eNote 002

ERROR CHECKING OF THE MINERALOGICAL DATA-FRAME: EAST COAST DELTAS OF PENINSULAR INDIA

BY

George F. Hart

This data-frame is used for the study of the mineralogical characteristics of the deltas of peninsular India by Hart, Ferrell, SetaRama Swamy, Banu Murthy and ? Gandhi.

CONCLUSIONS

All of the variables depart from normality and need to be transformed prior to further analysis. Aberrant values [outliers] exist in the plagioclase [#24], calcite [334], dolomite [185], CaCO3 [#142], OM [#53], and organic carbon [#236]. These need to be investigated prior to further analysis but the action is that they will be removed from the data-frame and coded as NA for the next phase of the preliminary study.

Examining the clustering of the samples along the variable vertical axis indicates that quartz and kfeldspar are present in all samples and each variable has a similar spread when the five deltas are compared. This suggests that either quartz or kfeldspar would be a good candidate as the divisor variable for **Aitchison's** log-ratio transform. The variables plagioclase, calcite, chlorite, illite, smectite and kaolinite show some separation amongst deltas and will probably be useful are discriminating variables when classification procedures are applied to the data-frame.

TABLES

Table 1: the number of samples examined from each delta				
Cauvery	**Godavari**	**Krishna**	**Mahanadi**	**Penner**
84	108	106	47	77

Table 2: the number of samples examined from each depositional environment													
Bay	Barrier island	Channel	Chenier	Distributary mouth bar	Mud flat	Foreshore	Lagoon	Levee	Mangrove	Shoal	Spit	Tidal creek	Unknown
7	21	52	2	2	15	12	25	5	29	2	2	21	227

Table 3: the number of samples examined from each delta by depositional environment					
Depositional environment	**Cauvery**	**Godavari**	**Krishna**	**Mahanadi**	**Penner**
Bay	0	0	0	7	0
Barrier island	7	5	6	1	2
Channel	6	16	22	6	2
Chenier	0	2	0	0	0
Distributary mouth bar	0	0	0	0	2
Mudflat	1	4	6	2	2
Foreshore	2	4	5	0	1
Lagoon	3	6	14	0	2
Levee	0	3	0	2	
Mangrove swamp	4	5	10	4	6
Shoal	0	0	0	0	2
Spit	0	0	0	0	2
Tidal creek	4	5	5	5	2
Unknown	57	58	38	20	54

Table 3: recorded variable and its abbreviation	
mineral	abbreviation
Quartz [XRD]	qtz
potassium feldspar[XRD]	kf
plagioclase[XRD]	plag
clay [XRD]	clay
amphibole[XRD]	amp
clinoptilite[XRD]	clin
gypsum[XRD]	gyp
calcite[XRD]	cal
dolomite[XRD]	dol
kaolinite[XRD]	kao
illite[XRD]	ill
smectite[XRD]	smec
chlorite[XRD]	chlor
pyrite [XRD]	py
CaCO3 [wet analysis]	ca
total organic matter [wet analysis]	tom
organic carbon [wet analysis]	oc

Error checking of original data using the Index plot

We use an Index plot to show outliers and aberrant values. The index plot graphs each sample according to its row position in the data-frame against the variable. The India deltas data-frame had each delta added in sequence. Thus if we look at the quartz index plot below we see the colors representing each delta in a sequence. This provides an initial view of the distribution of a variable by delta. The important value is the vertical axis which shows the distribution of the quartz: in this case there are no abnormalities as seen by the similar spread for each delta [see CaCO3 later for an abnormality].

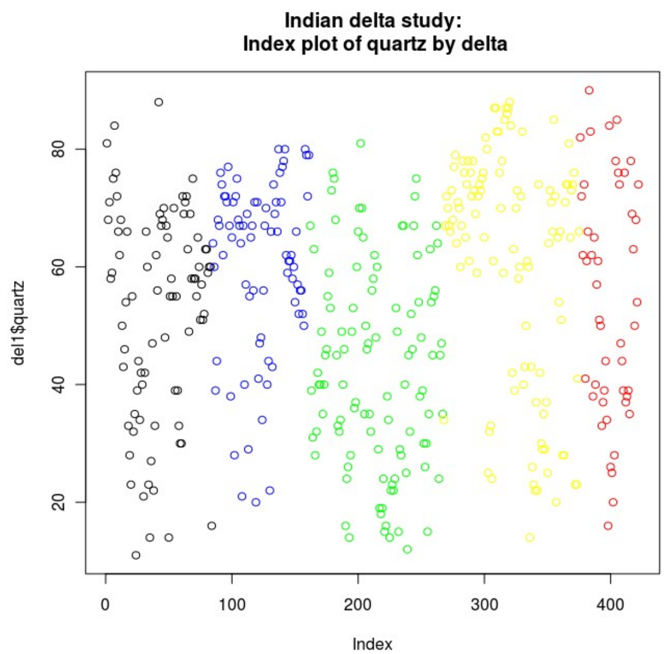
The R code for the index plot is:

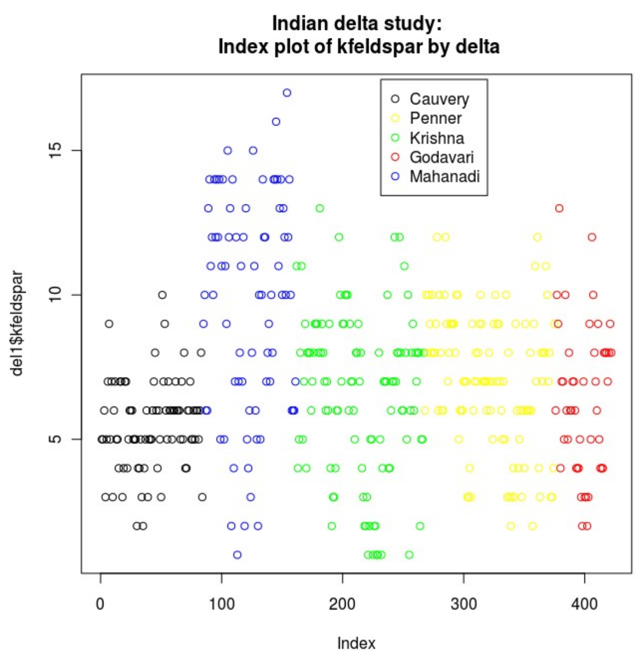
```
> my.colors<-c("black","yellow","green","red","blue") # set up a color array.

> plot(del1$pcchlorite,col=my.colors[del1$delta]) # plot the chlorite data.

> title(main="Indian delta study:\n Index plot of chlorite by delta") # add the title.

>legend(locator(1),c("Cauvery","Penner","Krishna","Godavari","Mahanadi"),pch=c(1,1),col=my.color
```

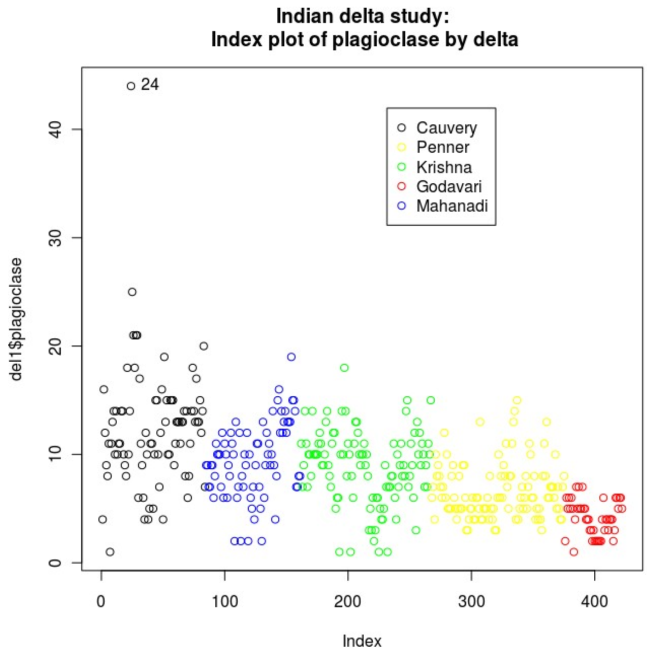
s) # add and position the legend using the locator.

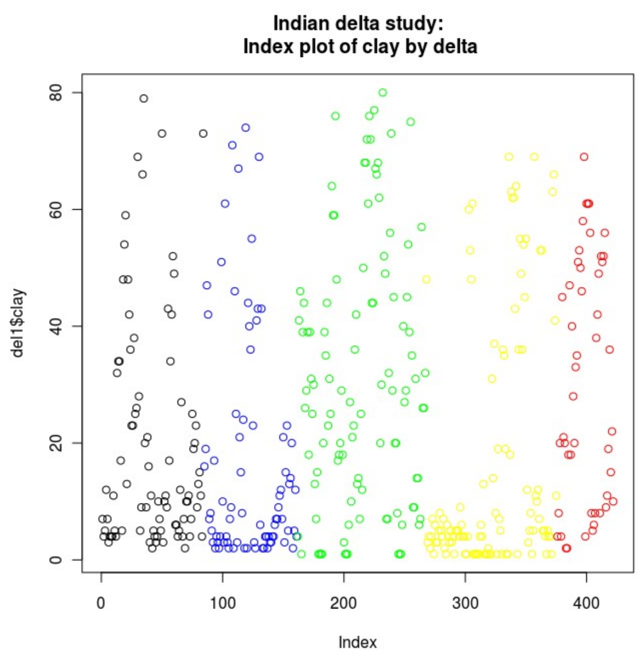
Individual points can be identified using:

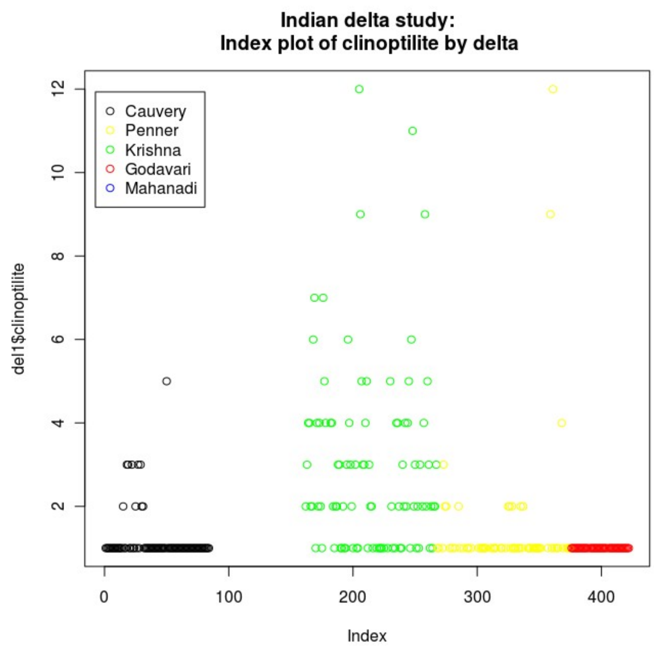
>identify(plagioclase,y=NULL) # place the pointer over the sample [see plagioclase below where point 24 is an outlier and possibly an error].



