HW13——Handout

Q1

In [5]:

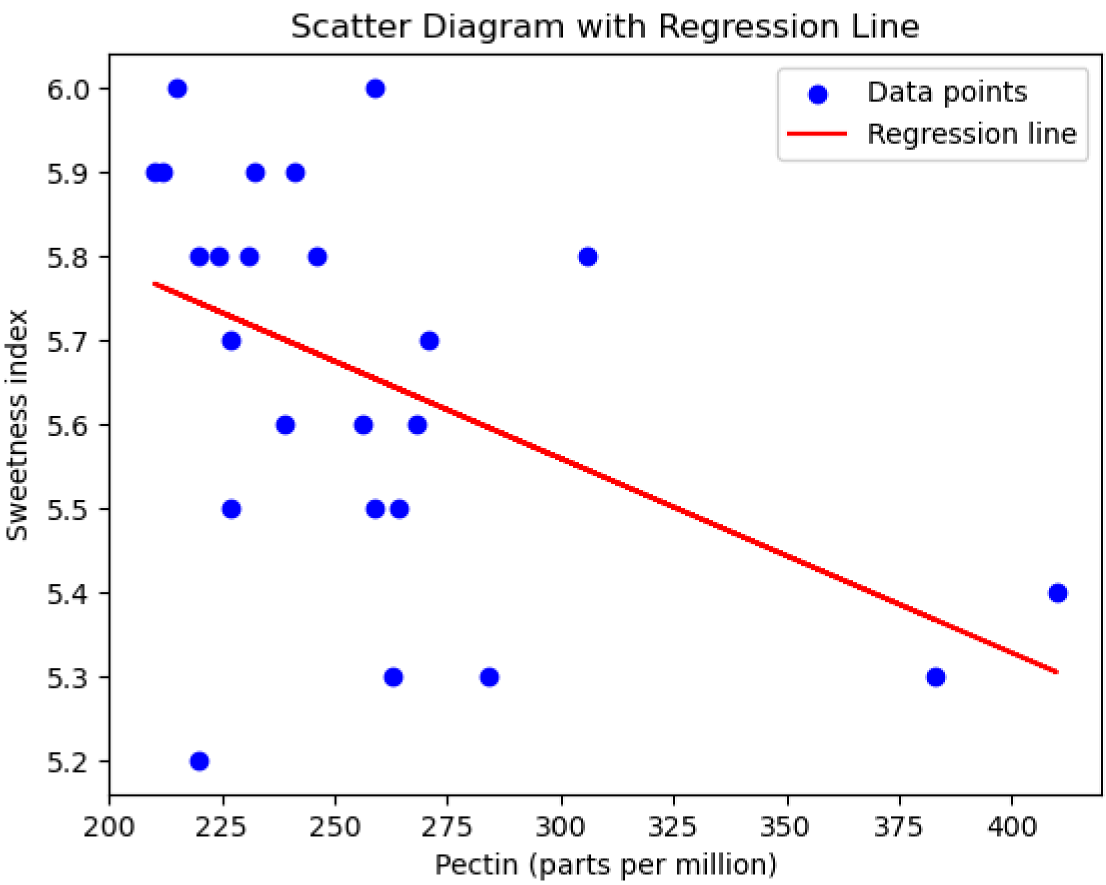
```
import pandas as pd
import numpy as np

data = {
    "Sweetness index (Y)": [5.20, 5.50, 6.00, 5.90, 5.80, 6.00, 5.80, 5.60, 5.60
    "Pectin (X)": [220.00, 227.00, 259.00, 210.00, 224.00, 215.00, 231.00, 268.0
}

df = pd.DataFrame(data)

x_bar = df["Pectin (X)"].mean()
y_bar = df["Sweetness index (Y)"].mean()

s_x2 = df["Pectin (X)"].var(ddof=1)
s_y2 = df["Sweetness index (Y)"].var(ddof=1)

s_xy = df.cov().iloc[0, 1]

r_xy = df.corr().iloc[0, 1]
r_xy2 = r_xy ** 2

print(f"Sample means: x̄ = {x_bar:.5f}, ȳ = {y_bar:.5f}")
print(f"Sample variances: s_x^2 = {s_x2:.5f}, s_y^2 = {s_y2:.5f}")
print(f"Sample covariance: s_xy = {s_xy:.5f}")
print(f"Sample correlation coefficient: r_xy = {r_xy:.5f}, r_xy^2 = {r_xy2:.5f}"
```

```
Sample means: x̄ = 256.95833, ȳ = 5.65833
Sample variances: s_x^2 = 2454.47645, s_y^2 = 0.05732
Sample covariance: s_xy = -5.67138
Sample correlation coefficient: r_xy = -0.47815, r_xy^2 = 0.22862
```

In [6]:

```
import matplotlib.pyplot as plt

b1 = s_xy / s_x2
b0 = y_bar - b1 * x_bar

print(f"Regression equation: Y = {b0:.5f} + {b1:.5f}X")

plt.scatter(df["Pectin (X)"], df["Sweetness index (Y)"], label="Data points", co
plt.plot(df["Pectin (X)"], b0 + b1 * df["Pectin (X)"], label="Regression line",
plt.xlabel("Pectin (parts per million)")
plt.ylabel("Sweetness index")
plt.title("Scatter Diagram with Regression Line")
plt.legend()
plt.show()
```

```
Regression equation: Y = 6.25207 + -0.00231X
```



Q2

I obtained a dataset containing 30 samples of red wine from the UCI Machine Learning Repository - Wine Quality Dataset. The variables I selected for analysis are sulphates (independent variable, X) and pH (dependent variable, Y).

