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.

ANS – To meet the criteria of a time series dataset it should show the following characteristics -

- It's over a continuous time interval
- There are sequential measurements across that interval
- There is equal spacing between every two consecutive measurements
- Each time unit within the time interval has at most one data point

Since our dataset displays all the above four characteristics it can be qualified as a time series dataset

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

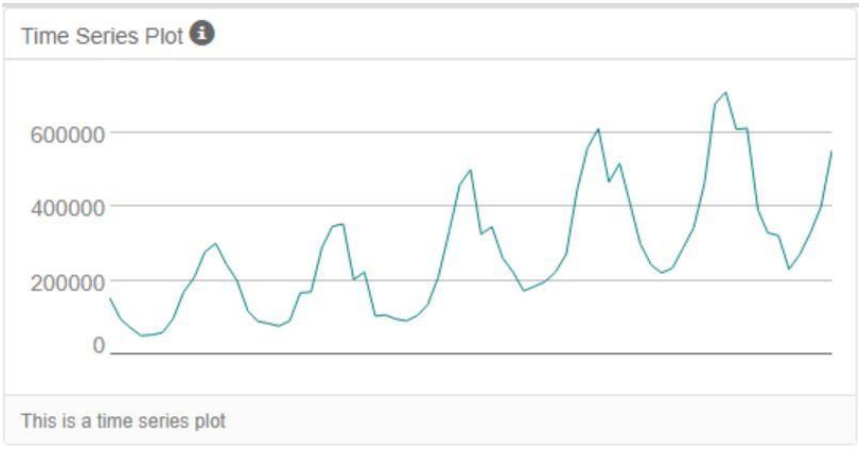
ANS - We need to use 4 records in the holdout sample. These will be the records – 2013-06 , 2013-07, 2013-08, 2013-09. Basically we used the last four records for the holdout sample because we need to predict the sales data for next 4 months.

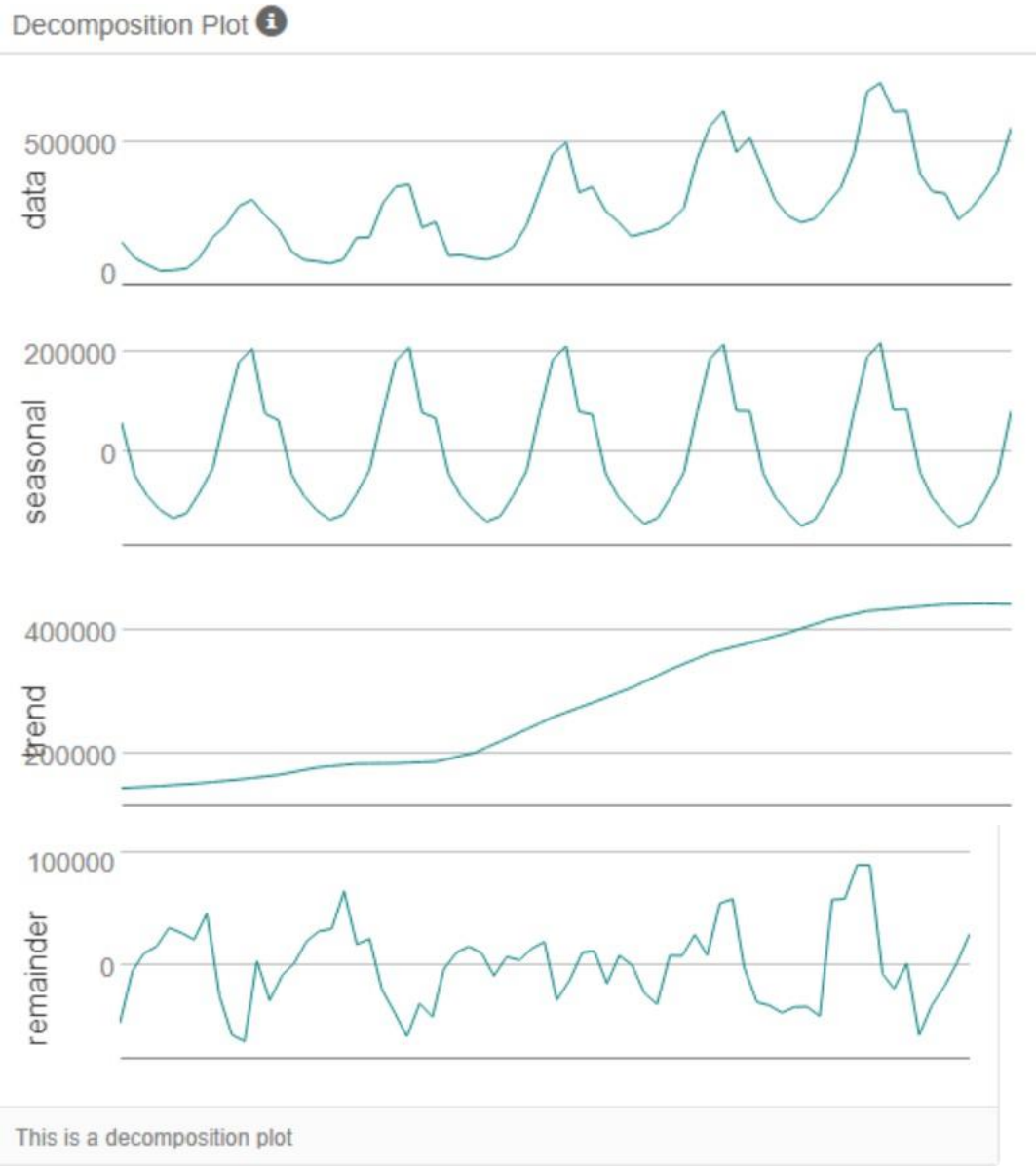
Step 2: Determine Trend, Seasonal, and Error components

