Store Item Demand Forecast

Ankita Bhanushali (M.Sc. Statistics)
Executive Analytics, Kantar - Worldpanel
INSAID, March’20 GCD Cohort

INTRODUCTION

Aim of the project is to predict item-level sales data for 3 months at different store locations.

Data description:

Serial No	Variable Name	Description	DataType
1	Date	Date of the sale data. There are no holiday effects or store closures.	Object
2	Store	Store ID	int64
3	Item	Item ID	int64
4	Sales	Number of items sold at a particular store on a particular date.	int64

TABLE OF CONTENTS

OBJECTIVE

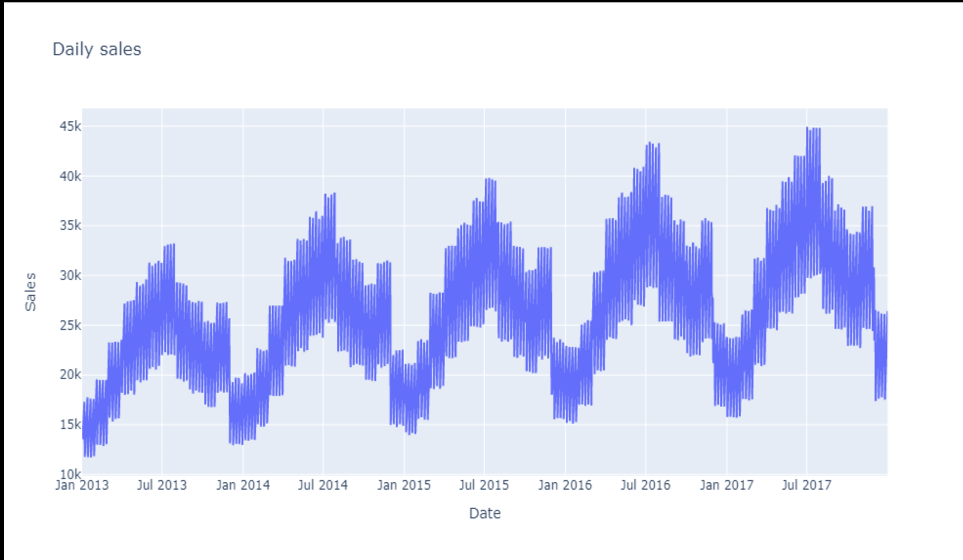
Build a model to **forecast** the demand of item sales in a Store using Time series analysis.

The data is classified in date and sales each day of every item in each store.

DATA VISUALIZATION

Daily Sales

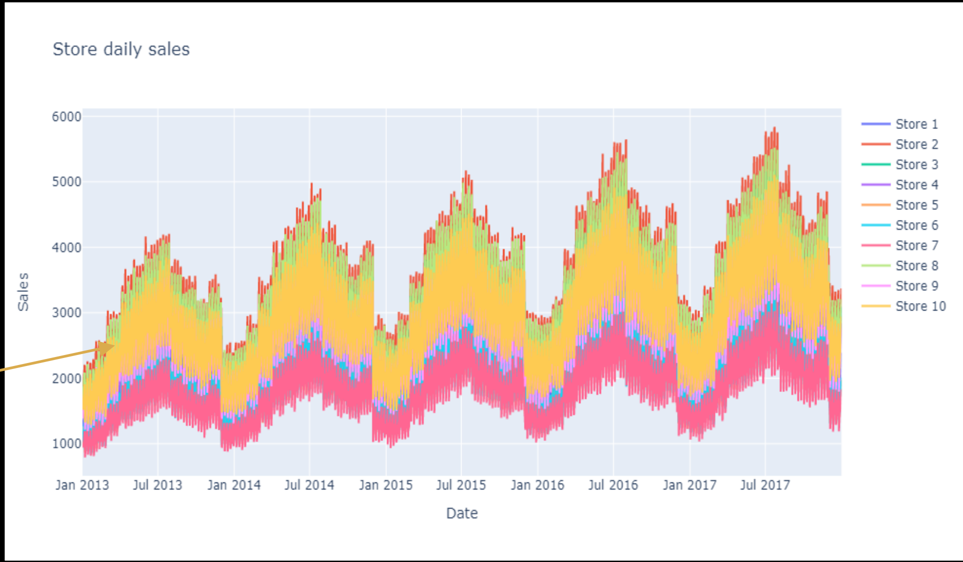
In this Graph it is clearly visible **July** is pick period across years .



