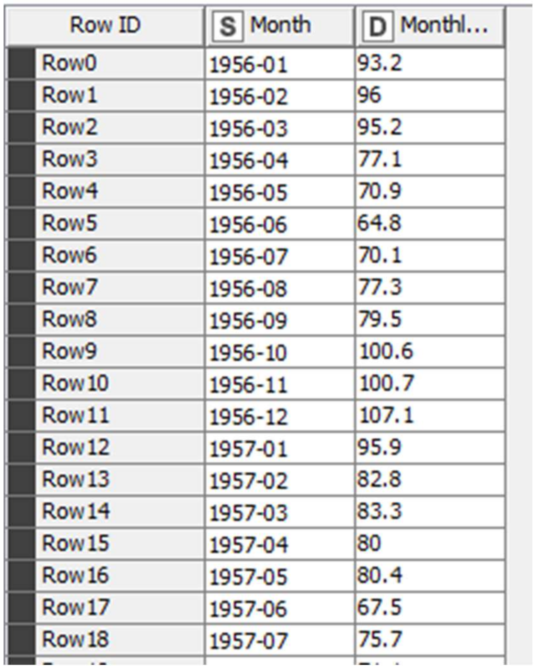
Beer Production

Data understanding

We get monthly beer production data. The data has 476 rows and 2 column (date and sales). We can do prediction regression for this data.

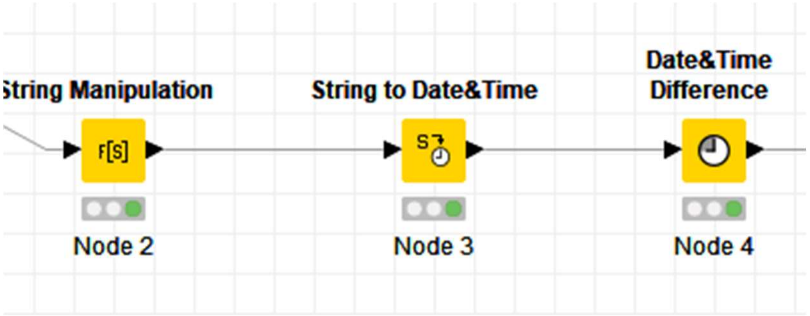


Row ID	S Month	D Monthl...
Row0	1956-01	93.2
Row1	1956-02	96
Row2	1956-03	95.2
Row3	1956-04	77.1
Row4	1956-05	70.9
Row5	1956-06	64.8
Row6	1956-07	70.1
Row7	1956-08	77.3
Row8	1956-09	79.5
Row9	1956-10	100.6
Row10	1956-11	100.7
Row11	1956-12	107.1
Row12	1957-01	95.9
Row13	1957-02	82.8
Row14	1957-03	83.3
Row15	1957-04	80
Row16	1957-05	80.4
Row17	1957-06	67.5
Row18	1957-07	75.7

Data Preparation

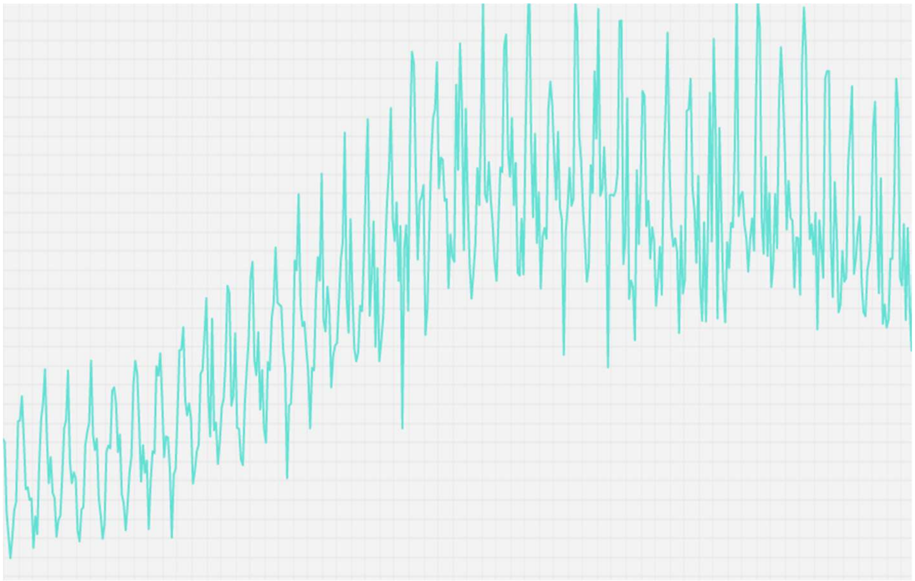
Cleaning data Format

In here we can see date string atribute and change it to normal date format and numerical date format

