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.

- Data covers full continues time interval: there are 5.9 years of continue data.
- Equal spacing between every 2 consecutive measures: all data is tracked monthly for all the provided data
- There is at most one data point for every time unit: there is only one record for every month in the full data

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

It has been used last 4 months for holdout to perform the forecast validation.

Step 2: Determine Trend, Seasonal, and Error components

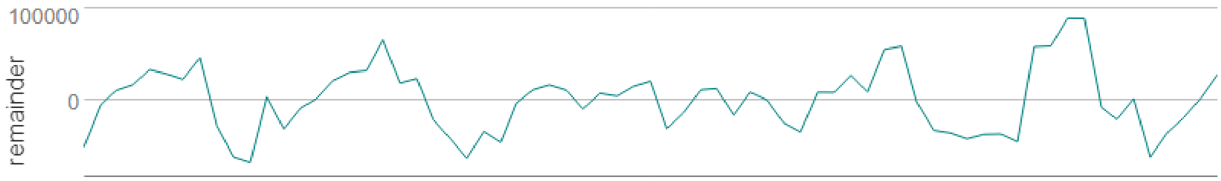
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

Answer this question:

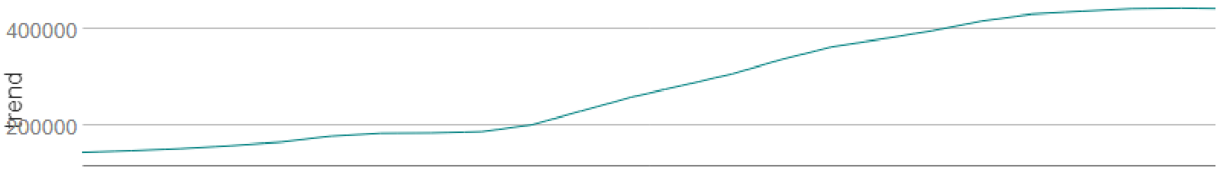
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

Using the TS Plot has been generated the next graphs for and Error, Trend and Seasonality respectively.

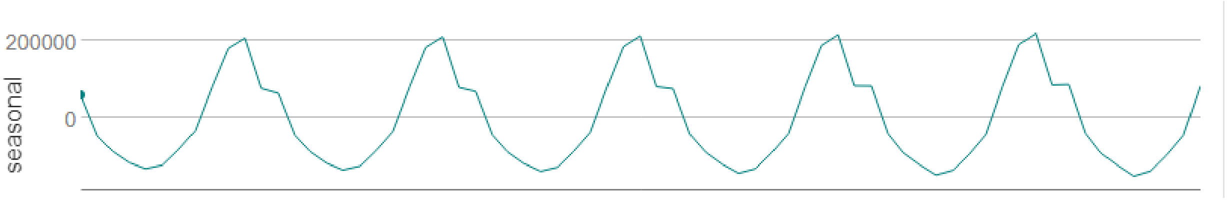
The graph for remainder or error shows a fluctuation of large and small errors over time, then has been considered as Multiplicatively (M).



The trend in the full graph can be considered as linear and therefore the Trend is Additive (A), although after the half of graph, it has been observed that trend grows up more considerably.



Per the graph, the fluctuation for every season seems to remains constant, but if observed with more detail, for last 3 seasons, there is a small increase in the peaks. So, Seasonality has been considered as Multiplicatively (M).



