**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**

****

**BÀI TẬP LỚN**

**HỌC MÁY CƠ BẢN**

**DỰ ĐOÁN BỆNH ALZHEIMER**

NGÀNH: KHOA HỌC MÁY TÍNH

SINH VIÊN: **1. VŨ QUANG PHÚC**

LỚP: **124221**

NGƯỜI HƯỚNG DẪN: **TS. HOÀNG QUỐC VIỆT**

**HƯNG YÊN – 2025**

**NHẬN XÉT**

**Nhận xét của giáo viên hướng dẫn**

**GIÁO VIÊN HƯỚNG DẪN**

**Hoàng Quốc Việt**

**~~LỜI CAM ĐOAN~~**

~~Em xin cam đoan bài tập lớn môn lập trình Python nâng cao có tên là “Phân tích dữ liệu về bộ dữ liệu thời trang Anh-Mỹ” là sản phẩm của bản thân em.~~

~~Những phần sử dụng tài liệu tham khảo trong bài tập lớn đã được nêu rõ trong phần tài liệu tham khảo. Các kết quả trình bày trong bài tập lớn hoàn toàn là kết quả do bản thân em thực hiện.~~

~~Nếu vi phạm lời cam đoan này, em xin chịu hoàn toàn trách nhiệm trước khoa và nhà trường.~~

*~~Hưng Yên, ngày … tháng 06 năm 2024~~*

~~Sinh viên~~

~~Nguyễn Văn X~~

**~~LỜI CẢM ƠN~~**

~~Để có thể hoàn thành bài tập lớn này, lời đầu tiên em xin phép gửi lời cảm ơn tới bộ môn Khoa học máy tính, Khoa Công nghệ thông tin – Trường Đại học Sư phạm Kỹ thuật Hưng Yên đã tạo điều kiện thuận lợi cho em thực hiện bài tập lớn môn học này.~~

~~Đặc biệt em xin chân thành cảm ơn thầy Hoàng Quốc Việt đã rất tận tình hướng dẫn, chỉ bảo em trong suốt thời gian thực hiện bài tập lớn vừa qua.~~

~~Em cũng xin chân thành cảm ơn tất cả các Thầy, các Cô trong Trường đã tận tình giảng dạy, trang bị cho em những kiến thức cần thiết, quý báu để giúp em thực hiện được bài tập lớn này.~~

~~Mặc dù em đã có cố gắng, nhưng với trình độ còn hạn chế, trong quá trình thực hiện đề tài không tránh khỏi những thiếu sót. Em hy vọng sẽ nhận được những ý kiến nhận xét, góp ý của các Thầy cô về những kết quả triển khai trong bài tập lớn.~~

~~Em xin trân trọng cảm ơn!~~

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN 2](#_Toc167950228)

[1.1 Bài toán 2](#_Toc167950229)

[1.2 Trình bày dữ liệu bài toán 2](#_Toc167950230)

[1.3 Tiền xử lý dữ liệu 4](#_Toc167950231)

[1.4 Thống kê dữ liệu 4](#_Toc167950232)

[1.5 Trực quan hoá dữ liệu 5](#_Toc167950233)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 6](#_Toc167950234)

[2.1 Pandas 6](#_Toc167950235)

[2.2 Matplotlib 6](#_Toc167950236)

[CHƯƠNG 3. GIẢI PHÁP 7](#_Toc167950237)

[3.1. Mã nguồn tiền xử lý dữ liệu 7](#_Toc167950238)

[3.2. Mã nguồn chức năng Thống kê dữ liệu 7](#_Toc167950239)

[3.3. Mã nguồn chức năng Trực quan hóa dữ liệu 13](#_Toc167950240)

[TÀI LIỆU THAM KHẢO 19](#_Toc167950241)

# GIỚI THIỆU BÀI TOÁN

## Bài toán

Hiện nay, Thị trường thời trang Anh-Mỹ là một trong những thị trường thời trang lớn nhất và có sự đa dạng về phong cách và xu hướng. Bằng cách phân tích dữ liệu thời trang Anh-Mỹ, chúng ta có thể thu thập thông tin về một phạm vi rộng hơn và cung cấp cái nhìn tổng quan về xu hướng mua sắm và ưa chuộng của người tiêu dùng trong lĩnh vực thời trang, bao gồm nhiều loại sản phẩm từ quần áo, giày dép, phụ kiện đến trang phục và phụ kiện thời trang cho nam, nữ và trẻ em. Dữ liệu thời trang Anh-Mỹ thường được thu thập từ các nguồn uy tín và có tính tin cậy cao. Ngoài ra, dữ liệu thời trang Anh-Mỹ thường được cập nhật thường xuyên, giúp đảm bảo tính sẵn có của dữ liệu và đáng tin cậy trong quá trình phân tích