DATA VISUALIZATION

Daily sales by store

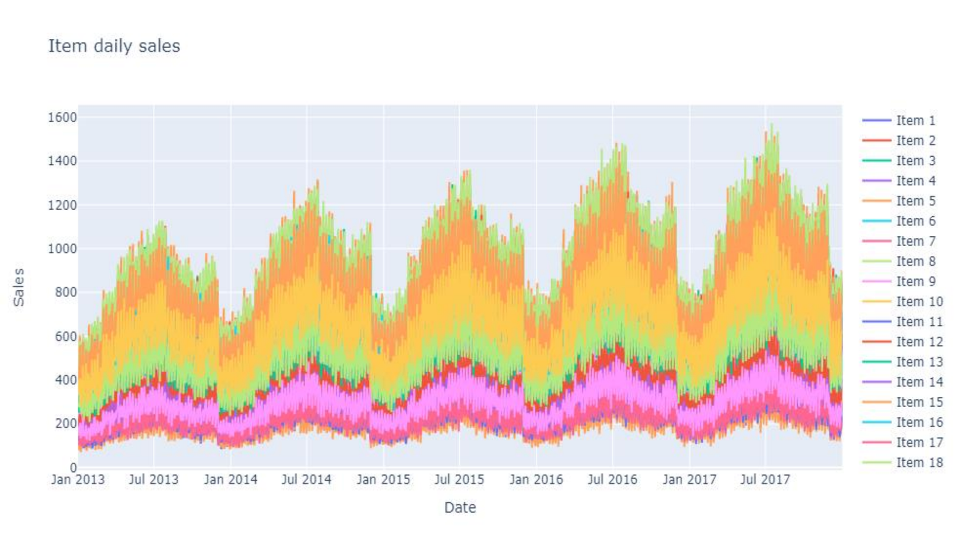
In this Graph store-wise distribution is given and major sales is coming from **Store 10** highlighted in Yellow color.



DATA VISUALIZATION

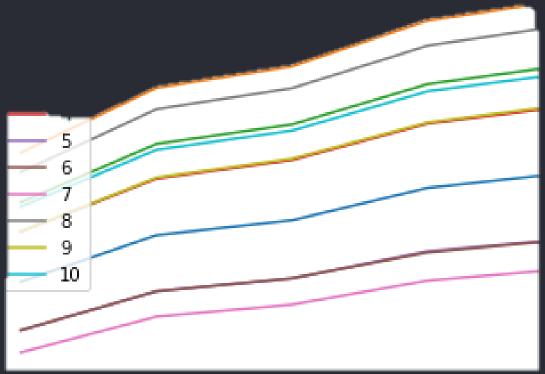
Daily Sales by Items

In this Graph item-wise distribution is given and major sales is coming from **Item 10** highlighted in Yellow color followed by **Item 9** highlighted in pink color.

