Dataset

Suresh Ooty

Table of Contents

ABC Dataset

Multiple sensors installed at machines and locations of plant convey time critical data to cloud based database. The sensor data are broadly classified at attribute level as "Physical" and "Virtual" (which is a calculated one from other attributes). Sensors capture every possible physical data such as temperature, volume, pressure, viscosity, volume, etc., In large, there are about 11M records of data with 33M metrics. The dataset chosen is specific to a plant over a period of time (year 2015)

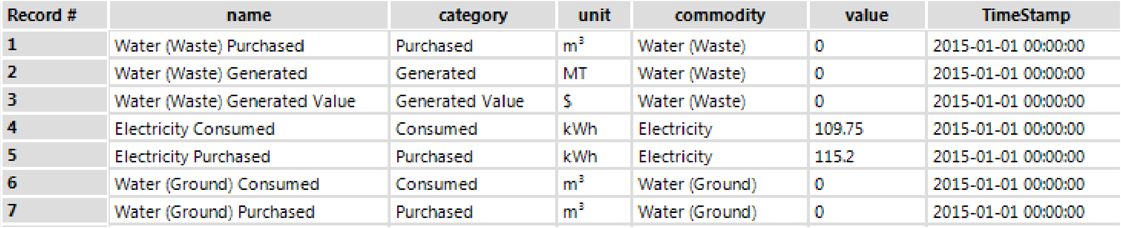
Attributes	Unit
Compressed Air Consumed	m^3
Compressed Air Generated	MT
Electricity Consumed	kWh
Electricity Purchased	kWh
Natural Gas Consumed	m^3
Natural Gas Converted	m^3
Natural Gas Purchased	m^3
Production	MT
Steam Consumed	kg
Steam Generated	MT
Steam Generated Value	$

Water (Ground) Consumed	m^3
Water (Ground) Purchased	m^3
Water (Municipal) Consumed	m^3
Water (Municipal) Purchased	m^3
Water (Waste) Generated	MT
Water (Waste) Generated Value	$
Water (Waste) Purchased	m^3
Hay Out Waste	MT

Data Structure

The original dataset was a raw dump from the database, which has records for multiple years in a linear order as in the image below. For a focused study, the records belong 2015 were chosen. Further, this dataset was filled up with the attribute *"Hay_out_waste"*. An *Alteryx* workflow that joins the records based on *TimeStamp* attribute was used to achieve this. As a final step, the complete dataset was transposed against the *TimeStamp* so that each record contains all the attributes captured in every 15 minutes.

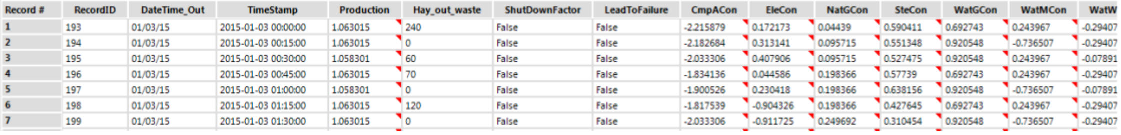
Original dataset



Record #	name	category	unit	commodity	value	TimeStamp
1	Water (Waste) Purchased	Purchased	m^3	Water (Waste)	0	2015-01-01 00:00:00
2	Water (Waste) Generated	Generated	MT	Water (Waste)	0	2015-01-01 00:00:00
3	Water (Waste) Generated Value	Generated Value	$	Water (Waste)	0	2015-01-01 00:00:00
4	Electricity Consumed	Consumed	kWh	Electricity	109.75	2015-01-01 00:00:00
5	Electricity Purchased	Purchased	kWh	Electricity	115.2	2015-01-01 00:00:00
6	Water (Ground) Consumed	Consumed	m^3	Water (Ground)	0	2015-01-01 00:00:00
7	Water (Ground) Purchased	Purchased	m^3	Water (Ground)	0	2015-01-01 00:00:00

Note*:Linear Order dataset*

Transposed dataset



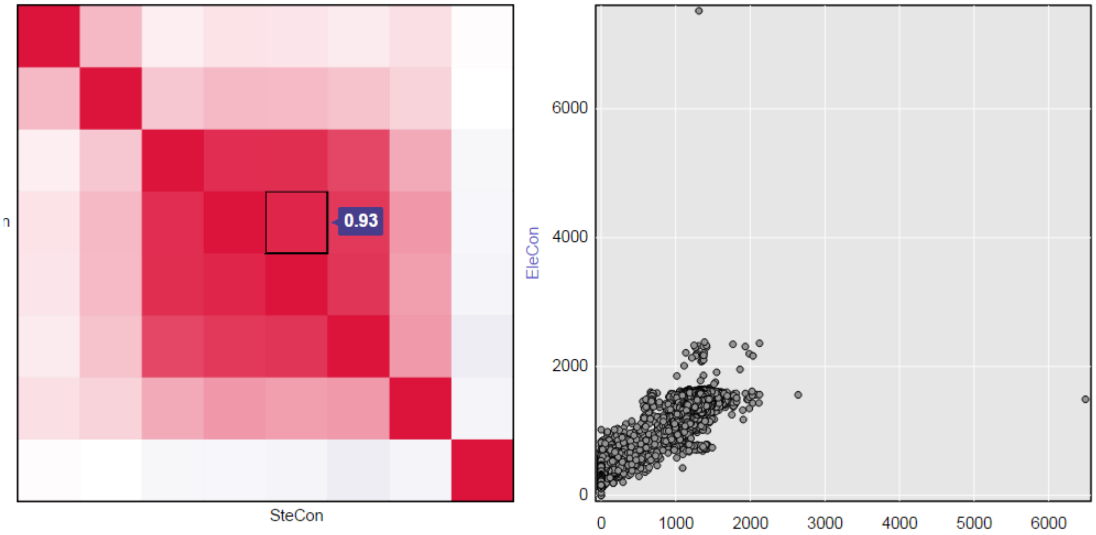
Record #	RecordID	DateTime_Out	TimeStamp	Production	Hay_out_waste	ShutDownFactor	LeadToFailure	CmpACon	EleCon	NatGCon	SteCon	WatGCon	WatMCon	WatW
1	193	01/03/15	2015-01-03 00:00:00	1.063015	240	False	False	-2.215879	0.172173	0.04439	0.590411	0.692743	0.243967	-0.29407
2	194	01/03/15	2015-01-03 00:15:00	1.063015	0	False	False	-2.182684	0.313141	0.095715	0.551348	0.920548	-0.736507	-0.29407
3	195	01/03/15	2015-01-03 00:30:00	1.058301	60	False	False	-2.033306	0.407906	0.095715	0.527475	0.920548	0.243967	-0.07891
4	196	01/03/15	2015-01-03 00:45:00	1.063015	70	False	False	-1.834136	0.044586	0.198366	0.57739	0.692743	0.243967	-0.29407
5	197	01/03/15	2015-01-03 01:00:00	1.058301	0	False	False	-1.900526	0.230418	0.198366	0.638156	0.920548	-0.736507	-0.07891
6	198	01/03/15	2015-01-03 01:15:00	1.063015	120	False	False	-1.817539	-0.904326	0.198366	0.427645	0.692743	0.243967	-0.29407
7	199	01/03/15	2015-01-03 01:30:00	1.063015	0	False	False	-2.033306	-0.911725	0.249692	0.310454	0.920548	-0.736507	-0.29407

Note*:Transposed dataset with 'Hay_out_waste' attribute*

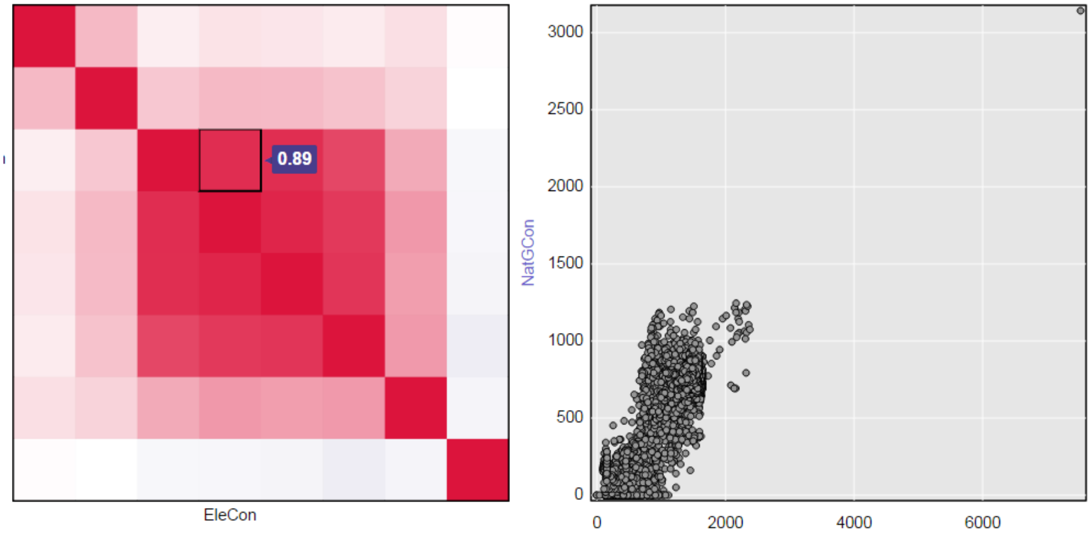
Initial Outliers & Correlation

The outliers have been identified easily with the help of **Error! Hyperlink reference not valid.** and the images captured are shown here. Appropriate decisions were made to drop these records from the dataset. The bivariate correlations between the predictors are not very clear with the presence of outliers

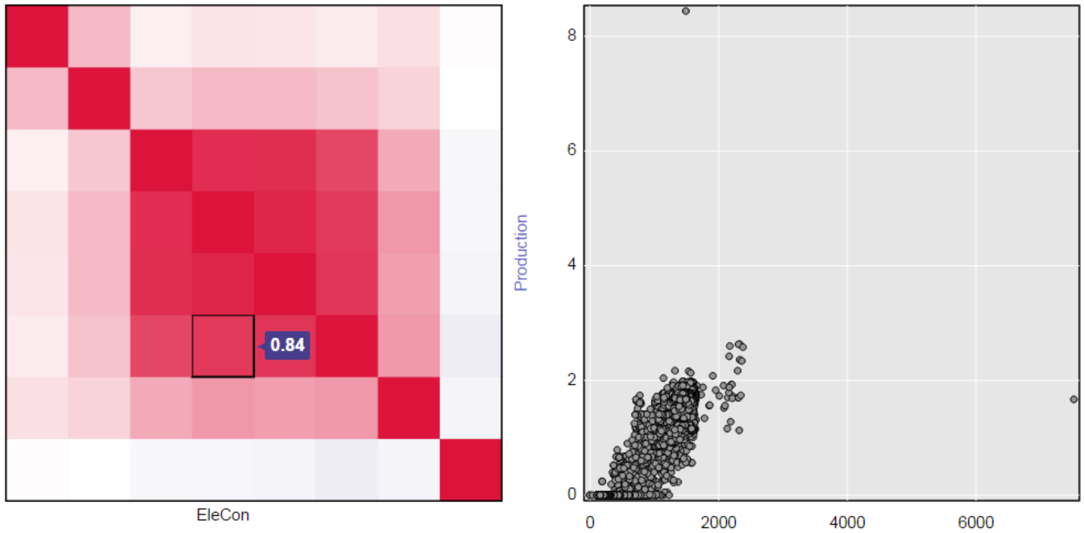
Electricity & Steam



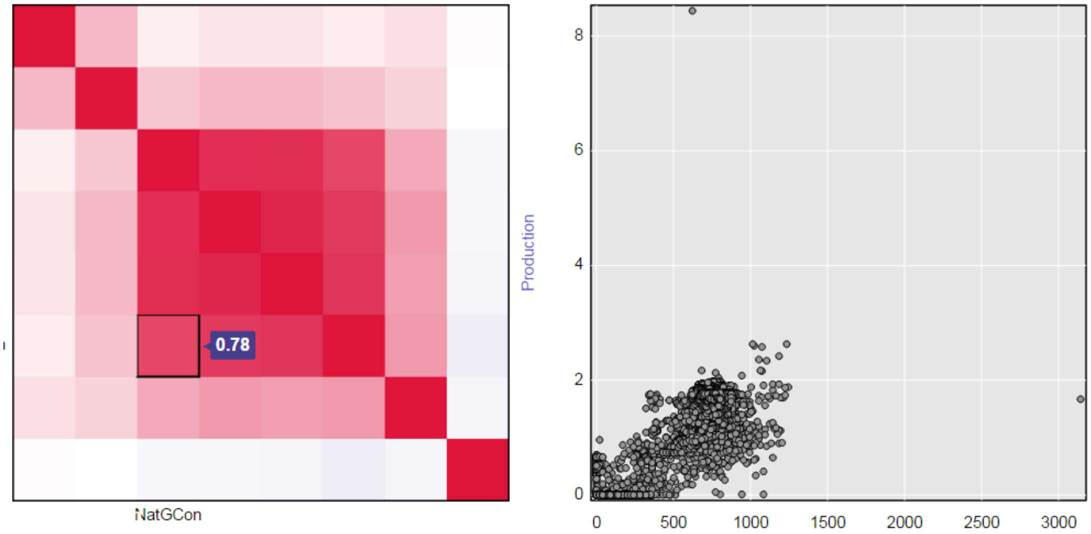
Natural Gas & Steam



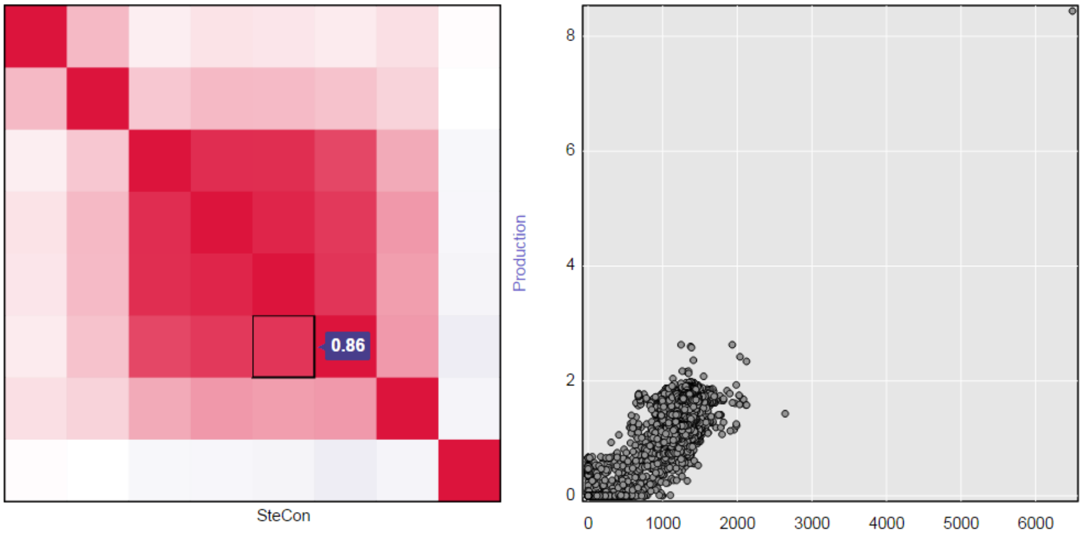
Production & Electricity



Production & Natural Gas



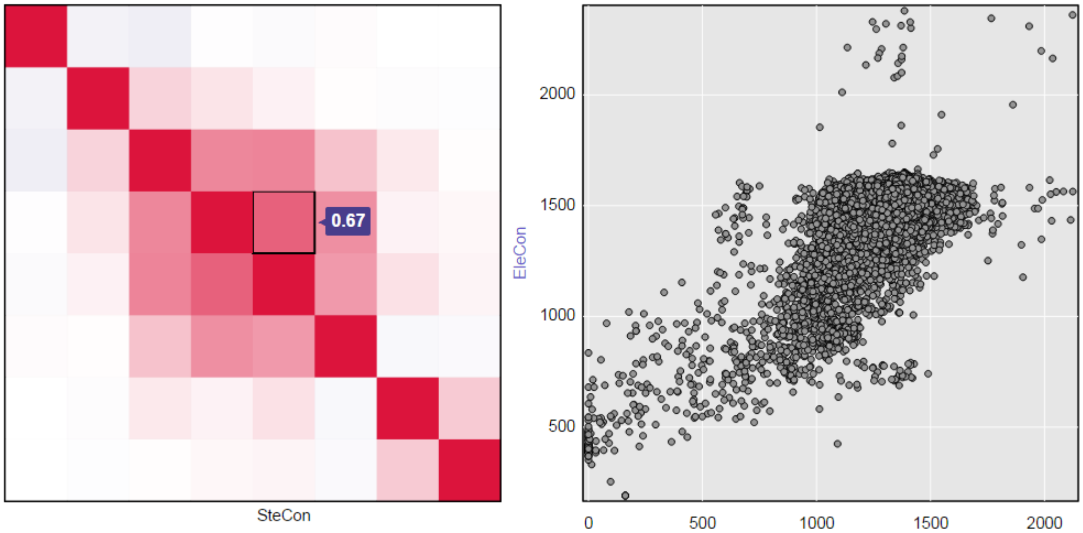
Production & Steam



After removing few Outliers

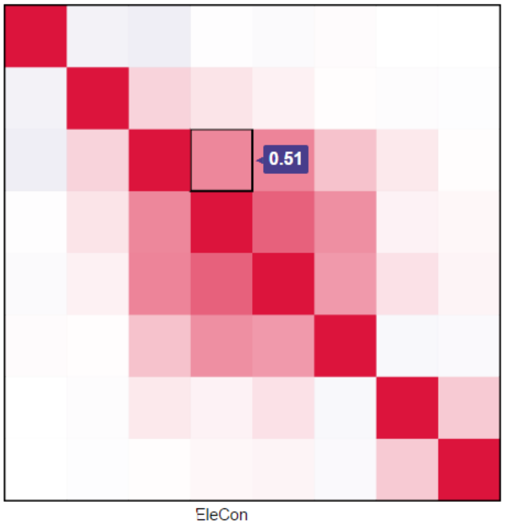
These OUTLIERS were removed using algorithm " [SteCon] < 2500 && [EleCon] < 6000 && [NatGCon] < 3000 && [Production] > 0 " and when the correlation plots were made.

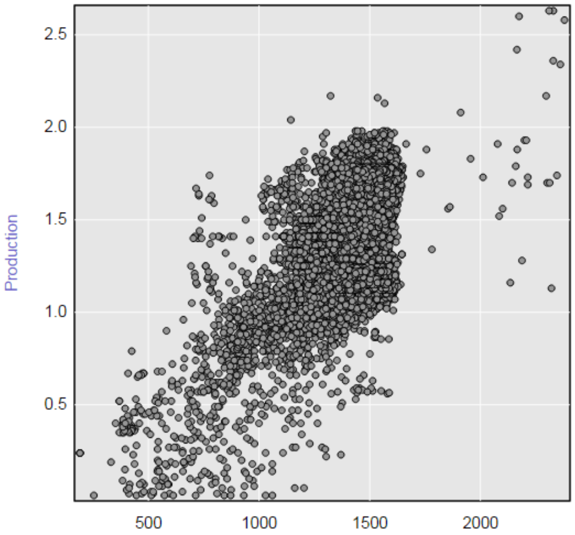
Electricity & Steam



Note:*One of the highest correlated variables @ 0.67*

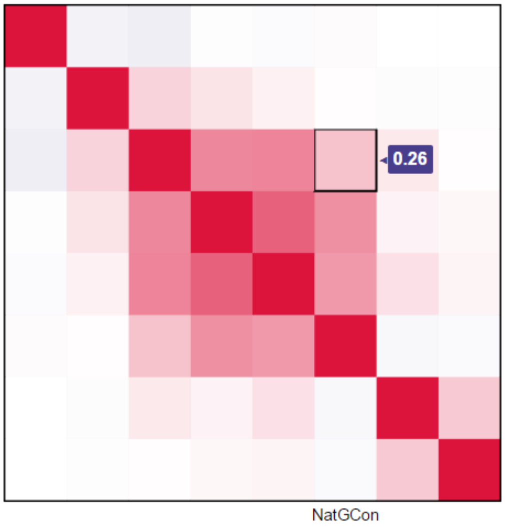
Production & Electricity

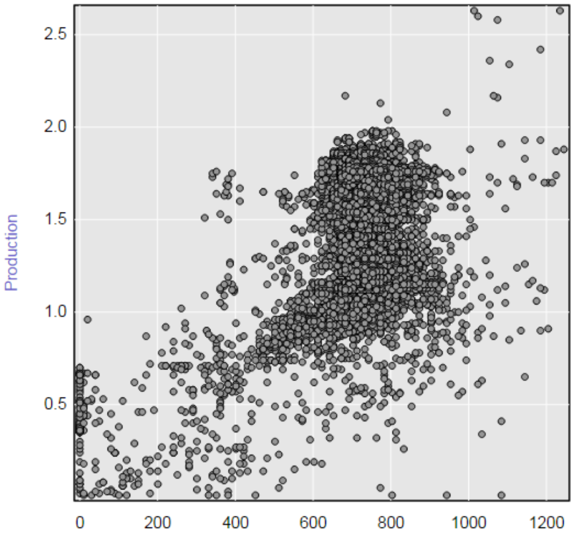




Note:*seems highly correlated @ 0.51, but there could be some outliers at high EleCon values*

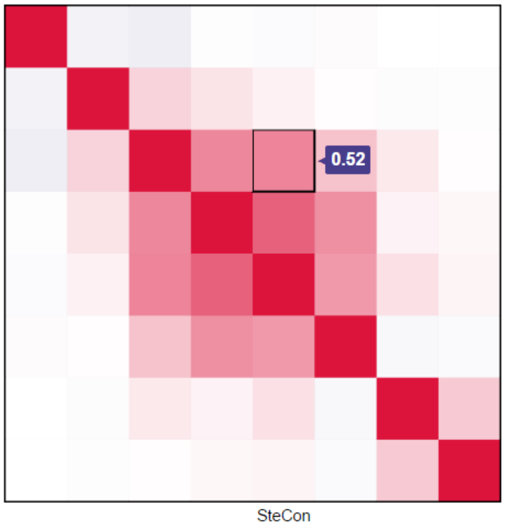
Production & Natural Gas

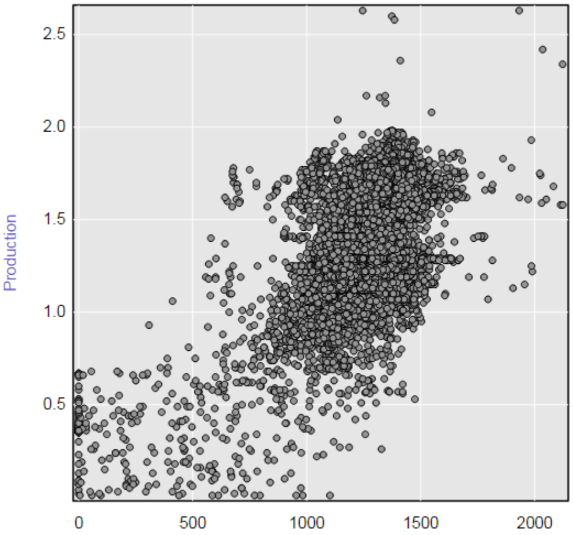




Note:*Not so much correlation between these two variable*

Production & Steam





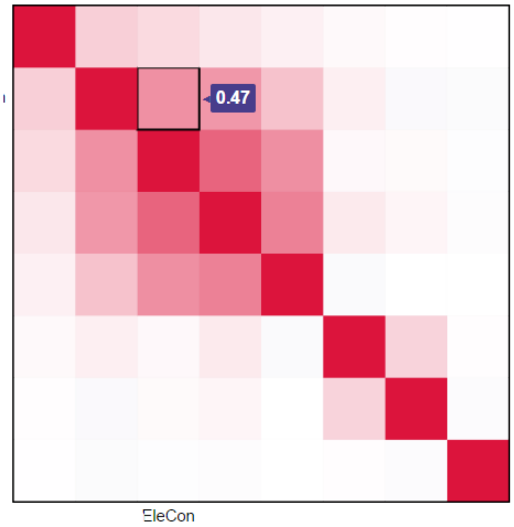
Note:*It is understandable that Steam consumption is required for Production, Noted for further study*

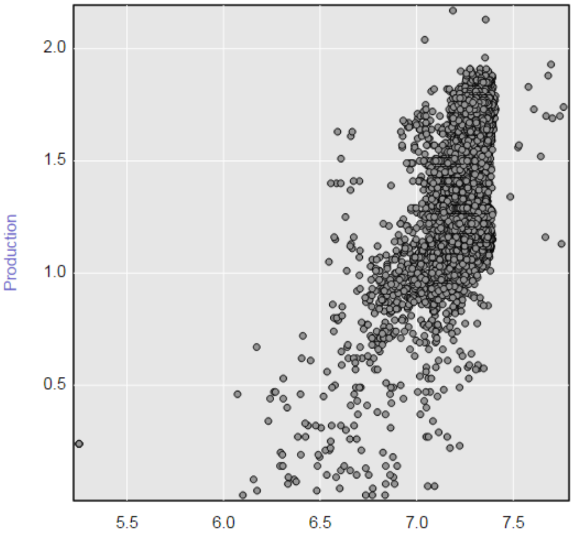
LOG transformation

Further, tuning the algorithm to " [SteCon] < 2500 && [EleCon] < 6000 && [NatGCon] < 3000 && [Production] < 2.5 && [Production] > 0" , as the production above 2.5 MT is lone observation. And the independent variables are transformed using LOG function. The correlation effect has been reduced drastically due to this transformation.

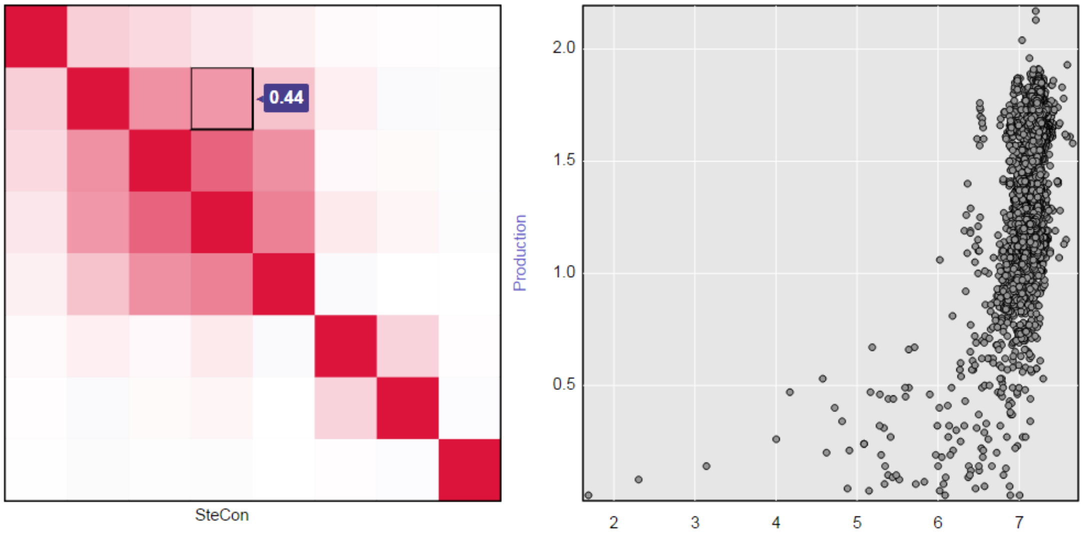
But this specific set does not have the '0' values of 'Production', which is necessary for analysis. However, these views were made to see the correlation.

Production & Electricity

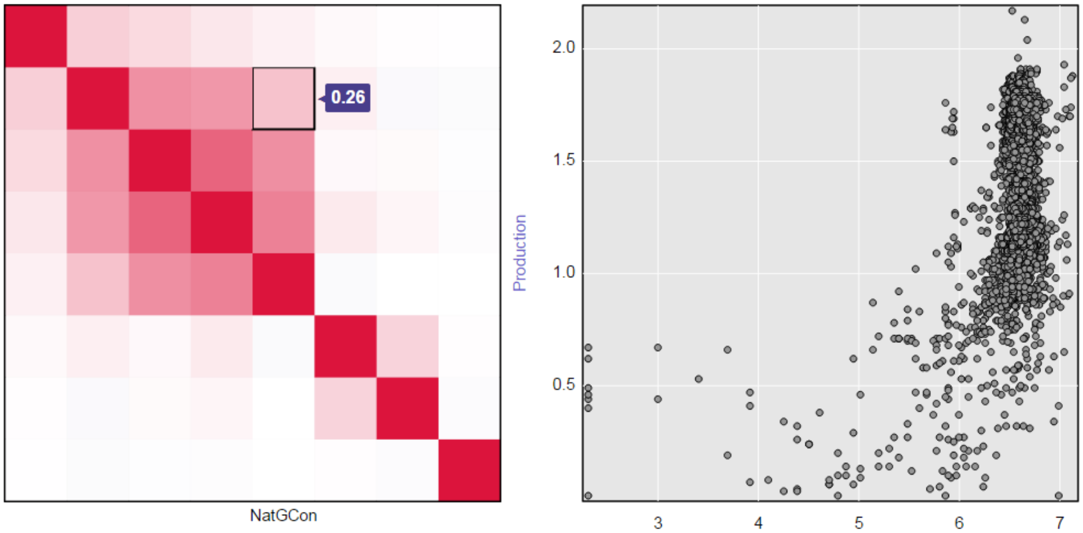




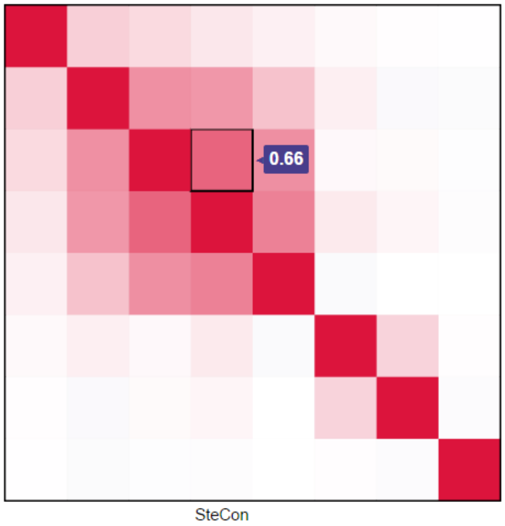
Production & Steam

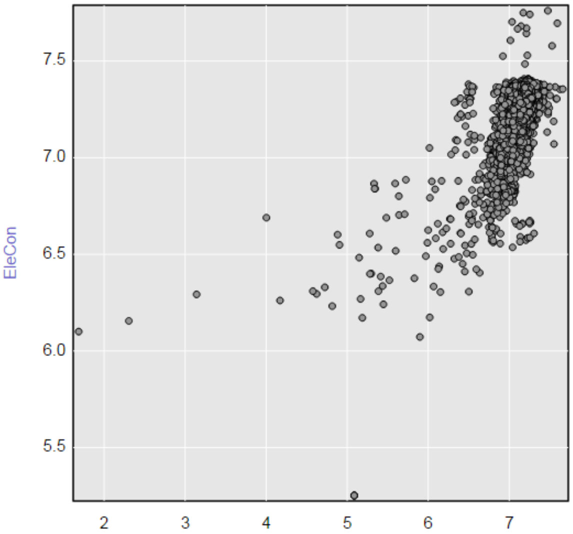


Production & Natural Gas



Electricity & Steam

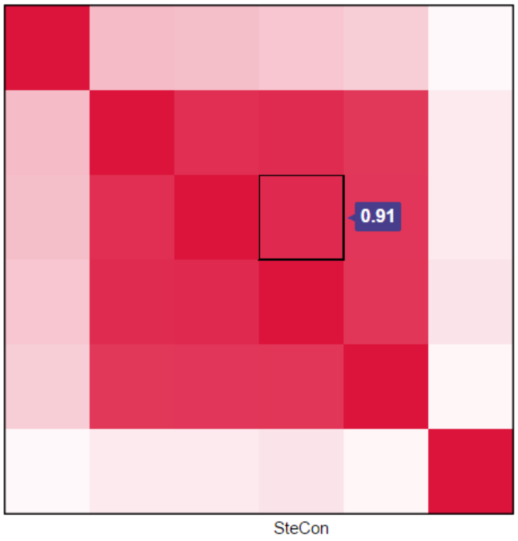


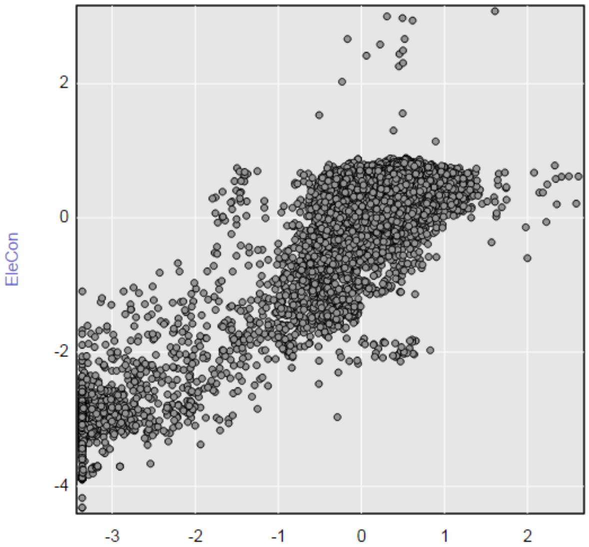


Min-Max Standardization

When the variables are standardized using min-max method and the production variable is transformed using SQRT, the following observations were made. Outliers were removed using [SteCon] < 2500 && [EleCon] < 6000 && [NatGCon] < 3000 && [Production] < 1.8 . This indicate that this transformation method does not remove the high bivariate correlations between variables.

