Water quality analysis of River Thames

S/16/499

2/28/2021

# Water Quality of River Thames

### Importing the dataset

waterQualitydf<-read.csv("F:\\rStudio Projects\\ST305\\Assignment\\Assignment 1\\River\_Thames\_Water\_Quality.csv",header = TRUE)

### Head of the data

head(waterQualitydf)

Site Sampling.date..dd.mm.yyyy. Sampling.time..hh.mm.  
1 River Thame at Wheatley 3/03/2009 9:25  
2 River Thame at Wheatley 9/03/2009 9:40  
3 River Thame at Wheatley 16/03/2009 10:00  
4 River Thame at Wheatley 24/03/2009 9:45  
5 River Thame at Wheatley 1/04/2009 9:46  
6 River Thame at Wheatley 6/04/2009 9:48  
 Water.temperature...C. pH Alkalinity..µ.equ.l.1.  
1 7.2 8.01 4915  
2 6.8 7.94 5637  
3 9.3 8.05 5393  
4 7.8 8.14 5351  
5 8.9 8.20 5129  
6 11.3 8.20 5067  
 Suspended.solids......mg.l.1. phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
1 7.7 438 0.2  
2 7.5 341 0.232  
3 5.3 415 0.176  
4 6.0 381 0.364  
5 4.4 480 0.384  
6 5.4 568 0.292  
 Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 5.8 6.93  
2 5.3 9.56  
3 4.4 8.88  
4 2.8 29.21  
5 2.3 17.63  
6 2.3 21.03  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 0.2 41.0  
2 0.2 42.5  
3 0.2 43.5  
4 0.2 46.0  
5 0.2 48.5  
6 0.2 47.5  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 34.0 77.0  
2 30.5 81.5  
3 30.5 80.5  
4 36.5 76.0  
5 34.5 70.0  
6 35.5 68.0  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 26.7 6.5  
2 29.7 6.5  
3 29.4 7.1  
4 34.5 8.0  
5 36.9 9.0  
6 34.2 8.9  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 140.0 6.0  
2 139.1 6.4  
3 142.9 6.3  
4 141.3 6.1  
5 145.8 6.3  
6 142.7 6.1  
 Dissolved.boron....µg.l.1.  
1 81  
2 88  
3 89  
4 83  
5 79  
6 91

### Changing the data type of last column(“Dissolved boron (µg l-1)”)

waterQualitydf$Ammonium..mg.l.1.NH4.<-as.numeric(waterQualitydf$Ammonium..mg.l.1.NH4.)

Warning: NAs introduced by coercion

waterQualitydf$Dissolved.silicon..mg.l.1.Si.<-as.numeric(waterQualitydf$Dissolved.silicon..mg.l.1.Si.)

Warning: NAs introduced by coercion

waterQualitydf$Dissolved.fluoride..mg.l.1.<-as.numeric(waterQualitydf$Dissolved.fluoride..mg.l.1.)

Warning: NAs introduced by coercion

waterQualitydf$Dissolved.boron....µg.l.1.<-as.numeric(waterQualitydf$Dissolved.boron....µg.l.1.)

Warning: NAs introduced by coercion

waterQualitydf$Sampling.date..dd.mm.yyyy.<-as.Date(waterQualitydf$Sampling.date..dd.mm.yyyy.,format="%d/%m/%Y")

### Counting missing values and removing them

waterQualitydf %>%  
 select(everything()) %>% # replace to your needs  
 summarise\_all(~(sum(is.na(.))))

Site Sampling.date..dd.mm.yyyy. Sampling.time..hh.mm. Water.temperature...C.  
1 0 21 0 58  
 pH Alkalinity..µ.equ.l.1. Suspended.solids......mg.l.1. phosphorus..µg.l.1.P.  
1 47 47 49 36  
 Ammonium..mg.l.1.NH4. Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 115 28 52  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 46 25  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 26 25  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 25 26  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 25 25  
 Dissolved.boron....µg.l.1.  
1 26

Removing missing values from the data set

df<-na.omit(waterQualitydf)  
head(df)

Site Sampling.date..dd.mm.yyyy. Sampling.time..hh.mm.  
1 River Thame at Wheatley 2009-03-03 9:25  
2 River Thame at Wheatley 2009-03-09 9:40  
3 River Thame at Wheatley 2009-03-16 10:00  
4 River Thame at Wheatley 2009-03-24 9:45  
5 River Thame at Wheatley 2009-04-01 9:46  
6 River Thame at Wheatley 2009-04-06 9:48  
 Water.temperature...C. pH Alkalinity..µ.equ.l.1.  
1 7.2 8.01 4915  
2 6.8 7.94 5637  
3 9.3 8.05 5393  
4 7.8 8.14 5351  
5 8.9 8.20 5129  
6 11.3 8.20 5067  
 Suspended.solids......mg.l.1. phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
1 7.7 438 0.200  
2 7.5 341 0.232  
3 5.3 415 0.176  
4 6.0 381 0.364  
5 4.4 480 0.384  
6 5.4 568 0.292  
 Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 5.8 6.93  
2 5.3 9.56  
3 4.4 8.88  
4 2.8 29.21  
5 2.3 17.63  
6 2.3 21.03  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 0.2 41.0  
2 0.2 42.5  
3 0.2 43.5  
4 0.2 46.0  
5 0.2 48.5  
6 0.2 47.5  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 34.0 77.0  
2 30.5 81.5  
3 30.5 80.5  
4 36.5 76.0  
5 34.5 70.0  
6 35.5 68.0  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 26.7 6.5  
2 29.7 6.5  
3 29.4 7.1  
4 34.5 8.0  
5 36.9 9.0  
6 34.2 8.9  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 140.0 6.0  
2 139.1 6.4  
3 142.9 6.3  
4 141.3 6.1  
5 145.8 6.3  
6 142.7 6.1  
 Dissolved.boron....µg.l.1.  
1 81  
2 88  
3 89  
4 83  
5 79  
6 91

Re-check

df %>%  
 select(everything()) %>% # replace to your needs  
 summarise\_all(~(sum(is.na(.))))