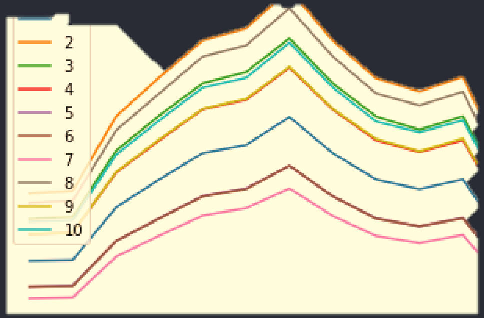


SALES PROFILE

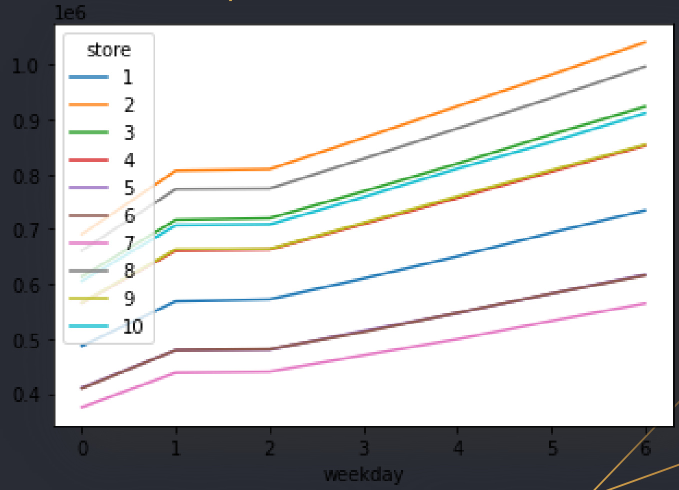
Annual Sale



Seasonal Sale

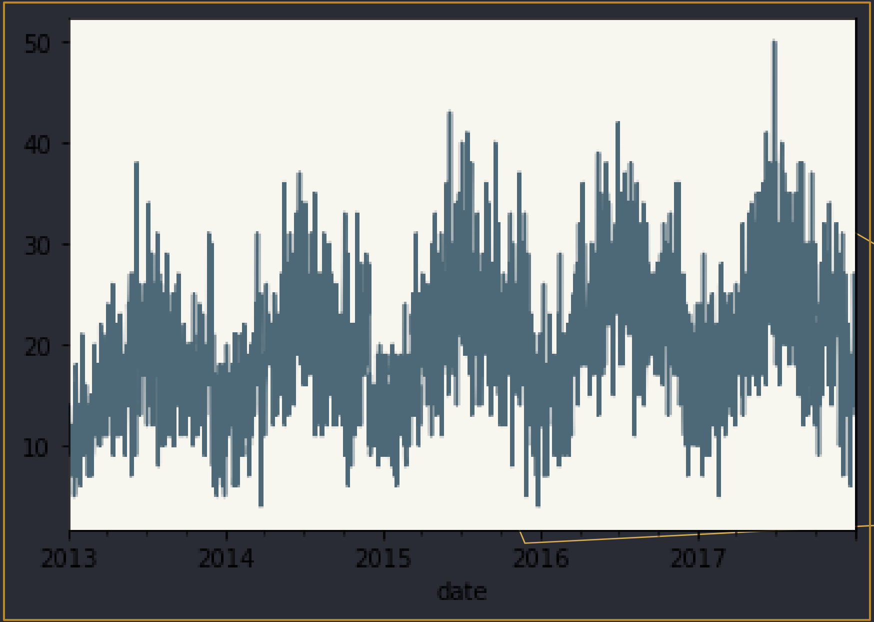


Weekly Sale



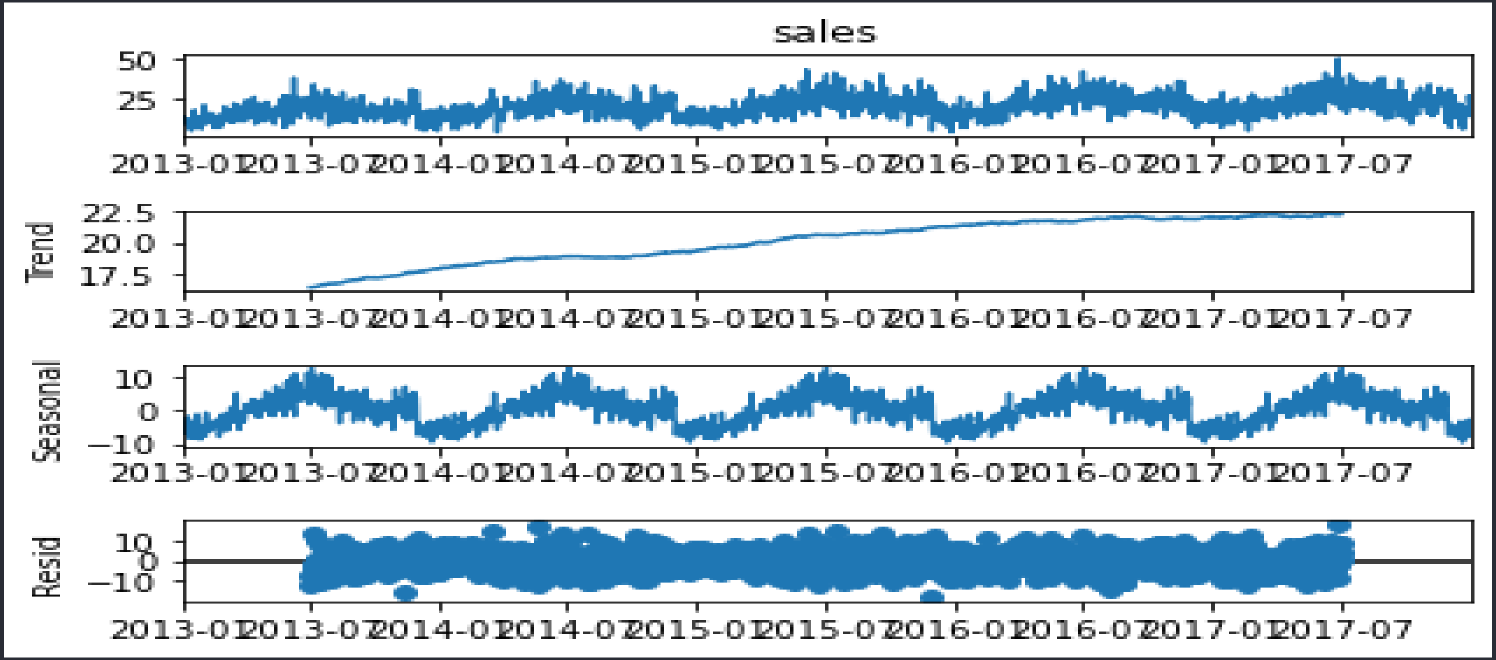
STATIONARITY CHECK PLOT

We will see two methods to check stationarity.
Making a function to check stationarity in one go using both **rolling statistics plot and ADF test**.



DECOMPOSITION

Below plot is very clearly showing **Trend** & **Seasonal** component in the series.



Lets see how to handle it ?

