Farida Muhammad

Project: Forecasting Sales

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** does **meet** the criteria of a **time series** dataset. The Monthly Sales variable in the dataset is **numeric** and is tracked and collected over time. The data covers a continuous time interval. The data in the dataset is ordered by date in **YYYYMM** order. **Order** matters in this dataset because there is a dependency on time and changing the order could change the meaning of the data. Each measurement of the data is taken across sequential and equal intervals. Each **time unit** has **one** data point.

2. **Which records should be used as the holdout sample?** Since, I have been tasked with providing a forecast for the next 4 months of sales, I used the last 4 months (4 records) as the holdout sample. Ideally, the size of the holdout sample should be at least the amount of periods I am forecasting for.

Step 2: Determine Trend, Seasonal, and Error components

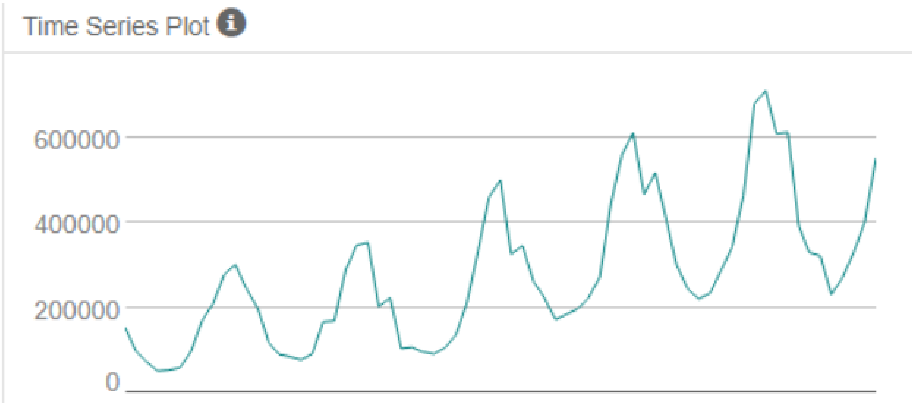
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

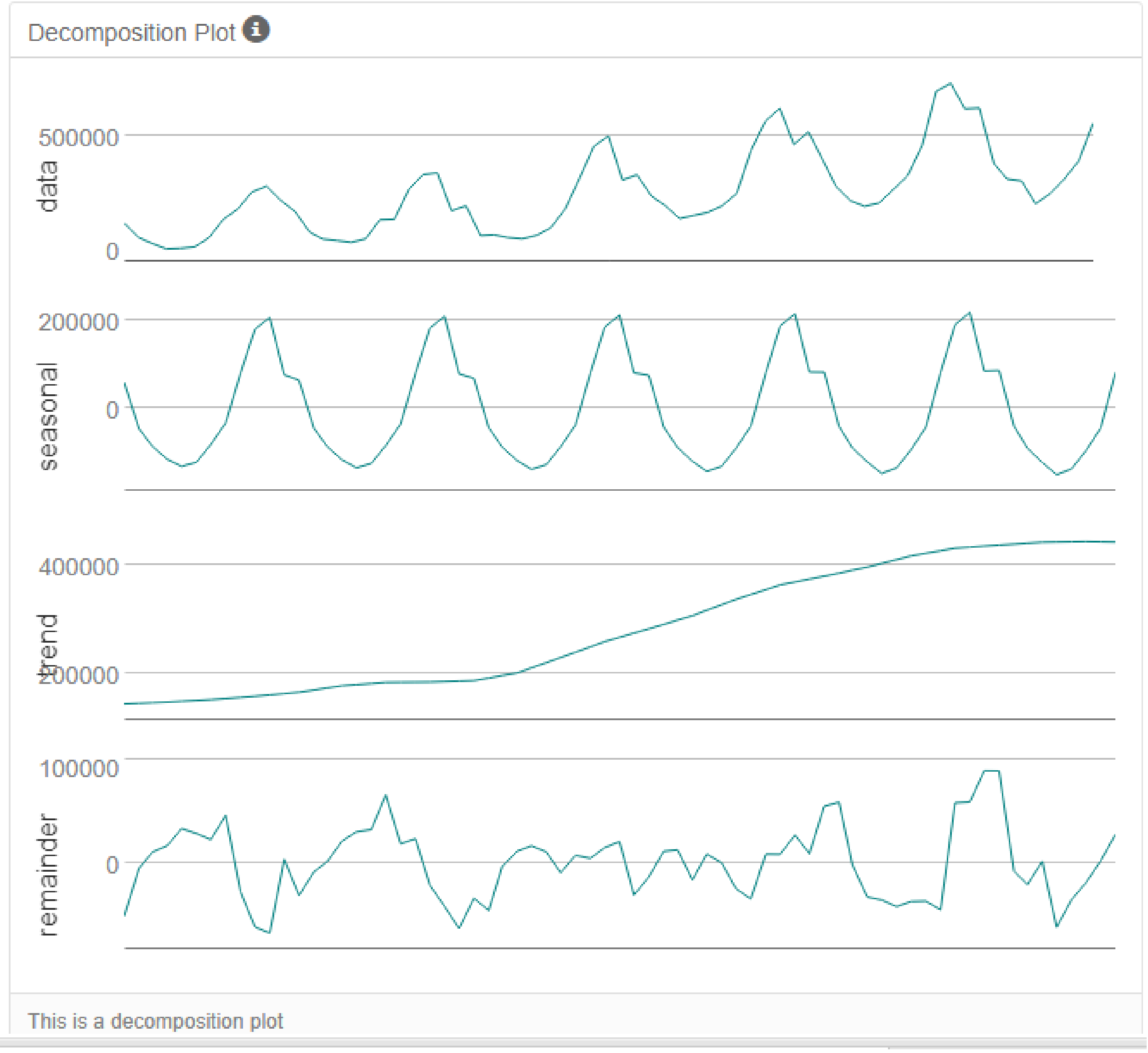
Answer this question:

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 plot tool in Alteryx was used to generate the time series and decomposition plots from the data set.

   The trend pattern exhibits an uptrend, a general direction that is upwards, where there are higher peaks and higher valleys.

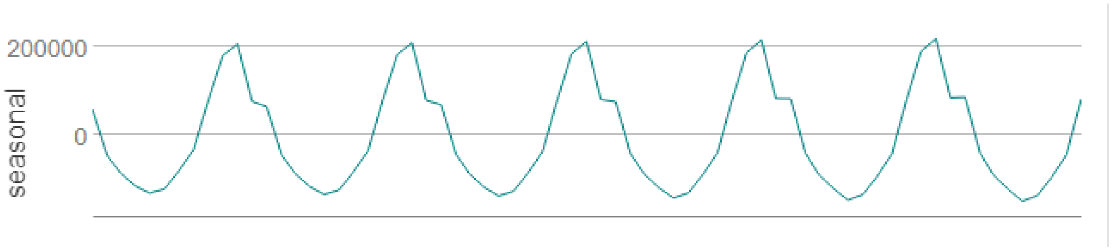


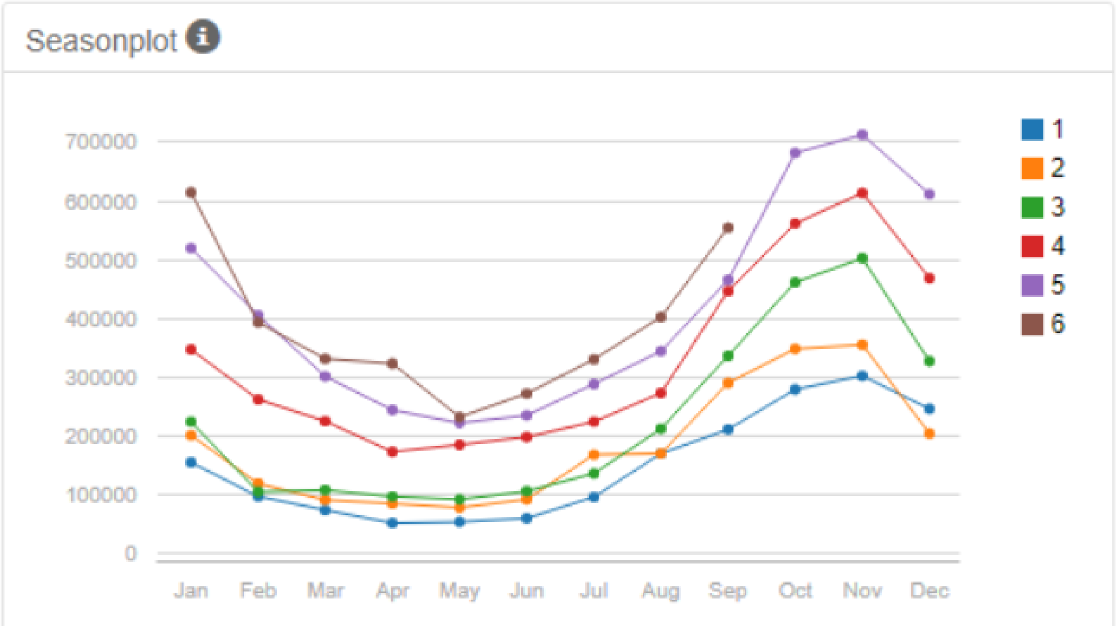


This is a decomposition plot

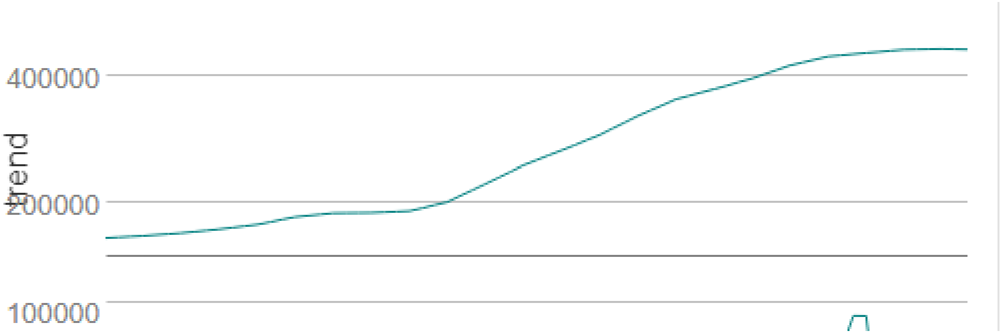
The **seasonal component** in the time series decomposition plot shows the following:

Peaks and valleys in monthly video game sales repeats each year during the same time intervals, therefore the time series have seasonality. **Seasonality** increases slightly over time, so the **multiplicative** method should be applied.