Site Sampling.date..dd.mm.yyyy. Sampling.time..hh.mm. Water.temperature...C.  
1 0 0 0 0  
 pH Alkalinity..µ.equ.l.1. Suspended.solids......mg.l.1. phosphorus..µg.l.1.P.  
1 0 0 0 0  
 Ammonium..mg.l.1.NH4. Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 0 0 0  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 0 0  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 0 0  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 0 0  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 0 0  
 Dissolved.boron....µg.l.1.  
1 0

Types of columns in the dataframe

glimpse(df)

Rows: 4,140  
Columns: 20  
$ Site <chr> "River Thame at Wheatley", "River ~  
$ Sampling.date..dd.mm.yyyy. <date> 2009-03-03, 2009-03-09, 2009-03-1~  
$ Sampling.time..hh.mm. <chr> "9:25", "9:40", "10:00", "9:45", "~  
$ Water.temperature...C. <dbl> 7.2, 6.8, 9.3, 7.8, 8.9, 11.3, 11.~  
$ pH <dbl> 8.01, 7.94, 8.05, 8.14, 8.20, 8.20~  
$ Alkalinity..µ.equ.l.1. <int> 4915, 5637, 5393, 5351, 5129, 5067~  
$ Suspended.solids......mg.l.1. <dbl> 7.70, 7.50, 5.30, 6.00, 4.40, 5.40~  
$ phosphorus..µg.l.1.P. <int> 438, 341, 415, 381, 480, 568, 568,~  
$ Ammonium..mg.l.1.NH4. <dbl> 0.200, 0.232, 0.176, 0.364, 0.384,~  
$ Dissolved.silicon..mg.l.1.Si. <dbl> 5.8, 5.3, 4.4, 2.8, 2.3, 2.3, 4.6,~  
$ Chlorophyll.a..µg.l.1. <dbl> 6.93, 9.56, 8.88, 29.21, 17.63, 21~  
$ Dissolved.fluoride..mg.l.1. <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2,~  
$ Dissolved.chloride..mg.l.1. <dbl> 41.0, 42.5, 43.5, 46.0, 48.5, 47.5~  
$ Dissolved.nitrate......mg.l.1.NO3. <dbl> 34.0, 30.5, 30.5, 36.5, 34.5, 35.5~  
$ Dissolved.sulphate.....mg.l.1.SO4. <dbl> 77.0, 81.5, 80.5, 76.0, 70.0, 68.0~  
$ Dissolved.sodium..mg.l.1. <dbl> 26.7, 29.7, 29.4, 34.5, 36.9, 34.2~  
$ Dissolved.potassium..mg.l.1. <dbl> 6.5, 6.5, 7.1, 8.0, 9.0, 8.9, 9.5,~  
$ Dissolved.calcium...........mg.l.1. <dbl> 140.0, 139.1, 142.9, 141.3, 145.8,~  
$ Dissolved.magnesium..mg.l.1. <dbl> 6.0, 6.4, 6.3, 6.1, 6.3, 6.1, 6.3,~  
$ Dissolved.boron....µg.l.1. <dbl> 81, 88, 89, 83, 79, 91, 94, 96, 10~

### Summary of the data set

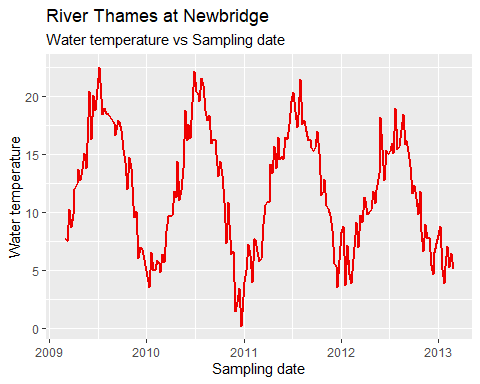
summary(df[-c(1,2,3)])

