

기초통계 과제

1. 데이터 로드 및 구조 확인

데이터를 불러온 후, `info()` 와 `head()` 를 통해 데이터의 기본 정보와 구조를 확인하였다.

`iris.info()` 결과

```
<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
```

`print(iris.head())` 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

Iris 데이터셋은 총 150개의 관측치와 5개의 변수로 구성되어 있으며, 변수는 4개의 연속형 변수 (sepal_length, sepal_width, petal_length, petal_width)와 1개의 범주형 변수 (species)로 이루어져 있는 것을 두 결과를 통해 알 수 있다.

2. 기술통계량

▼ 코드

```
iris.groupby("species")["petal_length"].describe()
```

`describe()`를 통해 species별 petal length에 대한 기술통계량을 확인한 결과는 다음과 같다.

	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

setosa, versicolor, virginica 세 종 모두 각각 50개의 관측치를 가지고 있어 그룹 간 표본 크기는 동일하다. 평균 petal length는 setosa가 1.462로 가장 짧았으며, versicolor는 4.260, virginica는 5.552로 가장 길게 나타났다. 표준편차 역시 setosa가 가장 작았고, versicolor, virginica 순으로 커지는 것을 알 수 있었다.

사분위수를 살펴보면, setosa의 3사분위수와 최대값이 versicolor와 virginica의 1사분위수보다도 작아, setosa와 다른 두 종 사이의 petal length 분포가 거의 겹치지 않음을 유추할 수 있다. 또한 versicolor와 virginica를 비교했을 때, virginica의 사분위수 값들이 전반적으로 versicolor보다 크게 나타나, virginica가 더 긴 분포를 갖는 경향을 보였다.

그룹별 데이터 개수 확인은 다음과 같이 `value_counts()` 함수를 통해서도 확인할 수 있다.

▼ 코드

```
iris["species"].value_counts()
```

```
species
setosa      50
versicolor  50
virginica   50
Name: count, dtype: int64
```

마찬가지로 각 species별로 50개가 나오는 것을 확인할 수 있다.

3. Box Plot

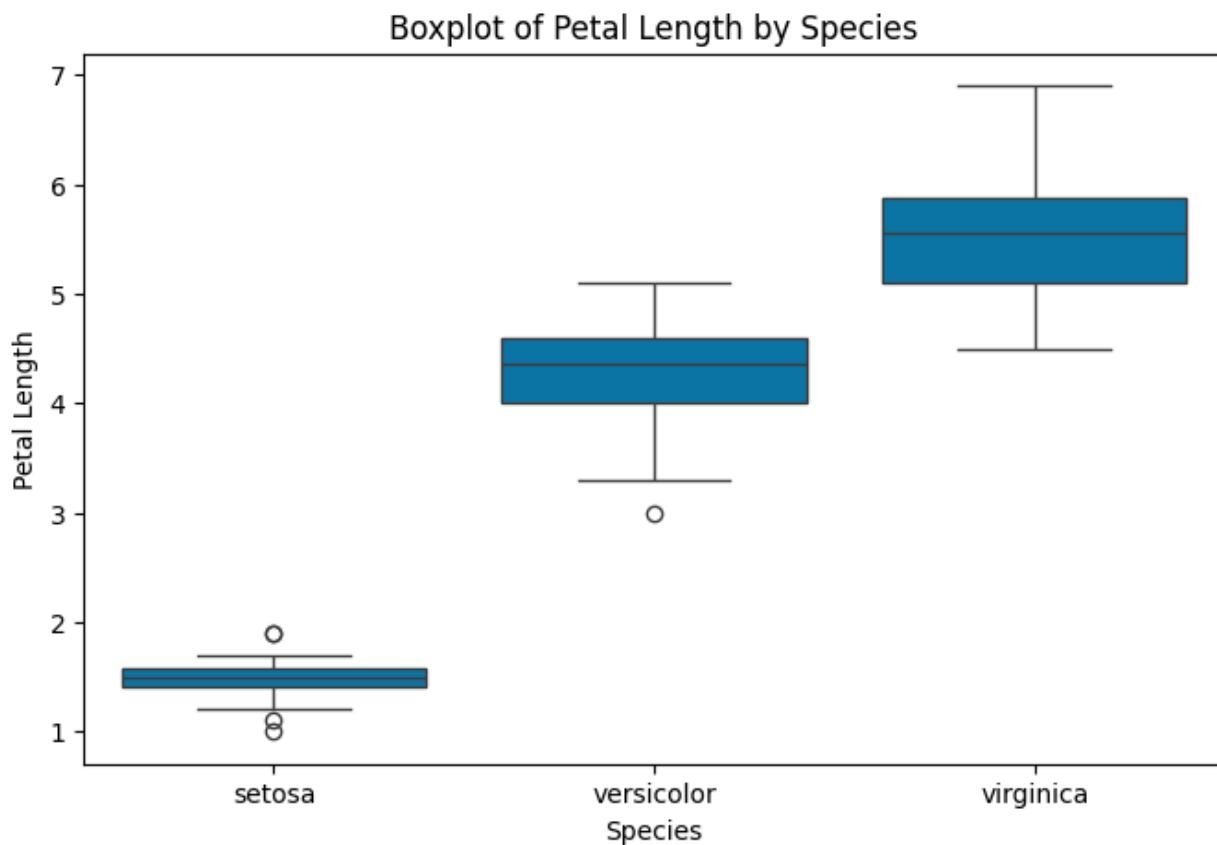
위에서 확인한 기술통계량을 box plot을 통해 시각화하여 보다 직관적으로 확인할 수 있다.

