**CHƯƠNG 3: PHÂN TÍCH ĐƠN BIẾN**

1. **Lý thuyết Phân tích Đơn biến và Hai biến**
2. **Phân tích đơn biến**

* **Định nghĩa:** là kỹ thuật phân tích dữ tập trung vào việc nghiên cứu và diễn giải một biến duy nhất để mô tả đặc điểm, tìm kiếm mô hình và đo lường xu hướng trung tâm, độ phân tán của dữ liệu mà không xem xét mối quan hệ với các biến khác.
* **Mục đích:**

+ Hiểu rõ đặt điểm cơ bản của biến đó

+ Phát hiện các xu hướng hoặc mẫu hình trong dữ liệu của biến đó

* **Các kỹ thuật phổ biến:**

+ **Thống kê tóm tắt**: Sử dụng các chỉ số như giá trị trung bình (mean), trung vị (median), giá trị lớn nhất (max), giá trị nhỏ nhất (min), độ lệch chuẩn (standard deviation) để tóm tắt dữ liệu.

+ **Biểu đồ:** Sử dụng biểu đồ cột, biểu đồ thanh, hoặc biểu đồ tần suất để trực quan hóa dữ liệu, giúp dễ dàng nhận thấy xu hướng.

1. **Phân tích hai biến**

* Định nghĩa: Phân tích hai biến là một phương pháp thống kê nhằm kiểm tra và xác định mối quan hệ giữa hai biến riêng biệt, cho phép nhà nghiên cứu hiểu được liệu chúng có liên quan với nhau hay không, mức độ mạnh yếu của mối liên hệ, và liệu một biến có thể được sử dụng để dự đoán biến kia hay không.
* Mục đích:

+ Kiểm tra giả thuyết: Giúp xác định các giả thuyết đơn giản về mối liên hệ giữa hai yếu tố.

+ Xác định mối quan hệ: Tìm ra liệu hai biến có tương quan (liên quan) với nhau hay không.

* **Đánh giá cường độ và hướng mối quan hệ:** Nếu có mối quan hệ, phân tích này cho biết mối quan hệ đó mạnh hay yếu, và nó là mối quan hệ thuận hay nghịch chiều.
* Vậy sự khác nhau chủ yếu là Phân tích đơn biết mục đích để kiếm tra chính biến đó thay đổi như thế nào, Phân tích 2 biến nhằm mục đích kiểm tra mối quan hệ coi chúng có liên quan, ảnh hưởng với nhau không.

1. **Thước đo thống kê phân tích đơn biến**

* **Thống kê mô tả trung tâm:**

+ **Trung bình (Mean)**: Là tổng tất cả các giá trị chia cho số lượng các giá trị. Đây là thước đo phổ biến nhất.

* **Công dụng:** Cho ta một con số đại diện cho toàn bộ tập dữ liệu.
* **Nhược điểm:** Rất nhạy cảm với các giá trị ngoại lai (outliers). Một giá trị quá lớn hoặc quá nhỏ có thể kéo giá trị trung bình lệch đi đáng kể.

+ **Trung vị (Median)**:Là giá trị nằm chính giữa của tập dữ liệu sau khi đã được sắp xếp theo thứ tự từ nhỏ đến lớn.

* **Công dụng:** Là thước đo tốt hơn giá trị trung bình khi dữ liệu có outliers hoặc bị lệch (skewed).
* **Cách tìm:** Nếu số lượng giá trị là lẻ, trung vị là số ở giữa. Nếu là chẵn, trung vị là trung bình của hai số ở giữa.

+ **Mode**: Giá trị xuất hiện thường xuyên nhất trong tập dữ liệu.

* **Công dụng:** Hữu ích nhất cho dữ liệu phân loại (categorical data). Một tập dữ liệu có thể không có mode, có một mode (unimodal) hoặc nhiều mode (bimodal, multimodal).
* **Các thước đo độ phân tán:** Nhóm này cho chúng ta biết dữ liệu trải rộng hay co cụm quanh giá trị trung tâm như thế nào.

+ Khoảng giá trị (Range):

* **Định nghĩa:** Là hiệu số giữa giá trị lớn nhất (max) và giá trị nhỏ nhất (min).
* **Công dụng:** Cung cấp cái nhìn nhanh nhất về sự biến thiên của dữ liệu.
* **Nhược điểm:** Giống như giá trị trung bình, nó rất nhạy cảm với outliers vì chỉ phụ thuộc vào hai giá trị đầu và cuối.

+ Phương sai (Variance):

* Là trung bình của bình phương khoảng cách từ mỗi điểm dữ liệu đến giá trị trung bình.
* **Công dụng:** Đo lường mức độ phân tán của dữ liệu. Phương sai lớn có nghĩa là dữ liệu trải rộng, phương sai nhỏ có nghĩa là dữ liệu co cụm gần giá trị trung bình.
* **Nhược điểm:** Đơn vị của nó là bình phương đơn vị của dữ liệu gốc (ví dụ: nếu dữ liệu là mét, phương sai là mét vuông), gây khó khăn trong việc diễn giải.

+ Độ lệch chuẩn (Standard Deviation):

* Là căn bậc hai của phương sai.
* **Công dụng:** Đây là thước đo độ phân tán phổ biến và quan trọng nhất. Nó cho biết trung bình mỗi điểm dữ liệu cách giá trị trung bình bao xa.
* **Ưu điểm:** Có cùng đơn vị với dữ liệu gốc, giúp việc diễn giải trở nên trực quan và dễ hiểu hơn nhiều so với phương sai.
* **Ví dụ:** Một độ lệch chuẩn nhỏ cho thấy các điểm dữ liệu có xu hướng rất gần với giá trị trung bình, trong khi độ lệch chuẩn lớn cho thấy các điểm dữ liệu trải rộng trên một phạm vi giá trị lớn hơn.

