**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**--------\*\*\*--------**

A blue button with white text

AI-generated content may be incorrect.

BÁO CÁO THỰC HÀNH

MÔN PHÂN TÍCH DỮ LIỆU

**TÊN ĐỀ TÀI PHÂN TÍCH KHÁM PHÁ DỮ LIỆU VỚI PYTHON**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | **ThS. Đỗ Như Tài** |
| **Nhóm sinh viên thực hiện:** |  |
| Hồ Thị Thanh Thảo | 3122410389 |
| Nguyễn Thị Hồng Thắm | 3122410392 |
| Phan Văn Thảo | 3122410391 |
| Nguyễn Hoàng Thiên Bảo | 3122410019 |

***Thành phố Hồ Chí Minh, tháng 09 năm 2025***

**LỜI MỞ ĐẦU**

Trong thời đại công nghệ 4.0, dữ liệu đã trở thành tài nguyên quý giá và đóng vai trò then chốt trong việc ra quyết định của các tổ chức và doanh nghiệp. Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA) là bước đầu tiên và quan trọng nhất trong quy trình khai thác dữ liệu, giúp chúng ta hiểu rõ bản chất, cấu trúc và các đặc điểm tiềm ẩn của tập dữ liệu trước khi tiến hành xây dựng mô hình học máy hay đưa ra các quyết định chiến lược.

Nhận thức được tầm quan trọng của EDA trong lĩnh vực khoa học dữ liệu, bài thực hành này tập trung vào việc trang bị cho sinh viên các kỹ năng cơ bản và cần thiết để thực hiện phân tích khám phá dữ liệu một cách có hệ thống và hiệu quả. Thông qua việc thực hành trên các tập dữ liệu thực tế, sinh viên sẽ được làm quen với các kỹ thuật thống kê mô tả, các phương pháp trực quan hóa dữ liệu, cũng như cách sử dụng các công cụ tự động hỗ trợ EDA.

Báo cáo này trình bày kết quả thực hành của chúng em về phân tích khám phá dữ liệu, bao gồm ba phần chính: Thống kê mô tả, Xử lý và trực quan hóa dữ liệu, và Phân tích đơn biến cùng hai biến. Mỗi phần không chỉ bao gồm kiến thức lý thuyết nền tảng mà còn có các bài tập thực hành cụ thể với code Python và phân tích kết quả chi tiết.

Qua quá trình thực hiện bài lab này, chúng em đã có cơ hội được áp dụng kiến thức đã học vào thực tế, từ đó hiểu sâu hơn về tầm quan trọng của việc khám phá và làm sạch dữ liệu trước khi xây dựng mô hình. Đồng thời, em cũng nắm vững cách sử dụng các thư viện Python phổ biến như Pandas, Matplotlib, Seaborn và các công cụ EDA tự động như SweetViz, AutoViz.

Chúng em mong rằng báo cáo này sẽ thể hiện được sự nỗ lực học tập và nghiên cứu của chúng em trong quá trình thực hành, đồng thời góp phần củng cố kiến thức nền tảng cho các môn học tiếp theo.

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến *Ths. Đỗ Như Tài*, người đã tận tình hướng dẫn và chia sẻ những kiến thức quý báu trong suốt quá trình thực hiện đề tài này. Sự chỉ bảo tận tình và những góp ý xây dựng từ thầy đã giúp chúng em hiểu rõ hơn về các vấn đề chuyên môn và hoàn thiện sản phẩm một cách tốt nhất.

Mặc dù đã rất cố gắng, nhưng do hạn chế về thời gian và kinh nghiệm, chắc chắn rằng không thể tránh khỏi những sai sót. Chúng em rất mong nhận được ý kiến đóng góp từ thầy để có thể cải thiện và hoàn thiện hơn nữa.

Một lần nữa, chúng em xin chân thành cảm ơn!

Contents

[PHẦN 1: GIỚI THIỆU VÀ MỤC TIÊU 5](#_Toc210769398)

[1.1. Giới thiệu 5](#_Toc210769399)

[1.1.1. Bối cảnh nghiên cứu 5](#_Toc210769400)

[1.1.2. Tầm quan trọng của EDA 5](#_Toc210769401)

[1.1.3. Công cụ và kỹ thuật 5](#_Toc210769402)

[1.2. Mục tiêu bài thực hành 6](#_Toc210769403)

[1.2.1. Mục tiêu chung 6](#_Toc210769404)

[1.2.2. Mục tiêu cụ thể 6](#_Toc210769405)

[1.2.3. Tập dữ liệu sử dụng 7](#_Toc210769406)

[1.2.4. Nội dung thực hiện 7](#_Toc210769407)

[PHẦN 2: THỐNG KÊ MÔ TẢ 8](#_Toc210769408)

[2.1. Lý thuyết 8](#_Toc210769409)

[2.2. Bài làm mẫu 11](#_Toc210769410)

[2.2.1. Khám phá dữ liệu COVID 11](#_Toc210769411)

[2.2.2. Khám phá và xử lý dữ liệu Marketing Campaign 12](#_Toc210769412)

[2.3. Bài tập thực hành 13](#_Toc210769413)

[2.3.1. Bài tập 1: Phân loại chất lượng rượu đỏ 13](#_Toc210769414)

[2.3.2. Bài tập 2: Dự đoán bệnh tiểu đường 13](#_Toc210769415)

[PHẦN 3: XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU 14](#_Toc210769416)

[3.1. Lý thuyết 14](#_Toc210769417)

[3.1.1. Vai trò của trực quan hóa dữ liệu trong phân tích và khám phá dữ liệu (EDA) 14](#_Toc210769418)

[3.1.2. Các loại biểu đồ phổ biến và ứng dụng 14](#_Toc210769419)

[3.1.3. Cách lựa chọn biểu đồ phù hợp với loại dữ liệu 15](#_Toc210769420)

[3.1.4. So sánh giữa các thư viện trực quan hóa trong Python 15](#_Toc210769421)

[3.1.5. Nguyên tắc thiết kế biểu đồ hiệu quả 16](#_Toc210769422)

[3.1.6. Ví dụ tạo biểu đồ trong Python 16](#_Toc210769423)

[3.2. Bài làm mẫu: Trực quan hóa dữ liệu giá nhà Amsterdam 17](#_Toc210769424)

[3.3. Bài tập thực hành 1 18](#_Toc210769425)

[3.4. Bài tập thực hành 2 18](#_Toc210769426)

[3.4.1. Bài tập 1 18](#_Toc210769427)

[3.4.2. Bài tập 2 18](#_Toc210769428)

[PHẦN 4: PHÂN TÍCH ĐƠN BIẾN VÀ HAI BIẾN 20](#_Toc210769429)

[4.1. Lý thuyết 20](#_Toc210769430)

[4.2. Bài làm mẫu 27](#_Toc210769431)

[4.2.1. Bài toán 1 27](#_Toc210769432)

[4.2.1. Bài toán 2 35](#_Toc210769433)

[4.2.1. Bài toán 3 39](#_Toc210769434)

[4.3. Bài tập thực hành 1 41](#_Toc210769435)

[4.4. Bài tập thực hành 2 50](#_Toc210769436)

[PHẦN 5: KẾT LUẬN 64](#_Toc210769437)

[5.1. Tổng kết kiến thức đã học 64](#_Toc210769438)

[5.1.1. Về thống kê mô tả 64](#_Toc210769439)

[5.1.2. Về trực quan hóa dữ liệu 64](#_Toc210769440)

[5.1.3. Về phân tích đơn biến và hai biến 65](#_Toc210769441)

[5.2. Kỹ năng đạt được 66](#_Toc210769442)

[5.2.1. Kỹ năng lập trình Python 66](#_Toc210769443)

[5.2.2. Kỹ năng xử lý dữ liệu 66](#_Toc210769444)

[5.2.3. Kỹ năng phân tích và tư duy 67](#_Toc210769445)

[5.3. Khó khăn gặp phải và cách khắc phục 68](#_Toc210769446)

[5.3.1. Khó khăn về kỹ thuật 68](#_Toc210769447)

[5.3.2. Khó khăn về phân tích 68](#_Toc210769448)

[5.3.3. Khó khăn về thời gian 69](#_Toc210769449)

[5.4. Bài học rút ra 69](#_Toc210769450)

[5.4.1. Về quy trình EDA 69](#_Toc210769451)

[5.4.2. Về kỹ năng phân tích 70](#_Toc210769452)

[5.4.3. Về công cụ và kỹ thuật 70](#_Toc210769453)

[5.5. Hướng phát triển 71](#_Toc210769454)

[LỜI KẾT 73](#_Toc210769455)

[TÀI LIỆU THAM KHẢO 74](#_Toc210769456)

# PHẦN 1: GIỚI THIỆU VÀ MỤC TIÊU

## 1.1. Giới thiệu

### 1.1.1. Bối cảnh nghiên cứu

Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA) là quá trình phân tích tập dữ liệu để tóm tắt các đặc điểm chính của chúng, thường sử dụng các phương pháp trực quan hóa. EDA được John Tukey phát triển vào những năm 1970 và hiện nay đã trở thành bước không thể thiếu trong quy trình khoa học dữ liệu.

Trong thực tế, trước khi xây dựng bất kỳ mô hình máy học nào, việc hiểu rõ dữ liệu là vô cùng quan trọng. EDA giúp chúng ta:

* Phát hiện các vấn đề về chất lượng dữ liệu (missing values, outliers, duplicates)
* Hiểu phân bố và mối quan hệ giữa các biến
* Xác định các biến quan trọng cho việc modeling
* Đưa ra giả thuyết ban đầu về dữ liệu

### 1.1.2. Tầm quan trọng của EDA

Theo một nghiên cứu của Forbes, các data scientist dành khoảng 80% thời gian cho việc làm sạch và chuẩn bị dữ liệu, trong đó EDA chiếm phần lớn. Một EDA tốt có thể:

* Tiết kiệm thời gian và chi phí cho các bước tiếp theo
* Nâng cao độ chính xác của mô hình
* Giúp đưa ra quyết định đúng đắn về việc chọn thuật toán phù hợp

### 1.1.3. Công cụ và kỹ thuật

Trong bài thực hành này, chúng em sử dụng ngôn ngữ Python cùng với các thư viện phổ biến:

* **Pandas**: Xử lý và thao tác dữ liệu
* **NumPy**: Tính toán thống kê
* **Matplotlib**: Tạo biểu đồ cơ bản
* **Seaborn**: Trực quan hóa thống kê nâng cao
* **SweetViz**: Công cụ EDA tự động
* **AutoViz**: Công cụ visualization tự động

## 1.2. Mục tiêu bài thực hành

### 1.2.1. Mục tiêu chung

Bài thực hành này nhằm giúp sinh viên nắm vững các kỹ thuật cơ bản trong khám phá dữ liệu để hiểu rõ đặc điểm và cấu trúc của tập dữ liệu, từ đó đặt nền tảng cho các bước phân tích sâu hơn hoặc xây dựng mô hình khai thác dữ liệu.

