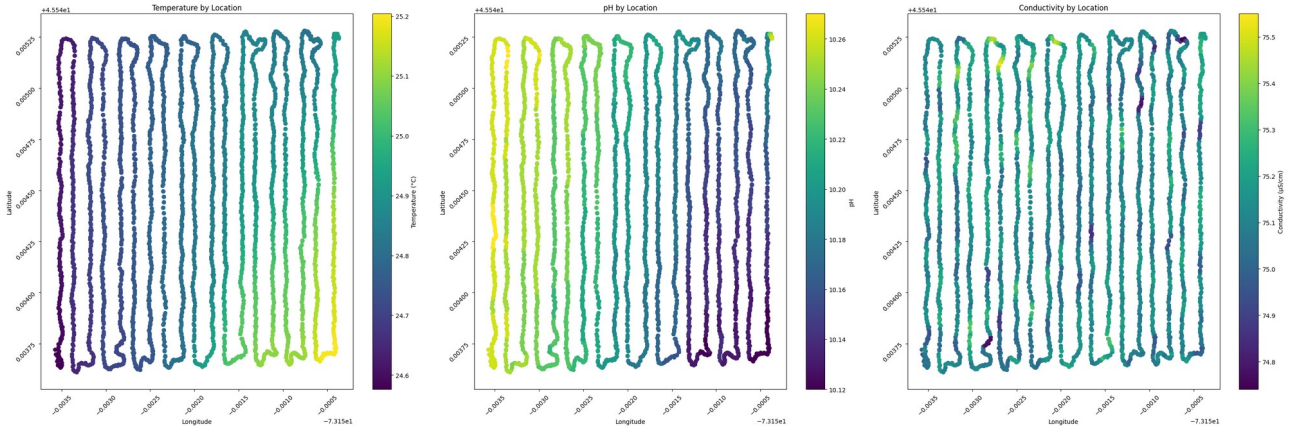
DATA SCIENCE
PLAKSHA UNIVERSITY
USMAN AKINYEMI
ASSIGNMENT 8.

Question 2 (1) :-

%time float64 – represent the time of measurement - Numerical/Continuous
field.header.seq int64 – Serial number of the measurement – Nominal
field.header.stamp float64 - Timestamp associated with the measurement - Numerical/Continuous
field.stamp_sonde float64 - Relative timestamp - Numerical/Continuous
field.temp_c float64 – Temperature in degrees Celsius - Numerical/Continuous
field.spcond_u float64 - Specific Conductance in microsiemens per centimeter (µS/cm). - Numerical/Continuous
field.sal float64 - Salinity in parts per thousand (PPT). - Numerical/Continuous
field.ph float64 -pH level - Numerical/Continuous
field.orp float64 - Oxidation-Reduction Potential (ORP) in millivolts (mV). - Numerical/Continuous
field.depth_m float64 - Depth in meters. - Numerical/Continuous
field.turbidity_ntu float64 - Turbidity in Nephelometric Turbidity Units (NTU). - Numerical/Continuous
field.turbidity_fnu float64 - Turbidity in Formazin Nephelometric Units (FNU) - Numerical/Continuous
field.odo_percsat float64 - Dissolved Oxygen (ODO) percentage saturation. - Numerical/Continuous
field.odo_m float64 - Dissolved Oxygen (ODO) concentration in milligrams per liter (mg/L). - Numerical/Continuous
field.latitude float64 - Latitude coordinate of the sampling platform. - Numerical/Continuous
field.longitude float64 - Longitude coordinate of the sampling platform. - Numerical/Continuous

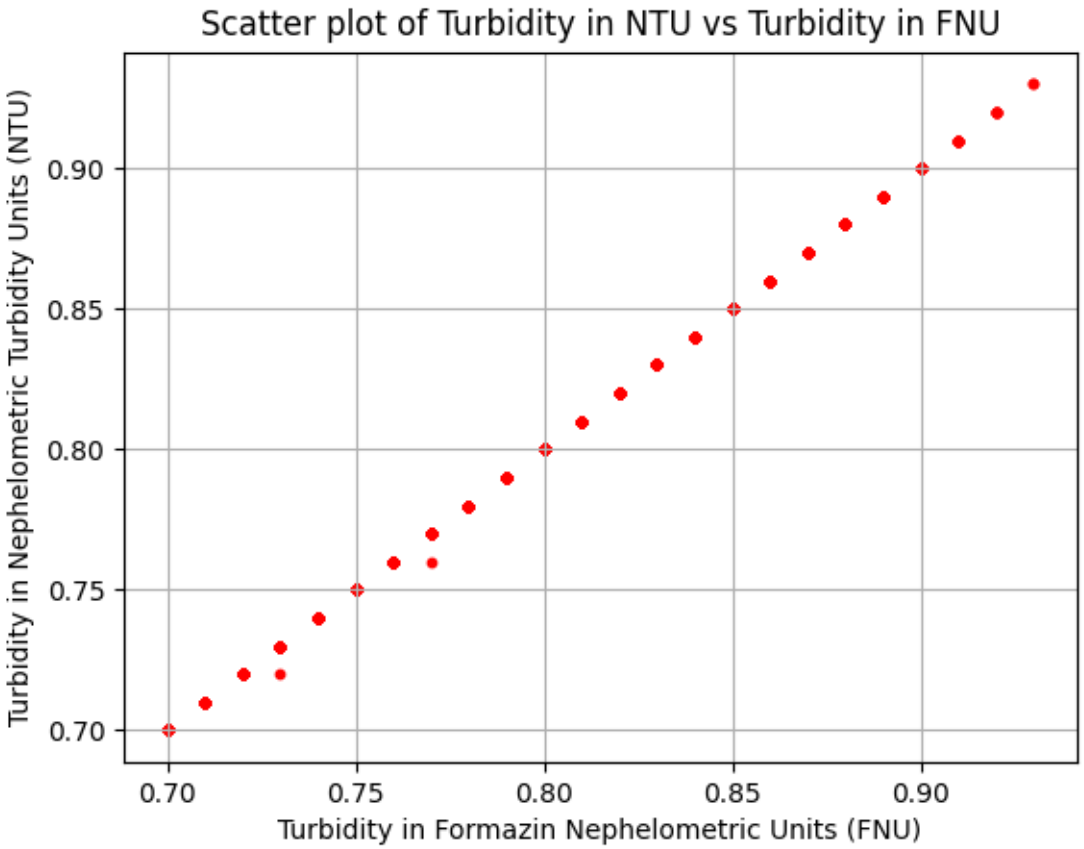
Question 2(2) – From my observation, I will say the unit time is in nanoseconds.

Question (3) 1



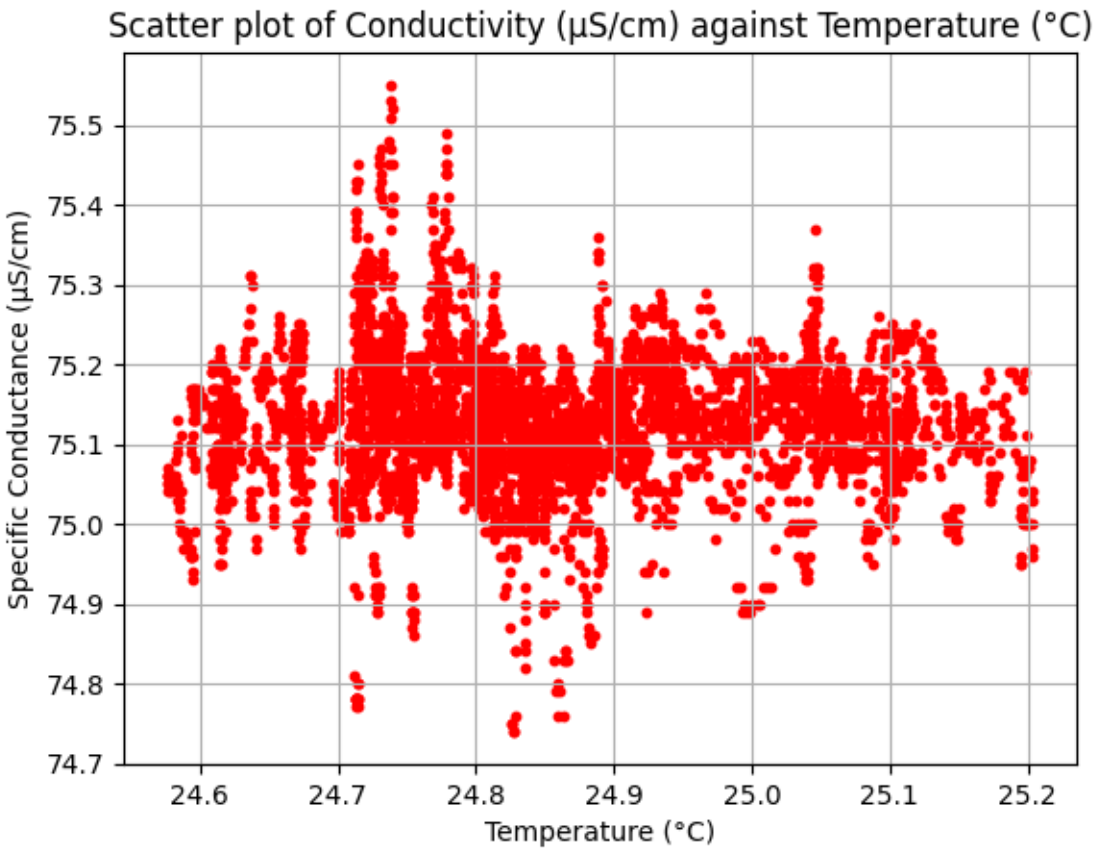
The range of the temperature in the location start from 24.6 – 25.2degree celcius which is around the normal room temperature. For the pH 10.12 – 10.26 and for the Conductivity it ranges from 74.8 to 75.2,

Question 3 (2)



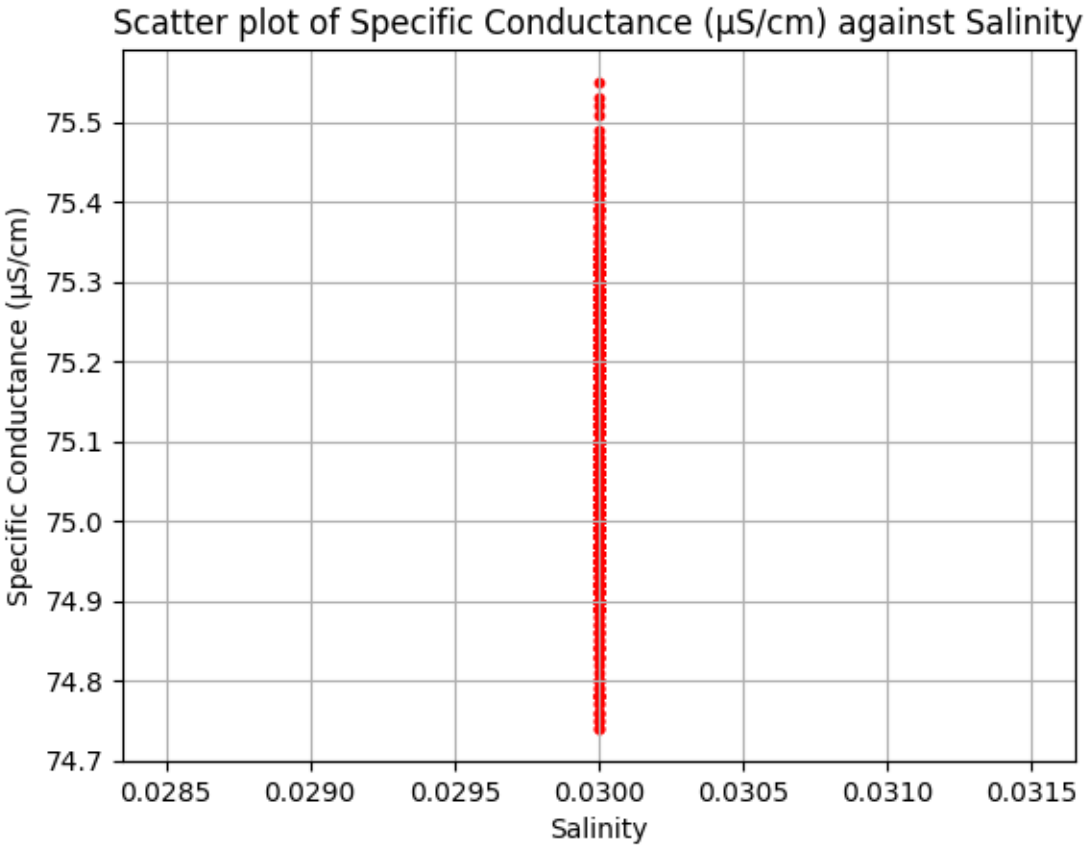
Observation:- The scatter plot shows that there is almost a perfect linear relationship between Turbidity NTU and Turbity FNU i.e both of them are linear combination of each other. Another way to explain this is that, the plot shows that if the Turbidity FNU increases, the Turbidity NTU also increases both in a positive direction.

Question 3 (3)a.



