**Data-driven Water Characteristics Analysis**

A Dataset which I do have is (WaterData.csv) from a wastewater treatment plant contains nine water characteristics (i.e., Total Solids, SS, BOD5, NH3, Org-N, P-TOT, SO4, TKN, PRCP\_NOOA) between 2001 and 2018.

Where is missing values are also presented in the dataset.

So what I did in the project for predicting the water characteristics.

The steps required to train a model in machine learning are :

Step 1: Load the data

There is a requirement to analyze the data and provide your best solution to predict each water characteristic. Where first we load the dataset.

Step 2: Prepare the data

After loading the dataset, the data may not be in the right format for a particular task hence the need for data preparation. Data preparation is made up of loading, cleaning, transforming, and rearranging the data. If there Missing data occurs often in data analysis. We will have those values which are Nan and null values. I customize them and fill them with appropriate values like the median and mean of the column or you can fill it by your good particular values. After filling in the missing values let I categorized the object values where most machine learning algorithms accept only numerical features. Hence, the need to convert the text object to numbers (dummy variables).

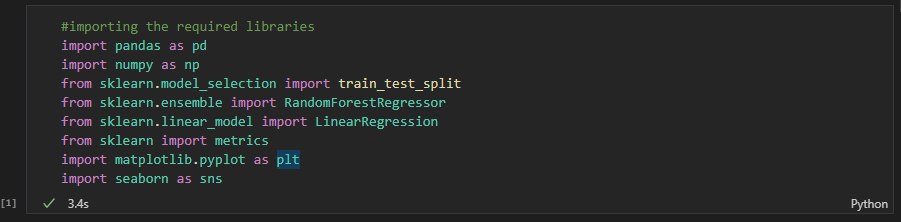
Step 3: Building the model

Lets I take Linear regression for prediction and classification, first I create the train and test dataset from the dataset, where I split 80% entire dataset will be used for training and 20% will be testing.

Step 4: Train your machine model

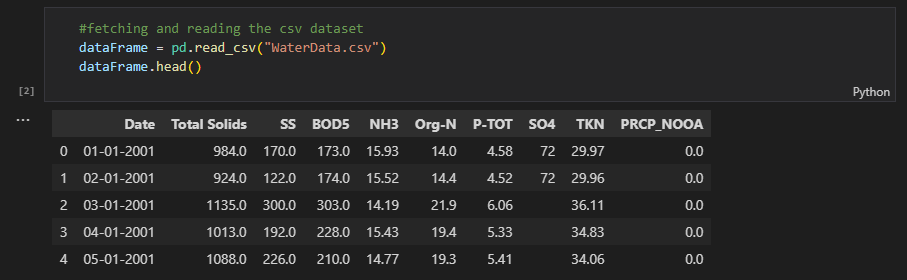
After the preparation of matrix data where I train our Linear Regression Model which tries to fit the best combination of weights and bias to minimize a loss function over the prediction.

First I will import the required libraries which are being used in water characteristics predictions.



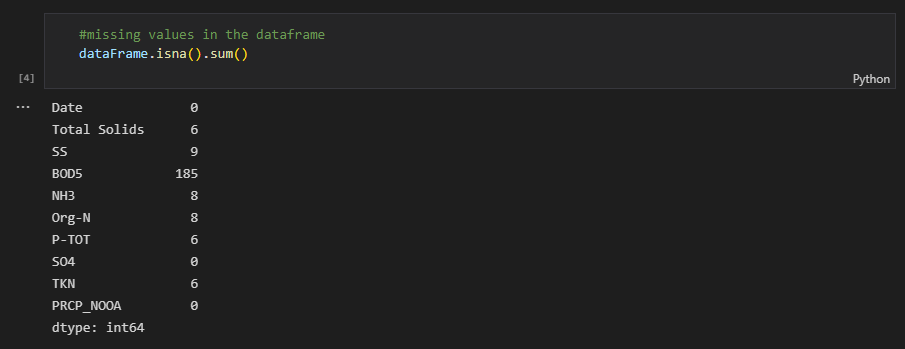
After that, I will fetch the water dataset and then I will extract all the elements and their values in the water dataset.

Let’s see what characteristics are in the water dataset.

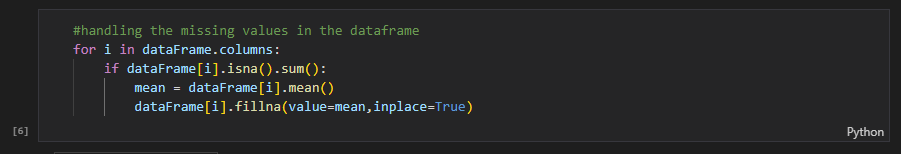


**Data Handling**

Let’s find is there are any missing values in our data frame or not, if it then I will handle the missing values in our dataset.

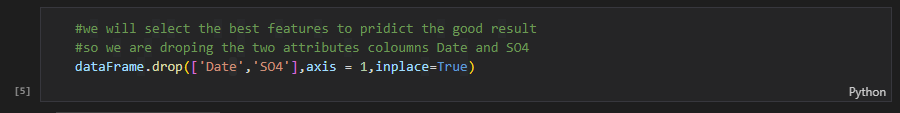


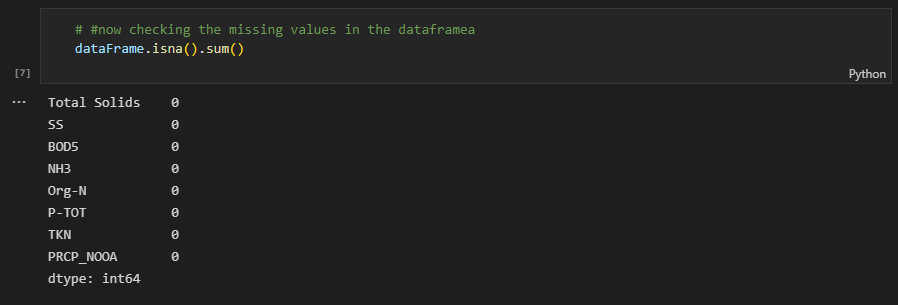
So after finding the missing values in our dataset lets fill the missing values.



What this will do is will checks the data frame column that is there any missing values if it then, we find the mean of that column and fill it with that mean value for all missing. Now check if is there any missing value left.

We drop the two columns from our dataset for better prediction of water characteristics.

