


통계패키지활용자료분석 (2021년 2학기)			
담당 교수 : 김 태 수			
강좌 번호	100982-31001	본인의 과제 자체 평가	5점
과제명 : 제2차 자료 실습 with R [IRIS]			

이름	강동현
	

제 출 일	2021년 10월 1일
학 과	컴퓨터공학과
학 번	강동현

2. 목차

-3. 자료 설명

-4. 자료 분석

-4-1. Sepal.Width , Sepal.Length , Petal.Width , Petal.Length의 특징

-4-2. Sepal.Width , Sepal.Length , Petal.Width , Petal.Length 요소들간의 특징

-4-3 .Sepal.Width , Sepal.Length , Petal.Width , Petal.Length 요소들간의 상관
관계

-4-4. setosa, versicolor, virginica를 분류하는 시나리오

-5. 결론

3. 자료 설명

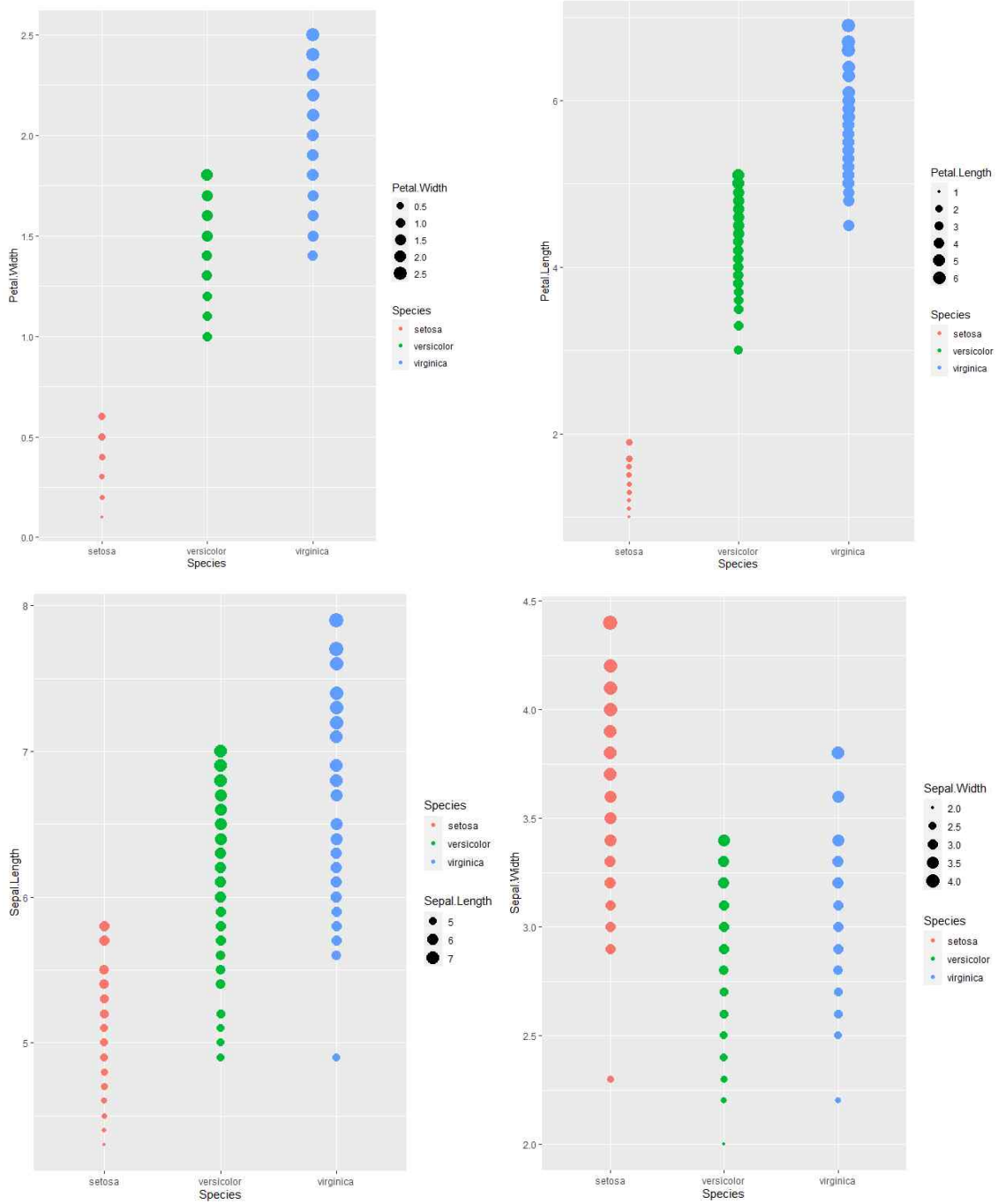
IRIS 데이터는 붓꽃의 3가지 종(setosa, versicolor, virginica)에 대해 꽃받침(sepal)과 꽃잎
[petal]의 길이를 정리한 데이터이다. 통계학자 피셔[Fisher]가 소개한 데이터로 각 붓꽃의 데
이터 마다 5개의 요소를 가지고 있다. 각각 Sepal.Length , Sepal.Width , Petal.Length ,
Petal.Width , Species 이다. 이 IRIS 데이터에는 붓꽃의 종별로 50행씩, 총 150개 행이 저
장되어 있다.

컬럼명	의미	데이터 타입
Species	붓꽃의 종. setosa, versicolor, virginica 세 가지 값 중 하나	Factor
Sepal.Width	꽃받침의 너비	Number
Sepal.Length	꽃받침의 길이	Number
Petal.Width	꽃잎의 너비	Number
Petal.Length	꽃잎의 길이	Number

4. 자료 분석

4-1. Sepal.Width , Sepal.Length , Petal.Width , Petal.Length의 특징

붓꽃의 데이터를 종을 기준으로 각 요소들이 어떤 특징을 띄고 있는지 살펴본다. 따라서 각
붓꽃종의 Sepal.Width , Sepal.Length , Petal.Width , Petal.Length 의 특징을 살펴보기
위해 Point 그래프를 사용한다.



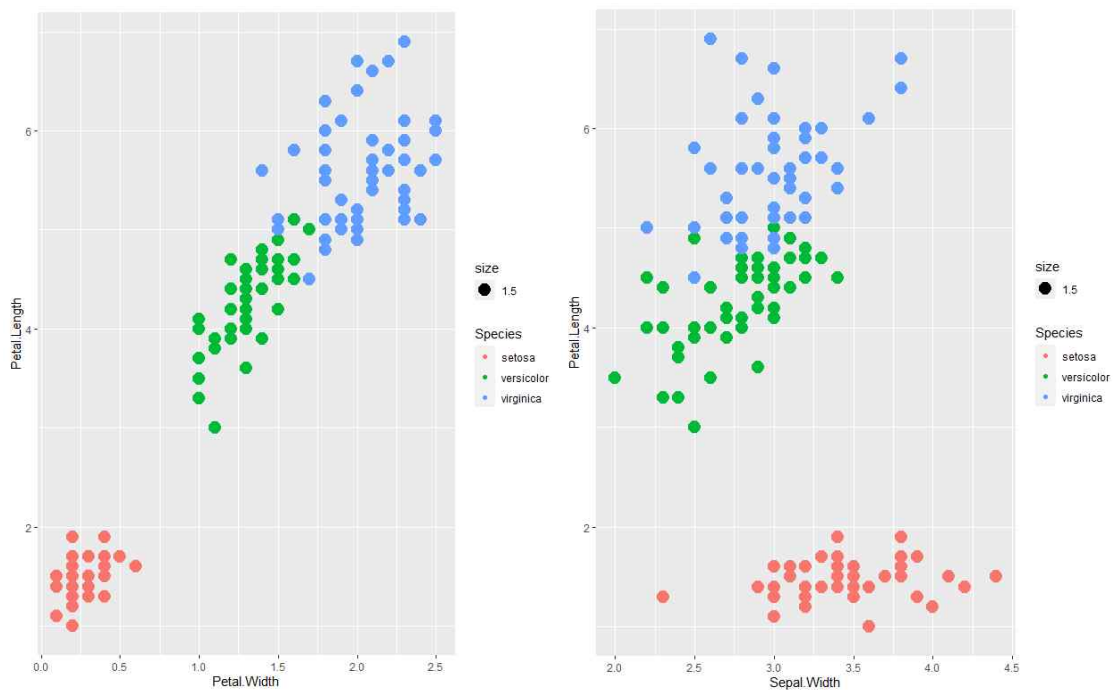
첫번째 그래프 , Petal.Width의 그래프에서 볼 수 있는 특징은 setosa가 다른 붓꽃의 종들에 비해 낮은 값을 갖는다는 점이다. 이 그래프를 통해 “붓꽃의 Petal.Width가 일정 수치 이하면 그 붓꽃의 종은 setosa 다” 라는 가설을 세울 수 있다.

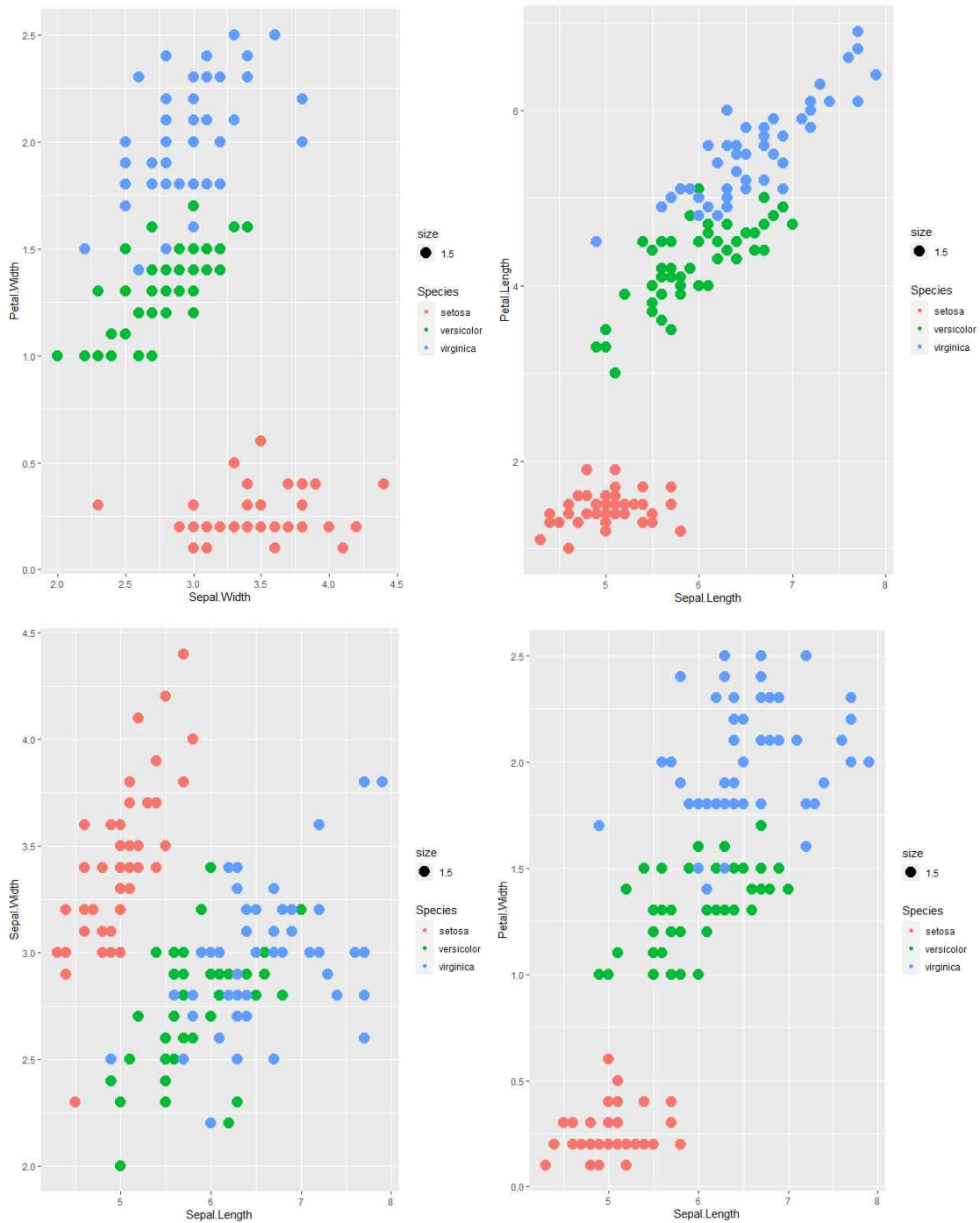
두번째 그래프 , Petal.Length의 경우에도 Petal.Width의 그래프와 같게 setosa가 다른 붓꽃의 종들에 비해 낮은 값을 갖는 모습을 보여주고 있다. 첫번째 그래프의 가설에 더불어 Petal과 관련된 수치가 일정 수치 이하이면 그 붓꽃을 setosa라고 분류 할 수 있다는 가설을 세울 수 있다.

세번째 그래프와 네번째 그래프 , Sepal.Length와 Sepal.Width의 경우에는 특별한 특징을 찾을 수는 없다. 세번째 그래프에서 각 붓꽃 품종의 Sepal.Length가 setosa , versicolor, virginica 순으로 커지는 경향을 찾을 수는 있지만 , 각 종의 Sepal.Length 범위가 상당 부분 겹치므로 이를 통해 특별한 가설을 세우기에는 어려워 보인다.

각 요소들을 각각 종과 연결시켜 살펴보았을 때 , Petal과 관련된 수치가 일정 수치 이하이면 그 붓꽃을 setosa라고 분류 할 수 있다는 가설을 세울 수 있었다. 종과 더불어 각 요소들간의 연결고리가 존재할 수도 있으므로 이번에는 종과 다른 두 요소를 묶어 함께 살펴보도록 한다.

4-2. Sepal.Width , Sepal.Length , Petal.Width , Petal.Length 요소들간의 특징





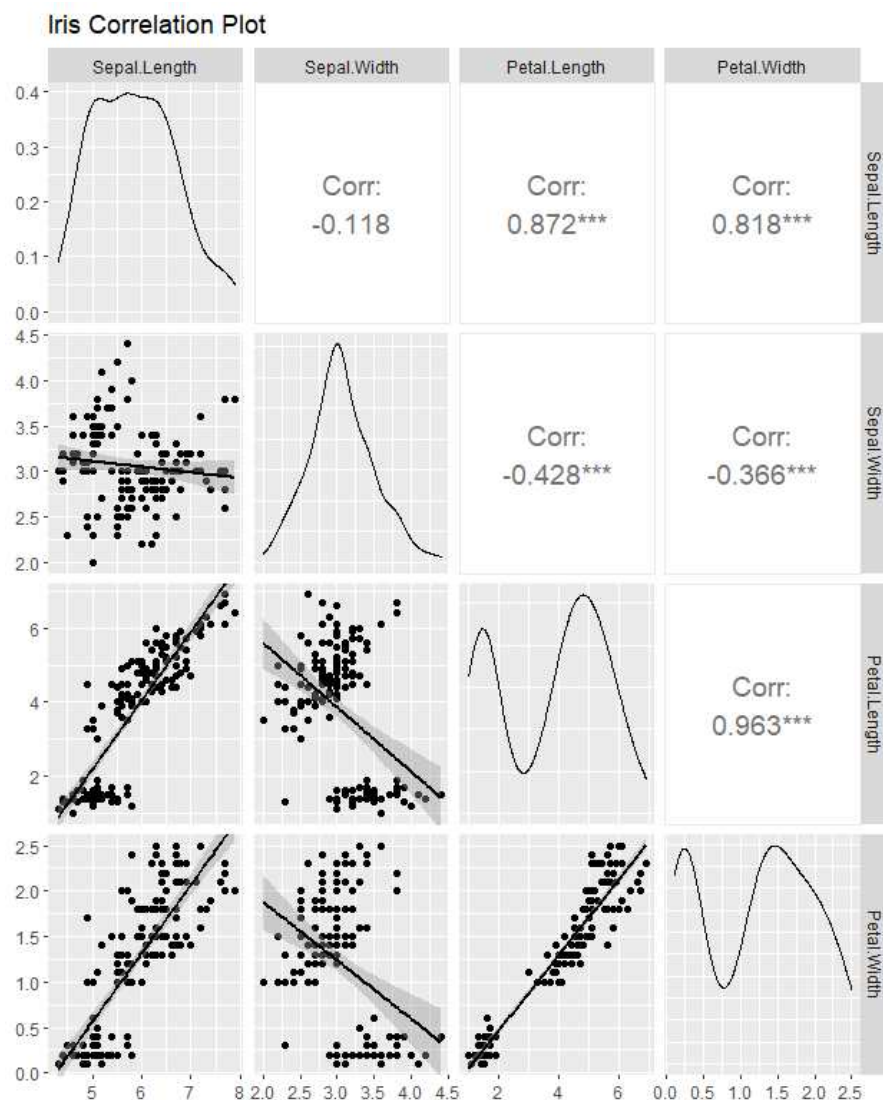
Petal.Length와 Petal.Width가 포함된 그래프들을 살펴보면, setosa가 타 붓꽃 종에 비해 해당수치들이 낮은 것을 볼 수 있다. 이러한 특징을 가장 잘 보여주는 그래프가 첫번째 Petal.Length와 Petal.Width 그래프이다. 첫번째 그래프를 확인해보면 setosa는 타 붓꽃 품종과 뚜렷히 구분되고 또 특정 값 이하인 것을 볼 수 있다. 현재 IRIS 데이터에 따르면 Petal.Width가 0.75 이하이고 Petal.Length가 2 이하인 경우에 대해서 붓꽃의 품종을 setosa로 구분할 수 있다. 이는 이 IRIS 데이터에 국한된 것이며, 데이터의 수가 커질 수록

더 정확한 결과를 얻을 수 있고 혹은 완전히 다른 결과를 얻을 수도 있다.

versicolor와 virginica의 경우 1번 그래프에서 virginica가 versicolor에 비해 높은 값을 보여주고 있음을 알 수 있다. 눈으로 보기에는 Petal.Length와 Petal.Width를 통해 virginica와 versicolor를 구분할 수 있을 것으로 보인다. 이 가설은 이후 확인해보도록 한다.

위의 그래프에서 눈여겨 보아야 할 점은, 각 그래프 별로 눈에 띄게 선형적인 관계를 보여주는 그래프들이 있다는 점이다. 위의 그래프에서는 1번과 4번, Petal.Length와 Petal.Width 그래프와 Sepal.Length와 Petal.Length 그래프 이다. 눈으로 보았을 때 두 그래프는 선형관계를 이루고 있는 것처럼 보여진다. 이번단계에서 추측한 선형관계에 대한 가설이 실제로 맞는지 모든 요소에 대한 Correlation Plot을 살펴보고 가설이 맞는지 실제로 살펴본다.

4-3 .Sepal.Width, Sepal.Length, Petal.Width, Petal.Length 요소들간의 상관관계



우리가 가설을 세운 부분에서 눈여겨 보아야 하는 그래프는 Petal.Length와 Petal.Width 그래프와 Sepal.Length와 Petal.Length 그래프 이다.

먼저 Petal.Length와 Petal.Width 그래프를 살펴보자. 이전에 Petal.Length와 Petal.Width 사이에는 선형관계가 존재한다고 가설을 세웠다. 위의 Correlation Plot을 통해 두 관계를 살펴보니 실제로 선형관계가 존재한다고 말할 수 있다. 즉 Petal 요소 , 꽃잎의 너비와 길이는 양적 선형관계라는 것을 알 수 있다. 즉 우리는 위의 그래프를 통해 꽃잎의 너비와 길이는 선형으로 함께 증가하고 함께 감소한다는 점을 알 수 있다.

이어서 Sepal.Length와 Petal.Length 그래프 역시 살펴보도록 하자. 이 역시도 선형관계가 나타남을 알 수 있다. 즉 Sepal.Length와 Petal.Length 꽃받침의 너비와 꽃잎의 길이는 선형적으로 함께 증가하고 함께 감소한다는 사실을 알 수 있다.

이 두가지 사실을 통해서 새로운 가설을 세울 수 있다. Petal.Length와 Petal.Width 사이에는 선형적으로 함께 증가하고 감소하는 관계가 존재하므로 Petal.Width가 증가하고 감소하면 Petal.Length 역시 선형적으로 증가하고 감소할 것이다. 여기에 이어 Sepal.Length와 Petal.Length 사이에도 선형관계가 증가하므로 Petal.Length가 증가하고 감소함에 따라 Sepal.Length도 선형적으로 증가하고 감소할 것이다. 이 사실을 통해 우리는 Petal.Width 증가/감소 -> Petal.Length 증가/감소 -> Sepal.Length 증가/감소 라는 가설을 세울 수 있다. 즉 Petal.Width와 Sepal.Length 사이에도 선형 관계가 존재할 것이라는 가설을 세울 수 있다. 이 가설에 대해 살펴보기 위해 위의 Correlation Plot에서 Petal.Width와 Sepal.Length에 관련된 그래프를 살펴보도록 한다.

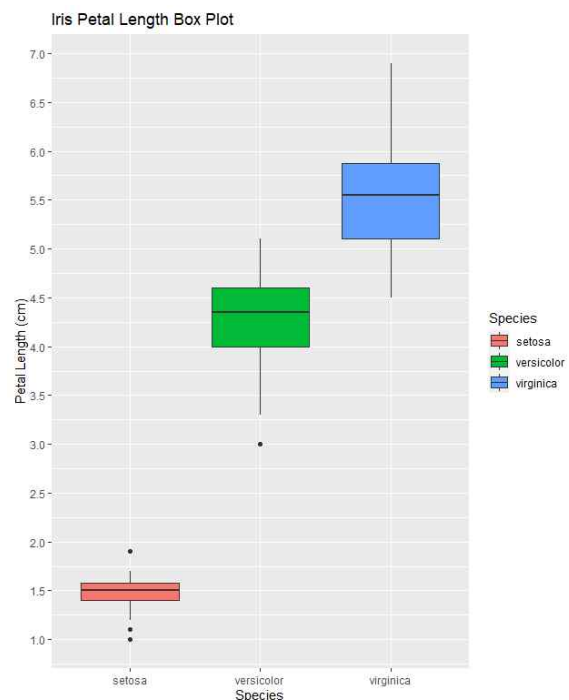
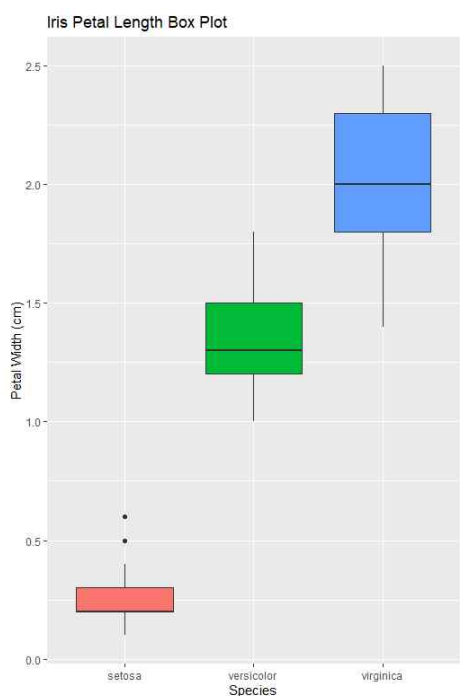
Correlation Plot에서 Petal.Width와 Sepal.Length에 관련된 그래프를 살펴보니 이 역시 선형관계가 나타나고 있음을 알 수 있다. 선형관계가 명확히 드러나는 타 그래프들과 다른 점은 중앙의 선형회귀선에서 값들이 Sepal.Length와 Petal.Length 그래프나 Sepal.Length와 Petal.Length 그래프에 비해 넓게 분포되어있다는 점이다.

위의 Correlation Plot에서 각 그래프들에 대한 Correlation을 살펴보았을 때 , Petal.Length와 Petal.Width 그래프 , Sepal.Length와 Petal.Length 그래프 , Petal.Width와 Sepal.Length 들의 Correlation은 각각 0.963 , 0.872 , 0.818을 가지고 있음을 알 수 있다. 이를 통해 우리는 위 요소들이 강한 양적 선형관계를 가진다는 것을 알 수 있다.

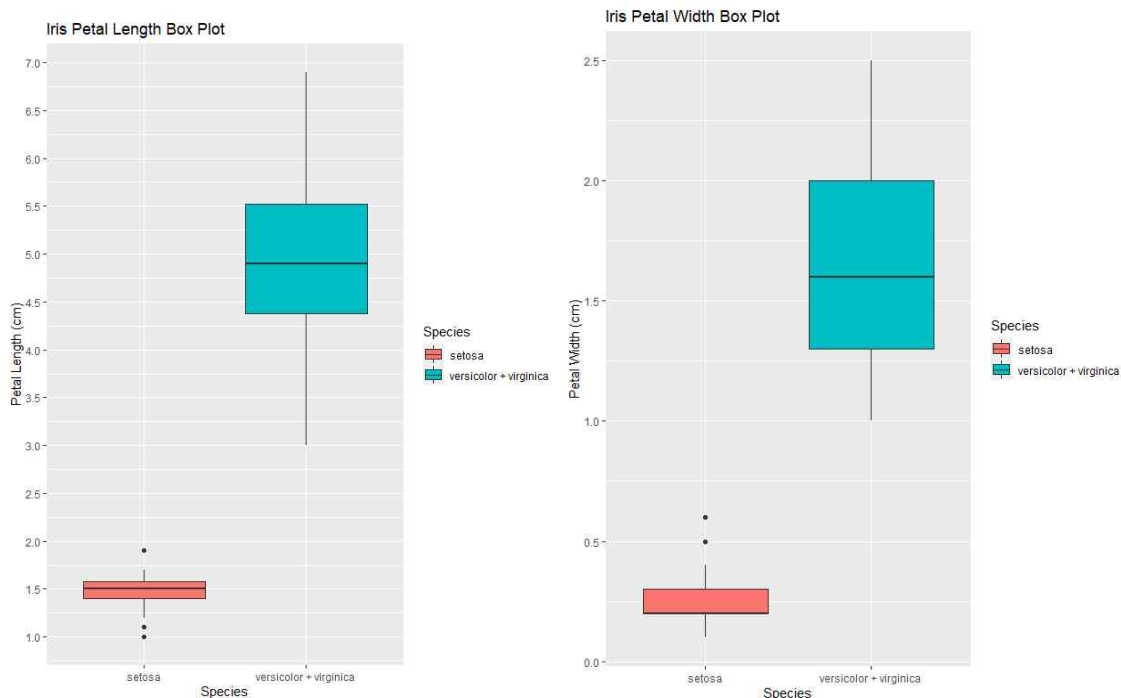
이외의 보다 낮은 값을 가지는 Correlation들은 -0.428 , -0.366 등으로 위에서 살펴본 경우와 다르게 보다 약한 음적 선형관계를 가지고 있음을 알 수 있다.

4-4. setosa, versicolor, virginica를 분류하는 시나리오

위의 내용에서 versicolor와 virginica를 Petal 요소로 구분할 수 있을 것이라는 가설이 있었다. 눈으로는 구별이 가능해보이지만, 실제 데이터들이 겹치는 부분이 존재하기 때문에 명확히 setosa 처럼 구분할 수 있어 보이지는 않는다. 구분이 실제로 가능할지, 혹은 대략적으로라도 가능할지 Box Plot을 통해 살펴보자.



Petal.Length와 Petal.Width 사이에는 강한 양적 상관관계가 있음을 확인했다. 그로부터 알 수 있듯이 위의 2개의 Petal.Length와 Petal.Width에 관한 Box Plot은 비슷한 모습을 보여주고 있다. 우리가 살펴보아야 할 점은 두 종을 이를 통해 확실히 구분할 수 있느냐는 점이다. 위의 두 Box Plot을 확인해보면 versicolor의 제 3사분위 이상의 값과 virginica의 제 1사분위 이하의 값이 겹치는 것을 확인 할 수 있다.



위의 Box Plot 그래프 들은 versicolor와 virginica를 하나의 종으로 합쳐서 Petal.Length와 Petal.Width를 살펴본 경우이다. 두 Box Plot 에서 볼 수 있듯이 , Petal.Length의 중앙값은 versicolor의 제 3사분위 이상의 값과 virginica의 제 1사분위 이하의 값을 알 수 있다. 또한 Petal.Width 역시 versicolor의 제 3사분위 이상의 값과 virginica의 제 1사분위 이하의 값을 알 수 있다. 이는 우리가 versicolor와 virginica를 Petal 요소와 Sepal 요소만을 가지고 분류할 때 , 어느정도의 대략적인 분류가 가능하다는 점을 보여준다.

종이 기록되지 않은 IRIS 데이터 셋이 주어지고 우리가 임의로 Petal 요소와 Sepal 요소의 특징을 보고 종을 예측한다고 생각해보자. 이때 사람의 노력으로 가능한 분류의 시나리오를 먼저 Petal.Length와 Petal.Width가 일정 값 이하인 것들을 모두 setosa로 분류할 것부터 시작할 것이다. 이후 setosa로 분류한 나머지 데이터들을 가지고 Petal.Length와 Petal.Width의 중앙값을 구한다. 이때 중앙값은 위에서 보았듯이 versicolor의 제 3사분위 이상의 값과 virginica의 제 1사분위 이하의 값을 알 수 있다. 이후 데이터들에 한해 중앙값 이상의 데이터들은 모두 virginica로 이하는 모두 versicolor로 분류하였다고 가정하자. 이러한 방법으로 분류하였을시 위 데이터에 한해서는 versicolor의 제 3사분위 이하와 virginica의 제 1사분위 이상의 경우는 대략적인 분류가 가능하다. 현재 이 시나리오에서 이렇게 대략적인 분류를 진행하는 것에는 근거가 있다. 바로 Petal 요소들이 강한 양의 상관관계를 가지고 있다는 점이다. 이는 Petal.Length 에서 중앙값 이상의 값을 가졌다면 Petal.Width 에서도 중앙값 이상의 값을 가질 가능성이 높다는 것을 의미한다.

물론 데이터셋의 크기가 더 커짐에 따라 더 정확한 분류 기준이 나타날 수도 혹은 완전 다른 분류 기준이 나타날 수도 , 혹은 아예 대략적인 분류도 어려워 질 수도 있다. 하지만 위의 데이터 셋을 가지고 판단한다면 , 위와 같은 방법으로 대략적인 분류가 가능하다는 것이다.

5. 결론

위의 IRIS 데이터 분석을 통해 크게 2가지의 결과를 얻을 수 있었다. 첫번째는 Petal.Width와 Petal.Length , Sepal.Length 사이에 강한 양적 상관관계가 존재한다는 것이다. 두번째로는 위의 데이터를 통해 컴퓨팅 능력을 이용하지 않은 사람의 대략적인 붓꽃 품종 분류 시나리오를 살펴보았다. 위의 데이터셋을 기반으로 한 이 가상의 시나리오는 먼저 Petal.Length와 Petal.Width가 일정 값 이하인 것들을 모두 setosa로 분류할 것이다. 이후 나머지 데이터에 대하여 중앙값을 찾고 중앙값 이상인 데이터들에 대해서는 virginica , 이하인 데이터들에 대해서는 versicolor 로 대략적으로 분류한다. 이렇게 대략적인 분류가 가능한 근거는 바로 Petal 요소들 사이에 강한 양적 상관관계가 존재하기 때문이다.