+ Khoảng tứ phân vị (Interquartile Range - IQR): Là khoảng cách giữa tứ phân vị thứ ba (Q3 - 75%) và tứ phân vị thứ nhất (Q1 - 25%) là "khoảng giá trị" của 50% dữ liệu ở giữa.

* **Công dụng:** Để đo độ phân tán khi dữ liệu có outliers hoặc bị lệch, vì nó không bị ảnh hưởng bởi 25% giá trị nhỏ nhất và 25% giá trị lớn nhất.

1. **Xác định mối quan hệ trong phân tích hai biến**

Phân tích hai biến là quá trình nghiên cứu mối quan hệ giữa hai biến (hay còn gọi là hai tập dữ liệu) để xem chúng có liên quan, ảnh hưởng, hoặc đồng biến với nhau hay không.

1. Xác định Mối quan hệ (Tương quan): tìm hiểu xem hai biến có "cùng nhau di chuyển" hay không. Tức là khi một biến thay đổi, biến kia có xu hướng thay đổi theo một quy luật nào đó không. Phương pháp sẽ phụ thuộc vào loại dữ liệu của hai biến mà ta xét.

+ Khi cả hai biến đều là biến số (Numerical)

* Công cụ trực quan hóa: Biểu đồ phân tán (Scatter Plot) giúp ta có thể **nhìn thấy** hình dạng của mối quan-hệ.
  + **Tương quan dương:** Các điểm dữ liệu có xu hướng đi lên từ trái sang phải (khi biến X tăng, biến Y cũng có xu hướng tăng).
  + **Tương quan âm:** Các điểm dữ liệu có xu hướng đi xuống từ trái sang phải (khi biến X tăng, biến Y có xu hướng giảm).
  + **Không tương quan:** Các điểm dữ liệu phân tán ngẫu nhiên như một đám mây, không có quy luật rõ ràng.
  + **Quan hệ phi tuyến:** Các điểm có thể tạo thành một đường cong (hình chữ U, chữ J, v.v.), cho thấy có mối quan-hệ nhưng không phải là đường thẳng.

Green dots in the sky

AI-generated content may be incorrect.

*Hình 1.3.1.1 Ví dụ về biểu đồ Scatter Plot*

* Thước đo thống kê: Hệ số tương quan (Correlation Coefficient): Hệ số này định lượng độ mạnh và hướng của mối quan-hệ **tuyến tính**.

**Hệ số tương quan Pearson (r):** Phổ biến nhất. Giá trị của nó dao động từ -1 đến +1.

* **r = +1:** Tương quan dương tuyến tính hoàn hảo.
* **r = -1:** Tương quan âm tuyến tính hoàn hảo.
* **r = 0:** Không có tương quan tuyến tính.

Quy ước về độ mạnh:

* |r| > 0.7: Tương quan mạnh.
* 0.5 < |r| < 0.7: Tương quan vừa phải.
* 0.3 < |r| < 0.5: Tương quan yếu.
* |r| < 0.3: Tương quan rất yếu hoặc không đáng kể.

+ Khi một biến là Phân loại (Categorical) và một biến là Số (Numerical): Mục tiêu là so sánh giá trị của biến số qua các nhóm khác nhau của biến phân loại.

Công cụ trực quan hóa:

* **Biểu đồ hộp theo nhóm (Grouped Box Plots):** Rất hiệu quả để so sánh sự phân phối (trung vị, IQR, outliers) của biến số giữa các nhóm.

A black background with red lines

AI-generated content may be incorrect.

*Hình 1.3.1.2 Ví dụ biểu đồ* ***Box Plots***

* **Biểu đồ Violin (Violin Plots):** Tương tự biểu đồ hộp nhưng thể hiện cả mật độ phân phối của dữ liệu.

A graph of a graph

AI-generated content may be incorrect.

*Hình 1.3.1.3 Ví dụ về Violin Plot*

* **Biểu đồ cột với thanh lỗi (Bar chart with error bars):** So sánh giá trị trung bình và độ không chắc chắn (ví dụ: độ lệch chuẩn hoặc sai số chuẩn) của các nhóm.

A diagram of a graph

AI-generated content may be incorrect.

*Hình 1.3.1.4 Bar chart with error bars*

Kiểm định thống kê:

* **T-test (2 nhóm):** Dùng để kiểm tra xem có sự khác biệt đáng kể về mặt thống kê giữa giá trị trung bình của hai nhóm hay không.
* **ANOVA (Analysis of Variance) (>2 nhóm):** Dùng khi có từ ba nhóm trở lên, để kiểm tra xem có ít nhất một nhóm có giá trị trung bình khác biệt so với các nhóm còn lại hay không.

+ Khi cả hai biến đều là biến Phân loại (Categorical): Mục tiêu là xem xét liệu có mối liên-kết (association) giữa các danh mục của hai biến này hay không.

1. Phân biệt Tương quan và Nhân quả: "Tương quan không bao hàm nhân quả"

**Ví dụ:** Vào **mùa hè**, **doanh số bán kem** và **số vụ đuối nước** *đều tăng lên*. Chúng có tương quan dương rất mạnh. Nhưng điều này không có nghĩa là ăn kem gây ra đuối nước.

* **Vậy Tại sao tương quan không phải là nhân quả?**

+ **Biến ẩn/Biến nhiễu (Confounding Variable):** Thường có một biến thứ ba (C) tác động lên cả A và B.

*Trong ví dụ trên*, biến ẩn là **nhiệt độ thời tiết**. Thời tiết nóng (C) khiến người ta đi bơi nhiều hơn (dẫn đến tăng nguy cơ đuối nước - A) và cũng khiến người ta ăn kem nhiều hơn (B).

