ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

TRỰC QUAN HÓA DỮ LIỆU

< Lab 03 - IRIS >

Sinh viên thực hiện: 21127115 - Trần Thanh Ngân
21127229 - Dương Trường Bình
21127616 - Lê Phước Quang Huy

Giảng viên hướng dẫn: TS. Bùi Tiến Lên

Lớp: 21KHDL

Mục lục

Thông tin nhóm và phân công công việc

MSSV	Họ và tên	Công việc được phân công	Mức độ hoàn thành
21127115	Trần Thanh Ngân	• A. Mô tả dữ liệu • C. Khám phá và đặt câu hỏi (câu hỏi 1)	100%
21127229	Dương Trường Bình	• B. Phân tích tổng quan (Categorical) • C. Khám phá và đặt câu hỏi (câu hỏi 2)	100%
21127616	Lê Phước Quang Huy	• B. Phân tích tổng quan (Numerical) • C. Khám phá và đặt câu hỏi (câu hỏi 3)	100%

Tiến độ công việc

Phần	Nội dung	Mức độ hoàn thành
A. Mô tả dữ liệu	1. Viết bảng mô tả về tập dữ liệu	100%
	2. Phân tích tỷ lệ missing rate	100%
A. Phân tích tổng quan	1. Categorical	100%
	2. Numerical	100%
	3. Phân tích 4 biến Numerical với biến Species	100%
C. Khám phá và đặt câu hỏi	Đặt câu hỏi về tập dữ liệu và trả lời (câu hỏi 1 - 3)	100%
D. Insights	Chia sẻ các phát hiện thú vị.	100%

1 Tổng quan

> **T**ập dữ liệu Iris Species bao gồm ba loài **iris** với 50 mẫu mỗi loài cũng như một số đặc tính về mỗi loài hoa. Một loài hoa có thể phân tách tuyến tính với hai loài còn lại, nhưng hai loài còn lại không thể phân tách tuyến tính với nhau.

1.1 Iris Species

Tập dữ liệu **Iris.csv** chứa 150 dòng và 6 cột, mỗi dòng tương ứng với đặc điểm của một loài hoa:

STT	Tên thuộc tính	Mô tả	Giá trị	Kiểu dữ liệu
1	Id	Mã của hoa	Nằm trong phạm vi từ 1 đến 150	Integer
2	SepalLengthCm	Chiều dài lá	Nằm trong khoảng từ 4.3 đến 7.9	Float
3	SepalWidthCm	Chiều rộng lá	Nằm trong khoảng từ 2 đến 4.4	Float
4	PetalLengthCm	Chiều dài cánh hoa	Nằm trong khoảng từ 1 đến 6.9	Float
5	PetalWidthCm	Chiều rộng cánh hoa	Nằm trong khoảng từ 0.1 đến 2.5	Float
6	Species	Giống loài hoa	**Iris-setosa**, **Iris-versicolor** và **Iris-virginica**	String

1.2 Phân phối cột Categorical

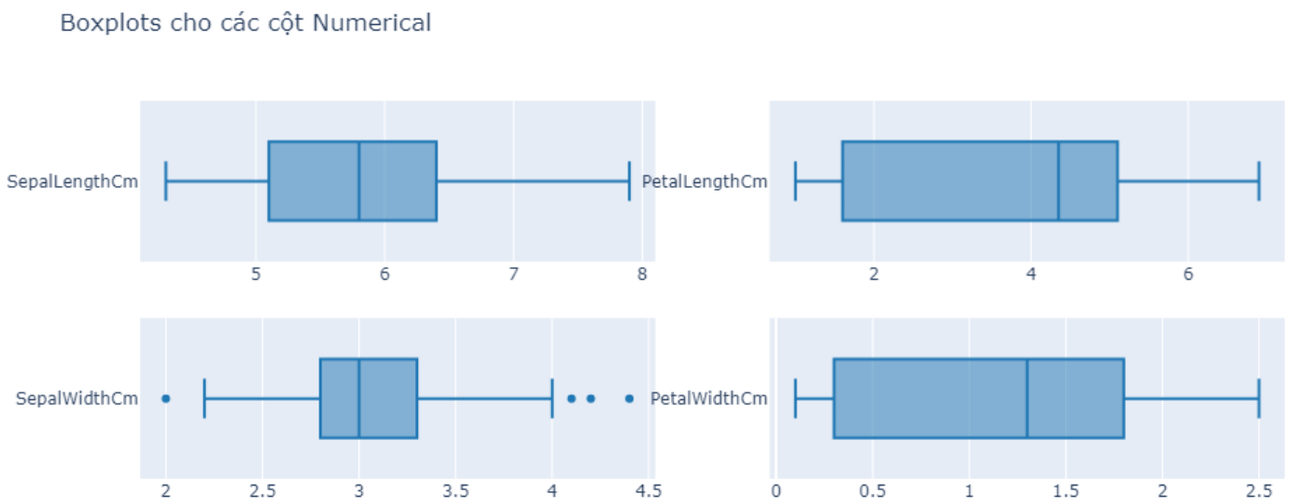
Nhóm sử dụng biểu đồ tròn để thể hiện phân phối các loài hoa.

