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.

The dataset captures the month wise sales for a defined time period. The dataset meets the criteria of a time series as:

1. Data is sequential and ordered
2. Data is countinous
3. There is equal spacing between two consecutive measurements
4. Each time unit has at most one value

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

We have to predict data for 4 periods. Hence, the hold out sample would also be for 4 months. This would be from 2013-06 to 2013-09.

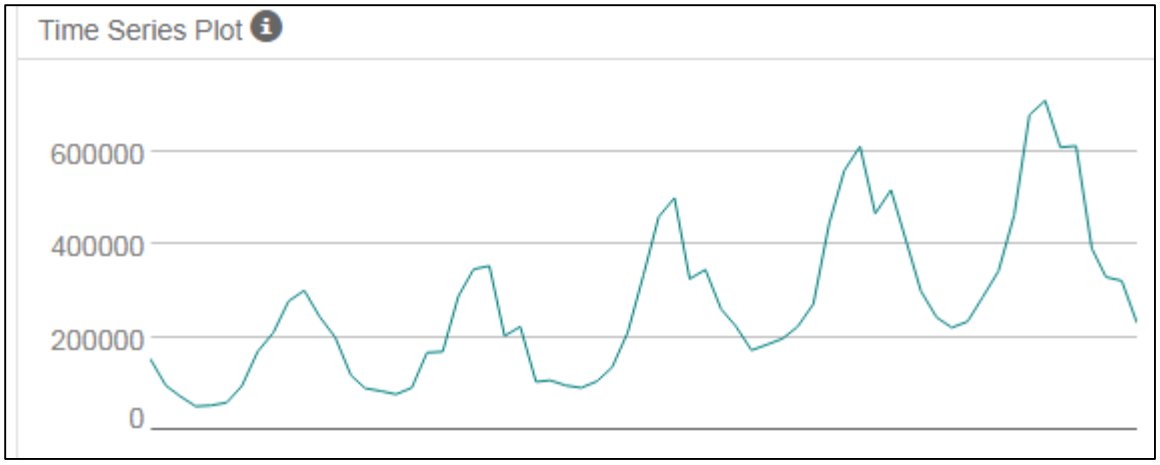
Step 2: Determine Trend, Seasonal, and Error components

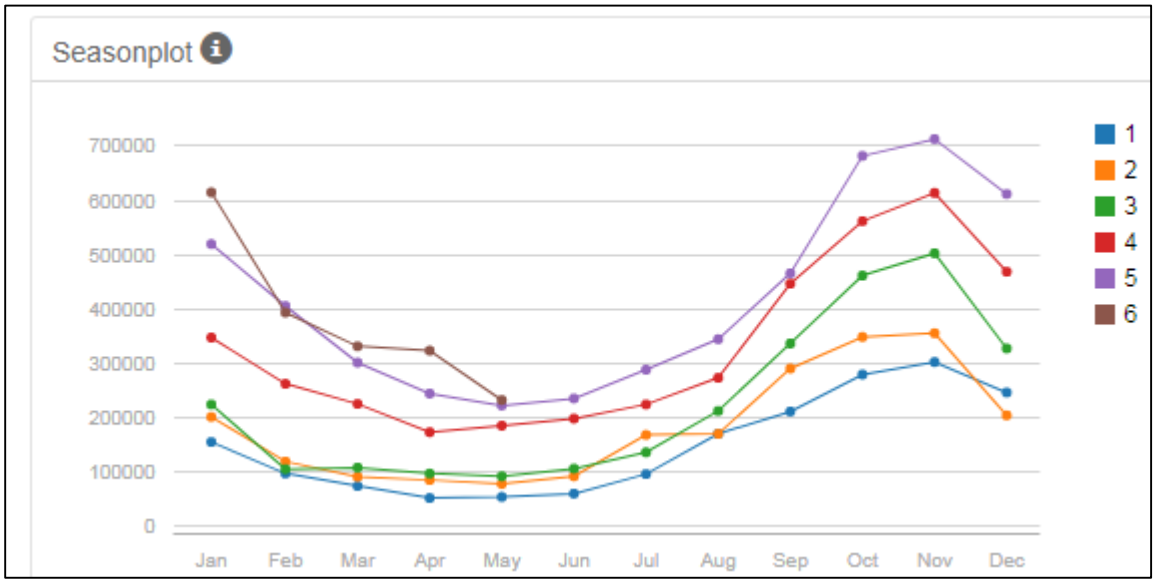
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

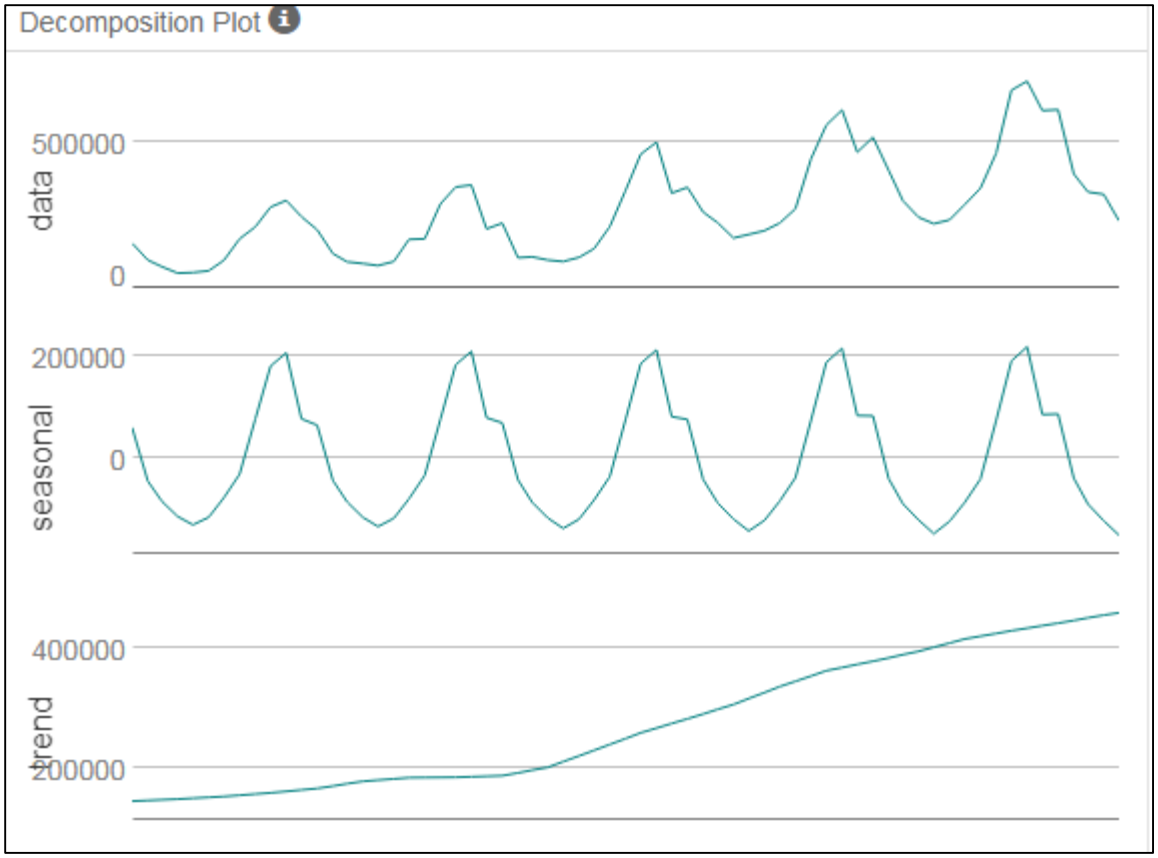
Answer this question:

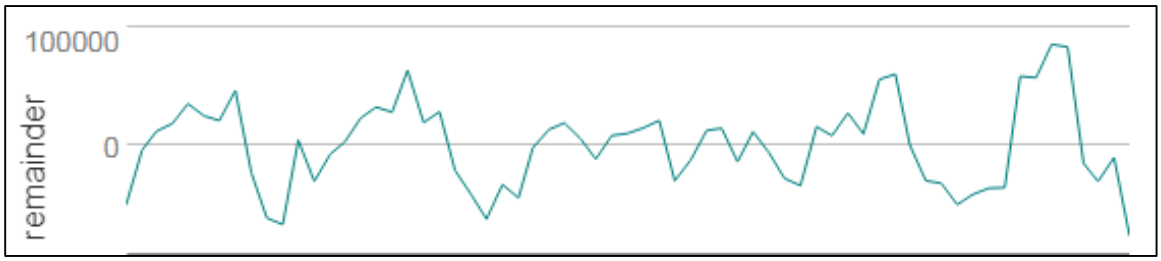
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

The time series plots are as below. They have been plotted using the TS Plot function.


Time Series Plot
600000
400000
200000
0


Seasonplot
700000
600000
500000
400000
300000
200000
100000
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
2
3
4
5
6





Key Inferences:

1. Trend- Increasing Linearly- Additive
2. Seasonality- Increasing - Multiplicative
3. Error- Increasing with time- Multiplicative

Step 3: Build your Models

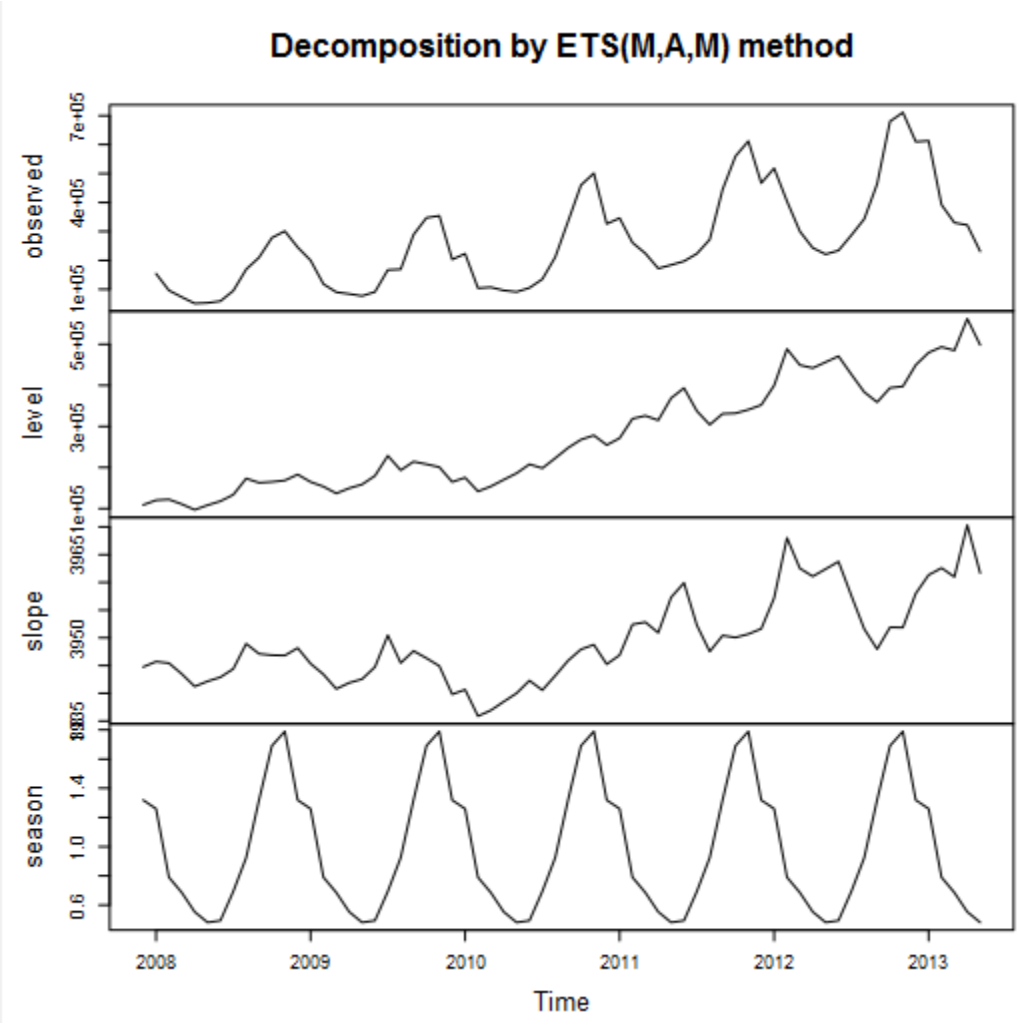
Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)

Answer these questions:

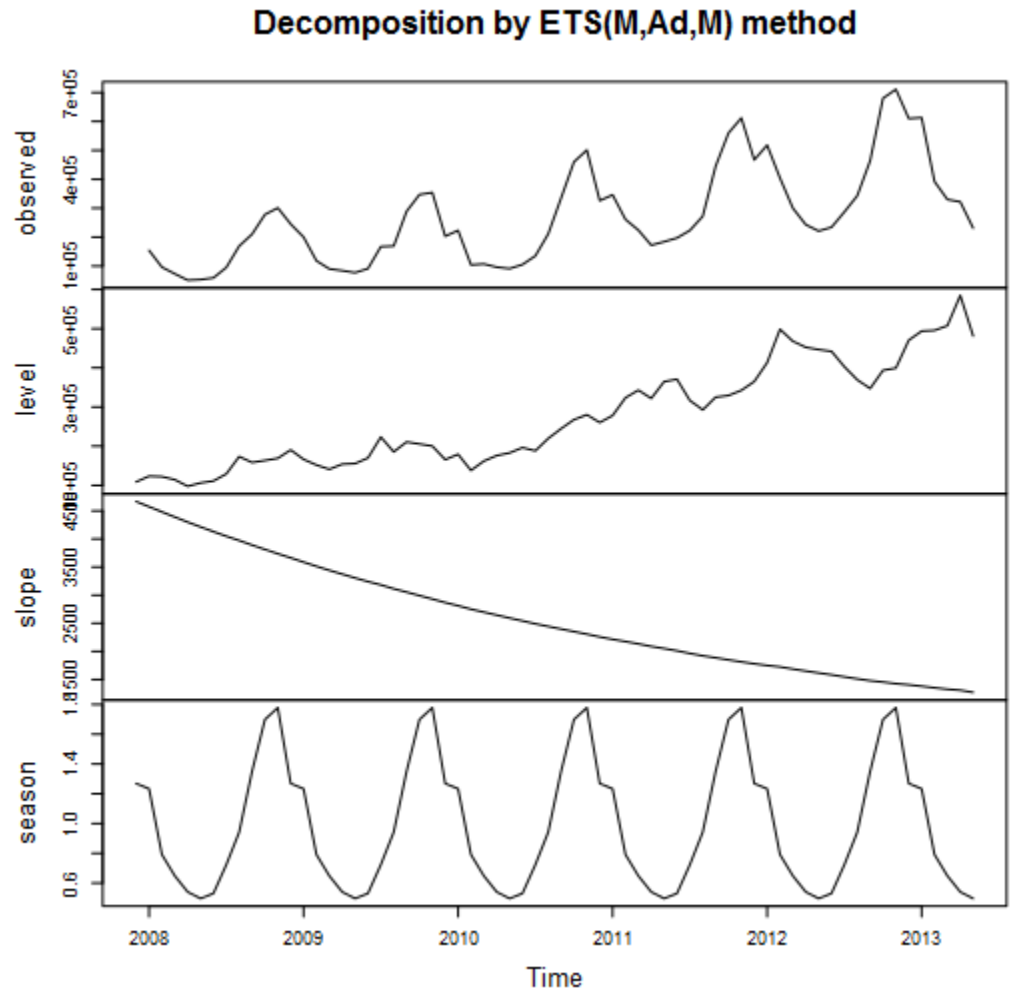
1. What are the model terms for ETS? Explain why you chose those terms.

I have chosen ETS(M,A,M) bases on the results of the decomposition plot. The dampened and non-dampened ETS models are as below. I have taken a hold out sample of 4 months.

Non-Dampened:



Dampened:



---

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

:

RMSE or Root Mean Square Error is a measure of how far from the regression line data points are. It indicates how concentrated data is around the best fit line.

MASE or Mean Absolute Scaled Error is also a measure of the accuracy of forecasts**.** This error is independent of the scale of the data, so can be used to compare forecasts across data sets with different scales.

In-Sample errors of Non-Dampened Model:

RMSE: 32992.73

MASE: 0.373 (This is less than 1 which is the threshold for model accuracy- hence this is a good forecast)

In-Sample errors of Dampened Model:

RMSE: 33153.53
MASE: 0.367 (This is less than 1 which is the threshold for model accuracy- hence this is a good forecast)

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

Dampened:

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

Accuracy Measures:

For the Dampened Model, both RMSE and MAE values are lower- as compared to Non-Dampened.

Non-Dampened

Actual and Forecast Values:

Actual	ETS_VideoSales
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

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

Dampened:

Actual and Forecast Values:

Actual	ETS_VideoSales_Dampen
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

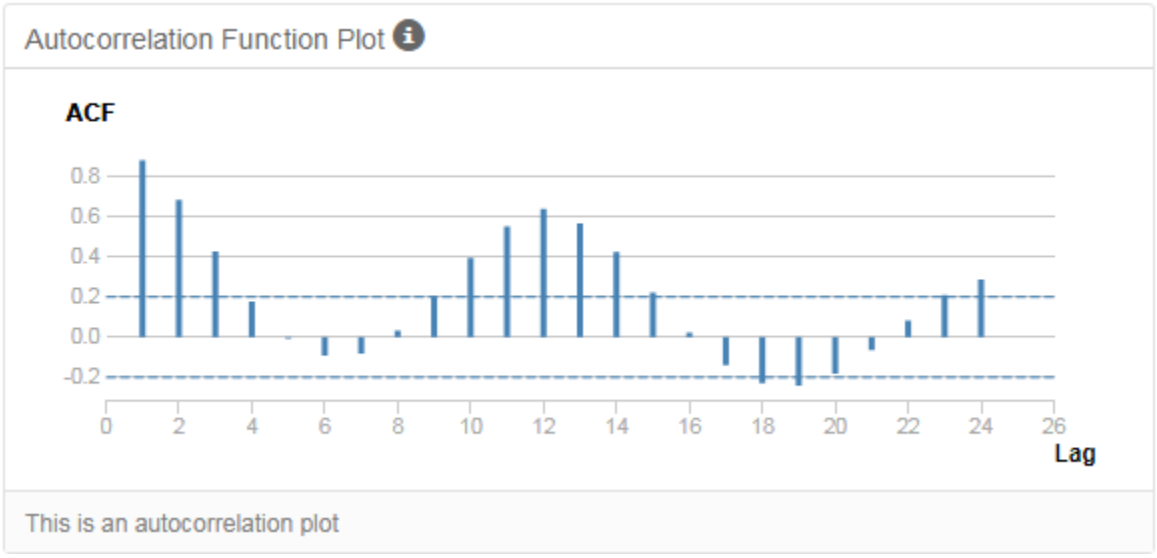
Accuracy Measures:

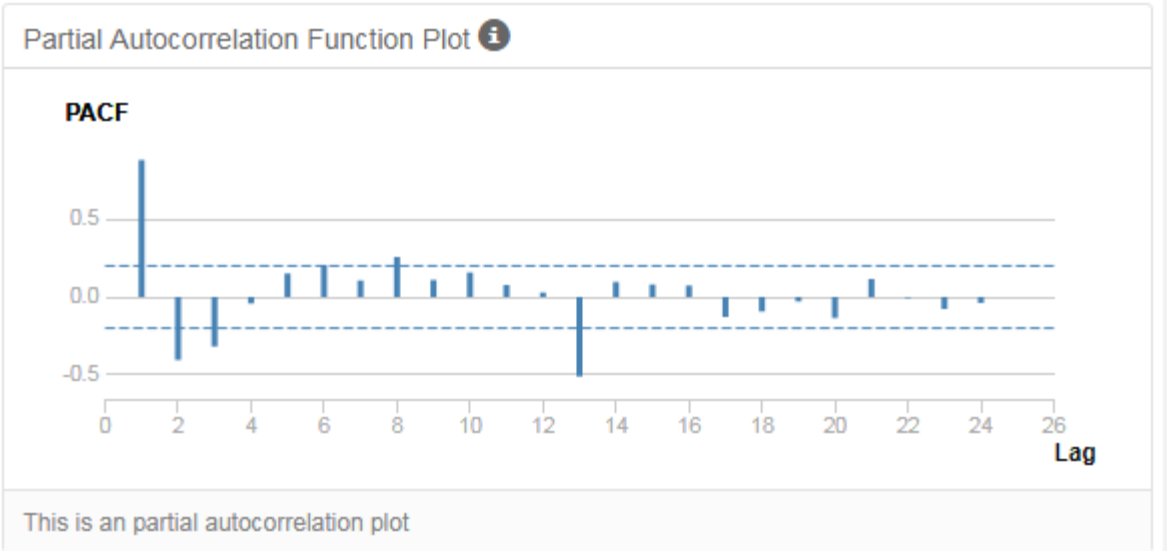
	Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_VideoSales_Dampen		-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116

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.

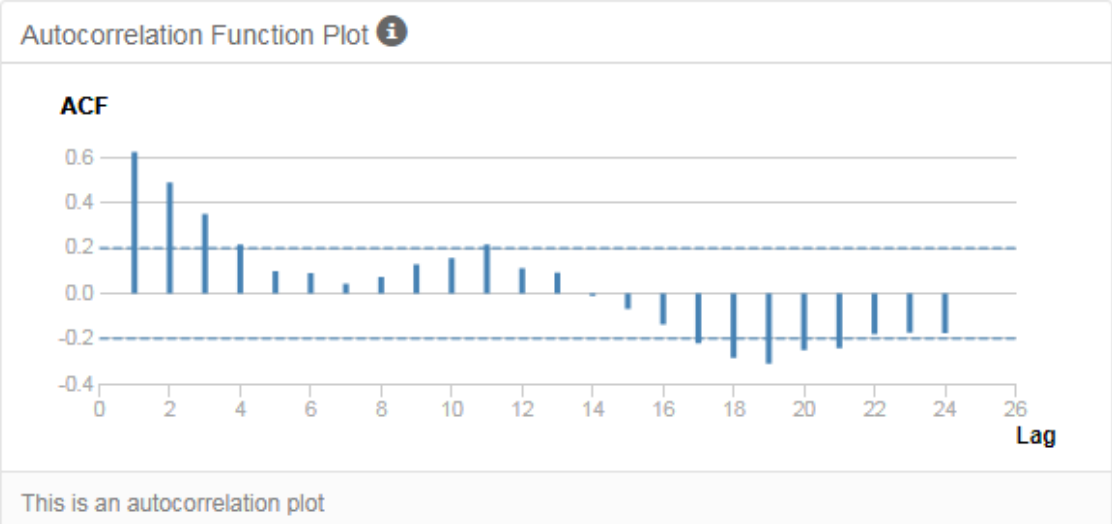
ARIMA models are displayed in the terms (p,d,q). These are explained as below:

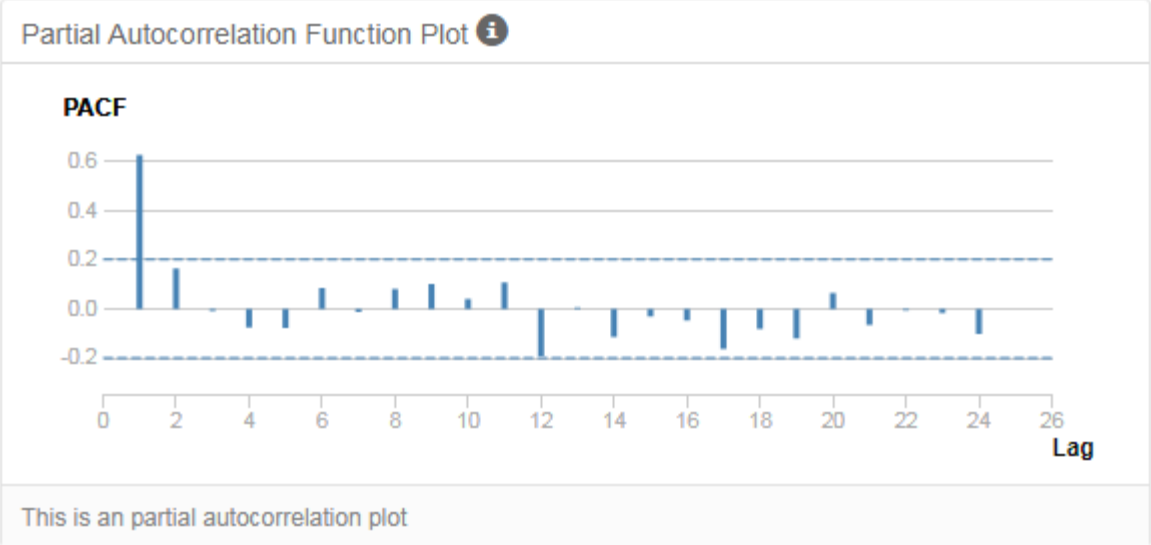
- p - periods to lag for
- d - number of transformations used to make the data stationary
- q - lags of the error component



