Retail Sales Forecasting

By Group 5 in section 2

SEONG KYOUNG           0725164

TEJAS MAHESH PATEL 0734912

SOORYA SURESH           0735168

RAJWINDER KAUR         0732989

YI ZHOU                         0730368

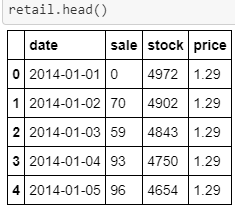
MOHSIN MEGHJANI         0730241

# Introduction

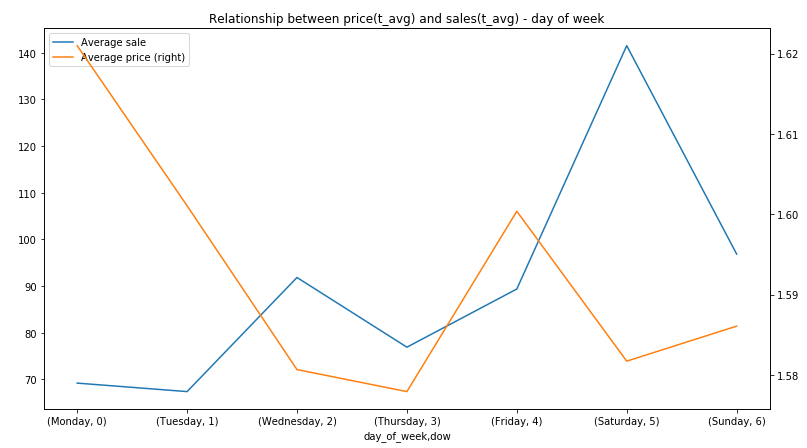
No matter where in the world, every retail business must face almost the same question: how much inventory should I carry? On one hand to mush inventory means working capital costs, operational costs and complex operation. On the other hand, lack of inventory leads to lost sales, unhappy customers and a damaged brand. So, choosing the appropriate method to calculate inventory, cost, which has a big benefit at customer and retailer.

# Related Work

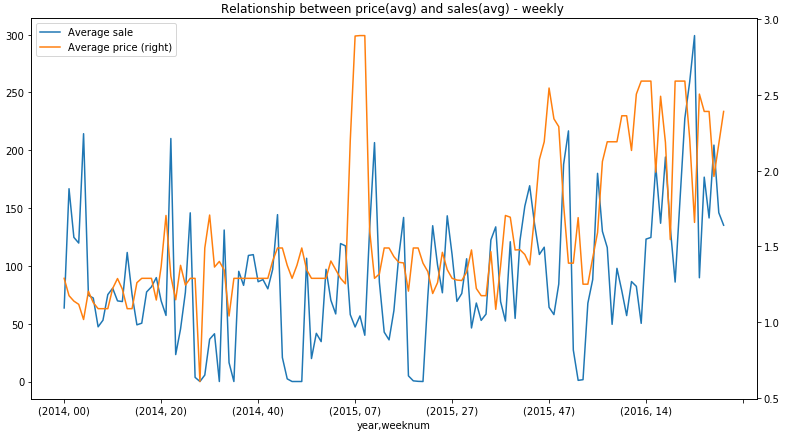
## Explorer Dataset



We taken this dataset from Kaggle. It was extracted from a Brazilian top retailer and has many SKUs and many stores. The data was transformed to protect the identity of the retailer, that including 4 columns date, sale, stock, and price from 2014 to 2016.

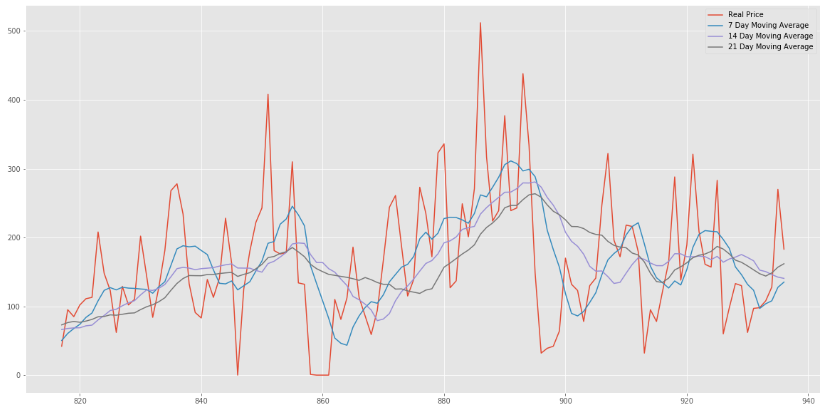


According to above dual weekly graph, we could find there is an inverse relationship between sales and price.



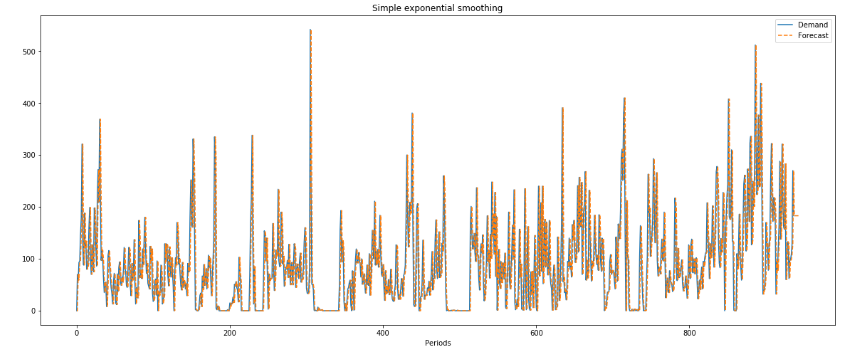
This graph clearly shows the relationship between price and sales group by yearly. That also has an inverse relationship.

## Moving Average



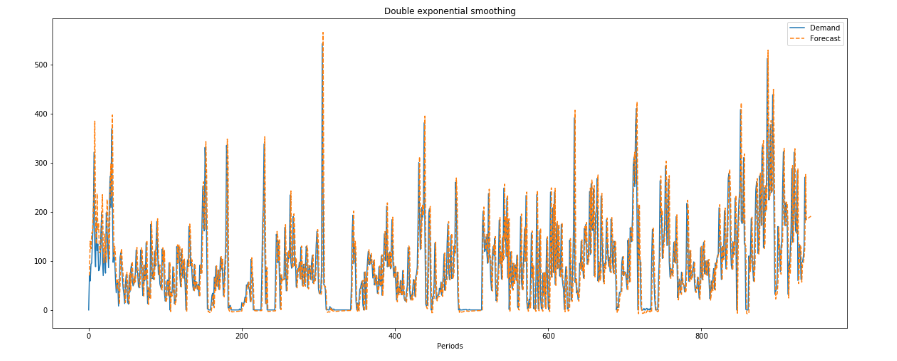
As considered that length of the data that we used 7,14, 21days. From the graph we can see there is no trend, but it has stationary.

## Simple Exponential Smoothing





## Double Exponential Smoothing





We are using dickey-fuller test in this model, by given alpha value which gave us MAE is 45.01, RMSE is 67.76 in single exponential smoothing. Otherwise, there are 46.76 MAE and 69.77 RMSE in double exponential smoothing.

## Machine Learning

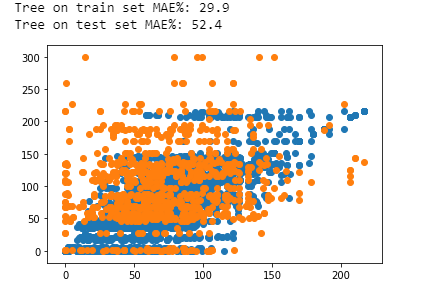
### Linear Regression



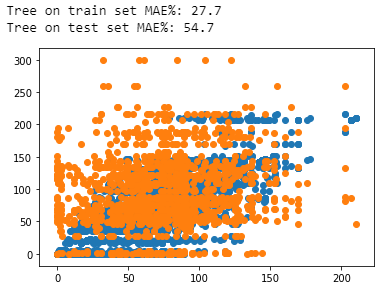
In linear regression, we got MAE 0, that means overfitting. We cannot use it.

### Decision Tree

* Split on 80-20



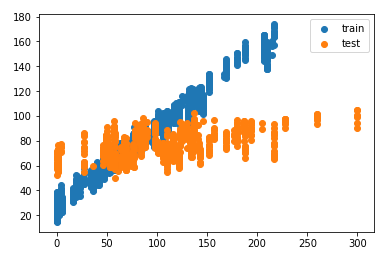
* Split on 70-30



When we set the max depth = 10, we got MAE is 29.9 on 80-20, but in 70-30, we got MAE 27.7.

