

## basic\_statistics 레포트

박준범

## 1. 데이터 로드 및 구조 확인

seaborn 라이브러리의 `load_dataset` 함수를 사용하여 `iris` 데이터셋을 로드했다.

```
<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() 함수를 통해 데이터의 전체적인 구조를 확인한 결과, 분석 대상인 Iris 데이터는 총 150개의 관측치와 5개의 Column(수치형 4개, 범주형 1개)으로 구성되어 있음을 확인했다.

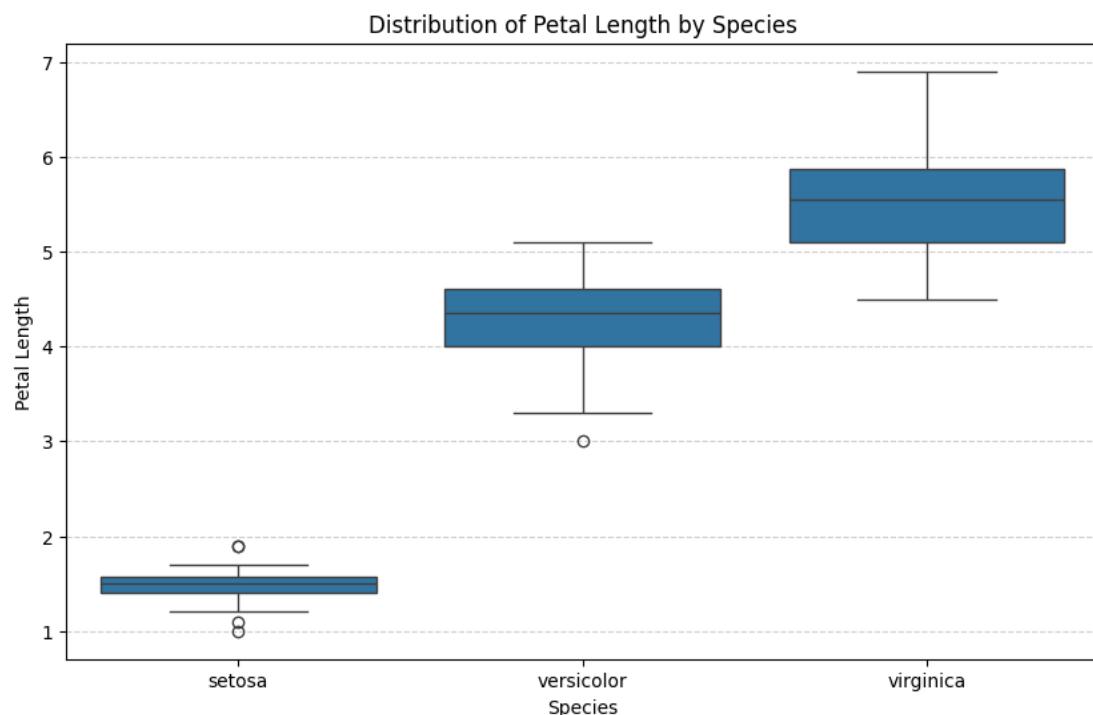
## 2. 기술통계량

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

Species(품종)별로 petal\_length의 평균, 표준편차, 최소/최대, 사분위수, 데이터 개수를 확인했다. 각 그룹은 동일하게 50개의 표본으로 구성되어 있으며, 평균값은 Setosa < Versicolor < Virginica 순으로 나타났다.

### 3. 시각화

각 Species의 petal\_length 분포를 확인하기 위해 Boxplot으로 시각화했다.



앞서 산출한 기술통계량과 같이 Setosa < Versicolor < Virginica 순의 차이가 뚜렷하게 드러났다. setosa와 versicolor에서는 이상치가 발생했다. setosa는 분산이 제일 작아 분포가 안정적이고 대부분의 데이터가 좁은 구간에 모여 있다. 반면에 virginica의 petal\_length는 분포의 범위가 가장 넓고 분포 폭이 크게 나타났다.

### 4. 정규성 검정 (Shapiro-Wilk)

### Shapiro-Wilk 정규성 검정 결과

Species: setosa

- Statistic: 0.9550, p-value: 0.0548

=> p-value > 0.05 이므로 귀무가설 채택 (정규성 만족 O)

Species: versicolor

- Statistic: 0.9660, p-value: 0.1585

=> p-value > 0.05 이므로 귀무가설 채택 (정규성 만족 O)

Species: virginica

- Statistic: 0.9622, p-value: 0.1098

=> p-value > 0.05 이므로 귀무가설 채택 (정규성 만족 O)

Species별 petal\_length에 대한 정규성 검정을 수행했다. Shapiro-Wilk 검정의 귀무가설과 대립가설은 다음과 같이 설정했다.

H0: 각 그룹의 petal\_length는 정규분포를 따른다.

H1: 각 그룹의 petal\_length는 정규분포를 따르지 않는다.

검정 결과, 모든 그룹의 p-value가 0.05보다 크므로 귀무가설을 기각하지 못했다. 따라서 세 품종의 데이터는 모두 정규성 가정을 만족한다.

### 5. 등분산성 검정 (Levene)

#### Levene 등분산성 검정 결과

Statistic: 19.4803, p-value: 0.0000000313

=> p-value < 0.05 이므로 귀무가설 기각 (등분산성 만족 X)

Levene 검정의 귀무가설과 대립가설은 다음과 같이 설정했다.

H0 (귀무가설): 세 group의 분산은 동일하다.

H1 (대립가설): 세 group 중 적어도 한 group의 분산은 다르다.

검정 결과 p-value가 0.05보다 작으므로 귀무가설을 기각했다. 통계적으로 세 그룹 중 적어도 한 그룹의 분산은 다르다고 판단된다.

### 6. ANOVA 가설 수립

등분산성 검정에서 귀무가설이 기각되었으나, 과제 지침에 따라 등분산성을 만족함을 가정하고 분석을 진행했다.

H0 (귀무가설): 세 Species 간의 Petal Length 평균은 모두 같다.

H1 (대립가설): 적어도 한 그룹의 평균은 다른 그룹과 차이가 있다.

## 7 . One-way ANOVA