### 1.2.2. Mục tiêu cụ thể

**Về kiến thức:**

* Hiểu và phân biệt được các khái niệm thống kê cơ bản (mean, median, mode, variance, standard deviation, quartiles, IQR)
* Nắm vững cách đọc và diễn giải các loại biểu đồ phổ biến (histogram, boxplot, scatter plot, heatmap)
* Phân biệt được phân tích đơn biến và phân tích hai biến
* Hiểu được tầm quan trọng của việc xử lý missing values và outliers

**Về kỹ năng:**

* Thực hiện thống kê mô tả trên tập dữ liệu thực tế
* Sử dụng thành thạo các thư viện Python: Pandas, Matplotlib, Seaborn
* Tạo các biểu đồ trực quan hóa dữ liệu hiệu quả
* Phát hiện và xử lý các vấn đề về chất lượng dữ liệu
* Sử dụng các công cụ EDA tự động (SweetViz, AutoViz)
* Phân tích mối quan hệ giữa các biến

**Về thái độ:**

* Có tư duy phân tích và khám phá dữ liệu có hệ thống
* Biết đặt câu hỏi đúng đắn về dữ liệu
* Cẩn thận, tỉ mỉ trong việc phân tích và đưa ra kết luận

### 1.2.3. Tập dữ liệu sử dụng

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên dataset** | **Nguồn** | **Mục đích** |
| 1 | Red Wine Quality | Kaggle | Thống kê mô tả & Trực quan hóa |
| 2 | Pima Indians Diabetes | Kaggle | Thống kê mô tả & Trực quan hóa |
| 3 | Marketing Campaign | Kaggle | Công cụ EDA tự động |

### 1.2.4. Nội dung thực hiện

Báo cáo được chia thành 3 phần chính:

**Phần 1: Thống kê mô tả**

* Ôn tập lý thuyết về các thước đo thống kê
* Thực hành tính toán các chỉ số thống kê
* Xử lý missing values và outliers

**Phần 2: Xử lý và trực quan hóa dữ liệu**

* Ôn tập lý thuyết về các loại biểu đồ
* Sử dụng Matplotlib và Seaborn
* Tạo các visualization có ý nghĩa

**Phần 3: Phân tích đơn biến và hai biến**

* Phân tích từng biến riêng lẻ
* Phân tích mối quan hệ giữa các biến
* Sử dụng công cụ EDA tự động

# PHẦN 2: THỐNG KÊ MÔ TẢ

## 2.1. Lý thuyết

**Câu hỏi 1: Thống kê mô tả là gì? Khác gì với thống kê suy luận?**

**Thống kê mô tả (Descriptive Statistics):**

* Là các phương pháp thu thập, tổ chức, trình bày và tóm tắt dữ liệu
* Mục đích: Mô tả đặc điểm cơ bản của tập dữ liệu hiện có
* Ví dụ: Tính trung bình, độ lệch chuẩn, vẽ biểu đồ

**Thống kê suy luận (Inferential Statistics):**

* Là phương pháp sử dụng dữ liệu mẫu để rút ra kết luận về tổng thể
* Mục đích: Dự đoán, kiểm định giả thuyết, ước lượng tham số
* Ví dụ: Kiểm định t-test, phân tích hồi quy, khoảng tin cậy

**Câu hỏi 2: Các thước đo thống kê mô tả chính (ví dụ: trung bình, trung vị, phương sai, độ lệch chuẩn) được sử dụng để làm gì? Trong trường hợp nào thì nên dùng trung vị thay vì trung bình?**

**Thước đo xu hướng trung tâm:**

* **Mean (Trung bình):** Tổng các giá trị chia cho số lượng quan sát
* **Median (Trung vị):** Giá trị ở giữa khi sắp xếp dữ liệu
* **Mode (Yếu vị):** Giá trị xuất hiện nhiều nhất

**Thước đo độ phân tán:**

* **Variance (Phương sai):** Đo lường mức độ phân tán quanh giá trị trung bình
* **Standard Deviation (Độ lệch chuẩn):** Căn bậc hai của phương sai

**Khi nào dùng Median thay vì Mean?**

* Khi dữ liệu có nhiều giá trị ngoại lai (outliers)
* Khi phân bố dữ liệu lệch (skewed distribution)
* Khi làm việc với dữ liệu thu nhập, giá nhà (thường có outliers)

**Câu hỏi 3: Làm thế nào để xác định phân bố của một tập dữ liệu? Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?**

**Các loại phân bố phổ biến:**

* **Phân bố chuẩn (Normal):** Đối xứng, hình chuông
* **Phân bố lệch phải (Right-skewed):** Đuôi dài bên phải, mean > median
* **Phân bố lệch trái (Left-skewed):** Đuôi dài bên trái, mean < median

**Cách xác định:**

* Sử dụng histogram để quan sát hình dạng
* So sánh mean, median, mode
* Tính hệ số skewness

**Câu hỏi 4: Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?**

**Độ lệch chuẩn (Standard Deviation):**

* Đo mức độ phân tán của dữ liệu xung quanh giá trị trung bình
* Giá trị nhỏ: Dữ liệu tập trung gần mean
* Giá trị lớn: Dữ liệu phân tán rộng

**Phạm vi (Range):**

* Range = Max - Min
* Cho biết khoảng biến thiên của dữ liệu
* Dễ bị ảnh hưởng bởi outliers

**Câu hỏi 5: Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì?**

**Các thước đo:**

* **Q1 (25th percentile):** 25% dữ liệu nhỏ hơn giá trị này
* **Q2 (Median, 50th percentile):** Trung vị
* **Q3 (75th percentile):** 75% dữ liệu nhỏ hơn giá trị này
* **IQR (Interquartile Range):** Q3 - Q1, đo mức độ phân tán của 50% dữ liệu giữa

**Xác định outliers:**

* Lower bound: Q1 - 1.5 × IQR
* Upper bound: Q3 + 1.5 × IQR

**Câu hỏi 6: Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả?**

**Phương pháp:**

* **Loại bỏ:** Xóa các dòng/cột có giá trị thiếu (nếu số lượng nhỏ)
* **Điền giá trị:** Mean, median, mode hoặc giá trị cụ thể
* **Dự đoán:** Sử dụng các thuật toán ML để dự đoán giá trị thiếu
* **Đánh dấu riêng:** Tạo category "Missing" cho biến phân loại

**Câu hỏi 7: Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không?**

**Histogram:**

* Trục X: Các khoảng giá trị (bins)
* Trục Y: Tần số xuất hiện
* Quan sát: Hình dạng phân bố, giá trị tập trung, độ phân tán

**Boxplot:**

* Hộp: Từ Q1 đến Q3
* Đường giữa hộp: Median
* Râu (whiskers): Kéo dài đến min/max trong khoảng hợp lệ
* Điểm ngoài râu: Outliers

**Câu hỏi 8: Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bạn sẽ xử lý chúng như thế nào trước khi thực hiện thống kê mô tả?**

**Các phương pháp:**

* Loại bỏ: Nếu là lỗi nhập liệu
* Giữ nguyên: Nếu là giá trị hợp lệ quan trọng
* Biến đổi: Log transformation, square root
* Winsorization: Thay thế outliers bằng giá trị gần nhất trong ngưỡng
* Binning: Gom nhóm dữ liệu

## 2.2. Bài làm mẫu

### 2.2.1. Khám phá dữ liệu COVID

Dataset: Dữ liệu COVID-19 từ Our World in Data

- Tổng số dòng: 429,435

- Số dòng có NaN: 19,276 (4.49%)

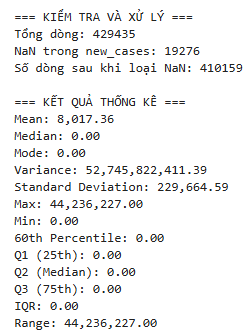
- Số dòng sau xử lý: 410,159

Các bước thực hiện:

- Load dữ liệu với Pandas

- Xem cấu trúc dữ liệu (shape, dtypes, head)

- Tính các chỉ số thống kê: mean, median, mode, variance, standard deviation, min, max, percentile, quartile, IQR



*Hình 2.1. Kết quả thống kê mô tả trên cột new\_case*

Từ hình ảnh trên ta thấy:

* Phân bố dữ liệu:

- Median = 0 và Q3 = 0 cho thấy ít nhất 75% dữ liệu có giá trị ≤ 0

- Điều này phản ánh thực tế: nhiều quốc gia trong nhiều ngày không có ca mới.

* Độ phân tán:

- Standard Deviation = 229,664.59 (rất lớn so với Mean)

- Range = 44,236,227 cho thấy có sự chênh lệch cực lớn

* Đặc điểm phân bố:

- Mean (8,017) lớn hơn rất nhiều so với Median (0), cho thấy dữ liệu có xu hướng lệch phải với một số giá trị cực đại kéo trung bình lên cao.

- Phù hợp với đặc điểm của dữ liệu dịch bệnh: đa số ngày ít ca, nhưng có đợt bùng phát với số ca rất lớn.

* Xử lý missing values:

- Đã loại bỏ 4.49% dữ liệu có NaN

### 2.2.2. Khám phá và xử lý dữ liệu Marketing Campaign

Dataset: Customer Personality Analysis từ Kaggle

Các kỹ thuật đã áp dụng:

- Loại bỏ dữ liệu trùng lặp với drop\_duplicates()

- Thay thế giá trị với replace()

- Điền missing values với fillna()

- Chuyển đổi kiểu dữ liệu với astype()

- Xử lý missing values với dropna()

**Kết quả:**

|  |  |  |
| --- | --- | --- |
| **Bước xử lý** | **Số dòng** | **Ghi chú** |
| Ban đầu | 2,240 | Dataset gốc |
| Sau loại duplicates | 2,240 | Không có dòng trùng |
| Missing values trong Income | 24 | 1.07% |
| Sau xử lý Income | 2,240 | Đã điền bằng median |
| Sau dropna() các cột khác | 2,216 | Mất 24 dòng (1.07%) |

## 2.3. Bài tập thực hành

### 2.3.1. Bài tập 1: Phân loại chất lượng rượu đỏ

Dataset: Dữ liệu về phân loại chất lượng rượu đỏ từ Kaggle

1. Import thư viện và nạp dữ liệu vào notebook

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Đọc dữ liệu

df = pd.read\_csv('winequality-red.csv', sep=',')

2. Thông tin cơ bản về dữ liệu

A screen shot of a computer

AI-generated content may be incorrect.

3. Thống kê mô tả các biến

A screenshot of a computer screen

AI-generated content may be incorrect.

4. Phân tích biến chất lượng

A screenshot of a computer

AI-generated content may be incorrect.

4. Tương quan với chất lượng

A screen shot of a computer

AI-generated content may be incorrect.

5. Phân tích theo nhóm chất lượng

A screenshot of a computer screen

AI-generated content may be incorrect.

7. Phát hiện ngoại lai

A screen shot of a computer

AI-generated content may be incorrect.

### 2.3.2. Bài tập 2: Dự đoán bệnh tiểu đường

Dataset: Dữ liệu về bệnh tiểu đường của người phụ nữ Pima

1. Import thư viện và nạp dữ liệu vào notebook

# Import thư viện

import pandas as pd

import numpy as np

# Đọc dữ liệu

df = pd.read\_csv('diabetes.csv')

2. Thông tin cơ bản về dataset

A screen shot of a computer code

AI-generated content may be incorrect.

