American University of Central Asia
7/6 Aaly Tokombaev, 720060,
Bishkek, Kyrgyzstan
ИНН 01407199310022 | 999 УККН
www.auca.kg

Understanding Water: How Temperature, Depth, and Oxygen Affect Salinity

Course: Theory of Probabilities and Statistics

Student: Aikokul Tashpulatova

Professor: Polina Dolmatova

Link for the source code: Here

Problem

Investigate the correlation between temperature, depth, oxygen levels, and salinity in water.

The problem at hand involves determining whether it is possible to forecast the salinity of a given environment using temperature, depth, and oxygen levels as parameters. Salinity, representing the concentration of dissolved salts in water, holds significance in aquatic ecosystems due to its influence on water density, buoyancy, and the distribution of marine organisms. The challenge is to decipher the complex relationships between temperature, depth, and oxygen levels and their combined impact on salinity. This predictive modeling task requires a multidisciplinary approach, combining insights from oceanography, environmental science, and data analysis to uncover nuanced patterns governing salinity dynamics in aquatic systems.

Mathematical Model

Regression.

The mathematical model for addressing this problem involves employing regression techniques. Regression analysis is a statistical method used to examine the relationship between a dependent variable, in this case, salinity, and one or more independent variables, such as temperature, depth, and oxygen levels. By fitting a regression model to the data, the goal is to establish a mathematical equation that can predict salinity based on the given input parameters. The challenge lies in identifying the most appropriate regression model and accurately estimating its parameters to achieve a reliable and robust predictive tool. This requires a rigorous analysis of the dataset, consideration of potential interactions between variables, and validation of the model's predictive performance. The mathematical model serves as a key tool in unraveling the complexities of the interplay between temperature, depth, and oxygen levels in predicting salinity in aquatic environments.

Population

The study focuses on water samples from oceans and lakes worldwide to investigate if salinity can be predicted based on temperature, depth, and oxygen levels. These samples represent diverse environments, helping us understand the relationships between these factors and salinity. Using a global dataset ensures the model's applicability across different water bodies and emphasizes the need for a comprehensive approach in predicting salinity variations.

Sampling method

Simple Random Sampling:

We are using simple random sampling, where every item in the population has an equal chance of being selected. This ensures fairness and represents the overall population characteristics in our sample.

Based on the existing dataset I am using, sampling methods differ. The CalCOFI data set uses more than 50,000 sampling stations by conducting quarterly cruises off southern & central California.

Collecting Data

In my project, I used an existing dataset from Kaggle. Here is the link for this dataset.

"The California Cooperative Oceanic Fisheries Investigations (CalCOFI) is a unique partnership of the California Department of Fish & Wildlife, NOAA Fisheries Service and Scripps Institution of Oceanography. The organization was formed in 1949 to study the ecological aspects of the sardine population collapse off California. The CalCOFI data set represents the longest (1949-present) and most complete (more than 50,000 sampling stations) time series of oceanographic and larval fish data in the world."
(Sohier Dane (Owner))

