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 meet the criteria of a time series because it simply shows monthly sales during 6 years so it's a numeric and time based data.
   The dataset covers a continuous time interval. Also there is equal spacing between every two monthly sales.
   Moreover, each time unit (Month), has at most one data point of sales.

2. Which records should be used as the holdout sample?
   Since we will forecast 4 periods, we will be using a holdout sample of 4 time intervals:

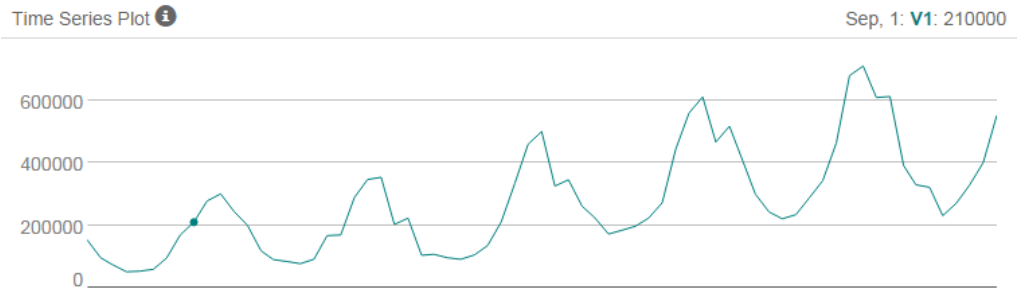
2013-06	**271000**
2013-07	**329000**
2013-08	**401000**
2013-09	**553000**

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)*

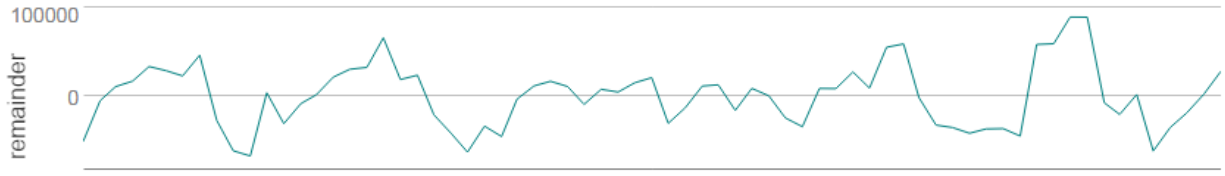
Answer this question:

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 following graph shows an uptrend in the data.



   Also the graph below clearly shows that the time series contains seasonality. Seasonality occurs at regular intervals

   Moreover, the error graph shows the difference between the observed value and the trend line estimate. It's showing the data which are not accounted for by combining the seasonal data and the trend data.



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.

    Since the time series has a fluctuating between large and small errors over time, linear trend, and an increase of seasonal fluctuations over time.. we would need to use an ETS model of: **ETS (M,A,M)**

    "Root Mean Squared Error (RMSE) represents the sample standard deviation of the differences between predicted values and observed values.
    Mean Absolute Scaled Error (MASE) is defined as the mean absolute error of the model divided by the mean absolute value of the first difference of the series. "

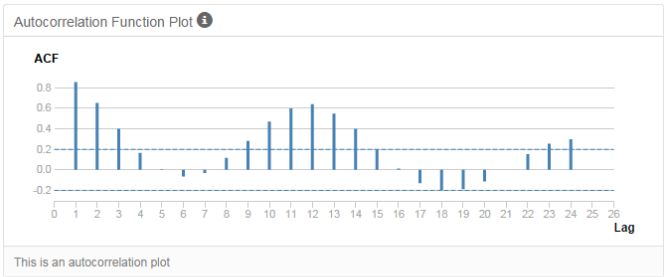
    The (RMSE) shows that the forecasted values are highly deviated from the mean. However the MASE error value shows that the model should be considered effective with a value of 0.36 which is < 1.