3. Kiểm tra kiểu dữ liệu

4. Kiểu tra giá trị Null

5. Thống kê mô tả các biến số

6. Phân phối biến mục tiêu

7. Kiểm tra giá trị 0 (có thể missing value)

8. Phân tích theo nhóm Outcome

9. Tương quan giữa các biến

10. Phân phối tuổi

11. Phân tích Glucose – chỉ số quan trọng cho tiểu đường

12. Phân tích BMI chi tiết

# PHẦN 3: XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU

## 3.1. Lý thuyết

### 3.1.1. Vai trò của trực quan hóa dữ liệu trong phân tích và khám phá dữ liệu (EDA)

Trực quan hóa dữ liệu (Data Visualization) là bước quan trọng trong quy trình phân tích dữ liệu, đặc biệt trong giai đoạn khám phá dữ liệu (Exploratory Data Analysis – EDA). Việc biểu diễn dữ liệu bằng đồ thị giúp nhà phân tích nhận diện nhanh các xu hướng, mối quan hệ, và điểm bất thường (outliers) mà các phương pháp thống kê thuần túy khó thể hiện được.

Trực quan hóa đóng vai trò là cầu nối giữa dữ liệu và hiểu biết, giúp:

* Nắm bắt cấu trúc tổng thể và phân bố dữ liệu.
* Phát hiện mối tương quan giữa các biến (dương hoặc âm).
* Phát hiện sai lệch, dữ liệu thiếu hoặc ngoại lai.
* Hỗ trợ xây dựng giả thuyết, chọn mô hình phù hợp.
* Trình bày kết quả một cách trực quan, dễ hiểu cho người không chuyên về kỹ thuật.

### 3.1.2. Các loại biểu đồ phổ biến và ứng dụng

* Histogram (biểu đồ tần suất):

Dùng để mô tả phân bố của một biến số liên tục.

Giúp xác định phân bố có chuẩn hay không, độ lệch (skewness), và có nhiều đỉnh (multimodal) hay không.

* Boxplot (biểu đồ hộp)

Dùng để so sánh phân bố của dữ liệu giữa các nhóm.

Thể hiện giá trị trung vị, khoảng tứ phân vị (IQR) và điểm ngoại lai (outliers).

* Scatter plot (biểu đồ phân tán)

Dùng để thể hiện mối quan hệ giữa hai biến số.

Giúp nhận diện mối tương quan tuyến tính hoặc phi tuyến, xu hướng và cụm điểm dữ liệu.

* Bar chart (biểu đồ cột)

Dùng cho dữ liệu phân loại (categorical).

Giúp so sánh tần suất hoặc tổng giá trị giữa các nhóm.

* Line chart (biểu đồ đường)

Phù hợp với dữ liệu theo thời gian (time series), giúp theo dõi xu hướng và biến động theo chu kỳ.

* Heatmap (bản đồ nhiệt)

Dùng để hiển thị ma trận tương quan giữa các biến.

Màu sắc thể hiện cường độ của mối quan hệ.

### 3.1.3. Cách lựa chọn biểu đồ phù hợp với loại dữ liệu

|  |  |  |
| --- | --- | --- |
| **Loại dữ liệu** | **Mục đích** | **Biểu đồ phù hợp** |
| Biến số (numerical) | Phân bố, độ lệch, outliers | Histogram, Boxplot, Violin |
| Phân loại (categorical) | So sánh tần suất, tỷ lệ | Bar chart, Countplot |
| Số ↔ Số | Mối tương quan, xu hướng | Scatter plot, Line plot |
| Số ↔ Phân loại | So sánh trung bình, phân bố giữa nhóm | Boxplot, Violin |
| Dữ liệu thời gian | Xu hướng, mùa vụ | Line chart |
| Ma trận tương quan | Mối quan hệ nhiều biến | Heatmap |

Việc lựa chọn đúng loại biểu đồ sẽ giúp dữ liệu được thể hiện chính xác, trực quan và truyền tải thông tin hiệu quả nhất.

### 3.1.4. So sánh giữa các thư viện trực quan hóa trong Python

|  |  |  |  |
| --- | --- | --- | --- |
| **Thư viện** | **Ưu điểm** | **Hạn chế** | **Khi nên dùng** |
| Matplotlib | Linh hoạt, mạnh mẽ, xuất ảnh chất lượng cao (PNG, PDF) | Cú pháp dài, khó tùy chỉnh thẩm mỹ | Khi cần biểu đồ tĩnh hoặc báo cáo in ấn |
| Seaborn | Dễ dùng, giao diện đẹp, tự động xử lý thống kê | Ít tương tác, tùy chỉnh sâu vẫn cần Matplotlib | Khi phân tích EDA nhanh, tạo biểu đồ đẹp cho báo cáo |
| Plotly | Tương tác cao, đẹp, có thể xuất HTML | File lớn, cần cài thêm kaleido để xuất ảnh | Khi cần biểu đồ tương tác, trình bày online |

### 3.1.5. Nguyên tắc thiết kế biểu đồ hiệu quả

* Đơn giản – Dễ hiểu: Tránh quá nhiều màu, hiệu ứng rối mắt.
* Chính xác: Không bóp méo tỷ lệ trục, không cắt trục làm sai lệch nhận thức.
* Rõ ràng: Tiêu đề, nhãn trục, đơn vị phải đầy đủ.
* Nhấn mạnh thông tin quan trọng: Dùng màu nổi bật, annotation hợp lý.
* Nhất quán: Cùng loại dữ liệu → dùng cùng loại biểu đồ và thang màu.
* Thân thiện thị giác: Dùng palette dễ nhìn, có độ tương phản phù hợp.

### 3.1.6. Ví dụ tạo biểu đồ trong Python

* **Histogram (biểu đồ tần suất)**

import matplotlib.pyplot as plt

import numpy as np

data = np.random.normal(50, 10, 1000)

plt.figure(figsize=(8,5))

plt.hist(data, bins=25, color='skyblue', edgecolor='black', alpha=0.8)

plt.title('Phân bố giá trị - Histogram')

plt.xlabel('Giá trị')

plt.ylabel('Tần suất')

plt.grid(alpha=0.3)

plt.tight\_layout()

plt.savefig('histogram.png', dpi=300)

plt.show()

* **Bar Chart (biểu đồ cột)**

categories = ['Sản phẩm A', 'Sản phẩm B', 'Sản phẩm C']

values = [230, 190, 320]

plt.bar(categories, values, color=['#2E86C1', '#AF7AC5', '#F1948A'])

plt.title('Doanh số theo sản phẩm')

plt.xlabel('Sản phẩm')

plt.ylabel('Số lượng bán')

plt.tight\_layout()

plt.savefig('bar\_chart.pdf')

plt.show()

* **Xuất biểu đồ ra các định dạng phổ biến**

PNG (ảnh tĩnh): plt.savefig('bieudo.png', dpi=300)

PDF (ảnh vector): plt.savefig('bieudo.pdf')

HTML (tương tác - Plotly):

import plotly.express as px

fig = px.scatter(x=[1,2,3,4], y=[10,15,13,17], title='Biểu đồ tương tác')

fig.write\_html('bieudo\_tuong\_tac.html')

## 3.2. Bài làm mẫu: Trực quan hóa dữ liệu giá nhà Amsterdam

Dataset: HousingPricesData.csv - Dữ liệu giá nhà tại Amsterdam

Mục đích:Làm quen với các công cụ trực quan hóa dữ liệu sử dụng Matplotlib và Seaborn.

Các bước thực hiện:

* Chuẩn bị dữ liệu cho trực quan hóa dữ liệu
* Trực quan hóa dữ liệu với thư viện Matplotlib
* Trực quan hóa dữ liệu với thư viện Seaborn

## 3.3. Bài tập thực hành 1

Dataset: Dữ liệu về phân loại chất lượng rượu đỏ từ Kaggle

## 3.4. Bài tập thực hành 2

### 3.4.1. Bài tập 1

### 3.4.2. Bài tập 2

# PHẦN 4: PHÂN TÍCH ĐƠN BIẾN VÀ HAI BIẾN

## 4.1. Lý thuyết

**Câu hỏi 1: Phân tích đơn biến (univariate analysis) là gì? Nó khác gì với phân tích hai biến (bivariate analysis) trong khám phá dữ liệu?**

**Phân tích đơn biến (Univariate analysis)** là phương pháp thống kê dùng để phân tích một biến số đơn lẻ trong tập dữ liệu. Mục đích là hiểu rõ đặc tính của biến đó thông qua các chỉ số thống kê tóm tắt (trung bình, trung vị, mode, độ lệch chuẩn…), bảng phân phối tần suất hoặc các biểu đồ trực quan (histogram, bar chart, pie chart...).

**Phân tích hai biến (Bivariate analysis)** là phương pháp thống kê dùng để phân tích đồng thời hai biến, nhằm khám phá mối quan hệ hoặc sự liên kết giữa chúng. Kỹ thuật này giúp xác định biến này thay đổi như thế nào khi biến kia thay đổi, đo lường cường độ và ý nghĩa thống kê của mối quan hệ. Các công cụ thường dùng gồm scatter plot, boxplot, heatmap, hệ số tương quan, bảng chéo.

**Khác biệt chính**:

**+ Univariate**: tập trung vào **một biến duy nhất** để hiểu rõ đặc điểm riêng của nó.

**+ Bivariate**: xem xét hai biến cùng lúc để tìm hiểu mối quan hệ, sự tương quan hoặc ảnh hưởng qua lại.

**Câu hỏi 2: Các thước đo thống kê nào thường được sử dụng trong phân tích đơn biến (ví dụ: trung bình, trung vị, mode, độ lệch chuẩn)?**

**Thước đo xu hướng trung tâm (Measures of Central Tendency):**

+ **Trung bình (Mean):**  
Giá trị trung bình cộng của tất cả các giá trị trong tập dữ liệu.

+ **Trung vị (Median):**  
Giá trị ở giữa khi dữ liệu được sắp xếp theo thứ tự tăng dần; nó ít bị ảnh hưởng bởi các giá trị ngoại lai hơn trung bình.

+ **Yếu vị (Mode):**  
Giá trị xuất hiện nhiều nhất trong tập dữ liệu; có thể có một hoặc nhiều mode (đa mode).

**Thước đo độ phân tán (Measures of Dispersion / Variability)**

**+ Độ lệch chuẩn (Standard Deviation):**Đo lường sự phân tán của dữ liệu so với trung bình; giá trị càng nhỏ chứng tỏ dữ liệu tập trung gần trung bình, giá trị càng lớn cho thấy dữ liệu phân tán rộng.

**+ Phương sai (Variance):**  
Bình phương của độ lệch chuẩn, đo lường sự phân tán trung bình của các giá trị so với trung bình.

**+ Khoảng biến thiên (Range):**

Sự khác biệt giữa giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.