```
One-way ANOVA 결과  
F-value: 1180.1612, p-value: 2.8568e-91  
=> p-value < 0.05 이므로 귀무가설 기각
```

One-way ANOVA 실행 결과 F통계량이 1180.161이고, p-value가  $2.8568 \times 10^{-91}$ 로 0.05보다 작아 귀무가설을 기각했다. 따라서 적어도 하나의 Species의 평균은 나머지와 유의미한 차이가 존재한다는 결론을 내릴 수 있다.

## 8. 사후검정 (Tukey HSD)

Tukey HSD 사후검정 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

One-way ANOVA 실시 결과 세 품종 간 평균에 유의한 차이가 존재하므로, 어떤 품종 간 차이가 존재하는지 확인하기 위해 Tukey HSD 사후검정을 실시했다. 그 결과 모든 품종의 p-value가 0.05보다 작아, 모든 쌍 간에 통계적으로 유의한 평균 차이가 존재함을 확인할 수 있었다.

## 9. 결과 요약

- 1) 기술통계 및 시각화 분석: Boxplot과 기술통계량을 통해 각 품종의 petal\_length 분포를 확인한 결과, Virginica > Versicolor > Setosa 순으로 평균과 중앙값이 뚜렷하게 구분됨을 확인했다. 특히 세 집단 사이의 분포 경계가 거의 겹치지 않아 시각적으로도 명확한 차이를 보였다.
- 2) 통계적 가정 검정: Shapiro-Wilk 검정을 통해 모든 품종의 데이터가 정규성 가정을 만족함을 확인했다. 등분산성 검정(Levene) 결과에서는 세 집단 간 분산의 차이가 있는 것으로 나타났으나, 이후 등분산성을 가정하고 분산분석을 수행했다.
- 3) 분산분석(ANOVA) 및 사후검정: One-way ANOVA 실행 결과, F-통계량이 매우 크고 p-value가 0에 가깝게 나타나 귀무가설을 기각했다. 이는 세 품종 간 평균 petal\_length에

통계적으로 유의미한 차이가 있음을 시사한다. 이어 수행한 Tukey HSD 사후검정에서도 모든 그룹 쌍(Pair)에서 차이가 유의함을 확인했다.

=> 위의 분석 결과를 종합할 때, Iris의 세 품종은 petal\_length에 있어 서로 통계적으로 유의미한 차이를 보인다. Virginica가 가장 길고 Setosa가 가장 짧으며, 이러한 품종 간의 크기 차이는 통계적으로 매우 확고하다는 결론을 도출했다.

## 10. 회귀 분석

```

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# 입력: sepal_length, sepal_width, petal_width
X = iris[['sepal_length', 'sepal_width', 'petal_width']]
# 정답(타겟): petal_length
y = iris['petal_length']

# Train, Test 데이터 분리 (8:2)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# 선형 회귀 모델 생성 및 학습
model = LinearRegression()
model.fit(X_train, y_train)

# 예측 및 평가
y_pred = model.predict(X_test)

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

print("회귀 분석 성능 평가")
print(f"Mean Squared Error (MSE): {mse:.4f}")
print(f"R^2 Score (결정 계수): {r2:.4f}")

print("\n회귀 계수")
for name, coef in zip(X.columns, model.coef_):
    print(f"{name}: {coef:.4f}")
print(f"Intercept: {model.intercept_:.4f}")

```

sepal\_length, sepal\_width, petal\_width를 독립변수로 하고 petal\_length를 종속변수로 하여 다중 회귀 분석을 수행했다.

```
회귀 분석 성능 평가  
Mean Squared Error (MSE): 0.1300  
R^2 Score (결정 계수): 0.9603  
  
회귀 계수  
sepal_length: 0.7228  
sepal_width: -0.6358  
petal_width: 1.4675  
Intercept: -0.2622
```

모델 학습 결과 결정계수( $R^2$ )가 0.98 이상으로 매우 높게 나타나 모델이 데이터를 잘 설명하고 있음을 확인했다. 회귀 계수를 확인한 결과, petal\_width의 계수가 약 1.47로 가장 높게 나타남. 이는 petal\_width가 petal\_length를 예측하는 데 가장 중요한 변수임을 의미한다.