Udacity Business Analyst Nanodegree Project: Forecasting Sales

Jason Grenig

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 dataset meets time series criteria based on 4 key components:
   a. Continuous data between January 2008 and September 2013.
   b. There are sequential measurements (ex. monthly sales).
   c. There is equal spacing between every 2 consecutive measurements.
   d. Each time unit has 1 data point (ex. each month has 1 value)
2. Which records should be used as the holdout sample?
   Since we're predicting the next 4 months in sales, the last 4 records (last 4 months of sales) should be used for the holdout sample. This is June 2013 to September 2013.

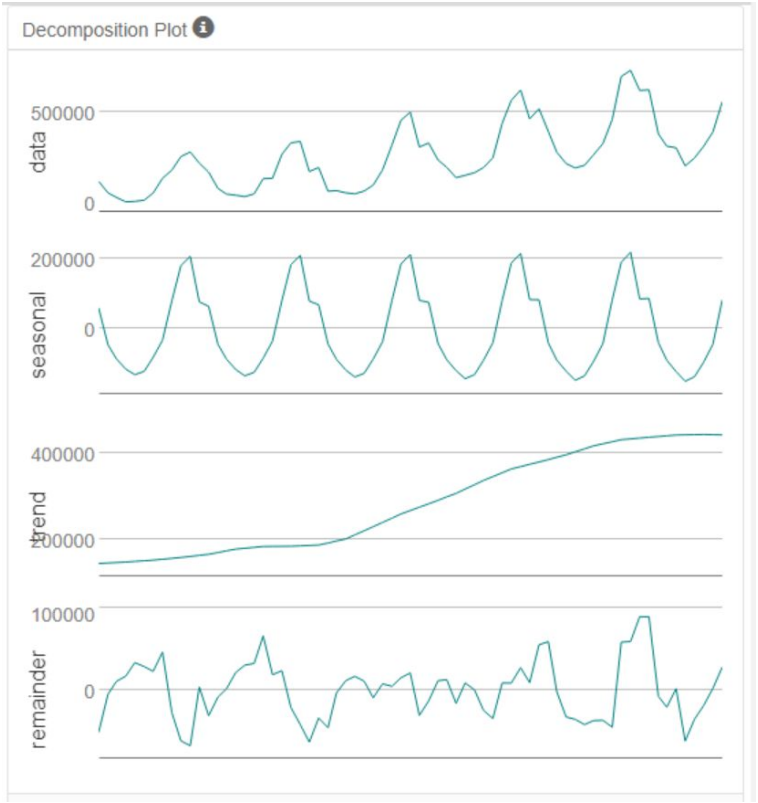
Step 2: Determine Trend, Seasonal, and Error components

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.
   Based on the below Time Series and Decomposition plots, we can interpret the following:
   - **Trend -** there is a linear, upward trend in our video game sales.
   - **Seasonality -** video game sales peak in November in all 6 years in our dataset. Therefore we have seasonality. The seasonality also increases year over year.
   - **Error -** Looking at the Decomposition Plot, we see that the error is variable over time.

Alteryx Time Series Plot and Seasonplot



Alteryx Decomposition Plot

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 for the ETS model are (M,A,M):
   Error - use multiplicative terms because the error is variable over time.
   Trend - use additive terms because the trend is linear.
   Seasonality - use multiplicative terms because the peaks increase year over year, implying an exponential relationship.
   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

   We'll run the ETS model with and without dampening, then perform an internal validation to see which is the better fit with less complexity.

   ETS(M,A,M), results without dampening:
   - Akaike Information Criterion (AIC) = 1639.7367
   - The Root-Mean-Square Error (RMSE) = 32992.7261
   - Mean Absolute Scaled Error (MASE) = 0.3727

Method:
ETS(M,A,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

Information criteria:

AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

Alteryx Summary of Time Series Exponential Smoothing Model - Without Dampening

ETS(M,Ad,M), results with dampening:
- Akaike Information Criterion (AIC) = 1639.4650
- The Root-Mean-Square Error (RMSE) = 33153.5268
- Mean Absolute Scaled Error (MASE) = 0.3675

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

Alteryx Summary of Time Series Exponential Smoothing Model - With Dampening

Internal Validation Summary:
The AIC value is used to measure the relative quality of a model. It deals with the tradeoff between 'Goodness of Fit' and Complexity of models.
The results show that the AIC is slightly lower on the dampened model suggesting it is a better fit with less complexity.

The RMSE value represents the sample standard deviation of the differences between predicted values and observed values. This measurement is good to use when comparing models because it shows how many deviations from the mean the forecasted values fall.
Examining the RMSE, the non-dampened model is lower by 160.8007. This means that the non-dampened model has a slightly narrower range of possible forecast values (meaning it's better in this regard).

The MASE value is the mean absolute error of the model divided by the mean absolute value of the first difference of the series. Therefore, it measures the relative reduction in error compared to a naive model. Ideally its value will be significantly less than 1.0, but is relative to comparison across other models for the same series. Since this error measurement is relative and can be applied across models, it's accepted as one of the best metrics for error measurement.
Comparing the MASE values of the dampened and non-dampened models, we can see the dampened model is slightly lower (better) by about 0.005, however both models have a MASE value less than 1.0.

Typically the best model has the lowest AIC value. However, when it's close we can compare the calculated errors. Based on the below accuracy measures, we now see that the dampened model has lower RMSE and MASE values, meaning it's a better overall fit.

Actual and Forecast Values:

Actual	ETS_M_A_M_
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_M_A_M_	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066

ETS(M,A,M) Alteryx Time Series Comparison of Actual and Forecast Values

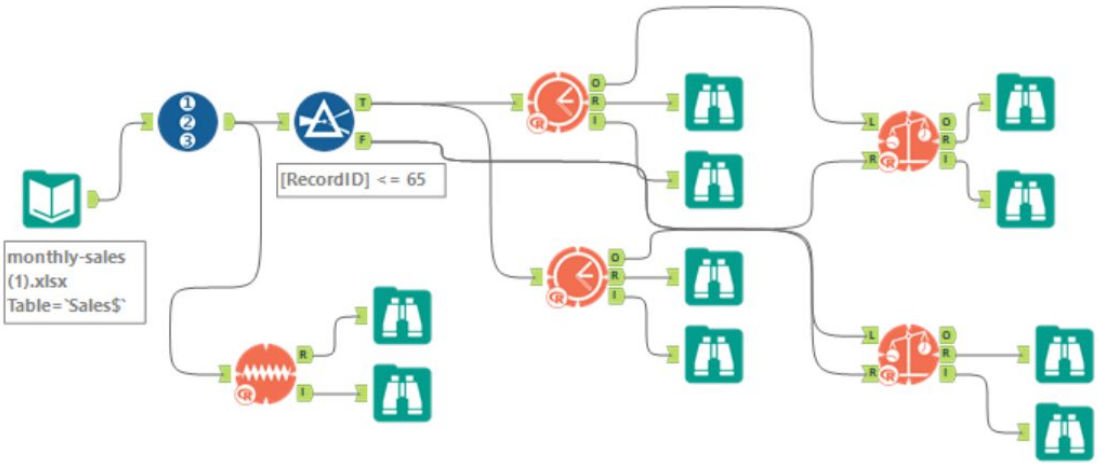
Actual and Forecast Values:

Actual	ETS_M_Ad_M_
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_M_Ad_M_	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116

