TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA TOÁN - ỨNG DỤNG

HỌC PHẦN: KHAI PHÁ DỮ LIỆU

LAB03: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU

SINH VIÊN THỰC HIỆN:

NGUYỄN ĐĂNG TIẾN MSSV: 3123590050

GIẢNG VIÊN: TS.ĐỖ NHƯ TÀI

Thành phố Hồ Chí Minh - 2025

Mục lục

CHƯƠNG I: GIỚI THIỆU 4

1. Lý do chọn đề tài 4

1.1. Lý do khoa học 4

1.2. Lý do thực tiễn 4

2. Ý nghĩa thực tiễn 4

3. Mục tiêu nghiên cứu 5

4. Phạm vi và giới hạn đề tài 5

4.1. Phạm vi nghiên cứu 5

4.2. Giới hạn đề tài 5

5. Phương pháp tiếp cận 6

CHƯƠNG II: CỞ SỞ LÝ THUYẾT 6

1. Tổng quan về khám phá dữ liệu 6

1.1. Khái niệm 6

1.2. Vai trò 6

1.3. Hạn chế 6

2. Giới thiệu các kỹ thuật liên quan 7

2.1. Kỹ thuật thống kê 7

3. Kỹ thuật quản lý và tiền sử lý dữ liệu 9

4. Kỹ thuật trực quan hóa dữ liệu 10

4.1. Khái niệm 10

4.2. Một số hình thức phổ biến như 10

CHƯƠNG III: DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU 17

1. Nguồn dữ liệu 17

1.1. Dữ liệu COVID-19 (Nguồn: COVID-19 Pandemic - Our World in Data) 17

1.2. Dữ liệu Marketing Campaign (Nguồn: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis) 17

1.3. Dữ liệu phân loại rươu đỏ (Nguồn: https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification) 17

1.4. Dữ liệu bệnh tiểu đường (Nguồn: https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906) 18

1.5. Dữ liệu giá nhà Amsterdam (Nguồn: https://www.kaggle.com/datasets/thomasnibb/amsterdam) 19

2. Thực hành khám phá dữ liệu 19

CHƯƠNG IV: TÓM TẮT THỰC HÀNH 32

CHƯƠNG I: GIỚI THIỆU

1. Lý do chọn đề tài

1.1. Lý do khoa học

Nhận thấy việc Khai phá dữ liệu chính là một trong những lĩnh vực cốt lõi đóng vai quan trọng trong Khoa học dữ liệu, giúp rút trích các tri thức tiềm ẩn từ dữ liệu lớn mà con người không thể tự quan sát.

Trong thời đại bối cảnh mà dữ liệu đang ngày càng được sử dụng nhiều cũng như sự phổ biến của Máy học cũng như Trí tuệ nhân tạo, Việc nghiên cứu và áp dụng các kỹ thuật khai phá dữ liệu không chỉ giúp **phát hiện các mẫu ẩn, mối quan hệ tiềm tàng trong dữ liệu**, mà còn **hỗ trợ dự đoán xu hướng tương lai**, góp phần củng cố vai trò của khai phá dữ liệu trong mọi lĩnh vực thực tiễn.

Đề tài báo cáo góp phần củng cổ ứng dụng của khai phá đòng thời giới thiệu mô tả các bước xử lý dữ liệu mục đích nhằm cho thấy tầm quan trọng của chúng trong mọi lĩnh vực thực tế

1.2. Lý do thực tiễn

Các tổ chức, công ty, xí nghiệp hiện nay đang có nguồn dữ liệu lớn được thu thập hằng ngày, nhưng chưa thể tận dụng hiểu quả tối đa để có thể đưa ra quyết định. Đồng thời cũng như việc tự động hóa phân tích dữ liệu bằng khai phá cũng chưa được tận dụng tốt. Do đó mà việc sử dụng khai phá dữ liệu cũng giúp tang hiệu quả, giảm chi phí và ra quyết định chính xác hơn

Đề tài này nhằm góp phần minh chứng cho **hiệu quả thực tiễn của việc áp dụng khai phá dữ liệu trong phân tích và dự đoán**, từ đó khẳng định tầm quan trọng của nó trong việc hỗ trợ ra quyết định trong doanh nghiệp và xã hội hiện đại.

2. Ý nghĩa thực tiễn

Nghiên cứu này tập trung vào việc phát triển và triển khai các giải pháp Khai phá Dữ liệu nhằm nâng cao năng lực khai thác và xử lý dữ liệu cho các đối tượng là sinh viên, nhà nghiên cứu, và các tổ chức. Việc ứng dụng Khai phá Dữ liệu (Data Mining) mang lại ý nghĩa thiết thực trong bối cảnh dữ liệu lớn (Big Data) ngày càng gia tăng, cụ thể như sau:

Nâng cao hiểu biết về Quy trình phân tích và trích xuất tri thức dự án nhằm cung cấp một khung khổ lý thuyết và thực tiễn toàn diện, giúp các đối tượng thụ hưởng nắm vững quy trình phân tích dữ liệu chuyên sâu, từ khâu làm sạch, biến đổi dữ liệu, cho đến áp dụng các thuật toán khai phá để khám phá các mẫu hình (patterns) và trích xuất tri thức tiềm ẩn từ các tập dữ liệu quy mô lớn. Điều này là nền tảng cho việc chuyển đổi dữ liệu thô thành thông tin có giá trị chiến lược.

3. Mục tiêu nghiên cứu

Mục tiêu của bài báo cáo này nhằm củng cố giúp cho người đọc nắm vững các kỹ thuật cơ bản trong khám phá dữ liệu nhằm mục đích không gì khác là hiểu rõ đặc điểm và cấu trúc của tập dữ liệu. Cụ thể là các kỹ thuật phân tích thống kê mô tả nhằm xác định các đặc trưng chính như giá trị trung bình, trung vị, độ lệch chuẩn và phân bố của dữ liệu, thêm vào đó bài báo cáo sử dụng các công cụ trực quan hóa như biểu đồ histogram, boxplot và scatterplot để phát hiện các mẫu, xu hướng hoặc bất thường trong dữ liệu. Và thư viện Pandas, Matplotlib, Seaborn để xử lý và trực quan hóa dữ liệu một cách hiệu quả và có hệ thống.

Ngoài ra bài báo cáo cũng giới thiệu đến các vấn đè như giá trị bị thiếu, giá trị ngoại lai, hoặc sự không nhất quán trong dữ liệu từ đó đề xuất các phương pháp tiền xử lý phù hợp. Giúp đưa ra các nhận định ban đầu về dữ liệu, đặt nền tảng cho các bước phân tính sâu hơn hoặc xây dựng mô hình khai thac dữ liệu trong các ứng dụng thực tiễn như phân tích khách hàng hoặc dự đoán xu hướng

4. Phạm vi và giới hạn đề tài

4.1. Phạm vi nghiên cứu

Phạm vi của bài báo cáo nằm trong khuôn khổ của môn học “Khai phá dữ liệu” - Trường Đại học Sài Gòn – Giảng viên: Đỗ Như Tài

Bộ dữ liệu được sử dụng trong báo cáo đươc cung cấp trong các bài giảng, bài tập thuộc trong phạm trù của môn học.

4.2. Giới hạn đề tài

Giới hạn của bài báo cáo chỉ dựa trên phần bài làm đươc đính kèm cùng với bài báo cáo này giúp giải thích, làm rõ một số các vấn đề cũng như đầu ra (output) mà trong bài thực hành chưa thể giải thích được

5. Phương pháp tiếp cận

Sử dụng ngôn ngữ lập trình Python để thực hiện việc phân tích và đánh giá dữ liệu thực hiện thống kê mô tả, sử dụng thư viện Pandas để có thể tương tác với các tập dữ liệu như khai báo hay xuất nhập tập dữ liệu, Numpy để chuyển đổi các kiểu dữ liệu, Matplotlib và Seaborn để trực quan hóa dữ liệu giúp cho việc khám phá dữ liệu trở nên dễ dàng hơn

Tất cả những thư viện các phương pháp đều được thực hiện trên Jupyter Lab trong Anaconda

CHƯƠNG II: CỞ SỞ LÝ THUYẾT

1. Tổng quan về khám phá dữ liệu

1.1. Khái niệm

Khám phá dữ liệu (Data Mining) là quá trình phân tích và trích xuất tri thức từ các tập dữ liệu lớn, phức tạp hoặc đa dạng nguồn. Mục tiêu của khai phá dữ liệu chính là phát hiện các mẫu, mối quan hệ, xu hướng hoặc quy luật tìm ẩn của dữ liệu mà trước đây chưa từng biết đến, từ đó hỗ trợ đưa ra quyết định trong kinh doanh, khoa học và đời sống.

