Project: Forecasting Sales

Step 1: Plan Your Analysis

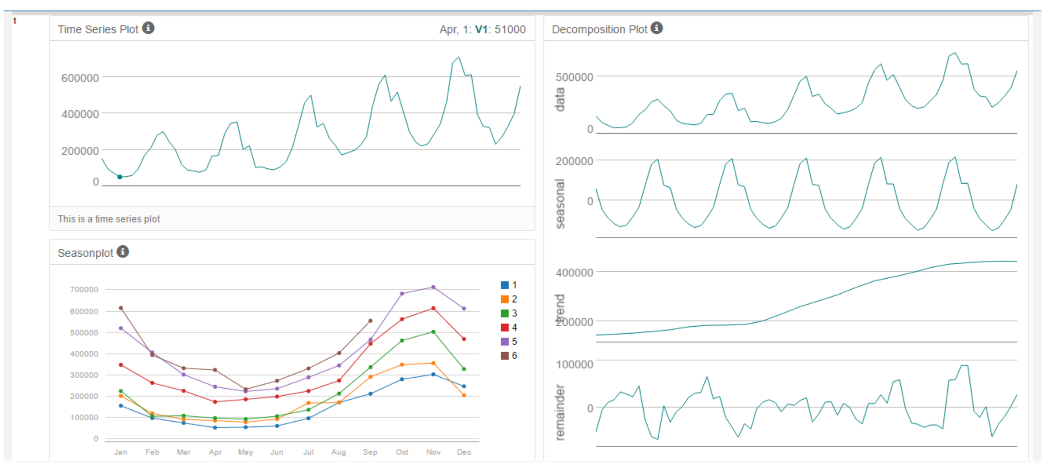
1. Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.
   Yes, the dataset meets the criteria of a time series dataset. It has the following four characteristics:
   - It is continuous data (time based)
   - It is sequential
   - It has equal intervals of time
   - For each time interval there is an associated value which is Monthly sales.

2. Which records should be used as the holdout sample?

4 records i.e.; 4 periods of data should be held for the holdout sample because we need forecast data for 4 periods.

Step 2: Determine Trend, Seasonal, and Error components

1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

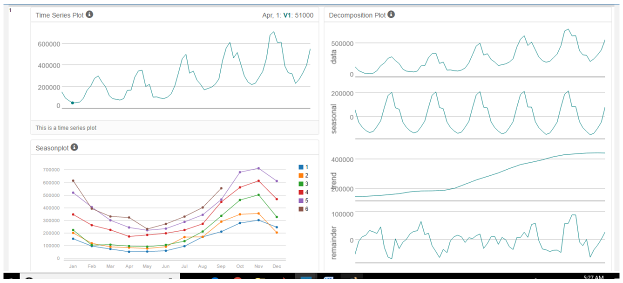


Looking at the decomposition plot, we can determine that the data has both upward linear trend and seasonality.
Looking at the plot, the data has an upward linear trend which means that we can add trend additively.
The seasonal component increases in magnitude each year so we can add seasonality multiplicatively.
The error component of the data increases in variance over time hence we can add error multiplicatively.

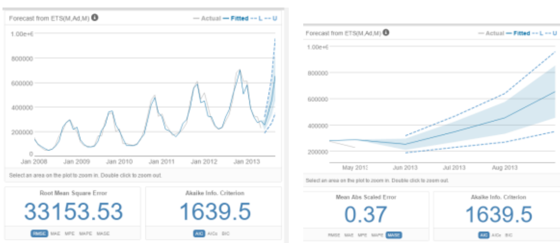
Step 3: Build your Models

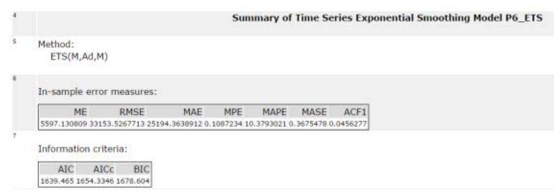
1. What are the model terms for ETS? Explain why you chose those terms.

a. Describe the in-sample errors. Use at least RMSE and MASE when examining results



Based on the decomposition plot above, we determine that the trend is added additively, seasonality and error are added multiplicatively. So the ETS model would be ETS (M, A, M).
I have used 'Auto' for dampening.





