

Iron Ore Quality Prediction Report

Tholang Chonelanga

2023-01-30

Introduction

Flotation is a process in which pulp containing some mineral ore deposit is fed into a plant with some reagents such that at the end of the process, pure mineral is collected while all impurities stay in the plant. In this project, the mineral of interest is iron ore. The quality of iron ore in a mining flotation plant is determined by the amount of silica (SiO_2) in the ore concentrate; a higher silica amount is indicative of a more impure sample. In this project the objective is to use other variables from the plant to try and predict the silica content. This will be helpful in improving productivity as there is at least one hour that has to be waited before obtaining the silica content reading of each sample sent to the lab, and many samples are sent in a day, creating a lot of dead time. Prediction using machine learning would save this time. The R code used to create this report is found in the `iron_ore_code.R` file that goes with this report.

Dataset

The `data` dataset taken from kaggle contains 7 37 453 entries of 24 variables from an iron ore mining flotation plant. The variables are attributes of the plant like pulp density, pulp pH and flow rate, amount of iron and silica in the feed, flow rates of the reagents used in the plant. Below are the first six lines of the dataset.

```
##           date X..Iron.Feed X..Silica.Feed Starch.Flow Amina.Flow
## 1 2017-03-10 01:00:00      55,2          16,98      3019,53      557,434
## 2 2017-03-10 01:00:00      55,2          16,98      3024,41      563,965
## 3 2017-03-10 01:00:00      55,2          16,98      3043,46      568,054
## 4 2017-03-10 01:00:00      55,2          16,98      3047,36      568,665
## 5 2017-03-10 01:00:00      55,2          16,98      3033,69      558,167
## 6 2017-03-10 01:00:00      55,2          16,98       3079,1      564,697
## Ore.Pulp.Flow Ore.Pulp.pH Ore.Pulp.Density Flotation.Column.01.Air.Flow
## 1      395,713      10,0664              1,74              249,214
## 2      397,383      10,0672              1,74              249,719
## 3      399,668      10,068              1,74              249,741
## 4      397,939      10,0689              1,74              249,917
## 5      400,254      10,0697              1,74              250,203
## 6      396,533      10,0705              1,74              250,73
## Flotation.Column.02.Air.Flow Flotation.Column.03.Air.Flow
## 1      253,235              250,576
## 2      250,532              250,862
## 3      247,874              250,313
## 4      254,487              250,049
## 5      252,136              249,895
## 6      248,906              249,521
## Flotation.Column.04.Air.Flow Flotation.Column.05.Air.Flow
## 1      295,096              306,4
```

## 2	295,096	306,4	
## 3	295,096	306,4	
## 4	295,096	306,4	
## 5	295,096	306,4	
## 6	295,096	306,4	
##	Flotation.Column.06.Air.Flow	Flotation.Column.07.Air.Flow	
## 1	250,225	250,884	
## 2	250,137	248,994	
## 3	251,345	248,071	
## 4	250,422	251,147	
## 5	249,983	248,928	
## 6	250,356	251,873	
##	Flotation.Column.01.Level	Flotation.Column.02.Level	Flotation.Column.03.Level
## 1	457,396	432,962	424,954
## 2	451,891	429,56	432,939
## 3	451,24	468,927	434,61
## 4	452,441	458,165	442,865
## 5	452,441	452,9	450,523
## 6	444,384	443,269	460,449
##	Flotation.Column.04.Level	Flotation.Column.05.Level	Flotation.Column.06.Level
## 1	443,558	502,255	446,37
## 2	448,086	496,363	445,922
## 3	449,688	484,411	447,826
## 4	446,21	471,411	437,69
## 5	453,67	462,598	443,682
## 6	439,92	451,588	433,539
##	Flotation.Column.07.Level	X..Iron.Concentrate	X..Silica.Concentrate
## 1	523,344	66,91	1,31
## 2	498,075	66,91	1,31
## 3	458,567	66,91	1,31
## 4	427,669	66,91	1,31
## 5	425,679	66,91	1,31
## 6	425,458	66,91	1,31

Cleaning The Dataset

All variables except date are numeric, but are not written in the correct format, with commas making them character strings. The date variable is also not in the datetime type. Cleaning this dataset will just be about making all variables numeric and datetime for the date variable, and also renaming the long variable names to shorter versions that will be quicker to write. Hourly entries will also have to be aggregated, because as seen in the dataset, there are different readings of all the other variables for each 60 entries with the same silica content reading. This is because all those entries belong to a single batch of ore taken to the lab to obtain the silica content reading, that comes after one hour. These values will be combined and an average taken for each variable to correspond with each unique value of silica concentrate. Below are the first six lines of the cleaned dataset.

