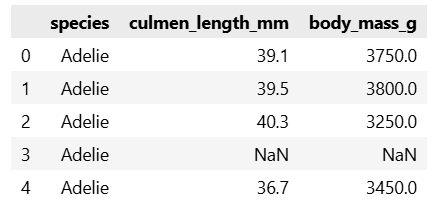
**1.3.2. Bài làm mẫu**

**Bài toán 1:** 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 đơn biến trên dữ liệu về chim cánh cụt

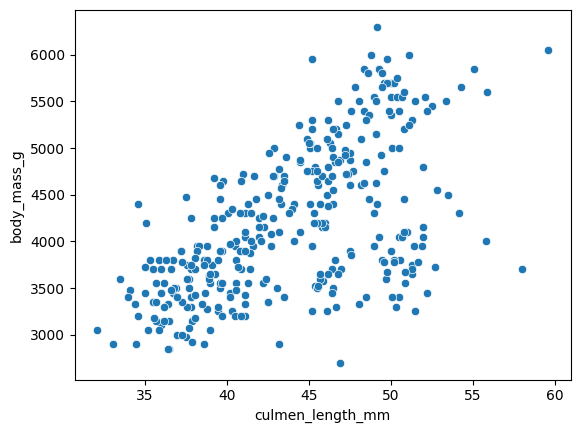
1. Import thư viện và chuẩn bị dữ liệu phân tích

Bảng 1: Dữ liệu Độ dài mỏ và khối lượng cơ thể



Nhận xét: Khi chiều dài mỏ tăng, khối lượng cơ thể chim cánh cụt thường cũng có xu hướng tăng.

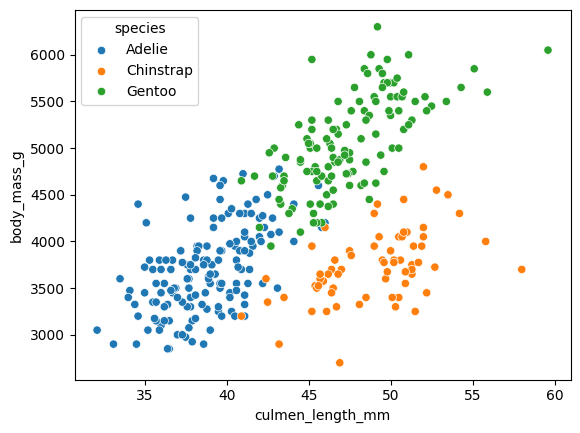
2. Phân tích dữ liệu 2 biến dựa vào phương pháp scatterplot



Hình 1: Biểu đồ phân tán giữa độ dài mỏ và khối lượng cơ thể

Dựa vào hình ta thấy:

Cho thấy mối tương quan dương giữa độ dài mỏ và khối lượng cơ thể.  
Chim cánh cụt có mỏ dài thường nặng hơn.



Hình 2: Biểu đồ phân tán giữa độ dài mỏ và khối lượng cơ thể theo loài

Dựa vào hình ta thấy:

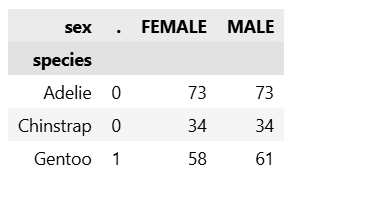
Loài Gentoo có xu hướng mỏ dài hơn và khối lượng cơ thể lớn hơn, cho thấy mối quan hệ tuyến tính khá mạnh.

Loài Adelie có mỏ ngắn và khối lượng nhẹ hơn, nhưng vẫn có sự phân bố tương đối đều theo xu hướng tăng.

Loài Chinstrap nằm giữa hai loài còn lại, với sự phân bố khá chặt chẽ, cho thấy độ tương quan cao giữa hai đặc tính.

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

Bảng 2: Phân bố giới tính theo loài



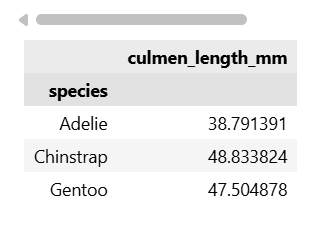
Nhận xét:

Adelie & Chinstrap: dữ liệu được thu thập rất đều, không thiên lệch giới tính.