**+ Khoảng tứ phân vị (Interquartile Range – IQR):**  
Sự khác biệt giữa tứ phân vị thứ ba (Q3)(Q\_3)(Q3​) và tứ phân vị thứ nhất (Q1)(Q\_1)(Q1​), cho biết sự phân tán của 50% dữ liệu ở giữa.

**Thước đo vị trí và hình dạng phân phối**

**+ Phân vị (Percentiles / Quartiles):**

* Phân vị (Percentiles): chia tập dữ liệu thành 100 phần bằng nhau.
* Tứ phân vị (Quartiles): chia tập dữ liệu thành 4 phần bằng nhau.

**+ Độ xiên (Skewness):**

Đo lường mức độ bất đối xứng của phân phối dữ liệu.

**+ Độ nhọn (Kurtosis):**

Đo lường mức độ "nhọn" hoặc "bằng phẳng" của đỉnh phân phối so với phân phối chuẩn.

**Câu hỏi 3: Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)?**

Để xác định mối quan hệ giữa hai biến, ta sử dụng các phương pháp như phân tích tương quan để đo lường sự cùng biến động, và phân tích nhân quả (thường thông qua thí nghiệm ngẫu nhiên có kiểm soát hoặc mô hình cao cấp) để tìm ra quan hệ nguyên nhân - kết quả. Phân tích tương quan chỉ ra rằng hai biến thay đổi cùng nhau, còn nhân quả chứng minh một biến gây ra sự thay đổi ở biến kia.

**Xác định sự tương quan:**

**+ Khái niệm:** Tương quan đo lường mức độ mà hai biến có xu hướng thay đổi cùng nhau.

**+ Cách thực hiện:**

* **Phân tích trực quan:** Vẽ biểu đồ phân tán (scatter plot) để xem liệu có một xu hướng chung giữa hai biến hay không.
* **Sử dụng hệ số tương quan:** Tính toán hệ số tương quan (như [Pearson](https://www.google.com/search?cs=0&sca_esv=a69f5a9c38853b18&sxsrf=AE3TifOUOGmjJrVjm1WGlxpBSfYe_akGvw%3A1759591293557&q=Pearson&sa=X&ved=2ahUKEwiWsNSv7IqQAxXH1zQHHT8EOnkQxccNegQIFBAB&mstk=AUtExfCFCO43RGlCbJlOXvihQesaCENiwabW94tfnuoXS9n1s3H-ZUupZLGc8adFd7hdDTwXZhKZARvy_vjFHbm50xz9ILwWfDF5MSTRit2xE0c9991XO6eY4ni7MaeKaW_41anVc6d1yGY0bFFDBkyVj_vfNZYrreXv94_LlFZIXK2jURc&csui=3)) để đo lường sức mạnh và hướng của mối quan hệ tuyến tính. Hệ số này dao động từ -1 (tương quan nghịch hoàn toàn) đến +1 (tương quan thuận hoàn toàn), với 0 nghĩa là không có tương quan tuyến tính.

**Xác định nhân quả:**

**+ Khái niệm:**

Nhân quả cho thấy một biến (biến độc lập) gây ra sự thay đổi ở biến kia (biến phụ thuộc).

**+ Cách thực hiện:**

* + **Thí nghiệm ngẫu nhiên có kiểm soát (RCT):** Là phương pháp “tiêu chuẩn vàng” để thiết lập nhân quả. Trong RCT, người tham gia được phân ngẫu nhiên vào nhóm đối chứng và nhóm can thiệp để loại bỏ ảnh hưởng của các yếu tố gây nhiễu.
  + **Thiết kế nghiên cứu khác:** Khi RCT không khả thi, có thể sử dụng các thiết kế nghiên cứu quan sát phức tạp hơn, như phân tích hồi quy có kiểm soát các biến nhiễu hoặc các mô hình [ca](https://www.google.com/search?cs=0&sca_esv=a69f5a9c38853b18&sxsrf=AE3TifOUOGmjJrVjm1WGlxpBSfYe_akGvw%3A1759591293557&q=ca&sa=X&ved=2ahUKEwiWsNSv7IqQAxXH1zQHHT8EOnkQxccNegQIKhAB&mstk=AUtExfCFCO43RGlCbJlOXvihQesaCENiwabW94tfnuoXS9n1s3H-ZUupZLGc8adFd7hdDTwXZhKZARvy_vjFHbm50xz9ILwWfDF5MSTRit2xE0c9991XO6eY4ni7MaeKaW_41anVc6d1yGY0bFFDBkyVj_vfNZYrreXv94_LlFZIXK2jURc&csui=3)o cấp (causal inference models).

**+ Tiêu chí để thiết lập nhân quả (theo**[**Bradford Hill**](https://www.google.com/search?cs=0&sca_esv=a69f5a9c38853b18&sxsrf=AE3TifOUOGmjJrVjm1WGlxpBSfYe_akGvw%3A1759591293557&q=Bradford+Hill&sa=X&ved=2ahUKEwiWsNSv7IqQAxXH1zQHHT8EOnkQxccNegQIIxAB&mstk=AUtExfCFCO43RGlCbJlOXvihQesaCENiwabW94tfnuoXS9n1s3H-ZUupZLGc8adFd7hdDTwXZhKZARvy_vjFHbm50xz9ILwWfDF5MSTRit2xE0c9991XO6eY4ni7MaeKaW_41anVc6d1yGY0bFFDBkyVj_vfNZYrreXv94_LlFZIXK2jURc&csui=3)**):**

Các tiêu chí như tính nhất quán, tính mạnh, tính đặc hiệu, tính thời gian (nguyên nhân phải xảy ra trước kết quả) và mối quan hệ liều lượng-phản ứng có thể hỗ trợ việc suy luận nhân quả.

**Câu hỏi 4: Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì?**

* Hiệp biến (Covariance):

Hiệp biến được dùng để đo lường mức độ mà hai biến thay đổi cùng nhau. Nếu giá trị hiệp biến lớn hơn 0, điều đó có nghĩa là khi một biến tăng thì biến còn lại cũng có xu hướng tăng theo (hai biến thay đổi cùng chiều). Ngược lại, nếu giá trị nhỏ hơn 0, thì khi một biến tăng, biến kia lại có xu hướng giảm (hai biến thay đổi ngược chiều). Tuy nhiên, giá trị hiệp biến phụ thuộc vào đơn vị đo lường của dữ liệu, nên khó dùng để so sánh trực tiếp giữa các cặp biến khác nhau.

* Tương quan (Correlation**):**

Tương quan có thể xem là dạng chuẩn hóa của hiệp biến, giúp khắc phục nhược điểm phụ thuộc vào đơn vị đo. Giá trị của hệ số tương quan luôn nằm trong khoảng từ -1 đến 1, nên dễ dàng so sánh giữa các cặp biến khác nhau. Nếu hệ số tương quan bằng +1, hai biến có mối quan hệ tuyến tính dương hoàn hảo (một biến tăng thì biến kia cũng tăng theo đúng tỉ lệ). Nếu bằng -1, hai biến có mối quan hệ tuyến tính âm hoàn hảo (một biến tăng thì biến kia giảm theo tỉ lệ). Còn nếu hệ số gần bằng 0, tức là gần như không tồn tại mối quan hệ tuyến tính rõ ràng giữa hai biến.

* Khác biệt nhau chính:

**+ Covariance**: cho biết hai biến thay đổi cùng/khác chiều, **nhưng khó so sánh** vì phụ thuộc đơn vị đo.

**+ Correlation**: là hiệp biến đã chuẩn hóa, **dễ so sánh và trực quan hơn**, luôn nằm trong [-1, 1]

**Câu hỏi 5: Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến?**

Trong phân tích dữ liệu, biểu đồ trực quan hóa được sử dụng để giúp nhìn rõ đặc điểm phân phối hoặc mối quan hệ giữa các biến. Với phân tích đơn biến, ta thường dùng biểu đồ khi muốn mô tả đặc trưng của một biến duy nhất, chẳng hạn như histogram để quan sát phân phối giá trị, bar chart hoặc pie chart để xem tần suất và tỷ lệ của các nhóm, hoặc boxplot để phát hiện ngoại lệ. Ngược lại, với phân tích hai biến, biểu đồ được dùng để khám phá mối quan hệ giữa hai biến, ví dụ scatter plot để xem xu hướng tuyến tính giữa hai biến số, heatmap để hiển thị ma trận tương quan, hay boxplot/violin plot để so sánh phân phối của một biến số theo từng nhóm của một biến phân loại.

**Câu hỏi 6: Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?**

* Tạo dữ liệu:

A screenshot of a computer code

AI-generated content may be incorrect.

* Mẫu code tạo biểu đồ scatter plot:

A screen shot of a computer code

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.

* Mẫu code tạo biểu đồ heatmap:

A computer screen shot of text

AI-generated content may be incorrect.

A red squares with white text

AI-generated content may be incorrect.

**Câu hỏi 7: Làm thế nào để trực quan hóa mối quan hệ giữa một biến số và một biến phân loại bằng biểu đồ boxplot hoặc violin plot trong Python?**

Để trực quan hóa mối quan hệ giữa một biến số (numeric variable) và một biến phân loại (categorical variable) trong Python, ta thường sử dụng boxplot hoặc violin plot với thư viện seaborn. Biểu đồ boxplot giúp thể hiện trung vị, khoảng tứ phân vị và các giá trị ngoại lệ của biến số theo từng nhóm. Trong khi đó, violin plot không chỉ hiển thị đặc điểm của boxplot mà còn cho thấy mật độ phân phối dữ liệu trong mỗi nhóm.

## 4.2. Bài làm mẫu

### 4.2.1. 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

**1. Giới thiệu dữ liệu:**

* Bộ dữ liệu được sử dụng trong nghiên cứu này là Palmer Archipelago (Antarctica) Penguin Data, thu thập tại quần đảo Palmer, Nam Cực và được công bố công khai trên Kaggle. Đây là một tập dữ liệu phổ biến trong giảng dạy và nghiên cứu khoa học dữ liệu, bao gồm thông tin về ba loài chim cánh cụt cùng nhiều đặc trưng sinh học khác (ví dụ: chiều dài mỏ, chiều rộng mỏ, khối lượng cơ thể, giới tính...).
* Trong phạm vi phân tích đơn biến, nghiên cứu tập trung vào hai biến chính:
  + Species: Loài chim cánh cụt (Adelie, Chinstrap, Gentoo) – biến định tính, dùng để phân tích tần suất và tỷ lệ phân bố.
  + Culmen Length (mm): Chiều dài mỏ – biến định lượng liên tục, dùng để khảo sát phân phối giá trị bằng histogram.
* Mục tiêu phân tích hai biến này nhằm làm rõ đặc điểm phân bố số lượng loài và xu hướng biến thiên chiều dài mỏ trong quần thể chim cánh cụt.

**2. Phân tích đơn biến bằng Histogram**

A graph of a number of blue bars

AI-generated content may be incorrect.

*Hình 1.1. Biểu đồ phân phối của biến Culmen Length (mm)*

Quan sát histogram cho thấy phân phối chiều dài mỏ tập trung chủ yếu trong khoảng 35–50 mm, chiếm phần lớn mẫu quan sát. Đỉnh tần suất rơi vào khoảng 40–45 mm với trên 60 cá thể, phản ánh giá trị phổ biến nhất trong tập dữ liệu.

