Project: Forecasting Sales (Resubmit)

Step 1: Plan Your Analysis

Look at your data set and determine whether the data is appropriate to use time series models. Determine which records should be held for validation later on (250 word limit).

Answer the following questions to help you plan out 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 given dataset meets the criteria of a time series dataset.
   The observations(records) start from Jan 2008 to Sep 2013 which is 69 months in total and the dataset contains all observations for 69 months in a proper continuous sequence with equal interval, which is one month. And also, there is only one data point for each observation (Monthly Sales).

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

   The amount of sample holdout should resemble the amount of our forecast. Since we want to forecast next 4 months, last 4 records (2013-06 to 2013-09) should be used as the holdout sample.

Step 2: Determine Trend, Seasonal, and Error components

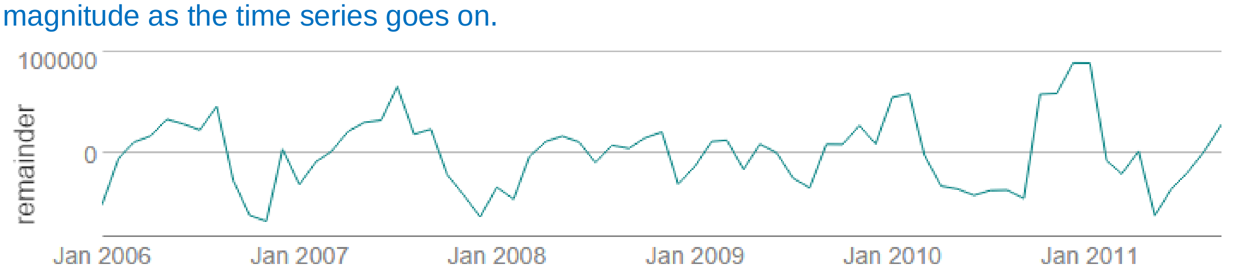
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

Answer this question:

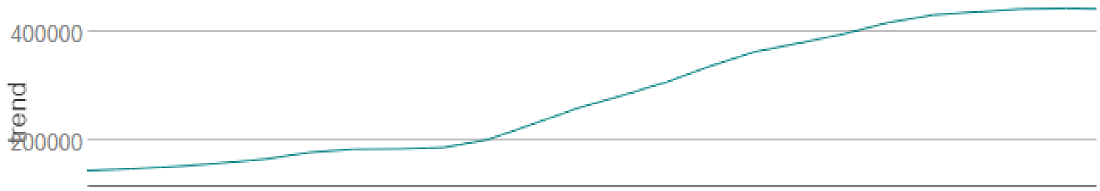
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

   The monthly sale data has an upward linear trend with seasonal pattern and constant variance remainder pattern.

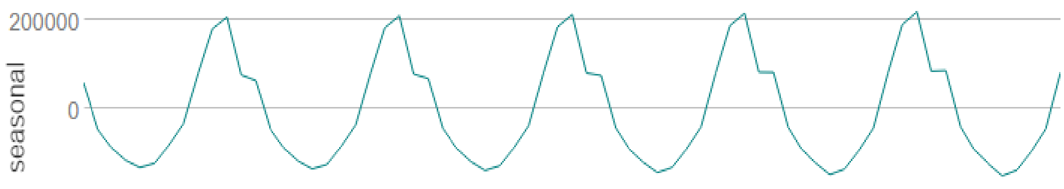
   There are fluctuations, between large and small errors, which are not consistent in magnitude as the time series goes on.



The monthly sale data has a trend that move upwards linearly.



Seasonal portion shows seasonal pattern with regularly occurring spikes in each seasonal turn and the magnitude of each spikes changes(increases) each year.



Step 3: Build your Models

Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)

Answer these questions:

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

   The model terms are
   - Multiplicative for error type
   - Additive for trend and
   - Multiplicative for seasonality.

   So, its ETS (M, A, M).
   The error type shows fluctuations with inconsistent magnitudes between small and large errors, so I chose multiplicative behavior.
   The seasonal pattern increases each seasonal turns, so I chose multiplicative behavior.
   The trend is linearly increased, so I chose additive behavior.

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

There is about 3729.29 or -0.95% average difference between actual and predicted values.

The standard deviation on the difference between actual and predicted values (RMSE) is 32883.83.

To detect forecast errors, I checked mean absolute error (MAE) and mean absolute percent error(MAPE) values, which are 24917.28 and 10.23 % respectively.

The mean absolute scaled error (MASE) is 0.36, so we can use this model to forecast future values, since MASE value less than 1 is usually considered significant.