Data Representation

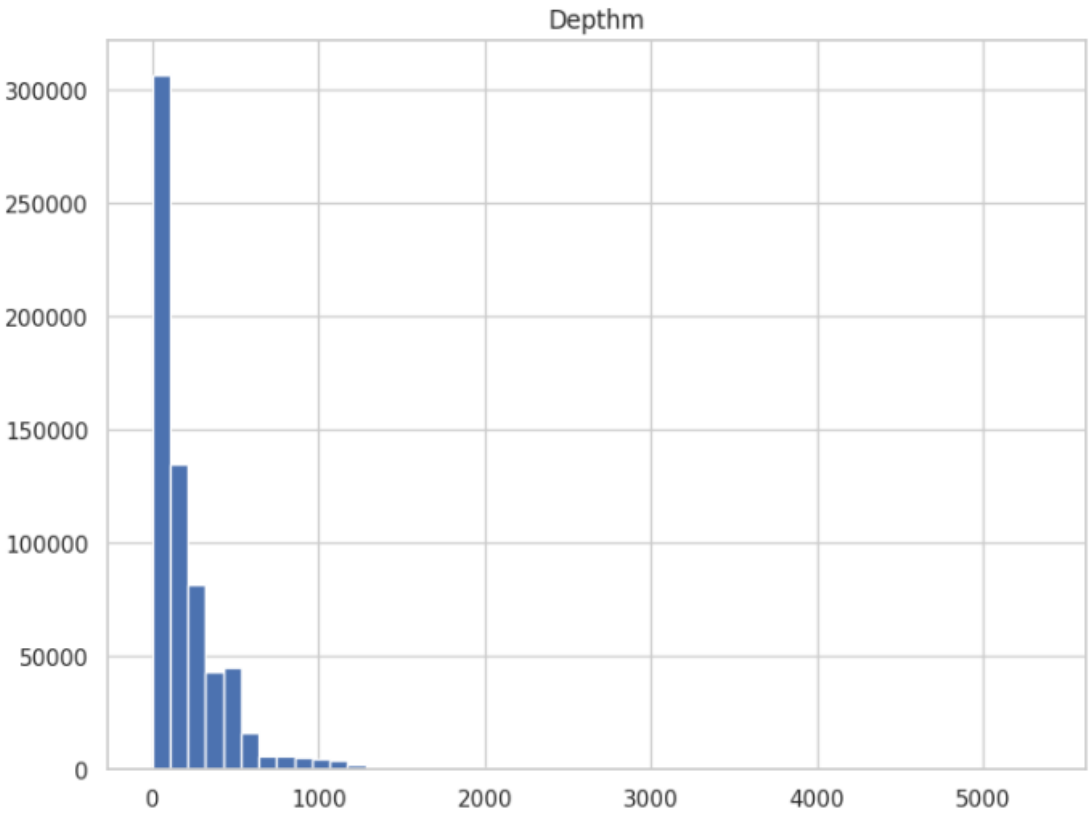
Rows: initial: 864863, after dropping not complete rows: 661489
Columns: 4

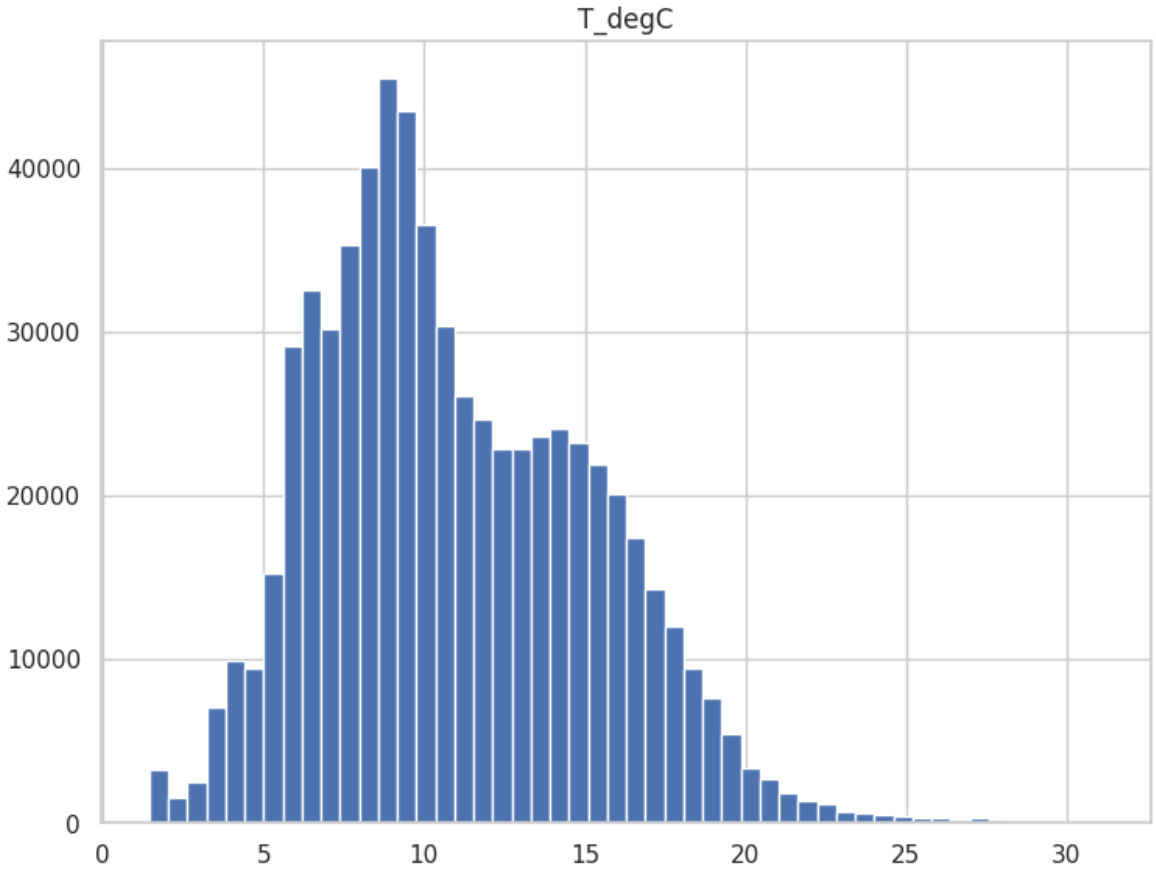
Depthm: Depth in meters
T_degC: Water temperature in degree Celsius
O2ml_L: O2 mixing ratio in ml/L
Salnty: Salinity in g of salt per kg of water (g/kg)

