Project: Forecasting Sales

Step 1: Plan Your Analysis

Answer the following questions to help you plan out your analysis:

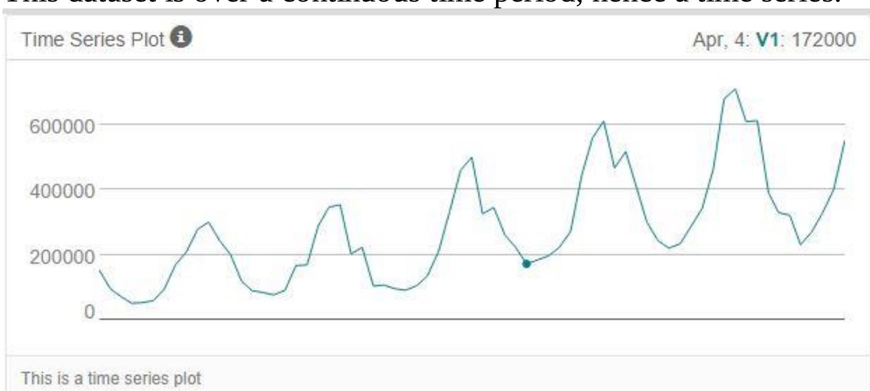
1. Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.

   Firstly, the dataset is cleaned and the years and months are separated to provide better distinction (sample below).

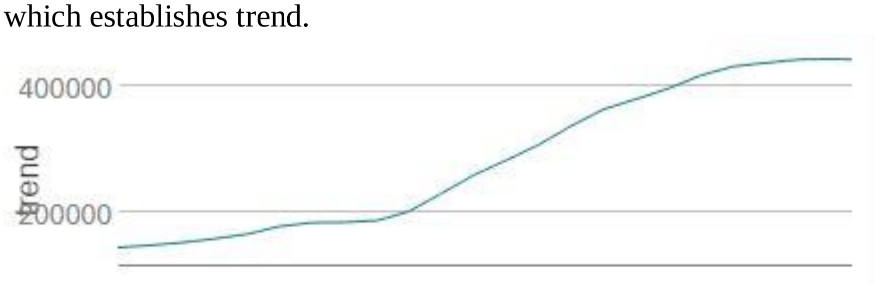
Record #	Year	Month	Monthly Sales
1	2008	January	154000
2	2008	February	96000
3	2008	March	73000
4	2008	April	51000

   The key characteristics of a time series dataset and its relation to this particular dataset are as follows:

   - This dataset is over a continuous time period, hence a time series.

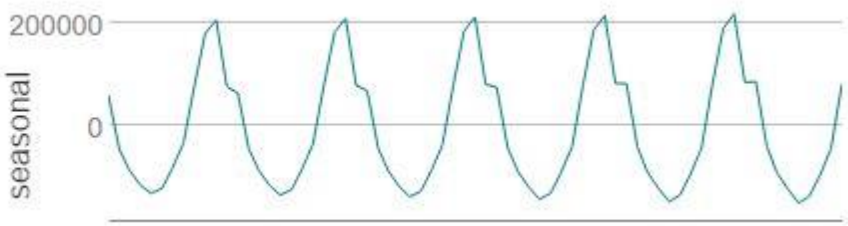


   - Sequential measurement over that time period reveals a regular directionality of the data points, which establishes trend.



   - Each time unit in the time period has only one data point.

- These patterns are regular in nature and display periodicity in a seasonal manner, hence seasonality.



- The dataset also exhibits remainder after accounting for trend and seasonality which represents error.

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

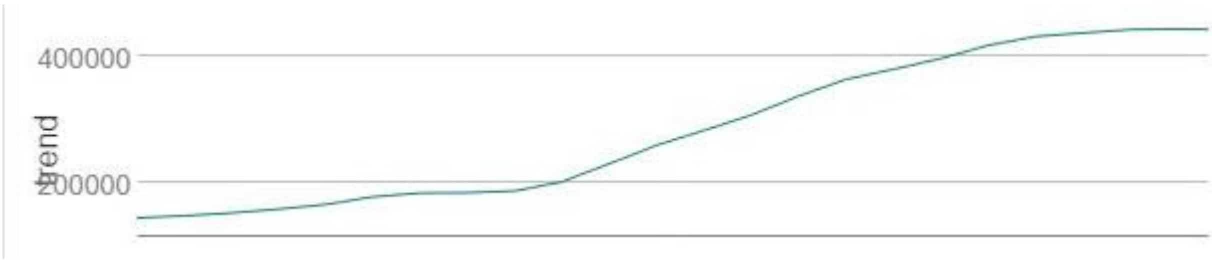
   The last 4 periods i.e. months are used as holdout samples as four months or period forecast is required.

Step 2: Determine Trend, Seasonal, and Error components

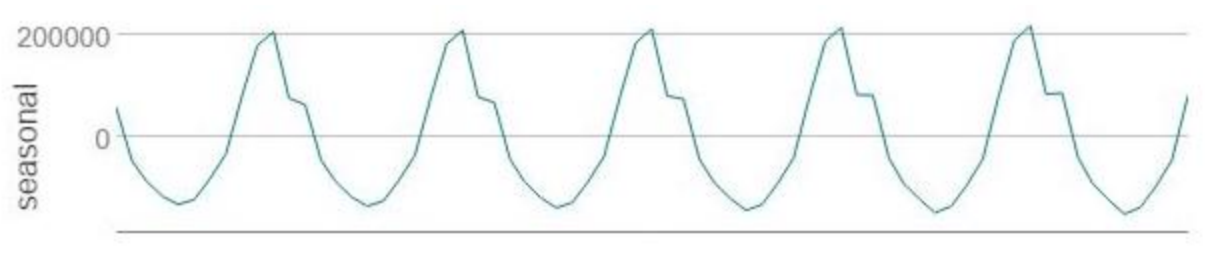
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

   A decomposition plot is determined to help us dive deep into the dataset. The data overall gives a picture of a seasonally increasing time series plot as determined from the decomposition plot.

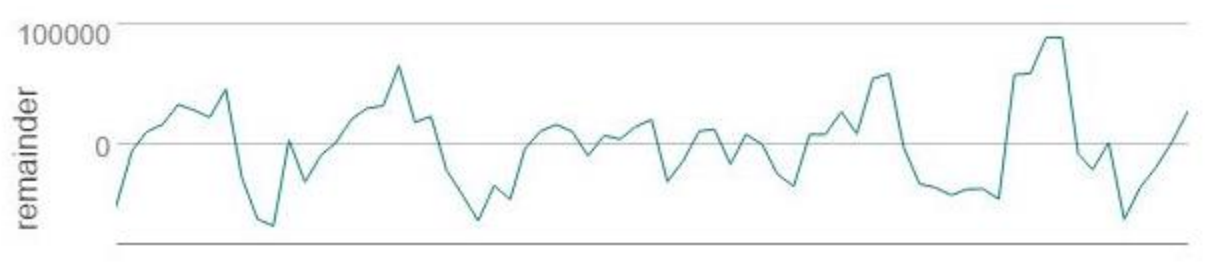
The dataset exhibits an uptrend or a steadily increasing direction of the dataset with the passage of time.



The dataset further exhibits a seasonal pattern establishing seasonality.



The remainder has peaks with varying altitudes. which indicate that the error factors are not uniform throughout the dataset.



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.
   Error: Multiplicative because the variation in the remainder shows a change in variance over time as seen in the decomposition plot above.
   Trend: Additive since the trend line is linear as seen in the decomposition plot above.
   Seasonality: Multiplicative as since the remainder is varying in magnitude every year.

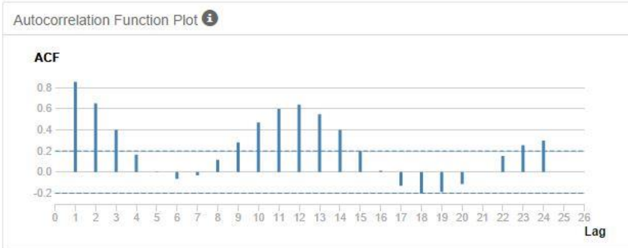
   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

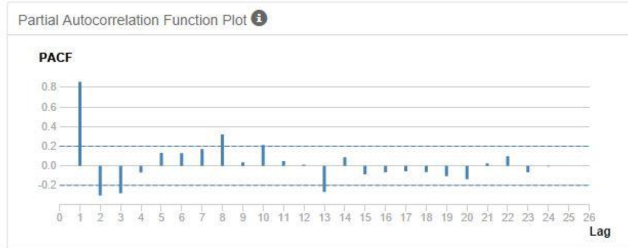
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056	0.1436491

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.

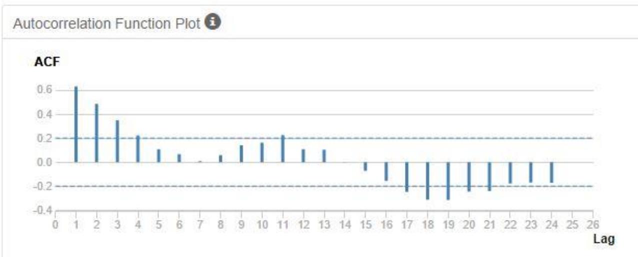
   To determine model terms, we will first determine time series ACF and PACF.

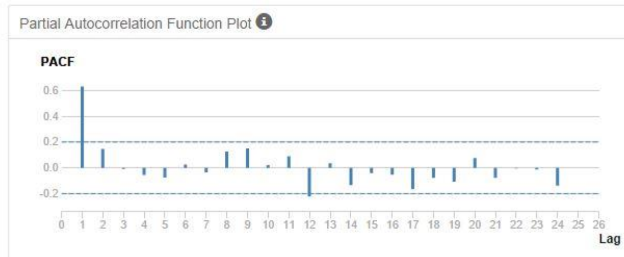




It is observed from the graph that ACF is slowly decreasing towards 0 with seasonal increases at the lags. This shows serial correlation. Hence, we need to difference the series.

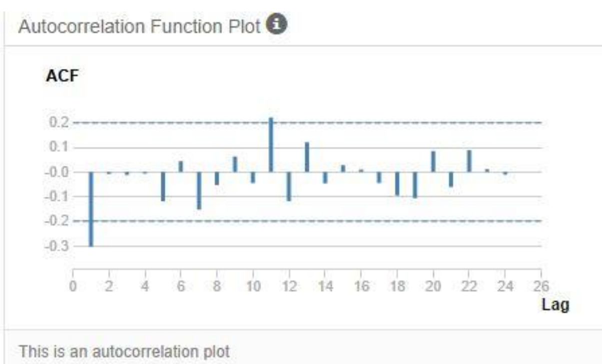
Now we determine the seasonal difference ACF and PACF.

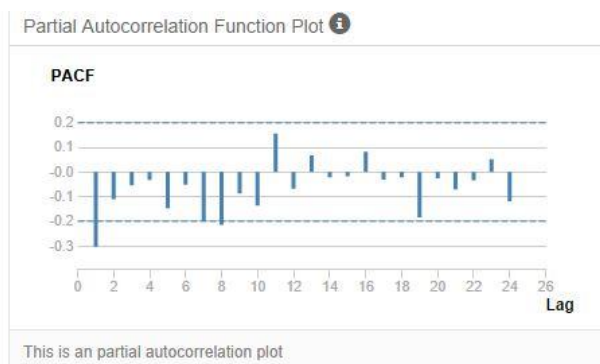




It is observed that the graphs are similar to the previous case. However, the correlation is lesser and hence first differencing is required. Seasonal First difference ACF and PACF is now performed.

Significant correlation is removed. Any remaining correlation can be accounted by the AR and MA terms.





Now, from the following discussion we can define the ARIMA terms

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

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

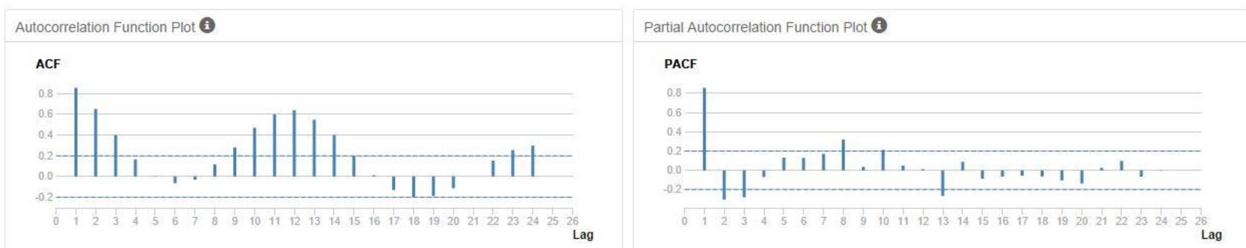
In-sample error measures:

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

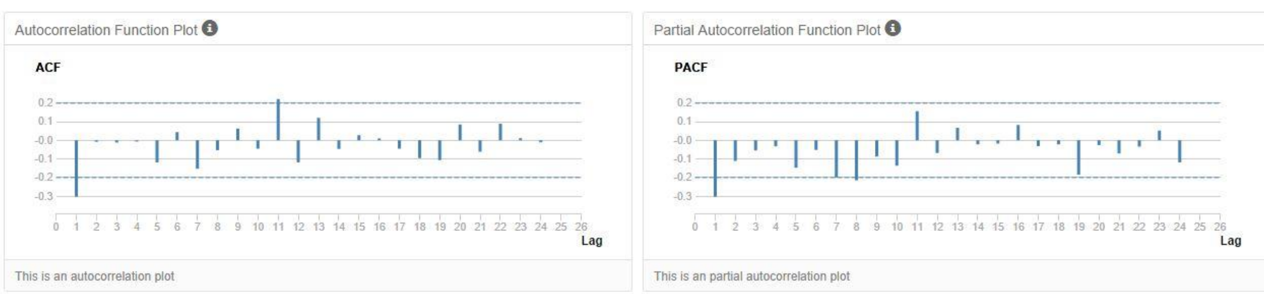
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

After establishing the correct ARIMA model, ACF and PACF is regraphed for both the Time Series and Seasonal Difference . 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.

Time Series



Seasonal 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)

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.

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

AIC

Information criteria:

AIC	AICc	BIC
1634.6435	1645.9768	1669.4337

(ets)

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

(arima)

AIC value of ARIMA model is lower indicating a better fit model.

In sample errors

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

(ets)

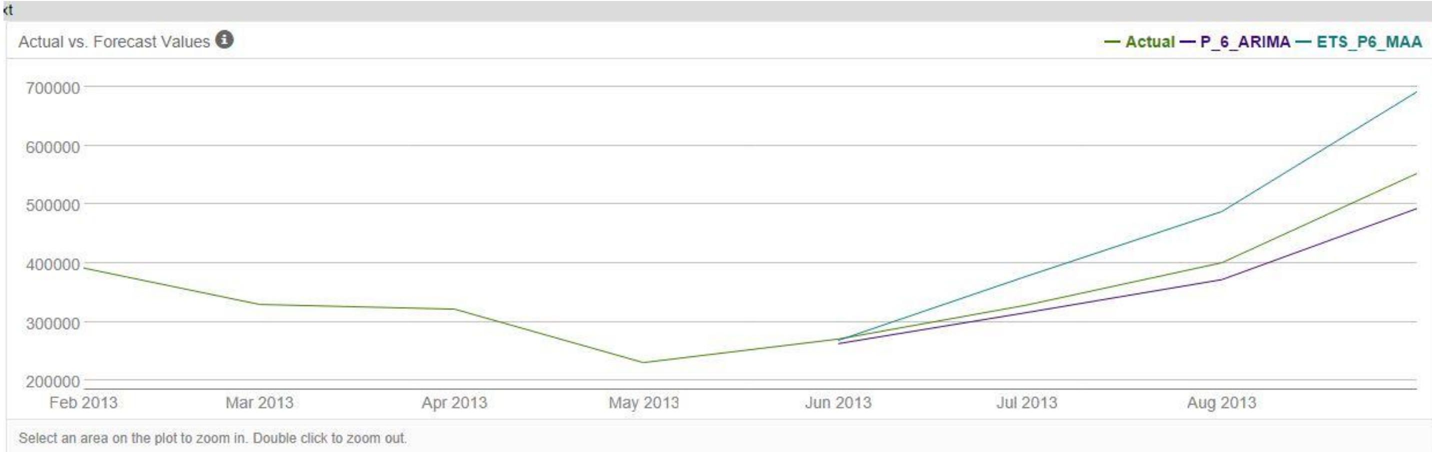
In-sample error measures:

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

(arima)

Forecast with holdout sample

Record #	Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
1	P_6_ARIMA	27271.5199	33999.7911	27271.5199	6.1833	6.1833	0.4532	[Null]
2	ETS_P6_MAA	-68257.4703	85623.175	69392.7195	-15.2446	15.6635	1.1532	[Null]



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.

7 of 7 Fields | Cell Viewer

Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2013	10	754854.460048	833335.856133	806170.686679	703538.233418	676373.063963
2	2013	11	785854.460048	878538.837645	846457.517118	725251.402978	693170.082452
3	2013	12	684854.460048	789837.592834	753499.24089	616209.679206	579871.327263
4	2014	1	687854.460048	803839.469806	763692.981576	612015.938521	571869.450291

`