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.
   Monthly sales data file which contains store information for the company's sales by month, has 2 fields and 69 records. The first field is Month and the second filed is monthly sales values. The provided dataset has the 4 characteristics of a Time Series. First, it's over a continuous time interval starting from 2008-01 up to 2013-09. Second, there are sequential measurements across the interval (Jan 2008 – Sep 2013). Third, there is equal spacing between every two consecutive measurements (a monthly space). And finally, every time step within the time interval has maximum of one data point (Monthly sales figure).

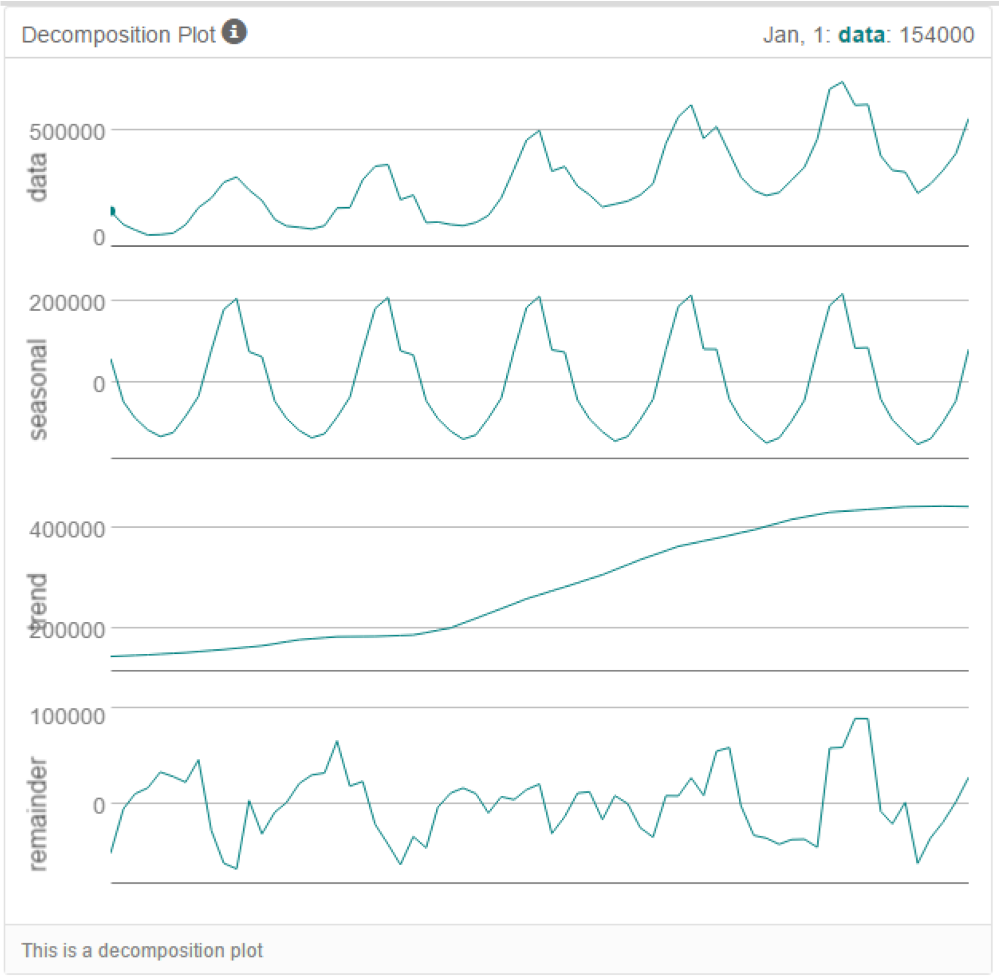
2. Which records should be used as the holdout sample?
   For external validation, we determine the accuracy measures by comparing the forecasted values with the holdout sample. This is important to compare ETS models to other types of models like ARIMA models. In this case we filter out the 4 most recent records as the holdout sample which is equal to what we are going to predict.

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.
   First, we use TS Plot Tool to understand the time series data and determine how to develop a forecasting model. The decomposition plots from this tool is presented below. As you can see from the graphs, the time series data has a clear trend. It also has seasonality portion and an increasing error portion. These plots suggest that Error component is fluctuating between large and small errors over time. We also see a linear trend rather than an exponential growth. From Seasonality chart, we see that the magnitudes of the waves are changing slightly over time, suggesting that our ARIMA model needs seasonal differencing.



