Name : Raghav Alagh
Reg No : 15BCE1249

Digital Assignment - 1

AutoRegressive Integrated Moving Averages (ARIMA)

I have chosen **sales of shampoo** dataset available on the kaggle to apply two algorithms that is **ARIMA** and **FBPROPHET**. This is basically a time series dataset on which sales of shampoo for three year period are given to us and we need to predict the future sales of the shampoo.

Data Preprocessing for Time Series Analysis

We basically need to check four main factors are :

Trend : It is referred as a linear or a non-linear component that changes over time and does not repeat within the span of time . Trend can be removed from the dataset using the process of smoothing (Integration method) . The secular trend is the main component of a time series which results from long term effects of socio-economic and political factors. This trend may show the growth or decline in a time series over a long period. This is the type of tendency which continues to persist for a very long period. Prices and export and import data, for example, reflect obviously increasing tendencies over time.

Seasonality : It is referred to as the component that has similar values after every fixed interval of time .These are short term movements occurring in data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities. For example, it is commonly observed that the consumption of ice-cream during summer is generally high and hence an ice-cream dealer's sales would be higher in some months of the year while relatively lower during winter months. Employment, output, exports, etc., are subject to change due to variations in weather. Similarly, the sale of garments, umbrellas, greeting cards and fire-works are subject to large variations during festivals like Valentine’s Day, Eid, Christmas, New Year's, etc. These types of variations in a time series are isolated only when the series is provided biannually, quarterly or monthly.Seasonality from the dataset can be removed by differencing (Integration) .

Randomness : Checking randomness basically help us determine whether learning is feasible or not . Autocorrelation plot can be very useful in determining the randomness of a dataset . In case this autocorrelation is 0 then there is large randomness in such case .

Cyclic : It basically involves checking if there are any kind of cycles existing in our dataset . These are long term oscillations occurring in a time series. These oscillations are mostly observed in economics data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated with the well known business cycles. These cyclic movements can be studied provided a long series of measurements, free from irregular fluctuations, is available.

AutoRegressive Integrated Moving Averages (ARIMA)

This model is basically a combination of three different models namely Autoregression , Integration and Moving Averages .

AutoRegression (p) : It basically states that the current value of an observation might depend upon a number of lagged observations . It is also called as the lag order.

Integration (d) : Sometimes our time-series dataset might not be stationary implying that the statistical properties of the dataset that is mean,median,mode might change over time . This varying nature may cause problems in making the future predictions .So this model basically performs the differencing operation in order to make a non-stationary time series stationary . It is also referred to as the degree of differencing , the number of times the raw observations are subtracted .

Moving Averages (q) : It basically depicts the relation between the current observation and the errors in the previous set of lagged observations .

FBprophet

ARIMA model is a completely automatic forecasting technique that can be brittle and is often too inflexible to incorporate several useful assumptions as well as heuristics . Analyst who can produce high quality forecasts are quite rare because forecasting is a specialized data skill requiring substantial experience .

Where Prophet shines :

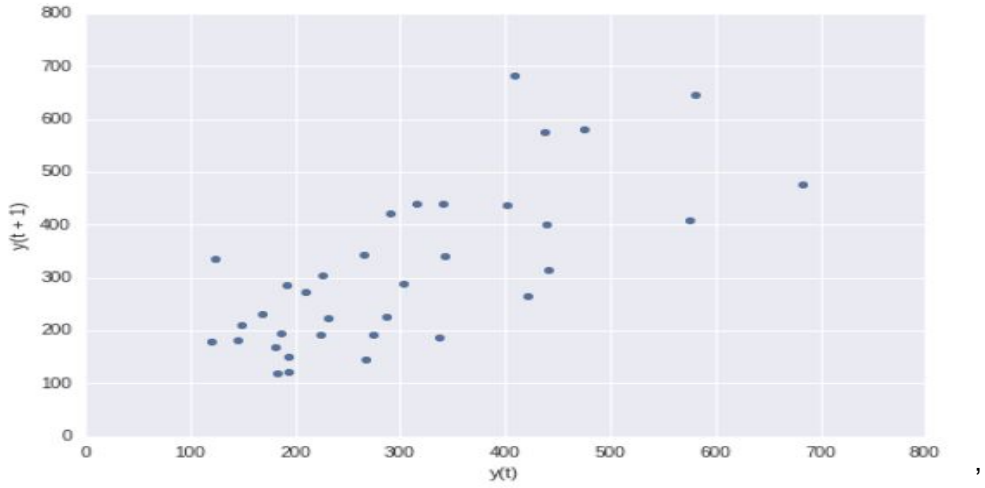
Hourly , daily or weekly observations with at least a few months os history .
Strong multiple “human-scale” seasonalities : day of week and time of year .
Important holidays that occur at irregular intervals that are known in advance .

