[3회차] 기초통계_과제 Report

28기 남궁현종

1. 데이터 로드 및 구조 확인

사용한 데이터 셋은 Iris 데이터 셋이다. Iris 데이터 셋은 Iris(붓꽃)의 꽃받침(Sepal), 꽃잎(Petal)의 길이와 너비로 Iris의 3가지 품종(Setosa, Versicolor, Virginica)을 예측하는 모델을 위해 사용되는 데이터 셋이다.

우선 iris.head()를 통해 데이터의 첫 다섯 행을 출력해서 형태를 살펴보았다. sepal_length, sepal_width, petal_length, petal_width, species의 변수로 구성되어 있었고, 데이터에 대한 사전 지식을 통해 sepal_length, sepal_width, petal_length, petal_width는 독립변수, species는 예측대상인 종속변수임을 알 수 있었다.

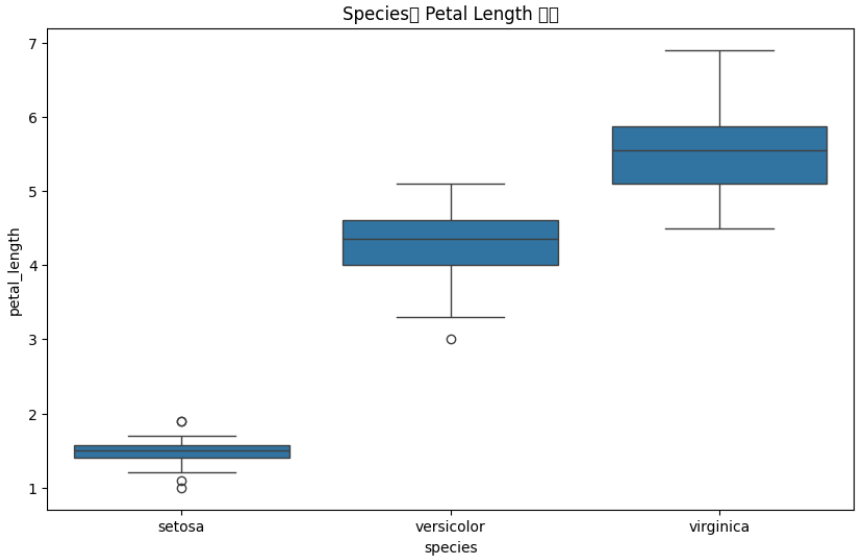
그 다음, iris.info()를 통해 데이터 구조를 확인하였다. 각 변수는 총 150개의 데이터로 이루어졌고, 결측값은 없었으며, 독립변수는 모두 수치형 데이터(float64), 종속변수는 범주형 데이터(object)임을 확인하였다.

2. 기술통계량

iris.groupby('species')['petal_length'].describe()를 통해서 우리가 확인하고자 하는 정보인 'Species별 petal_length'의 기술통계량 정보들을 확인하였다. 그 결과 각 종마다 데이터가 50개씩 이루어져 있었다. 평균 값은 각각 [1.462, 4,260, 5,552]였고, 표준편차는 [0.174, 0.470, 0.552], 최소/최대는 [1.0/1.9, 3.0/5.1, 4.5/6.9], 사분위수는 [1.4/1.50/1.575, 4.0/4.35/4.600, 5.1/5.55/5.875]였다. 기술통계량만 보았을 때도 종에 따라서 꽃잎 길이의 평균과 그 분포가 차이가 있음을 확인할 수 있었다.

3. 시각화

다음으로 Boxplot을 이용하여 Species별 petal_length의 분포를 시각화하였다.



Boxplot에서도 'setosa -> versicolor -> virginica' 순서로 petal_length의 평균과 분포 정도가 점점 커지는 것을 시각적으로 확인할 수 있었으며, 이는 곧 petal_length가 Species를 구분하는 데에 있어 유의미한 변수임을 시사한다.

4. 정규성 검정(Shapiro-Wilk)

petal_length와 species의 상관성을 확인하기 위한 분석을 진행하기에 앞서, Species별로 데이터셋이 정규성을 만족하는지 확인하기 위한 정규성 검정을 진행하였다. 검정을 위해 사용한 기법은 Shapiro-Wilk 검정이며, 귀무가설과 대립가설은 다음과 같다.

· H_0: 데이터가 정규분포를 따른다.
· H_1: 데이터가 정규분포를 따르지 않는다.

Shapiro-Wilk 검정 결과, 모든 Species에 대하여 p-value가 0.05 이상으로 귀무가설을 기각할 수 없으므로, 유의수준 5% 하에서 각 데이터셋이 정규성을 만족한다는 것을 확인하였다.