##	date	iron_feed	silica_feed	starch_flow	amina_flow	pulp_flow
## 1	2017-03-10 01:00:00	55.2	16.98	3162.625	578.7867	398.7534
## 2	2017-03-10 02:00:00	55.2	16.98	3133.256	537.2197	399.8718
## 3	2017-03-10 03:00:00	55.2	16.98	3479.483	591.9067	398.7638
## 4	2017-03-10 04:00:00	55.2	16.98	3228.036	593.1701	399.8670
## 5	2017-03-10 05:00:00	55.2	16.98	3327.281	619.7108	399.6151
## 6	2017-03-10 06:00:00	55.2	16.98	3405.162	621.8785	399.7493

```

##      pulp_ph pulp_density air_flow_1 air_flow_2 air_flow_3 air_flow_4 air_flow_5
## 1 10.113487    1.729558   251.1667   250.2261   250.1783   295.096   306.4
## 2 10.129742    1.667784   249.8806   250.2140   250.0333   295.096   306.4
## 3 10.048403    1.732711   250.1613   250.1042   250.0463   295.096   306.4
## 4  9.918614    1.731056   250.2088   250.2048   250.1209   295.096   306.4
## 5  9.746029    1.765879   249.9178   250.1605   250.0135   295.096   306.4
## 6  9.892237    1.765064   249.8983   250.1110   250.0754   295.096   306.4
##      air_flow_6 air_flow_7 column_level_1 column_level_2 column_level_3
## 1    251.2325   250.2082     450.3838     446.8918     450.4745
## 2    249.9095   249.8976     449.3734     450.2494     450.0812
## 3    250.2422   250.4842     449.9729     450.8687     450.9018
## 4    249.8251   250.1576     487.9407     491.4621     487.3872
## 5    250.2496   250.0786     549.0315     549.9832     549.4596
## 6    249.9425   250.1499     550.5996     549.9291     549.0892
##      column_level_4 column_level_5 column_level_6 column_level_7 iron_concentrate
## 1         449.9123         455.7922         464.3833         450.5327         66.91
## 2         450.3288         448.7230         455.5015         451.3877         67.06
## 3         451.1458         451.1342         459.9813         450.2967         66.97
## 4         494.5282         495.6640         502.7638         494.9399         66.75
## 5         549.9755         549.5125         560.6963         550.2718         66.63
## 6         549.6097         549.2207         561.0516         551.0908         66.85
##      silica_concentrate
## 1              1.31
## 2              1.11
## 3              1.27
## 4              1.36
## 5              1.34
## 6              1.15

```

Exploratory Data Analysis

The first step is to check the summary statistics of each variable, to see how variable each one is. Then explore the distribution of the response variable, and then checking how it changes with time. This can be done by plotting silica content against time and observe whether or not there is an obvious pattern. The first plot is the distribution of silica content in the concentrate, is it changing at all and if yes how? The second plot shows silica content over the course of the three months in which these data were collected.

```

##      date                iron_feed      silica_feed
## Min.   :2017-03-10 01:00:00.00  Min.   :42.74  Min.   : 1.31
## 1st Qu.:2017-05-04 23:00:00.00  1st Qu.:52.67  1st Qu.: 8.94
## Median :2017-06-16 15:00:00.00  Median :56.08  Median :13.85
## Mean   :2017-06-16 03:26:05.82  Mean   :56.29  Mean   :14.65
## 3rd Qu.:2017-07-29 07:00:00.00  3rd Qu.:59.72  3rd Qu.:19.60
## Max.   :2017-09-09 23:00:00.00  Max.   :65.78  Max.   :33.40
##      starch_flow      amina_flow      pulp_flow      pulp_ph
## Min.   : 54.59  Min.   :242.9  Min.   :376.8  Min.   : 8.753
## 1st Qu.:2168.97  1st Qu.:436.0  1st Qu.:398.9  1st Qu.: 9.541
## Median :2908.34  Median :502.5  Median :399.8  Median : 9.796
## Mean   :2869.14  Mean   :488.1  Mean   :397.6  Mean   : 9.768
## 3rd Qu.:3528.73  3rd Qu.:549.5  3rd Qu.:400.6  3rd Qu.:10.031
## Max.   :6270.16  Max.   :737.0  Max.   :418.1  Max.   :10.807
##      pulp_density      air_flow_1      air_flow_2      air_flow_3
## Min.   :1.520  Min.   :175.9  Min.   :178.2  Min.   :177.2