**+ Quan hệ ngược (Reverse Causality):** Có thể B gây ra A thay vì A gây ra B.

+ **Quan hệ ngẫu nhiên (Spurious Correlation):** Hai biến hoàn toàn không liên-quan nhưng lại cho thấy một mối tương quan mạnh chỉ do sự trùng hợp ngẫu nhiên trong dữ liệu.

1. **Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance)**

Tương quan (Correlation) chính là phiên bản được chuẩn hóa của Hiệp phương sai (Covariance).

1. Định nghĩa và Mục đích

* Hiệp phương sai (Covariance):

+ **Mục đích:** Đo lường **chiều hướng** của mối quan hệ tuyến tính giữa hai biến. Nó cho chúng ta biết liệu hai biến có xu hướng *cùng nhau thay đổi* hay không.

+ **Cách hoạt động:** Nó tính toán mức độ mà hai biến cùng nhau di chuyển ra khỏi giá trị trung bình của chúng.

* **Hiệp phương sai dương (> 0):** Khi biến X tăng, biến Y cũng có xu hướng tăng.
* **Hiệp phương sai âm (< 0):** Khi biến X tăng, biến Y có xu hướng giảm.
* **Hiệp phương sai bằng 0:** Không có mối quan hệ tuyến tính nào giữa hai biến.
* Tương quan (Correlation):

+ **Mục đích:** Đo lường cả **chiều hướng và độ mạnh** của mối quan hệ tuyến tính giữa hai biến.

+ **Cách hoạt động:** Về bản chất, tương quan là hiệp phương sai được "chuẩn hóa" bằng cách chia cho tích độ lệch chuẩn của hai biến.

Correlation(X,Y)= Covariance(X,Y)​/ Standard Deviation(X, Y)

1. Thang đo và Đơn vị

* Hiệp phương sai:

+ **Khoảng giá trị:** Không giới hạn, có thể chạy từ -∞ đến + ∞.

+ **Đơn vị:** Bị phụ thuộc vào đơn vị của hai biến. Ví dụ: tính hiệp phương sai giữa chiều cao (mét) và cân nặng (kg), đơn vị của hiệp phương sai sẽ là "mét-kg".

+ **Hạn chế:** Vì giá trị của nó không được chuẩn hóa, nên không thể so sánh trực tiếp hiệp phương sai giữa các cặp biến khác nhau. Một hiệp phương sai là 2000 không nhất thiết chỉ ra mối quan hệ mạnh hơn một hiệp phương sai là 0.5 nếu chúng được tính từ các bộ dữ liệu với đơn vị và thang đo khác nhau.

* Tương quan:

+ **Khoảng giá trị:** Luôn luôn nằm trong khoảng từ **-1 đến +1**.

+ **Đơn vị:** **Không có đơn vị** (dimensionless).

**+ Ưu điểm:** Chính vì được chuẩn hóa và không có đơn vị, hệ số tương quan cho phép chúng ta:

* **Đánh giá độ mạnh** của mối quan hệ một cách khách quan.
* **So sánh độ mạnh** của các mối quan hệ giữa các cặp biến khác nhau (ví dụ: so sánh mối tương quan giữa "chiều cao-cân nặng" với "thu nhập-chi tiêu").

1. Kết luận

- Hiệp phương sai:

+ Chỉ có thể diễn giải **dấu** của nó (dương, âm, hoặc bằng 0) để biết chiều hướng của mối quan hệ.

+ Và **Không thể** dựa vào độ lớn của hiệp phương sai để kết luận về độ mạnh của mối quan hệ.

* Tương quan có thể diễn giải cả **dấu và độ lớn** của nó:

+ **Dấu:** Cho biết chiều hướng (dương là cùng chiều, âm là ngược chiều).

+ **Độ lớn (giá trị tuyệt đối):** Cho biết độ mạnh của mối quan hệ tuyến tính.

* Gần 1 (hoặc -1): Mối quan hệ tuyến tính rất mạnh.
* Gần 0: Mối quan hệ tuyến tính rất yếu hoặc không tồn tại.

1. **Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến?**

* **Phân tích đơn biến:** Sử dụng khi muốn có một cái nhìn sơ bộ về nó như sự biến động, hình dạng, trung tâm, phân tán.
* Sử dụng khi:

+ Muốn hiểu phân phối của một biến

+ Tìm kiếm outliers

+ Kiểm tra tính chuẩn của dữ liệu

+ So sánh các nhóm khác nhau của cùng một biến

* **Phân tích hai biến:** Sử dụng trực quan hóa trong phân tích hai biến khi bạn muốn hiểu mối quan hệ, tương quan, và cách một biến thay đổi khi biến kia thay đổi.
* Sử dụng khi:

+ Muốn khám phá mối quan hệ giữa hai biến

+ So sánh các nhóm khác nhau

+ Tìm kiếm patterns hoặc clusters

+ Chuẩn bị cho modeling

1. **Mẫu code ví dụ về tạo biểu đồscatter plot**

Biểu đồ này là công cụ chính để phân tích mối quan hệ giữa **hai biến định lượng** (Quantitative Variables), giúp thấy rõ **tương quan** (correlation).

A screenshot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.1.5 Code để tạo Biểu đồ*

* Kết quả:

A graph with blue dots

AI-generated content may be incorrect.

*Hình 1.3.1.5 Kết quả khi chạy code tạo biểu đồ*

+ Biểu đồ sẽ hiển thị các điểm gần như nằm trên một đường thẳng có độ dốc dương.

+ Điều này cho thấy có một mối tương quan dương mạnh giữa hai biến: Kích thước nhà càng lớn, giá bán có xu hướng càng cao (đồng biến).