In-sample error measures:

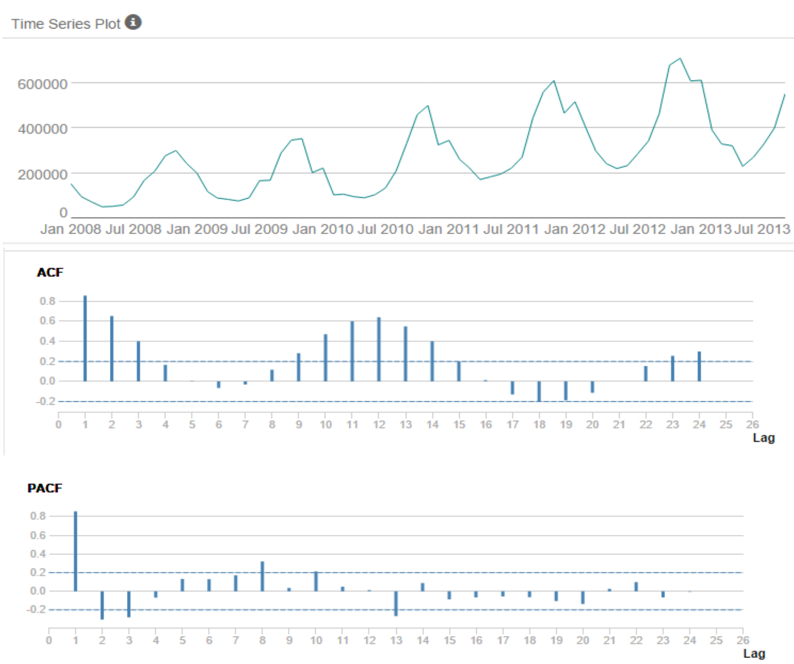
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056	0.1436491

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.

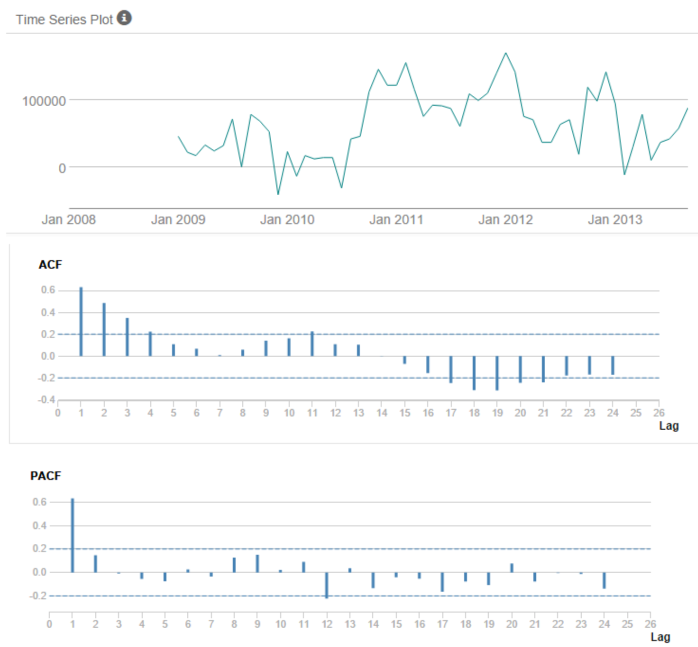
Model terms for Non-seasonal ARIMA are

- AR 1 and MA 0 (ARIMA (0,1,1)(0,1,0)[12])

TS Plot of time series before differencing



TS plot after seasonal differencing (d=1)



The seasonal difference ACF and PACF plots look similar to the Time Series ACF PACF plot.
After first differentiation, time Series plot lost its seasonal pattern, and there is no significant correlation in seasonal lags in the ACF and PCAC plot. So it becomes a stationary, and further differencing is not needed. ACF plot decays slowly while PACF plot displays a sharp cutoff. There is a positive correlation in lag of ACF which is confirmed by PACF lag1. So, the stationarized series shows 'AR signature'. Only one significant lag indicates that we have a MA(1) term. There are no significant correlations in seasonal lags, so we don't need further seasonal terms.
I used first seasonal differentiation, so d and D is equal to 1.
Lag 1 of ACF and PACF shows positive correlation, so q is equal to 0.
There is no significant correlation except lag 1, so Q is equal to 0.

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

There is about -356.27 or -1.8% average difference between actual and predicted values.
The standard deviation on the difference between actual and predicted values is 36762.
To detect forecast error, I checked mean absolute error (MAE) and mean absolute percent error (MAPE) values which are 24993 and 9.82% respectively.

The mean absolute scaled error (MASE) is 0.36, so we can use this model to predict forecast future values, since MASE value less than 1 is normally considered significant.