Theo Han và Kamber (2011), khai phá dữ liệu là ***bước cốt lõi trong quy trình khám phá tri thức trong cơ sở dữ liệu (Knowledge Discovery in Databases – KDD)****. Nó không chỉ đơn thuần là việc thu thập hoặc lưu trữ dữ liệu, mà là* ***quá trình chuyển đổi dữ liệu thành tri thức hữu ích****.*

1.2. Vai trò

Vai trò của khai phá dữ liệu trong thời đại dữ liệu lớn chính là tự động phát hiện tri thức trong kho dữ liệu lớn mà con người khó nhận ra bằng cách thủ công, hỗ trợ ra quyết định dựa trên dữ liệu, dự đoán xu hướng và hành vi giúp tối ưu các hoạt động của doanh nghiệp, cuối cùng chính là nền tảng cgo các lĩnh vực mới như trí tuệ nhân tạo, học máy, phân tích dự đoán và khoa học dữ liệu

1.3. Hạn chế

Khai phá dữ liệu đa dạng như vậy, hữu dụng như vậy. Tuy nhiên, chúng cũng gặp phải nhiều thách thức chẳng hạn như: Chất lượng dữ liệu kém (thiếu, bị lệch hoặc không đồng nhất giữa các nguồn), khối lượng dữ liệu khổng lồ (Dữ liệu ngày càng lớn, đa dạng và thay đổi liên tục,đòi hỏi khả năng tính toán cao),Tính

bảo mật và quyền riêng tư (Việc thu thập và khai thác dữ liệu cá nhân có thể xâm phạm đến quyền riêng tư)

2. Giới thiệu các kỹ thuật liên quan

2.1. Kỹ thuật thống kê

Thống kê mô tả: Là một nhánh của thống kê liên quan đến việc tóm tắt sắp xếp và trình bày dữ liệu, giúp tô tả đặc điểm chỉnh của tập dữ liệu mà không đưa ra bất kỳ suy diễn nào trong tổng thể lớn hơn

Sự khác nhau giữa thống kê mô tả và thống kê suy luận chính là thống kê mô tả chỉ tóm tắt và mô tả các tính năng chính của tập dữ liệu chứ không đi sâu vào suy luận dự đoán như thống kê suy luận

Các thước đo thống kê mô tả bao gồm:

Trung vị: là thước đo khuynh hướng trung tâm trong thống kê,dung để biểu diễn giá trị ở giữa của một tập dữ liệu sau khi đã được sắp xép theo thứ tự tang giảm.

Cách tính trung vị;

Trường hợp 1: n là số lẻ thì

$$Median = \frac{x_{n+1}}{2}$$

Trường hợp 2: n là số chẵn

$$Median = \frac{x_{n+1} + x_{n+2}}{2}$$

Trung bình: là thước đo khuynh hướng trung tâm phổ biến nhất trong thống kê, dung đẻ mô tả giá trị đại diện cho một tập dữ liệu

$$Mean = \sum \frac{x_i}{n}$$

Phương sai: là một thước đo độ phân tán (mức độ biến động) của dữ liệu so với giá trị trung bình (các giá trị trong tập dữ liệu khác nhau bao nhiêu so với trung bình)

+ Nếu các giá trị gần trung bình -> Phương sai nhỏ

+ Nếu các giá trị xa trung bình -> Phương sai lớn

Cách tính phương sai:

Trường hợp 1: Với mẫu

Type equation here.

Trường hợp 2: Với toàn bộ dữ liệu

Type equation here.

Độ lệch chuẩn: là thước đo mức đo phân tán của dữ liệu quanh giá trụ trung bình. Nói cách khác, nó cho biết trung bình mỗi giá trị tổng tập dữ liệu lệch bao nhiêu so với trung bình

+ Nếu độ lệch chuẩn nhỏ -> các giá trị tập trung gần trung bình

+ Nếu độ lệch chuẩn lớn -> dữ liệu phân tán rộng, không ổn định

Cách tính độ lệch chuẩn

Trường hợp 1: Với mẫu

Type equation here.

Trường hợp 2: Với toàn bộ dữ liệu

Type equation here.

Đa phần những thước đo trên được sử dụng đẻ có thể tóm tắt, mô tả và cung cấp thông tin tổng quát của bộ dữ liệu, tùy theo nhu cầu trong bảng dữ liệu mà người ta phân chia chúng thành hai nhóm:

+ Nếu muốn đo lường xu hướng trung tâm ta sử dụng: ***Trung bình hoặc trung vị***

+ Nếu muốn đo lường độ phân tán ta có thể sử dụng: ***Phương sai, Độ lệch chuẩn***

Một lưu ý nhỏ rằng ta nên sử dụng trung vị thay vì trung bình khi tập dữ liệu có giá trị ngoại lai hoặc dữ liệu bị lệch (*Giá trị ngoại lai: Thông thường được xem như là giá trị quá cao hoặc quá thấp so với các giá trị còn lại, có thẻ nhìn thấy giá trị này khi sử dụng biểu đồ Boxplot*)

3. Kỹ thuật quản lý và tiền sử lý dữ liệu

Trước khi khái phá, dữ liệu thường ở dạng thô, thiếu, nhiễu hoặc không đồng nhất do đó cần qua giai đọna tiền xử lý để đảm bảo chất lượng.

Các bước chính:

1. *Làm sạch dữ liệu: Loại bỏ dữ liệu trùng, sửa giá trị sai, điền dữ liệu thiếu*

Trước khi đến với những phần thống kê mô tả ta cũng không thể quên bước xử lý các giá trị bị thiếu. Có nhiều cách để ta có thể xử lý các giá trị bị thiếu như:

+ **Loại bỏ:** Thực hiện xóa các dòng hoặc cột chứa giá trị thiếu. Phương pháp này đơn giản những có thẻ làm mất mát thông tin quan trọng neus dữ liệu thiếu nhiều

+ **Điền giá trị:** Ta có thẻ thay thế giá trị bị thiếu bằng giá trung binfhm trung vị hoặc mode, điền giá trị trước và sau, Sử dụng mô hình học máy đẻ điền giá trị thiếu

Trong thực hiện khai phá dữ liệu ta sẽ gặp rất nhiều cấc điểm giá trị ngoại lại. Như vậy làm sao để ta có thẻ xử lý được nó:

- **Giữ nguyên:** Nếu giá trị ngoại lai là hợp lệ và thể hiện sự biến động thực tế của dữ liệu (ví dụ: thu nhập của người giàu), bạn có thể giữ chúng lại, đặc biệt khi sử dụng các thước đo thống kê ít nhạy cảm với ngoại lai như trung vị hoặc IQR.
- **Loại bỏ:** Nếu giá trị ngoại lai là do lỗi nhập liệu hoặc đo lường, bạn có thể loại bỏ các dòng chứa chúng. Tuy nhiên, cần cân nhắc cẩn thận để không làm mất mát quá nhiều dữ liệu.
- **Chuyển đổi:** Áp dụng các phép biến đổi toán học (ví dụ: logarit, căn bậc hai) để giảm ảnh hưởng của các giá trị lớn.
- **Winsorizing:** Thay thế các giá trị ngoại lai bằng các giá trị cận trên/dưới của "râu" trong boxplot hoặc một ngưỡng nhất định.

