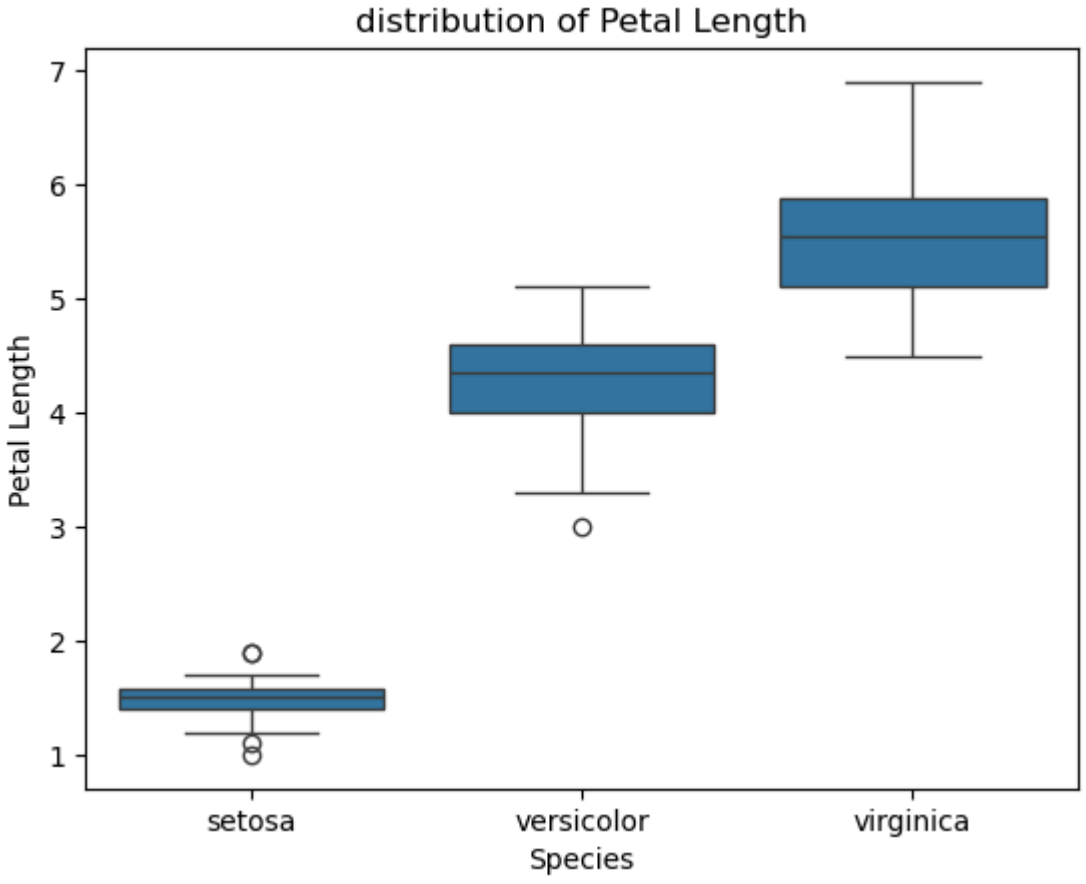
Iris 데이터셋을 이용한 꽃잎 길이 통계 분석 및 예측

분석 개요

- 목적: Iris 품종별 꽃잎 길이의 통계적 차이 규명과 예측 모델 수립
- 데이터셋: Iris dataset
    - 150행, 5개의 컬럼
        - 꽃받침 너비
        - 꽃받침 길이
        - 꽃잎 너비
        - 꽃잎 길이
        - 종

품종별 데이터 분포 시각화



- setosa: 짧은 길이, 작은 분산. 하지만 이상치 존재
- versicolor: 중간 길이, 중간 분산
- virginica: 긴 길이, 큰 분산. 하지만 이상치가 없음

통계적 전제조건 검정

정규성 검정

- Shapiro-Wilk Test
  - H_0 : 데이터가 정규분포를 따른다.
  - H_1 : 데이터가 정규분포를 따르지 않는다.
  - $p_{setosa} = 0.0548, p_{versicolor} = 0.1585, p_{virginica} = 0.1098$
  - 세 품종 모두 귀무가설을 채택하고, 데이터가 정규분포를 따른다고 가정한다.

등분산성 검정

- Levene Test
  - H_0 : 그룹 간 분산이 같다
  - H_1 : 적어도 하나의 그룹이 다른 그룹과 분산이 다르다
  - $p = 3.1288 * 10^{-8}$
  - p-value < 0.05로 등분산성을 만족하지 않는다.

품종간 평균 차이 분석

분산분석

- ANOVA
  - H_0 : 세 species 간 petal_length의 평균이 같다.
  - H_1 : 적어도 하나의 그룹이 다른 그룹과 분산이 다르다
  - $p = 2.8568 * 10^{-91}$
  - p-value < 0.05로 $H1$ 채택

사후검정

- Tukey HSD
  - $p_{0-1} < 0.05, p_{0-2} < 0.05, p_{1-2} < 0.05$
  - 세 그룹 간 Tukey HSD 수행 결과, 모든 관계에서 p-value < 0.05인것으로 볼때 각 종 사이의 꽃잎 길이는 통계적으로 유의미한 차이가 있다.

회귀분석을 통한 꽃잎 길이 예측

분석 설정

- 독립변수: Sepal Length, Sepal Width, Petal Width
- 종속변수: Petal Length

모델 성능 및 회귀 계수

- MSE : 0.1300
- R^2 : 0.9603
- $y = 0.7228x_{sl} - 0.6358x_{sw} + 1.4675x_{pw} + \beta_0$

- Petal_width가 꽃잎 길이에 가장 큰 영향을 미침

최종 결론

1. Iris 품종은 꽃잎 길이에 따라 명확히 구분되는 통계적 특성을 가짐.
2. ANOVA 및 사후검정을 통해 세 품종 간의 길이 차이가 유의미함을 증명함.
3. 선형 회귀 모델을 통해 타 변수들로 꽃잎 길이를 96%의 높은 정확도로 예측할 수 있음을 확인함.