
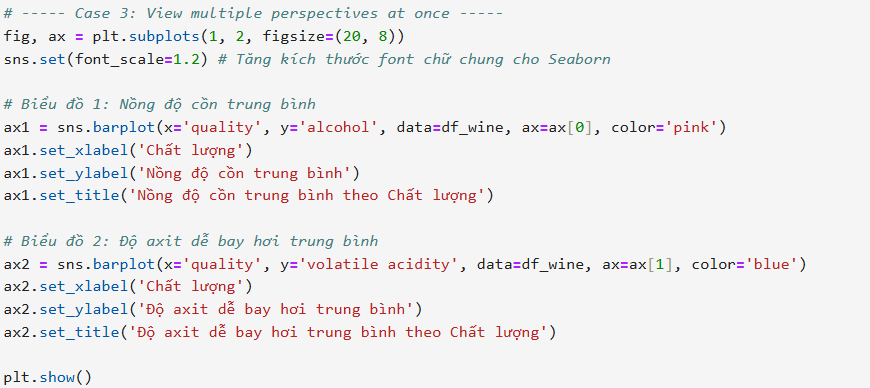
### 3. Trực quan hóa dữ liệu với thư viện Seaborn

Bây giờ, chúng ta sẽ thực hiện lại các tác vụ tương tự bằng Seaborn để so sánh.









1.2.3 Bài thực hành 2

***Thực hiện trực quan hóa dữ liệu trên tập dữ liệu về bệnh tiểu đường.Dữ liệu lấy tại***

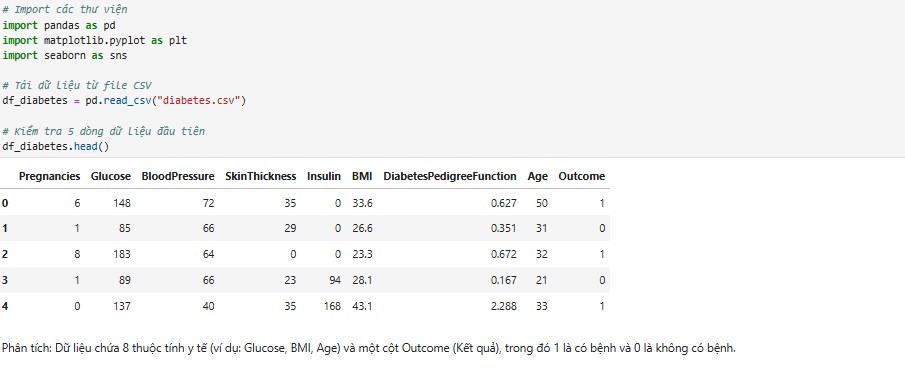
[***https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-****prediction-0-906*](https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906)

***Bài Làm***

*Mục tiêu: Thực hành các kỹ thuật trực quan hóa để phân tích các yếu tố y tế liên quan đến bệnh tiểu đường.*

### *1. Chuẩn bị dữ liệu cho trực quan hóa dữ liệu*

*Tải dữ liệu và các thư viện cần thiết.*

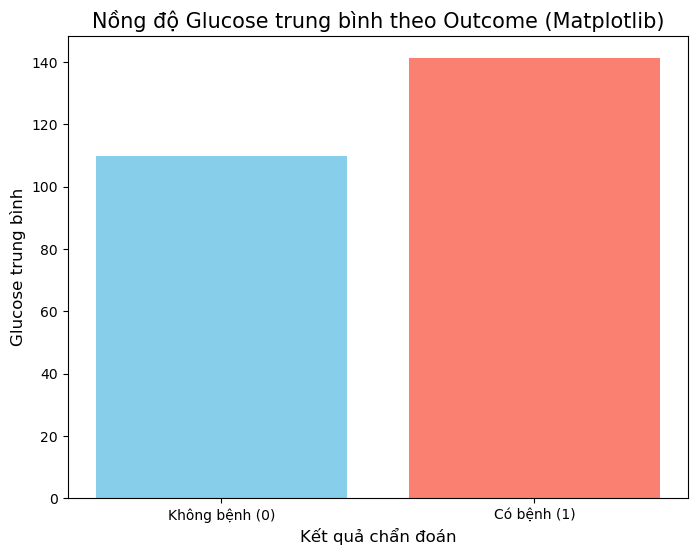
**

### 2. Trực quan hóa dữ liệu với thư viện Matplotlib

2.1. Case 1 & 2: Vẽ biểu đồ cơ bản và nâng cao

Thay vì so sánh giá, chúng ta hãy so sánh giá trị trung bình của các chỉ số y tế giữa hai nhóm: người có bệnh (Outcome=1) và người không có bệnh (Outcome=0).  
Trước tiên, chúng ta cần chuẩn bị dữ liệu bằng cách nhóm và tính trung bình.





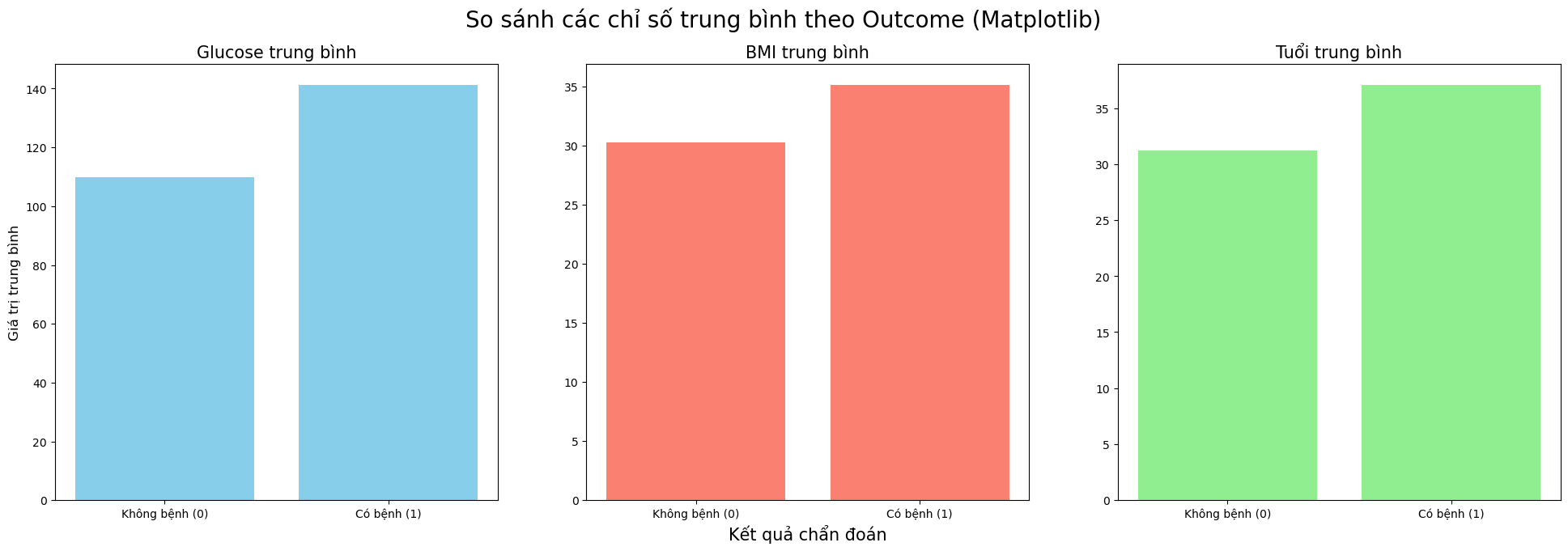
*Phân tích:*

*Chúng ta đã sử dụng groupby().mean() để chuẩn bị dữ liệu, một bước cần thiết khi dùng Matplotlib cho loại phân tích này.*

*Biểu đồ cho thấy rõ ràng nhóm người có bệnh (Outcome=1) có nồng độ Glucose trung bình cao hơn đáng kể so với nhóm không có bệnh.*

2.2. Case 3: Xem nhiều góc nhìn cùng lúc

Chúng ta sẽ dùng subplots để so sánh đồng thời nhiều chỉ số quan trọng giữa hai nhóm. Ví dụ: Glucose, BMI, và Age.

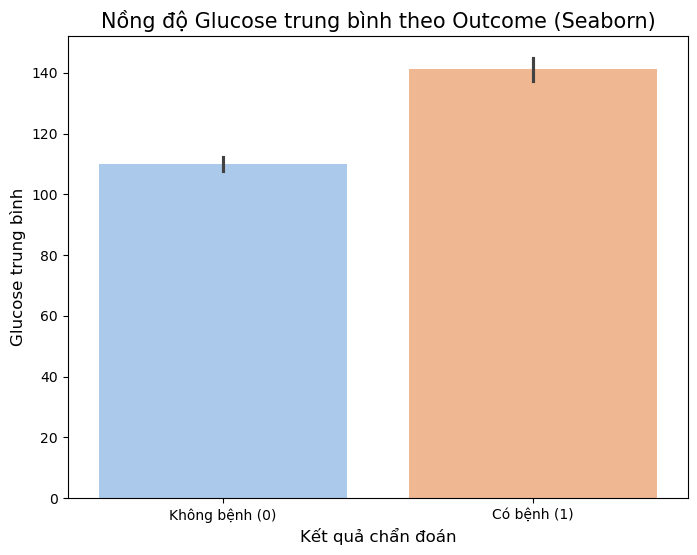


### 3. Trực quan hóa dữ liệu với thư viện Seaborn

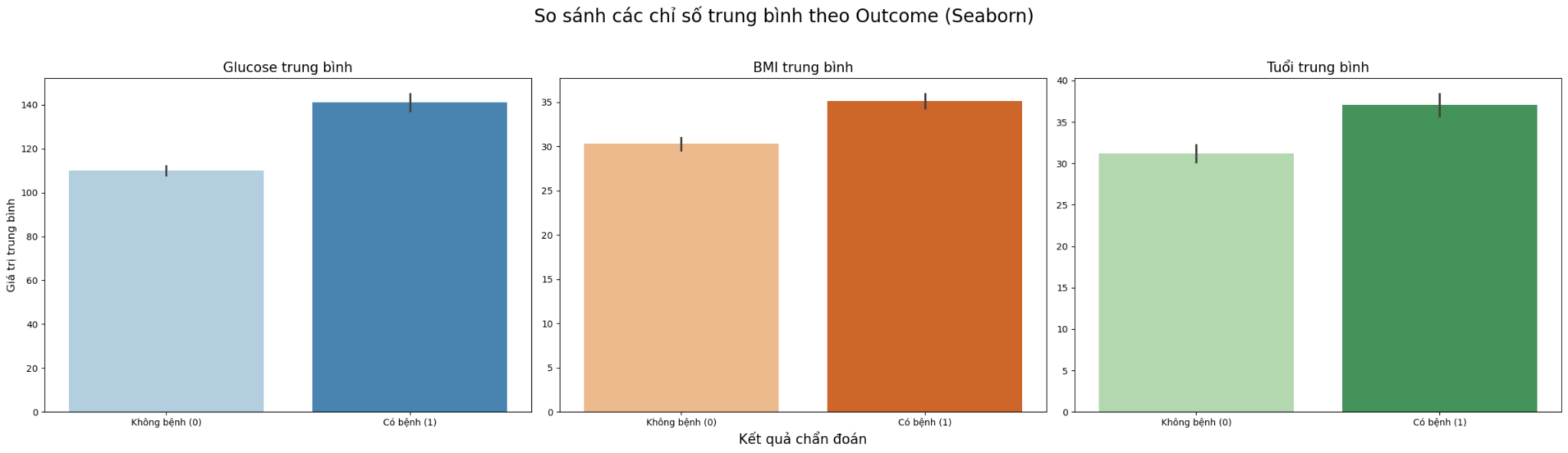
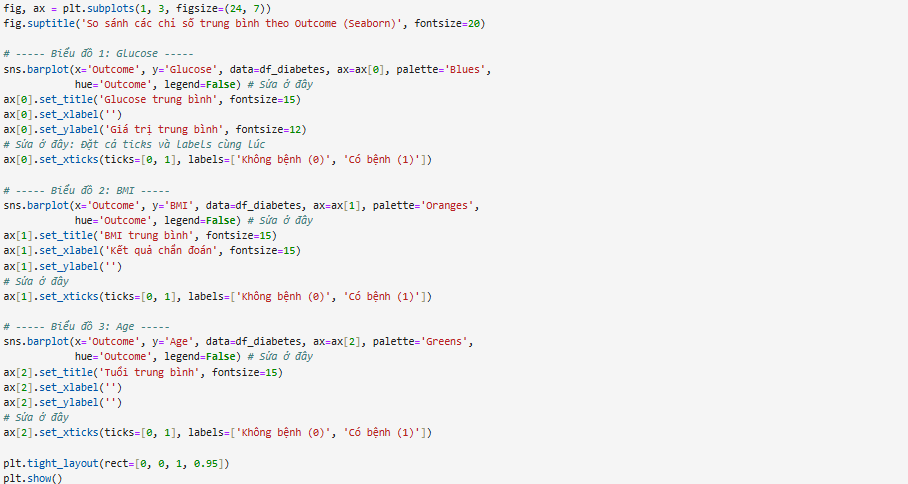
Bây giờ, chúng ta sẽ thực hiện lại các tác vụ trên bằng Seaborn.

Case 1 & 2: Vẽ biểu đồ cơ bản và nâng cao





Case 3: Xem nhiều biểu đồ cùng lúc



### 

### 

### 

### 

### 

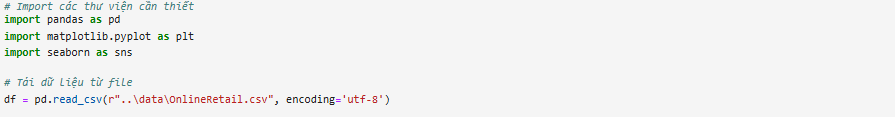
### 

### *Thực hiện EDA trên tập dữ liệu mua sắm tại siêu thị.Tập dữ liệu lấy từ: https://www.kaggle.com/code/rajatkumar30/eda-online-retail*

### Bài làm

Mục tiêu biến dữ liệu giao dịch thô thành những hiểu biết kinh doanh có giá trị, giúp doanh nghiệp trả lời các câu hỏi quan trọng về khách hàng, sản phẩm, và xu hướng thị trường để đưa ra các quyết định tốt hơn.

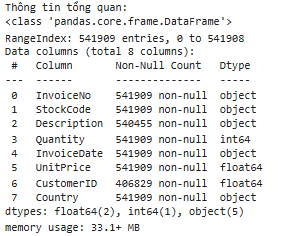
### Bước 1: Tải và Khám phá dữ liệu ban đầu[¶](http://localhost:8888/notebooks/Desktop/BTL01/1.2.3%20B%C3%A0i%20t%E1%BA%ADp%20th%E1%BB%B1c%20h%C3%A0nh%202/code/code-1.2.3-EDA-OnlineRetail.ipynb?#B%C6%B0%E1%BB%9Bc-1:-T%E1%BA%A3i-v%C3%A0-Kh%C3%A1m-ph%C3%A1-d%E1%BB%AF-li%E1%BB%87u-ban-%C4%91%E1%BA%A7u)



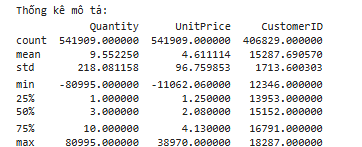












Phân tích ban đầu:

- Cho thấy có rất nhiều giá trị thiếu (`null`) trong cột `Description` và đặc biệt là `CustomerID`.

