Data Science Final Project Report

Carlos Blanco

A sample of the original data (along with a description of variables and where the data was obtained)

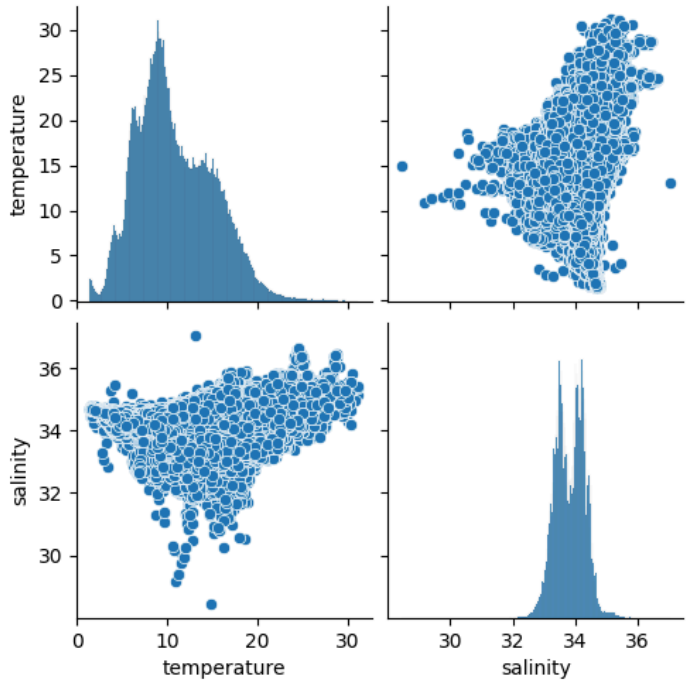
Original data link: https://www.kaggle.com/datasets/sohier/calcofi

The original dataset contains over 74 variables for each row, but I selected Temperature and Salinity. The description for temperature: Water temperature in degrees Celsius. The description for salinity: Salinity in g of salt per kg of water (g/kg)

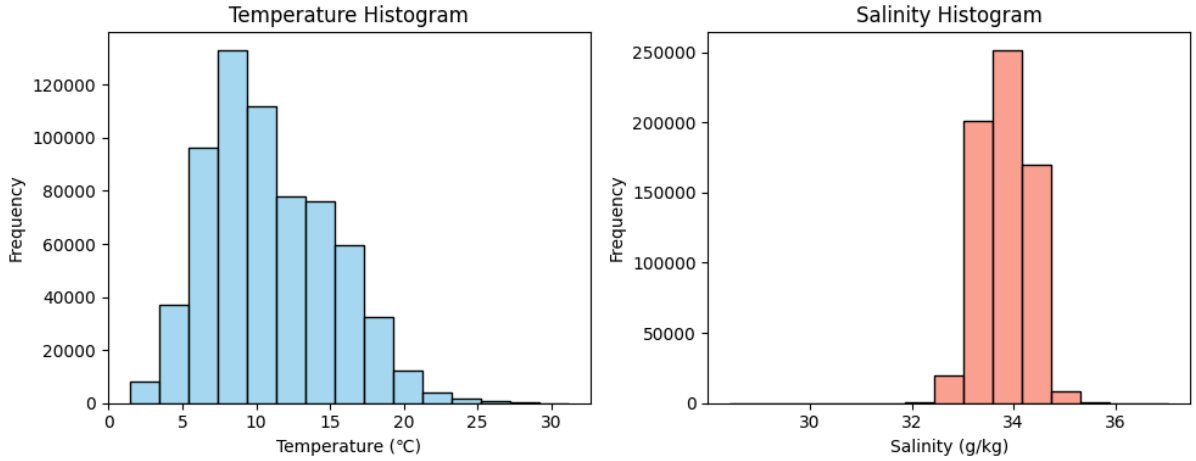
Some of the challenges with the data (Was there missing data? How did you wrangle the data?)