Ngoài ra thị trường thời trang Anh-Mỹ có sự ảnh hưởng toàn cầu đến các thị trường khác trên thế giới. Việc phân tích dữ liệu thời trang Anh-Mỹ giúp hiểu rõ hơn về ảnh hưởng của xu hướng và thị trường thời trang Anh-Mỹ đến các thị trường quốc tế khác, từ đó có thể áp dụng những kết quả phân tích để tối ưu hóa hoạt động kinh doanh và phát triển thị trường thời trang.

Chúng ta đã có dữ liệu và bây giờ chúng ta sẽ phải khai thác dữ liệu đó giúp người dùng biết rõ hơn về thông tin của các sản phẩm thời trang của Anh Mỹ ,…Giúp người dùng dễ dàng lựa chọn được trang phục và phụ kiện thời trang phù hợp nhất cho bản thân.

## Trình bày dữ liệu bài toán

Dữ liệu được lấy từ trang web này:

[Fashion Dataset UK-US | Kaggle](https://www.kaggle.com/datasets/a23bisola/fashion-dataset-uk-us)

A picture containing text, screenshot, number, font

Description automatically generated

A screenshot of a computer code

Description automatically generated with medium confidence

-Dữ liệu bài toán gồm các feature sau:

+ Tên sản phẩm

+ Giá sản phẩm

+ Hãng sản phẩm

+ Loại sản phẩm

+ Mô tả

+ Tổng đánh giá

+ Kiểu phong cách

+ Tổng size

+ Size có sẵn

+ Màu sắc

+ Lịch sử mua hàng

+ Tuổi

+ Mùa

+ Đánh giá khách hàng

+ Phản hồi

- Dữ liệu bài toán là 1 file csv gồm 599999 rows × 15 columns

+ Có 15 feature và mỗi feature có 599999 dữ liệu đầu vào

-Sau khi mô tả dữ liệu ta có:

A screenshot of a computer

Description automatically generated with low confidence

+ Giá sản phẩm trong dữ liệu này Max là 99,9999 và Min là 10

+Tổng đánh giá sản phẩm Max là 499 và Min là 0

+ Độ tuổi lớn nhất mua sản phẩm là 64 và độ tuổi nhỏ nhất là 18

## Tiền xử lý dữ liệu

## Thống kê dữ liệu

1. Thống kê loại sản phẩm bán chạy nhất
2. Thống kê số lượng sản phẩm Theo nhãn hiệu (articleType)
3. Thống kê số lượng các sản phẩm mỗi mùa
4. Thống kê độ tuổi mua nhiều sản phẩm nhất
5. Thống kê số lượng sản phẩm theo từng phong cách
6. Thống kê top 10 hãng được bán nhiều nhất
7. Thống kê số lượng sản phẩm theo đánh giá khách hàng
8. Thống kê số lượng sản phẩm theo trạng thái phản hồi
9. Thống kê 10 sản phẩm có giá cao nhất
10. Thống kê số lượng sản phẩm theo nhóm tuổi (trẻ từ 1 đến 30) và (già từ 31 đến 60)

## Trực quan hoá dữ liệu

1. Biểu đồ thể hiện tỉ lệ số lượng sản phẩm mỗi loại
2. Hiển thị top 10 sản phẩm có giá bán cao nhất
3. Biểu đồ hiển thị top 10 hãng được bán nhiều nhất
4. Biểu đồ thể hiện số lượng sản phẩm theo nhóm tuổi (trẻ từ 1 đến 30) và (già từ 31 đến 60)
5. Biểu đồ thể hiện số lượng các phản hồi của khách hàng