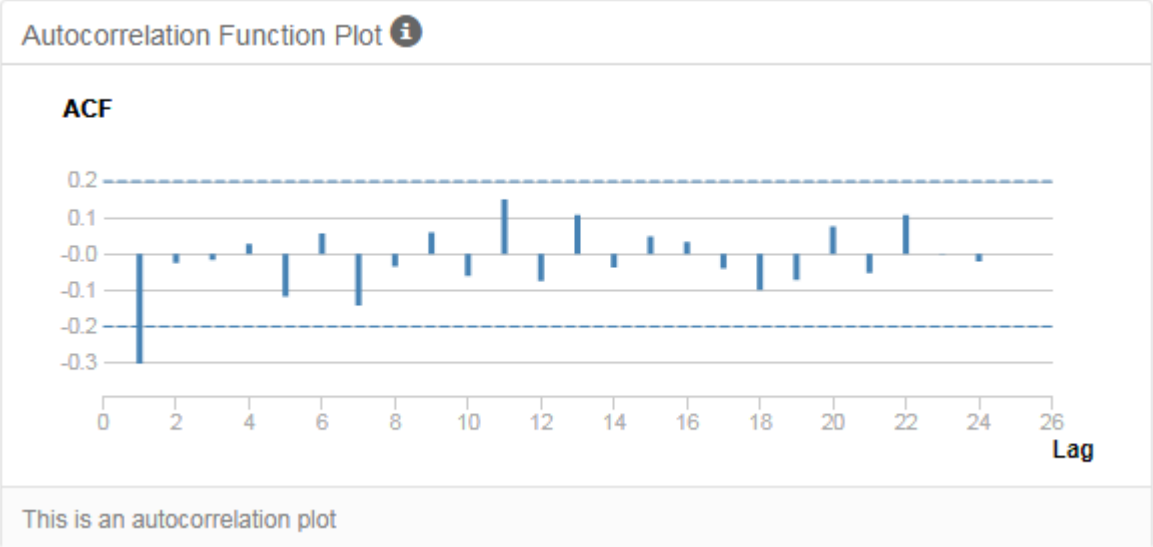


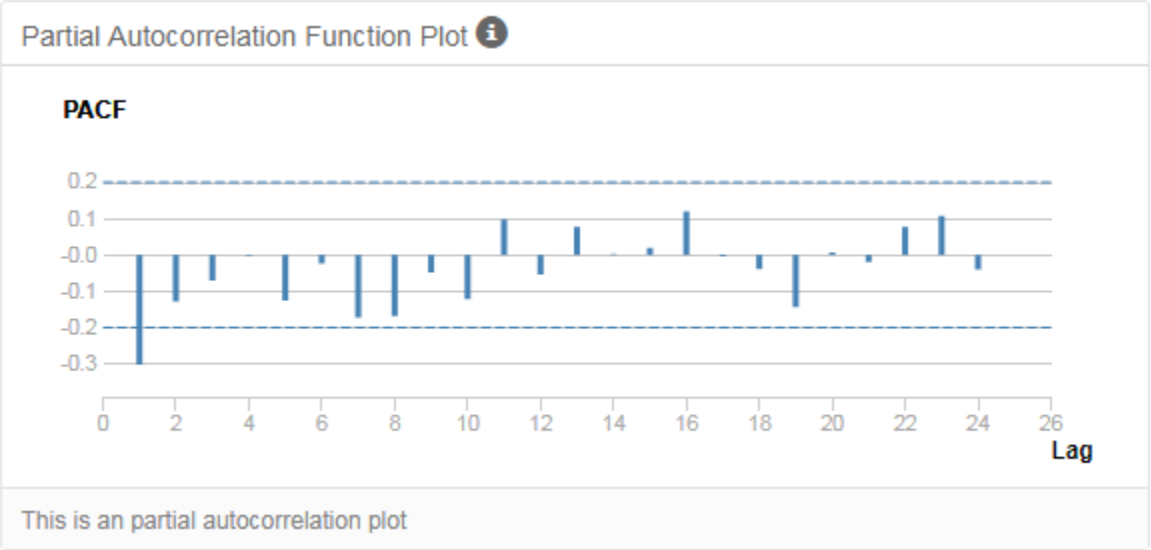
I have re-plotted the ACF and PACF graphs after taking 1 seasonal difference. Even after this, the ACF still shows high co-relation.





I have re-plotted the ACF and PACF graphs after taking the 1st difference. There is no significant co-relation now.





ARIMA(0,1,1)(0,1,0)12 is used as lag-1 is negative and the number of period is 12 months.

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

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

Call:
auto.arima(Monthly.Sales)

Coefficients:

	ma1
Value	-0.378032
Std Err	0.146228

sigma^2 estimated as 1722385234.94439: log likelihood = -626.29834

Information Criteria:

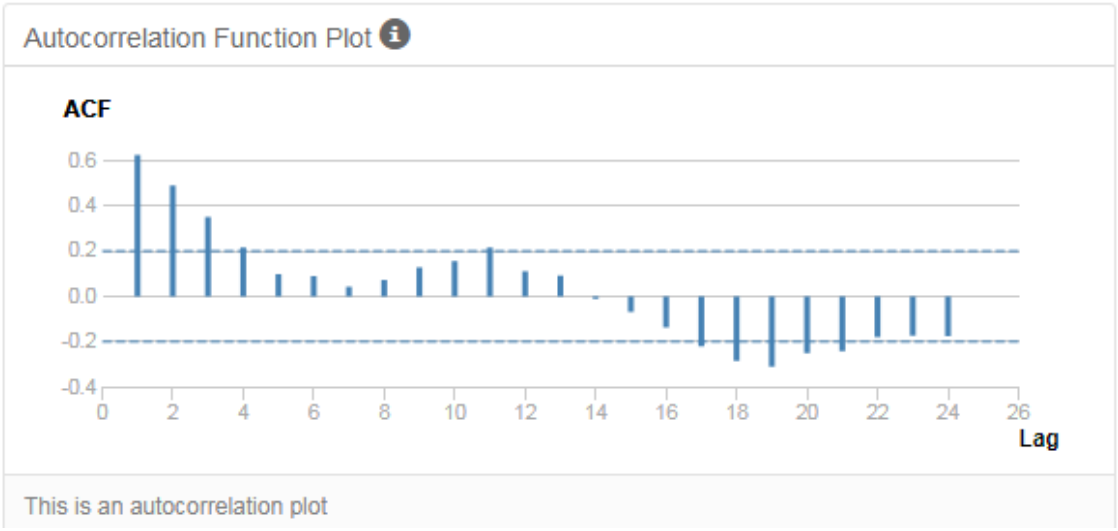
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