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.





ANS – As can be seen in above plots

1. Seasonality has shown an increasing trend hence it should be applied multiplicatively
2. Trend appears to be linear so it should be applied additively
3. For error plot, there isn't a trend but rather fluctuations and thus should beapplied multiplicatively as well

Step 3: Build your Models

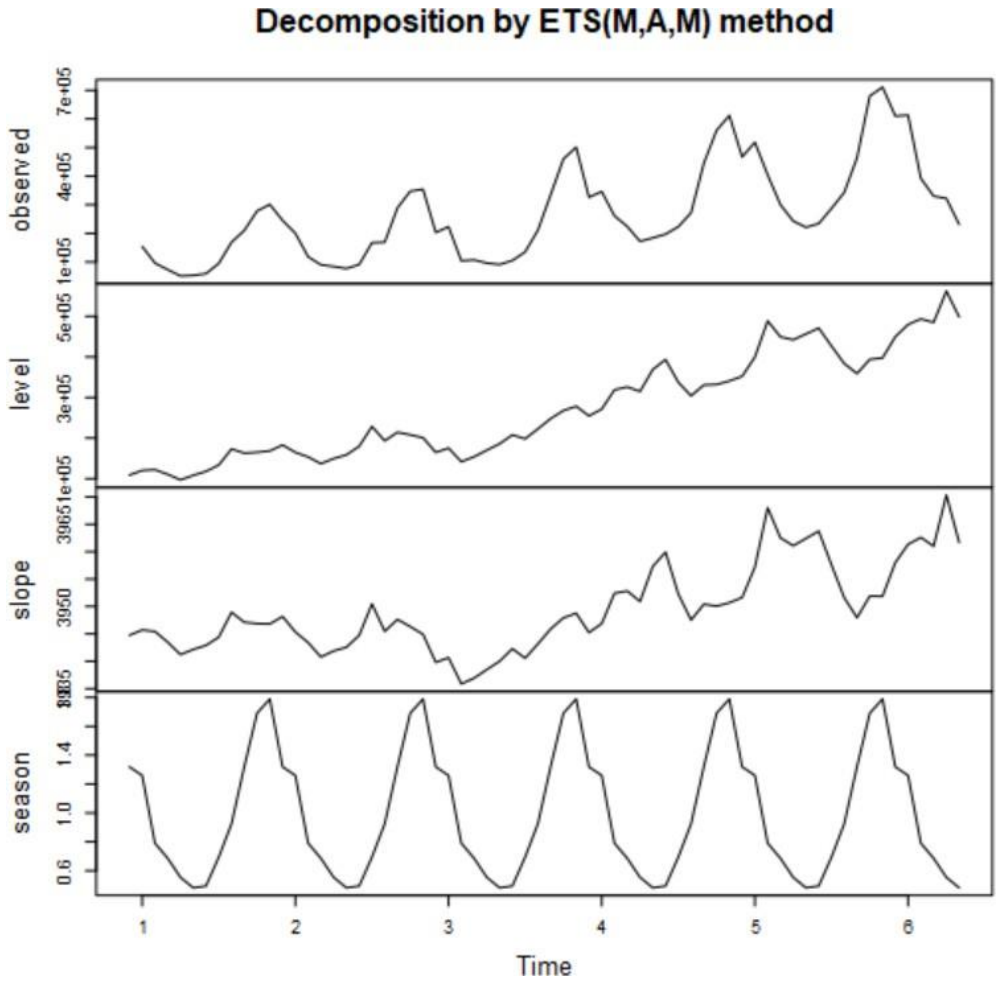
Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)