2. *Tích hợp dữ liệu: Két hợp dữ liệu từ nhiều nguồn khác nhau*
3. *Biến đổi dữ liệu: Chuẩn hóa, rút gọn hoặc mã hóa dữ liệu về dạng phù hợp*
4. *Rút trích đặc trưng: Lựa chọn hoặc tạo ra các thuộc tính có ý nghĩa nhất để phục vụ mô hình học*

4. Kỹ thuật trực quan hóa dữ liệu

4.1. Khái niệm

Trực quan hóa dữ liệu là kỹ thuật hỗ trợ hiểu và diễn giải kết quả khai phá dữ liệu một cách trực quan, các công cụ trực quan có thể giúp người phân tích nhận biết nhanh các mẫu, xu hướng và ngoại lệ trong dữ liệu

Nếu hỏi tại sao trực quan hóa lại có vai trò quan trọng vì

Thứ nhất, hiểu phân bố dữ liệu nhanh chóng nhìn thấy hình dạng, xu hướng trung tâm và sự phân tán của các biến

Thứ hai, phát hiện mẫu và xu hướng, nhận diện các mối quan hệ,nhóm hoặc xu hướng tiềm ẩn trong dữ liệu mà có thể khó thấy trong dữ liệu thô

Thứ ba, nhận diện giấ trị ngoại lai và bất thường dễ dàng phát hiện các điểm dữ lieu nằm xa phần còn lại, có thể là lỗi hoặc thống tin quan trọng

Thứ tư, truyền đạt kết quảm trình bày các phát biểu một cách rõ rang, dễ hiểu và thuyết phục cho người khác

4.2. Một số hình thức phổ biến như

Biểu đồ mô tả phân bố dữ liệu:

+ **Histogram:** Là biểu đồ tần sất biểu diễn phân bố của một biến số liên tục. Nó trả lời cho câu hỏi dữ liệu được phân bố như thế nào – tập trùng, lệch trái, lệch phải hay trải rộng

+ Trục X (hoành): các khoảng giá trị của dữ liệu
+ Trục T (tung): tàn suất xuất hiện hoặc mật độ xác suất của dữ liệu trong mỗi khoảng

+ Boxplot: Là biểu đồ mô tả phân bố phân bố dữ liệu dựa trên 5 số thống kê chính

+ Hộp thể hiện dữ liệu nằm giữa Q1 và Q3 (50% trung tâm)

+ Đường giữa hộp là **trung vị (median)**.

+ Râu (whiskers) kéo dài đến giới hạn của dữ liệu không phải outlier.

+ Các điểm ngoài râu là **outliers**.

Để có thể đọc hiểu và diễn giải một biểu đồ Histogram hoặc Boxplot từ dữ liệu thự tế ta có thể xem theo các bước sau:

- *Đối với Histogram:*
  + **Hình dạng:** Xem xét hình dạng tổng thể (đối xứng, lệch trái/phải, có nhiều đỉnh không) để hiểu về phân bố dữ liệu.
  + **Vị trí đỉnh:** Đỉnh biểu đồ cho biết giá trị hoặc khoảng giá trị xuất hiện nhiều nhất (mode).
  + **Phạm vi:** Quan sát khoảng giá trị mà dữ liệu trải dài.
  + **Khoảng trống/Bất thường:** Tìm kiếm các khoảng trống hoặc các thanh đơn lẻ xa các thanh khác, có thể là dấu hiệu của các nhóm dữ liệu riêng biệt hoặc giá trị ngoại lai.

Biểu đồ mô tả mối quan hệ

+ **Scatterplot**: là một trong những biểu đồ cơ bản và quan trọng nhất trong phân tích dữ liệu, khoa học dữ liệu và khai phá dữ liệu thể hiện quan hệ giữa hai biến số và một biến phụ thuộc Y (Ứng với mỗ một điểm trên X sẽ có một điểm dối chiếu với nó trên Y)

+ **Heatmap:** Biểu đồ thể hiện giá trị bằng màu sắc, dung để hiển thị mối tương quan giữa nhiều biến hoặc ma trận dữ liệu. Giusp ta phân tích tương quan giữa các biến trong tập dữ liệu

+ **Pairplot:** Là tập hợp nhiều scatter plot và histogram hiển thị mọi cặp biến trong tập dữ liệu giúp quan sát nhanh các mối quan hệ giữa các biến và phân bố của từng biến đó

- Đối với các biến khác nhau ta sử dụng scatter để chỉ ra các mối quan hệ
- Còn với chính nó sẽ hiện thị histogram thể hiện phân bố của chúng ta có thể quan sát theo đường chéo

Biểu đồ mô tả thành phần:

+ **Pie chart:** Biểu đồ tròn thể hiện tỷ lệ phần trăm của các hạng mục trong tổng thể tức chúng chiếm bao nhiêu phần tram so với tổng thể, mỗi lát cắt biểu diên tỷ lệ phần tram của một nhóm

+ **Stacked bar chart**: Là biểu đồ cột trong đó các thành phần nhỏ dduwoc chồng lên nhau trong cũng một cột thể hiện tổng giá trị và cấu trúc thành phần bên trong. Giusp so sánh tổng và thành phần giữa các nhóm

Biểu đồ mô tả xu hướng

+ **Line chart**: Là bieru đồ thể hiện xu hướng của dữ liệu theo thời gian với mỗi điểm nối với nhau bằng đường thẳng trên biểu đồ. Giúp phân tích xu hướng tang/giảm theo thời gian, theo dõi sự thay đổi liên tục

+ **Area chart**: Giống biểu đồ đường, nhưng phần dưới đường được tổ màu, thể hiện mức độ tích lũy hoặc tổng hợp theo thời gian. Giúp só sánh tỷ trọng thay đổi qua thời gian giữa các danh mục, diễn giải trực quan hơn khi dữ liệu có xu hướng tang dần

Biểu đồ mô tả so sánh

+ **Bar chart**: Là loại biểu đồ dung các thanh (cột) để biểu diễn giá trị của các danh mục (categories).

+ *Trục hoành:* biểu diễn các danh mục

+ *Trục tung:* biểu diễn giá trị số

+ *Chiều cao cột:* thể hiện mức độ hoặc tần suất của mỗi danh mục

+ *Màu sắc:* có thể dung để phân nhóm hoặc so sánh nhiều biến cùng lúc

+ **Column chart:** Là một dạng của **biểu đồ cột (bar chart)**, nhưng các **cột được vẽ theo chiều dọc (vertical)** thay vì nằm ngang.
Mỗi cột biểu diễn **giá trị của một danh mục (category)**, và chiều cao của cột phản ánh **mức độ hoặc tần suất** của biến số tương ứng.

+ Trục hoành: chứa các danh mục

+ Trục tung: biểu diễn giá trị

+ Cột đứng: Mỗi cột biểu diễn giá trị của một danh mục cụ thể

Chẳng hạn như nếu ta muốn xác định phân bố của tập dữ liệu ta có thể sử dụng các biểu đồ như histogram hoặc biểu đồ mật độ (KDE). Tuy nhiên khi ta quan sát các biểu đồ ta sẽ nhận thấy có những phân bố như:

- **Phân bố chuẩn:** Có dạng đối xứng, hình chuông,ở đỉnh nằm ở trung tâm
- **Phạm vi:** Là hiệu số giữa giá trị lớn nhất và giá trị nhỏ nhất trong tập dữ liệu. Nó cung cấp
- **Phân bố lệch phải:** Đuôi phân bố dài hơn về phái. Gía trị trung bình thường lớn hơn trung vị

- **Phân bố lệch trái:** Đuôi phân bố dài hơn về trái. Gía trị trung bình thường nhỏ hơn trung vị
- **Phân bố đều:** Tất cả các giá trị có tần suất xuất hiện xấp xỉ nhau
- **Phân bố binomial:** Có hai hoặc nhiều đỉnh

Ngoài những biểu đồ trên ta còn có một biểu đồ khác như Boxplot mục đích của biểu đồ này nhằm xác định các giá trị ngoại lai một cách rõ ràng đồng thời cũng thể hiện các phân bố của dữ liệu (IQR, Max, Min, Mean):

+ **Giá trị Max:** Gía trị lớn nhất trong bảng dữ liệu

+ **Giá trị Min:** Giá trị nhỏ nhất trong tập dữ liệu

+ **Giá trị Mean:** Giá trị trung bình trong tập dữ liệu

+ **Giá trị IQR:** là thước đo độ phân tán của dữ liệu thường được tính bằng hiệu giữa tứ phân vi thứ **(Q3)** và **tứ phân vị thứ nhất (Q1)**:

- Nếu IQR nhỏ dữ liệu tập trung gần trung vị (ít biến động)
- Nếu IQR **lớn**, dữ liệu phân tán rộng (nhiều biến động).