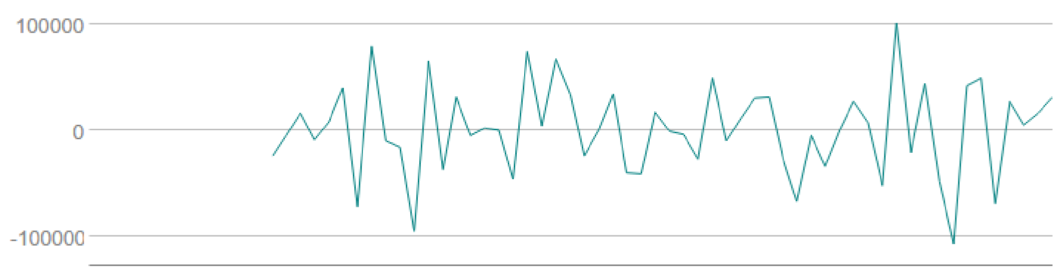
In-sample error measures:

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

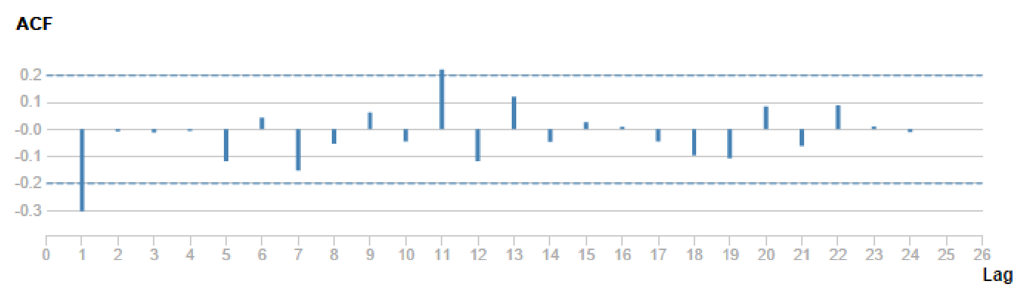
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

After seasonal first differencing

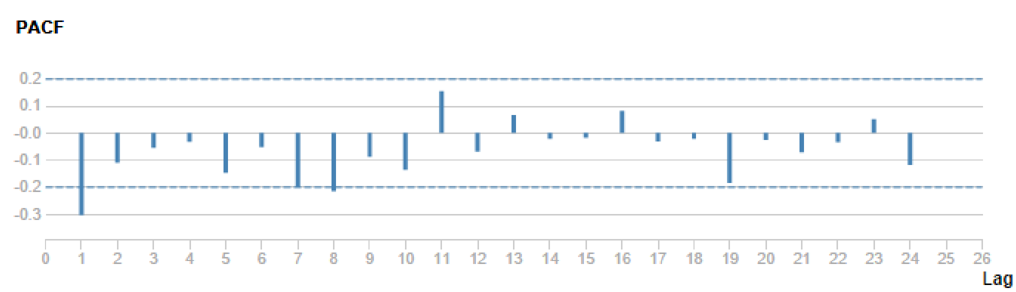
Time Series Plot



Autocorrelation Function Plot

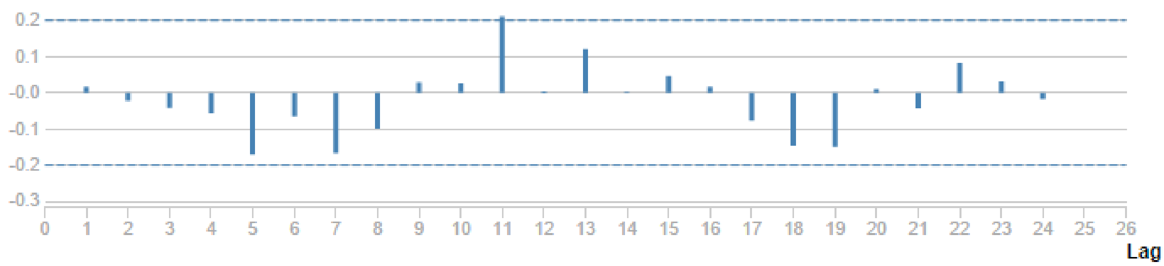


Partial Autocorrelation Function Plot

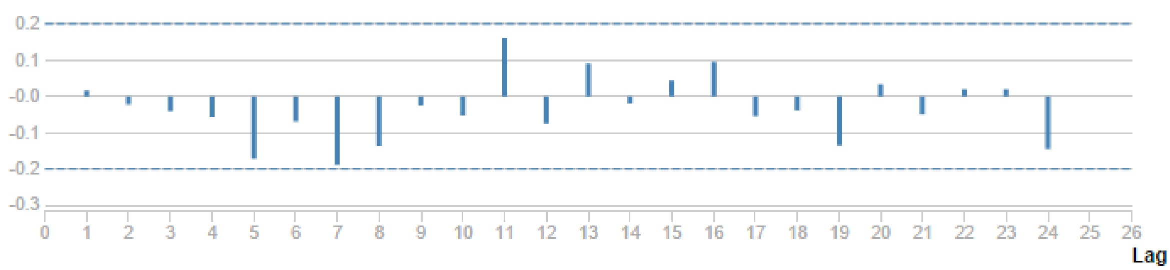


This is ACF and PACF plots after one MA term is added. There is no significantly correlated lags in ACF and PACF plots. No additional AR() or MA() terms is needed.

ACF



PACF



Step 4: Forecast

Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)

Answer these questions.

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

   I chose ARIMA model.

   Comparing in-sample errors

   In-sample error measures:

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

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

   Method: ETS(M,A,M)

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056	0.1436491

   MASE value of both ARIMA and ETS model is less than 1, which means that both model can be used to predict future values.
   The average error (ME) of ARIMA is smaller than that of the ETS model.
   Standard deviation of ARIMA model is larger than that of ETS.
   The average of the absolute values of the residuals (MAE) of the ARIMA model is slightly larger than that of the ETS model.
   MAPE value of ARIMA model is smaller than that of ETS model.
   So by comparing in-sample errors, I think ARIMA model is more suitable to forecast future.

Then, I forecast for holdout records and compare accuracy of both models.

ETS (M, A, M)

Actual and Forecast Values:

Actual	ETS
271000	268729.50166
329000	378187.04023
401000	488199.64792
553000	691913.69155

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	-68257.47	85623.18	69392.72	-15.2446	15.6635	1.1532	NA

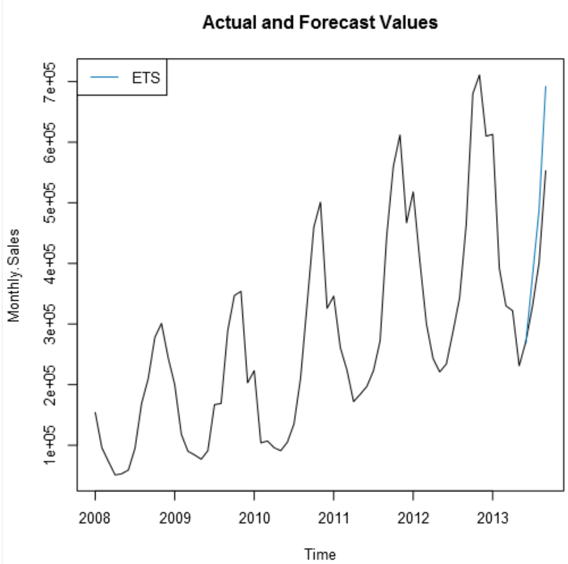
ARIMA(0,1,1)(0,1,0)[12]

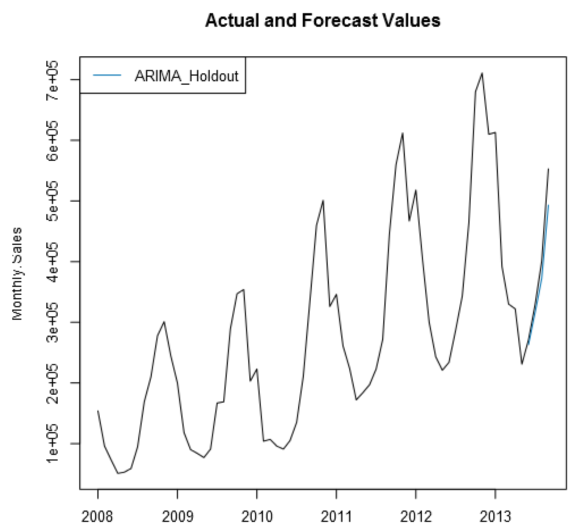
Actual and Forecast Values:

Actual	ARIMA_Holdout
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_Holdout	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA





The average mean error of ARIMA is smaller than that of ETS.
The standard deviation of ARIMA is also smaller than that of ETS.
The average of the absolute values of the residuals (MAE) of the ARIMA model is smaller than that of ETS model.
And also the MASE value of ARIMA is much smaller than that of ETS and it is also significant to forecast future values.
So, according to these two comparison results, I chose ARIMA model as the best fit to forecast future values.

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

Month	Forecast	Forecast_High_95	Forecast_High_80	Forecast_Low_80	Forecast_Low_95
2013-10	754,854	833,336	806,171	703,538	676,373
2013-11	785,854	878,539	846,458	725,251	693,170
2013-12	684,854	789,838	753,499	616,210	579,871
2014-01	687,854	803,839	763,693	612,016	571,869