ADP_Red\stats\2.anova.py

```python
# %% Chapter 6. Statistics - ANOVA
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import scipy.stats as stats


# %% 1. 데이터 수집
df = pd.read_csv('../../ADP_Python/data/iris.csv')

print(df.shape)
print(df.info())

# Check binary variable
for i, var in enumerate(df.columns):
    print(i, var, len(df[var].unique()))

# Check data summary
print(df.describe())

# 2. 데이터 결측치 보정
print(df.isna().sum())

# 3. 라벨 인코딩
# 4. 데이터타입, 더미변환 (One-Hot Encoding)
# 5. 파생변수 생성
# 6. 정규화 또는 스케일 작업
# Boxplot for scaling check
sns.boxplot(df)
plt.tight_layout()
plt.show()

# 7. 데이터 분리
# 8. 모델 학습 - ML


# %%
mtcars = pd.read_csv('../../ADP_Python/data/mtcars.csv')
print(mtcars.shape)
print(mtcars.info())

for i, var in enumerate(mtcars.columns):
    print(i, var, len(mtcars[var].unique()))

print(mtcars.describe())
print(mtcars.isna().sum())

df2_sample = mtcars[['mpg', 'am', 'cyl']]

from pandas.plotting import scatter_matrix

scatter_matrix(df2_sample)
plt.show()


```

```

# %% 9. 모델 학습 - Stats
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.formula.api as smf

from statsmodels.stats.anova import anova_lm

df_A = df[df.target == 'Iris-setosa']['sepal width']
df_B = df[df.target == 'Iris-versicolor']['sepal width']
df_C = df[df.target == 'Iris-virginica']['sepal width']

# ANOVA 정규성 체크 Shapiro (위반 시 Kruskal)
print(stats.shapiro(df_A))
print(stats.shapiro(df_B))
print(stats.shapiro(df_C))

# ANOVA 등분산성 체크 Levene (위반 시 Welch)
print(stats.levene(df_A, df_B, df_C))

# one-way ANOVA using statsmodels
df_sample = df[['sepal width', 'target']]
df_sample.columns = ['sepal_width', 'target']

model = smf.ols(
    formula='sepal_width ~ 1 + target', data=df_sample
).fit()

print(model.summary())

# one-way ANOVA using scipy.stats
anova = stats.f_oneway(df_A, df_B, df_C)
print(anova)

# one-way ANOVA using pinguoin

# %% two-way ANOVA using statsmodels
model2 = smf.ols(
    formula='mpg ~ 1 + am*cyl', data=df2_sample
).fit()

print(model2.summary())
print(anova_lm(model2, typ=2))

sns.lineplot(data=df2_sample, x='cyl', y='mpg', hue='am')
plt.show()

# %% 10. 앙상블
# 11. 모델 평가
# one-way ANOVA post-hoc
from statsmodels.stats.multicomp import pairwise_tukeyhsd, MultiComparison

mc = MultiComparison(data=df['sepal width'], groups=df['target'])
tukeyhsd = mc.tukeyhsd(alpha=0.05)
tukeyhsd.plot_simultaneous()

print(tukeyhsd.summary())

from statsmodels.graphics.factorplots import interaction_plot
interaction_plot(x=df2_sample['cyl'], trace=df2_sample['am'], response=df2_sample['mpg'],
                 colors=['red', 'blue'], markers=['D', 'o'])

plt.show()
# 12. 하이퍼파라미터 튜닝
# 13. 예측값 저장
# References
```