1. Thu thập dữ liệu

→ Dữ liệu thô (chưa qua xử lý)

1. Tiền xử lý dữ liệu

→ Dữ liệu sạch hơn, chuẩn hơn,...

Công cụ: numpy, pandas, matplotlib, seaborn

**Numpy, pandas:**

Link slide: <https://www.canva.com/design/DAGZ3FGZ9-M/4SNYPJCWnTJL7JH3J6I8Lw/edit>

Link demo: <https://colab.research.google.com/drive/1nZD1sDfQjOwrkyjb622-UTBXLWgqScME?usp=sharing>

**Matplotlib, seaborn:**

Link demo: <https://colab.research.google.com/drive/1EvYxTXCmrsb--Grb9t0ypJ8FF-SPxhE3>

**Xử lý ngôn ngữ tự nhiên: NLTK, TextBlob**

**Dự báo chuỗi thời gian: ARIMA, Facebook Prophet, LSTM.**

**Pandas:**

dataframe: df

df.info() → hiển thị thông tin cơ bản: tên cột, bao nhiêu không null, kiểu dữ liệu

df.head() → hiển thị các dòng dữ liệu

df.describe() → mô tả những dữ liệu kiểu số

df.mean(), df.sum() → tính toán

df.isnull(), df.isna() → kiểm tra xem có trống dữ liệu không? → df.fillna(value)

df[col].value\_counts() → kiểm tra xem cột có bao nhiêu giá trị

df[col].unique() → các giá trị duy nhất của cột

Matplotlib, seaborn

countplot()

pieplot()

lineplot()

…

**→ Xác định features (các đặc trưng để đưa vào mô hình), output (kết quả mong muốn)**

Yêu cầu: EDA dữ liệu - Phân tích dữ liệu khám phá

→ Khi code chạy xong, đưa ra nhận xét

Dataset: <https://drive.google.com/drive/folders/1bTB4W-Sp-ZxgypjIi5seXl7kzmviV4CU?usp=drive_link>

Link code: <https://colab.research.google.com/drive/1BhdGlG9QMVGSEM6e6Co1-AM1b6kkV-y1>

Khang 150325

Dataset: <https://drive.google.com/drive/folders/1bTB4W-Sp-ZxgypjIi5seXl7kzmviV4CU?usp=drive_link>

Yêu cầu:

### **1. Tổng quan dữ liệu**

* Có bao nhiêu dòng và bao nhiêu cột trong bộ dữ liệu?
* Có bao nhiêu giá trị bị thiếu trong mỗi cột?
* Các kiểu dữ liệu của từng cột là gì?

### **2. Phân phối dữ liệu**

* Loại trang phục nào phổ biến nhất trong tập dữ liệu?
* Phong cách thời trang nào đang có độ phổ biến cao nhất?
* Xu hướng thời trang nào có điểm phổ biến (**popularity\_score**) trung bình cao nhất?
* Màu sắc nào xuất hiện nhiều nhất trong các xu hướng thời trang?
* Chất liệu vải nào được sử dụng phổ biến nhất?
* Phân phối giá trị trong cột **popularity\_score** có dạng như thế nào (bị lệch về một phía hay phân phối đều)?

### **3. Mối quan hệ giữa các yếu tố**

* Xu hướng thời trang có khác nhau theo từng mùa không? (Ví dụ: Mùa hè thường có phong cách nào phổ biến nhất?)
* Phong cách thời trang nào thường xuất hiện với mức giá **cao cấp (high-end)** nhất?
* Các thương hiệu có ảnh hưởng (**brand\_influence**) tác động như thế nào đến độ phổ biến?
* Những nền tảng mạng xã hội nào đang ảnh hưởng mạnh đến xu hướng thời trang?
* Có sự khác biệt nào giữa các màu sắc được ưa chuộng theo từng phong cách thời trang không?

### **4. Phân tích nâng cao**

* Có xu hướng nào nổi bật trên nhiều nền tảng mạng xã hội cùng lúc không?
* Những chất liệu vải nào thường gắn liền với các phong cách cụ thể?
* Có mối tương quan nào giữa **popularity\_score** và **brand\_influence** không?
* Xu hướng thời trang có bị ảnh hưởng bởi **season (mùa)** không?