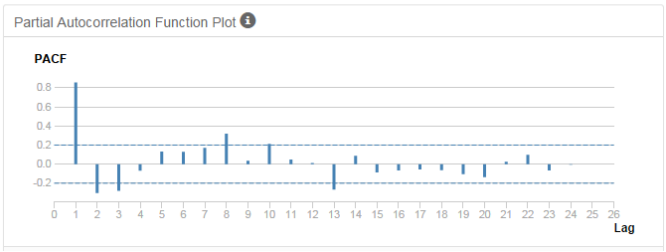
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056	0.1436491

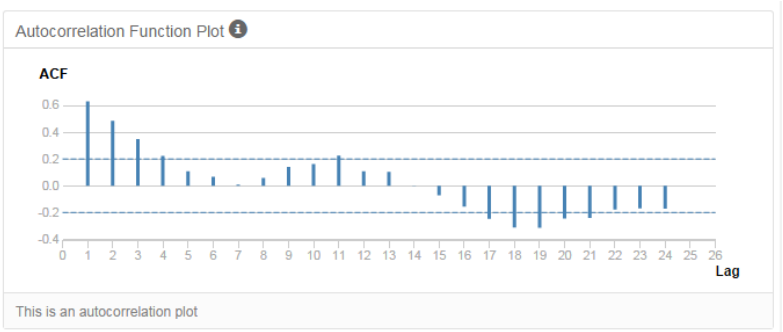
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
    a. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

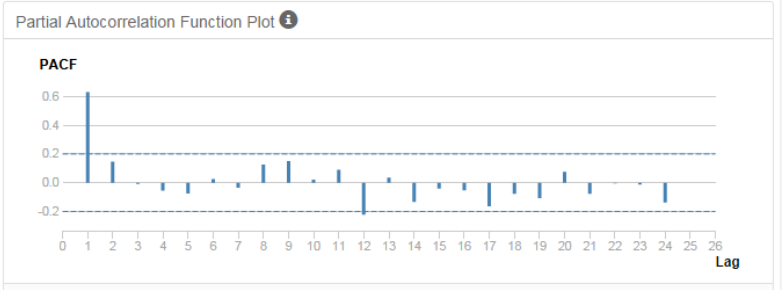
The graphs, shows the ACF and PACF plots, of original dataset:



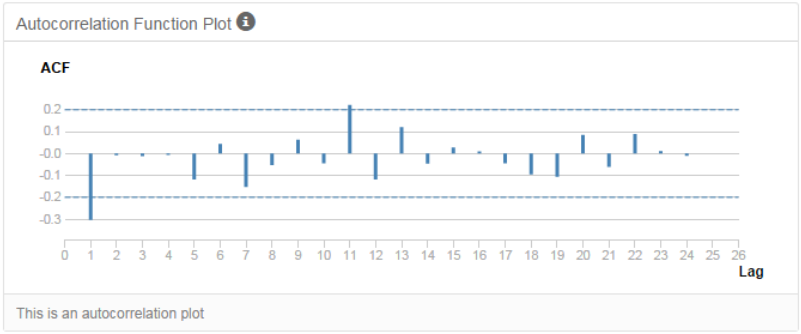


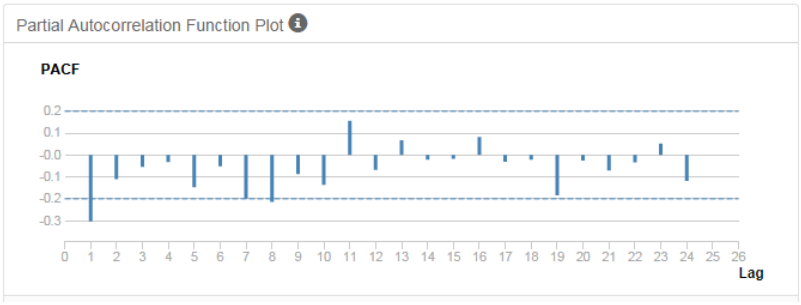
The data is not stationary, and we should take the seasonal differencing to remove the seasonal effect. (The following graphs show, the ACF and PACF pots after applying seasonal differencing)





After, seasonal difference, the data is still not stationary, thus we have to take first differencing: (Following plots show, ACF and PACF, after first difference)





The data is finally stationary after first seasonal difference.