### Random Forest

* Spilt on 80-20





* Split on 70-30

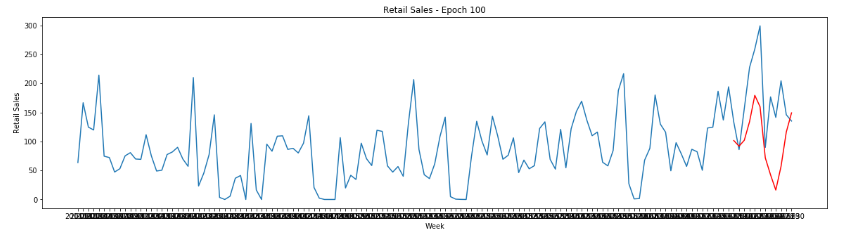


comparing those two different splitting models on random forest, as we expected, both, we got high MAE on testing set than training set. By contrast, there is low MAE on 80-20 model.

## LSTM

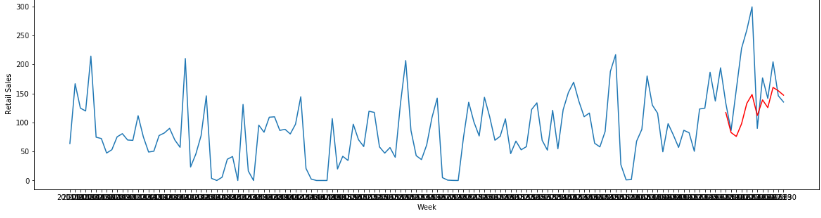
Long short-term memories are based on neural network. We use the past value to make future predictions, time as an important parameter, so we will need to split the data in time.

* Epoch 100



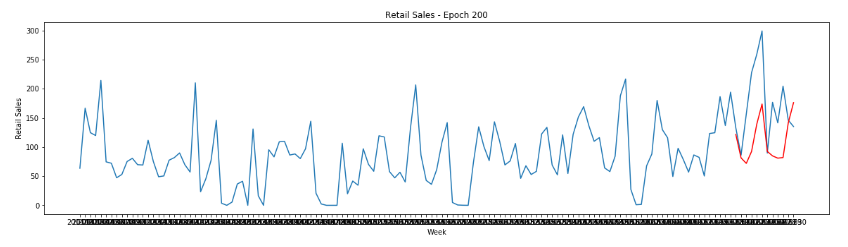
In epoch 100 the data is stationary and the prediction which is done on this model is under fitted as does not give good prediction the error is more between actual and predicted value

* Epoch 150



In epoch 150 the data is stationary and the prediction which is done on this model is more generalize as prediction is better and the error between actual and predicted value is less.

* Epoch 200

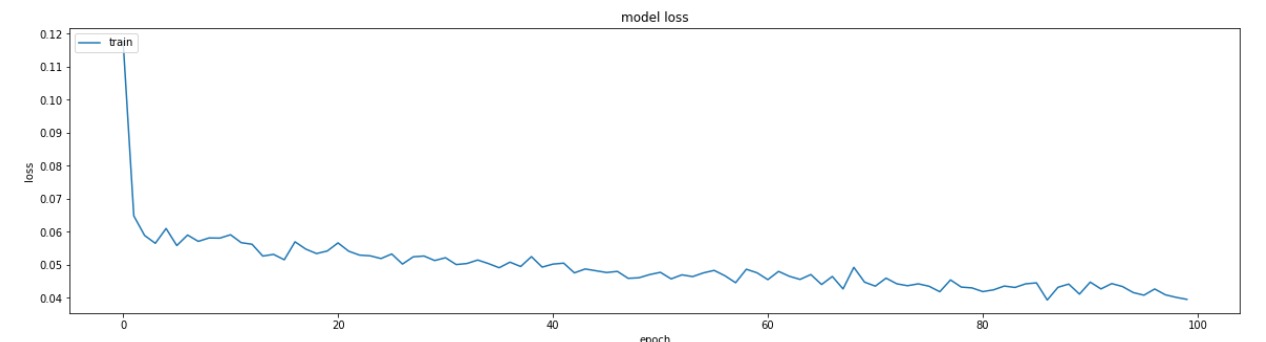


In epoch 200 the data is stationary and the prediction which is done on this model is somewhat generalize as it' is prediction is looking smoothed version of the actual result.

Model Loss:

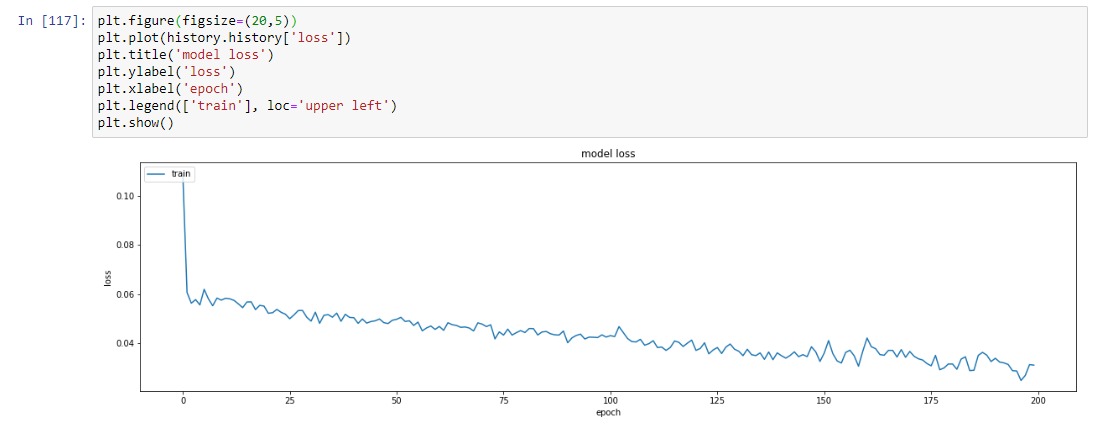
Epoch 100

this is the loss graph per epoch use for prediction as we can see the early epoch has high loss but as our model trains the loss starts to decrease



Epoch 200

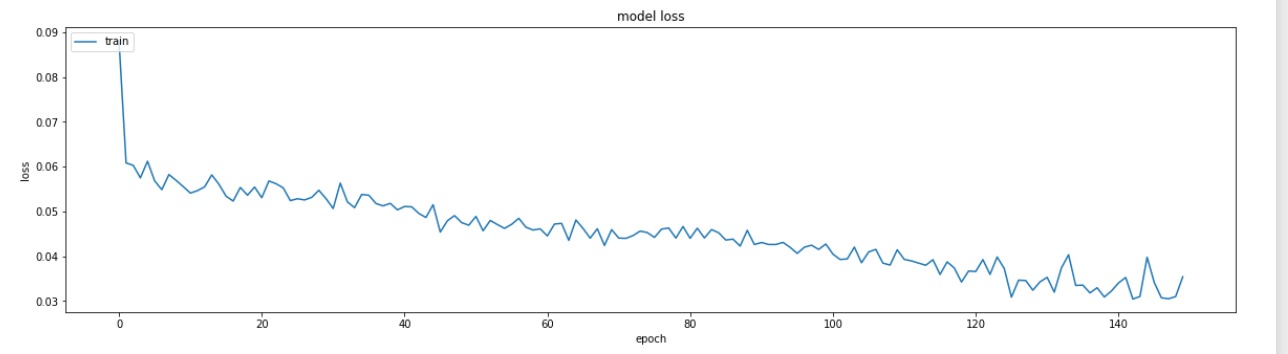
this is the loss graph per epoch use for prediction as we can see the early epoch has high loss but as our model trains the loss starts to decrease.



Epoch 150:

the loss graph per epoch use for prediction as we can see the early epoch has high loss but as our model trains the loss starts to decrease gradually.

After Epoch 150 we can see that it gradually Increases.



# Conclusion

As we are comparing all the methods, we implement when epoch equal to 150 in LSTM, we have the best prediction with low MAE. Linear regression assuming each point are independent, which means there is no relationship between time and sales, so we cannot use it in this case.

## Reference

1. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
2. <https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/>
3. <https://medium.com/@KodiakRating/can-machine-learning-shape-the-future-of-supply-chain-optimization-bdcb8e7ab49b>
4. <https://medium.com/analytics-vidhya/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d>