Introduction to Data Science

Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Trường Đại học Công nghiệp thành phố Hồ Chí Minh-IUH

Cho dữ liệu Iris như sau:

https://archive.ics.uci.edu/dataset/53/iris

Hãy thực hiện các yêu cầu sau:

1- Đọc dữ liệu trên Google Colab bằng cách share cả thư mục và trên máy tính cá nhân.

2- Viết chương trình (tự viết code không sử dụng thư viện) mã hóa cột class (3 loại hoa) thành các con số tương ứng với các số 1, 2, 3

3- Viết hàm tính giá trị lớn nhất, nhỏ nhất và trung bình (tự viết code không sử dụng thư viện).

4- Áp dụng 3 hàm vừa viết ở 3 vào từng cột 1,2,3,4 và hiển thị kết quả ra màn hình.

5- Viết hàm tính Độ lệch chuẩn (standard deviation) theo công thức sau:

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

6- Áp dụng 3 hàm vừa viết ở 5 vào từng cột 1,2,3,4 và hiển thị kết quả ra màn hình.

7- Viết hàm tính mối độ tương quan pearson theo công thức sau:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

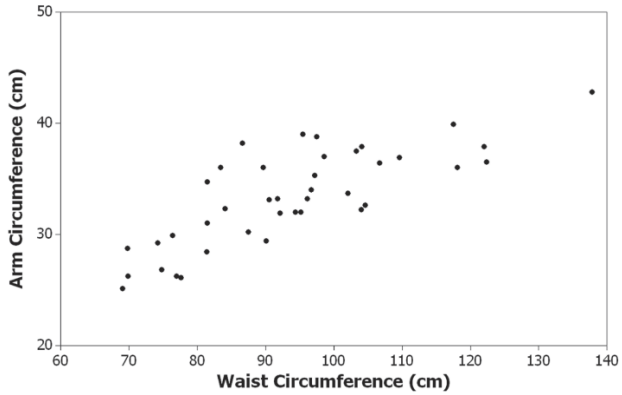
8- Tính mối tương quan của từng cột 1, 2, 3 và 4 với cột cuối cùng.

Trực quan hóa kết quả này bằng biểu đồ tự vẽ và nhận xét về kết quả

9- Trực quan hóa dữ liệu với biểu đồ dạng hộp (box) (sử dụng thư viện matplotlib)

10- Trực quan hóa dữ liệu với biểu đồ scatter (sử dụng thư viện matplotlib)

như ví dụ sau



11- Trực quan hóa dữ liệu với biểu đồ time serius graph (sử dụng thư viện matplotlib) như ví dụ sau: