기초통계 과제 보고서: Iris 데이터셋 분석

1. '데이터 로드 및 구조 확인' 셀 결과

=== 데이터 구조 확인 (Head) ===

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

=== 데이터 정보 확인 (Info) ===

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
None
```

-> Seaborn 라이브러리에 내장된 Iris 데이터셋을 불러오고, head()와 info() 함수를 사용하여 데이터의 전반적인 구조를 파악하였다. 4개의 독립변수(sepal_length, sepal_width, petal_length, petal_width)는 실수형 데이터이고, 종속변수인 species는 문자형 데이터이다. 모든 컬럼에서 결측치(Null)가 발견되지 않아 별도의 전처리 없이 분석이 가능하다.

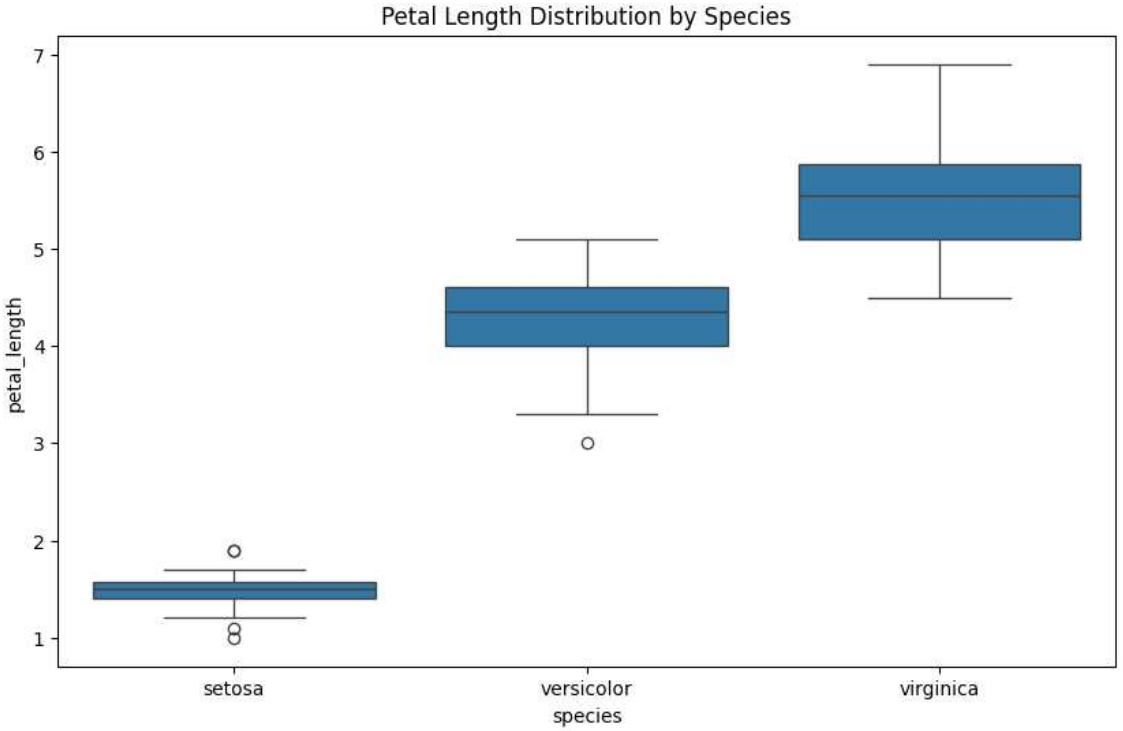
2. '기술통계량 확인' 셀 결과

=== Species별 Petal Length 기술통계량 ===

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

-> 종별 꽃잎 길이의 통계량을 계산하였다. 각 종별 꽃잎 길이 평균과 표준편차가 확연히 차이가 있음을 확인할 수 있다.

3. ‘시각화’ 셀 결과



-> Boxplot 작성 결과, 세 종의 박스(Box) 높이가 서로 겹치지 않고 계단식으로 명확하게 구분되었다. Setosa는 매우 낮은 위치에 좁게 분포하며 일부 이상치가 관측되었고, Virginica는 가장 높은 위치에 넓게 분포하는 특성을 보였다. 이는 통계적 검정 전부터 그룹 간 차이가 큼을 시사한다.

4. ‘정규성 검정’ 셀 결과
=== 정규성 검정 (Shapiro-Wilk) ===
Species: setosa, P-value: 0.0548
Species: versicolor, P-value: 0.1585
Species: virginica, P-value: 0.1098
-> ANOVA 분석의 전제 조건인 '정규성'을 확인하기 위해, 세 그룹 각각에 대해 Shapiro-Wilk 검정을 실시하였다.
그 결과, 세 그룹 모두 P-value가 0.05 이상으로 산출되어 귀무가설(=정규분포를 따른다)을 기각하지 못했다. 따라서 세 집단 모두 정규성을 만족하는 것으로 판단하였다.

5. ‘등분산성 검정’ 셀 결과
=== 등분산성 검정 (Levene) ===
Levene Result - P-value: 0.0000
-> 먼저 귀무가설(H0)는 ‘그룹 간 분산이 같다.’. 대립가설(H1)은 ‘적어도 한 그룹의 분산은 다른 그룹과 차이가 있다.‘이다. 검정 결과 P-value가 0.0000으로 산출되어 유의수준 0.05보다 현저히 작았다. 따라서 귀무가설이 기각되어 등분산 가정은 만족하지 못하는 것으로 나타났다. (이에 따르면 ANOVA 검정을 시행하기 어려우나, 본 과제 명세서의 가이드라인에 따라 등분산성을 만족한다고 가정하고 이후 분석을 진행하였다.)

6. 'ANOVA 가설 수립' 셀 결과
=== ANOVA 가설 ===
H0: 세 종(Species) 간의 Petal Length 평균에는 차이가 없다.
H1: 적어도 한 그룹의 평균은 다른 그룹과 차이가 있다.
-> 따로 분석한 코드가 있진 않으며, ANOVA Test를 위해 먼저 가설을 수립한다

7. 'One-way ANOVA' 셀 결과