Important holidays that occur at irregular intervals that are known in advance .
A reasonable number of missing observations .
Historical trend changes , for instance due to product launches or logging chances .

Experiment

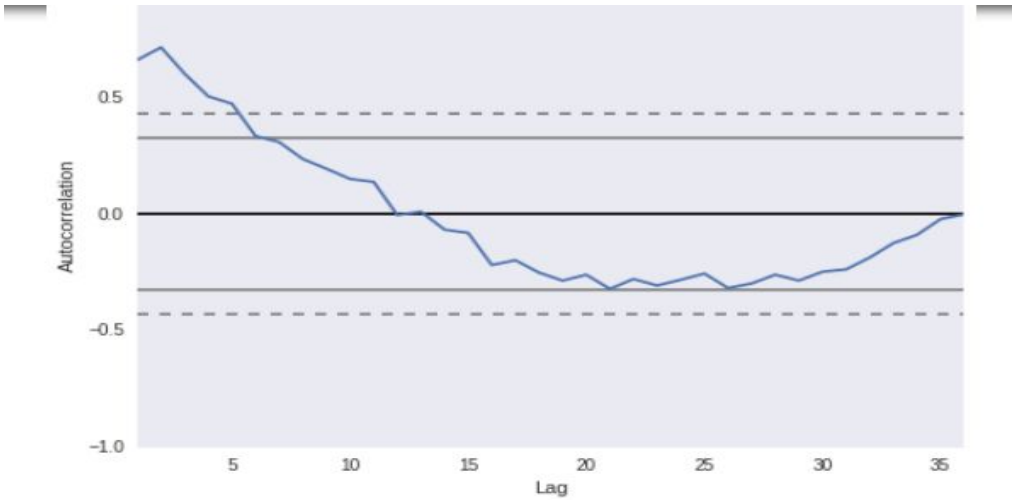
About dataset : I have a very small dataset with 36 rows having the shampoo sales for the last three years and our task is train a model to predict the future sale values of these shampoo sales .
The **initial plot of the dataset** shows the existence of a clear trend in the dataset .

The l**ag plot** clearly shows the existence of autocorrelation between the time series data .

