Project: Forecasting Sales

Step 1: Plan Your Analysis

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.

   The four key characteristics of a time series data set are:
   - *"The series is over a continuous time interval*
   - *Sequential measurements across that interval*
   - *There is equal spacing between every two-consecutive measurement*
   - *Each time unit within the time interval has at most one data point"*

   Hence, yes, the dataset meets the above criteria of a time series dataset as follows: the dataset holds sixty-nine records which represent the time from first entry in January 2008 till September 2013. Moreover, by transforming the 'month' column into 'year' column and 'month' column, the other characteristics of a time series data can be verified.

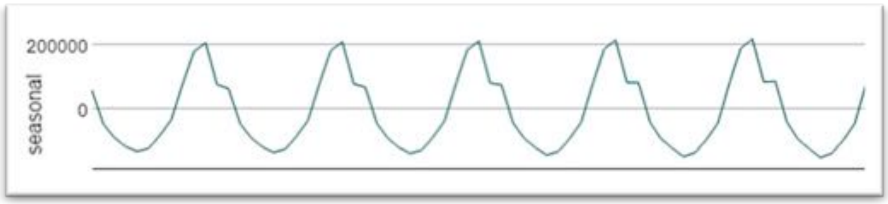
2. Which records should be used as the holdout sample?
   The records 66-69 should be used as the holdout sample, which satisfied the business requirement of four-month forecast.

Step 2: Determine Trend, Seasonal, and Error components

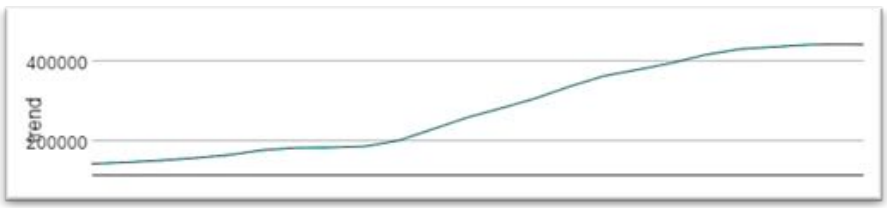
Answer this question:

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 trend, seasonality, and error (i.e. remainder) are components of the time series, they can be decomposed via the ***TS Plot tool*** as shown in the graphs below:

This graph shows the time series before using the TS Plot tool.



This graph shows the annual seasonal pattern where the annual highest sales occur in Novembers while the lowest are in Mays, with continuous increment and decrement over the years (e.g. November 2012 is higher than November 2011).



This graph shows the steady rise of sales over the years except between July 2009 and January 2010 where sales were low.

This graph shows the error of the time series. The occurrence of constant variance in the middle is more often the rest of the graph.

Step 3: Build your Models

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

       The model terms for ETS are (M, A, M) for the following reasons:
       - First, the occurrence of constant variance of error is more often and smaller in the middle than the rest of the graph, the error type should be applied as multiplicatively (M), consequently.
       - Also, because of the linear upward trend, the trend type should be applied as additively (A).
       - Finally, seasonality type should be applied as multiplicatively (M) due the (slight) growth of sales.

Nevertheless, testing the ETS (M, A, M) as a damper model and undamped model (separately) on a selected period of time for both model, the result of damped ETS (M, A, M) is better in term of lower AIC and MASE.

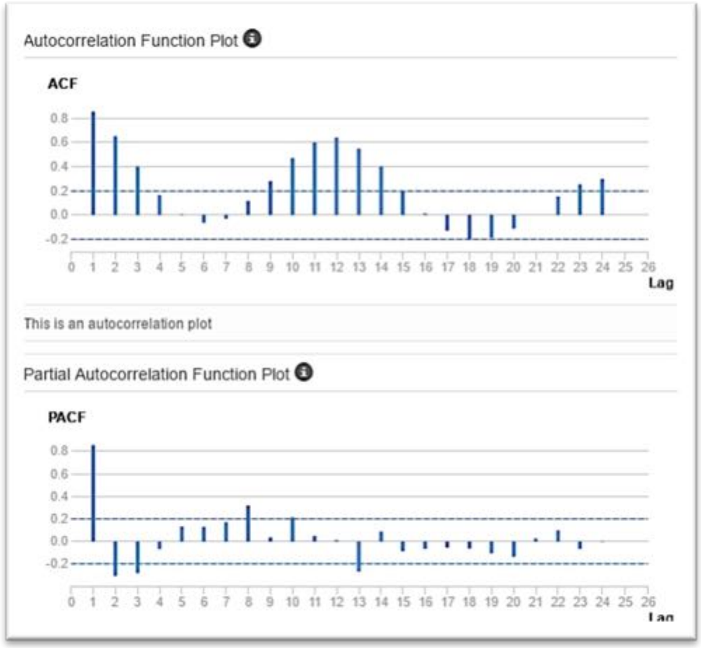
The RMSE assists in quantitatively measure the closeness between the forecasted variable and the actual data. It measures by the same unit as the original data.

In the graph below, the MASE is less than one, which reflects that the prediction model is good.

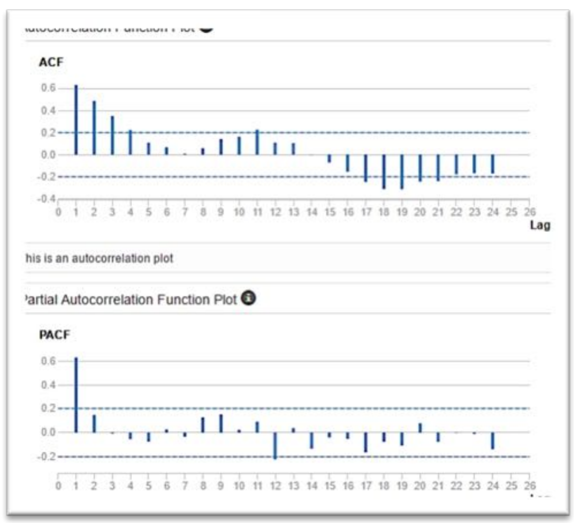
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

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

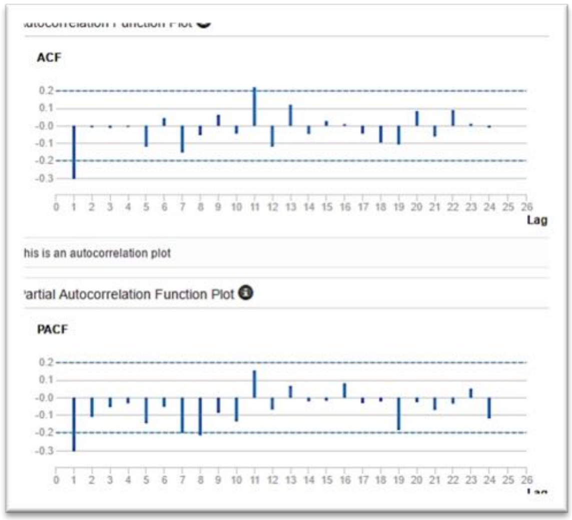
       The model is built as ARIMA (p, d, q)(P, d, Q)m since the given time series has seasonality, and thereby the model terms are seven. m=12 as the seasonal period equals 12 months. Some ACF and PACF graphs are explained below:



In this graph, serial correlation appears at the 12 and 24 lags (0 reduction in-between)



trend/pattern returns after the series was differenced to be stationary.



The final stationary time series after differencing the seasonal difference.

The ARIMA model is configured as ARIMA(0,1,1)(0,1,0)12 for the following reasons:

- Since ACF lag-1 term is negative and the a sharp is obvious, then p=0 and q=1
- Since ACF lag-1 is negative, then no other SAR terms is needed (i.e. P=0).
- Since all seasonal legs present no spike, then no SMA term is needed (i.e. Q=0).
- Since both seasonal and non-seasonal differencing are used, then d=1 and D=1

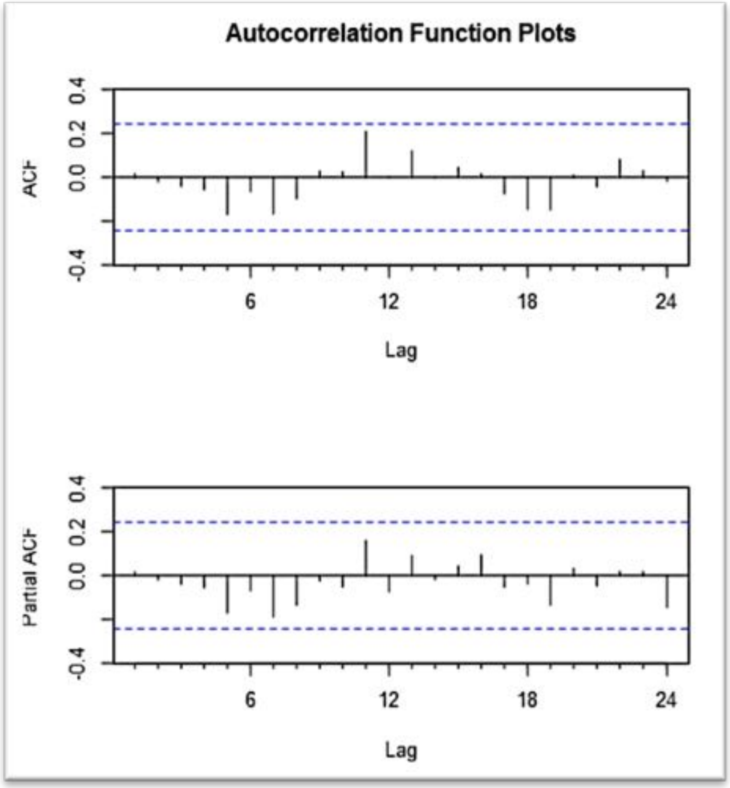
Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

The model is good since MASE (of ARIMA) is less than 1.

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



Step 4: Forecast

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.
   From step 3, the in-sample errors of ETS are shown in the graph below:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

   While the in-sample errors of ARIMA are:

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

   However, the ARIMA model is chosen (over the ETS model) since it predicts values that are closer to the original points of the forecasted period. Also, ARIMA model has lower error measurements for both of RMSE and MASE (see the graphs above). Nevertheless, a comparison of time series models is shown below:

Actual and Forecast Values:

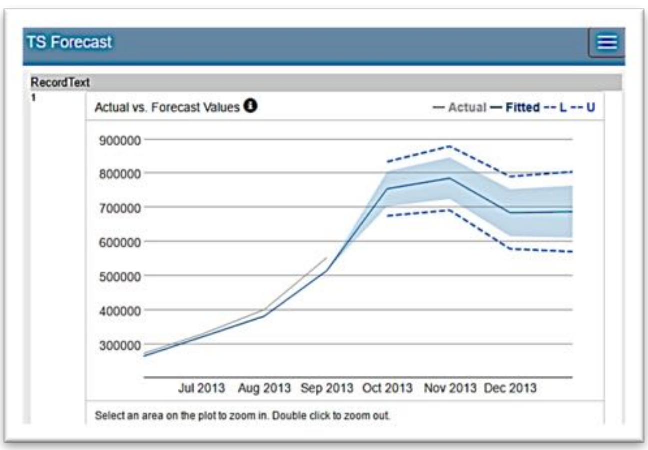
Actual	ETS_Model_MAM_Damped	ARIMA_0_1_1__0_1_0_12
271000	255966.17855	263228.48013
329000	350001.90227	316228.48013
401000	456886.11249	372228.48013
553000	656414.09775	493228.48013

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Model_MAM_Damped	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA
ARIMA_0_1_1__0_1_0_12	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

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

The following graphs show the desired results:



Period	Sub_Period	ARIMA_Forecast	ARIMA_Forecast_high_95	ARIMA_Forecast_high_80	ARIMA_Forecast_low_80	ARIMA_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