Project: Forecasting Sales

Step 1: Plan 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.

We consider a company that creates and sells video games. The file “monthly_sales” contains information about the company's sales. It's a time series given that is a list of observations where the ordering matters, each measurement of data taken across sequential and equal intervals, in fact, historical sales data contains monthly data from 2008-01 to 2013-09, and moreover, with each time unit have at most one data point.

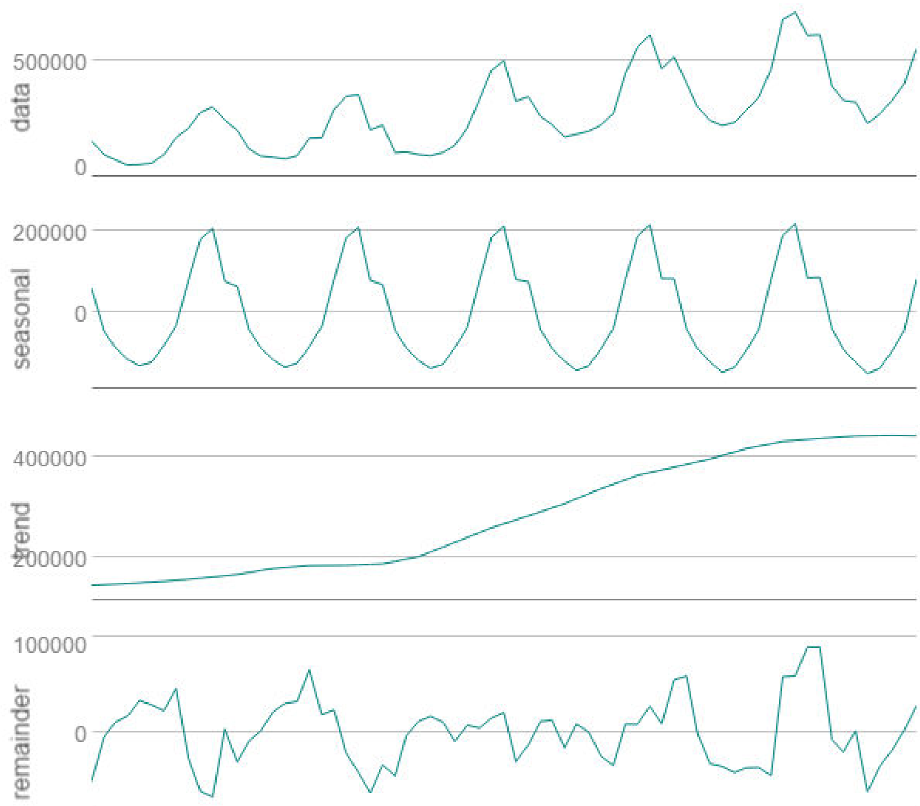
2. Which records should be used as the holdout sample?

We need to forecast monthly sales data in order to synchronize supply with demand. The size of the holdout sample depends on how long the time series is and how far you would like to forecast. The goal is to provide a forecast for the next 4 months of sales. Then, the size of the holdout sample should be the number of periods we want to forecast, 4 months, from 2013-06 to 2013-09.

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.

The original time series is decomposed into three sub-time series, that is the seasonal component, the trend component and the remainder. Below, we report the time series decomposition plot.

