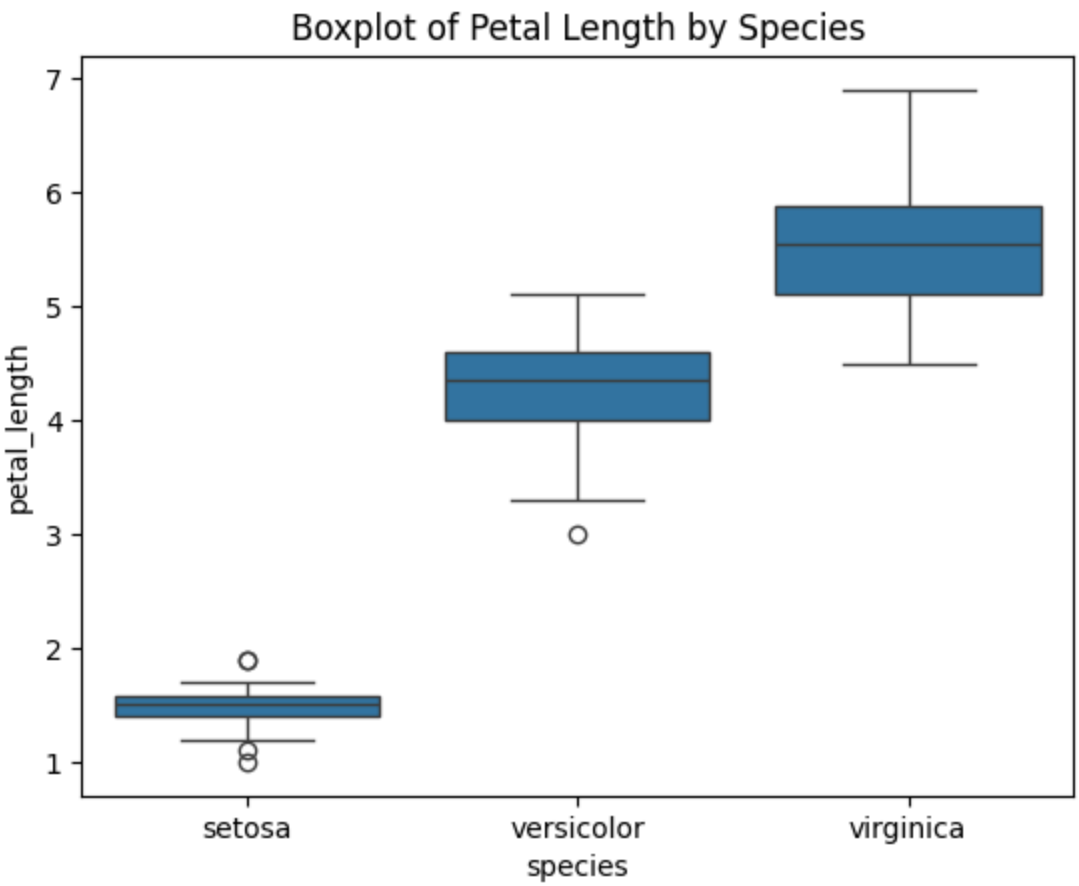
YBIGTA hw 3-1

데이터 파악:

기본정보:

Data columns (total 5 columns):

```
#   Column        Non-Null Count  Dtype
---
0   sepal_length  150 non-null    float64
1   sepal_width   150 non-null    float64
2   petal_length  150 non-null    float64
3   petal_width   150 non-null    float64
4   species       150 non-null    object
```



해석: 종류별 petal length의 평균값을 차이가 난다. Setosa, Versicolor Species중에 이상치도 있는것 같다. Variance within group으로 살펴보면, Setosa< Versicolor < Virginica를 관찰될 수 있다

각종 test

=== Shapiro-Wilk Test for 'petal length' ===
H0: Data follows Normal Distribution (p > 0.05)