Ngược lại, số cá thể có chiều dài mỏ dưới 35 mm hoặc trên 55 mm là khá ít, được coi là ngoại biên (outliers). Đặc biệt, phần đuôi phải của phân phối có kéo dài nhẹ, với một số cá thể đạt chiều dài mỏ vượt quá 55 mm.

Điều này gợi ý rằng phân phối của biến Culmen Length gần cân đối quanh trung tâm nhưng có độ lệch phải nhẹ (slight right skew). Kết quả này hàm ý rằng đa số cá thể có mỏ với chiều dài trung bình, trong khi chỉ một nhóm nhỏ sở hữu mỏ dài vượt trội.

Hiện tượng này phản ánh đa dạng sinh học trong quần thể, đồng thời cho thấy chiều dài mỏ có thể đóng vai trò phân tách loài. Ví dụ, Gentoo Penguin thường có chiều dài mỏ lớn hơn đáng kể so với Adelie Penguin, từ đó tạo nên những giá trị ngoại biên ở phía đuôi phải. Ngoài đặc điểm loài, giới tính và tuổi sinh học cũng có thể ảnh hưởng đến đặc điểm hình thái này. Một số giá trị ngoại biên hiếm gặp có thể bắt nguồn từ sai số đo đạc hoặc điều kiện thu thập.

**3. Phân tích đơn biến bằng bar chart**

Biểu đồ Bar Chart được sử dụng để phân tích biến định tính *Species* (loài chim cánh cụt). Biểu đồ thể hiện số lượng quan sát của ba loài chính trong tập dữ liệu: Adelie Penguin (Pygoscelis adeliae), Chinstrap Penguin (Pygoscelis antarctica) và Gentoo Penguin (Pygoscelis papua).

A graph of blue squares

AI-generated content may be incorrect.

*Hình 1.2. Biểu đồ tần suất loài chim cánh cụt (Species)*

Qua biểu đồ cột cho thấy sự phân bố tần suất không đồng đều giữa ba loài được khảo sát. Loài Adelie Penguin là loài chiếm ưu thế rõ rệt trong tập dữ liệu với số lượng cá thể cao nhất, xấp xỉ 150 cá thể đã được thu thập. Loài Gentoo Penguin đứng ở vị trí thứ hai với khoảng 125 cá thể, cho thấy mật độ thu thập dữ liệu khá gần với loài Adelie. Ngược lại, loài Chinstrap Penguin được ghi nhận với tần suất thấp nhất, chỉ khoảng 70 cá thể.

Sự chênh lệch tần suất này báo hiệu một mức độ mất cân bằng lớp (class imbalance) đáng kể trong tập dữ liệu. Tỷ lệ số lượng của loài Adelie so với loài Chinstrap là hơn 2:1, điều này cần được xem xét nghiêm túc trong các giai đoạn phân tích tiếp theo, đặc biệt là khi xây dựng mô hình phân loại. Sự mất cân bằng này có thể dẫn đến thiên vị mô hình (model bias), trong đó mô hình học máy có xu hướng dự đoán chính xác hơn cho các lớp đa số (Adelie và Gentoo) và kém hiệu quả hơn đối với lớp thiểu số (Chinstrap).

Do đó, ở các bước xử lý tiếp theo, cần cân nhắc áp dụng các kỹ thuật cân bằng dữ liệu như oversampling, undersampling, hoặc sử dụng các thuật toán thích ứng với dữ liệu mất cân bằng, nhằm đảm bảo tính công bằng và nâng cao độ chính xác của mô hình trên toàn bộ các lớp.

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

Biểu đồ tròn (Pie Chart) được sử dụng để thể hiện tỷ lệ phần trăm quan sát của từng loài chim cánh cụt (*Species*) trong tập dữ liệu. Biểu đồ này giúp dễ dàng nhận diện tỷ trọng tương đối giữa các nhóm, thay vì số lượng tuyệt đối như Bar Chart.

A pie chart with different colored circles

AI-generated content may be incorrect.

*Hình 1.3. Biểu đồ tròn thể hiện tỷ lệ phân bố loài*

Kết quả cho thấy loài Adelie Penguin (Pygoscelis adeliae) chiếm tỷ lệ cao nhất, khoảng 44% tổng số mẫu. Tiếp đến là loài Gentoo Penguin (Pygoscelis papua) với khoảng 36%, trong khi Chinstrap Penguin (Pygoscelis antarctica) có tỷ lệ thấp nhất, chỉ khoảng 20%.

Quan sát này cho thấy dữ liệu tập trung chủ yếu ở hai loài Adelie và Gentoo, trong khi số lượng mẫu của loài Chinstrap ít hơn đáng kể. Thông tin từ biểu đồ tròn góp phần củng cố nhận định từ biểu đồ cột, rằng phân bố mẫu giữa các loài không đồng đều.

**5. Nhận xét và giải thích**

Kết quả phân tích đơn biến cho thấy chiều dài mỏ (Culmen Length) của chim cánh cụt chủ yếu tập trung trong khoảng 35–50 mm, với đỉnh tần suất ở khoảng 40–45 mm. Phân phối có dạng gần cân đối quanh trung tâm nhưng vẫn tồn tại một số giá trị ngoại biên, chủ yếu nằm ở phía đuôi phải trên 55 mm. Hiện tượng này có thể bắt nguồn từ sự khác biệt hình thái giữa các loài, trong đó loài Gentoo thường có chiều dài mỏ lớn hơn đáng kể so với loài Adelie, từ đó tạo nên những cá thể có mỏ dài vượt trội so với trung bình. Ngoài đặc trưng loài, các yếu tố sinh học như giới tính, độ tuổi cũng có khả năng tác động đến sự biến thiên này, trong khi một số ngoại biên có thể xuất phát từ sai số đo đạc hoặc điều kiện thu thập dữ liệu thực địa.

Đối với biến định tính Species, kết quả từ cả Bar Chart và Pie Chart đều phản ánh sự phân bố mẫu không đồng đều giữa ba loài. Adelie là loài chiếm ưu thế với khoảng 44% số mẫu, tiếp theo là Gentoo với khoảng 36%, trong khi Chinstrap chỉ chiếm 20%. Sự mất cân bằng này có thể giải thích từ hai khía cạnh: thứ nhất, yếu tố sinh thái tự nhiên khi Adelie và Gentoo có phạm vi phân bố rộng hơn, mật độ quần thể cao hơn tại khu vực khảo sát; thứ hai, yếu tố thiết kế nghiên cứu và thiên lệch trong quá trình lấy mẫu, khi dữ liệu có xu hướng tập trung nhiều hơn vào các loài phổ biến. Việc nhận diện sự mất cân bằng này đặc biệt quan trọng vì nó có thể ảnh hưởng trực tiếp đến các phân tích tiếp theo, chẳng hạn làm tăng nguy cơ thiên lệch trong mô hình dự đoán hoặc giảm độ chính xác khi phân loại loài thiểu số như Chinstrap.

Phân tích đơn biến đã cung cấp cái nhìn tổng quan về đặc điểm phân bố của các biến trong tập dữ liệu, đồng thời gợi ý những hướng phân tích sâu hơn ở mức nhị biến hoặc đa biến nhằm làm rõ hơn mối quan hệ giữa chiều dài mỏ và loài chim cánh cụt.

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

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

Trong nghiên cứu này, nhóm tiến hành phân tích dữ liệu giá nhà tại Amsterdam với mục tiêu khảo sát đặc trưng phân bố và nhận diện các giá trị bất thường. Sau khi nạp dữ liệu và lựa chọn các biến số quan trọng như mã vùng, giá, diện tích và số phòng, nhóm xây dựng biểu đồ boxplot để mô tả sự phân bố giá trị của biến Price.

A graph of a house price

AI-generated content may be incorrect.

*Hình 2.1. Biểu đồ boxplot thể hiện phân bố giá nhà tại Amsterdam*

Biểu đồ boxplot cho thấy phần lớn giá nhà tập trung trong khoảng dưới một triệu (theo đơn vị dữ liệu), thể hiện qua phần hộp chính của biểu đồ. Tuy nhiên, bên cạnh đó xuất hiện rất nhiều điểm ngoại lai (outliers) nằm rải rác về phía bên phải, kéo dài đến mức giá hơn sáu triệu. Hiện tượng này cho thấy dữ liệu giá nhà có phân bố lệch phải (right-skewed), trong đó đa số bất động sản thuộc nhóm giá trung bình hoặc thấp, nhưng tồn tại một số ít bất động sản có giá trị đặc biệt cao, vượt trội so với mặt bằng chung.

Phát hiện này phản ánh đặc điểm thực tế của thị trường bất động sản tại Amsterdam: phần lớn giao dịch tập trung ở phân khúc phổ thông, trong khi phân khúc cao cấp tuy hiếm nhưng có giá trị rất lớn, góp phần tạo nên sự chênh lệch rõ rệt trong phân phối dữ liệu. Đồng thời, sự tồn tại của nhiều ngoại lai cũng cho thấy các thống kê mô tả như trung bình hoặc phương sai có thể bị ảnh hưởng mạnh, do đó cần được cân nhắc khi sử dụng ở các bước phân tích sau.

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

A blue line graph with black lines

AI-generated content may be incorrect.

*Hình 2.2. Biểu đồ violin plot thể hiện mật độ phân bố giá nhà*

Quan sát từ violin plot, ta nhận thấy phần lớn giá nhà tập trung rất dày ở khoảng dưới 1 triệu. Đây là phân khúc chiếm ưu thế trên thị trường, với mật độ cao và trải rộng trong vùng giá thấp đến trung bình. Khi giá tăng dần, mật độ dữ liệu giảm rõ rệt và kéo dài về phía bên phải đến mức hơn 6 triệu. Phần “đuôi” bên phải của biểu đồ khá dài nhưng mật độ rất thưa, điều này cho thấy chỉ có một số rất ít bất động sản có giá trị vượt trội, đóng vai trò là các điểm ngoại lai.

Violin plot giúp minh họa trực quan sự lệch phải (right-skewed) trong phân bố giá nhà: đa số bất động sản tập trung ở vùng giá vừa phải, trong khi một nhóm nhỏ thuộc phân khúc cao cấp có giá cực kỳ cao, tạo ra sự phân hóa rõ rệt. So với boxplot, violin plot mang lại cái nhìn trực quan hơn về hình dạng phân bố, giúp ta thấy rõ không chỉ sự tồn tại của ngoại lai mà còn cả mật độ dữ liệu giữa các mức giá khác nhau.

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

A screenshot of a graph

AI-generated content may be incorrect.