STATIONARITY CHECKING

ADF Test(Augmented Dickey fuller) Test

- For a Time series to be **stationary**, its ADF test should have:
- **low p-value** (according to the null hypothesis)
- **Critical values** at 1%, 5%, 10% confidence intervals should be as **close** as possible **to** the **Test Statistic**

Rolling Statistics Methodology

- Rolling mean has a **trend** component
- Rolling standard deviation is fairly **constant** with time.
- For our time series to be stationary, we need to ensure that both the **rolling statistics** i.e.: **mean & standard deviation remain time invariant** or constant with time. Thus the curves for both of them have to be parallel to the x-axis, which in our case is not so.

STATIONARITY CHECKING

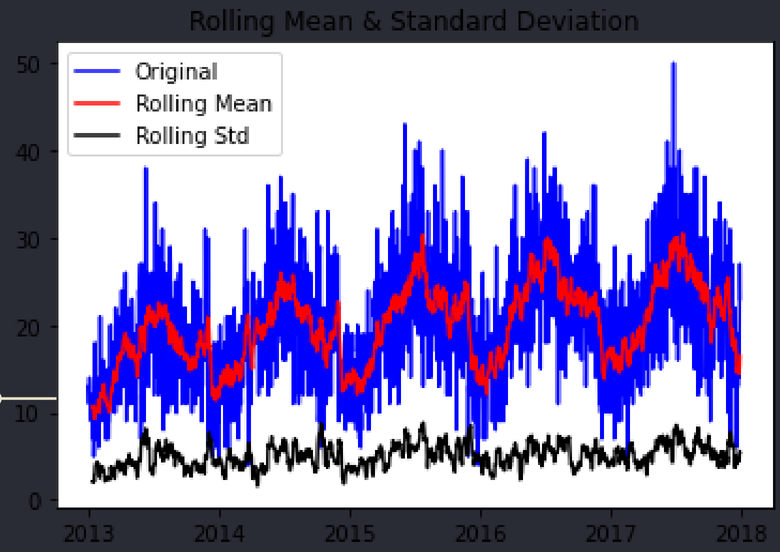
ADF Test

Results of Dickey Fuller Test:

- Test Statistic -3.157671
- p-value 0.022569
- Critical Value (1%) -3.433984

The Dickey-Fuller test the time series is not considered stationary as Test Statistic greater than Critical Value and we can see visually that there is an upwards trend.

Rolling Statistics



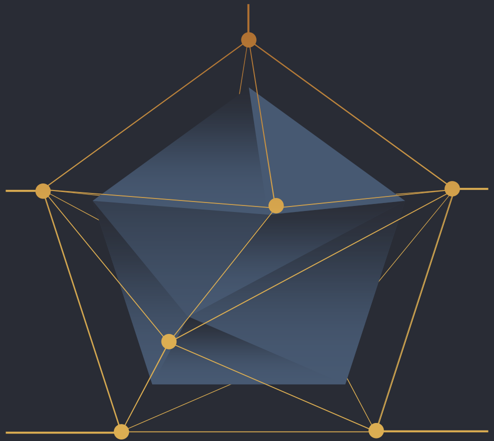
- Rolling mean has a **trend** component
- Rolling standard deviation is fairly **constant** with time.

DATA TRANSFORMATION TO ACHIEVE STATIONARITY

There are a couple of ways to achieve stationarity through data transformation like taking $log10$log10,$loge$loge, square, square root, cube, cube root, exponential decay, time shift and so on ... We will use this 4 methods.

Log scale transform

Exponential Decay



Log Scale Minus Moving Average

Log Scale Shifting

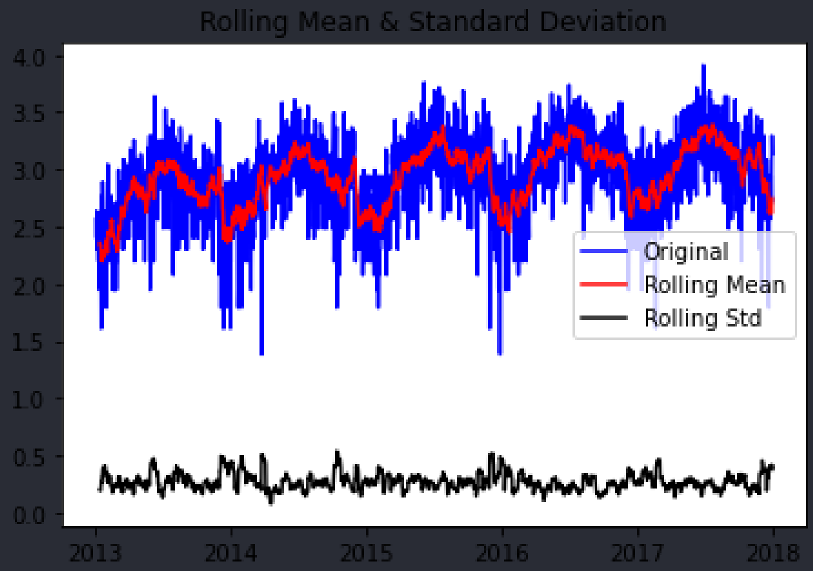
LOG SCALE TRANSFORMATION

ADF Test

Results of Dickey Fuller Test:

- Test Statistic: -3.594424
- p-value: 0.005869
- Critical Value (1%): -3.433984

Rolling Statistics



From above graph, we see that even though rolling mean is not stationary, it is still better than the previous case, where no transformation were applied to series. So we can at least say that we are heading in the right direction.