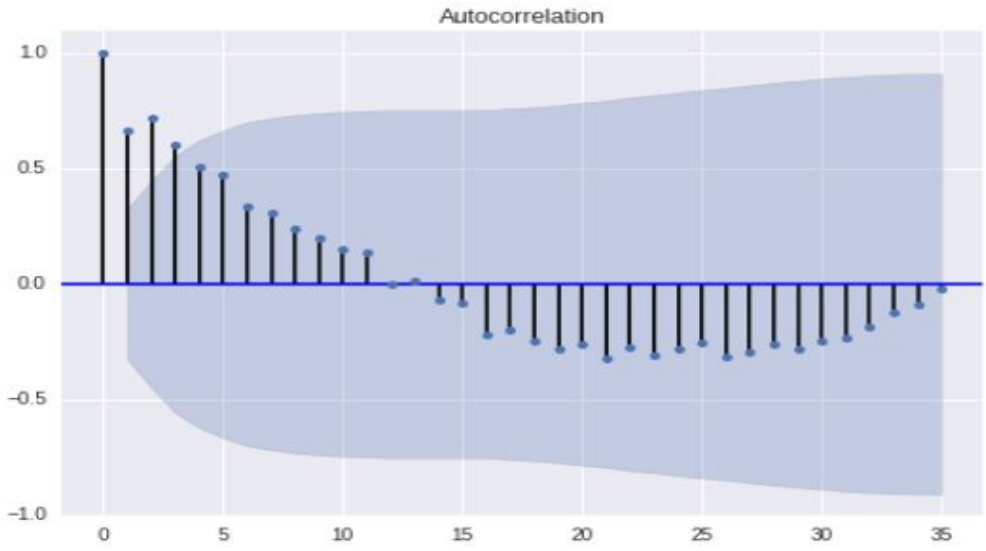


,

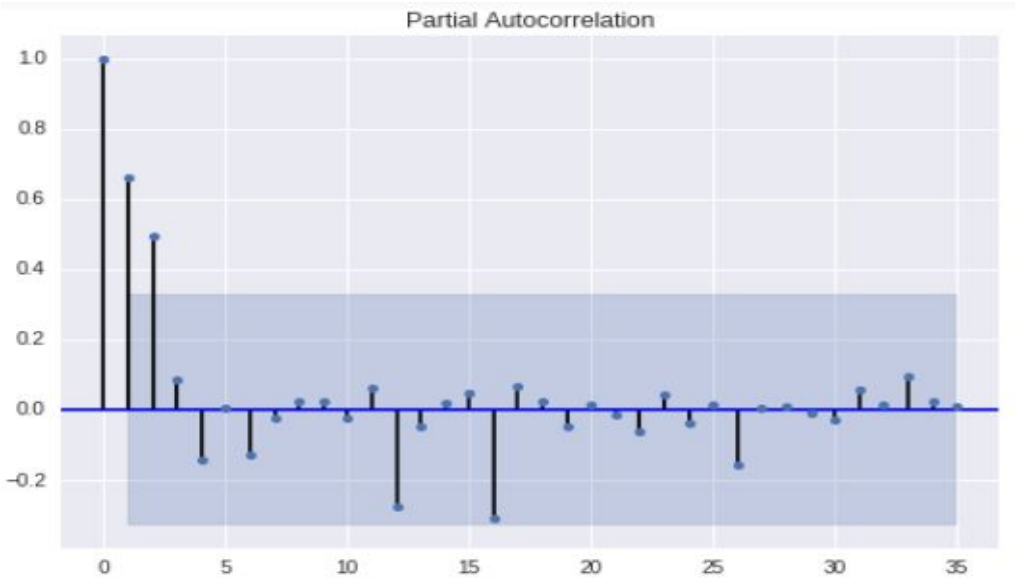
The **autocorrelation plot** clearly shows that our time series have a high positive autocorrelation for almost first 5 to 10 lag values . So a lag of nearly 5 can be a good starting point .



Following is another view of the **autocorrelation plot** :



The **partial autocorrelation** graph of the data is given as follows :



ARIMA Model Implementation

Now I did the hypertuning for various parameters of ARIMA out of which p=2,q=2 and d=1 showed the best results giving the mean squared error of 31,776 (approx) .

Following is the summary of our best trained ARIMA model .