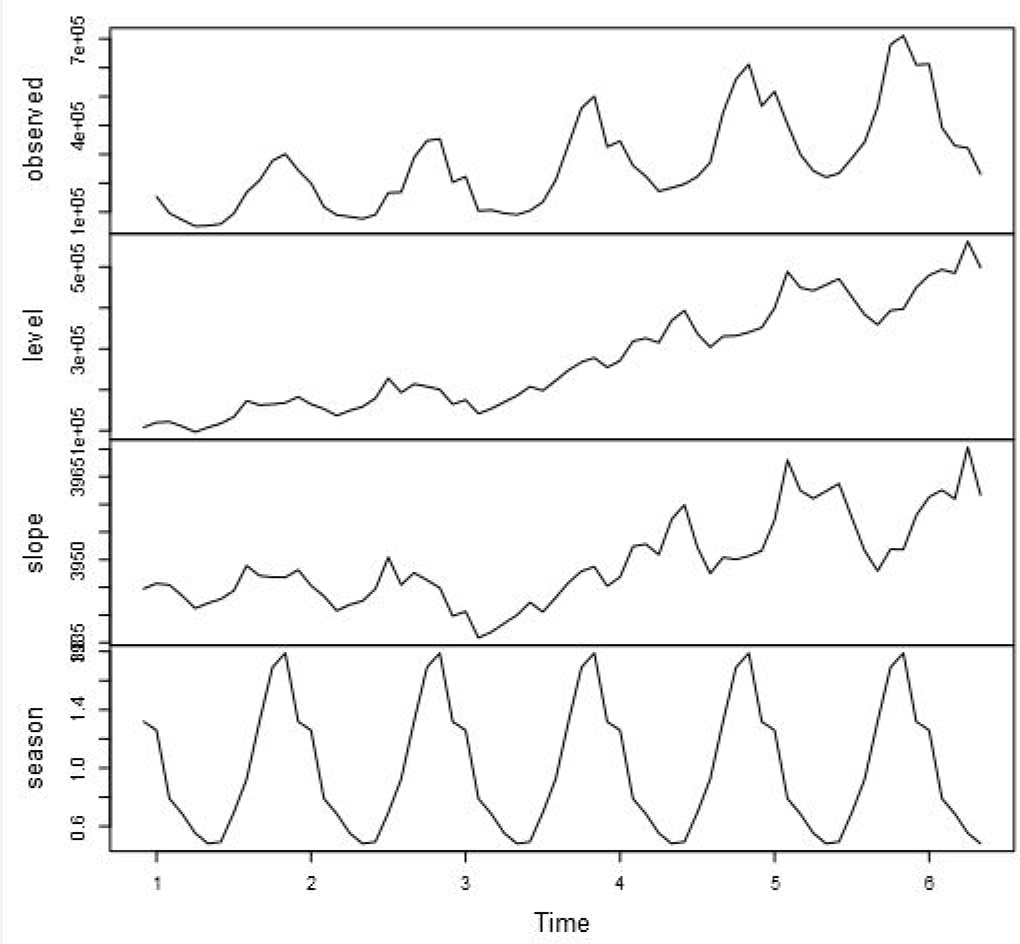


The time series is characterized by a pattern that repeats with a fixed period of time, so-called seasonality. It is a seasonality of 12 months. Furthermore, it is present a trend component that reflects the long-term progression of the series.