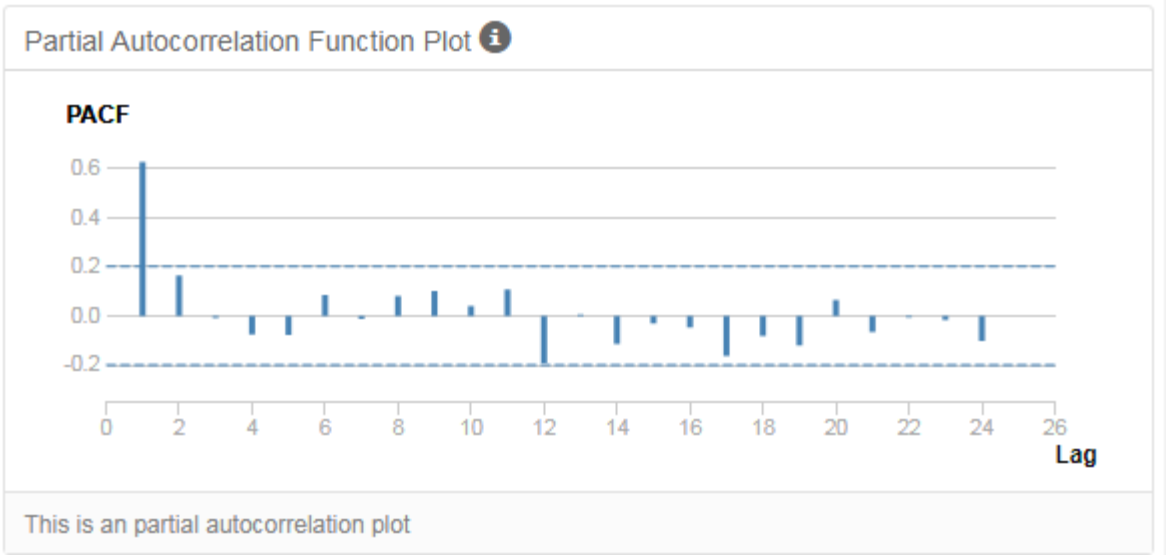
In-sample error measures:

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

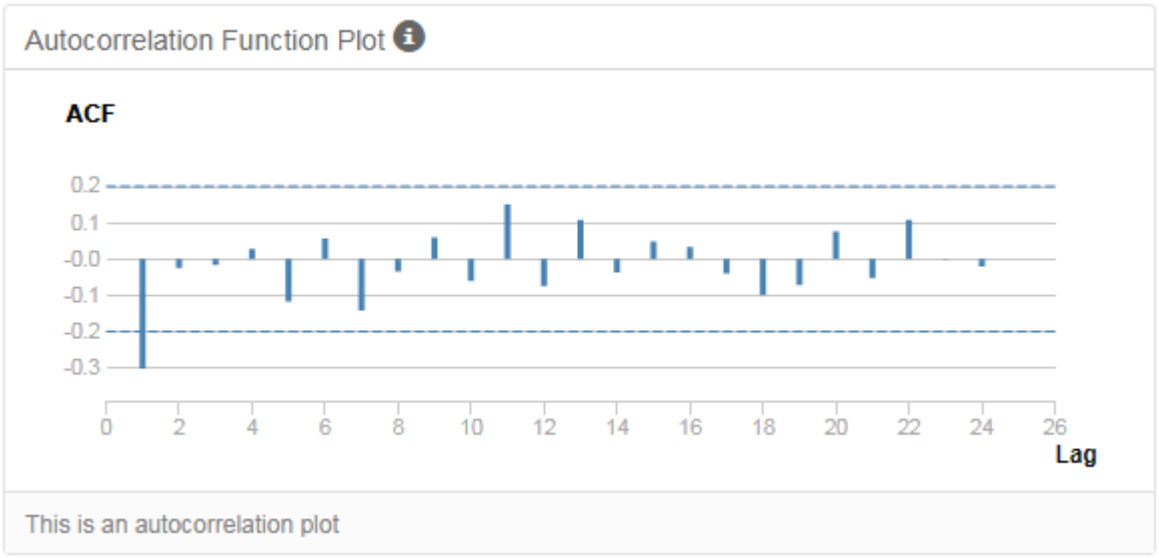
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

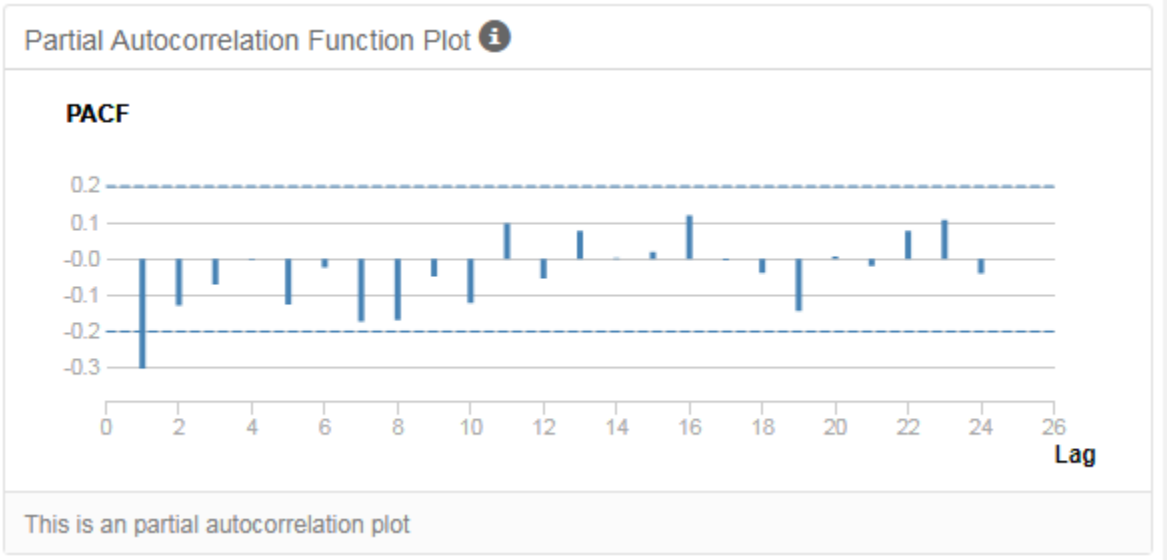
Post Seasonal Difference:





Post Time Series difference:





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)

ETS Model:

Actual and Forecast Values:

Actual	ETS_VideoSales_Dampen
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Accuracy Measures:

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

ARIMA Model:

Actual and Forecast Values:

Actual	Video_Arima
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

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

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.