Gentoo: Có sự chênh lệch nhẹ

4. Phân tích 2 biến sử dụng pivot\_table

Bảng 3: Độ dài mỏ trung bình theo loài

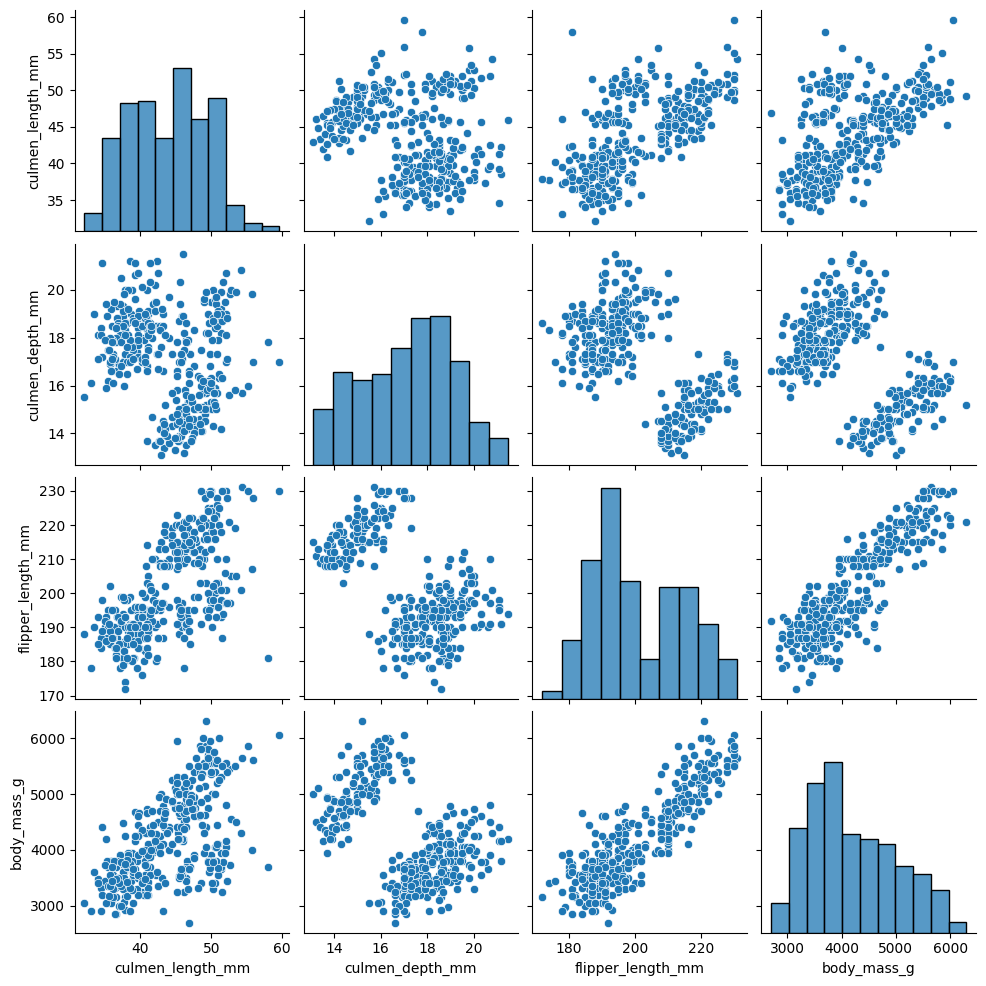


Nhận xét:

Gentoo có độ dài mỏ trung bình lớn nhất, tiếp theo là Chinstrap, và Adelie có mỏ ngắn nhất.

Điều này phù hợp với các biểu đồ bạn đã cung cấp trước đó, cho thấy Gentoo thường có kích thước cơ thể lớn hơn và mỏ dài hơn.