There were a total of 864863 rows. 10963 rows were missing the temperature value and 47354 were missing their salinity value. The rows missing both values won't show up in the model. However, the data points missing only one of these 2 values will impact the model. For wrangling I considered standardizing the data. However, the range difference between salinity and temperature wasn't significant enough to have an impact. Even though the variables had different units, they ranged from 1 - 40, so one variable wouldn't outweigh the other. Moreover, I didn't explore pivot tables or data melting techniques since the dataset was already structured in an optimal format.

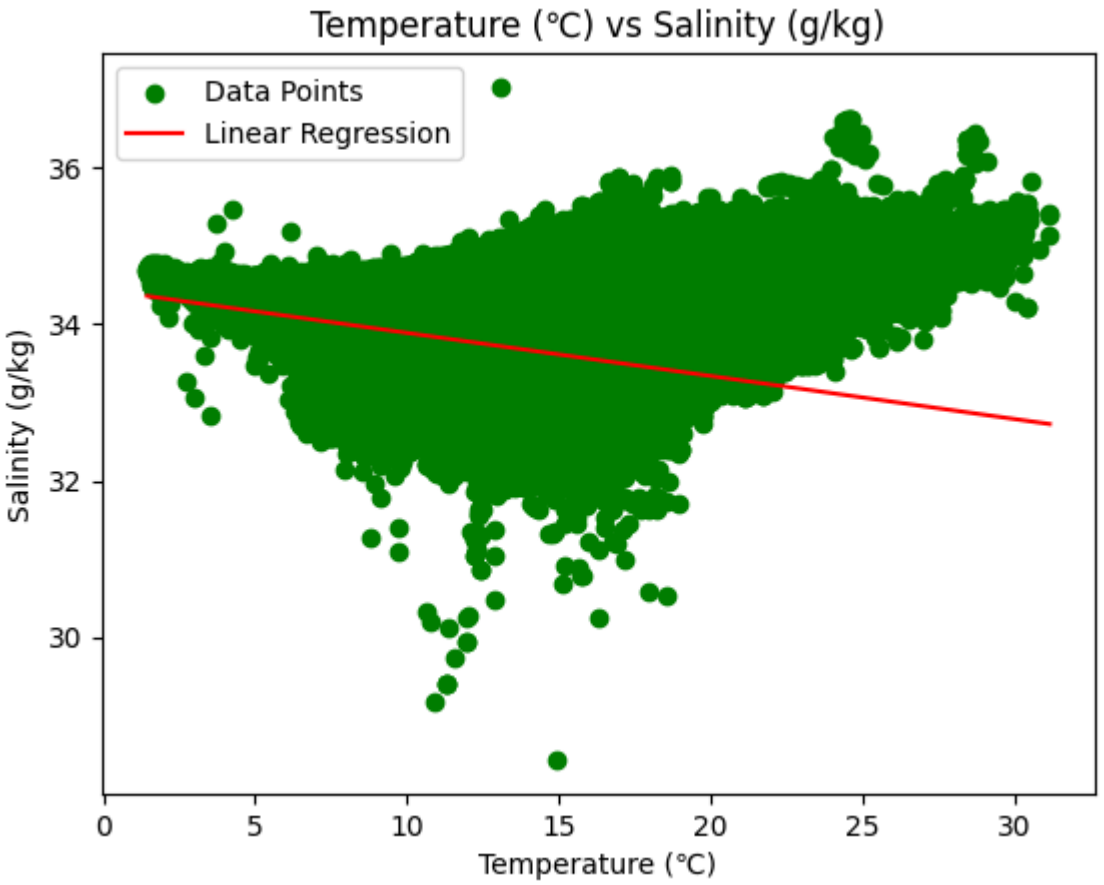
Show the figures and explain what certain features of the graphs tell us



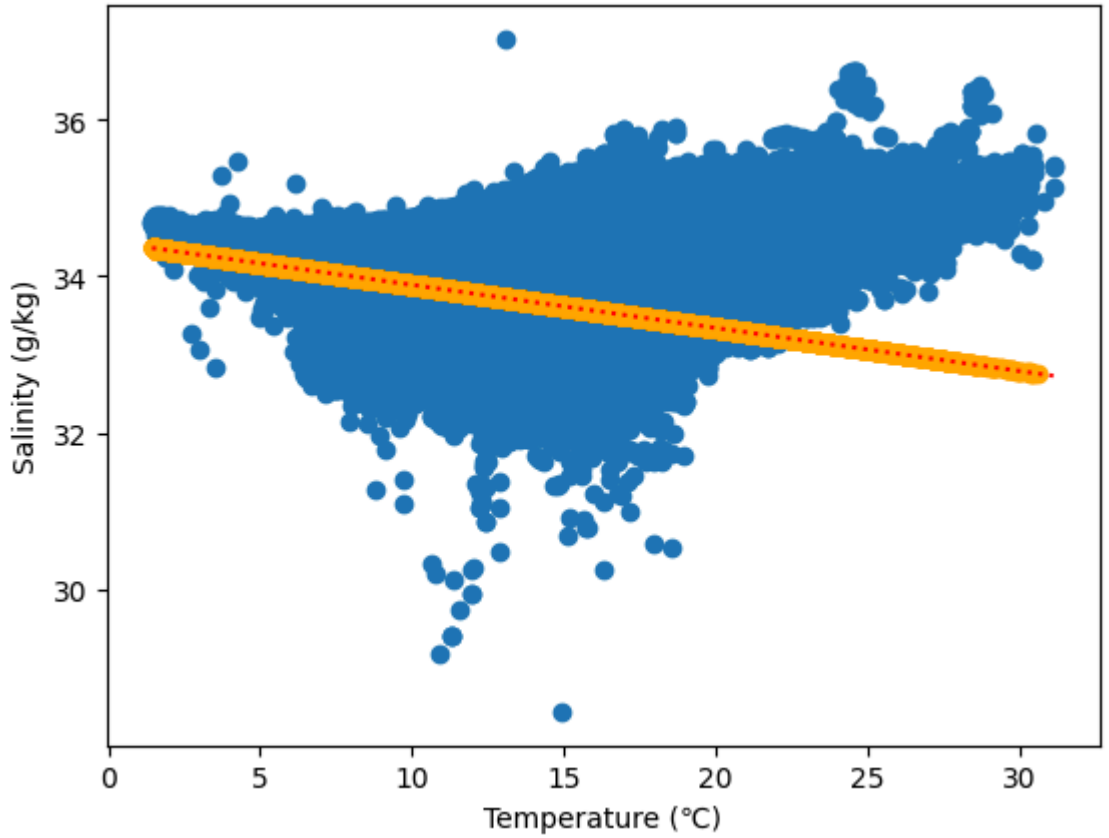
This graph portrays the overall data trend. We can see there's not a strong correlation between salinity and temperature. Additionally, it's hard to describe the trend because of the amount of data points.



These histograms portray the distribution of data points for both temperature and salinity. Temperature is close to a normal distribution that's positively skewed. Salinity has a similar case, however, it's not as skewed as temperature. Additionally, temperature values exhibit a greater variance than salinity.



This graph portrays a scatter plot of temperature against salinity and includes the linear regression model. As you can see, the data doesn't have a strong correlation and there's no sign of a relationship between the salinity and temperature of the water. Nevertheless, the overall relationship between the two seems to be a negative correlation.



However, the predictions made from our linear regression model are generally accurate. We can also see this when we evaluate our model. Because of our large dataset of almost a million points, the model is overfitted. Even though there's no correlation between salinity and temperature, the predictions tend to be accurate. So I would conclude our model is biased and that leads to a false impression of better model accuracy.

Make some conclusions from your analysis, and describe where you could go in the future (another analysis, create a model, …)

In conclusion, we cannot predict the salinity of ocean water based on its temperature. The correlation between the 2 is moderate and negative. Our model was accurate when making predictions from the test data. However, this might've been a cause of overfitting, since the test data came from the same dataset as the training data. From here, I believe it would be best to develop another type of model, such as a multilinear regression model, and take other variables into account for predicting water salinity. Another option would be to develop another linear regression model with a different variable than temperature. Also, it would be optimal to not use the same dataset for model training and evaluation.