basic statistics

주제

1. Iris 의 종별 Petal Length 평균 차이 검정
2. Iris 의 Petal Length 회귀 예측 모델 구축

1. 데이터 로드 및 구조 확인

- 코드 결과

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
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 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

2. 기술통계량

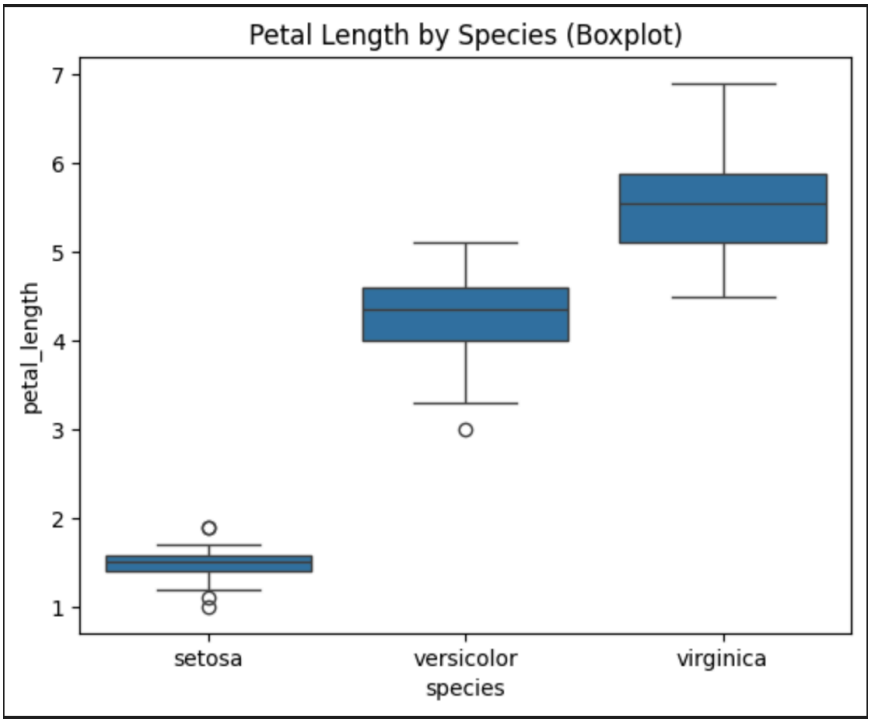
Species별 Petal Length의 평균, 개수, 표준편차, 최소/최대, 사분위수, 그룹별 데이터 개수

- 코드 결과

species	count	mean	std	min	25%	50%	75%	max
setosa	50.0	1.462	0.173664	1.0	1.4	1.50	1.575	1.9
versicolor	50.0	4.260	0.469911	3.0	4.0	4.35	4.600	5.1
virginica	50.0	5.552	0.551895	4.5	5.1	5.55	5.875	6.9

3. 시각화

- 코드 결과



- 해석

종 (species) 에 따라 꽃잎 길이 (petal length) 에 차이가 존재할 가능성을 시각적으로 확인할 수 있음.

세 종 간의 꽃잎의 길이의 분포가 잘 겹치지 않기 때문임.

각 종 별 꽃잎 길이의 분포를 보여줌.

setosa 는 꽃잎 길이가 전반적으로 가장 짧고, 분산도 작은 편이며 이상치가 존재함.

versicolor는 setosa 보다 꽃잎 길이가 뚜렷하게 크며, 약한 이상치가 존재함.

virginica는 세 종 중 꽃잎 길이가 제일 긴 경향이 있음. 중앙값과 범위도 가장 큼.

4. 정규성 검정 (Shapiro-Wilk)

- 코드 결과

	species	W-statistic	p-value
0	setosa	0.954977	0.054811
1	versicolor	0.966004	0.158478
2	virginica	0.962186	0.109775

- 해석

  Shpiro-Wilk 검정을 통해 species 별 petal_length 의 정규성을 검정했음.

  유의수준 0.05 하, 세 종 모두 p값이 0.05 보다 커서, 귀무가설 (정규분포를 따름) 을 기각하지 못함.

  세 종 모두 정규성을 만족한다는 가정 하에 이하 분석을 진행함. (ANOVA 가정 만족)

5. 등분산성 검정 (Levene)

- 코드 결과

  (np.float64(19.480338801923573), np.float64(3.1287566394085397e-08))

  - statistic : 19.48...
  - p-value = 3.13e-08
- 해석

  species 별 petal_length 의 등분산성을 검정함.

p-value = 3.13e-08 < 0.05 따라서 귀무가설 (세 그룹의 분산은 같음) 을 기각함.

세 종의 꽃잎 길이의 분산이 동일하지 않음.

하지만 과제 명세에 따라 등분산성을 만족한다는 가정 하에 이하 분석을 진행함. (ANOVA 가정 만족)

6. ANOVA 가설 수립

- 가설 수립

  귀무가설 (H0) :

  세 species (setosa, versicolor, virginica) 간 petal_length 평균은 모두 같음.

  대립가설 (H1):

  적어도 한 species는 petal_lengthd 의 평균이 다름.

7. One-way ANOVA

- 코드 결과

  (np.float64(1180.1611822529785), np.float64(2.8567766109619814e-91))

  - F statistic = 1180.16
  - p-value = 거의 0
- 해석

  귀무가설을 기각함. 세 종 간의 꽃잎 길이 차이가 통계적으로 유의미함.

8. 사후검정 (Tukey HSD)

- 코드 결과

```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1     group2   meandiff p-adj lower  upper  reject
-----------------------------------------------------
    setosa versicolor    2.798   0.0 2.5942 3.0018   True
    setosa  virginica     4.09   0.0 3.8862 4.2938   True
versicolor  virginica    1.292   0.0 1.0882 1.4958   True
-----------------------------------------------------
```

- 결과 해석

  one-way ANOVA 결과가 유의하여 Tukey HSD 사후검정을 실시함.

  사후검정 결과, 모든 종 쌍에서 p-adj <0.05 이어서, petal_length 평균 차이가 통계적으로 유의미함.

  또, meandiff 값을 보아, 꽃잎의 길이값은 setosa < versicolor < virginica 임을 알 수 있음.

9. 결과 요약

iris 데이터셋을 이용하여 종(species) 별 꽃잎 길이(petal_length) 평균 차이를 검정했음.

기술통계량과 박스플롯 시각화를 통해 species에 따라 petal_length 분포가 서로 다름을 시각적으로 확인함.

petal_length의 크기는
setosa < versicolor < virginica 순으로 증가하는 경향을 보임.

Shapiro-Wilk 검정 결과 세 종 모두 정규성을 만족하였으며,
Levene 검정 결과 등분산성은 만족하지 않았으나,
과제 명세에 따라 등분산성 만족한다는 가정 하에 이하 분석을 수행함.

One-way ANOVA 분석 결과,
세 종 간 petal_length 평균 차이는 통계적으로 유의미하였으며,
Tukey HSD 사후검정을 통해 모든 종 쌍 간 평균 차이가 유의함을 확인하였음.

10. 회귀분석

- 코드 결과

```
MSE: 0.13001626031382701
R^2: 0.9603293155857663

Intercept: -0.2621959025887084
```

	feature	coefficient
0	sepal_length	0.722815
1	sepal_width	-0.635816
2	petal_width	1.467524

- 결과 해석

  꽃잎 길이 (sepal_length) , 꽃받침 너비 (sepal_width), 꽃잎 너비 (petal_width) 를 독립변수로 하고, petal_length를 종속변수로 하는 선형 회귀분석을 수행함.

  분석 결과, 예측오차 (MSE) 는 0.13 으로 비교적 작게 나왔으며,

  R**2 값은 0.96으로 독립변수들이 petal_length 변동의 약 96%를 설명함.

  각 회귀계수 중에는, petal_length의 회귀계수가 가장 큰 값인 1.47 으로 가장 중요한 변수임을 알 수 있음.

  또한 sepal_width의 회귀계수는 -0.64 으로, 증가할수록 petal_length가 감소하는 경향을 보임.