Project: Forecasting Sales

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.

Initial findings of the time series showed a complete series exhibiting the 4 key characteristics of time series data. The series is over a continuous time interval, of sequential measurements across that interval, using equal spacing between every two consecutive measurements and each time unit within the time interval has at most one data point.

The data collected is composed of monthly sales data dating back to 2008 and going until September 2013. A sample image of the data is shown below:

	Month	Monthly Sales
1	2008-01	154000
2	2008-02	96000
3	2008-03	73000
4	2008-04	51000
5	2008-05	53000
6	2008-06	59000
7	2008-07	95000
8	2008-08	169000
9	2008-09	210000
10	2008-10	278000
11	2008-11	301000
12	2008-12	245000

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

In preparation for construction of a predictive models, I have filtered out the last 4 records, 2013-06 to 2013-09, as a holdout sample so that I can check the accuracy of my model to forecast predicted values against the actual values.

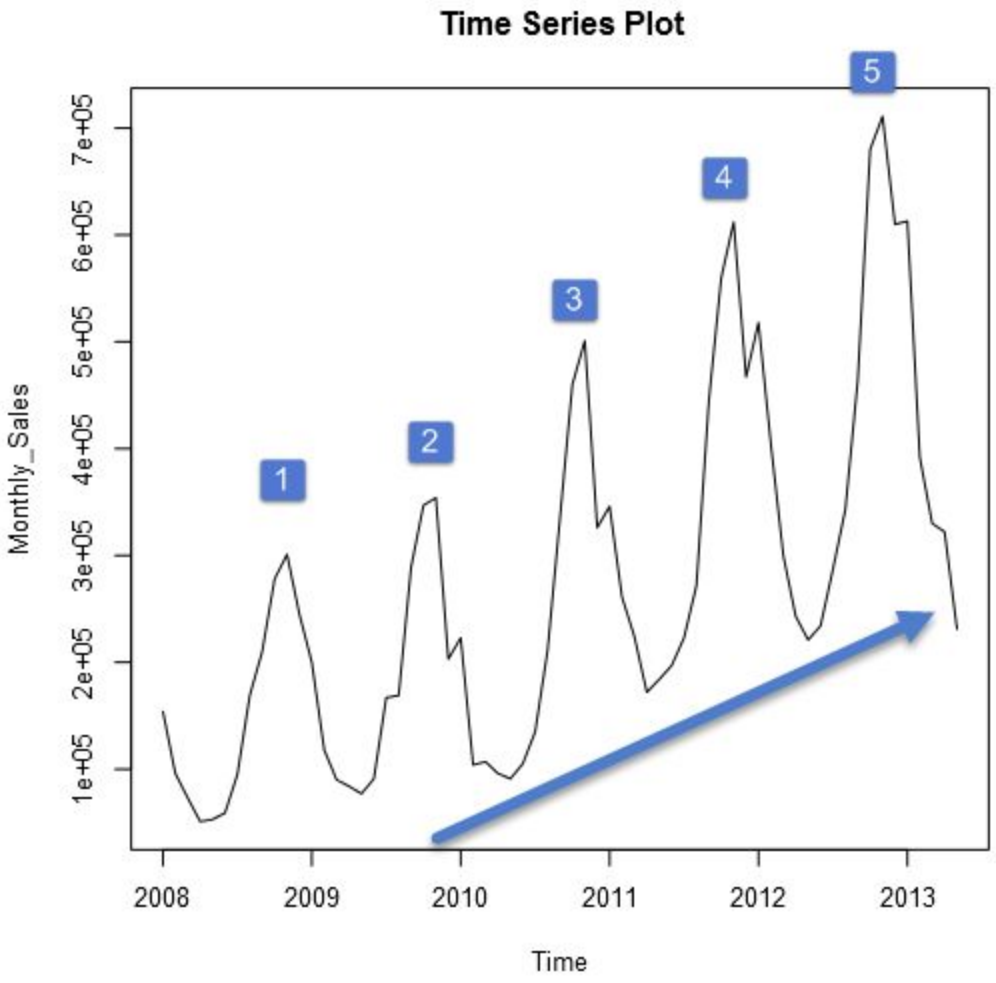
Step 2: Determine Trend, Seasonal, and Error components

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

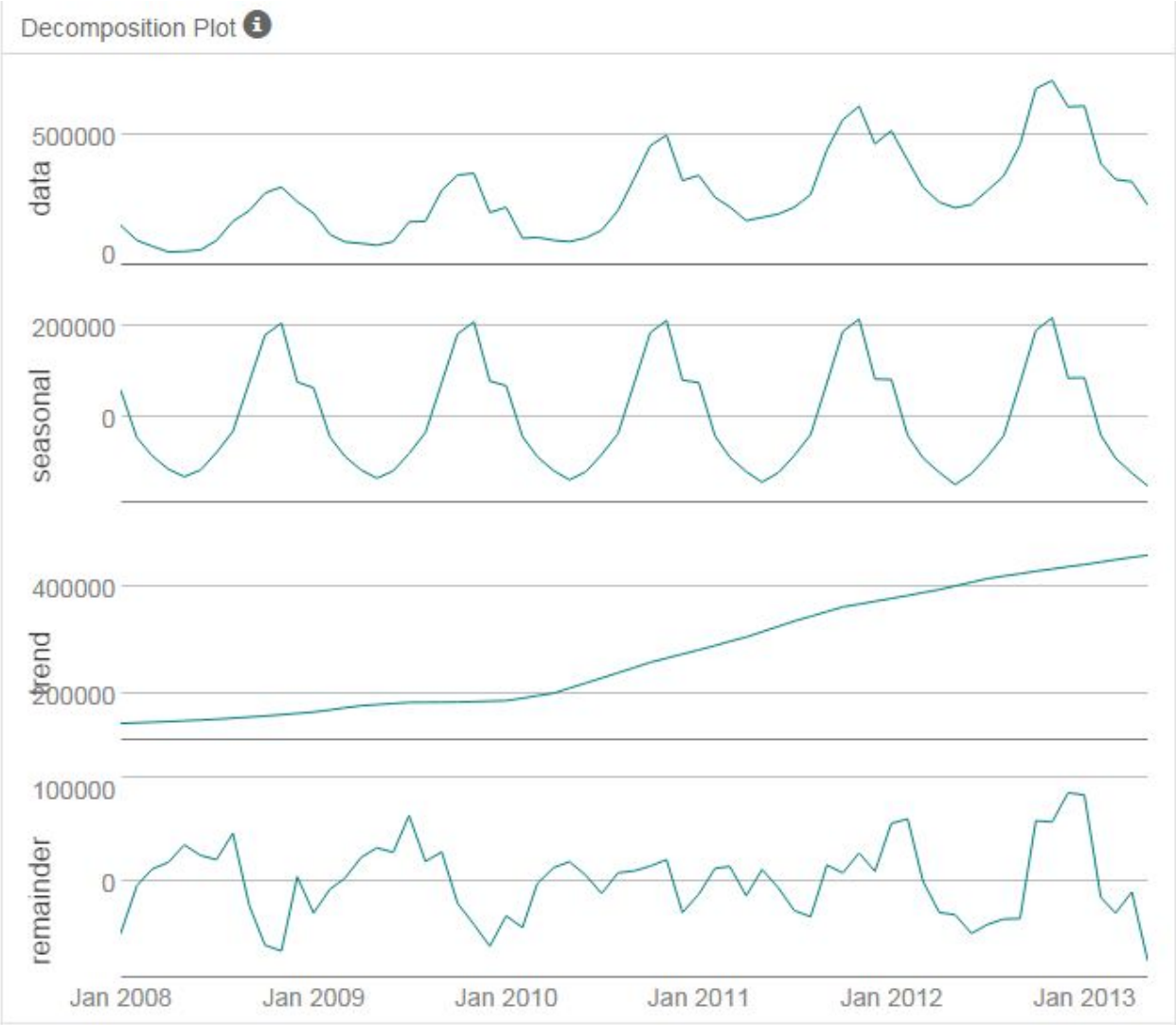
Answer this question:

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 initial findings of the time series plot shows an upward rising trend with a regularly occurring spike in sales each year reported at the end of the year. This pattern shows that we have seasonality in our time series. There are no patterns within the series suggesting cyclicity.



The decomposition plot shows our time series broken down into its three components: trend, seasonal and the error. Each of these components makes up our time series and helps us confirm what we saw in our initial time series plot.



Our trend line is confirmed as upward trending.

The seasonal portion shows that the regularly occurring spike in sales each year changes in magnitude, ever so slightly. 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.

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.

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

From our decomposition plot we can obtain the necessary information to define our terms for the ETS model.

Our trend line exhibits linear behavior so we will use an additive method.

The seasonality changes in magnitude each year so a multiplicative method is necessary.

The error changes in magnitude as the series goes along so a multiplicative method will be used.

This leaves us with an **ETS(M, A, M) model**.

Error Terms:
The in-sample error measures give us a look at how well our model is able to predict future values.

THESE NUMBER HAVE CHANGED WITH NEWER VERSIONS OF ALTERYX, IF YOU HAVE SELECTED THE CORRECT MODEL TERMS DO NOT WORRY IF THEY DO NOT MATCH UP EXCATLY

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

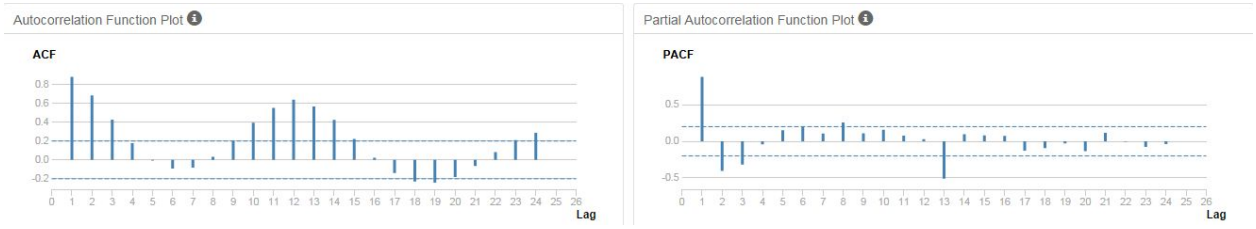
Two key components to look at are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models. We can see that our variance is about 33000 units around the mean.

The MASE shows a fairly strong forecast at .36 with its value falling well below the generic 1.00, the commonly accepted MASE threshold for model accuracy.

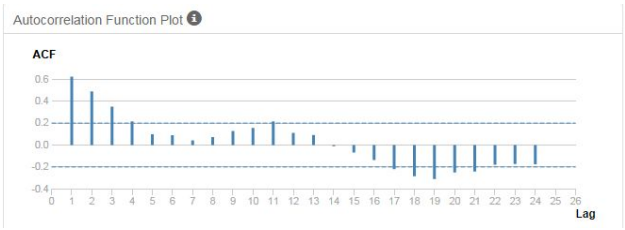
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
    a. Describe the in-sample errors. 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 and show that the graphs have no autocorrelated lag anymore.

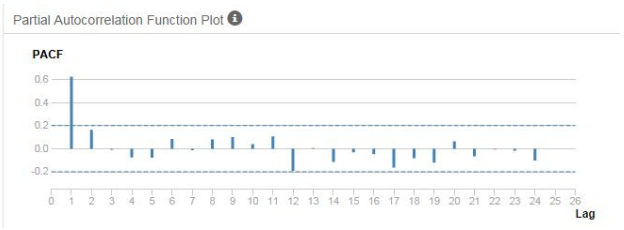
Since there are seasonal components found in the time series I will use an ARIMA(p, d, q)(P, D, Q)S model for forecasting.

Time Series ACF and PACF:
The ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high I will need to seasonally difference the series.



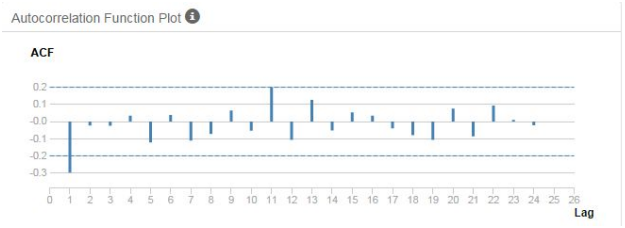
Seasonal Difference ACF and PACF:
The seasonal difference presents similar ACF and PACF results as the initial plots without differencing, only slightly less correlated. In order to remove correlation we will need to difference further.

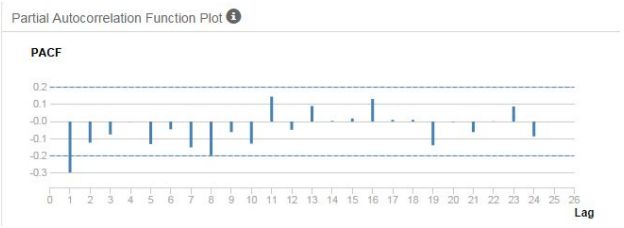




Seasonal First Difference:
The seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be d(1) and D(1).

The ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model since there is only 1 significant lag. The seasonal lags (lag 12, 24, etc.) in the ACF and PACF do not have any significant correlation so there will be no need for seasonal autoregressive or moving average terms.



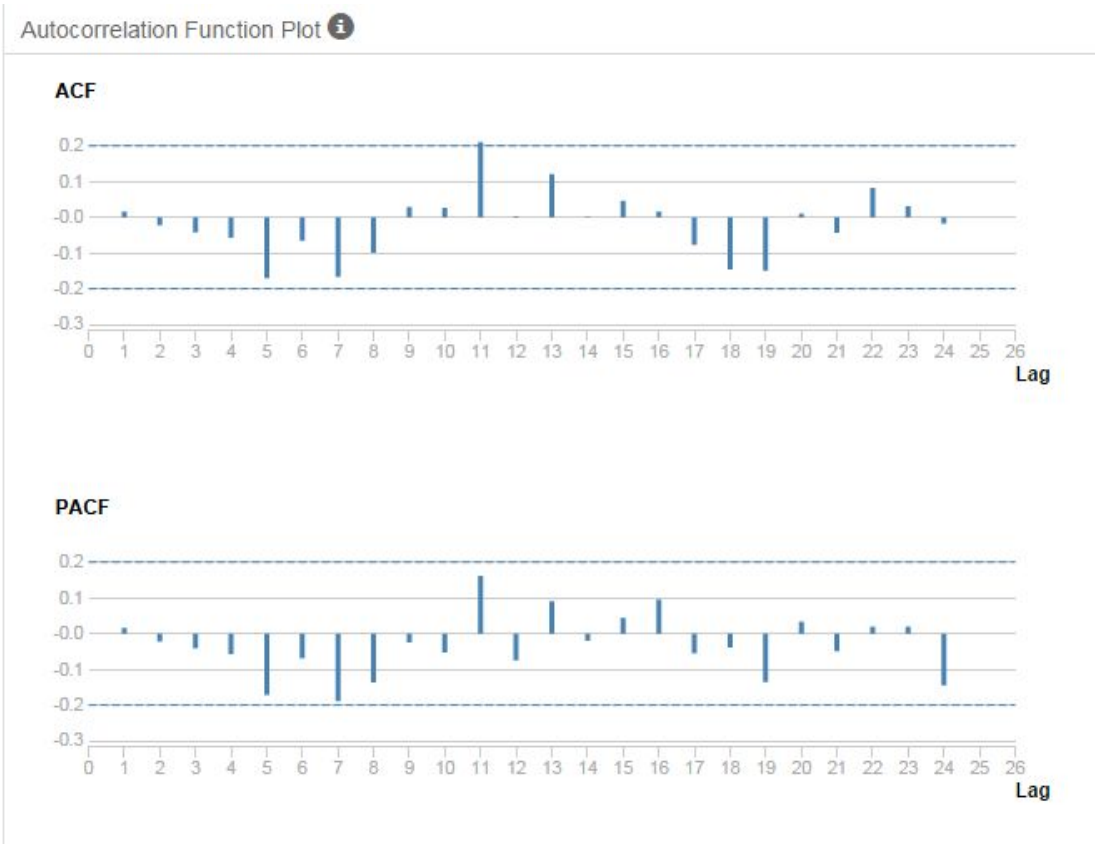


Therefore the model terms for my ARIMA model are:

ARIMA(0, 1, 1)(0, 1, 0)[12]

Error Terms:

The ACF and PACF results for the ARIMA(0, 1, 1)(0, 1, 0)[12] model shows no significantly correlated lags suggesting no need for adding additional AR() or MA() terms.



The in-sample error provides a closer look at the model accuracy.

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. We can see that our variance is about 37000 units around the mean.

The MASE shows a fairly strong forecast at .36 with its value falling well below the generic 1.00, the commonly accepted MASE threshold for model accuracy.

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.

When fitting a forecasting model we can use a series of identifiers that help us choose the best model.

When comparing the two in-sample error measures we used, the RMSE and MASE, we see very similar results. The ETS model does have a narrower standard deviation but only by a few thousand units.

ETS

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

ARIMA

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

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.

When looking at the model's ability to predict the holdout sample, we see that the ARIMA model has better predictive qualities in just about every metric.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-73257.47	89012.35	74392.72	-17.1046	17.5235	1.2363
ARIMA	22271.52	33589.74	25885.76	4.628	5.7976	0.4302

For our forecast, we will use the ARIMA model.

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

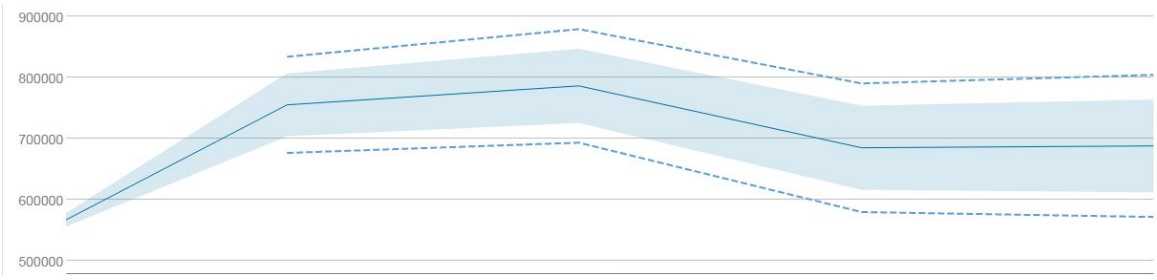
Forecast results using 95% and 80% confidence intervals:

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

From NEWER VERSION (see table below) of ALTERYX numbers are slightly different, both are acceptable.

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

Shown graphically here:



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.