In [9]:

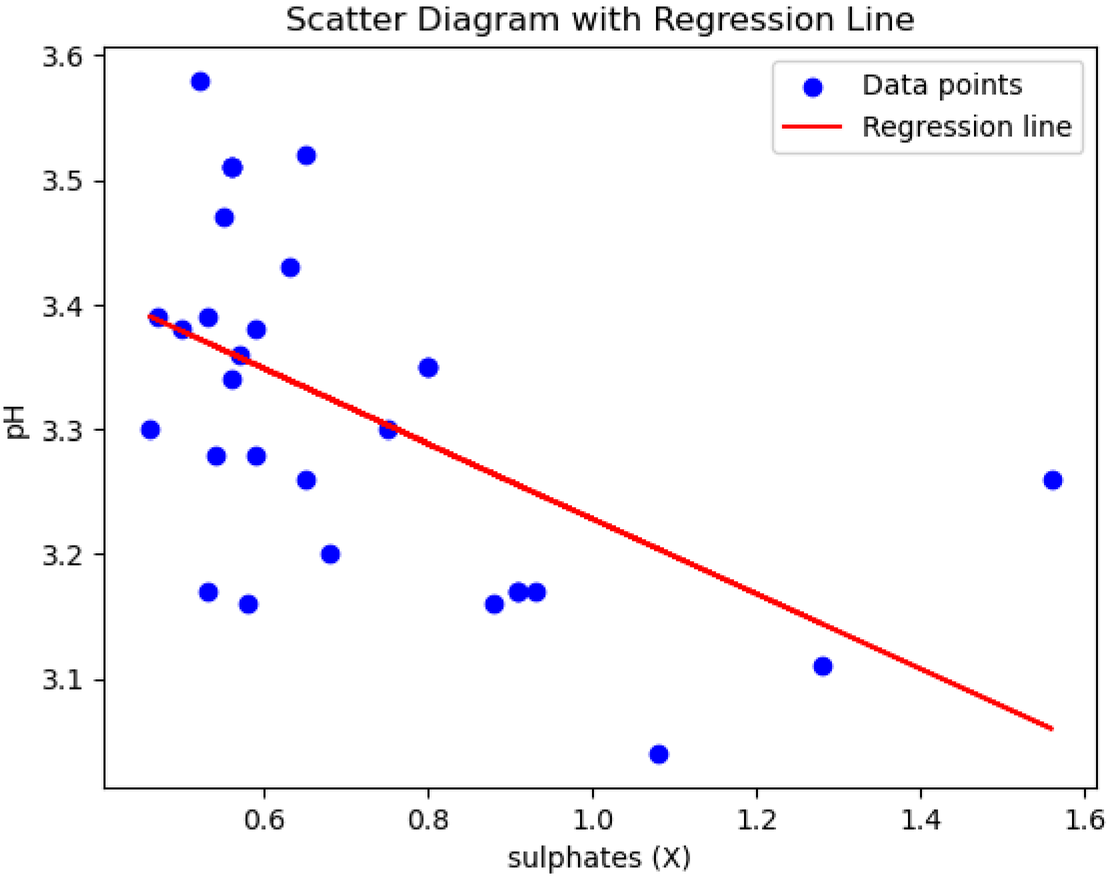
```
data = {
    "pH (Y)": [3.51, 3.2, 3.26, 3.16, 3.51, 3.51, 3.3, 3.39, 3.36, 3.35, 3.28, 3
    "sulphates (X)": [0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0.57, 0.8,
}
```

In [10]:

```
df = pd.DataFrame(data)

x_bar = df["sulphates (X)"].mean()
y_bar = df["pH (Y)"].mean()

s_x2 = df["sulphates (X)"].var(ddof=1)
s_y2 = df["pH (Y)"].var(ddof=1)

s_xy = df.cov().iloc[0, 1]

r_xy = df.corr().iloc[0, 1]
r_xy2 = r_xy ** 2

print(f"Sample means: x̄ = {x_bar:.5f}, ȳ = {y_bar:.5f}")
print(f"Sample variances: s_x^2 = {s_x2:.5f}, s_y^2 = {s_y2:.5f}")
print(f"Sample covariance: s_xy = {s_xy:.5f}")
print(f"Sample correlation coefficient: r_xy = {r_xy:.5f}, r_xy^2 = {r_xy2:.5f}"

import matplotlib.pyplot as plt
```

```
b1 = s_xy / s_x2
b0 = y_bar - b1 * x_bar

print(f"Regression equation: Y = {b0:.5f} + {b1:.5f}X")

plt.scatter(df["sulphates (X)"], df["pH (Y)"], label="Data points", color="blue"
plt.plot(df["sulphates (X)"], b0 + b1 * df["sulphates (X)"], label="Regression l
plt.xlabel("sulphates (X)")
plt.ylabel("pH")
plt.title("Scatter Diagram with Regression Line")
plt.legend()
plt.show()
```

```
Sample means: x̄ = 0.70600, ȳ = 3.31667
Sample variances: s_x^2 = 0.06374, s_y^2 = 0.01932
Sample covariance: s_xy = -0.01918
Sample correlation coefficient: r_xy = -0.54657, r_xy^2 = 0.29874
Regression equation: Y = 3.52912 + -0.30092X
```


Scatter Diagram with Regression Line
Data points
Regression line
pH
sulphates (X)