Water quality of river Thames

S/16/806

Abstract

Water is perhaps the most precious natural resource after air. Though the surface of the earth is mostly consists of water, only a small part of it is usable, which makes this resource limited. This precious and limited resource, therefore, must be used with care. As water is required for different purposes, the suitability of it must be checked before use. Also, sources of water must be monitored regularly to determine whether they are in sound health or not. Poor condition of water bodies are not only the indictor of environmental degradation, it is also a threat to the ecosystem. In industries, improper quality of water may cause hazards and severe economic loss. Thus, the quality of water is very important in both environmental and economic aspects. Thus, water quality analysis is essential for using it in any purpose.

The River Thames known alternatively in parts as the River Isis, is a river that flows through southern England including London. At 215 miles (346 km), it is the longest river entirely in England and the second-longest in the United Kingdom, after the River Severn.It flows through Oxford, Reading, Henley-on-Thames and Windsor. The lower reaches of the river are called the Tideway, derived from its long tidal reach up to Teddington Lock. It rises at Thames Head in Gloucestershire, and flows into the North Sea via the Thames Estuary. The Thames drains the whole of Greater London.

The River Thames has some of the highest recorded levels of microplastics for any river in the world.Scientists have estimated that 94,000 microplastics per second flow down the river in places.The quantity exceeds that measured in other European rivers, such as the Danube and Rhine.

The quality of the water is a one of the modern global issue. This study mainly focuses on the quality of the water in the River Thames. We consider Water Temperature, PH value, Alkalinity, Suspended Solids, Phosphorus, Ammonium, Dissolved Silicon, Chlorophyl, Dissolved Fluoride, Dissolved chloride, Dissolved Nitrate, Dissolved Sulphate, Dissolved Sodium, Dissolved Potassium, Dissolved Calcium, Dissolved Magnesium and Dissolved Boron as a chemical parameters to analyze the quality of the water. We divided River Thames for a 22 Sampling point to collect data and we have collected 4300 samples for this analyze.

Introduction

water quality standards are put in place to ensure the suitability of efficient use of water for a designated purpose. Water quality analysis is to measure the required parameters of water, following standard methods, to check whether they are in accordance with the standard. Water quality analysis is required mainly for monitoring purpose. To check whether the water quality is in compliance with the standards, and hence, suitable or not for the designated use, to monitor the efficiency of a system, working for water quality maintenance, to check whether upgradation / change of an existing system is required and to decide what changes should take place and to monitor whether water quality is in compliance with rules and regulations.

Univariate analysis

Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved.

Multivariate analysis

Multivariate analysis consists of a collection of methods that can be used when several variables are measured on each subject or experimental unit. In some cases, it may be productive to isolate each variable in a system and study it separately. In general, however, the variables are interrelated and therefore, when analysed variables individually, they yield little information about the system. Using multivariate analysis, the variables can be examined simultaneously to access the key features of the process that produced them. The multivariate approach enables us to

(1) explore the joint performance of the variables.

(2) determine the effect of each variable in the presence of the others.

Univariate statistics summarize only one variable at a time. Multivariate statistics compare more than two variables. Univariate analysis can yield misleading results in cases in which multivariate analysis is more appropriate. This study is containing more than two variables and large number of samples then most suitable analysis is multivariate analysis.

Theory

Cluster analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Multivariate analysis of variance (MANOVA)

Multivariate analysis of variance (MANOVA) is an extension of the univariate analysis of variance (ANOVA). In an ANOVA, we examine for statistical differences on one continuous dependent variable by an independent grouping variable. The MANOVA extends this analysis by taking into account multiple continuous dependent variables, and bundles them together into a weighted linear combination or composite variable. The MANOVA will compare whether or not the newly created combination differs by the different groups, or levels, of the independent variable. In this way, the MANOVA essentially tests whether or not the independent grouping variable simultaneously explains a statistically significant amount of variance in the dependent variable.

Det 9.348845e+22

Remove 9-fluoride,14- pottasium,calcium-15

Analysis

In this data set have empty rows , factor variables and null values. Firstly we remove the empty rows. And NULL values , factor variable with replace numerical value .

Handling Missing Values

-Delete complete empty rows

-Replace factor variables using FIND &SELECT tab and find all "<" notation. We found <0.004, <0.01, <0.2, <5 and replace all factor variablewith 0.004, 0.01, 0.2, 5 respectively .

-Obtain Average of all variable column
Use AVERAGE function

-Replace all NULL value with relevant their column mean value .

Import data set into rstudio

-Using read.csv command we can import the data set.

using "sum(is.na(dataset)==TRUE)" command we can find NULL value.

Very important replace NULL value because NULL value effect the final result.

Proof Assumption of MANOVA

Get the Covariance values and determinant of covariance matrix

```
> det(cov(data2[2:18]))
[1] 9.348845e+22
```

Determinant is positive. Therefore first assumption is true.

MANOVA

Prove the following assumptions before MANOVA test

1. The independent variable are categorical and dependent variables are continuous.

2. The dependent variables are normally distributed in factor group.

3. There is no multicollinearity.

4. There should be no univariate or multivariate outlier.

Covariance matrix

The covariance matrix provides a useful tool for separating the structured relationships in a matrix of random variables. This can be used to decorrelate variables or applied as a transform to other variables. It is a key element used in the Principal Component Analysis data reduction method. Then we can look at relationship between the variables.

Table 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		*Y1*	*Y2*	*Y3*	*Y4*	*Y5*	*Y6*	*Y7*	*Y8*	*Y9*	*Y10*	*Y11*	*Y12*	*Y13*	*Y14*	*Y15*	*Y16*	*Y17*
2	Y1	25.53276																
3	Y2	0.038214	0.033701															
4	Y3	27.48073	69.57859	473848.8														
5	Y4	-7.33738	-0.39649	-2113.74	221.831													
6	Y5	183.9324	-18.1465	-35480.7	556.3475	52080.44												
7	Y6	-0.06768	-0.00726	-14.9226	0.185285	9.887074	0.03675											
8	Y7	0.157518	-0.07784	67.12301	-0.48354	130.5651	0.042644	4.489689										
9	Y8	25.01899	0.614838	-2061.2	78.28246	108.9726	-0.18956	-14.5762	798.1164									
10	Y9	-0.00991	-0.00158	1.279969	0.114139	3.052364	0.000588	0.017256	0.000988	0.002639								
11	Y10	10.35858	-1.51144	-4926.22	-28.0608	3372.21	0.76742	10.78038	18.38556	0.301579	523.2856							
12	Y11	-1.93077	-0.79332	-3444.35	-30.5454	1585.438	0.348351	3.96213	-34.1021	0.073235	204.0981	243.0072						
13	Y12	4.848896	-1.3736	-2880.35	-29.5134	3353.344	0.952333	-2.52434	28.36209	0.543059	427.5812	204.2014	621.8146					
14	Y13	13.76141	-1.30315	-3904.48	-22.4689	2923.458	0.550796	6.875939	18.10016	0.231023	408.9426	174.3106	356.9681	335.7346				
15	Y14	3.473662	-0.29594	-766.118	-2.20101	641.5139	0.117387	1.707194	2.777589	0.054316	72.12043	33.77809	71.36838	61.39959	12.86394			
16	Y15	-7.44214	1.283337	8803.551	-56.1701	-103.91	0.002422	0.768654	-36.7255	0.174783	-16.3985	11.4066	88.76427	-8.83994	-1.73948	256.7839		
17	Y16	-0.29069	-0.11647	-458.945	-2.16793	144.1512	0.037727	-0.72357	0.535196	0.028348	23.74808	14.05023	32.94502	19.70139	3.609542	-2.00827	3.280932	
18	Y17	31.96426	-1.69162	-2320.49	-8.97376	4272.192	0.709281	3.553044	93.75943	0.606748	457.772	169.2062	573.4973	395.8676	83.85993	70.47112	26.88445	742.5948

Second Assumption –Check normality of variables

Draw the plots and their respective regression lines. Normal Q-Q plots were constructed based on each variable and the plots that violate the assumptions, Regression line should have an angle of nearly 45 degrees. If curve must go through the regression line then the variables normal. Else not normal the variable.plot is a graphical technique for representing a data set, usually as a graph showing the

relationship between two or more variables. The plot can be drawn by hand or by a computer. In the past, sometimes mechanical or electronic plotters were used.

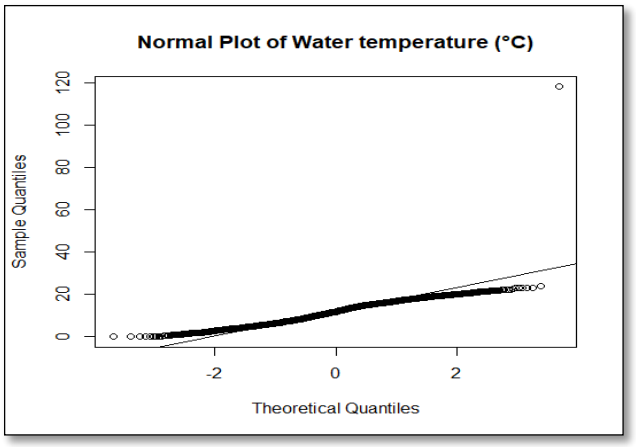

Figure 1

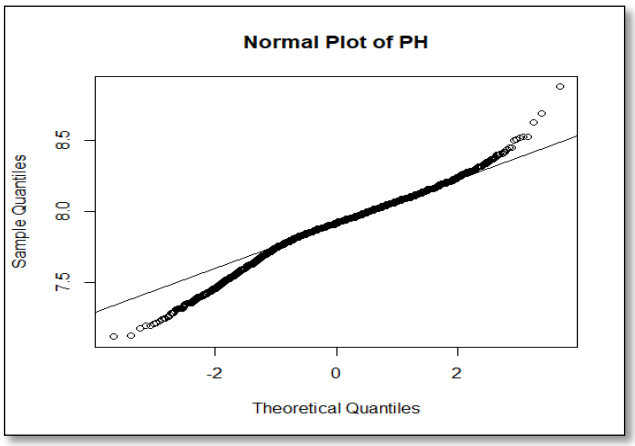

Figure 2

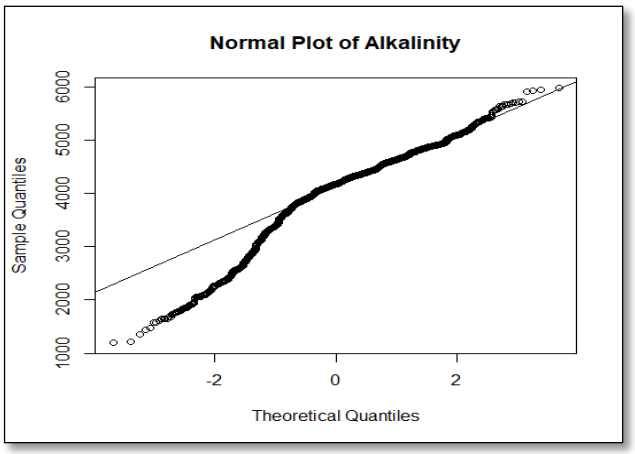

Figure 3

u

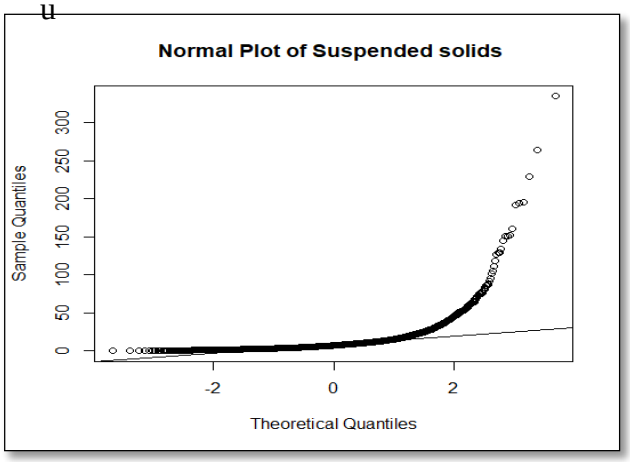

Figure 4

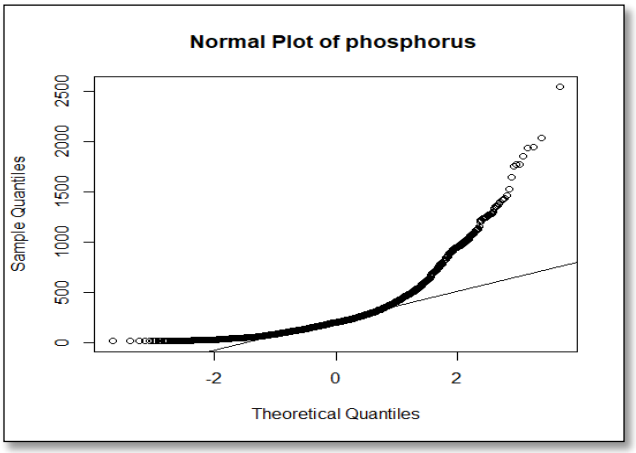

Figure 5

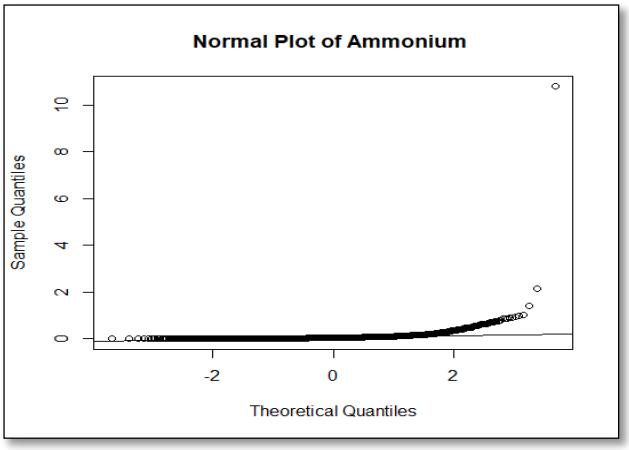

Figure 6

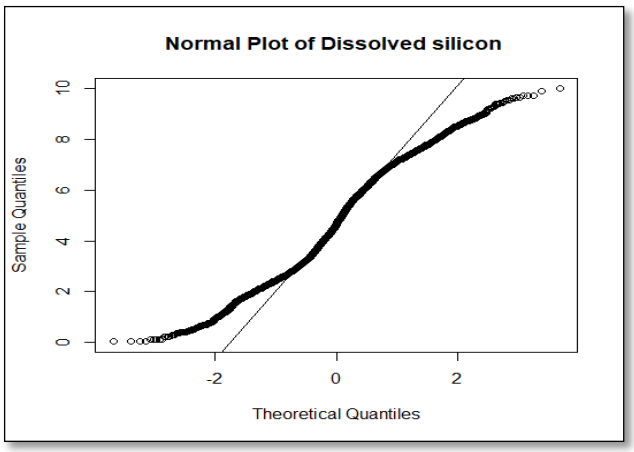

Figure 7

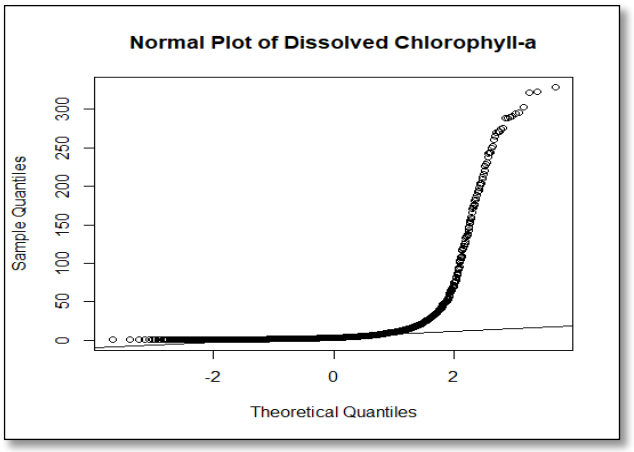

Figure 8

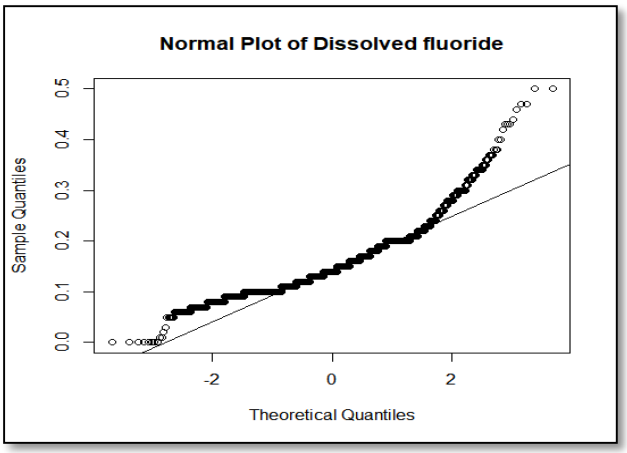

Figure 10

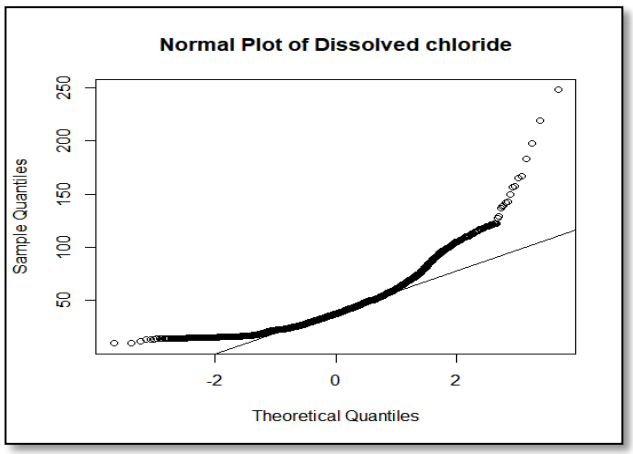

Figure 9

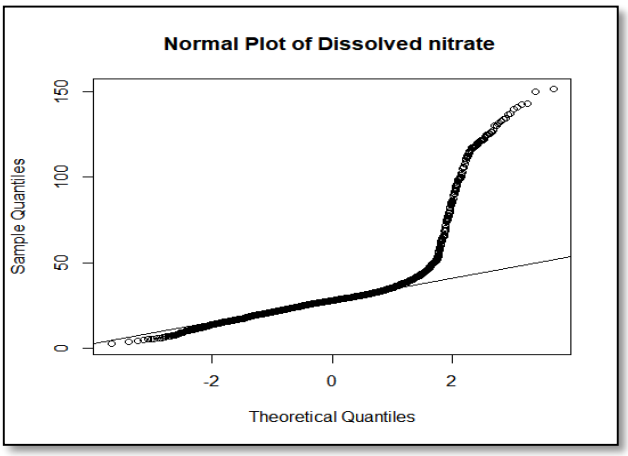

Figure 11

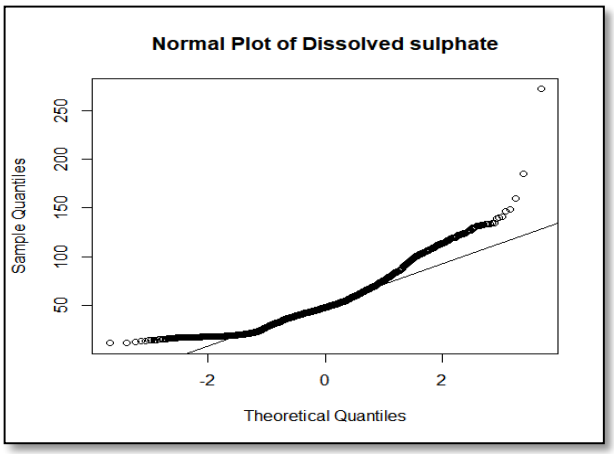

Figure 12

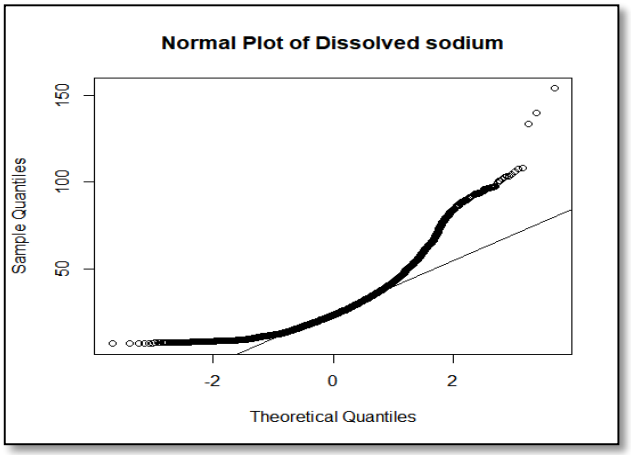

Figure 13

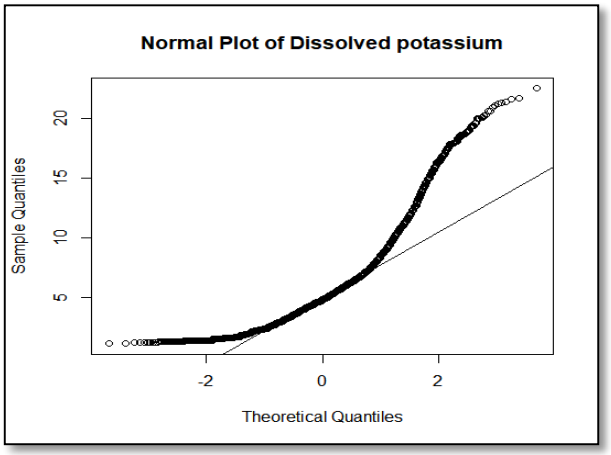

Figure 14

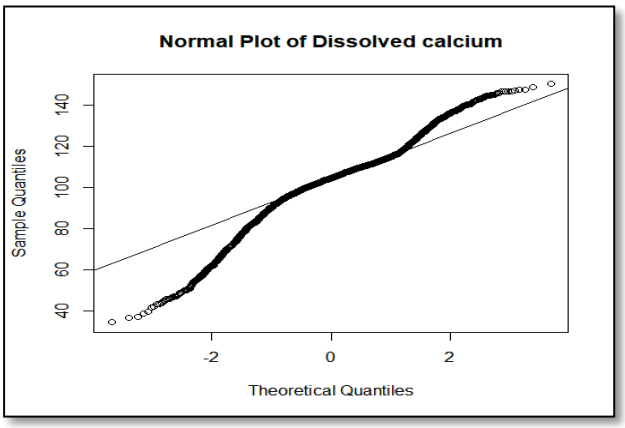

Figure 15

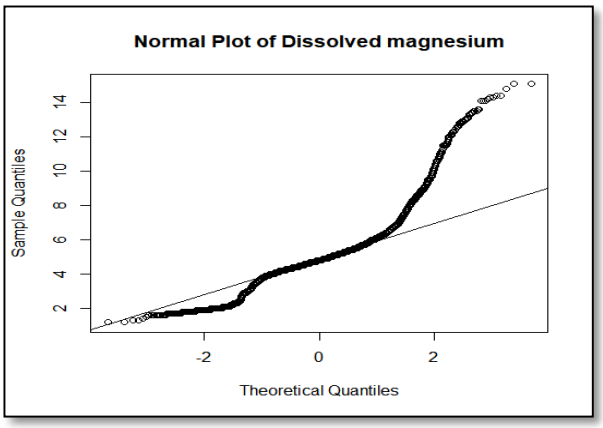

Figure 16

Graphs are a visual representation of the relationship between variables.

A **histogram** is an approximate representation of the distribution of numerical data.

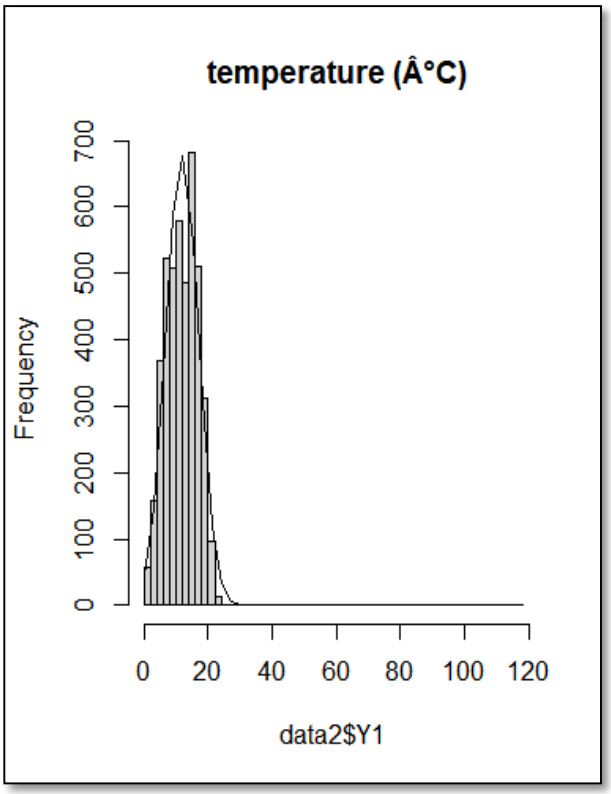

Figure 17

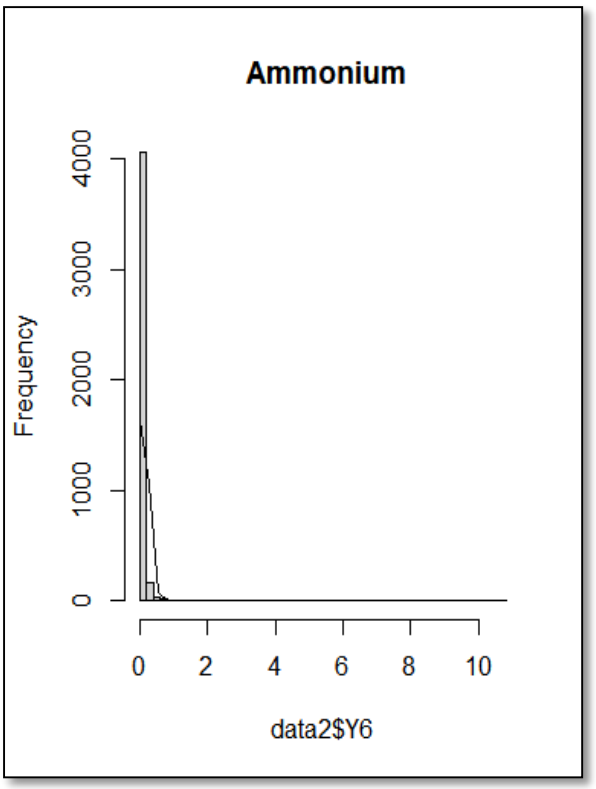

Figure 18

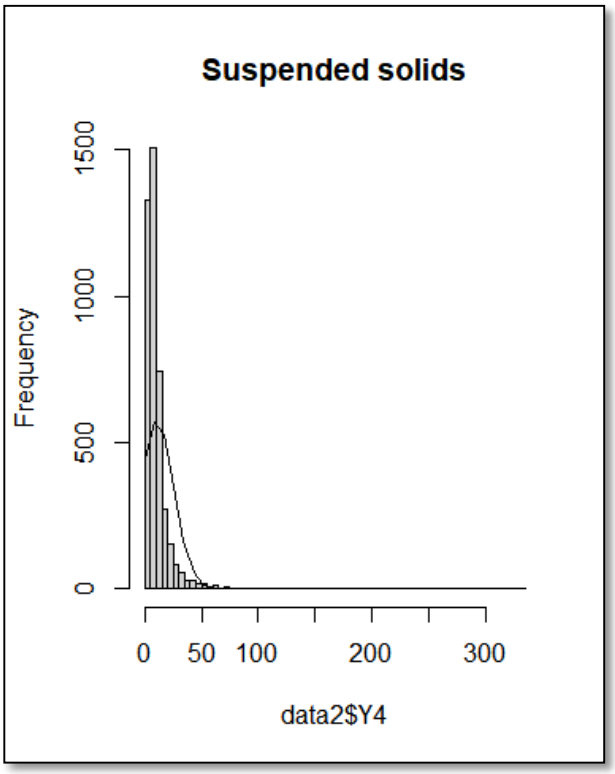

Figure 19

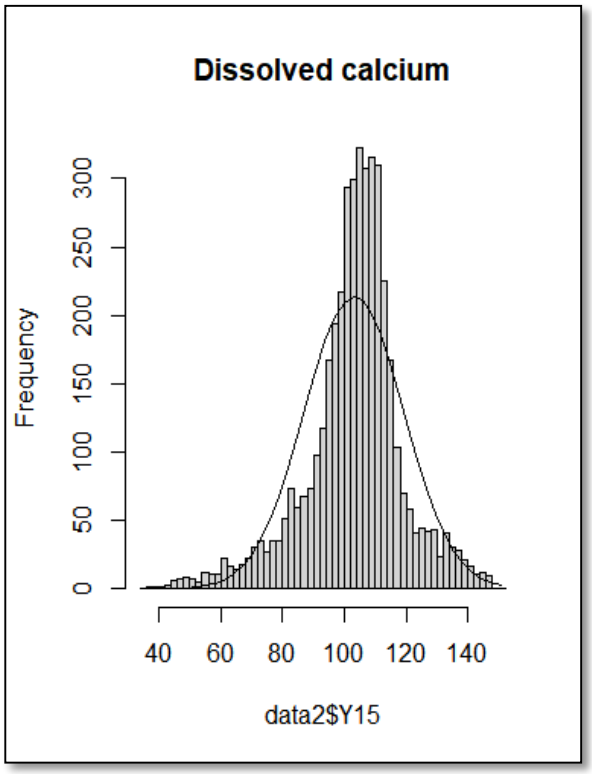

Figure 20

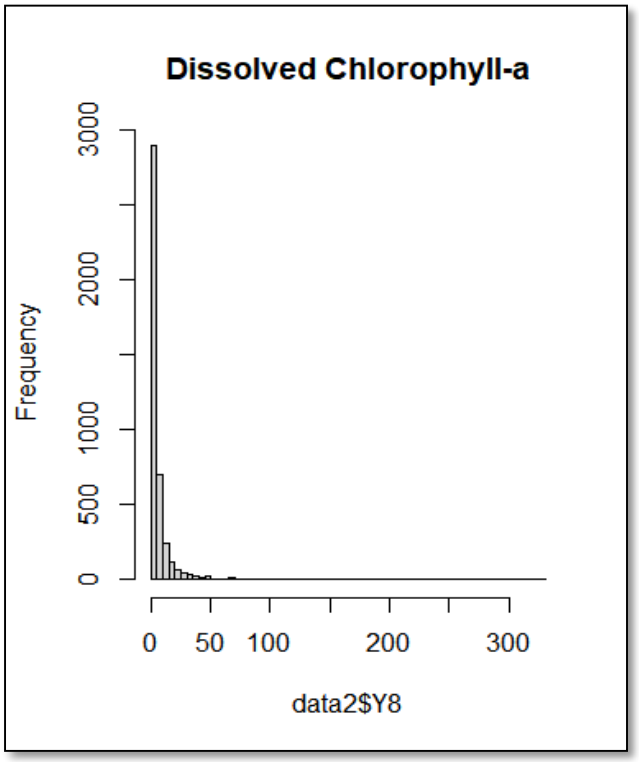

Figure 21

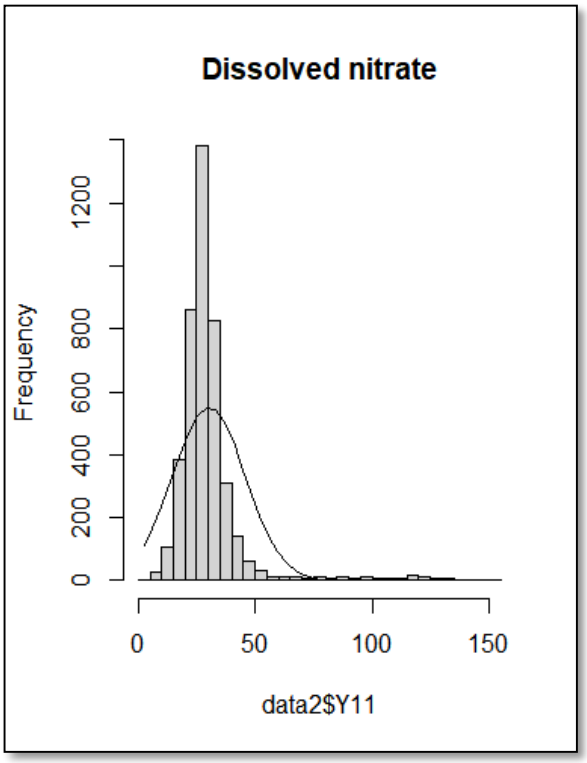

Figure 22

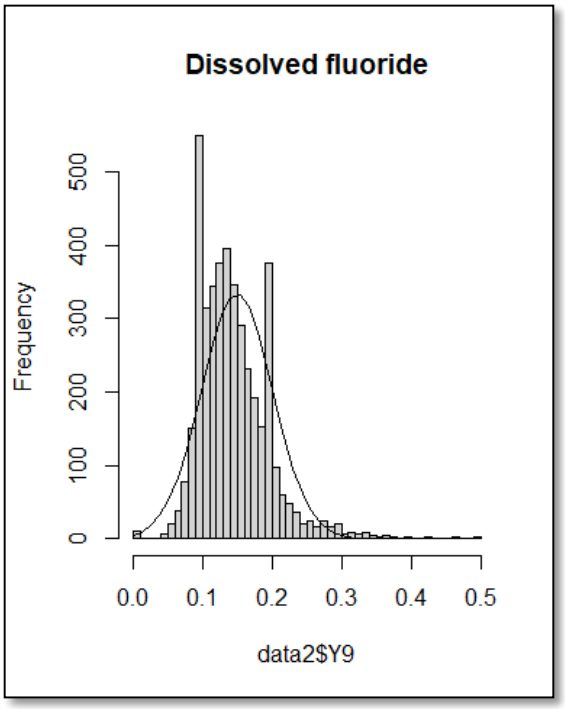

Figure 23

We can see shape of ***Dissolved Calcium*** and ***Dissolved Fluoride*** are bell shaped. Therefore we can assume these two are variables normal.

Correlation Matrix

```
#Reject variables from non normality

data3<-data2[,!(names(data2)%in%c("Y1","Y4","Y6","Y8","Y11","date","time"))]
```

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.

		Y3	*Y5*	*Y7*	*Y9*	*Y10*	*Y12*	*Y13*	*Y14*	*Y15*	*Y16*	*Y17*
2	Y3	1										
3	Y5	-0.22586	1									
4	Y7	0.04602	0.270011	1								
5	Y9	0.036194	0.260347	0.158523	1							
6	Y10	-0.31284	0.645963	0.222411	0.256617	1						
7	Y12	-0.1678	0.589265	-0.04778	0.423907	0.749581	1					
8	Y13	-0.30956	0.699136	0.177103	0.24542	0.975652	0.78127	1				
9	Y14	-0.3103	0.783757	0.22464	0.294778	0.879026	0.797973	0.934287	1			
10	Y15	0.798094	-0.02841	0.022638	0.212309	-0.04474	0.222138	-0.03011	-0.03027	1		
11	Y16	-0.36808	0.348724	-0.18853	0.304638	0.57314	0.729391	0.593609	0.555605	-0.06919	1	
12	Y17	-0.1237	0.68697	0.061534	0.433397	0.734351	0.843966	0.792823	0.858008	0.161381	0.544661	1

Table 2

If correlation is greater than 0.06 reject them. Then reobtain the correlation matrix after removing rejected variables.

```
data4<data3[,!(names(data3)%in%c("Y17","Y10","Y12","Y13","Y15","Y5"))]

write.csv(data4,"preprocessed_new_1.csv")
```

	A	B	C	D	E	F	G
1		*Y2*	*Y3*	*Y7*	*Y9*	*Y14*	*Y16*
2	Y2	1					
3	Y3	0.550595	1				
4	Y7	-0.20012	0.04602	1			
5	Y9	-0.16782	0.036194	0.158523	1		
6	Y14	-0.44946	-0.3103	0.22464	0.294778	1	
7	Y16	-0.35027	-0.36808	-0.18853	0.304638	0.555605	1

Table 3

pH, Alkalinity, Dissolved Silicon, Dissolved Fluoride, Dissolved Potassium, Dissolved Magnesium are the independent variables. After that going to obtain MANOVA table.

MANOVA Table

The term MANOVA comes from Multiple Analysis Of Variance, and refers to a well established technique for comparing multivariate population means of several groups. This is done by essentially comparing the variance covariance between variables to test the statistical significance of the mean differences.

I used to four MANOVA test for Analysis Of Variance, and refers to a well established technique for comparing multivariate population means of several groups

1. Pillai test
2. Wilk's test
3. Roy's test
4. Lawley-Hotelling test

```
> ocdmodel1 = manova(cbind(Y2,Y3,Y7,Y9,Y14,Y16)~Sites,data = data2)
> summary(ocdmodel1)
           Df Pillai approx F num Df den Df    Pr(>F)
Sites      21 2.8604   185.43    126  25644 < 2.2e-16 ***
Residuals 4274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ocdmodel1,test = "Roy")
           Df   Roy approx F num Df den Df    Pr(>F)
Sites      21 5.462   1111.6     21   4274 < 2.2e-16 ***
Residuals 4274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ocdmodel1,test = "Wilks")
           Df     Wilks approx F num Df den Df    Pr(>F)
Sites      21 0.0075899   259.48    126  24763 < 2.2e-16 ***
Residuals 4274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ocdmodel1,test = "Hotelling-Lawley")
           Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
Sites      21           10.494    355.4    126  25604 < 2.2e-16 ***
Residuals 4274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

According to F test all the variables are significant

ANOVA Table

```
 Response Y2 :
             Df Sum Sq Mean Sq F value    Pr(>F)
Sites        21 59.780 2.84668  143.14 < 2.2e-16 ***
Residuals  4274 85.001 0.01989
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y3 :
             Df     Sum Sq  Mean Sq F value    Pr(>F)
Sites        21 1217356837 57969373  302.78 < 2.2e-16 ***
Residuals  4274  818297668   191459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y5 :
             Df    Sum Sq Mean Sq F value    Pr(>F)
Sites        21 128469617 6117601  274.45 < 2.2e-16 ***
Residuals  4274  95267954   22290
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y7 :
             Df Sum Sq Mean Sq F value    Pr(>F)
Sites        21  12337  587.46  361.22 < 2.2e-16 ***
Residuals  4274   6951    1.63
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y12 :
             Df  Sum Sq Mean Sq F value    Pr(>F)
Sites        21 1967346   93683  568.78 < 2.2e-16 ***
Residuals  4274  703969     165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y13 :
             Df Sum Sq Mean Sq F value    Pr(>F)
Sites        21 894945   42616  332.76 < 2.2e-16 ***
Residuals  4274 547371     128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y16 :
             Df Sum Sq Mean Sq F value    Pr(>F)
Sites        21  11255  535.95  806.58 < 2.2e-16 ***
Residuals  4274   2840    0.66
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Y17 :
             Df  Sum Sq Mean Sq F value    Pr(>F)
Sites        21 2366077  112670  584.33 < 2.2e-16 ***
Residuals  4274  824110     193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to F test all the variables are significant.

Results and Conclusion

By using MANOVA,

checked whether mean difference between thames river sites at least one variable. From MANOVA output I can say that river sites have mean difference. From MANOVA test p-value is close to zero ($p<0.05$) . therefore reject the sample.

We have found independent variable from the Q-Q flot and the histogram. Then we created MANOVA table from those independent variables and there we found mean difference in MANOVA table but we could not find what are the independent variables caused to mean difference.

By using ANOVA,

Check whether what sites have the mean difference.

I found ANOVA for all variable separately .and found that mean difference between sites for all variables. And their p-values are close to zero ($p<0.05$) .

For the river Thames to support a variety of wildlife, the water mustn't be too acid or alkali. In the past, the pH of the Thames would have been affected by pollution from industry, killing all wildlife.

Reference

https://www.wwdmag.com/channel/casestudies/monitoring-water-quality-thames

https://en.wikipedia.org/wiki/River_Thames

http://www.environmentdata.org/archive/ealit:1710/OBJ/20000756.pdf

http://nora.nerc.ac.uk/id/eprint/520909/