Tìm hiểu các tính năng và cách sử dụng sản phẩm SweetViz

1. **Giới thiệu**

* **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. **Cách cài đặt và sử dụng SweetViz**
   1. Cài đặt

```

pip install sweetviz

```

Trong Terminal của Vscode hoặc PowerShell

A screenshot of a black screen

AI-generated content may be incorrect.

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.

```

Python -m pip install numpy==1.23.4

```

* 1. 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.
* Import thư viện pandas để đọc data:

```

import pandas as pd

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

```

* Xem thông tin cơ bản về dữ liệu:

A computer screen with white and blue text

AI-generated content may be incorrect.

* Thông tin các cột của dữ liệu:

**A. Thông tin nhân khẩu học (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

**B. 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

**C. 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

**D. 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

* 1. Tạo báo cáo
* Phân tích một tập dữ liệu:

```

report = sv.analyze(df)

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

```

Sweetviz sẽ:

+ Quét và phân loại dữ liệu (Data Profiling)

# 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 như thế nào,

+ 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.

Biểu đồ Associations được hiển thị khi bạn bấm vào:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

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.

* 1. Kết quả với biến giá trị:
* Các thông số trên thẻ biến:

A screenshot of a computer

AI-generated content may be incorrect.

+ **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.

+ **Histogram (thanh dọc)** thể hiện phân phối giá trị theo khoảng (bins).

* 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.
  1. 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.

+ 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).

* 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.

**Biến mục tiêu: Response**

* **Tổng số dòng:** 2,240
* **Phân bố:**
  + 0 (không phản hồi): **1,906 khách (≈ 85%)**
  + 1 (có phản hồi): **334 khách (≈ 15%)**
    - Dữ liệu **bất cân bằng** (imbalanced), nhóm không phản hồi chiếm đa số.

**Biến Income (Thu nhập)**

* **Số liệu chính:**
  + Trung bình (AVG): ~52k
  + Trung vị (Median): ~51k
  + Phân bố khá lệch phải (có outlier tới **667k**).
  + ~1% dữ liệu bị thiếu (Missing = 24).
    - Những khách hàng thu nhập cao hơn có thể ảnh hưởng đến khả năng phản hồi, nhưng hệ số tương quan chỉ ở mức thấp.

**Biến MntWines (Chi tiêu cho rượu vang)**

* **Số liệu chính:**
  + Trung bình: ~304
  + Trung vị: ~174
  + Có khách chi đến gần **1,500** cho rượu vang.
  + Phân bố lệch phải (đa số chi thấp, một số ít chi rất cao).
    - MntWines có vẻ là biến **phân biệt mạnh hơn** giữa nhóm phản hồi và không phản hồi.
* 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.