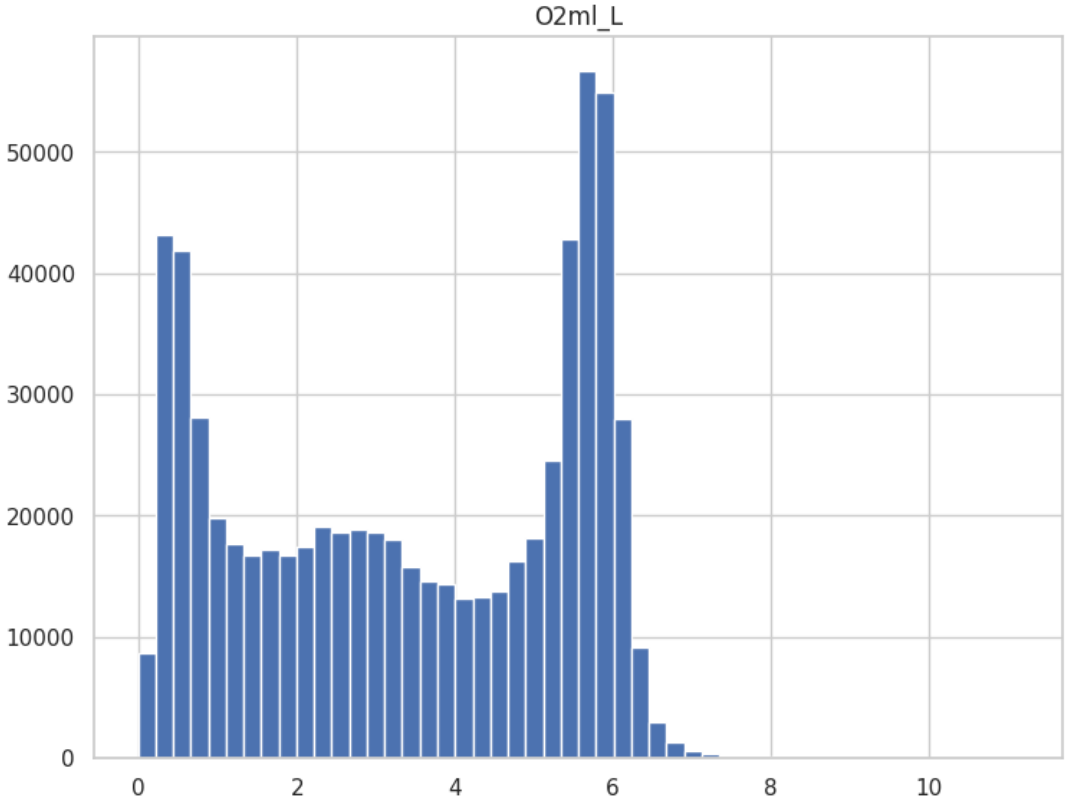
```
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Depthm  661489 non-null  int64
 1   T_degC  661489 non-null  float64
 2   O2ml_L  661489 non-null  float64
 3   Salnty  661489 non-null  float64
dtypes: float64(3), int64(1)
```