Water.temperature...C. pH Alkalinity..µ.equ.l.1.  
 Min. : 0.00 Min. :7.120 Min. :1191   
 1st Qu.: 7.90 1st Qu.:7.810 1st Qu.:3789   
 Median : 11.90 Median :7.920 Median :4179   
 Mean : 11.83 Mean :7.907 Mean :4047   
 3rd Qu.: 15.60 3rd Qu.:8.020 3rd Qu.:4465   
 Max. :118.00 Max. :8.880 Max. :5976   
 Suspended.solids......mg.l.1. phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
 Min. : 0.00 Min. : 11.0 Min. :0.00000   
 1st Qu.: 4.42 1st Qu.: 115.0 1st Qu.:0.03200   
 Median : 7.30 Median : 199.0 Median :0.05000   
 Mean : 10.95 Mean : 258.6 Mean :0.07788   
 3rd Qu.: 12.01 3rd Qu.: 317.0 3rd Qu.:0.08500   
 Max. :334.62 Max. :2545.0 Max. :2.16000   
 Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
 Min. : 0.020 Min. : 0.210   
 1st Qu.: 2.920 1st Qu.: 1.800   
 Median : 4.660 Median : 3.050   
 Mean : 4.741 Mean : 9.509   
 3rd Qu.: 6.562 3rd Qu.: 6.372   
 Max. :10.000 Max. :328.500   
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
 Min. :0.0000 Min. : 9.63   
 1st Qu.:0.1100 1st Qu.: 25.51   
 Median :0.1400 Median : 37.48   
 Mean :0.1501 Mean : 42.13   
 3rd Qu.:0.1800 3rd Qu.: 51.43   
 Max. :0.5000 Max. :248.06   
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
 Min. : 2.39 Min. : 10.70   
 1st Qu.: 23.36 1st Qu.: 35.79   
 Median : 27.79 Median : 47.51   
 Mean : 30.26 Mean : 51.85   
 3rd Qu.: 32.02 3rd Qu.: 64.39   
 Max. :151.33 Max. :184.98   
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
 Min. : 6.50 Min. : 1.1   
 1st Qu.: 14.40 1st Qu.: 3.1   
 Median : 23.10 Median : 4.8   
 Mean : 27.64 Mean : 5.6   
 3rd Qu.: 34.60 3rd Qu.: 6.9   
 Max. :154.20 Max. :22.5   
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
 Min. : 34.2 Min. : 1.200   
 1st Qu.: 96.3 1st Qu.: 4.200   
 Median :104.5 Median : 4.800   
 Mean :103.0 Mean : 4.973   
 3rd Qu.:111.4 3rd Qu.: 5.500   
 Max. :150.5 Max. :15.100   
 Dissolved.boron....µg.l.1.  
 Min. : 5.00   
 1st Qu.: 31.00   
 Median : 52.00   
 Mean : 53.91   
 3rd Qu.: 67.20   
 Max. :184.00

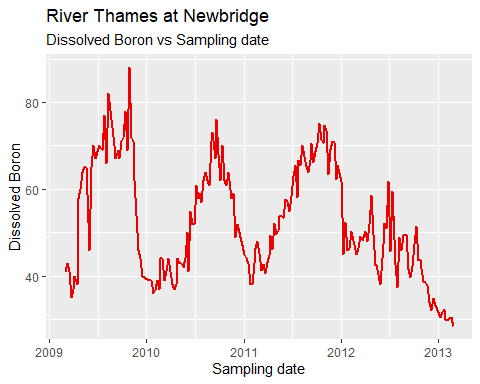
options(scipen=100)  
options(digits=2)  
a<-stat.desc(df[-c(1,2,3)])  
  
write.csv(a,"F:\\rStudio Projects\\ST305\\Assignment\\Assignment 1\\summary1.csv",row.names = TRUE)

### Plots and diagrams drawn versus time(Month)

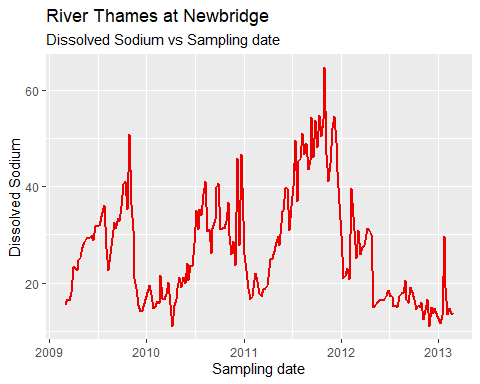
#unique(df$Site)  
  
df1<-df %>%  
 select(everything()) %>%  
 filter(Site=="River Thames at Newbridge")  
#head(df1)   
  
ggplot(df1)+geom\_line(aes(x=Sampling.date..dd.mm.yyyy.,y=Water.temperature...C.),color="red2",size=1)+  
 labs(x="Sampling date",y="Water temperature",title = "River Thames at Newbridge",subtitle = "Water temperature vs Sampling date")



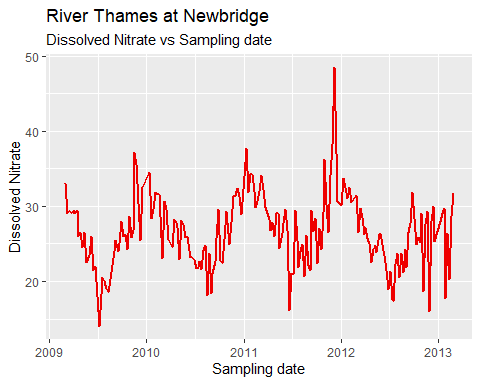
ggplot(df1)+geom\_line(aes(x=Sampling.date..dd.mm.yyyy.,y=Dissolved.boron....µg.l.1.),color="red2",size=1)+  
 labs(x="Sampling date",y="Dissolved Boron",title = "River Thames at Newbridge",subtitle = "Dissolved Boron vs Sampling date")