LOG SCALE MINUS MOVING AVERAGE

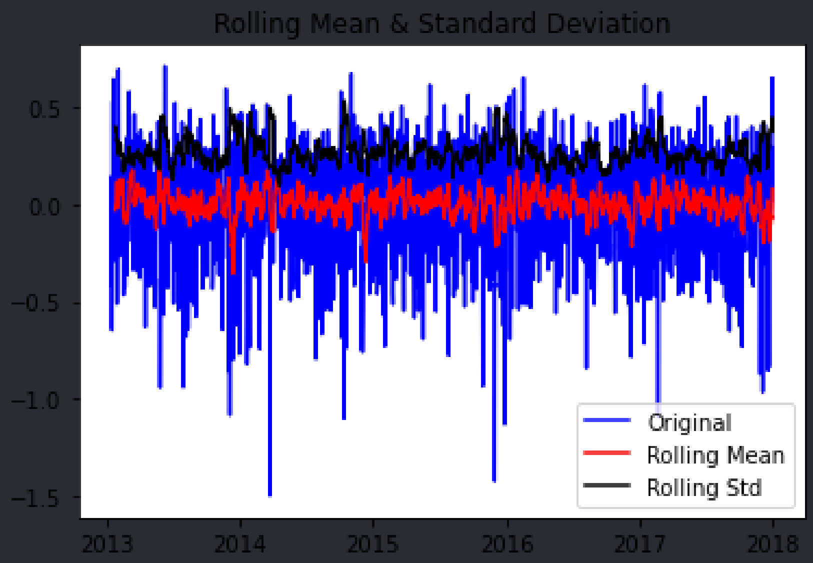
ADF Test

Results of Dickey Fuller Test:
- Test Statistic: -1.031707e+01
- p-value: 3.077488e-18
- Critical Value (1%): -3.434000e+00

The **critical values** at confidence intervals are pretty **close to the Test Statistic**.

Thus, we can say that our given series is stationary.

Rolling Statistics



From above graph, we observe that our intuition that subtracting two related series having similar trend components will make the result stationary"* is true.

EXPONENTIAL DECAY WEIGHTED AVERAGE

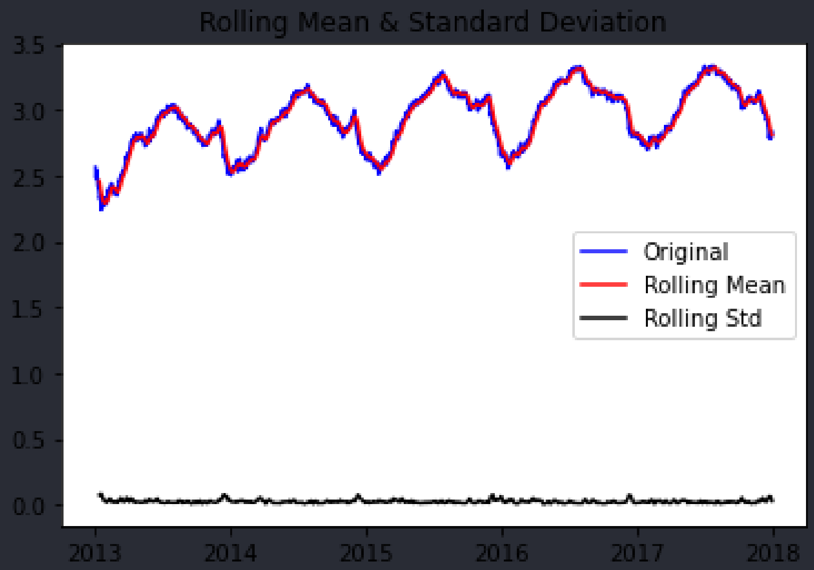
ADF Test

Results of Dickey Fuller Test:

- Test Statistic: -3.666915
- p-value: 0.004603
- Critical Value (1%): -3.433986

p-value has **decreased and** Test Statistic value is very **much closer** to the Critical values. Both the points say that our current transformation is better than the previous logarithmic transformation. Even though, we couldn't observe any differences by visually looking at the graphs, the tests confirmed decay to be much better.

Rolling Statistics



From above graph, it seems that exponential decay is not holding any advantage over log scale as both the corresponding curves are similar.