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

Ans – For ETS I chose error being applied multiplicatively, trend being applied additively and seasonality being applied multiplicatively. This MAM form of model was chosen based on the seasonality and trend plots showing increasing trend, whereas with error plot there were no trend but there were fluctuations which suggested that error should be applied multiplicatively. We took to ETS models with these parameters – one with dampening on and other with no dampening to make a comparison.

Part 1 – ETS non dampened model.



In-sample error measures:

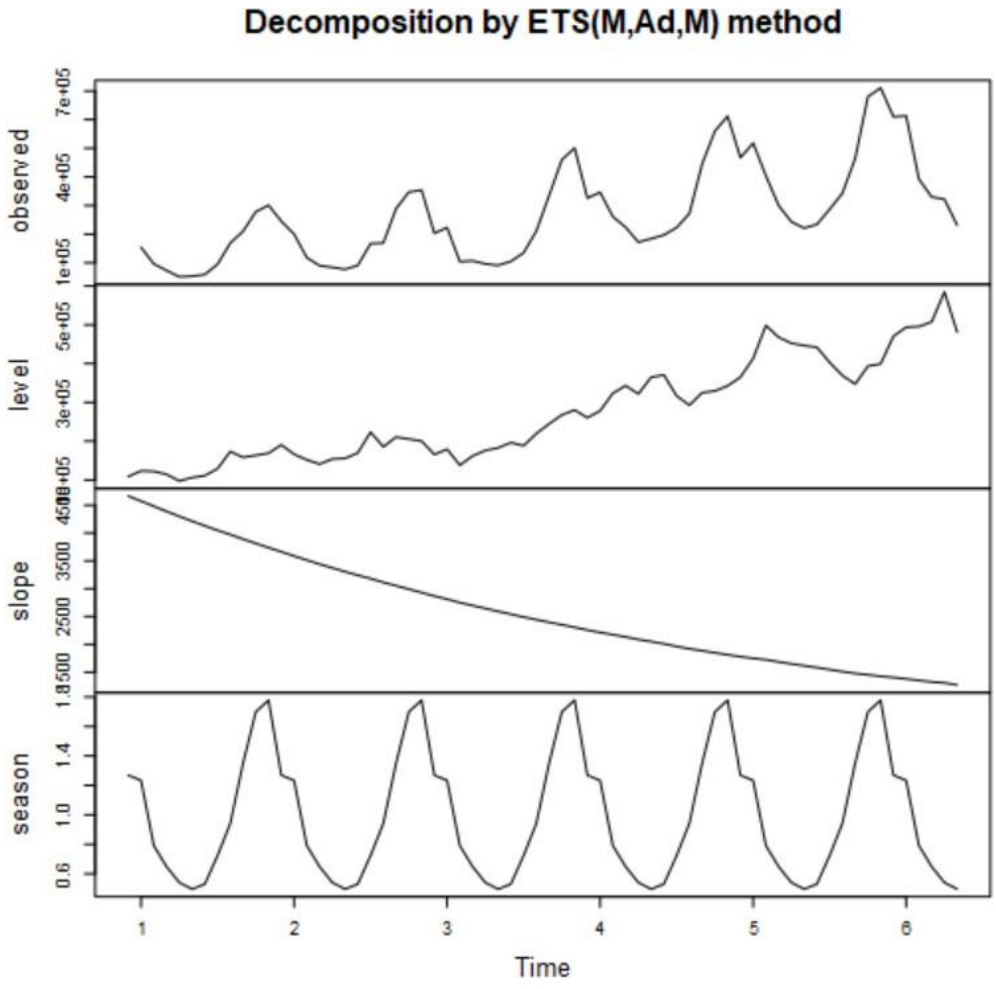
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

Information criteria:

AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

AIC value is 1639.7367, MASE is 0.372685, RMSE is 32992.7261011

Part 2 ETS dampened model



In-sample error measures:

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

Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

ETS non dampened comparison plot

Actual and Forecast Values:

Actual	etsnondamp
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
etsnondamp	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066

ETS dampened comparison plot

Actual and Forecast Values:

Actual	etsdamp
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
etsdamp	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116

Hence we will choose ETS dampened model for calculations because -

a) ETS dampened model has a lower MASE. MASE is defined as the mean absolute error of the model divided by the the mean absolute value of the first difference of

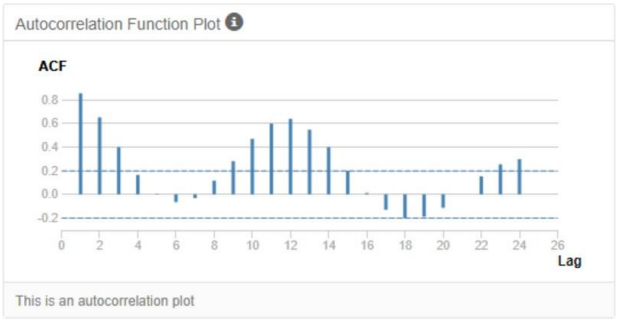
the series. Thus, it measures the relative reduction in error compared to a naive model. Ideally its value will be significantly less than 1 but is relative to comparison across other models for the same series. In our case, comparing the 2 models this is lower for the ETS dampened model.

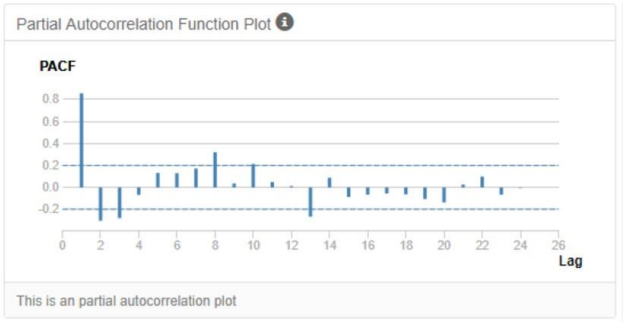
b) ETS dampened model has a lower MAPE. Mean Absolute Percentage Error (MAPE) is also often useful for purposes of reporting, because it is expressed in generic percentage terms it will make sense even to someone who has no idea what constitutes a "big" error in terms of dollars spent or widgets sold.

c) RMSE for ARIMA model incase of the holdout sample by comparison is lower. RMSE represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. This is a good measurement to use when comparing models as it shows how many deviations from the mean the forecasted values fall.

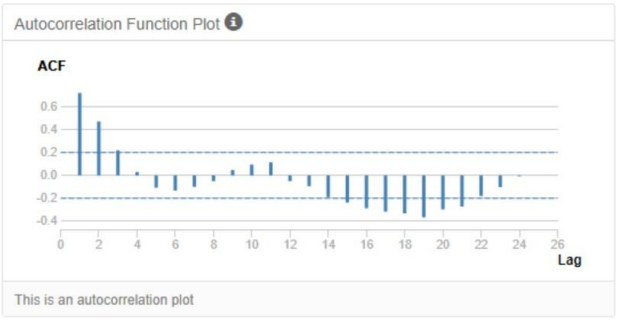
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
    a. Describe the in-sample errors. 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.

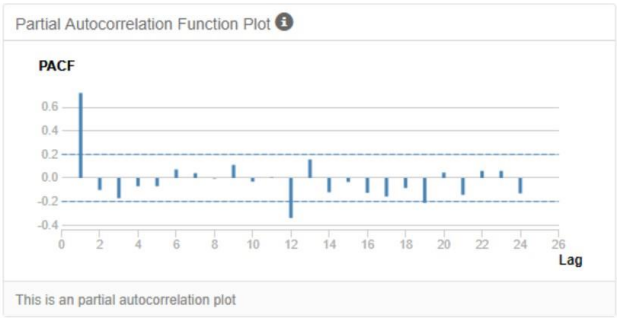
ANS-



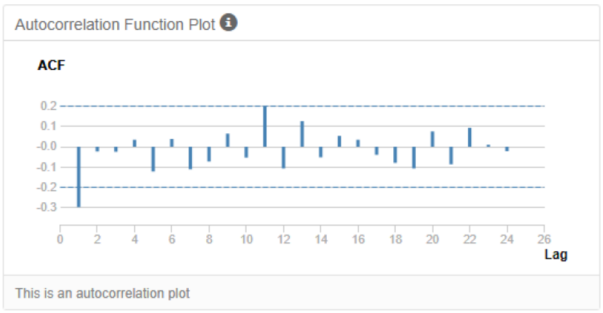


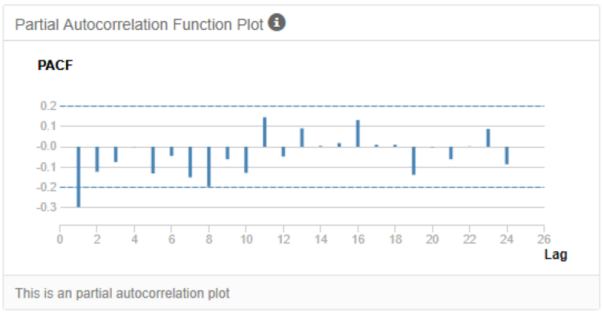
First we are showing the ACF and PACF plots of the undifferenced series. We can clearly see the series is not stationarized because of the strong correlation values that we can see from several occurring significant values. Hence our first task will be to stationarize the series
After taking the first difference in the series these are the acf and pacf plots





The series is still not stationarized as can be seen from the strong correlation values so we will take the first seasonal difference.





Based on the above acf and pacf plots we will take values - ARIMA(0,1,1)(0,1,0)12 is used as lag-1 is negative and the number of period is 12 months.

Information Criteria:

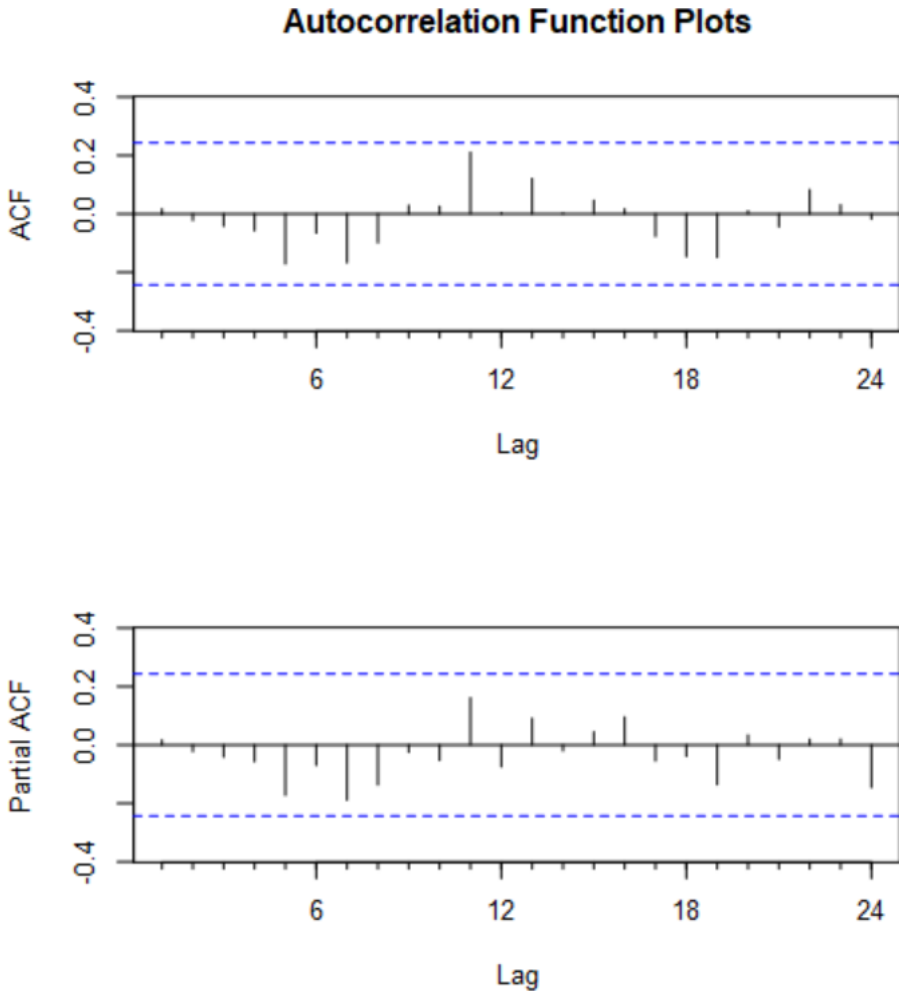
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

1. AIC is 1256.5967 . This is a measure of the quality of the statistical model and the lower this value is the better the model
2. RMSE is 36761.5281724 . Root Mean Squared Error (RMSE) represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. This is a great measurement to use when comparing models as it shows how many deviations from the mean the forecasted values fall.
3. MASE is 0.3646109. **Mean Absolute Scaled Error (MASE)** is another relative measure of error that is applicable only to time series data. It is defined as the mean absolute error of the model divided by the the mean

absolute value of the first difference of the series. Thus, it measures the relative reduction in error compared to a naive model. Ideally its value will be significantly less than 1 but is relative to comparison across other models for the same series. In our models case this value is significantly less than 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.

ANS - I am chosing the ARIMA (0,1,1)(0,1,0)12 Model because it is better based on in sample error and forecast error measurements against the holdout sample.

In-sample error measurements for ARIMA Model.

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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 ETS dampened model.

In-sample error measures:

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

Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

We can see from both that arima model has better accuracy measures such as a lower AIC, lower MASE.

ARIMA holdout sample comparison

Actual and Forecast Values:

Actual	arima1
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
arima1	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

ETS holdout sample comparison

Actual and Forecast Values:

Actual	etsdamp
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

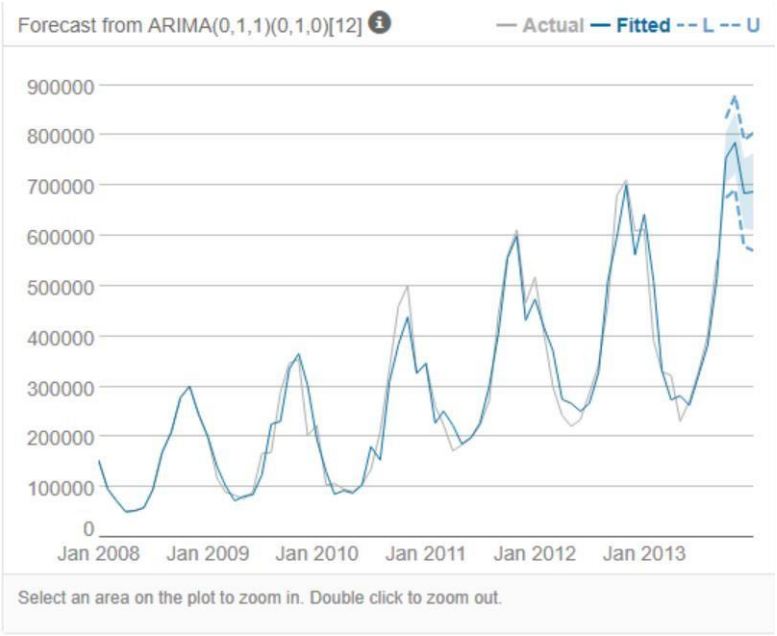
Accuracy Measures:

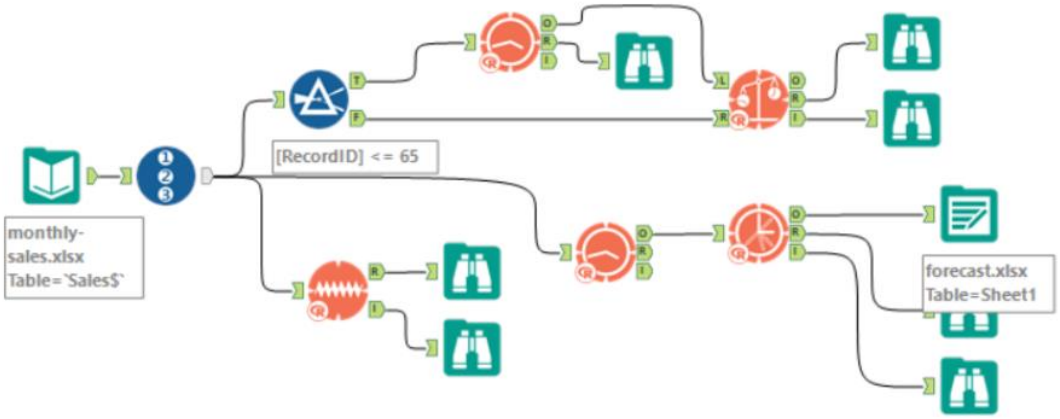
Model	ME	RMSE	MAE	MPE	MAPE	MASE
etsdamp	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116

From the above two plots we can see that ARIMA model is better at forecasting sales using holdout sample as validation data as the MAPE and ME value are lower than ETS model.

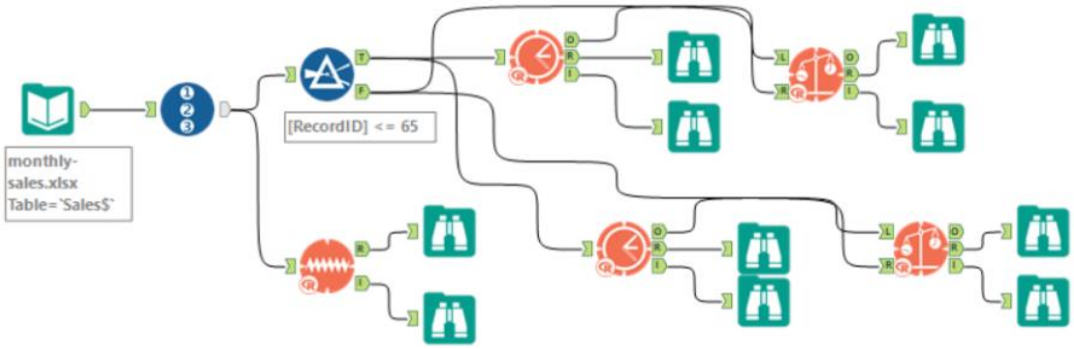
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

ANS - The forecast for the next four months is 754854.46, 785854.46, 684854.46, 687854.46





Workflow 1: ARIMA Workflow



Workflow 2: ETS Workflow

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.