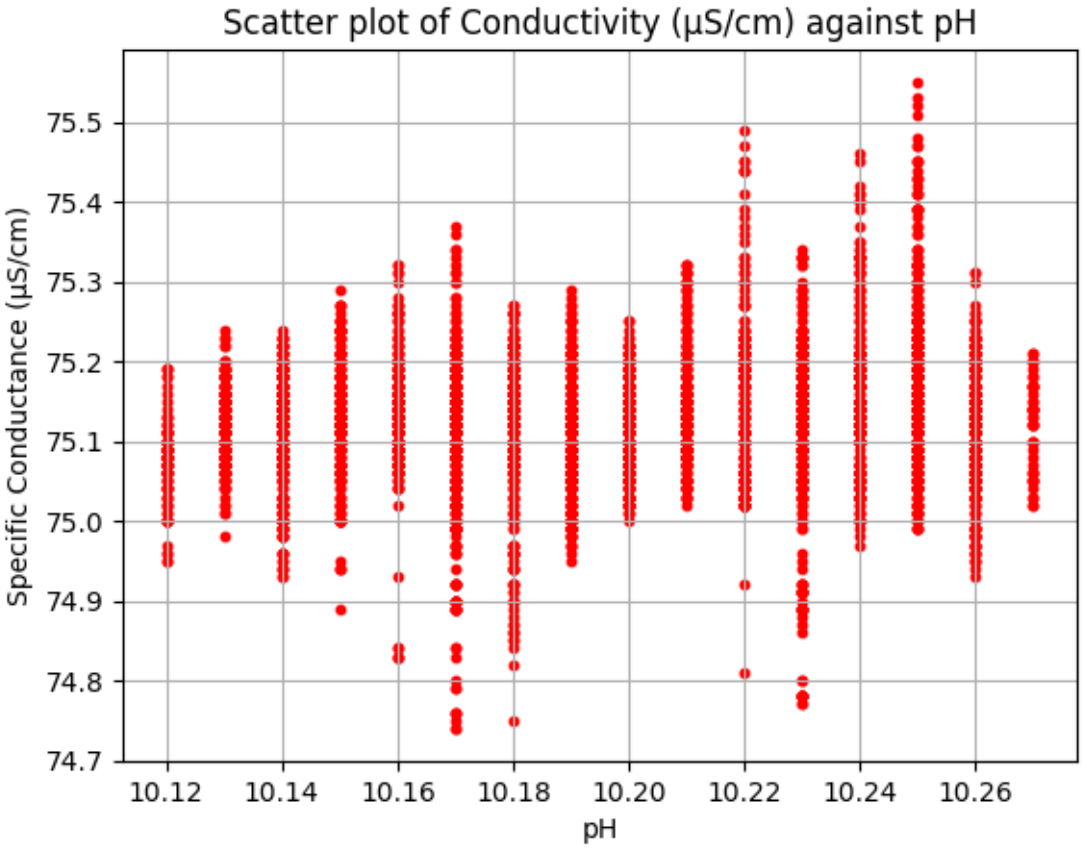
From this graph, I will say there is no really any visible relationship between the conductivity and the temperature

b.



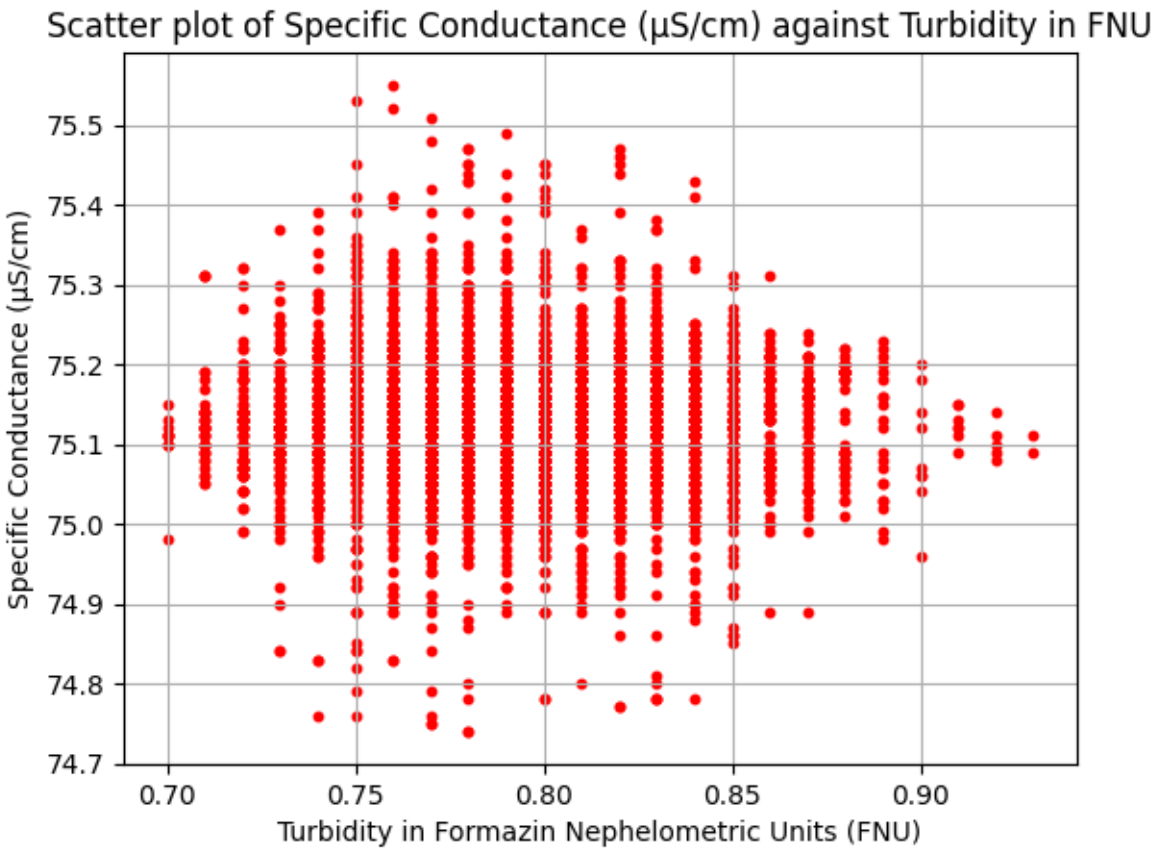
What I can infer from this plot is that, the Salinity is constant for all of the change in the specific conductance. This is also visible in the data values of Salinity. It is a constant value.

3c.



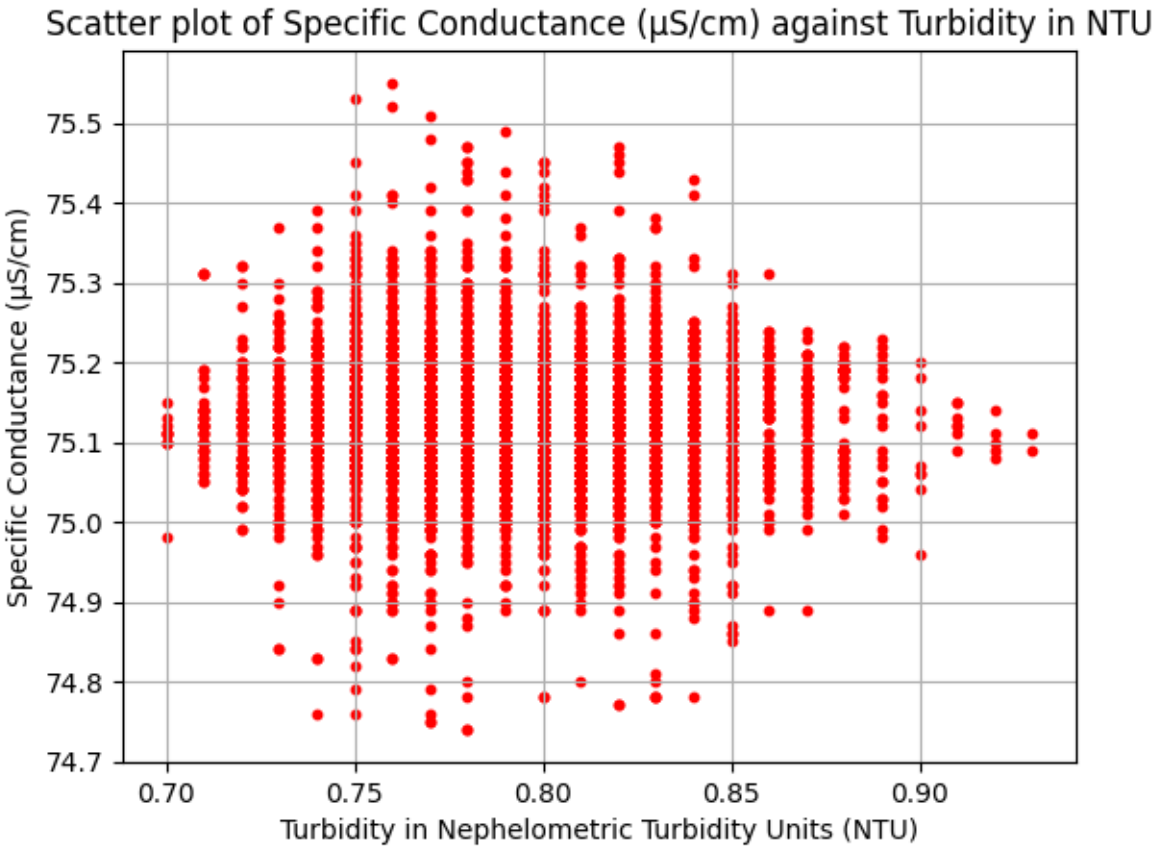
From my observation, there is no really any visible/visible relationship between the pH and the specific conductance.

