

PE 5970: Data Mining for Petroleum Engineers

R Assignment 1: Due 18th February.

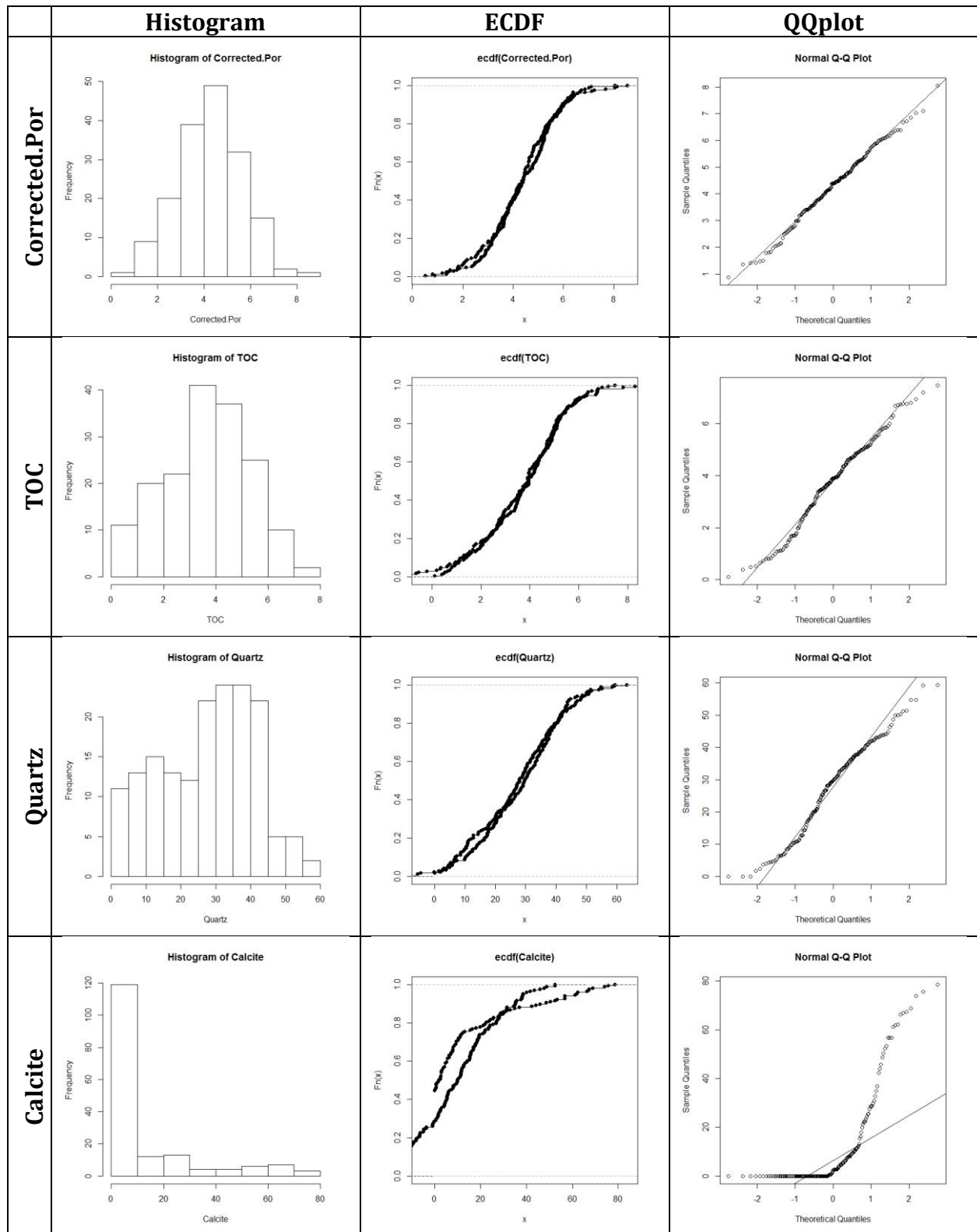
113102152 Jiwon Jeon

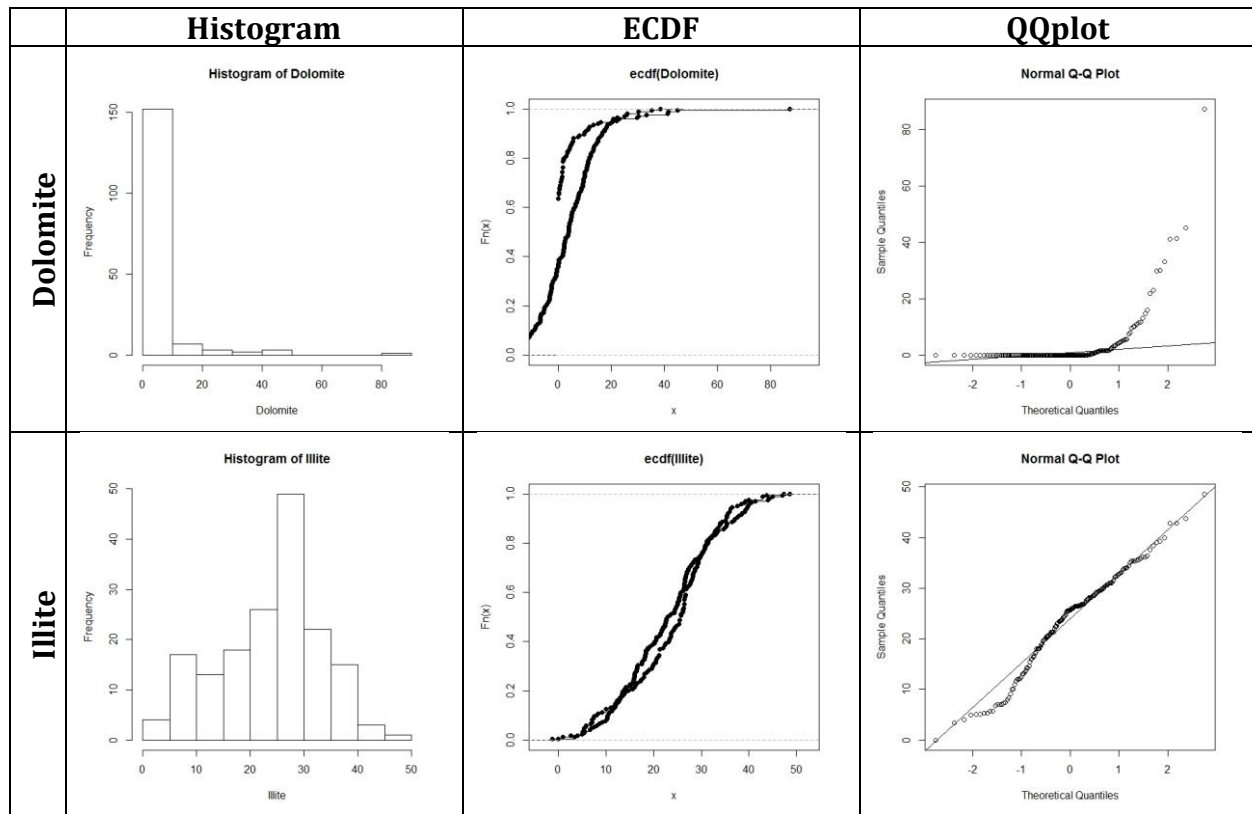
1. Show that the slope is related to the covariance.
2. Download the file, ASW.dat from D2L. Analyze the distribution of each variable. Are they normal?
 - a. Generate linear regression plots for each combination of variables except dolomite and illite. Analyze the residuals. Is linear regression doing a good job in each of the cases?
 - b. Also determine the covariance and correlation values between each pair of variables. Construct the covariance and correlation matrix.
 - c. Predict Porosity using multiple linear regression with TOC, Quartz and Calcite as inputs.
3. For the specific case of Porosity versus TOC, determine the 'best' value of the slope and intercept using Bayes' theorem. This approach is called Bayes' Linear Regression.

Questions 1 and 2 should be a hardcopy submission in class next week. Question 1 can be handwritten. Question 3 should be a copy of your R script uploaded to the dropbox on D2L.

Please turn in your work. Discuss among yourselves as much as you like.