```

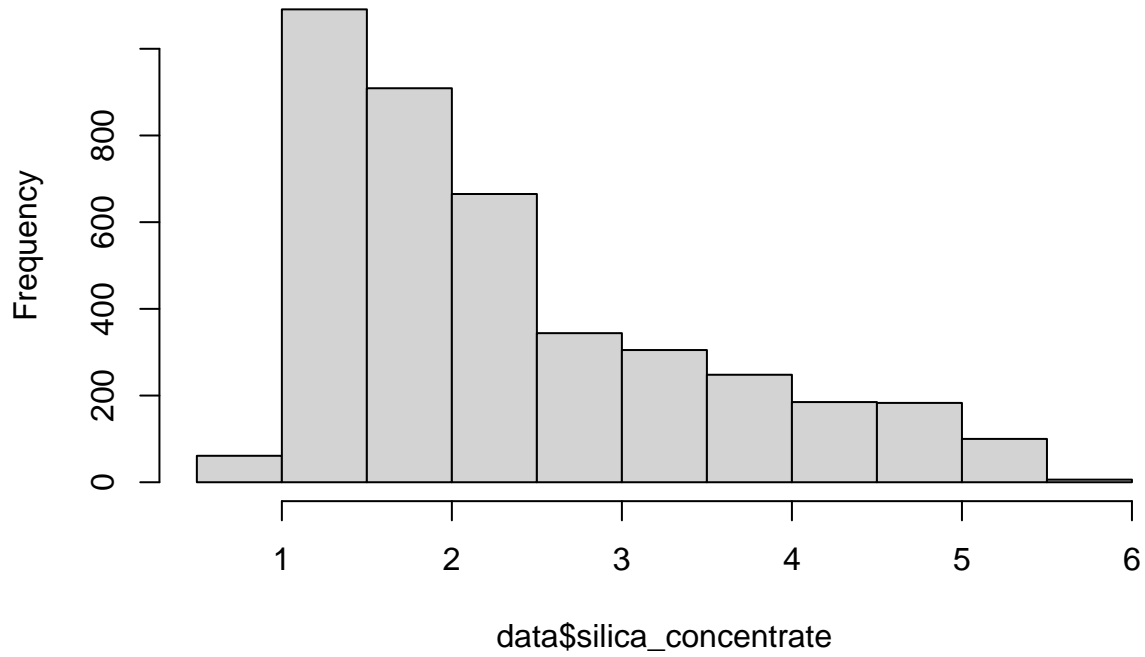
```

## 1st Qu.:1.651 1st Qu.:250.1 1st Qu.:250.1 1st Qu.:250.1
## Median :1.696 Median :299.8 Median :299.5 Median :299.9
## Mean :1.680 Mean :280.2 Mean :277.2 Mean :281.1
## 3rd Qu.:1.722 3rd Qu.:300.0 3rd Qu.:300.0 3rd Qu.:299.9
## Max. :1.832 Max. :312.3 Max. :309.9 Max. :302.8
## air_flow_4 air_flow_5 air_flow_6 air_flow_7
## Min. :293.3 Min. :287.1 Min. :196.5 Min. :199.7
## 1st Qu.:299.7 1st Qu.:299.7 1st Qu.:268.7 1st Qu.:283.2
## Median :299.9 Median :299.9 Median :299.9 Median :299.9
## Mean :299.4 Mean :299.9 Mean :292.1 Mean :290.8
## 3rd Qu.:300.0 3rd Qu.:300.1 3rd Qu.:300.1 3rd Qu.:300.1
## Max. :305.6 Max. :307.0 Max. :355.0 Max. :351.3
## column_level_1 column_level_2 column_level_3 column_level_4
## Min. :181.9 Min. :224.9 Min. :135.2 Min. :165.7
## 1st Qu.:416.5 1st Qu.:449.2 1st Qu.:405.4 1st Qu.:351.5
## Median :499.6 Median :499.8 Median :499.6 Median :401.3
## Mean :520.2 Mean :522.6 Mean :531.4 Mean :420.3
## 3rd Qu.:599.7 3rd Qu.:599.3 3rd Qu.:600.2 3rd Qu.:496.2
## Max. :859.0 Max. :827.8 Max. :884.8 Max. :675.6
## column_level_5 column_level_6 column_level_7 iron_concentrate
## Min. :214.7 Min. :203.7 Min. :185.1 Min. :62.05
## 1st Qu.:351.0 1st Qu.:354.1 1st Qu.:350.9 1st Qu.:64.37
## Median :401.1 Median :407.5 Median :401.0 Median :65.21
## Mean :425.3 Mean :429.9 Mean :421.0 Mean :65.05
## 3rd Qu.:497.8 3rd Qu.:497.8 3rd Qu.:462.3 3rd Qu.:65.86
## Max. :674.1 Max. :698.5 Max. :655.5 Max. :68.01
## silica_concentrate
## Min. :0.600
## 1st Qu.:1.440
## Median :2.000
## Mean :2.327
## 3rd Qu.:3.010
## Max. :5.530