This is a decomposition plot

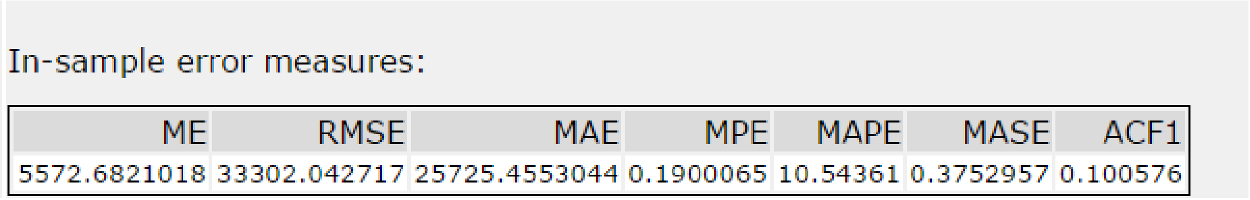
Step 3: Build your Models

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

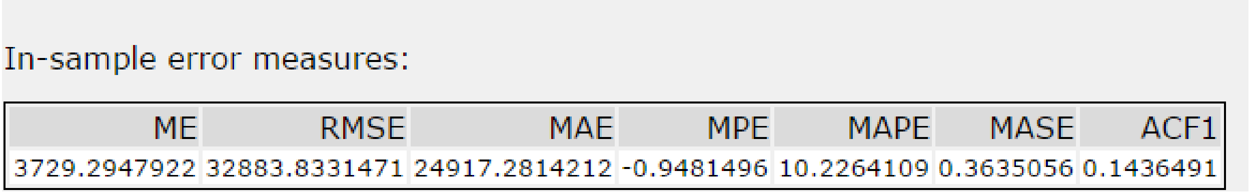
As you can see from the graphs above, the time series data has a clear trend. It also has seasonality portion and an increasing error portion. These plots suggest that we have Multiplicative Error Portion as the graph is fluctuating between large and small errors over time. We also see an additive trend, since the trend is linear and there is not an exponential growth. From Seasonality chart, we see that the ranges vary from early waves to the later ones. The size of the seasonal fluctuations tends to slightly increase over time, suggesting a multiplicative seasonality portion. Therefore, I would start with the ETS (M, A, M) at first and then test the Trend Dampening effect using 2 different model set-ups.



In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5572.6821018	33302.042717	25725.4553044	0.1900065	10.54361	0.3752957	0.100576

In_sample error measures for ETS(M,Ad,M)



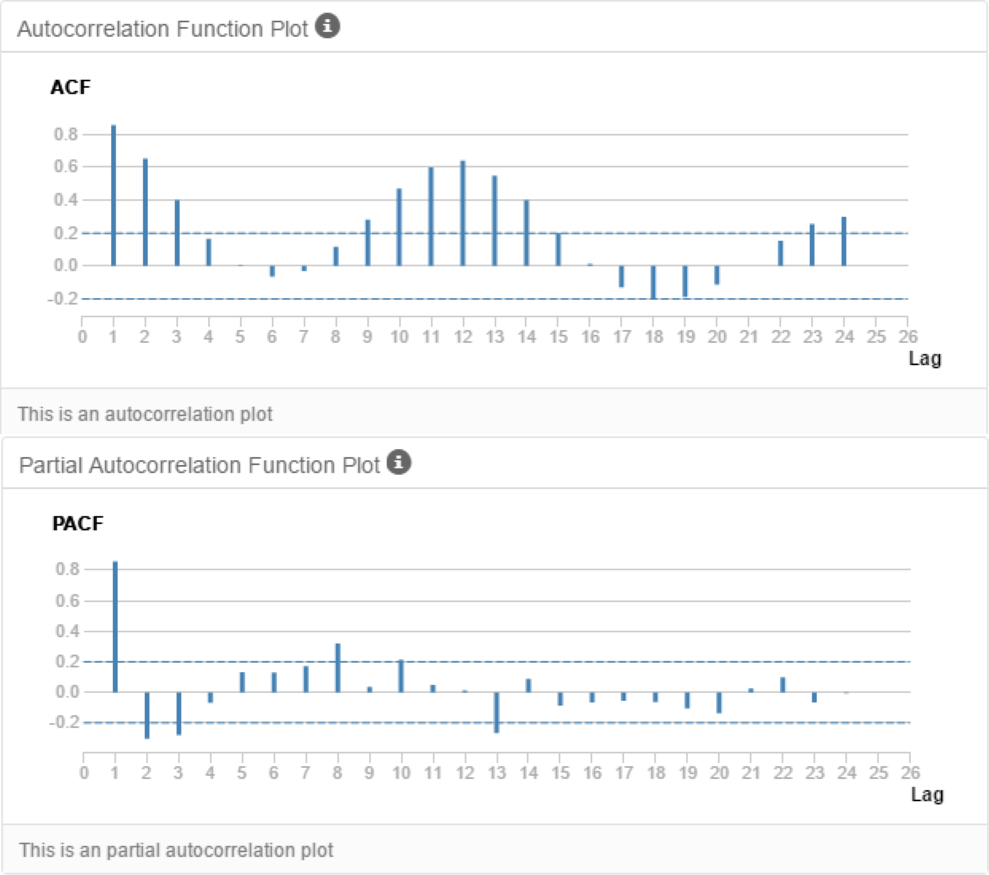
In-sample error measures:

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 for ETS(M,A,M)

Above, we have brought the in-sample error measures that we used to compare the 2 different settings of the ETS model. The first row shows the accuracy measures for the ETS (M,Ad,M) ETS model with Trend Dampening effects and the second is for the ETS (M,A,M) without Trend Dampening effects. The average error (ME) of Dampened model is smaller than that of the ETS(M,A,M). RMSE is a great measurement to use, as it shows how many deviations from the mean the forecasted values fall. The model without trend dampening has a smaller standard deviation from the mean (RMSE) than that of the ETS_MAdM. The further the MASE value from 1, the better the model is. Mean Absolute Percentage Error (MASE) for the Dampened model is further from 1 than the non-trend dampened model.
After reviewing the report output for the Dampened model and the non-dampened one, we see that the latter showed a lower AIC (Akaike Info. Criterion) of 1634.6 suggesting that the Non-Dampened setup will produce better results moving forward. So, we use the model without the Trend Dampening effects.

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.
   We first started with a TS Plot to understand the time series data and see the Trend and seasonality of the time series. The ACF and PACF before any differencing are presented below. As indicated before, we have seasonality component and therefore, we need to seasonally difference the time series to remove the seasonal portion. We will analyze the ACF and PACF throughout the process.


Autocorrelation Function Plot

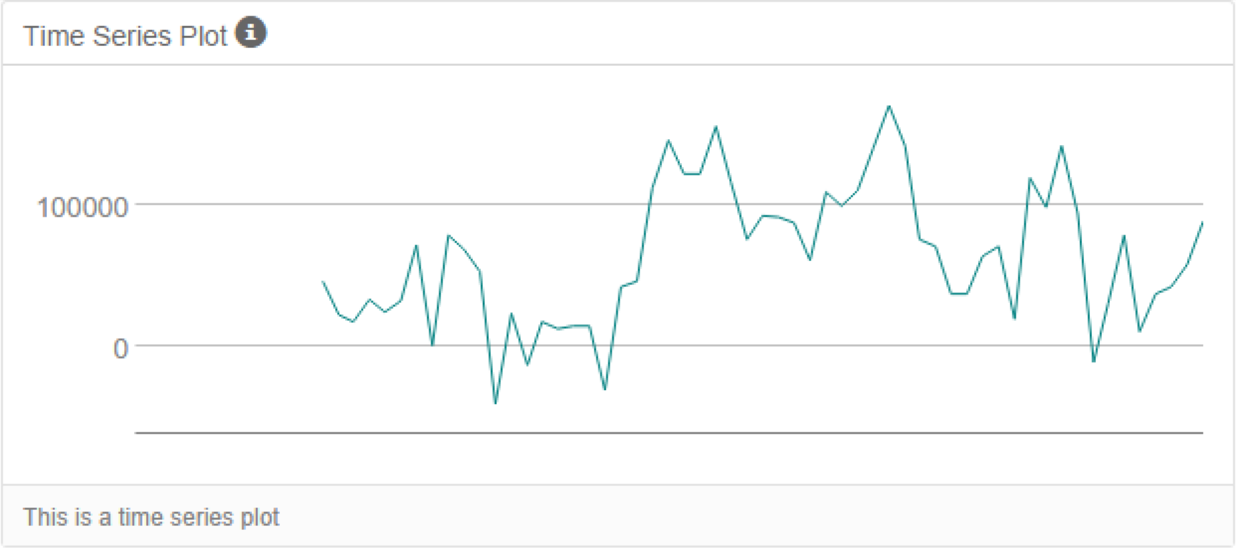
This is an autocorrelation plot

Partial Autocorrelation Function Plot

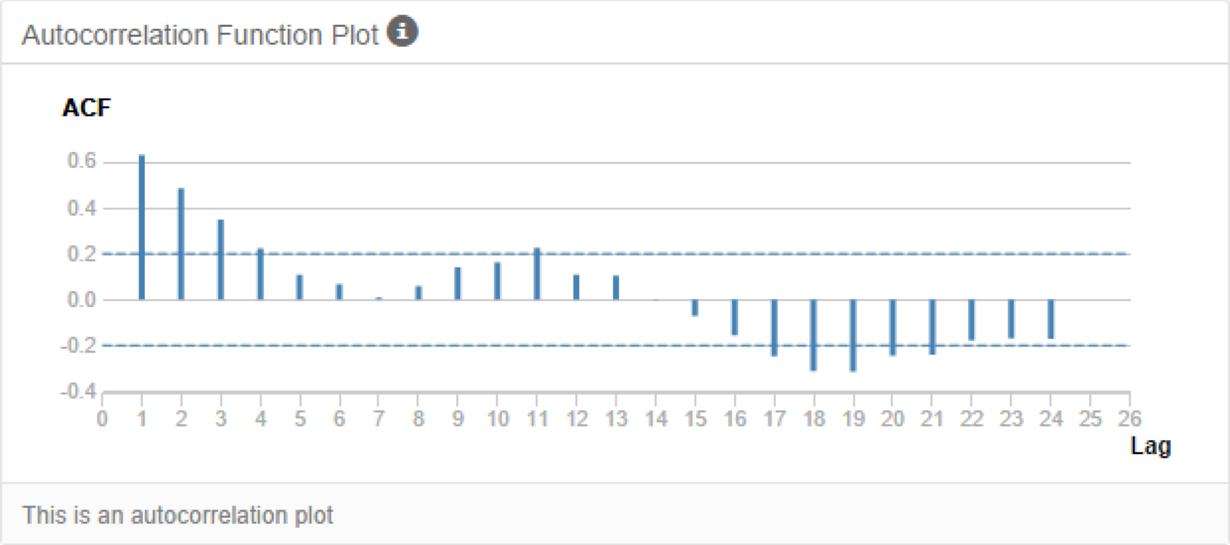
This is an partial autocorrelation plot

ACF and PACF Before any differencing

We conducted seasonal differencing using a Formula tool and a TS Plot to see the Time Series Plot after seasonal differencing. The plots are shown below. As you can see the timeseries is not stationary yet, so we use differencing to stationarize the data.