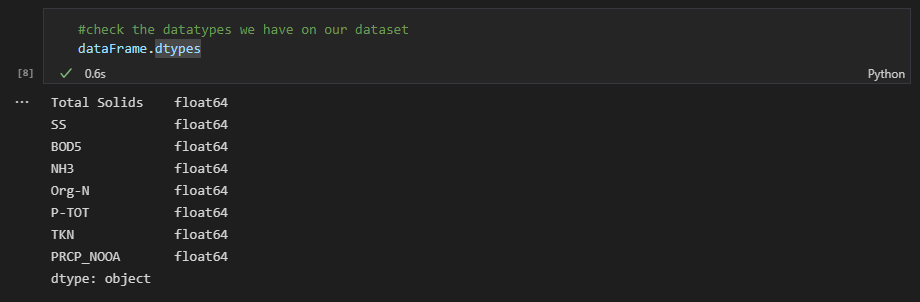




As you can see there are no null values is located in any column of the dataset.

After handling the missing, null, and Nan values in our water characteristics dataset we convert the data types of it for getting better results.

Let’s check what column feature has what datatype after converting to float elements.

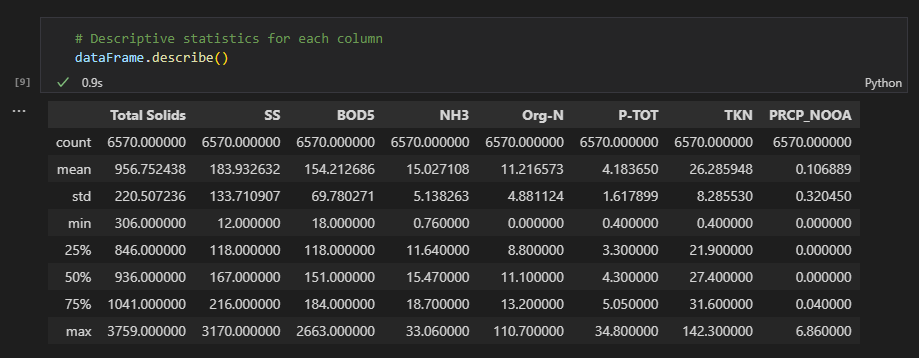


**Description of the data in the Data Frame.**

This method will compute and displays summary statistics for a data frame, where we’re working, the description contains information for each column.

count - The number of not-empty values.  
mean - The average (mean) value.  
std - The standard deviation.  
min - the minimum value.  
25% - The 25% percentile\*.  
50% - The 50% percentile\*.  
75% - The 75% percentile\*.  
max - the maximum value

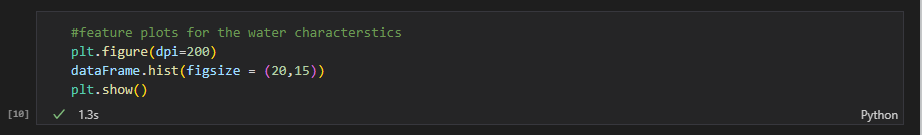
\*Percentile meaning: how many of the values are less than the given percentile.

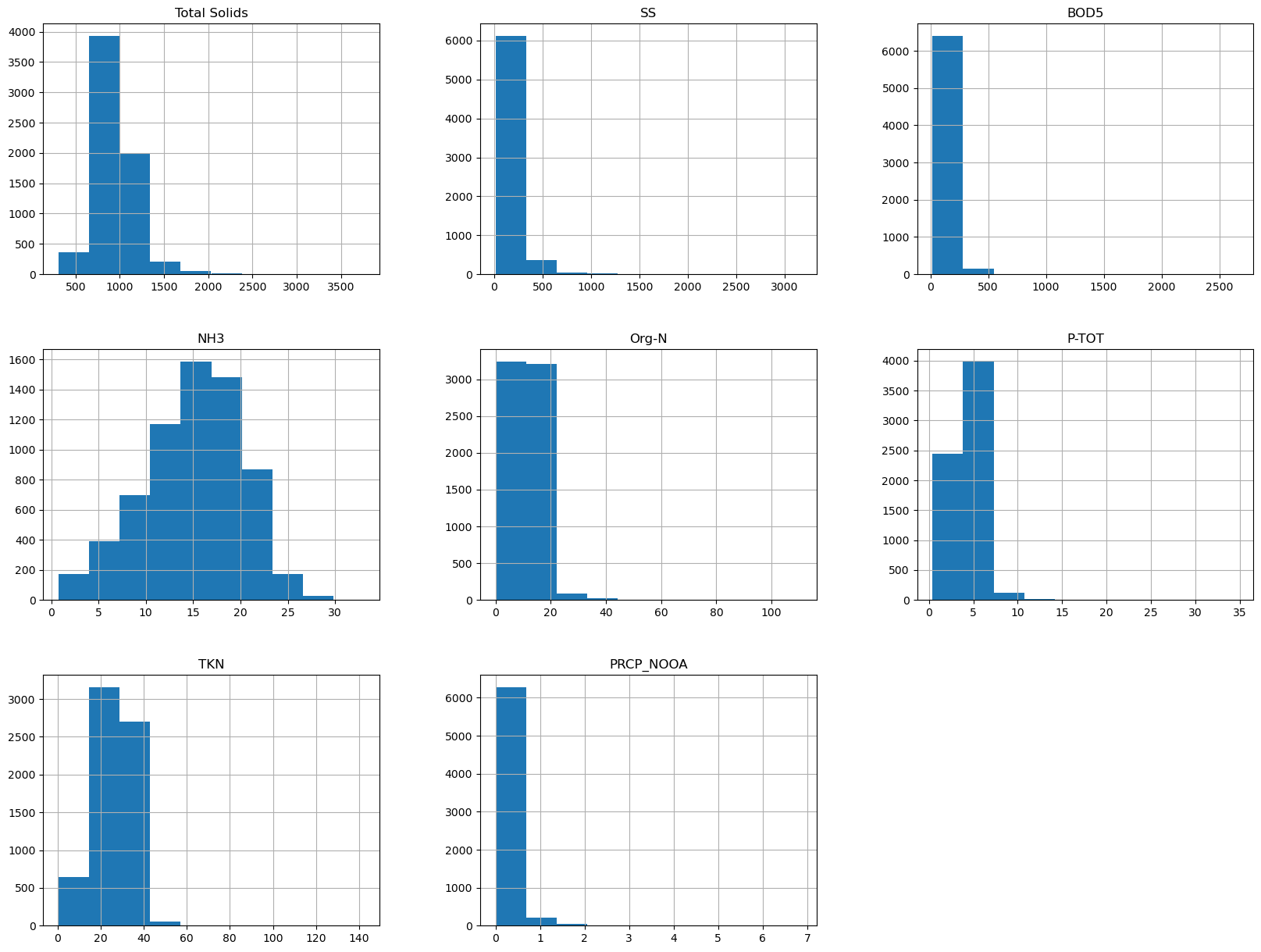


**Data Visualization**

Data visualization is a way to represent information graphically, highlighting patterns and trends in data and helping the reader to achieve quick insights. Also known as “interactive visual exploration,” it enables the exploration of data via the manipulation of chart images, with the color, brightness, size, shape and motion of visual objects representing aspects of the dataset being analyzed. It includes an array of visualization options that go beyond those of pie, bar and line charts, including heat and tree maps, geographic maps, scatter plots, and other special-purpose visuals. These tools enable users to analyze the data by interacting directly with a visual representation of it.