LOG SCALE SHIFTING

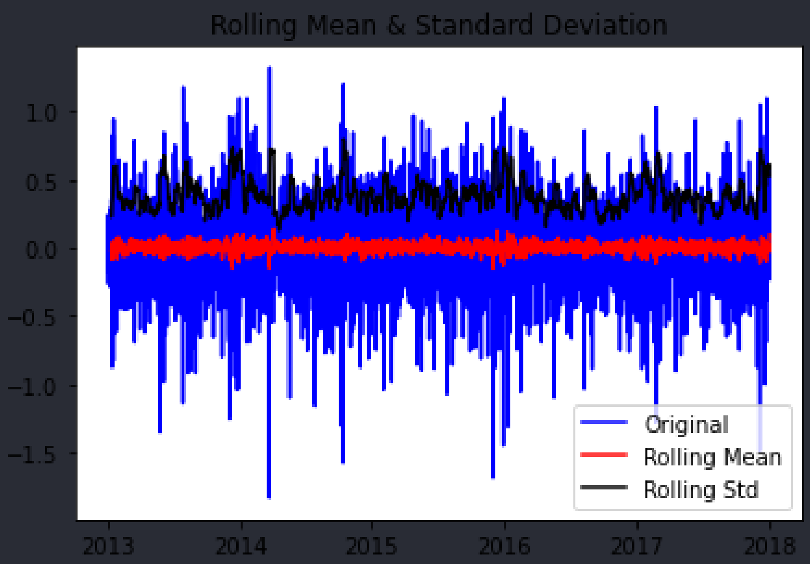
ADF Test

Results of Dickey Fuller Test:

- Test Statistic: -1.259629e+01
- p-value: 1.775858e-23
- Critical Value (1%):-3.433984e+00

p-value is extremely small. Thus this series is very likely to be stationary.

Rolling Statistics

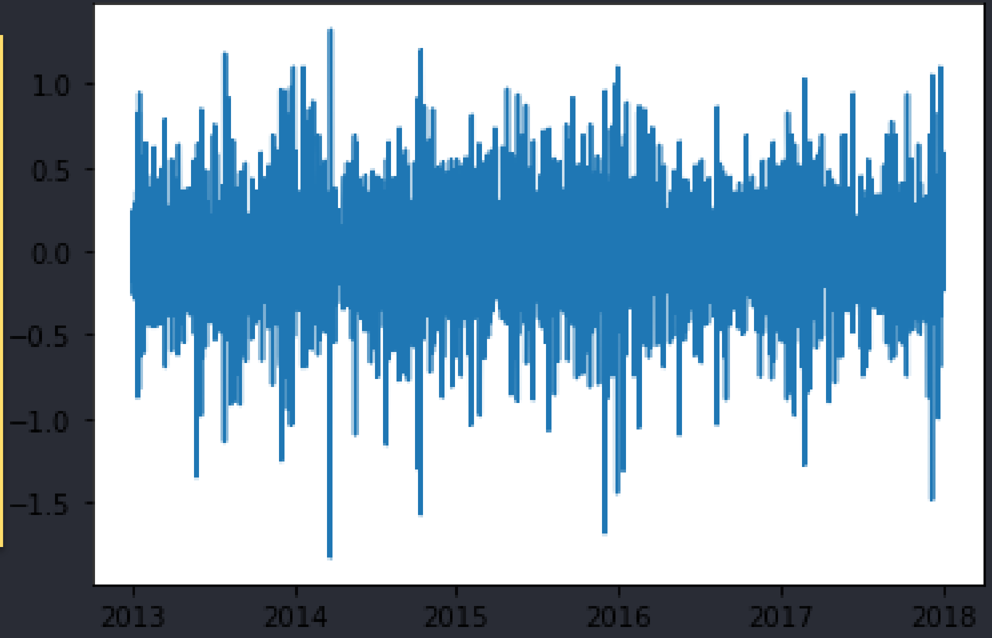


From above graph, we can see that, visually this is the best result as our series along with rolling statistic values of moving avg & moving std. dev. is very much flat & stationary.

