1. Iris 데이터셋 기반 기초 통계 분석

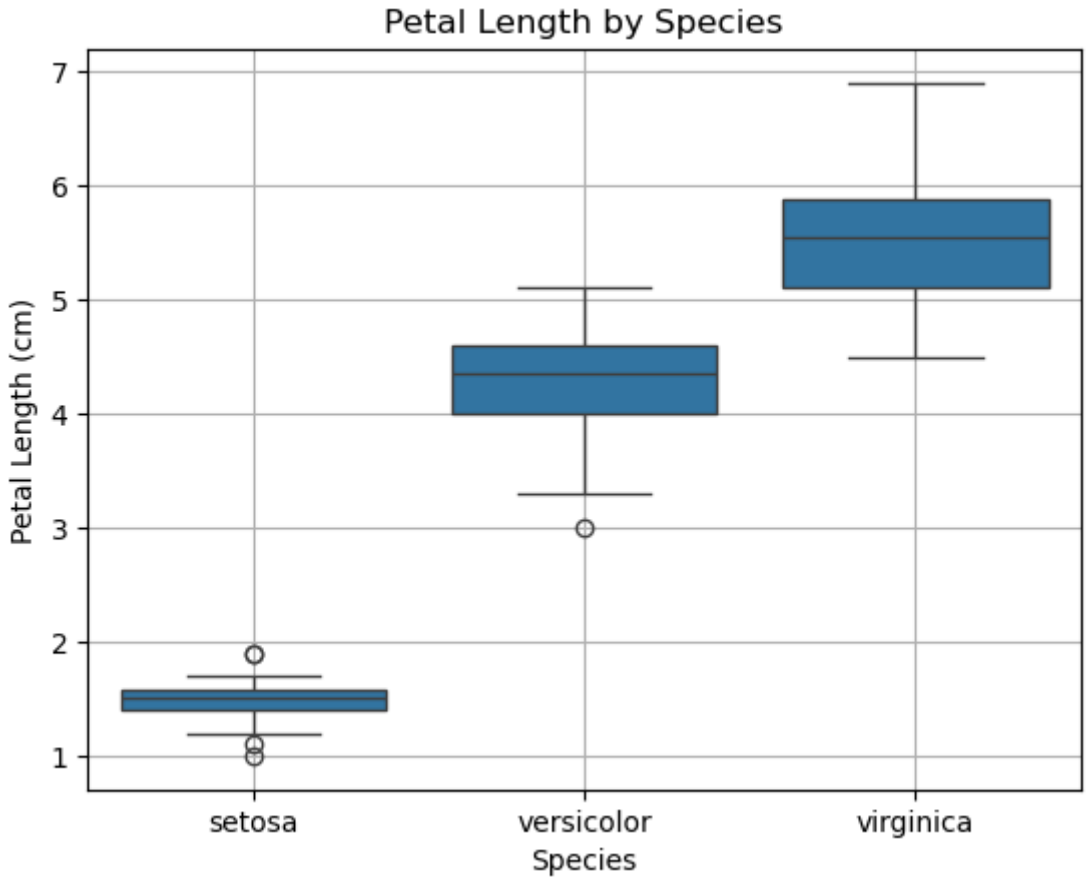
1-1. 종별 Petal_length 기술통계량 및 Boxplot

Species	Mean	Std	Min	Q1	Median	Q3	Max
Setosa	1.464	0.174	1.0	1.4	1.5	1.575	1.9
Versicolor	4.260	0.470	3.0	4.0	4.35	4.6	5.1
Virginica	5.552	0.552	4.5	5.1	5.55	5.875	6.9

1. Iris 데이터셋 기반 기초 통계 분석

1-1. 종별 Petal_length 기술통계량 및 Boxplot

Species	Mean	Std	Min	Q1	Median	Q3	Max
Setosa	1.464	0.174	1.0	1.4	1.5	1.575	1.9
Versicolor	4.260	0.470	3.0	4.0	4.35	4.6	5.1
Virginica	5.552	0.552	4.5	5.1	5.55	5.875	6.9



- Virginica 그룹의 평균 및 중앙값이 가장 높다.
- setosa 그룹의 평균 및 중앙값이 가장 낮다.
- versicolor Species의 최댓값이 Virginica의 최솟값보다 크다.

1-2. 정규성 검정 (Shapiro-Wilk)

- setosa: p-value = 0.0548
- versicolor: p-value = 0.1585
- virginica: p-value = 0.1098
- setosa의 p-value가 0.05 이상이므로 귀무가설을 기각할 수 없다. 즉 정규성을 만족한다.
- versicolor의의 p-value가 0.05 이상이므로 귀무가설을 기각할 수 없다. 즉 정규성을 만족한다.
- virginica의의 p-value가 0.05 이상이므로 귀무가설을 기각할 수 없다. 즉 정규성을 만족한다.