# CƠ SỞ LÝ THUYẾT

## Pandas

## Matplotlib

# CHƯƠNG 3. GIẢI PHÁP

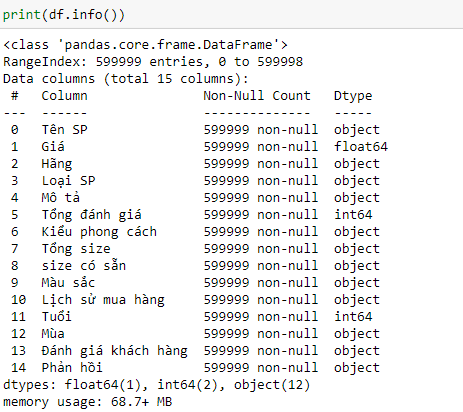
## Mã nguồn tiền xử lý dữ liệu







Đầu tiên em sử dụng hàm df.info() để kiểm tra các feature nào bị thiếu và có kiểu dữ liệu thế nào.Em hiển thị lại hàm df và dùng lại hàm info để kiểm tra xem dữ liệu. Và hình ảnh dưới là kết quả sau khi đã điền các dữ liệu:



## Mã nguồn chức năng Thống kê dữ liệu

a, Thống kê loại sản phẩm bán chạy nhất

A picture containing text, font, screenshot, line

Description automatically generated

- Đầu tiên em sử dụng hàm groupby(),hàm này cho phép chúng ta nhóm các dữ liệu giống nhau theo nhãn feature mà ta chọn trong dataframe ,vì ở đây là thống kê số lượng sản phẩm nên em sẽ nhóm các dữ liệu liên quan đến Tên loại sản phẩm vào với nhau;em sẽ chọn dữ liệu “Loại SP” và “Tên SP” từ dataframe để nhóm với dữ liệu “Loại SP” giống nhau .

-Sử dụng hàm assign() phương thức gán các cột mới cho DataFrame, trả về một đối tượng mới (bản sao) với các cột mới được thêm vào các cột ban đầu. Các cột hiện có được gán lại sẽ bị ghi đè. Em sẽ gán cột mới tên “soluong” cho cột “Tên SP” trong dataframe để về sau chúng ta sẽ cột dữ liệu mới ứng dữ liệu như với các dữ liệu giống của “Loại SP”

-Sau đó em sử dụng hàm count() để đếm xem số lượng phim mà có cùng dữ liệu “Loai SP”giống nhau.

-Sau khi đã có được dataframe dữ liệu tổng hợp lại rồi em sử dụng hàm sort\_values để sắp xếp dữ liệu lại theo thứ tự giảm dần và ta sẽ có dược kết quả như sau:

A picture containing text, screenshot, font, number

Description automatically generated

b, Thống kê số lượng sản phẩm Theo nhãn hiệu (articleType)

A picture containing text, screenshot, font, line

Description automatically generated

- Đầu tiên em sử dụng hàm groupby(), hàm này cho phép chúng ta nhóm các dữ liệu giống nhau theo nhãn feature mà ta chọn trong dataframe ,vì ở đây là thống kê số lượng sản phẩm nên em sẽ nhóm các dữ liệu liên quan đến nhãn hiệu sản phẩm vào với nhau;em sẽ chọn dữ liệu “Nhãn hiệu” và “Tên SP” từ dataframe để nhóm với dữ liệu “Nhãn hiệu” giống nhau .

- Sử dụng hàm assign() phương thức gán các cột mới cho DataFrame, trả về một đối tượng mới (bản sao) với các cột mới được thêm vào các cột ban đầu. Các cột hiện có được gán lại sẽ bị ghi đè. Em sẽ gán cột mới tên “soluong” cho cột “Tên SP” trong dataframe để về sau chúng ta sẽ cột dữ liệu mới ứng dữ liệu như với các dữ liệu giống của “Nhãn hiệu”

- Sau đó em sử dụng hàm count() để đếm xem số lượng phim mà có cùng dữ liệu “Nhãn hiệu”giống nhau.

- Sau khi đã có được dataframe dữ liệu tổng hợp lại rồi em sử dụng hàm sort\_values để sắp xếp dữ liệu lại theo thứ tự giảm dần và ta sẽ có dược kết quả như sau:

A picture containing text, screenshot, font, number

Description automatically generated

c, Thống kê số lượng các sản phẩm mỗi mùa

A picture containing text, screenshot, font, line

Description automatically generated

- Sử dụng groupby, hàm này cho phép chúng ta nhóm các dữ liệu giống nhau theo cột mà ta chọn trong Dataframe, vì đây là thống kê số lượng sản phẩm liên quan đến mùa nên e chọn 2 cột mùa và cột tên sản phẩm để nhóm với dữ liệu mùa giông nhau.

- Sử dụng hàm assign phương thức gán cột mới cho Dataframe. Trả về một đối tượng mới với các cột mới được thêm vào các cột ban đầu. Các cột hiện có được gán sẽ bị ghi đè.

- Sau khi đã có Datafame dữ liệu tổng hợp em sử dụng hàm sort\_values để sắp xếp dữ liệu theo thứ tự giảm dần(ascending=False), ignore\_index=True để giúp cho các thú tự của index không bị thay đổi vị trí sau khi sắp xếp và được kết quả sau:

A picture containing text, screenshot, font, number

Description automatically generated

d, Thống kê độ tuổi mua nhiều sản phẩm nhất

A picture containing text, screenshot, font, line

Description automatically generated

- Sử dụng groupby , hàm này cho chúng ta gom nhóm các thuộc tính giống nhau theo cột mà ta chọn trong dataframe. Vì đây là thống kê sản phẩm theo độ tuổi người mua nên ta chọn 2 cột là “Tuổi” và Số lượng để gom nhóm với dữ liệu tuổi giống nhau.

- Sử dụng hàm assign phương thức gán cột mới cho dataframe. Trả về một đối tượng gồm các cột mới cùng với các cột cũ, các cột cũ hiện có bị gán sẽ được ghi đè.

- Vì đây là thống kê ra độ tuổi mua nhiều sản phẩm nhất nên cho cột soluong bằng soluong lớn nhất ở trong dataframe bằng cách sử dùng hàm max .Từ có ta có được kết quả sau:

A picture containing text, font, screenshot, white

Description automatically generated

e, Thống kê số lượng sản phẩm theo từng phong cách



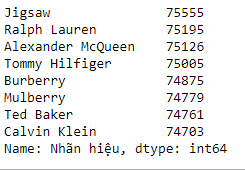
A picture containing text, screenshot, font, number

Description automatically generated

f, Thống kê top 10 hãng được bán nhiều nhất



- Sử dụng groupby để gom nhóm các thuộc giống nhau theo cột . Vì đây là thống kê sản phẩm theo màu sắc nên e chọn cột “Sản phẩm” và “nhãn hiệu”.để gom nhóm với dữ liệu nhãn hiệu



g, Thống kê số lượng sản phẩm theo đánh giá khách hàng



- Sử dụng groupby để gom nhóm các thuộc giống nhau theo cột . Vì đây là thống kê sản phẩm theo màu sắc nên e chọn cột “Đánh giá khách hàng” và “Tên sản phẩm”.để gom nhóm với dữ liệu màu sắc giống nhau.

A picture containing text, font, screenshot, algebra

Description automatically generated

h, Thống kê số lượng sản phẩm theo trạng thái phản hồi

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

i, Thống kê top 10 sản phẩm có giá bán cao nhất

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

j , Thống kê thể hiện số lượng sản phẩm theo nhóm tuổi (trẻ từ 1 đến 30) và (già từ 31 đến 60)

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, font, screenshot, white

Description automatically generated

## Mã nguồn chức năng Trực quan hóa dữ liệu



Khai báo các thư viện cần thiết để trực quan hóa dữ liệu

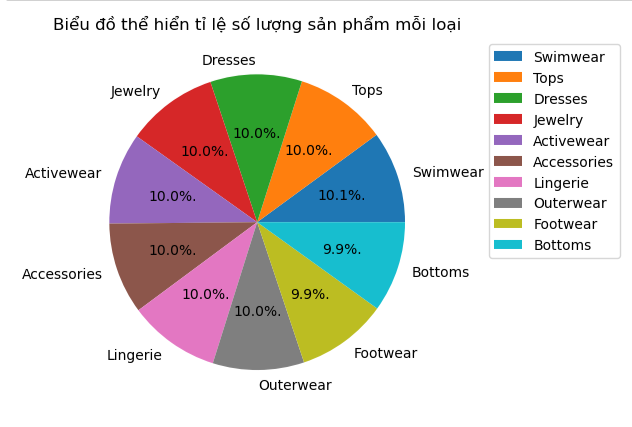
a, Biểu đồ thể hiện tỉ lệ số lượng sản phẩm mỗi loại

