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.

Yes, the dataset meets criteria of a time series dataset.
It is taken across a continuous time interval and in yyyy-mm format.
It is sequential.
It is separated by monthly intervals.
It contains at most one data point for each interval.

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

The final prediction will be forecasting for the next 4 months of sales.
Hence, the holdout sample will be the last 4 months of sales data from 2013-06 to 2013-09.

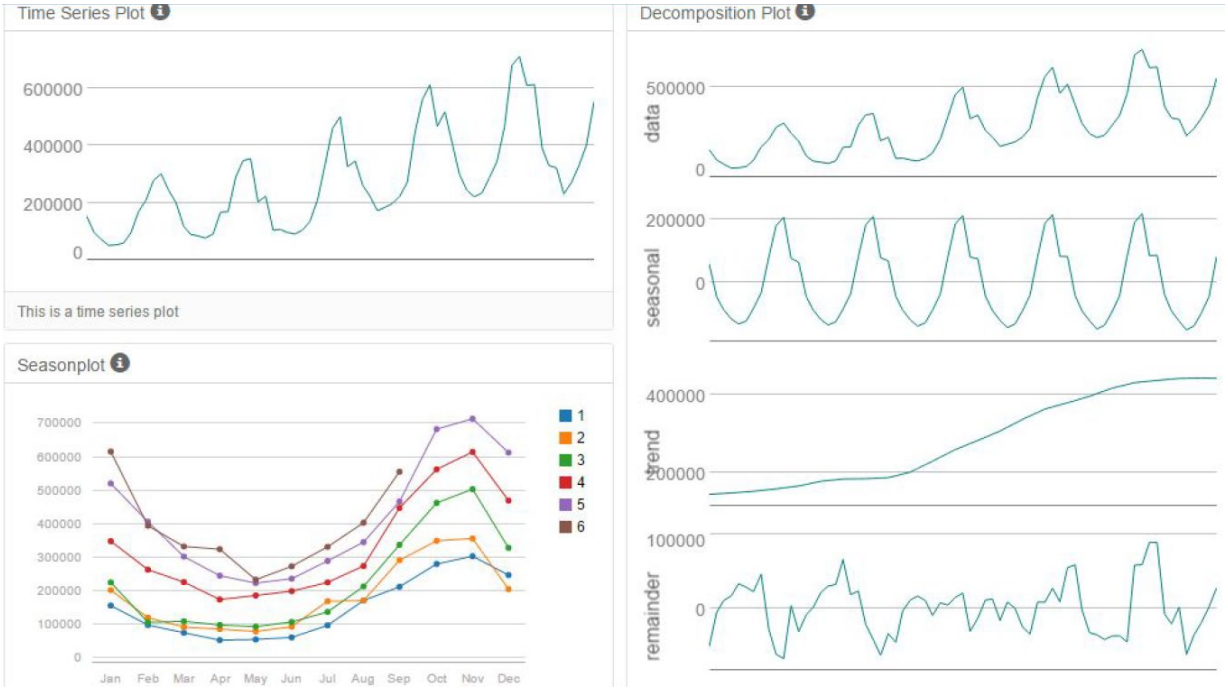
Step 2: Determine Trend, Seasonal, and Error components

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 time series and decomposition plots are generated using TS plot tool.



Picture 2.1 Alteryx-generated plots.

The trend is upward and linear.
The seasonality has peaks and valleys and increasing over time.
The error has irregular pattern.

Step 3: Build your Models

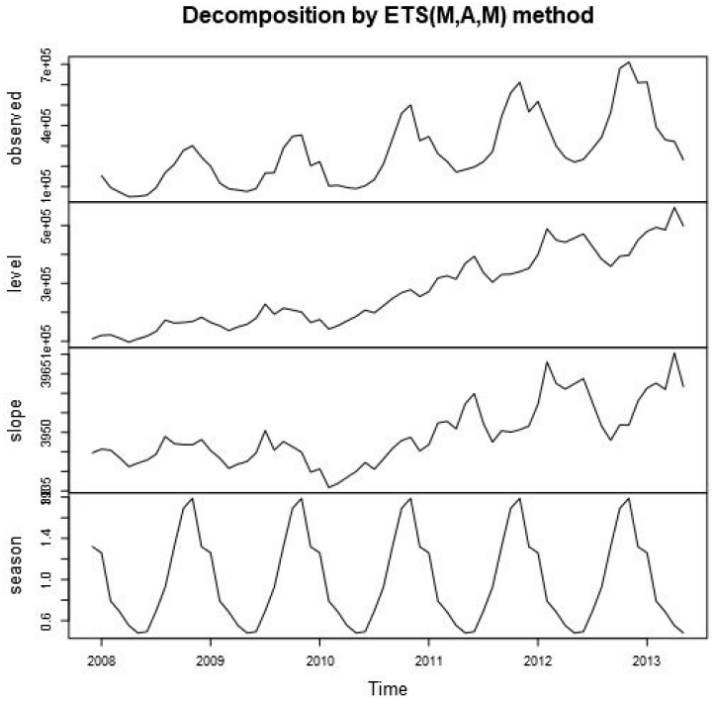
Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)

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.

ETS Model Terms
Error - Multiplicative - Irregular pattern
Trend - Additive - Upward and Linear
Seasonality - Multiplicative - Increasing peaks and valleys

I chose ETS Model (M,A,M).
ETS model is run with a holdout sample of 4 months.



Summary of Time Series Exponential Smoothing Model MAM

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

Picture 3.1 ETS Model Part 1.

Two key components to look at are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models.
ETS model has RMSE 32992.7 and MASE 0.37 and AIC 1639.7.

MASE value is under the 1.0 threshold(commonly accepted threshold for model accuracy), which is good.

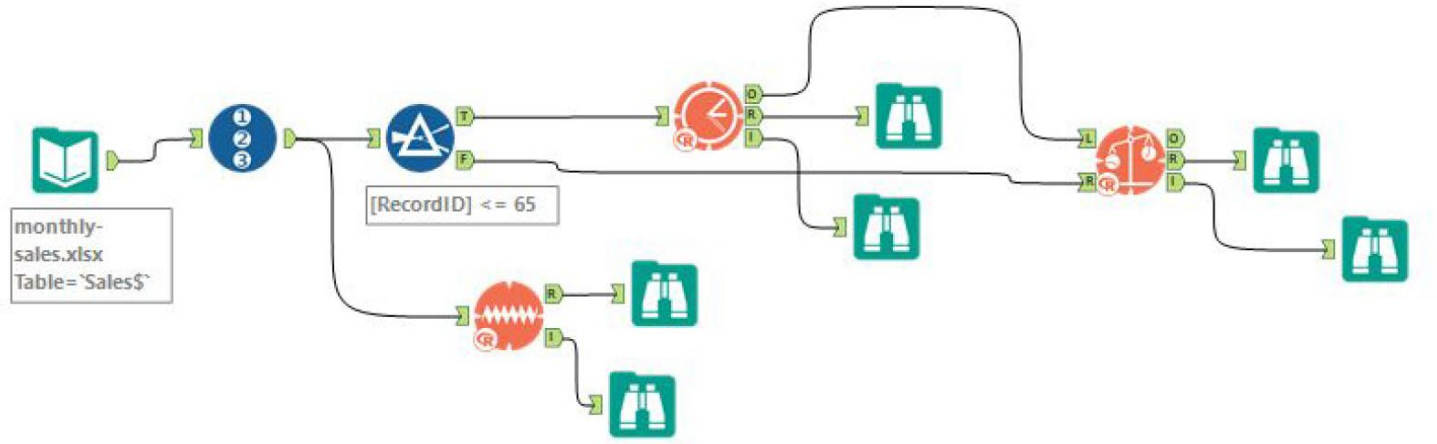
Actual and Forecast Values:

Actual	MAM
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

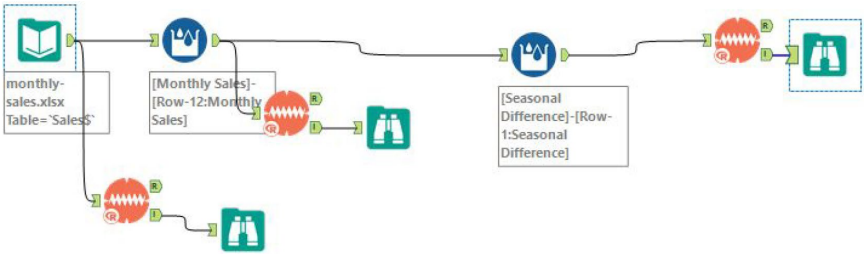
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MAM	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066	NA

Picture 3.2 ETS Model Part 2.



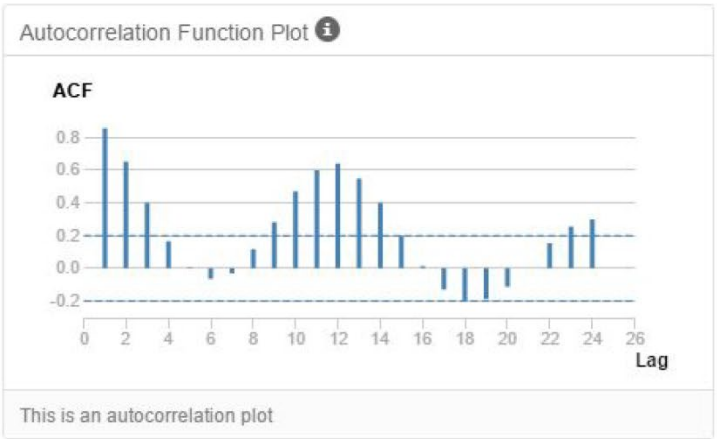
Picture 3.3 Alteryx Workflow to generate ETS Model Report.

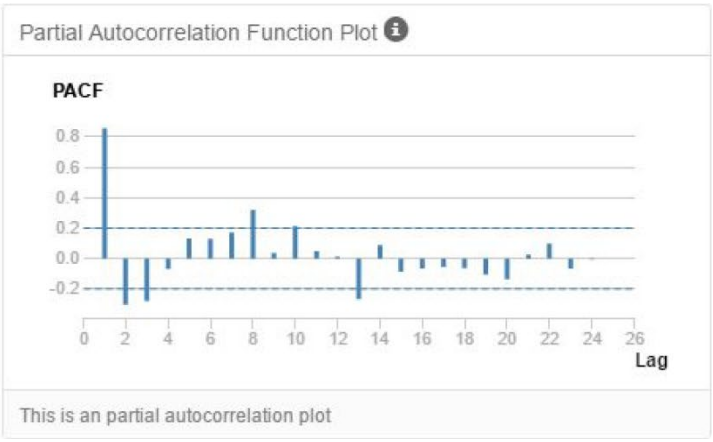
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
    a. Describe the in-sample errors.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



Picture 3.4 Alteryx Workflow to generate reports.

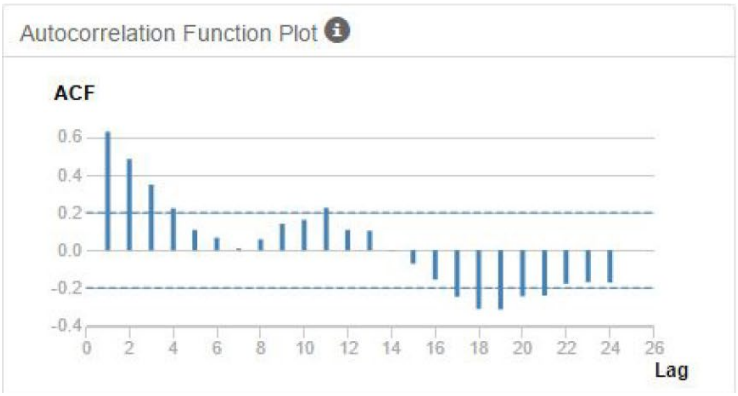
The original time series plot shows seasonality.

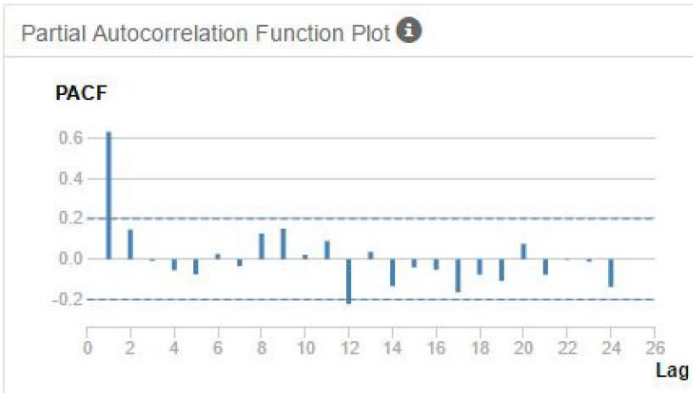




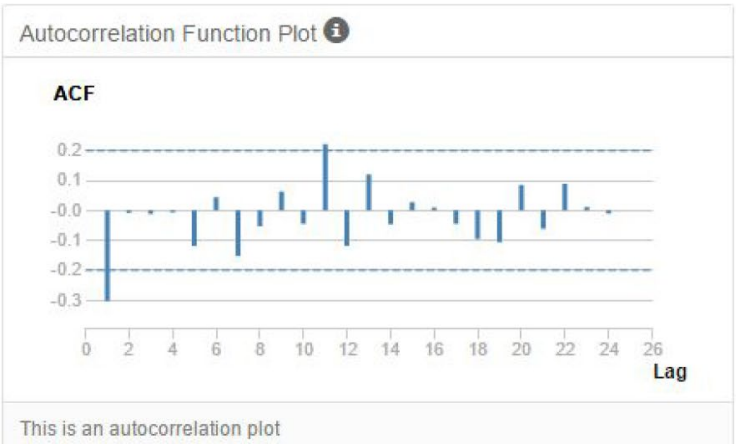
The ACF plot and PACF plot show positive correlation at lag 1.

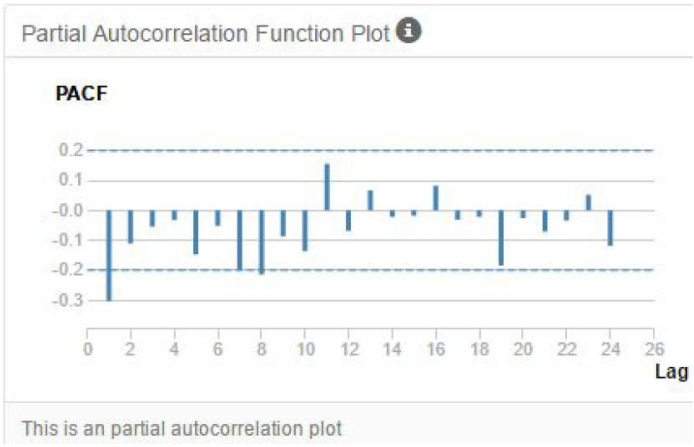
So differencing for the 12 months is done. This will remove seasonality.
ACF and PACF plot after seasonal difference is applied.





ACF and PACF still show positive correlation at lag 1.
A seasonal first difference is applied and ACF plot does not show strong correlation.





ARIMA model (0,1,1) (0,1,0) is used as lag 1 is negative and the number of period is 12 months.

Information Criteria:

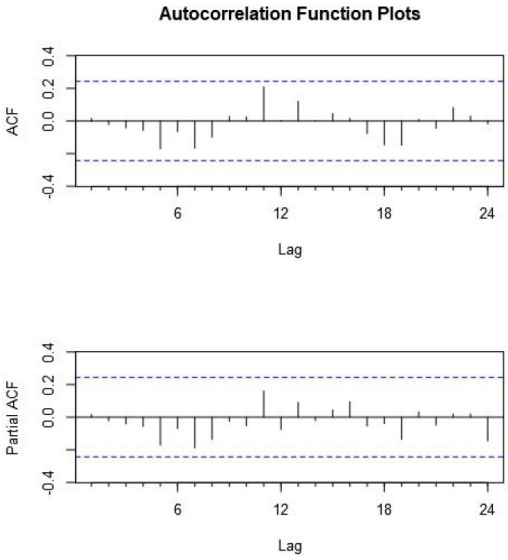
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

Two key components to look at are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models.

ARIMA model has RMSE 36761.5 and MASE 0.36 and AIC 1256.59
MASE is at 0.36 which means it under the commonly accepted threshold of 1.0.



Actual and Forecast Values:

Actual	ARIMA
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

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

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.

ETS Model

Actual and Forecast Values:

Actual	MAM
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

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

ARIMA Model

Actual and Forecast Values:

Actual	ARIMA
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

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

ARIMA model is better at forecasting sales using holdout sample.

The RMSE for ARIMA is **33999.79** compared to ETS' RMSE at **74101.16** . ARIMA's MASE value of **0.4532** is also lower than ETS' MASE value of **1.0066** . ARIMA model has lower AIC of **1256.59** and ETS has higher AIC of **1639.73**.

ARIMA

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

ETS

Summary of Time Series Exponential Smoothing Model MAM

Method:
ETS(M,A,M)

In-sample error measures:

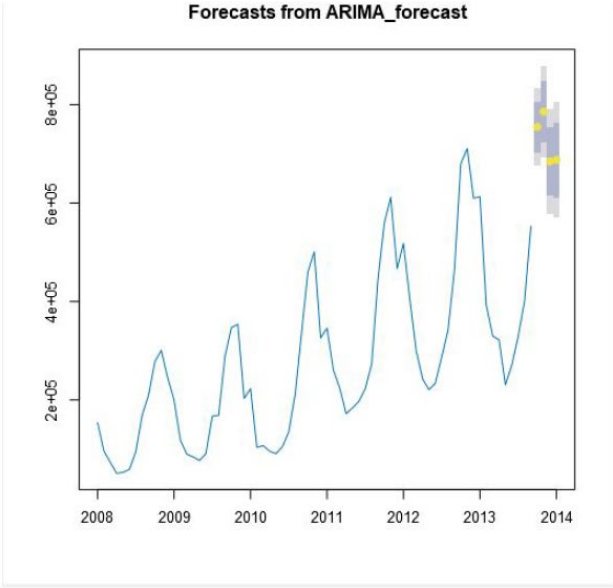
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