Lets visualize the feature characteristics in our dataFrame

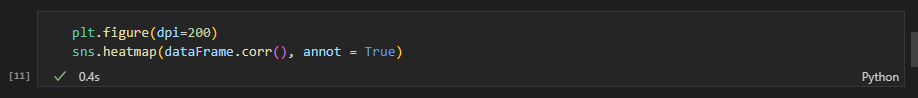




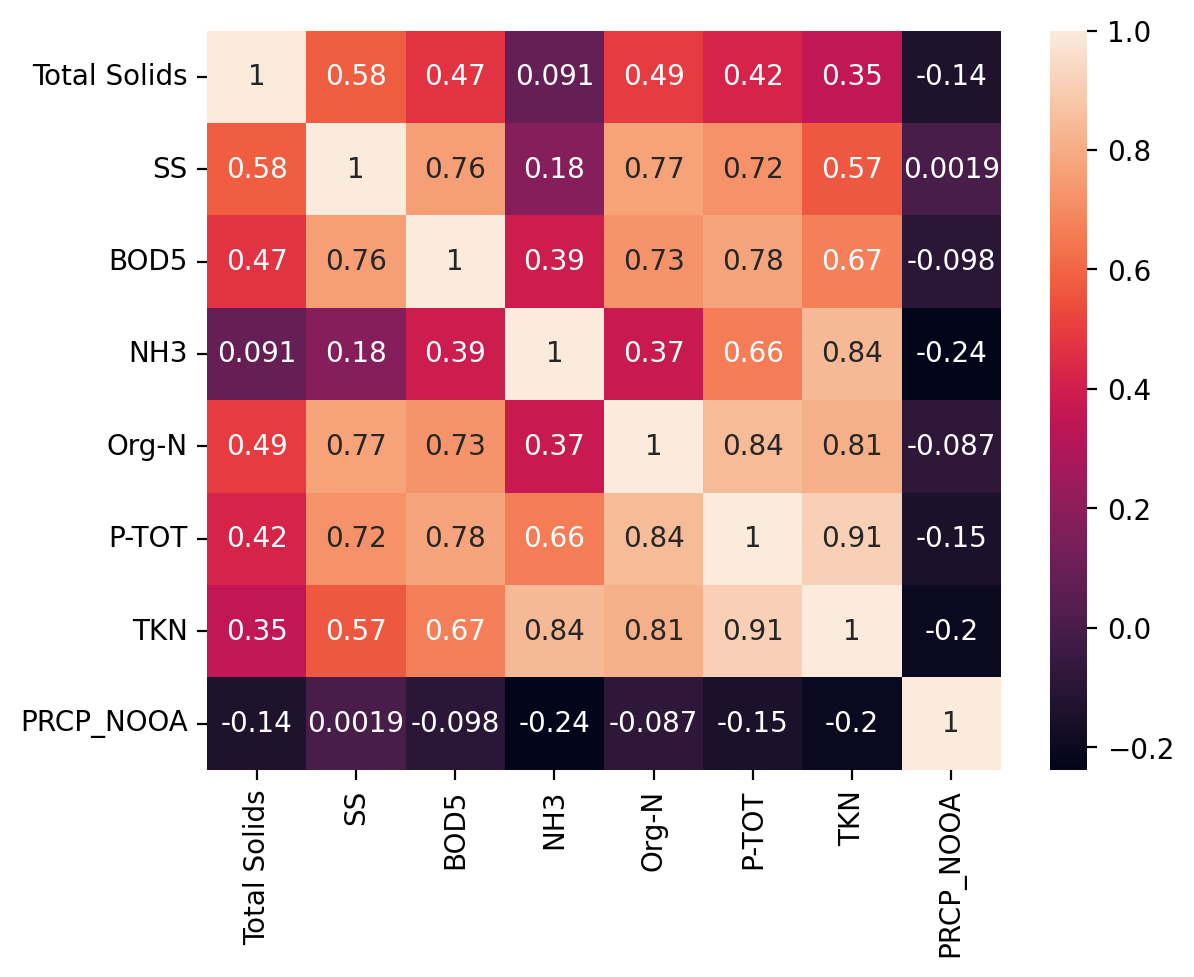
**Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