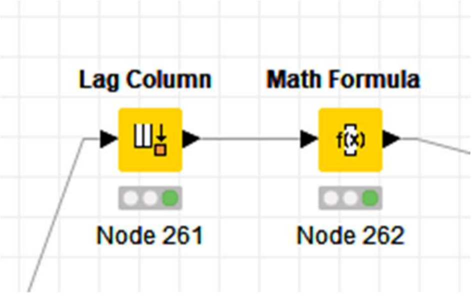


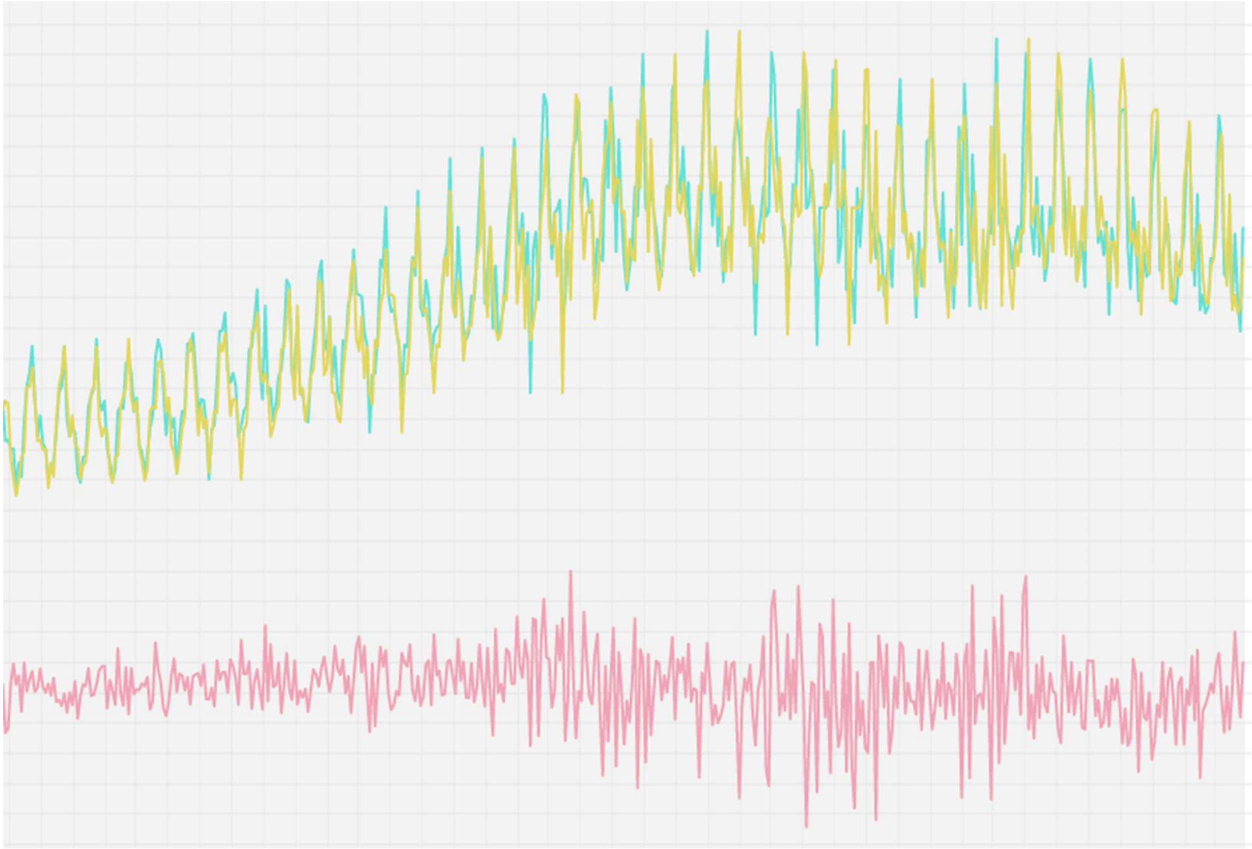
Removing Seasonality

For regression model to get good result, we need to set the data stationary and remove any unused on the data.



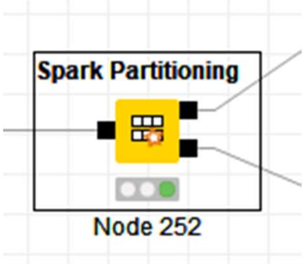
As we can see the pattern of the data repeats every year, so we need to preprocess it.





The red one is the cleaned and scale down data.

