Project: Forecasting Sales

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/edd0e8e8-158f-4044-9468-3e08fd08cbf8/project

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

ANSWER 1: The dataset meets the criteria of a time series dataset as it covers a continuous period, data is at sequential and at equal intervals between two measurements, each time unit within the period has one data point.
ANSWER 2: Data of the latest 4 months' records should be used as holdout sample (06/13 - 09/13)

Step 2: Determine Trend, Seasonal, and Error components

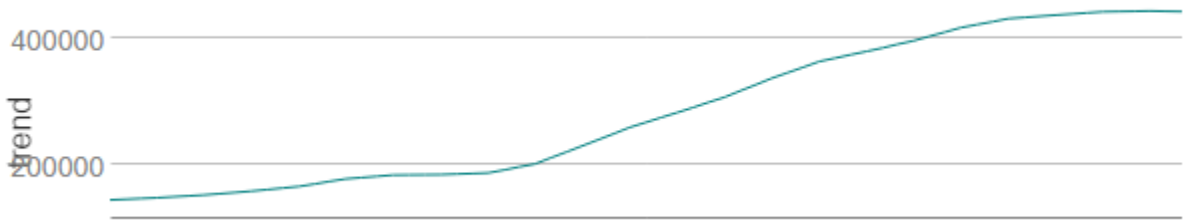
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

Answer this question:

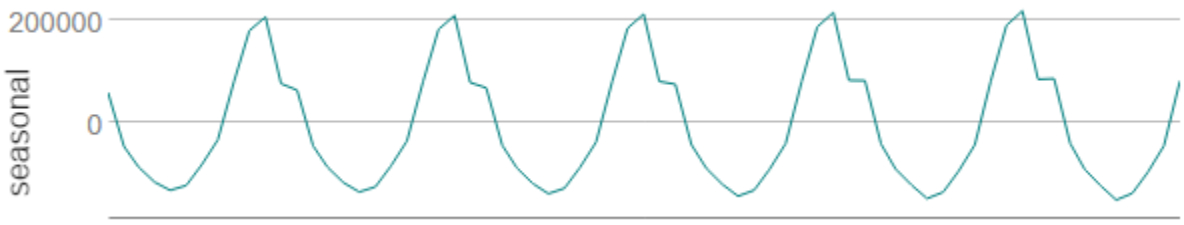
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

   Attached below the plots of error, trend and seasonality. It can be directly derived through the monthly sales and time data provided in the project.

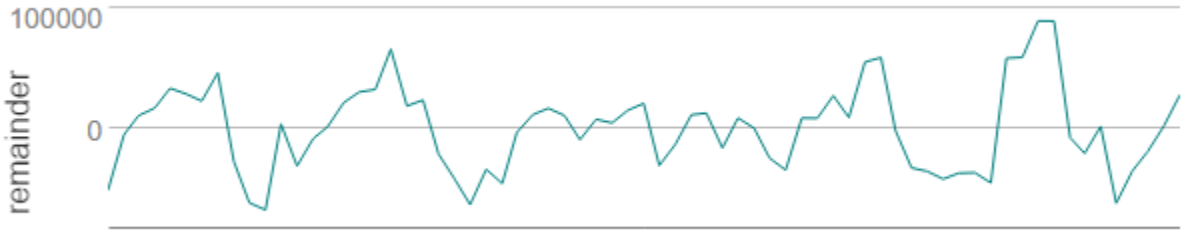
   Trend- Sales are in an uptrend as the graph shows, sales keep on increasing month on month. Also, the trend increases in a linear manner rather than exponentially. Thereby it indicates that we apply additive method in the ETS Model