Step 3: Build your Models

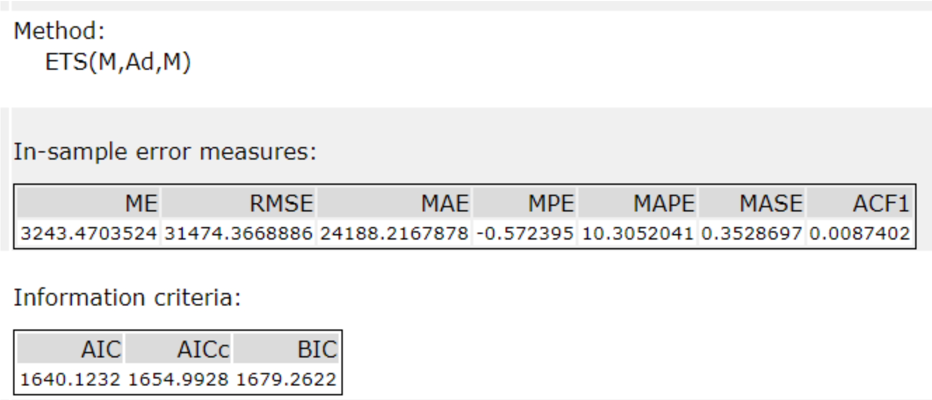
Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)

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

The components set for the ETS are (M, A, M) as result of the analysis from TS Plot graphs for Error, Trend and Seasonality.

The Internal Validation has been done using the RMSE and MASE errors, as well is being considered the AIC value:



Method:
ETS(M,Ad,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3243.4703524	31474.3668886	24188.2167878	-0.572395	10.3052041	0.3528697	0.0087402

Information criteria:

AIC	AICc	BIC
1640.1232	1654.9928	1679.2622

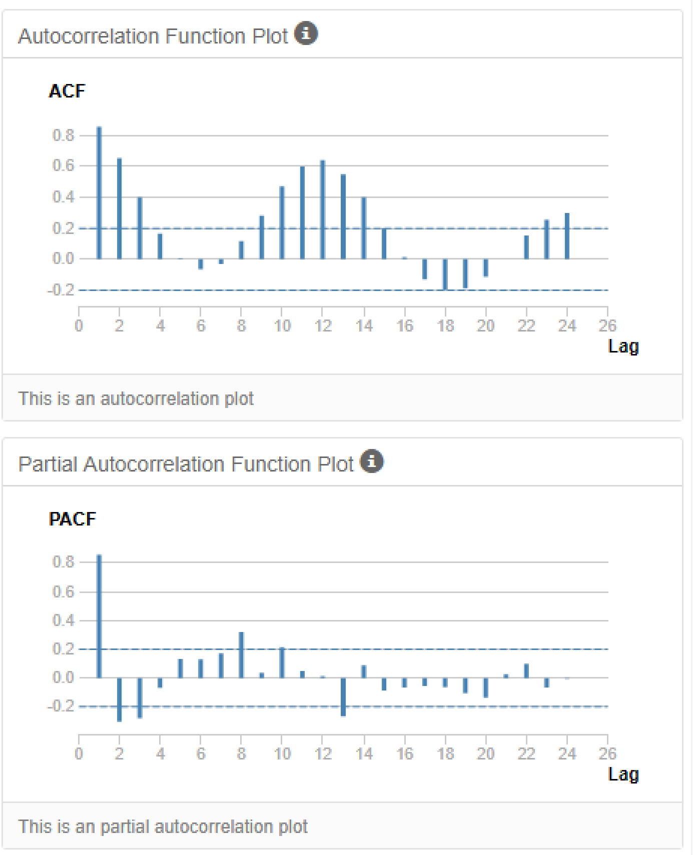
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
   a. Describe the in-sample errors.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

After analysis of graphs from the TS Plot has been observed that there is a seasonality. And therefore, has been chosen to configure a Seasonal ARIMA Model.

For the ARIMA time series model has been chosen the next terms ARIMA (p,d,q) (P,D,Q) m considering the graphs for ACF and PACF from the TS Plots using the original data and calculating the seasonal differences. After this analysis, has determined to use the next values for the ARIMA MODEL.

- p = 0
- d = 1
- q = 1
- P = 0
- D = 1
- Q = 0
- m = 12

By the initial ACF and PACF graphs, can be observed that lag 1 is positive and slowly decrease for lags 2,3,4 and 5. And increase again at seasonal lags in 12 and 24. This can be considered as highly correlation and therefore has been required to calculate the seasonal difference and generate ACF and PACF graphs.



