**Mô tả tập dữ liệu**

* **Business operations survey 2022: information and communications technology**: file CSV ghi lại dữ liệu kinh doanh hoạt động khảo sát doanh nghiệp trong Công nghệ thông tin và truyền thông trong năm 2022
  + Tập dữ liệu gồm có các trường:
    - Description (Mô tả)
    - Industry (Doanh nghiệp)
    - Level (mức độ)
    - Size (kích thước nhân sự)
    - line\_code
    - value
* **Environmental tax account 1999-2021**: file CSV ghi lại dữ liệu về thuế môi trường trong khoảng năm 1999-2021
  + Tập dữ liệu gồm có các trường:
    - Year (Năm)
    - Industry\_Code (Mã doanh nghiệp)
    - Industry\_Desc (Mô tả doanh nghiệp)
    - Nation\_Account\_Var (Tài khoản quốc gia)
    - Units (Đơn vị tiền tệ)
    - Magnitude (Độ lớn tiền tệ)
    - Source (Nguồn tài khoản)
    - Data\_value (Trị giá dữ liệu)
    - Flag (Đánh dấu)

**Mô tả và diễn giải quá trình tiền xử lý dữ liệu**

* Tiền xử lý dữ liệu với file **Business operations survey 2022: information and communications technology:** Giả sử ta có yêu cầu chỉ xử lý dữ liệu có value trên 100 và ta tiến hành tiền xử lý dữ liệu bằng phương pháp **Lọc dữ liệu** để loại bỏ các bộ có value bằng hoặc dưới 100 và chỉ lấy những bộ có value trên 100.

A screenshot of a computer program

Description automatically generated

Dữ liệu khi chưa được xử lý tiền dữ liệu

Sau khi xử lý dữ liệu ta có còn lại các bộ dữ liệu như sau:

A screenshot of a computer screen

Description automatically generated

Dữ liệu sau tiền xử lý dữ liệu

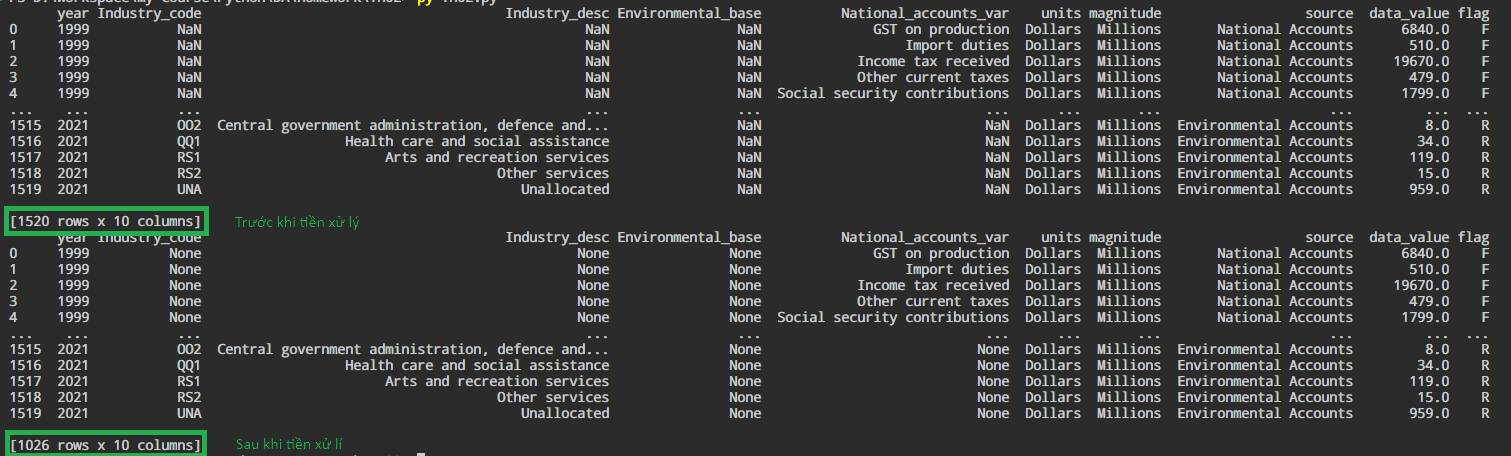
Sau khi tiền xử lý dữ liệu số bộ dữ liệu là 8699 giảm còn 4437

* Tiền xử lý dữ liệu với file **Environmental tax account 1999-2021**: Do dataset của bộ dữ liệu này chứa rất nhiều bộ có chứa giá trị rỗng và các bộ thường không thống nhất nên ta sử dụng phương pháp **Lọc dữ liệu** và **Xử lí dữ liệu thiếu** đối với dataset trên.
  + Lọc dữ liệu: loại bỏ các bộ có giá trị rỗng của cột data\_value
  + Xử lí dữ liệu thiếu: thay thế các dữ liệu còn thiếu trong các bộ bằng giá trị None

A screenshot of a computer

Description automatically generated

Một số bộ dữ liệu ban đầu



Dữ liệu sau tiền xử lý dữ liệu

Sau khi tiền xử lý dữ liệu số bộ ban đầu là 1520 giảm còn 1026 bộ

**Hình ảnh trực quan hóa**

* Trực quan hóa dữ liệu **Business operations survey 2022: information and communications technology** thông qua cột level và value (lấy 20 dòng đầu) A graph with blue rectangles

  Description automatically generated
* Trực quan hóa dữ liệu **Environmental tax account 1999-2021** thông qua cột National\_accounts\_var và data\_value (lấy 5 dòng đầu)

A graph of blue bars

Description automatically generated with medium confidence