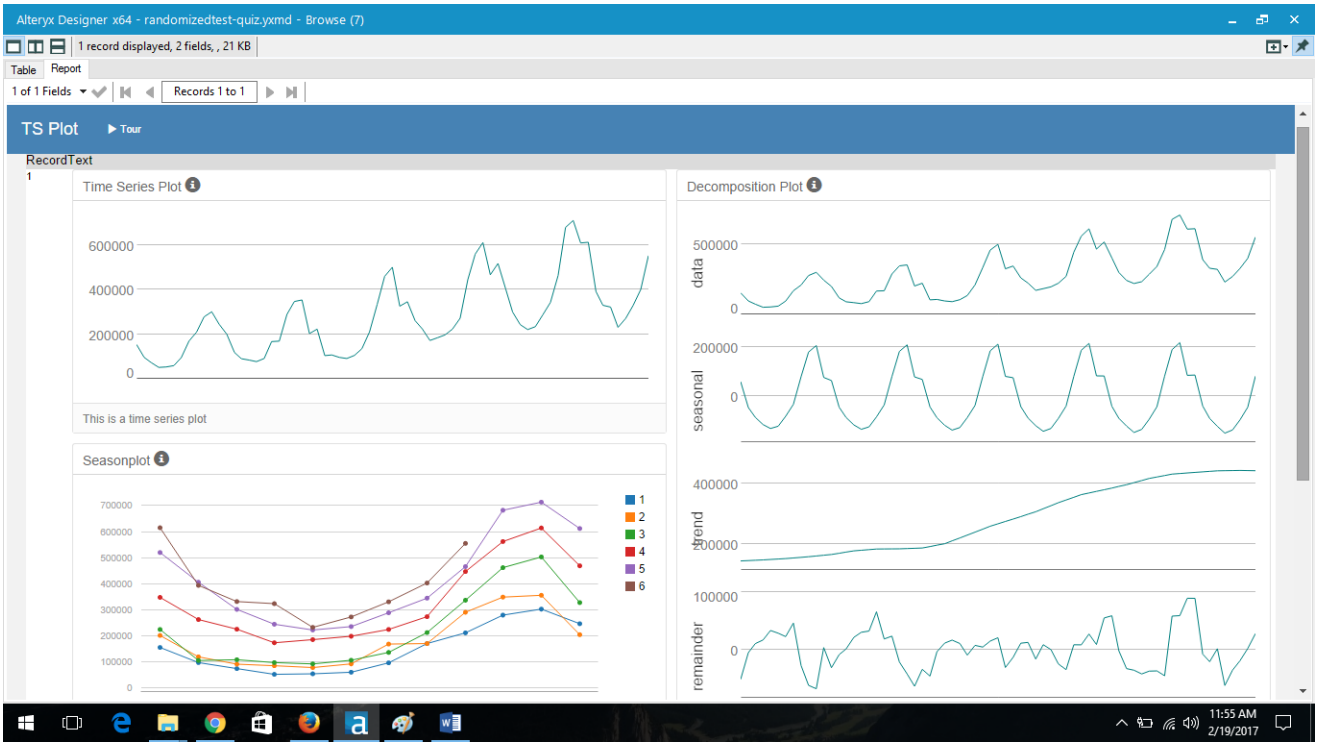
Seasonality- The sales fluctuate up and down every month of every year in a similar pattern/ similar time intervals, thereby indicating presence of seasonality. Also visible is that the magnitude of sales is increasing, thereby suggesting a Multiplicative method for ETS model



Error – The error graph indicates increase in variance, as time goes by (visually determined)



Combined image:



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)

