Report

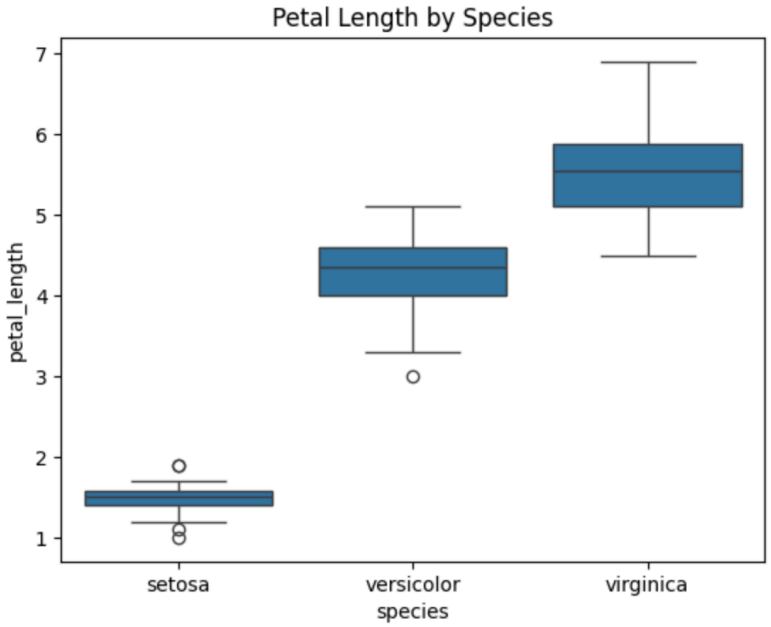
엄준서

Part 1. Iris 데이터셋 기반 기초 통계 분석

통계량

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

petal length boxplot



정규성 검증

- H0: 각 그룹은 정규분포를 따른다
- H1: 정규분포를 따르지 않는다

setosa 정규성 p-value: 0.0548
versicolor 정규성 p-value: 0.1585
virginica 정규성 p-value: 0.1098

→ 모두 유의수준 0.05보다 크므로 **정규성 만족**

등분산성 검정

- H0: 등분산성 만족
- H1: 등분산성 불만족

p-value: 0.0000000313

→등분산성 **불만족**

ANOVA

- H0: 세 종의 평균 Petal Length는 같다
- H1: 적어도 하나의 그룹 평균은 다르다

F=1180.16, p=0.000

→ 유의수준 0.05에서 귀무가설 기각.

차이 있음

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

세 종 간 Petal Length는 통계적으로 유의한 차이가 있으며,

Virginica > Versicolor > Setosa 순으로 길다.

Part 2. 신용카드 사기 거래 탐지

원본 Class 분포

Class
0(정상) 284315
1(사기) 492

분할된 데이터의 Class 비율

Train class ratio: Class
0 0.953056
1 0.046944

SMOTE 적용

사기 거래 수 적어 성능 저하 → SMOTE로 오버샘플링

✅ SMOTE 적용 전 클래스 분포 (학습 데이터 기준):
Class
0 7999
1 394
Name: count, dtype: int64

✅ SMOTE 적용 후 클래스 분포:
Class
0 7999
1 7999

모델 학습 및 결과

```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      2001
           1       0.95      0.89      0.92        98

    accuracy                           0.99      2099
   macro avg       0.97      0.94      0.96      2099
weighted avg       0.99      0.99      0.99      2099

PR-AUC: 0.9537
```

목표 Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90 모두 **달성**