ARIMA Terms:
The correlation between lags 12, in the ACF and PACF, and with a negative lag at 12 this indicates using seasonal terms.
From the graphs above, we can see that for the ACF there is a strong cut off to 0 at lag-1 and on the PACF there is a gradual drop to 0. Also since there is a negative correlation at -1, then MA terms are best.
The number of times we did differencing is 1, thus d is 1.

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

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

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-358.1274828	36758.4027043	24996.5435416	-1.800917	9.8272386	0.3646619	0.0166958

"**Root Mean Squared Error (RMSE)** represents the sample standard deviation of the differences between predicted values and observed values.
Mean Absolute Scaled Error (MASE) is defined as the mean absolute error of the model divided by the mean absolute value of the first difference of the series. "

The **(RMSE)** shows that the forecasted values are highly deviated from the mean with a value of (**36758**).
However the MASE error value shows that the model should be considered effective with a value of 0.36 which is **< 1**.

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.

If we will consider the AIC value, ARIMA model will be selected which has a lower value of AIC.

ARIMA AIC:

Information Criteria:

AIC	AICc	BIC
1258.5932	1259.0932	1264.447

ETS AIC:

Information criteria:

AIC	AICc	BIC
1634.6435	1645.9768	1669.4337

Also looking at the error measurements below, ARIMA model showed better results in terms of RMSE and MASE errors.
The MASE error value shows that the ARIMA model should be considered effective with a value of 0.45 which is < 1. It's lower than the MASE value of the ETS model (1.15).
Moreover, the RMSE is showing also that the ARIMA model is more effective with a value of 34010 which is less than the RMSE value of ETS (85623), which means a lower variance between forecasted data and the mean of actual data.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_MAM_	-68257.47	85623.18	69392.72	-15.2446	15.6635	1.1532	NA

Accuracy Measures:

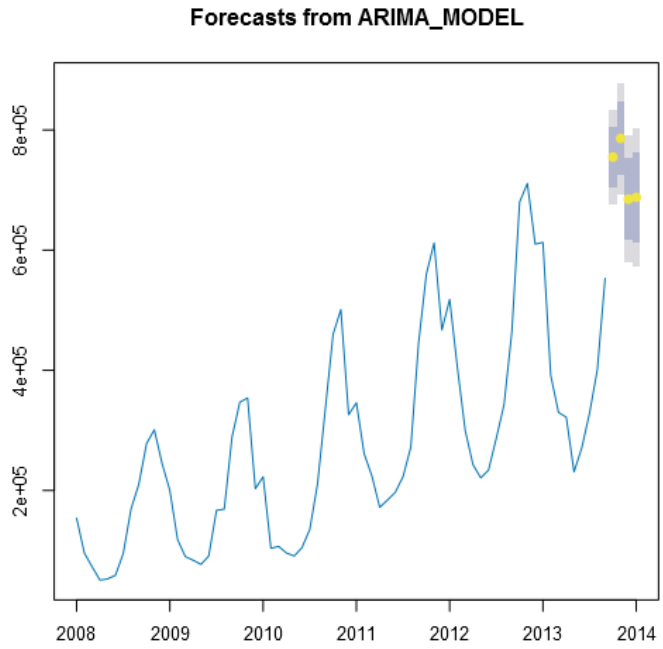
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_MODEL	27184.45	34010.92	27184.45	6.1547	6.1547	0.4518	NA

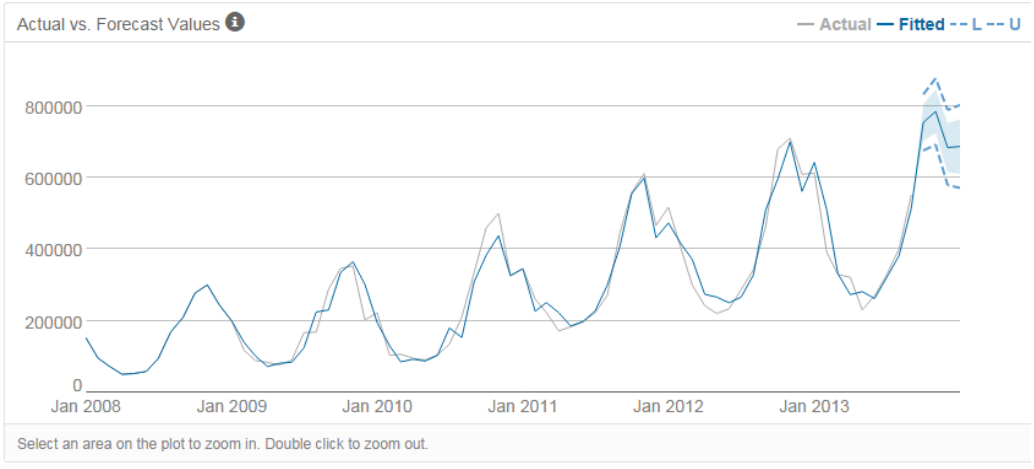
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

The forecasted values for the next four months are:

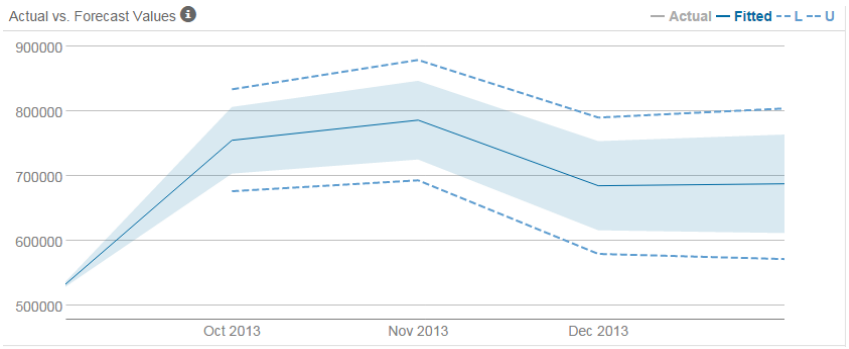
Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2013	10	754854.460048	833335.856133	806170.686679	703538.233418	676373.063963
2013	11	785854.460048	878538.837645	846457.517118	725251.402978	693170.082452
2013	12	684854.460048	789837.592834	753499.24089	616209.679206	579871.327263
2014	1	687854.460048	803839.469806	763692.981576	612015.938521	571869.450291

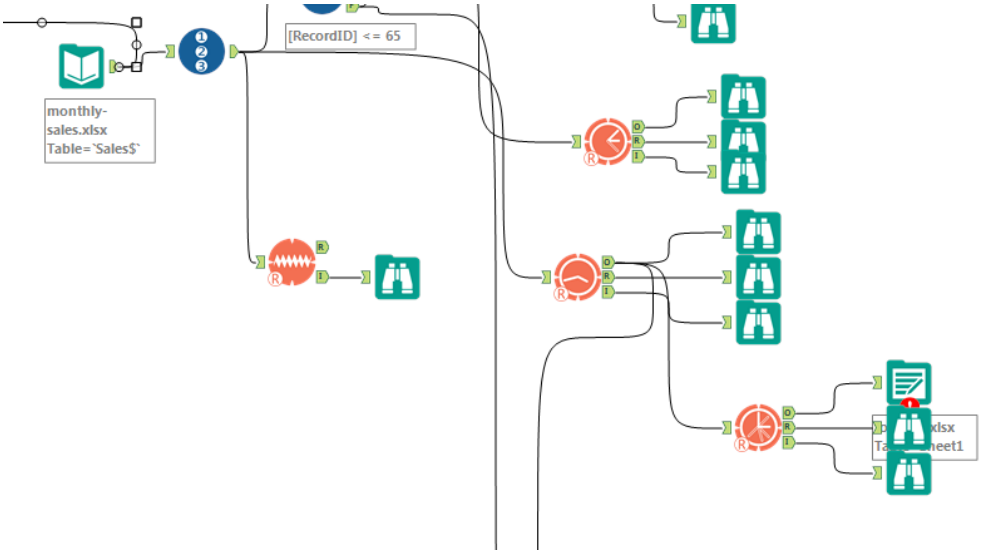
The following graph shows, all dataset with forecasted values:





The following graph shows Forecasts of the next 4 months:





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.