*Bảng 2.3. Thống kê mô tả biến Price trong dữ liệu giá nhà tại Amsterdam*

Dữ liệu mô tả về biến Price cho thấy một sự phân bố giá trị đáng chú ý trong tập dữ liệu gồm 920 quan sát. Giá trị trung bình (Mean) của giá nhà là , trong khi trung vị (50th Percentile) chỉ là . Sự chênh lệch lớn này—trung bình cao hơn trung vị tới hơn 155,000—là một dấu hiệu rõ ràng cho thấy sự phân bố lệch phải dương (positively skewed). Điều này ngụ ý rằng tồn tại một số lượng nhỏ các căn nhà có giá rất cao đang kéo giá trị trung bình lên. Hơn nữa, với độ lệch chuẩn (Standard Deviation) là , gần bằng giá trị trung bình, cho thấy sự biến động và phân tán giá rất lớn. Phạm vi giá trị trải dài từ mức tối thiểu là đến mức tối đa là . Giá trị tối đa này cao gấp hơn 12 lần giá trị trung vị, cung cấp bằng chứng mạnh mẽ về sự tồn tại của các giá trị ngoại lai (outliers) cực kỳ đắt đỏ. Phân tích tứ phân vị cho thấy 50% giá nhà tập trung trong khoảng từ (Q1) đến (Q3), xác nhận rằng đại đa số các giao dịch nhà nằm dưới mức , trong khi phần còn lại của dữ liệu bị chi phối bởi các bất động sản cao cấp.

**4. Nhận xét và giải thích**

Kết quả phân tích từ boxplot, violin plot và thống kê mô tả đều nhất quán, khẳng định rằng phân phối giá nhà tại Amsterdam có dạng lệch phải rõ rệt. Phần lớn giao dịch diễn ra ở mức giá trung bình và phổ thông (dưới 1 triệu), trong khi các bất động sản cao cấp hiếm hoi nhưng có giá trị vượt trội đã kéo dài đuôi phân phối.

Hiện tượng này phản ánh bối cảnh thực tiễn của thị trường bất động sản Amsterdam. Là một trung tâm kinh tế – tài chính quan trọng của châu Âu với mật độ dân cư cao nhưng quỹ đất hạn chế, Amsterdam có thị trường nhà ở phổ thông chiếm ưu thế về số lượng giao dịch. Ngược lại, các căn hộ hạng sang hoặc biệt thự cao cấp thường phục vụ giới thu nhập rất cao hoặc mục đích đầu tư quốc tế, dẫn đến sự xuất hiện của những giá trị ngoại lai cực lớn.

Ngoài ra, chính sách quy hoạch đô thị cũng góp phần định hình phân bố này: việc hạn chế phát triển dự án quy mô lớn tại trung tâm thành phố làm gia tăng mật độ giao dịch trong phân khúc vừa túi tiền, trong khi nguồn cung nhà ở cao cấp tuy ít nhưng vẫn tồn tại và tạo ra sự chênh lệch mạnh mẽ trong phân phối giá.

Tổng hợp lại, phân tích đơn biến không chỉ giúp mô tả đặc trưng của dữ liệu mà còn đưa ra những gợi ý quan trọng cho các bước tiếp theo. Đặc biệt, sự tồn tại của ngoại lai và phân phối lệch là những yếu tố cần được cân nhắc kỹ lưỡng khi xây dựng các mô hình dự báo giá nhà, nhằm tránh sai lệch trong ước lượng và cải thiện tính chính xác.

### 4.2.1. 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 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

A screenshot of a computer

AI-generated content may be incorrect.

**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

A blue dots on a white background

AI-generated content may be incorrect.

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.

A diagram of different colored dots

AI-generated content may be incorrect.

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

A table with numbers and text

AI-generated content may be incorrect.

**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

A screenshot of a computer

AI-generated content may be incorrect.

**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

A collage of blue and white bars

AI-generated content may be incorrect.

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.

### 4.2.1. 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>)

A close-up of a text

AI-generated content may be incorrect.

Hình 4: Hướng dẫn tải thư viện ydata\_profiling

2. Sử dụng công cụ

A green and white striped flag

AI-generated content may be incorrect.

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:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 5:

## 4.3. 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

A white rectangular object with black text

AI-generated content may be incorrect.

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.

Nhóm khách hàng có học vấn cao hơn có xu hướng chi nhiều hơn cho rượu vang (MntWines) với correlation≈ 0.21.

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

A screenshot of a computer

AI-generated content may be incorrect.

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.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 10: Phân bố thu nhập khách hàng (Income)

**Dựa vào hình ta thấy:**

Cột Income (thu nhập) có 24 giá trị bị thiếu.

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

AI-generated content may be incorrect.

Hình 11: 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 hàng qua catalog(0.52), 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.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 12: Phân bố số trẻ vị thành niên (Teenhome)

**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 liên hệ cao với NumDealsPurchases (0.40), cho thấy các gia đình có trẻ vị thành niên quan tâm hơn đến các chương trình khuyến mãi.

Có liên hệ trung bình với MntMeatProducts (0.28) và thấp hơn với MntFishProducts (0.21), MntFruits (0.18) và MntSweetProducts (0.17), gợi ý rằng các gia đình này chi tiêu thực phẩm nhiều hơn.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 13: Phân bố ngày đăng ký khách hàng (Dt\_Customer)

**Dựa vào hình ta thấy:**

Phần lớn khách hàng gia nhập với 53% vào năm 2013, 25% vào năm 2014 và 22% vào năm 2012.

Có mối liên hệ yếu nhất định với NumWebVisitsMonth (0.25), cho thấy thời điểm khách hàng gia nhập ảnh hưởng một phần tới tần suất truy cập web.  
Có liên hệ khá yếu với NumDealsPurchases (0.19) và NumWebPurchases (0.17), gợi ý rằng khách hàng gia nhập sớm hơn có xu hướng tương tác với các khuyến mãi và mua sắm trực tuyến nhiều hơn.

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.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 14: 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.

**A screenshot of a computer

AI-generated content may be incorrect.**

Hình 15: Phân bố số lần mua trên website (NumWebPurchases)

**Dựa vào hình ta thấy:**

Trong các biến phân loại, Kidhome (0.37) có liên hệ mạnh nhất, cho thấy số con nhỏ trong gia đình ảnh hưởng rõ rệt đến hành vi/mục tiêu.

Trong các biến số liên tục, MntWines (0.54), NumStorePurchases (0.50) và MntGoldProds (0.42) có liên hệ tích cực mạnh, trong khi Year\_Birth (-0.15) và NumWebVisitsMonth (-0.06) có liên hệ âm yếu.

**A screenshot of a computer

AI-generated content may be incorrect.**

Hình 16: Phân bố số lần mua qua catalog (NumCatalogPurchases)

**Dựa vào hình ta thấy:**

Phần lớn khách hàng mua qua catalog từ 2–6 lần, một số ít mua nhiều hơn, lên tới khoảng 10 lần.

Trong các biến phân loại, Kidhome (0.52) có liên hệ mạnh nhất, cho thấy số con nhỏ trong gia đình ảnh hưởng rõ rệt đến hành vi/mục tiêu.

Trong các biến số liên tục, MntMeatProducts (0.72), MntWines (0.64) và Income (0.59) có liên hệ tích mạnh.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 17: 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.

Trong các biến phân loại, Kidhome (0.51) có liên hệ mạnh nhất, cho thấy số con nhỏ trong gia đình ảnh hưởng đáng kể đến số lần mua tại cửa hàng.

Trong các biến số liên tục, MntWines (0.64), Income (0.53) và NumCatalogPurchases (0.52) có liên hệ tích cực mạnh, trong khi NumWebVisitsMonth (-0.43) có liên hệ âm vừa phải.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 18: 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.

Trong các biến phân loại, Kidhome (0.46) có liên hệ mạnh nhất, cho thấy số con nhỏ trong gia đình ảnh hưởng đáng kể đến tần suất truy cập website.

Trong các biến số liên tục, Income (-0.55), MntMeatProducts (-0.54) và NumCatalogPurchases (-0.52) có liên hệ âm mạnh, nghĩa là khách hàng có thu nhập cao hoặc chi tiêu nhiều cho thực phẩm và catalog thường ít truy cập website hơn.



Hình 19: 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.

## 4.4. 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

A close-up of a white background

AI-generated content may be incorrect.

Hình 20: Hướng dẫn tải thư viện autoviz

2. Sử dụng công cụ và EDA:

Bảng 4: Nhận xét tổng quan về chất lượng dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.

**Nhận xét:**

**ID**: Không thiếu dữ liệu nhưng là cột định danh, nên loại bỏ trước khi mô hình hóa.

**Year\_Birth**: Không thiếu dữ liệu nhưng có 3 giá trị ngoại lệ, cần xử lý outlier (cắt hoặc loại bỏ).

**Education:** Không có vấn đề.

**Marital\_Status:** Có 3 giá trị hiếm ['Alone', 'Absurd', 'YOLO'], nên gộp hoặc loại bỏ.

**Income:** 24 giá trị thiếu, có 8 outlier, cần impute và xử lý outlier.

**Kidhome, Teenhome, Recency, AcceptedCmp1-5, Complain, Response**: Không có vấn đề.

Dt\_Customer: Có thể biến đổi thành datetime trước khi mô hình hóa.

**MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds:** Có nhiều outlier, nên cắt hoặc loại bỏ.

**NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumWebVisitsMonth:** Một số outlier, cần xử lý.

**NumStorePurchases:** Không có vấn đề.

**Z\_CostContact, Z\_Revenue:** Giá trị không đổi hoặc ít biến thiên, nên loại bỏ trước khi mô hình hóa.

A diagram of a graph

AI-generated content may be incorrect.

**Dựa vào ảnh ta thấy:**

Biểu đồ Distplot cho thấy phân phối thu nhập lệch phải mạnh, phần lớn khách hàng có thu nhập dưới 100,000.

Boxplot cho thấy thu nhập khách hàng chủ yếu ở mức thấp với trung vị thấp, râu hộp kéo dài về phía trên và chỉ có ít ngoại lệ ở nhóm thu nhập cao.

Probability Plot có skew = 6.8 và các điểm lệch xa đường chéo, chứng tỏ dữ liệu Income không tuân theo phân phối chuẩn và bị ảnh hưởng bởi các giá trị cực đoan.

A group of colored rectangular bars

AI-generated content may be incorrect.

Hình 21: Biểu đồ cột thể hiện tỷ lệ phản hồi các chiến dịch marketing

**Dựa vào ảnh ta thấy:**

Các biến phản hồi chiến dịch (AcceptedCmp1–5) đều có tỉ lệ 0 chiếm ưu thế (>80–90%), cho thấy đa số khách hàng không tham gia hoặc không phản hồi tích cực trong các chiến dịch trước.

Biến Complain cũng chủ yếu bằng 0, chứng tỏ phần lớn khách hàng không khiếu nại.

A graph of different colored squares

AI-generated content may be incorrect.

Hình 22: Biểu đồ cột phân bố tỷ lệ phản hồi chiến dịch và trình độ học vấn của khách hàng

**Dựa vào ảnh ta thấy:**

Biến Response cho thấy tỉ lệ phản hồi thấp, chứng tỏ xu hướng tương tác của khách hàng với các chiến dịch marketing nhìn chung chưa cao.

Phần lớn khách hàng có trình độ học vấn cao (Graduation chiếm tỷ trọng lớn).

A graph and a chart

AI-generated content may be incorrect.