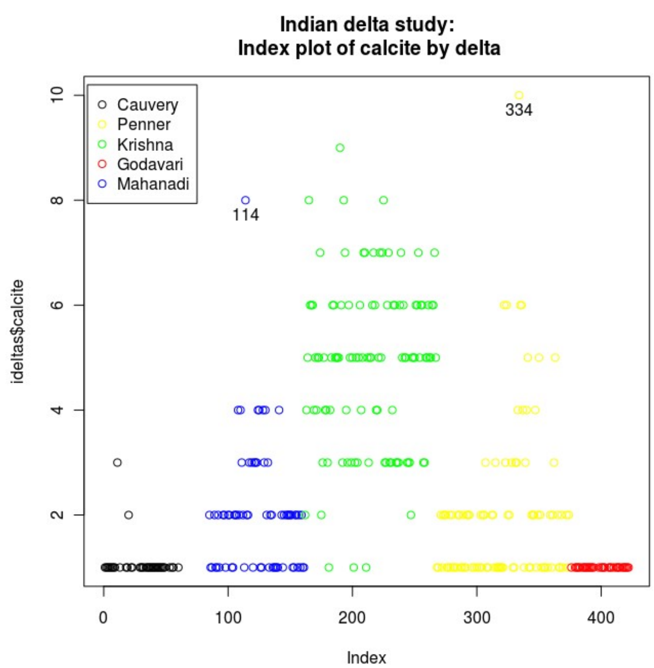



Indian delta study:
Index plot of plagioclase by delta
24
Cauvery
Penner
Krishna
Godavari
Mahanadi
del1$plagioclase
Index
0
10
20
30
40
0
100
200
300
400


Indian delta study:
Index plot of clay by delta
del1$clay
Index
0
20
40
60
80
0
100
200
300
400

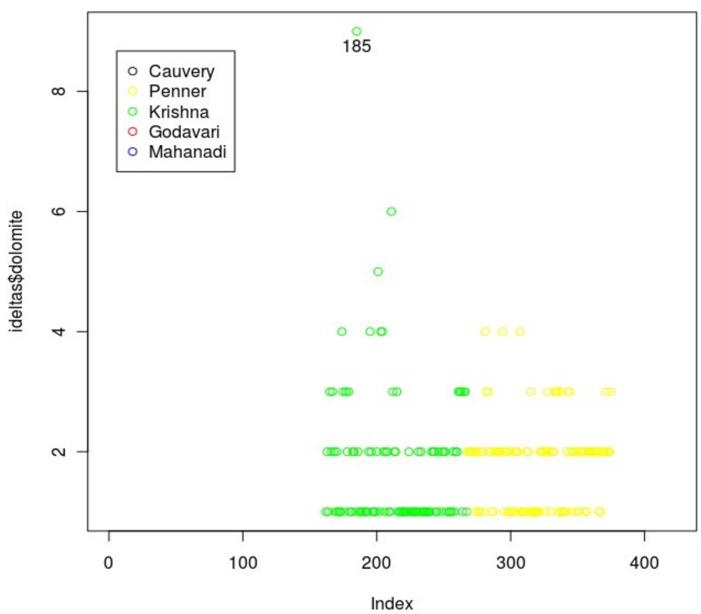


Indian delta study:
Index plot of clinoptilite by delta
Cauvery
Penner
Krishna
Godavari
Mahanadi
del1$clinoptlite
Index

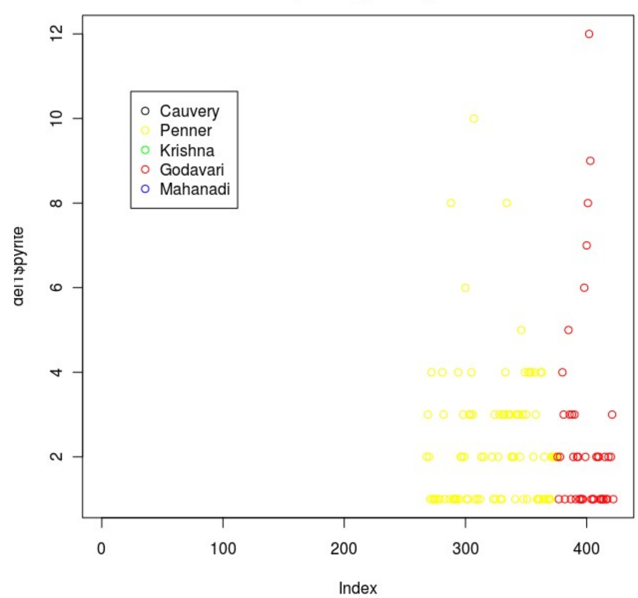


Indian delta study:
Index plot of calcite by delta
Cauvery
Penner
Krishna
Godavari
Mahanadi
334
114
ideltas$calcite
Index

Indian delta study:
Index plot of dolomite by delta



Indian delta study:
Index plot of pyrite by delta



ANALYSES OF COMMON MINERALS

Table 4: SUMMARY STATISTICS FOR COMMON MINERALS							
Mineral	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	NA's
qtz	11	39	58	54	70	90	0
kf	1	5	7	7.03	9	17	0
plag	1	5	8	8.6	11	44	0
clay	1	4	13	22.7	39	80	0
amph	1	1	2	3.5	3.8	47	40
clin	1	1	1	1.8	2	12	137
gyp	1	1	1	1	1	2	283
cal	1	1	2	2.7	4	10	80
dol	1	1	2	1.8	2	9	211
py	1	1	2	2.5	3	12	291

The normality tests combine graphical tests (Q-Q plot, histograms plus density line overlay, and the Shapiro-Wilks test statistic which is based on H_0 that the data are normally distributed . Because the the sample size [n] is larger than 50 the D'Agnostino's test also is calculated. The D'Agnostino statistic measures the linearity of the points on the normal probability plot. "*If the normal probability plot is approximately linear (the data follow a normal curve), the correlation coefficient will be relatively high. If the normal probability plot contains significant curves (the data do not follow a normal curve), the correlation coefficient will be relatively low*". [**EPA, 2008**]

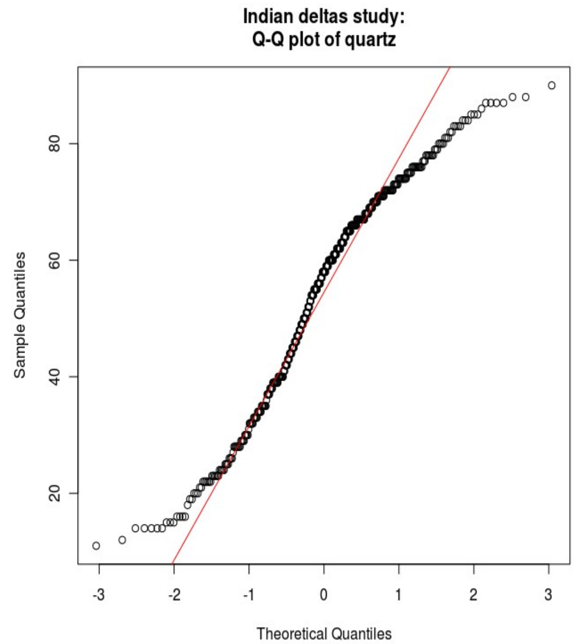
The assumption of normality is important because it is a requirement of many statistical tests. The normal distribution is a reasonable model for many variables in the natural sciences. The central limit theorem shows that as the sample size gets large, many of the sample summary statistics, such as the sample mean, behave as if they are from a normally distributed variable. Thus it is assumed that parametric tests or statistical models have associated errors that follow a normal distribution. **EPA Report EM 1110-1-4014, 31-Jan-08 [p:F-1]** points out:

"*statistical tests for normality do not actually demonstrate normality but the lack of normality. They rely on the probability a given data set is normal (e.g., statistical software typically reports a "p value" for the hypothesis that the population distribution is normal). If the probability is low (e.g. p < 0.01), one "rejects the assumption of normality," that is, one concludes, based upon weight of evidence, that the data set is not normal. However, if the assumption of normality is not rejected, then, strictly speaking, the statistical test is inconclusive; the data may or may not be normal. This constitutes an additional reason to visually examine the data set for normality and to decide whether to proceed with* *a statistical test that requires normality. In practice, if the assumption of normality is not rejected and graphical plots suggest normality, the statistical tests that rely upon normality are typically used*" ,

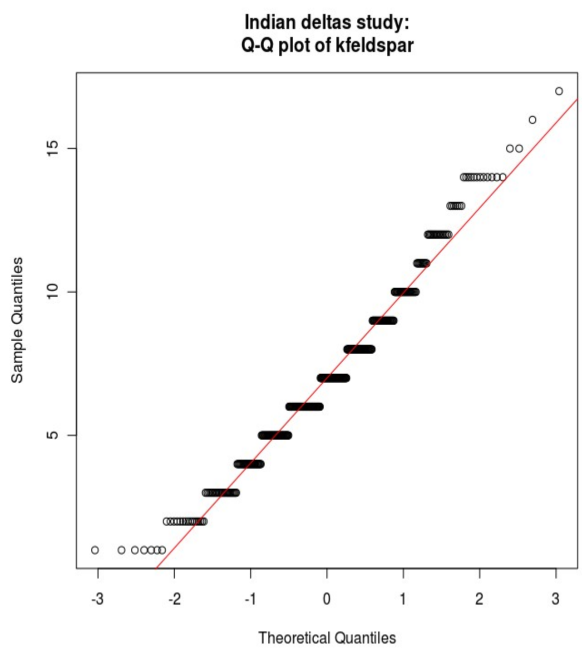
and

"The assumption of normality should not be rejected on the basis of a statistical test alone. In particular, when a large number of data are available, statistical tests for normality can be sensitive to very small (i.e., negligible) deviations in normality. Therefore, if a very large number of data are available, a statistical test may reject the assumption of normality when the data set, as shown using graphical methods, is essentially normal and the deviation from normality too small to be of practical significance."

