

BENJAMIN OWUSU BEDIAKO

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

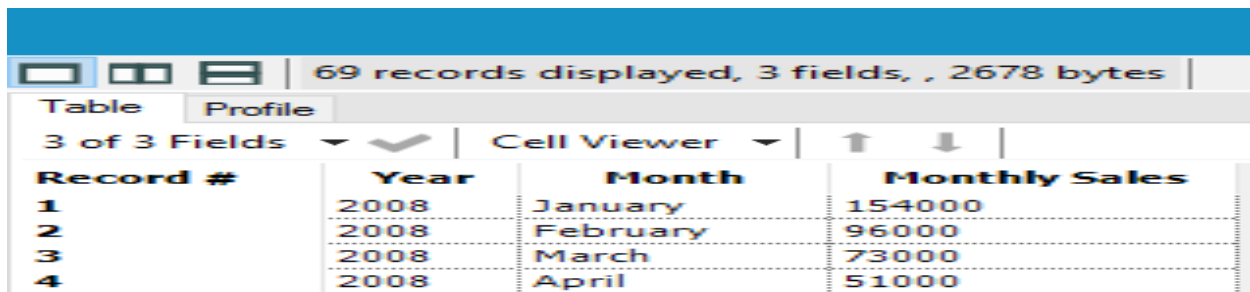
Look at your data set and determine whether the data is appropriate to use time series models. Determine which records should be held for validation later on (250 word limit).

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.

Answer:

Yes, dataset really does meet the criteria of a time series dataset by grouping the year, month and monthly sales for better clarification and the visualization can be shown below:



Record #	Year	Month	Monthly Sales
1	2008	January	154000
2	2008	February	96000
3	2008	March	73000
4	2008	April	51000

All four key characteristics of a time series data and the data is a time series because of the following characteristics:

- Continuous time interval
- Sequential measurements across that interval
- Equal spacing between every two consecutive measurements
- Each time unit within the time interval has at most one data point

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

Answer:

The last four periods (four months) should be used as the holdout samples for the periodic forecast.

Step 2: Determine Trend, Seasonal, and Error components

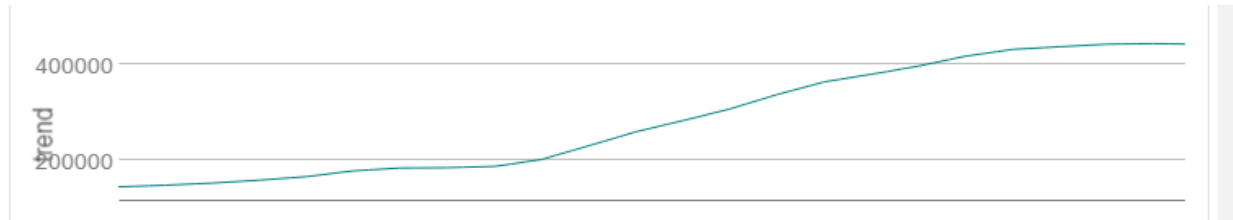
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. (250 word limit)

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.

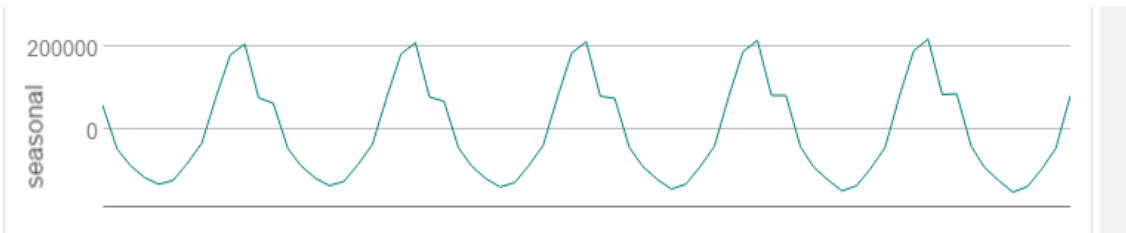
Answer:

Trend component: - Looking at the trend time series plot below, shows the dataset upward tendency within a time frame.



