Linear Regression Model

Contents :

- ## Linear Regression Model

linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables.

more than one independent variable, the process is called multiple linear regression

multiple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

,
where

$$\hat{Y}$$

is the predicted or expected value of the dependent variable,

X_1 through X_p are p distinct independent or predictor variables,
b_0 is the value of Y when all of the independent variables (X_1 through X_p) are equal to zero, and b_1 through b_p are the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable.

In the multiple regression situation, b_1, f

Example, change in Y relative to a one unit change in X_1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed).

Identifying & Controlling for Confounding With Multiple Linear Regression

linear regression equation relating the risk factor (the independent variable) to the dependent variable as follows:

$$\hat{Y} = b_0 + b_1 X_1$$

where b_1 is the estimated regression coefficient that quantifies the association between the risk factor and the outcome

to **assess whether a third variable is a confounder,** we can denote the potential confounder X_2, and then estimate a multiple linear regression equation as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

In the multiple linear regression equation, b_1 is the estimated regression coefficient that quantifies the association between the risk factor X_1 and the outcome, adjusted for X_2 (b_2 is the estimated regression coefficient that quantifies the association between the potential confounder and the outcome)

Least Squares to determine the best fit line to data

residual error in ith observation. to minimize the total residual error.

To determine the value of m and c, that gives the minimum error for the given dataset. We will be doing this by using the Least Squares method.

Quadratic Loss Function helps to calculate the loss or error in our model. It can be defined as:

$$L(x) = \sum_{i=1}^{n} (y_i - p_i)^2$$

Least Squares method

we determined the loss function and need to minimize it. This is done by finding the partial derivative of **L**, equating it to 0 and then finding an expression for **m** and **c**. After equation:

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$c = \bar{y} - m\bar{x}$$

Here $\bar{x}$ is the mean of all the values in the input **X**
$\bar{y}$ is the mean of all the values in the desired output **Y**. This is the Least Squares method.

Perform multiple linear regression

We will use multiple linear regression to predict the stream demand **prediction** (i.e., the dependent variable) by using 4 independent/ input variables:

- PitchTankTemp
- PitchTankVol
- SterolTankTemp
- SterolTankVol

we validated several assumptions are met before apply linear regression models. identifying linear relationship exists between the dependent variable and the independent variables

linearity check

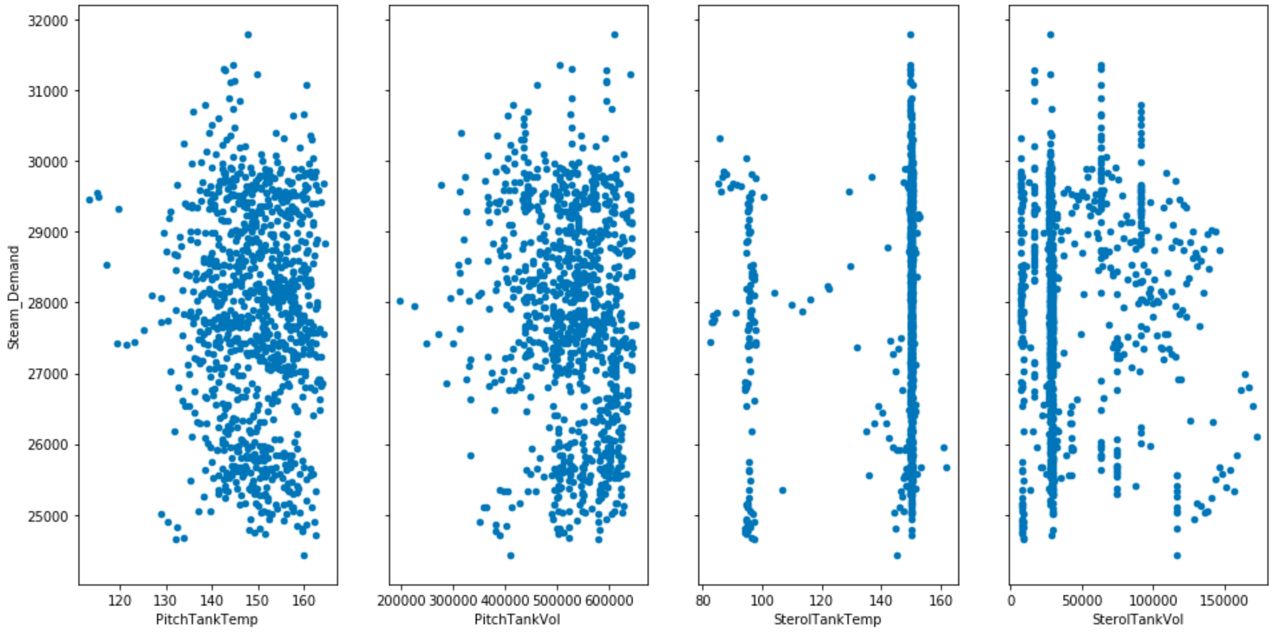
To check that a linear relationship exists between the dependent variable and the independent variable/s.