1. **Cách trực quan hóa mối quan hệ giữa một biến số và một biến phân loại bằng biểu đồ boxplot**

* Biến định lượng: là biến giá trị, thể hiện bằng những con số
* Biến phân loại: ví dụ: giới tính, khu vực, loại sản phẩm
* Ở đây chúng ta sẽ sử dụng biểu đồ hộp để trực quan hóa mối quan hệ này:

+ Giả sử chúng ta có một tập dữ liệu về các bữa ăn tại một nhà hàng (tips dataset, có sẵn trong Seaborn) và chúng ta muốn tìm hiểu mối quan hệ giữa ngày trong tuần (biến phân loại) và tổng hóa đơn (biến số).

**Bước 1:** Chuẩn bị môi trường và dữ liệu

Cài đặt thư viện cần thiết **pip install pandas seaborn matplotlib**

A screenshot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.1.6 Code khai báo tậo dữ liệu*

**Bước 2:** Trực quan hóa bằng Biểu đồ Hộp (Box Plot)

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.1.7 Code tạo biểu đồ*

* Kết quả:

A diagram of a graph

AI-generated content may be incorrect.

*Hình 1.3.1.8 Biểu đồ hộp*

* Diễn giải kết quả từ Box Plot:

+ **So sánh Trung vị (Median):** Đường kẻ ngang bên trong mỗi hộp là giá trị trung vị. Nhìn vào biểu đồ, chúng ta có thể thấy trung vị của tổng hóa đơn vào cuối tuần (Thứ 7 - Sat, Chủ Nhật - Sun) cao hơn so với các ngày trong tuần (Thứ 5 - Thur, Thứ 6 - Fri).

+ **So sánh Độ phân tán (Spread):** Chiều dài của hộp (khoảng tứ phân vị - IQR) cho biết mức độ phân tán của 50% dữ liệu ở giữa. Hộp của ngày Thứ 7 và Chủ Nhật dài hơn, cho thấy tổng hóa đơn vào những ngày này có sự biến động lớn hơn.

+ **Phát hiện Outliers:** Các chấm tròn nằm ngoài "râu" của biểu đồ là các giá trị ngoại lai. Chúng ta thấy rằng tất cả các ngày đều có một vài hóa đơn cao bất thường.

**Bước 3:** Trực quan hóa bằng Biểu đồ Violin (Violin Plot) với cùng tập dataset ở trên: Biểu đồ violin kết hợp ưu điểm của biểu đồ hộp và biểu đồ mật độ (density plot). Ngoài việc hiển thị các thông số như trung vị, IQR, nó còn cho thấy hình dạng phân phối đầy đủ của dữ liệu.

A computer screen shot of a code

AI-generated content may be incorrect.

*Hình 1.3.1.9 Code để tạo biểu đồ Violin*

+ Kết quả:

A diagram of a blue leaf

AI-generated content may be incorrect.

*Hình 1.3.1.10 Biểu đồ Violin*

* Diễn giải kết quả từ Violin Plot:

+ **Hình dạng Phân phối:** "Thân" của violin cho thấy mật độ dữ liệu. Phần thân càng rộng, mật độ dữ liệu tại giá trị đó càng cao. Ví dụ, vào Thứ 5 (Thur), có một sự tập trung lớn các hóa đơn quanh khoảng 12-18$.

+ **So sánh với Box Plot:** Biểu đồ violin cung cấp nhiều thông tin hơn về hình dạng. Chúng ta có thể thấy rằng vào Thứ 7 (Sat), dữ liệu có vẻ tập trung ở nhiều mức giá khác nhau, trong khi vào Thứ 5 thì tập trung rõ rệt hơn ở một khoảng hẹp.

+ **Các thành phần bên trong:** Dấu chấm trắng là trung vị. Thanh màu đen dày là khoảng tứ phân vị (IQR). Đường kẻ đen mỏng là khoảng tin cậy 95%.

1. **Làm quen với các hàm và thư viện hỗ trợ phân tích dữ liệu đơn biến.**

**1. Bài toán 1:** 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

* Phân tích đơn biến bằng Histogram

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.1 Import thư viện và tạo biểu đồ*

A graph of a number of blue bars

AI-generated content may be incorrect.

*Hình 1.3.2.2 Biểu đồ Histogram của cột culmen\_length\_mm*

* Phân bố của culmen\_length\_mm có dạng xấp xỉ chuẩn (bell-shaped).
* Giá trị phổ biến nhất (mode) rơi vào khoảng 40–50 mm — đây là chiều dài mỏ trung bình của đa số chim.
* Phạm vi dữ liệu dao động từ khoảng 30 mm đến 60 mm, cho thấy sự đa dạng giữa các loài (Adelie, Chinstrap, Gentoo).
* Có thể xuất hiện nhiều đỉnh (multi-modal distribution). Ví dụ:

+ Một đỉnh quanh 38 mm (loài Adelie).

+ Một đỉnh khác quanh 48–55 mm (loài Gentoo hoặc Chinstrap).

* Điều này cho thấy biến culmen\_length\_mm có thể dùng để phân biệt loài chim
* Phân tích đơn biến bằng Bar chart

A screen shot of a computer code

AI-generated content may be incorrect.

*Hình 1.3.2.3 Code tọ biểu đồ*

A graph of different species

AI-generated content may be incorrect.

*Hình 1.3.2.4 Biểu đồ Bar chart*

+ Loài chim Adelie: Khoảng 150 mẫu, Là loài phổ biến nhất trong dataset.

+ Loài chim Chinstrap: Khoảng 68 mẫu, Ít hơn Adelie, nhưng đủ đại diện.

+ Loài chim Gentoo: Khoảng 124 mẫu, Đứng thứ hai về số lượng.

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

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.5 Code tạo biểu đồ*

A pie chart with different colored circles

AI-generated content may be incorrect.