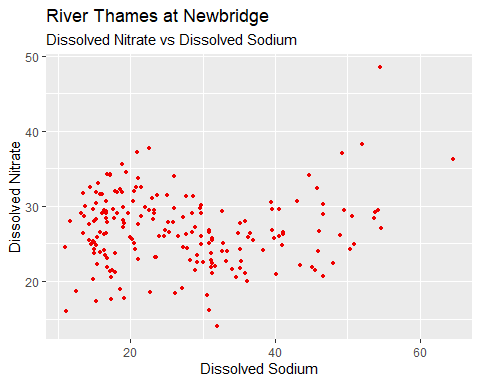
ggplot(df1)+geom\_line(aes(x=Sampling.date..dd.mm.yyyy.,y=Dissolved.sodium..mg.l.1.),color="red2",size=1)+  
 labs(x="Sampling date",y="Dissolved Sodium",title = "River Thames at Newbridge",subtitle = "Dissolved Sodium vs Sampling date")



ggplot(df1)+geom\_line(aes(x=Sampling.date..dd.mm.yyyy.,y=Dissolved.nitrate......mg.l.1.NO3.),color="red2",size=1)+  
 labs(x="Sampling date",y="Dissolved Nitrate",title = "River Thames at Newbridge",subtitle = "Dissolved Nitrate vs Sampling date")



ggplot(df1)+geom\_point(aes(x=Dissolved.sodium..mg.l.1.,y=Dissolved.nitrate......mg.l.1.NO3.),color="red2",size=1)+  
 labs(x="Dissolved Sodium",y="Dissolved Nitrate",title = "River Thames at Newbridge",subtitle = "Dissolved Nitrate vs Dissolved Sodium")



## Manova

MANOVA analysis was done at 5% significance level H0 : Mean water properties of each month is equal vs H1: Mean water properties of at least 2 months are not equal

Creating a new data frame to do MANOVA

df\_manova<-df %>%   
 select(everything()) %>%  
 group\_by(Site)  
  
samplingYM<-format(df\_manova$Sampling.date..dd.mm.yyyy.,"%Y-%m")  
  
df\_manova$samplingYM<-samplingYM  
  
df\_manova= subset(df\_manova, select = -c(Sampling.time..hh.mm.,Sampling.date..dd.mm.yyyy.) )  
head(df\_manova)

# A tibble: 6 x 19  
# Groups: Site [1]  
 Site Water.temperature.~ pH Alkalinity..µ.equ~ Suspended.solids...~  
 <chr> <dbl> <dbl> <int> <dbl>  
1 River Thame~ 7.2 8.01 4915 7.7  
2 River Thame~ 6.8 7.94 5637 7.5  
3 River Thame~ 9.3 8.05 5393 5.3  
4 River Thame~ 7.8 8.14 5351 6   
5 River Thame~ 8.9 8.2 5129 4.4  
6 River Thame~ 11.3 8.2 5067 5.4  
# ... with 14 more variables: phosphorus..µg.l.1.P. <int>,  
# Ammonium..mg.l.1.NH4. <dbl>, Dissolved.silicon..mg.l.1.Si. <dbl>,  
# Chlorophyll.a..µg.l.1. <dbl>, Dissolved.fluoride..mg.l.1. <dbl>,  
# Dissolved.chloride..mg.l.1. <dbl>,  
# Dissolved.nitrate......mg.l.1.NO3. <dbl>,  
# Dissolved.sulphate.....mg.l.1.SO4. <dbl>, Dissolved.sodium..mg.l.1. <dbl>,  
# Dissolved.potassium..mg.l.1. <dbl>,  
# Dissolved.calcium...........mg.l.1. <dbl>,  
# Dissolved.magnesium..mg.l.1. <dbl>, Dissolved.boron....µg.l.1. <dbl>,  
# samplingYM <chr>

Mean values when grouped by Site and Sample taken date

dependent variable extraction

d.v<-as.matrix(df\_manova  
 [2:18])

***By SITE***

df\_groupedbySite<-aggregate(d.v~df\_manova$Site,data = df\_manova, function(x)round(mean(x),2))  
colnames(df\_groupedbySite)[1]<-"Site"  
head(df\_groupedbySite,n=10L)

Site Water.temperature...C. pH  
1 Jubilee River at Pocock's Bridge 13 8.0  
2 River Cherwell at Hampton Poyle 12 7.9  
3 River Cole at Lynt Bridge 12 7.9  
4 River Coln at Whelford 12 8.0  
5 River Enborne at Brimpton 11 7.8  
6 River Evenlode at Cassington Mill 11 7.9  
7 River Kennet at Woolhampton 11 8.0  
8 River Leach at Mill Lane,Lechlade 11 7.9  
9 River Lodden at Charvil 12 7.8  
10 River Ock at Abingdon 12 8.0  
 Alkalinity..µ.equ.l.1. Suspended.solids......mg.l.1. phosphorus..µg.l.1.P.  
1 4088 8.4 192  
2 4134 13.3 193  
3 4335 15.2 307  
4 4247 5.4 84  
5 2819 9.5 183  
6 4028 15.7 252  
7 4500 9.3 78  
8 4367 3.0 34  
9 3209 7.3 209  
10 4702 11.1 320  
 Ammonium..mg.l.1.NH4. Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 0.07 5.2 18.7  
2 0.04 3.3 14.1  
3 0.05 6.4 5.7  
4 0.04 2.6 3.0  
5 0.08 6.9 2.5  
6 0.04 2.7 12.4  
7 0.05 6.8 8.2  
8 0.06 2.4 1.9  
9 0.08 5.4 3.9  
10 0.06 7.1 3.9  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 0.15 44  
2 0.20 54  
3 0.19 46  
4 0.13 17  
5 0.12 34  
6 0.12 26  
7 0.12 24  
8 0.10 16  
9 0.12 60  
10 0.20 39  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 26 47  
2 25 65  
3 18 53  
4 26 34  
5 17 26  
6 25 46  
7 24 20  
8 31 35  
9 34 48  
10 30 72  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 27.4 5.4  
2 35.6 6.2  
3 27.4 5.3  
4 8.8 1.7  
5 17.8 3.6  
6 16.2 3.5  
7 12.4 2.4  
8 8.3 1.5  
9 38.6 7.5  
10 25.0 5.9  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 102 4.4  
2 104 7.6  
3 110 4.4  
4 101 5.7  
5 68 4.4  
6 102 4.2  
7 107 2.2  
8 109 5.1  
9 83 5.2  
10 126 4.6  
 Dissolved.boron....µg.l.1.  
1 54  
2 73  
3 55  
4 20  
5 26  
6 51  
7 22  
8 25  
9 56  
10 62