DATA IS NOW STATIONARY

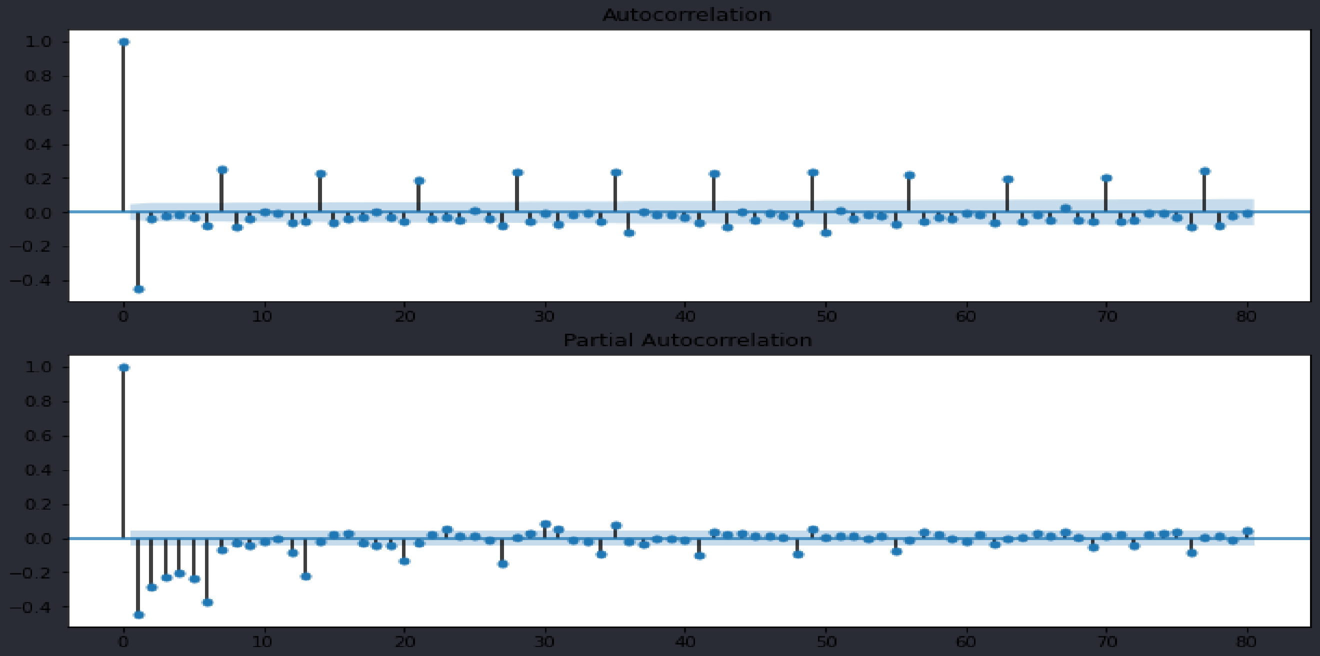
From graphs, we can see that, visually this is the best result as our series along with rolling statistic values of moving avg & moving std. dev. is very much flat & stationary.

Test Statistic value not as close to the critical values as that for exponential decay.
We have thus tried out 4 different transformation: log, exp decay & time shift. We will go ahead with the time shifted dataset.



PLOTTING ACF & PACF

Here we can see the ACF and PACF both has a recurring pattern every 7 periods. Indicating a weekly pattern exists. With both P & Q=7

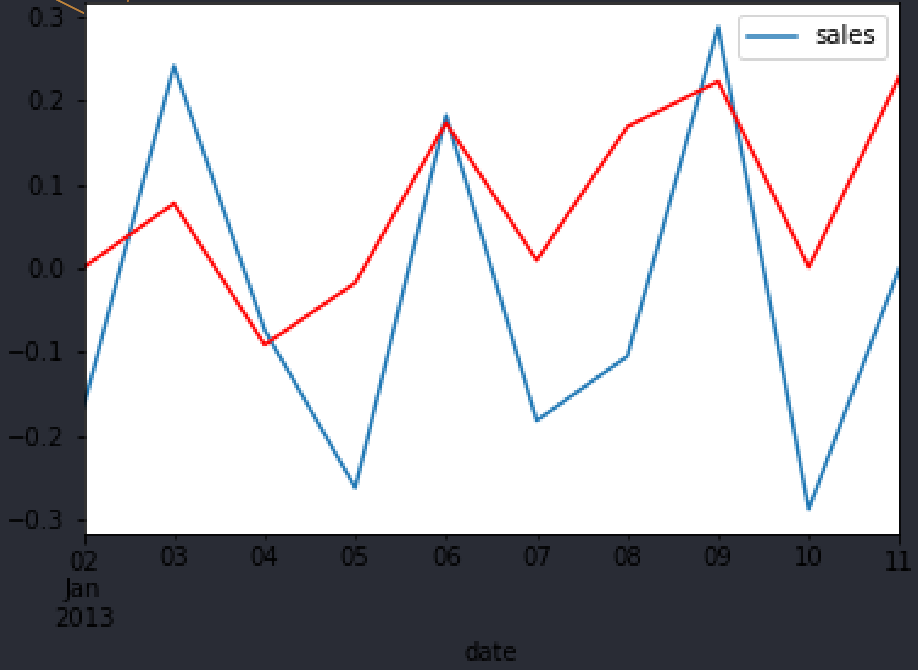


TRAIN TEST SPLIT & BUILDING MODELS

Summary of ARMA Model:
Forecasting using the developed model and printing out the **7 month forecast**.