*Hình 1.3.2.6 Biểu đồ Pie-Chart*

+ Loài chim Adelie: Khoảng 150 mẫu, ~ 44%

+ Loài chim Chinstrap: Khoảng 68 mẫu, ~ 20%

+ Loài chim Gentoo: Khoảng 124 mẫu, ~ 36%

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

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

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.7 Import thư viện và tạo biểu đồ*

A graph of a bar graph

AI-generated content may be incorrect.

*Hình 1.3.2.8 Biểu đồ Box*

* Phân tích dữ liệu đơn biến dựa vào violin plot

A screen shot of a computer code

AI-generated content may be incorrect.

*Hình 1.3.2.9 Code tạo biểu đồ Violin*

A graph of a function

AI-generated content may be incorrect.

*Hình 1.3.2.10 Biểu đồ Violin*

* Phân tích dữ liệu đơn biến dựa vào bản tóm tắt dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.2.11 Bảng thống kê mô tả*

* Nhận xét:
  1. Biến Price (Giá nhà)

+ Biến Price đại diện cho giá bán của từng căn nhà trong tập dữ liệu.

+ Giá nhà trung bình là 622,065, trong khi giá trung vị chỉ đạt 467,000, cho thấy phân phối dữ liệu bị lệch phải (right-skewed) — tức là tồn tại một số căn nhà có giá rất cao làm kéo trung bình lên.

+ Khoảng giá dao động lớn (từ 175,000 đến 5,950,000) chứng tỏ dữ liệu bao gồm nhiều phân khúc thị trường từ nhà bình dân đến cao cấp.

+ Giá trị lớn nhất (5,950,000) có thể được xem là outlier, đại diện cho nhà biệt thự hoặc bất động sản đặc biệt.

* Kết luận: Dữ liệu giá nhà có sự phân tán cao, lệch phải mạnh, nên cần xem xét chuẩn hóa hoặc log-transform trước khi xây dựng mô hình dự đoán.
  1. Biến Area (Diện tích nhà)

+ Biến Area biểu thị diện tích của ngôi nhà (đơn vị m²).

Giá trị trung bình đạt 95.95 m², với độ lệch chuẩn 57.45, thể hiện mức biến động lớn giữa các loại hình nhà ở.

+ Phần lớn nhà có diện tích từ 60–120 m², phù hợp với nhóm nhà phố hoặc căn hộ trung bình.

+ Tuy nhiên, diện tích tối đa lên tới 623 m², cho thấy có sự xuất hiện của các căn biệt thự cao cấp — đây cũng là các outlier tiềm năng cần lưu ý trong quá trình phân tích.

* Kết luận: Biến Area có phân phối lệch phải nhẹ, tập trung chủ yếu quanh 80–100 m², nhưng tồn tại một số mẫu diện tích lớn cần được xử lý.
  1. Biến Room (Số lượng phòng)

+ Biến Room phản ánh số phòng trong mỗi căn nhà.

+ Trung bình, mỗi căn có 3.57 phòng, độ lệch chuẩn 1.59 cho thấy mức phân tán tương đối thấp — đa số nhà tập trung quanh 3–4 phòng.

+ Giá trị nhỏ nhất là 1 phòng (studio hoặc căn hộ nhỏ), trong khi giá trị lớn nhất là 14 phòng, có thể là biệt thự hoặc nhà nhiều tầng.

* Kết luận: Phân phối của biến Room gần chuẩn (near-normal), không có sự lệch rõ rệt, phù hợp để sử dụng làm biến dự đoán trong mô hình hồi quy.
  1. Nhận xét tổng thể

+ Price và Area có tương quan dương tiềm năng — nhà diện tích lớn thường có giá cao.

+ Room là biến rời rạc, nhưng phản ánh khá tốt quy mô nhà ở.

+ Tồn tại một số outliers ở biến Price và Area cần xử lý trong các bước tiền xử lý dữ liệu.

+ Có thể áp dụng các kỹ thuật như log-transformation, standardization hoặc winsorization để cải thiện tính ổn định cho mô hình học máy.

**2. 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.

A screenshot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.12 Chuẩn bị dữ liệu và tạo scatterplot*

A diagram of a number of blue dots

AI-generated content may be incorrect.

*Hình 1.3.2.13 Kết quả biểu đồ scatterplot*

Phân tích kết quả quan sát được:

+ Xu hướng các điểm tăng dần từ trái sang phải (càng dài mỏ → càng nặng cân), ta có thể rút ra:

+ Mối quan hệ dương (positive correlation) giữa chiều dài mỏ và khối lượng cơ thể.

* Chim có mỏ dài thường có khối lượng lớn hơn.

+ Nếu các điểm nằm rải rác nhưng vẫn có xu hướng tăng nhẹ theo đường chéo, đó là mối quan hệ tuyến tính yếu (weak positive correlation).

+ Nếu dữ liệu tạo thành các cụm rõ ràng, mỗi cụm có thể tương ứng với một loài chim khác nhau (ví dụ: Adelie, Chinstrap, Gentoo).

+ Một vài điểm nằm xa khỏi vùng tập trung → outliers, có thể do lỗi đo hoặc cá thể khác biệt.

* Phân tích 2 biến dựa vào bảng crosstab/two-way

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.14 Thống kê crosstab/two-way*

* Thống kê giới tính của từng loài

A screenshot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.15 Code tạo bảng Pivot table*

* Phân tích 2 biến sử dụng pairplot

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.2.16 Code tạo biểu đồ Pairplot*

Nhận xét

+ Biến culmen\_length\_mm có phân phối tập trung chủ yếu trong khoảng 40–50 mm.

+ Có sự xuất hiện của một số cá thể đặc biệt với mỏ dài hơn 55 mm.

+ Dữ liệu này phù hợp để sử dụng cho phân tích thống kê hoặc huấn luyện mô hình phân loại loài chim dựa trên đặc điểm hình thái.

**3. 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.

* Giả sử ta có tệp tin dữ liệu marketing\_campaign, file này chứa thông tin mô phỏng của một chiến dịch marketing. Các bản ghi thể hiện thông tin khách hàng và hành vi tham gia chiến dịch.
* Các trường dữ liệu có trong file:

+ CustomerID – Mã định danh duy nhất của khách hàng.

+ Age – Tuổi của khách hàng.

+ Gender – Giới tính (Male / Female).

+ Income – Mức thu nhập hàng năm (USD).

+ Campaign\_Joined – Tên chiến dịch marketing mà khách hàng tham gia (Campaign\_A, Campaign\_B, v.v.).

+ Purchase\_Amount – Số tiền mà khách hàng đã chi tiêu trong chiến dịch.

+ Channel – Kênh tiếp thị mà khách hàng tiếp cận (Email, Social Media, TV, Web, v.v.).

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.17 Code tạo báo cáo html*

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.2.18 Giao diện báo cáo html*

A screen shot of a computer program

AI-generated content may be incorrect.

*Hình 1.3.2.19 Code để mở giao diện D-Tale*

1. **Giới thiệu các tính năng và cách sử dụng thư viện SweetViz**
   1. **Giới thiệu thư viện SweetViz**

* **Sweetviz** là thư viện Python mã nguồn mở dùng để **Exploratory Data Analysis (EDA) tự động**. Nó tạo ra báo cáo HTML trực quan, tương tác với các phân tích thống kê chi tiết.
* Mục tiêu của nó là hỗ trợ để có một cái nhìn tổng quan dữ liệu mình có: phân bố, mối liên hệ giữa biến, giá trị thiếu, so sánh giữa các tập dữ liệu (ví dụ train vs test), và phân tích biến mục tiêu nếu có.
  1. **Những tính năng chính của SweetViz**
* **Phân tích biến mục tiêu (Target analysis):** Điều này cho thấy giá trị mục tiêu liên quan như thế nào đến các tính năng khác.
* **So sánh hai tập dữ liệu:** Bạn có thể so sánh hai DataFrame (ví dụ: train và test) để xem phân bố biến, giá trị thiếu, mối quan hệ giữa các biến trong mỗi tập.
* **Liên kết giữa các loại biến hỗn hợp:** Với dữ liệu mà có biến số (numerical), biến phân loại (categorical) hoặc hỗn hợp, SweetViz dùng các phép đo tương thích (ví dụ Pearson, hệ số không chắc chắn, correlation ratio) để đánh giá mối liên hệ.
* **Phát hiện kiểu dữ liệu tự động / ghi đè thủ công:** SweetViz tự nhận biết biến số, phân loại, văn bản, nhưng bạn có thể ghi đè nếu muốn.
* **Báo cáo thống kê & tổng hợp:** Cho mỗi biến: số lượng, giá trị thiếu, giá trị duy nhất, giá trị hay gặp nhất, min, max, trung vị, trung bình, độ lệch chuẩn, độ lệch (skewness), độ nhọn (kurtosis), hệ số biến động, v.v.

* 1. **Cài đặt và sử dụng**

**Bước 1:** Cài đặt thư viện

```

*pip install sweetviz*

```

Ở trong Terminal hoặc PowerShell

A screenshot of a black screen

AI-generated content may be incorrect.

*Hình 1.3.3.1 Giao diện cài đặt thư viện*

**Lưu ý:** Phiên bản sweetviz lần cuối được cập nhật là khoảng tháng 11/2023, dẫn đến việc các phiên bản python và numpy (do sweetviz yêu cầu numpy) mới hơn xảy ra vấn đề không tương thích. Khuyến nghị khi sử dụng sweetviz nên cài python khoảng ver 3.9 và numpy ver 1.23.4 để đảm bảo hoạt động đúng.

**Bước 2:** Nhập thư viện & chuẩn bị dữ liệu

* Tải dữ liệu mẫu tại: [kaggle](https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis)

+ **Bối cảnh dữ liệu:** Phân tích Tính cách Khách hàng là một phân tích chi tiết về khách hàng lý tưởng của một công ty. Nó giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình và giúp họ dễ dàng điều chỉnh sản phẩm theo nhu cầu, hành vi và mối quan tâm cụ thể của từng nhóm khách hàng khác nhau.

+ Phân tích tính cách khách hàng giúp doanh nghiệp điều chỉnh sản phẩm dựa trên khách hàng mục tiêu từ các phân khúc khách hàng khác nhau. Ví dụ, thay vì chi tiền để tiếp thị sản phẩm mới cho tất cả khách hàng trong cơ sở dữ liệu của công ty, doanh nghiệp có thể phân tích phân khúc khách hàng nào có khả năng mua sản phẩm cao nhất và sau đó chỉ tiếp thị sản phẩm cho phân khúc cụ thể đó.

* Một số thông tin về tập dữ liệu:

+ Thông tin về người (People):

* ID: Mã định danh khách hàng
* Year\_Birth: năm sinh khách hàng
* Education: trình độ học vấn của khách hàng
* Marital\_Status: Tình trạng hôn nhân (8 loại bao gồm các giá trị bất thường như "YOLO", "Absurd", "Alone")
* Income: Thu nhập hàng năm (có giá trị missing)
* Kidhome: Số trẻ em nhỏ trong gia đình
* Teenhome: Số thanh thiếu niên trong gia đình
* Dt\_customer: Ngày khách hàng đăng ký với công ty
* Recency: Số ngày kể từ lần mua hàng cuối cùng của khách hàng
* Complain: 1 nếu khách hàng đã khiếu nại trong 2 năm qua, 0 nếu không

+ Hành vi mua hàng (products):

* MntWines: Số tiền chi cho rượu vang trong 2 năm qua
* MntFruits: Số tiền chi cho trái cây trong 2 năm qua
* MntMeatProducts: Số tiền chi cho thịt trong 2 năm qua
* MntFishProducts: Số tiền chi cho hải sản trong 2 năm qua
* MntSweetProducts: Số tiền chi cho đồ ngọt trong 2 năm qua
* MntGoldProds: Số tiền chi cho vàng trong 2 năm qua

+ Khuyến mãi (promotion):

* NumDealsPurchases: Số lượng mua hàng được giảm giá
* AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise. 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch đầu tiên, 0 nếu không
* AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise. Lần thứ 2
* AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise. Lần thứ 3
* AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise. Lần thứ 4
* AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise. Lần thứ 5
* Response: 1 if customer accepted the offer in the last campaign, 0 otherwise. Lần cuối.