***By Sample taken date***

head(aggregate(d.v~df\_manova$samplingYM,data = df\_manova,function(x)round(mean(x),2)),n=10L)

df\_manova$samplingYM Water.temperature...C. pH Alkalinity..µ.equ.l.1.  
1 2009-03 8.8 8.1 4464  
2 2009-04 11.9 8.1 4128  
3 2009-05 14.6 7.9 3890  
4 2009-06 18.4 7.9 3790  
5 2009-07 18.4 7.9 3953  
6 2009-08 18.3 7.8 4014  
7 2009-09 16.9 8.0 4144  
8 2009-10 13.8 7.9 4128  
9 2009-11 10.4 7.8 3699  
10 2009-12 7.2 7.8 3713  
 Suspended.solids......mg.l.1. phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
1 6.8 159 0.05  
2 7.6 222 0.07  
3 11.7 302 0.06  
4 12.0 357 0.05  
5 9.5 318 0.04  
6 9.0 324 0.05  
7 9.0 361 0.04  
8 7.1 370 0.06  
9 21.7 317 0.09  
10 23.3 215 0.08  
 Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 3.9 9.6  
2 2.8 32.0  
3 3.4 57.7  
4 4.3 43.6  
5 4.8 21.8  
6 5.2 9.7  
7 4.9 11.2  
8 5.0 6.7  
9 5.0 4.6  
10 4.8 2.5  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 0.13 36  
2 0.15 39  
3 0.13 43  
4 0.15 44  
5 0.18 45  
6 0.18 44  
7 0.13 50  
8 0.10 55  
9 0.15 38  
10 0.20 29  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 35 53  
2 35 52  
3 35 55  
4 33 54  
5 30 51  
6 31 55  
7 32 58  
8 34 62  
9 30 53  
10 32 49  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 23 4.3  
2 27 5.4  
3 31 6.2  
4 30 6.3  
5 31 6.3  
6 30 6.6  
7 36 7.4  
8 41 8.5  
9 25 6.1  
10 17 4.3  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 115 5.4  
2 116 5.6  
3 108 5.6  
4 104 5.2  
5 99 4.9  
6 102 5.0  
7 104 5.1  
8 108 5.7  
9 98 5.3  
10 103 5.2  
 Dissolved.boron....µg.l.1.  
1 50  
2 58  
3 70  
4 67  
5 69  
6 75  
7 76  
8 79  
9 60  
10 49

### Manova Test

waterqualitymodel<-manova(d.v~df\_manova$samplingYM\*df\_manova$Site)  
  
summary(waterqualitymodel,test = "Pillai")

Df Pillai approx F num Df den Df  
df\_manova$samplingYM 47 5.03 27.9 799 52989  
df\_manova$Site 21 6.78 98.5 357 52989  
df\_manova$samplingYM:df\_manova$Site 954 7.76 2.7 16218 52989  
Residuals 3117   
 Pr(>F)   
df\_manova$samplingYM <0.0000000000000002 \*\*\*  
df\_manova$Site <0.0000000000000002 \*\*\*  
df\_manova$samplingYM:df\_manova$Site <0.0000000000000002 \*\*\*  
Residuals   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(waterqualitymodel,test = "Wilk")

Df Wilks approx F num Df den Df  
df\_manova$samplingYM 47 0.0002881 41.3 799 49713  
df\_manova$Site 21 0.0000002 251.9 357 41169  
df\_manova$samplingYM:df\_manova$Site 954 0.0000076 3.3 16218 52827  
Residuals 3117   
 Pr(>F)   
df\_manova$samplingYM <0.0000000000000002 \*\*\*  
df\_manova$Site <0.0000000000000002 \*\*\*  
df\_manova$samplingYM:df\_manova$Site <0.0000000000000002 \*\*\*  
Residuals   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(waterqualitymodel,test = "Roy")

Df Roy approx F num Df den Df  
df\_manova$samplingYM 47 7.7 510 47 3117  
df\_manova$Site 21 34.5 5116 21 3117  
df\_manova$samplingYM:df\_manova$Site 954 4.7 15 954 3117  
Residuals 3117   
 Pr(>F)   
df\_manova$samplingYM <0.0000000000000002 \*\*\*  
df\_manova$Site <0.0000000000000002 \*\*\*  
df\_manova$samplingYM:df\_manova$Site <0.0000000000000002 \*\*\*  
Residuals   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(waterqualitymodel,test = "Hotelling-Lawley")

Df Hotelling-Lawley approx F num Df  
df\_manova$samplingYM 47 17.8 69 799  
df\_manova$Site 21 82.6 717 357  
df\_manova$samplingYM:df\_manova$Site 954 21.1 4 16218  
Residuals 3117   
 den Df Pr(>F)   
df\_manova$samplingYM 52685 <0.0000000000000002 \*\*\*  
df\_manova$Site 52685 <0.0000000000000002 \*\*\*  
df\_manova$samplingYM:df\_manova$Site 52685 <0.0000000000000002 \*\*\*  
Residuals   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Clustering

### Estimating the optimal number of clusters

Creating a new data frame for clustering

df\_forClustering<-aggregate(d.v~df\_manova$Site,data = df\_manova, function(x)round(mean(x),2))[,-1]  
rownames(df\_forClustering)<-aggregate(d.v~df\_manova$Site,data = df\_manova, function(x)round(mean(x),2))[,1]  
  
head(df\_forClustering)

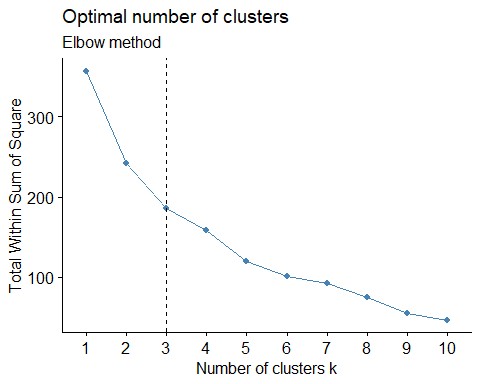
Water.temperature...C. pH  
Jubilee River at Pocock's Bridge 13 8.0  
River Cherwell at Hampton Poyle 12 7.9  
River Cole at Lynt Bridge 12 7.9  
River Coln at Whelford 12 8.0  
River Enborne at Brimpton 11 7.8  
River Evenlode at Cassington Mill 11 7.9  
 Alkalinity..µ.equ.l.1.  
Jubilee River at Pocock's Bridge 4088  
River Cherwell at Hampton Poyle 4134  
River Cole at Lynt Bridge 4335  
River Coln at Whelford 4247  
River Enborne at Brimpton 2819  
River Evenlode at Cassington Mill 4028  
 Suspended.solids......mg.l.1.  
Jubilee River at Pocock's Bridge 8.4  
River Cherwell at Hampton Poyle 13.3  
River Cole at Lynt Bridge 15.2  
River Coln at Whelford 5.4  
River Enborne at Brimpton 9.5  
River Evenlode at Cassington Mill 15.7  
 phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
Jubilee River at Pocock's Bridge 192 0.07  
River Cherwell at Hampton Poyle 193 0.04  
River Cole at Lynt Bridge 307 0.05  
River Coln at Whelford 84 0.04  
River Enborne at Brimpton 183 0.08  
River Evenlode at Cassington Mill 252 0.04  
 Dissolved.silicon..mg.l.1.Si.  
Jubilee River at Pocock's Bridge 5.2  
River Cherwell at Hampton Poyle 3.3  
River Cole at Lynt Bridge 6.4  
River Coln at Whelford 2.6  
River Enborne at Brimpton 6.9  
River Evenlode at Cassington Mill 2.7  
 Chlorophyll.a..µg.l.1.  
Jubilee River at Pocock's Bridge 18.7  
River Cherwell at Hampton Poyle 14.1  
River Cole at Lynt Bridge 5.7  
River Coln at Whelford 3.0  
River Enborne at Brimpton 2.5  
River Evenlode at Cassington Mill 12.4  
 Dissolved.fluoride..mg.l.1.  
Jubilee River at Pocock's Bridge 0.15  
River Cherwell at Hampton Poyle 0.20  
River Cole at Lynt Bridge 0.19  
River Coln at Whelford 0.13  
River Enborne at Brimpton 0.12  
River Evenlode at Cassington Mill 0.12  
 Dissolved.chloride..mg.l.1.  
Jubilee River at Pocock's Bridge 44  
River Cherwell at Hampton Poyle 54  
River Cole at Lynt Bridge 46  
River Coln at Whelford 17  
River Enborne at Brimpton 34  
River Evenlode at Cassington Mill 26  
 Dissolved.nitrate......mg.l.1.NO3.  
Jubilee River at Pocock's Bridge 26  
River Cherwell at Hampton Poyle 25  
River Cole at Lynt Bridge 18  
River Coln at Whelford 26  
River Enborne at Brimpton 17  
River Evenlode at Cassington Mill 25  
 Dissolved.sulphate.....mg.l.1.SO4.  
Jubilee River at Pocock's Bridge 47  
River Cherwell at Hampton Poyle 65  
River Cole at Lynt Bridge 53  
River Coln at Whelford 34  
River Enborne at Brimpton 26  
River Evenlode at Cassington Mill 46  
 Dissolved.sodium..mg.l.1.  
Jubilee River at Pocock's Bridge 27.4  
River Cherwell at Hampton Poyle 35.6  
River Cole at Lynt Bridge 27.4  
River Coln at Whelford 8.8  
River Enborne at Brimpton 17.8  
River Evenlode at Cassington Mill 16.2  
 Dissolved.potassium..mg.l.1.  
Jubilee River at Pocock's Bridge 5.4  
River Cherwell at Hampton Poyle 6.2  
River Cole at Lynt Bridge 5.3  
River Coln at Whelford 1.7  
River Enborne at Brimpton 3.6  
River Evenlode at Cassington Mill 3.5  
 Dissolved.calcium...........mg.l.1.  
Jubilee River at Pocock's Bridge 102  
River Cherwell at Hampton Poyle 104  
River Cole at Lynt Bridge 110  
River Coln at Whelford 101  
River Enborne at Brimpton 68  
River Evenlode at Cassington Mill 102  
 Dissolved.magnesium..mg.l.1.  
Jubilee River at Pocock's Bridge 4.4  
River Cherwell at Hampton Poyle 7.6  
River Cole at Lynt Bridge 4.4  
River Coln at Whelford 5.7  
River Enborne at Brimpton 4.4  
River Evenlode at Cassington Mill 4.2  
 Dissolved.boron....µg.l.1.  
Jubilee River at Pocock's Bridge 54  
River Cherwell at Hampton Poyle 73  
River Cole at Lynt Bridge 55  
River Coln at Whelford 20  
River Enborne at Brimpton 26  
River Evenlode at Cassington Mill 51