we checked the linear relationship exists between:

- The Stream_Demand (dependent variable) and the PitchTankTemperature (independent variable)
- The Stream_Demand (dependent variable) and the PitchTankVolume (independent variable)
- The Stream_Demand (dependent variable) and the SterolTankTemperature(independent variable)
- The Stream_Demand (dependent variable) and the SterolTankVolume (independent variable.

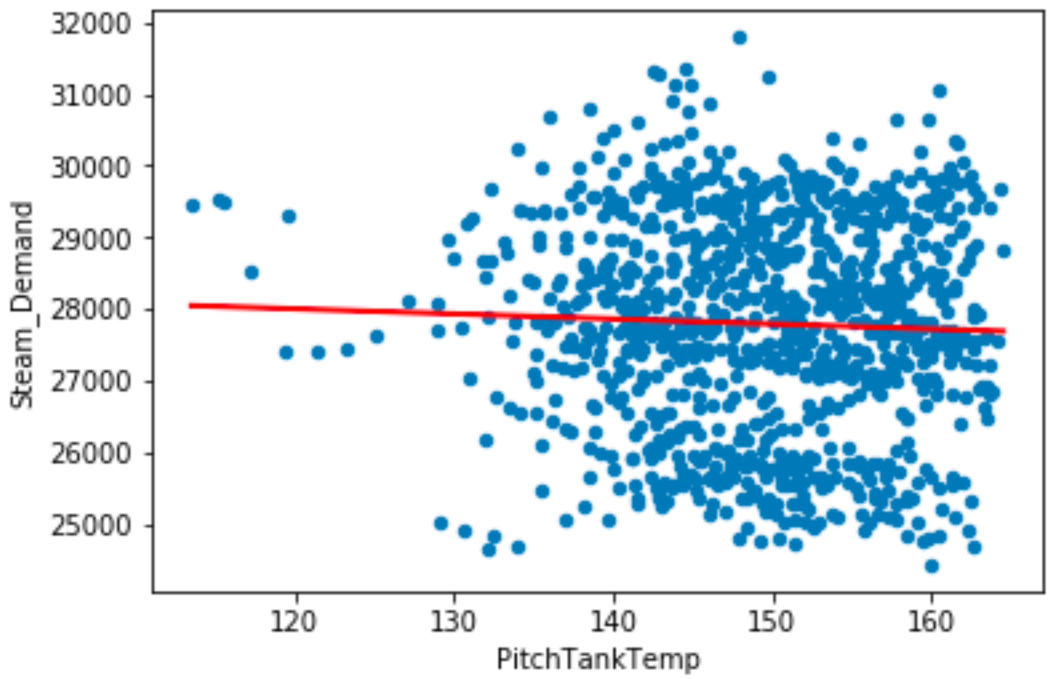
To perform a quick linearity check, we used scatter diagrams Fig-1

Fig-1

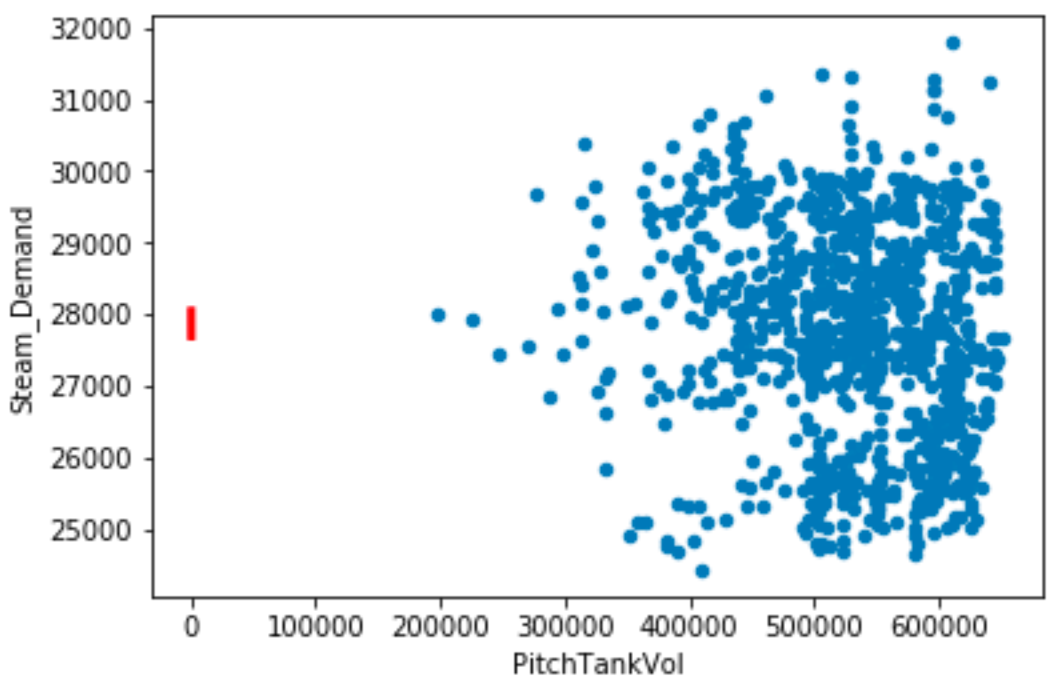


we can see, a linear relationship exists in both cases:

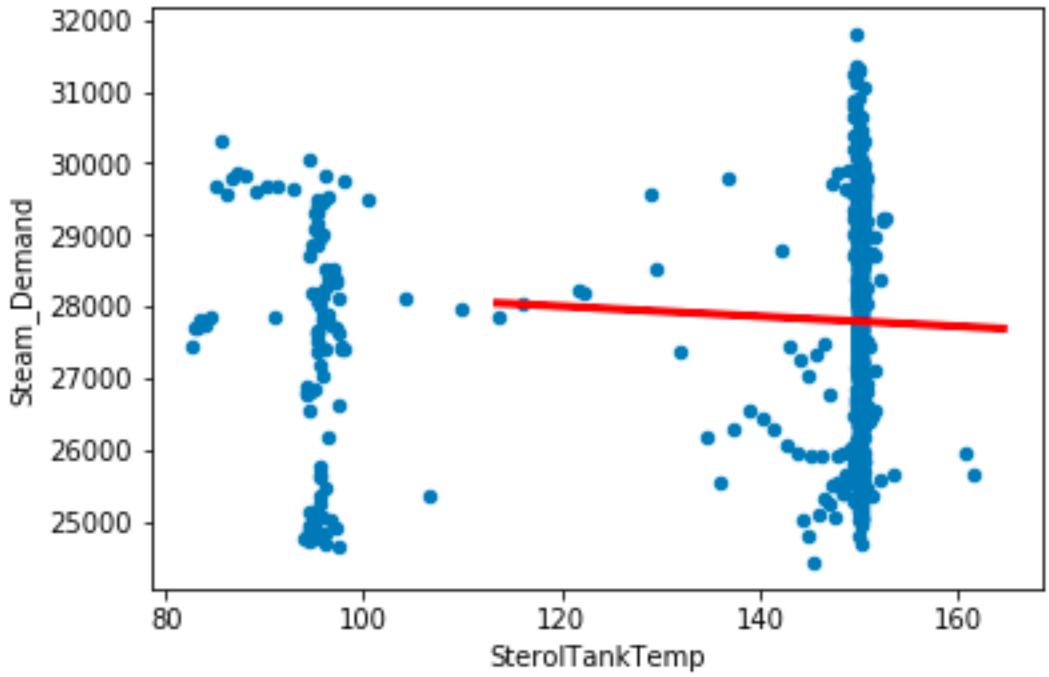
- In the first case, when PitchTankTemp goes up, the Stream_Demand goes up slightly.



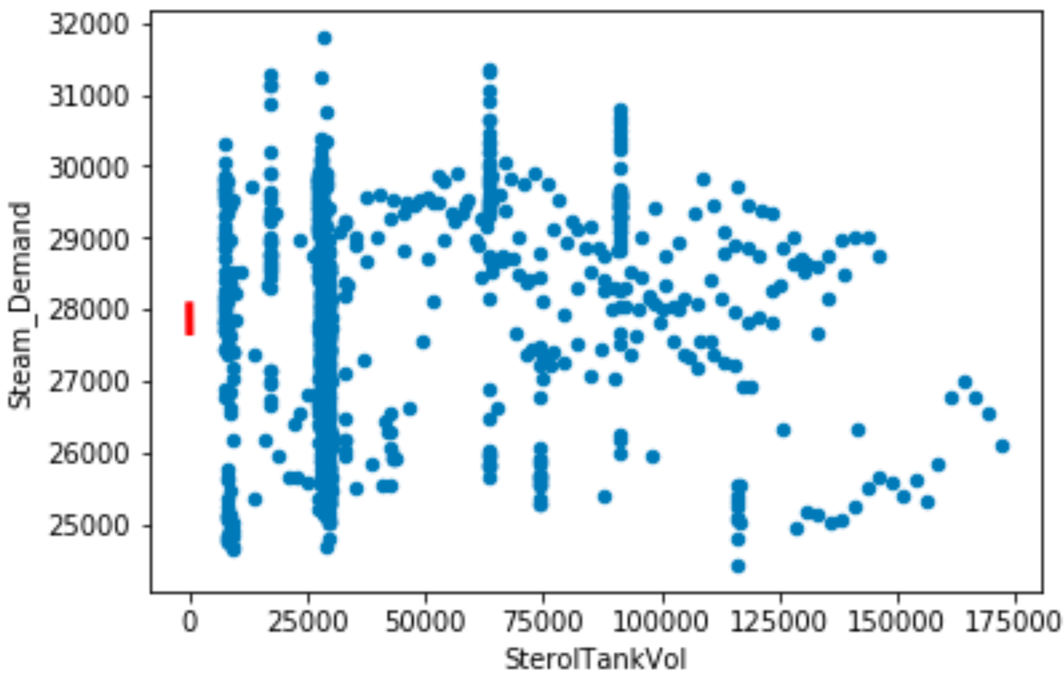
- In the second case, when PitchTankVolume goes up, there is no change in Stream_Demand.



In the 3rd case, when SterolTankTemp goes up, the Stream_Demand goes up slightly .



- In the 4th case, when SterolTankVolume goes down or up , there is no change in Stream_Demand.



- **UseCase Information**

1. shows the output information :

- PitchTankTemp
- PitchTankVol
- SterolTankTemp
- SterolTankVol

intercept : 30840.56895678324
coefficients : [-1.72950365e+01 -3.13835842e-03 8.32582460e+00 1.80900546e-05]

Stream_Demand = (Intercept) + (PitchTankTemp)*X1 + (PitchTankVol)*X2 + (SterolTankTemp)*X3 + (SterolTankVol)*X4

Stream_Demand = (30840.56895678324) + (PitchTankTemp coefficient)*X1 + (PitchTankVol coefficient)*X2 + (SterolTankTemp coefficient)*X3 + (SterolTankVol coefficient)*X4

Stream_Demand = (30840.56895678324) + (-1.72950365e+01)*X1 +(-3.13835842e-03)*X2+(8.32582460e+00)*X3+ (1.80900546e-05)*X4

2. Displays the stream demand prediction

Predict the Stream Demand : 27186.438414230535

X1=149.140014	X2=512659.244865	X3=94.067842	X4=8169.35773

to predict the **steam demand** after you collected the following data:

- PitchTankTemp = 149.140014 (X1=149.140014)
- PitchTankVol = 512659.244865 (X2=512659.244865)
- SterolTankTemp = 94.067842 (X3=94.067842)
- SterolTankVol = 8169.35773 (X4=8169.35773)

Plug that data into the regression equation, you'll get the exact same predicted results as displayed in the second part:

Steam_Demand = (30840.56895678324) + (-3.97750965e+01)* 149.140014 + (-3.05173578e+00)* 512659.244865 +(4.20839003e+03)* 94.067842 + (1.14234164e+02)* 8169.35773

Stream_Demand = XXXXXXXX

3. **displays a comprehensive table with statistical info**

OLS Regression Results

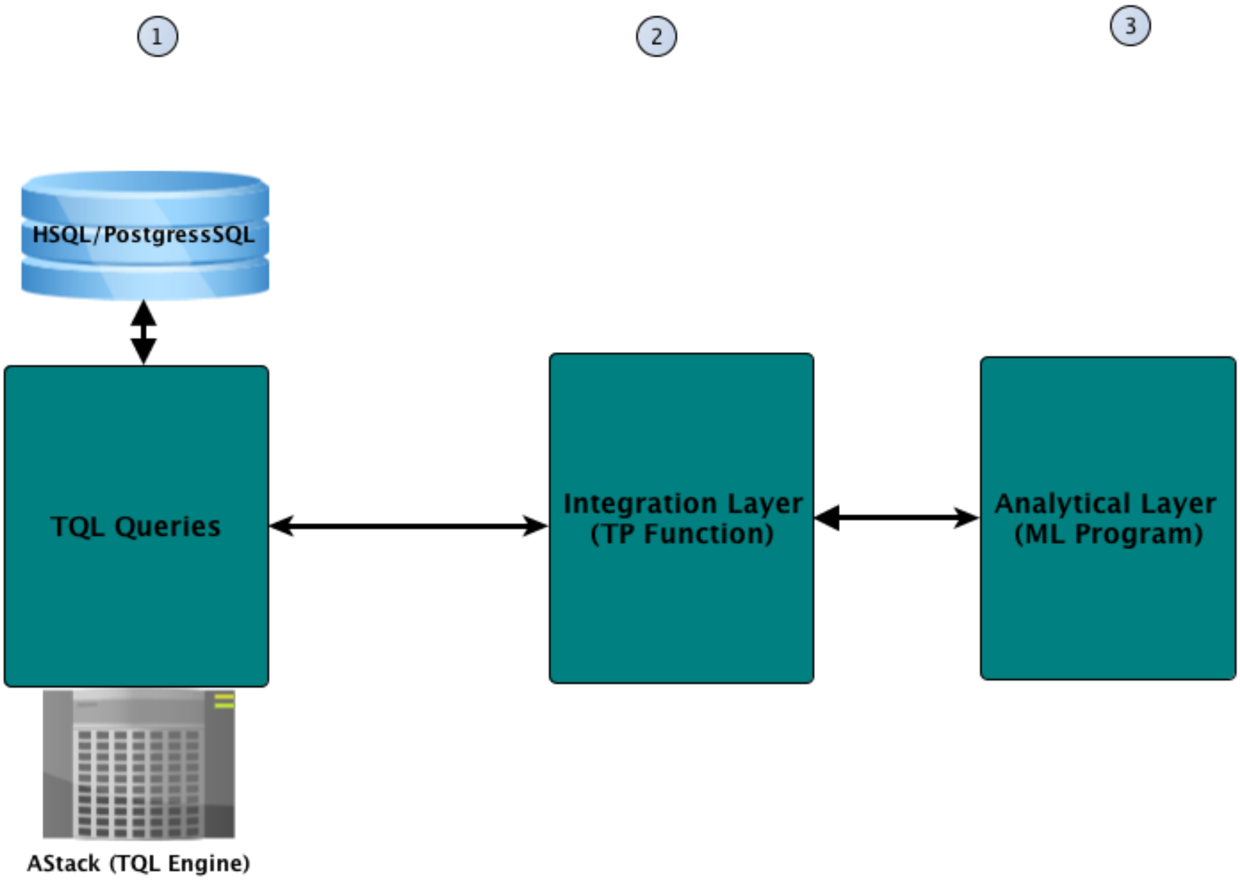
Dep. Variable:	Steam_Demand	**R-squared:**	0.031
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	7.875
Date:	Tue, 12 Mar 2019	**Prob (F-statistic):**	3.00E-06
Time:	03:35:06	**Log-Likelihood:**	-8705.4
No. Observations:	1000	**AIC:**	1.742E+04
Df Residuals:	995	**BIC:**	1.745E+04
Df Model:	4		

Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.084E+04	1000.524	30.824	0.000	2.89E+04	3.28E+04
PitchTankTemp	-17.2950	5.841	-2.961	0.003	-28.757	-5.833
PitchTankVol	-0.0031	0.001	-4.870	0.000	-0.004	-0.002
SterolTankTemp	8.3258	2.961	2.811	0.005	2.515	14.137
SterolTankVol	1.809E-05	0.002	0.012	0.990	-0.003	0.003

Omnibus:	58.468	**Durbin-Watson:**	0.148
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24.094
Skew:	-0.120	**Prob(JB):**	5.86E-06
Kurtosis:	2.279	**Cond. No.**	1.16E+07

TQL Integration Architecture :



Analytical Layer : Implemented (Regression analysis class GetMultiLinearRegressionCoeff) predictive modelling technique which investigates the relationship between a dependent and independent variables.

```
// Apply OLSMultipleLinearRegression on X[] and X[][]  data
                OLSMultipleLinearRegression ols = new OLSMultipleLinearRegression();
                NumberFormat numberFormat = NumberFormat.getNumberInstance();
                ols.newSampleData(YValues, XValues);
                parameters = ols.estimateRegressionParameters();
```

Integration Layer:

```
public class CustomTPFactory implements SffTpFunctionsSvc {

		private static final String GetMultiLinearRegressionCoeff = "GetMultiLinearRegres-
sionCoeff";

	@Override
	public ListMap getInfo() {
			info.put(GetMultiLinearRegressionCoeff, new GetMultiLinearRegressionCo-
eff());

		return info;
	}
}
```

TQL Layer:

Sample TQL Query data :


```
<Data>
	<Demand>24750.163986044714</Demand>
	<Temp>65.04</Temp>
	<Humidity>0.98</Humidity>
	<Rain>3.55</Rain>
	<Windspeed>65.04</Windspeed>
	<PitchTankTemp>0.98</PitchTankTemp>
	<PitchTankVol>3.55</PitchTankVol>
	<SterolTankTemp>0.98</SterolTankTemp>
	<SterolTankVol>3.55</SterolTankVol>
</Data>
<Data>
	<Demand>24807.677974906637</Demand>
	<Temp>64.5</Temp>
	<Humidity>0.99</Humidity>
	<Rain>4.3</Rain>
	<Windspeed>64.5</Windspeed>
	<PitchTankTemp>0.99</PitchTankTemp>
	<PitchTankVol>4.3</PitchTankVol>
	<SterolTankTemp>0.99</SterolTankTemp>
	<SterolTankVol>4.3</SterolTankVol>
</Data>
<Data>
	<Demand>24737.304867405295</Demand>
	<Temp>63.99</Temp>
	<Humidity>0.95</Humidity>
	<Rain>3.12</Rain>
	<Windspeed>63.99</Windspeed>
	<PitchTankTemp>0.95</PitchTankTemp>
	<PitchTankVol>3.12</PitchTankVol>
	<SterolTankTemp>0.95</SterolTankTemp>
	<SterolTankVol>3.12</SterolTankVol>
</Data>
<Data>
	<Demand>24737.304867405295</Demand>
	<Temp>63.45</Temp>
```

<Humidity>0.96</Humidity>
<Rain>4.44</Rain>
<Windspeed>63.45</Windspeed>
<PitchTankTemp>0.96</PitchTankTemp>
<PitchTankVol>4.44</PitchTankVol>
<SterolTankTemp>0.96</SterolTankTemp>
<SterolTankVol>4.44</SterolTankVol>
</Data>
<Data>
<Demand>24937.752426349307</Demand>
<Temp>64.24</Temp>
<Humidity>0.95</Humidity>
<Rain>4.89</Rain>
<Windspeed>64.24</Windspeed>
<PitchTankTemp>0.95</PitchTankTemp>
<PitchTankVol>4.89</PitchTankVol>
<SterolTankTemp>0.95</SterolTankTemp>
<SterolTankVol>4.89</SterolTankVol>
</Data>
<Data>
<Demand>25153.019191240655</Demand>
<Temp>65.69</Temp>
<Humidity>0.93</Humidity>
<Rain>4.9</Rain>
<Windspeed>65.69</Windspeed>
<PitchTankTemp>0.93</PitchTankTemp>
<PitchTankVol>4.9</PitchTankVol>
<SterolTankTemp>0.93</SterolTankTemp>
<SterolTankVol>4.9</SterolTankVol>
</Data>
<Data>
<Demand>24911.444912672854</Demand>
<Temp>65.67</Temp>
<Humidity>0.94</Humidity>
<Rain>4.45</Rain>
<Windspeed>65.67</Windspeed>
<PitchTankTemp>0.94</PitchTankTemp>
<PitchTankVol>4.45</PitchTankVol>
<SterolTankTemp>0.94</SterolTankTemp>
<SterolTankVol>4.45</SterolTankVol>
</Data>
<Data>
<Demand>26539.62918439706</Demand>
<Temp>66.0</Temp>
<Humidity>0.94</Humidity>
<Rain>3.76</Rain>
<Windspeed>66.0</Windspeed>
<PitchTankTemp>0.94</PitchTankTemp>
<PitchTankVol>3.76</PitchTankVol>
<SterolTankTemp>0.94</SterolTankTemp>
<SterolTankVol>3.76</SterolTankVol>
</Data>
<Data>
<Demand>26807.064624054197</Demand>
<Temp>65.94</Temp>
<Humidity>0.95</Humidity>
<Rain>4.46</Rain>
<Windspeed>65.94</Windspeed>
<PitchTankTemp>0.95</PitchTankTemp>
<PitchTankVol>4.46</PitchTankVol>

<SterolTankTemp>1</SterolTankTemp>
<SterolTankVol>4.46</SterolTankVol>
</Data>
<Data>
<Demand>26807.064624054197</Demand>
<Temp>65.94</Temp>
<Humidity>90.4</Humidity>
<Rain>4.46</Rain>
<Windspeed>65.94</Windspeed>
<PitchTankTemp>0.95</PitchTankTemp>
<PitchTankVol>4.46</PitchTankVol>
<SterolTankTemp>0</SterolTankTemp>
<SterolTankVol>0</SterolTankVol>
</Data>