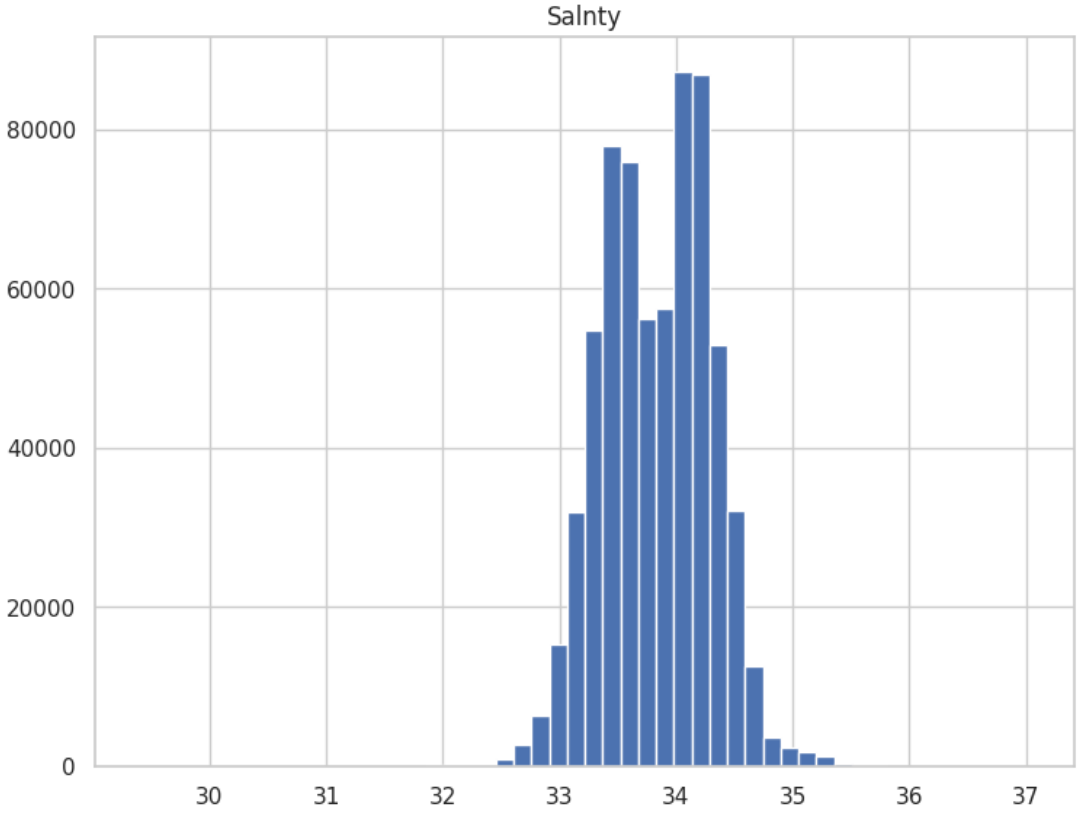
	A	B	C	D
1	Depthm	T_degC	O2ml_L	Salnty
2	0	10.5		33.44
3	8	10.46		33.44
4	10	10.46		33.437
5	19	10.45		33.42
6	20	10.45		33.421
7	30	10.45		33.431
8	39	10.45		33.44
9	50	10.24		33.424
10	58	10.06		33.42
11	75	9.86		33.494
12	78	9.83		33.51
13	100	9.67		33.58
14	117	9.5		33.64
15	125	9.32		33.689
16	150	8.76		33.847