Scaling data frame (standardizing the data to make variables comparable)

df\_scaled<-scale(df\_forClustering)

Determining the optimal number of clusters for k-means clustering by ***Elbow method***

fviz\_nbclust(df\_scaled, kmeans, method = "wss") +  
 geom\_vline(xintercept = 3, linetype = 2)+  
 labs(subtitle = "Elbow method")



Therefore we select k=3 as the number of clusters

### Clustering using K-means

set.seed(123)  
  
km.res <- kmeans(df\_scaled, 3,nstart = 25)

Details

print(km.res)

K-means clustering with 3 clusters of sizes 8, 11, 3  
  
Cluster means:  
 Water.temperature...C. pH Alkalinity..µ.equ.l.1.  
1 -0.7659 0.29 0.136  
2 0.5550 0.28 0.084  
3 0.0073 -1.81 -0.670  
 Suspended.solids......mg.l.1. phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
1 -0.328 -0.65 -0.43  
2 0.235 -0.11 -0.22  
3 0.012 2.15 1.95  
 Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 -0.046 -0.60  
2 -0.056 0.48  
3 0.329 -0.17  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 -0.94 -0.91  
2 0.36 0.22  
3 1.20 1.60  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 -0.33 -0.97  
2 -0.20 0.24  
3 1.62 1.69  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 -0.89 -0.92  
2 0.18 0.15  
3 1.74 1.92  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 -0.22 -0.629  
2 0.12 0.089  
3 0.17 1.349  
 Dissolved.boron....µg.l.1.  
1 -1.02  
2 0.28  
3 1.71  
  
Clustering vector:  
 Jubilee River at Pocock's Bridge River Cherwell at Hampton Poyle   
 2 2   
 River Cole at Lynt Bridge River Coln at Whelford   
 2 1   
 River Enborne at Brimpton River Evenlode at Cassington Mill   
 1 1   
 River Kennet at Woolhampton River Leach at Mill Lane,Lechlade   
 1 1   
 River Lodden at Charvil River Ock at Abingdon   
 2 2   
 River Pang at Tidmarsh River Ray at Islip   
 1 3   
 River Thame at Wheatley River Thames at Hannington Wick   
 3 2   
 River Thames at Newbridge River Thames at Runnymede   
 2 2   
 River Thames at Sonning River Thames at Swinford   
 2 2   
 River Thames at Wallingford River Windrush at Newbridge   
 2 1   
 River Wye at Bourne End The Cut at Paley Street   
 1 3   
  
Within cluster sum of squares by cluster:  
[1] 69 66 48  
 (between\_SS / total\_SS = 48.7 %)  
  
Available components:  
  
[1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
[6] "betweenss" "size" "iter" "ifault"

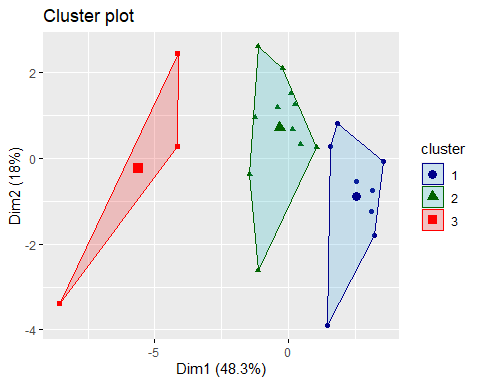
km.res$centers

Water.temperature...C. pH Alkalinity..µ.equ.l.1.  
1 -0.7659 0.29 0.136  
2 0.5550 0.28 0.084  
3 0.0073 -1.81 -0.670  
 Suspended.solids......mg.l.1. phosphorus..µg.l.1.P. Ammonium..mg.l.1.NH4.  
1 -0.328 -0.65 -0.43  
2 0.235 -0.11 -0.22  
3 0.012 2.15 1.95  
 Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1 -0.046 -0.60  
2 -0.056 0.48  
3 0.329 -0.17  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1 -0.94 -0.91  
2 0.36 0.22  
3 1.20 1.60  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1 -0.33 -0.97  
2 -0.20 0.24  
3 1.62 1.69  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1 -0.89 -0.92  
2 0.18 0.15  
3 1.74 1.92  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1 -0.22 -0.629  
2 0.12 0.089  
3 0.17 1.349  
 Dissolved.boron....µg.l.1.  
1 -1.02  
2 0.28  
3 1.71

Visualizing

fviz\_cluster(km.res, data = df\_scaled,  
 palette = c("#2E9FDF", "#00AFBB", "#ed0000"),   
 geom = "point",  
 ellipse.type = "convex"  
 )+scale\_colour\_manual(values = c("darkblue", "darkgreen", "red"))

Scale for 'colour' is already present. Adding another scale for 'colour',  
which will replace the existing scale.



pc<-princomp(df\_scaled)  
plot3d(pc$scores[,1:3],col = km.res$cluster,size = 20)

### cluster summarising

Attaching clusters to each observation accordingly

df\_forClustering<-cbind(df\_forClustering,cluster=km.res$cluster)

setDT(df\_forClustering,keep.rownames = "Site")  
head(df\_forClustering)