3(3)d1



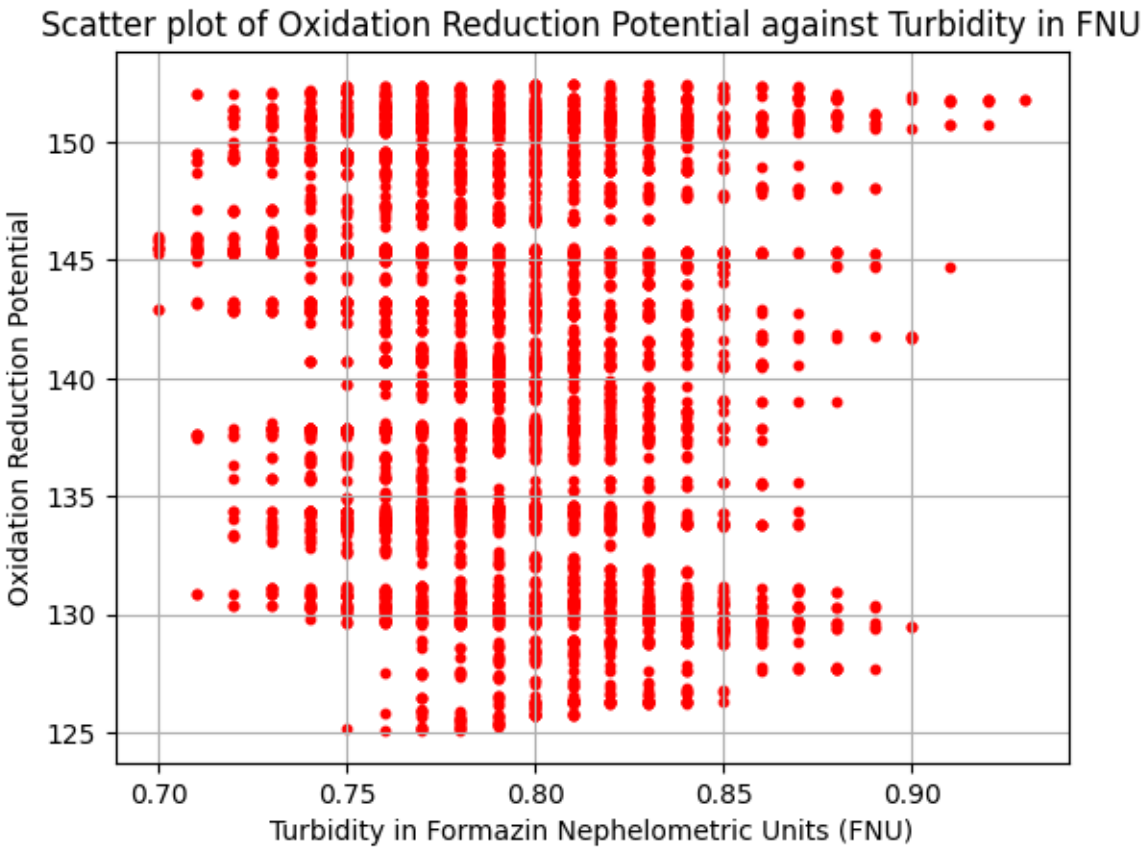
From my observations, there is not visible/linear relationship between the two parameters

3(3)d2



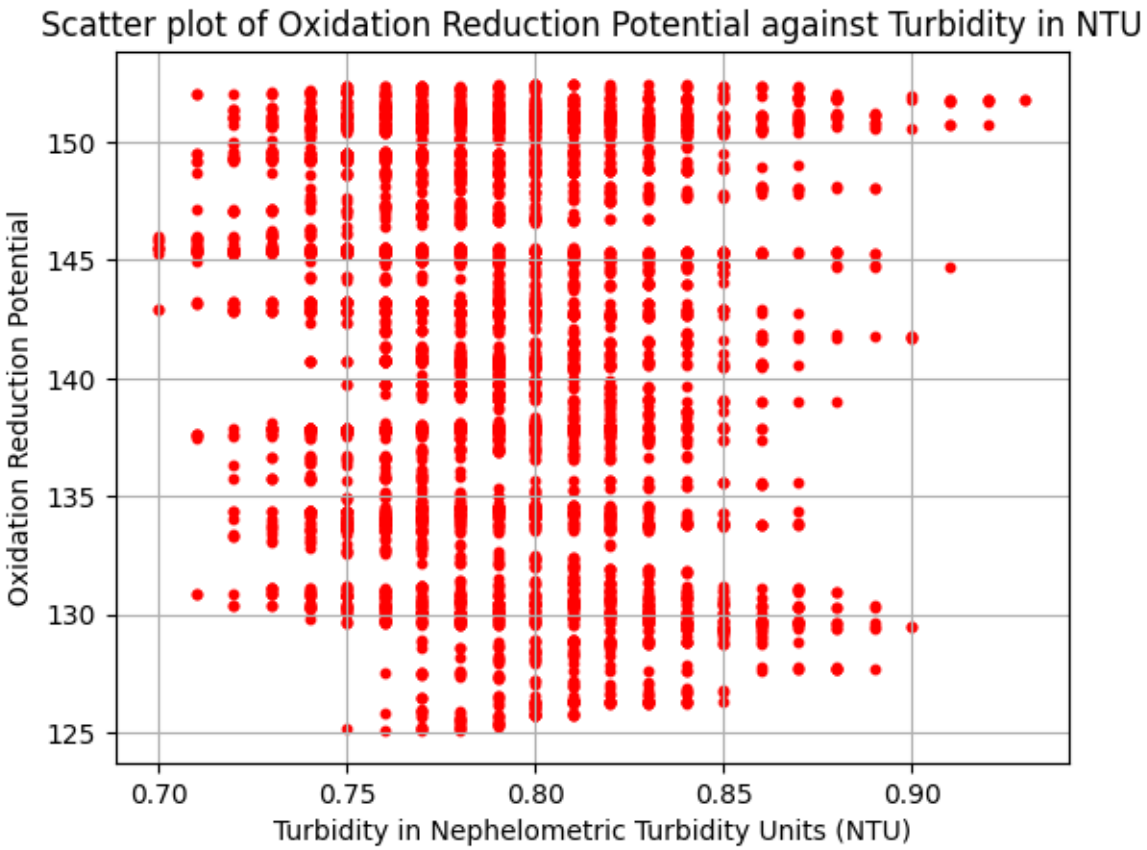
I cannot also see any visible/linear relationship between the two variables.