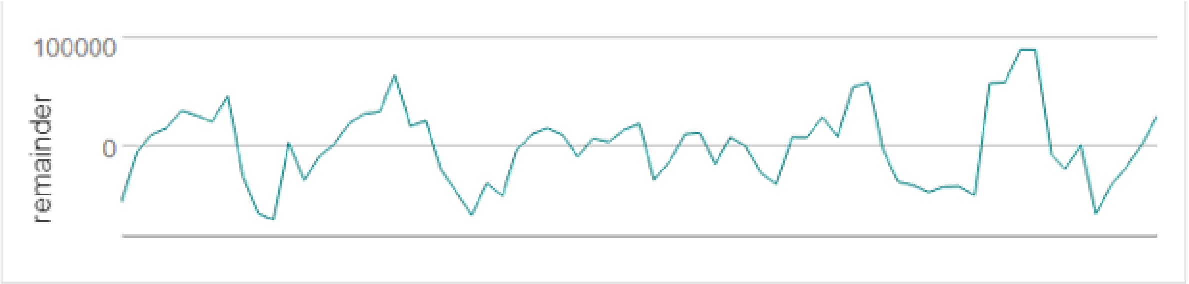


The **trend component** in the time series decomposition plot shows the following: following:



The trend appears to be linear, therefore the **additive** method will be applied.

The **error component** (labeled remainder) in the time series decomposition plot shows the following: following:



This is a decomposition plot

The data fluctuates in the error plot, so the so the **multiplicative** method should be applied.

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)

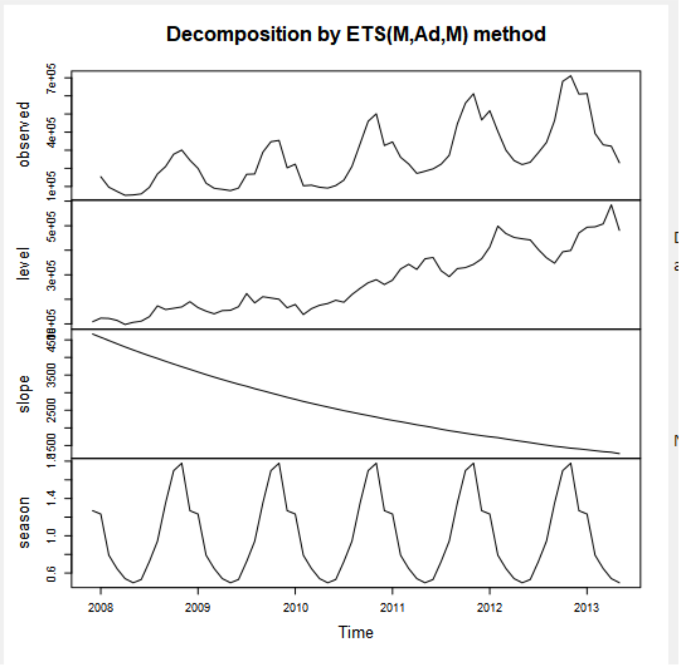
Answer these questions:

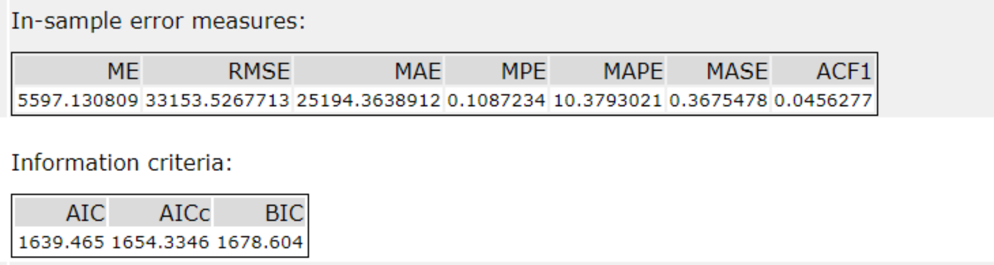
1. **What are the model terms for ETS? Explain why you chose those terms.**
    a. **Describe the in-sample errors. Use at least RMSE and MASE when examining results**

Based on the decomposition plot above, I chose to use model terms M,A,M for the ETS model. Dampened and non-dampened ETS models were run with a holdout sample of 4 periods (months).