1-3. 가설 수립 및 등분산성 검정 (Levene Test)

- 가설
  - H_0 : 3개 Species 간 분산이 같다.
  - H_1 : 적어도 한 Species의 분산이 다르다.
- 해석
  - Levene test p-value = 0.0000000313
  - Levene test 결과 p-value 값이 0.0000000313으로 0.05보다 훨씬 작으므로, 등분산성을 만족하지 않는다.

1-4. 일원분산분석 (ANOVA)

- F-statistic: 1180.1612
- p-value: 2.8567766109615584e-91
- 해석
  - 유의수준 0.05에서 p-value < 0.05이므로, 귀무가설을 기각한다.
  - 따라서 세 species 간 petal_length의 평균에는 유의미한 차이가 있다.

1-5. 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
-----------------------------------------------------
```

- 해석 - 모든 그룹쌍 사이의 p-adj 값이 0.0으로 0.05보다 작으며 reject = true인 것을 통해 세 종 사이의 petal_length 평균에 유의미한 차이가 존재한다고 볼 수 있다. - setosa vs versicolor / setosa vs virginica / versicolor vs virginica 모두 유의미한 차이를 가진다.

1-6. 결론

- Boxplot 및 ANOVA 결과, 세 종(setosa, versicolor, virginica) 간의 petal length 평균에는 통계적으로 유의미한 차이가 있음을 알 수 있다.
- Tukey HSD 사후검정 결과, 모든 그룹 쌍(setosa–versicolor, setosa–virginica, versicolor–virginica) 간에 유의미한 차이가 존재함을 확인할 수 있다.
- 세 그룹 중 virginica가 가장 긴 petal length를 보였으며, setosa는 가장 짧은 petal length를 가진다고 할 수 있다.

---

2. 신용카드 사기 탐지 모델 (Logistic Regression)

2-1. 데이터 기본 탐색

- Class 분포 (전체): 정상 거래(0): 284,315 (99.8273%), 사기 거래(1): 492 (0.001727%)

2-2. 전처리 및 샘플링

- 사기 거래(Class=1)는 전체 유지
- 정상 거래(Class=0)는 10,000건 무작위 샘플링
- `Amount` 변수는 StandardScaler로 정규화한 `Amount_Scaled`로 대체 후 삭제
- 샘플링 후 Class 비율 재출력 결과
  - Class 0 : 0.953107
  - Class 1 : 0.046893

2-3. 데이터 분할

- 학습:테스트 = 8:2 (`stratify=y` 적용)
- random_state는 42로 설정
- 학습 데이터에 대해 SMOTE 적용
- Train Class 분포:

Class	Count
0	7999
1	394

- Test Class 분포:

Class	Count
0	2001
1	98

2-4. SMOTE 적용

- SMOTE 적용 전 사기 거래 건수 분포:

Class	Count
0	7999
1	394

- SMOTE 적용 후 사기 거래 건수 분포:

Class	Count
0	7999
1	7999

- SMOTE를 적용해야 하는 이유
    - 전체 284,807 건 중 사기 거래는 단 492건으로, 이와 같은 클래스 불균형은 머신 러닝 모델이 대부분의 샘플을 '정상거래'로 예측해도 높은 정확도를 얻게 만드는 허상을 야기한다.
    - 이러한 상황에서는 정확도보다 소수 클래스 (사기 거래)의 Recall 값을 향상하는 것이 중요하다.
    - 이를 해결하기 위해 SMOTE 적용이 필요하다.
    - SMOTE는 소수 클래스의 데이터를 단순 복제하는 것이 아니라, 기존 소수 클래스의 샘플 사이에서 합성 샘플을 생성하여 모델의 학습 균형을 맞춰준다.

2-5. 모델 학습 및 Threshold 조정

- 모델: `LogisticRegression(max_iter=5000, random_state=42)`
- 예측 확률 기반으로 Threshold = 0.8 설정 후 이진 분류 수행

2-6. 예측값 및 예측확률, 평가 지표 (테스트셋 기준)

- 예측 확률 (Class=1, 처음 10개만 출력):
    - [0.140133283
      0.07740872
      0.03678127
      0.02188825
      0.29228138
      0.00212237
      1.0
      0.00716832
      0.00123913
      0.0144795]
- 예측 결과 (threshold=0.8 기준, 처음 10개만 출력):
    - [0 0 0 0 0 1 0 0 0 0]

지표 유형	Class 0 (정상 거래)	Class 1 (사기 거래)
Precision	0.9940	0.9247

지표 유형	Class 0 (정상 거래)	Class 1 (사기 거래)
Recall	0.9965	0.8776
F1-score	0.9953	0.9005
Support	2001	98

- PR-AUC (average_precision_score): 0.9551

2-7. 성능 해석 및 목표 충족 여부

평가 항목	값	목표 기준	충족 여부
Recall (Class 0)	0.9965	≥ 0.80	만족
Recall (Class 1)	0.8776	≥ 0.80	만족
F1-score (Class 0)	0.9953	≥ 0.88	만족
F1-score (Class 1)	0.9005	≥ 0.88	만족
PR-AUC	0.9551	≥ 0.90	만족

→ 모든 기준을 충족하므로, 모델은 목표 성능을 달성함