Question 3(4)a



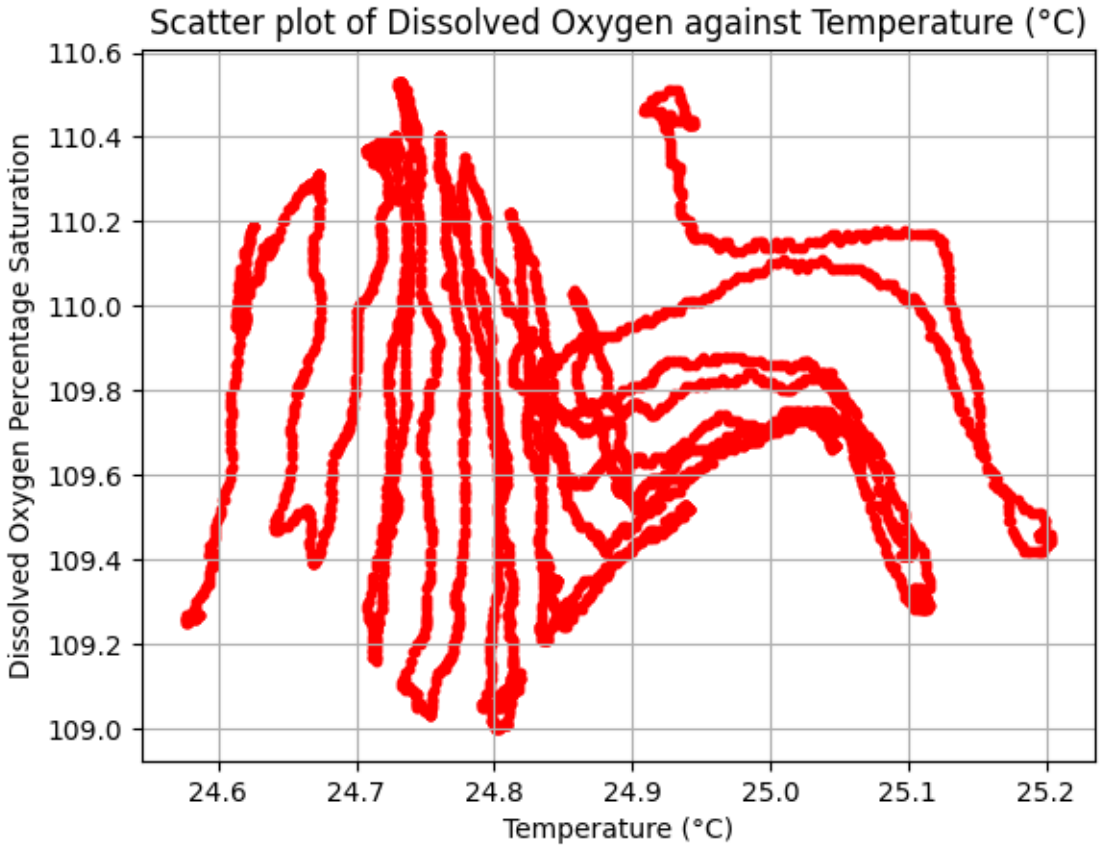
There is no visible linear relationship between the two plot.

Question 3(4)b



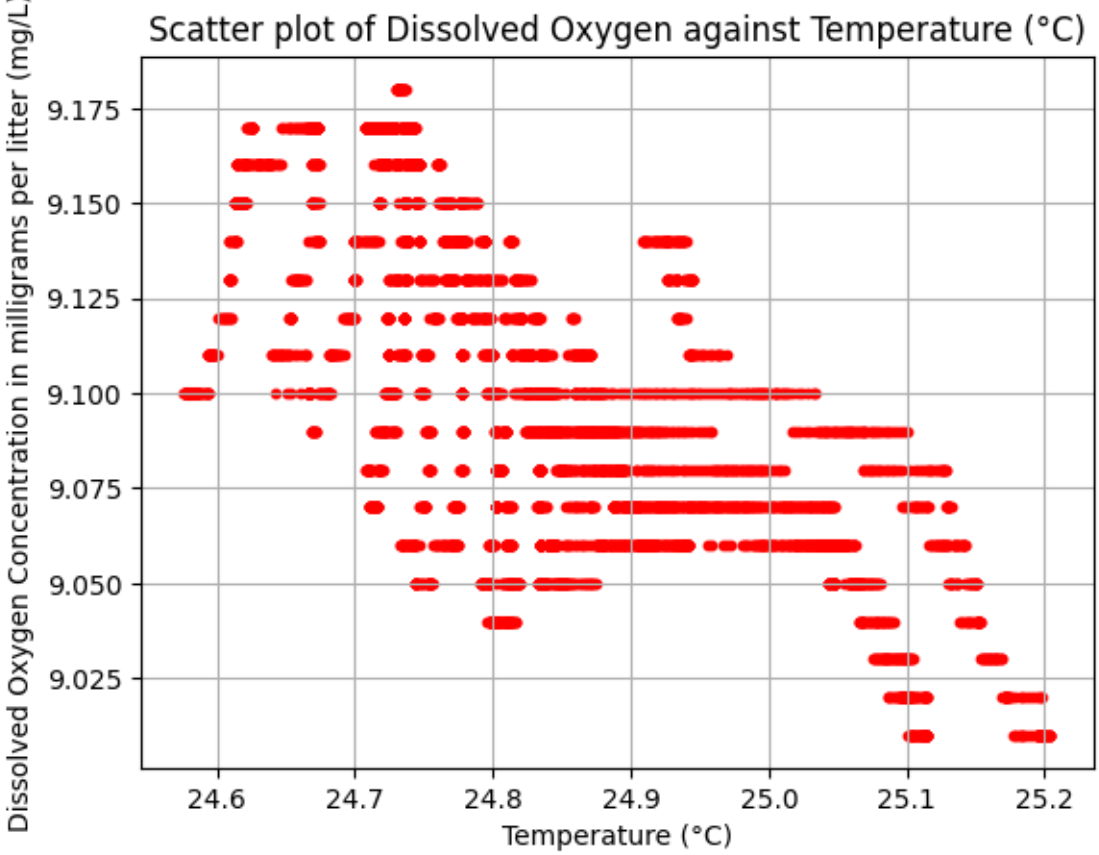
No visible linear relationship between the two variables.

