ADP_Red\stats\4.reg.py

```python
# %% Chapter 6. Statistics - Regression
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/Cars93.csv')
sample = df[['EngineSize', 'RPM', 'Weight', 'Length', 'MPG.city', 'MPG.highway', 'Price']]
sample.columns = ['EngineSize', 'RPM', 'Weight', 'Length', 'MPGcity', 'MPGhighway', 'Price']

print(sample.shape)
print(sample.info())

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

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

# Check scatterplot
from pandas.plotting import scatter_matrix

scatter_matrix(sample)
plt.show()


# %% 2. 데이터 결측치 보정
print(sample.isna().sum())


# %% 6. 정규화 또는 스케일 작업
# Boxplot for scaling check
sns.boxplot(sample)
plt.tight_layout()
plt.show()

# Multicollinearity
sns.heatmap(sample.corr(), annot=True)
plt.show()

print(sample.corr())

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

# Simple Linear Regression
model = smf.ols(
    formula='Price ~ 1 + Length', data=sample
).fit()
```

```python

print(model.summary())

# Residual Plot
sns.scatterplot(model.resid)
plt.show()


# Multiple Linear Regression
model2 = smf.ols(
    formula='Price ~ EngineSize + RPM + Weight + Length + MPGcity + MPGhighway', data=sample
).fit()

model2.summary()
```