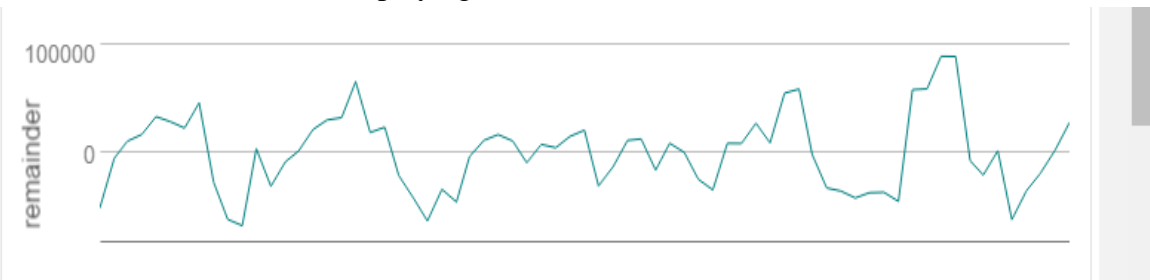
Seasonal component

The dataset shows a seasonal pattern of time series displaying a periodic fluctuations.



Error component

The dataset is inconsistent displaying a variation of error.



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

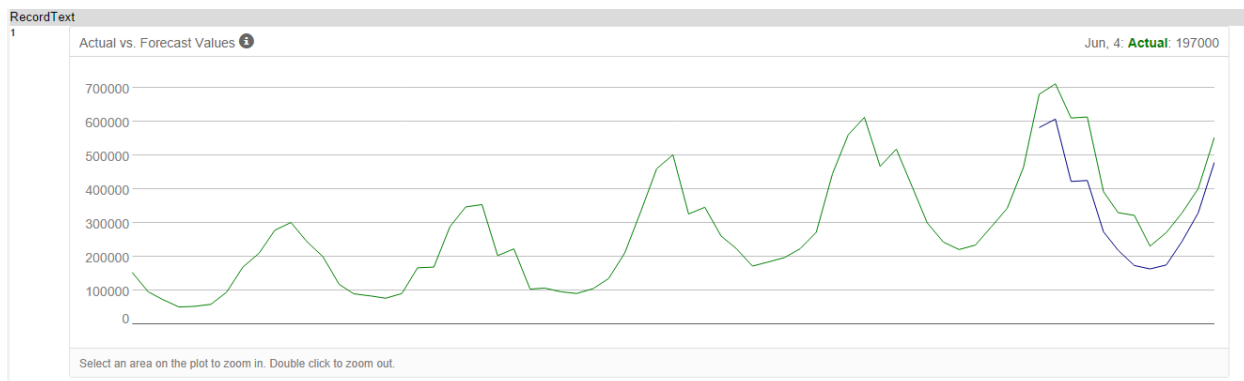
Answer:

ETS framework shows an automatic way of selecting the method and the model terms are **Error** (it is a multiplicative since there is a change in variance over time in the remainder plot), **Trend** (it is additive since there is a linear trend plot) and **Seasonal** (it is a multiplicative since there is a variations in seasonal plot).

(a): In-sample errors

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

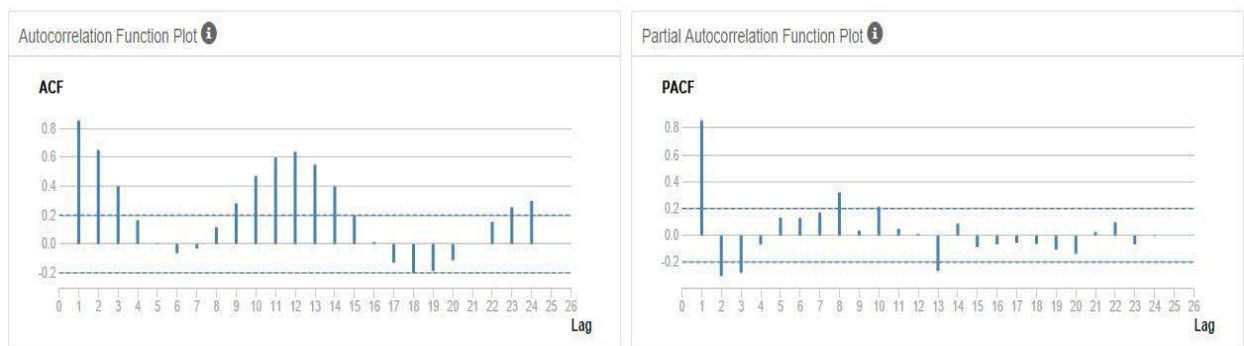


2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.

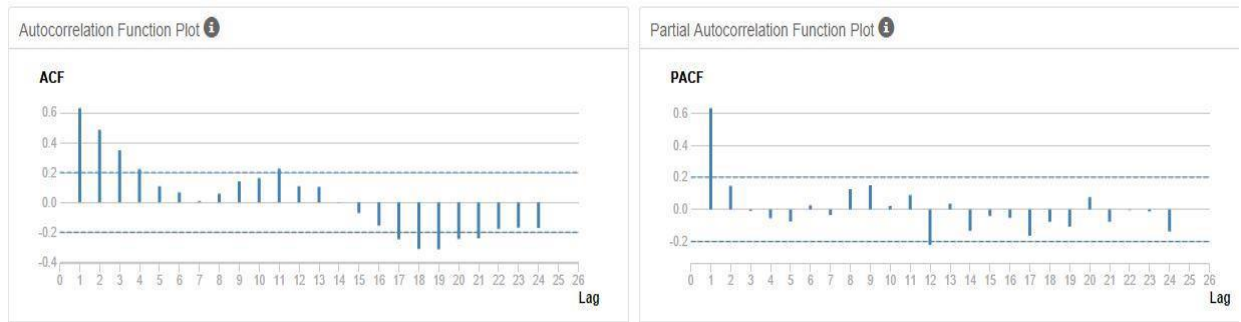
Answer:

In order to answer:

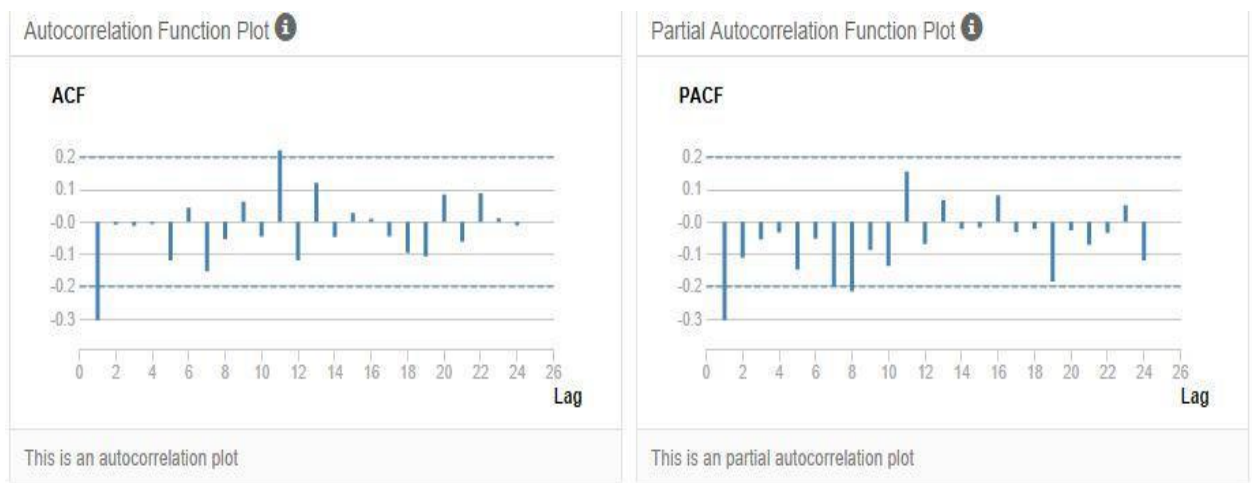
- (i) The ACF and PACF time series visualizations show that the ACF decreased slightly whiles the seasonal increased resulting in a serial correlation.



- (ii) Furthermore, the difference on seasonal of ACF and PACF visualizations are below:



- (iii) After removing the significant correlation, AR and MA terms will take care of the rest of the correlation.



After taking the seasonal difference and the difference of the season difference, the data is stationary. The seasonal pattern in the series without taking difference has been eliminated. Thus the data need first difference and 1 seasonal difference. The graph above shows that we only need MA (1) term as the ACF and PACF all have negative value and only significant at the first lag. Finally the ARIMA (p, d, q) * (P, D, Q)*m model, where m= 12 as the lag repeats after 12 periods. Therefore ARIMA (0, 1, 1) (0, 1, 0)12.

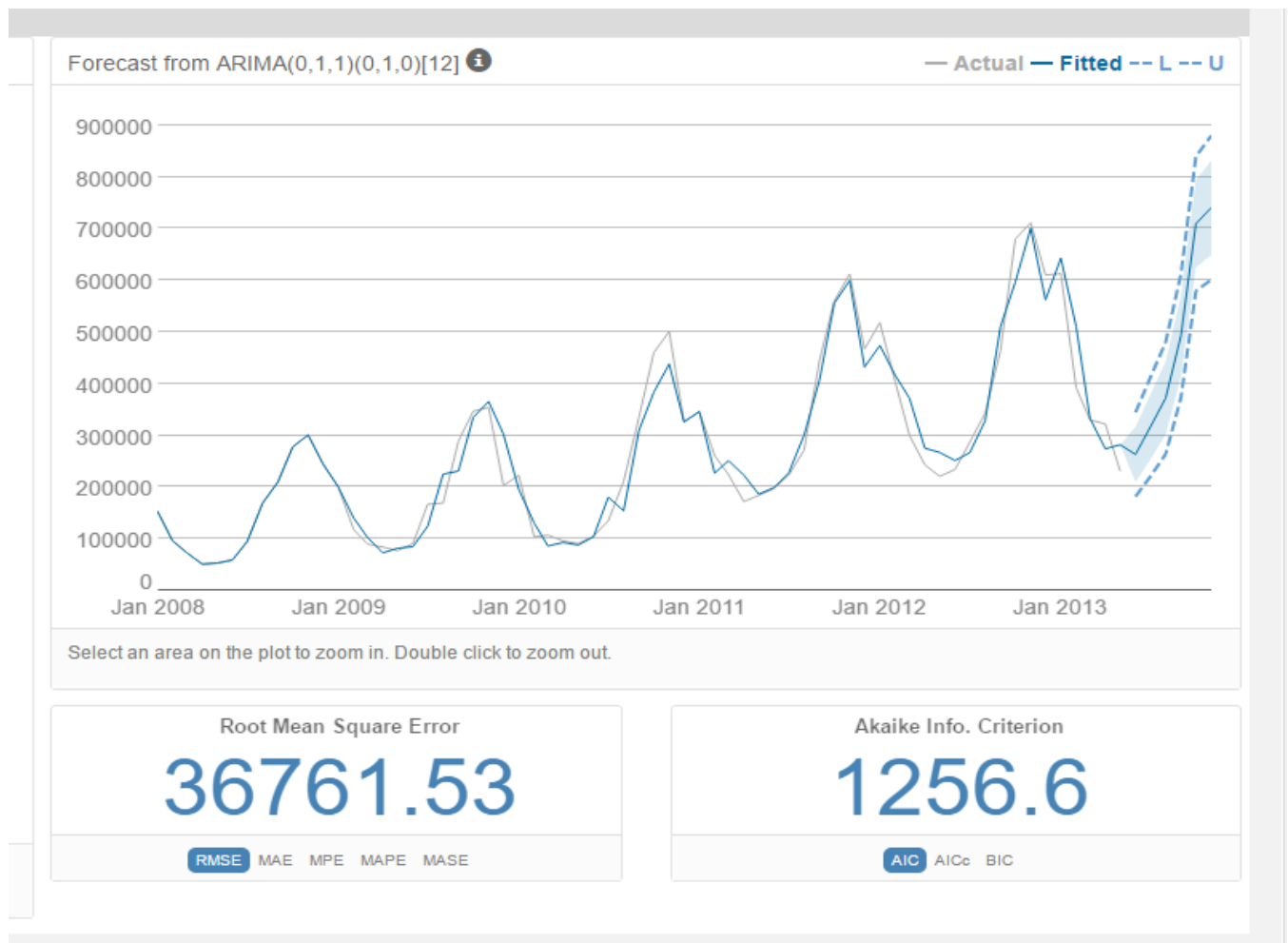
- a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