5. Phân tích 2 biến sử dụng pairplot



Hình 3: Biểu đồ pair-plot

Dựa vào hình ta thấy:

Chim có mỏ dài thường có khối lượng cơ thể lớn hơn.

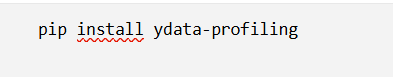
Chim có vây dài thì thường nặng hơn.

Chim có mỏ dài thì vây cũng thường dài.

**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.

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>)



Hình 4: Hướng dẫn tải thư viện ydata\_profiling

2. Sử dụng công cụ

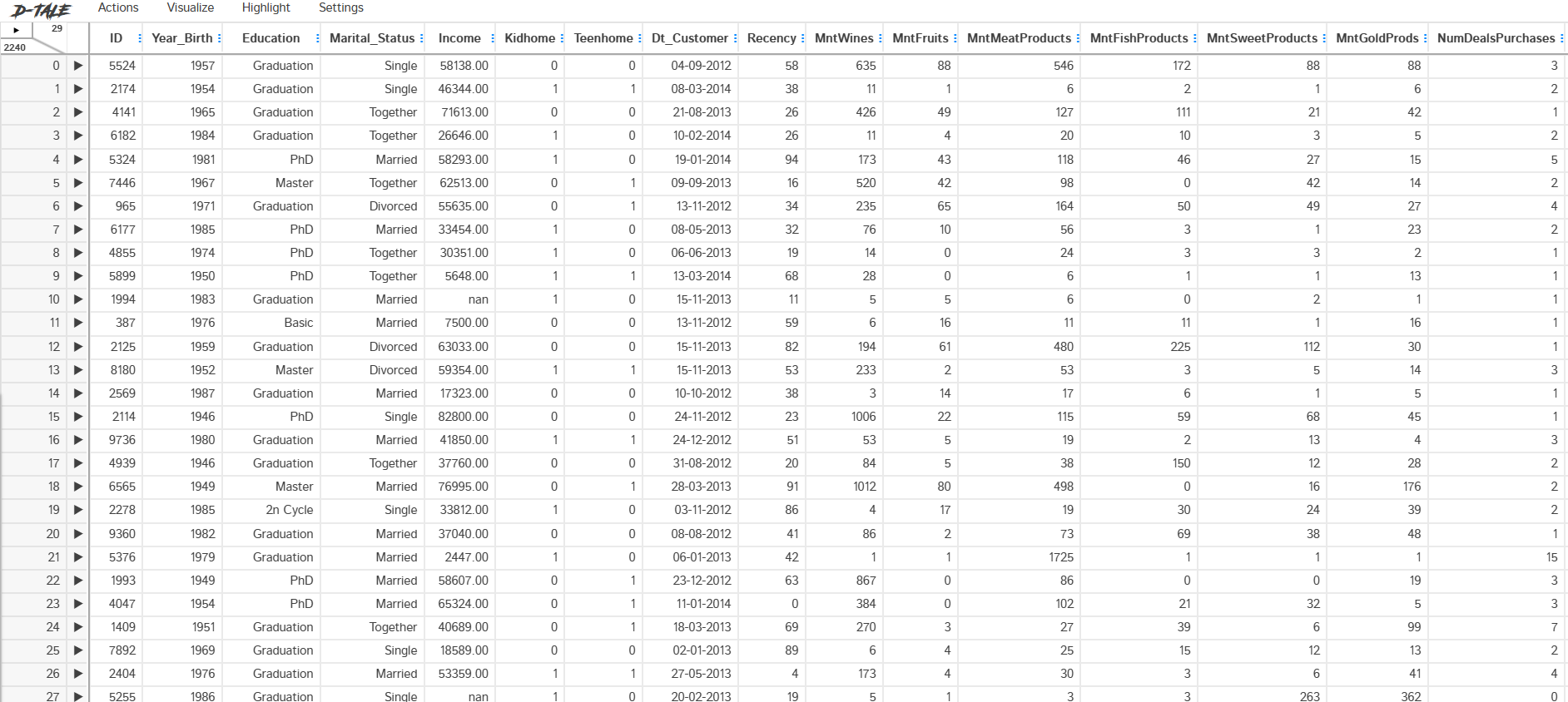


3. Tiến hành EDA trên trang tập tin profile\_output.html

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

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

2. Sử dụng công cụ:

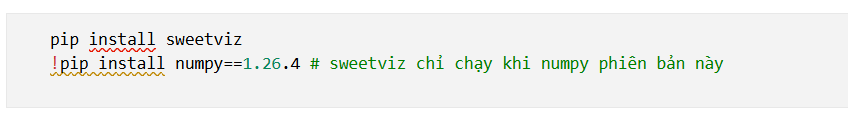


Hình 5:

**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. Cài đặt sweetviz



Hình 6: Hướng dẫn tải thư viện sweetviz