=== One-way ANOVA 결과 ===
F-statistic: 1180.1612, P-value: 2.8568e-91
-> 분석 결과 F-statistic 값이 매우 높고, P-value가 0에 가까운 매우 작은 값이므로 귀무가설을 강력하게 기각한다. 세 종의 꽃잎 길이 평균에는 통계적으로 매우 유의미한 차이가 존재한다고 볼 수 있다.

8. '사후검정 (Tukey HSD)' 셀 결과
=== 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

-> ANOVA에서 그룹 간 평균 차이가 존재함을 확인했으므로, 구체적으로 어떤 그룹 간에 차이가 있는지 파악하기 위해 Tukey HSD 사후검정을 수행하였다.
그 결과, 모든 그룹 쌍의 비교에서 reject 값이 True로 나타났다. 이는 어떤 조합이든 상관없이 꽃잎 길이의 평균 차이가 유의하다, 즉 평균 차이가 존재한다는 것을 의미한다.

9. 결과 요약
지금까지 수행한 분석 결과를 종합하면 다음과 같다.
- 시각화 (Boxplot): 종별 꽃잎 길이의 분포(Box) 높이가 서로 겹치지 않고 계단식으로 뚜렷하게 구분됨을 시각적으로 확인하였다.
- ANOVA (분산분석): 검정 결과 P-value가 0에 가까운 값으로 나타나, 세 종의 평균 간에 통계적으로 매우 유의미한 차이가 있음을 입증하였다.
- 사후검정 (Tukey HSD): 모든 종의 조합)에서 P-adj 값이 0.05 미만으로 나타나, 어떤 두 종을 비교하더라도 통계적으로 확실한 차이가 있음을 확인하였다.

-> 시각화 결과 꽃잎 평균 길이는 Virginica > Versicolor > Setosa 순으로 나타났고, ANOVA와 Tukey HSD를 통해 평균 차이가 통계적으로 존재함을 증명하였으므로,
"Virginica 종의 꽃잎 길이가 통계적으로 유의하게 가장 길며, Setosa 종이 가장 짧다"고 결

론지을 수 있다.
10. '회귀 분석' 셀 결과
=== 회귀 분석 결과 ===
MSE: 0.1300
R2 Score: 0.9603
회귀 계수(Coefficients): [0.72281463 -0.63581649 1.46752403]
절편(Intercept): -0.2621959025887066
-> 꽃받침 길이(Sepal Length), 꽃받침 너비(Sepal Width), 꽃잎 너비(Petal Width) 정보를 이용하여 꽃잎 길이(Petal Length)를 예측하는 선형 회귀 모델을 구축하였다.
모델 성능 평가 결과 결정계수(R2)가 높게 나왔고, 전체 데이터 변동성의 약 96.0%를 설명한다고 한다. 즉 해당 독립변수들이 종속변수(꽃잎 길이)를 설명하는 데 매우 적합하고 모델의 예측 신뢰도가 매우 높다고 볼 수 있다.
또한 평균제곱오차(MSE)가 작게 나타나, 모델이 데이터를 정밀하게 적합(Fitting)시켰음을 알 수 있다.
도출된 회귀 계수로 모델 표현시, 다음과 같다:

$$\text{Petal Length} = 0.72 * \{\text{Sepal Length}\} - 0.64 * \{\text{Sepal Width}\} + 1.47 * \{\text{Petal Width}\} - 0.26$$

이를 해석해보면 꽃받침 너비(Sepal Width)는 그 크기가 커질수록 꽃잎 길이는 짧아지고, 꽃받침 길이(Sepal Length), 꽃잎 너비(Petal Width)는 크기가 커질수록 꽃일 길이가 길어지나 꽃잎 너비가 좀 더 크게 영향을 준다고 해석할 수 있다.