	Depthm	T_degC	O2ml_L	Salnty
count	661489.000000	661489.000000	661489.000000	661489.000000
mean	219.685874	10.918816	3.416704	33.832682
std	311.150576	4.224876	2.068723	0.460343
min	0.000000	1.440000	-0.010000	29.402000
25%	49.000000	7.809000	1.400000	33.479200
50%	125.000000	10.160000	3.470000	33.853000
75%	300.000000	14.020000	5.513000	34.184000
max	5351.000000	31.140000	11.130000	37.034000


Depthm
300000
250000
200000
150000
100000
50000
0
0
1000
2000
3000
4000
5000

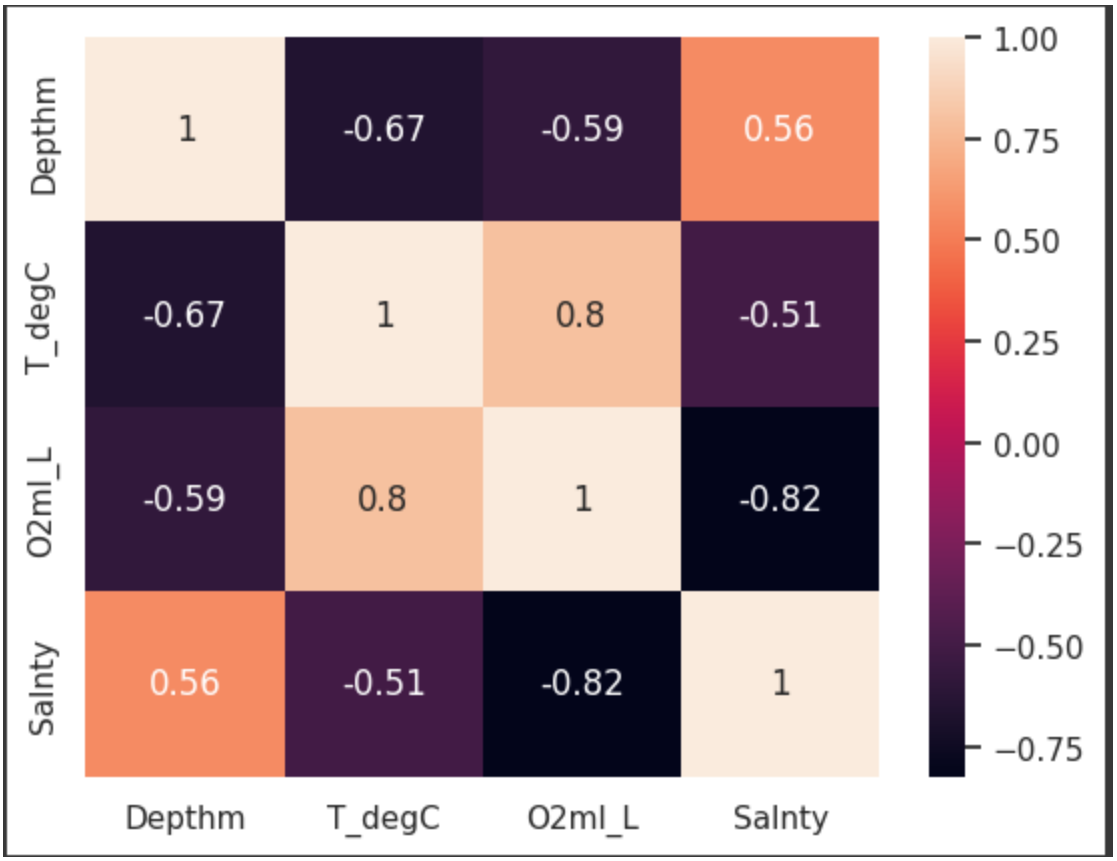

T_degC
40000
30000
20000
10000
0
0
5
10
15
20
25
30


O2ml_L
50000
40000
30000
20000
10000
0
0
2
4
6
8
10


Salnty
80000
60000
40000
20000
0
30
31
32
33
34
35
36
37

Data analysis

Data correlation heatmap: The lighter the stronger connection, the darker, the weaker connection



P-Value calculation:

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	33.9064	0.0011	30296.2282	0.0000	33.9042	33.9086
Depthm	0.0004	0.0000	377.3689	0.0000	0.0004	0.0004
T_degC	0.0613	0.0001	549.4534	0.0000	0.0610	0.0615
O2ml_L	-0.2449	0.0002	-1164.7731	0.0000	-0.2453	-0.2445

According to the table P>|t| all of the features are above 0.05 which means that each attribute contributes to the model prediction. Therefore, we include all attributes in the model.