+ Địa điểm (Place):

* NumWebPurchases: Số lượng mua hàng được thực hiện thông qua trang web của công ty
* NumCatalogPurchases: Số lần mua hàng qua Catalog
* NumStorePurchases: Số lần mua hàng trực tiếp tại cửa hàng
* NumWebVisitsMonth: Số lần truy cập web mỗi tháng

**Bước 3:** Đọc dữ liệu bằng Pandas

```

*import pandas as pd*

*df = pd.read\_csv('marketing\_campaign.csv', delimiter='\t')*

```

**Bước 4:** Tạo báo cáo

```

*report = sv.analyze(df)*

*report.show\_html("marketing\_campaign.html")*

```

* SweetViz sẽ:

+ # Sweetviz tự động:

* Đếm số lượng: rows, columns
* Xác định kiểu dữ liệu: numeric, categorical, text, datetime
* Phát hiện duplicate rows
* Tính tỷ lệ missing values cho từng cột

Phân tích từng biến (Per-Feature Analysis):

Với biến số (Numerical) là các thống kê mô tả, Biểu đồ phân phối

Với biến phân loại (Categorical) là các count của biến phân loại

Tạo cấu trúc báo cáo (Report Generation)

Render HTML

Sau khi tạo xong chỉ cần mở nó chạy trên trình duyệt.

* Kết quả

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.2 Giao diện tổng quan của báo cáo*

* Biểu đồ Associations được hiển thị khi bạn bấm vào:

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.3 Vị trí nhấn để xem Associations*

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.4 Giao diện của Associations*

* Mức ý nghĩa:

**+ Hình tròn (circles)**: là *correlation* giữa hai biến **số** (Pearson).

* Kích thước vòng / kích thước điểm → độ lớn tuyệt đối của hệ số tương quan (|r|).
* Màu sắc cho biết **dấu**: trong báo cáo của bạn, **xanh đậm ≈ +1**, **đỏ ≈ -1**, màu nhạt ≈ gần 0.
* Vòng to + xanh dương: tương quan dương mạnh; vòng to + đỏ: tương quan âm mạnh.

**+ Hình vuông (squares)**: là chỉ số hiệp hội cho **biến phân loại / hỗn hợp** (SweetViz dùng uncertainty coefficient / correlation ratio).

* Vuông càng đậm/càng to → mối liên hệ mạnh (phạm vi 0..1). **Lưu ý: đây là chỉ số bất đối xứng** — nghĩa là hàng → cột chưa chắc bằng cột → hàng; SweetViz hiển thị theo quy ước: **ROW LABEL** cho biết bao nhiêu thông tin được cung cấp cho LABEL ở **TOP**.
* **Đường chéo (diagonal)** thường để trống (trivial) — không có thông tin hữu ích ở đó.
* **Thanh màu (colorbar)** bên phải: cho biết quy ước màu (xanh = +1, đỏ = -1).
* Như đã đề cập trước đó Sweetviz sẽ Phân tích từng biến và hiển thị. Để xem được biểu đồ đơn biến chỉ cần ấn vào 1 trong danh sách các biến bên trái:

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.5 Phần giao diện chính của biểu đồ dựa trên biến*

* Kết quả với biến giá trị:

+ Các thông số trên thẻ biên:

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.6 Phần giao diện chính của biểu đồ dựa trên biến*

+ **VALUES: 2,240 (100%):** Số lượng bản ghi không rỗng.

+ **MISSING:** Số hoặc kí hiệu nếu có giá trị thiếu.

+ **DISTINCT: 776 (35%)** Số giá trị khác nhau (unique) và tỉ lệ so với tổng. Ví dụ MntWines có 776 giá trị khác nhau, chiếm 35% tổng bản ghi.

+ **ZEROS: 13 (<1%)**: Số bản ghi có giá trị = 0 (và tỉ lệ%). Quan trọng để biết có nhiều khách hàng không mua mặt hàng đó hay không.

**+ MIN / 5% / Q1 / MEDIAN / AVG / Q3 / 95% / MAX**

+ Các chỉ số phân vị và thống kê cơ bản:

* **MIN**: giá trị nhỏ nhất.
* **5%**: phân vị 5% (giá trị dưới 5% thấp nhất).
* **Q1**: 25% quantile. Đây là tứ phân vị thứ 1 (25%) nghĩa là **25% khách hàng chi tiêu ≤ 24** cho rượu.
* **MEDIAN**: 50% quantile (giá trị ở giữa).
* **AVG**: trung bình (mean).
* **Q3**: Đây là tứ phân vị thứ 3 (75%). Có nghĩa là **75% khách hàng chi tiêu ≤ 504**, và chỉ 25% khách hàng chi tiêu nhiều hơn 504.
* **95%**: phân vị 95% (giúp thấy mức giá trị rất cao).
* **MAX**: giá trị lớn nhất.

+ **RANGE** = MAX – MIN(khoảng giá trị).

+ **IQR** = Q3 − Q1 (inter-quartile range — độ rộng khoảng giữa 25% và 75%).