Autocorrelation Function Plot

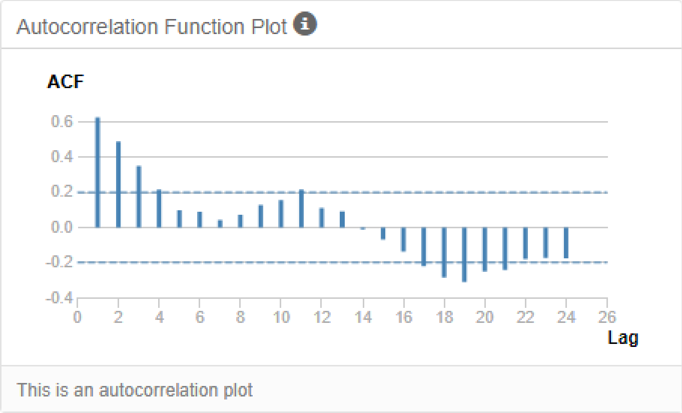
This is an autocorrelation plot

Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

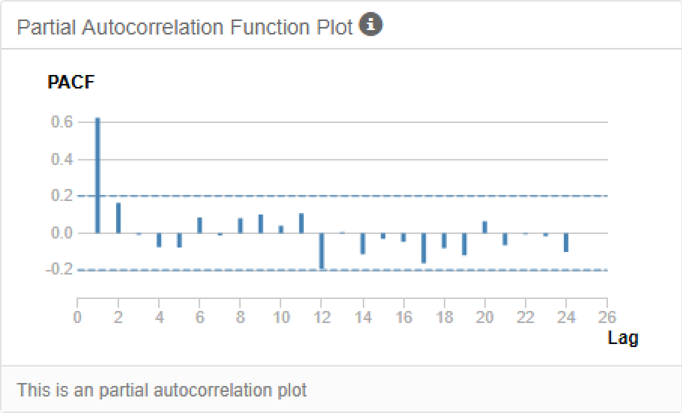
Seasonal Difference for ACF and PACF

After applying the Season Difference it is observed that ACF and PACF graphs have reduced the value for lag 1 and decreased slowly near to 0 at lag 4. But this is still an indicator or high correlation, and is required to calculate first seasonal difference and plot again the ACF and PACF graphs to determine better the values for the ARIMA termns.



Autocorrelation Function Plot

This is an autocorrelation plot



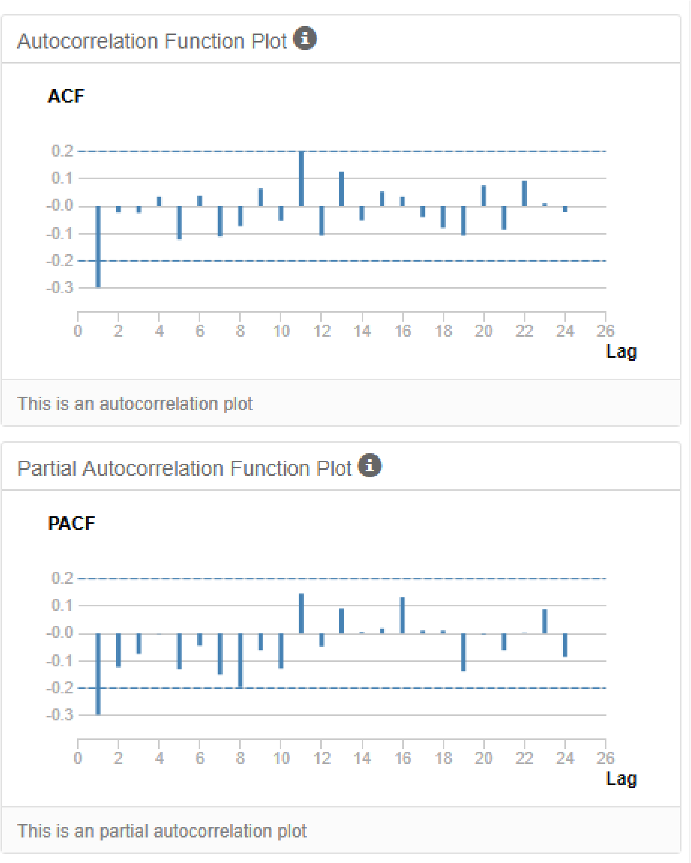
Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