Summary of Time Series Exponential Smoothing Model P6_ETS

Method:
ETS(M,Ad,M)

In-sample error measures:

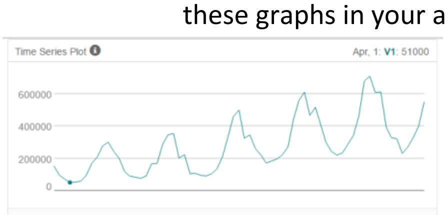
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

Information criteria:

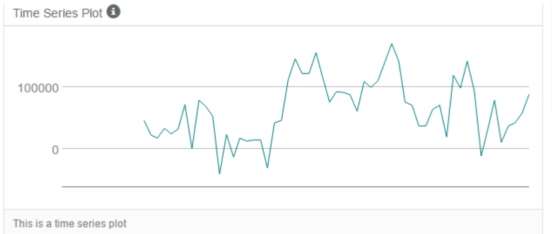
AIC	AICc	BIC
1639.465	1654.3346	1678.604

It has low RMSE, which suggest this has narrow possible range of set of values.
The RMSE value is 33153.53 which suggests that the forecasted values are deviated by this amount from the mean.
Also MASE is 0.37 which is less than one and hence we can say that this is a good ETS model.

2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
   b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

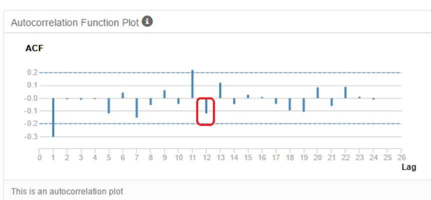


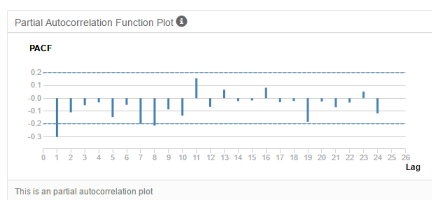
The above plot is the time series data plot, which suggests that the mean is not close to zero and doesn't have equal variance. The mean and variance are not constant over time.
This means we have to stationarize the data hence taking the seasonal difference.



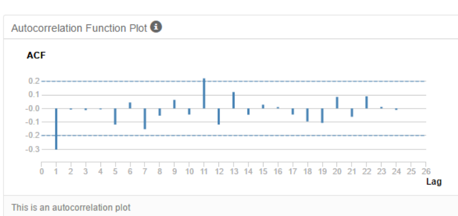
Even after taking the seasonal difference, the mean and variance are not constant. So in order to stationrize the data we need to the first differencing of seasonal difference.

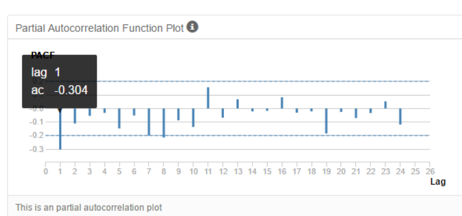
After, first differencing the seasonal difference, we are able to stationarize the data as shown below with constant mean and variance. Also ACF and PACF plots show that most of the lags are not correlated significantly.



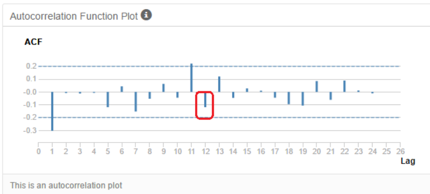


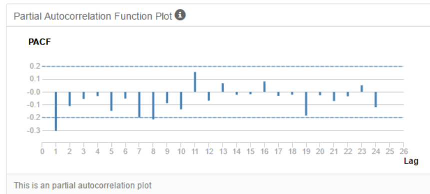
Now constructing the ARIMA model, I have used ARIMA (0, 1, 1) (0, 1, 0) [12] model.
Considering that ACF at lag-1is negatively correlated based on the below plot, I decided to choose MA.





Based on the ACF and PACF plots, lag-1 is negatively correlated with an ACF value of -0.3 which is significant hence considering 1 for q and 0 for p since we are considering MA due to negative correlation which is also confirmed in the PACF plot.



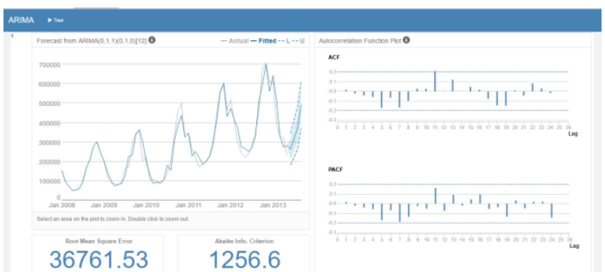


Now looking at the seasonal lag at lag-12 which is also negatively correlated however since the value is insignificant Q is 0. And since this is negatively correlated P is 0 as well.

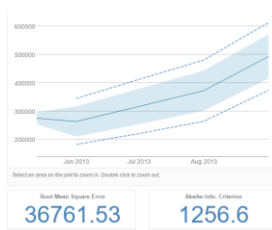
I have done one seasonal differencing and one first differencing of seasonal difference values. Hence assigning 1, 1 to d, D respectively.

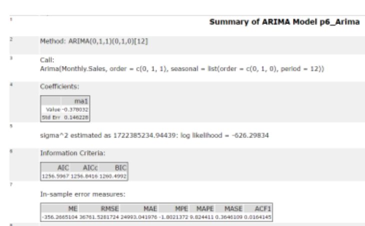
In conclusion, I have used ARIMA (0, 1, 1) (0, 1, 0) with 12 months seasonal lag.

Now considering this, below are the results of the ARIMA model



Now if we look at the ACF and PACF plot we can see that there is no significant correlation between any ACF values.





Summary of ARIMA Model p6_Arima

Method: ARIMA(0,1,1)(0,1,0)[12]

Call:
Arima(Monthly.Sales, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1
Value	-0.378032
Std Err	0.146228

sigma^2 estimated as 1722385234.94439: log likelihood = -626.29834

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

In-sample error measures:

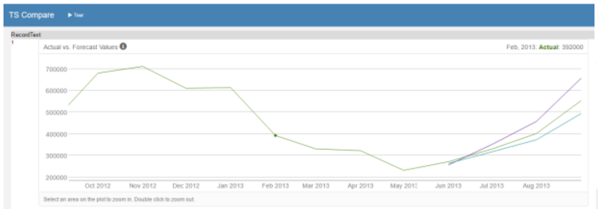
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

The MASE value for this ARIMA model is 0.36 which is less than 1 and is significant which suggests this is a good value. The RMSE is low, which suggests that the deviation from mean for forecasted values is only 36761.53.

Step 4: Forecast

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.

Below are the forecast error measures:



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
P6_ETS	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA
p6_Arima	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

Below are the in-sample error measures for ETS and ARIMA models:

Summary of Time Series Exponential Smoothing Model P6_ETS

Method:
ETS(M,Ad,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

Summary of ARIMA Model p6_Arima

Method: ARIMA(0,1,1)(0,1,0)[12]

Call:
Arima(Monthly.Sales, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1
Value	-0.378032
Std Err	0.146228

sigma^2 estimated as 1722385234.94439: log likelihood = -626.29834

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

Considering the AIC values of ETS (1639.465) and ARIMA (1256.5967) we can conclude that ARIMA is better model for forecasting the values.

Now looking at the accuracy measures, RMSE and MAPE are lower for ARIMA model when compared to ETS. Also the MASE for ARIMA is 0.45 which is significant as it is less than one.

Keeping together all the factors, we can say that ARIMA would be a better model to consider than ETS here. When looking at predicting values for holdout sample ARIMA model works better which suggest that it has better predicting capability than ETS here.

2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

Forecasting the values using ARIMA (1, 0, and 1) (1, 0, and 0) [12] below are the results:

Period	Sub Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2013	10	754854.5	834046.2	806635.2	703073.8	675662.7
2013	11	785854.5	879377.8	847006.1	724702.9	692331.2
2013	12	684854.5	790787.8	754120.6	615588.4	578921.1
2014	1	687854.5	804889.3	764379.4	611329.5	570819.6