Site Water.temperature...C. pH  
1: Jubilee River at Pocock's Bridge 13 8.0  
2: River Cherwell at Hampton Poyle 12 7.9  
3: River Cole at Lynt Bridge 12 7.9  
4: River Coln at Whelford 12 8.0  
5: River Enborne at Brimpton 11 7.8  
6: River Evenlode at Cassington Mill 11 7.9  
 Alkalinity..µ.equ.l.1. Suspended.solids......mg.l.1. phosphorus..µg.l.1.P.  
1: 4088 8.4 192  
2: 4134 13.3 193  
3: 4335 15.2 307  
4: 4247 5.4 84  
5: 2819 9.5 183  
6: 4028 15.7 252  
 Ammonium..mg.l.1.NH4. Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1: 0.07 5.2 18.7  
2: 0.04 3.3 14.1  
3: 0.05 6.4 5.7  
4: 0.04 2.6 3.0  
5: 0.08 6.9 2.5  
6: 0.04 2.7 12.4  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1: 0.15 44  
2: 0.20 54  
3: 0.19 46  
4: 0.13 17  
5: 0.12 34  
6: 0.12 26  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1: 26 47  
2: 25 65  
3: 18 53  
4: 26 34  
5: 17 26  
6: 25 46  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1: 27.4 5.4  
2: 35.6 6.2  
3: 27.4 5.3  
4: 8.8 1.7  
5: 17.8 3.6  
6: 16.2 3.5  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1: 102 4.4  
2: 104 7.6  
3: 110 4.4  
4: 101 5.7  
5: 68 4.4  
6: 102 4.2  
 Dissolved.boron....µg.l.1. cluster  
1: 54 2  
2: 73 2  
3: 55 2  
4: 20 1  
5: 26 1  
6: 51 1

df\_forClustering %>%  
 select(everything()) %>%  
 filter(cluster==1)

Site Water.temperature...C. pH  
1: River Coln at Whelford 12 8.0  
2: River Enborne at Brimpton 11 7.8  
3: River Evenlode at Cassington Mill 11 7.9  
4: River Kennet at Woolhampton 11 8.0  
5: River Leach at Mill Lane,Lechlade 11 7.9  
6: River Pang at Tidmarsh 11 7.9  
7: River Windrush at Newbridge 11 8.1  
8: River Wye at Bourne End 12 8.1  
 Alkalinity..µ.equ.l.1. Suspended.solids......mg.l.1. phosphorus..µg.l.1.P.  
1: 4247 5.4 84  
2: 2819 9.5 183  
3: 4028 15.7 252  
4: 4500 9.3 78  
5: 4367 3.0 34  
6: 4495 8.3 68  
7: 3880 14.0 132  
8: 4593 13.3 290  
 Ammonium..mg.l.1.NH4. Dissolved.silicon..mg.l.1.Si. Chlorophyll.a..µg.l.1.  
1: 0.04 2.6 3.0  
2: 0.08 6.9 2.5  
3: 0.04 2.7 12.4  
4: 0.05 6.8 8.2  
5: 0.06 2.4 1.9  
6: 0.04 7.0 2.7  
7: 0.04 2.5 4.0  
8: 0.11 6.7 3.7  
 Dissolved.fluoride..mg.l.1. Dissolved.chloride..mg.l.1.  
1: 0.13 17  
2: 0.12 34  
3: 0.12 26  
4: 0.12 24  
5: 0.10 16  
6: 0.14 25  
7: 0.11 23  
8: 0.11 42  
 Dissolved.nitrate......mg.l.1.NO3. Dissolved.sulphate.....mg.l.1.SO4.  
1: 26 34  
2: 17 26  
3: 25 46  
4: 24 20  
5: 31 35  
6: 28 19  
7: 28 42  
8: 27 20  
 Dissolved.sodium..mg.l.1. Dissolved.potassium..mg.l.1.  
1: 8.8 1.7  
2: 17.8 3.6  
3: 16.2 3.5  
4: 12.4 2.4  
5: 8.3 1.5  
6: 12.1 2.9  
7: 13.3 2.7  
8: 26.3 4.2  
 Dissolved.calcium...........mg.l.1. Dissolved.magnesium..mg.l.1.  
1: 101 5.7  
2: 68 4.4  
3: 102 4.2  
4: 107 2.2  
5: 109 5.1  
6: 107 3.2  
7: 98 4.5  
8: 107 1.9  
 Dissolved.boron....µg.l.1. cluster  
1: 20 1  
2: 26 1  
3: 51 1  
4: 22 1  
5: 25 1  
6: 21 1  
7: 33 1  
8: 35 1

dfcl<-df\_forClustering %>%  
 select(-Site) %>%  
 group\_by(cluster) %>%  
 summarise\_all("mean")  
  
write.csv(dfcl,"F:\\3-1\\ST305\\Assignment\\meanclusters.csv", row.names = FALSE)

df\_forClustering %>%  
 select(Site,cluster) %>%  
 group\_by(cluster)

# A tibble: 22 x 2  
# Groups: cluster [3]  
 Site cluster  
 <chr> <int>  
 1 Jubilee River at Pocock's Bridge 2  
 2 River Cherwell at Hampton Poyle 2  
 3 River Cole at Lynt Bridge 2  
 4 River Coln at Whelford 1  
 5 River Enborne at Brimpton 1  
 6 River Evenlode at Cassington Mill 1  
 7 River Kennet at Woolhampton 1  
 8 River Leach at Mill Lane,Lechlade 1  
 9 River Lodden at Charvil 2  
10 River Ock at Abingdon 2  
# ... with 12 more rows