Step 3: Build your Models

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

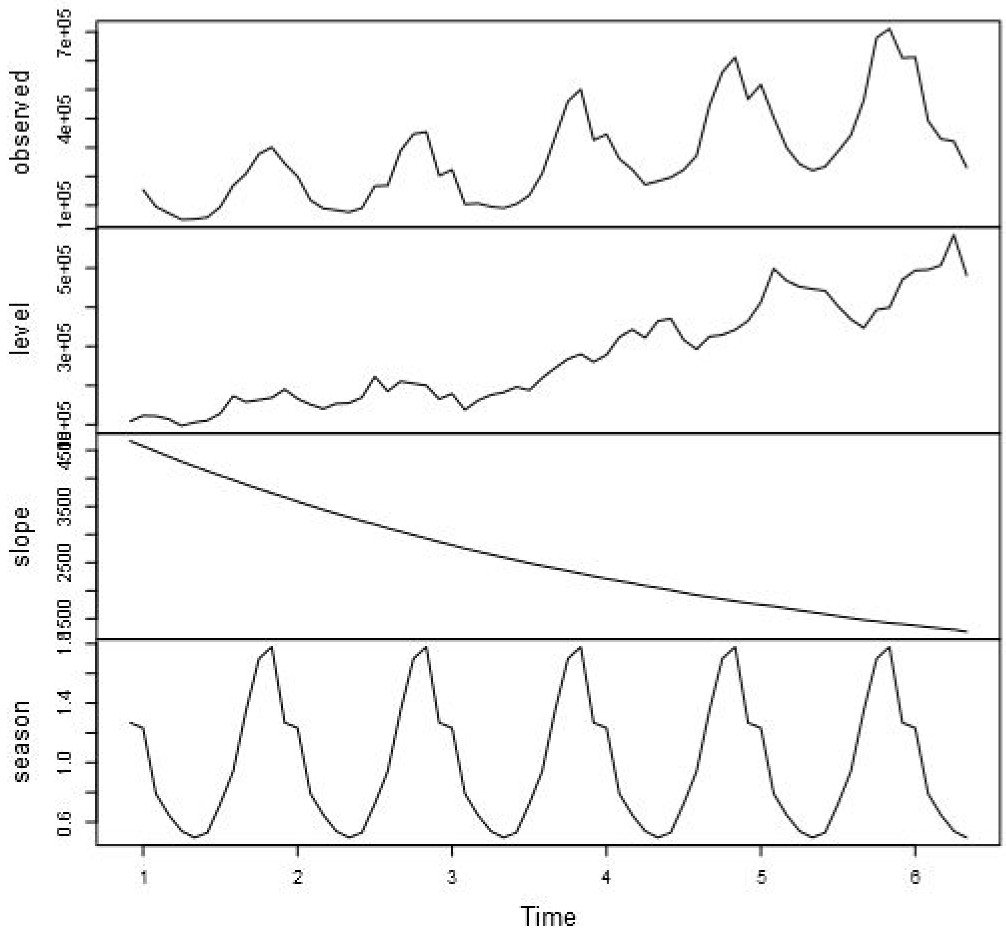
Looking the seasonal component, trend component and noise/remainder component in the time series decomposition plot in Alteryx, we build the ETS model. Seasonality is growing slightly overtime (the peaks are increasing ever so slowly), so we apply it multiplicatively. There is an increasing direction in the data then a trend exists. The trend is linear then we apply it additively. Furthermore, we consider the remainder that is the residuals of the time series after allocation into the seasonal and trends time series. The error is growing or shrinking over time, we apply the error multiplicatively. Then, we run both dampened and non-dampened ETS (M, A, M).
Below, we report the results of non-dampened ETS in the estimation period:



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

AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

Below, we report the results of dampened ETS in the estimation period:



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

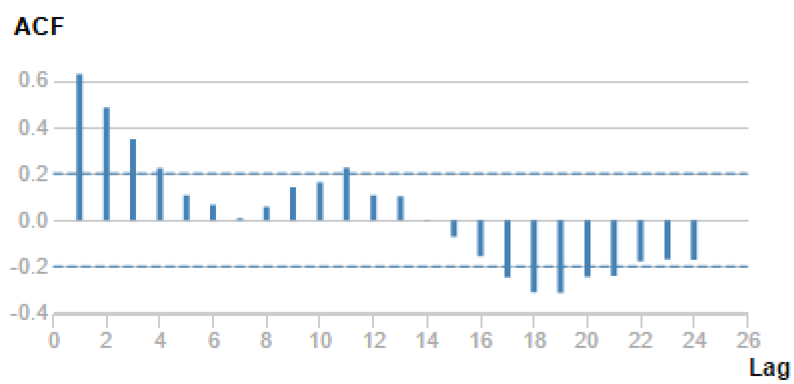
AIC	AICc	BIC
1639.465	1654.3346	1678.604

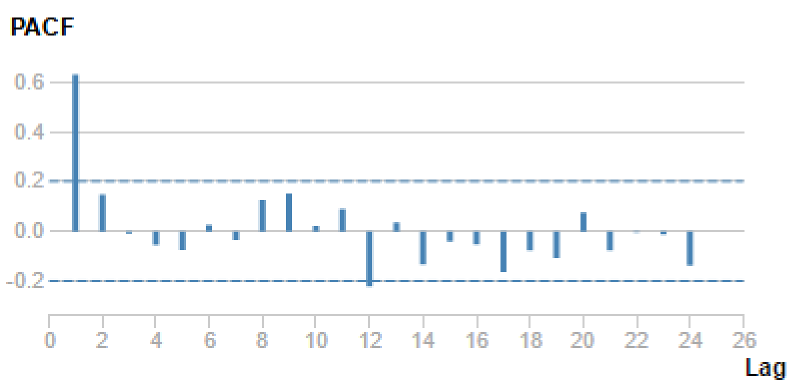
We consider two accuracy measure, RMSE and MASE, to judge what a good forecasting model is. For non-dampened ETS model, RMSE is 32992.72 and MASE is 0.37. For dampened ETS model, RMSE is 33153.52 and MASE 0.36. The RMSE of dampened ETS model is greater than non-dampened ETS model instead the MASE of dampened ETS model are lower than non-dampened ETS model. Dampened ETS model is chosen given that has AIC lower than non-dampened 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.