ETS(M,Ad,M) Alteryx Time Series Comparison of Actual and Forecast Values

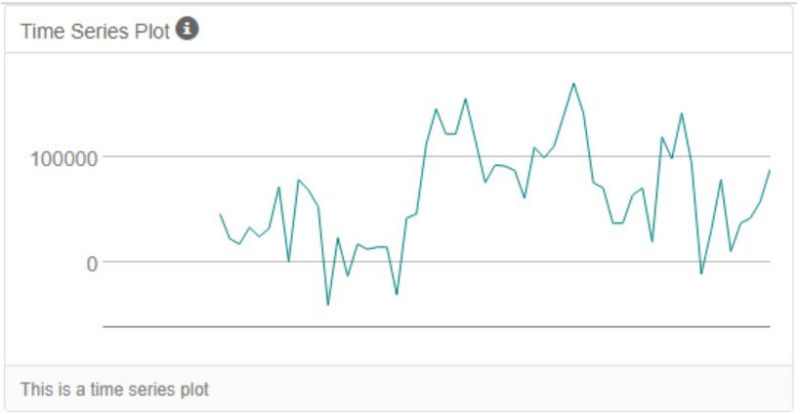


Alteryx Workflow for ETS Model

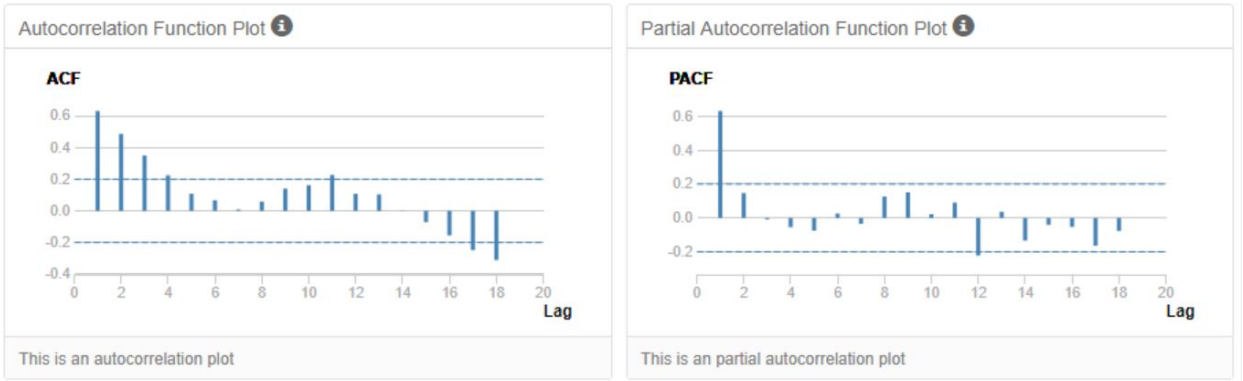
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.

   Since out data exhibits seasonality, we'll need to use a seasonal ARIMA model denoted **ARIMA(p,d,q)(P,D,Q)m**, where m refers to the number of periods in each season, the lowercase p,d,q refer to the autoregressive, differencing, and moving avg terms, and the uppercase P,D,Q refer to the same terms, but for the seasonal part of the ARIMA model.

   Below are the Time Series plot along with the ACF and PACF plots for the Seasonal Differencing of the dataset. After taking the seasonal difference of the data, we can see in the ACF and PACF plots that the data is not yet stationarized. Many of the lags (1, 2, 3, 4, 17, and 18 on the ACF, and 1 and 12 on the PACF) extend beyond the dashed line, meaning there is autocorrelation in the residuals. We also know the data isn't stationarized because the mean of the Time Series plot is not 0. We'll need to take the First Seasonal Difference of the data to see if it stationarizes.



Alteryx Time Series Plot of the Seasonal Differencing



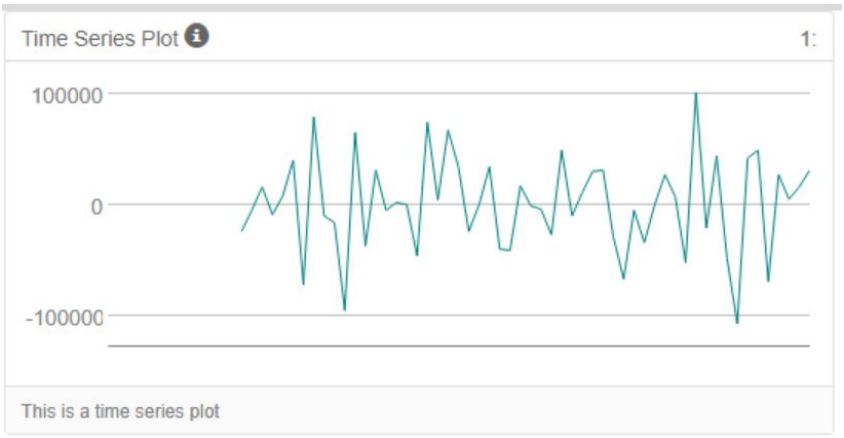
Alteryx ACF and PACF of the Seasonal Differencing