2. EDA trên tập dữ liệu Marketing Campaign:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 7: Phân bố năm sinh(Year\_Birth)

**Dựa vào hình ta thấy:**

Dữ liệu năm sinh phân bố chủ yếu trong khoảng 1950 – 1980, nhiều nhất là năm 1970.

Độ lệch chuẩn nhỏ cho thấy nhóm khách hàng khá tập trung ở trung niên.

A screenshot of a computer

AI-generated content may be incorrect.

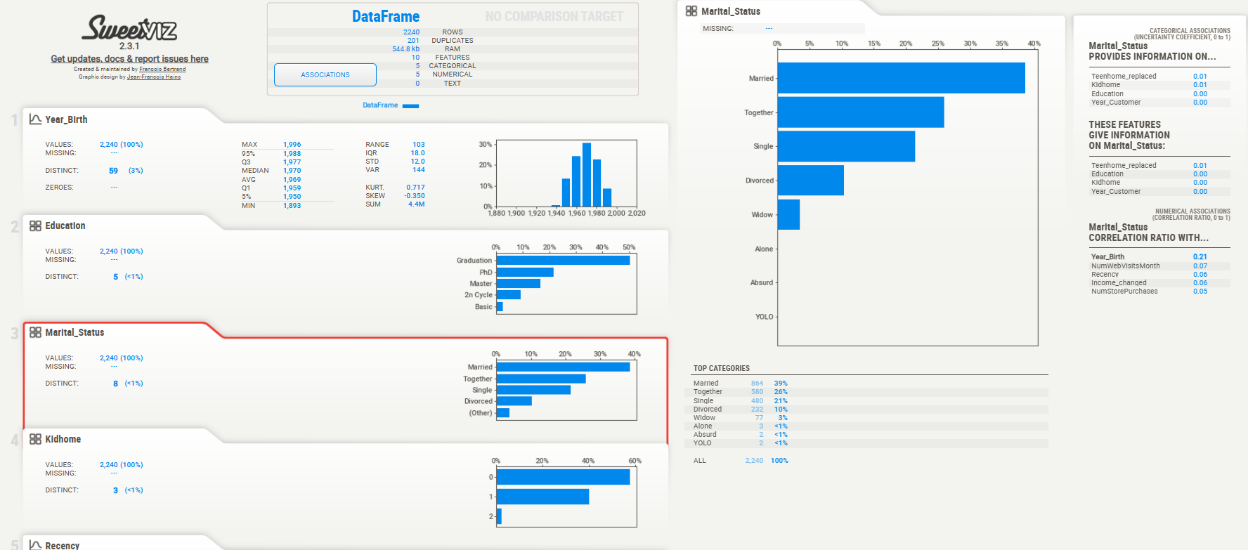
Hình 8: Phân bố trình độ học vấn(Education)

**Dựa vào hình ta thấy:**

Nhóm khách hàng có trình độ Graduation (Đại học) chiếm đa số (~50%).

Tiếp theo là PhD (~22%), Master (~17%), các trình độ thấp hơn (2nd Cycle, Basic) chiếm tỉ lệ nhỏ.

Trình độ học vấn có tương quan vừa phải với thu nhập (corr ≈ 0.21), cho thấy học vấn cao hơn thường có thu nhập cao hơn.