```

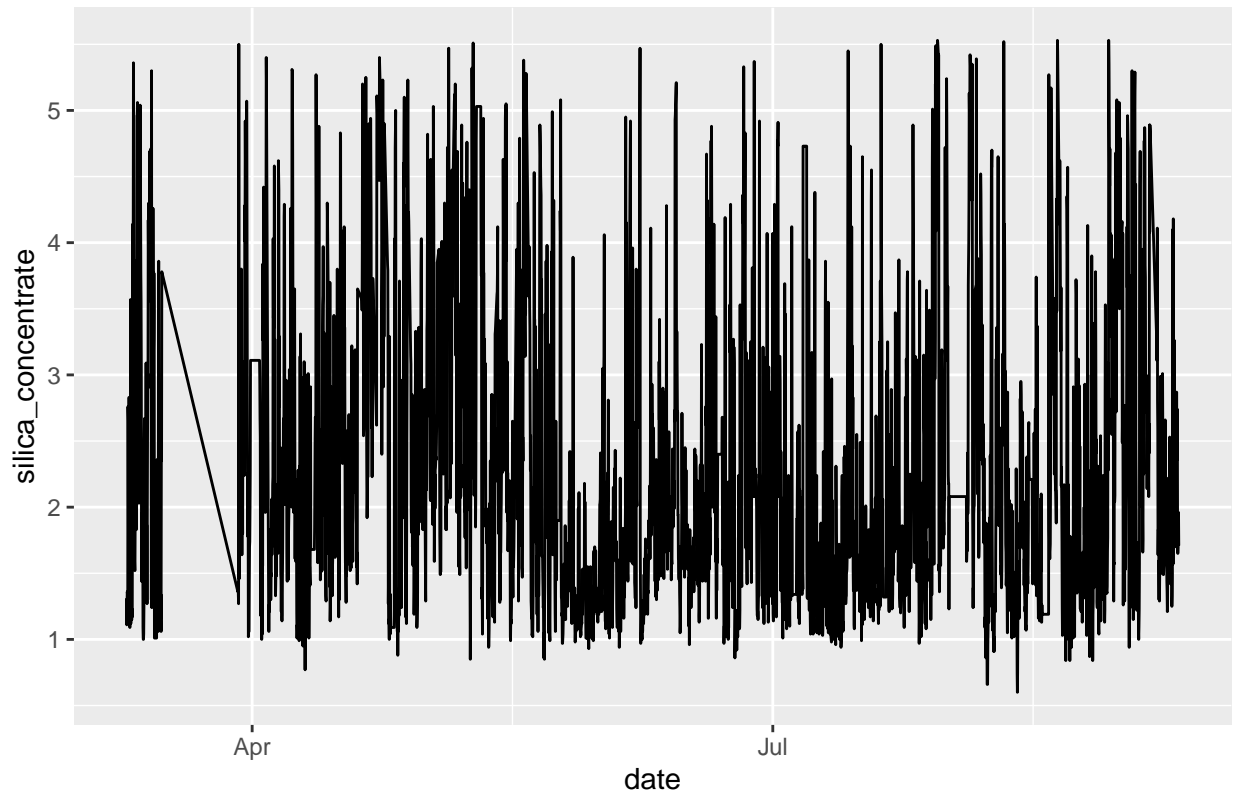
The summary shows that although almost all these variables are in the same order of magnitude, they are not exactly the same. The numbers are similar. Starch flow is the only variable in which the minimum is in the different order of magnitude from the rest of the other numbers.