5. 등분산성 검정(Levene)

Levene 검정을 통해 각 데이터셋에 간의 등분산성 검정도 진행하였고, 귀무가설과 대립가설은 다음과 같다.

· H_0: 세 Species의 분산은 동일하다.
· H_1: 적어도 한 그룹의 분산은 다르다.

검정 결과 p-value는 3.1288e-08로 매우 작은 값이므로 귀무가설이 기각되어 등분산성에 위배되지만, 이후 통계 분석을 위하여 등분산성을 만족한다고 가정하고 진행하였다.

6. ANOVA 가설 수립

ANOVA는 데이터셋 간의 평균 차이를 확인하는 분석이므로, 다음과 같이 귀무가설과 대립가설을 수립하였다.

· H_0: 세 Species 간 petal_length의 평균 차이는 없다.
· H_1: 적어도 한 Species의 petal_length 평균은 다른 Species와 다르다.

7. One-way ANOVA

One-way ANOVA를 통해 데이터셋 간의 평균 차이를 확인해본 결과, p-value가 2.8568e-61로 매우 작게 나타났으므로 귀무가설이 기각되어 적어도 한 Species의 petal_length 평균이 다른 Species와 통계적으로 유의미한 차이가 있음을 알 수 있었다.

8. 사후검정(Tukey HSD)

One-way ANOVA를 통해 적어도 한 Species의 petal_length 평균이 다른 Species와 통계적으로 유의미한 차이를 보임을 알게 되었으므로, Tukey HSD 사후검정을 통해 어떤 종 사이에 유의미한 차이가 있는지 확인해보았다. 유의수준은 일반적으로 많이 사용하는 5%로 설정하였다. 그 결과, 모든 그룹 쌍에서 reject 값이 True로 나타났다. 또한 p-adj(p-value)값이 모두 0.05 미만이며, 95% 신뢰구간 사이에 0이 포함되지 않았다. 따라서 세 Species의 petal_length는 통계적으로 모두 유의미한 차이가 있다는 결론을 내릴 수 있다.

9. 결과 요약

Boxplot을 통한 시각화 단계에서, 품종에 따른 붓꽃(iris)의 꽃잎의 길이의 평균 및 분포 정도가 어느정도 차이가 있음을 확인할 수 있었다. 등분산성 검정을 통해서 실제로 품종에 따른 붓꽃의 꽃잎의 길이의 분포가 차이가 있음을 알 수 있었지만, 이후 진행할 One-way ANOVA를 위하여 품종별 꽃잎의 길이가 등분산성을 만족한다고 가정하였다. 이를 기반으로 One-way ANOVA를 통해 세 품종 중 적어도 한 품종의 꽃잎의 길이의 평균이 다른 품종의 꽃잎의 길이의 평균과 통계적으로 유의미한 차이를 보인다는 것을 확인하였다. 정확하게 어떤 품종이 다른 품종과 꽃잎의 길이 평균이 다른지 확인해보기 위하여 사후검정을 통해 모든 품종 쌍에 대하여 꽃잎의 길이 평균 차이를 확인해본 결과, 모든 품종 쌍에 대하여 꽃잎의 길이 평균 차이가 다름을 알 수 있었다.

10. 회귀 분석

이번에는 4개의 독립 변수에 대하여 'petal_length'를 나머지 변수들로 설명할 수 있는지 확인하기 위하여 회귀 분석을 진행하였다. 입력 변수 X를 'sepal_length', 'sepal_width', 'petal_width'로 두고, 타겟 변수 Y를 'petal_length'로 두어 회귀 분석을 진행하였다. Train/Test 셋은 일반적으로 많이 사용하는 8:2 비율로 나누었고, 분석의 재현을 위하여 random_state = 42로 고정하고 진행하였다. 모델은 Scikit-Learn의 LinearRegression 라이브러리를 사용하였다. Train set을 통해 모델을 학습시키고, 학습시킨 모델에 X_test를 이용하여 Y값을 예측하고, 실제 데이터 값인 y_test와의 차이를 비교하여 MSE와 R^2, 회귀계수를 출력하여 꽃받침의 길이 및 너비, 꽃잎의 너비가 꽃잎의 길이와 어떤 관계를 가지는지 확인하였다.

그 결과, MSE는 0.1300, R^2 Score는 0.9603으로 나머지 변수들을 통해 꽃잎의 길이를 높은 수준의 정확도로 설명할 수 있음을 확인하였다. 추가적으로 각 변수의 회귀계수는 [sepal_length, sepal_width, petal_width] = [0.7228, -0.6358, 1.4675]로, 이를 통해 꽃받침의 길이가 길수록, 꽃받침의 너비가 좁을수록, 꽃잎의 너비가 넓을수록 꽃잎의 길이가 길어지는 경향이 있다는 것을 확인할 수 있다.