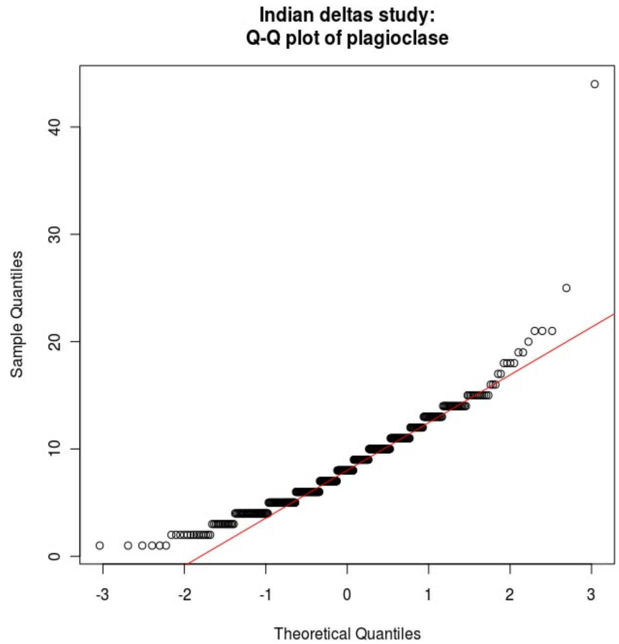
TESTS FOR NORMALITY:Q=Q plots



D'Agostino skewness test [two tailed],data: quartz skew = -0.343, z = -1.879, p-value = 0.06032

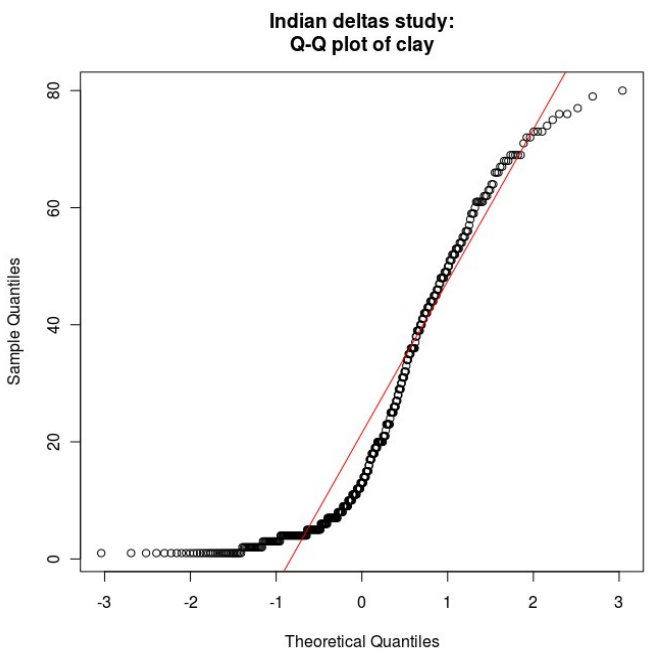


D'Agostino skewness test data: kfeldspar, skew = 0.467, z = 2.503, p-value = 0.01230



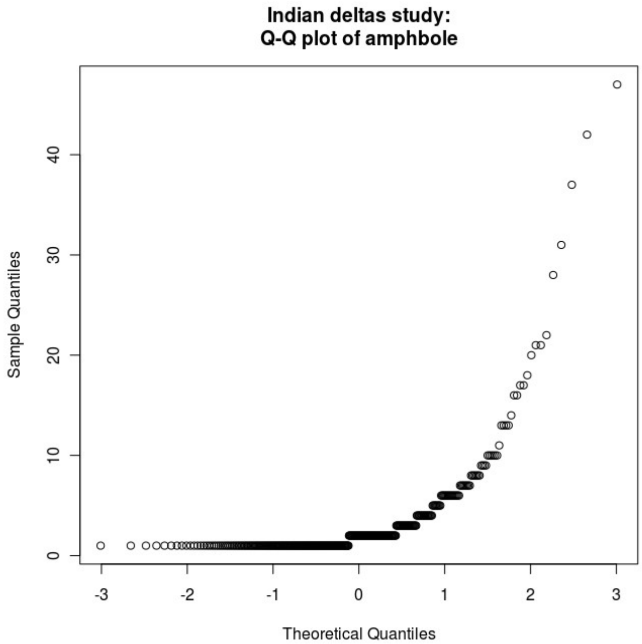
D'Agostino skewness test data: plagioclase, skew = 1.64, z = 6.72, p-value = 1.825e-11

Variable has skewness.



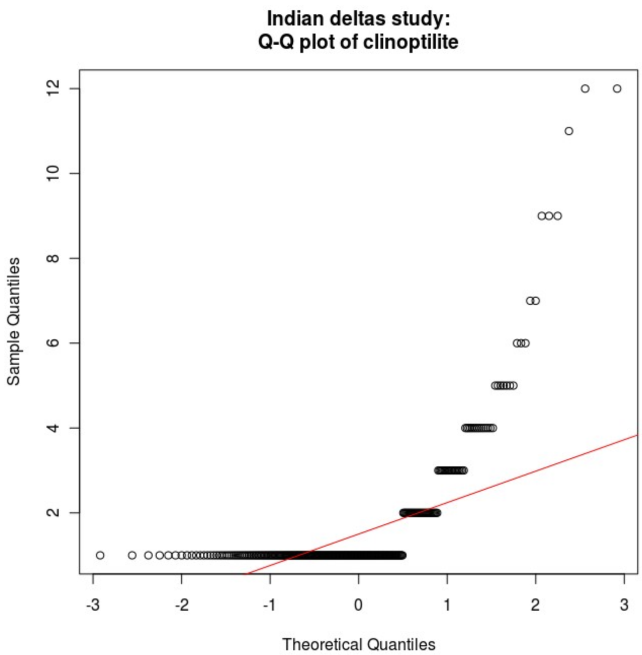
D'Agostino skewness test data: clay, skew = 0.884, z = 4.331, p-value = 1.486e-05

Variable has skewness.



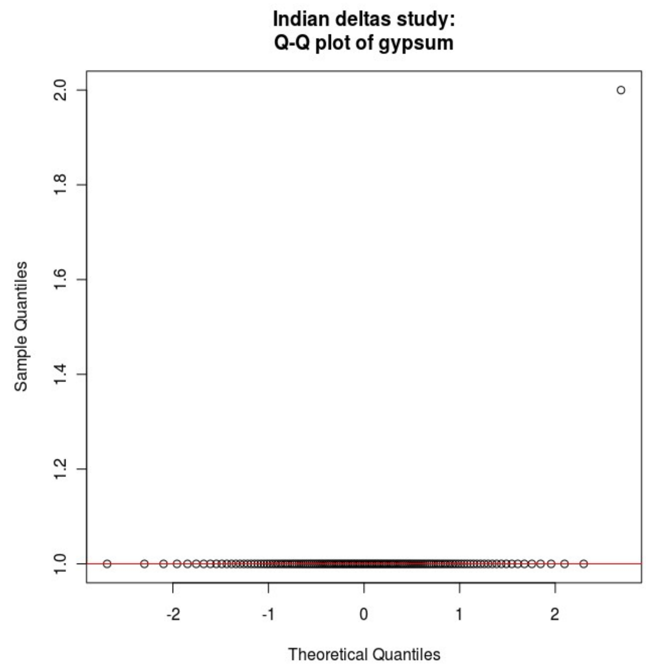
D'Agostino skewness test data: amphibole, skew = 4.56, z = 10.78, p-value < 2.2e-16

Variable has skewness.

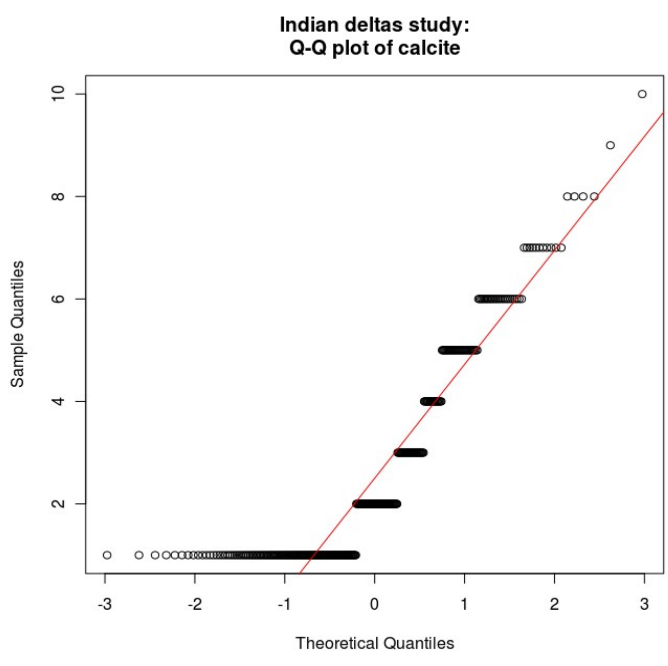


D'Agostino skewness test data: clinoptilite, skew = 3.25, z = 8.11, p-value = 5.165e-16

Variable has skewness.