Model:

Train and Test data proportion: 66% - 33%
I used the LinearRegression model from the Sklearn library.

Excerpt from the model code:

```
X = data.loc[:, data.columns != "Salnty"]
y = data.loc[:, data.columns == "Salnty"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

cols=["T_degC", "Depthm", "O2ml_L"]
X=X_train[cols]
y=y_train["Salnty"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
lin = LinearRegression()
lin.fit(X_train, y_train)
y_pred = lin.predict(X_test)
```

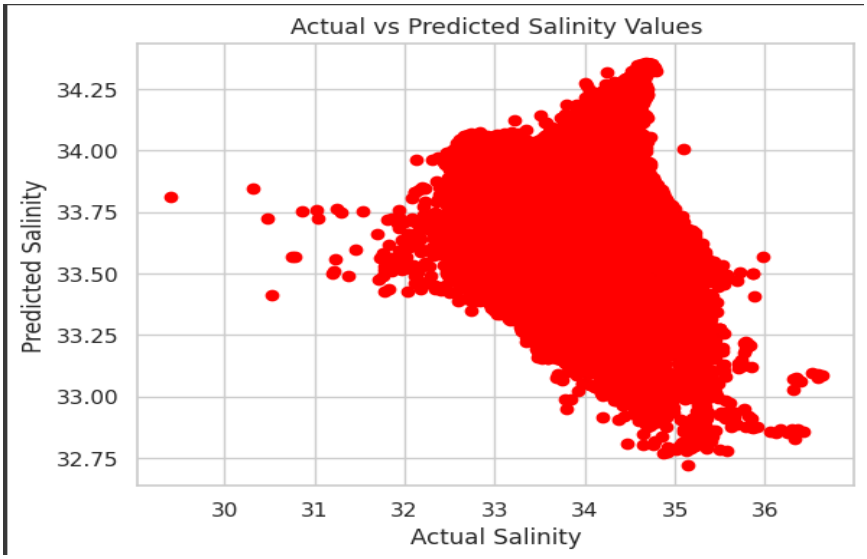
Prediction examples:
When Depth Temperature O2 are [0, 25.0, 2.5] respectively, the prediction value for Salinity is: [33.3037091]
When Depth Temperature O2 are [100, 10.0, 2.0] respectively, the prediction value for Salinity is: [39.55612661]

Evaluation:

R squared -> The more value means better prediction
Mean Squared Error(MSE) -> The less value means better prediction
Root Mean Squared Error(RMSE) -> The less value means better prediction
Mean Absolute Error(MAE) -> The less value means better prediction

R squared measures the proportion of the variance in the dependent variable (salinity) that is predictable from the independent variables (temperature, depth, and oxygen levels). A higher R squared indicates a better predictive model. Additionally, we will consider MSE, RMSE, and MAE. Lower values for MSE, RMSE, and MAE signify better predictive accuracy, reflecting how closely the model's predictions align with the actual salinity values. These evaluation metrics collectively provide a comprehensive assessment of the model's effectiveness in predicting salinity based on the given parameters.

Evaluation: Temperature - Salinity relationship model prediction



R squared of test set: 0.26

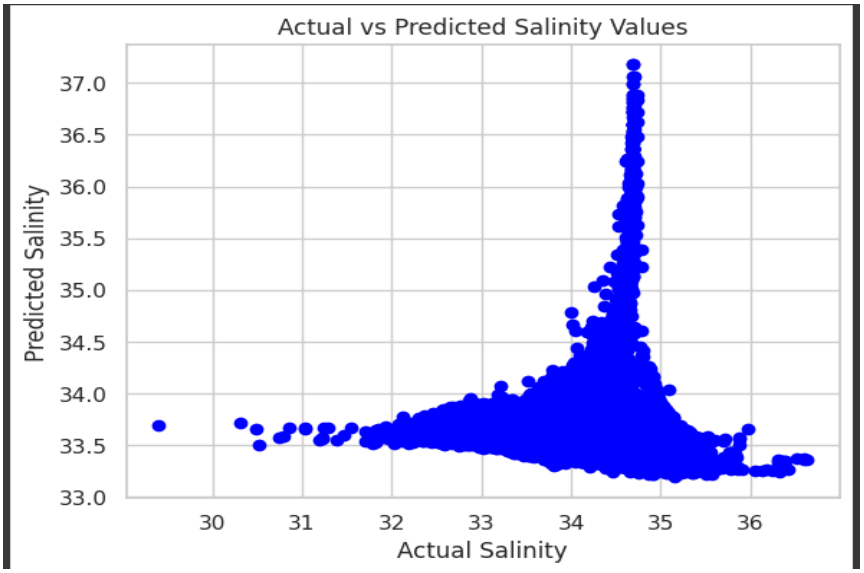
Mean Squared Error(MSE) of the test set: 0.16

Root Mean Squared Error(RMSE) of the test set: 0.40

Mean Absolute Error(MAE) of the test set: 0.28

Inference: The relationship between Temperature and Salinity is very weak.

Evaluation: Temperature + Depth - Salinity relationship model prediction



R squared of test set: 0.35

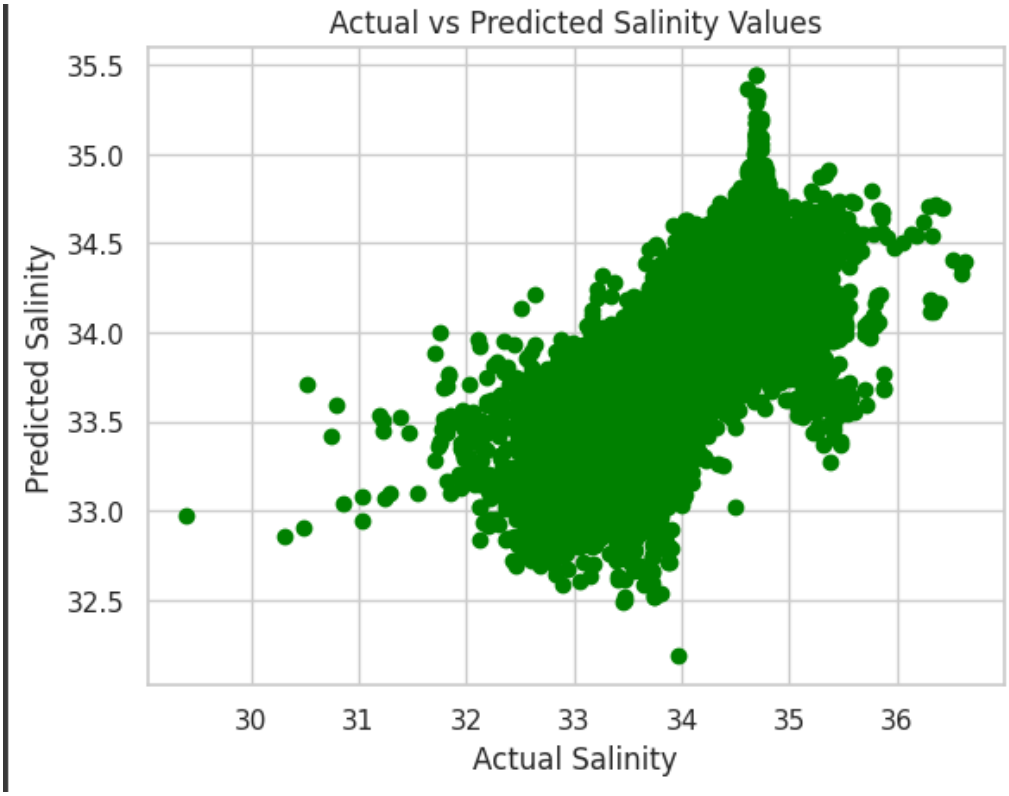
Mean Squared Error(MSE) of the test set: 0.14

Root Mean Squared Error(RMSE) of the test set: 0.37

Mean Absolute Error(MAE) of the test set: 0.26

Inference: The relationship between Temperature, Depth and Salnty is weak, but better than the previous model.

