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.

   This dataset meets the criteria of a time series dataset. Each data point is taken across sequential and equal intervals. Each month has at least one data point. We are trying to predict future numeric values based on past results.

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

   We are trying to predict 4 months of sales. The hold out sample should typically be at least 4 months, however because of seasonality, I chose to hold out at least 12 months to try to predict one seasonal peak.

: Required: Just 3 of the key 4 characteristics were listed. Please include the 4th one. For a refresh on this topic please look at this lesson -(https://classroom.udacity.com/courses/ud980/lessons/7bcd2c42-a582-4e07-967c-f93a52de59f1/concepts/2423dfb7-8f1f-46c4-b8fc-4d7ca44f8ea1#)

: Required: Please note that the holdout sample should contain the last four moths since we are forecasting for the next 4 months. Please make sure that these four months are - 2013-06 to 2013-09,

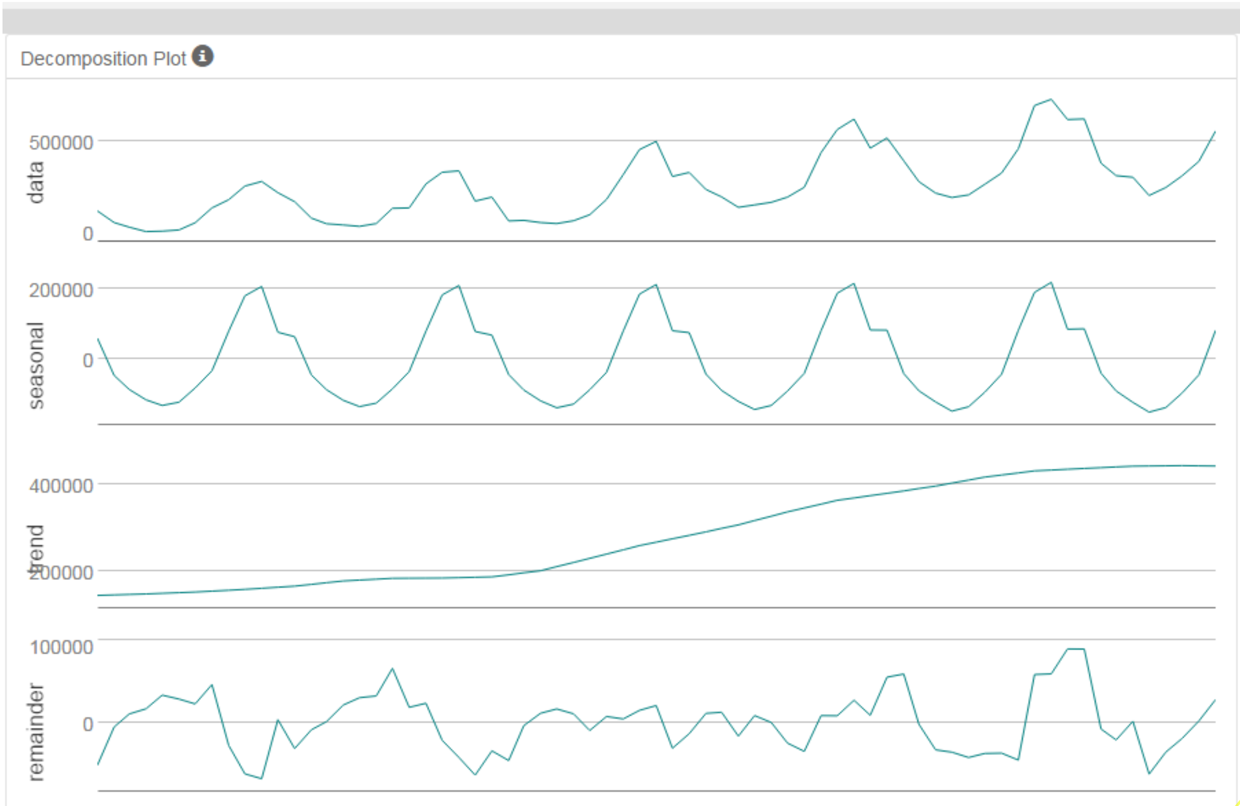
Indeed, the number of periods in each season is equal to 12. And we have monthly data therefore the "m" term in the ARIMA model should be 12 - ARIMA(0, 1, 1)(0, 1, 0)[12]

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.



Using the decomposition plot, the data has a seasonal component. It has a very slight increase over time which means multiplicatively.

Trend shows the general course and tendency of the time series. It is the centered moving average of the time series. It is moving in a linear fashion suggesting applying trend additively.

The remainder plot show what the residual error that seasonal and trend plots cannot explain. It shows changing variance as the time series moves along, so this suggest multiplicatively.

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:

MAM	42292.7436	62285.862	51111.7179	10.6898	12.4551	0.8738
ARIMA	-29521.8126	52097.6537	41468.5925	-10.4569	12.3784	0.709

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
      The RMSE was 62285.86 and the MASE is 0.8738.
      I chose ETS MAM, because of the decomposition plot above.

: Suggestion: Nice work! The seasonal portion shows that the regularly occurring spike in sales each year changes in magnitude, even so slightly rather than being constant. In Alteryx, we will need to hover our mouse over the seasonal graph in Interface mode to be able to see that the seasonal numbers are slightly increasing. This is important because:
- Having seasonality suggests that any ARIMA models used for analysis will need seasonal differencing.
- The change in magnitude suggests that any ETS models will use a multiplicative method in the seasonal component.

: Awesome: Indeed, the trend line is confirmed as upward linear trending.

: Awesome: The error plot of the series presents a fluctuations between large and smaller errors as the time series goes on. Since the fluctuations are not consistent in magnitude then we will apply error in a multiplicative manner for any ETS models.

: Required: The in sample errors that are listed here and included in the table above are incorrect.

Please, as I mentioned above use just the last 4 months as a holdout sample and recalculate the in sample errors and later the forecast error measurements against the holdout sample which are necessary for answering the questions in Step 4.

: Awesome; Correct! It would be still nice to include a few words here as well of why you chose this 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.

   I chose model term (0,1,1)(0,1,0)

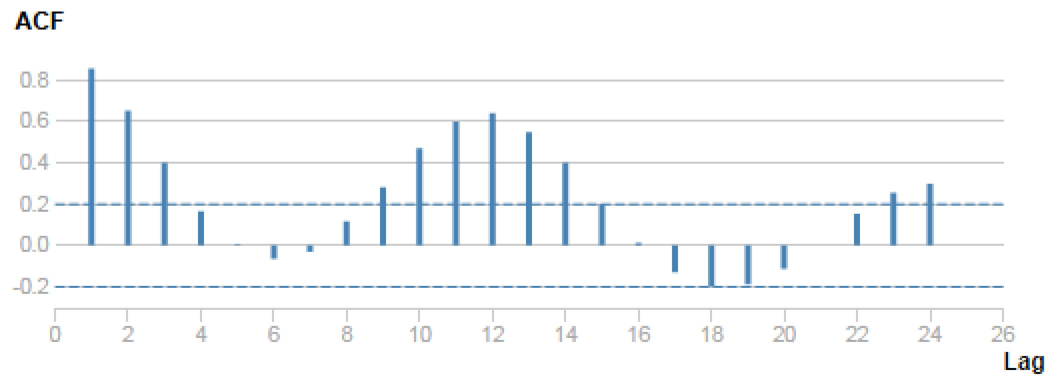
   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
      The RMSE is 52097.65.04 and the MASE is 0.709
   b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

: Required; The seasonal and the non seasonal terms of the ARIMA model are correct - great job!

Just please also include the "m" term which represent the number of periods in each season here as well.

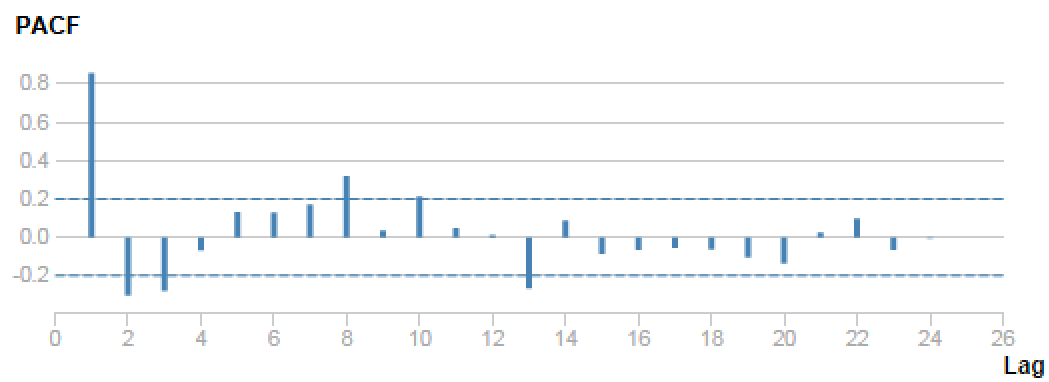
Without differencing, the ACF plot shows seasonality because the upward and downward peaks at lag 1, 12, and 24. The data points are outside of the .2 and -.2 range.

Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot

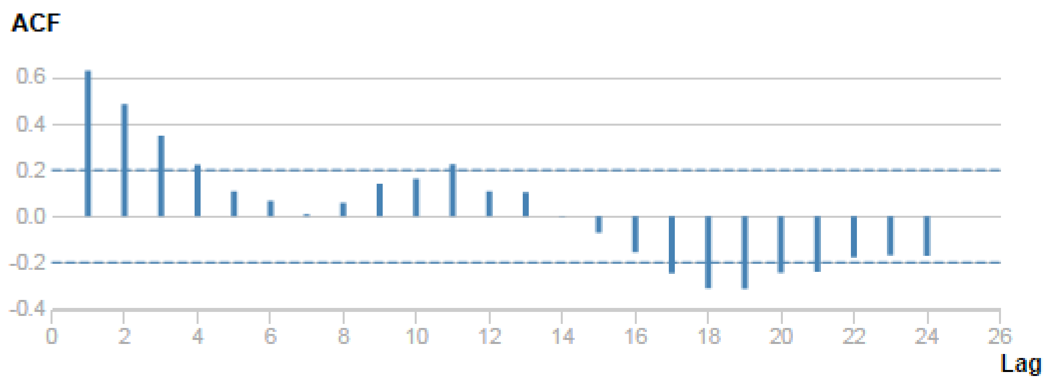


This is an partial autocorrelation plot

: Awesome: Nice work with the Time Series ACF and PACF plots.

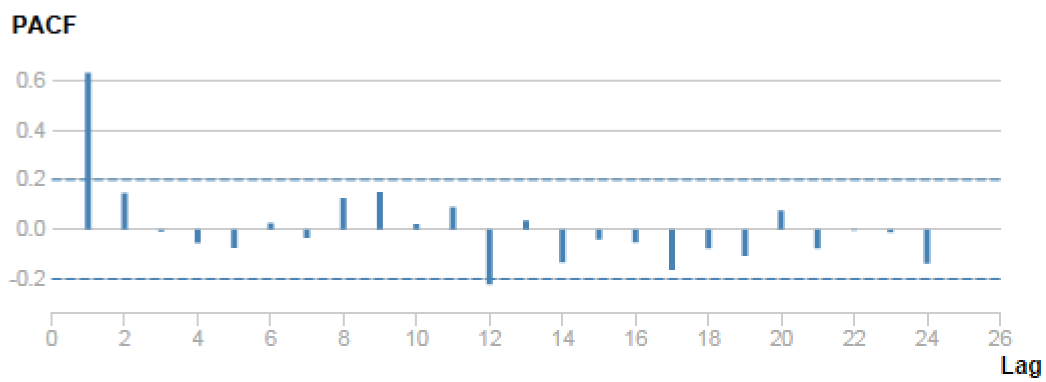
After taking the first seasonal difference, lag 12 and 24 no longer show seasonality and are within the acceptable range. However, there are still many points outside of .2 and -.2. A non-seasonal difference is needed for the data.

Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot



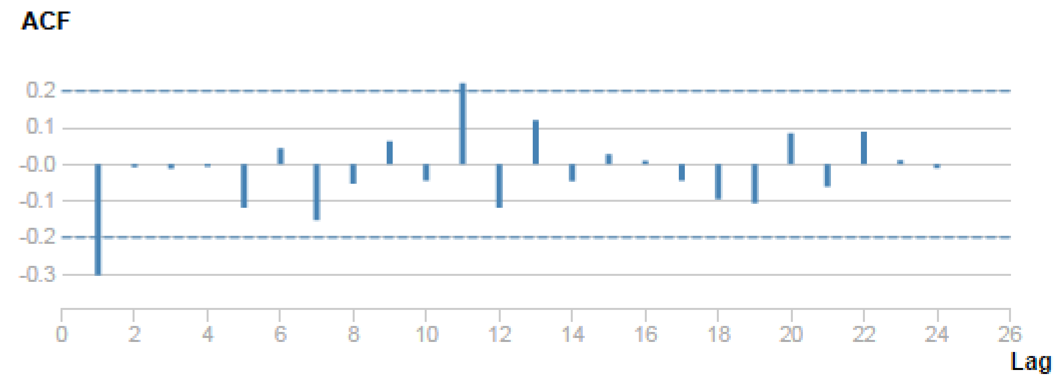
This is an partial autocorrelation plot

: Awesome; Nice work with the Seasonal Difference ACF and PACF plots!

After taking performing a non-seasonal difference, our data is now stationary. There is a small peak at lag 11 showing that there is some residual correlation in our data, but this is close enough and doesn't require another difference.

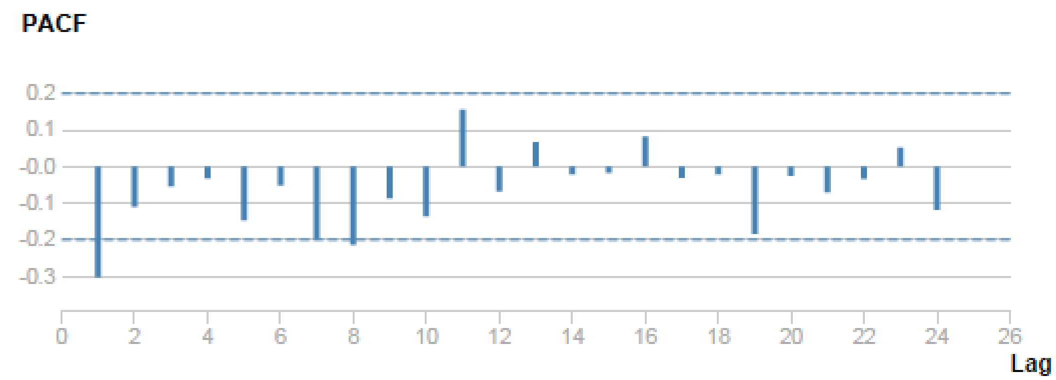
With a negative value at lag 1, a quick drop to 0 in the ACF and a more gradual trend to 0 in the PACF, the arima model is MA.

Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot



This is an partial autocorrelation plot

: Awesome; Nice work with the Seasonal First Difference ACF and PACF plots!

: Required: Great work with the model so far! Just one last thing left here - after establishing that the ARIMA(0, 1, 1)(0, 1, 0)[12] is the best one we should regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in our answer. The ACF and PACF results for the ARIMA(0, 1, 1)(0, 1, 0)[12] model should show no significantly correlated lags suggesting no need for adding additional AR() or MA() terms.

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.

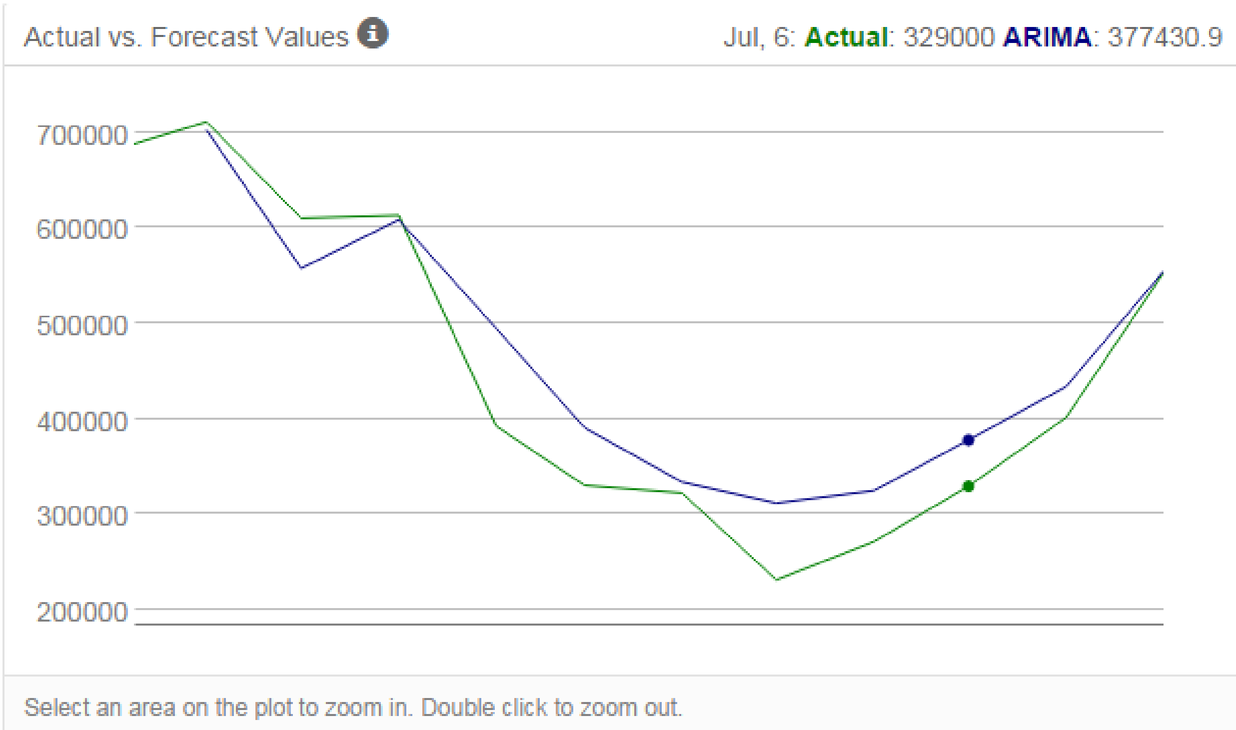
The arima (0, 1, 1)(0, 1, 0) performed the best against the holdout sample. It had an AIC of 1081.4 and overall smaller errors in all of the accuracy measures.

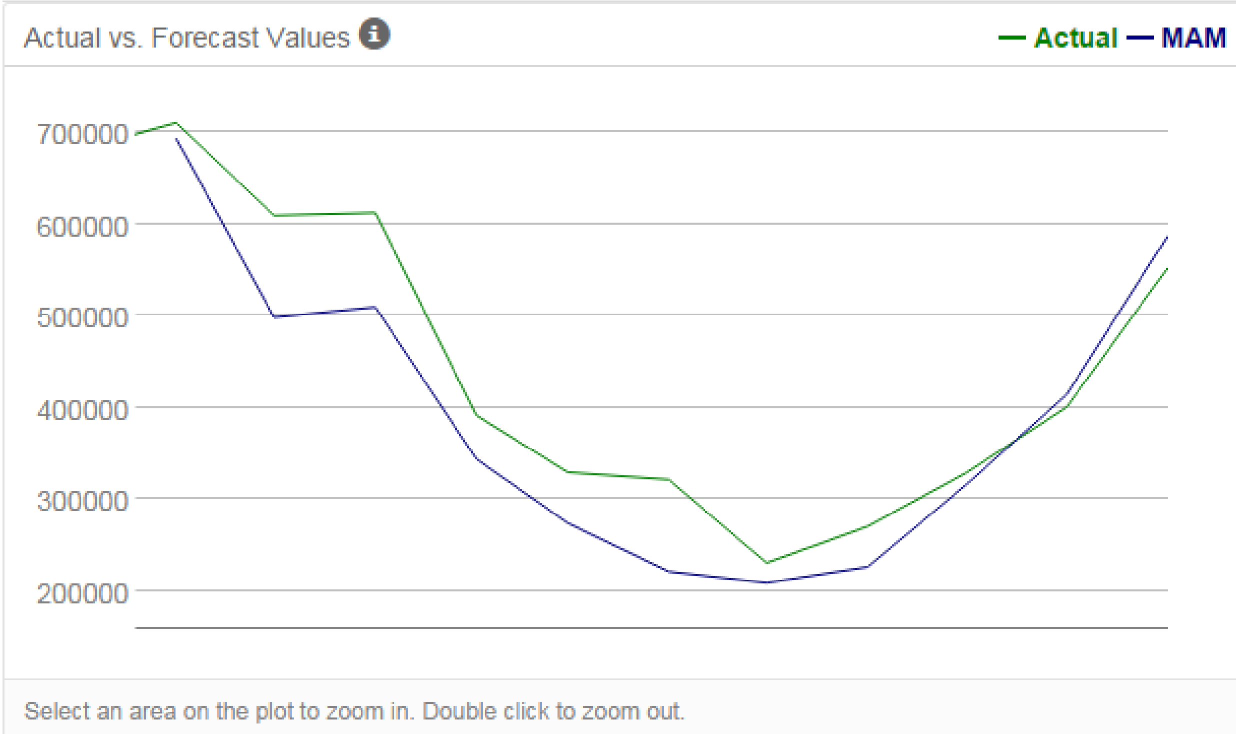
: Awesome: Indeed, the best performing model is the ARIMA model. Just please correct the issue with the holdout sample as I mentioned above.

MAM	42292.7436	62285.862	51111.7179	10.6898	12.4551	0.8738
ARIMA	-29521.8126	52097.6537	41468.5925	-10.4569	12.3784	0.709

: Required: As I said the in sample errors are incorrect. Please recalculate them after correcting the issue with the holdout sample.

ARIMA

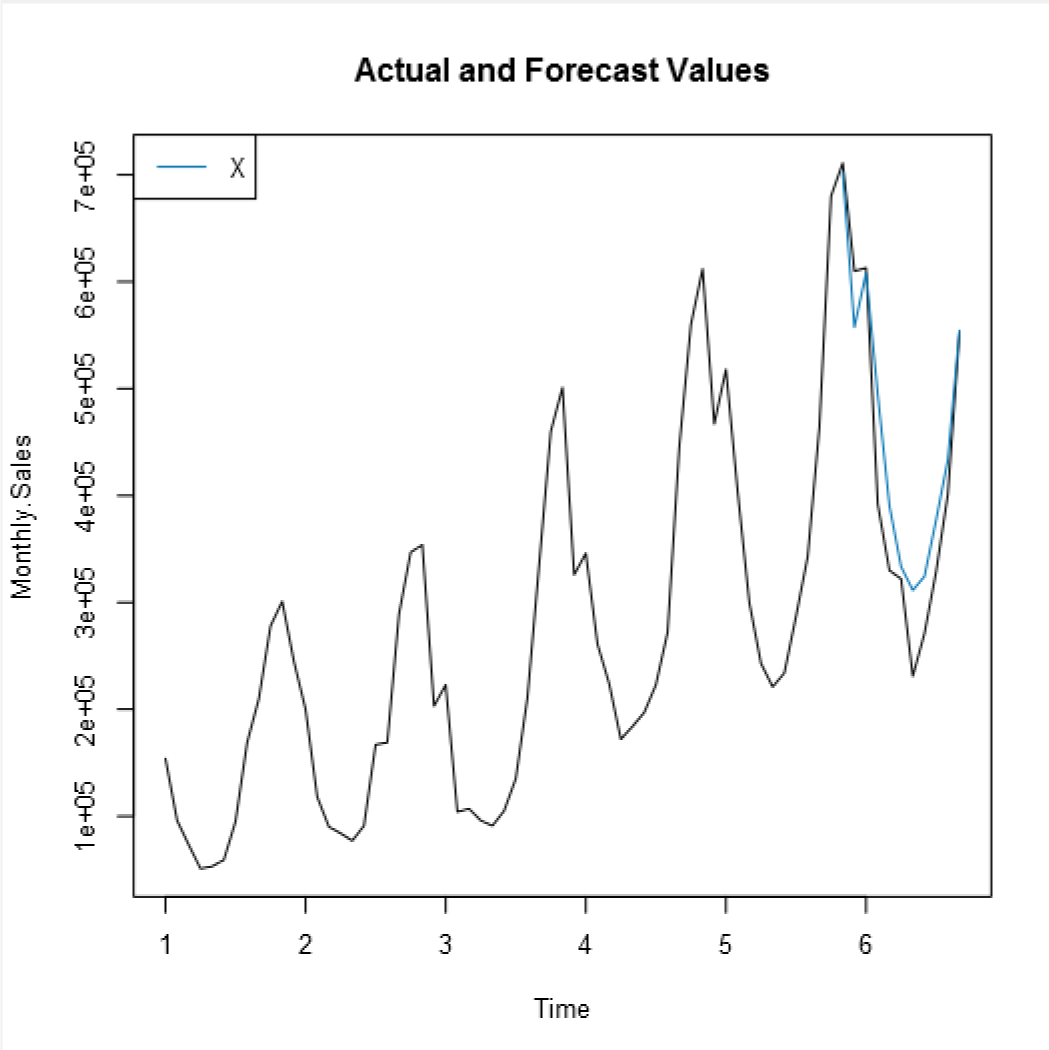



Actual vs. Forecast Values
— Actual — MAM
700000
600000
500000
400000
300000
200000
Select an area on the plot to zoom in. Double click to zoom out.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
X	-29521.81	52097.65	41468.59	-10.4569	12.3784	0.709	NA

: Required: Please recalculate the forecast error measurements against the holdout sample and please do that for both of the models.

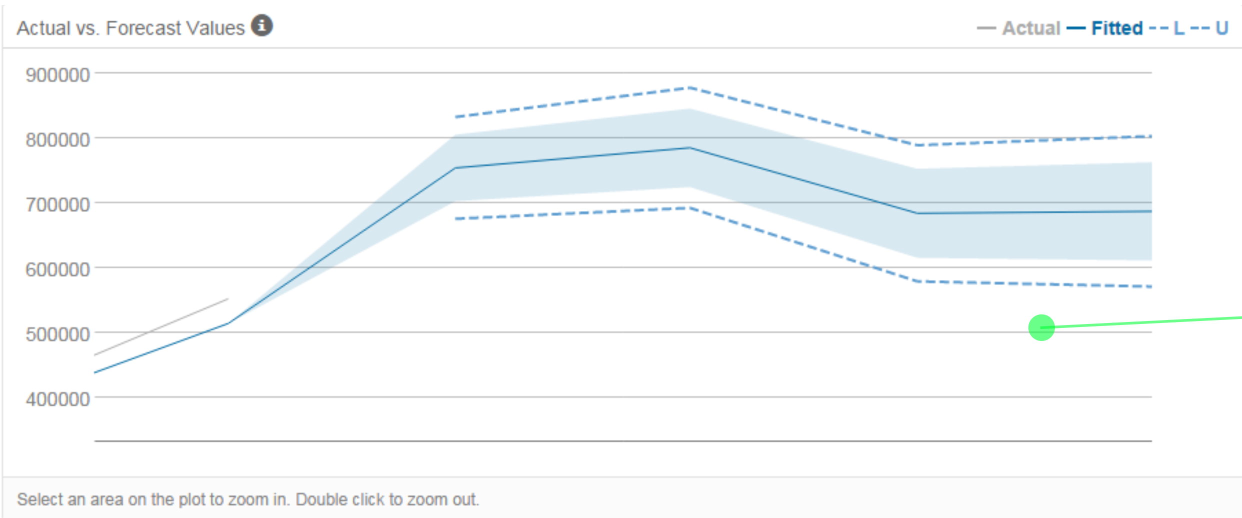


2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

The confidence for the next four periods is as follows:

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
6	10	754854.460048	833335.856133	806170.686679	703538.233418	676373.063963
6	11	785854.460048	878538.837645	846457.517118	725251.402978	693170.082452
6	12	684854.460048	789837.592834	753499.24089	616209.679206	579871.327263
7	1	687854.460048	803839.469806	763692.981576	612015.938521	571869.450291

: Awesome; Excellent! The forecasts are correct!



: Awesome: Great work with the plot!

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.