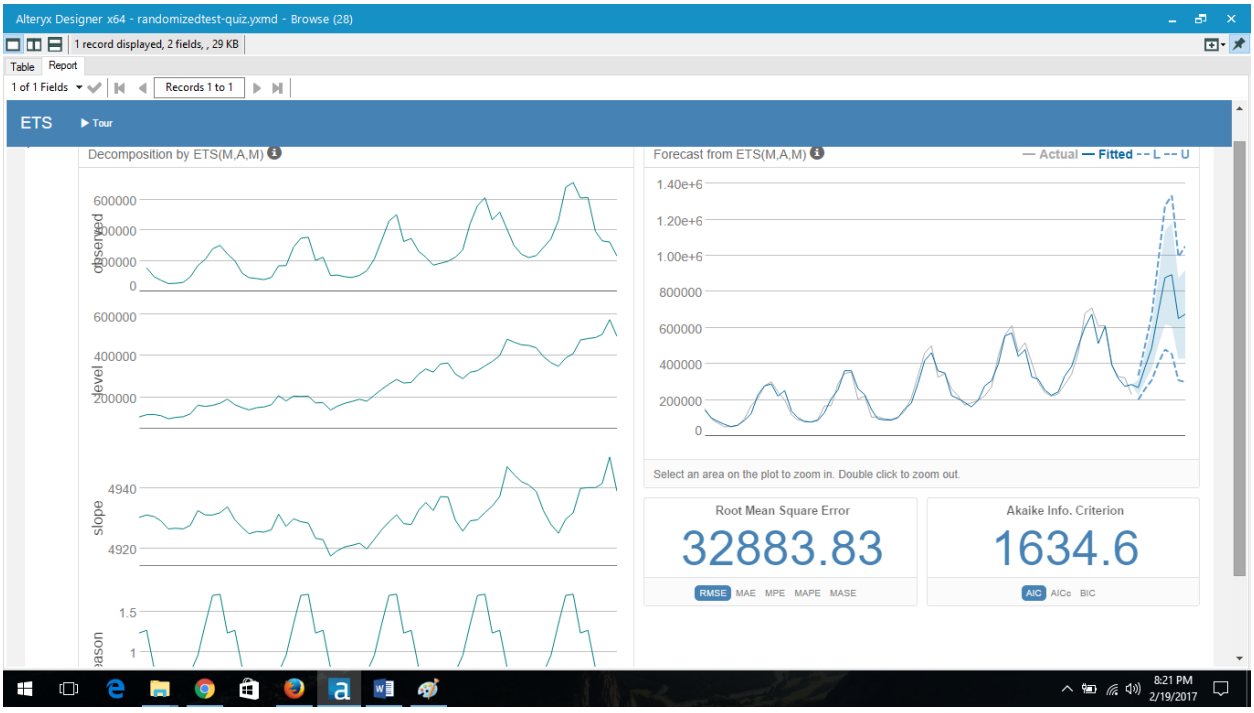
Answer these questions:

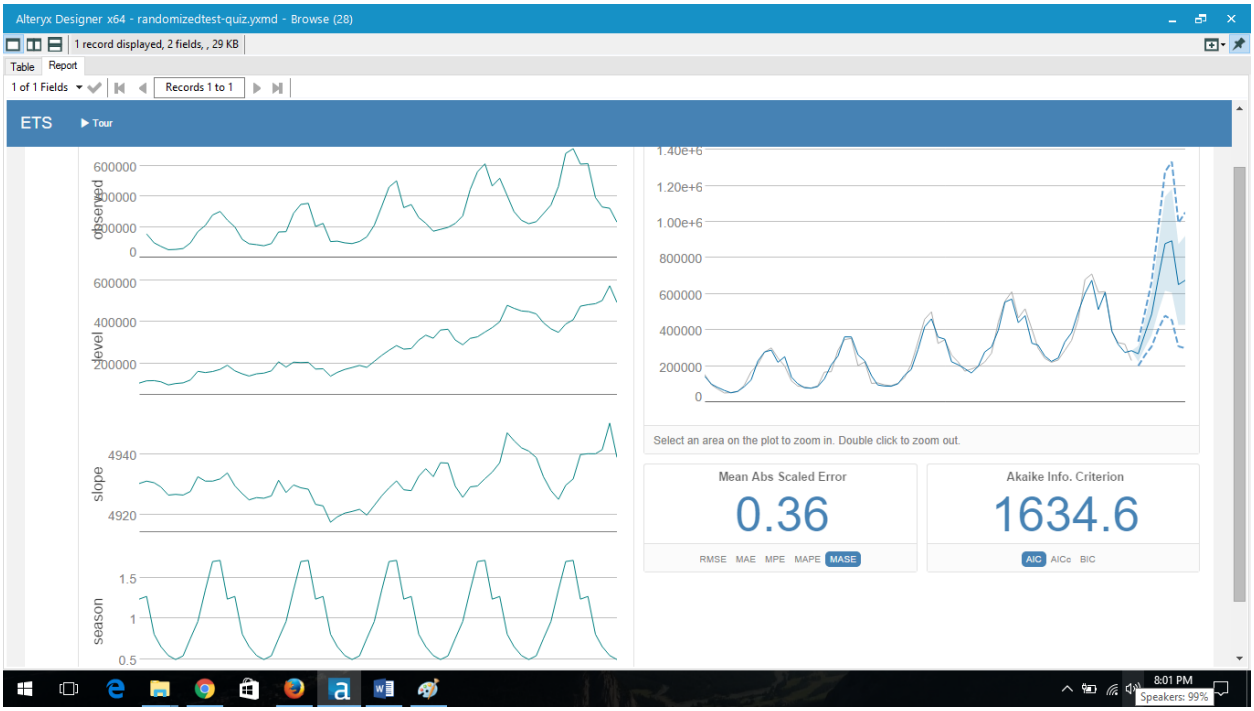
1. What are the model terms for ETS? Explain why you chose those terms.
    a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