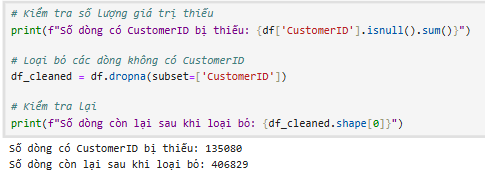
- `.describe()` cho thấy `Quantity` có giá trị tối thiểu là số âm (ví dụ: -80995), và `UnitPrice` có giá trị tối thiểu là 0. Đây là những điểm bất thường cần được xử lý.

### Bước 2: Làm sạch và Tiền xử lý dữ liệu

Đây là bước cốt lõi để đảm bảo dữ liệu của chúng ta đáng tin cậy

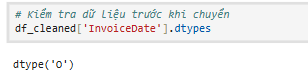
**2.1. Xử lý giá trị bị thiếu**

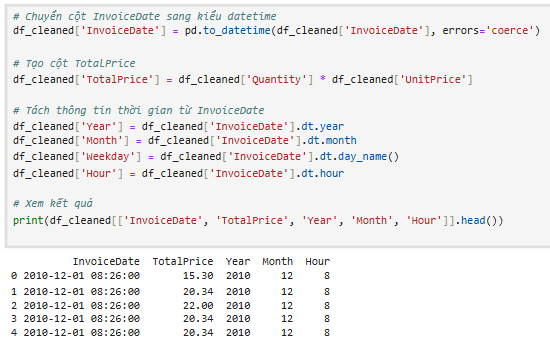
Chúng ta sẽ tập trung vào cột CustomerID vì các phân tích về khách hàng sẽ không thể thực hiện nếu thiếu thông tin này.



**2.3. Chuyển đổi kiểu dữ liệu**

Chuyển đổi CustomerID và InvoiceDate về đúng định dạng.





### Bước 4: Trực quan hóa và Phân tích

Bây giờ dữ liệu đã sạch, chúng ta sẽ bắt đầu đặt câu hỏi và tìm câu trả lời bằng biểu đồ.

**4.1. Phân tích tổng quan về Doanh thu và Khách hàng**

Câu hỏi: Những quốc gia nào đóng góp nhiều doanh thu nhất?



### 

### Nhận xét:

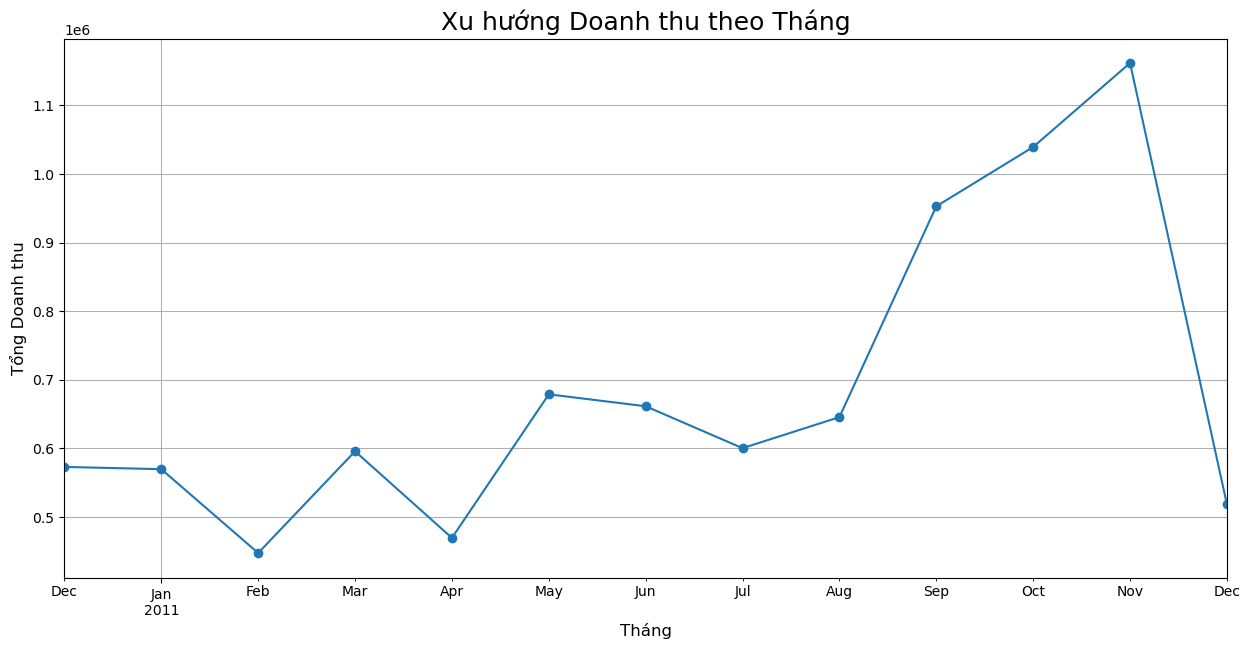
### - United Kingdom chiếm danh thu cao nhất, tiếp đến là Netherlands,EIRE,Germany

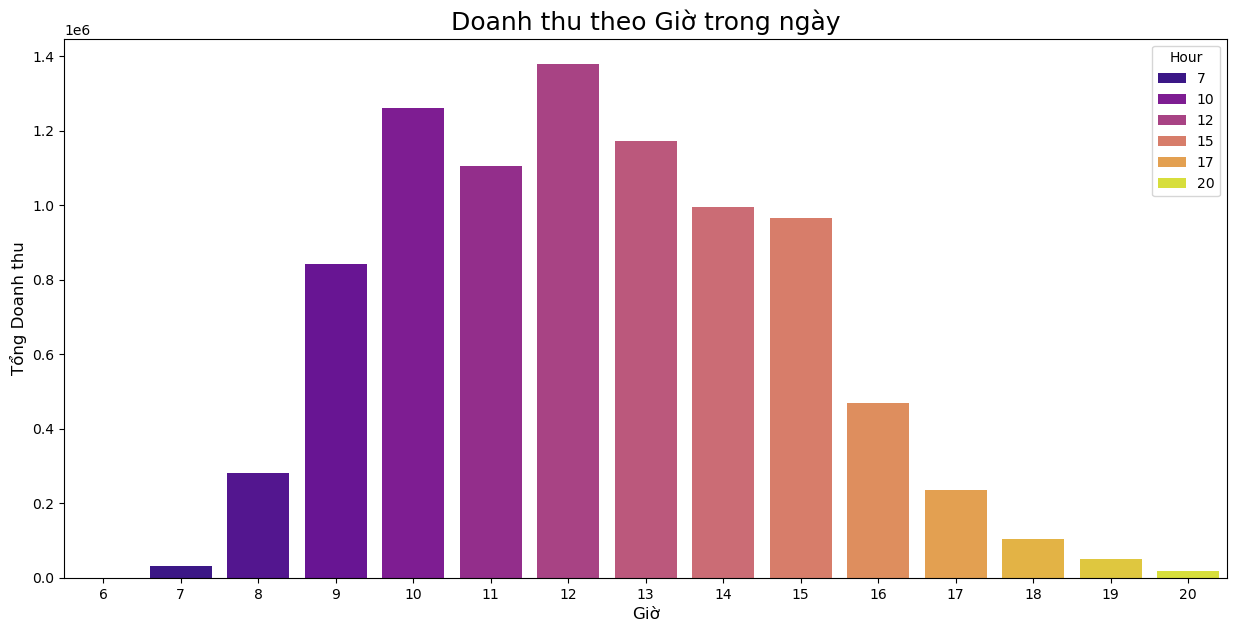
### - Thấp nhất là Sweden

**4.2. Phân tích theo Thời gian**

Câu hỏi: Xu hướng doanh thu theo tháng và theo giờ trong ngày như thế nào?



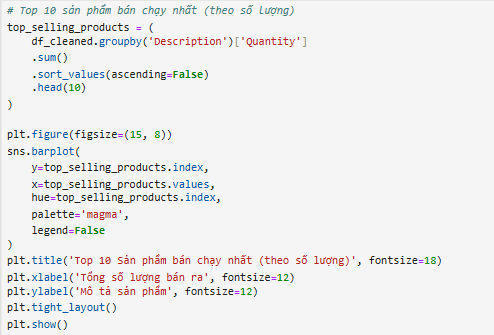


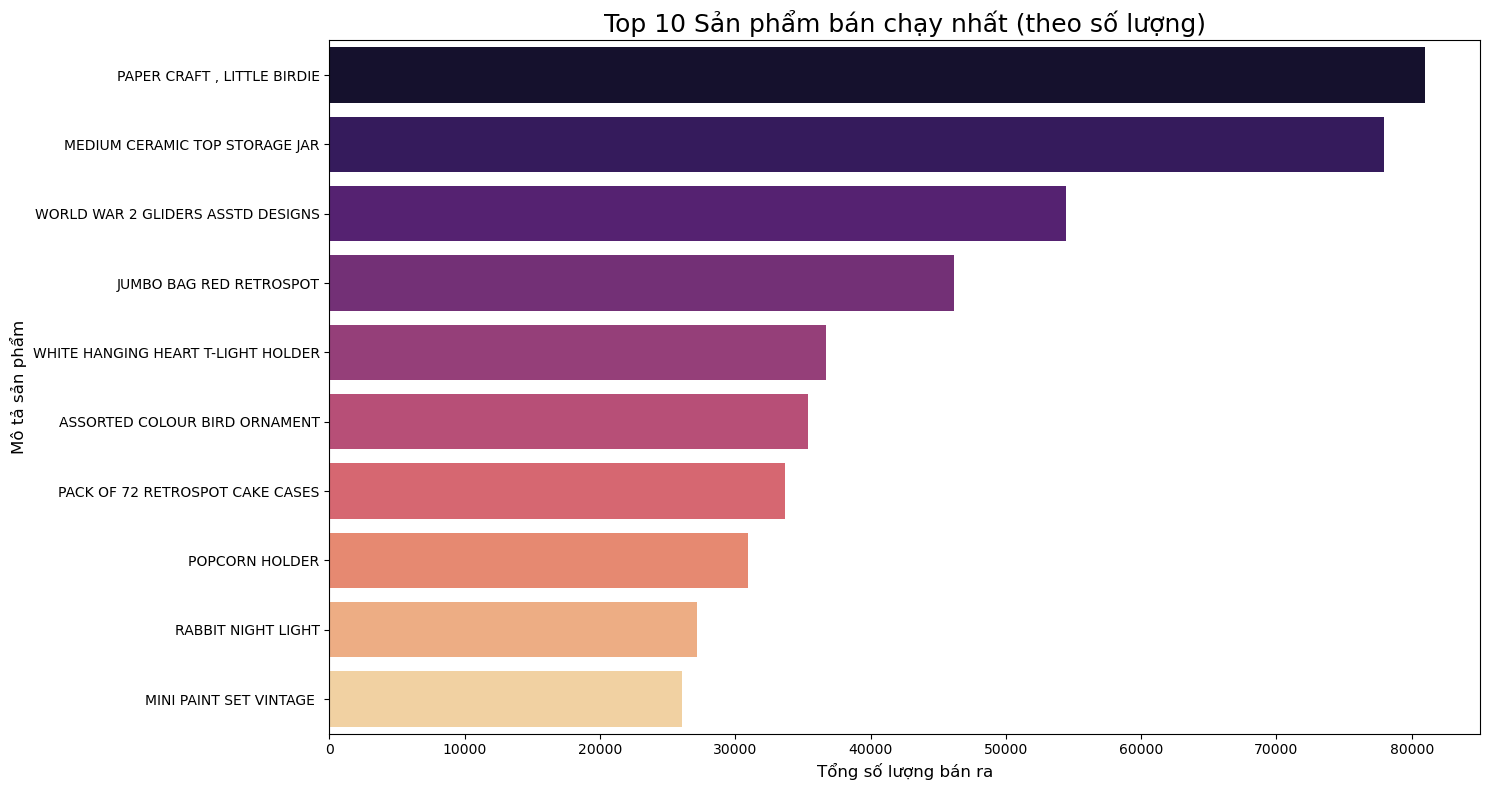


*Nhận xét*- Theo tháng: tăng mạnh vào cuối năm đặc biệt từ tháng 9 đến tháng 11 do dịp lễ.  
- Theo giờ : khung giờ cao nhất từ 10 giờ sáng đến 3h chiều, doanh thu rất thấp vào sáng sớm và chiều muộn.

**4.3. Phân tích Sản phẩm**

Câu hỏi: Những sản phẩm nào được bán chạy nhất?

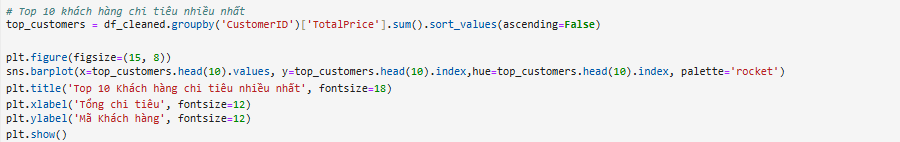


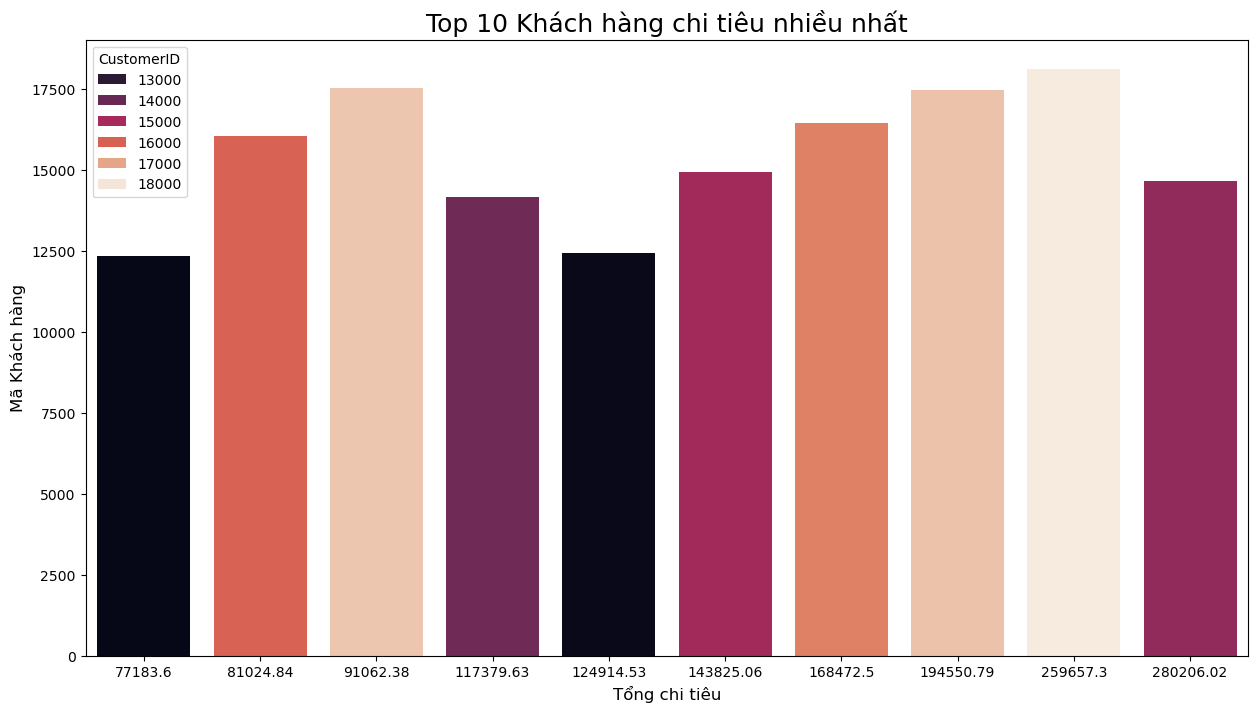


**Phân tích**Các sản phẩm bán chạy nhất chủ yếu là hàng trang trí, đồ gia dụng nhỏ, và đồ thủ công.  
Đây có khả năng là các mặt hàng có đơn giá thấp, khuyến khích khách hàng mua với số lượng lớn để làm quà tặng hoặc bán sỉ.  
Nhu cầu về các sản phẩm lưu trữ, đóng gói (như STORAGE JAR, JUMBO BAG) là rất cao.

**4.4. Phân tích Khách hàng (Sơ bộ)**

Câu hỏi: Những khách hàng nào chi tiêu nhiều nhất?





**Nhận xét:**

Có một nhóm nhỏ khách hàng đóng góp một phần rất lớn vào tổng doanh thu. Khách hàng có ID 14646.0, 18102.0, 17450.0 là những khách hàng giá trị nhất.

### Bước 5: Kết luận & Tổng kết các Insight chính

Qua quá trình Phân tích Dữ liệu Khám phá (EDA) trên tập dữ liệu bán lẻ trực tuyến, chúng tôi đã rút ra được những kết luận và insight quan trọng sau về hoạt động kinh doanh của doanh nghiệp:

**1.Chất lượng dữ liệu** là yếu tố then chốt: Dữ liệu gốc chứa một lượng lớn các giao dịch bị hủy (số lượng âm) và các dòng thiếu thông tin khách hàng (CustomerID). Việc làm sạch và tiền xử lý dữ liệu ở Bước 2 là bước bắt buộc để đảm bảo các phân tích sau này có độ chính xác và tin cậy.

**2. Đặc điểm thị trường và khách hàng:**

- Thị trường cốt lõi: Vương quốc Anh (United Kingdom) là thị trường lớn nhất, chiếm phần lớn áp đảo trong tổng doanh thu. Các thị trường tiềm năng tiếp theo ở châu Âu bao gồm Hà Lan, Ireland, Đức và Pháp.

- Phân bổ khách hàng: Số lượng giao dịch giữa khách hàng Nam/Nữ và giữa khách hàng Thành viên/Vãng lai là khá cân bằng.

**3. Xu hướng kinh doanh theo thời gian:**

- Xu hướng tháng: Doanh thu có xu hướng tăng trưởng mạnh vào cuối năm, đạt đỉnh vào tháng 11. Đây có thể là mùa mua sắm cao điểm cho các dịp lễ, đòi hỏi sự chuẩn bị kỹ lưỡng về hàng hóa và marketing.

- Xu hướng giờ: Khung giờ mua sắm sôi động nhất trong ngày là từ 10 giờ sáng đến 3 giờ chiều. Đây là thông tin quan trọng để tối ưu hóa việc vận hành website và phân bổ nhân sự hỗ trợ.

**4. Phân tích Sản phẩm và Khách hàng:**

- Sản phẩm phổ biến: Các sản phẩm bán chạy nhất về mặt số lượng chủ yếu là các mặt hàng trang trí, quà tặng có giá trị thấp. Điều này cho thấy chiến lược sản phẩm "mồi" hoặc bán sỉ có thể đang hoạt động hiệu quả.

- Khách hàng giá trị: Có một nhóm nhỏ khách hàng (CustomerID như 14646.0, 18102.0) đóng góp một phần rất lớn vào tổng doanh thu. Đây là những khách hàng VIP cực kỳ quan trọng cần được chăm sóc đặc biệt.

Nhận xét sau 2 bài thực hành

**1. Matplotlib**

* **Đặc điểm**: Thư viện nền tảng, mạnh mẽ và cực kỳ linh hoạt, cho phép kiểm soát đến từng chi tiết nhỏ nhất của biểu đồ. Tuy nhiên, cú pháp thường dài dòng và yêu cầu người dùng phải tự xử lý dữ liệu trước khi vẽ.
* **Điểm mạnh**: Tùy biến sâu, linh hoạt.
* **Điểm yếu**: Cú pháp phức tạp, biểu đồ mặc định không đẹp.

**2. Seaborn**

* **Đặc điểm**: Thư viện cấp cao được xây dựng trên Matplotlib, hỗ trợ vẽ biểu đồ thống kê đẹp mắt một cách nhanh chóng. Seaborn tích hợp tốt với Pandas DataFrame và tự động thực hiện nhiều phép tính thống kê phức tạp.
* **Điểm mạnh**:
* Cú pháp đơn giản, dễ sử dụng.
* Biểu đồ mặc định đẹp, trực quan.
* Lý tưởng cho phân tích dữ liệu khám phá (EDA).
* **Điểm yếu**: Ít linh hoạt hơn khi cần các tùy chỉnh đặc biệt.

**3. Kết luận**

* **Seaborn** phù hợp để vẽ nhanh các biểu đồ chính trong giai đoạn EDA.
* **Matplotlib** phù hợp để tinh chỉnh chi tiết và kiểm soát biểu đồ ở mức độ cao.
* Hai thư viện **bổ trợ cho nhau** thay vì thay thế nhau.

## 1.3. Phân tích đơn biến và hai biến

### 1.3.1. Ôn lý thuyết

Phân tích đơn biến (Univariate Analysis):

Định nghĩa: Là loại phân tích khám phá dữ liệu (EDA) tập trung vào việc mô tả và tóm tắt một biến duy nhất tại một thời điểm. Mục tiêu là hiểu sự phân bố, xu hướng trung tâm và sự phân tán của biến đó.

Khác biệt với phân tích hai biến: Phân tích đơn biến chỉ xem xét đặc điểm của từng biến riêng lẻ mà không xem xét mối quan hệ giữa các biến.

Các thước đo thống kê thường được sử dụng:

Trung bình (Mean): Giá trị trung bình cộng của tất cả các điểm dữ liệu.

Trung vị (Median): Giá trị nằm chính giữa tập dữ liệu khi đã được sắp xếp. Ít bị ảnh hưởng bởi các giá trị ngoại lai (outliers).

Mode: Giá trị xuất hiện nhiều nhất trong tập dữ liệu.

Độ lệch chuẩn (Standard Deviation): Đo lường mức độ phân tán của dữ liệu so với giá trị trung bình. Độ lệch chuẩn lớn cho thấy dữ liệu phân tán rộng.

Phương sai (Variance): Bình phương của độ lệch chuẩn, cũng đo lường sự phân tán.

Phần tư vị (Quartiles) và Khoảng cách giữa các phần tư vị (Interquartile Range - IQR): Chia dữ liệu thành bốn phần bằng nhau và đo lường sự phân tán của 50% dữ liệu ở giữa.

Phạm vi (Range): Sự khác biệt giữa giá trị lớn nhất và nhỏ nhất.

Độ xiên (Skewness): Đo lường sự bất đối xứng của phân phối dữ liệu.

Độ nhọn (Kurtosis): Đo lường "độ phồng" hoặc "độ phẳng" của phân phối dữ liệu so với phân phối chuẩn.

Phân tích hai biến (Bivariate Analysis):

Định nghĩa: Là loại phân tích khám phá dữ liệu tập trung vào việc khám phá mối quan hệ giữa hai biến khác nhau. Mục tiêu là hiểu cách một biến thay đổi khi biến kia thay đổi.

Xác định mối quan hệ giữa hai biến:

Tương quan (Correlation): Đo lường mức độ và hướng của mối quan hệ tuyến tính giữa hai biến số. Hệ số tương quan Pearson (r) là thước đo phổ biến nhất, có giá trị từ -1 đến +1.

r = +1: Tương quan dương hoàn hảo.

r = -1: Tương quan âm hoàn hảo.

r = 0: Không có mối quan hệ tuyến tính.

Nhân quả (Causation): Xác định xem một biến có gây ra sự thay đổi ở biến kia hay không. Điều này phức tạp hơn tương quan và thường cần

Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance):

Hiệp biến (Covariance): Đo lường mức độ hai biến thay đổi cùng nhau. Giá trị dương cho thấy khi một biến tăng thì biến kia cũng có xu hướng tăng, giá trị âm cho thấy khi một biến tăng thì biến kia có xu hướng giảm. Tuy nhiên, giá trị hiệp biến phụ thuộc vào đơn vị đo của các biến, làm cho việc so sánh giữa các cặp biến khác nhau trở nên khó khăn.

Tương quan (Correlation): Là phiên bản chuẩn hóa của hiệp biến. Nó chia hiệp biến cho tích của độ lệch chuẩn của hai biến. Kết quả là một giá trị không có đơn vị nằm trong khoảng từ -1 đến +1, giúp dễ dàng so sánh mức độ và hướng của mối quan hệ giữa các cặp biến khác nhau, bất kể đơn vị đo của chúng là gì. Tóm lại, tương quan là thước đo chuẩn hóa của hiệp biến.

Khi nào nên sử dụng biểu đồ trực quan hóa:

Phân tích đơn biến: Sử dụng để hiển thị phân bố của một biến duy nhất. Các biểu đồ phổ biến bao gồm:

Biểu đồ tần suất (Histograms): Hiển thị tần suất của dữ liệu trong các khoảng giá trị khác nhau.

Biểu đồ hộp (Box Plots): Hiển thị phân vị, trung vị và các giá trị ngoại lai.

Biểu đồ mật độ (Density Plots): Hiển thị ước lượng phân bố xác suất của biến.

Biểu đồ thanh (Bar Charts): Sử dụng cho các biến phân loại để hiển thị tần suất hoặc tỷ lệ của từng danh mục.

Phân tích hai biến: Sử dụng để hiển thị mối quan hệ giữa hai biến. Các biểu đồ phổ biến bao gồm:

Biểu đồ phân tán (Scatter Plots): Hiển thị mối quan hệ giữa hai biến số.

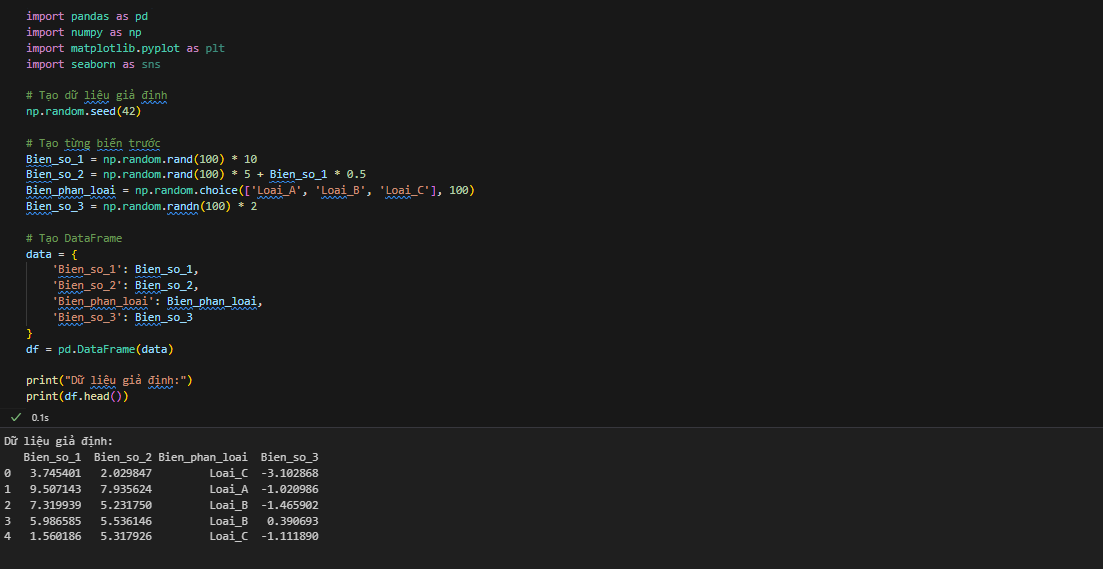
Biểu đồ đường (Line Plots): Thường sử dụng để hiển thị xu hướng của một biến số theo thời gian hoặc một biến số liên tục khác.

Biểu đồ nhiệt (Heatmaps): Hiển thị ma trận tương quan giữa nhiều cặp biến.

Biểu đồ hộp (Box Plots) hoặc Biểu đồ violin (Violin Plots): Sử dụng để so sánh phân bố của một biến số giữa các danh mục khác nhau của một biến phân loại.

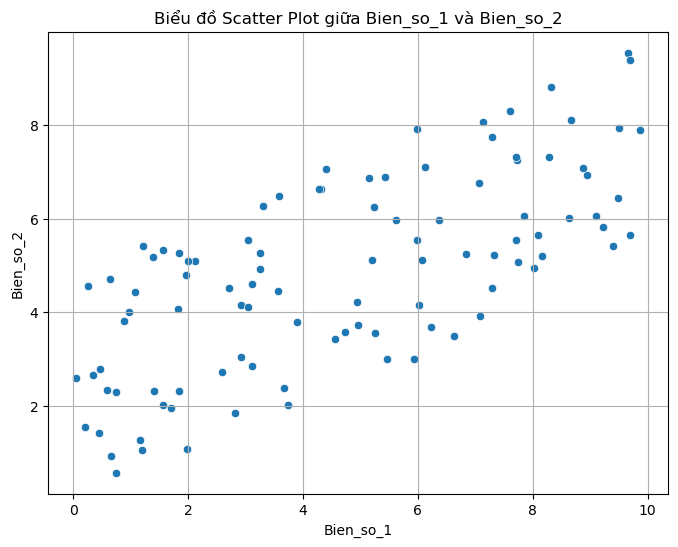
Đoạn code mẫu tạo biểu đồ Scatter Plot (Phân tán) và Heatmap:

Tạo dữ liệu giả định:



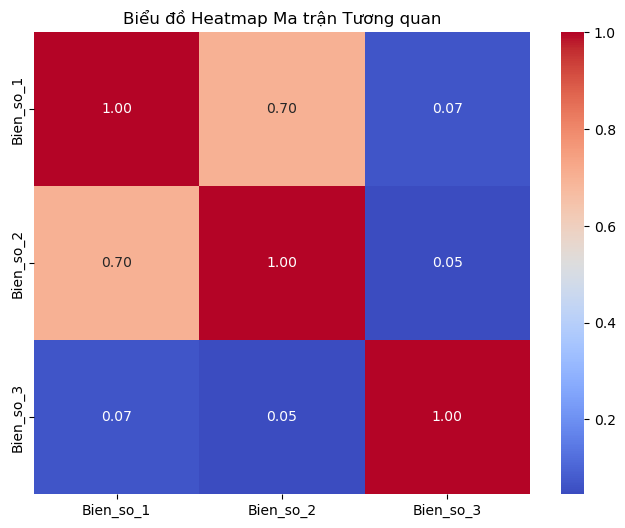
Biểu đồ Scatter Plot (Phân tán):

Biểu đồ phân tán được sử dụng để hiển thị mối quan hệ giữa hai biến số.



Biểu đồ heatmap:

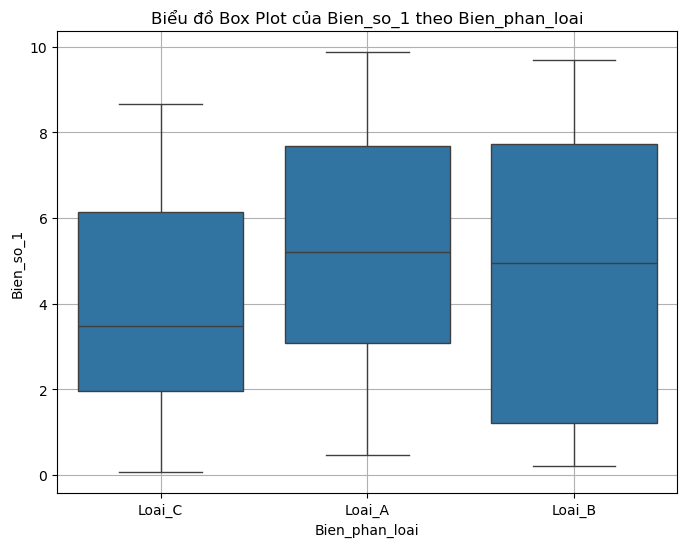
Biểu đồ Heatmap thường được sử dụng để hiển thị ma trận tương quan giữa nhiều biến số. Các giá trị tương quan được mã hóa bằng màu sắc.



Trực quan hóa mối quan hệ giữa một biến số và một biến phân loại bằng biểu đồ boxplot hoặc violin plot trong Python:

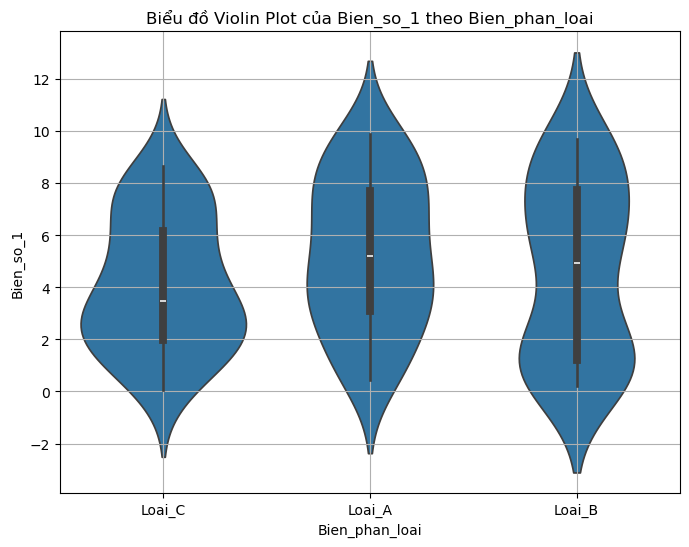
Biểu đồ boxplot:

Biểu đồ Boxplot hữu ích để so sánh sự phân bố của một biến số giữa các nhóm trong một biến phân loại.



Biểu đồ violin plot:

Biểu đồ Violin Plot cũng cho thấy sự phân bố của một biến số theo các nhóm phân loại, tương tự như Box Plot nhưng cung cấp thêm thông tin về mật độ phân phối.



### 1.3.2 Bài làm mẫu

**Bài toán 1:** Thực hiện các nhiệm vụ trong bài toán 1 để làm quen với các hàm và thư viện hỗ trợ phân tích

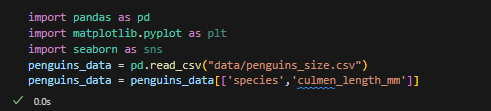
dữ liệu đơn biến. Bài toán này được thực hiện trên 2 tập dữ liệu là tập dữ liệu về chim cánh cụt và tập dữ

liệu giá nhà.

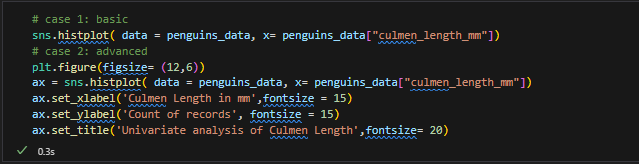
**Nhiệm vụ 1:** phân tích dữ liệu đơn biến trên dữ liệu về chim cánh cụt lấy tại

<https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data>

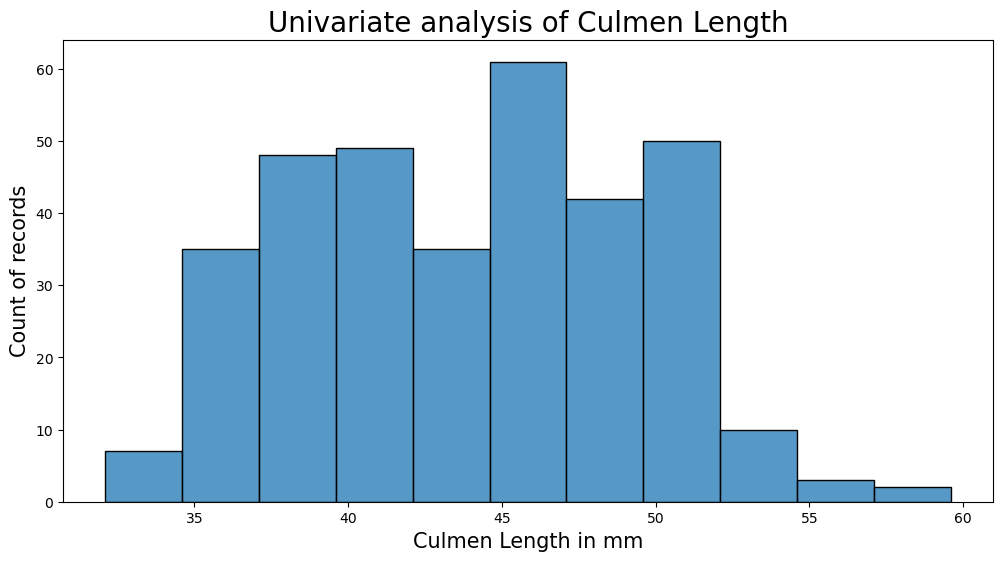
1. Import thư viện và nhập dữ liệu



2. Phân tích đơn biến bằng Histogram

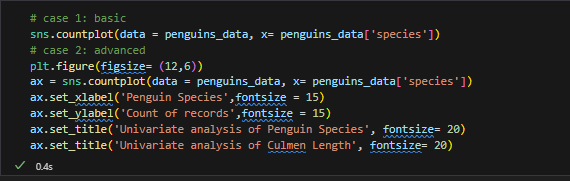


Kết quả thực hiện case 2:

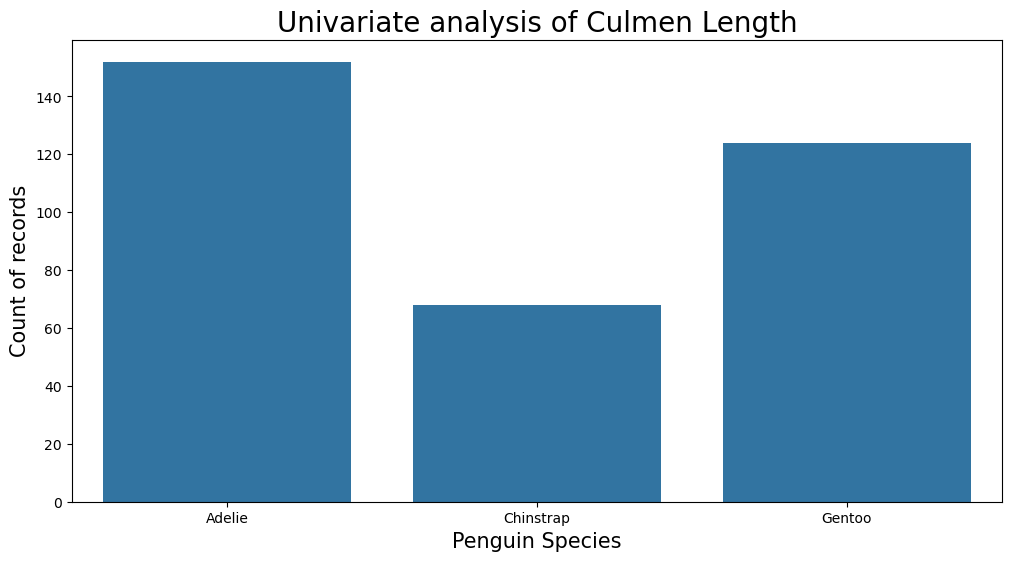


Hình 1.3 - Biểu đồ histogram

3. Phân tích đơn biến bằng bar chart

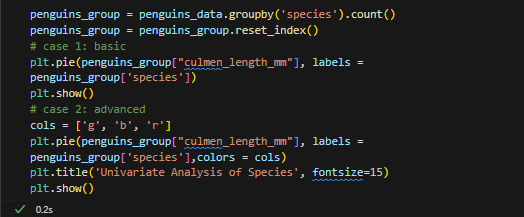


Kết quả thực hiện case 2:

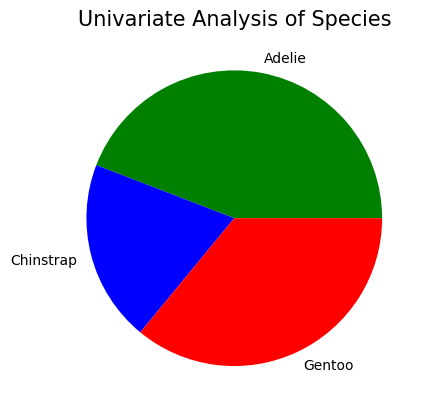


Hình 1.4 - Biểu đồ thanh (bar chart)

4. Phân tích đơn biến bằng biểu đồ tròn (Pie-chart)



Kết quả thực hiện case 2:

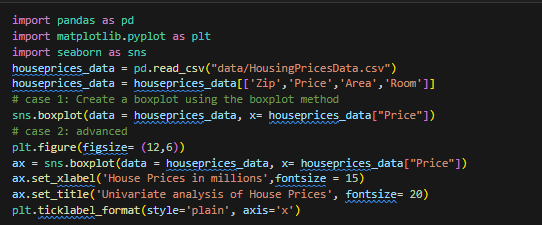


Hình 1.5 - Biểu đồ tròn (pie-chart)

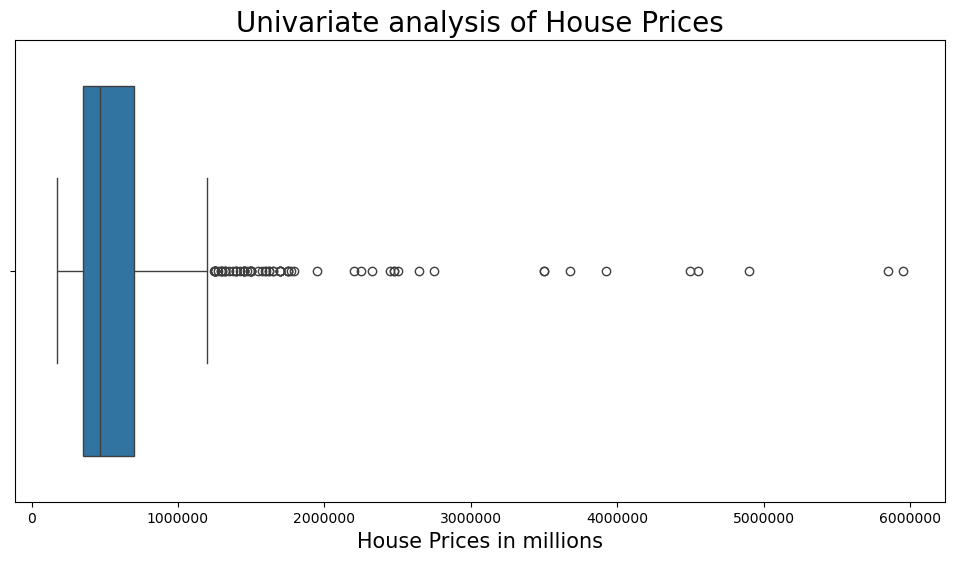
Nhiệm vụ 2: Phân tích dữ liệu đơn biến trên dữ liệu giá nhà lấy từ

<https://www.kaggle.com/datasets/thomasnibb/amsterdam-house-price-prediction>

1. Import thư viện, nạp dữ liệu giá nhà và phân tích đơn biến dựa vào boxplot

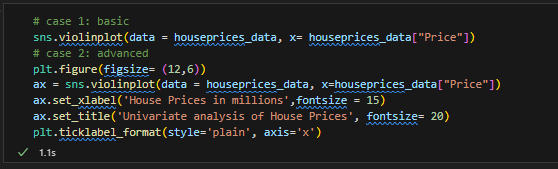


Kết quả thực hiện case 2:

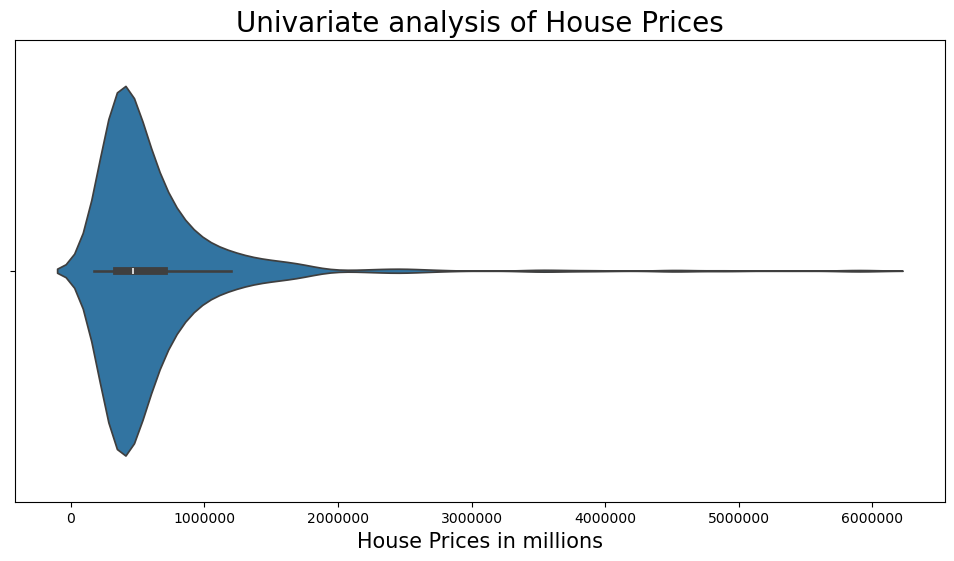


Hình 1.6 - Biểu đồ boxplot (được trình bày phương ngang)

2. Phân tích dữ liệu đơn biến dựa vào violin plot

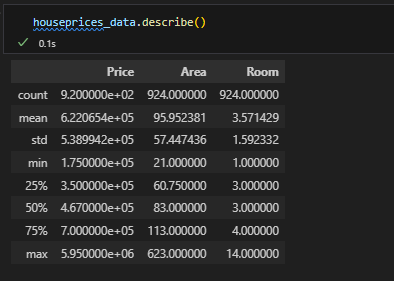


Kết quả thực hiện case 2:



Hình 1.7 - Biểu đồ violin

3. Phân tích dữ liệu đơn biến dựa vào bản tóm tắt dữ liệu



Hình 1.8 - Dữ liệu thống kê mô tả

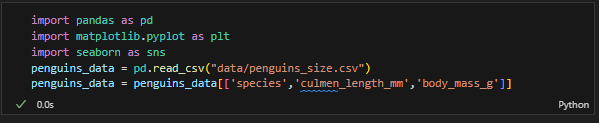
**Bài toán 2:** Thực hiện các nhiệm vụ trong bài toán 2 để làm quen với việc phân tích hai biến với các hàm

trong thư viện scikit-learn.

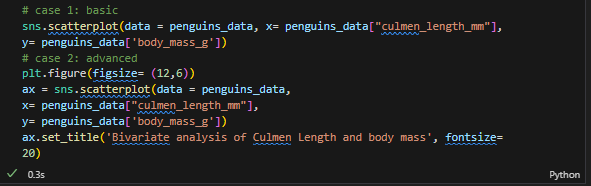
**Nhiệm vụ 1:** phân tích dữ liệu hai biến trên dữ liệu về chim cánh cụt. Dữ liệu lấy tại

<https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data>

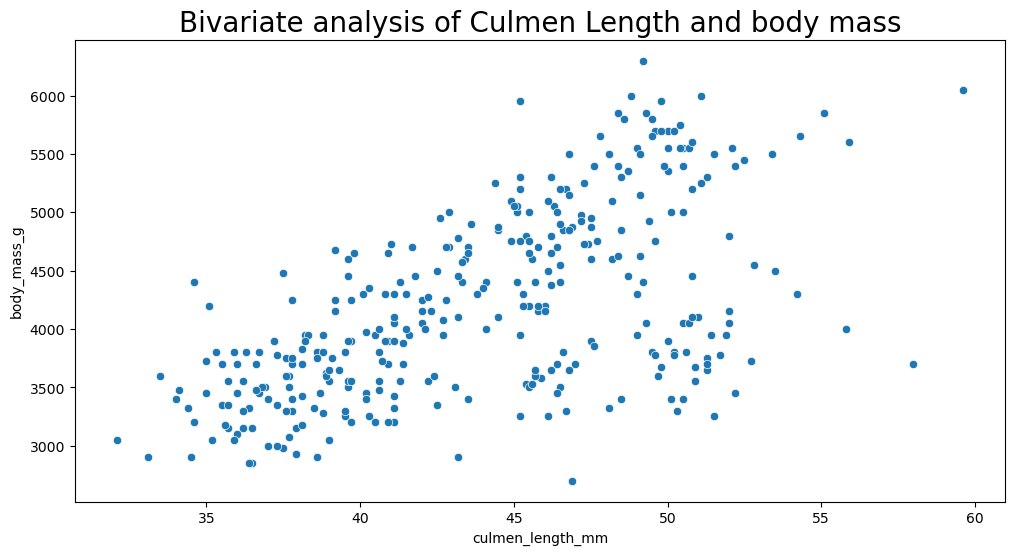
1. Import thư viện và chuẩn bị dữ liệu phân tích



2. Phân tích dữ liệu 2 biến dựa vào phương pháp scatterplot



Kết quả thực hiện case 2:



Hình 1.9- Biểu đồ scatter plot

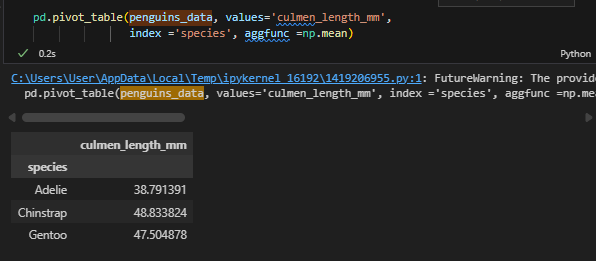
3. Phân tích 2 biến dựa vào bảng crosstab/two\_way

Kết quả thực hiện:

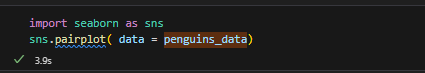


4. Phân tích hai biến sử dụng pivot\_table

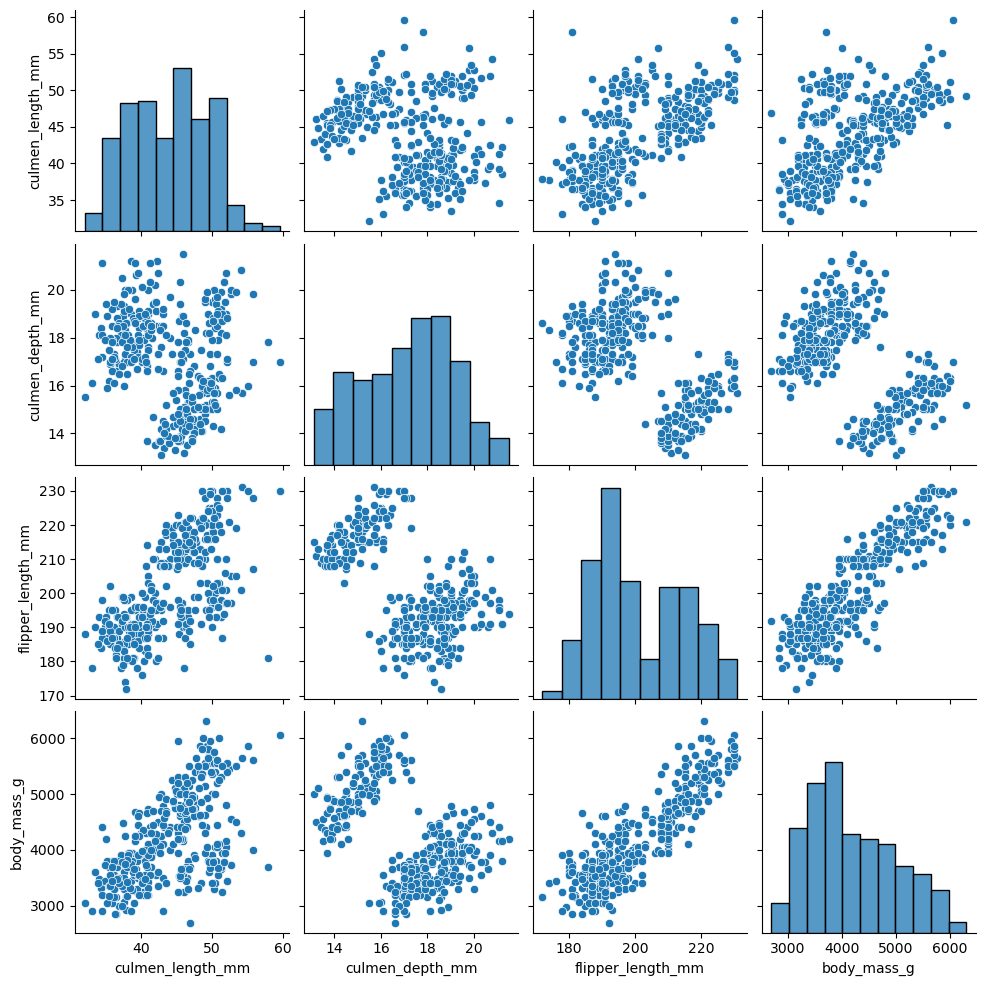
Kết quả thực hiện:



5. Phân tích 2 biến sử dụng pairplot



Kết quả thực hiện:



Hình 1.10-Biểu đồ pair-plot

**Bài toán 3:** Thực hiện các nhiệm vụ trong bài toán 3 để làm quen với việc sử dụng các công cụ hỗ trợ EDA

tự động.

**Nhiệm vụ 1:** Sử dụng pandas profiling trên dữ liệu Customer Personality Analysis. Dữ liệu lấy tại

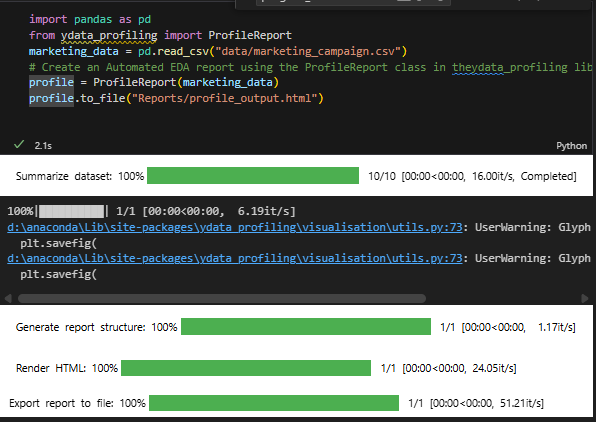
<https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

1. Cài đặt pandas\_profiling sau này đổi tên thành ydata\_profiling (xem thông tin chi tiết tại

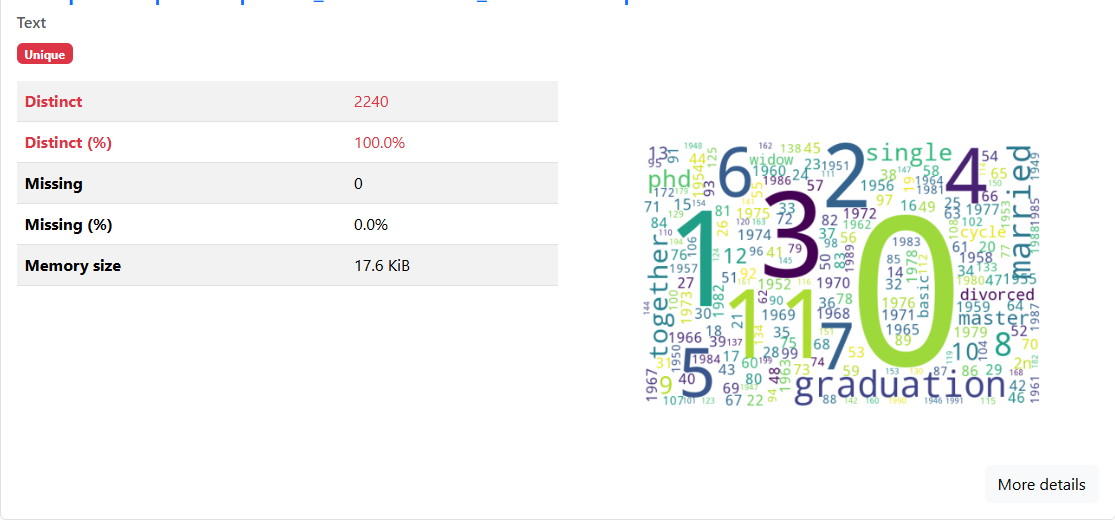
<https://pypi.org/project/pandas-profiling/3.1.0>)

2. Sử dụng công cụ

Kết quả thực hiện



3. Tiến hành EDA trên trang tập tin profile\_output.html



Hình 1.11 - EDA Report của công cụ ydata\_profiling

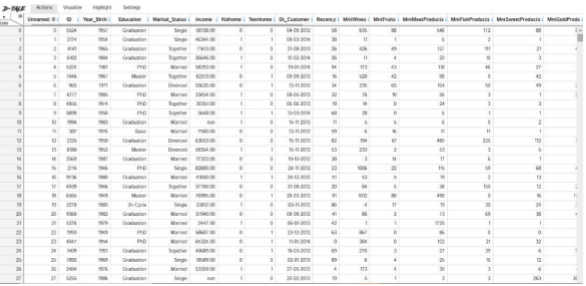
**Nhiệm vụ 2:** Sử dụng dtale trên dữ liệu Marketing Campaign. Dữ liệu lấy từ

<https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

1. Cài đặt dtale (xem thông tin chi tiết tại <https://dtale.readthedocs.io/en/latest/>)

2. Sử dụng công cụ

Kết quả thực hiện



Hình 1.12 - EDA Report của công cụ dtale

### 1.3.3. Bài tập thực hành 1

Tìm hiểu các tính năng và cách sử dụng sản phẩm SweetViz (<https://pypi.org/project/sweetviz>) áp dụng

trên tập dữ liệu Marketing Campaign

1. Giới thiệu về SweetViz

SweetViz là một thư viện Python mã nguồn mở được phát triển để tự động hóa và đơn giản hóa quá trình Phân tích Dữ liệu Khám phá (EDA). Được thiết kế với mục tiêu giúp các nhà khoa học dữ liệu và nhà phân tích nhanh chóng hiểu rõ cấu trúc, phân phối và mối quan hệ trong dữ liệu của họ, SweetViz tạo ra các báo cáo trực quan, mật độ cao và tương tác chỉ với vài dòng mã.

Thư viện này đặc biệt hữu ích trong các giai đoạn đầu của dự án phân tích dữ liệu, khi cần nhanh chóng hiểu được bản chất của dữ liệu trước khi đi sâu vào các phân tích phức tạp hơn. SweetViz tự động phân tích các kiểu dữ liệu khác nhau và tạo ra các biểu đồ phù hợp, giúp tiết kiệm thời gian đáng kể so với việc tạo thủ công các biểu đồ và phân tích.

2. Các tính năng chính của SweetViz

SweetViz cung cấp một loạt các tính năng mạnh mẽ để hỗ trợ quá trình EDA:

2.1 . Phân tích tự động và trực quan hóa

SweetViz tự động phân tích các kiểu dữ liệu khác nhau (số, phân loại, văn bản) và tạo ra các biểu đồ phù hợp. Thư viện này tạo ra các biểu đồ trực quan phong phú, bao gồm biểu đồ phân phối, biểu đồ thanh, biểu đồ tương quan và biểu đồ mối quan hệ giữa các biến.

2.2 . Báo cáo HTML tương tác

Một trong những tính năng nổi bật của SweetViz là khả năng xuất báo cáo dưới dạng tệp HTML độc lập, cho phép người dùng tương tác với các biểu đồ, phóng to, thu nhỏ và xem chi tiết dữ liệu. Báo cáo này có thể dễ dàng chia sẻ với các bên liên quan khác mà không cần cài đặt thêm phần mềm.

2.3. So sánh tập dữ liệu

SweetViz cho phép so sánh hai tập dữ liệu (ví dụ: tập huấn luyện và tập kiểm tra) hoặc so sánh một tập dữ liệu với chính nó dựa trên một biến mục tiêu. Tính năng này đặc biệt hữu ích trong các dự án học máy để đảm bảo rằng tập huấn luyện và tập kiểm tra có phân phối tương tự.

2.4. Phát hiện mối quan hệ

Thư viện này tự động phát hiện và trực quan hóa các mối quan hệ tiềm ẩn giữa các biến, bao gồm cả mối quan hệ tương quan và mối quan hệ phân loại. Điều này giúp nhà phân tích nhanh chóng xác định các biến có ảnh hưởng lớn đến biến mục tiêu.

2.5. Tóm tắt thống kê

SweetViz cung cấp các số liệu thống kê tóm tắt cho từng biến, bao gồm giá trị trung bình, trung vị, độ lệch chuẩn, giá trị thiếu, v.v. Thông tin này được trình bày một cách trực quan và dễ hiểu [2].

Bảng dưới đây tóm tắt các tính năng chính của SweetViz:

| Tính năng | Mô tả |
| --- | --- |
| Phân tích tự động | Tự động phân tích các kiểu dữ liệu khác nhau và tạo biểu đồ phù hợp |
| Trực quan hóa mật độ cao | Tạo các biểu đồ phong phú như phân phối, thanh, tương quan, v.v. |
| Báo cáo HTML tương tác | Xuất báo cáo dạng HTML cho phép tương tác với biểu đồ |
| So sánh tập dữ liệu | So sánh hai tập dữ liệu hoặc phân tích dựa trên biến mục tiêu |
| Phát hiện mối quan hệ | Tự động phát hiện và trực quan hóa mối quan hệ giữa các biến |
| Tóm tắt thống kê | Cung cấp các số liệu thống kê cho từng biến |

3. Cách sử dụng SweetViz

Việc sử dụng SweetViz rất đơn giản, thường chỉ với hai dòng mã chính:

3.1. Cài đặt

Trước tiên, cần cài đặt thư viện SweetViz:



3.2. Phân tích dữ liệu cơ bản

Để phân tích một DataFrame, sử dụng hàm analyze() :



3.3. Phân tích với biến mục tiêu

Phân tích dữ liệu với một biến mục tiêu cụ thể, bạn có thể chỉ định biến đó trong hàm analyze() :



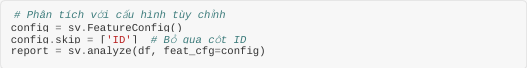
3.4. So sánh hai tập dữ liệu

SweetViz cho phép so sánh hai tập dữ liệu, ví dụ như tập huấn luyện và tập kiểm tra:



3.5. Phân tích với cấu hình tùy chỉnh

SweetViz cũng cho phép tùy chỉnh các cấu hình phân tích:

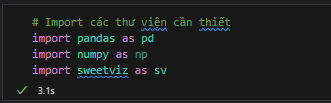


**4. Áp dụng SweetViz vào tập dữ liệu Marketing Campaign**

Tập dữ liệu Marketing Campaign chứa thông tin về các chiến dịch tiếp thị và phản hồi của khách hàng. Dưới đây là cách SweetViz có thể được áp dụng để phân tích tập dữ liệu này:

4.1. Đọc và phân tích dữ liệu

Cài đặt thư viện:



Đọc và khám phá dữ liệu:

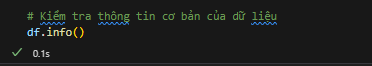


Kết quả thục hiện:

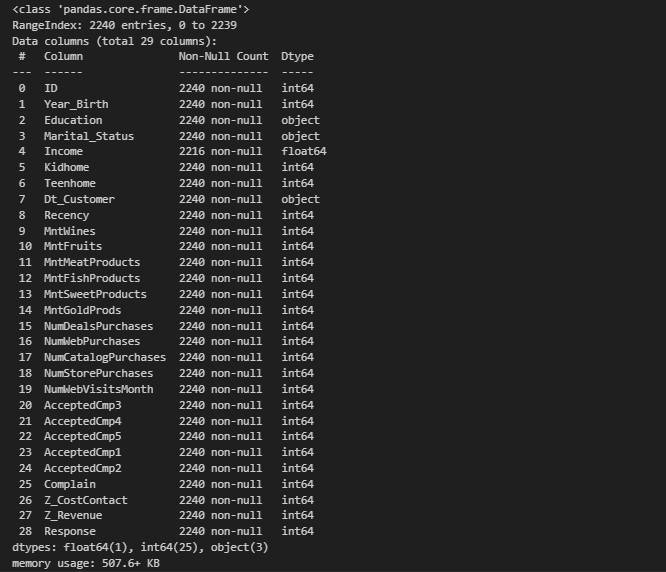




Kiểm tra thông tin dữ liệu:



Kết quả thực hiện:



Nhận xét ban đầu:

Tập dữ liệu có khoảng 2.240 dòng và 29 cột.

Mỗi dòng đại diện cho một khách hàng.

Biến mục tiêu (Response) thể hiện khách hàng có phản hồi (1) hay không (0) với chiến dịch.

Danh sách biến trong tập dữ liệu Marketing Campaign:

1. ID - Mã định danh duy nhất của khách hàng

2. Year\_Birth - Năm sinh của khách hàng

3. Education - Trình độ học vấn (Basic, Graduation, Master, PhD, ...)

4. Marital\_Status - Tình trạng hôn nhân (Single, Married, Together, Divorced, ...)

5. Income - Thu nhập hàng năm của khách hàng

6. Kidhome - Số trẻ em trong gia đình

7. Teenhome - Số thiếu niên trong gia đình

8. Dt\_Customer - Ngày khách hàng gia nhập hệ thống

9. Recency - Số ngày kể từ lần mua hàng gần nhất

10. Complain - Khách hàng đã từng khiếu nại (1 = Có, 0 = Không)

11. MntWines - Số tiền chi tiêu cho rượu vang

12. MntFruits - Số tiền chi tiêu cho trái cây

13. MntMeatProducts - Số tiền chi tiêu cho thịt

14. MntFishProducts - Số tiền chi tiêu cho cá

15. MntSweetProducts - Số tiền chi tiêu cho đồ ngọt

16. MntGoldProds - Số tiền chi tiêu cho sản phẩm vàng / cao cấp

17. NumDealsPurchases - Số lần mua hàng khi có khuyến mãi / giảm giá

18. NumWebPurchases - Số lần mua qua website

19. NumCatalogPurchases - Số lần mua qua catalog (sổ giới thiệu)

20. NumStorePurchases - Số lần mua trực tiếp tại cửa hàng

21. NumWebVisitsMonth - Số lần truy cập website trong tháng gần nhất

22. AcceptedCmp1 - Khách hàng phản hồi Chiến dịch 1 (1 = Có, 0 = Không)

23. AcceptedCmp2 - Khách hàng phản hồi Chiến dịch 2

24. AcceptedCmp3 - Khách hàng phản hồi Chiến dịch 3

25. AcceptedCmp4 - Khách hàng phản hồi Chiến dịch 4

26. AcceptedCmp5 - Khách hàng phản hồi Chiến dịch 5

27. Response - Phản hồi chiến dịch hiện tại (biến mục tiêu)

Biến có thể tạo thêm:

28. Age - Tuổi khách hàng (2025 - Year\_Birth)

29. Children - Tổng số con (Kidhome + Teenhome)

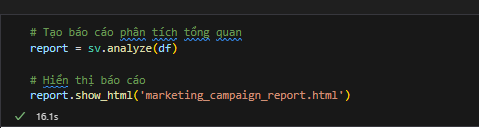
30. TotalSpent - Tổng chi tiêu (cộng các biến Mnt\*)

31. TotalAccepted - Tổng số chiến dịch đã phản hồi (AcceptedCmp1–5)

32. CustomerTenure - Thời gian gắn bó với công ty (ngày hiện tại - Dt\_Customer)

4.2. Phân tích tổng quan

Tạo báo cáo phân tích tổng quan



Các biến được chia thành 4 nhóm chính:

(1) Nhóm nhân khẩu học (Demographics)

| Biến | Ý nghĩa | Kiểu dữ liệu |
| --- | --- | --- |
| ID | Mã định danh khách hàng | Số nguyên |
| Year\_Birth | Năm sinh khách hàng | Số nguyên |
| Education | Trình độ học vấn | Phân loại |
| Marital\_Status | Tình trạng hôn nhân | Phân loại |
| Income | Thu nhập hàng năm | Số thực |
| Kidhome, Teenhome | Số trẻ em, thiếu niên trong nhà | Số nguyên |

Nhóm này giúp phân khúc khách hàng theo độ tuổi, thu nhập, và hoàn cảnh gia đình.

(2) Nhóm hành vi mua hàng (Purchase Behavior)

| **Biến** | **Ý nghĩa** |
| --- | --- |
| MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds | Số tiền chi tiêu cho từng nhóm sản phẩm |
| NumDealsPurchases | Số lần mua hàng khi có khuyến mãi |
| NumWebPurchases, NumCatalogPurchases, NumStorePurchases | Kênh mua hàng (online, catalog, cửa hàng) |
| NumWebVisitsMonth | Số lần truy cập website mỗi tháng |
| Recency | Số ngày kể từ lần mua hàng gần nhất |

Giúp phân tích mức độ tiêu thụ, hành vi mua hàng, mức độ tương tác với các kênh.

(3) Nhóm chiến dịch Marketing

| Biến | Ý nghĩa | Kiểu dữ liệu |
| --- | --- | --- |
| ID | Mã định danh khách hàng | Số nguyên |
| Year\_Birth | Năm sinh khách hàng | Số nguyên |
| Education | Trình độ học vấn | Phân loại |
| Marital\_Status | Tình trạng hôn nhân | Phân loại |
| Income | Thu nhập hàng năm | Số thực |
| Kidhome, Teenhome | Số trẻ em, thiếu niên trong nhà | Số nguyên |

Nhóm này phản ánh mức độ hiệu quả của các chiến dịch marketing và lòng trung thành của khách hàng.

(4) Nhóm kỹ thuật / hệ thống

| Biến | Ý nghĩa |
| --- | --- |
| Z\_CostContact, Z\_Revenue | Biến giả lập (constant) — dùng để đảm bảo tính ẩn danh |
| Dt\_Customer | Ngày khách hàng bắt đầu tham gia chương trình |

Nhận xét:

Không có thiếu dữ liệu nghiêm trọng (MISSING: rất ít hoặc không có).

Các biến định lượng (Income, Recency, MntWines, v.v.) có phân bố lệch phải (right-skewed) → đa số khách hàng chi tiêu ở mức trung bình, ít khách hàng chi tiêu cao.

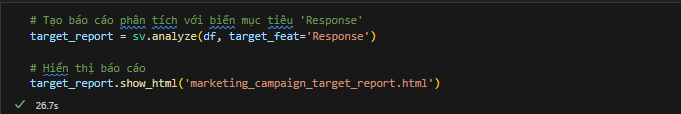
Các biến định tính có số lượng nhóm hợp lý (Education, Marital\_Status).

Biến Recency có tương quan âm nhẹ (-0.05) với tổng chi tiêu → khách hàng mua hàng gần đây có xu hướng chi tiêu nhiều hơn.

Biến NumDealsPurchases và MntFishProducts có tương quan nhỏ nhưng cùng chiều → nhóm thích giảm giá thường chi tiêu cho thực phẩm.

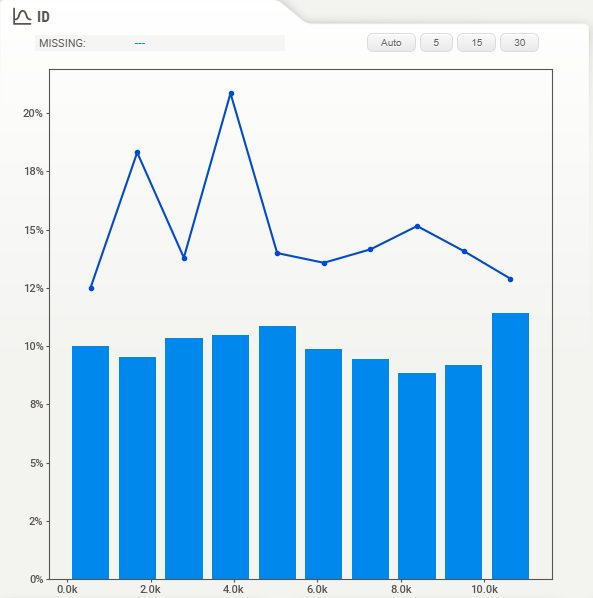
Biến Marital\_Status và AcceptedCmp3 có tương quan danh mục nhỏ (≈ 0.05–0.04) → cho thấy nhóm hôn nhân có thể ảnh hưởng đến hành vi phản hồi.

4.3. Phân tích với biến mục tiêu



Nhận xét mối quan hệ với các biến quan trọng:

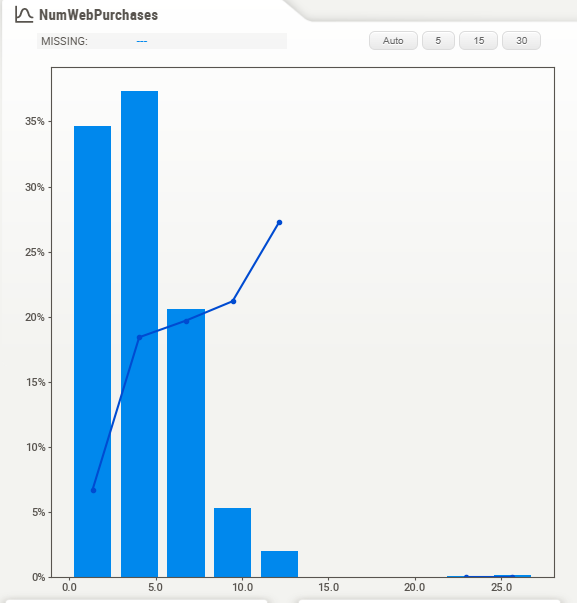
Thu nhập (Income)

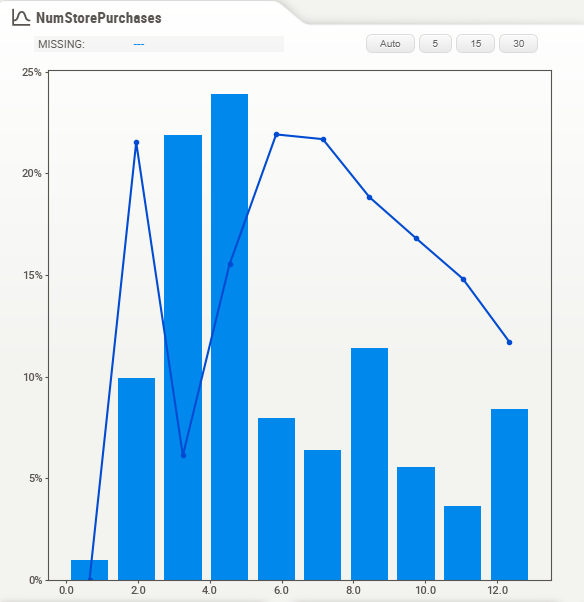


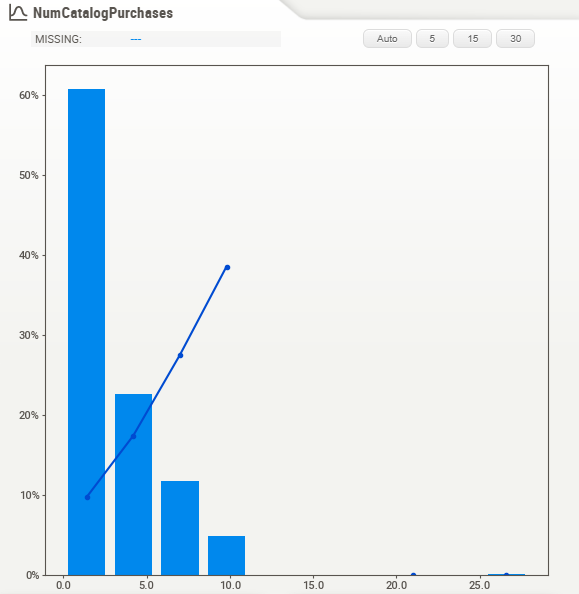
Trung bình thu nhập của nhóm phản hồi cao hơn (~60.000 so với 45.000).

=> Người có thu nhập cao có xu hướng mua hàng sau khi được tiếp cận chiến dịch → tiềm năng “High Value Target”.

Tần suất mua (NumWebPurchases, NumStorePurchases, NumCatalogPurchases):



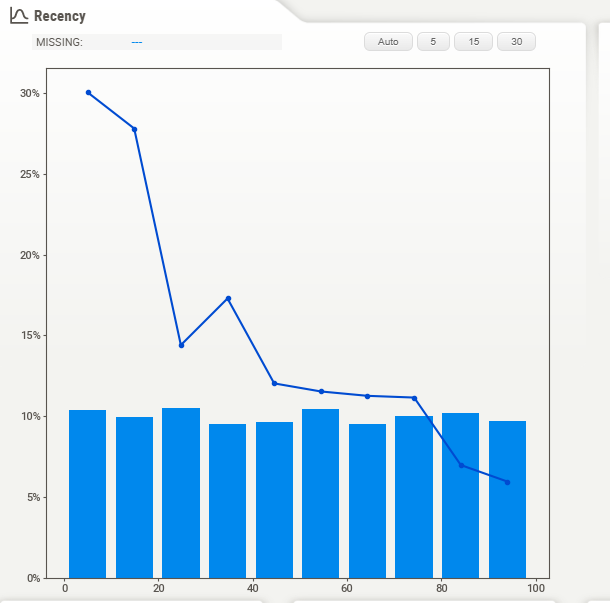




Nhóm phản hồi (Response=1) có tần suất mua cao hơn ở mọi kênh.

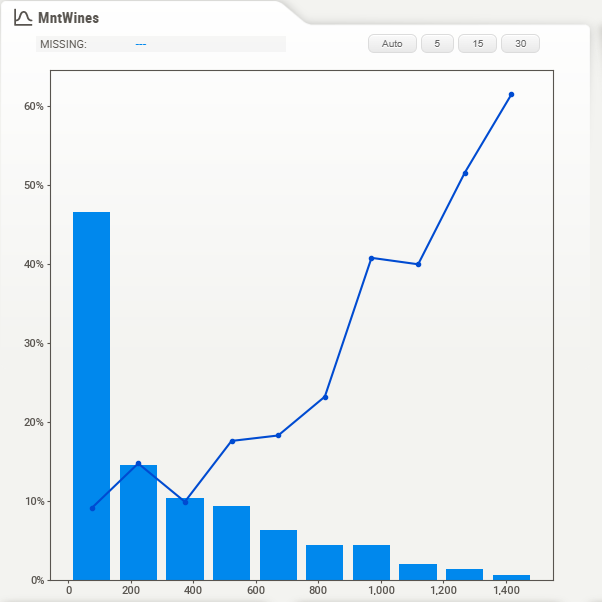
Đặc biệt, họ thường có số lần mua qua web cao hơn trung bình → nhóm am hiểu công nghệ, dễ bị ảnh hưởng qua digital marketing.

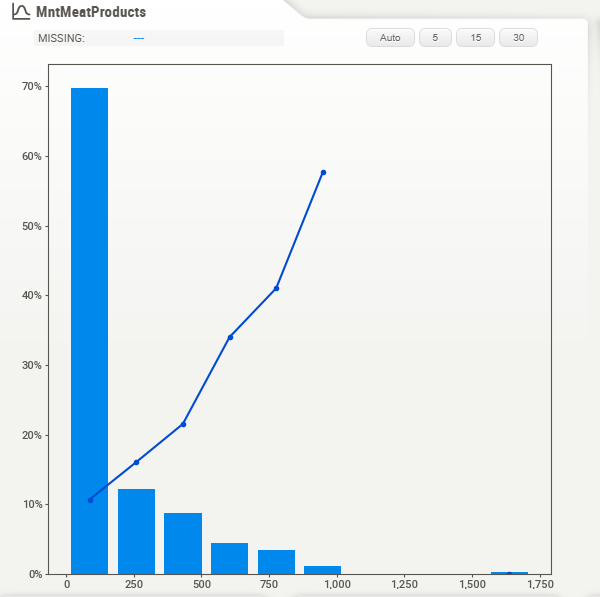
Thời gian từ lần mua gần nhất (Recency):

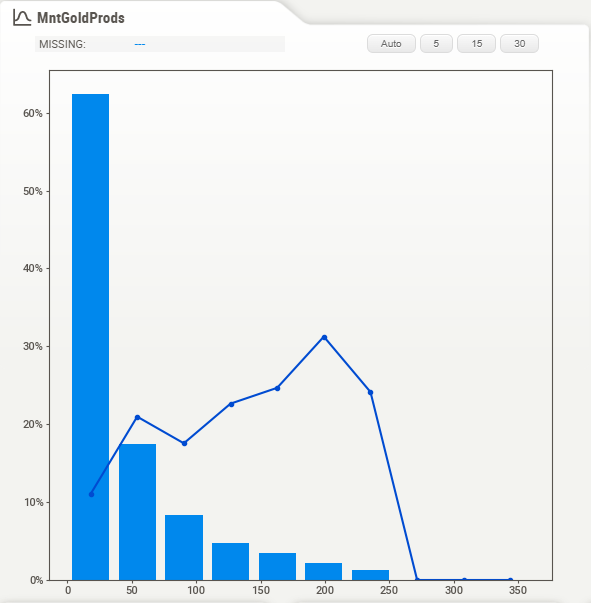


* Recency thấp (tức mới mua gần đây) có tỷ lệ phản hồi cao hơn.  
   → Khách hàng “mới hoạt động” dễ tương tác lại.
* Có thể phân nhóm:
  + Recency < 30 ngày → nhóm tiềm năng cao
  + Recency > 90 ngày → nhóm cần tái kích hoạt

Tổng chi tiêu (MntWines, MntMeatProducts, MntGoldProds, …):

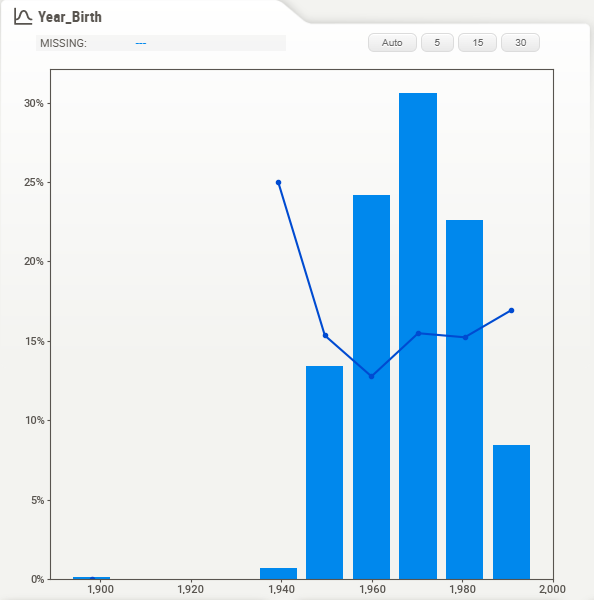


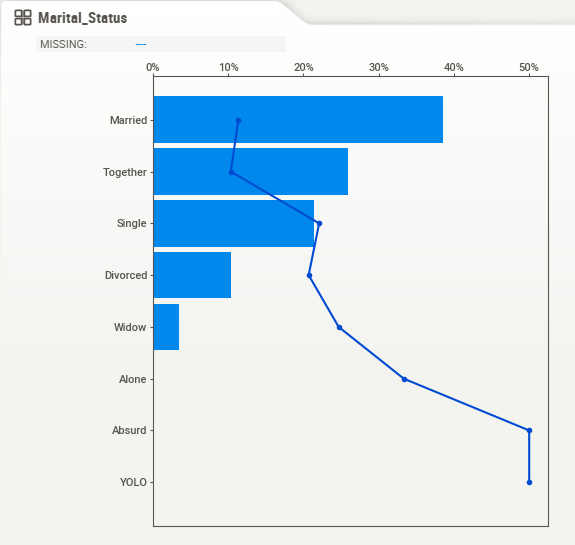


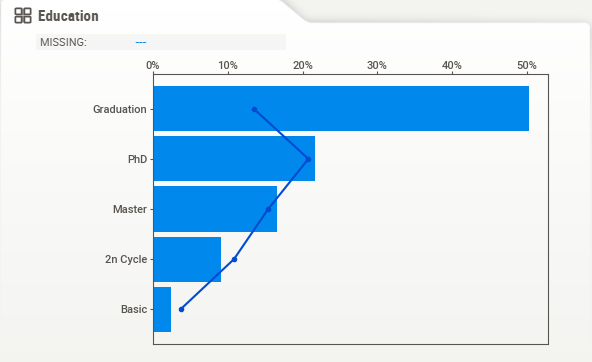


* Nhóm phản hồi có tổng chi tiêu cao hơn rõ rệt, đặc biệt ở:
  + Rượu vang (MntWines)
  + Thịt (MntMeatProducts)
  + Sản phẩm cao cấp (MntGoldProds)
* => Nhóm khách hàng cao cấp là đối tượng chủ yếu phản hồi chiến dịch.

Nhân khẩu học (Year\_Birth, Marital\_Status, Education):







* Tuổi trung bình của nhóm phản hồi: 35–50 tuổi.
* Trình độ: nhóm có bằng Graduation hoặc Master phản hồi nhiều nhất.
* Hôn nhân: nhóm Married / Together phản hồi cao hơn nhóm Single.

4.4. So sánh các nhóm khách hàng

Chia dữ liệu thành 2 nhóm khách hàng có con (Kidhome) và không có con (Teenhome):

Kết quả so sánh hai nhóm khách hàng cho thấy:

Nhóm không có con có thu nhập trung bình cao hơn khoảng 15.000 đơn vị so với nhóm có con.

Họ cũng chi tiêu nhiều hơn cho rượu vang và sản phẩm cao cấp, đồng thời phản hồi chiến dịch marketing cao hơn (~12% so với 7%).

Ngược lại, nhóm có con có xu hướng chi tiêu cho sản phẩm thiết yếu như thịt, trái cây và đồ ngọt.

Do đặc thù thời gian bận rộn, họ ít mua hàng online hơn.

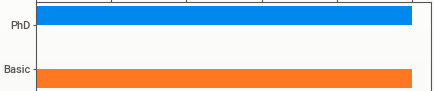
Đề xuất chiến lược marketing:

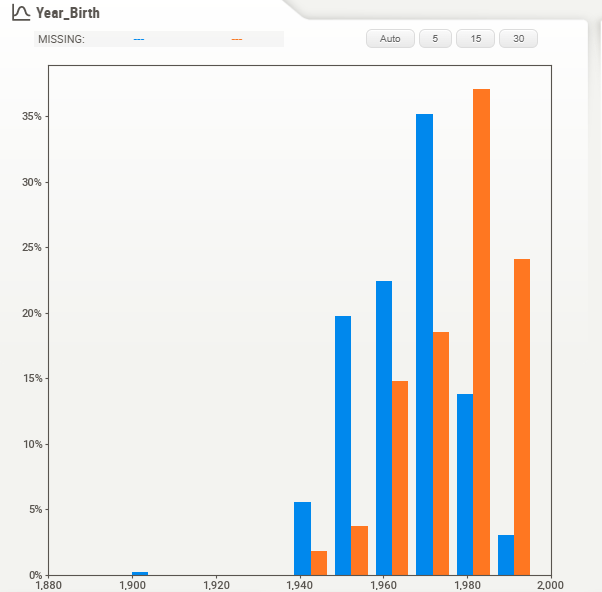
Với nhóm có con: tập trung chiến dịch “gia đình – tiết kiệm” hoặc combo sản phẩm thiết yếu.

Với nhóm không con: ưu tiên quảng cáo sản phẩm cao cấp, khuyến mãi online, và digital marketing cá nhân hóa.

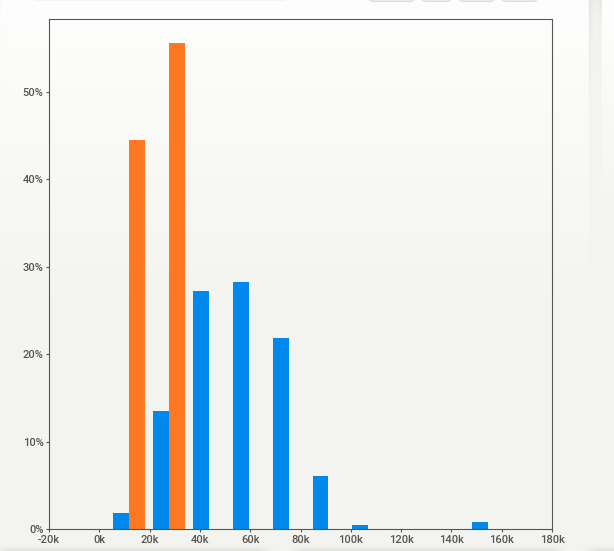
4.5. Phân tích theo trình độ học vấn

Số lượng khách hàng mỗi nhóm → Nhóm “Basic” thường chiếm tỉ lệ lớn hơn (do phổ biến hơn).

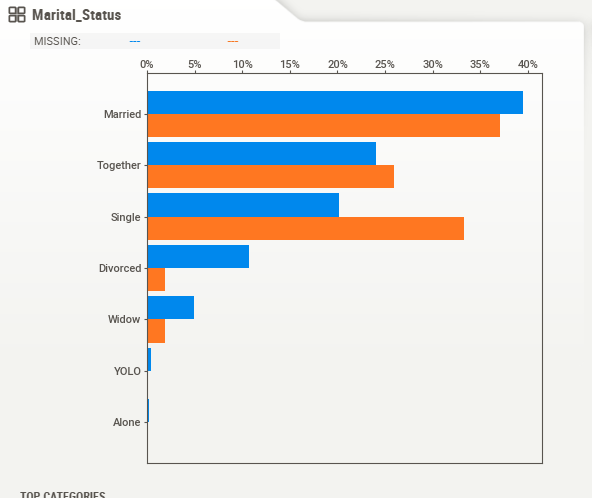




Tuổi (Year\_Birth) → Nhóm PhD có xu hướng lớn tuổi hơn, thường đã ổn định về tài chính.



Thu nhập (Income) → PhD thường có thu nhập trung bình cao hơn rõ rệt.



Tình trạng hôn nhân (Marital\_Status) → PhD có thể có tỷ lệ “Married” cao hơn.

**5. Nhận xét và kết luận**

**\* Nhận xét về SweetViz**

Mặc dù không thể chạy trực tiếp SweetViz trong môi trường hiện tại, dựa trên tài liệu và kinh nghiệm chung, có thể rút ra một số nhận xét về công cụ này:

1. Tiết kiệm thời gian: SweetViz giúp tự động hóa quá trình EDA, tiết kiệm thời gian đáng kể cho nhà phân tích.

2. Trực quan hóa phong phú: SweetViz tạo ra các biểu đồ trực quan phong phú, giúp hiểu rõ hơn về dữ liệu.

3. Phát hiện mối quan hệ: Công cụ này giúp phát hiện các mối quan hệ tiềm ẩn giữa các biến, đặc biệt là với biến mục tiêu.

4. So sánh dữ liệu: Khả năng so sánh các tập dữ liệu hoặc các phân đoạn dữ liệu là một tính năng mạnh mẽ của SweetViz.

5. Hạn chế tùy chỉnh: SweetViz có ít tùy chọn tùy chỉnh so với các thư viện trực quan hóa khác như Matplotlib hoặc Plotly.

6. Vấn đề tương thích: Như đã gặp phải, SweetViz có thể gặp vấn đề tương thích với các phiên bản mới của các thư viện phụ thuộc.

\* Kết luận

SweetViz là một công cụ mạnh mẽ cho phân tích dữ liệu khám phá, cung cấp một cách nhanh chóng và hiệu quả để hiểu cấu trúc và mối quan hệ trong dữ liệu. Với khả năng tạo ra các báo cáo trực quan tương tác chỉ với vài dòng mã, SweetViz là một công cụ quý giá trong bộ công cụ của bất kỳ nhà khoa học dữ liệu nào.

Đối với tập dữ liệu Marketing Campaign, SweetViz có thể cung cấp những hiểu biết quý giá về hành vi khách hàng, hiệu quả của các chiến dịch tiếp thị và các yếu tố ảnh hưởng đến tỷ lệ phản hồi. Những hiểu biết này có thể giúp các nhà tiếp thị tối ưu hóa chiến lược của họ và cải thiện kết quả kinh doanh.

### 1.3.4. Bài tập thực hành 2

Tìm hiểu các tính năng và cách sử dụng sản phẩm AutoViz (<https://pypi.org/project/autoviz>) áp dụng trên

tập dữ liệu Marketing Campaign

1. Giới thiệu AutoViz

AutoViz là thư viện Python dành cho visualization tự động (auto visualization / exploratory data analysis tự động), giúp người dùng nhanh chóng tạo ra các biểu đồ trực quan từ một DataFrame hoặc file CSV / JSON mà không cần viết nhiều code vẽ từng biểu đồ thủ công.

Bạn chỉ cần một dòng lệnh (với một số tùy chọn cấu hình) là AutoViz sẽ tự phân tích dữ liệu, xác định kiểu biến, vẽ histogram, scatter plot, box plot, heatmap, biểu đồ tương quan, và thậm chí đánh giá chất lượng dữ liệu.

AutoViz cũng có khả năng đánh giá chất lượng dữ liệu (data quality) — nhận diện missing values, outliers và đề xuất sửa chữa nhanh (FixDQ).

Ngoài ra, AutoViz có thể hỗ trợ lựa chọn biến (feature selection) bằng cách “nhấn mạnh” các biến có ảnh hưởng lớn qua các biểu đồ mà nó vẽ.

2. Sử dụng AutoViz

Khi chạy AutoViz, bạn có thể thấy:

Biểu đồ histogram, density, box plot cho mỗi biến số liên tục.  
 Biểu đồ bar plot / count plot cho các biến phân loại.