Distribution of silica content in the concentrate



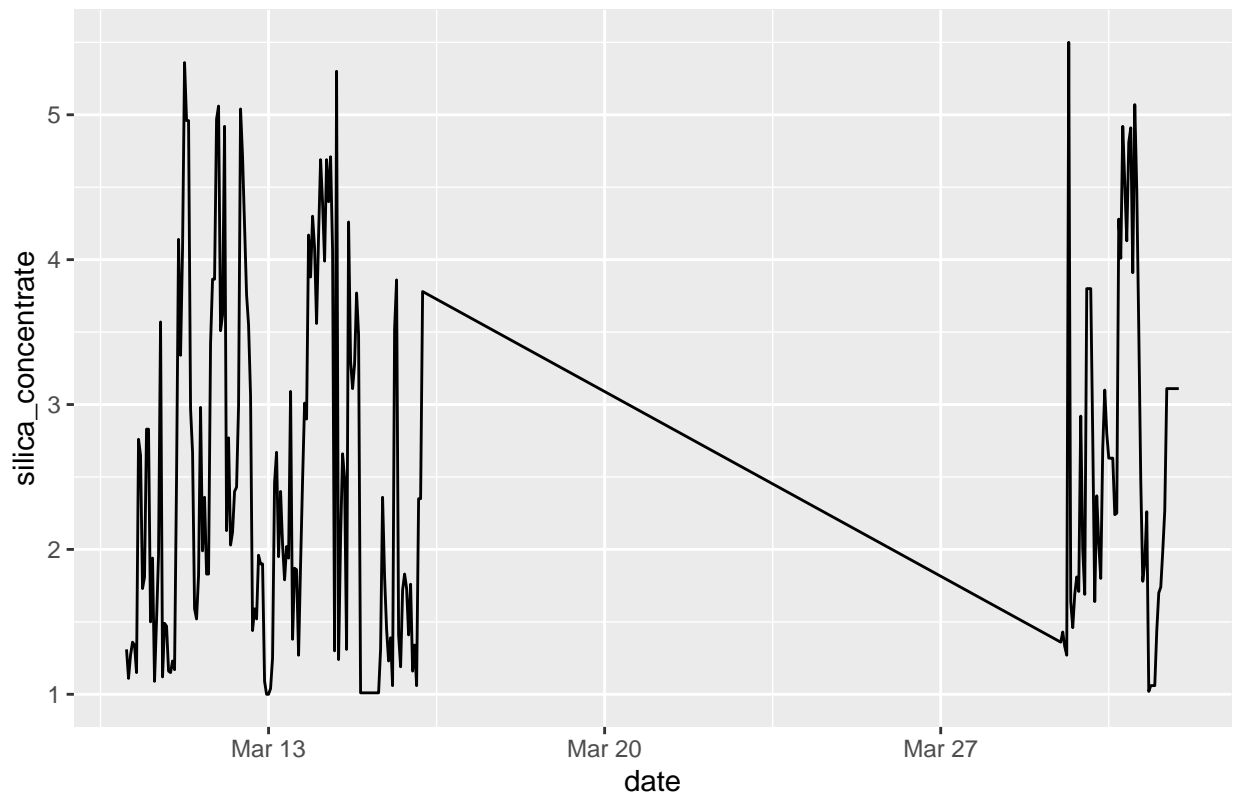
It can be seen that the distribution of the response variable is highly variable.

Silica content in concentrate over time



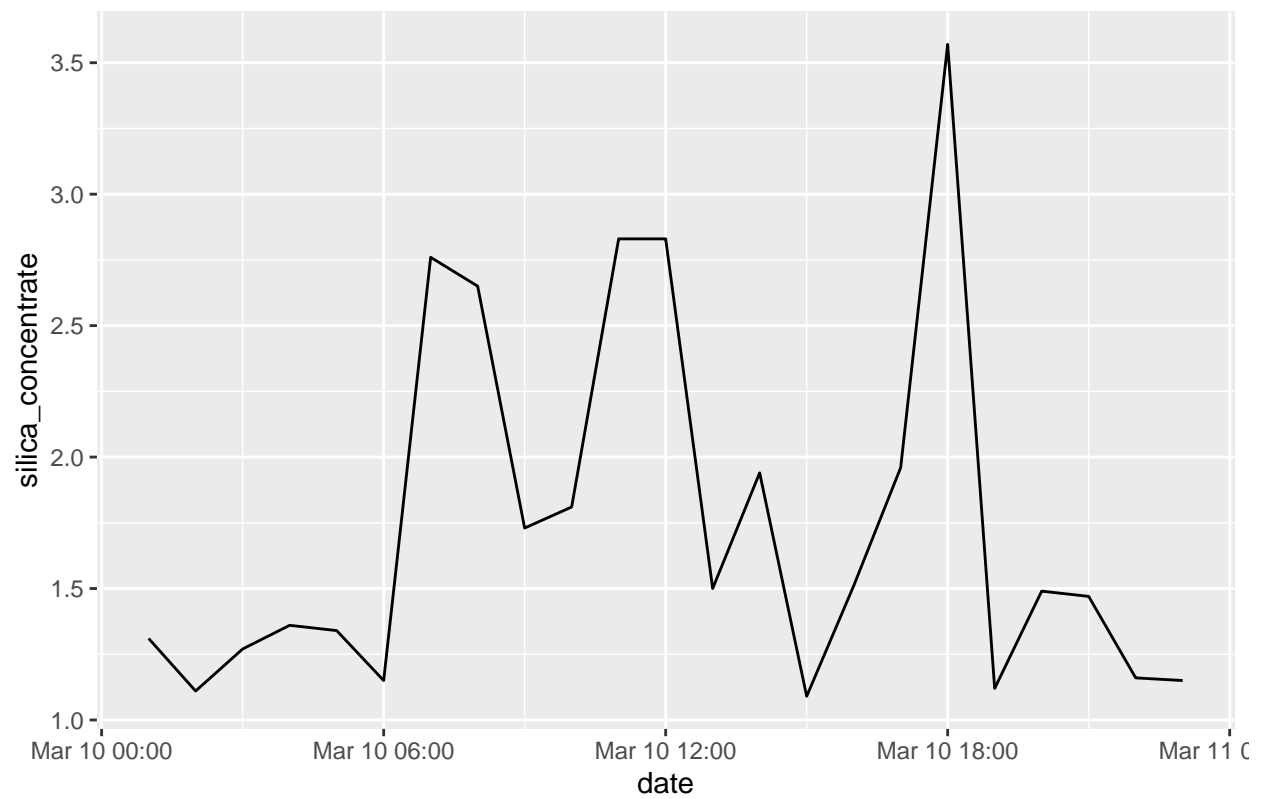
Looking at this plot shows that there is no clear pattern between the response variable and time in the whole of these three months. Maybe zooming in on just one month will be clearer :