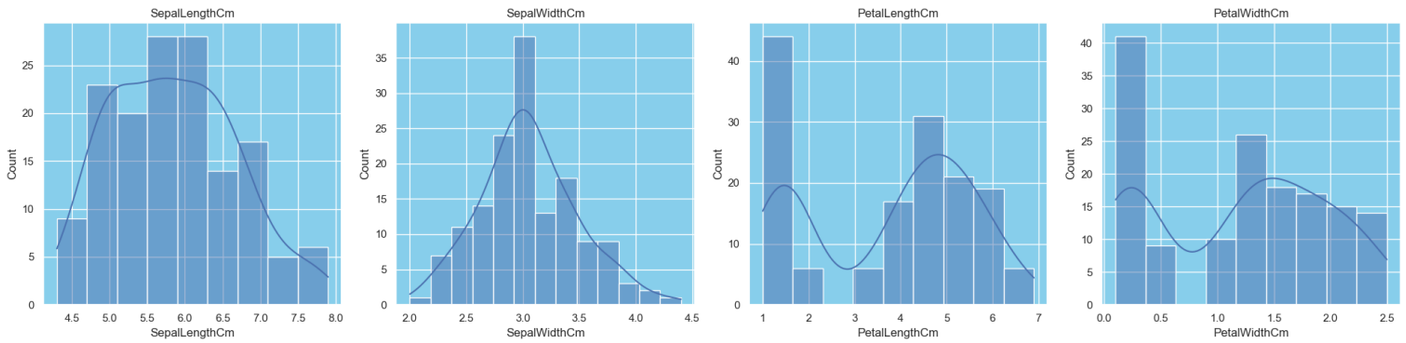
Nhận xét:

- Ba giống loài hoa iris là **Iris-setosa**, **Iris-virginica** và **Iris-versicolor** phân bố đồng đều, với 50 mẫu dữ liệu cho mỗi giống loài (khoảng 33.3%).

1.3 Phân phối cột Numerical

Nhóm sử dụng biểu đồ Boxplot và Histogram để thể hiện phân phối các đặc điểm của các loài hoa.





Nhận xét:

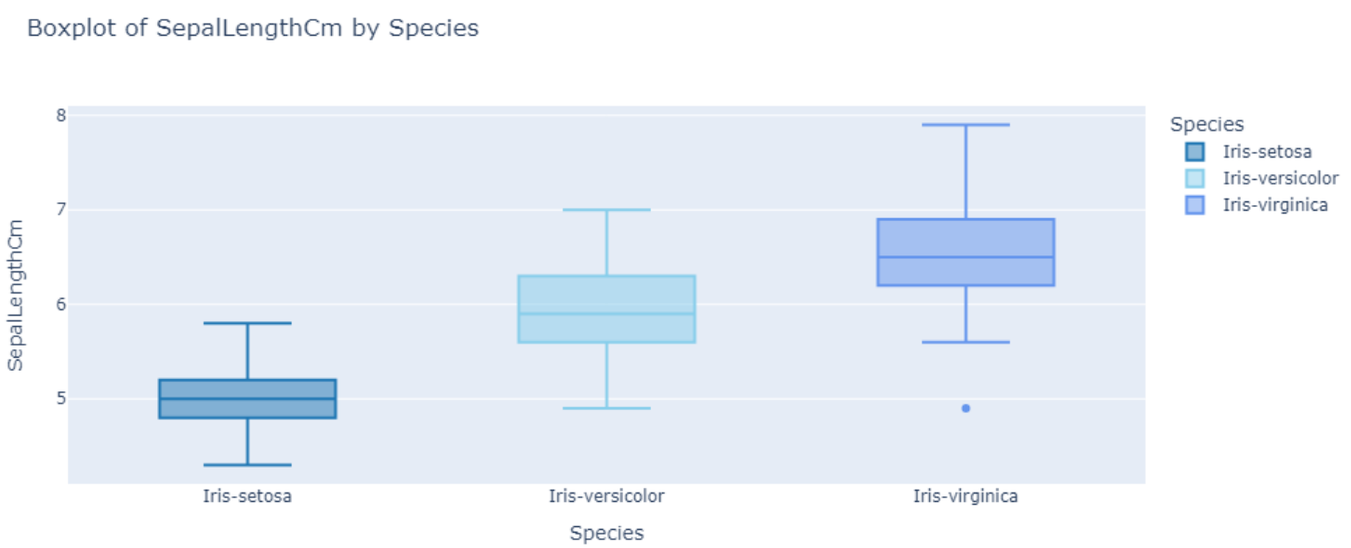
- Từ biểu đồ Boxplot ở trên chúng ta có thể thấy rằng SepalWidthCm chứa một số ngoại lệ và tất cả các giá trị khác đều hoàn toàn ổn
- **Chiều dài của đài hoa (SepalLengthCm)**: Phân phối này có dạng phân phối chuẩn tương đối, với nhiều quan sát tập trung trong khoảng từ 5 đến 7 cm.