2. Normal Distribution





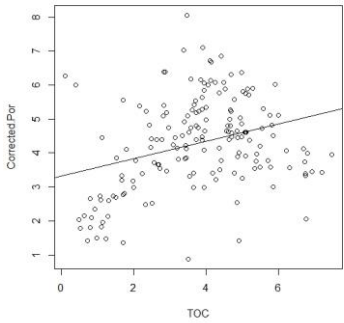
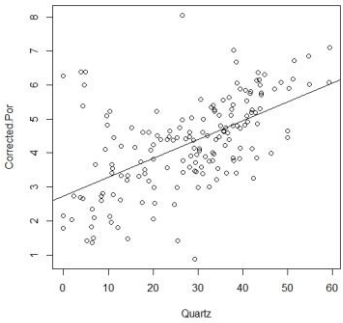
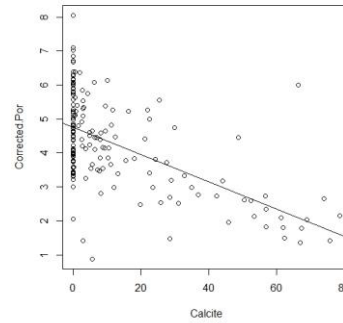
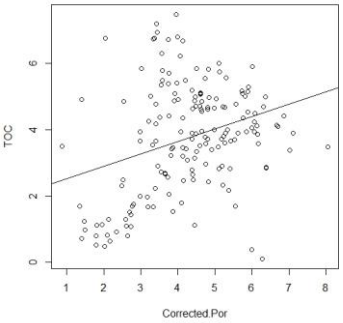
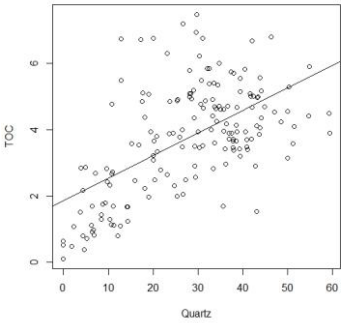
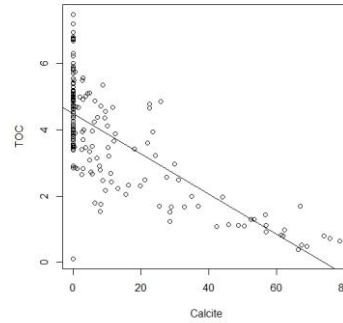
From the histogram of each variable, it can be said that Porosity and TOC are well fitted to normal distribution, Quartz and Illite show less normal distribution, however, Calcite and Dolomite do not follow the Gaussian (normal) distribution.

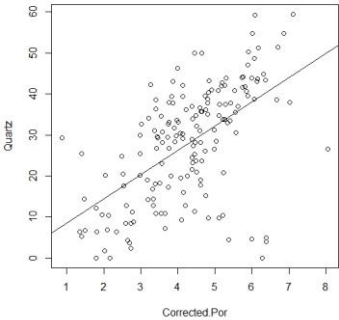
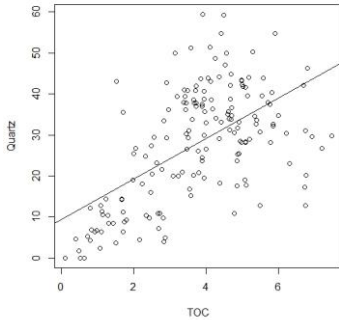
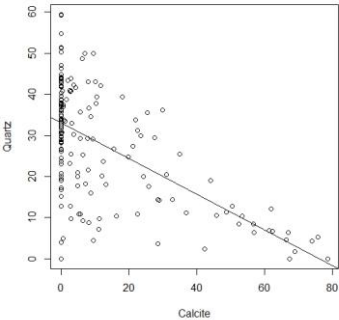
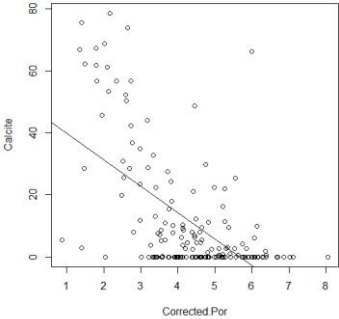
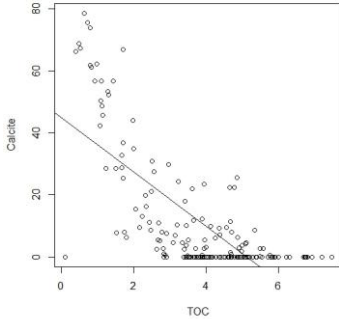
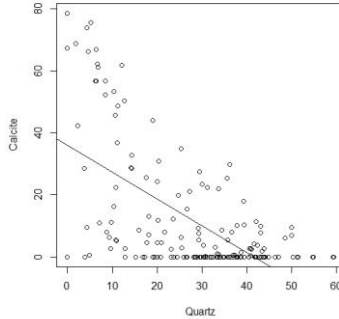
This trend is also shown in ECDF profiles; ECDF plots for Porosity and TOC have S-shape curves which indicate normally distributed data. The plots are almost overlapped with ECDF lines of normal distribution function, therefore, it can be concluded that Porosity and TOC data show very good normal distribution. Quartz and Illite ECDFs display proper profiles, however, they have bigger deviations from the normal distribution ECDF lines compared to those of Porosity and TOC. For Calcite and Dolomite, ECDF profiles confirm that they are not normally distributed.

In QQplots, normalized Porosity and TOC points have very good qqlines which pass through the data, followed by fair trends of Quartz and Illite. Meanwhile, QQplots for Calcite and Dolomite data show bad matching with corresponding qqlines.

2.a Linear Regression & Residuals

1) Linear Regression Plots

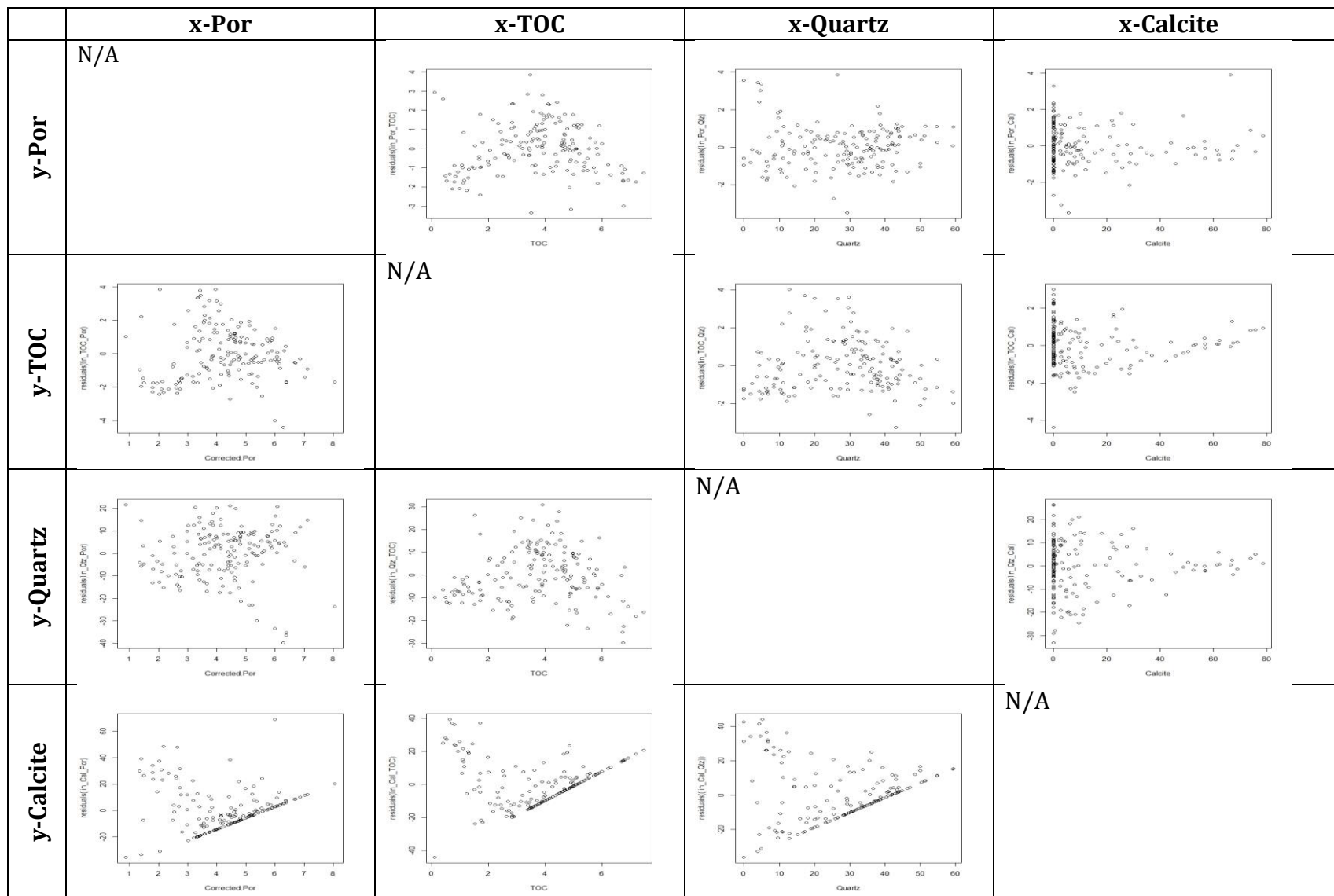
	x-Por	x-TOC	x-Quartz	x-Calcite
y-Por	N/A	 <p>Por = $3.323 + 0.254\text{TOC}$</p>	 <p>Por = $2.7362 + 0.0554\text{Quartz}$</p>	 <p>Por = $4.7578 - 0.0401\text{Calcite}$</p>
y-TOC	 <p>TOC = $2.147 + 0.375\text{Por}$</p>	N/A	 <p>TOC = $1.858 + 0.068\text{Quartz}$</p>	 <p>TOC = $4.4801 - 0.0607\text{Calcite}$</p>

	x-Por	x-TOC	x-Quartz	x-Calcite
y-Quartz	 $\text{Quartz} = 2.51 + 5.92\text{Por}$	 $\text{Quartz} = 9.32 + 4.93\text{TOC}$	N/A	 $\text{Quartz} = 33.022 - 0.434\text{Calcite}$
y-Calcite	 $\text{Calcite} = 48.59 - 8.55\text{Por}$	 $\text{Calcite} = 44.93 - 8.77\text{TOC}$	 $\text{Calcite} = 36.111 - 0.866\text{Quartz}$	N/A

Considering the linear regression lines passing through the data points and the residuals plots displayed below, it can be said that the linear regression fits well for Por ~ TOC, Por ~ Quartz, TOC ~ Por, and TOC ~ Quartz relationship. The Residuals profiles show that the data are close to 0 line. Even though the linear regression also seems to work well for Por ~ Calcite and TOC ~ Calcite, the Residuals plots show that the regression does not fit well for small values of Calcite.

For Quartz ~ Por, Quartz ~ TOC, Quartz ~ Calcite as well as Calcite ~ Por, Calcite ~ TOC and Calcite ~ Quartz relationships, the Residual plots show big deviations from 0 line, which indicates that the linear regression does not function properly.

2) Residual Plots



2.b. Covariance and Correlation

1) Covariance Matrix

	Porosity	TOC	Quartz	Calcite	Dolomite	Illite
Porosity	1.8346239	0.6876901	10.85195	-15.680913	-0.6045053	3.50102
TOC	0.6876901	2.7071876	13.33902	-23.744422	-5.0941657	9.78373
Quartz	10.8519489	13.3390209	196.04113	-169.694641	-43.1941296	30.72633
Calcite	-15.6809132	-23.7444222	-169.69464	391.337964	-1.4816766	-139.49186
Dolomite	-0.6045053	-5.0941657	-43.19413	-1.481677	100.5877887	-25.61804
Illite	3.5010200	9.7837298	30.72633	-139.491856	-25.6180368	92.34299

2) Correlation Matrix

	Porosity	TOC	Quartz	Calcite	Dolomite	Illite
Porosity	1.00000000	0.3085745	0.5722168	-0.585223599	-0.044499383	0.2689794
TOC	0.30857451	1.0000000	0.5790168	-0.729501919	-0.308703295	0.6187900
Quartz	0.57221681	0.5790168	1.0000000	-0.612658607	-0.307594453	0.2283679
Calcite	-0.58522360	-0.7295019	-0.6126586	1.000000000	-0.007468009	-0.7337885
Dolomite	-0.04449938	-0.3087033	-0.3075945	-0.007468009	1.000000000	-0.2658100
Illite	0.26897945	0.6187900	0.2283679	-0.733788482	-0.265809955	1.0000000

2.c. Multiple linear regression for Porosity

Porosity = **4.7849 - 0.306*TOC + 0.0407*Quartz - 0.041*Calcite**

Code for Question 2.c

```
lin_Por = lm(Corrected.Por ~ TOC+Quartz+Calcite)
```

```
print(lin_Por)
```