Evaluation: Temperature, Depth, Oxygen - Salinity relationship model prediction



R squared of test set: 0.78
Mean Squared Error(MSE) of the test set: 0.05
Root Mean Squared Error(RMSE) of the test set: 0.21
Mean Absolute Error(MAE) of the test set: 0.14

Inference: The relationship between Temperature, Depth, Oxygen, and Salinity is strong, as evidenced by the evaluation metrics. The high R squared value of 0.78 and low values for Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) on the test set (0.05, 0.21, and 0.14, respectively) indicates a high proportion of the variability in salinity can be explained by the combined influence of temperature, depth, and oxygen levels. This implies that considering all three variables significantly improves the model's performance in capturing the complex interplay that determines salinity in aquatic environments. The integration of oxygen levels enhances the predictive capacity, demonstrating the importance of a multidimensional approach for a more accurate understanding of salinity dynamics.

Comparison with actual data/information:

Dataset I used for this research project is actual data collected from The California Cooperative Oceanic Fisheries Investigations (CalCOFI) . The organization has been present since 1949 and was collecting various data related to the ocean.

Even though I couldn't find specific studies on all four things I'm looking at (depth, temperature, oxygen, and salinity), I did find research on how each of these things relates to salinity separately. This is really helpful for my study because it gives me a better understanding of how all three factors together affect salinity.

For example, when we look at how oxygen and salinity are connected, we find that when salinity goes up, the amount of oxygen in the water goes down (Atlas Scientific, 2022).

In terms of depth and salinity, a study mentioned that as the depth increases, the salinity value also goes up (Leidonald et al., 2018). This helps us see how salinity changes at different depths in the water.

And when it comes to temperature and salinity, one study found that as temperature goes down, both salinity and dissolved oxygen go up (Khan & Rajshekhar, 2020). This gives us insight into how temperature affects the balance of these important factors in the water.

In essence, each parameter—depth, temperature, and oxygen—holds a unique link to salinity. When considered together, these connections provide a comprehensive understanding of what influences the saltiness of water. *It's akin to solving a puzzle, where each piece contributes to the overall picture of how salinity operates in the ocean, revealing the intricate interplay of these factors.*

List of Reference

Tashpulatova, Aikokul. "Understanding Water: How Temperature, Depth, and Oxygen Affect Salinity." Source code: https://colab.research.google.com/drive/16Ye_yxBLnTH5F-hTo43tcZL029rHt7N2#scrollTo=zCv9nEAd2Nup

Sohier Dane (Owner). "The California Cooperative Oceanic Fisheries Investigations (CalCOFI)." Kaggle Dataset. Available https://www.kaggle.com/datasets/sohier/calcofi

Atlas Scientific. (2022). "What Is The Relationship Between Dissolved Oxygen And Salinity?" https://atlas-scientific.com/blog/dissolved-oxygen-and-salinity/#:~:text=The%20solubility%20of%20oxygen%20

Leidonald, R., et al. A Muhtadi, I Lesmana, Z A Harahap and A Rahmadya (2018). "Profiles of temperature, salinity, dissolved oxygen, and pH in Tidal Lakes." https://www.researchgate.net/publication/333663616_Profiles_of_temperature_salinity_dissolved_oxygen_and_pH_in_Tidal_Lakes

Khan, M. Z. H., & Rajshekhar, A. (2020). "Relation Among Temperature, Salinity, pH and DO of Seawater Quality" https://media.neliti.com/media/publications/354071-relation-among-temperature-salinity-ph-a-3e52c28c.pdf