- **Chiều rộng của đài hoa (SepalWidthCm)**: Phân phối này cũng khá chuẩn nhưng có một đỉnh rõ rệt xung quanh giá trị 3 cm. Điều này có thể cho thấy rằng chiều rộng của đài hoa trong tập dữ liệu này có xu hướng tập trung nhiều vào một giá trị cụ thể.
- **Chiều dài của cánh hoa (PetalLengthCm)**: Phân phối này cho thấy có hai nhóm dữ liệu rõ rệt, một nhóm có chiều dài ngắn hơn (từ 1 đến 2 cm) và nhóm còn lại có chiều dài dài hơn (từ 3 đến 6 cm). Điều này giúp nhận biết sự phân chia trong đặc tính của chiều dài cánh hoa và có thể liên quan đến sự phân biệt giữa các loài trong tập dữ liệu iris.
- **Chiều rộng của cánh hoa (PetalWidthCm)**: Phân phối này cũng cho thấy sự tồn tại của hai nhóm dữ liệu. Một nhóm quan sát tập trung chủ yếu dưới 1 cm và nhóm còn lại tập trung trong khoảng từ 1 đến 2,5 cm. Đây cũng có thể là dấu hiệu của sự phân biệt giữa các loài trong tập dữ liệu.

1.4 Phân tích giữa 4 biến Numerical và biến Species

Việc phân tích hai biến là một phương pháp thống kê được sử dụng để khám phá mối quan hệ giữa hai biến cùng một lúc. Phân tích này rất quan trọng để xác định các mối quan hệ tiềm ẩn, xác định mối liên hệ giữa các biến số và hiểu rõ hơn về đặc tính chung của chúng.

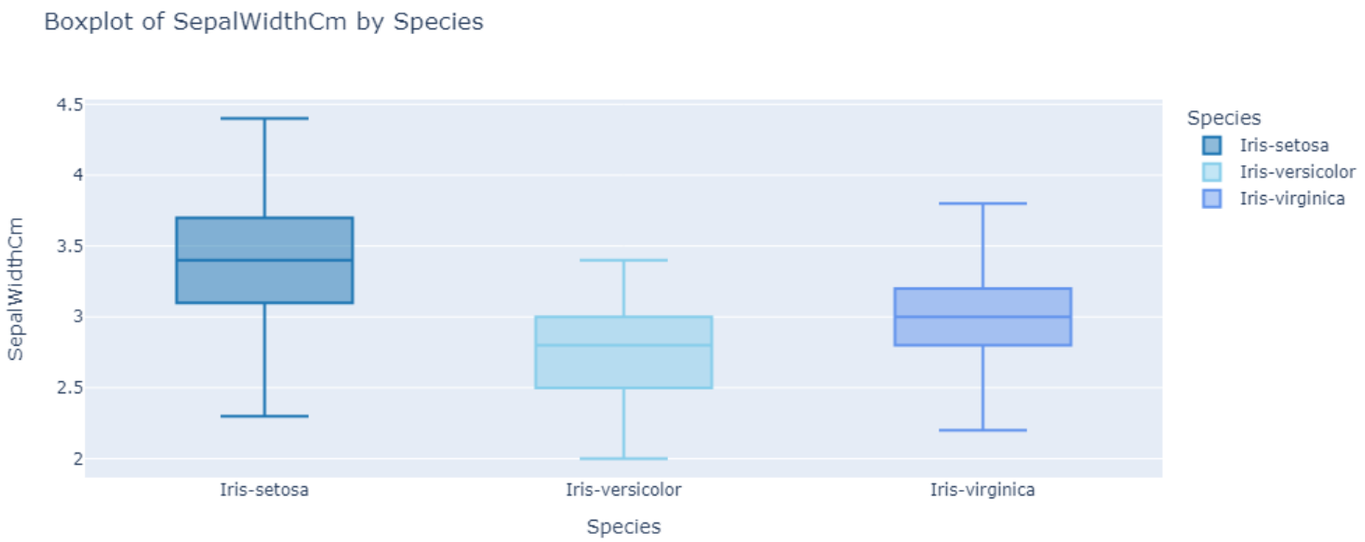
SepalLengthCm vs. Species



Quan sát:

- Từ biểu đồ trên, có thể thấy rõ rằng loài Virginica có giá trị trung bình về Chiều dài đài hoa (SepalLengthCm) cao hơn so với các loài khác. Ngược lại, loài Setosa thể hiện giá trị trung bình của Chiều dài đài hoa thấp hơn.
- Giá trị tối đa của Chiều dài đài hoa của loài Setosa là 5,8cm và giá trị thấp hơn của loài Setosa là 4,3cm.
- Giá trị tối đa của Chiều dài đài hoa của các loài Versicolor là 7cm và giá trị thấp hơn là 4,9.
- Giá trị tối đa của Chiều dài đài hoa của loài Virginica là 7,9cm và giá trị thấp hơn là 5,6.

SepalWidthCm vs. Species

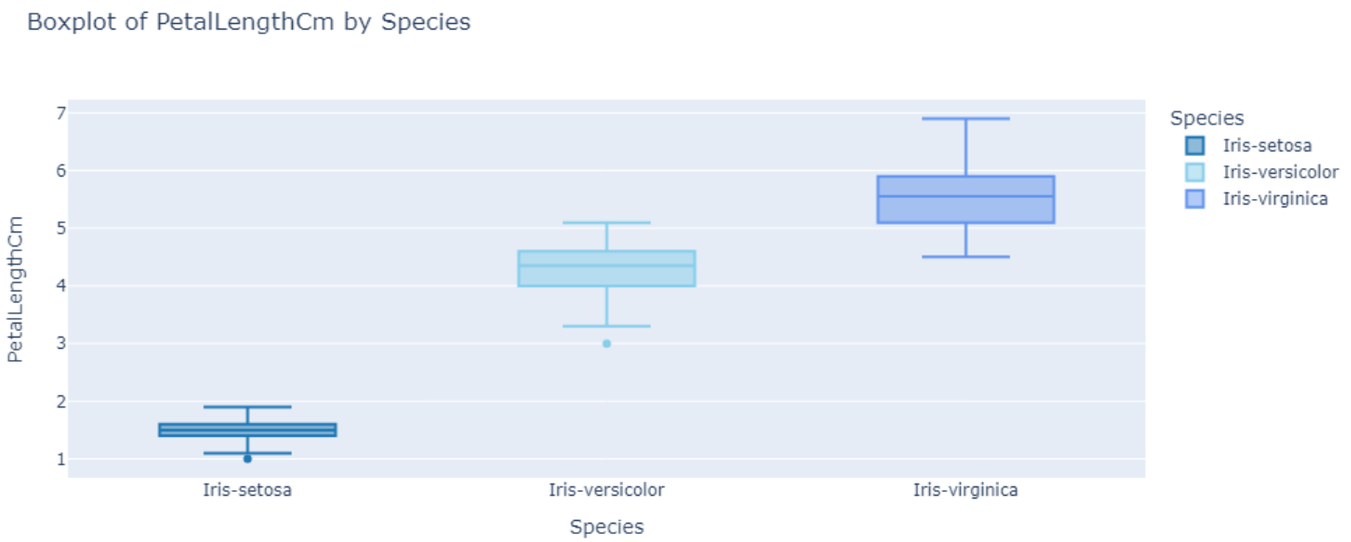


Quan sát:

- Ta có thể thấy rõ rằng loài Setosa có giá trị trung bình về Chiều rộng đài hoa (SepalWidthCm) cao hơn so với các loài khác. Ngược lại, loài Versicolor thể hiện giá trị trung bình của Độ rộng đài hoa thấp hơn.
- Giá trị tối đa của chiều rộng đài hoa của loài Setosa là 4,4cm và giá trị trung bình thấp hơn của chiều rộng đài hoa là 2,3cm.

- Giá trị tối đa của chiều rộng đài hoa của loài Versicolor là 3,4cm và giá trị thấp hơn là 2cm.
- Giá trị tối đa của chiều rộng đài hoa của loài Virginica là 3,8cm và giá trị thấp hơn là 2,2cm.

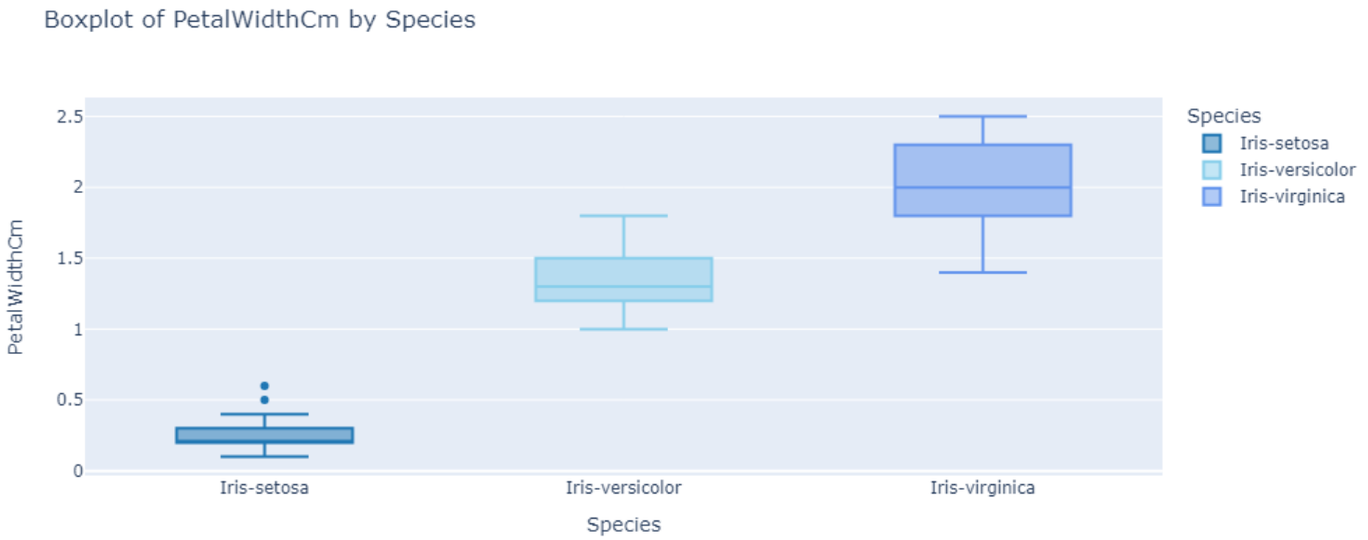
PetalLengthCm vs. Species



Quan sát:

- Từ biểu đồ trên, có thể thấy rõ loài Virginica có giá trị trung bình về Chiều dài cánh hoa (PetalLengthCm) cao hơn so với các loài khác. Ngược lại, loài Setosa thể hiện giá trị trung bình của Chiều dài cánh hoa thấp hơn.
- Giá trị tối đa của Chiều dài cánh hoa của loài Setosa là 1,9cm và giá trị trung bình thấp hơn của Setosa là 1.
- Giá trị tối đa của chiều dài cánh hoa của loài Versicolor là 5,1 và giá trị trung bình thấp hơn của loài nhiều màu là 3,3.
- Giá trị chiều dài cánh hoa tối đa của các loài Virginica là 6,9 và giá trị trung bình thấp hơn của Virginica là 4,5.

PetalWidthCm vs. Species

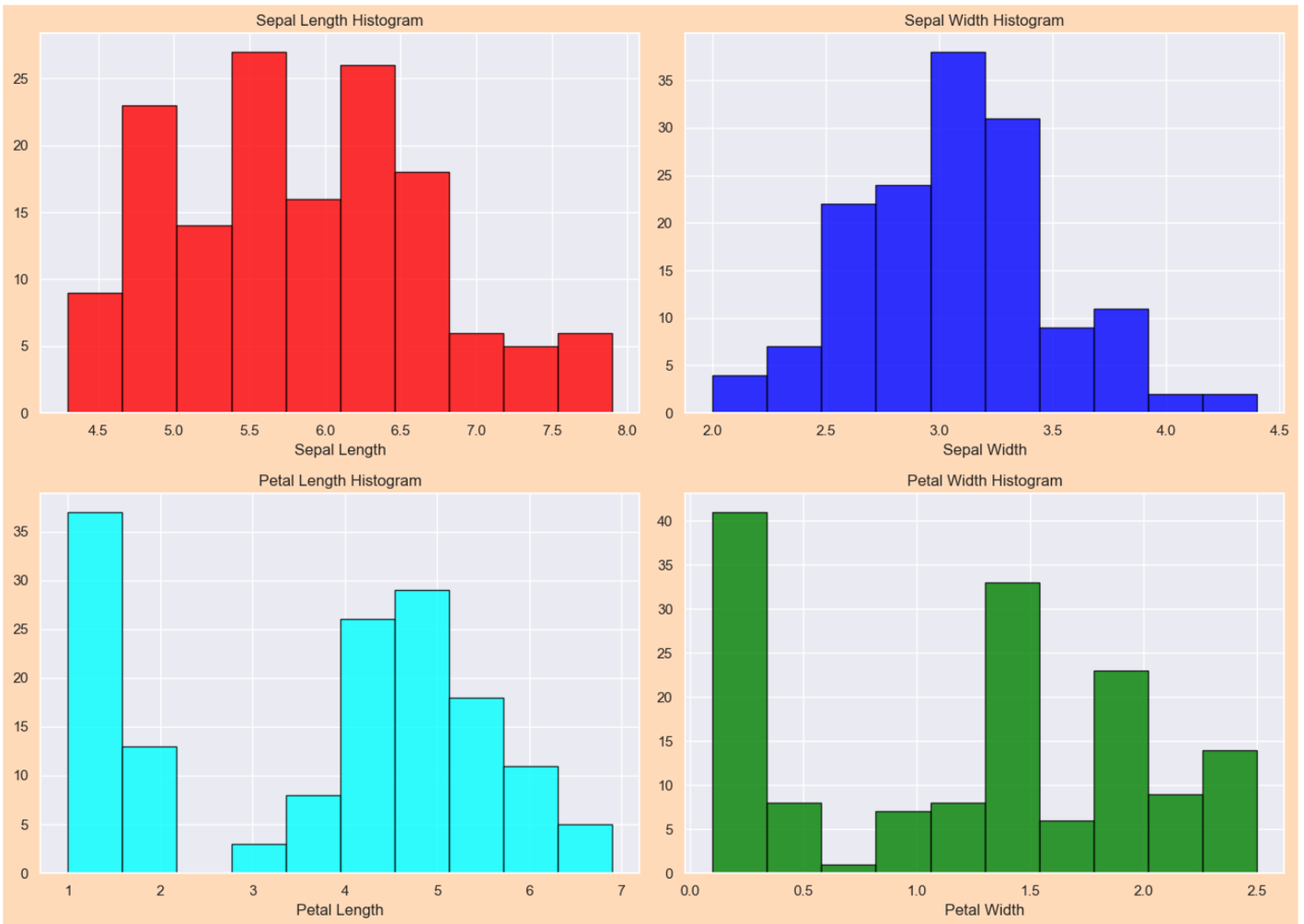


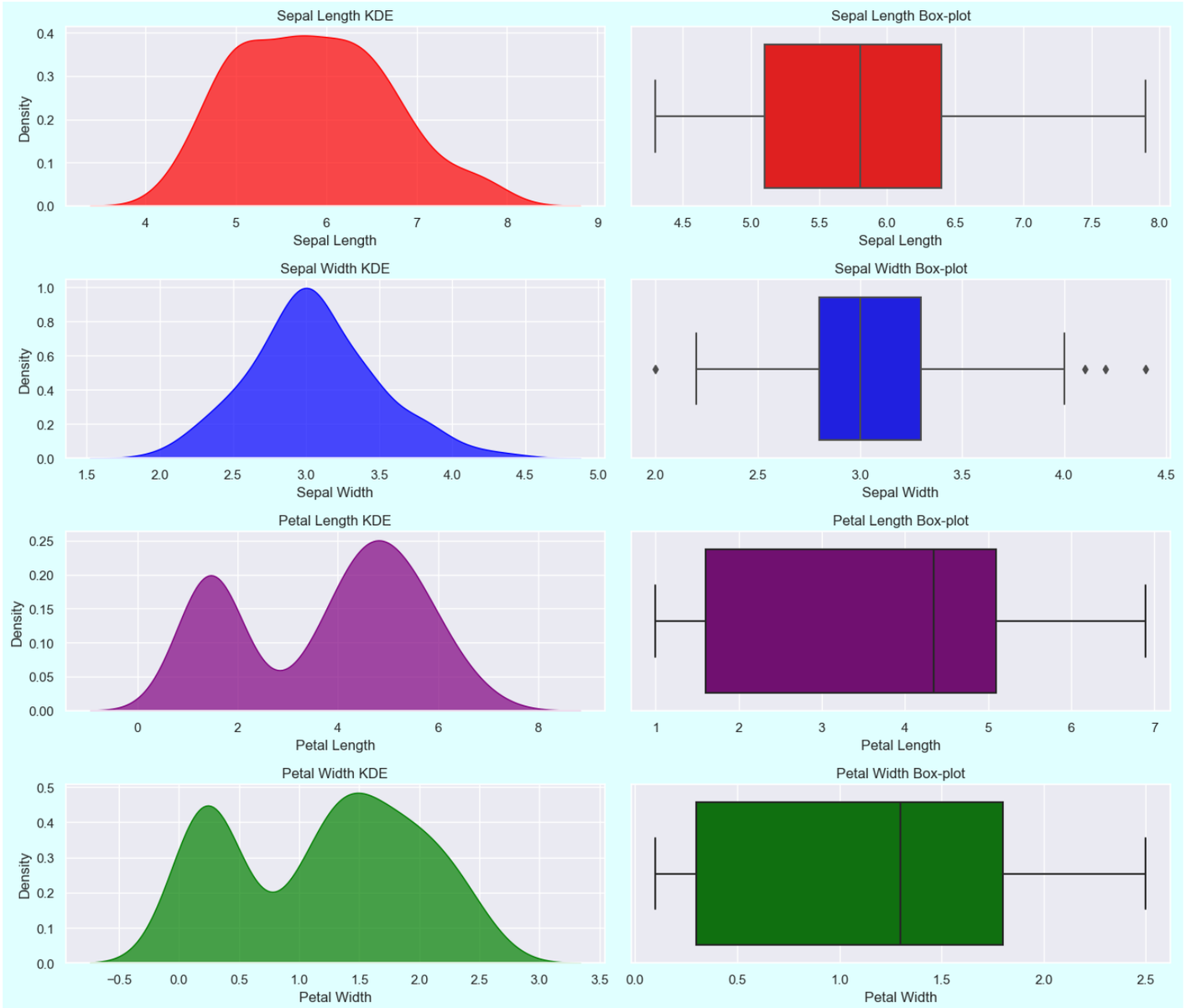
Quan sát:

- Từ biểu đồ trên, có thể thấy rõ rằng loài Virginica có giá trị trung bình về Chiều rộng cánh hoa (PetalWidthCm) cao hơn so với các loài khác. Ngược lại, loài Setosa thể hiện giá trị trung bình của Chiều rộng cánh hoa thấp hơn.
- Giá trị tối đa của Chiều dài cánh hoa của loài Setosa là 0,6cm và giá trị trung bình thấp hơn của Setosa là 0,1.
- Giá trị tối đa của chiều dài cánh hoa của loài Versicolor là 1,8cm và giá trị trung bình thấp hơn của loài nhiều màu là 1cm.
- Giá trị chiều dài cánh hoa tối đa của loài Virginica là 2,5cm và giá trị trung bình thấp hơn của loài Virginica là 1,4cm.

2 Khám phá và phân tích dữ liệu

2.1 Câu hỏi 1: Phân phối của các đặc tính của hoa Iris như thế nào? Có sự khác biệt nào rõ rệt khi xem xét phân bố của các đặc tính không?





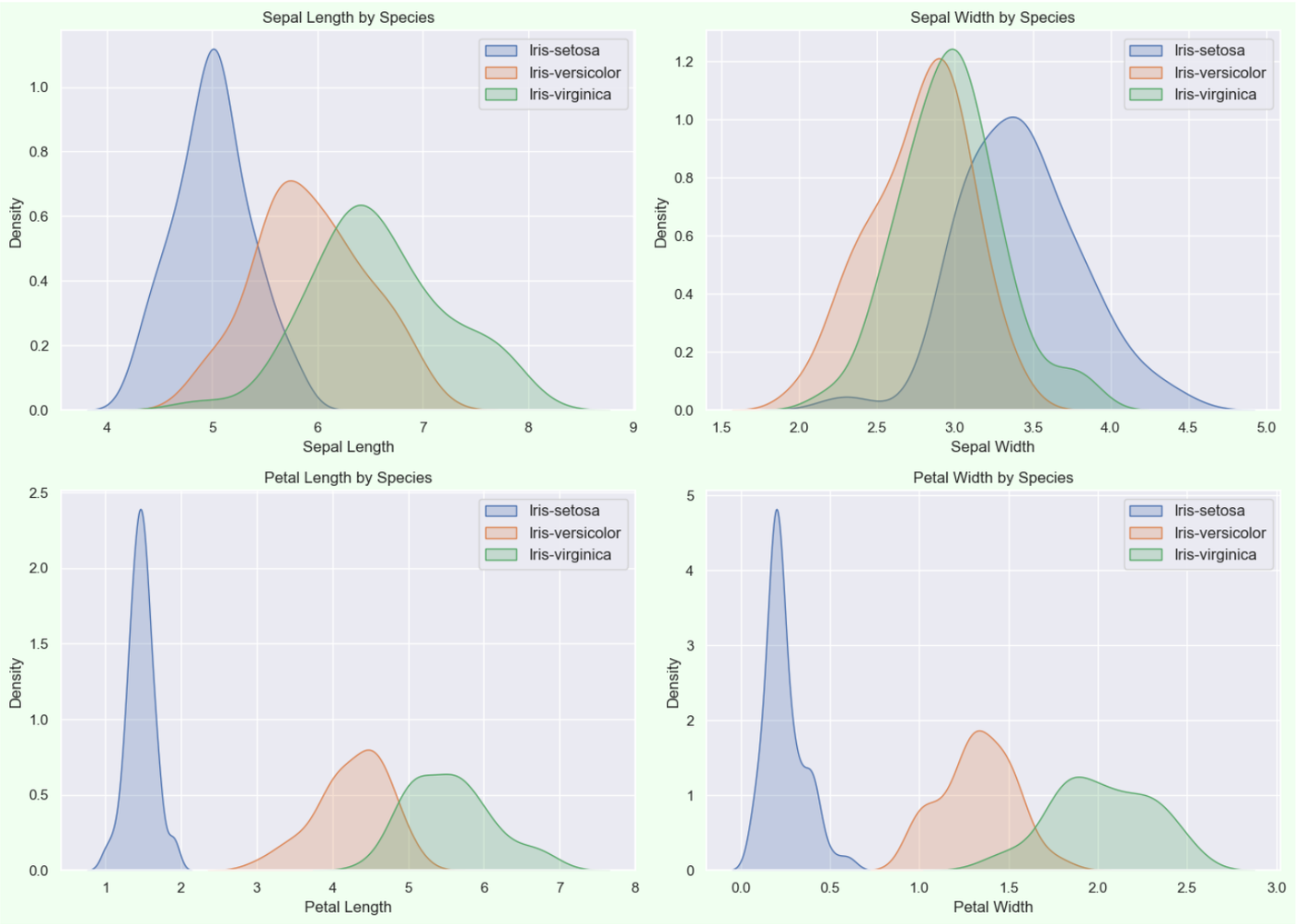
Trả lời:

- Ở câu hỏi đầu tiên, nhóm đã trực quan biểu đồ historgram và boxplot để xem xét về tổng quan về phân bố của các đặc tính của hoa Iris.
- Đối với đặc tính SepalLengthCm, SepalWidthCm, tức là chiều dài và chiều rộng của lá, ta thấy rằng phân phối của chúng khá giống nhau và boxplot cho thấy khoảng giá trị hẹp hơn so với 2 đặc trưng còn lại, , khó để thấy sự khác biệt rõ rệt giữa các loại hoa.
- Đối với đặc tính PetalLengthCm, PetalWidthCm, tức là chiều dài và chiều rộng của cánh hoa, ta thấy rằng có sự tách biệt rõ rệt trong phân phối của 2 dặc tính này qua historgram. Biểu đồ boxplot cũng cho thấy một khoảng giá trị khá rộng. Do đó ta có thể

nhận xét: có thể có một loài hoa có chiều dài và chiều rộng cánh hoa nhỏ hơn so với các loài hoa còn lại.

- Để biết được rõ hơn về sự khác biệt giữa các loài hoa, nhóm sẽ tiếp tục khám phá ở các câu hỏi tiếp theo.

2.2 Câu hỏi 2: Có sự khác biệt của phân bố các đặc tính giữa các loài hoa không?



Trả lời:

- SepalLengthCm: Chúng ta có thể thấy có sự giao nhau giữa cả iris-setosa và iris-versicolor; giữa iris-versicolor và iris-virginica; cũng như giữa cả 3 loài hoa trong biểu đồ histogram trên. Điều đó làm cho việc phân biệt giữa các loài hoa trở nên khó khăn hơn.
- SepalWidthCm: Có sự giao nhau khá nhiều giữa cả 3 loài hoa, đặc biệt là giữa iris-

versicolor và iris-virginica, phân bố của chúng dường như trùng vào nhau. Phân bố của loài iris-setosa nằm về phía bên phải hơn so với phân bố ở SepalLengthCm. Điều này cho thấy nếu chỉ dựa vào SepalWidthCm để phân biệt giữa các loài hoa thì còn khó khăn hơn nữa.

- PetalLengthCm và PetalWidthCm: Có sự phân biệt rõ rệt giữa các loài hoa, đặc biệt là giữa iris-setosa và 2 loài hoa còn lại. Phân bố của iris-setosa nằm ở phía bên trái hơn so với 2 loài hoa còn lại, còn phân bố của iris-versicolor và iris-virginica có một sự giao nhau nhẹ, nhưng vẫn phân biệt được. Điều này cho thấy rằng PetalLengthCm và PetalWidthCm là 2 đặc tính quan trọng để phân biệt giữa các loài hoa.

2.3 Câu hỏi 3: Những yếu tố nào hình thành nên sự khác biệt của các loài hoa

Trả lời:

- Từ biểu đồ 1 ta thấy được:
  - Loài Setosa có chiều dài lá đài ngắn hơn nhưng chiều rộng lá đài lớn hơn.
  - Loài Versicolor có chiều dài và rộng của lá đài đều nhỏ hơn so với 2 loài còn lại
  - Loài Virginica có lá đài dài hơn nhưng chiều rộng lá đài nhỏ hơn.
- Đối với biểu đồ 2:

  - Loài Setosa có chiều dài và chiều rộng cánh hoa nhỏ hơn hẳn so với 2 loài còn lại.
  - Loài Versicolor nằm ở giữa đường chéo phụ của hai loài còn lại về chiều dài và chiều rộng của cánh hoa.
  - Loài Virginica có chiều dài và chiều rộng cánh hoa lớn nhất.

  => Từ những thông tin trên ta có một số kết luận như:

- Loài hoa có cánh hoa nhỏ, ngắn và lá đài ngắn nhưng chiều rộng lá đài lớn thì ta có thể suy ra đây là loài Setosa
- Loài mà có cả lá đài dài và cánh hoa vừa rộng vừa dài thì đó là loài Virginica

3 Insights

Trong quá trình khám phá toàn diện tập dữ liệu **Iris** bằng các công cụ trực quan phân tích như Matplotlib, Seaborn và Plotly, nhóm đã tìm hiểu sâu về các đặc tính của các loài hoa Iris. Trong phần này, nhóm sẽ chia sẻ những phát hiện thú vị nhất, bao gồm các đặc tính riêng biệt, sự khác nhau giữa các loài hoa:

- **Tập dữ liệu:**
  - Ngoài giá trị SepalWidthCm chứa một số ngoại lệ vthì tất cả các giá trị khác còn lại đều hoàn toàn ổn
  - Cả SepalLengthCm và SepalWitdhCm đều có dạng phân phối chuẩn, SepalWitdhCm đạt đỉnh rõ rệt xung quanh giá trị 3cm còn với SepalLengthCm thì nhiều quan sát tập trung trong khoảng từ 5 đến 7 cm.
  - PetalLengthCm và PetalWidthCm cả 2 phân phối của 2 cột này đều cho thấy có 2 nhóm dữ liệu rõ rệt.
- **Đặc tính của các loài hoa:**
  - Ta có loài hoa Virginica đứng đầu trong 3 đặc tính là chiều dài đài hoa (SepalLengthCm), cả chiều dài và rộng của cánh hoa (Petal).
  - Ngược lại với loài Virginica thì loài Setosacó giá trị 3 đặc tính trên thấp nhất.
  - Loài Virsicolor các đặc tính đều ở mức trung bình.
- **Sự khác biệt giữa các loài hoa:**
  - Từ một số đặc tính đã được rút trích ở trên, ta thấy rằng loài hoa có cánh hoa nhỏ, ngắn và lá đài ngắn nhưng chiều rộng lá đài lớn thì ta có thể suy ra đây là loài Setosa
  - Loài mà có cả lá đài dài và cánh hoa vừa rộng vừa dài thì đó là loài Virginica