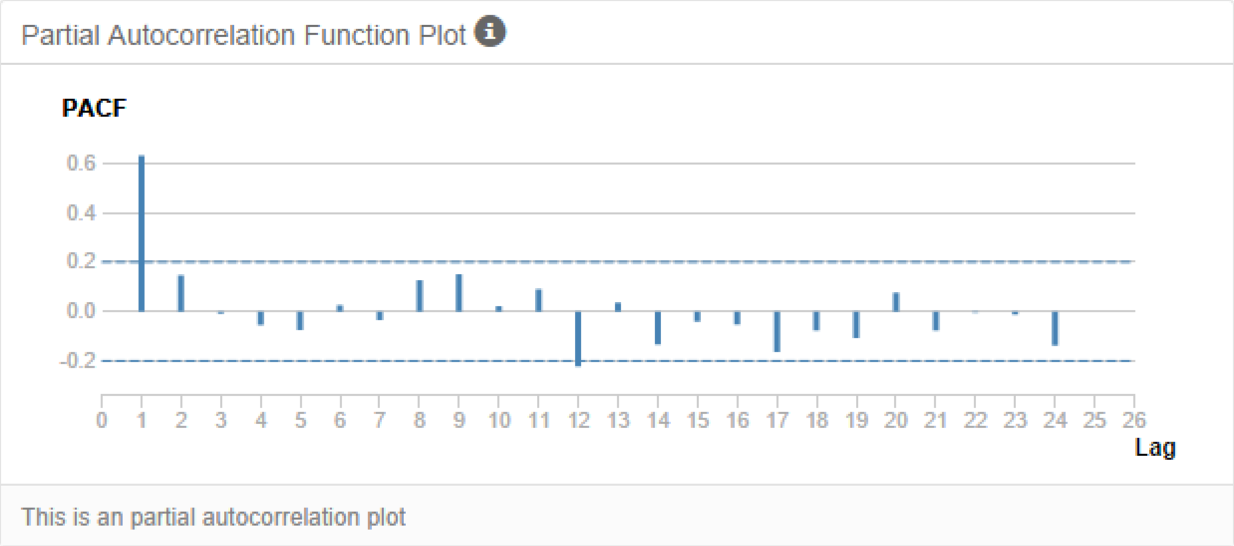

Time Series Plot

This is a time series plot



Autocorrelation Function Plot

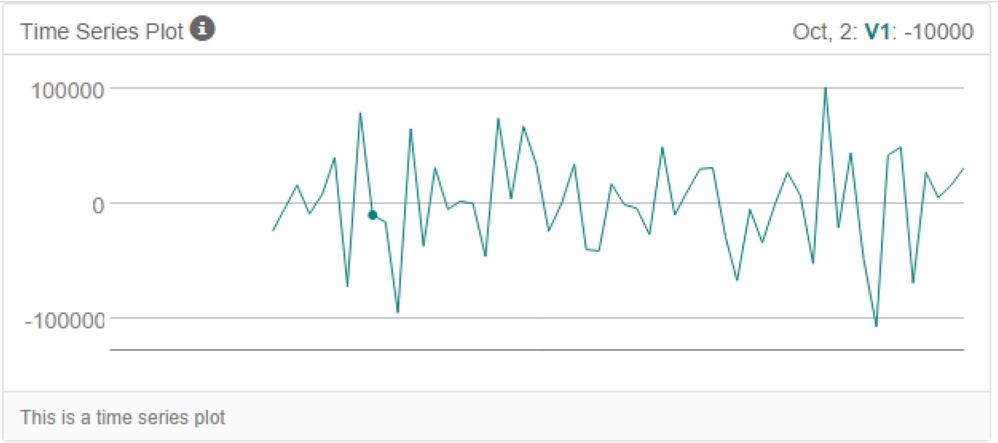
This is an autocorrelation plot



Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

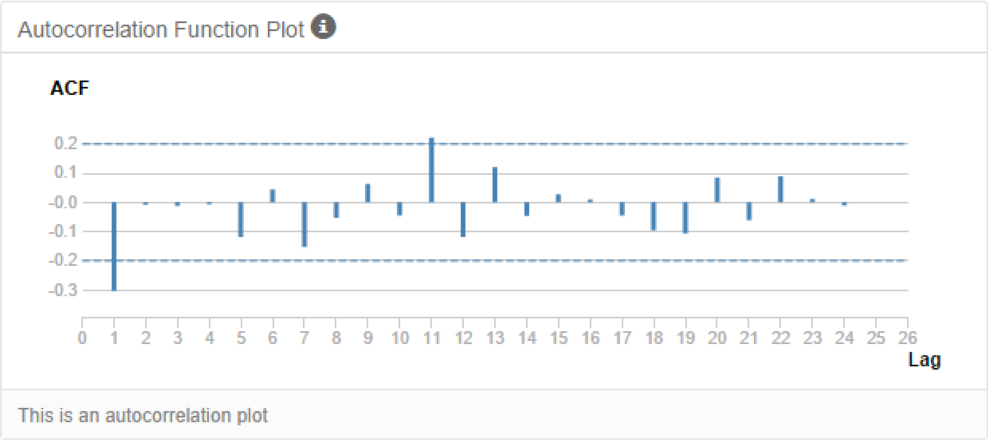
The results of the first differencing of the seasonal portion is presented below. The dataset is obviously stationary, as it fluctuates around zero line and has a constant range for its peaks to valleys. Therefore, the Capital “D” term is 1.
As you can see from the ACF and PACF plots below, there is a significant spike at lag 1 and then it drops off to 0 at lag 2. The PACF graph also shows a decay towards zero from the first lag. Because there is no other significant spike other than the one at lag 1, they both suggest a MA and AR term of 0. So, our seasonal terms are (0,1,0).



Time Series Plot

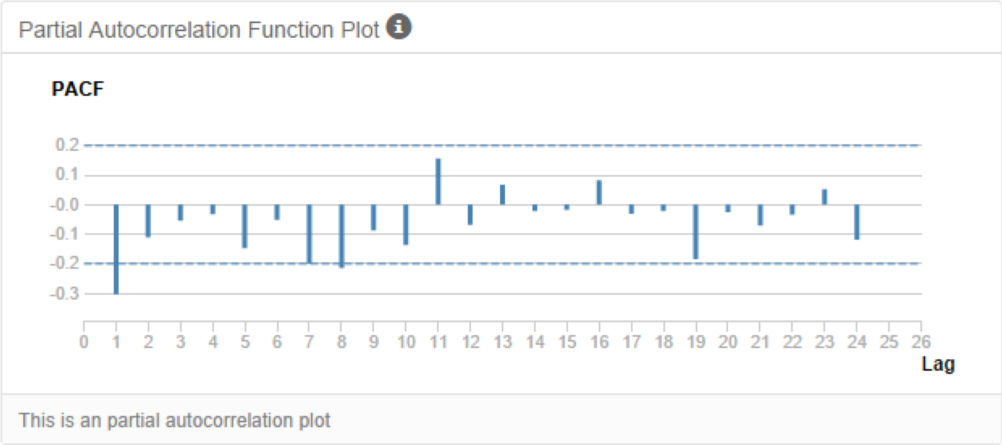
Oct, 2: **V1**: -10000

This is a time series plot



Autocorrelation Function Plot

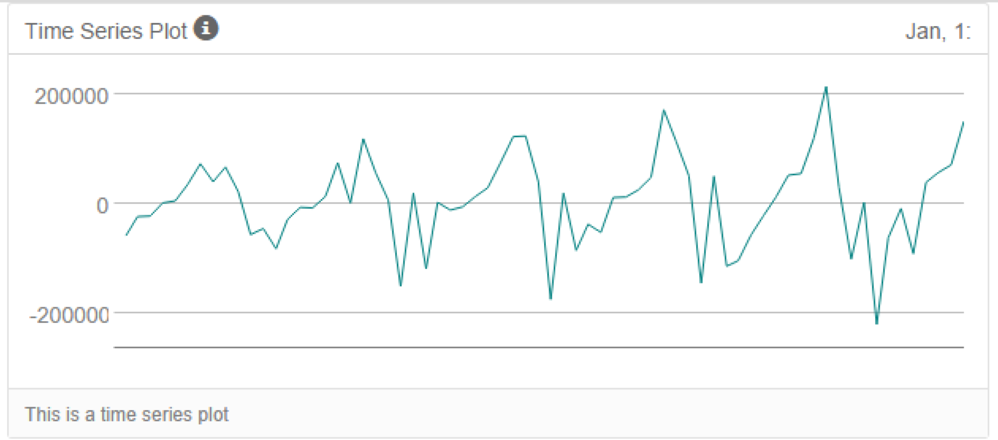
This is an autocorrelation plot



Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

Now, we calculate the first difference of the dataset to stationarize it. As you can see below, the dataset fluctuates around 0 median line suggesting that the data set is stationarized after one round of differencing.



Time Series Plot

This is a time series plot

Analyzing the plots, we have a suggestion of a MA(1) model because after we stationarized the dataset through differencing, there is strong negative autocorrelation (spikes) which is confirmed in PACF. As for the differencing term (d), we conclude that after the first differencing the dataset was stationary and therefore d=1.

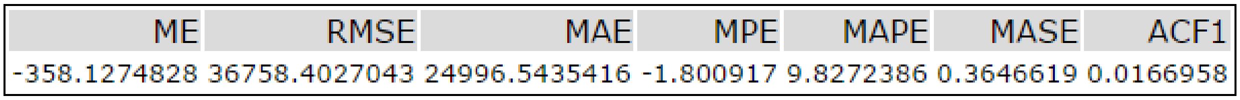
Considering the presented plots and discussions above, our final terms for our Seasonal ARIMA model would be ARIMA(0,1,1)(0,1,0)12.

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

In order to compare the different settings and to figure out which one produces the best results, we have compared the following 2 models.

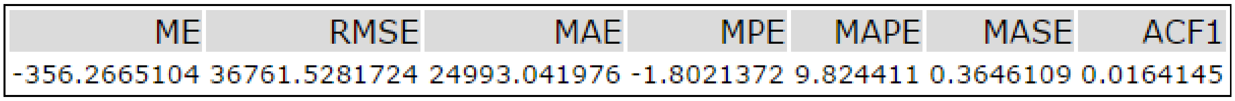
- ARIMA(0,1,1)(0,1,1)
- ARIMA(0,1,1)(0,1,0)

The In-sample error measures for both models are presented below:



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

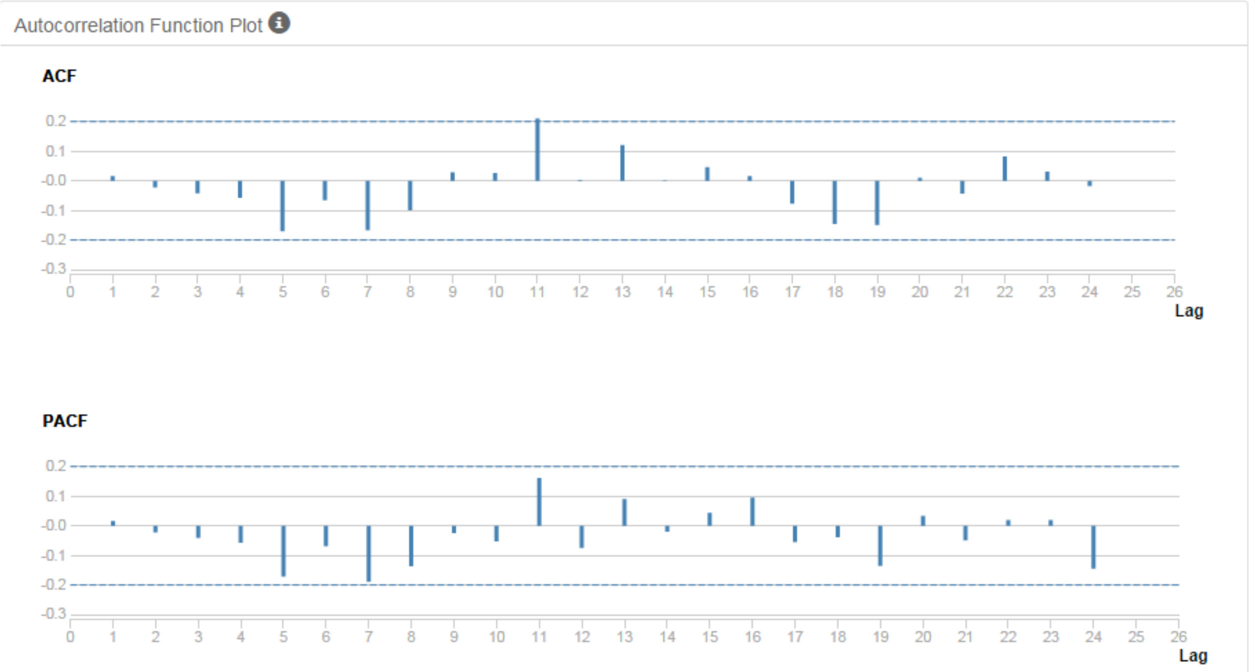
In-sample errors for ARIMA (0,1,1)(0,1,1)



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

In-sample errors for ARIMA(0,1,1)(0,1,0)

Looking at important in-sample error measures for the 2 models, the model presented in the second row (ARIMA (0,1,1)(0,1,0)) has shown smaller ME, MAE, MASE and also smaller Akaike Information Criterion (AIC) of 1256.84 compared to the AIC of 1259.09 for the ARIMA (0,1,1)(0,1,1) model. Hence we choose ARIMA(0,1,1)(0,1,0)12.
The plots of ACF and PACF after adding MA and AR terms are presented below:



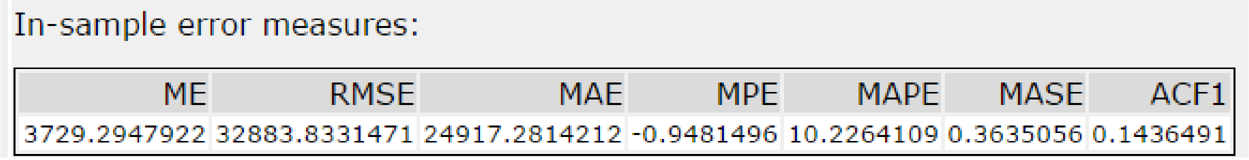
ACF and PACF after adding AR and MA terms

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.

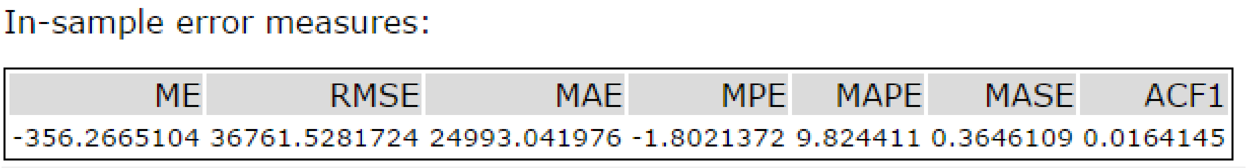
1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.
   The results of in-sample error measures for ETS and ARIMA models are shown below. They both have a very low MASE suggesting that they are reliable models. The AIC value for the ARIMA model is 1256.84 and for the ETS model is 1645.97. Comparing the 2 AIC values, we conclude that the ARIMA model is the better model moving forward.



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 for ETS model

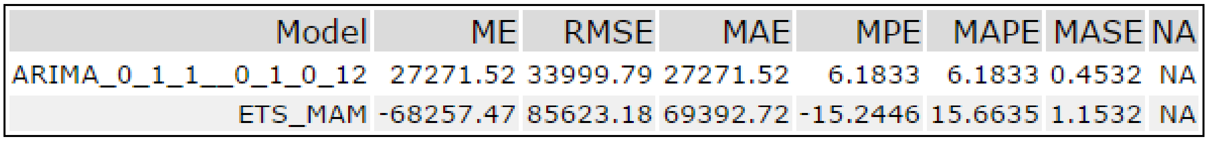


In-sample error measures:

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

In-sample error measures for ARIMA

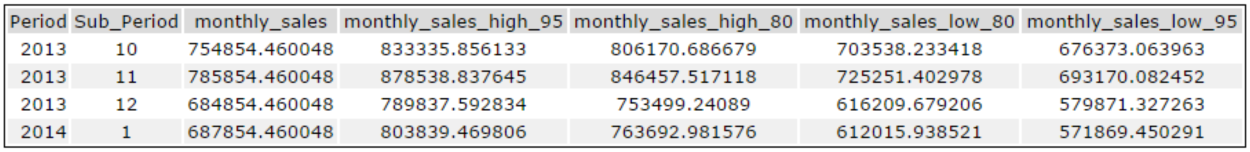
To compare the ETS against the ARIMA model using the holdout sample, we use union tool to create a dataset of the Object output for the 2 models and compare them side by side. The holdout sample of 4 month is then brought to TS compare tool to determine which model produces better results. The report output from the TS compare tool is presented below:



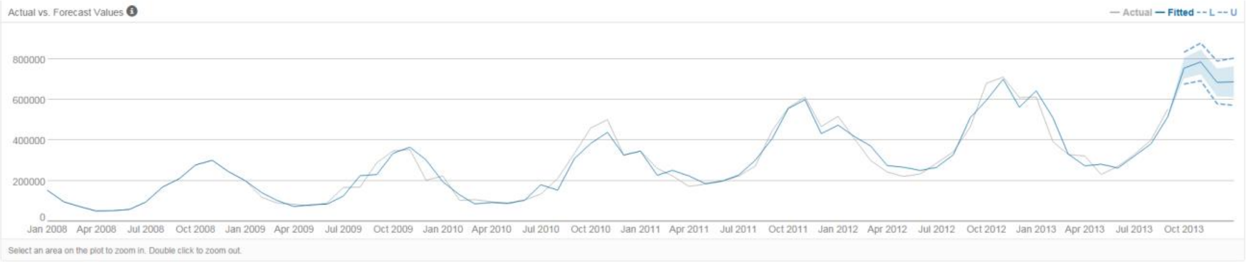
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_0_1_1__0_1_0_12	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA
ETS_MAM	-68257.47	85623.18	69392.72	-15.2446	15.6635	1.1532	NA

From the accuracy measures (table above), it is evident that the ARIMA model has better predictive qualities in almost all metrics. Hence ARIMA is the best fitting model (smaller RMSE, smaller MASE, smaller ME, MAE, MPE, and smaller MAPE). Therefore, we will use the ARIMA model to forecast the next 4 periods.

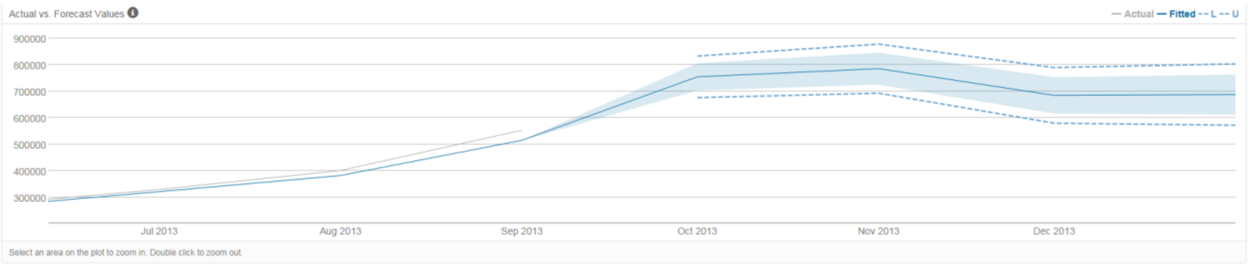
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.
   Using TS Forecast tool and ARIMA model Object output, we bring the entire dataset of 69 months into the TS Forecast tool and create a forecast for the next 4 months. Below is a table showing the results from the forecasting tool using 95% and 80% confidence intervals.



Period	Sub_Period	monthly_sales	monthly_sales_high_95	monthly_sales_high_80	monthly_sales_low_80	monthly_sales_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



We zoomed into the area of the last 4 periods for better clarity. Please see the graph below:



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.