Below are the Time Series plot along with the ACF and PACF plots for the First Seasonal Differencing of the dataset. We can now consider the data stationarized since

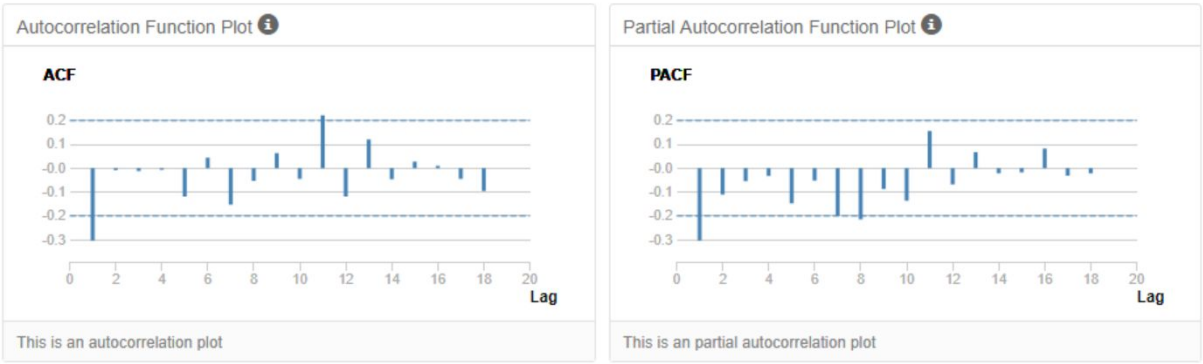
the Time Series plot mean is closer to 0 and lags on the ACF and PACF plots are generally within the dashed lines after lag 1.

So in determining the terms for the model, we have:

- MA (1) terms for the non-seasonal component, and p (0), P (0), q (1), Q (0), due to negative autocorrelations at lag 1 in both the ACF and PACF plots, then the PACF plot having a gradual smoothing from lag 1 and no other major autocorrelation.
- We used First Seasonal Differencing to stationarize the dataset, so we have d (1) and D (1)
- Then m = 12 since we have 12 periods (12 months)
- Finally our model looks like: ARIMA(0,1,1)(0,1,0)12



Alteryx Time Series Plot of the First Seasonal Differencing



Alteryx ACF and PACF of the First Seasonal Differencing

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

   After running the ARIMA(0,1,1)(0,1,0)12 model, our results are:

   - RMSE = 36761.5282
   - MASE = 0.3646. This is less than 1, which indicates a good fit.

