CP2403 - Assignment – Part 2 – Task 3: Linear Regression

First Name: Kantapong
Last Name: Wongsanguan

1: Data Selection

- CalCOFI bottle dataset
  - Water temperatures (T_degC) subset of temperature up to 25 degrees.
  - Water Salinity (Salinty) subset of oxygen saturation up to 140%.
  - Null values dropped

2: Scatter plot with regression line

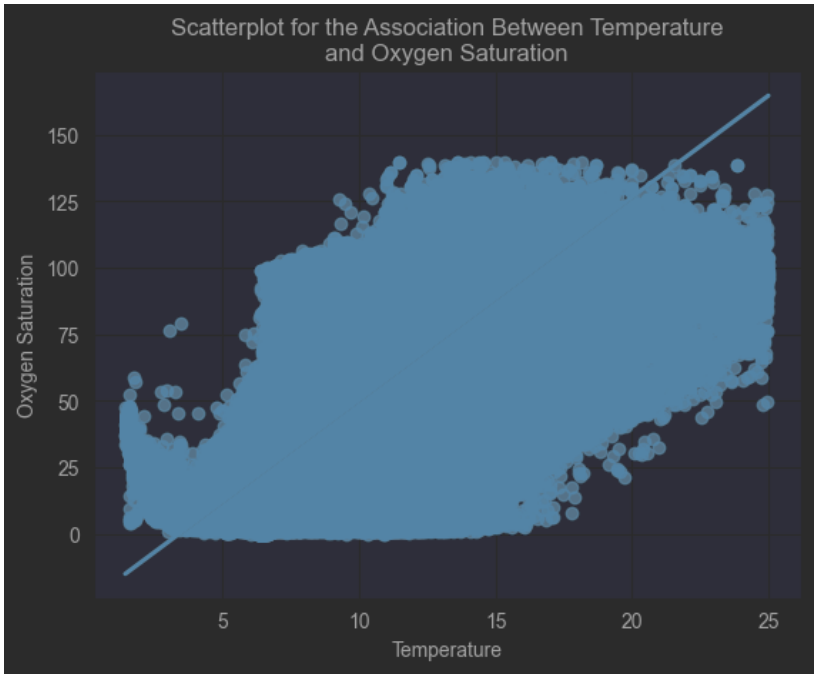

Figure 1: Scatter plot of water temperature vs oxygen saturation.

The scatter plot above represents a relationship between the water temperature and the oxygen saturation. Roughly, we can estimate that as the water temperature increase, the oxygen saturation also rises, as supported by the line of best fit through the plot.

Association between temperature and oxygen saturation.
"PearsonRResult(statistic=0.8565216058262785, pvalue=0.0)"

- The Pearson correlation coefficient between temperature ('T_degC') and oxygen saturation ('O2Sat') is approximately 0.857, indicating a strong positive correlation between the two variables. Additionally, the p-value is extremely small (close to zero), suggesting that this correlation is statistically significant.

3: Regression Analysis results

OLS Regression Results

```
==============================================================================
=====
Dep. Variable:            O2Sat_c        R-squared:                 0.734
Model:                        OLS        Adj. R-squared:            0.734
Method:             Least Squares        F-statistic:           1.817e+06
Date:            Thu, 28 Sep 2023        Prob (F-statistic):         0.00
Time:                    18:11:45        Log-Likelihood:      -2.8827e+06
No. Observations:          659639        AIC:                   5.765e+06
Df Residuals:              659637        BIC:                   5.765e+06
Df Model:                       1
Covariance Type:        nonrobust
==============================================================================
=====
              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.143e-14     0.024   4.85e-13      1.000      -0.046      0.046
T_degC_c       7.6335     0.006   1347.869      0.000       7.622      7.645
==============================================================================
=====
Omnibus:                45614.265   Durbin-Watson:              0.131
Prob(Omnibus):              0.000   Jarque-Bera (JB):       75845.337
Skew:                      -0.539   Prob(JB):                    0.00
Kurtosis:                   4.264   Cond. No.                    4.16
==============================================================================
=====
```

4: Regression line – if valid

O2Sat_c = 1.143e-14 + 7.6335(T_degC_c)

4: Residual plot – if required

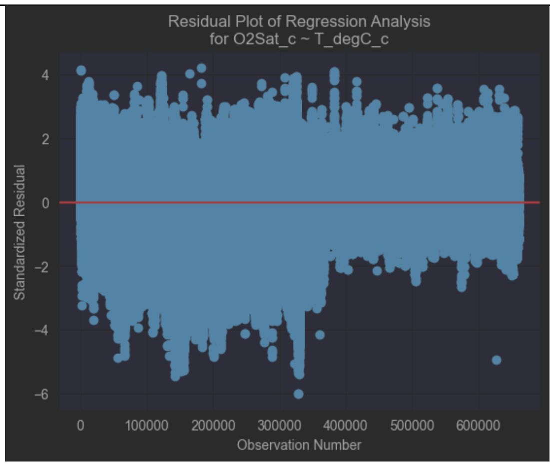

Figure 2: Residual plot of regression

A residual plot was generated to assess the goodness of fit of the regression model. The plot appears to show residuals evenly distributed around zero, which is a positive sign indicating that the assumptions of linear regression are met.

Percentage of Residuals Beyond 2 Standard Deviations:

- Approximately 5.04% of the residuals are more than 2 standard deviations away from the mean. While this suggests some deviation from normality, it's within an acceptable range for a good fit.

Percentage of Residuals Beyond 2.5 Standard Deviations:

- Approximately 2.62% of the residuals are more than 2.5 standard deviations away from the mean. This percentage is even lower, indicating that the model's fit is better when using a stricter threshold.

5: Conclusion from residual plot – if valid

To conclude, the examination of these two variables reveals a robust positive connection between temperature and oxygen saturation within the dataset. The linear regression model indicates the statistical significance of temperature's impact on oxygen saturation, indicating that for the water temperature unit, as there is an associated rise in oxygen saturation of roughly 7.6335 units. The model explains a substantial portion of the variance in oxygen saturation (73.4%).

Implications:
In terms of biological Significance, the strong temperature-oxygen saturation correlation suggests an important biological link. It warrants further exploration in fields like marine biology and healthcare to understand its implications for ecosystems and human health.
In addition, as for predictive value, the regression model can predict oxygen levels based on temperature. This can aid environmental forecasting and medical monitoring, enhancing decision-making.
Lastly, data quality matters as the analysis underscores the significance of data quality and outlier handling. Proper data preprocessing is vital for reliable results, emphasizing the need for rigorous data management practices.