I will review the RMSE, MASE in-sample errors and the AIC information criteria for the ETS models. The **RMSE** represents the sample standard deviation from the mean. The Mean Absolute Scaled Error **(MASE)** is the mean absolute error of the model divided by the mean absolute value of the first difference of the series. It measures the relative reduction in error compared to a naive model. Ideally its value will be significantly less than 1. The **AIC** measures the relative quality of a statistical model. When comparing two models, the one with the lower AIC is generally “better”.

Dampened ETS Model





In-sample error measures:

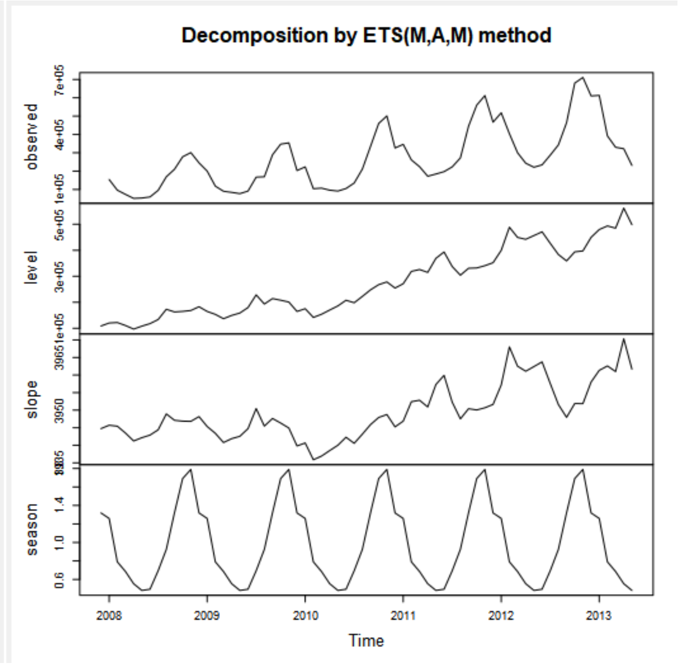
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

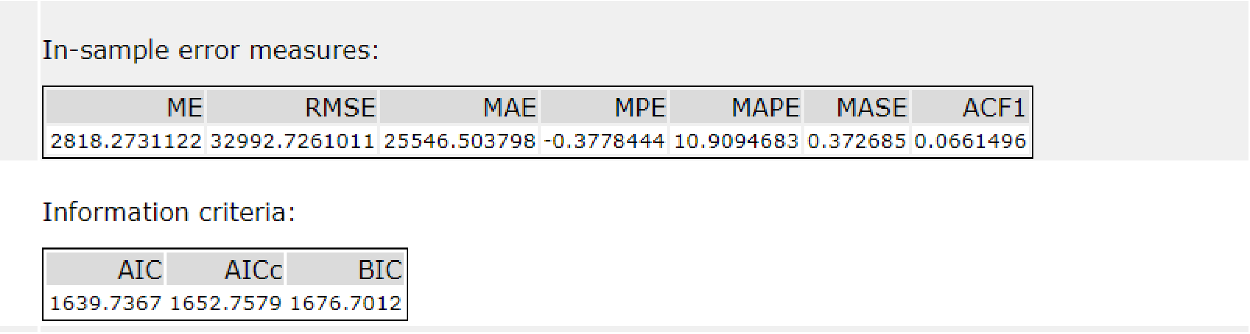
Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

RMSE is **33153.53** and MASE is **0.3675**. The AIC value is **1639.47**.

Non-Dampened ETS 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:

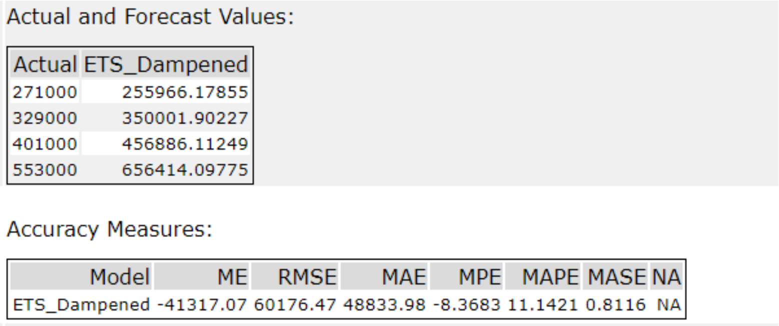
AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

RMSE (Moot Mean Square Error) is **32992.73** and MASE (Mean Absolute Percentage Error) is **0.3727.** The AIC value is **1639.74**.

Validating the Forecast

After comparing the actual results with the forecasted results, I chose the dampened model since it is more accurate. Also, the dampened model’s RMSE, MASE, and AIC are lower.