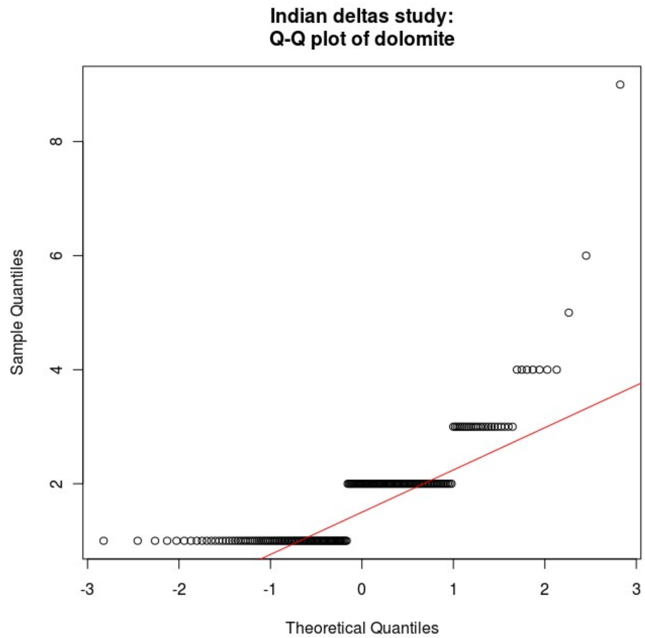


Variable should be discarded: [all values are 1.0]



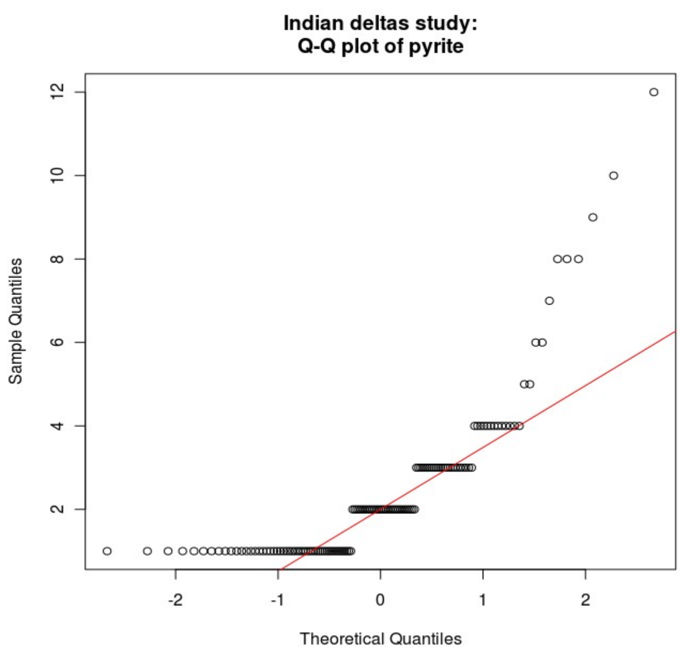
D'Agostino skewness test data: calcite, skew = 1.02, z = 4.38, p-value = 1.162e-05

Variable has skewness.



D'Agostino skewness test data: dolomite, skew = 2.58, z = 6.32, p-value = 2.667e-10

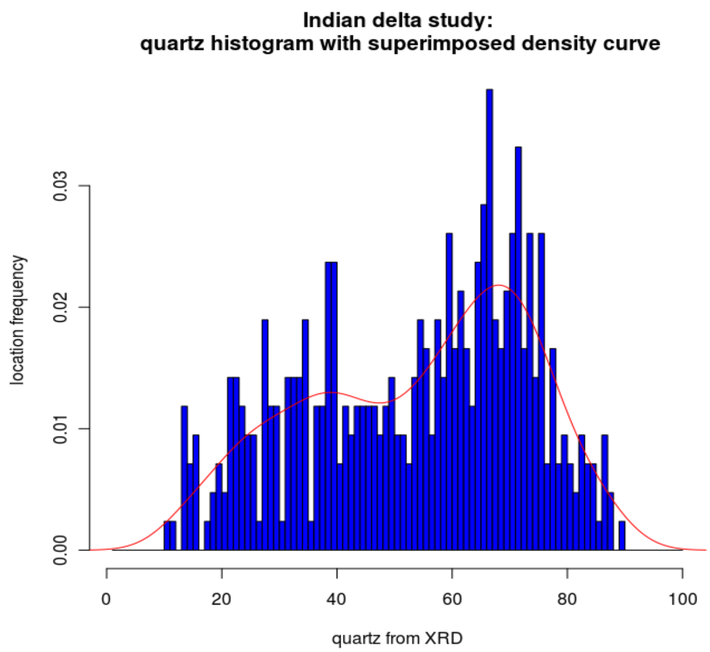
Variable has skewness.



D'Agostino skewness test data: pyrite, skew = 2.27, z = 4.79, p-value = 1.683e-06

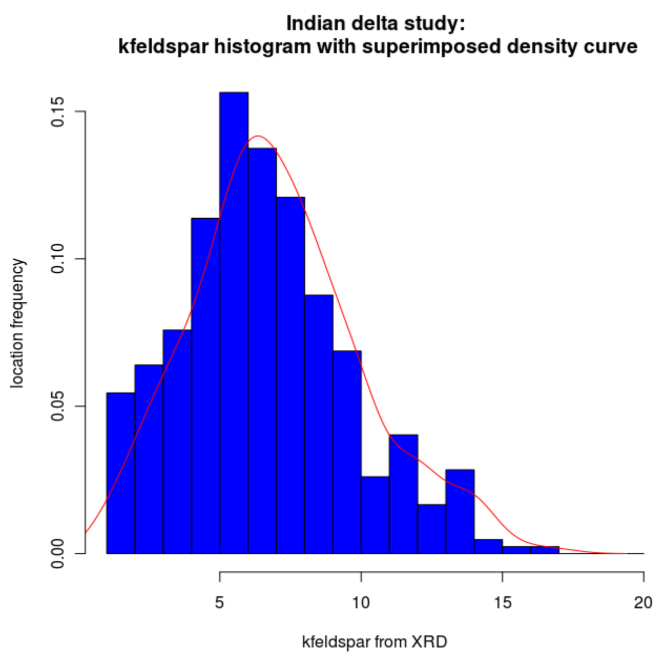
Variable has skewness.

Histograms with overlain density plot (n=1,000)



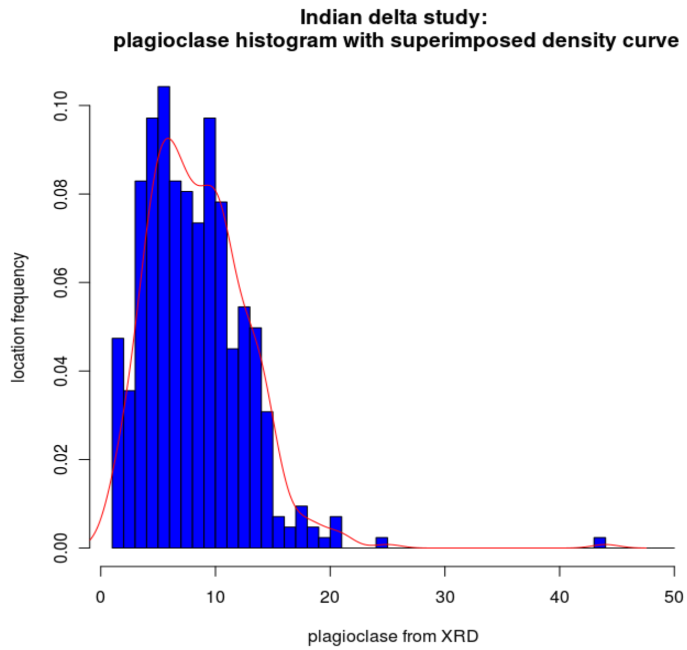
Shapiro-Wilk normality test , data: quartz : W = 0.959, p-value = 1.762e-09

Variable is not Gaussian.



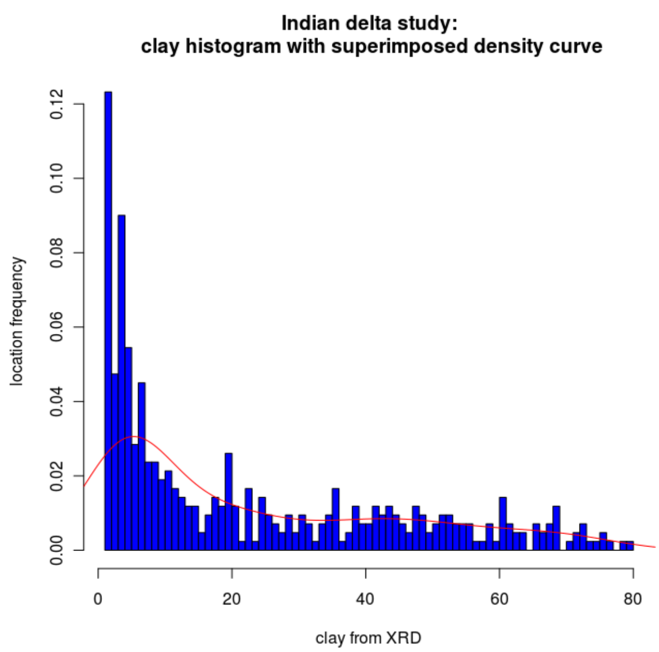
Shapiro-Wilk normality test , data: kfeldspar W = 0.973, p-value = 5.149e-07

Variable is not Gaussian.



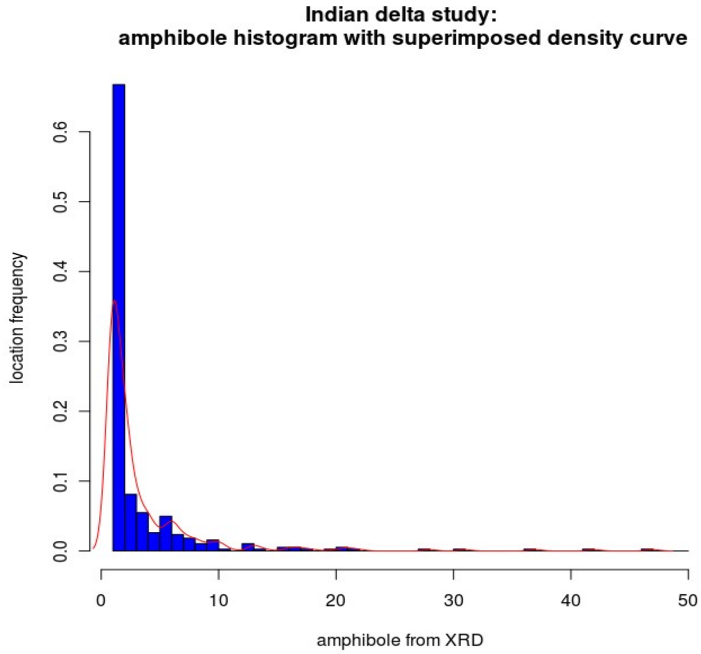
Shapiro-Wilk normality test , data: plagioclase W = 0.91, p-value = 3.208e-15

Variable is not Gaussian



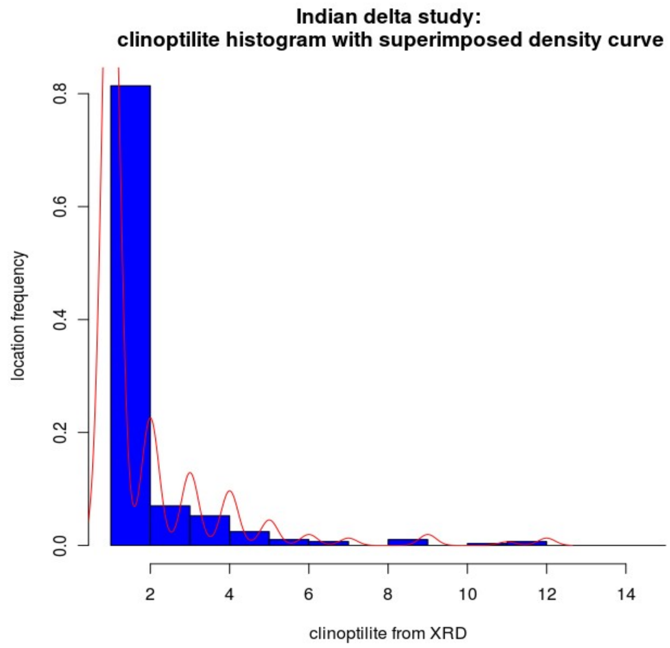
Shapiro-Wilk normality test , data: clay W = 0.854, p-value < 2.2e-16

Variable is not Gaussian



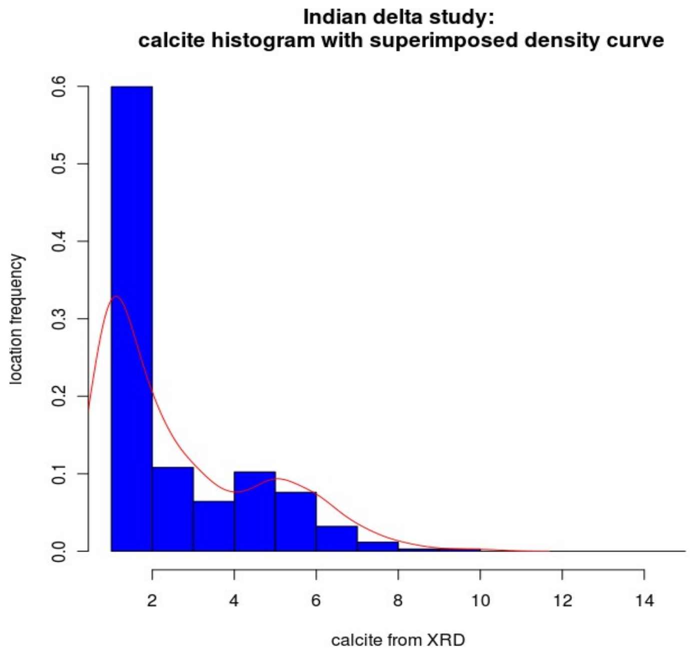
Shapiro-Wilk normality test , data: amphibole W = 0.491, p-value < 2.2e-16

Variable is not Gaussian



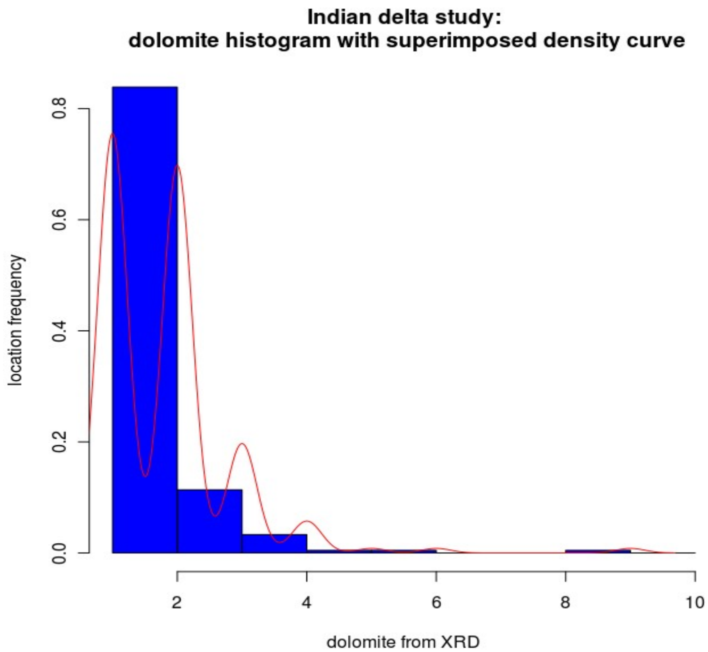
Shapiro-Wilk normality test , data: clinoptilite W = 0.532, p-value < 2.2e-16

Variable is not Gaussian



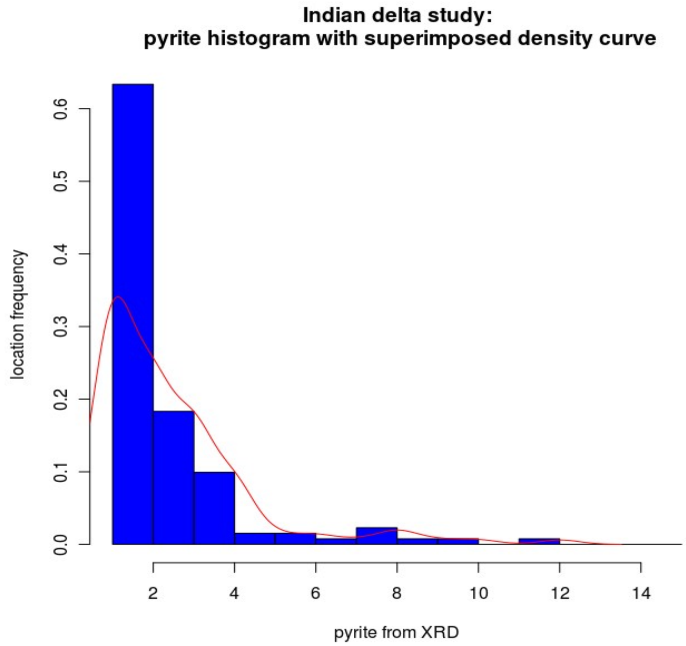
Shapiro-Wilk normality test, data: calcite W = 0.815, p-value < 2.2e-16

Variable is not Gaussian



Shapiro-Wilk normality test , data: dolomite W = 0.71, p-value < 2.2e-16

Variable is not Gaussian



Shapiro-Wilk normality test , data: pyrite W = 0.724, p-value = 2.204e-14

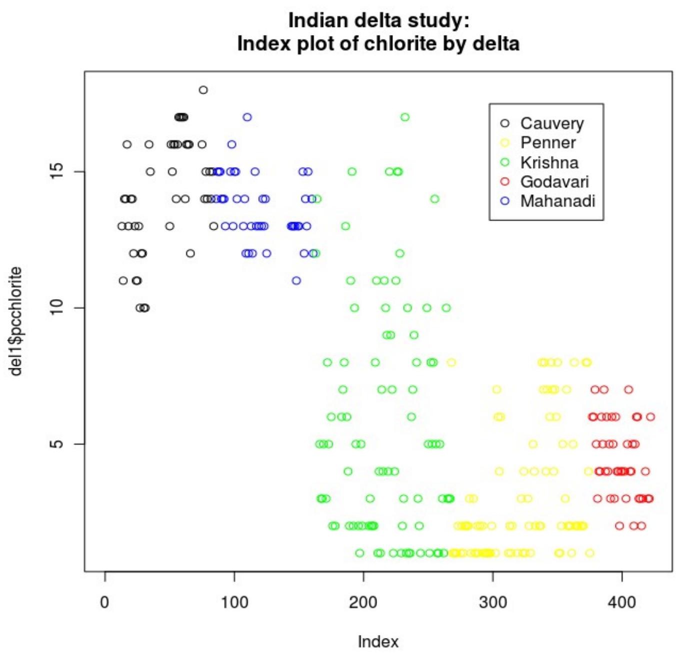
Variable is not Gaussian

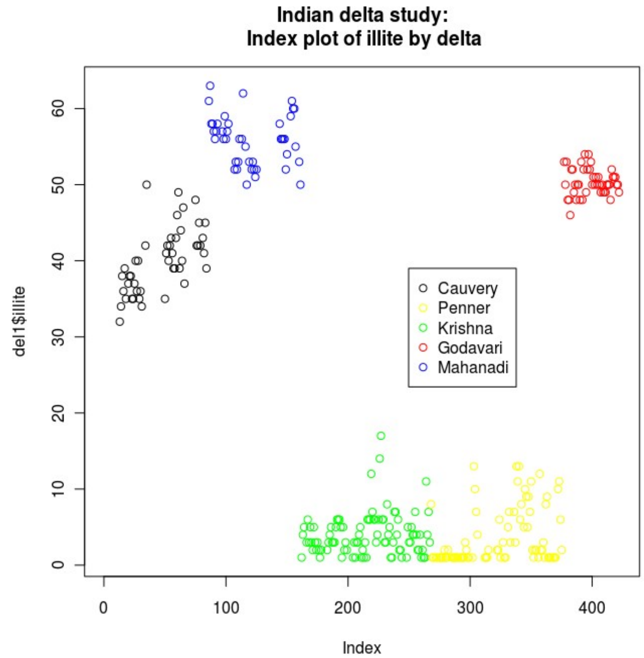
ANALYSES OF CLAY MINERALS

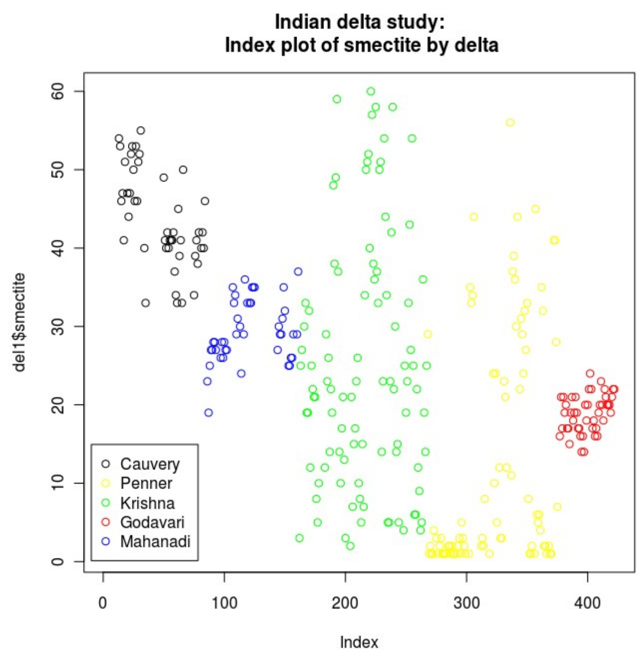
Table 5: SUMMARY STATISTICS FOR CLAY MINERALS							
Mineral	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	NA"s
kaol	1	1	2	6.2	4	30	169
ill	1	2.5	8	23.3	49	63	111
smec	1	12	23	24.2	35	60	108
chlor	1	2	6	7.2	13	18	120

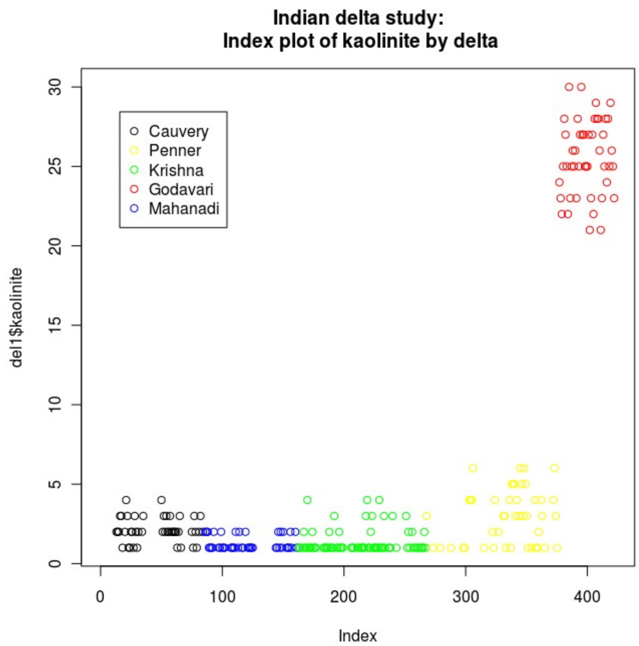
Error checking of original data using the Index plot

Chlorite does not show any abnormalities. The initial view of the distribution of the variable by delta indicates the Cauvery and Mahanadi have similar levels as do the Penner and Godavari. The Krishna data spans the two groups



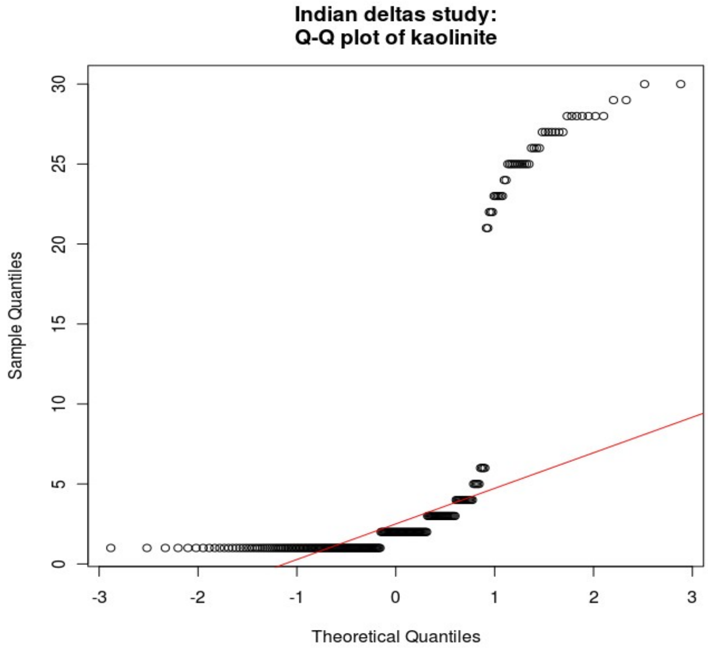

Indian delta study:
Index plot of illite by delta
del1$illite
Index
Cauvery
Penner
Krishna
Godavari
Mahanadi


Indian delta study:
Index plot of smectite by delta
del1$smectite
Index
Cauvery
Penner
Krishna
Godavari
Mahanadi



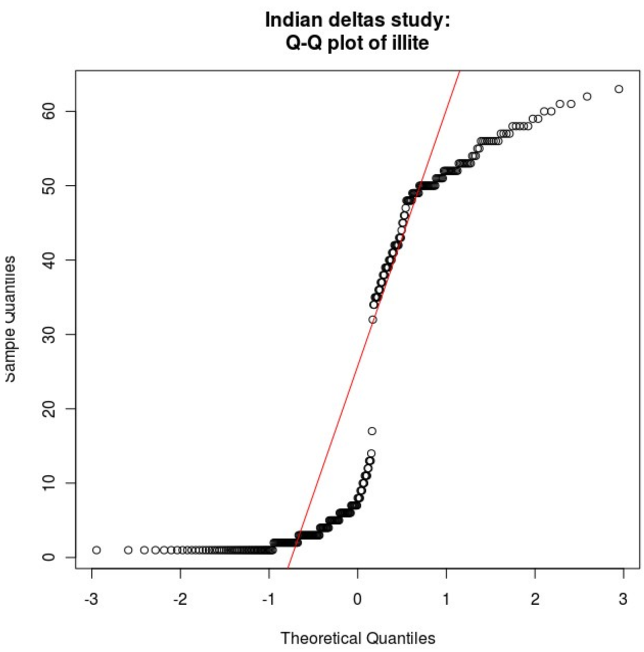
Indian delta study:
Index plot of kaolinite by delta
Cauvery
Penner
Krishna
Godavari
Mahanadi
del1$kaolinite
0
5
10
15
20
25
30
0
100
200
300
400
Index

TESTS FOR NORMALITY: Q=Q plots

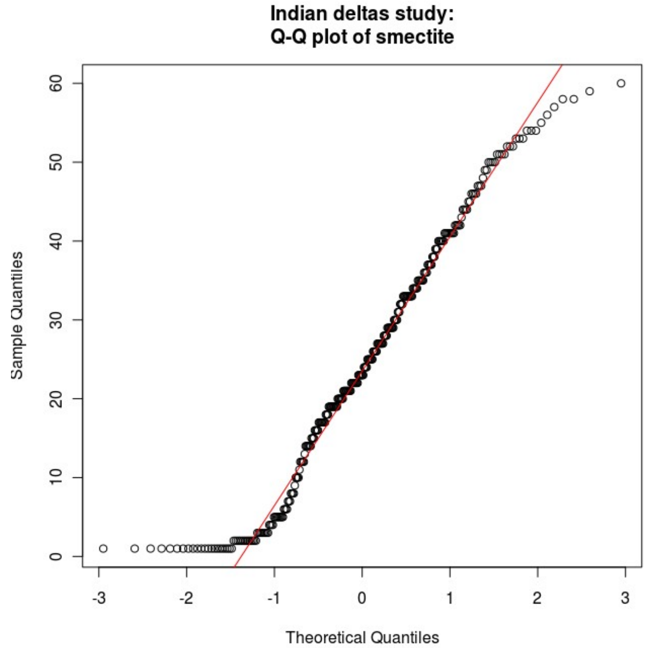


D'Agostino skewness test data: kaolinite, skew = 1.64, z = 5.29, p-value = 1.234e-07

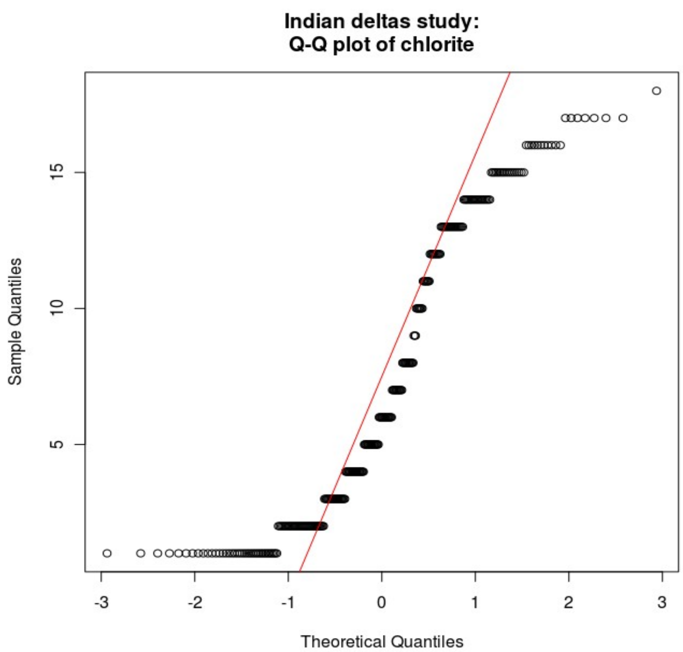
Variable has skewness.



D'Agostino skewness test data: illite, skew = 0.359, z = 1.693, p-value = 0.0904

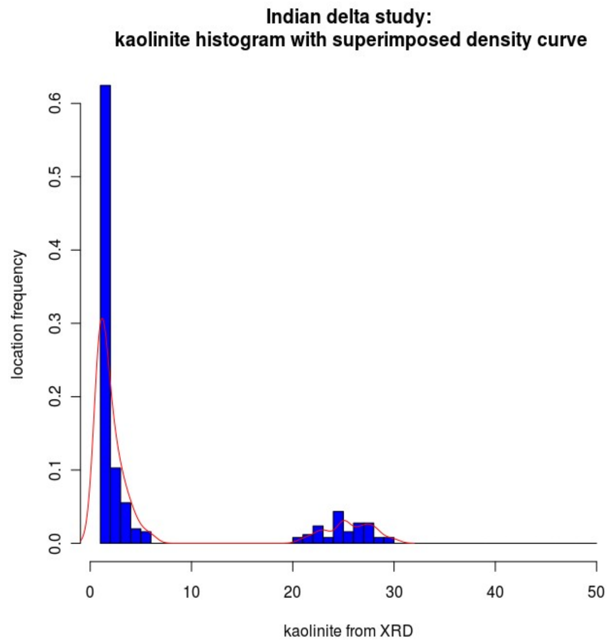


D'Agostino skewness test data: smectite, skew = 0.196, z = 0.946, p-value = 0.3441



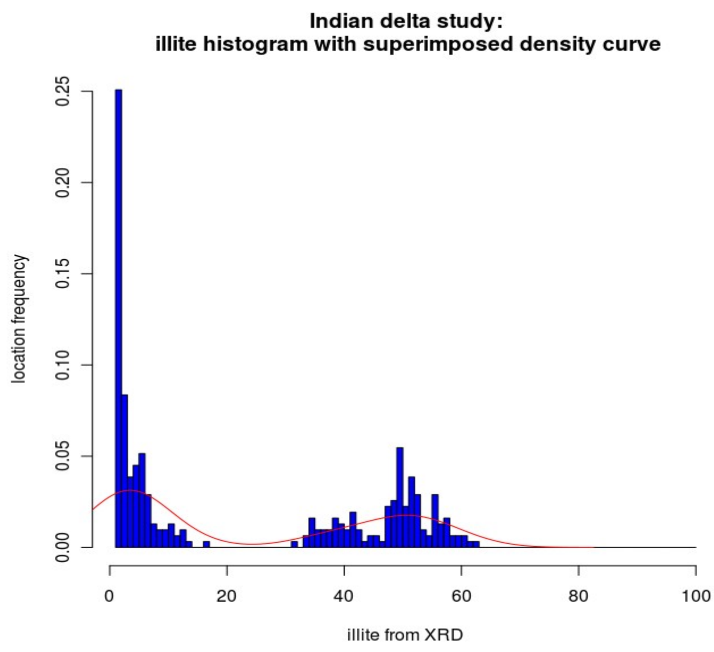
D'Agostino skewness test data: chlorite,skew = 0.401, z = 1.851, p-value = 0.06411

Histograms with overlain density plot (n=1,000)



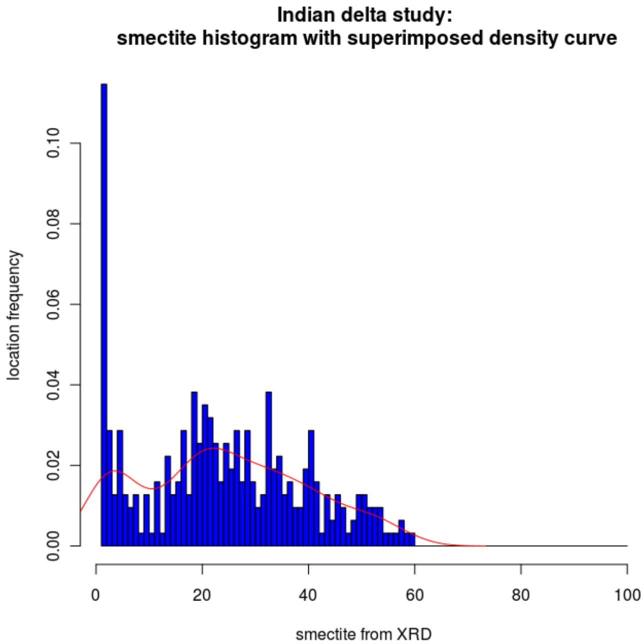
Shapiro-Wilk normality test , data: kaolinite W = 0.575, p-value < 2.2e-16

Variable is not Gaussian



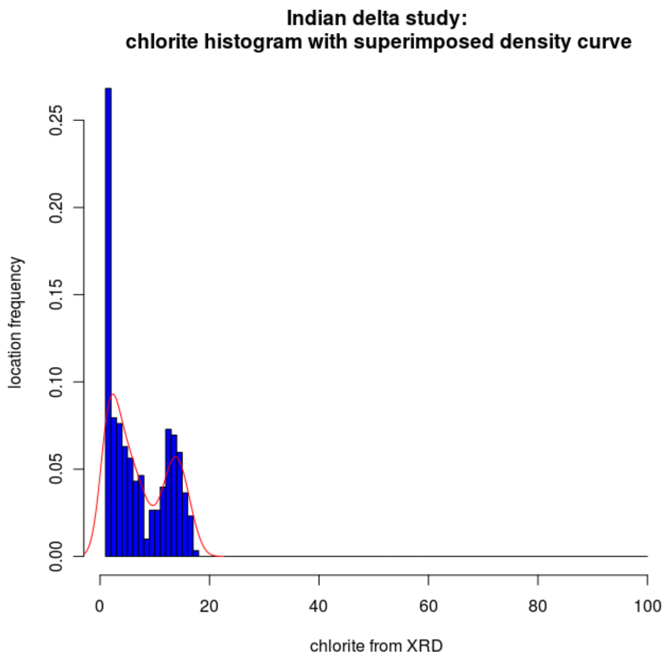
Shapiro-Wilk normality test , data: illite W = 0.781, p-value < 2.2e-16

Variable is not Gaussian



Shapiro-Wilk normality test , data: smectite W = 0.962, p-value = 2.624e-07

Variable is not Gaussian



Shapiro-Wilk normality test , data: chlorite W = 0.887, p-value = 3.833e-14

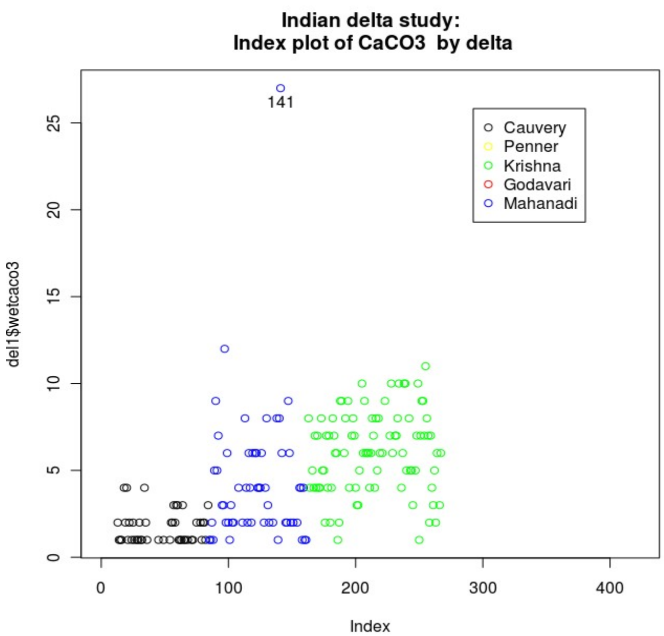
Variable is not Gaussian

ANALYSES OF OTHER CHEMICAL SPECIES

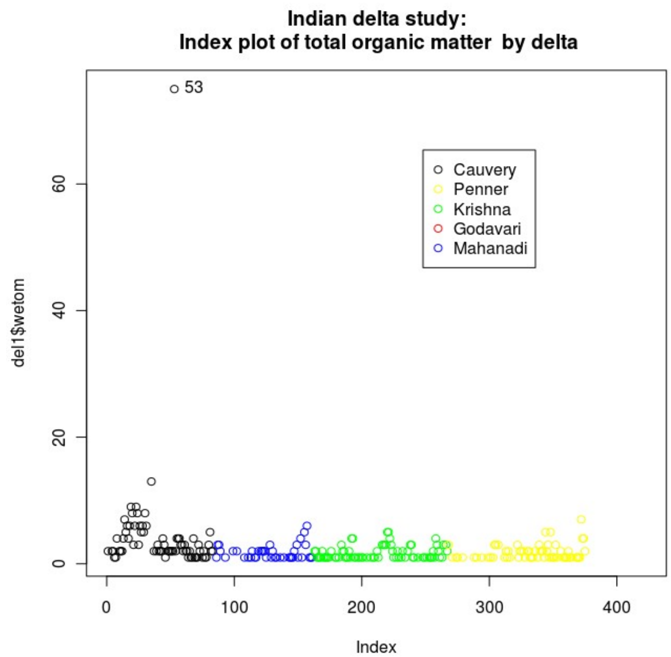
Table 6: SUMMARY STATISTICS FOR OTHER CHEMICAL SPECIES							
Mineral	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	NA's
caco3	1	2	4	4.5	7	27	230
tom	1	1	2	2.5	3	75	170
orgc	1	1	1	1.5	2	15	315

Error checking of original data using the Index plot

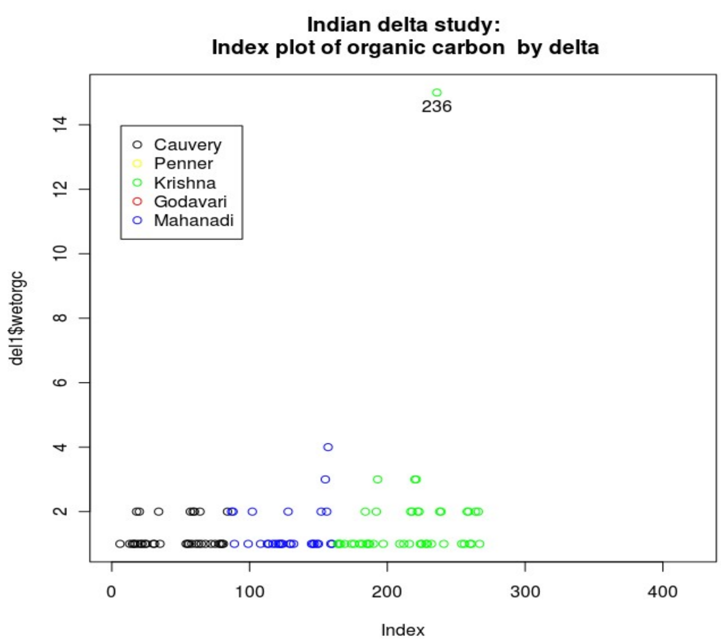
Index plot for **CaCO3** showing aberrant value of sample 141 from the Mahanadi delta.



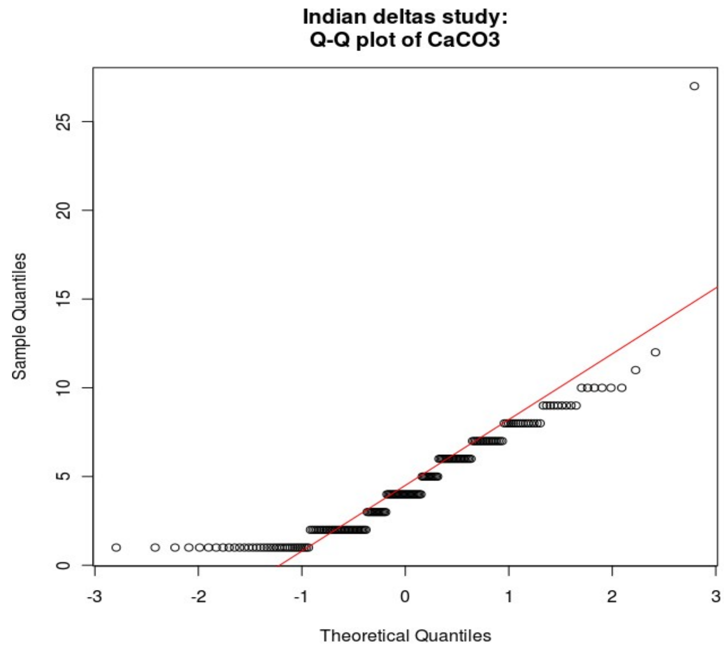
Index plot for **total organic matter** showing aberrant value of sample >53 from the Cauvery delta.



Index plot for **organic carbon** showing aberrant value of sample 236.

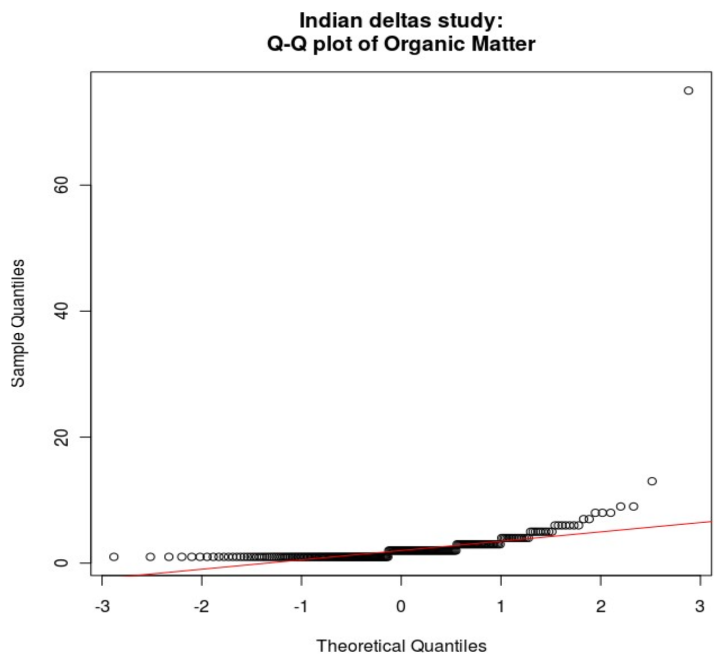


TESTS FOR NORMALITY: Q=Q plots



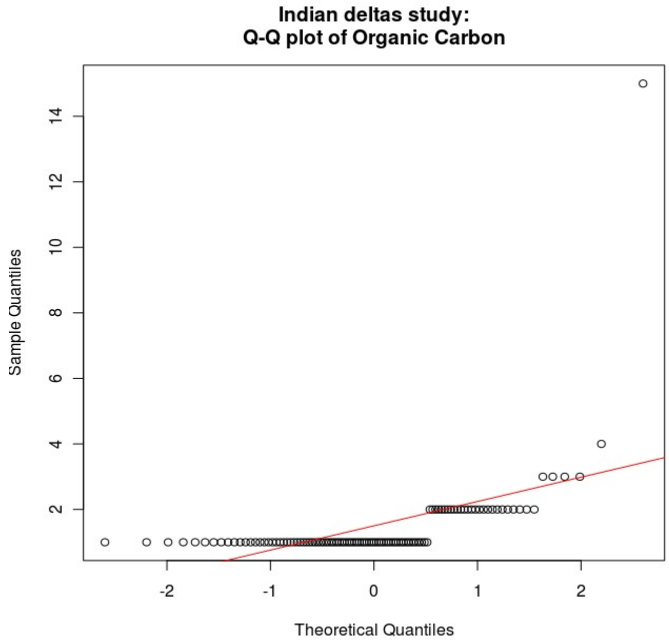
D'Agostino skewness test data: CaCO3, skew = 2.00, z = 5.25, p-value = 1.514e-07

Variable has skewness.



D'Agostino skewness test data: Organic Matter, skew = 13.0, z = 12.8, p-value < 2.2e-16

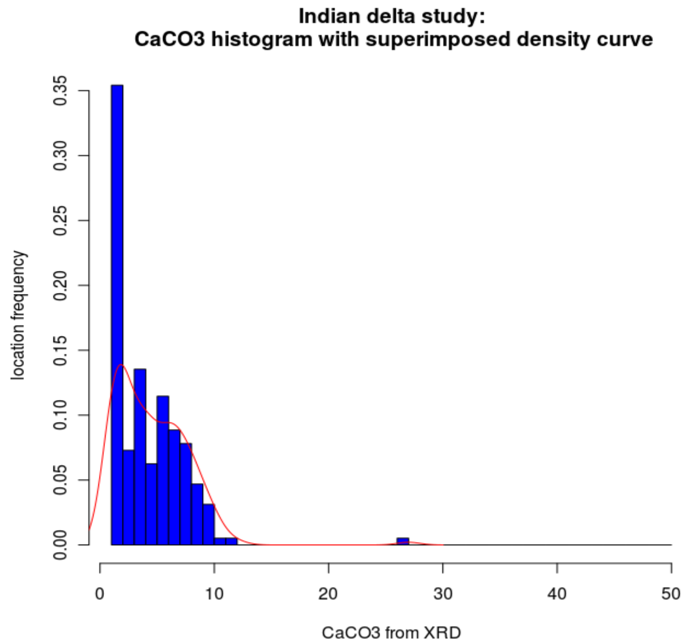
Variable has skewness.



D'Agostino skewness test data: Organic Carbon, skew = 7.78, z = 7.59, p-value = 3.318e-14

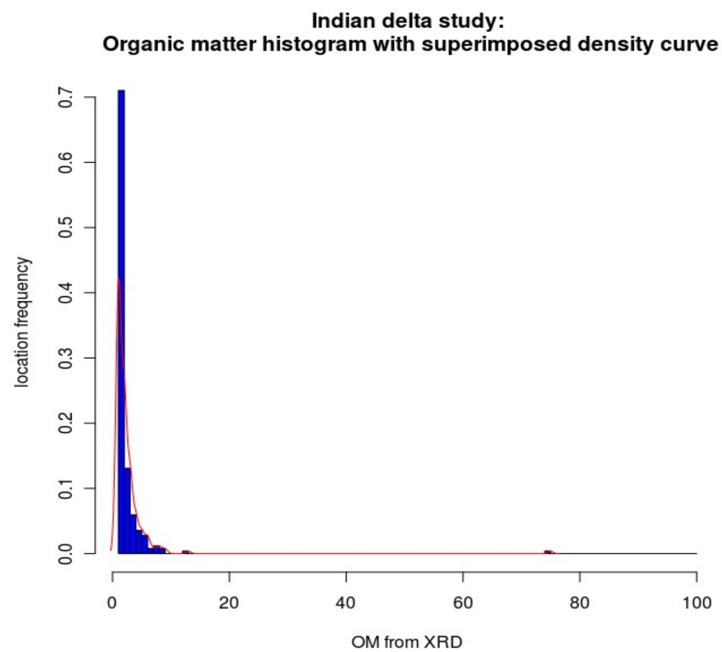
Variable has skewness.

Histograms with overlain density plot (n=1,000)



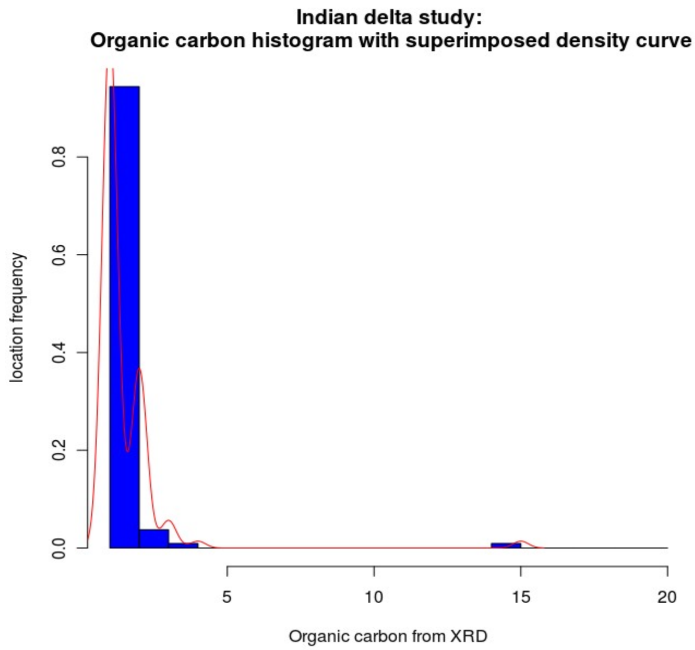
Shapiro-Wilk normality test , data: CaCO3 W = 0.839, p-value = 2.561e-13

Variable is not Gaussian



Shapiro-Wilk normality test , data: Organic Matter W = 0.22, p-value < 2.2e-16

Variable is not Gaussian



Shapiro-Wilk normality test , data: Organic Carbon W = 0.299, p-value < 2.2e-16

Variable is not Gaussian