ARIMA Model Results

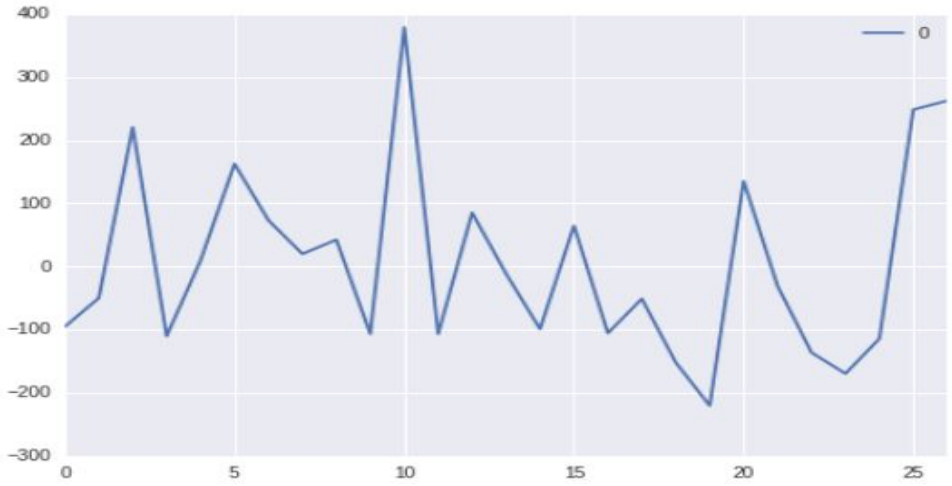
Dep. Variable:	D.y	No. Observations:	27
Model:	ARIMA(2, 1, 1)	Log Likelihood	-173.662
Method:	css-mle	S.D. of innovations	140.796
Date:	Fri, 16 Feb 2018	AIC	357.323
Time:	02:09:56	BIC	363.802
Sample:	1	HQIC	359.250

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.3772	3.197	0.744	0.465	-3.889	8.643
ar.L1.D.y	-0.0844	0.215	-0.393	0.698	-0.506	0.337
ar.L2.D.y	-0.0286	0.236	-0.121	0.905	-0.491	0.434
ma.L1.D.y	-1.0000	0.126	-7.966	0.000	-1.246	-0.754

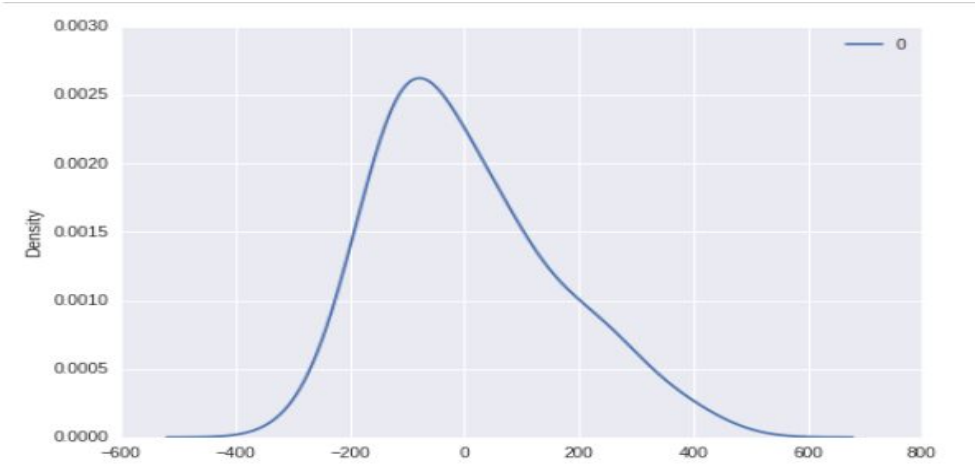
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.4775	-5.7286j	5.9161	-0.2902
AR.2	-1.4775	+5.7286j	5.9161	0.2902
MA.1	1.0000	+0.0000j	1.0000	0.0000

The residual plot of the errors of the ARIMA model will be given as follows :



And the required gaussian plot of the errors in this case will be given as :



Implementation Of FBprophet Model

Following are the 4 next predicted values of the FBProphet Model :

	ds	yhat	yhat_lower	yhat_upper
10	1902-10-01	300.758196	198.831708	410.058889
11	1903-01-01	491.058263	385.506895	603.899831
12	1903-04-01	525.701306	414.420246	635.603253
13	1903-05-01	458.441784	345.979933	576.114742
14	1903-08-01	488.168282	380.883854	585.267835

And I got the best result when I had put the container width as 80% . Following is the result :