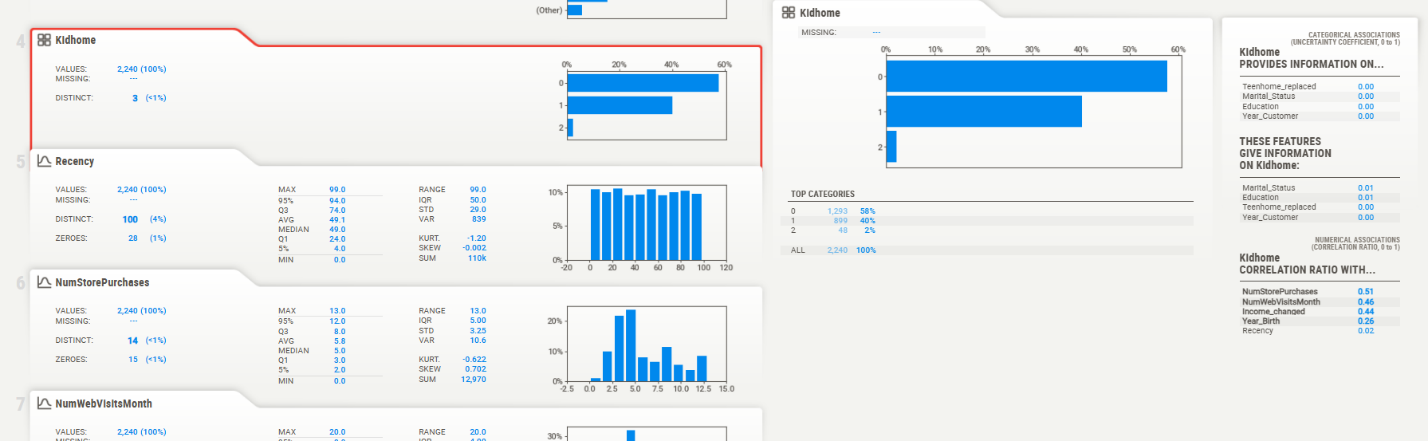


Hình 9: Phân bố tình trạng hôn nhân(MaritalStatus)

**Dựa vào hình ta thấy:**

Phân bố chính: Married (39%), Together (26%), Single (21%), Divorced (10%).

Tương quan với các biến khác khá yếu (cao nhất với Year\_Birth là 0.21), cho thấy tình trạng hôn nhân không ảnh hưởng mạnh tới các chỉ số mua sắm.