I have chosen the ARIMA model as error values like MAPE and ME are lessor in ARIMA compared to ETS (Dampened). Hence this misses its forecast by a lessor amount.

Hence, as forecast error measurements are smaller in ARIMA and that is the chosen model.

In-sample errors of 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

In-sample errors of ETS (Dampened Model):

In-sample error measures:

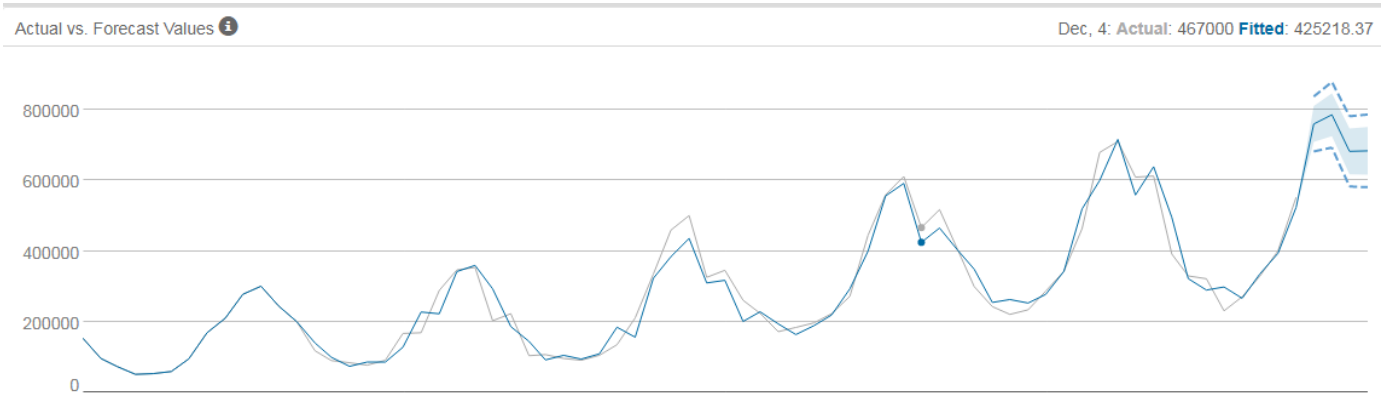
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

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

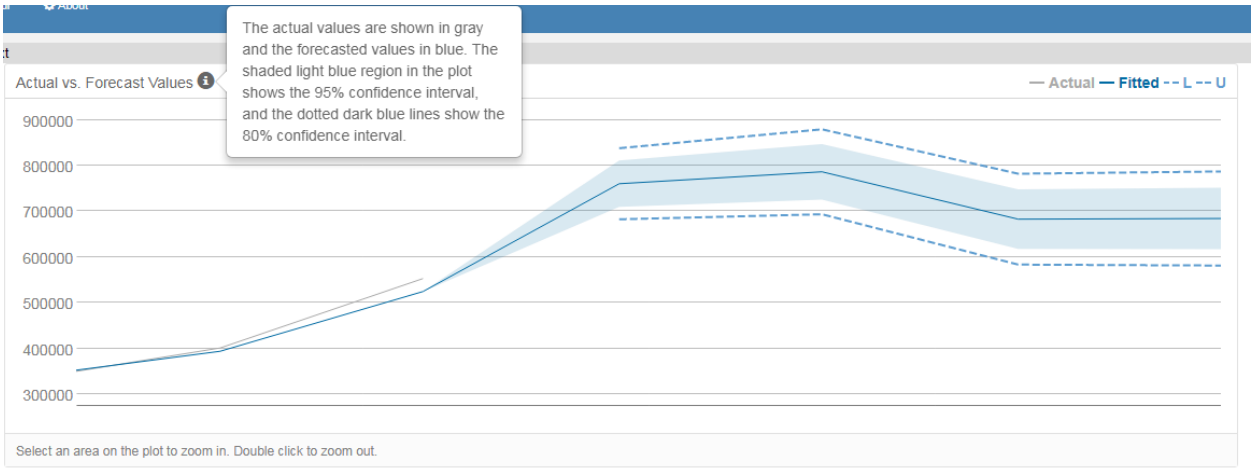
The forecast for the next four periods is a as below:

Period	Sub-period	Forecast
6	10	760617.2
6	11	786812.7
6	12	683059.1
7	1	684481.0