Answer:

The visualization shows the result of RMSE and the forecast from ARIMA (0, 1, 1) (0, 1, 0) (12)

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

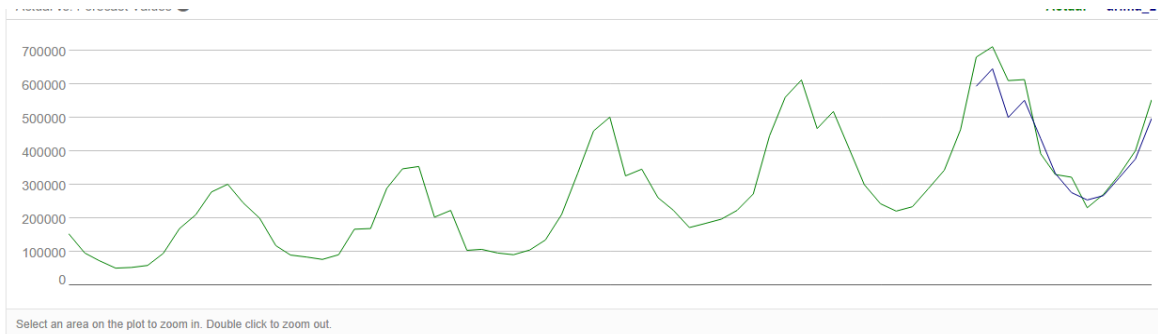


The result is pretty good in the sense that RMSE has the highest measure of predicting errors than the other in-sample error measures. Even though the MASE is less sensitive to outliers and more easily interpreted, RMSE was able to measure greater spread of the residuals (prediction errors) and has a better forecasting ability.

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

Answer:

Below are the re-graph visualizations of ACF and PACE for time series and seasonal difference. It could be seen that the ACF and PACF results for the correct ARIMA model exhibits no significant correlated lags suggesting no need for additional AR() or MA() terms, but must be noted that AR term works best when there is positive autocorrelation at lag 1 while MA term works best when there is negative autocorrelation at lag 1.



The ARIMA result is used to forecast the holdout sample seems it looks very promising, as the predicted value follows closely the real ones.

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.

Answer:

- (i) The ETS and ARIMA model were chosen.
- (ii) ETS for AIC:

Information criteria:

AIC	AICc	BIC
1634.6435	1645.9768	1669.4337

ARIMA for AIC

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

The AIC value for the ARIMA show the best fit model since it is lesser (lower) than the ETS for AIC

- (iii) **In sample error measures of the ETS model**

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

In sample error measures of the ARIMA model

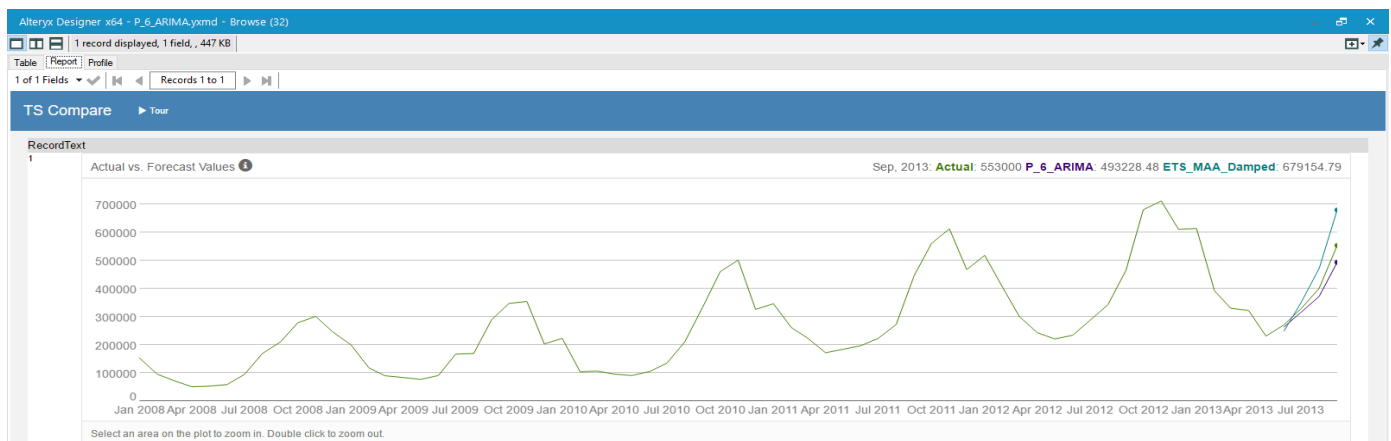
In-sample error measures:

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

(iv) **Forecast error measurements against the holdout sample**

2 records displayed, 8 fields, , 1718 bytes 1 record selected								
Table Profile								
8 of 8 Fields Cell Viewer ↑ ↓								
Record #	Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
1	P_6_ARIMA	27271.5199	33999.7911	27271.5199	6.1833	6.1833	0.4532	[Null]
2	ETS_MAA_Damped	-49103.3322	74101.1581	60571.8226	-9.7018	13.9337	1.0066	[Null]

Looking at the above table, the MASE value for EST is greater than the ARIMA with 1.0066 and 0.4532 respectively but the ARIMA (MASE) value is a bit closer to the actual value of the holdout sample. Therefore the ARIMA is best fitting model that will be use to forecast the next four periods.



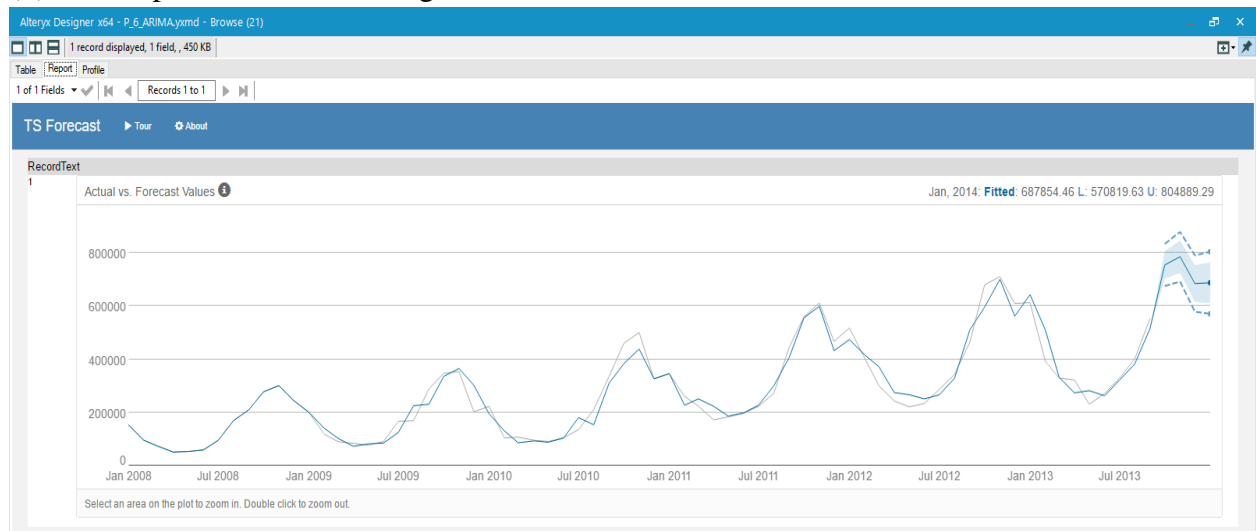
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

Answer:

(i) Forecasting the next four periods will be:

4 records displayed, 7 fields, 1743 bytes							
Table	Profile						
7 of 7 Fields	Cell Viewer						
Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2013	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
2	2013	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
3	2013	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
4	2014	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462

(ii) Graph of the results using 95% and 80% confidence intervals



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.