+ **STD** (standard deviation) Độ lệch chuẩn — đo độ phân tán.

+ **VAR** (variance) Phương sai (STD^2).

+ **KURT. (kurtosis)** Độ nhọn/đuôi của phân phối.

* Kurtosis > 0: đuôi dày (outliers nhiều hơn so với chuẩn).
* Kurtosis < 0: phân phối bằng phẳng hơn.

+ **SKEW (skewness)** Độ lệch trái/phải:

* **> 0**: *right-skewed* (đa số nhỏ, vài giá trị lớn — đuôi sang phải).
* **< 0**: *left-skewed*.

+ **SUM** Tổng giá trị (ví dụ tổng tiền chi cho MntWines).

Biểu đồ Histogram:

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.7 Phần giao diện biểu đồ Histogram và các loại thống kê khác*

+ **Histogram** thể hiện phân phối giá trị.

+ Trục hoành = giá trị biến (tiền chi tiêu), trục tung = tần suất (số lượng hoặc phần trăm).

+ Nếu histogram có **cột cao bên trái** và đuôi kéo dài sang phải → phân phối lệch phải (right-skew).

+ Ở đây: phần lớn khách hàng chi tiêu thấp (cột cao ở vùng gần 0–200 và giảm dần), có một “đuôi dài” tới các giá trị lớn (một vài khách bỏ ra rất nhiều tiền).

Các biểu đồ “most frequent / smallest / largest values”:

**+ Most frequent values**: danh sách các giá trị xuất hiện nhiều nhất (và tỉ lệ). Thường áp dụng cho biến rời rạc.

* **Smallest values**: các giá trị nhỏ xuất hiện (và tần suất).
* **Largest values**: các giá trị lớn nhất (và tần suất — thường rất ít).

+ Numerical associations:

* Hiển thị các **Pearson correlations** (từ −1 đến +1) giữa biến đang xem (ví dụ MntWines) và các biến số khác.
* Giá trị gần +1: tương quan dương mạnh (ký hiệu màu xanh).
* Giá trị gần −1: tương quan âm mạnh (màu đỏ).
* Trị số giúp biết biến nào “liên quan” với MntWines — ví dụ MntWines có thể tương quan tốt với NumStorePurchases, NumCatalogPurchases, Income…

+ Categorical associations:

* Cho biết sự liên hệ giữa biến đang xem và **các biến phân loại** (dùng correlation ratio / uncertainty coefficient; thang 0..1).
* Ví dụ Kidhome 0.51 nghĩa là có mức liên hệ đáng kể giữa có con nhỏ và giá trị MntWines (nếu đó là cách SweetViz hiển nghĩa).
* Dùng để xem các biến category giải thích được bao nhiêu biến khác nhau.
* Với biến phân loại

+ Các thông số trên thẻ biến được rút gọn chỉ còn: Values, Missing, Distinct

+ Biểu đồ:

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.8 Phần giao diện biểu đồ Histogram và các loại thống kê khác của biến phân loại*

+ Top categories: Hiển thị tần suất xuất hiện trong data của biến phân loại

* Phần lớn khách hàng là Married hoặc Together (tổng cộng ~65%), Single chiếm 21%. Những giá trị hiếm như *YOLO, Absurd, Alone* có thể được coi là **outliers / lỗi nhập liệu**.

Categorical associations:

+ Marital\_Status PROVIDES INFORMATION:

Xem *biến Marital\_Status* cung cấp thông tin cho biến nào khác.

Ở đây, các giá trị cao nhất:

* Z\_CostContact = 1.00 và Z\_Revenue = 1.00 (nhưng hai biến này trong dataset mặc định là hằng số = không có ý nghĩa).
* Response = 0.03 → tình trạng hôn nhân có chút ảnh hưởng đến việc khách hàng phản hồi chiến dịch, nhưng rất yếu.
* Các biến khác <0.02 → gần như không có tác động.

+ THESE FEATURES GIVE INFORMATION ON Marital\_Status:

Ngược lại: biến nào giải thích cho Marital\_Status.

* Tất cả hệ số rất thấp (max chỉ 0.01). → Nghĩa là tình trạng hôn nhân gần như độc lập so với các biến khác trong dataset.

+ Numerical associations:

* Hiển thị mức độ liên hệ (0..1) giữa biến phân loại và các biến số.
* Year\_Birth = 0.21 → tuổi (năm sinh) có liên quan vừa phải đến tình trạng hôn nhân (hợp lý: người trẻ thường Single / Together, lớn tuổi thường Married / Widow).
* Tạo báo cáo với biến mục tiêu: Sweetviz có thể tạo báo cáo trên biến mục tiêu với các cột chỉ định

```

*data = df[["Response", "Income", "MntWines"]].copy()*

*# Tạo báo cáo SweetViz, với 'Response' là biến mục tiêu*

*report = sv.analyze([data, "Marketing Data"], target\_feat="Response")*

*#show báo cáo*

*report.show\_html("sweetviz\_response\_report.html")*

```

* Kết quả:

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.9 Phần giao diện biểu đồ Histogram và các loại thống kê khác khi tạo báo cáo kèm biến mục tiêu*

* Ngoài ra Sweetviz cũng có thể trực quan hợp biến:

```

*# Chọn 2 biến cần phân tích*

*data1 = df[["Income", "MntWines"]].copy()*

*median\_wine = data1["MntWines"].median()*

*report = sv.compare\_intra(*

*data1,*

*data1["MntWines"] > median\_wine,*

*["MntWines > median", "MntWines <= median"]*

*)*

*report.show\_html("sweetviz\_income\_wine.html")*

```

* Kết quả

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 1.3.3.10 Phần giao diện biểu đồ Histogram và các loại thống kê khác khi tạo báo cáo với hợp biến*