Information criteria:

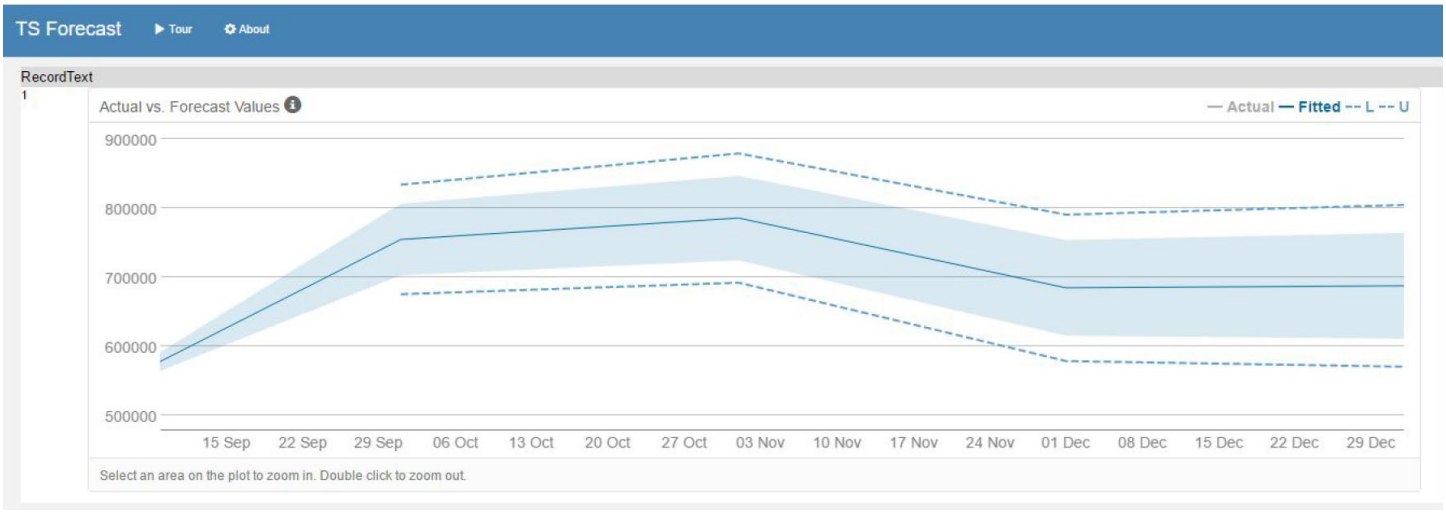
AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

Further investigation shows that the MAPE and ME of the ARIMA model are lower than the ETS. This suggests that, on average, the ARIMA model misses its forecast by a lesser amount.
ARIMA model AIC is lower than that of ETS. Lower AIC reflects better model accuracy.
Hence, ARIMA model is chosen.

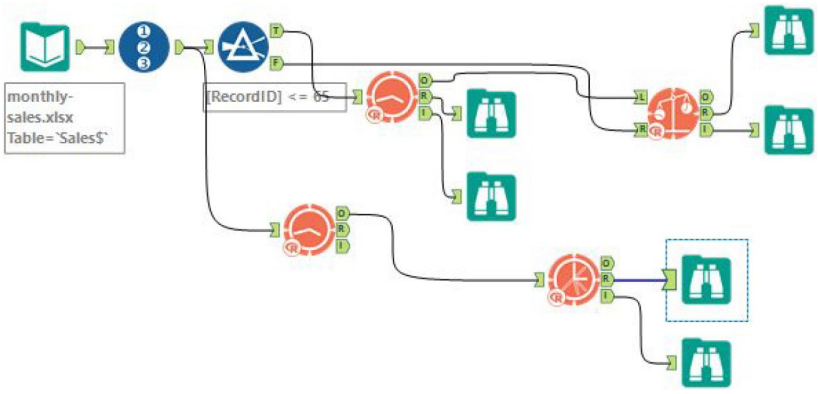
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2013	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
2013	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
2013	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
2014	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462



Alteryx Workflow ARIMA Model



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.