Dampened ETS Model



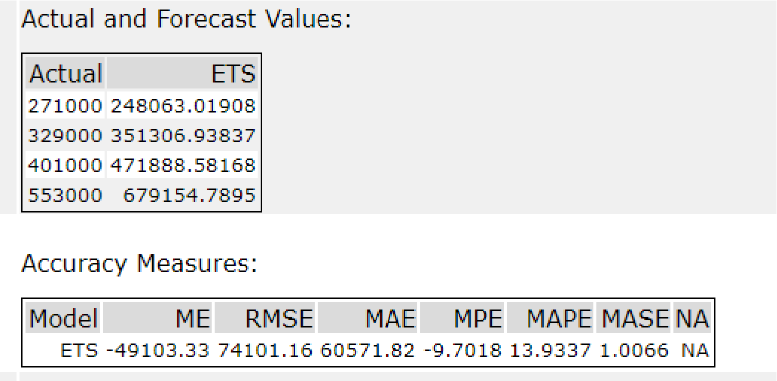
Actual and Forecast Values:

Actual	ETS_Dampened
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Accuracy Measures:

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

Non-Dampened ETS Model



Actual and Forecast Values:

Actual	ETS
271000	248063.01908
329000	351306.93837
401000	471888.58168
553000	679154.7895

Accuracy Measures:

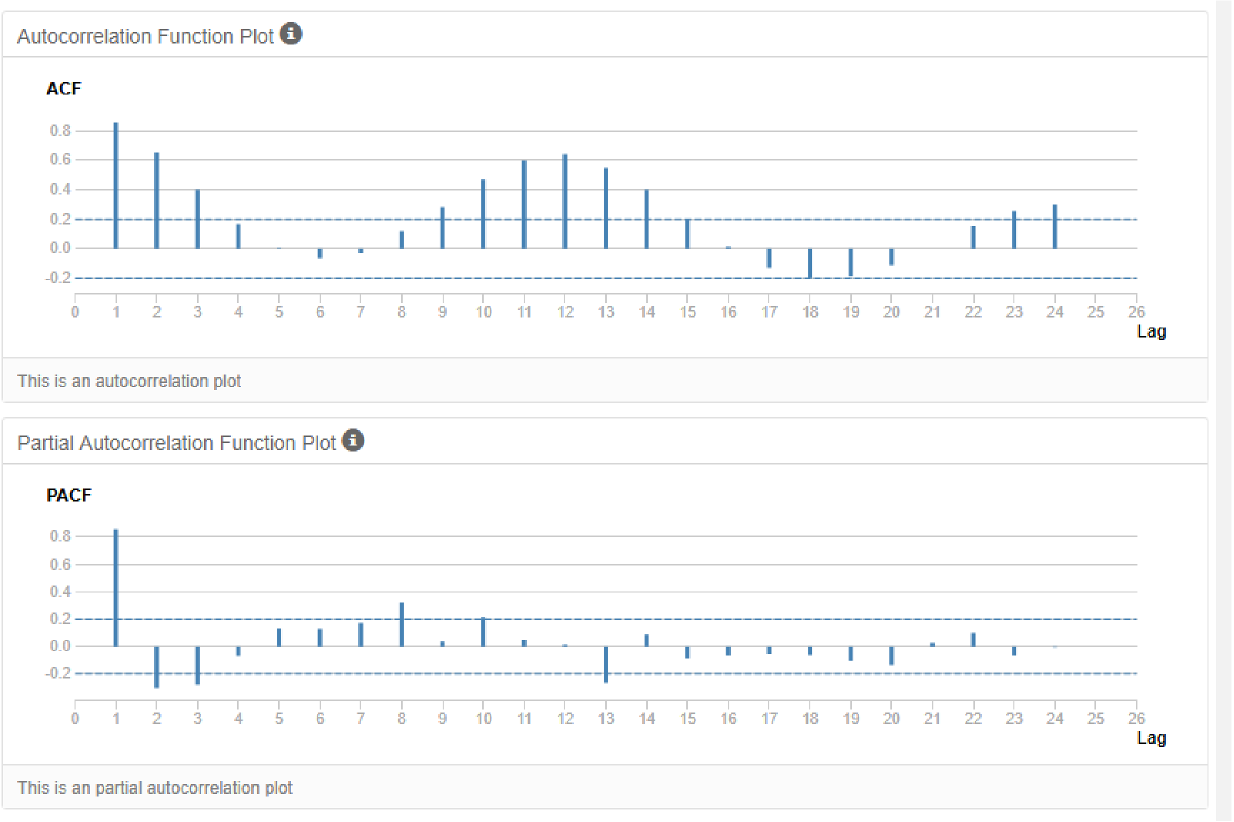
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066	NA

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.

The model terms for ARIMA: ARIMA(0,1,1)(0,1,0)12 is used since lag-1 is negative and 12 months is the number of periods.

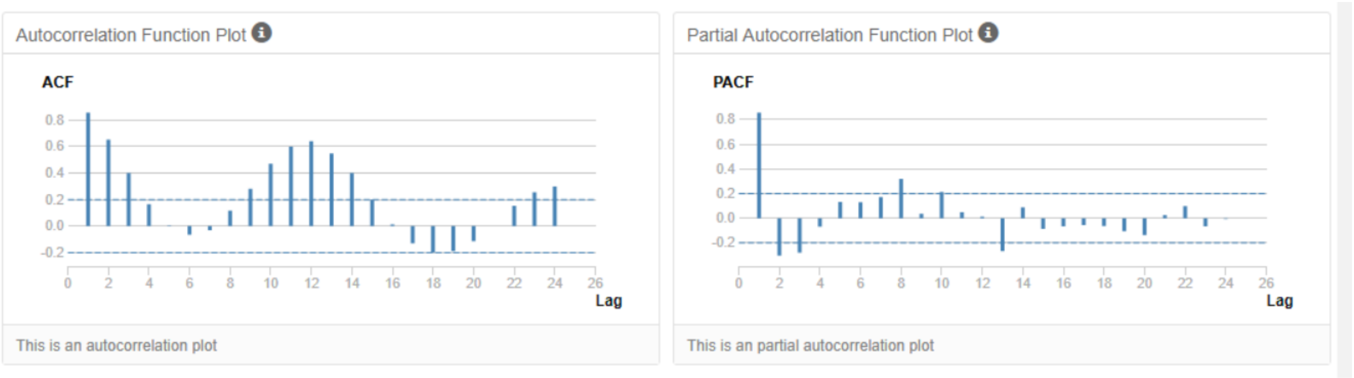
Without seasonal differencing, the time series and seasonal component's Auto-Correlation Function (ACF) shows high correlation and the Partial Autocorrelation Function (PACF) shows a significant lag at period 13, which is a result of the seasonal effect.



This is an autocorrelation plot

This is an partial autocorrelation plot

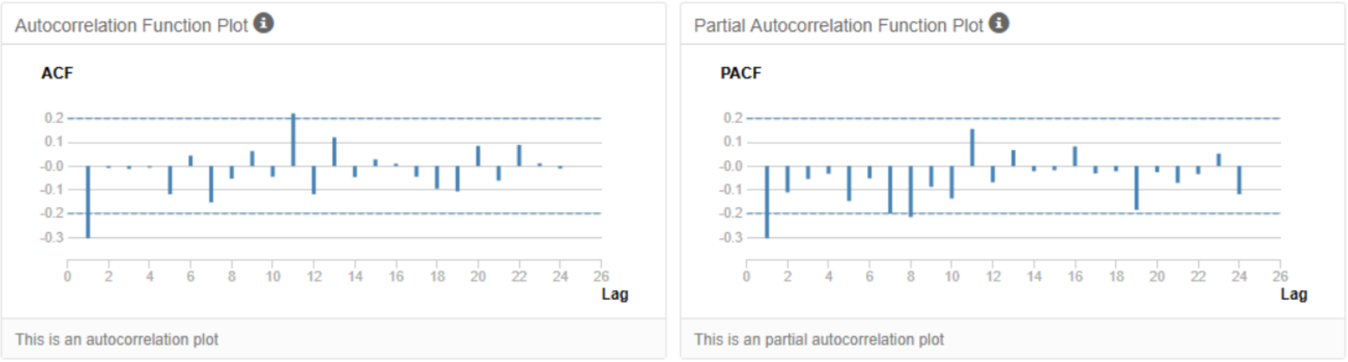
When a seasonal difference is applied, the ACF graph still shows high correlation. While, the PACF graph does not show a strong correlation.



This is an autocorrelation plot

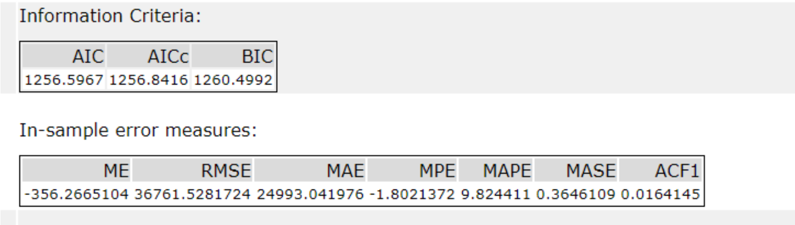
This is an partial autocorrelation plot

A seasonal first difference is applied and serial correlation has disappeared in the ACF plot.



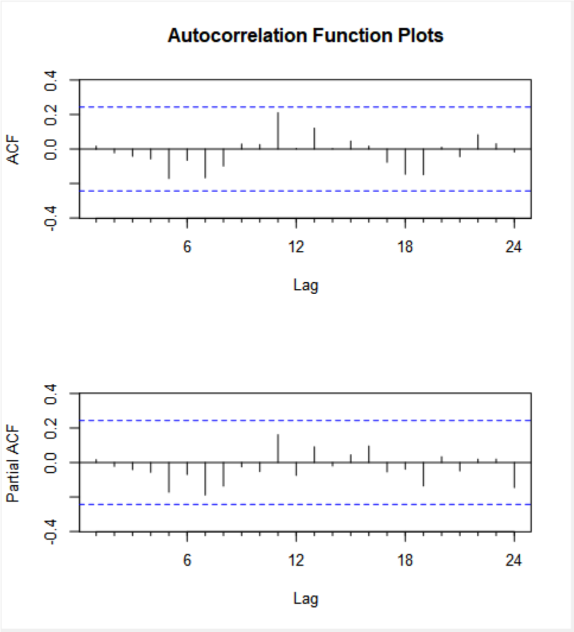
In-Sample Errors (ARIMA)

As can be seen below, the RMSE is **36761.53** and **MASE is 0.3646**. The **AIC** is 1256.6



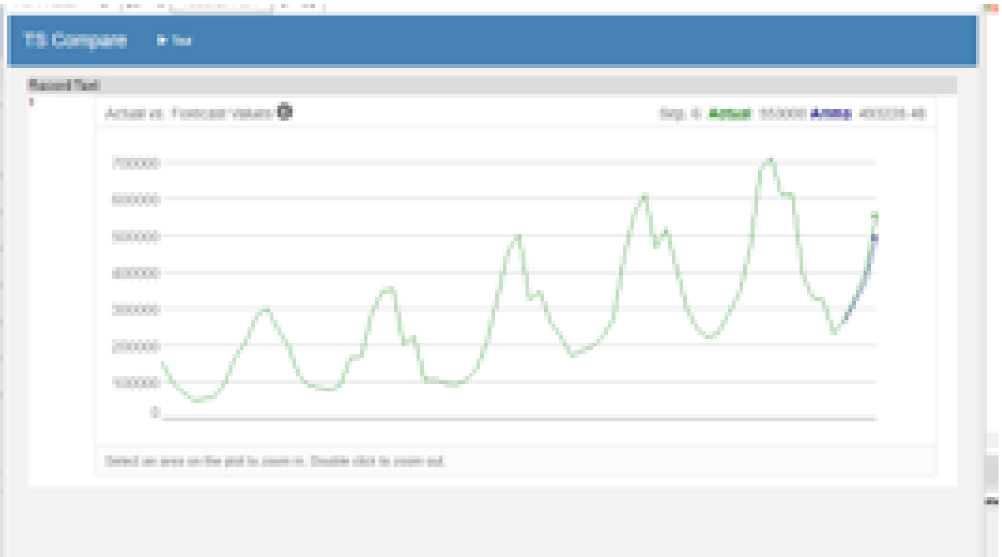
```
Information Criteria:

      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
```

Based on the ACF and PACF charts below, no additional AR or MR terms are need since neither chart show significant correlation. The bars are also between the dashed lines.



Model Validation (ARIMA)

The predicted values were compared to the actuals in the validation sample using the *TS Compare* tool. For the final period, the ARIMA model forecasted 493,228.48 vs the actual of 553,000.


TS Compare
Actual vs. Forecast Values
Actual 553000 Arima 493228.48
700000
600000
500000
400000
300000
200000
100000
0
Select an area on the plot to zoom in. Double click to zoom out.

Step 4: Forecast

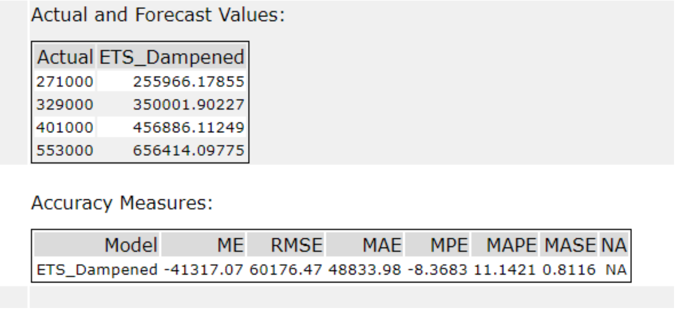
Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)

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**.

   I chose the ARIMA model because it is better at forecasting sales against the holdout sample. The MAPE and ME values are lower than ETS dampened model values. In addition, the RMSE for ARIMA is **33999.79** compared to RMSE for ETS dampened model at **60176.47**. The MASE value for the ARIMA model is **0.4532,** which is also lower than MASE value for the ETS dampened model at **0.8116**. Therefore, the **ARIMA** model is **better**, since its forecast error and in-sample error measurements are smaller. The ARIMA **AIC value,** 1256.6, is also lower than the **AIC** for the ETS dampened model, **1639.47**.

Dampened ETS Model



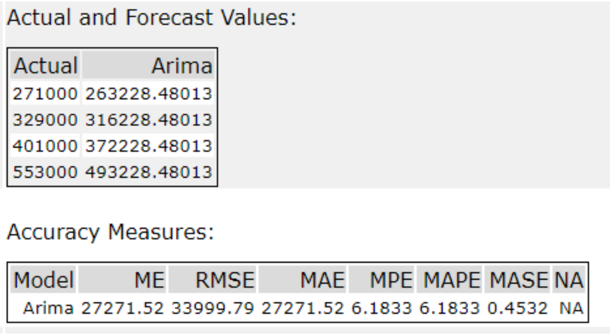
Actual and Forecast Values:

Actual	ETS_Dampened
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Accuracy Measures:

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

ARIMA Model



Actual and Forecast Values:

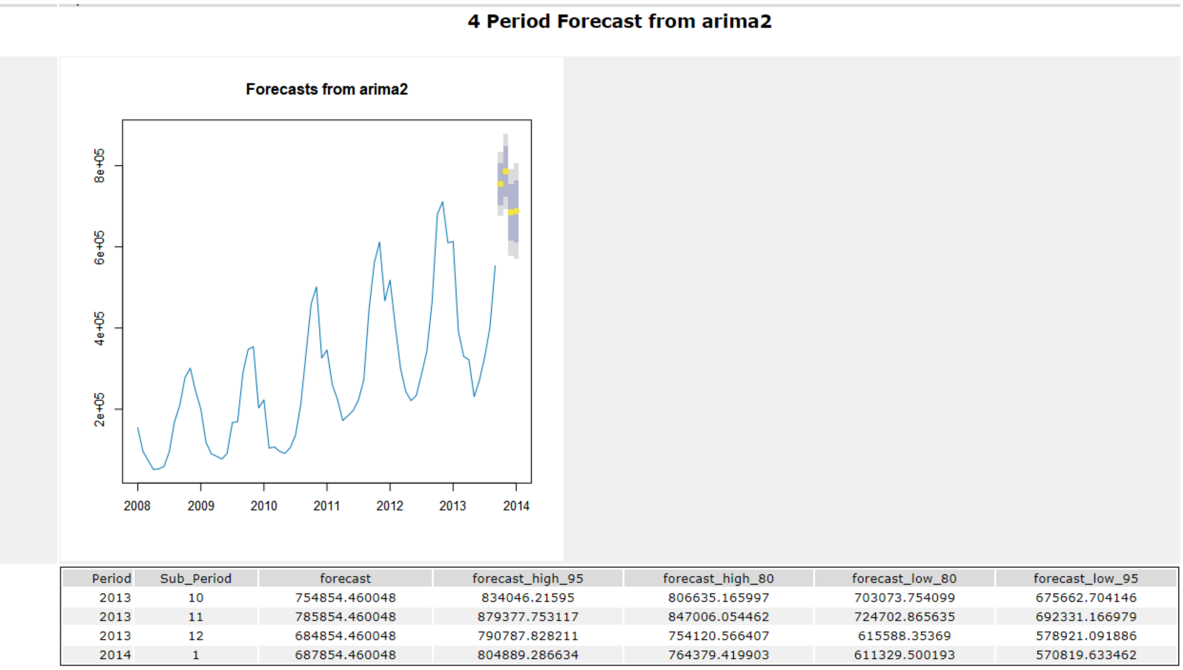
Actual	Arima
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

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

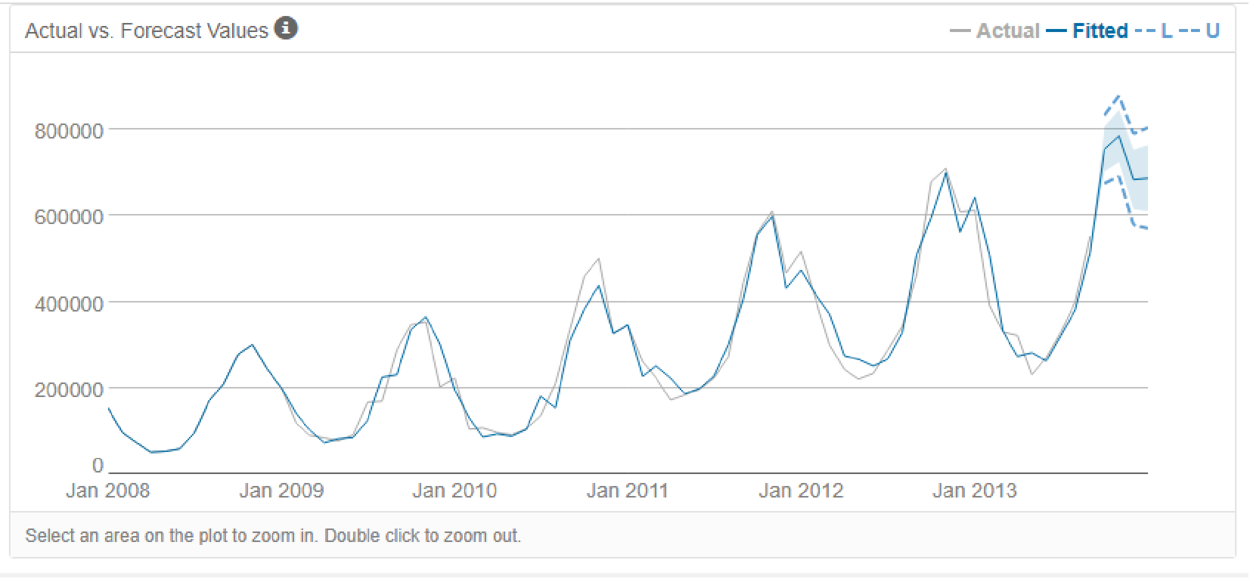
2. **What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.**

   The forecast for the next 4 periods, October 2013 to January 2014, are **754,854**, **785,854**, **684,854** and **687,854**.



4 Period Forecast from arima2

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



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.