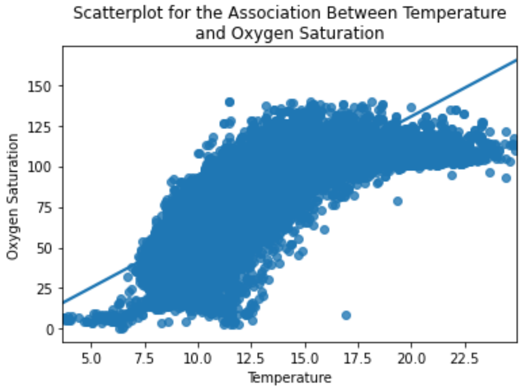
CP2403 - Project – Part 2 - REGRESSION

First Name: Caleb
Last Name: Webster

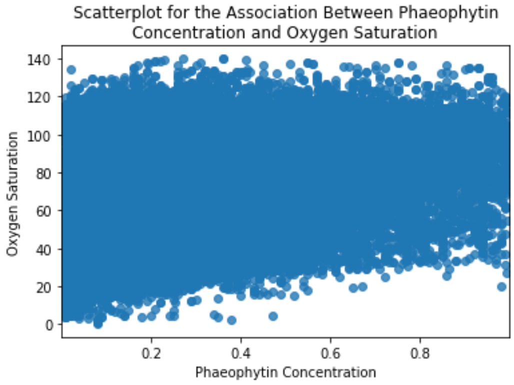
1: Scatter plots between each explanatory variable and response variable

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



The plot above reveals an increasing correlation between water temperature and oxygen saturation.

Figure 2: Scatter plot of phaeophytin concentration vs oxygen saturation.



This plot shows a weaker yet still increasing correlation between phaeophytin concentration and oxygen saturation.

2: List all the explanatory variables selected for regression analysis. Justify your selection

Water Temperature (T_degC): temperature was chosen as an explanatory variable because the temperature of the water is mostly dependent on the environment, location, and season. Most of the other variables in the dataset should have very little effect on the temperature, so it is an excellent choice.

Phaeophytin Concentration (Phaeop): phaeophytin is a chemical compound that assists with the photosynthetic reaction in plants and purple bacteria. Photosynthesis is the process used to create oxygen, therefore changes in phaeophytin concentration should directly affect the oxygen saturation of the water.

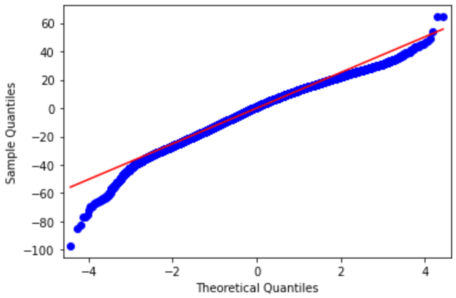
3: Regression analysis results

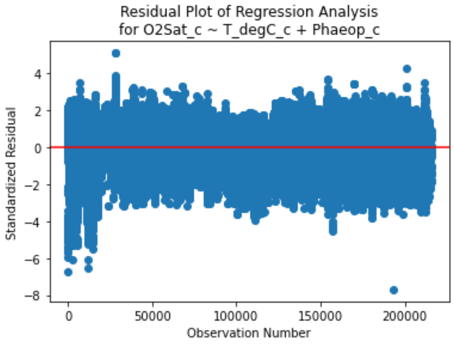
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                O2Sat_c   R-squared:                       0.752
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                 3.266e+05
Date:                Tue, 01 Jun 2021   Prob (F-statistic):               0.00
Time:                        14:01:37   Log-Likelihood:            -8.5234e+05
No. Observations:              215536   AIC:                         1.705e+06
Df Residuals:                  215533   BIC:                         1.705e+06
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -8.618e-13      0.027  -3.17e-11      1.000      -0.053       0.053
T_degC_c       7.0153      0.009    793.812      0.000       6.998       7.033
Phaeop_c      25.5296      0.165    154.545      0.000      25.206      25.853
==============================================================================
Omnibus:                     4549.238   Durbin-Watson:                   0.257
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4841.712
Skew:                          -0.365   Prob(JB):                         0.00
Kurtosis:                       3.070   Cond. No.                         18.7
==============================================================================
```

4: Regression equation/line

O2Sat_c = -8.618e-13 + 7.0153(T_degC_c) + 25.5296(Phaeop_c)

5: qqplot





6: Conclusion from qqplot

The values shown on the qqplot fit the line well for -2 < x < 2, but trail away from the line beyond these values. The values converge on the line again at x = 4. There are no major outliers, and overall, the line appears to be a good fit.

7: percentage of observations over 2 standardized deviation

%3.925

8: percentage of observations over 2.5 standardized

%0.997

9: Conclusion from observations over 2 std and 2.5 std

Less than 5% of observations from the residual plot have > 2 standard deviation (%3.925), and less than 1% of observations have > 2.5 standard deviation (%0.997), which means that the regression line is a good fit for the correlation between temperature, phaeophytin concentration, and oxygen saturation.

This means that the correlation between temperature, oxygen saturation, and phaeophytin concentration can be fairly accurately modelled using the equation stated in (4). Any increases in either temperature or phaeophytin concentration will relate to an increase in oxygen concentration; any decreases will have the opposite effect.

Implications: the oxygen concentration of the water should have a direct impact on sardine population; sardines need oxygen to breathe, therefore any significant decrease will cause a reduction in population. Using water temperature and phaeophytin concentration, a model was constructed to predict oxygen saturation which could aid in understanding the changes in sardine population using the CalCOFI dataset.