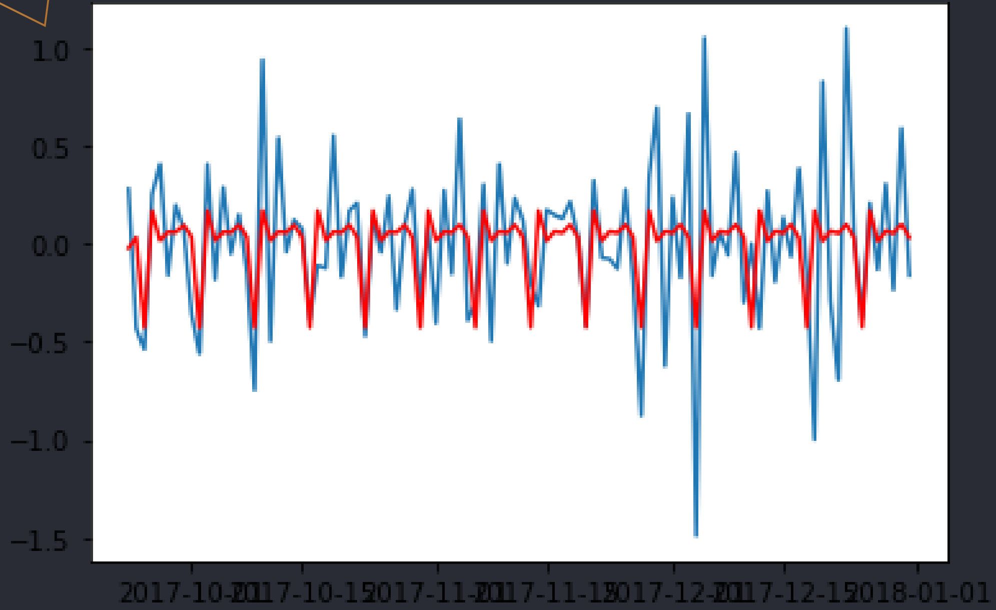
```
                              ARMA Model Results
==============================================================================
Dep. Variable:                  sales   No. Observations:               1725
Model:                     ARMA(7, 7)   Log Likelihood                -99.607
Method:                       css-mle   S.D. of innovations             0.255
Date:                Tue, 08 Dec 2020   AIC                           231.213
Time:                        11:06:23   BIC                           318.461
Sample:                    01-02-2013   HQIC                          263.489
                         - 09-22-2017
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const           0.0004      0.001      0.687      0.492      -0.001       0.002
ar.L1.sales    -0.9856      0.000  -3052.021      0.000      -0.986      -0.985
ar.L2.sales    -0.9850      0.001  -1754.320      0.000      -0.986      -0.984
ar.L3.sales    -0.9857      0.001  -1342.580      0.000      -0.987      -0.984
ar.L4.sales    -0.9851      0.001  -1323.222      0.000      -0.987      -0.984
ar.L5.sales    -0.9856        nan        nan        nan         nan         nan
ar.L6.sales    -0.9852      0.001  -1749.848      0.000      -0.986      -0.984
ar.L7.sales     0.0146      0.001     26.889      0.000       0.014       0.016
ma.L1.sales     0.1062      0.011      9.426      0.000       0.084       0.128
ma.L2.sales     0.0997      0.011      8.784      0.000       0.077       0.122
ma.L3.sales     0.1084      0.011      9.605      0.000       0.086       0.131
ma.L4.sales     0.1019      0.011      9.161      0.000       0.080       0.124
ma.L5.sales     0.1026      0.010     10.088      0.000       0.083       0.123
ma.L6.sales     0.1023      0.011      9.044      0.000       0.080       0.124
ma.L7.sales    -0.8913      0.011    -82.741      0.000      -0.912      -0.870
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            0.6235           -0.7818j            1.0000           -0.1429
AR.2            0.6235           +0.7818j            1.0000            0.1429
AR.3           -0.9012           -0.4336j            1.0000           -0.4286
AR.4           -0.9012           +0.4336j            1.0000            0.4286
AR.5           -0.2225           -0.9750j            1.0000           -0.2857
AR.6           -0.2225           +0.9750j            1.0000            0.2857
AR.7           68.2578           -0.0000j           68.2578           -0.0000
MA.1           -0.9017           -0.4325j            1.0001           -0.4288
MA.2           -0.9017           +0.4325j            1.0001            0.4288
MA.3           -0.2228           -0.9766j            1.0017           -0.2857
MA.4           -0.2228           +0.9766j            1.0017            0.2857
MA.5            0.6236           -0.7826j            1.0007           -0.1429
MA.6            0.6236           +0.7826j            1.0007            0.1429
MA.7            1.1166           -0.0000j            1.1166           -0.0000
```

TRAIN THE MODEL



Test MSE: 0.066
Observing the plot of **expected** vs. the **predicted**. The forecast does look pretty good with slightly large deviation.

TEST THE MODEL



Test MSE: 0.134

We see that our predicted forecasts are very close to the real time series values indicating a fairly accurate model.

CONCLUSION

We have built a model to forecast the demand of item 1 sales in a Store 1 Using ARIMA model with P value and Q value 7 Since we observed the ACF and PACF both has a recurring pattern every 7 periods. Indicating a weekly pattern exists.

ANNEXURES

Insaid.co

https://towardsdatascience.com

https://otexts.com/fpp2/data-methods.html

https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

THANKS!