First Seasonal Difference for ACF and PACF

According to the ACF and PACF graphs below can be observed that the seasonal correlation has been eliminated and there is no required to perform any aditional difference. Therefore d and D has been set equal to 1.

Also the graphs for ACF and PACF have a negative correlation at lag 1, and therefore MA has been set to 1.


Autocorrelation Function Plot

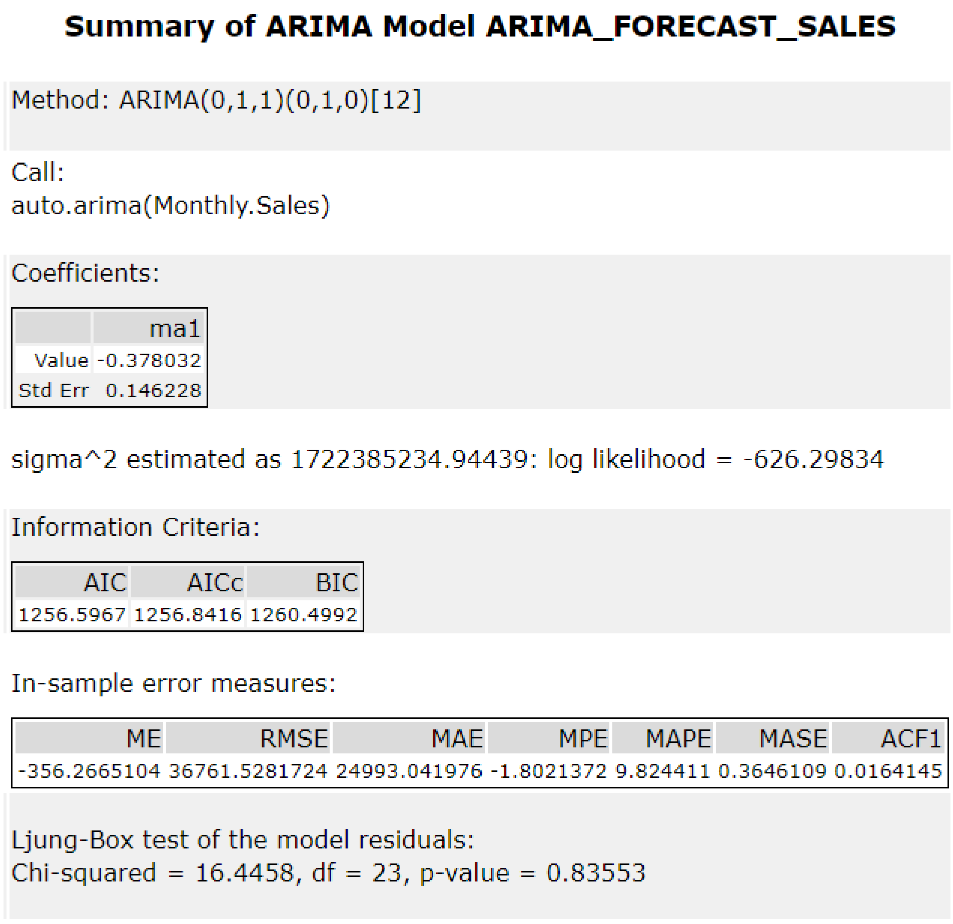
This is an autocorrelation plot

Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

Internal validation considering the RMSE, MASE

- RMSE: 36,761, it is near to the median value.
- MASE: 0.36, since it is lower than 1, it is a good indicator of a good level of accuracy.



