Shampoo Sales

Data understanding

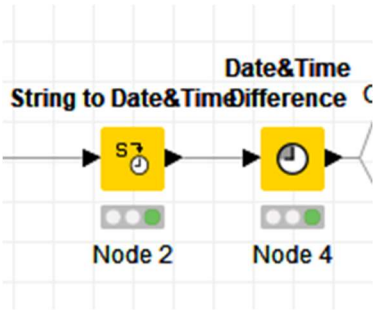
we get monthly shampoo sales. The data has 36 row and 2 column (date and sales). We can do prediction regression for this data.

Row ID	S Month	D Sales o...
Row0	1/1/2020	266
Row1	2/1/2020	145.9
Row2	3/1/2020	183.1
Row3	4/1/2020	119.3
Row4	5/1/2020	180.3
Row5	6/1/2020	168.5
Row6	7/1/2020	231.8
Row7	8/1/2020	224.5
Row8	9/1/2020	192.8
Row9	10/1/2020	122.9
Row10	11/1/2020	336.5
Row11	12/1/2020	185.9
Row12	1/2/2020	194.3
Row13	2/2/2020	149.5
Row14	3/2/2020	210.1
Row15	4/2/2020	273.3
Row16	5/2/2020	191.4
Row17	6/2/2020	287
Row18	7/2/2020	226

Data Preparation

Cleaning data Format

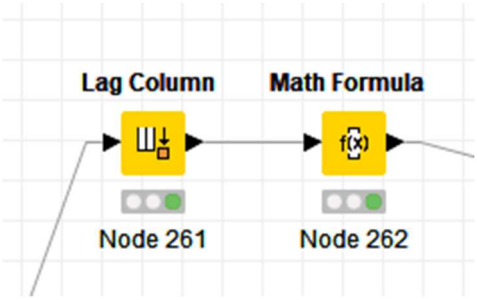
In here we can see date string atribute and change it to normal date format and numerical date format



Creating Stasionary

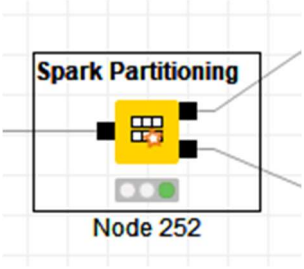
For regression model to get good result, we need to set the data stationary and remove any unused on the data.

we can look at the pattern of the data repeat every year, so we need to preprocess it.



The red one is the cleaned and scale down data.

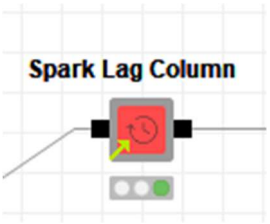
Split training and Testing



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

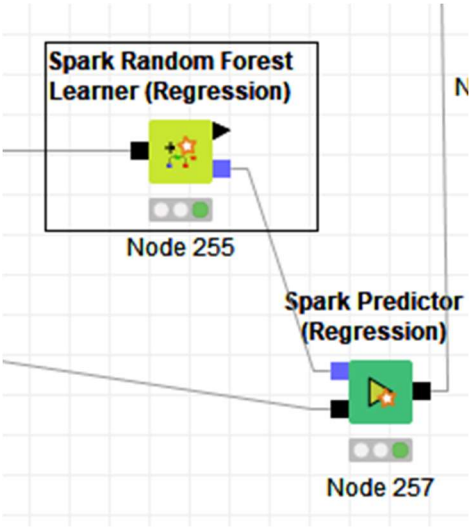
Modeling

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



Month	Sales o...	date&time	Sales o...	stationary	station...	station...	station...	station...	station...	station...	station...	station...	station...	station...	station...	D
20-12-03	646.9	-18599	581.3	65.6	?	?	?	?	?	?	?	?	?	?	?	?
20-12-02	342.3	-18598	264.5	77.8	65.6	?	?	?	?	?	?	?	?	?	?	?
20-11-03	581.3	-18569	475.3	106	77.8	65.6	?	?	?	?	?	?	?	?	?	?
20-11-02	264.5	-18568	421.6	-157.1	106	77.8	65.6	?	?	?	?	?	?	?	?	?
20-10-02	421.6	-18537	289.9	131.7	-157.1	106	77.8	65.6	?	?	?	?	?	?	?	?
20-10-01	122.9	-18536	192.8	-69.9	131.7	-157.1	106	77.8	65.6	?	?	?	?	?	?	?
20-09-03	682	-18508	407.6	274.4	-69.9	131.7	-157.1	106	77.8	65.6	?	?	?	?	?	?
20-09-02	289.9	-18507	303.6	-13.7	274.4	-69.9	131.7	-157.1	106	77.8	65.6	?	?	?	?	?
20-08-02	303.6	-18476	226	77.6	-13.7	274.4	-69.9	131.7	-157.1	106	77.8	65.6	?	?	?	?
20-07-03	575.5	-18446	437.4	138.1	77.6	-13.7	274.4	-69.9	131.7	-157.1	106	77.8	65.6	?	?	?

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



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.

```
1 $Prediction (stationary)$+$Sales of shampoo over a three year period(-1)$
```

Statistics - ...

File

R²:	-0.123
Mean absolute error:	107.971
Mean squared error:	16,193.199
Root mean squared error:	127.252
Mean signed difference:	74.723