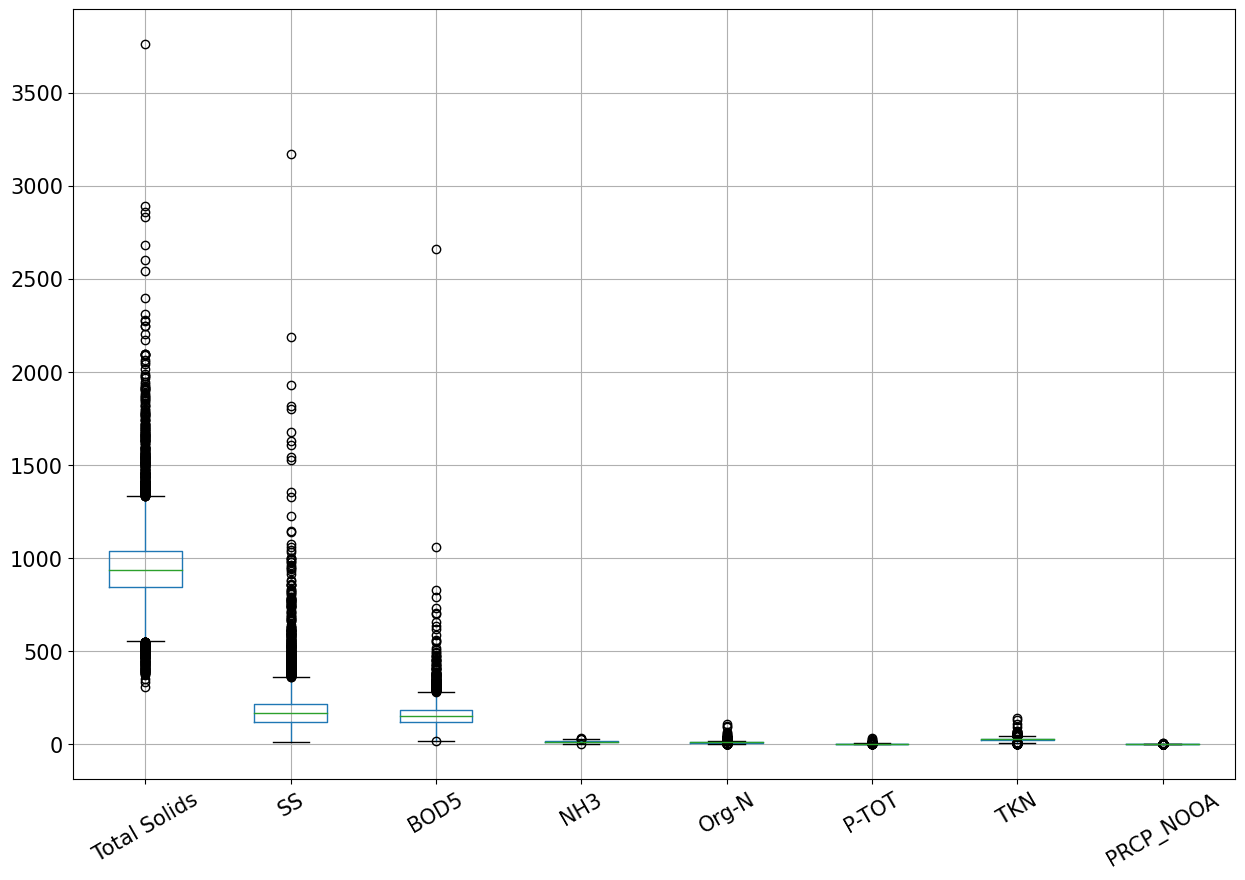
Lets have the heatmap for our dataFrame



Heatmap



Box Plot of the Water characteristics



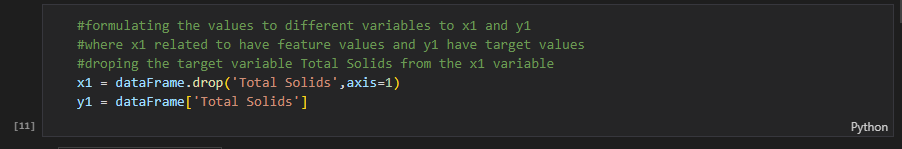
The boxplot shows that some of the variables exert dominance over the others hence the need for scaling.

**Building the machine learning model**

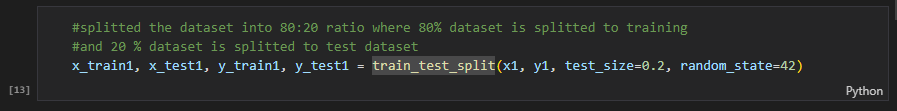
The characteristics we have now for predict the water characterstics.



As you can see x1 variable I dropped the Total Solids Feature column from the dataset because Total Solids will be the target variable which I am going to predict the total solids water characteristics. So I assigned the y1 for the target variable where only the Total Solids feature values be there.

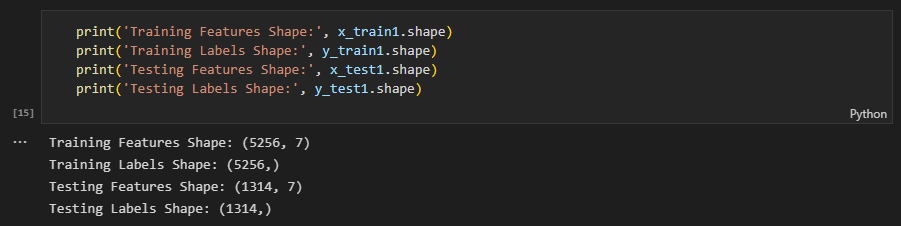


**Split the dataset**



Here I splitted the dataset in the ratio of 80:20 where the 80 percent data is for training and 20 percent data will there for testing and validation.

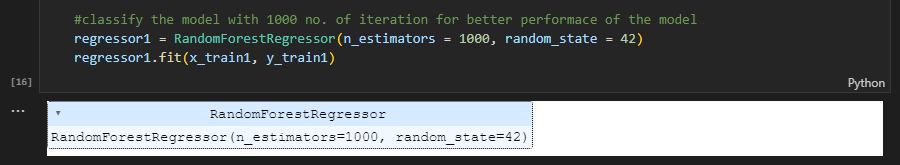
Here you can see the result after splitting the dataset in 80:20 ratio.



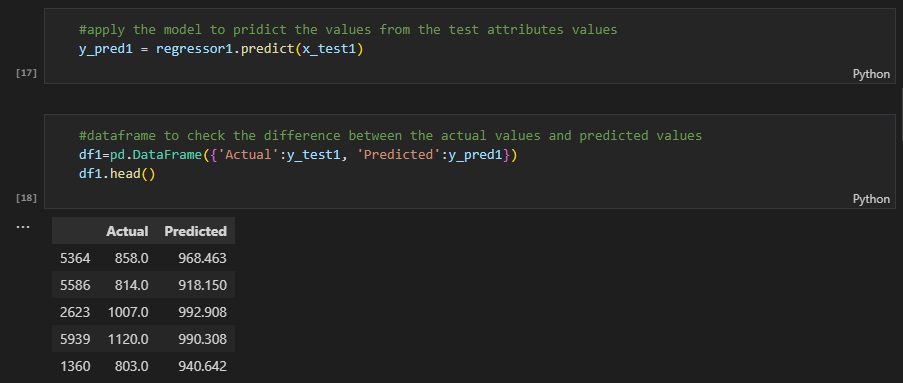
**Creating the Model**

I used the RandomForestRegressor Model which is a supervised learning based algorithm model that uses group learning method for regression. This learning method is a algorithm which combines or group the predictions from multiple learning algorithms to make a more accurate prediction then the single model.

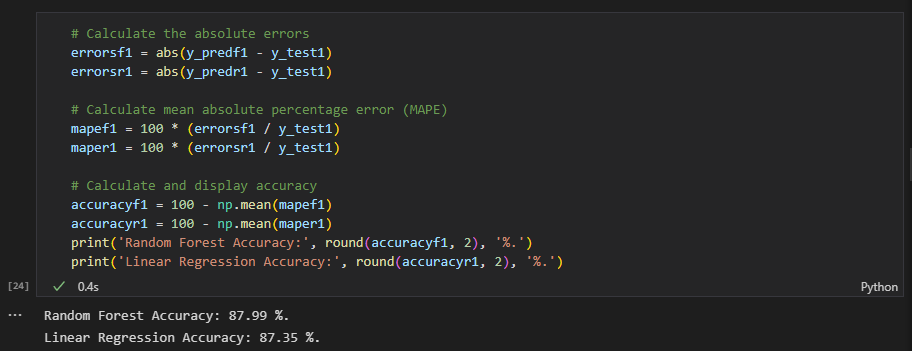
That’s why I chose the Random Forest Regression model for better prediction.