Question 3 (5)a



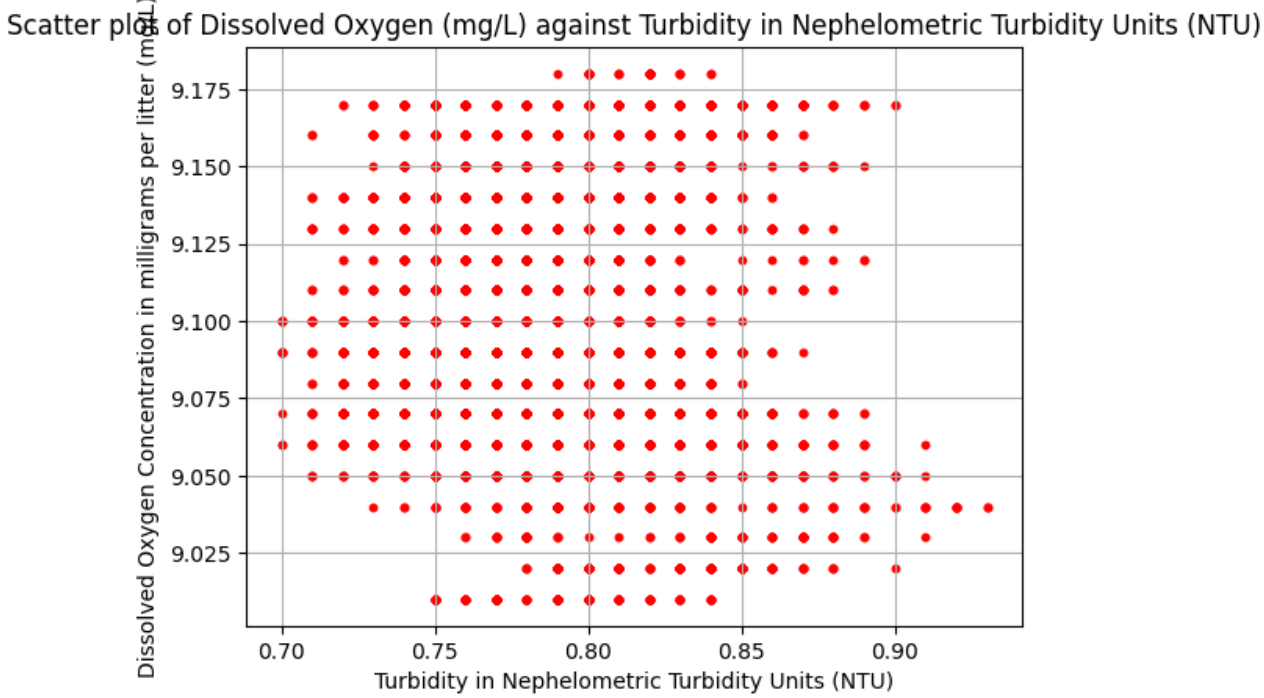
No visible linear relationship between the two variables.

Question 3 (5)b



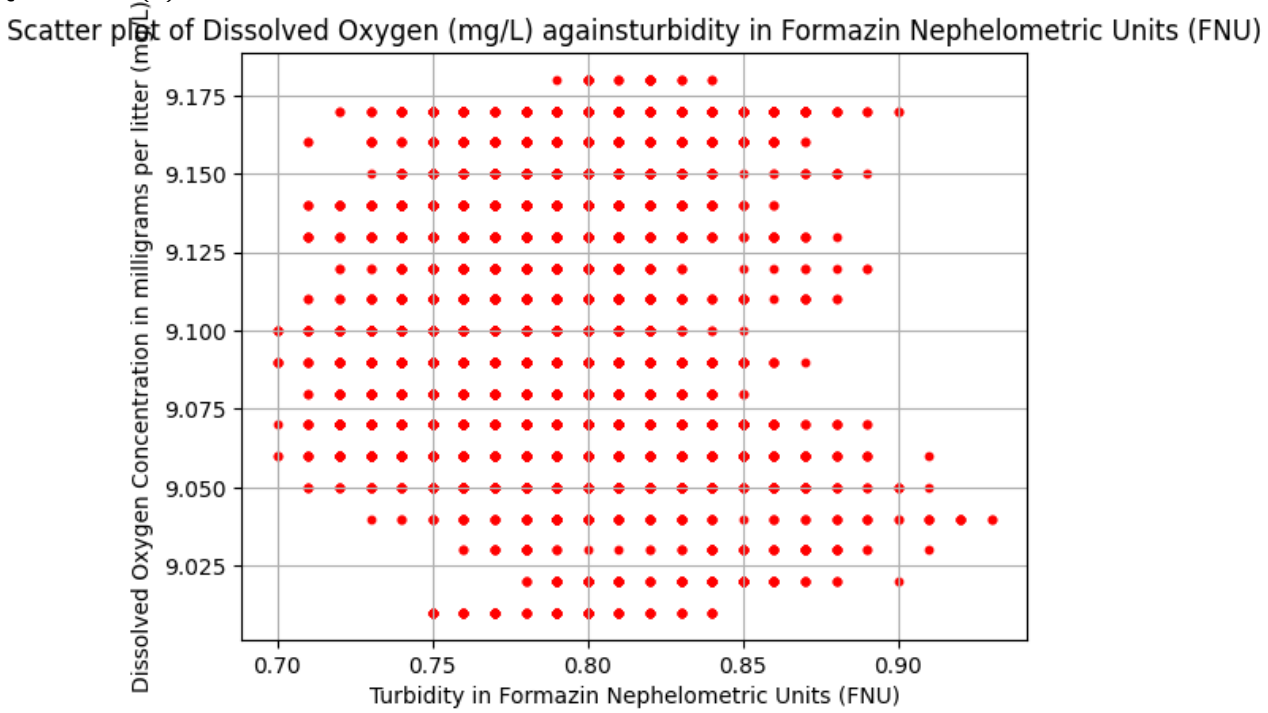
No visible linear relationship between the two variables.