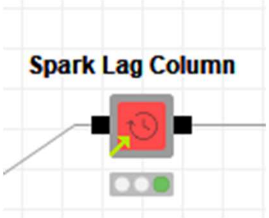
Split training and Test

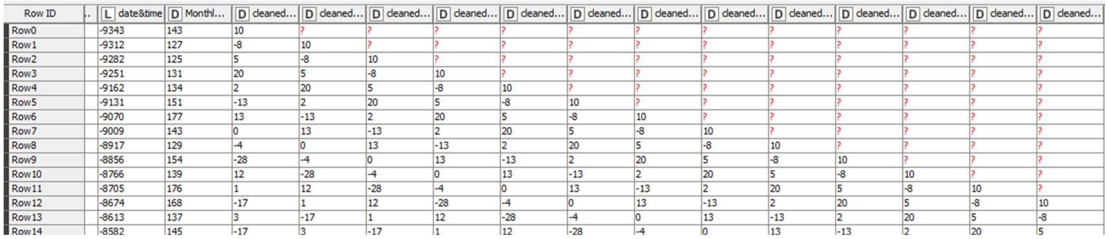


We used spark partitioning to split the data into training and testing data.

Modeling

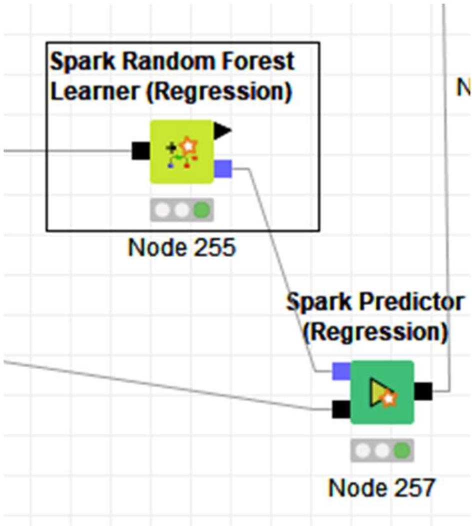
We will use spark Lag to make set of feature from past data.





Row ID	..	L date&time	D Monthl...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...	D cleaned...
Row0		-9343	143	10	?	?	?	?	?	?	?	?	?	?	?	?
Row1		-9312	127	-8	10	?	?	?	?	?	?	?	?	?	?	?
Row2		-9282	125	5	-8	10	?	?	?	?	?	?	?	?	?	?
Row3		-9251	131	20	5	-8	10	?	?	?	?	?	?	?	?	?
Row4		-9162	134	2	20	5	-8	10	?	?	?	?	?	?	?	?
Row5		-9131	151	-13	2	20	5	-8	10	?	?	?	?	?	?	?
Row6		-9070	177	13	-13	2	20	5	-8	10	?	?	?	?	?	?
Row7		-9009	143	0	13	-13	2	20	5	-8	10	?	?	?	?	?
Row8		-8917	129	-4	0	13	-13	2	20	5	-8	10	?	?	?	?
Row9		-8856	154	-28	-4	0	13	-13	2	20	5	-8	10	?	?	?
Row10		-8766	139	12	-28	-4	0	13	-13	2	20	5	-8	10	?	?
Row11		-8705	176	1	12	-28	-4	0	13	-13	2	20	5	-8	10	?
Row12		-8674	168	-17	1	12	-28	-4	0	13	-13	2	20	5	-8	10
Row13		-8613	137	3	-17	1	12	-28	-4	0	13	-13	2	20	5	-8
Row14		-8582	145	-17	3	-17	1	12	-28	-4	0	13	-13	2	20	5

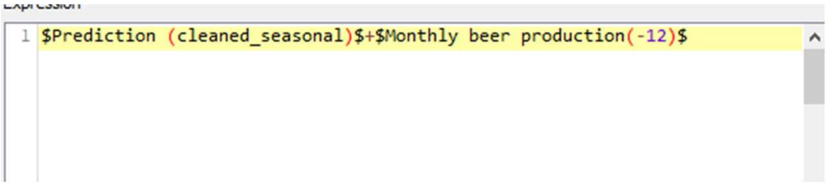
Past data added into new column. We use last 12 month for data features


Spark Random Forest Learner (Regression)
Node 255
N
Spark Predictor (Regression)
Node 257

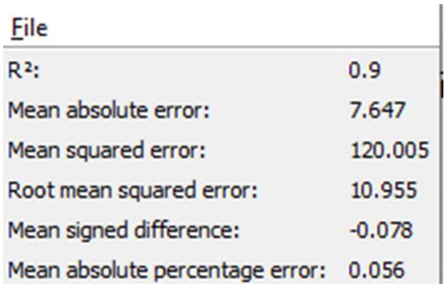
We used random forest learner to train the model.

Evaluation

Before we evaluate the data, we need to restore the cleaned data into a normal one first.



```
$Prediction (cleaned_seasonal)$+$Monthly beer production(-12)$
```



```
R²:                              0.9
Mean absolute error:             7.647
Mean squared error:              120.005
Root mean squared error:         10.955
Mean signed difference:          -0.078
Mean absolute percentage error:  0.056
```