After creating and customizing the model then I had trained the model, let have the prediction values and their result.

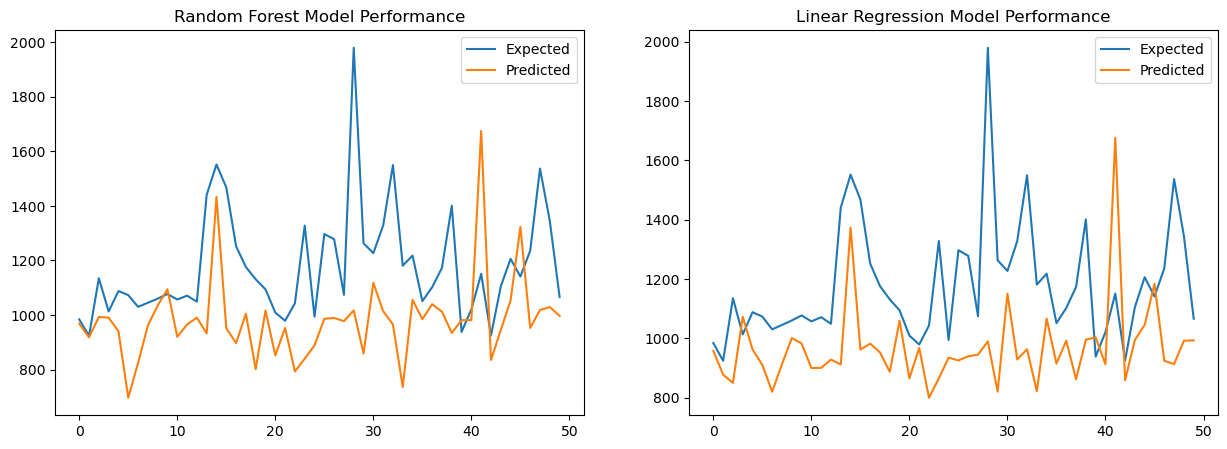


Accuracy of the both models.



Their accuracy I got is 87.99% and 87.35% for predicting the one characterstic Total Solids.

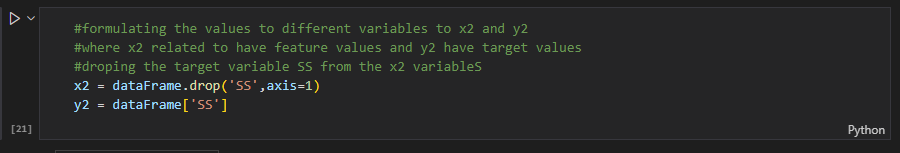
I applied these values to same for linear regression model for comparing the both results among between the Random Forest Regression and Linear Regression.

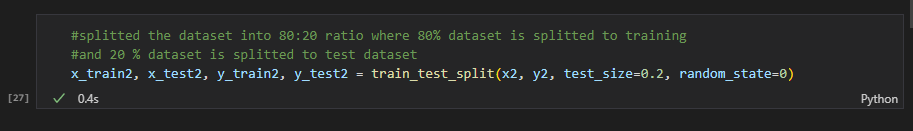




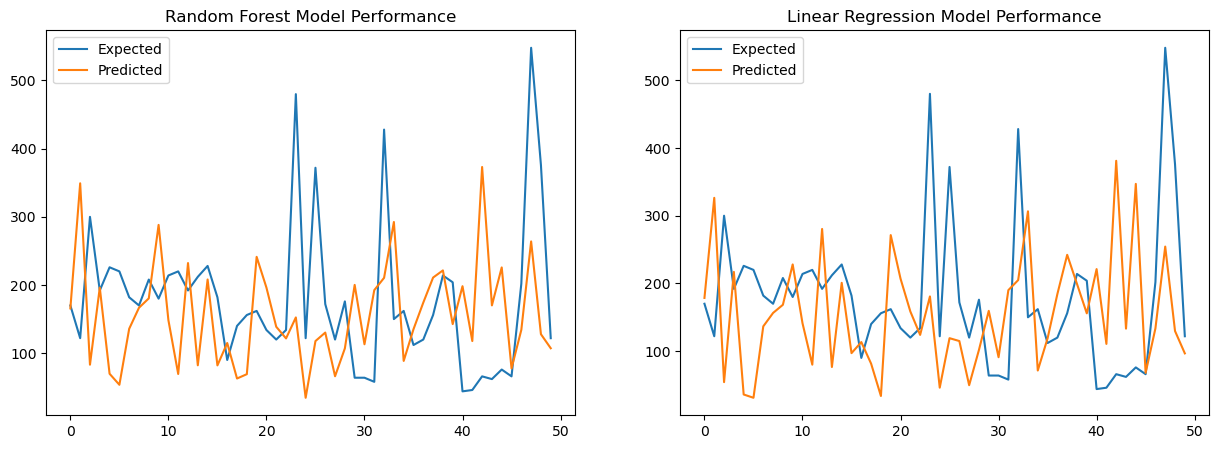
Where I did same for all the characteristics where now predicting the “SS” water characterstic same I did for this for the prediction for this.

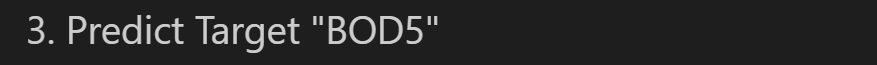
From starting assigning the x2 variable and dropped the “SS” characteristic from the dataset because this is target feature characteristic which is assigned in y2.





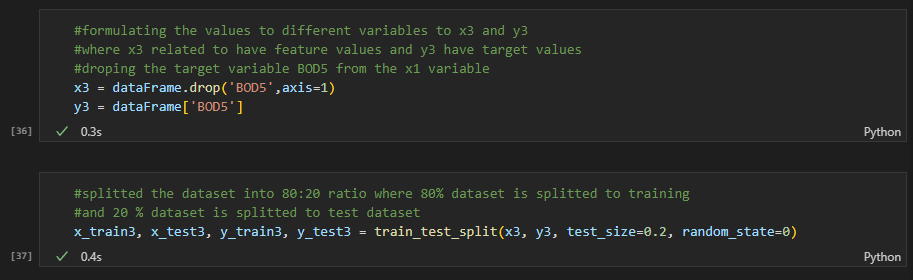
Performance between the models in Second SS water Characteristics.



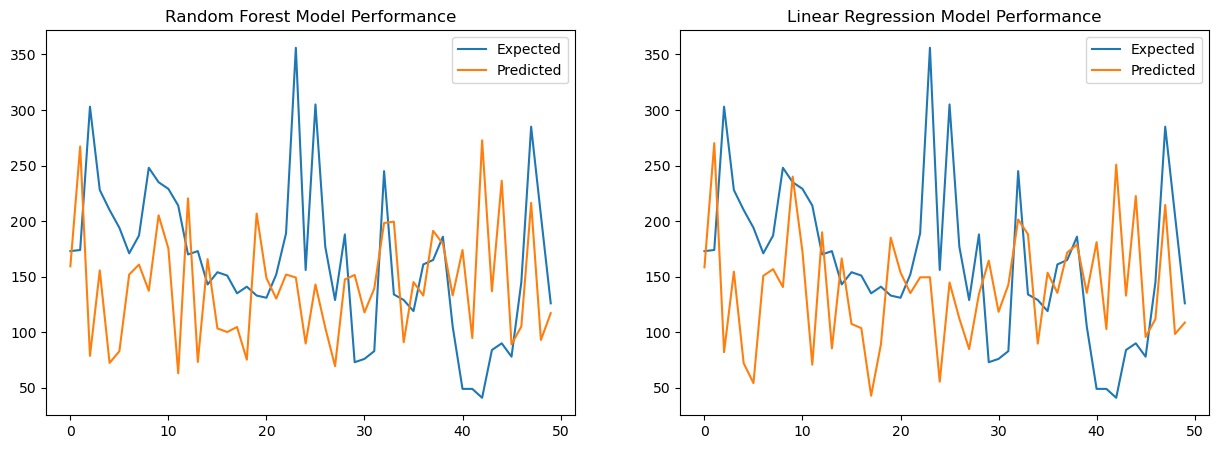


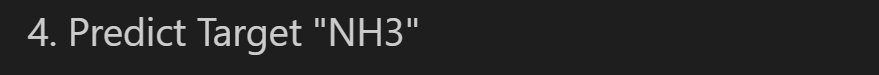
Where I did same for all the characteristics where now predicting the “BOD5” water characteristic same I did for this for the prediction for this.

From starting assigning the x3 variable and dropped the “BOD5” characteristic from the dataset because this is target feature characteristic which is assigned in y3.



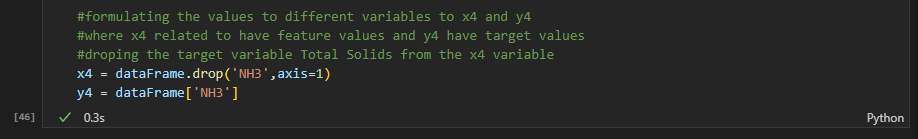
Performance between the models in BOD5 water Characteristics.

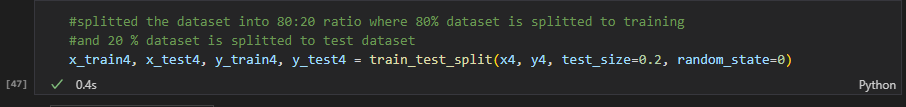




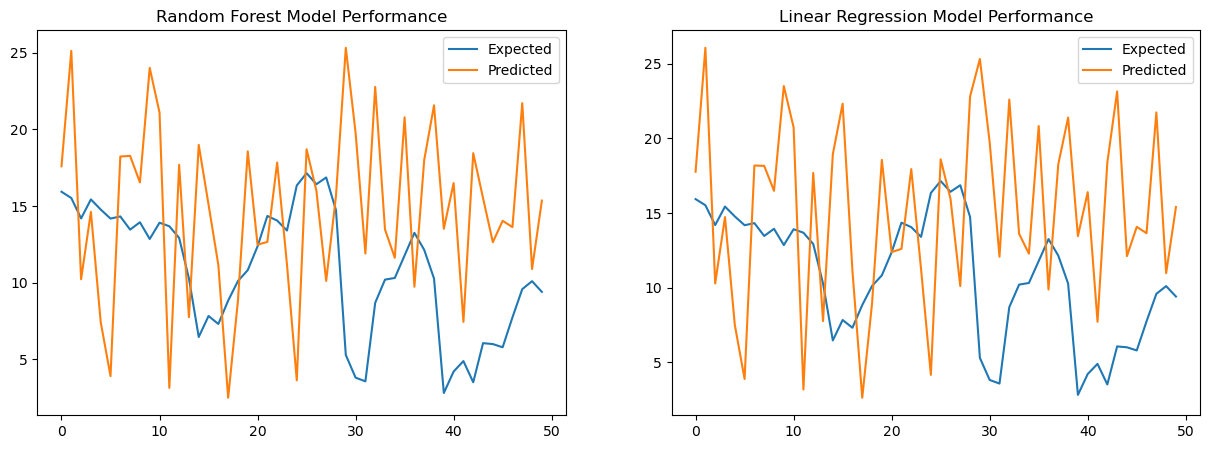
Where I did same for all the characteristics where now predicting the “NH3” water characteristic same I did for this for the prediction for this.

From starting assigning the x4 variable and dropped the “NH3” characteristic from the dataset because this is target feature characteristic which is assigned in y4.





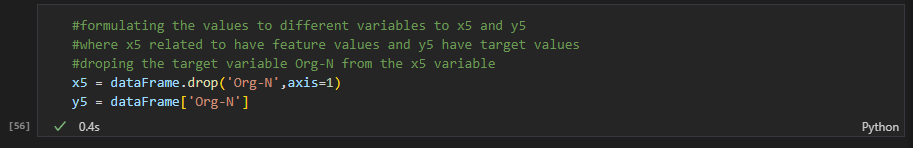
Performance between the model in NH3 water Characterstics.

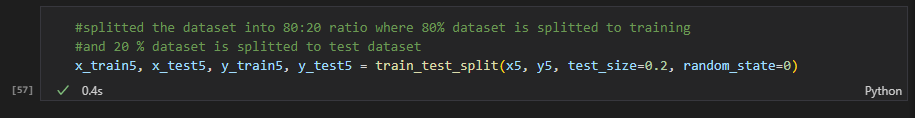




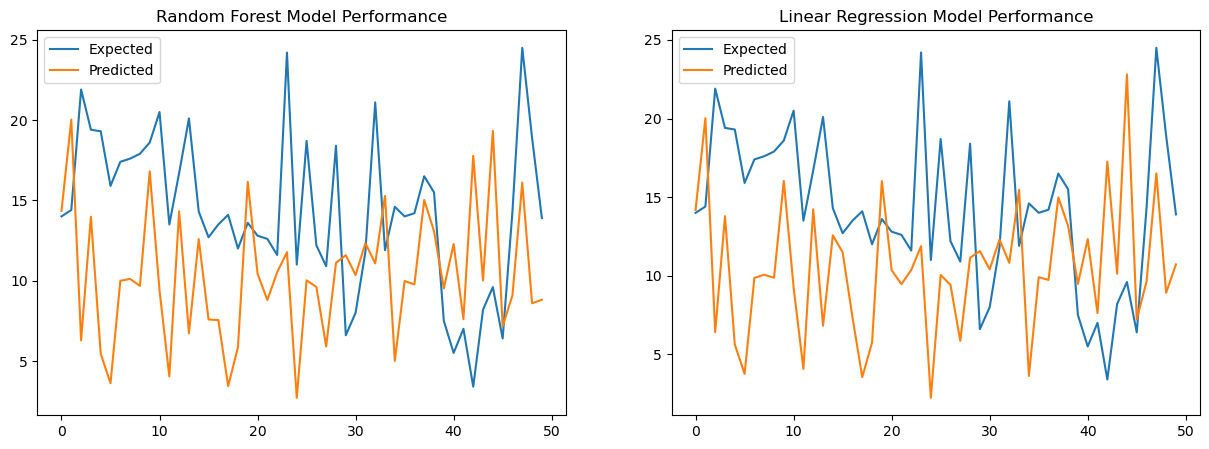
Where I did same for all the characteristics where now predicting the “Org-N” water characteristic same I did for this for the prediction for this.

From starting assigning the x5 variable and dropped the “Org-N” characteristic from the dataset because this is target feature characteristic which is assigned in y5.





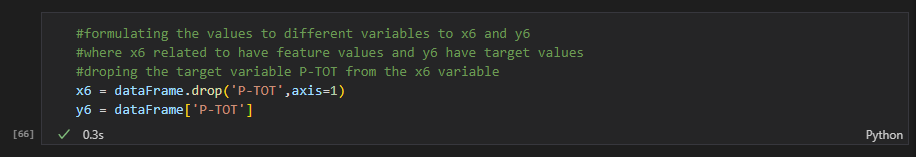
Performance between the models in Org-N water Characteristics.

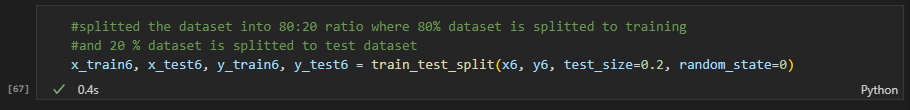




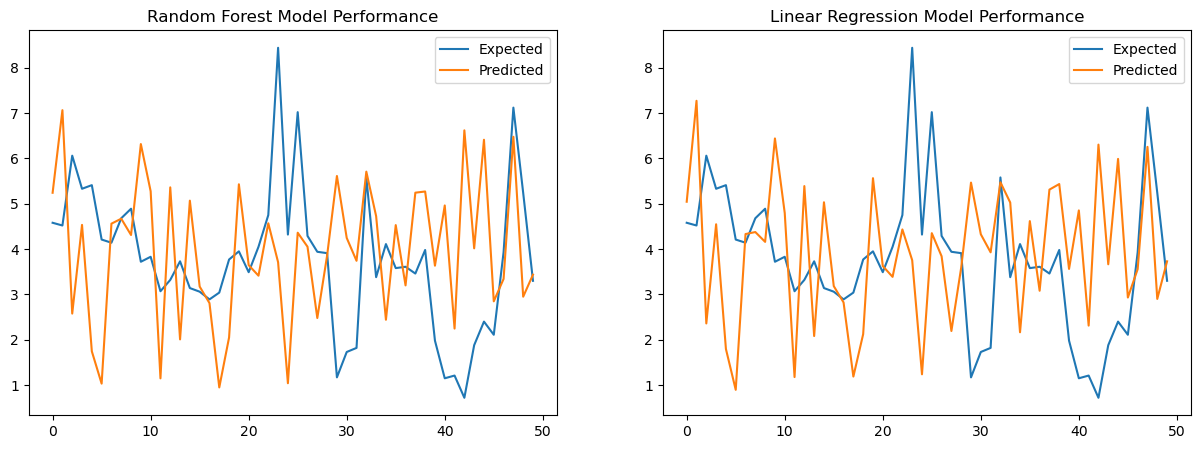
Where I did same for all the characteristics where now predicting the “P-TOT” water characteristic same I did for this for the prediction for this.

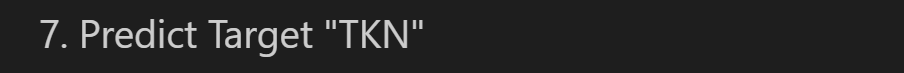
From starting assigning the x6 variable and dropped the “P-TOT” characteristic from the dataset because this is target feature characteristic which is assigned in y6.





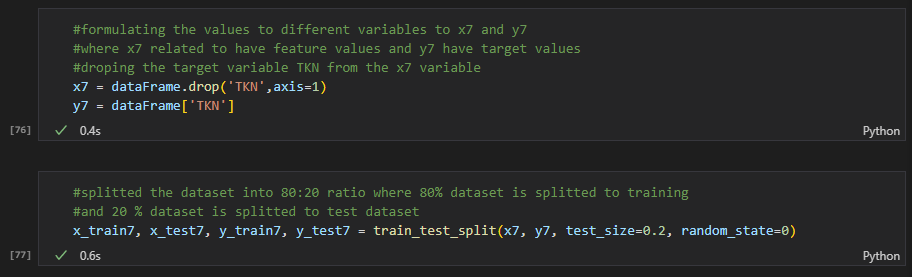
Performance between the models in P-TOT water Characterstics.



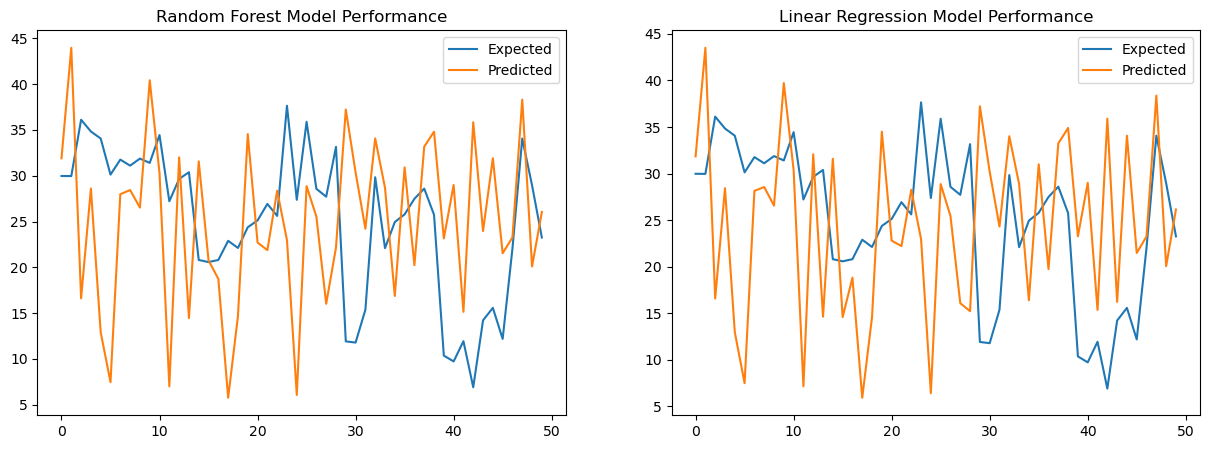


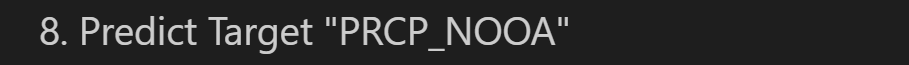
Where I did same for all the characteristics where now predicting the “TKN” water characteristic same I did for this for the prediction for this.

From starting assigning the x7 variable and dropped the “TKN” characteristic from the dataset because this is target feature characteristic which is assigned in y7.



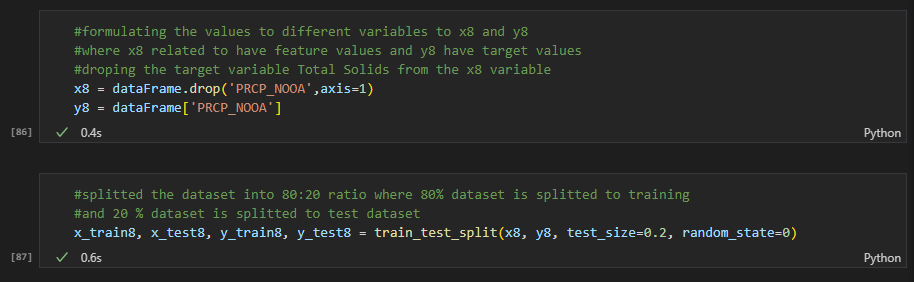
Performance between the models in TKN water Characteristics.



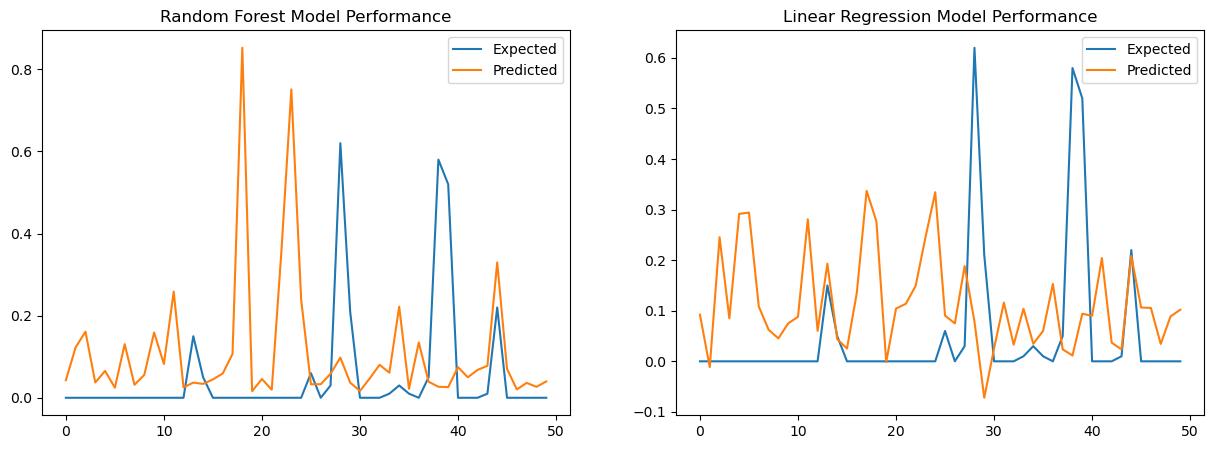


Where I did same for all the characteristics where now predicting the “PRCP\_NOOA” water characteristic same I did for this for the prediction for this.

From starting assigning the x8 variable and dropped the “PRCP\_NOOA” characteristic from the dataset because this is target feature characteristic which is assigned in y8.

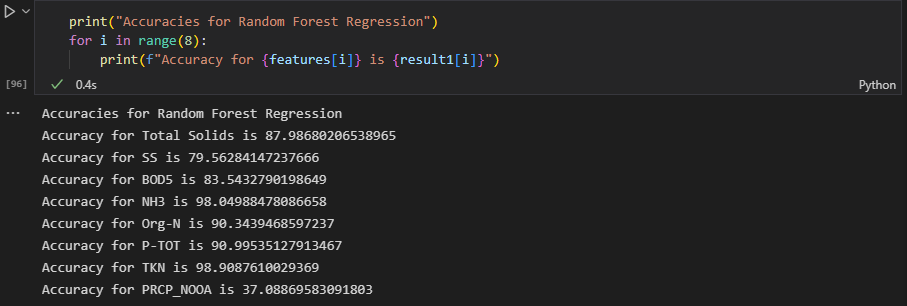


Performance between the models in TKN water Characteristics.

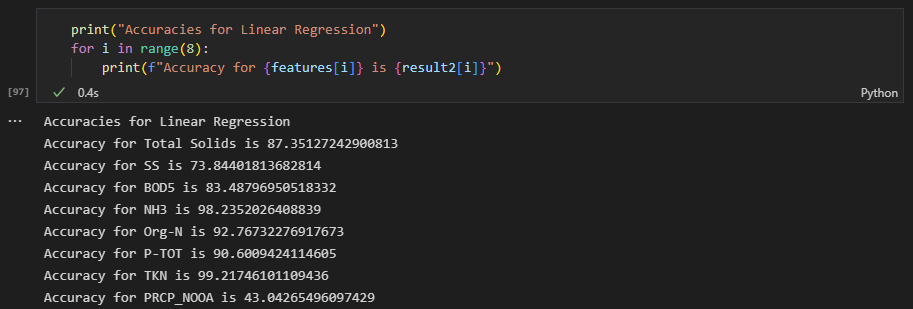


After training, prediction and testing of all the 8 models I got the accuracies for all the models are given below:-

**Random Forest Regression**



**Linear Regression**



Compare the Models with the visualization