ANSWER : ETS MODEL terms can be obtained directly through the Auto function in Alteryx. In this case the applied model is ETS (M, A, M) because
Error keeps increasing in variance -Thus Multiplicative method
Trend keeps increasing linearly rather than exponentially, thus additive method
Seasonality – Increases in magnitude, with passage of time, thereby Multiplicative method

The RMSE, MASE and AIC values are as given in the images below (ETS MODEL). Also attached is a comparison table of actual and forecast values

The RMSE error indicates the difference between values (Sales in this case) predicted by the ETS model and the values actually observed. Closer the range of values, the better is the model MASE error is a measure of the accuracy of forecasts. The lower the value, the better the forecast. This will be useful for comparing this model with the Arima model.





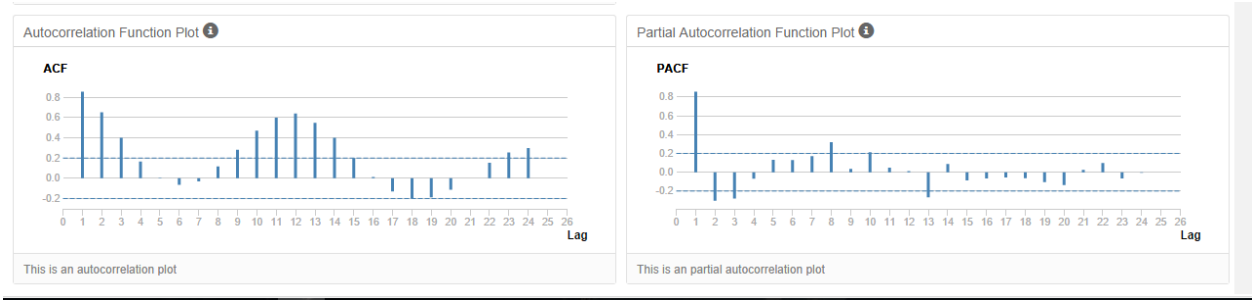
Forecast Graph and values

Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	6	10	718108.632513	893531.884348	832811.731547	603405.533478	542685.380678
2	6	11	750234.986343	986202.59396	904525.91176	595944.060926	514267.378726
3	6	12	559499.665177	767133.101138	695263.871568	423735.458786	351866.229215
4	7	1	566561.867458	804301.884068	722011.708823	411112.026092	328821.850848

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

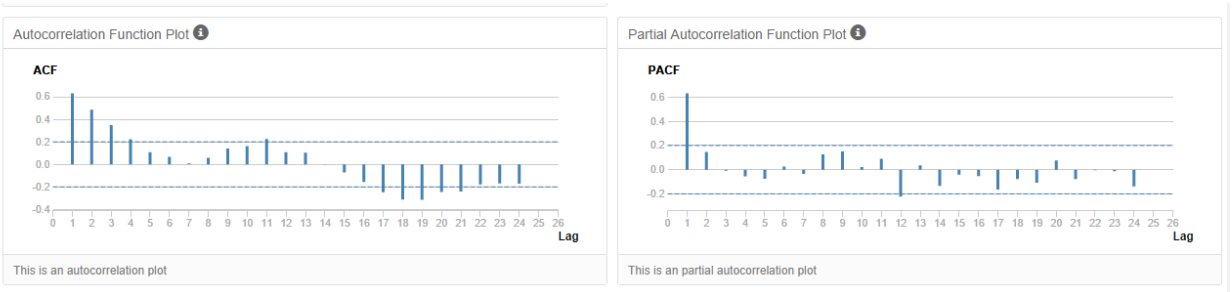
       First let's understand the ACF and PACF graphs,