- AIC = 1256.5967

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

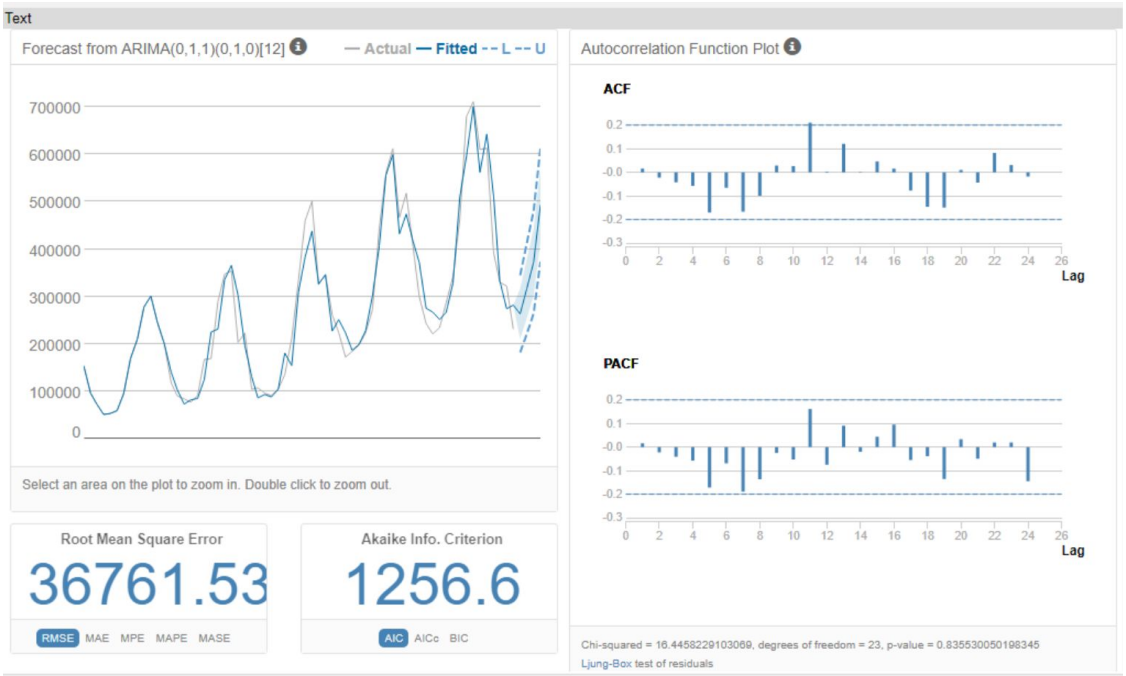
In-sample error measures:

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

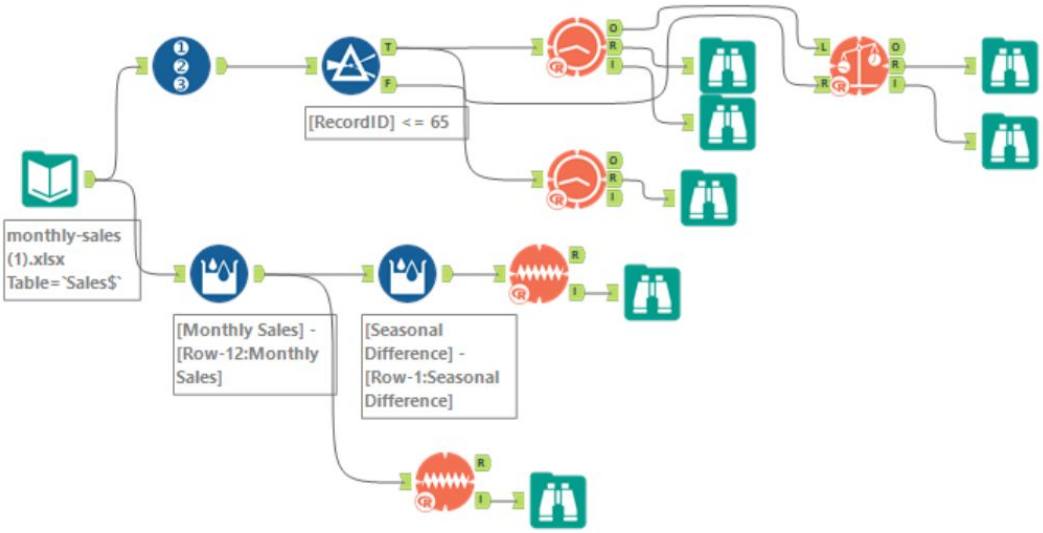
Alteryx results for ARIMA(0,1,1)(0,1,0)[12]

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

Looking at the ACF and PACF plots, all of the lags fall within the dashed lines (lag 11 of the ACF is just barely outside), therefore we have removed autocorrelation and don't need to further adjust the model.