Question 3 (5)C



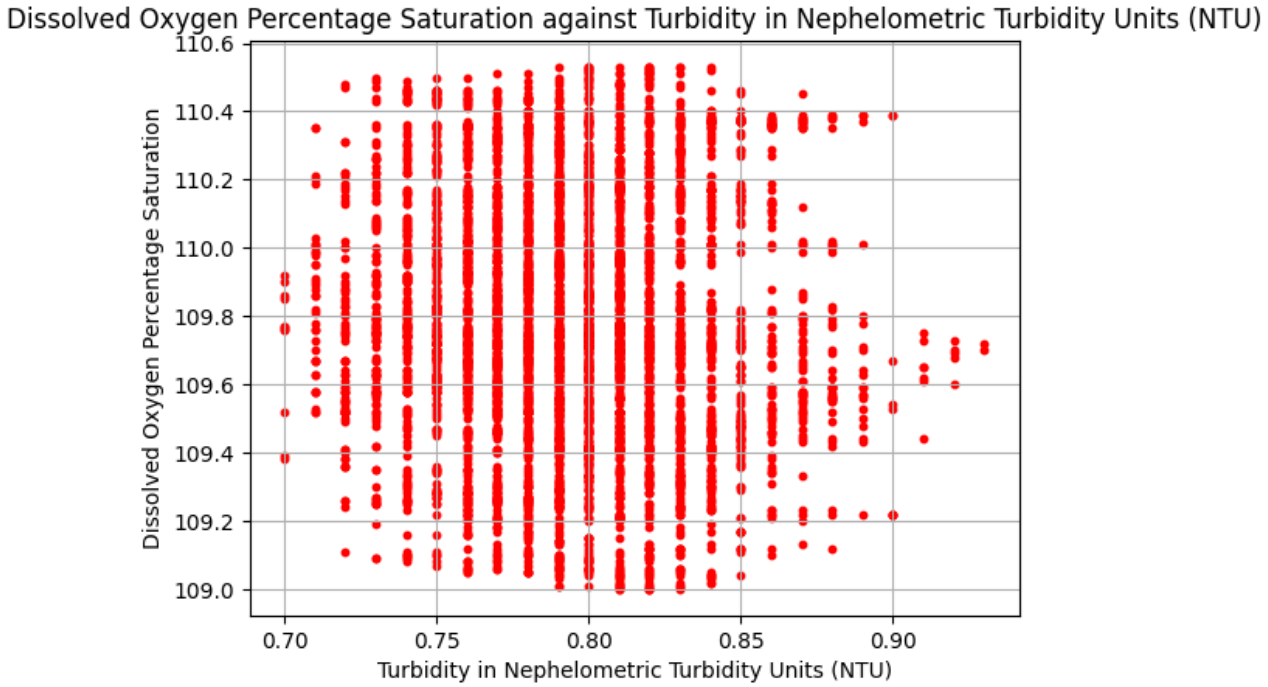
No visible linear relationship between the two variables.

Question 3 (5)d



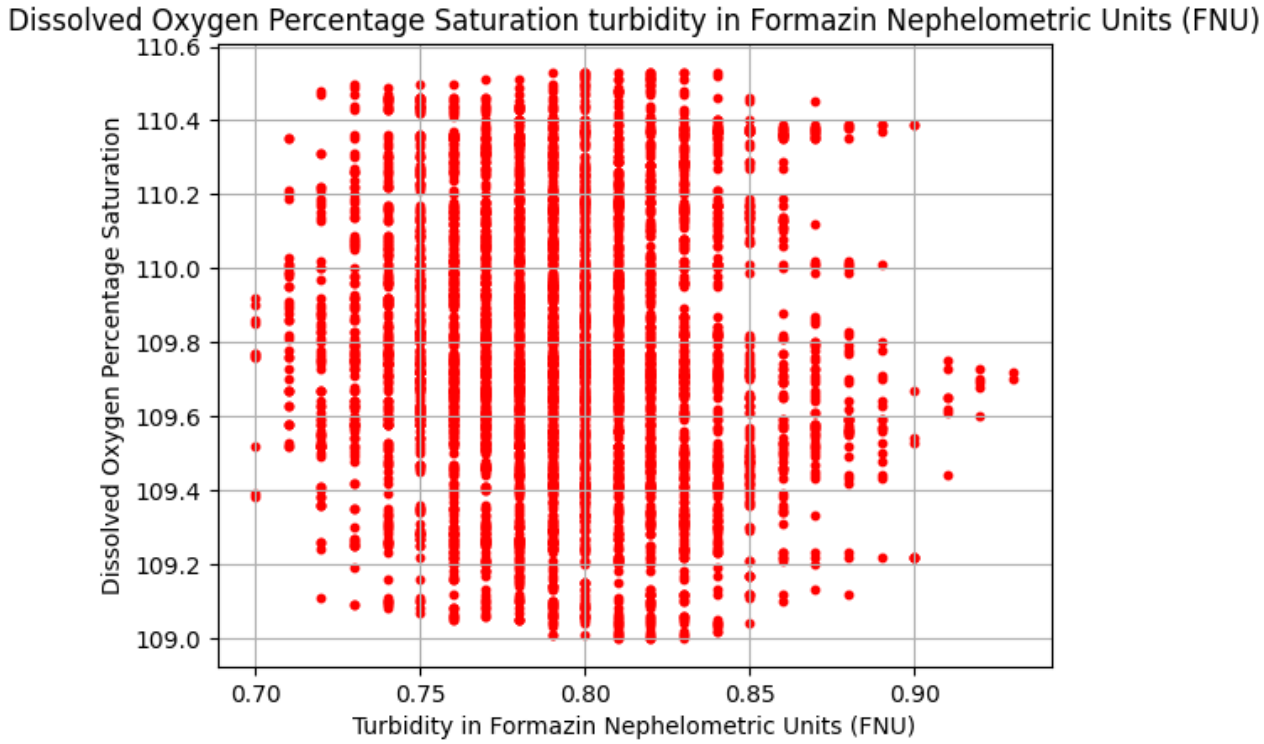
No visible linear relationship between the two variables.

Question 3 (5)e



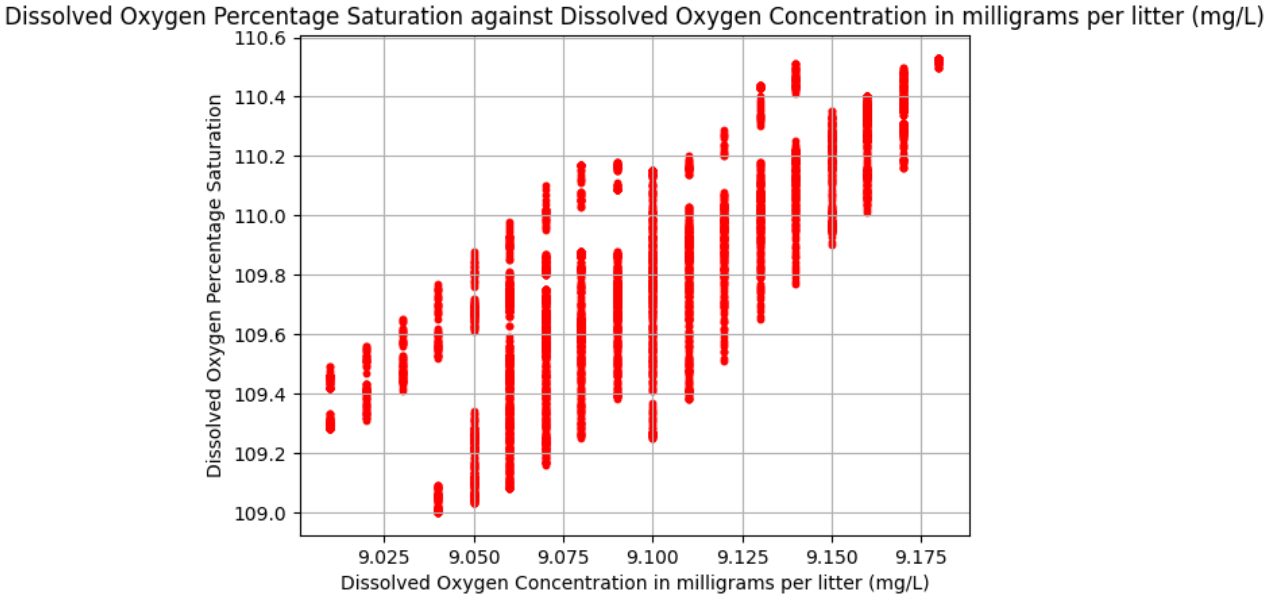
No visible linear relationship between the two variables.