Biểu đồ scatter plot giữa các cặp biến liên tục, có thể thêm đường hồi quy nếu lowess=True.

Heatmap tương quan giữa các biến số để xem mối quan hệ tuyến tính.

Biểu đồ violin plots hoặc distribution plots cho biến số phân phối.

Biểu đồ outlier để nhận diện điểm ngoại lai.

Đánh giá missing / null values (báo cáo số lượng giá trị thiếu, tỷ lệ) và đề xuất sửa.

Nếu có biến mục tiêu (depVar), AutoViz sẽ làm nổi bật các tương quan giữa biến mục tiêu và các biến khác — hiển thị mối liên hệ, phân phối biến khác theo từng nhãn mục tiêu.

Tóm tắt thống kê mô tả (mean, median, std, min, max) của mỗi biến.

Report HTML tương tác: bạn có thể mở file HTML để tương tác với biểu đồ, zoom, tiêu đề, lọc.

3. Ưu điểm và hạn chế của AutoViz

Ưu điểm:

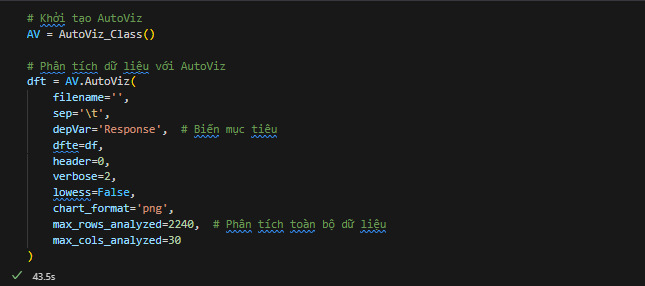
| Ưu điểm | Mô tả |
| --- | --- |
| Tự động hóa cao | Chỉ cần một lệnh là vẽ được hàng chục biểu đồ mà không cần viết từng biểu đồ thủ công |
| Dễ dùng | Phù hợp với người mới, giúp tiết kiệm thời gian so với dùng trực tiếp Matplotlib / Seaborn |
| Phân tích dữ liệu chất lượng | Tự phát hiện missing, outliers, mô tả dữ liệu và đưa ra báo cáo nhanh |
| Tích hợp biến mục tiêu | Nếu bạn đặt depVar, AutoViz sẽ làm nổi bật các mối quan hệ với biến mục tiêu |
| Có tùy chọn giới hạn | Qua max\_rows\_analyzed, max\_cols\_analyzed giúp chạy nhanh hơn nếu dữ liệu lớn |

Han chế:

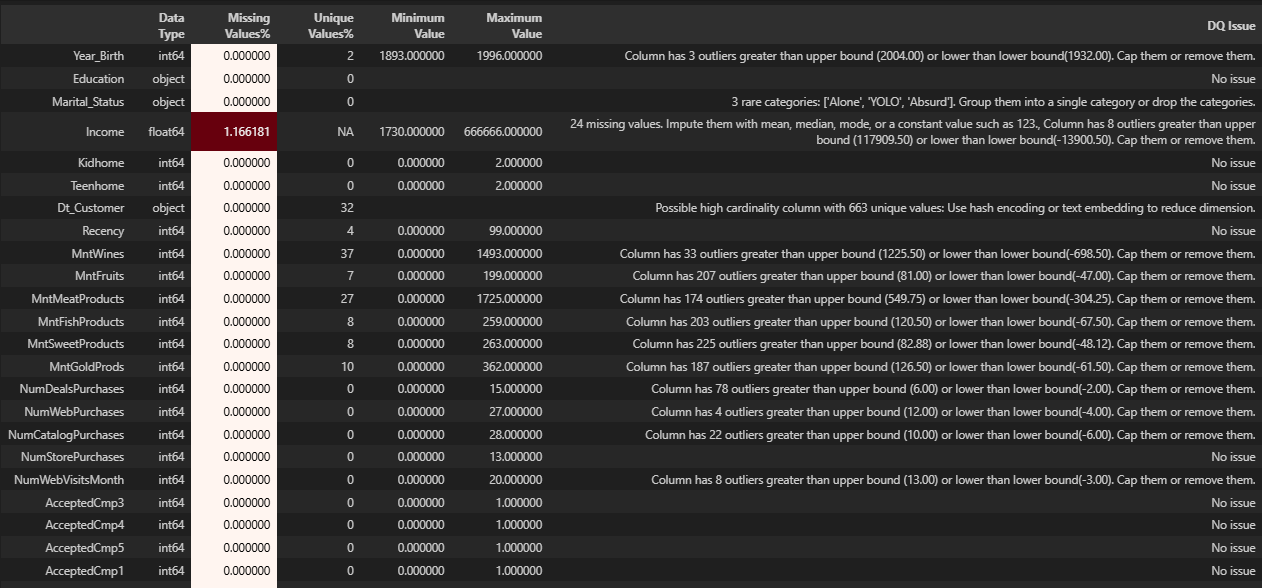
| **Hạn chế** | **Ghi chú** |
| --- | --- |
| Tùy chỉnh hạn chế | Nếu bạn muốn biểu đồ cực kỳ tùy chỉnh (color, theme, annotation), AutoViz không linh hoạt như viết tay bằng Seaborn / Plotly |
| Có thể vẽ quá nhiều | Với dữ liệu nhiều biến, AutoViz có thể sinh ra rất nhiều biểu đồ, đôi khi gây loãng thông tin |
| Hiệu suất với dữ liệu rất lớn | Mặc dù có giới hạn, nếu dữ liệu quá lớn, việc xử lý và vẽ có thể chậm hoặc bị out-of-memory |
| Không thay thế việc hiểu dữ liệu | AutoViz giúp nhanh, nhưng bạn vẫn cần có tư duy để chọn biến quan trọng, kiểm tra kỹ lưỡng kết quả |

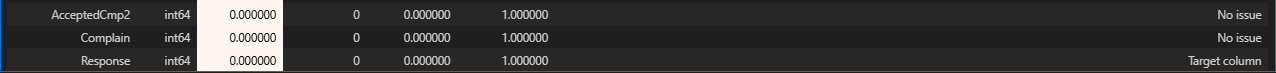
4. Áp dụng thực hành cho tập dữ liệu Marketing Campaign

4.1 Khời tạo dữ liệu tổng quan

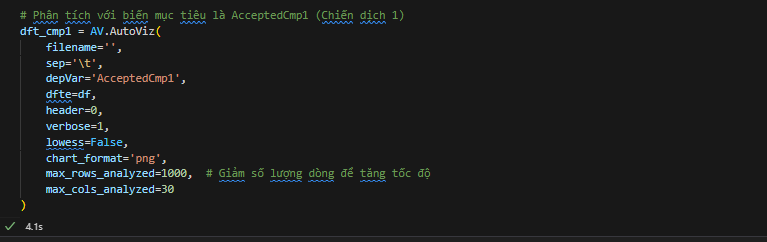


Kết quả thực hiện:

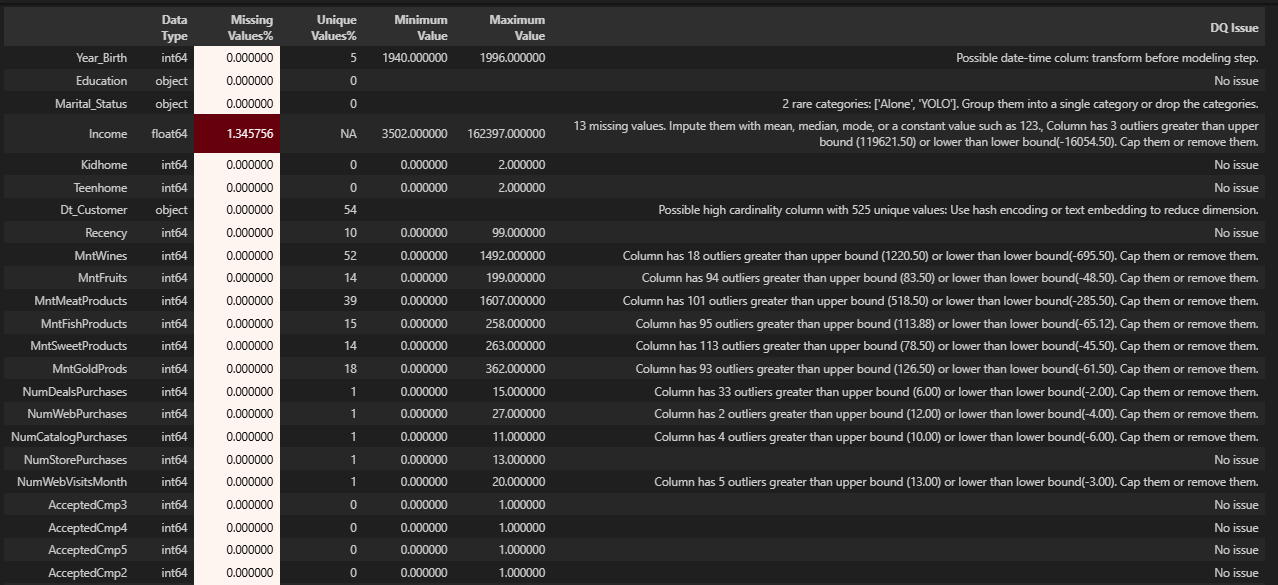


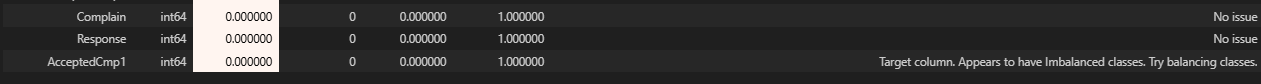


4.2. Phân tích với các biến mục tiêu khác nhau

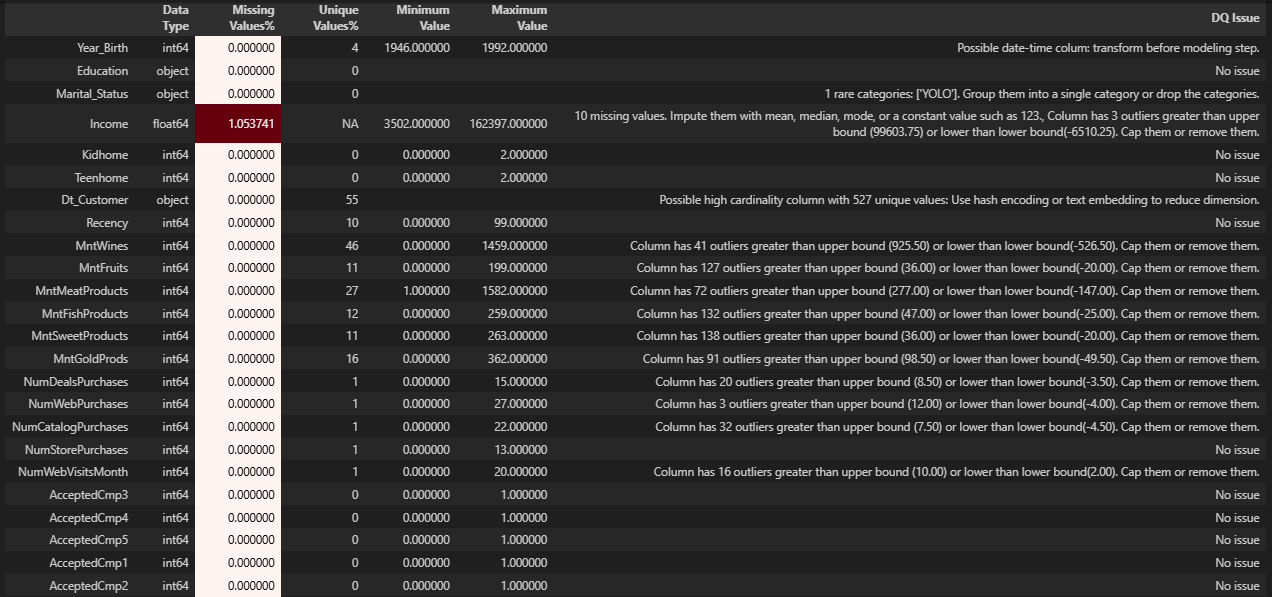


Kết quả thực hiện:





4.3. Phân tích với tập dữ liệu con





5. Nhân xét và phân tích

5.1. Ưu điểm của AutoViz

Tự động hóa cao: AutoViz tự động tạo ra nhiều loại biểu đồ khác nhau chỉ với một dòng mã, giúp tiết kiệm thời gian đáng kể.

Phát hiện vấn đề chất lượng dữ liệu: AutoViz tự động phát hiện và báo cáo các vấn đề về chất lượng dữ liệu như giá trị thiếu, giá trị trùng lặp, v.v.

Đa dạng biểu đồ: AutoViz tạo ra nhiều loại biểu đồ khác nhau như biểu đồ phân phối, biểu đồ hộp, biểu đồ thanh, biểu đồ nhiệt, biểu đồ scatter, v.v.

Phân tích theo biến mục tiêu: AutoViz tự động phân tích mối quan hệ giữa các biến và biến mục tiêu, giúp hiểu rõ hơn về các yếu tố ảnh hưởng.

Tùy chỉnh linh hoạt: AutoViz cung cấp nhiều tùy chọn để điều chỉnh quá trình phân tích, từ việc chọn biến mục tiêu đến việc bỏ qua các biến không cần thiết.

5.2. Phân tích kết quả

Từ các biểu đồ được tạo ra bởi AutoViz, chúng ta có thể rút ra một số nhận xét về tập dữ liệu Marketing Campaign:

1. Phân phối biến mục tiêu: Biến Response có sự mất cân bằng, với số lượng khách hàng phản hồi (1) ít hơn nhiều so với số lượng khách hàng không phản hồi (0).

2. Mối quan hệ giữa các biến: Biểu đồ nhiệt cho thấy có mối tương quan giữa các biến liên quan đến chi tiêu (MntWines, MntMeatProducts, v.v.) và biến mục tiêu Response.

3. Phân phối thu nhập: Biểu đồ phân phối cho thấy thu nhập của khách hàng có phân phối lệch phải, với một số lượng nhỏ khách hàng có thu nhập cao. 4. Ảnh hưởng của việc có con: Biểu đồ hộp cho thấy khách hàng có con có xu hướng chi tiêu ít hơn và ít phản hồi chiến dịch hơn so với khách hàng không có con.

5. Hiệu quả của các chiến dịch trước đó: Biểu đồ thanh cho thấy khách hàng đã phản hồi các chiến dịch trước đó có xu hướng phản hồi chiến dịch hiện tại cao hơn.

6. Kết luận

AutoViz là một công cụ mạnh mẽ cho phân tích dữ liệu khám phá tự động, đặc biệt hữu ích khi cần nhanh chóng hiểu về một tập dữ liệu mới. Với khả năng tạo ra nhiều loại biểu đồ khác nhau và phát hiện vấn đề chất lượng dữ liệu, AutoViz giúp tiết kiệm thời gian đáng kể trong quá trình EDA.

Đối với tập dữ liệu Marketing Campaign, AutoViz đã giúp chúng ta hiểu rõ hơn về các yếu tố ảnh hưởng đến tỷ lệ phản hồi của khách hàng, từ đó có thể đưa ra các quyết định marketing hiệu quả hơn.

Tuy nhiên, AutoViz cũng có một số hạn chế như không cung cấp báo cáo tương tác như SweetViz và có thể chậm khi xử lý các tập dữ liệu lớn. Vì vậy, tùy thuộc vào nhu cầu cụ thể, chúng ta có thể lựa chọn giữa AutoViz và SweetViz hoặc kết hợp cả hai để có cái nhìn toàn diện nhất về dữ liệu.