Hình 23: Biểu đồ Histogram phân bố tình trạng hôn nhân và năm sinh của khách hàng

**Dựa vào ảnh ta thấy:**

Biến Marital\_Status cho thấy đa số khách hàng đã kết hôn hoặc đang sống chung, phản ánh nhóm khách hàng ổn định về gia đình và tài chính.

Biến Year\_Birth tập trung chủ yếu trong giai đoạn 1950–1980, cho thấy khách hàng trung niên chiếm ưu thế, đây là nhóm có khả năng chi tiêu cao.

A graph of a bar and a bar

AI-generated content may be incorrect.

Hình 24: Biểu đồ cột phân bố Kidhome và Teenhome

**Dựa vào ảnh ta thấy:**

Biến Kidhome và Teenhome cho thấy đa số khách hàng có từ 0–1 con, chứng tỏ phần lớn họ sống trong gia đình nhỏ hoặc không có con phụ thuộc, ổn định và có khả năng chi tiêu cá nhân cao.

A group of graphs with different colored bars

AI-generated content may be incorrect.

Hình 25: Biểu đồ Histogram phân bố Recency

**Dựa vào ảnh ta thấy:**

Biến Recency cho thấy phần lớn khách hàng đã lâu không mua hàng, phân bố nghiêng về các giá trị cao (khoảng 50–60), phản ánh mức độ tương tác gần đây thấp.

A graph of different sizes and colors

AI-generated content may be incorrect.A group of graphs with different colored bars

AI-generated content may be incorrect.

Hình 26: Biểu đồ Histogram phân bố tần suất mua hàng

**Dựa vào ảnh ta thấy:**

Các biến chi tiêu (MntWines, MntFishProducts, MntMeatProducts, MntSweetProducts, MntGoldProds, MntFruits,) đều có phân phối lệch phải, cho thấy đa số khách hàng chi tiêu ở mức thấp, chỉ một số ít chi tiêu cao hơn, đại diện cho nhóm khách hàng giá trị cao.

A group of graphs with different colors

AI-generated content may be incorrect.A group of graphs with different colors

AI-generated content may be incorrect.

**Dựa vào ảnh ta thấy:**

Các biến tần suất mua hàng (NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth) đều có phân phối lệch phải, cho thấy đa số khách hàng mua sắm với tần suất thấp, chỉ một nhóm nhỏ có hành vi mua hàng thường xuyên hơn.

A red propeller with black text

AI-generated content may be incorrect.

Hình 27: Biểu đồ Violin (Income)

**Dựa vào ảnh ta thấy:**

**Biểu đồ violin** cho thấy thu nhập (Income) phân bố không đồng đều, phần lớn khách hàng có thu nhập ở mức thấp đến trung bình, trong khi một số ít có thu nhập rất cao (outliers), thể hiện phân phối lệch phải rõ rệt.

A colorful grid with different colors

AI-generated content may be incorrect.

**Dựa vào hình ta thấy:**

Các biến chi tiêu như MntWines, MntGoldProds và MntMeatProducts có tương quan dương khá mạnh với nhau (r = 0.5–0.7), thể hiện hành vi mua sắm ổn định và chi tiêu cao ở cùng nhóm khách hàng.

Biến Response tương quan dương nhẹ với MntWines (r ≈ 0.25) và với các chiến dịch AcceptedCmp3, AcceptedCmp5 (r = 0.26–0.32), cho thấy khách hàng từng chi tiêu lớn và có phản hồi tích cực trước đây là nhóm tiềm năng cho chiến dịch mới.  
Ngoài ra, NumWebVisitsMonth có tương quan âm với NumWebPurchases (r ≈ -0.45), gợi ý rằng khách hàng truy cập web nhiều nhưng ít mua có thể cần chiến lược khuyến khích chuyển đổi tốt hơn.

A comparison of a bar graph

AI-generated content may be incorrect.

Hình 28: Biểu đồ Bar plots thể hiện Average Income theo Education và Marital\_Status

**Dựa vào hình ta thấy:**

Thu nhập trung bình tăng dần theo trình độ học vấn. Nhóm có bằng PhD và Master có thu nhập cao nhất, trong khi nhóm Basic có thu nhập thấp hơn đáng kể. Cho thấy học vấn cao là yếu tố gắn liền với thu nhập cao.

Các nhóm Single và Together có thu nhập cao hơn so với Divorced và Widow. Người đang trong mối quan hệ hoặc độc thân có xu hướng thu nhập tốt hơn nhóm đã ly hôn hoặc góa.

A graph with numbers and lines

AI-generated content may be incorrect.

Hình 29: Biểu đồ Bar plots thể hiện Average Income theo Year\_Birth và Kidhome

**Dựa vào hình ta thấy:**

Nhìn chung, khách hàng trẻ hơn có thu nhập cao hơn so với các nhóm lớn tuổi. Điều này có thể phản ánh xu hướng thu nhập theo thế hệ hoặc giai đoạn nghề nghiệp.

Thu nhập trung bình thấp hơn khi số con trong nhà tăng. Gia đình có ít con thường có khả năng tài chính mạnh hơn.

A close-up of a graph

AI-generated content may be incorrect.

Hình 30: Biểu đồ Bar plots thể hiện Average Income theo Teenhome và Recency

**Dựa vào hình ta thấy:**

Thu nhập trung bình thấp hơn khi số con trong nhà tăng. Gia đình có ít con thường có khả năng tài chính mạnh hơn.

Khách hàng có Recency thấp (gần đây mua hàng) có xu hướng thu nhập cao hơn nhẹ. Người thu nhập cao có thể thường xuyên hơn mua hàng.

A group of different colored bars

AI-generated content may be incorrect.

Hình 31: Biểu đồ Bar plots thể hiện Average Income theo việc Chi tiêu cho các loại sản phẩm

**Dựa vào hình ta thấy:**

Thu nhập trung bình tăng rõ rệt theo mức chi tiêu, đặc biệt là với Wines và Gold Products. Các sản phẩm có giá trị cao (rượu, vàng) thường gắn liền với nhóm khách hàng thu nhập cao.

A group of different colored squares

AI-generated content may be incorrect.

A group of graphs with different colors

AI-generated content may be incorrect.

Hình 32: Biểu đồ Bar plots thể hiện Average Income theo Tần suất mua hàng

**Dựa vào hình ta thấy:**

CatalogPurchases và WebPurchases: nhóm có thu nhập cao chi tiêu nhiều qua kênh này.

DealsPurchases (mua giảm giá): chủ yếu tập trung ở nhóm thu nhập thấp.

WebVisitsMonth: nhóm thu nhập cao có xu hướng truy cập web ít hơn.

A group of different colored squares

AI-generated content may be incorrect.

A group of different colored squares

AI-generated content may be incorrect.

Hình 33: Biểu đồ Bar plots thể hiện Average Income theo Mức phản hồi chiến dịch (AcceptedCmp1–5)

**Dựa vào hình ta thấy:**

Khách hàng chấp nhận hoặc phản hồi tích cực chiến dịch marketing có thu nhập cao hơn so với nhóm không phản hồi. Chiến dịch marketing thu hút mạnh hơn nhóm khách hàng thu nhập cao.

A close-up of a graph

AI-generated content may be incorrect.

Hình 34: Biểu đồ Bar plots thể hiện Average Income theo Complain và Response

**Dựa vào hình ta thấy:**

Thu nhập trung bình của hai nhóm có và không khiếu nại khá tương đồng, cho thấy mức thu nhập không ảnh hưởng rõ rệt đến hành vi phàn nàn của khách hàng.

Nhóm khách hàng đã phản hồi có thu nhập cao hơn, phản ánh xu hướng khách hàng giàu có hơn quan tâm và phản ứng với chiến dịch marketing tốt hơn.

# PHẦN 5: KẾT LUẬN

## 5.1. Tổng kết kiến thức đã học

Qua quá trình thực hiện bài thực hành Lab 01 về Phân tích khám phá dữ liệu (EDA), chúng em đã nắm vững và củng cố được nhiều kiến thức quan trọng trong lĩnh vực khoa học dữ liệu.

### 5.1.1. Về thống kê mô tả

Chúng em đã hiểu rõ vai trò và cách sử dụng các thước đo thống kê cơ bản:

**Thước đo xu hướng trung tâm:**

* **Mean (Trung bình):** Cho biết giá trị trung tâm của dữ liệu, nhưng dễ bị ảnh hưởng bởi outliers. Qua dataset COVID-19, em thấy mean = 8,017 trong khi median = 0, cho thấy mean bị kéo lên bởi các giá trị cực đại.
* **Median (Trung vị):** Ổn định hơn mean khi có outliers. Là thước đo tốt hơn cho dữ liệu lệch như thu nhập, giá nhà, số ca COVID.
* **Mode (Yếu vị):** Giá trị xuất hiện nhiều nhất, hữu ích cho dữ liệu phân loại.

**Thước đo độ phân tán:**

* **Standard Deviation:** Đo mức độ phân tán quanh mean. Giá trị cao (như 229,664 trong COVID data) cho thấy dữ liệu rất phân tán.
* **IQR (Interquartile Range):** Đo độ phân tán của 50% dữ liệu giữa, ít bị ảnh hưởng bởi outliers hơn standard deviation.
* **Range:** Khoảng biến thiên từ min đến max, giúp hiểu phạm vi dữ liệu.

**Quartiles và Boxplot:**

* Q1, Q2 (median), Q3 chia dữ liệu thành 4 phần bằng nhau
* Giúp phát hiện outliers (giá trị ngoài khoảng [Q1 - 1.5×IQR, Q3 + 1.5×IQR])
* Quan trọng trong việc đánh giá phân bố dữ liệu

### 5.1.2. Về trực quan hóa dữ liệu

Em đã làm quen và thành thạo việc sử dụng các loại biểu đồ phổ biến:

**Các loại biểu đồ và ứng dụng:**

* **Histogram:** Xem phân bố của biến số, phát hiện skewness
* **Bar chart:** So sánh giá trị giữa các categories
* **Box plot:** Phát hiện outliers, so sánh phân bố giữa các nhóm
* **Scatter plot:** Xem mối quan hệ giữa 2 biến số
* **Heatmap:** Visualize correlation matrix
* **Pair plot:** Xem tất cả mối quan hệ cùng lúc

**So sánh Matplotlib vs Seaborn:**

* Matplotlib: Linh hoạt, kiểm soát chi tiết, phù hợp khi cần customize sâu
* Seaborn: Đơn giản, đẹp hơn, tích hợp tốt với Pandas, phù hợp EDA nhanh
* Cả hai đều quan trọng và bổ sung cho nhau

### 5.1.3. Về phân tích đơn biến và hai biến

**Phân tích đơn biến (Univariate):**

* Phân tích từng biến riêng lẻ để hiểu phân bố, xu hướng, outliers
* Sử dụng histogram, bar chart, boxplot, violin plot
* Tính các chỉ số thống kê mô tả

**Phân tích hai biến (Bivariate):**

* Tìm mối quan hệ giữa 2 biến: correlation, patterns
* Sử dụng scatter plot, heatmap, crosstab, pivot table
* Phân biệt được correlation và causation