Summary of ARIMA Model ARIMA_FORECAST_SALES

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

Ljung-Box test of the model residuals:
Chi-squared = 16.4458, df = 23, p-value = 0.83553

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)

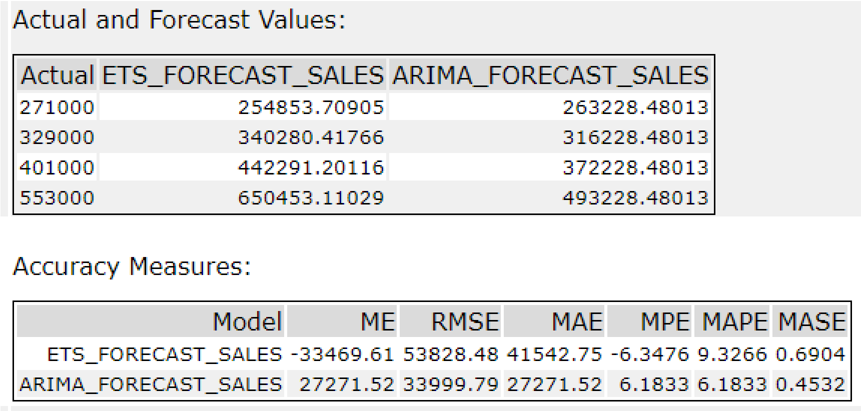
Answer these questions.

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.

The External Validation using the hold out sample with ETS and ARIMA models

It can be observed that the values for RMSE and MASE are lower for the ARIMA model, which indicates is performing better at making the predictions again the hold out sample. Also, the MASE value can be considered as a good indicator of the accuracy for the forecast prediction, because is lower than 1.

Comparison of Time Series Models

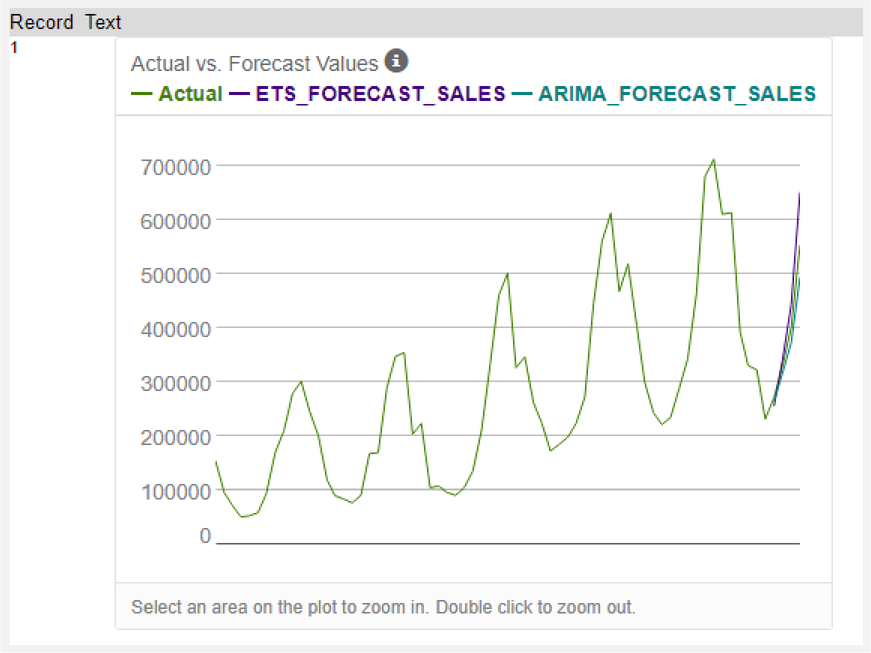


Actual and Forecast Values:

Actual	ETS_FORECAST_SALES	ARIMA_FORECAST_SALES
271000	254853.70905	263228.48013
329000	340280.41766	316228.48013
401000	442291.20116	372228.48013
553000	650453.11029	493228.48013

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_FORECAST_SALES	-33469.61	53828.48	41542.75	-6.3476	9.3266	0.6904
ARIMA_FORECAST_SALES	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

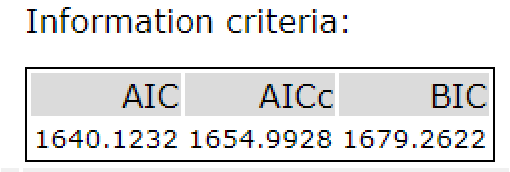


Record	Text
1	Actual vs. Forecast Values

Select an area on the plot to zoom in. Double click to zoom out.

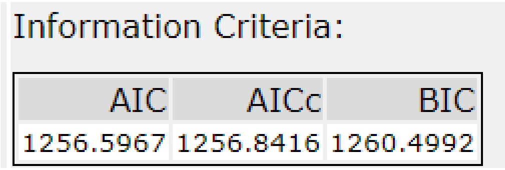
AIC values for ETS and ARIMA

Also, has been compared the AIC values for the ETS and ARIMA models respectively, for which can be confirmed that ARIMA model is performing better because it has a lower value.



Information criteria:

AIC	AICc	BIC
1640.1232	1654.9928	1679.2622



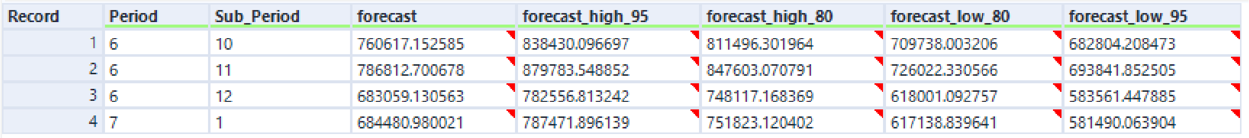
Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

Therefore, has been selected the ARIMA model as the best option, because it has lower RMSE, MASE and AIC values as well it has generated more accurate values for the test forecast.

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

Using the ARIMA model with the 69 periods of data the result for the prediction of the next f4 months are:



Record	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	6	10	760617.152585	838430.096697	811496.301964	709738.003206	682804.208473
2	6	11	786812.700678	879783.548852	847603.070791	726022.330566	693841.852505
3	6	12	683059.130563	782556.813242	748117.168369	618001.092757	583561.447885
4	7	1	684480.980021	787471.896139	751823.120402	617138.839641	581490.063904

Last year of data including 4 months forecast:

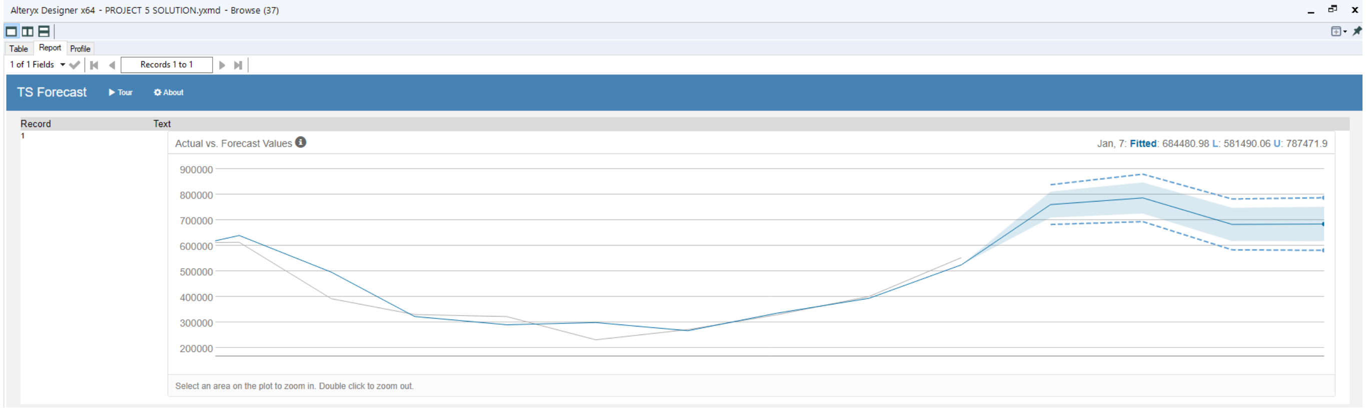
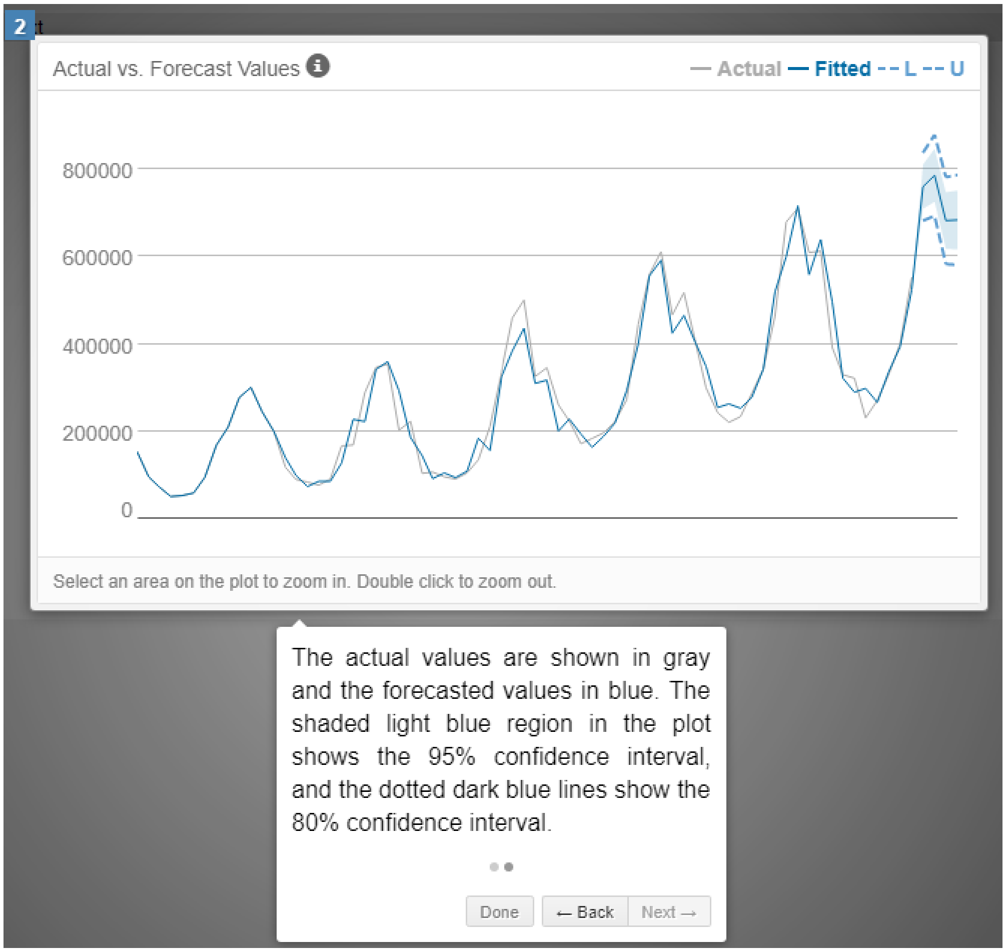

Alteryx Designer x64 - PROJECT 5 SOLUTION.yxmd - Browse (37)

Table
Report
Profile
1 of 1 Fields
Records 1 to 1
TS Forecast
Tour
About
Record
Text
1
Actual vs. Forecast Values
Jan, 7: Fitted: 684480.98 L: 581490.06 U: 787471.9
900000
800000
700000
600000
500000
400000
300000
200000
Select an area on the plot to zoom in. Double click to zoom out.

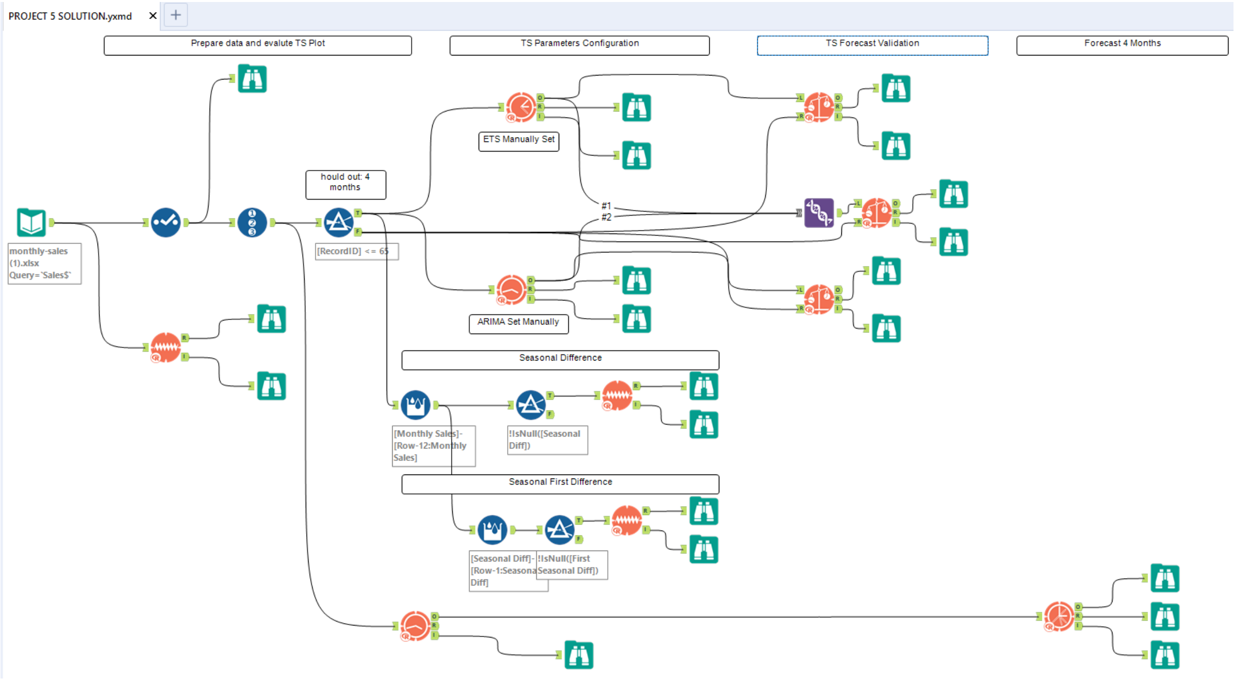
All periods including 4 months forecast:



Select an area on the plot to zoom in. Double click to zoom out.

The actual values are shown in gray and the forecasted values in blue. The shaded light blue region in the plot shows the 95% confidence interval, and the dotted dark blue lines show the 80% confidence interval.

Alteryx Workflow


PROJECT 5 SOLUTION.yxmd
Prepare data and evalute TS Plot
TS Parameters Configuration
TS Forecast Validation
Forecast 4 Months
hould out: 4 months
ETS Manually Set
ARIMA Set Manually
monthly-sales (1).xlsx Query=`Sales$`
[RecordID] <= 65
#1
#2
Seasonal Difference
[Monthly Sales]-[Row-12:Monthly Sales]
!IsNull([Seasonal Diff])
Seasonal First Difference
[Seasonal Diff]-[Row-1:Seasonal Diff]
!IsNull([First Seasonal Diff])

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.