iris_data

사용 데이터: iris (붓꽃 데이터)

꽃잎의 각 부분의 너비와 길이들을 측정한 데이터

데이터 출처| https://www.kaggle.com/datasets/saurabh00007/iriscsv

데이터 colum 의미

id : 단순 순서 표시

SepalLengthCm : 꽃받침의 길이 정보 / 단위: cm

SepalWidthCm : 꽃받침의 너비 정보 / 단위: cm

PetalLengthCm : 꽃잎의 길이 정보 / 단위: cm

PetalWidthCm : 꽃잎의 너비 정보 / 단위: cm

Species : 꽃의 종류 정보 / 종류:setosa, versicolor, virginica

데이터 확인

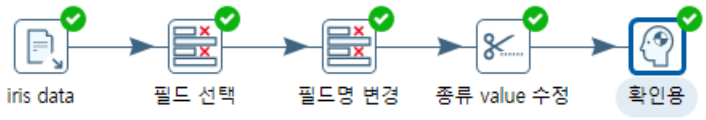
#	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	1	5.1	3.5	1.4	0.2	Iris-setosa
2	2	4.9	3.0	1.4	0.2	Iris-setosa
3	3	4.7	3.2	1.3	0.2	Iris-setosa
4	4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	5.0	3.6	1.4	0.2	Iris-setosa
6	6	5.4	3.9	1.7	0.4	Iris-setosa
7	7	4.6	3.4	1.4	0.3	Iris-setosa
8	8	5.0	3.4	1.5	0.2	Iris-setosa
9	9	4.4	2.9	1.4	0.2	Iris-setosa
10	10	4.9	3.1	1.5	0.1	Iris-setosa
11	11	5.4	3.7	1.5	0.2	Iris-setosa
12	12	4.8	3.4	1.6	0.2	Iris-setosa
13	13	4.8	3.0	1.4	0.1	Iris-setosa
14	14	4.3	3.0	1.1	0.1	Iris-setosa
15	15	5.8	4.0	1.2	0.2	Iris-setosa
16	16	5.7	4.4	1.5	0.4	Iris-setosa
17	17	5.4	3.9	1.3	0.4	Iris-setosa
18	18	5.1	3.5	1.4	0.3	Iris-setosa
19	19	5.7	3.8	1.7	0.3	Iris-setosa
20	20	5.1	3.8	1.5	0.3	Iris-setosa
21	21	5.4	3.4	1.7	0.2	Iris-setosa
22	22	5.1	3.7	1.5	0.4	Iris-setosa

원본 csv파일

할 일

1. **데이터 분석시, 필요없는 Id컬럼을 없애고, Species에서 종류만 나오도록 변경하기**
2. **결측값(null)이 있는지 확인하기**

1. 데이터 필드 선택, 필드명 변경, value 수정

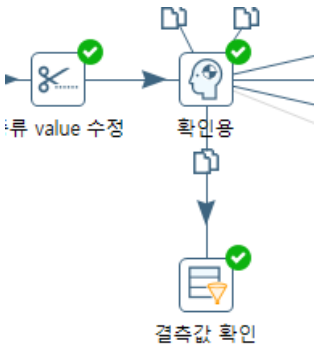


수정을 위한 flow

#	꽃받침_길이	꽃받침_너비	꽃잎_길이	꽃잎_너비	종류
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa

수정 결과

2. 결측값 확인



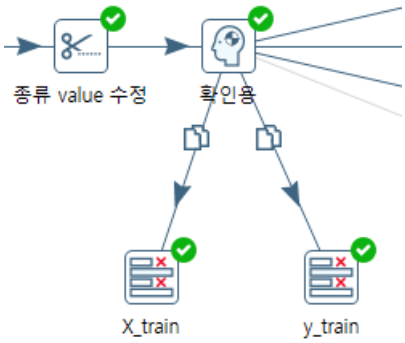
Filter rows를 통해 결측값이 존재하는지 확인하기

⇒ 결측값이 없다는 것 확인

#	꽃받침_길이	꽃받침_너비	꽃잎_길이	꽃잎_너비	종류
1					

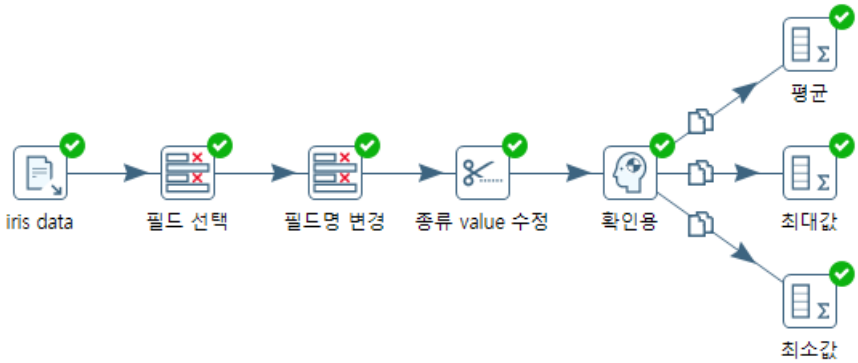
X_train과 y_train으로 나누기

이후, 모델링을 할 경우를 위해 X_train과 y_train으로 나누기



+ 추가적으로, 각 종류별로 평균,max,min 구하기

Group by를 통해 종류별로 평균, 최대값, 최소값 구하기



평균

#	종류	평균_꽃받침_길이	평균_꽃받침_너비	평균_꽃잎_길이	평균_꽃잎_너비
1	setosa	5.006	3.418	1.464	0.244
2	versicolor	5.936	2.77	4.26	1.326
3	virginica	6.588	2.974	5.552	2.026

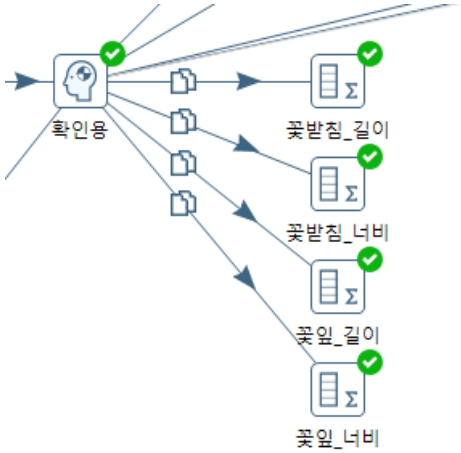
최대값

#	종류	max_꽃받침_길이	max_꽃받침_너비	max_꽃잎_길이	max_꽃잎_너비
1	setosa	5.8	4.4	1.9	0.6
2	versicolor	7.0	3.4	5.1	1.8
3	virginica	7.9	3.8	6.9	2.5

최소값

#	종류	min꽃받침_길이	min_꽃받침_너비	min_꽃잎_길이	min_꽃잎_너비
1	setosa	4.3	2.3	1.0	0.1
2	versicolor	4.9	2.0	3.0	1.0
3	virginica	4.9	2.2	4.5	1.4

+ 각 colum별 종류에 따른 개수,평균,중위수,최소,최대,표준편차 다시 구한 것



꽃받침_길이

#	종류	개수_꽃받침_길이	평균_꽃받침_길이	중위수_꽃받침_길이	최소_꽃받침_길이	최대_꽃받침_길이	표준편차_꽃받침_길이
1	setosa	50	5.006	5.0	4.3	5.8	0.3489469874
2	versicolor	50	5.936	5.9	4.9	7.0	0.5109833657
3	virginica	50	6.588	6.5	4.9	7.9	0.6294886814

꽃받침_너비

#	종류	개수_꽃받침_너비	평균_꽃받침_너비	중위수_꽃받침_너비	최소_꽃받침_너비	최대_꽃받침_너비	표준편차_꽃받침_너비
1	setosa	50	3.418	3.4	2.3	4.4	0.3771949098
2	versicolor	50	2.77	2.8	2.0	3.4	0.3106444913
3	virginica	50	2.974	3.0	2.2	3.8	0.3192553837

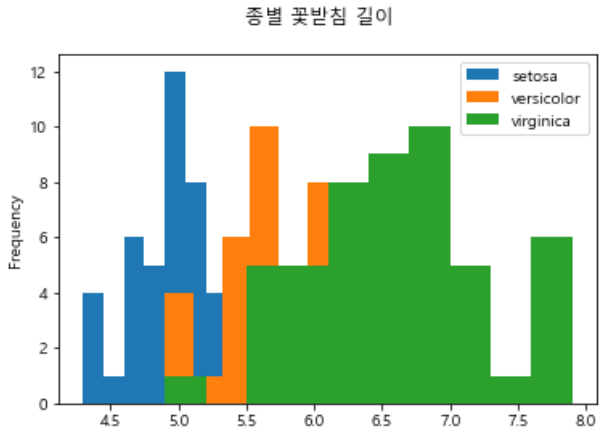
꽃잎_길이

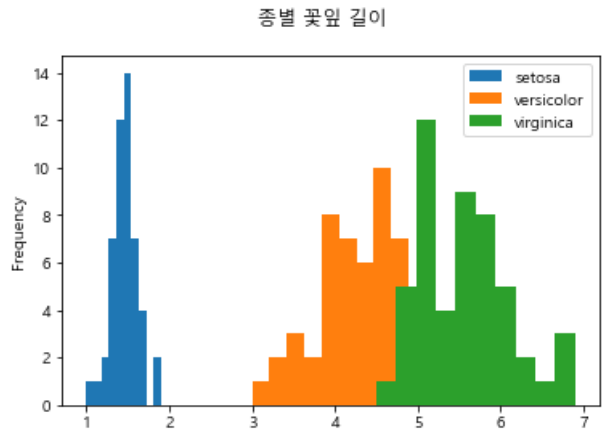
#	종류	개수_꽃잎_길이	평균_꽃잎_길이	중위수_꽃잎_길이	최소_꽃잎_길이	최대_꽃잎_길이	표준편차_꽃잎_길이
1	setosa	50	1.464	1.5	1.0	1.9	0.1717672844
2	versicolor	50	4.26	4.35	3.0	5.1	0.465188134
3	virginica	50	5.552	5.55	4.5	6.9	0.5463478745

꽃잎_너비

#	종류	개수_꽃잎_너비	평균_꽃잎_너비	중위수_꽃잎_너비	최소_꽃잎_너비	최대_꽃잎_너비	표준편차_꽃잎_길이
1	setosa	50	0.244	0.2	0.1	0.6	0.1061319933
2	versicolor	50	1.326	1.3	1.0	1.8	0.1957651654
3	virginica	50	2.026	2.0	1.4	2.5	0.2718896835

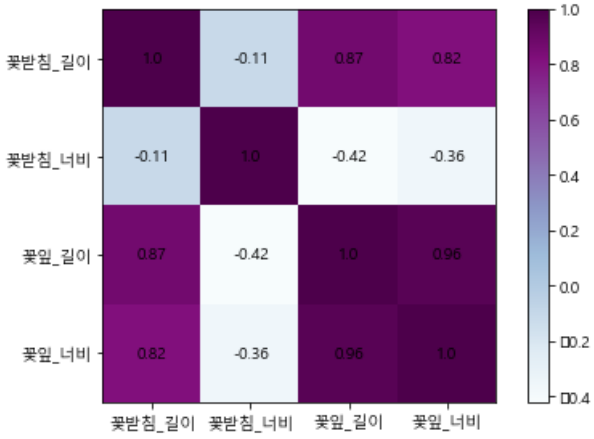
위 데이터를 파이썬 matplotlib을 통해 간단하게 시각화 진행





크기는 보통 setosa < versicolor < virginica 순인걸 알 수 있고,

꽃잎의 경우 setosa가 다른 두종에 비해 많이 작다는 것을 알 수 있었다.



상관관계를 살펴보면,

꽃잎_길이와 꽃잎_너비가 0.96으로 관계가 깊고,

꽃받침_길이와 꽃잎_길이는 0.87으로 높은 편이다.

꼽받침_길이와 꽃받침_너비는 -0.11로 0에 가깝기때문에 관계가 낮다.