+ Q1: Là giá trị tại đó 25% dữ liệu hoăc bằng cạnh dưới của “box”

+ Q2: Là giá trị tại đó 50% dữ liệu hoặc bằng cạnh ngang của “box”

+ Q3: Là giá trị tại đó 75% dữ liệu hoặc trên cạnh ngang của “box”

Mỗi giá trị kiểu dữ liệu với tùy mục đích khác nhau có thể lựa chọn loại biểu đô khác nhau nhằm đáp ứng mục tiêu. Cụ thể

- **Một biến số liên tục:** Histogram, KDE plot, Boxplot, Violin plot.
- **Một biến phân loại:** Bar chart, Countplot, Pie chart (thận trọng khi dùng pie chart với nhiều danh mục).
- **Mối quan hệ giữa hai biến số liên tục:** Scatter plot, Line plot (nếu có yếu tố thời gian).
- **Mối quan hệ giữa một biến số liên tục và một biến phân loại:** Boxplot, Violin plot, Grouped bar chart (nếu biến số được tóm tắt theo danh mục).
- **Mối quan hệ giữa hai biến phân loại:** Stacked bar chart, Grouped bar chart, Heatmap (cho bảng tần suất).
- **Mối quan hệ giữa nhiều biến số:** Pair plot, Heatmap (cho ma trận tương quan).

- **Dữ liệu thời gian:** Line plot (cho xu hướng theo thời gian), Time series plot

Trong bài báo cáo này ta sẽ sử dụng thư viện thuộc ngôn ngữ lập trình Python chính là Matplotlb và Seaborn hay Pyplot để có thể thục hiện trực quan hóa. Sụ khác biệt của cả ba thư viện này nằm ở

- **Matplotlib:** Là thư viện trực quan hóa cơ bản và linh hoạt nhất trong Python. Nó cung cấp khả năng kiểm soát chi tiết mọi yếu tố của biểu đồ. Tuy nhiên, việc tạo ra các biểu đồ phức tạp hoặc đẹp mắt có thể đòi hỏi nhiều dòng mã hơn.
- **Seaborn:** Dựa trên Matplotlib và cung cấp giao diện cấp cao hơn để tạo các biểu đồ thống kê hấp dẫn và cung cấp thông tin. Seaborn đặc biệt hữu ích cho việc trực quan hóa mối quan hệ giữa các biến và làm việc với DataFrame của Pandas. Nó thường yêu cầu ít mã hơn Matplotlib cho các biểu đồ thống kê phổ biến.
- **Plotly:** Là một thư viện tương tác, cho phép tạo các biểu đồ động, có thể phóng to, thu nhỏ, hiển thị thông tin khi di chuột (tooltips) và nhúng vào web. Plotly không chỉ tạo biểu đồ tĩnh mà còn tập trung vào khả năng tương tác.

Bên cạnh đó khi thực hiện trực quan hóa ta cũng đã đặt ra những quy tắc để có thể thiết kế một biểu đồ sao cho vữa đẹp vừa mang nhiều ý nghĩa và bao quát toàn bộ nội dung mang lại hiệu quả tối ưu cho doanh nghiệp. Ta có thể tham khảo qua các nguyên tắc sau như:

  - **Rõ ràng và đơn giản:** Biểu đồ nên dễ hiểu ngay từ cái nhìn đầu tiên. Tránh sự lộn xộn không cần thiết.
  - **Tiêu đè rõ ràng:** Biểu đồ cần có tiêu đề mô tả nội dung.
  - **Nhãn trục rõ ràng:** Các trục X và Y cần có nhãn mô tả đơn vị và ý nghĩa.
  - **Chú thích:** Nếu biểu đồ có nhiều yếu tố hoặc nhóm, hãy sử dụng chú thích để phân biệt chúng.
  - **Chọn màu sắc phù hợp:** Sử dụng màu sắc một cách nhất quán và có ý nghĩa. Tránh sử dụng quá nhiều màu hoặc các màu khó phân biệt.
  - **Đúng tỷ lệ:** Đảm bảo các tỷ lệ trên biểu đồ phản ánh đúng dữ liệu. Tránh làm méo mó biểu đồ.
  - **Nhấn mạnh thông điệp chính:** Thiết kế biểu đồ sao cho thông điệp hoặc phát hiện quan trọng nhất được làm nổi bật.

- **Sử dụng loại biểu đồ phù hợp:** Như đã đề cập ở câu trên, chọn loại biểu đồ thể hiện tốt nhất mối quan hệ hoặc phân bố dữ liệu.

Một code mẫu ví dụ cho việc trực quan hóa dữ liệu bằng việc sử dụng biểu đồ histogram và bar để biểu diễn phân bố

```
import matplotlib.pyplot as plt
import numpy as np

# Dữ liệu mẫu cho histogram
data_for_hist = np.random.randn(1000)

# Tạo Histogram
plt.figure(figsize=(8, 5)) # Điều chỉnh kích thước của biểu đồ
plt.hist(data_for_hist, bins=30, edgecolor='black') # bins là số lượng cột
plt.title('Histogram của dữ liệu mẫu') # Tiêu đề cột
plt.xlabel('Giá trị') # Tiêu đề cột hoành
plt.ylabel('Tần suất') # Tiêu đề cột tung
plt.show()

# Dữ liệu mẫu cho bar chart
categories = ['A', 'B', 'C', 'D', 'E']
values = [23, 45, 56, 12, 30]

# Tạo Bar Chart
plt.figure(figsize=(8, 5)) # Điều chỉnh kích thước biểu đồ
plt.bar(categories, values) # Tạo bar
plt.title('Bar Chart của các danh mục')# Tiêu đề biểu đồ
plt.xlabel('Danh mục') # Tiêu đề cột hoành
plt.ylabel('Giá trị') # Tiêu đề cột tung
plt.show()
```

CHƯƠNG III: DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU

1. Nguồn dữ liệu

1.1. Dữ liệu COVID-19 (Nguồn: COVID-19 Pandemic - Our World in Data)

Là bộ dữ liệu công khai tổng hợp số liệu về địa dịch COVID-19 trên toàn cầu. Bộ dữ liệu được sử dụng, cung cấp dựa trên những thu thập từ **Our World in Data (OWID) –** một dự án dữ liệu mở cung cấp các thống kê toàn cầu vê xét nghiêm, và tỷ lệ tiêm chủng theo từng quốc gia

1.2. Dữ liệu Marketing Campaign (Nguồn: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis)

Tập dữ liệu có 541,909 hàng và 8 cột

_ InvoiceNo: Số hóa đơn

_ StockCode: Mã sản phẩm

_ Description: Mô tả sản phẩm

_ Quantity: Số lượng sản phẩm

_ InvoiceDate: Ngày giờ hóa đơn

_ UnitPrice: Đơn giá

_ CustomerID: Mã khách hàng

_ Country: Quốc gia

1.3. Dữ liệu phân loại rươu đỏ (Nguồn: https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification)

Tập dữ liệu Rượu đỏ chứa 1599 mẫu với 12 thuộc tính (11 thuộc tính hóa học và 1 thuộc tính mục tiêu là quality).

Hầu hết các thuộc tính hóa học được lưu trữ dưới dạng số thực (float64), trong khi điểm chất lượng (quality) là số nguyên (int64), hoàn toàn phù hợp với bản chất của dữ liệu.

_fixed acidity: Độ axit cố định

_ volatile acidity: Độ axit dễ bay hơi

_ citric acid: Axit citric

_ residual sugar: Lượng đường dư

_ chlorides: Clorua

_ free sulfur dioxide: Lưu huỳnh đioxit tự do

_ total sulfur dioxide: Tổng lưu huỳnh đioxit

_ density: Mật độ

_ pH: Độ pH

_ sulphates: Sulfat

_ alcohol: Nồng độ cồn

_ quality: Điểm chất lượng (3-8)

1.4. Dữ liệu bệnh tiểu đường (Nguồn: https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906)

Tập dữ liệu này (Pima Indians Diabetes) có 768 mẫu với 9 thuộc tính.

_Pregnancies: Số lần mang thai

_ Glucose: Nồng độ glucose trong huyết tương

_ BloodPressure: Huyết áp tâm trương
_ SkinThickness: Độ dày da (cơ thể)
_ Insulin: Mức Insulin huyết thanh
_ BMI: Chỉ số khối cơ thể
_ DiabetesPedigreeFunction: Chức năng phả hệ tiểu đường
_ Age: Tuổi
_ Outcome: Biến mục tiêu: 1 (có tiểu đường) / 0 (không tiểu đường)

_ Tập dữ liệu này nổi tiếng với việc có các giá trị 0 cho các biến không thể bằng 0 trong thực tế (Glucose, BloodPressure, BMI, v.v.). Đây là những giá trị bị thiếu (missing values) được mã hóa bằng 0, và bước tiền xử lý sẽ cần giải quyết chúng (ví dụ: thay thế bằng giá trị trung vị hoặc trung bình) trước khi thực hiện các phép tính thống kê sâu hơn.

1.5. Dữ liệu giá nhà Amsterdam (Nguồn: https://www.kaggle.com/datasets/thomasnibb/amsterdam)

Là bộ dữ liệu chứa thông tin chi tết về thị trường bất động sản tại Amsterdam, bao gôm các cột chính cần dươc quan tâm như:

_ Zip: Mã vùng hoặc khu vực (quận, phường).

_ Price: Giá bán của căn nhà (EUR).

_ Area: Diện tích (m^2).

_ Room: Số phòng.

_ Biến mới PriceperSqm: Giá trung bình trên mỗi mét vuông.

2. Thực hành khám phá dữ liệu

2.1. Thống kê mô tả

Ta có code mẫu dưới đây thực hiện các nhiệm vụ như sau được ghi chú trong đoạn code như bên dưới nhìn vào code bên dưới ta có thể giải thích sơ rằng file csv được gán vào biến covid_data bằng thư viện pandas dung hàm read_csv() sau đó được lấy các cột có tên [['iso_code','continent','location','date','total_cases','new_cases']] thuộc trong bảng dữ liệu đến với những dòng tiếp theo ta có thể thấy sử dụng hàm head() để có thể xuất ra màn hình 5 dòng đầu giúp cho ta có cái nhìn tổng quát về dữ liệu, có thể thấy bên cạnh đó hàm dtypes giúp in ra các kiểu dữ liệu ứng với từng cột trong bảng. Nếu dữ liệu trong cột thuộc kiểu dữ liệu số liên tục thì sẽ hiện output là (int64 hoặc float64) còn nếu ngược lại thì dữ liệu trong cột sẽ được xem là kiểu dữ liệu phân loại (object). Và dòng cuối cùng phía dưới shape giúp ta có thể nhận biết được tập dữ liệu của ta có bao nhiều dòng và bao nhiêu cột

```
# Khai báo thư viện
# Khai báo thư viện numpy và pandas dưới tên tạm là np và pd
import numpy as np
import pandas as pd
from scipy import stats # Thư viện stats từ gói scipy
# Nạp dữ liệu dạng file csv bằng read_csv
covid_data = pd.read_csv("covid-data.csv")
covid_data = covid_data[['iso_code','continent',
'location','date','total_cases','new_cases']]
# Có cái nhìn nhanh về bộ dữ liệu
covid_data.head(5)
covid_data.dtypes
covid_data.shape
```

Tiếp theo ta cùng đến với đoạn code mẫu sau miêu tả giới thiệu cho chúng ta biết một quá trì để có thể thực hiện thống kê mô tả để từ đó ta có thể có thêm được cái nhìn tổng quan về dữ liệu

Cụ thể như sau ta có thể sử dụng hàm

- np.mean(): Để có thể lấy giá trị trung bình của một đặc trưng (ở đây là “new_cases”)
- np.median(): Giusp thực hiện lấy các giá trị trung vị trong tập dữ liệu
- stats.mode(): Là một hàm có sẵn trong thư viện stats giúp ta có thểm tìm kiếm giá trị mode thuộc trong tập dữ liệu
- np.var(): Để có thể trả về phương sai trong tập dữ liệu
- np.std(): Trả về giá trị độ lệch chuẩn trong tập dữ liệu
- np.max(): Trả về giá trị lớn nhất
- np.min(): Trả về giá trị nhỏ nhất
- np.quantile(): Trả về các giá trị mà tại đó các giá trị thuộc mỗi khoảng chiếm nhiều nhất
- stats.iqr(): Giusp xác định giá trị IQR có trong bảng dữ liệu

Như vậy nhìn chung ta đã biết được khái niệm ban đầu về thống kê mô tả cũng như những thước đo thường được sử dụng nhiều khi thực hiện thống kê cũng như dựa vào bài toán mẫu nhiệm vụ 1 trong bài thực hành đưa ra. Tiếp đến ta cũng đến với một dạng dự liệu khác để có thể xem thêm về sự phân bố cũng như tuonrg quan của nó

```
# Lấy giá trị trung bình trong tập dữ liệu
data_mean = np.mean(covid_data["new_cases"])
# Lấy giá trị trung vị trong tập dữ liệu
data_median = np.median(covid_data["new_cases"])
# Lấy mốt của tập dữ liệu
data_mode = stats.mode(covid_data["new_cases"])
# Xác định phương sai trong tập dữ liệu
data_variance = np.var(covid_data["new_cases"])
# Xác định đô lệch chuẩn trong tập dữ liệu
data_sd = np.std(covid_data["new_cases"])
# So sánh giá trị lớn nhất và nhỏ nhất trong tập dữu liệu
data_max = np.max(covid_data["new_cases"])
data_min = np.min(covid_data["new_cases"])
# Tìm các khoảng xác định trong tập dữ liệu
data_percentile = np.percentile(covid_data["new_cases"],60)
# Tìm khoảng xác định phân bố trong tập dữ liệu
data_quartile = np.quantile(covid_data["new_cases"],0.75)
# Lấy giá trị IQR trong tập dữ liệu
data_IQR = stats.iqr(covid_data["new_cases"])
```

Đến với bảng dữ liệu "Marketing Campaign" ta lại có những vấn đề mứi hơn cụ thể chính là các bươc tiền xử lý dữ liệu (Nhiệm vụ 2 trong bài thực hành)

Sử dụng thư viện pandas đẻ có thể nạp dữ liệu vào môi trường cũng như đặt tên cho các cột có trong bảng dữ liệu. Tuong tự giống với nhiệm vụ 1 ta cũng có các bước như nạp dữ liệu và dặ tên cho các cột. Các bước này tuy không giúp ích gì nhưng nó lại là bước đầu để cho ta có thể làm quen cũng như xác định rõ các đặc trưng trong tập dữ liệu. Tuy nhiên sau những dòng code này ta có thể đến với những dòng code kế tiếp điều này có thể tạo sự khác biệt.

```
import pandas as pd

marketing_data = pd.read_csv("data/marketing_campaign.csv")

marketing_data = marketing_data[['ID','Year_Birth','Education',

                  'Marital_Status','Income','Kidhome',
'Teenhome', 'Dt_Customer','Recency','NumStorePurchases',
'NumWebVisitsMonth']]
```

```
# Xóa dữ liệu lặp ra khỏi bảng dữ liệu

wine_quality_red_data_without_duplicates =
wine_quality_red_data.drop_duplicates()

# Trình bày hình dạng tập dữ lieu sau khi đã loại bỏ

print("Shape of data after removing duplicates:",
wine_quality_red_data_without_duplicates.shape)
```

Đọc dữ liệu bằng thư viện Pandas sau đó sử dụng hàm drop_duplicated để có thể loại bỏ các giá tị bị trùng rồi dung shape để có thể xác nhận lại số hang và số cột của tập dữ liệu sau khi đã hoàn thành loại bỏ dữ lieuj trùng.

Tiếp đến chinh là bước thực hiện loại bỏ các giá rị bị thiếu sử dụng hàm isnull().sum() để có thể đếm trong bảng dữ liệu rằng có bao nhiêu dữ liệu là bị thiếu. Cuối cùng khi đã xác định được giá tri bị theieus rồi thì sử dụng phương thức dropna() để loại bỏ hoặc thay thế hoàn toàn giá trị bị thiếu đó

```
# Kiểm tra giá trị bị thiết và hàm sum

marketing_data.isnull().sum()

# Sử dụng dropna để loại bỏ giá trị bị thiếu

marketing_data_withoutna = marketing_data.dropna(how = 'any')

marketing_data_withoutna.shape
```

Sau khi đã tìm hiểu cũng như phân tích các chức năng của những đoạn code mẫu ta sẽ cùng đi vào phân tích và đnáh giá khi thực hiện những bài tập trong bài thực hành đính kèm cùng báo cáo náy

Cụ thể gồm 2 phần bài thực hành:

Bài thực hành 1: Thực hiện thống kế mô tả trên tập dữ liệu về phân loại chất lượng rượu đỏ. Các bước thực hiện như những đoạn code trên khi chạy thì có thể thấy rõ giá trị đầu ra như sau