▼ 코드

```
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 5))
sns.boxplot(
    data=iris,
    x="species",
    y="petal_length"
)

plt.title("Boxplot of Petal Length by Species")
plt.xlabel("Species")
plt.ylabel("Petal Length")
plt.show()
```



앞서 확인한 것 처럼, species별 petal length 분포 차이가 나는 것을 확인할 수 있다. setosa는 매우 낮고 좁게 분포하여 petal length가 짧고 변동성이 작은 반면 virginica는 세 종 중 가장 높은 중앙값과 가장 큰 분포 범위를 보여주고 있다.

4. 정규성 검정 (Shapiro-Wilk)

Species별 Petal Length의 정규성을 검정하기 위해, 각 species에 대해 다음과 같은 가설을 설정할 수 있다.

- 귀무가설 H_0 : 해당 species의 petal length는 정규분포를 따른다.
- 대립가설 H_1 : 해당 species의 petal length는 정규분포를 따르지 않는다.

▼ 코드

```
from scipy.stats import shapiro

for species in iris["species"].unique():
    data = iris.loc[iris["species"] == species, "petal_len"]
```

```
gth"]
    stat, p_value = shapiro(data)
    print(f"{species}: p-value = {p_value:.5f}")
```

Shapiro-Wilk 정규성 검정을 진행한 결과는 다음과 같다.

```
setosa: p-value = 0.05481
versicolor: p-value = 0.15848
virginica: p-value = 0.10978
```

모든 species에 대하여 p-value가 유의수준 0.05보다 크게 나타났기 때문에 귀무가설을 기각하지 못한다. 즉, 모든 species의 petal length는 정규성을 위배한다고 볼 근거가 없다.

따라서 이후 분석에서는 각 species의 petal length가 정규분포를 따른다는 가정 하에 분석을 진행할 수 있다.

5. 등분산성 검정(Levene)

다음은 등분산성 검정이다. 등분산성 검정에 대한 가설은 다음과 같다.

- 귀무가설 (H_0): 세 species(setosa, versicolor, virginica)의 petal length 분산은 동일하다.
- 대립가설 (H_1): 적어도 한 species의 petal length 분산이 다르다.

▼ 코드

```
from scipy.stats import levene

setosa = iris.loc[iris["species"] == "setosa", "petal_length"]
versicolor = iris.loc[iris["species"] == "versicolor", "petal_length"]
virginica = iris.loc[iris["species"] == "virginica", "petal_length"]
```

```
stat, p_value = levene(setosa, versicolor, virginica)
print(p_value)
```

Levene 등분산성 검정을 진행한 결과 p-value가 3.129e-08로 유의수준 0.05보다 작기 때문에 귀무가설을 기각할 수 있다. 즉, 세 species 간 petal length의 분산은 동일하지 않다고 판단할 수 있다. 이는 box plot에서 species별로 변동 범위가 다르게 나타난 것과 같이, species 간 변동성 차이가 존재함을 시각적으로 확인한 결과와도 일치한다.

6. ANOVA 가설 수립

ANOVA는 세 개 이상의 집단 간 평균 차이를 검정하기 위한 통계적 방법이다.

ANOVA의 기본 가정은 독립성, 정규성, 등분산성이다. 앞서 수행한 정규성 검정에서는 각 species의 petal length가 정규성을 위배한다고 볼 근거가 없음을 확인하였다. 반면 Levene 등분산성 검정 결과, 등분산성 가정은 충족되지 않는 것으로 나타났다. 다만 본 과제에서는 등분산성을 만족한다고 가정하여 ANOVA를 실시한다.

Species별 petal length 평균 차이를 검정하기 위한 가설은 다음과 같다.

- 귀무가설 (H_0): 세 species 간 petal length 평균은 모두 같다. $\mu_{\text{setosa}} = \mu_{\text{versicolor}} = \mu_{\text{virginica}}$
- 대립가설 (H_1): 적어도 하나의 species의 petal length 평균은 다르다.

7. One-way ANOVA

One-way ANOVA를 실시하여 species별 petal length 평균 차이를 검정하였다.