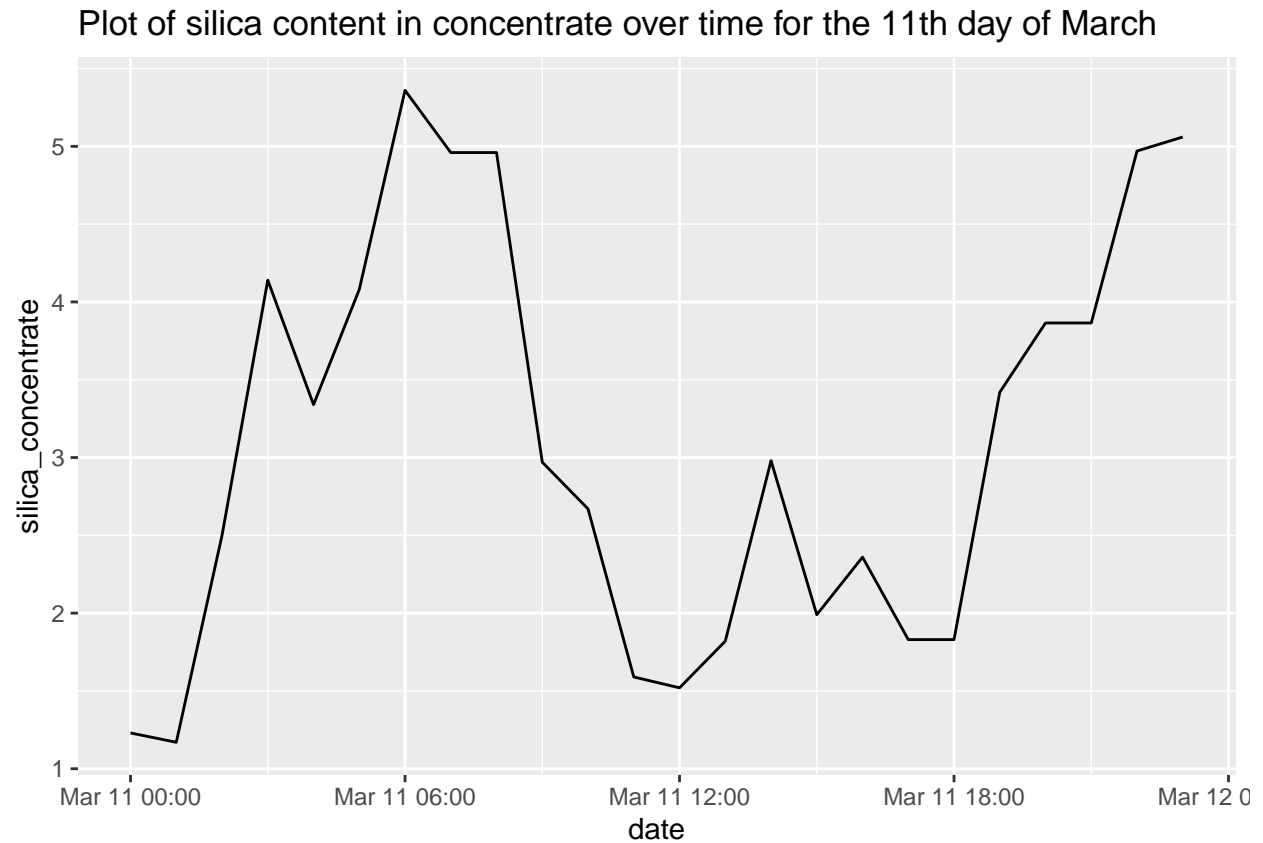
Plot of silica content in concentrate over time for the month of March



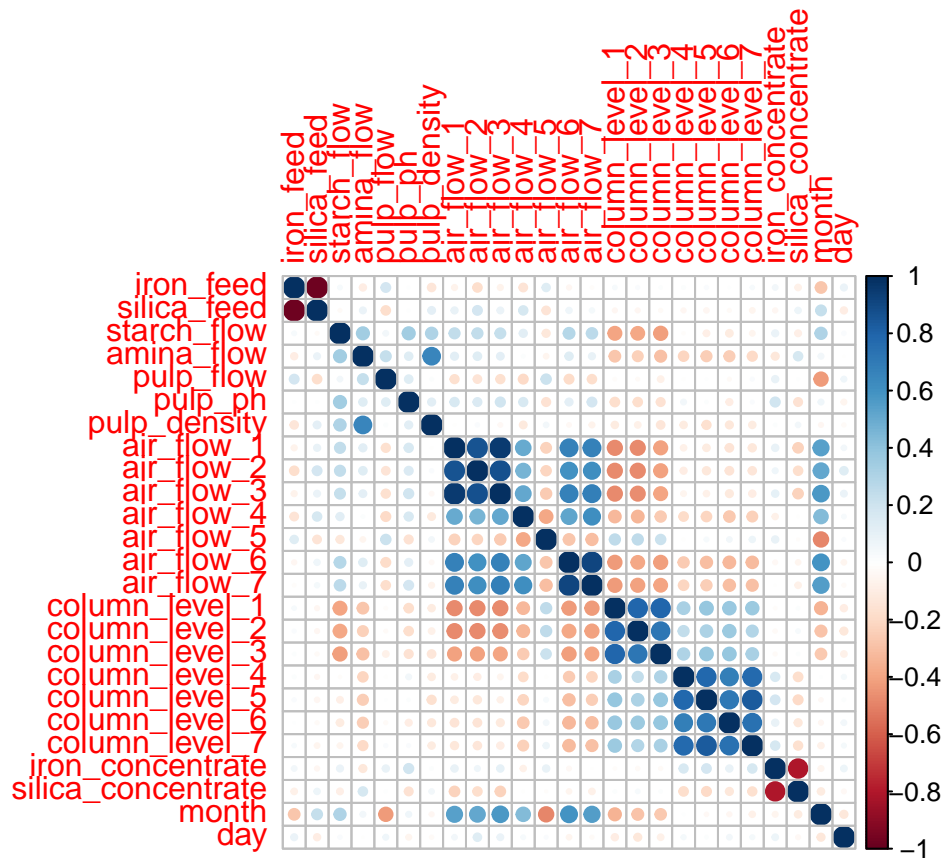
There still is no clear pattern. Below I check by a single day, will pick two random days to see if there is any relationship between silica content and say time of day.

Plot of silica content in concentrate over time for the 10th day of March





Both plots show that there is no obvious pattern. Next I move on to checking if all these variables are correlated with the response variable and with one another. This is in the correlation plot below:



The only variable that is clearly strongly correlated with the response variable as seen from the plot is iron concentrate, which is the amount of iron in the concentrate. It is a negative correlation and it makes sense because the more iron there is the less contaminant there will be and vice versa. All the variables do not seem to individually have a direct correlation with the response variable, but there are some variables that positively or negatively, correlated with one another.

Building A Predictive Model

Since all of the variables except date, in the dataset are numeric, the first option algorithm to explore is linear regression. The dataset is split into the training and test sets, and then a model is trained using linear regression and its performance tested on the test set. Because it has already been seen that many of the variables are correlated with one another, it is likely that the model will not fit the data with high efficiency due to the redundance in the features. Therefore, principal component analysis will be used to reduce the dimensions in the features and see whether or not the model fit will improve. Below is the summary of the model run with linear regression:

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05853 -0.39769 -0.03191  0.36265  2.33681
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.445e+02  5.287e+02   1.408 0.159172
## date          -4.661e-07  3.569e-07  -1.306 0.191688
## iron_feed      -2.122e-03  1.035e-02  -0.205 0.837468
## silica_feed     1.287e-02  7.587e-03   1.696 0.089973 .
## starch_flow    -6.466e-06  1.575e-05  -0.410 0.681491
## amina_flow      1.102e-03  2.065e-04   5.337 1.01e-07 ***
## pulp_flow       7.456e-04  1.658e-03   0.450 0.652982
## pulp_ph         1.716e-02  3.534e-02   0.486 0.627270
## pulp_density    -2.167e-01  2.583e-01  -0.839 0.401546
## air_flow_1      -4.626e-03  1.479e-03  -3.128 0.001775 **
## air_flow_2      -2.995e-03  8.744e-04  -3.426 0.000621 ***
## air_flow_3       1.136e-03  1.642e-03   0.692 0.489094
## air_flow_4      -2.083e-03  6.507e-03  -0.320 0.748945
## air_flow_5       5.048e-03  4.105e-03   1.230 0.218897
## air_flow_6       3.844e-04  1.035e-03   0.371 0.710411
## air_flow_7       2.547e-03  1.133e-03   2.248 0.024615 *
## column_level_1  -4.941e-04  1.854e-04  -2.665 0.007735 **
## column_level_2  -1.499e-04  1.722e-04  -0.871 0.383828
## column_level_3  -3.869e-07  1.412e-04  -0.003 0.997813
## column_level_4  -1.048e-04  2.555e-04  -0.410 0.681719
## column_level_5  -2.380e-05  3.055e-04  -0.078 0.937891
## column_level_6  -2.229e-04  2.395e-04  -0.931 0.351982
## column_level_7   2.979e-04  3.097e-04   0.962 0.336128
## iron_concentrate -7.982e-01  1.087e-02 -73.438 < 2e-16 ***
## month           1.143e+00  9.417e-01   1.214 0.225008
## day             4.324e-02  3.075e-02   1.406 0.159747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6308 on 3250 degrees of freedom
## Multiple R-squared:  0.6895, Adjusted R-squared:  0.6871
## F-statistic: 288.6 on 25 and 3250 DF, p-value: < 2.2e-16
```

As anticipated, the model does not fit the data well enough, the adjusted R-squared value is 0.6871, so there might still be room for improvement. BUt first the model is tested on test data to see its performance on new data. Below is the RMSE value obtained from comparing predictions on the test set with real values:

```
## [1] 0.6143063
```

This RMSE value, 0.61 is almost equal to the minimim value of this variable in the dataset, it is a big error that the model makes when trying to predict. Maybe applying pca will bring it down. Below is the summary of model fit after applying pca:

```
## Importance of components:
##               PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation  2.6341 1.8483 1.54072 1.43414 1.24837 1.1336 1.08018
## Proportion of Variance 0.2775 0.1367 0.09495 0.08227 0.06234 0.0514 0.04667
## Cumulative Proportion 0.2775 0.4142 0.50914 0.59141 0.65375 0.7052 0.75182
##               PC8    PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation  0.99261 0.93657 0.84308 0.79394 0.71872 0.64511 0.56774
## Proportion of Variance 0.03941 0.03509 0.02843 0.02521 0.02066 0.01665 0.01289
## Cumulative Proportion 0.79123 0.82632 0.85475 0.87996 0.90062 0.91727 0.93016
```

```

##              PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation    0.54341 0.50216 0.49677 0.45738 0.42678 0.40593 0.38187
## Proportion of Variance 0.01181 0.01009 0.00987 0.00837 0.00729 0.00659 0.00583
## Cumulative Proportion 0.94198 0.95206 0.96193 0.97030 0.97759 0.98418 0.99001
##              PC22    PC23    PC24    PC25
## Standard deviation    0.35570 0.26215 0.1800 0.14861
## Proportion of Variance 0.00506 0.00275 0.0013 0.00088
## Cumulative Proportion 0.99507 0.99782 0.9991 1.00000

##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56545 -0.19537  0.01189  0.19806  1.23198
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.327561   0.005701  408.259 < 2e-16 ***
## PC1          0.030802   0.002165   14.229 < 2e-16 ***
## PC2          0.287542   0.003085   93.208 < 2e-16 ***
## PC3          0.310894   0.003701   84.005 < 2e-16 ***
## PC4         -0.023537   0.003976   -5.920 3.56e-09 ***
## PC5         -0.578237   0.004568 -126.596 < 2e-16 ***
## PC6         -0.168639   0.005030  -33.527 < 2e-16 ***
## PC7          0.101148   0.005279   19.161 < 2e-16 ***
## PC8          0.201552   0.005744   35.086 < 2e-16 ***
## PC9         -0.026557   0.006088   -4.362 1.33e-05 ***
## PC10         -0.196808   0.006763  -29.099 < 2e-16 ***
## PC11          0.056181   0.007182    7.822 6.94e-15 ***
## PC12          0.030186   0.007934    3.805 0.000145 ***
## PC13          0.101358   0.008839   11.467 < 2e-16 ***
## PC14         -0.016569   0.010043   -1.650 0.099103 .
## PC15          0.068952   0.010493    6.571 5.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3263 on 3260 degrees of freedom
## Multiple R-squared:  0.9166, Adjusted R-squared:  0.9163
## F-statistic: 2390 on 15 and 3260 DF, p-value: < 2.2e-16

```

The results show that the adjusted R-squared is now 0.92, a big improvement from a linear regression without pca. They also show that after 15 principal components, 95% of the variance in the data has already been accounted for, so the dimensions can definately be reduced. Below we check how the linear regression model after reducing the dimensions using pca performs on new data by checking the RMSE value returned.

```
## [1] 0.3232659
```

The RMSE value for the liner regression model after first performing pca is now 0.32, almost half of that without pca.

Conclusion and further work

The linear regression model build here fits the data fairly well and predicts the target variable on new data with fairly good accuracy too. It makes an error of about 0.32 units when predicting the silica content in the concentrate. However, it is unknown whether or not this error is acceptable for this specific plant, because no further information is given regarding this flotation process. What margin of error can be acceptable for them? And what are the implications of making an error? More information is needed to better understand the process and this could be achieved by talking with the people concerned, about the process. With more information available then maybe more algorithms may be deployed to build a model with even better prediction accuracy.