Thực hiện việc khai báo thư viện sau đó dung read_csv() nhằm mục đích để nhập vào dữ liệu

```
import pandas as pd

import numpy as np

from scipy import stats


wine_quality_red_data = pd.read_csv("winequality-red.csv")

wine_quality_red_data.head() # In ra 5 dòng đầu của bảng dữ liệu
```

head() dung để xác định có cái nhìn tóm tắt về dạng dữ liệu. Gía trị đầu ra như sau:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Bảng bên trên miêu tả 5 dòng đầu của tập dữ liệu. Nhìn chung đây đều là dạng dữ liệu số luên tục và không hệ chưa chắc gì đã có tương. Tuy nhiên nếu muốn chắc hơn, ta có thể sử dụng hàm tail() để có thể xuất ra 5 dòng cuối của tập dữ liệu. Khi ta chạy lệnh shape ta sẽ có kết quả

	0
fixed acidity	*float64*
volatile acidity	*float64*
citric acid	*float64*
residual sugar	*float64*
chlorides	*float64*
free sulfur dioxide	*float64*
total sulfur dioxide	*float64*
density	*float64*
pH	*float64*
sulphates	*float64*
alcohol	*float64*
quality	*int64*

dtype: *object*

Điều này cho phép cũng như thông báo rằng tất cả những cột dữ liệu của ta đề là kiểu dữ liệu số và có thể biểu diễn được bằng biểu đồ phân bố

Tiếp đến khi ta sử dụng câu lệnh như những dòng code dưới đây ta sẽ nhận được một tin vui rằng Các cột dữ liệu thuộc trong tập dữu liệu của ta là không hề có giá trị bị thiếu nào và hoàn toàn sạch sẽ sẵn sàng để có thể trực quan hóa. Sau đó ta sẽ chạy sáng việc thống kê mô tả sử dụng stats mode và stats.iqr để tiifm kiếm mốt và giá tri iqr trong tập dữ liệu.

"Mode của chất lượng rượu: ModeResult(mode=np.int64(5), count=np.int64(577))

IQR của chất lượng rượu: 1.0"

	0
fixed acidity	*0*
volatile acidity	*0*
citric acid	*0*
residual sugar	*0*
chlorides	*0*
free sulfur dioxide	*0*
total sulfur dioxide	*0*
density	*0*
pH	*0*
sulphates	*0*
alcohol	*0*
quality	*0*

dtype: *int64*

```
wine_quality_red_data.duplicated().sum() # Kiểm tra dữ liệu bị trùng lặp

wine_quality_red_data = wine_quality_red_data.drop_duplicates() # Xóa dữ liệu trùng lặp

wine_quality_red_data.shape # In ra số dòng và số cột của bảng dữ liệu sau khi xóa dữ liệu trùng lặp

wine_quality_red_data.isnull().sum() # Kiểm tra dữ liệu bị thiếu
```

Cuối cùng ta có thể sử dụng describe() để có thể thực hiện thống kê mô tả trong tập dữ liệu. Gía trị trong phân này chính là thống kế đem qua cho người khác

	count	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
fixed acidity	*1359.0*	*8.310596*	*1.736990*	*4.60000*	*7.1000*	*7.9000*	*9.20000*	*15.90000*
volatile acidity	*1359.0*	*0.529478*	*0.183031*	*0.12000*	*0.3900*	*0.5200*	*0.64000*	*1.58000*
citric acid	*1359.0*	*0.272333*	*0.195537*	*0.00000*	*0.0900*	*0.2600*	*0.43000*	*1.00000*
residual sugar	*1359.0*	*2.523400*	*1.352314*	*0.90000*	*1.9000*	*2.2000*	*2.60000*	*15.50000*
chlorides	*1359.0*	*0.088124*	*0.049377*	*0.01200*	*0.0700*	*0.0790*	*0.09100*	*0.61100*
free sulfur dioxide	*1359.0*	*15.893304*	*10.447270*	*1.00000*	*7.0000*	*14.0000*	*21.00000*	*72.00000*
total sulfur dioxide	*1359.0*	*46.825975*	*33.408946*	*6.00000*	*22.0000*	*38.0000*	*63.00000*	*289.00000*
density	*1359.0*	*0.996709*	*0.001869*	*0.99007*	*0.9956*	*0.9967*	*0.99782*	*1.00369*
pH	*1359.0*	*3.309787*	*0.155036*	*2.74000*	*3.2100*	*3.3100*	*3.40000*	*4.01000*
sulphates	*1359.0*	*0.658705*	*0.170667*	*0.33000*	*0.5500*	*0.6200*	*0.73000*	*2.00000*
alcohol	*1359.0*	*10.432315*	*1.082065*	*8.40000*	*9.5000*	*10.2000*	*11.10000*	*14.90000*
quality	*1359.0*	*5.623252*	*0.823578*	*3.00000*	*5.0000*	*6.0000*	*6.00000*	*8.00000*

Đánh giá kết quả đầu ra có trong bảng sau:

- Lệch phải mạnh ($\bar{x} \gg$ Median): Các biến total_SO2 (46.8 vs 38.0), free_SO2 (15.9 vs 14.0), và residual_sugar (2.52 vs 2.20) cho thấy $\bar{x}$ cao hơn đáng kể so với Median. Điều này khẳng định sự hiện diện của các mẫu rượu vang có nồng độ hóa chất cực cao (ngoại lai), kéo Trung bình lên.
- Lệch phải nhẹ: Hầu hết các biến khác, bao gồm fixed acidity (8.31 vs 7.90) và chlorides (0.088 vs 0.079), cũng cho thấy xu hướng lệch phải nhẹ.
- Phân phối Gần Chuẩn: Biến pH (3.31 vs 3.31) có $\bar{x}$ và Median gần như bằng nhau, cho thấy phân phối rất đối xứng và có thể tuân theo phân phối chuẩn.
- Độ Biến động cao: Biến total_SO2 có Độ lệch chuẩn (33.4) và Phương sai (1116.1) cao nhất, chỉ ra sự khác biệt lớn nhất về nồng độ tổng sulfur dioxide giữa các mẫu rượu.

- Mật độ (density): Biến này có độ phân tán cực kỳ thấp ($\sigma \approx 0.002$), xác nhận rằng mật độ của rượu vang đỏ trong bộ dữ liệu là rất đồng nhất.
- IQR và Ngoại lai: IQR của total_SO2 (41.0) và free_SO2 (14.0) là tương đối lớn, nhưng Phạm vi (Range) của chúng còn lớn hơn nhiều (283.0 và 71.0). Khoảng cách lớn giữa IQR và Range cho thấy các biến này chứa nhiều ngoại lai nghiêm trọng ở hai đầu (theo lý thuyết 1.5×IQR).

Phân tích đa biến xác nhận rằng hầu hết các thuộc tính hóa học của rượu vang bị lệch phải và có ngoại lai. Đặc biệt, các biến total_SO2 và residual_sugar cần được xử lý triệt để (ví dụ: Winsorizing hoặc biến đổi logarit) để làm giảm độ lệch trước khi tiến hành mô hình hóa suy luận, trong khi biến pH đã sẵn sàng cho các mô hình dựa trên giả định phân phối chuẩn.

Bài tập 2: Thống kê mô tả với bệnh tiểu đường

Các bước thực hiện tương tự như với bài tập trước, tuy nhiên điểm đặc biệt của tập dữu liệu này chính là dược vệ sinh làm sạch hoàn toàn khiến cho ta cảm thấy an toàn hơn. Ta có đầu ra như sau

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42

	count	mean	std	min	25%	50%	75%	max
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

- Phát hiện Độ lệch (Skewness): Biến Insulin có sự chênh lệch lớn nhất giữa Trung bình (155.5) và Trung vị (125.0). Điều này là dấu hiệu rõ ràng của phân phối lệch phải cực kỳ mạnh và sự hiện diện của ngoại lai cực lớn kéo giá trị Trung bình lên cao.
- Thước đo Phù hợp: Đối với các biến lệch như Insulin, Trung vị (Q2) là thước đo xu hướng trung tâm mạnh mẽ (*robust*) và đáng tin cậy hơn so với Trung bình.
- Độ Biến động: Insulin có Độ lệch chuẩn ($\sigma \approx 118.8$) và Khoảng Tứ phân vị (IQR ≈ 128.75) cao nhất, cho thấy độ biến động (*volatility*) lớn nhất giữa các bệnh nhân.
- Ngoại lai: Sự chênh lệch giữa Range (Giá trị lớn nhất - Giá trị nhỏ nhất) và IQR cho hầu hết các biến (ví dụ, Insulin có Range khoảng 846 nhưng IQR chỉ 128) xác nhận ngoại lai đang hiện diện ở các đuôi phân phối.

2.2. Xử lý và trực quan hóa

```
import pandas as pd
import matplotlib.pyplot as plt
houseprices_data = pd.read_csv("data/HousingPricesData.csv")
houseprices_data = houseprices_data[['Zip', 'Price', 'Area', 'Room']]
# Create a PriceperSqm variable based on the Price and Area variables:
```

Đối với dữ liệu trong phần làm mẫu 1(giá nhà ở Amstedam) ta có thể quan sát được như sau:

Trước hết, thư viện Pandas được import để hỗ trợ đọc và xử lý dữ liệu dạng bảng. Sau đó, tập dữ liệu "HousingPricesData.csv" được đọc vào chương trình bằng lệnh pd.read_csv(), giúp ta có thể truy cập và thao tác với từng cột như trong Excel. Tập dữ liệu ban đầu có thể chứa nhiều thông tin khác nhau, vì vậy chỉ những cột quan trọng nhất là Zip, Price, Area và Room được giữ lại để phục vụ cho việc phân tích giá nhà.

Tiếp theo, chương trình tạo thêm một biến mới có tên PriceperSqm, được tính bằng cách chia giá nhà (Price) cho diện tích (Area). Biến này thể hiện giá trung bình trên mỗi mét vuông, là chỉ số quan trọng giúp đánh giá mức độ "đắt rẻ" thật sự của từng khu vực. Nhờ đó, ta có thể so sánh giá trị các căn nhà một cách công bằng hơn, bất kể diện tích lớn hay nhỏ.

Tiếp theo ta cùng đi đến với các dạng biểu đồ thể hiện sự phân tán của dữ liệu

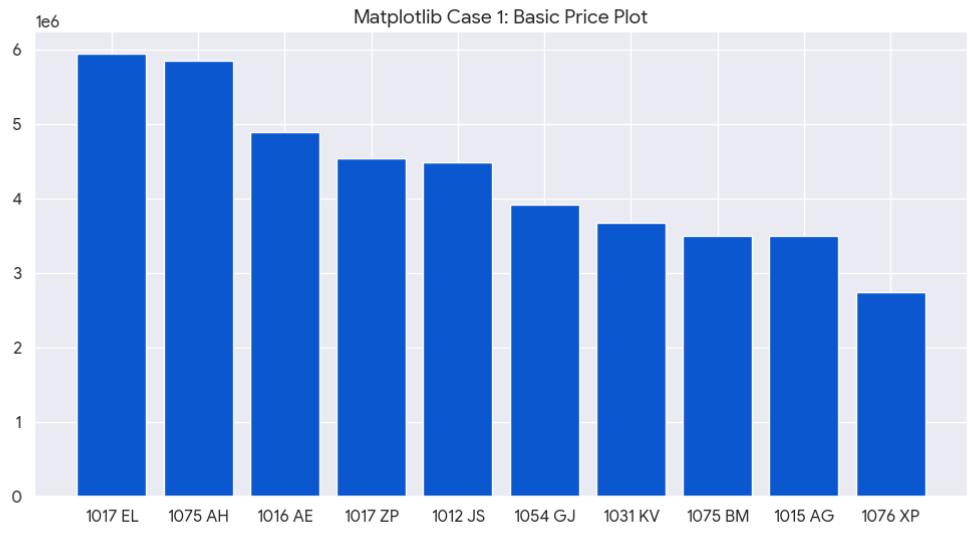
```
plt.figure(figsize= (12,6))

plt.bar(x,y)

plt.title('Top 10 Areas with the highest house prices', fontsize=15)

plt.xlabel('Zip code', fontsize = 12)

plt.xticks(fontsize=10)
```



_ Do dữ liệu đã được sắp xếp giảm dần theo 'Price', biểu đồ cho thấy một xu hướng giảm dần rõ ràng từ trái sang phải, bắt đầu từ Mã Zip 1017 EL (cột cao nhất) đến 1076 XP (cột thấp nhất).

_ Chênh lệch giữa khu đầu và cuối top 10 khoảng 2–3 triệu EUR.

_ Thể hiện phân hóa giá nhà rõ rệt theo khu vực.

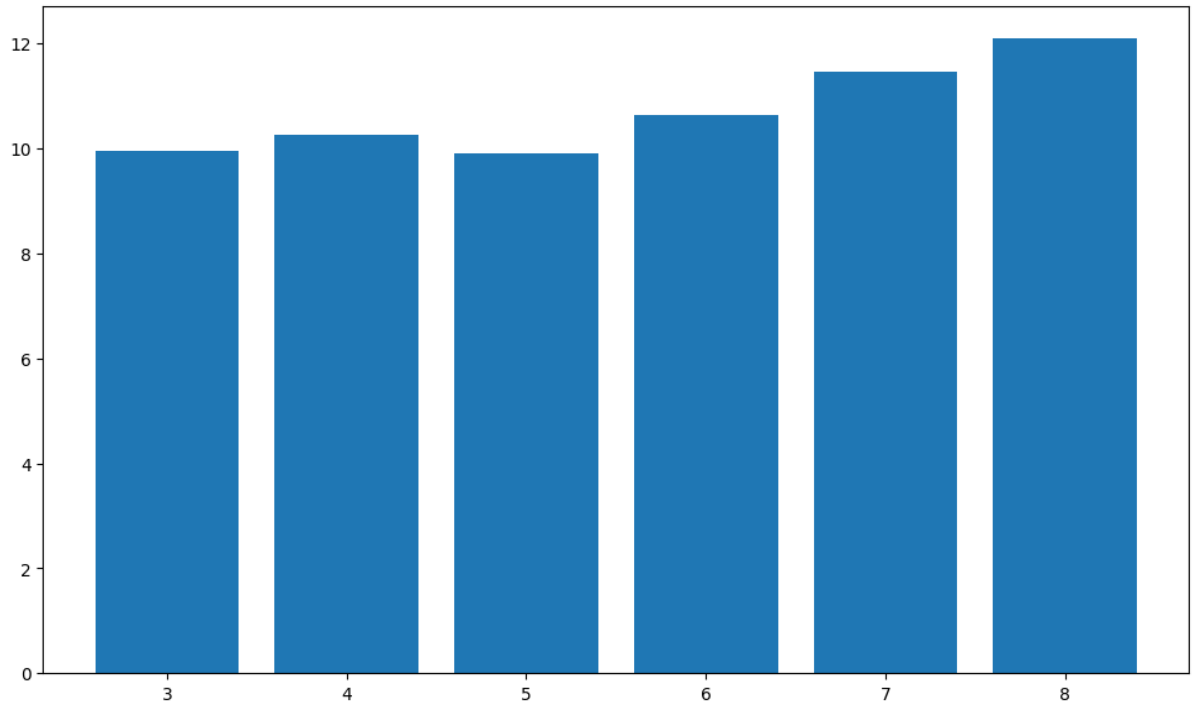
Kết luận: Sự khác biệt giữa hai biểu đồ chứng minh rằng tổng giá nhà và giá trên mỗi mét vuông là hai yếu tố độc lập và cần được phân tích cùng nhau để đưa ra kết luận chính xác về giá trị bất động sản

_ Hạn chế: Biểu đồ này thiếu các thành phần trực quan quan trọng như tiêu đề, nhãn trục, và đơn vị đo lường, khiến người đọc khó hiểu chính xác nội dung đang được trình bày (ví dụ: đơn vị của giá là gì, trục hoành đại diện cho gì).

Đến với bài tiếp theo ta cùng phân tisch chất lượng rượu đỏ có trong bài thực hành

```
# Chuẩn bị dữ liệu để trực quan hóa
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
wine_quality_red_data = pd.read_csv("winequality-red.csv")
```

```
wine_quality_avg =
dfsort.groupby('quality').mean().reset_index()
x = wine_quality_avg['quality']
y_alcohol = wine_quality_avg['alcohol']
y_acidity = wine_quality_avg['volatile acidity']
plt.figure(figsize=(12, 7))
```

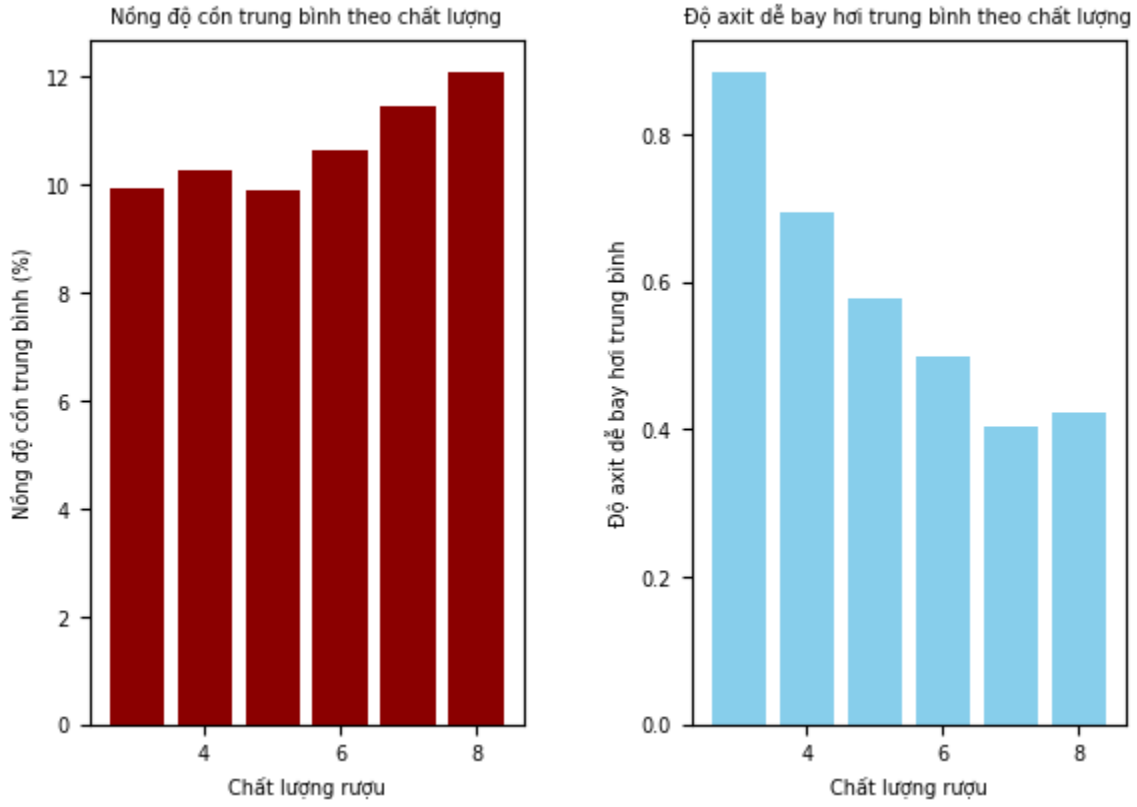


_ Đặc điểm Matplotlib: Ở mức cơ bản, biểu đồ chỉ là một tập hợp các cột đơn giản.

_ Ưu điểm: Cực kỳ linh hoạt, cho phép người dùng vẽ nhanh mối quan hệ.

_ Hạn chế: Thiếu tính thẩm mỹ và thiếu thông tin thống kê (như thanh lỗi/error bar) một cách mặc định. Người dùng cần phải tính toán các giá trị trung bình và độ lệch chuẩn trước khi vẽ.

_ Biểu đồ thể hiện mối quan hệ đồng biến (tăng dần): Nồng độ Cồn trung bình tăng lên khi Chất lượng rượu tăng từ 3 lên 8.



Tạo hai biểu đồ con (subplot) để so sánh Nồng độ Cồn (tích cực) và Axit dễ bay hơi (tiêu cực) đối với Chất lượng.

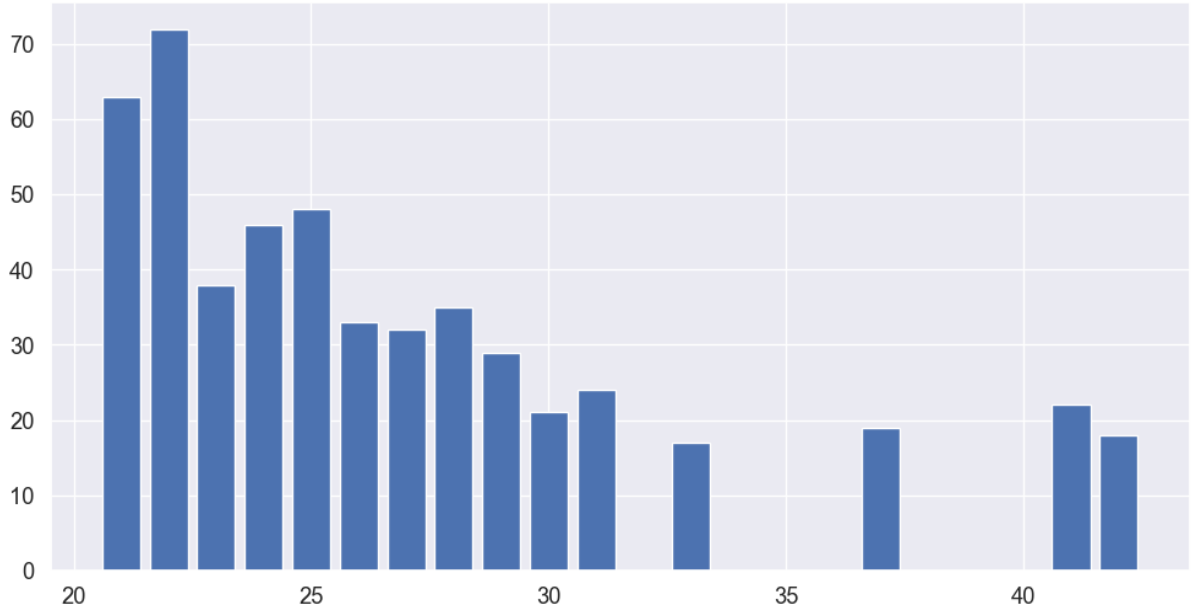
_ Việc tạo các subplot (plt.subplot hoặc plt.subplots) yêu cầu người dùng quản lý thủ công từng trục (ax[0], ax[1]). Cần phải đặt tiêu đề, nhãn trục, và kích thước font chữ cho từng biểu đồ con một cách riêng biệt.

_ Biểu đồ 1 (Nồng độ Cồn): Củng cố xu hướng tăng dần.

_ Biểu đồ 2 (Axit dễ bay hơi): Cho thấy xu hướng nghịch biến (giảm dần). Axit dễ bay hơi cao nhất ở rượu điểm thấp (3, 4) và thấp nhất ở rượu điểm cao (7, 8).

Kết luận: Matplotlib cho phép tùy chỉnh bố cục mạnh mẽ, nhưng yêu cầu nhiều dòng lệnh hơn để đạt được một biểu đồ đa góc nhìn có chất lượng trình bày cao.

Quá trình này nhằm mục đích làm gọn và tập trung trực quan hóa vào các nhóm tuổi có ý nghĩa thống kê nhất trong tập dữ liệu tiểu đường. Cụ thể, dữ liệu về Số lượng bệnh nhân và Mức Glucose trung bình được tính toán cho tất cả các độ tuổi, sau đó được lọc để chỉ giữ lại 15 nhóm tuổi có số lượng bệnh nhân đông nhất. Việc này tạo ra hai tập dữ liệu con (top_15_ages và top_15_glucose) đảm bảo rằng các biểu đồ trực quan (như Case 3) sẽ dễ đọc hơn và hạn chế nhiễu từ các độ tuổi chỉ có rất ít mẫu, đồng thời làm nổi bật mối quan hệ giữa tần suất dân số và chỉ số sức khỏe (Glucose) ở các nhóm tuổi đại diện.



CHƯƠNG IV: TÓM TẮT THỰC HÀNH

Khám phá dữ liệu (Exploratory Data Analysis - EDA) là một bước quan trọng trong phân tích dữ liệu và khai thác dữ liệu, nhưng quá trình này không tránh khỏi những khó khăn. Một trong những thách thức lớn nhất là chất lượng dữ liệu không đảm bảo, bao gồm giá trị thiếu, giá trị ngoại lai hoặc dữ liệu không nhất quán, đòi hỏi kỹ năng tiền xử lý phức tạp và tốn thời gian. Bên cạnh đó, việc xử lý khối lượng dữ liệu lớn có thể gây khó khăn trong việc xác định các mẫu hoặc xu hướng có ý nghĩa, đặc biệt khi sử dụng các công cụ không được tối ưu hóa cho dữ liệu lớn. Chương đã trình bày một số kỹ thuật cơ bản khi sử dụng Python và các công cụ phát triển bằng Python giúp thực hiện việc khám phá dữ liệu được hiệu quả hơn.