**Công cụ EDA tự động:**

* **SweetViz:** Tạo báo cáo HTML đẹp, tương tác, phù hợp presentation
* **AutoViz:** Nhanh, nhiều biểu đồ, phù hợp khám phá với dataset lớn
* Cả hai giúp tiết kiệm thời gian nhưng vẫn cần phân tích thủ công để hiểu sâu

## 5.2. Kỹ năng đạt được

### 5.2.1. Kỹ năng lập trình Python

**Thư viện Pandas:**

* Load dữ liệu từ CSV với pd.read\_csv()
* Khám phá dữ liệu: head(), shape, dtypes, describe()
* Xử lý dữ liệu: drop\_duplicates(), dropna(), fillna(), replace()
* Lọc và chọn dữ liệu: indexing, boolean masking
* Group và aggregate: groupby(), pivot\_table(), crosstab()

**Thư viện NumPy và SciPy:**

* Tính toán thống kê: np.mean(), np.median(), np.std(), np.var()
* Percentiles và quantiles: np.percentile(), np.quantile()
* Mode và IQR: stats.mode(), stats.iqr()
* Xử lý NaN: np.nanmean(), np.nanmedian()

**Thư viện Matplotlib:**

* Tạo figure và axes: plt.figure(), plt.subplot()
* Plot cơ bản: plt.bar(), plt.plot(), plt.scatter()
* Customize: plt.title(), plt.xlabel(), plt.ylabel()
* Font và size: fontsize, figsize

**Thư viện Seaborn:**

* Statistical plots: sns.histplot(), sns.boxplot(), sns.violinplot()
* Categorical plots: sns.barplot(), sns.countplot()
* Relational plots: sns.scatterplot(), sns.pairplot()
* Heatmap: sns.heatmap()

### 5.2.2. Kỹ năng xử lý dữ liệu

**Phát hiện vấn đề dữ liệu:**

* Kiểm tra missing values: isnull().sum()
* Phát hiện duplicates: duplicated().sum()
* Tìm outliers: IQR method, boxplot
* Kiểm tra kiểu dữ liệu: dtypes

**Xử lý dữ liệu:**

* Missing values:
  + Drop: khi tỷ lệ thấp (< 5%)
  + Fill: median cho numeric, mode cho categorical
  + Interpolate: cho time series
* Outliers:
  + Loại bỏ nếu là lỗi nhập liệu
  + Giữ lại nếu là giá trị thật
  + Transform: log, sqrt
* Duplicates: drop\_duplicates()
* Data type conversion: astype()

### 5.2.3. Kỹ năng phân tích và tư duy

**Đọc và diễn giải biểu đồ:**

* Hiểu ý nghĩa của histogram: phân bố, skewness, modality
* Đọc boxplot: median, quartiles, outliers, spread
* Phân tích scatter plot: correlation, clusters, trends
* Diễn giải heatmap: correlation strength và direction

**Tư duy phản biện:**

* Không tin vào số liệu một cách mù quáng
* Luôn kiểm tra chất lượng dữ liệu trước khi phân tích
* Đặt câu hỏi: Tại sao? Có hợp lý không? Có outliers không?
* So sánh nhiều thước đo (mean vs median) để hiểu rõ hơn

**Rút ra insights:**

* Từ COVID data: Phần lớn ngày có 0 ca, nhưng có đợt bùng phát mạnh
* Từ Marketing data: Dataset sạch, chỉ 1% missing
* Từ House prices: Cần xem cả Price và PriceperSqm để đánh giá đầy đủ

## 5.3. Khó khăn gặp phải và cách khắc phục

### 5.3.1. Khó khăn về kỹ thuật

**Vấn đề 1: Kết quả tính toán bị NaN**

* **Nguyên nhân:** Dữ liệu có missing values, NumPy/SciPy không thể tính toán với NaN
* **Cách khắc phục:**
  + Kiểm tra missing values trước khi tính: isnull().sum()
  + Loại bỏ NaN: dropna() hoặc dùng np.nanmean(), np.nanmedian()
  + Luôn xử lý missing values trước khi phân tích

**Vấn đề 2: Biểu đồ không hiển thị hoặc bị lỗi**

* **Nguyên nhân:** Thiếu plt.show(), hoặc lỗi syntax
* **Cách khắc phục:**
  + Luôn có plt.show() ở cuối
  + Kiểm tra import: import matplotlib.pyplot as plt
  + Test trên dữ liệu nhỏ trước

**Vấn đề 3: Code chạy chậm với dataset lớn**

* **Nguyên nhân:** Dataset COVID có > 400k dòng
* **Cách khắc phục:**
  + Filter dữ liệu trước khi visualize
  + Dùng sample: df.sample(n=1000)
  + Tối ưu code, tránh vòng lặp không cần thiết

### 5.3.2. Khó khăn về phân tích

**Vấn đề 1: Không biết chọn biểu đồ nào**

* **Cách khắc phục:**
  + Học thuộc quy tắc: numeric → histogram/boxplot, categorical → bar chart, 2 numeric → scatter
  + Tham khảo bài mẫu và documentation
  + Thử nhiều loại biểu đồ khác nhau

**Vấn đề 2: Khó diễn giải kết quả**

* **Cách khắc phục:**
  + So sánh mean vs median để hiểu phân bố
  + Xem std để biết độ phân tán
  + Vẽ biểu đồ để visualize
  + Đọc thêm tài liệu và case studies

**Vấn đề 3: Không biết xử lý outliers như thế nào**

* **Cách khắc phục:**
  + Hiểu ngữ cảnh: outlier có hợp lý không?
  + Kiểm tra xem có phải lỗi nhập liệu
  + Tham khảo domain knowledge
  + Thử cả 2 cách: giữ và loại bỏ, so sánh kết quả

### 5.3.3. Khó khăn về thời gian

**Vấn đề:** Mất nhiều thời gian để làm quen với syntax

* **Cách khắc phục:**
  + Đọc documentation của Pandas, Matplotlib, Seaborn
  + Thực hành với dataset nhỏ trước
  + Tham khảo các tutorial và example code
  + Ghi chú lại code hay dùng để tái sử dụng

## 5.4. Bài học rút ra

### 5.4.1. Về quy trình EDA

**EDA là bước quan trọng nhất:**

* Không thể bỏ qua, phải làm trước khi modeling
* 80% thời gian Data Science dành cho data cleaning và EDA
* Một EDA tốt giúp tiết kiệm thời gian và nâng cao chất lượng model

**Quy trình chuẩn:**

1. Load và xem dữ liệu (head(), info(), shape)
2. Kiểm tra chất lượng (missing, duplicates, outliers)
3. Thống kê mô tả (describe())
4. Trực quan hóa (histogram, boxplot, scatter...)
5. Xử lý vấn đề (clean data)
6. Phân tích sâu hơn (correlation, patterns)
7. Rút ra insights và đề xuất

### 5.4.2. Về kỹ năng phân tích

**Cần nhiều góc nhìn:**

* Không chỉ nhìn 1 thước đo (ví dụ: chỉ mean)
* So sánh mean vs median, std vs IQR
* Xem cả số liệu và biểu đồ
* Phân tích đơn biến trước, sau đó hai biến

**Hiểu ngữ cảnh quan trọng:**

* Dataset COVID: 0 ca là bình thường ở giai đoạn đầu
* Dataset Marketing: Missing 1% là chấp nhận được
* Dataset House: Cần xem cả Price và PriceperSqm

**Tư duy phản biện:**

* Luôn hỏi "Tại sao?"
* Kiểm tra tính hợp lý của số liệu
* Không tin vào outliers ngay, phải xác minh
* So sánh với kiến thức thực tế

### 5.4.3. Về công cụ và kỹ thuật

**Không có công cụ hoàn hảo:**

* Matplotlib: Mạnh nhưng phức tạp
* Seaborn: Đẹp nhưng ít linh hoạt
* SweetViz: Nhanh nhưng ít customize
* AutoViz: Tự động nhưng thiếu context
* → Cần biết khi nào dùng gì

**Kết hợp là tốt nhất:**

* Dùng SweetViz/AutoViz để overview nhanh
* Dùng Seaborn cho biểu đồ đẹp trong báo cáo
* Dùng Matplotlib khi cần customize chi tiết
* Dùng Pandas cho data manipulation

**Học hỏi liên tục:**

* Documentation là nguồn tài liệu tốt nhất
* Stackoverflow giúp giải quyết lỗi
* Kaggle có nhiều notebook hay để học
* Practice makes perfect

## 5.5. Hướng phát triển

**Thống kê nâng cao:**

* Hypothesis testing (t-test, chi-square, ANOVA)
* Probability distributions chi tiết hơn
* Confidence intervals
* Statistical significance

**Trực quan hóa nâng cao:**

* Interactive visualizations với Plotly, Bokeh
* Dashboard với Dash, Streamlit
* Geospatial visualization với Folium
* 3D plots

**Kỹ thuật EDA nâng cao:**

* Feature engineering
* Dimensionality reduction (PCA, t-SNE)
* Time series analysis
* Text data analysis

**Machine Learning:**

* Supervised learning (Regression, Classification)
* Unsupervised learning (Clustering, Anomaly detection)
* Feature selection
* Model evaluation

# LỜI KẾT

Bài thực hành Lab 01 về Phân tích khám phá dữ liệu đã mang lại cho chúng em những kiến thức nền tảng vô cùng quan trọng trong hành trang sau này. Chúng em đã không chỉ học được các kỹ thuật thống kê và trực quan hóa, mà còn rèn luyện được tư duy phản biện và kỹ năng phân tích dữ liệu một cách có hệ thống.

Qua quá trình thực hành với các dataset thực tế (COVID-19, Marketing Campaign, House Prices), chúng em nhận ra rằng EDA không chỉ là việc chạy code và tạo biểu đồ, mà là nghệ thuật kể chuyện bằng dữ liệu (storytelling with data). Mỗi con số, mỗi biểu đồ đều chứa đựng những thông tin quý giá, và nhiệm vụ của người phân tích là phát hiện và diễn giải chúng một cách chính xác.

Chúng em cũng nhận thức sâu sắc hơn về tầm quan trọng của việc làm sạch và chuẩn bị dữ liệu. Câu nói "Garbage in, garbage out" là hoàn toàn đúng - nếu dữ liệu đầu vào kém chất lượng, thì bất kỳ phân tích hay model nào cũng sẽ cho kết quả sai lệch.

Mặc dù còn nhiều kiến thức cần học và kỹ năng cần rèn luyện, nhưng bài thực hành này đã trang bị cho chúng em một nền tảng vững chắc để tiếp tục khám phá thế giới khoa học dữ liệu. Chúng em cam kết sẽ tiếp tục học hỏi, thực hành và áp dụng những kiến thức đã học vào các dự án thực tế.

Cuối cùng, em xin chân thành cảm ơn Thầy đã tận tình hướng dẫn và tạo điều kiện để chúng em hoàn thành tốt bài thực hành này. Những kiến thức và kỹ năng chúng em học được sẽ là hành trang quý giá cho sự nghiệp tương lai.

# TÀI LIỆU THAM KHẢO