[3회차] 기초통계 과제

iris 데이터셋을 활용해 다양한 통계 기법들을 실습해본 결과이다.

데이터 로드 및 구조 확인

```
[2]  iris.head()
     ✓ 0.0s
```

	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

- .head()를 활용해 데이터가 어떻게 구성되었는지 대략적으로 파악할 수 있다.

```
[3]  iris.info()
     ✓ 0.0s

<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
```

- .info()를 통해 열별 결측치 여부, 자료형 등을 파악할 수 있다.

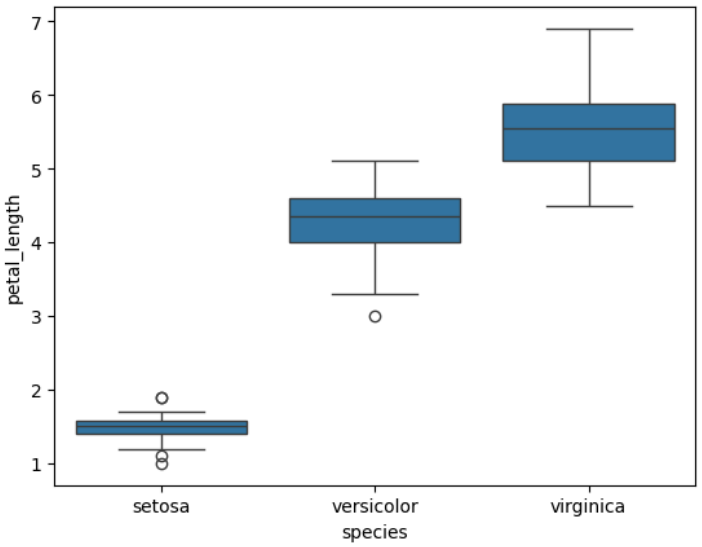
기술통계량

```
[4]  print(iris.groupby('species')['petal_length'].describe())
     ✓ 0.0s

             count   mean       std  min  25%   50%    75%  max
species
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
```

- .describe()를 통해 Petal Length의 평균, 개수, 표준편차, 최소/최대, 사분위수 등을 출력할 수 있다.

시각화



- Species별 Petal Length의 분포를 Boxplot으로 시각화 한 결과이다.
- setosa 종이 다른 종에 비해 petal_length가 짧고 분산이 더 적은 것을 확인할 수 있다.
- petal_length는 분포상으로 virginica, versicolor, setosa 순서대로 길다.
- setosa와 versicolor에서는 이상치를 확인할 수 있다.

정규성 검정 (Shapiro-Wilk)

가설

각 species들에 대해

- 귀무가설(H0): {species}의 petal_length 데이터는 정규분포를 따른다.
- 대립가설(H1): {species}의 petal_length 데이터는 정규분포를 따르지 않는다.

결과

[setosa] p-value: 0.0548
[versicolor] p-value: 0.1585
[virginica] p-value: 0.1098

- 분석 결과 모든 species들의 p값이 유의수준 0.05 이상이므로 petal_length는 정규분포를 따른다.

등분산성 검정 (Levene)

- 귀무가설(H0): 모든 species의 petal_length의 분산은 같다.
- 대립가설(H1): 적어도 하나의 species의 petal_length의 분산은 다르다.

p-value가 약 3.12 * 10^(-8)이므로 귀무가설을 기각, 적어도 한 species의 분산이 다르다고 할 수 있다.

ANOVA 가설 수립

- 귀무가설(H0): 세 species 간 petal_length의 평균은 모두 같다.
- 대립가설(H1): 적어도 하나의 species의 petal_length 평균은 다르다.

결과: ANOVA F-statistic: 1180.1612, p-value: 0.0000

- F-statistic 값이 약 1180으로 종별로 petal_length의 차이가 매우 크다는 것을 알 수 있다.
- p-value가 약 2.85 * 10^(-91)이므로 귀무가설을 기각, 적어도 한 species의 petal_length의 평균이 다르다고 할 수 있다.

사후검정 (Tukey HSD)

```
   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
-----------------------------------------------------
```

- 모든 species 쌍에 대해 reject 값이 True 이므로 세 species 간 모두 petal_length 평균 간 유의미한 차이가 있음을 알 수 있다.

결과 요약

- Boxplot, ANOVA, 사후검정 결과를 종합하였을 때
  petal_length는 virginica, versicolor, setosa 순서대로 통계적으로 유의하게 길다.

회귀 분석

입력값을 sepal_length, sepal_width, petal_width, 타겟값을 petal_length로 설정하고 회귀분석을 진행한 결과이다.

Mean Squared Error: 0.1300
R-squared: 0.9603

Coefficient
sepal_length 0.722815
sepal_width -0.635816
petal_width 1.467524

- MSE 값이 0.13으로 모델의 예측 오차가 상당히 작음을 알 수 있다.
- R-squared 값이 약 0.96이라는 것은 해당 모델이 전체 petal_length 데이터의 96% 이상을 설명하는 것을 의미한다.
- 각 회귀계수를 검토하였을 때, sepal_length와 petal_width는 petal_length와 양의 관계, sepal_width와 petal_length는 음의 관계에 있음을 알 수 있다.

- sepal_length의 회귀계수가 약 0.72라는 것은 sepal_length가 1cm 길어질 때 petal_length는 약 0.72cm 길어지는 경향이 있음을 의미한다.
- petal_width의 경우 회귀계수의 절댓값이 가장 높으므로 세 변수 중 petal_length에 미치는 영향이 가장 크다고 할 수 있다.