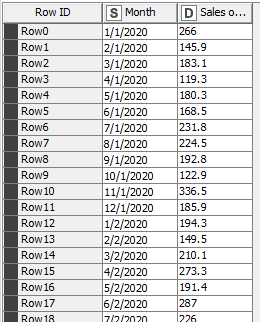
# Shampoo Sales

## Data understanding

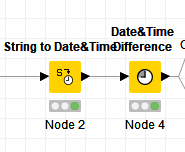
we get monthly shampoo sales. The data has 36 row and 2 column ( date and sales). We can do prediction regression for this data.



## Data Preparation

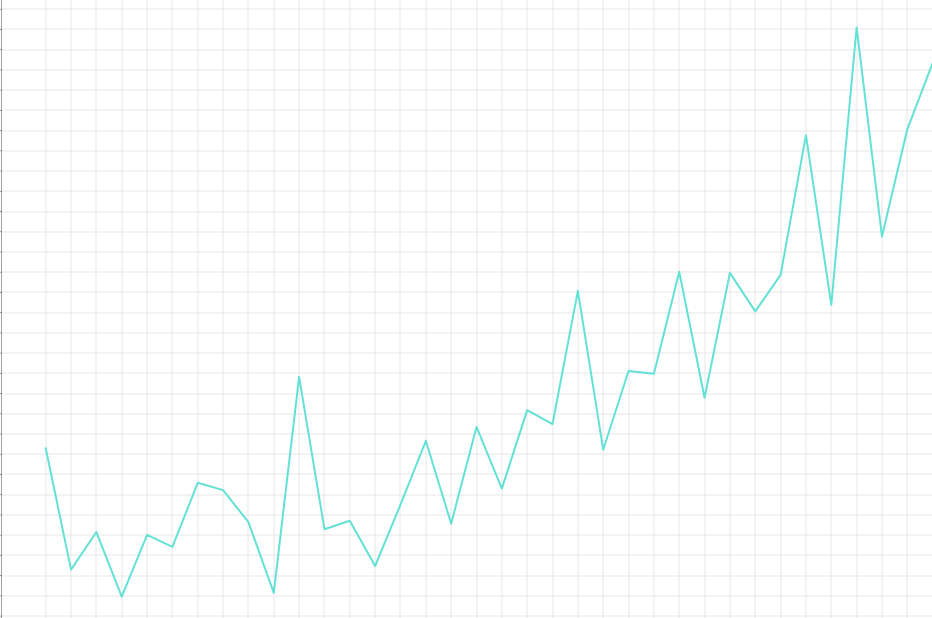
### Cleaning data Format

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

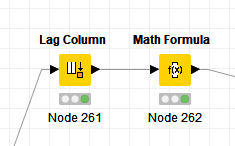


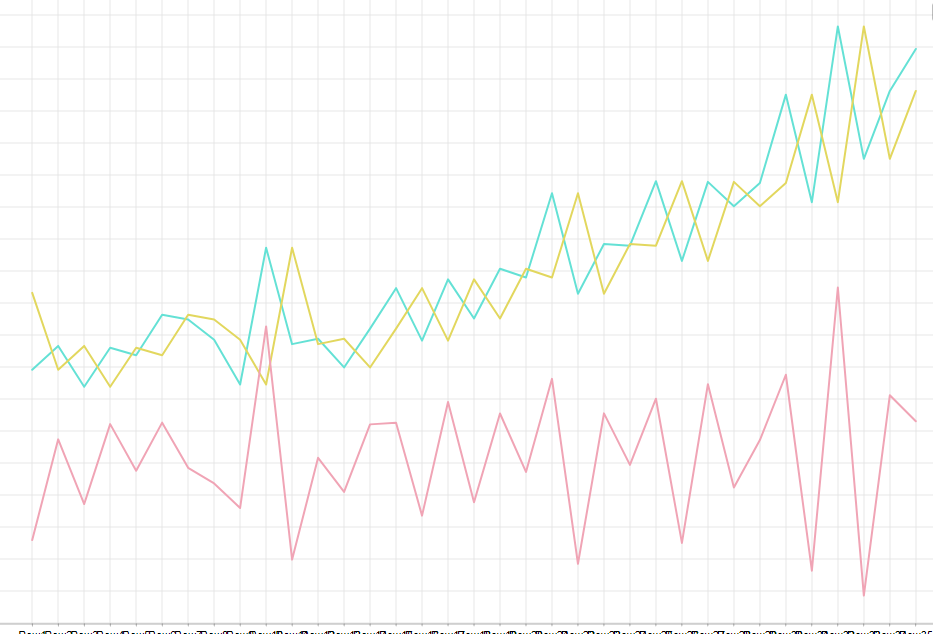
### Creating Stasionary

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



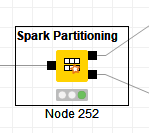
we can look at the pattern of the data repeat every year, so we need to preprocess it.





The red one is the cleaned and scale down data.

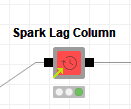
### Split training and Testing

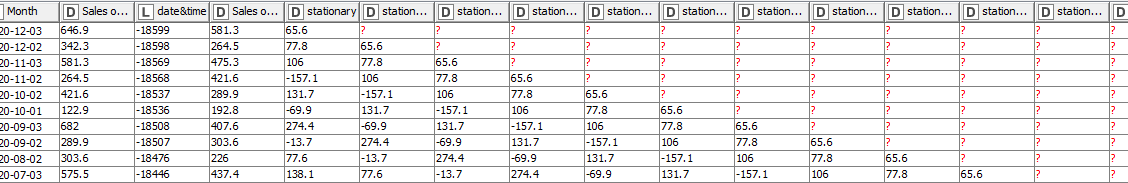


We used spark partitioning to split the data into training and testing data.

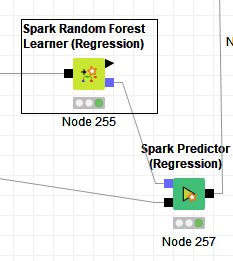
## Modeling

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





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.

