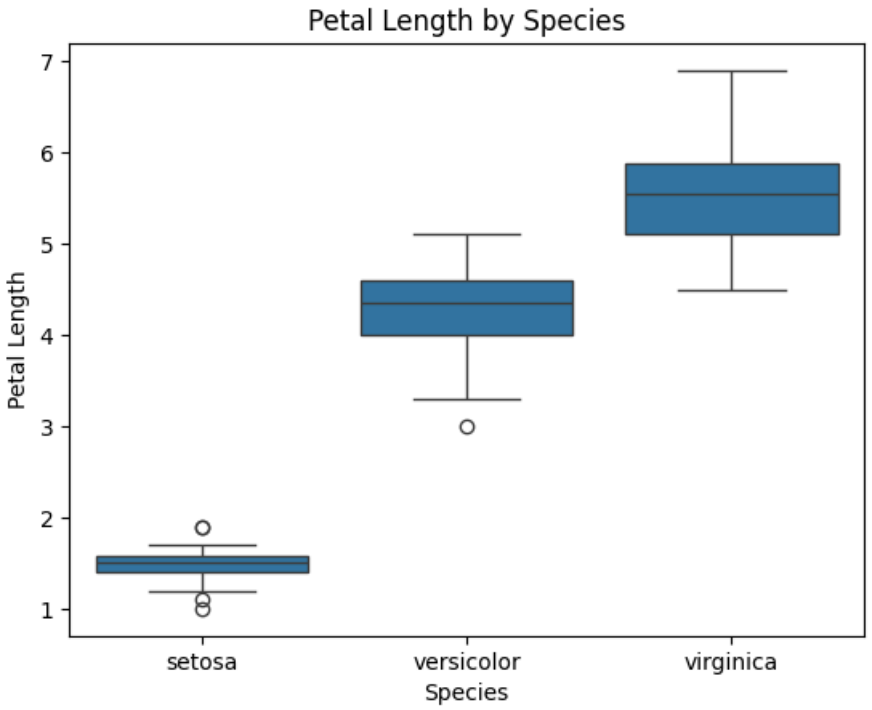
3(1)-Basic_Statistics_ML Report

📊 기초 통계 과제 결과 요약

1. Boxplot 분석 결과



종(Species)에 따른 `petal_length` 분포를 Boxplot으로 시각화한 결과,

Setosa < Versicolor < Virginica 순으로 꽃잎 길이 평균이 점차 커지는 경향을 보였다.

특히 Setosa는 분포가 좁고 중앙값도 낮아, 세 그룹 중 가장 짧은 petal length 값을 가진다.

2. Shapiro-Wilk 정규성 검정 & Levene 등분산성

유의수준 α = 0.05 하에서 ANOVA 분석을 위해 정규성 검정과 등분산성 검정을 하면 다음과 같다,

하나, **Shapiro-Wilk 정규성 검정** 결과, 아래와 같다.

- setosa의 `p-value: 0.0548` → 정규성 있음
- versicolor의 `p-value: 0.1585` → 정규성 있음
- virginica의 `p-value: 0.1098` → 정규성 있음

따라서 모든 그룹은 정규성을 지닌 데이터라 판단할 수 있다.

둘, **Levene 등분산성 검정** 결과, `p-value: 0.0000` 으로 등분산성이 있다고 판단하기 어렵다.

그러나 문제의 요건에 따라, 결과와 무관하게 ANOVA 분석을 진행한다.

3. ANOVA 분석 결과

일원분산분석(One-way ANOVA) 결과,

F 값은 1180.1612, **p-value는 0.0000**으로 유의수준 0.05보다 작게 나타났다.

이에 따라 **귀무가설("세 그룹의 평균은 같다")은 기각되며**,

3개 그룹 간 `petal_length` 평균에는 통계적으로 유의미한 차이가 존재함을 확인할 수 있다.

4. Tukey HSD 사후검정 결과

ANOVA 결과에 따라 Tukey HSD 사후검정을 실시한 결과,

모든 그룹 쌍(setosa–versicolor, setosa–virginica, versicolor–virginica) 사이에서

p-value가 0.05보다 작게 나타났으며, 평균 차이도 모두 통계적으로 유의미한 것으로 확인되었다.

비교 그룹	평균 차이	p-value	유의
Setosa – Versicolor	2.798	0.000	O
Setosa – Virginica	4.090	0.000	O
Versicolor – Virginica	1.292	0.000	O

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

🤖 기초 머신러닝 과제 결과 요약

1. SMOTE 적용 전후 클래스 비율

신용카드 사기 탐지 데이터는 클래스 1(사기 거래)의 비율이 전체의 약 4.7%로 매우 불균형한 상태였다.

이를 해결하기 위해 학습 데이터에 SMOTE 기법을 적용하여 클래스 간의 샘플 수를 균형 맞췄다.

```
SMOTE 전 y_train 클래스 분포:
Class
0    7999
1     394
Name: count, dtype: int64

SMOTE 후 y_train_sm 클래스 분포:
Class
0    7999
1    7999
Name: count, dtype: int64
```

단계	Class 0 개수	Class 1 개수
SMOTE 적용 전	7,999	394
SMOTE 적용 후	7,999	7,999

---

2. 모델별 성능 비교

총 6개의 분류 모델(RandomForest, XGBoost, LightGBM, CatBoost, LogisticRegression, NaiveBayes)을

기본 설정으로 학습한 뒤, Class 1(사기 거래)에 대한 성능을 비교하였다.

평가지표는 Precision, Recall, F1-score, PR-AUC를 사용하였다.

	precision	recall	f1-score	PR-AUC
RandomForest	0.975904	0.826531	0.895028	0.915689
XGBoost	0.921348	0.836735	0.877005	0.905101
LightGBM	0.952941	0.826531	0.885246	0.909188
CatBoost	0.931818	0.836735	0.881720	0.912981
LogisticRegression	0.875000	0.857143	0.865979	0.905700
NaiveBayes	0.771084	0.653061	0.707182	0.761097

모델	Precision	Recall	F1-score	PR-AUC
RandomForest	0.976	0.827	0.895	0.916
XGBoost	0.921	0.837	0.877	0.905
LightGBM	0.953	0.827	0.885	0.909
CatBoost	0.932	0.837	0.882	0.913
LogisticRegression	0.875	0.857	0.866	0.906
NaiveBayes	0.771	0.653	0.707	0.761

→ 이 중 **RandomForestClassifier**가 가장 높은 F1-score와 PR-AUC를 기록하여 최종 모델로 선정하였다.

3. 최종 모델 성능 평가

최종 모델(RandomForestClassifier)의 Class 1 기준 성능은 다음과 같다.

- Precision: 0.976
- Recall: 0.827
- F1-score: 0.895
- PR-AUC: 0.916

→ 과제에서 제시된 목표 성능 기준(Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90)을 모두 충족하였다.

4. 추가 개선 제안

더 높은 성능이 필요할 경우 아래 방법들을 고려할 수 있다.

- RandomForest의 하이퍼파라미터 튜닝
- Threshold 조정을 통한 precision–recall 균형 개선
- XGBoost, LightGBM 등과의 앙상블 모델 구축