Đại học Khoa học Tự nhiên,
Đại học quốc gia HCM

Trực quan hóa dữ liệu

Lab3

Sinh viên

21127592 – Nguyễn Minh Đạt

21127191 – Nguyễn Nhật Truyền

Lớp: 21 KHDL
Giảng viên lý thuyết: Bùi Tiến Lên
Giảng viên thực hành: Lê Nguyễn Nhựt Trường

Công việc	Người thực hiện	Mức độ hoàn thành
Data Understanding	21127592 + 21127191	100%
EDA with Num	21127191 – Nguyễn Nhật Truyền	100%
EDA with Num and Cate	21127592 – Nguyễn Minh Đạt	100%
Ask a question	21127592 – Nguyễn Minh Đạt	100%
Insight	21127592 + 21127191	100%
Report	21127191 – Nguyễn Nhật Truyền	100%

I. Thông tin dataset:

Tên bộ dữ liệu: Iris Specices
Nguồn: Kaggle
File: Iris.csv
Danh sách thuộc tính:

STT	Tên Cột	Ý Nghĩa
1	Id	Id
2	SepalLengthCm	Length of the sepal (in cm)
3	SepalWidthCm	Width of the sepal (in cm)
4	PetalLengthCm	Length of the petal (in cm)
5	PetalWidthCm	Width of the petal (in cm)
6	Species	Species name

Cách thức xử lý dữ liệu:

- Đầu tiên, chúng em lưu bộ dữ liệu từ trên Kaggle về máy tính cá nhân.
- Tiếp theo, chúng em đẩy dữ liệu lên Github Repo của một thành viên trong nhóm (Nguyễn Minh Đạt).
- Cuối cùng, chúng em thực hiện lệnh “git clone" trong file notebook để lưu và sử dụng dữ liệu. Trong đó:
    - Môi trường mà nhóm chọn để viết code là Google Colab.
    - Ngôn ngữ Python phiên bản 3.0.

II. Phân tích và Đặt câu hỏi:

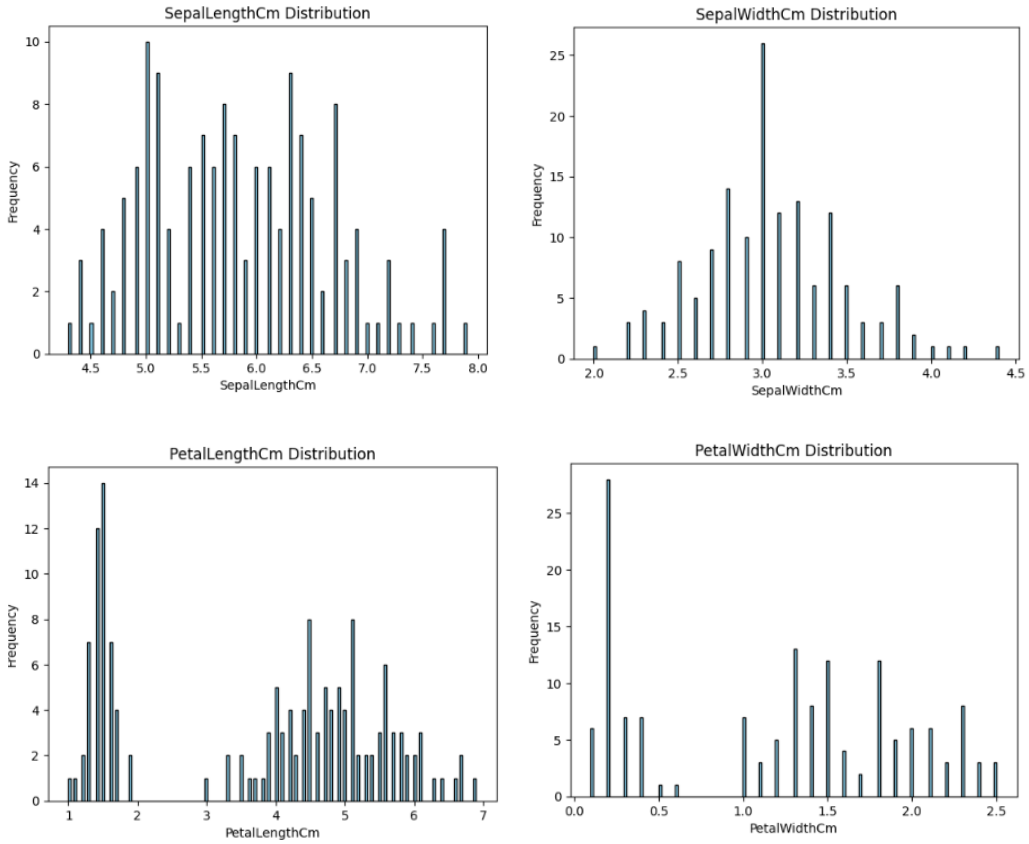
Vì đây là một data set cho tác vụ classification nên chúng ta sẽ có một số câu hỏi liên quan đến tác vụ này:

- Phân phối cái giá trị của các biến Num như thế nào?
- Tính tương quan giữa các biến Num như thế nào?
- Phân phối các giá trị Num ảnh hưởng đến Cate(label) thế nào?
    - Có phan chia cụm hay không?
    - Nếu phân chia được cụm thì nó có rõ ràng hay không?

- Mỗi lớp (label) có đặt điểm gì, phân bố các điểm ngoại lệ ra sao?
- Có đặc điểm gì nổi bật đối với khả năng phân loại của các biến Num hay không.

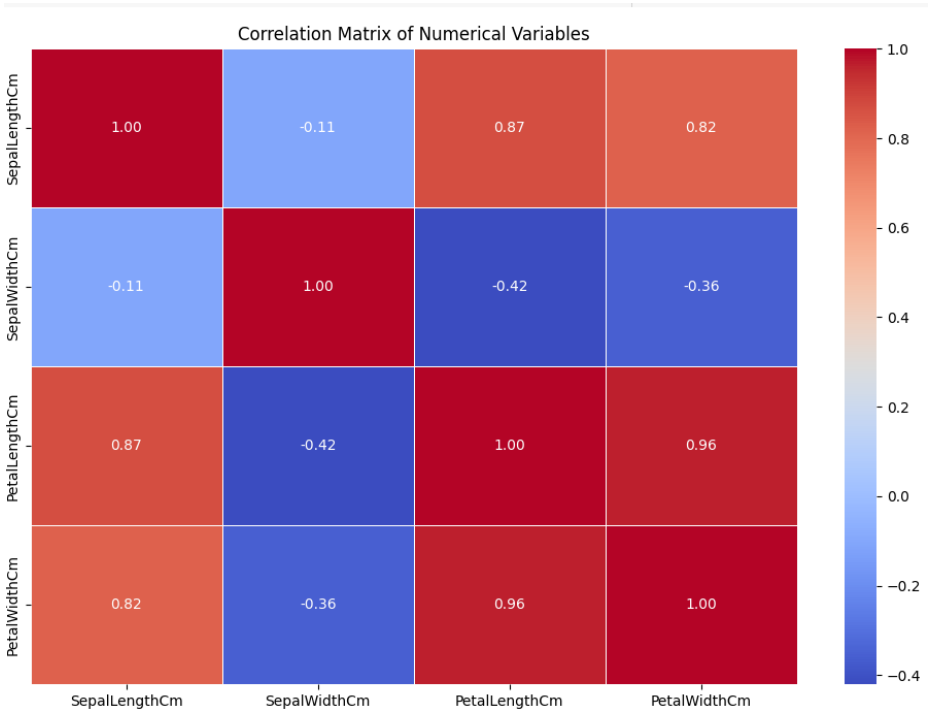
III. Cài đặt và Trả lời câu hỏi:

3.1 Phân phối cái giá trị của các biến Num như thế nào?



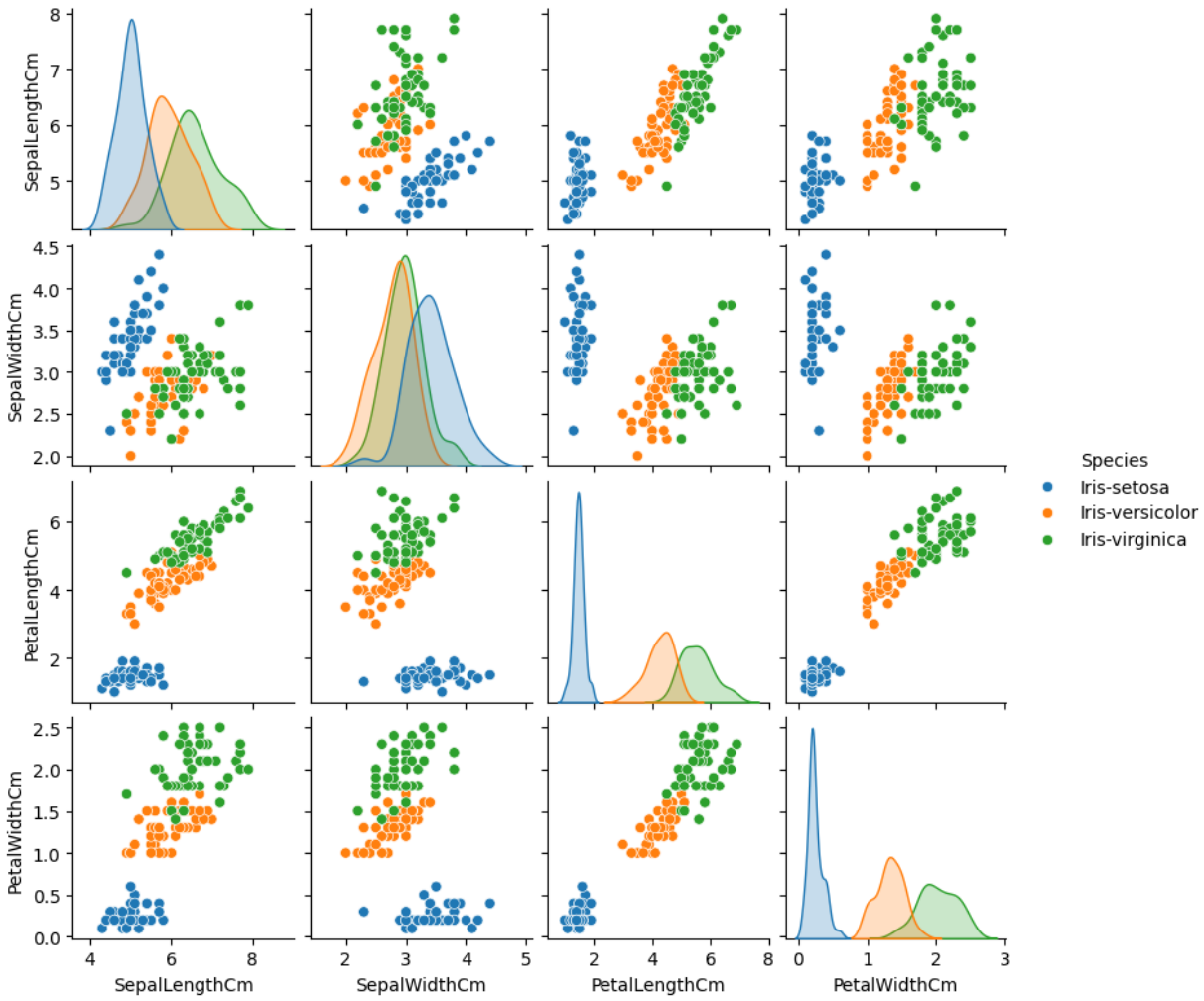
- Các giá trị của các biến Num phân bố không đều.
- Đối với 2 biến PentalLengthCm và PetalWidthCm phân bố giá trị 2 biến này bị tách ra làm 2 phần.