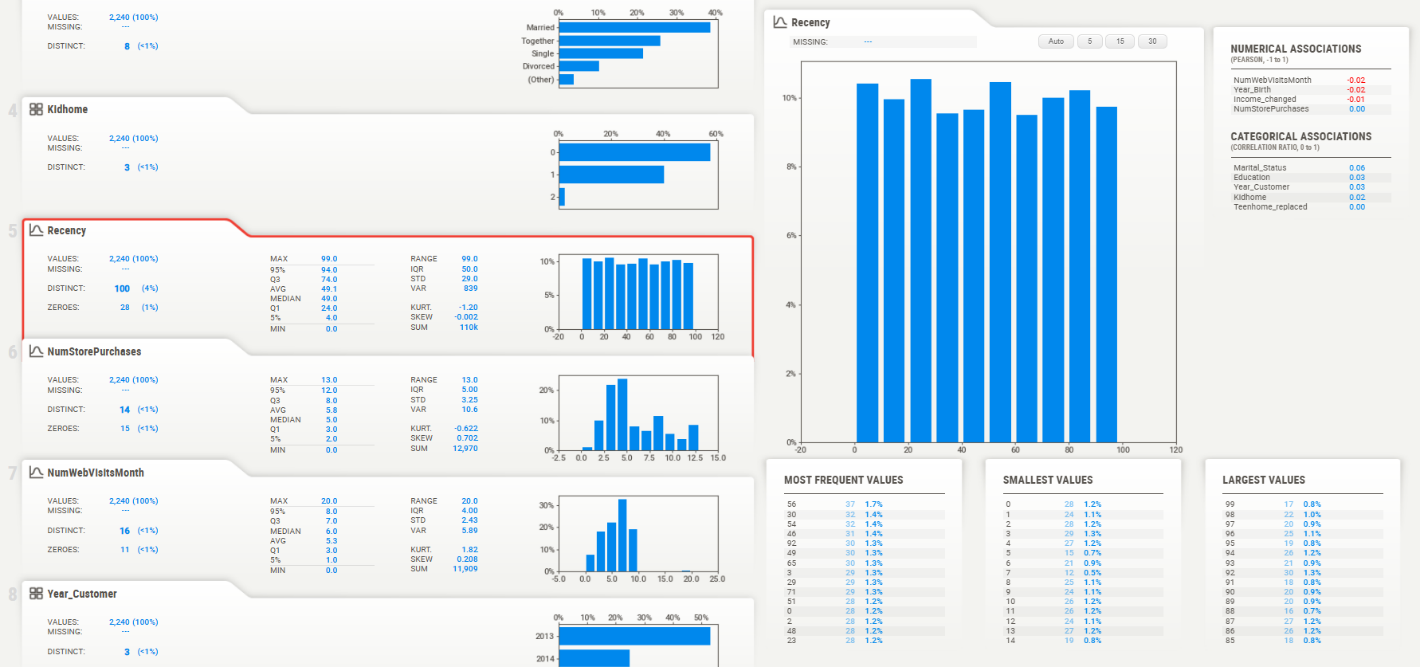


Hình 10: Phân bố số trẻ nhỏ trong gia đình (Kidhome)

**Dựa vào hình ta thấy:**

Phân bố 58% khách hàng không có con nhỏ, 40% có 1 con, chỉ 2% có 2 con.

Biến này có tương quan vừa với số lần mua tại cửa hàng (0.51) và mua qua web (0.46), cho thấy gia đình có trẻ nhỏ có thể ảnh hưởng hành vi mua sắm.



Hình 11: Phân bố số ngày kể từ lần mua cuối cùng (Recency)

**Dựa vào hình ta thấy:**

Phân bố khá đồng đều từ 0–100 ngày, không bị lệch hẳn về phía nào.

Tương quan với các biến khác rất thấp (<0.05), cho thấy tần suất mua không phụ thuộc nhiều vào thu nhập hay tình trạng hôn nhân.



Hình 12: Phân bố số lần mua tại cửa hàng(NumStorePurchases)

**Dựa vào hình ta thấy:**

Phần lớn khách hàng mua từ 3–8 lần, một số rất ít mua tới 13 lần.

Liên quan đáng kể với biến Kidhome (0.51), có thể vì khách hàng có gia đình thường thích mua trực tiếp tại cửa hàng.