Question 3 (5)f



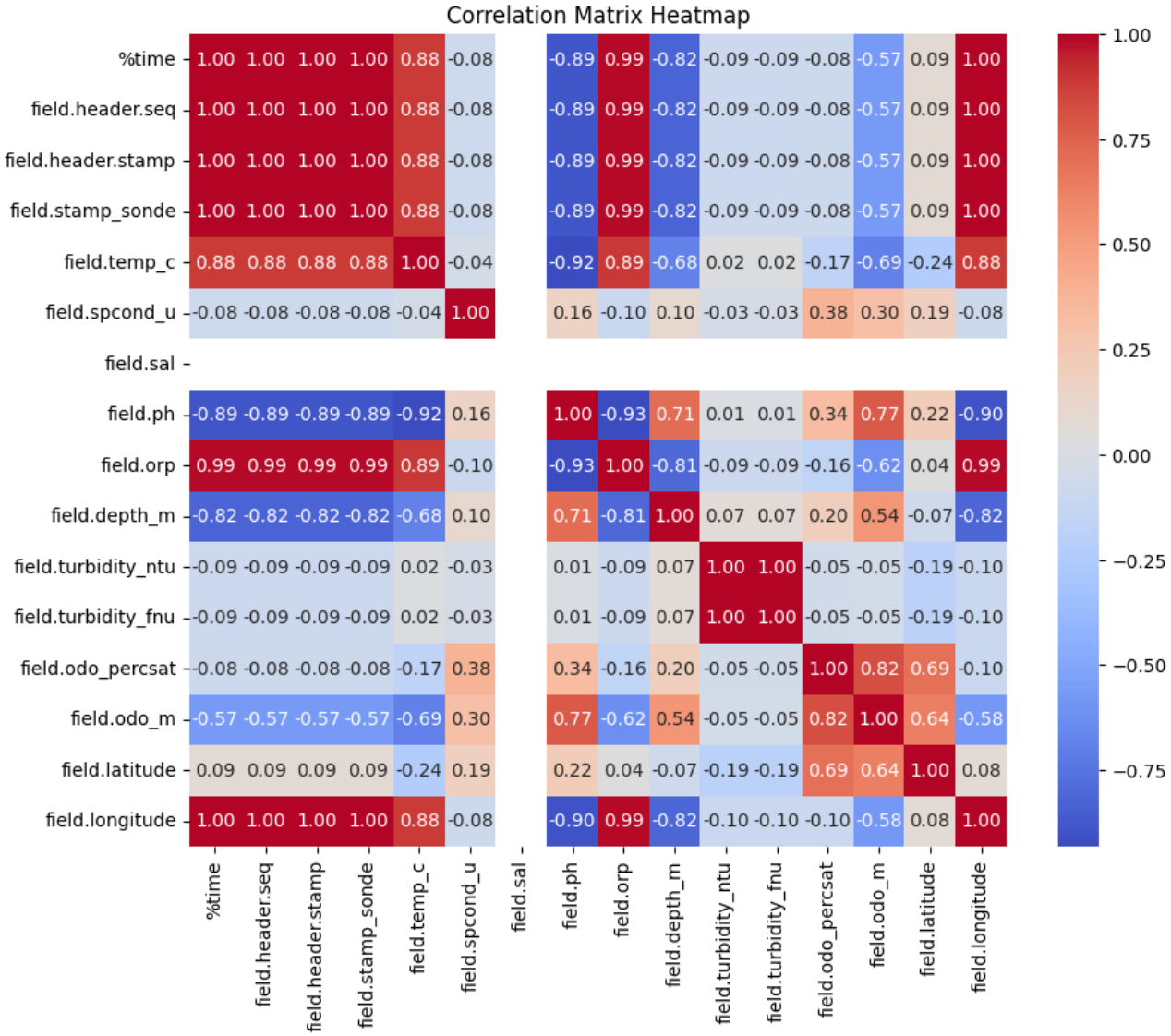
No visible linear relationship between the two variables.

Question 3_6



In this plot, I will say there is a linear relationship between the two Dissolved oxygen measurement, though the relationship might not be perfect. The relationship is a positive one, I.e when there is increase in dissolved oxygen percentage saturation, there will be increase in Dissolved oxygen concetration in milligrams per liter(mg/L)

Question 4.
To know what factors could influence the oxygen level and to check for collinearity among the independent variables., I find the correlation between each of the variables/factors in the data by plotting correlation matrix heat map. This shows the correlation values (R-valued)of each of the value against each other. The value ranges between -1 to 1. If the absolute value is around 0.8 to 1. It shows a strong relationship between two variables and if it is around 0.2 to 0, it shows a weak relationship between the two value.



The heat shows that turbidity nfu and ntu are linear combination of each other since the value is 1. This shows that, we have to drop one of them to avoid multicollinearity.

I used the field.odo_m which is the Dissolved Oxygen (ODO) concentration in milligrams per liter (mg/L) as the target i.e the dependent variable.

I also dropped %time, field.header.seq, field.header.stamp as they all are linear combination of the field.stamp.sonde. Each of their correlation value is one, which shows they have linear relationship. To avoid the collinearity, I have to drop the three and use field.stamp.sonde as it represent all of the three time measurement.

I also dropped field.sal as it has a constant value therefore It cannot exhbit any relationship with the target value.

From the correlation value, both the two turbidity gave -0.05 which is almost zero showing that they do not have correlation with the target value, so I dropped the two.

So the factors remained after all the analysis that could influence the oxygen level are :-
0 field_stamp_sonde
1 field_temp_c
2 field_spcond_u
3 field_ph
4 field_orp
5 field_depth_m
6 field_odo_percsat
7 field_latitude
8 field_longitude

target => field_odo_m

After developing a multiple linear regression. I got this coefficient, mean squared error and intercept for each of the value.
Mean Squared Error: 8.367117998600006e-06

	Feature	Coefficient
0	field_stamp_sonde	7.212098e-16
1	field_temp_c	-1.705226e-01
2	field_spcond_u	5.263107e-04
3	field_ph	-1.947453e-02
4	field_orp	-2.022716e-04
5	field_depth_m	-5.425856e-02
6	field_odo_percsat	8.322529e-02
7	field_latitude	2.388705e-03
8	field_longitude	-2.457427e-04

Intercept: -1100.2200639186606

During the training, I also employ forward selection and backward elimination to ensure I got the best model.
The mean square error value is almost 0 which shows that the model is a good model.

To also verified my model.
I also find the r_squared value and the adjusted r_squared value and they both gave
R-squared: 0.9952789293968191 and Adjusted R-squared: 0.9952694956411733.

This two values shows that my model is a good one.