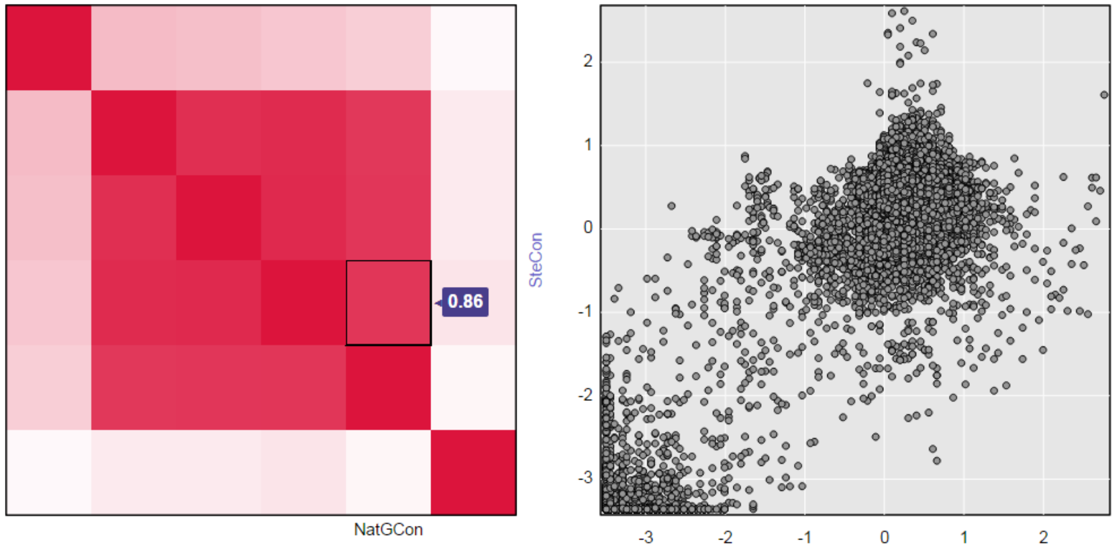
Electricity & Steam





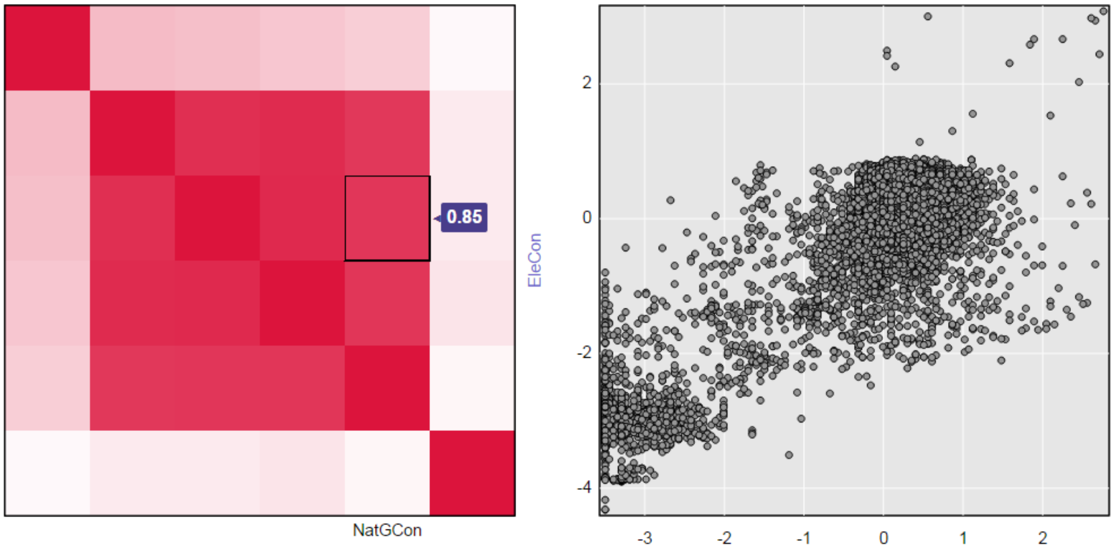
***Note**:High Correlation @0.91*

Steam & Natural Gas



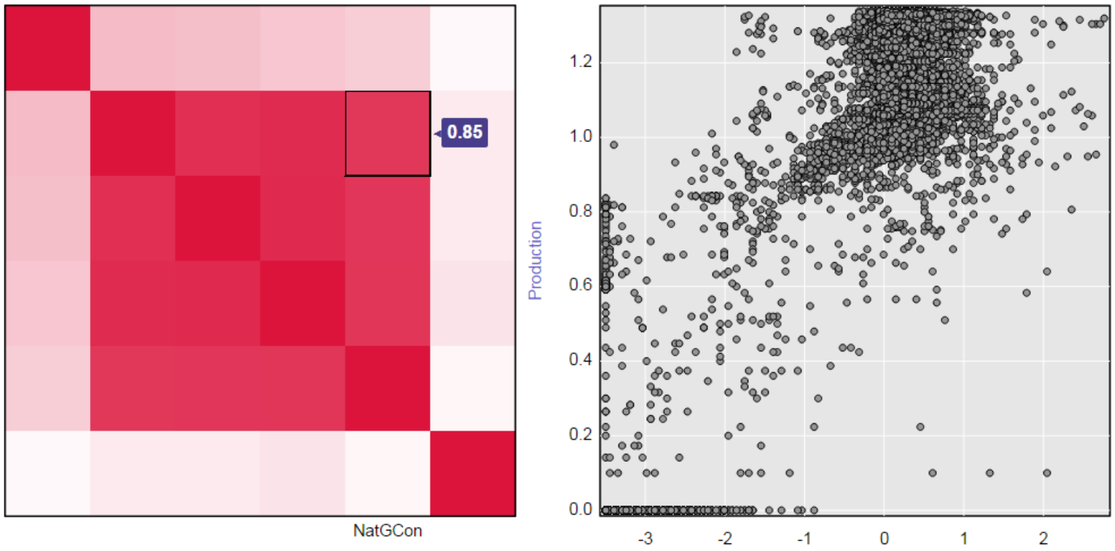
Note:*Multiple values of Steam at same position indicates influence of other variables*

Electricity & Natural Gas



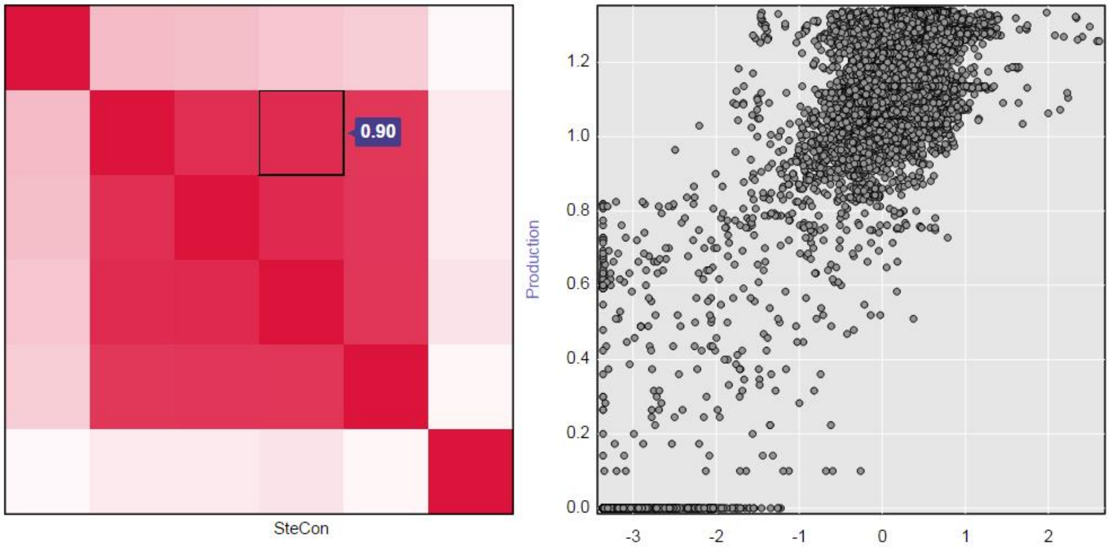
Note:*Multiple levels of Electricity consumption at 0 Natural Gas indicates that Electricity is used on some other attribute*

Production & Natural Gas



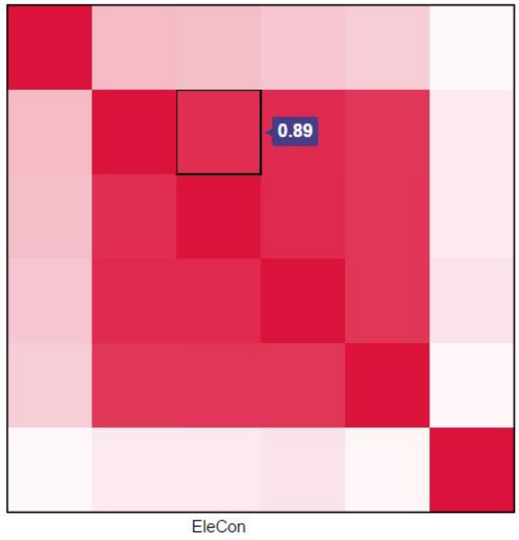
Note*:Moderate Production at very low Natural Gas consumption indicates that Natural Gas is not consumed occasionally, or it is involved in ealier stages of Production*

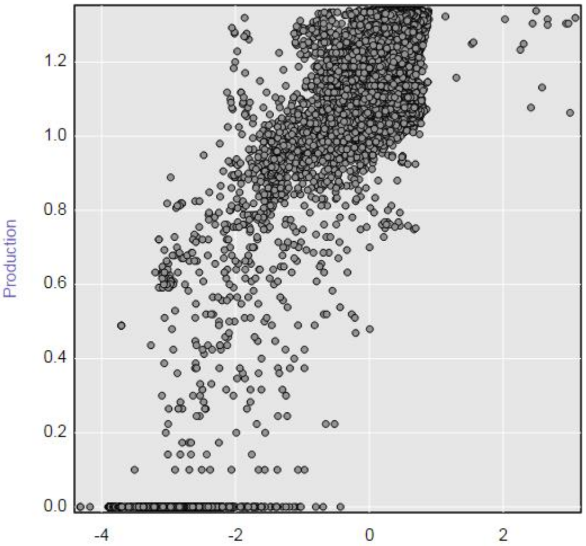
Production & Steam



Note*:Zero production at different values of Steam indicate that Steam is stored or used at different levels of paper manufacturing*

Production & Electricity





Note*:Electricity is always consumed, few 0,0 values indicate Plant shut downs*

Conclusions on Dataset

From the overall perspective both the transformations appear good at high level. However, a choice between Min-Max and LOG transformed dataset need to be made at the time of research. LOG transformation obviously scales down the values to similar scales.

- For identifying pattern appropriate slice and dice of dataset need to be chosen for research, as different type of subset of data would give different results.
- The correlation identified need to be applied while building the model. For example, on WatMCon is highly correlated to other water attributes.
- This is a valid Time-Series data in a real time production unit. Specific seasonality, trends, sensor errors, human errors are unavoidable. Need to be closely looked for each level of study.