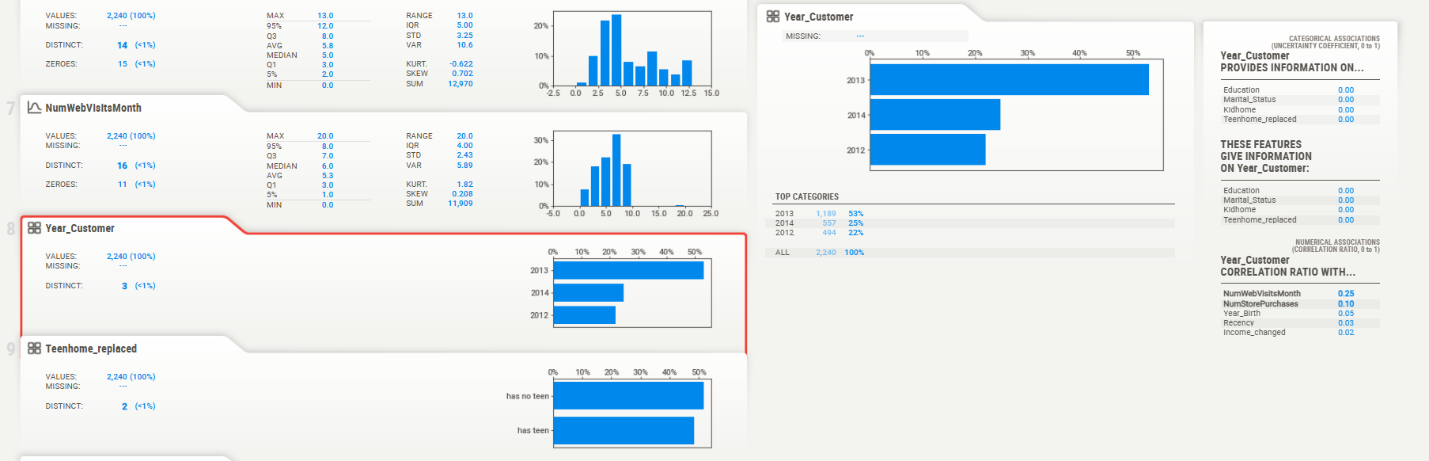


Hình 13: Phân bố số lần truy cập web mỗi tháng(NumWebVisitsMonth)

**Dựa vào hình ta thấy:**

Đa số khách hàng truy cập web từ 3–8 lần/tháng.

Tương quan nghịch nhẹ với thu nhập (-0.54), gợi ý rằng khách hàng có thu nhập cao có thể ít truy cập website hơn.

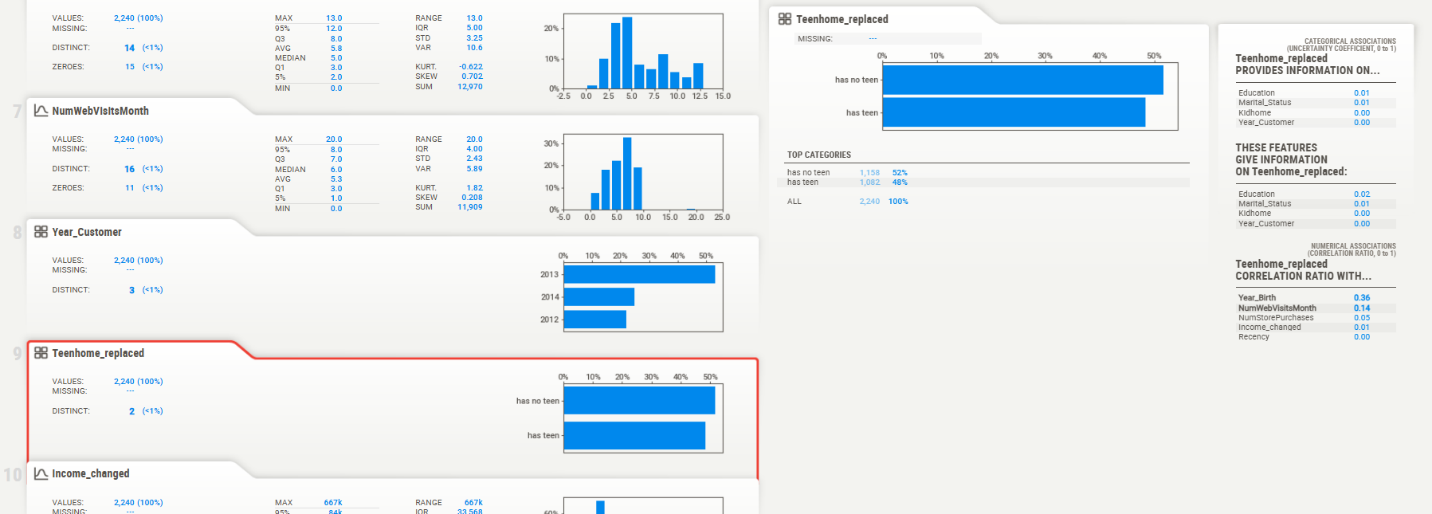


Hình 14: Phân bố ngày đăng ký khách hàng(Year\_Customer)

**Dựa vào hình ta thấy:**

Phần lớn khách hàng gia nhập năm 2013 (~50%), tiếp theo là 2014 và 2012.

Biến này có tương quan rất thấp với các biến khác, đóng vai trò như một đặc trưng về thời gian, có thể hữu ích trong phân tích xu hướng.

