

```
# -----
# 와인 품질 데이터 분석 (Wine Quality Analysis)
# 葡萄酒质量数据分析
# -----



import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
from statsmodels.formula.api import ols
import statsmodels.api as sm


# 1. 데이터 불러오기
# 1. 读取数据
red_df = pd.read_csv('winequality-red.csv', sep=';', header=0, engine='python')
white_df = pd.read_csv('winequality-white.csv', sep=';', header=0, engine='python')


# 2. 각 데이터에 type 변수 추가
# 2. 为每个数据集添加“type”列
red_df.insert(0, column='type', value='red')
white_df.insert(0, column='type', value='white')


# 3. 두 데이터를 하나로 합치기
# 3. 将两个数据集合并
wine = pd.concat([red_df, white_df])


# 4. 데이터 기본 구조 확인
# 4. 查看数据的基本结构
print(wine.info())


# 5. 기술통계량 확인
# 5. 查看描述性统计量
print(wine.describe())


# 6. 품질별 평균 확인
# 6. 查看不同质量等级的平均值
print(wine.groupby('type')['quality'].mean())


# 7. 박스플롯으로 품질 분포 시각화
# 7. 使用箱线图可视化质量分布
```

```

sns.boxplot(x='type', y='quality', data=wine)
plt.title('Wine Quality by Type / 按类型划分的葡萄酒质量')
plt.show()

# 8. 와인 종류별 품질 평균 비교 (t-검정)
# 8. 按葡萄酒类型比较质量平均值 ( t检验 )
red_quality = wine.loc[wine['type'] == 'red', 'quality']
white_quality = wine.loc[wine['type'] == 'white', 'quality']

t_stat, p_value = stats.ttest_ind(red_quality, white_quality, equal_var=False)
print("T통계량(T-statistic): ", t_stat)
print("p값(p-value): ", p_value)

# 9. p값이 0.05보다 작으면 → 두 그룹 간 품질 평균이 유의하게 다름
# 9. 如果p值小于0.05 → 两组间的质量平均值存在显著差异
if p_value < 0.05:
    print("두 와인 품질의 평균은 통계적으로 유의한 차이가 있습니다.")
    print("红葡萄酒与白葡萄酒的质量平均值存在显著差异。")
else:
    print("두 와인 품질의 평균 차이는 유의하지 않습니다.")
    print("红葡萄酒与白葡萄酒的质量平均值差异不显著。")

# 10. 산점도 행렬로 변수 간 관계 확인
# 10. 使用散点矩阵查看变量之间的关系
sns.pairplot(wine, hue='type', vars=['fixed acidity', 'volatile acidity', 'citric acid',
'residual sugar', 'quality'])
plt.suptitle('Pairplot of Wine Variables / 葡萄酒变量散点图', y=1.02)
plt.show()

# 11. 회귀분석 (예: fixed acidity가 quality에 미치는 영향)
# 11. 回归分析 (例：固定酸度对质量的影响)
model = ols('quality ~ fixed_acidity', data=red_df).fit()
print(model.summary())

# 12. 부분 회귀 플롯 (Partial Regression Plot)
# 12. 部分回归图 ( Partial Regression Plot )
fig = plt.figure(figsize=(10, 6))
sm.graphics.plot_partregress('quality', 'fixed_acidity', ['volatile_acidity', 'citric_acid'],
data=red_df, obs_labels=False)
plt.title('Partial Regression Plot / 部分回归图')
plt.show()

```