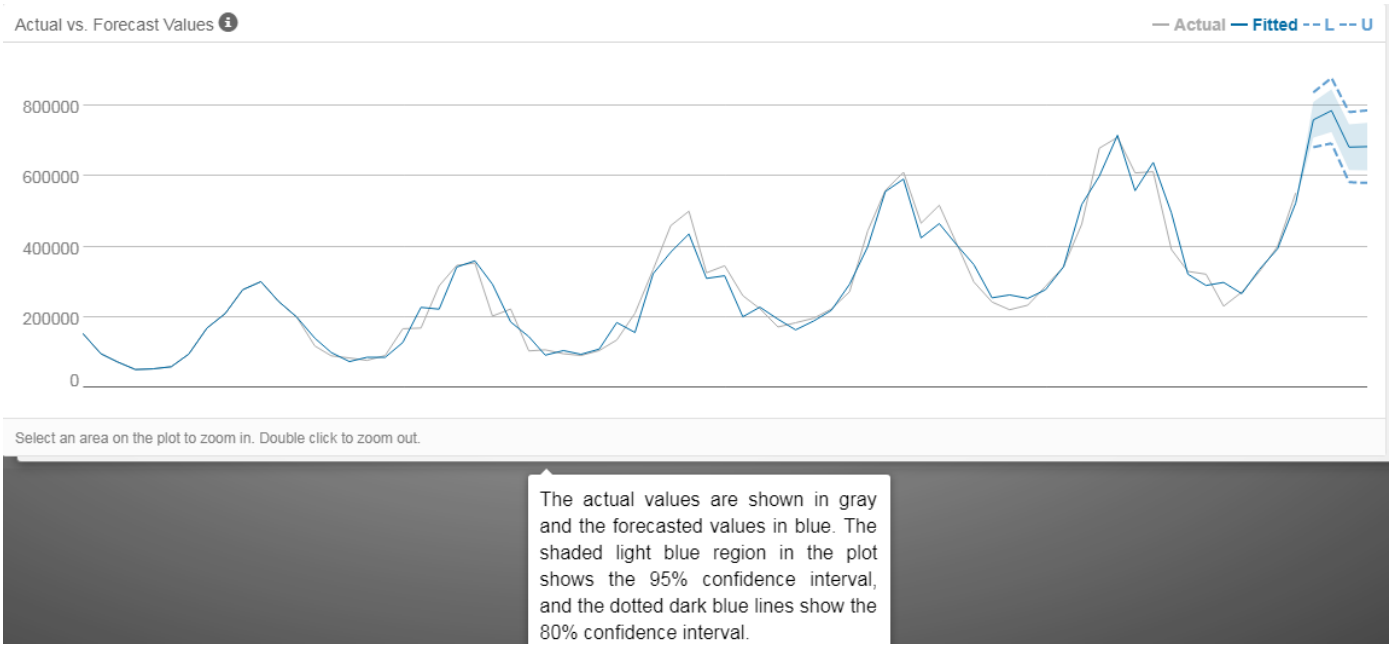
Forecasts:



Zoomed version of the forecast:

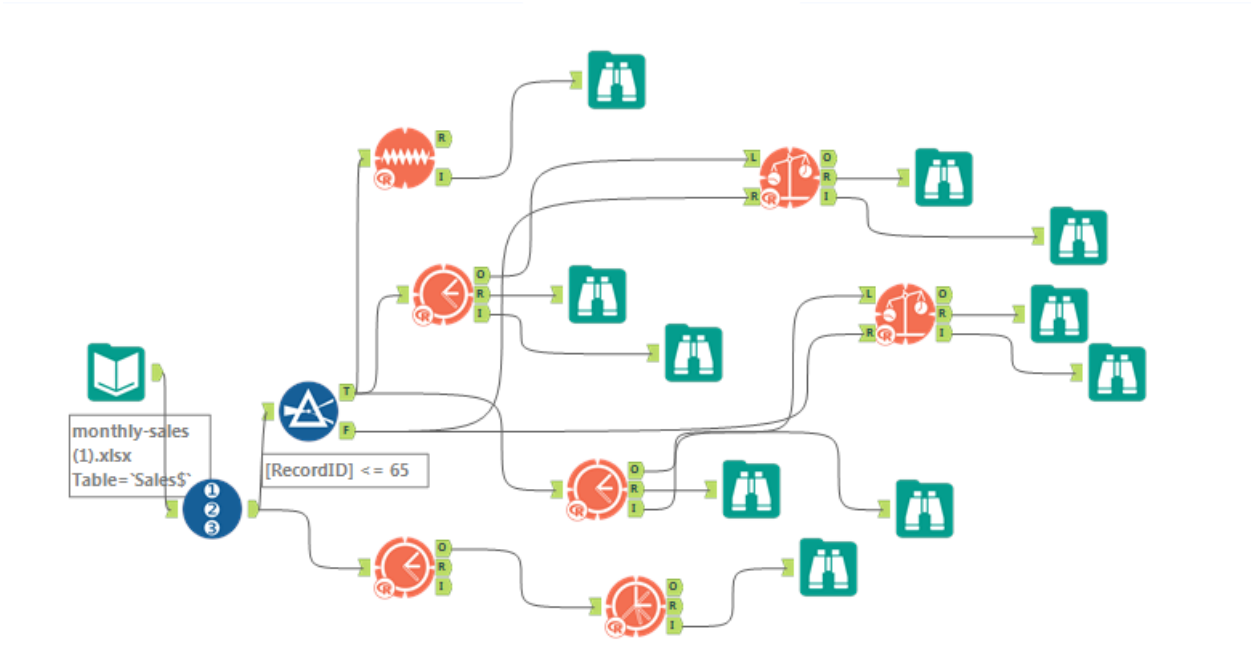


Confidence Intervals:

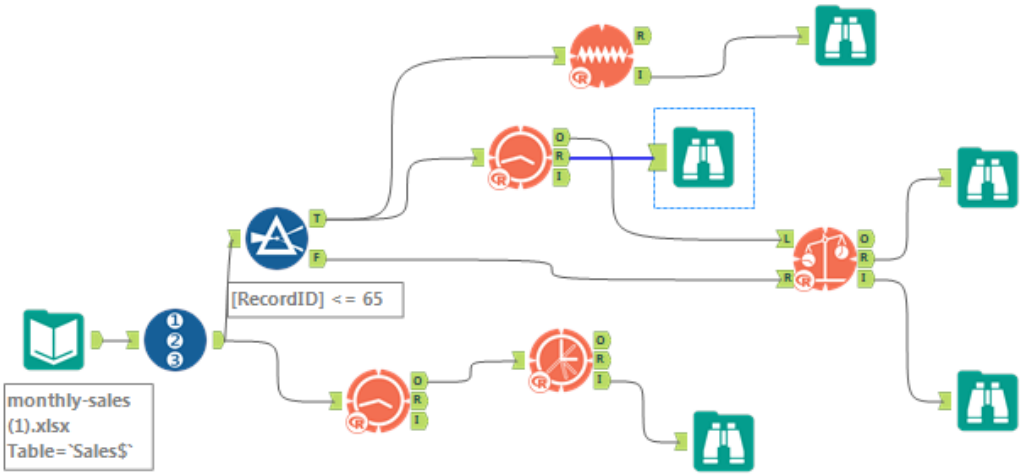


Workflow:

ETS



ARIMA



Removing seasonality: