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

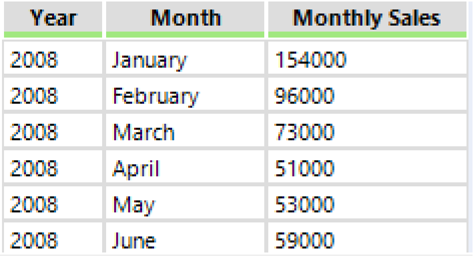
Step 1: Plan Your Analysis

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: Firstly, the year and month were separated in order to make the dataset more readable

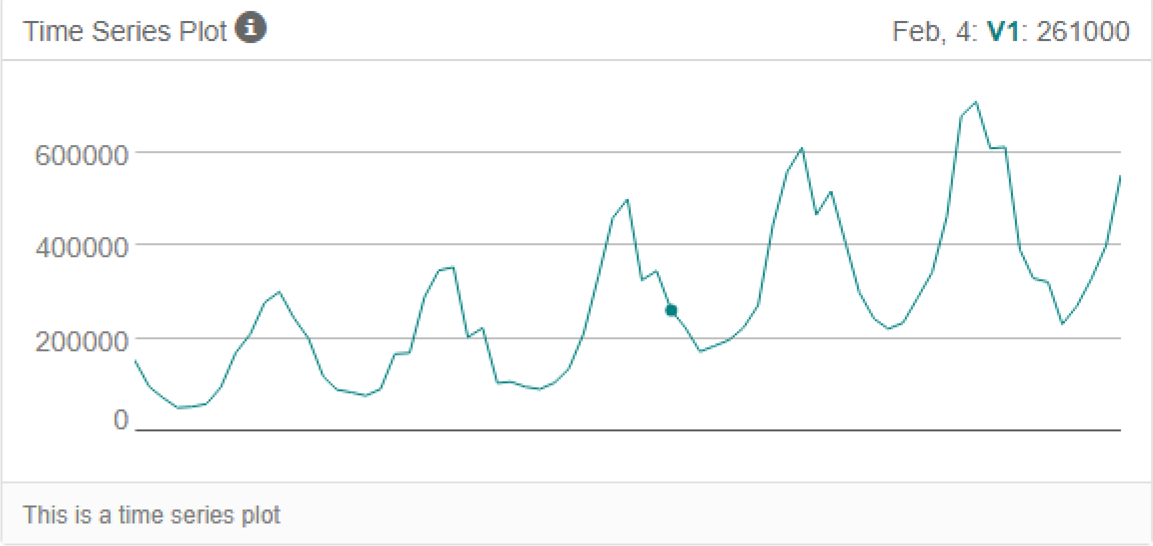


Year	Month	Monthly Sales
2008	January	154000
2008	February	96000
2008	March	73000
2008	April	51000
2008	May	53000
2008	June	59000

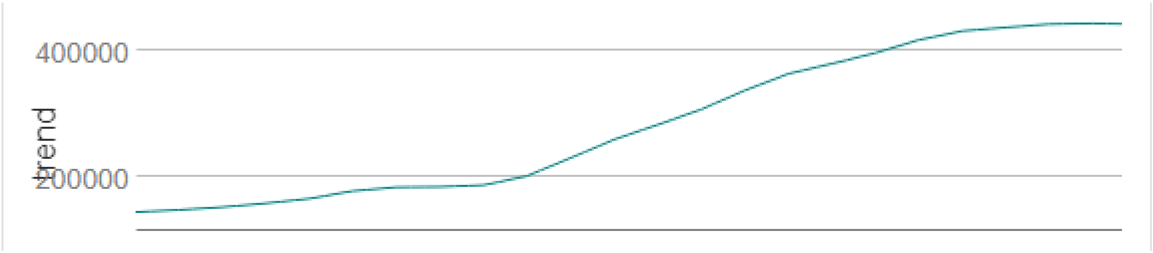
Figure 1: Dataset Cleaned Sample

The key characteristics of a Time Series Dataset and its relation with this dataset are described with figured below:

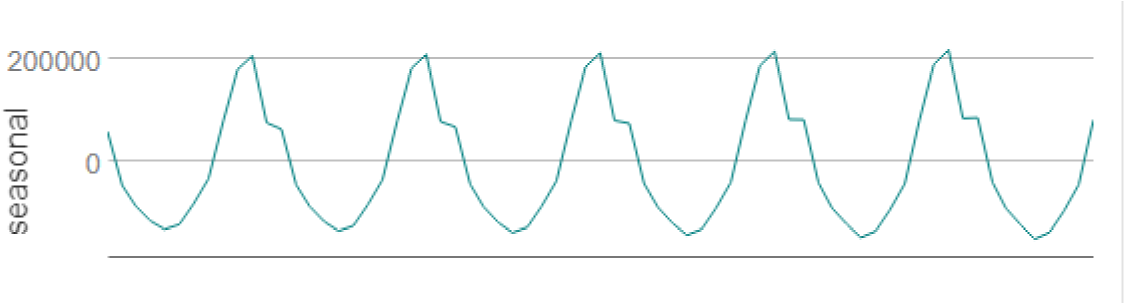
- This dataset is over a continuous time period, thus it is a time series



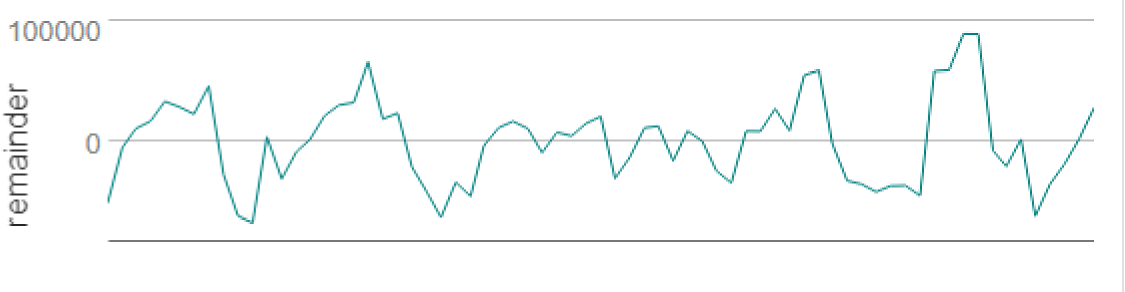
- Sequential measurement over this time period reveals a regular direction of the data points, which establishes a linear trend



- Each time period unit in the dataset has only one data point.
- These peaks and valleys are regular in nature and display period in a seasonal manner, hence seasonality exist



- The dataset also exhibits remainder after calculating Trend and Seasonality which represents error.



- There is equal spacing between every two consecutive measurement.

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

Ans: Last four periods of the dataset should be used as a holdout sample. The period of holdout sample is 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.

Ans: From Figure 2 we can see, the trend increasing in a linear fashion so we need to use Trend additively in this analysis. In Seasonality plot we can see, seasonality increases in volume in each seasonal period suggesting applying seasonality in a multiplicative manner.

Lastly, the remainder or error changing variance as the time series moves along. It means we should use error multiplicatively.

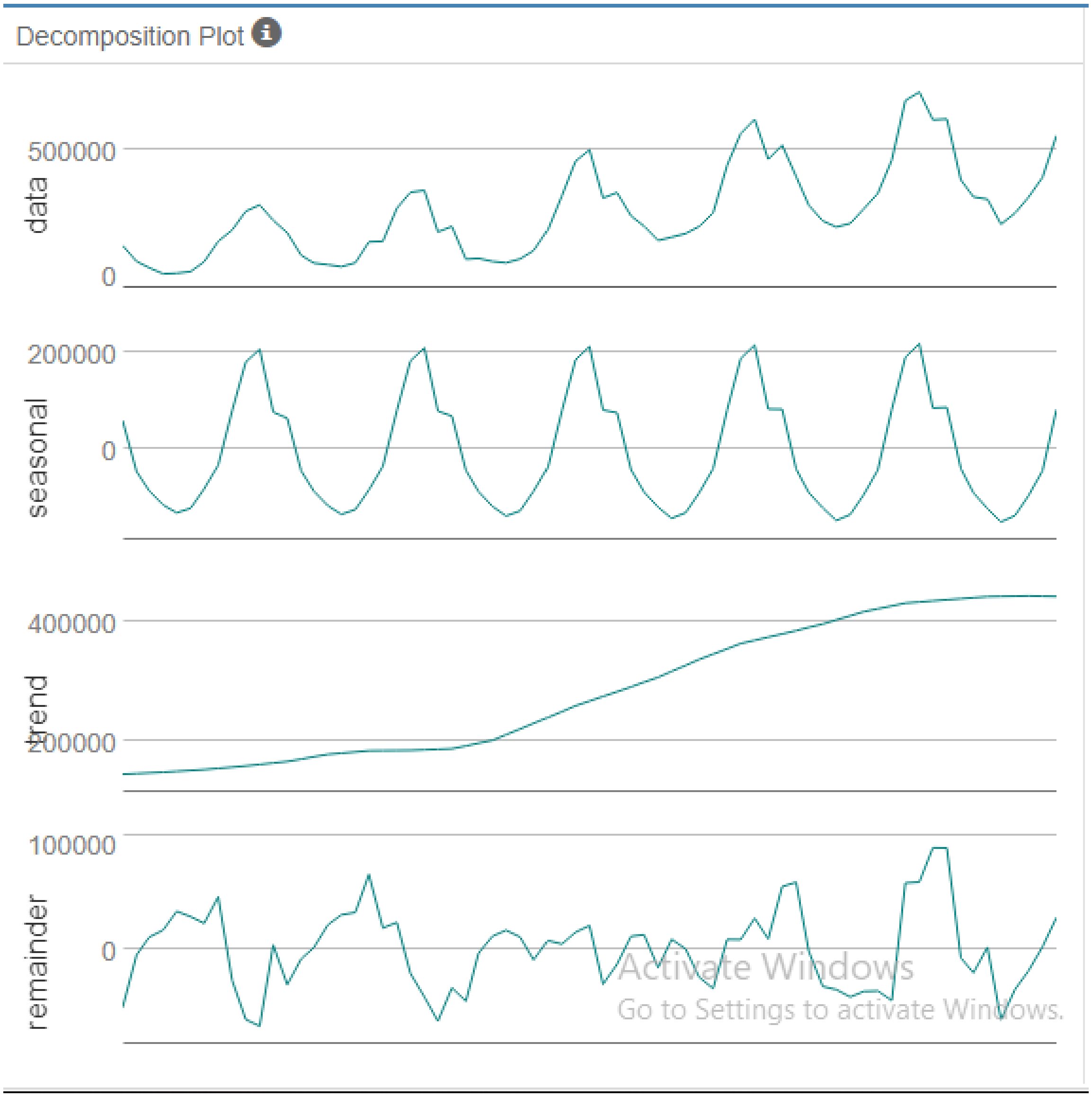


Figure 2: Decomposition Plot

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.

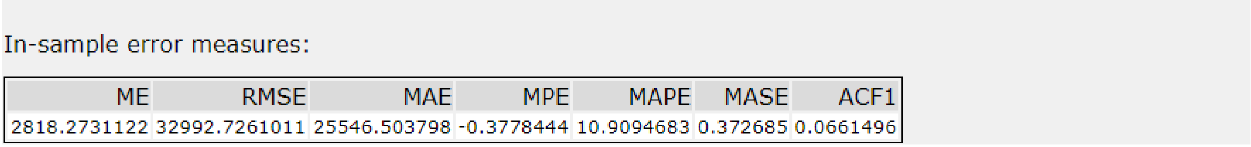
a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

Ans:

Error: As seen in the Decomposition Plot above (Figure:2), remainder changing variance as the time series moves along. That is why it will be Multiplicative term.

Term: Additive because the trend is increasing in a linear fashion as seen in the decomposition plot.

Seasonality: Multiplicative because increases in volume in each seasonal period and varying in every year.



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

Figure 3: In Sample Error Measure (ETS Model)

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.

Ans: In order to determine ARIMA model terms, we first need to determine the ACF and PACF of time series plot.

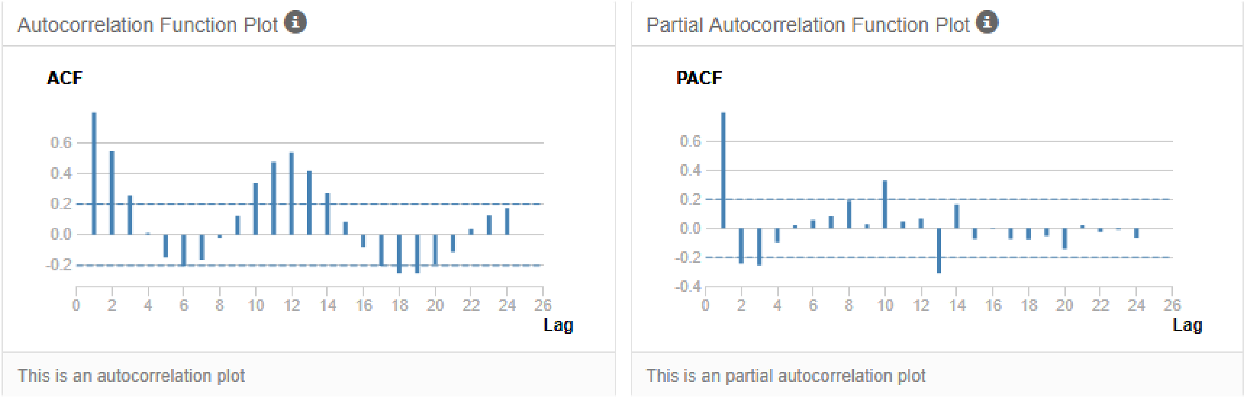
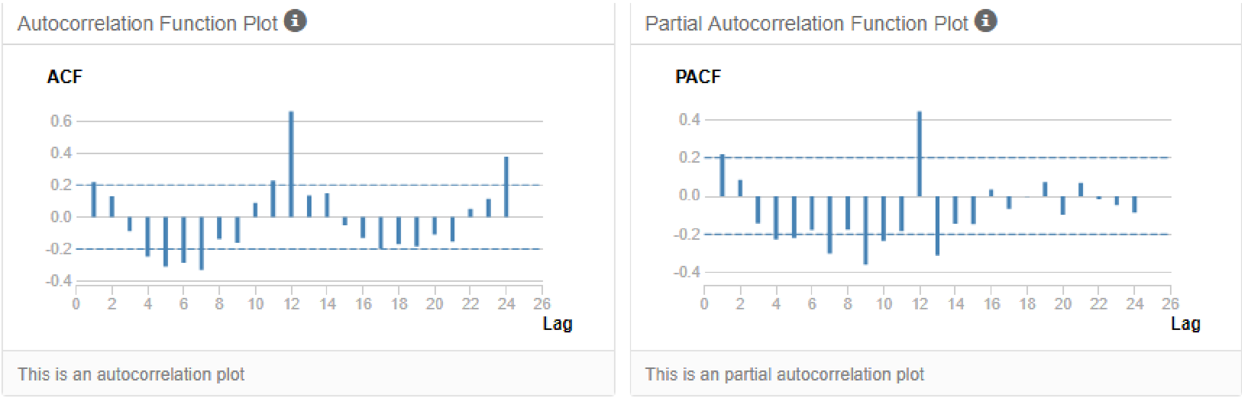


Figure 4: Time Series ACF and PACF

From figure 4 it is observed that, ACF is slowly is slowly decreasing towards 0 with seasonal decreases towards lags. This shows serial correlation. Hence, we need to difference the series.

Now we determine the ACF and PACF seasonal plot


Autocorrelation Function Plot

This is an autocorrelation plot

Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

Figure 5: Seasonal ACF and PACF

From figure 5 we can determine, the plot is very much similar to the previous one however the correlation is lesser and first seasonal difference has been used here. First Seasonal Difference ACF and PACF now has been performed.

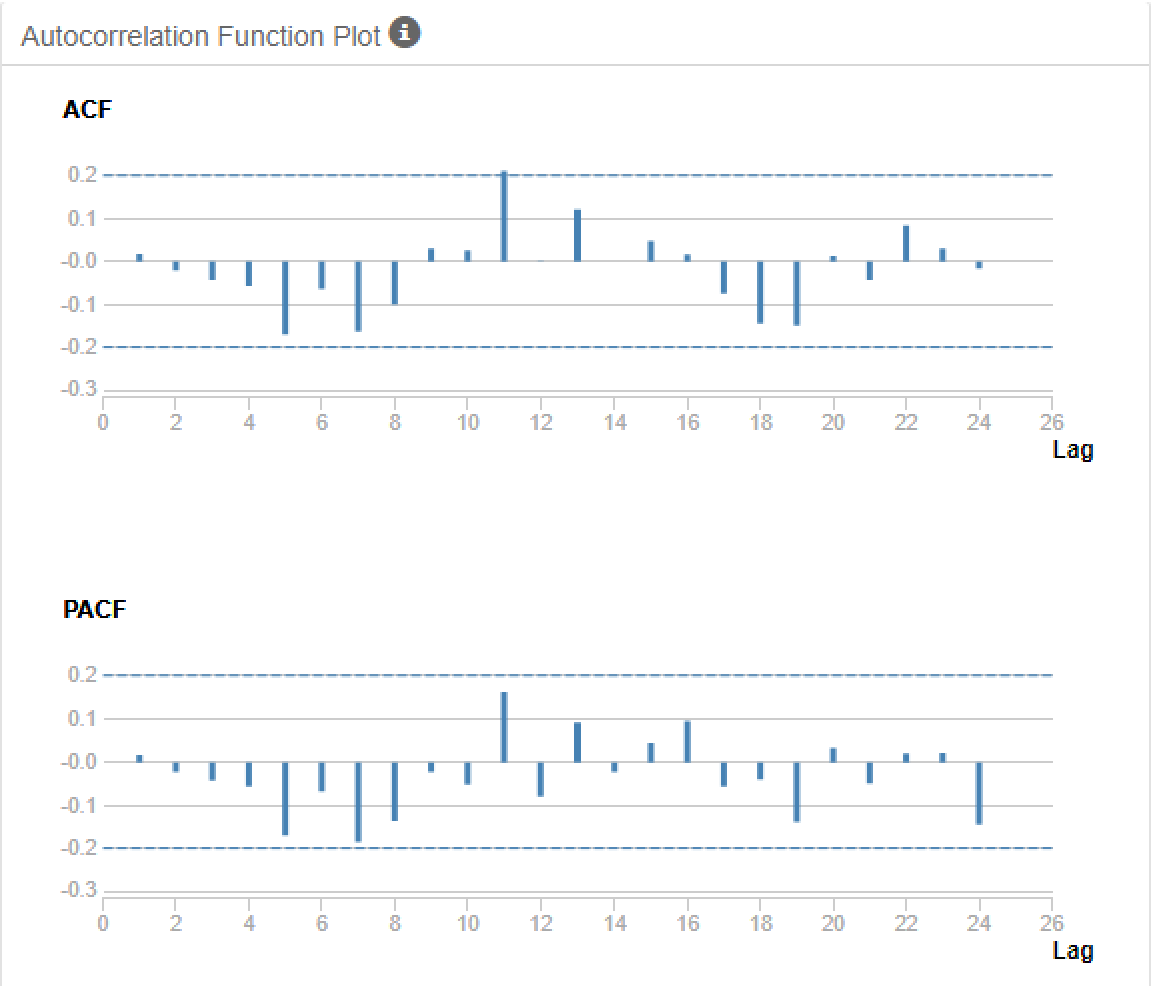

Autocorrelation Function Plot

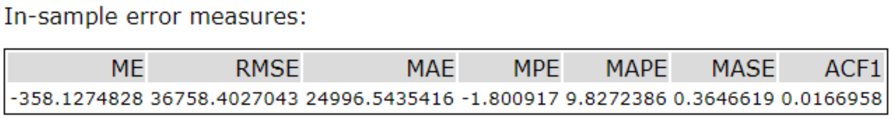
Figure 6: ARIMA model ACF and PACF

All the significant correlation has been removed. Any remaining correlation would be taken care of by the AR and MA terms.

p = 0, q = 1 and d = 1,