Alteryx re-graphed ACF and PACF results for ARIMA(0,1,1)(0,1,0)[12]



Alteryx Workflow for ARIMA Model

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.
   To find the best model, we need to compare the Actual and Forecast values for both the ETS and ARIMA models.

   Based on the below error measurements we can see that:
   - Only the ARIMA model has a MASE value of less than 1. The ETS MASE value is close to 1, but ideally should be less than 1 to be deployed.
   - For the in-sample error measures, the RMSE is lower in the ETS, however looking at the forecast error measurements, the RMSE value is almost twice as low in the ARIMA model compared to the ETS model. The ETS performed better on the training data, but worse against the holdout data.
   - The AIC value of the ETS is 1639.465 compared with 1256.5967 of the ARIMA. The ARIMA model performed better here as well.

   The ARIMA model is more accurate, therefore we should deploy this model to forecast the next 4 months of video game sales.

Actual and Forecast Values:

Actual	ETS_M_A_M_
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

	Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_M_A_M_		-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066

Alteryx ETS Model Actual and Forecast Values

Actual and Forecast Values:

Actual	ARIMA_0_1_1__0_1_0_12
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

	Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_0_1_1__0_1_0_12		27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

Alteryx ARIMA Model Actual and Forecast Values

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

Alteryx Summary of Time Series Exponential Smoothing Model - With Dampening

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:

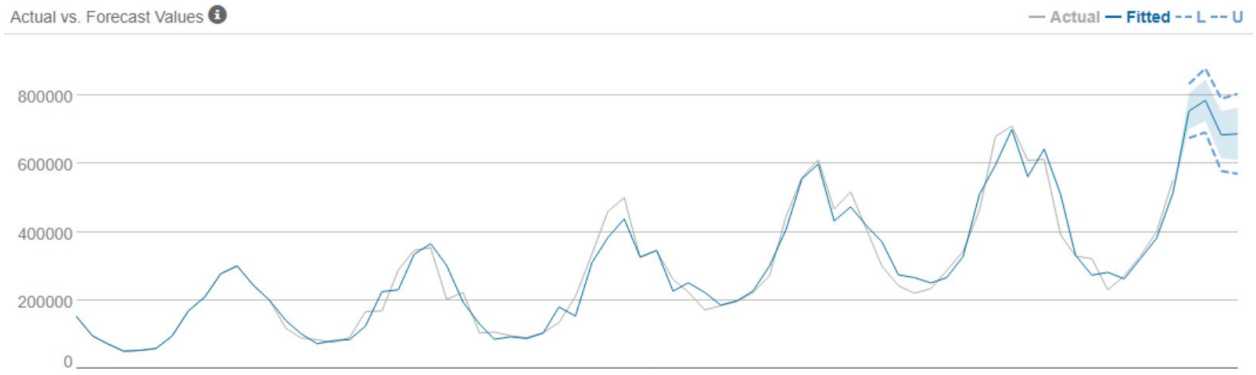
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

Alteryx results for ARIMA(0,1,1)(0,1,0)[12]

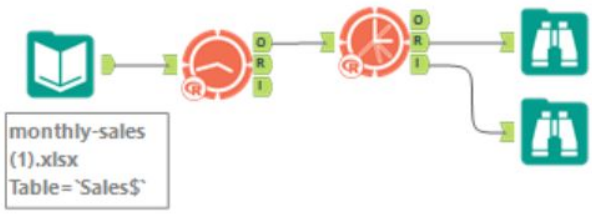
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.
   The next 4 periods are October 2013 - January 2014, so we'll use our ARIMA model to forecast these periods.
   The forecasted values for the next 4 periods are as per the below table.



Alteryx Time Series Forecast for Oct. 2013 - Jan. 2014 using our ARIMA Model

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
6	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
6	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
6	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
7	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462

Alteryx ARIMA Model Forecast Results based on 95% and 80% Confidence Intervals



Alteryx Workflow for ARIMA Model Forecast