A picture containing text, screenshot, font

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated



Đầu tiên chọn dataframe mà em vừa thống kê tỉ lệ số lượng sản phẩm mỗi loại ở trên. Sau đó chúng ta sử dụng pl.pie() để vẽ biểu đồ Pie. Truyền vào số lượng từ cột values của product\_count\_by\_type và nhãn từ cột index. autopct='%1.1f%%' được sử dụng để hiển thị tỷ lệ phần trăm trên biểu đồ.

b, Hiển thị top 10 sản phẩm có giá bán cao nhất

A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, screenshot, plot, number

Description automatically generated

- pl.bar() được sử dụng để vẽ biểu đồ cột. Chúng ta truyền vào danh sách tên sản -phẩm từ cột 'Tên SP' và danh sách giá bán từ cột 'Giá' của DataFrame top\_10\_expensive\_products.

pl.xlabel() được sử dụng để đặt nhãn cho trục x, trong trường hợp này là 'Tên sản phẩm'.

pl.ylabel() được sử dụng để đặt nhãn cho trục y, trong trường hợp này là 'Giá bán'.

pl.title() được sử dụng để đặt tiêu đề cho biểu đồ, trong trường hợp này là 'Top 10 sản phẩm có giá bán cao nhất'.

Cuối cùng, plt.show() được sử dụng để hiển thị biểu đồ cột trên màn hình.

c, Biểu đồ hiển thị top 10 hãng được bán nhiều nhất

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, screenshot, font, number

Description automatically generated

- pl.barh() được sử dụng để vẽ biểu đồ cột ngang. Chúng ta truyền vào danh sách tên hãng từ cột 'Hãng' và danh sách số lượng sản phẩm từ cột 'Số lượng sản phẩm' của DataFrame product\_count\_by\_brand.

- pl.xlabel() được sử dụng để đặt nhãn cho trục x, trong trường hợp này là 'Số lượng sản phẩm'.

- pl.ylabel() được sử dụng để đặt nhãn cho trục y, trong trường hợp này là 'Hãng'.

- pl.title() được sử dụng để đặt tiêu đề cho biểu đồ, trong trường hợp này là 'Top 10 hãng bán được nhiều nhất'.

- Cuối cùng, pl.show() được sử dụng để hiển thị biểu đồ cột ngang trên màn hình.

d, Biểu đồ thể hiện số lượng sản phẩm theo nhóm tuổi (trẻ từ 1 đến 30) và (già từ 31 đến 60)

A picture containing text, font, screenshot, white

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

- pl.bar() được sử dụng để vẽ biểu đồ cột. Chúng ta truyền vào danh sách nhóm tuổi từ cột 'Nhóm tuổi' và danh sách số lượng sản phẩm từ cột 'Số lượng' của DataFrame product\_count\_by\_age.

- pl.xlabel() được sử dụng để đặt nhãn cho trục x, trong trường hợp này là 'Nhóm tuổi'.

- pl.ylabel() được sử dụng để đặt nhãn cho trục y, trong trường hợp này là 'Số lượng'.

- pl.title() được sử dụng để đặt tiêu đề cho biểu đồ, trong trường hợp này là 'Thống kê số lượng sản phẩm theo nhóm tuổi'.

Cuối cùng, pl.show() được sử dụng để hiển thị biểu đồ cột trên màn hình.

e, Biểu đồ thể hiện số lượng các phản hồi của khách hàng

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, screenshot, diagram, font

Description automatically generated

- pl.pie() được sử dụng để vẽ biểu đồ Pie. Chúng ta truyền vào danh sách các giá trị từ cột 'Phản hồi' của DataFrame feedback\_count và danh sách các nhãn tương ứng từ cột 'index' của feedback\_count. Định dạng hiển thị tỉ lệ phần trăm sử dụng %1.2f%% để làm tròn đến 2 chữ số thập phân.

- pl.title() được sử dụng để đặt tiêu đề cho biểu đồ, trong trường hợp này là 'Thống kê phản hồi của khách hàng'.

- pl.legend() được sử dụng để đặt chú thích cho biểu đồ, ở đây là đặt chú thích vị trí bên phải của biểu đồ.

Cuối cùng, pl.show() được sử dụng để hiển thị biểu đồ Pie trên màn hình.

# TÀI LIỆU THAM KHẢO

1. Dữ liệu

[Fashion Dataset UK-US | Kaggle](https://www.kaggle.com/datasets/a23bisola/fashion-dataset-uk-us)

1. Tài liệu

* Các tài liệu slide của thầy Nguyễn Văn Quyết
* Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython 2nd Edition