P, D, Q = 0,1,0
m= 12 as lag repeats after 12 months

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

Ans:



In-sample error measures:

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

Figure 7: Errors of ARIMA model

MASE is well below 1 which indicates a strong model. Other depends on the comparison with different model.

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

Ans:

After establishing the correct ARIMA model, ACF and PACF is re-graphed. The ACF and PACF results for the correct ARIMA model shows no significantly correlated lags suggesting no need for adding additional AR or MA terms.

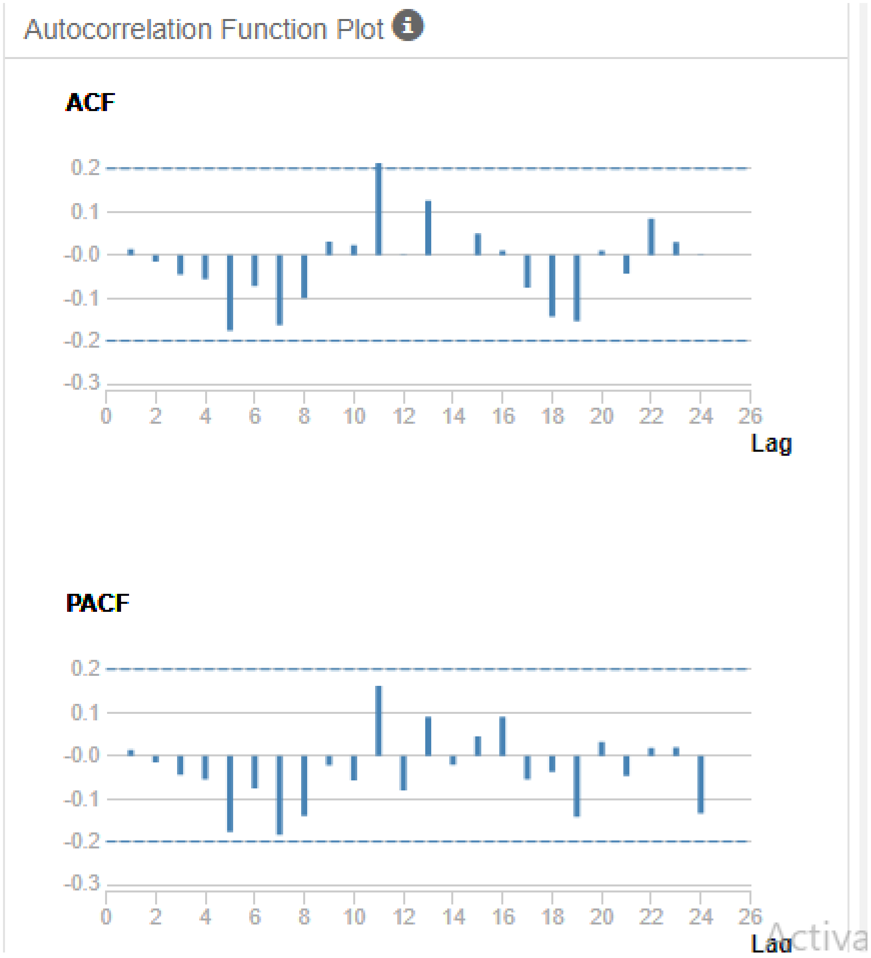


Figure 8: Final ACF-PACF Plot (After ARIMA model)