3.2 Tính tương qua giữa các biến Num như thế nào?



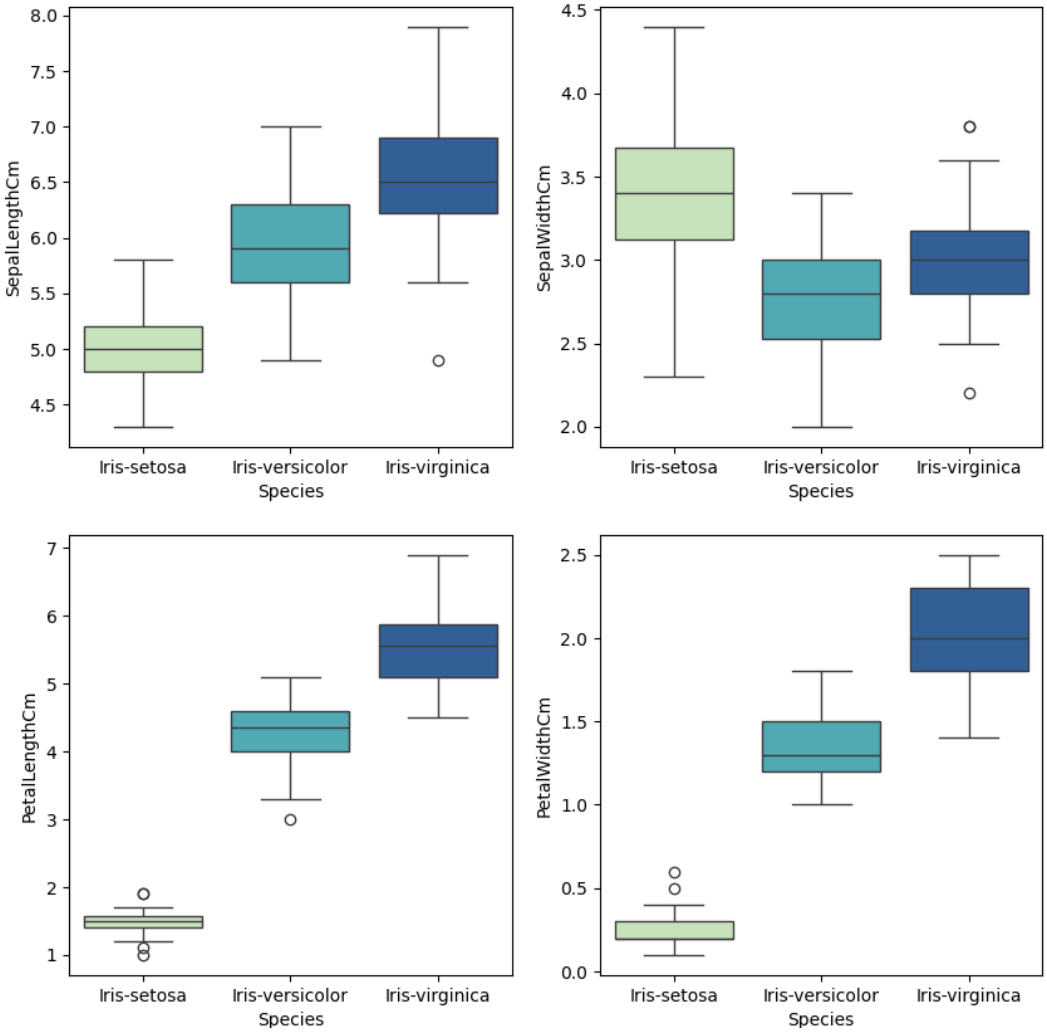
- Petal width và petal length có tương quan cao.
- Petal length và sepal width có tương quan tốt.
- Petal Width và Sepal length có tương quan tốt.

3.3 Phân phối các giá trị Num ảnh hưởng đến Cate(label) thế nào?



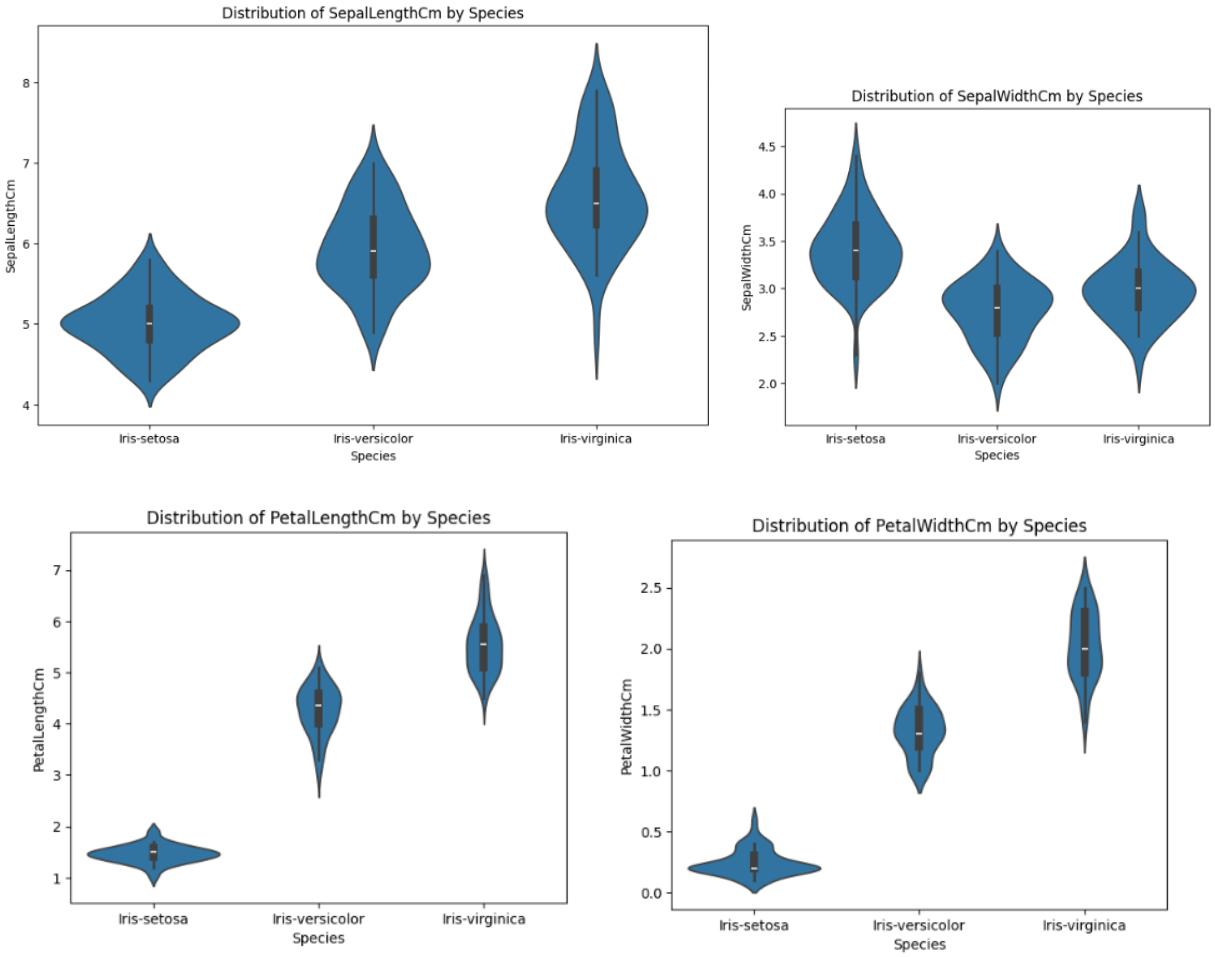
- Chúng ta có thể thấy nhiều loại mối quan hệ từ biểu đồ này, chẳng hạn như loài Setosa có chiều rộng và chiều dài cánh hoa nhỏ nhất. Nó cũng có chiều dài lá đài nhỏ nhất nhưng chiều rộng lá đài lớn. Thông tin như vậy có thể được thu thập đối với các loài khác.
- Nhìn sơ chỉ với 2 tham số ta có thể thấy các giá trị có khả năng phân cụm khá cao:
  - Trong hầu hết các biểu đồ thì Iris-setosa có khả năng phân cụm rất cao (gần như là phân biệt hoàn toàn với 2 lớp còn lại).
  - Đối với 2 lớp Iris-versicolor và Iris-virginica thì cũng có khả năng phân cụm tương đối tốt (có một số điểm ngoại lệ, tuy nhiên tổng quan thì khả năng phân cụm này khá tốt)

3.4 Mỗi lớp (label) có đặt điểm gì, phân bố các điểm ngoại lệ ra sao?



- Loài Setosa có đặc điểm nhỏ nhất và ít phân bố hơn với một số ngoại lệ.
- Loài Versicolor có đặc điểm trung bình.
- Loài Virginica có những đặc tính cao nhất

3.5 Có đặc điểm gì nổi bật đối với khả năng phân loại của các biến Num



- Iris-setosa có khả năng phân loại khá rõ với 2 thuộc tính PetalLengthCm và PentalWidthCm.
- Đối với 2 biến SepalLengthCm và SepalWidthCm thì khả năng phân loại ít rõ ràng hơn.

IV. Tổng kết và Nêu phát hiện nổi bật:

Công nghệ:

- pandas:
  - Lưu dữ liệu (Dataframe)
  - Tính toán trên Dataframe
- matplotlib, seaborn:
  - Vẽbiểu đồ

Phát hiện:

- Đây là một data set cho tác vụ classification, và số lượng các mẫu thuộc các lớp khác nhau là cân bằng.
- Các lớp phân cụm tương đối tốt và ít có điểm ngoại lệ.
- Đối với lớp Iris-setosa việc phân loại khá rõ ràng (2 thuộc tính PetalLengthCm và PentalWidthCm).

➢ Đánh giá đây là một bài toán phân loại khá đơn giản và sẽ có ít sai số.