       Given below is the Time series ACF and PACF graph



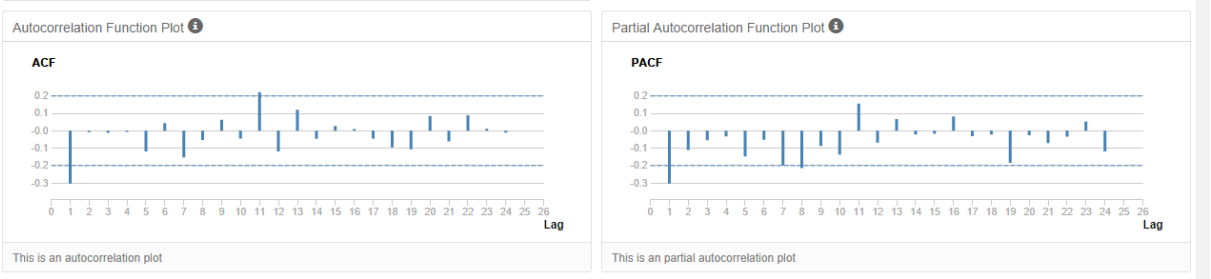
The ACF graph shows gradually decline to zero with seasonal lag but there is high serial correlation, therefore it is required to be seasonally differenced. The PACF graph also shows high serial correlation after lag1, reaffirming the need to difference the dataset.

Seasonal differenced graph



The ACF graph shows positive co-relation at lag 1 and slow decay towards 0 with seasonal lags. However, the serial co-relation of is still high. Therefore, we will need to seasonally difference it, to stationarize the time-series