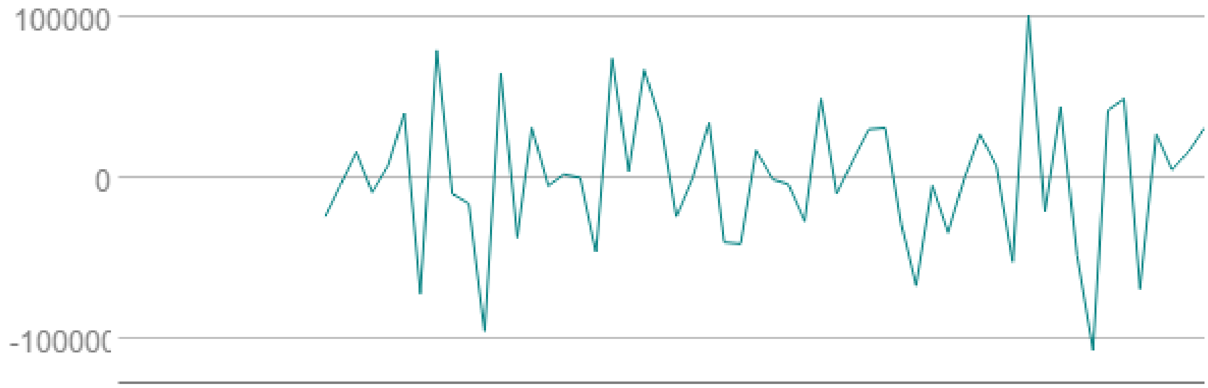
a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

Fitting an ARIMA model requires the series to be stationary. The monthly sales data is non-stationary. The plot of time series shows an upward trend and seasonality.
Furthermore, there are some trend or seasonal components and therefore its statistical properties are not constant over time if the series is correlated with its lags. Autocorrelation plots (ACF) or partial autocorrelation (PACF) plots are a visual tool in determining the existence of autocorrelation for any particular lag.

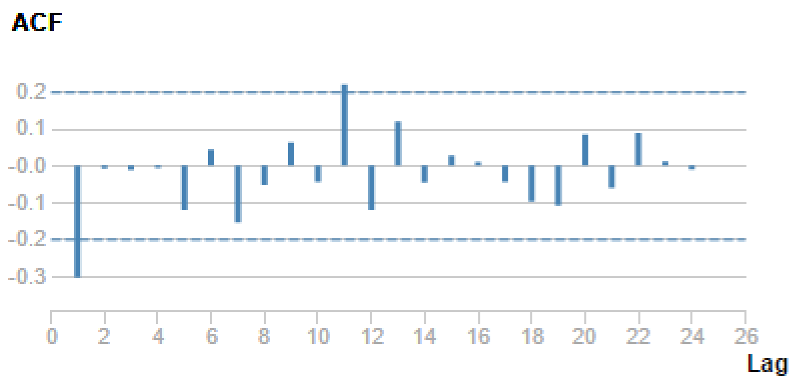


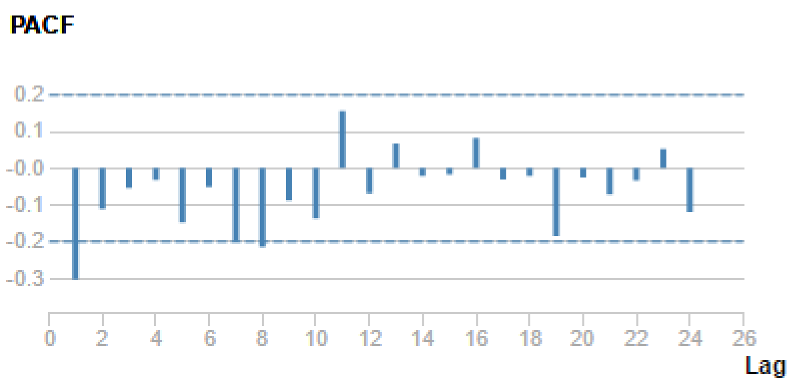


In particular, we note that the ACF shows an oscillation, indicative of a seasonal series. Examine the patterns across lags that are multiple seasonal periods. For monthly data, look at lags 12, 24, we note the peaks occur at lags of 12 months and 24 months. Furthermore, we observe a spike at lag 1 in an ACF plot indicates a strong correlation between each series value and the preceding value. Then, we fit the times series with seasonal ARIMA model.
Non-stationary series can be corrected by a simple transformation such as the differencing. We consider seasonal first difference, we can observe in the plot below and the time series has been stationarized.



By looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the seasonal first difference, we can identify the numbers of AR and/or MA terms that are needed.



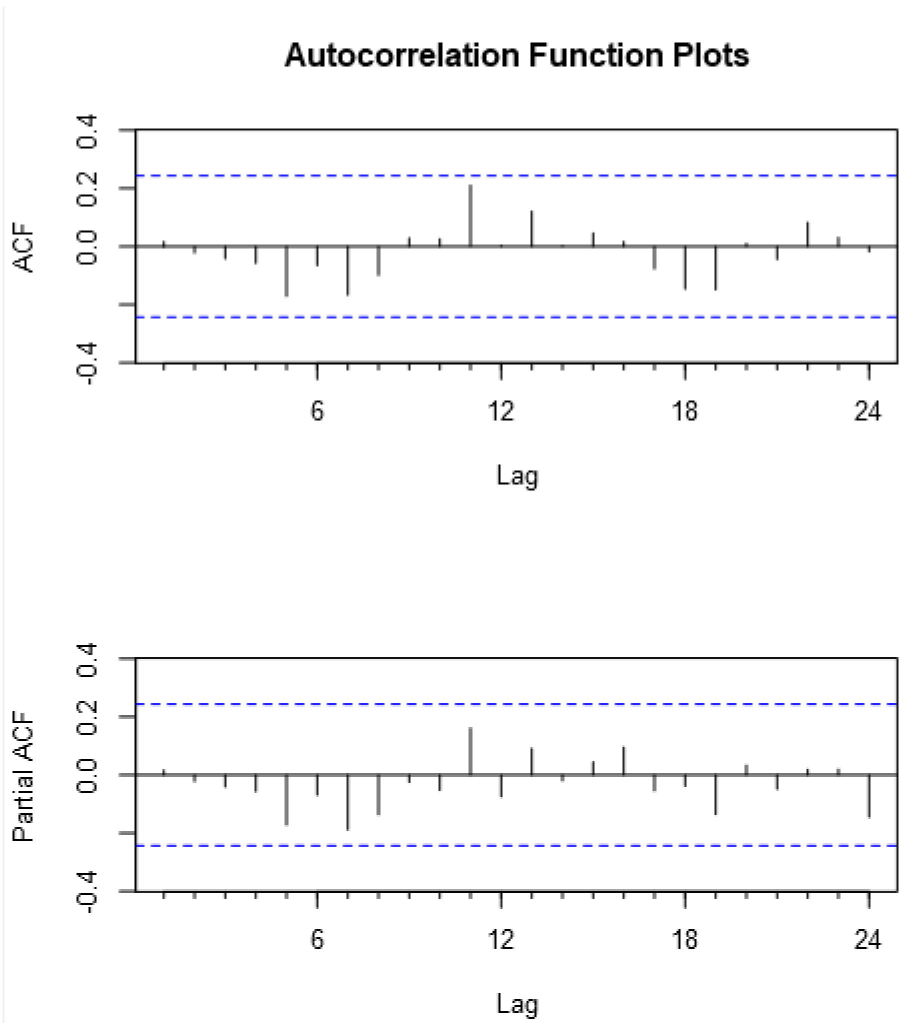


For non-seasonal terms, we examine the early lags and we observe a spike in the ACF at lag 1 value of the series, this indicates non-seasonal MA terms.
For seasonal terms, we note that there are no more peaks occur at lags of 12 months and 24 months.
Then, the model that fits is ARIMA (0, 1, 1) (0, 1, 0) 12. Below, we report the results of seasonal ARIMA model in the estimation period:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

To understand the forecasting errors and accuracy measurements of the model, RMSE and MASE are 36761.52 and 0.36, respectively.
Below, we report ACF and PACF:



We observe that the model doesn't present correlation. This confirms stationary of the built ARIMA model.

Step 4: Forecast

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

We report the results of dampened ETS model and seasonal ARIMA model in the estimation period:

- **dampened ETS**

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

- **seasonal ARIMA**

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

We report the results of dampened ETS model and seasonal ARIMA model in the validation period:

- **dampened ETS**

Actual	MAM_damped
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

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

- **seasonal ARIMA**

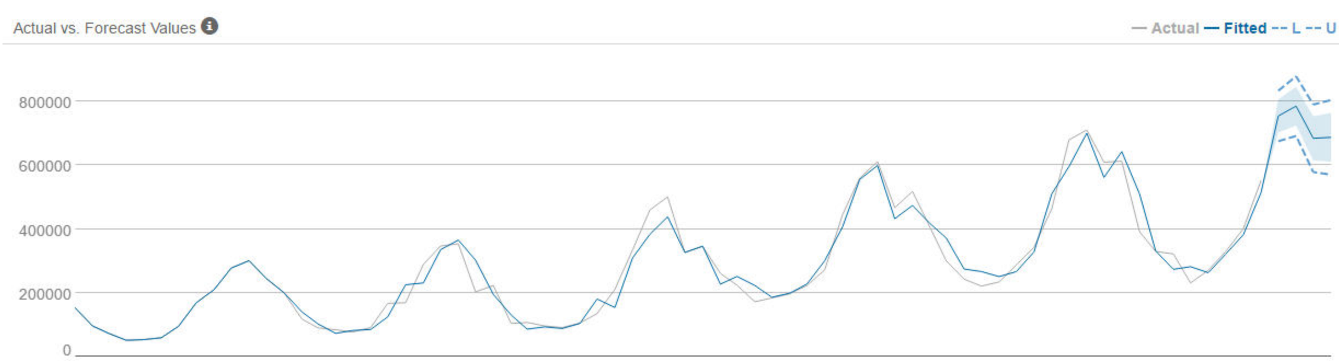
Actual	seasonal_ARIMA
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
seasonal_ARIMA	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

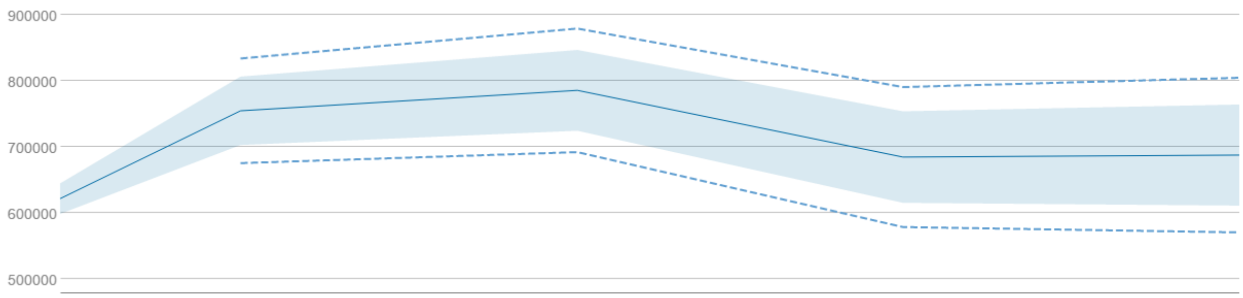
If we consider in-sample error measurements. For the dampened ETS model, RMSE is 33153.52 and MASE 0.367. For the ARIMA model, RMSE is 36761.52 and MASE is 0.364. RMSE of the dampened ETS model is lower than the seasonal ARIMA model. Instead, we consider MASE of the two models the difference is small.
Nevertheless, if we consider the forecasts of sales of models against the holdout sample, the seasonal ARIMA model is better. RMSE and MASE of the ARIMA model are lower than the dampened ETS model. For the ARIMA model, RMSE is 33999.79 and MASE 0.453. For the dampened ETS model, RMSE is 60176.47 and MASE is 0.81. Therefore, the ARIMA model is chosen.

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

Below, we report the forecast for the next period, 2013-10 to 2014-01:



Below, the zoom graph where we see dotted lines.



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
6	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
6	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
6	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
7	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462