Step 4: Forecast

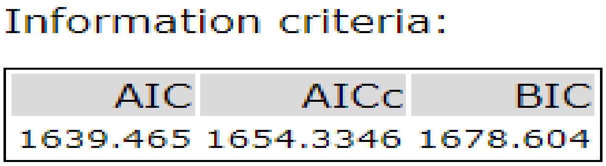
Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)

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: A holdout sample of 4 months was created to test the models, as the prediction of sales is required for 4 months. Both the ETS and ARIMA models are tested using following different criteria.

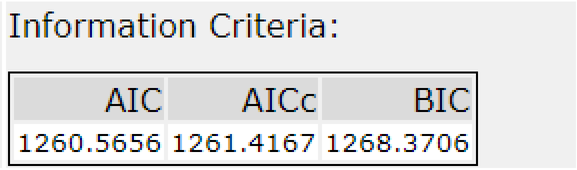
AIC



Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

Figure 9: Information Criteria ETS Model

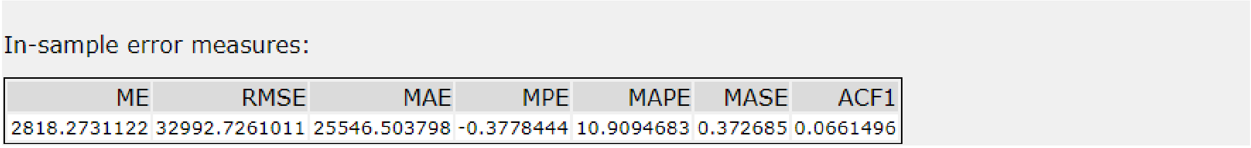


Information Criteria:

AIC	AICc	BIC
1260.5656	1261.4167	1268.3706

Figure 10: Information Criteria ARIMA model

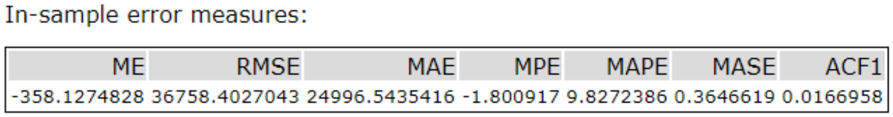
From figure 9 and 10 we can see the AIC values of both ETS model and ARIMA model. Lower AIC value of ARIMA model indicating a better fit model.



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

Figure 11: In sample error ETS model



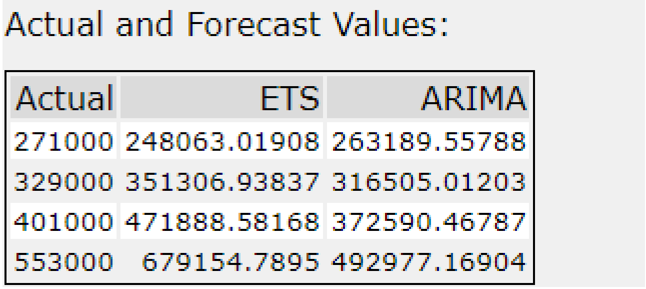
In-sample error measures:

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

Figure 12: In sample error ARIMA model

From figure 11 and 12 we can see MASE of ARIMA model is lower than the ETS model making it more efficient and better fit. Besides, ARIMA model also have lower ME and MAE.

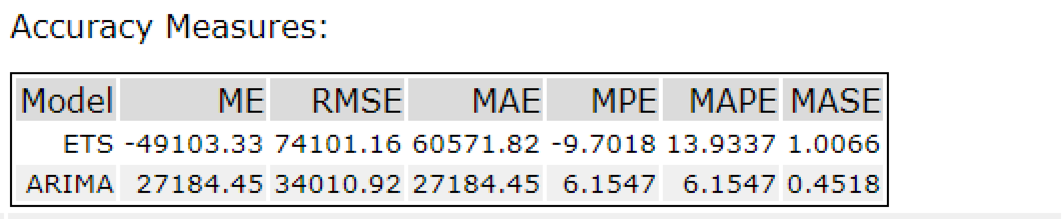
Forecast with holdout sample



Actual and Forecast Values:

Actual	ETS	ARIMA
271000	248063.01908	263189.55788
329000	351306.93837	316505.01203
401000	471888.58168	372590.46787
553000	679154.7895	492977.16904

Figure 13: Actual and Forecast Values



Accuracy Measures:

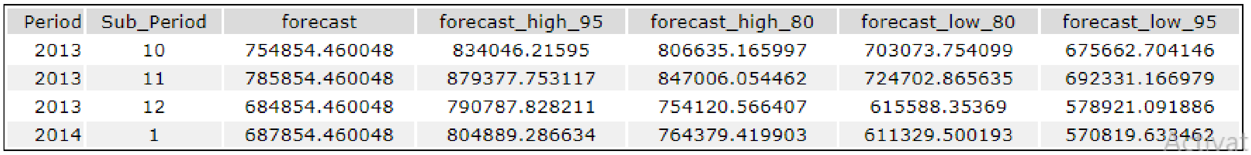
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066
ARIMA	27184.45	34010.92	27184.45	6.1547	6.1547	0.4518

Figure 14: Accuracy measures comparison between ETS and ARIMA model

It is clear that ARIMA has lower MASE value and is closer to the actual figures while compared against the holdout sample and has lower AIC than ETS, Hence ARIMA model is chosen to forecast the result.

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

Ans:



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	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

Figure 15: Forecast Dataset

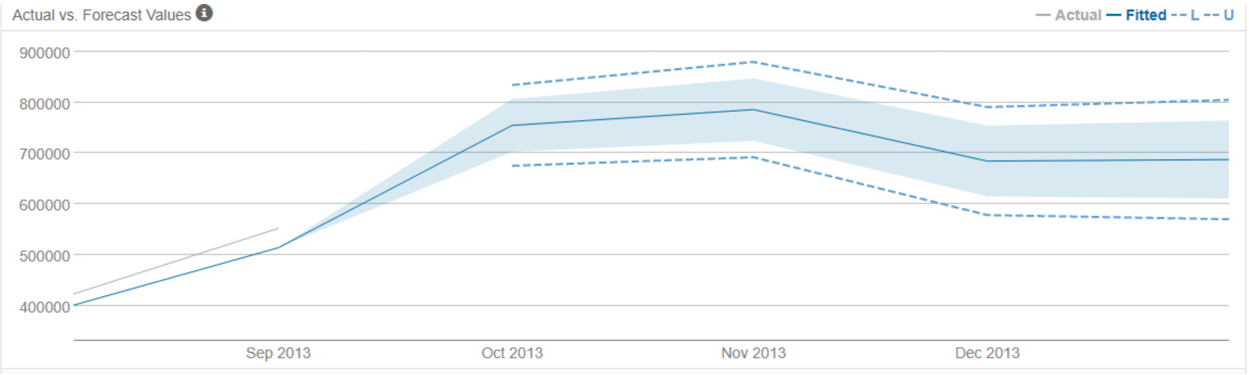


Figure 16: Forecast Plot

Figure 15 and 16 indicating the forecast of sales of the last four month of the dataset. In figure 15, the figures of different confidence interval are shown and in figure 16 the blue lie indicating the forecasted sales.

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.