Species	Statistic	P-Value	Result
setosa	0.9550	0.0548	Normal (H0 채택)
versicolor	0.9660	0.1585	Normal (H0 채택)
virginica	0.9622	0.1098	Normal (H0 채택)

결론: H0을 채택임으로, 데이터 'petal length'변수는 정규성을 있다

=== Levene Test for 'petal length' ===
H0: 각 그룹(Species) 간 분산은 동일하다.
Variable | Statistic | P-Value | Result

petal_length | 19.4803 | 0.0000 | Not Equal (H0 기각)
결론: H0을 기각함으로, 데이터 ‘petal length’변수가 그룹간의 분산성은 다른다, 앞 그림에서도 확인했다

=== ANOVA Hypothesis for 'petal length' ===

H0: 세 species 간 'petal_length'의 평균에는 차이가 없다. (모두 동일하다)
H1: 적어도 한 species의 'petal_length' 평균은 다른 species와 통계적으로 유의한 차이가 있다.
=== One-way ANOVA Results ===
Variable | F-Value | P-Value | Conclusion (alpha=0.05)
petal_length | 1180.1612 | 2.8568e-91 | H0 기각 (유의한 차이 있음)
결론: H0을 기각함으로, 데이터 ‘petal length’변수가 그룹간의 평균값도 다르다, 앞 그림에서도 확인했다

===Tukey HSD (Scipy) for 'petal length'===

Pairwise Group Comparisons (95.0% Confidence Interval)
Comparison Statistic p-value Lower CI Upper CI
(0 - 1)   -2.798   0.000   -3.002   -2.594
(0 - 2)   -4.090   0.000   -4.294   -3.886
(1 - 0)   2.798   0.000   2.594   3.002
(1 - 2)   -1.292   0.000   -1.496   -1.088
(2 - 0)   4.090   0.000   3.886   4.294
(2 - 1)   1.292   0.000   1.088   1.496
※ Group Index Mapping:
Group 0: setosa
Group 1: versicolor
Group 2: virginica

[Significant Differences (p < 0.05)]

- setosa vs versicolor (p=0.0000) → 차이 있음
- setosa vs virginica (p=0.0000) → 차이 있음
- versicolor vs virginica (p=0.0000) → 차이 있음

결론: 세 그룹 간 평균값은 거로 다르다

결과 요약 (Conclusion)

분석 결과, 세 Species 간의 Petal Length는 모두 통계적으로 유의미한 차이가 있습니다.

1. 크기 순서 평균값을 기준으로 꽃잎 길이는 다음 순서대로 깁니다.

Virginica > Versicolor > Setosa
(ingroup variance는 반대로)

2. 통계적 근거

ANOVA: P-value < 0.05로 귀무가설이 기각되어, 세 그룹 중 적어도 하나는 차이가 있음을 확인했습니다.

사후검정 (Tukey HSD): 모든 조합에서 유의미한 차이(Reject H0)가 확인되었습니다. 즉, 세 그룹은 서로 명확하게 구분되는 집단입니다.

Boxplot: 시각적으로도 세 그룹의 분포(Box)가 겹치지 않고 높이 차이가 뚜렷하게 나타납니다.

예측(회귀)

=== Regression Performance Metrics ===

MSE (Mean Squared Error): 0.1300 (낮을수록 좋음)
R2 Score (결정계수) : 0.9603 (높을수록 좋음)

=== Regression Coefficients ===

Intercept (절편): -0.2622
Feature Coefficient
0 sepal_length 0.722815
1 sepal_width -0.635816
2 petal_width 1.467524

결과 분석:

모델이 전체 데이터 변동성의 약 96%를 설명하고 있음을 의미한다. 아주 좋은 성능이다.
Petal length는 Sepal length, Petal width와 **양(+)의 상관관계**를 갖고 있으며, Sepal width와는 **음(-)의 상관관계가 있다.**