▼ 코드

```
from scipy.stats import f_oneway

f_stat, p_value = f_oneway(setosa, versicolor, virginica)

print(f"F-statistic: {f_stat:.4f}")
print(f"p-value: {p_value:.5e}")
```

검정 결과는 다음과 같다.

```
F-statistic: 1180.1612  
p-value: 2.85678e-91
```

F 통계량 값은 1180.1612, p-value는 2.85678e-91이다.

p-value를 기준으로 살펴보면 p-value가 0.05보다 작기 때문에 귀무가설을 기각할 수 있다. 따라서, species에 따라 petal length 평균에는 통계적으로 유의한 차이가 존재한다고 판단할 수 있다.

8. 사후 검정 (Tukey-HSD)

ANOVA 결과를 통해 세 species의 petal length 평균이 같지 않음을 확인하였지만, 이 결과만으로는 어떤 그룹끼리 차이가 나는지 알 수 없다. 따라서, 이를 확인하기 위해 Tukey 사후 검정을 진행할 수 있다.

먼저, 가설은 다음과 같다.

- 귀무가설 (H_0): 비교되는 두 species 간 petal length 평균의 차이는 0이다. $\mu_a - \mu_b = 0$
- 대립가설 (H_1): 비교되는 두 species 간 petal length 평균의 차이는 0이 아니다. $\mu_a - \mu_b \neq 0$

Tukey HSD 사후검정에서는 위 가설을 각 species 쌍 (setosa-versicolor, setosa-virginica, versicolor-virginica)에 대해 개별적으로 검정한다.

▼ 코드

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd  
tukey = pairwise_tukeyhsd(  
    endog=iris["petal_length"],  
    groups=iris["species"],  
    alpha=0.05  
)  
print(tukey)
```

검정 결과는 다음과 같이 나왔다.

```
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

세 쌍의 검정 결과 모두에서 p-value가 유의수준 0.05보다 작고 reject=True가 나왔기 때문에 귀무가설을 기각할 수 있다. 즉, 모든 species 쌍에서 평균 차이가 통계적으로 유의한 것을 알 수 있다.

9. 결과 요약

Boxplot 시각화 결과, setosa, versicolor, virginica 순으로 petal length가 증가하는 경향이 관찰되었으며, 특히 setosa는 다른 두 종과 분포가 거의 겹치지 않으며 명확하게 다른 분포를 띠는 것을 확인할 수 있었다. One-way ANOVA 분석 결과, p-value가 0.05보다 작으면서 species에 따른 petal length 평균 차이가 유의한 것을 통계적으로 확인할 수 있었다. 이후 실시한 Tukey HSD 사후검정 결과, setosa-versicolor, setosa-virginica, versicolor-virginica 모든 species 쌍에서 petal length 평균 차이가 유의하게 나타났다. 종합하면, virginica의 petal length가 가장 길고, setosa가 가장 짧으며, versicolor는 그 중간에 위치하고, 이는 통계적으로 차이가 나는 것을 확인할 수 있었다.

10. 회귀 분석

마지막으로 sepal_length, sepal_width, petal_width를 설명변수로 사용하여 petal_length를 예측하는 선형회귀분석을 수행하였다. 데이터는 train/test 비율을 7:3으로 분리하였고, 학습된 모델을 통해 예측 성능을 평가하였다.

▼ 코드

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

```

from sklearn.metrics import mean_squared_error, r2_score

X = iris[["sepal_length", "sepal_width", "petal_width"]]
y = iris["petal_length"]

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("MSE:", mse)
print("R^2:", r2)

coef_df = pd.DataFrame({
    "Variable": X.columns,
    "Coefficient": model.coef_
})

print(coef_df)
print("Intercept:", model.intercept_)

```

```

MSE: 0.10913071951125887
R^2: 0.967636014590789
      Variable  Coefficient
0  sepal_length      0.745236
1  sepal_width       -0.651708

```

```
2    petal_width      1.453794  
Intercept: -0.32381864485354317
```

회귀 분석 결과, 테스트 데이터 기준 MSE는 0.1091로 나타나 예측값과 실제값 간의 평균 제곱 오차가 비교적 작음을 확인하였다. 또한 결정계수 R^2 는 0.9676으로, 본 회귀모형이 Petal Length 변동성의 약 96.8%를 설명하는 것으로 나타나 높은 설명력을 보였다.

회귀계수를 살펴보면, petal_width의 회귀계수는 1.4538로 가장 크게 나타나, 다른 변수들이 일정할 때 petal_width가 증가할수록 petal_length가 크게 증가하는 경향이 있음을 알 수 있다. sepal_length의 회귀계수는 0.7452로, 마찬가지로 petal_length와 양의 선형 관계가 있음을 확인하였다. 반면 sepal_width의 회귀계수는 -0.6517로 음의 값을 보여, 다른 변수들이 동일할 경우 sepal_width가 증가하면 petal_length는 감소하는 경향을 보였다.

절편(intercept)은 -0.3238로, 모든 설명변수가 0일 때의 예측값을 의미하나 실제로 petal_length가 음수가 될 수는 없기 때문에 해석상의 의미는 제한적이다.