First seasonal difference

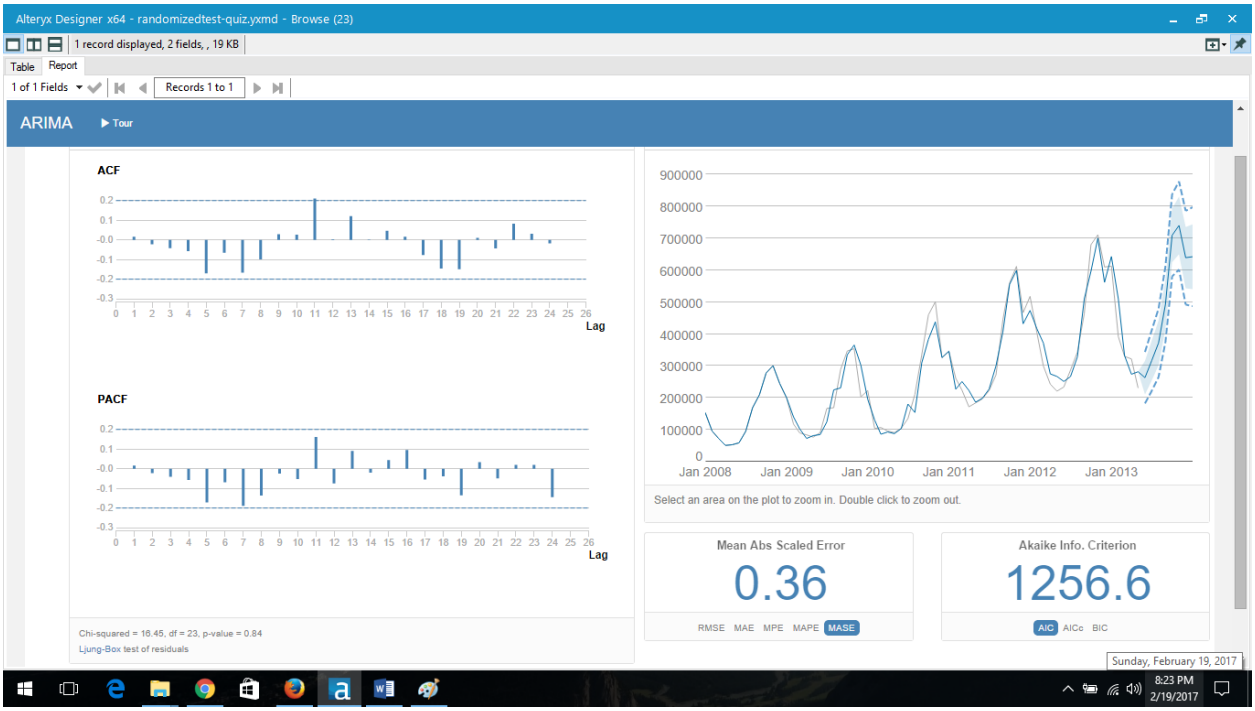


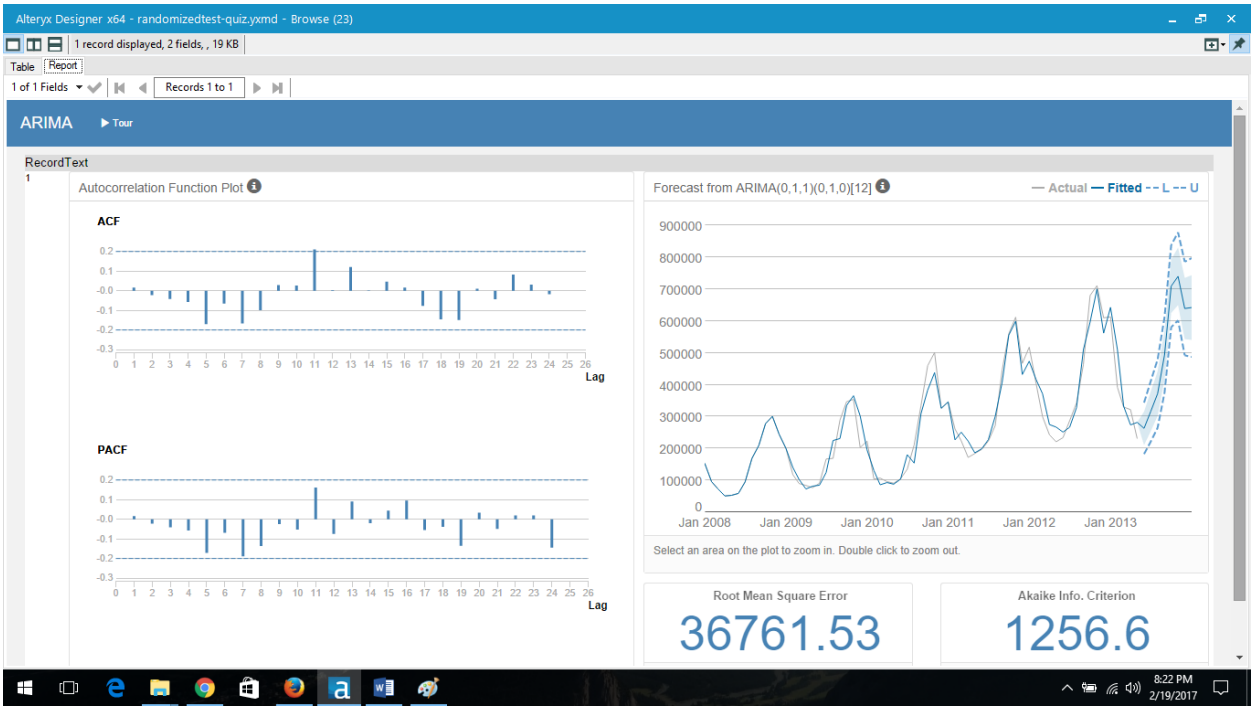
As we can see the serial correlation has now subsidized and the time series is now stationary. Besides The ACF is negative at Lag-1 Thereby suggesting an MR(1) Model and since it took us 1 time to difference & stationarize the data, I = 1,Besides we also seasonally differenced the data once Therefore we get ARIMA(0,1,1) (0,1,0) m. “m” signifies Months which is 12 here.

Also, below is the Time series ACF and DCF plots

ARIMA MODEL terms can also be obtained directly through the Auto function in Alteryx. In this case the applied model is ARIMA(0,1,1)(0,1,0){12}

& the RMSE and MASE values are as given in the images below. Also attached is a comparison table of actual and forecast values





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)

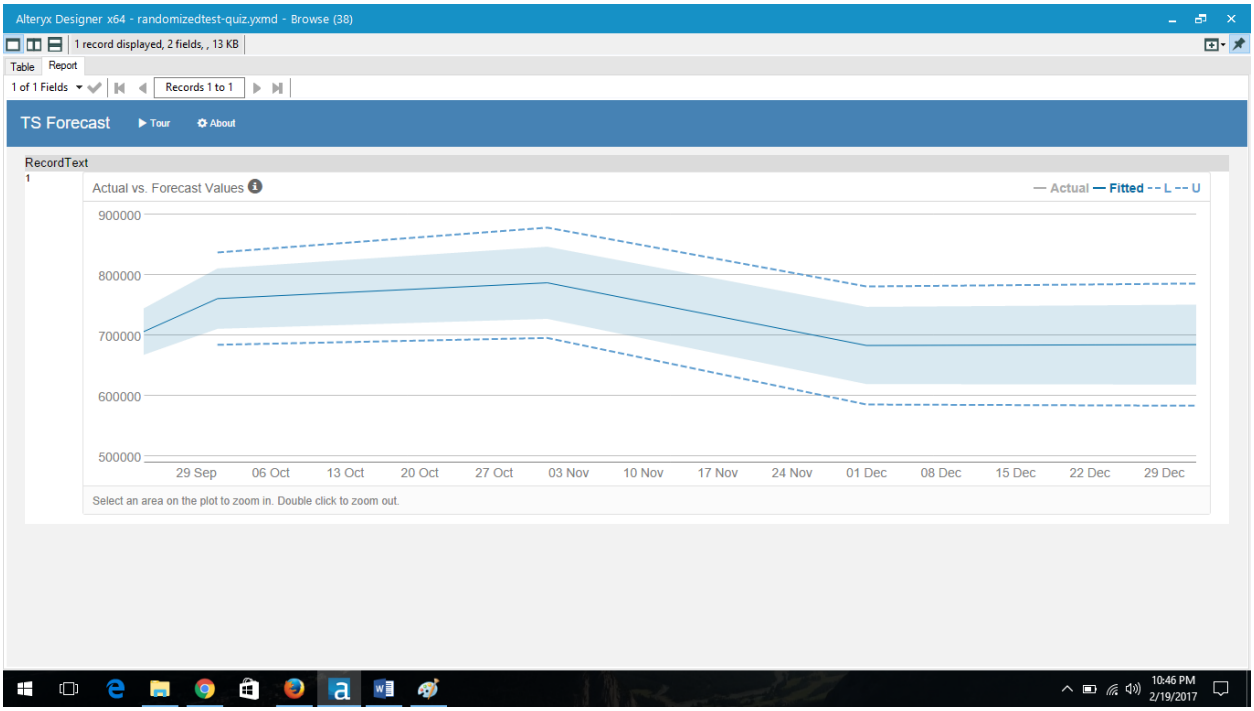
Answer these questions.

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

I chose the ARIMA Model, because of a lower AIC Value signifying it is the best fit & since the accuracy of both the models are same (MASE value)

ARIMA Forecasts (95% & 80%)

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2013	10	754854.460048	833335.856133	806170.686679	703538.233418	676373.063963
2013	11	785854.460048	878538.837645	846457.517118	725251.402978	693170.082452
2013	12	684854.460048	789837.592834	753499.24089	616209.679206	579871.327263
2014	1	687854.460048	803839.469806	763692.981576	612015.938521	571869.450291



.

ARIMA& ETS Model Validation against Holdout sample & measures (against holdout sample only)

Actual and Forecast Values:

Actual	ARIMA
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

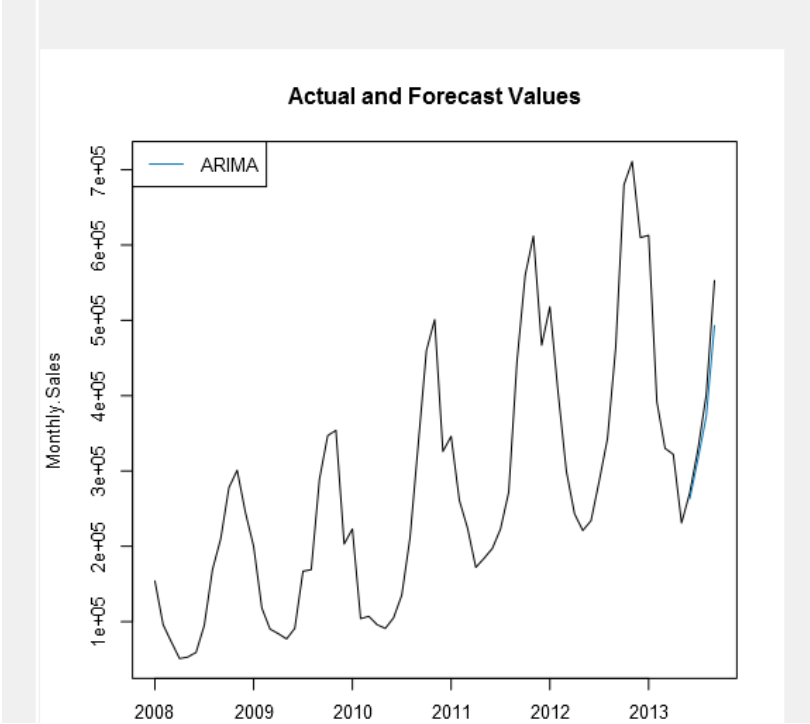
Actual and Forecast Values:

Actual	ETS
271000	268729.50166
329000	378187.04023
401000	488199.64792
553000	691913.69155

Accuracy measures for ARIMA model

Accuracy Measures:

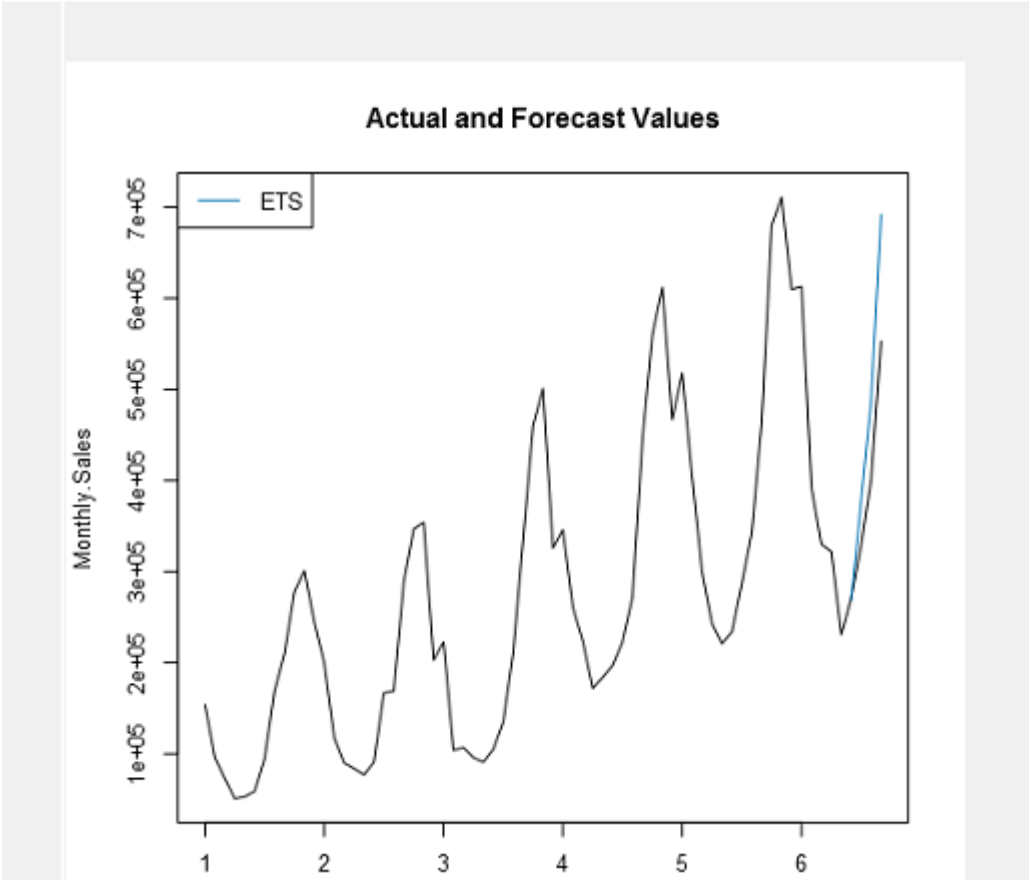
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA



Accuracy Measures for ETS Model

Accuracy Measures:

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



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.