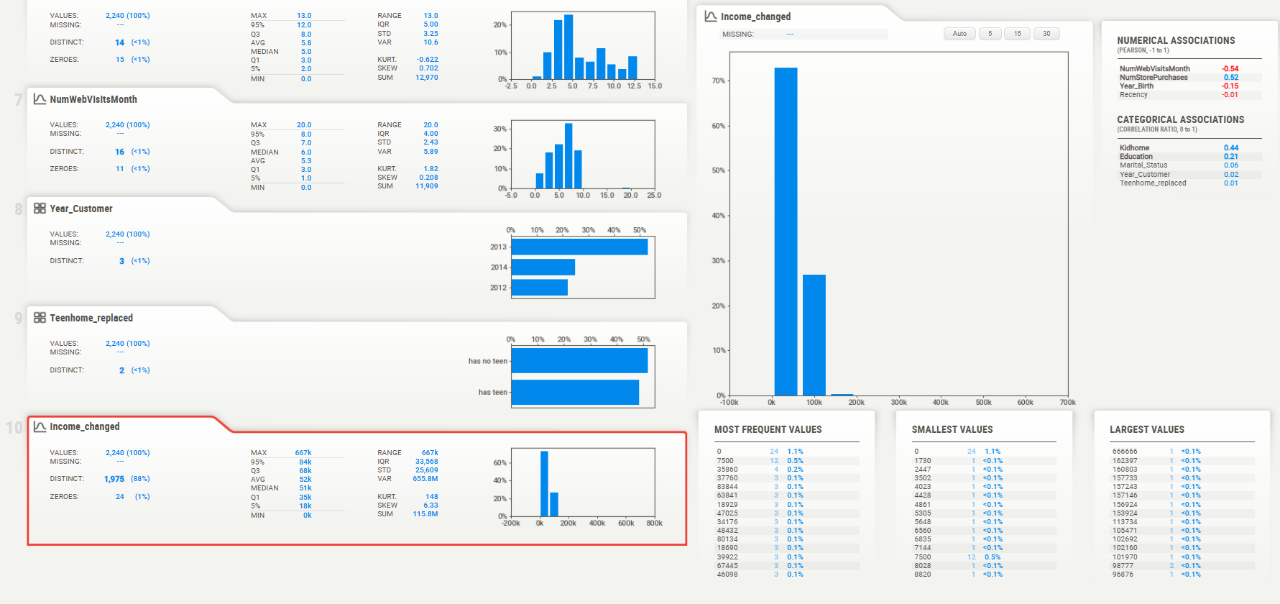


Hình 15: Phân bố số trẻ vị thành niên(Teenhome\_replaced)

**Dựa vào hình ta thấy:**

Có 52% hộ gia đình không có trẻ vị thành niên, 46% có 1 trẻ, và chỉ 2% có 2 trẻ.

Có mối liên hệ khá mạnh với Year\_Birth (0.36), hợp lý vì phụ huynh trẻ tuổi hơn có xu hướng có con vị thành niên.



Hình 16: Phân bố thu nhập khách hàng (Income)

**Dựa vào hình ta thấy:**

Phân bố thu nhập lệch phải, đa số tập trung dưới 100,000.

Thu nhập có quan hệ mạnh với số con nhỏ (Kidhome) và số lần mua hàng qua cửa hàng.

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 17: Ma trận tương quan

**Dựa vào hình ta thấy:**

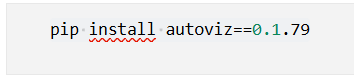
Phân bố năm sinh tập trung chủ yếu trong khoảng 1950–1980, với năm 1970 là giá trị xuất hiện nhiều nhất, cho thấy nhóm khách hàng thuộc thế hệ trung niên chiếm ưu thế.

Độ lệch chuẩn thấp (~12) phản ánh độ tuổi khách hàng khá đồng đều, không quá phân tán, thuận lợi cho việc xây dựng chiến lược tiếp thị tập trung.

**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. Cài đặt autoviz



Hình 18: Hướng dẫn tải thư viện autoviz

2. Sử dụng công cụ: