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

Four Key Characteristics of a time series data:

* Time Series is a list of observation where the ordering matters. There is a dependency on time and changing the order could change the meaning of the data
* Time series data are sequential
* The data points have equal interval
* Each time unit having at most one data point

The data set of video game sales have met all these four requirements. Therefore, it is a solid time series data.

The record that should be used as the hold out sample should be the most recent records. And the number of data point should be at least same amount f the number of data point we would like to forecast

In this case, we want to forecast the sales of video game for the next 4 months, therefore we will take the 4 months of latest sales which is : Year 2013 month 6,7,8,9 to predict sales of year 2013 month 10,11,12 & year 2014 month 1

## 

## 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.

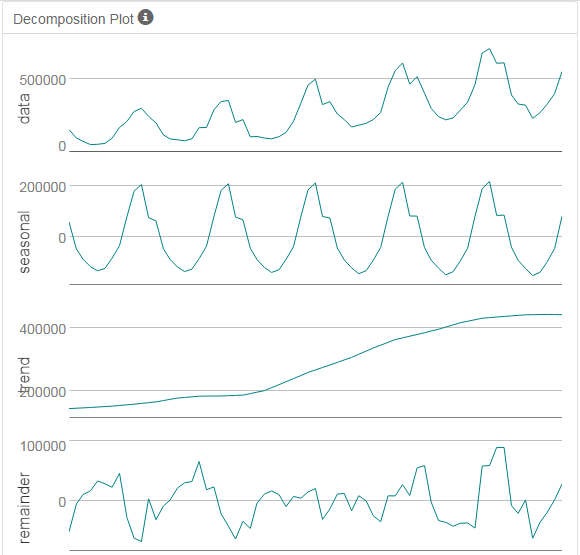
Based from the decomposition plot we can see that there is

An increase in error (M)

Exponential trend (M)

And an increase in seasonality (M)

(In seasonality at a glance it looked like a constant trend but if we take a look deeper into the data point, the peaks slightly increases over time, therefore we have to include seasonality as increase (Multiplicative model))

**

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)*

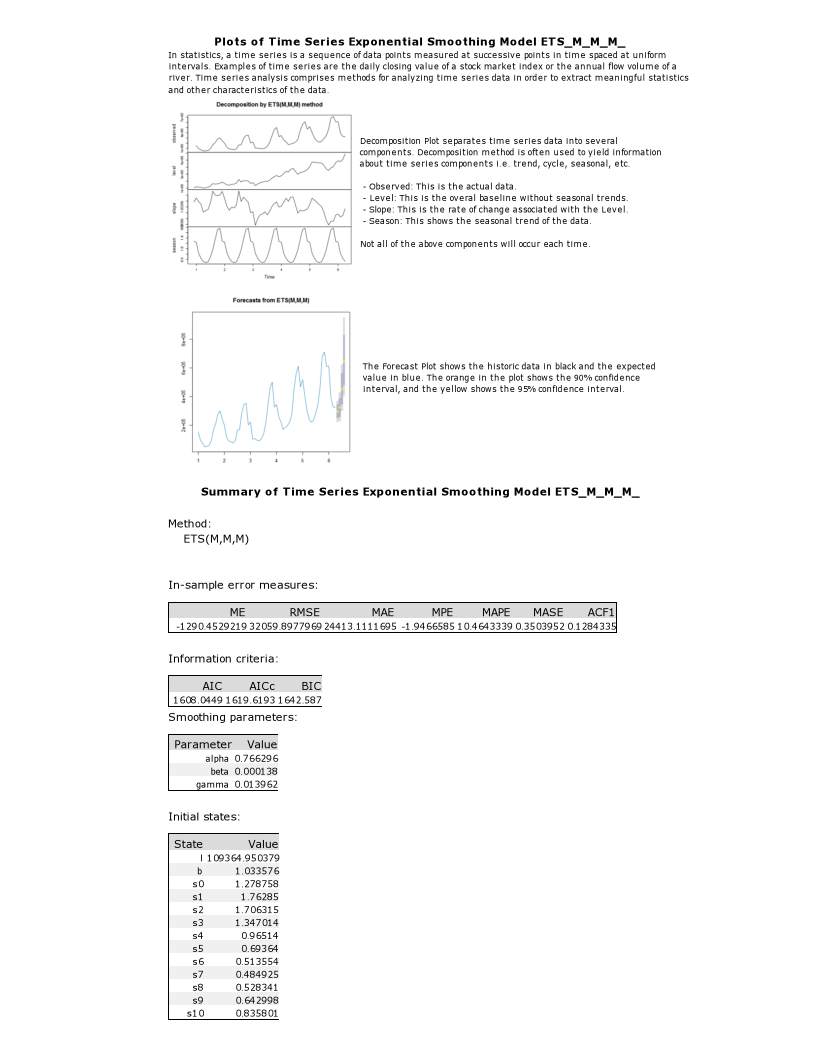
*Answer these questions:*

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

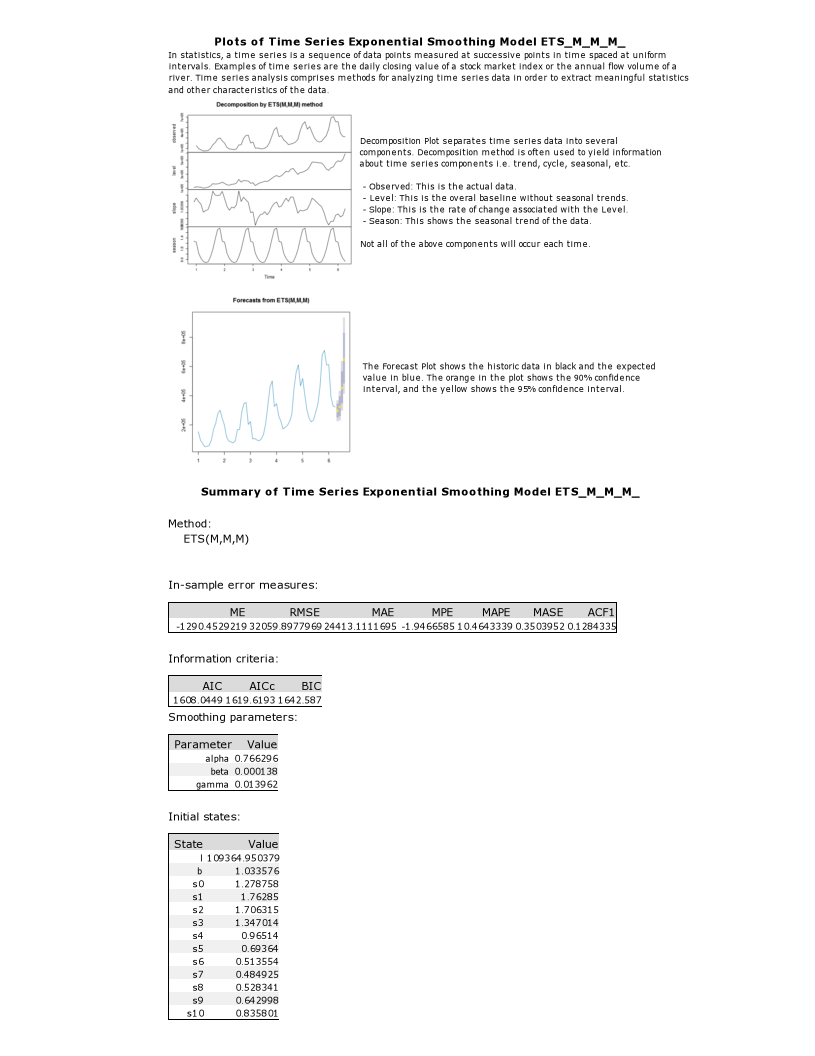
The model term for ETS is ETS (M,M,M)

Because it has increasing error, exponential increase in trend, and increasing seasonality

**ETS M,M,M Result**



**ETS (M,M,M) Dampen Result**



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
| ETS\_M\_M\_M\_ | -175503 | 216070.6 | 175502.6 | -43.5058 | 43.5058 | 2.9406 |
| ETS\_M\_M\_M\_DAMPEN TREND | -128640 | 152964.6 | 128639.7 | -32.7399 | 32.7399 | 2.1554 |

ETS (M,M,M)

Average Error (ME) = -175503

Std. Dev.of Mean (RMSE) = 216070.6

Average Absolute Value (MAE) = 175502.6

Average Percentage Error (MPE) = -43.5058

Mean Absolute Scale Error (MASE) = 2.9406

The Average percentage error is very high at -43%

And MASE is way above 1.0 which is 2.94

Personally I don’t think it is a good model to use as a prediction

Then I tried to use trend dampen to improve its accuracy. The result improves but still at the state where the model is still not good enough for use as prediction model

ETS (M,M,M) Dampen

Average Error (ME) = -128640

Std. Dev.of Mean (RMSE) = 152964.6

Average Absolute Value (MAE) = 128639.7

Average Percentage Error (MPE) = -32.7399

Mean Absolute Scale Error (MASE) = 2.1554

1. 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.
   1. Describe the in-sample errors. Use at least RMSE and MASE when examining results
   2. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

## Since the data set contains seasonality, we need to use seasonality ARIMA Model ARIMA(p,d,q) (P,D,Q) m

## p is the number of autoregressive

d is the degree of differencing,

q is number of moving average term

PDQ is similar with pdq but refer to the seasonality differencing

p is the number of autoregressive

d is the degree of differencing,

q is number of moving average term

m is umber of period for each season, which is 12

From the excel calculation, we can find that we need 2 times of differencing in order to make the dataset become stationary. So d=2

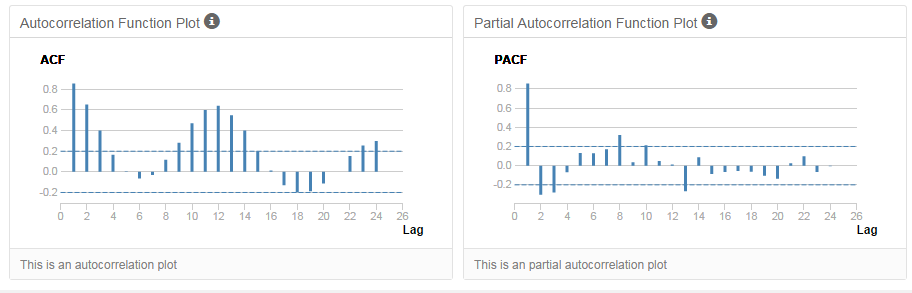


From ACF we can find there’s a negative autocorrelation after lag-1, so we use MA model as the best mode

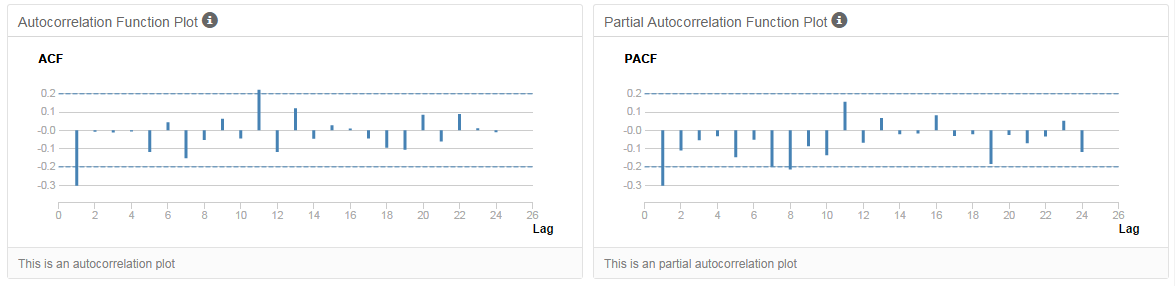
(AR 0, MA 1)

So the term would br (p,d,f,) -> (0,2,1)

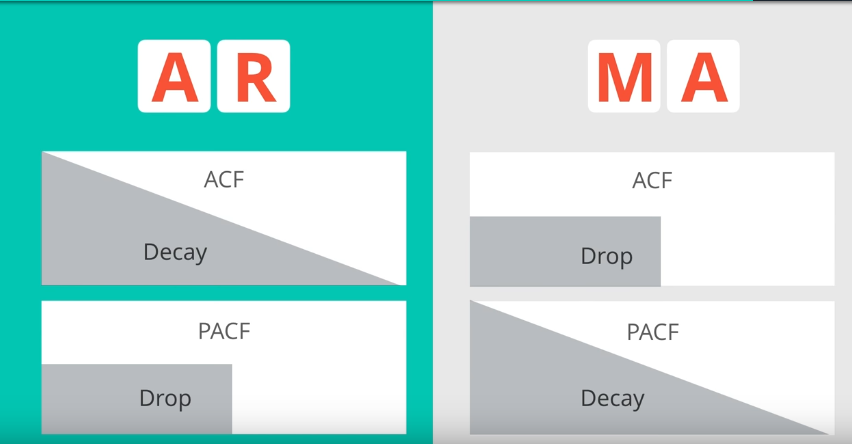
For Seasonal Difference we will take a look at ACF & PACF graph below



At the seasonal difference we can see that ACF is not stationary , so we need to create a seasonal difference. At Seasonal Difference -1 , the ACF graph is now become stationary



The outline for determining AR or MA can be determined from below figure

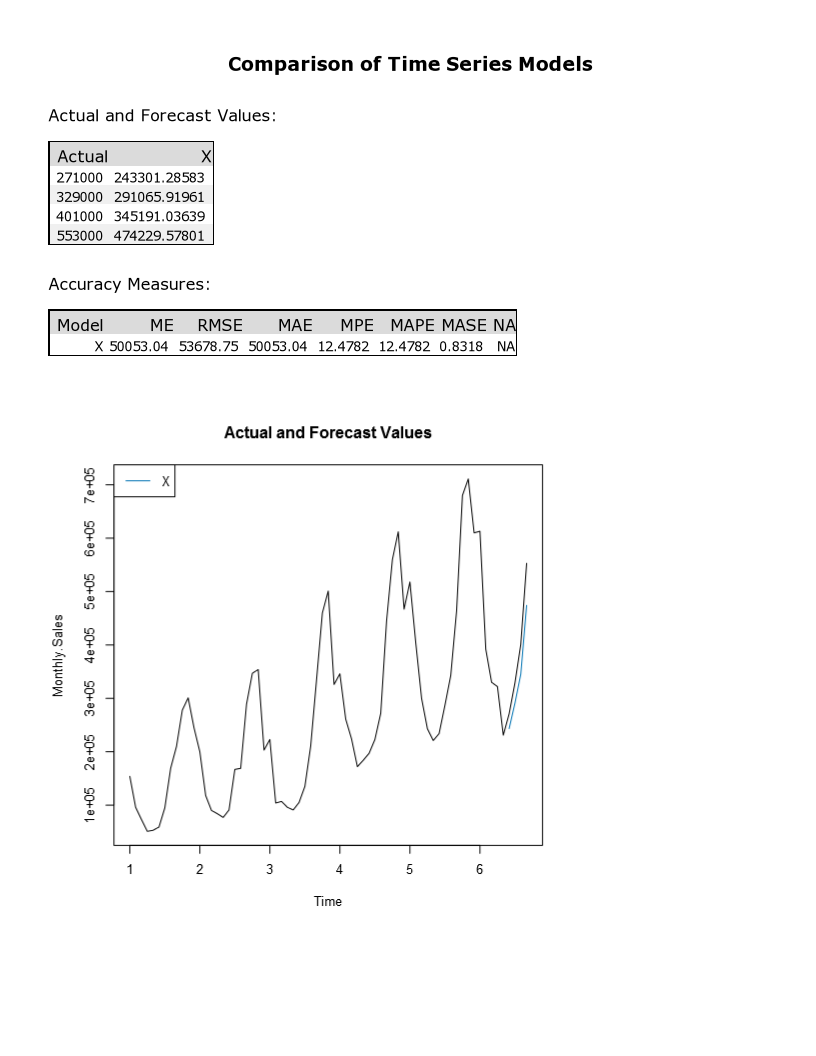


Since the Seasonal ACF decay, and PACF drops , we will use AR model (AR 1, MA 0), and difference 1

So the term would be (P,D,F) -> (1,1,0)

Since we know m =12

The ARIMA Model would be ARIMA (0,2,1) (1,1,0) 12

z

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
| ARIMA (0,2,1) (1,1,0) 12 | 50053.05 | 53678.75 | 50053.05 | 12.4782 | 12.4782 | 0.8318 |

Average Error (ME) = 50053.05

Std. Dev.of Mean (RMSE) = 53678.75

Average Absolute Value (MAE) = 50053.05

Average Percentage Error (MPE) = 12.4782

Mean Absolute Scale Error (MASE) = 0.8318

This is much better model than ETS model with Average percentage error (MPE) 12.48% and Mean Absolute Scale Error well below 1.0 which, is 0.8318

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
| ETS\_M\_M\_M\_DAMPEN TREND | -128640 | 152964.6 | 128639.7 | -32.7399 | 32.7399 | 2.1554 |
| ARIMA (0,2,1) (1,1,0) 12 | 50053.05 | 53678.75 | 50053.05 | 12.4782 | 12.4782 | 0.8318 |

**AIC ETS Trend Dampen Method**

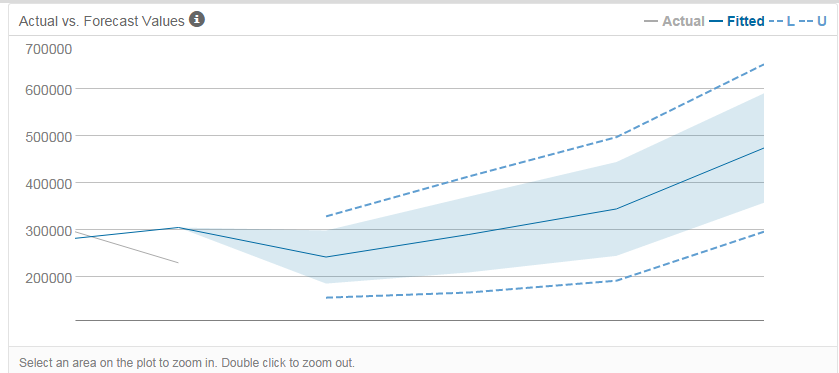


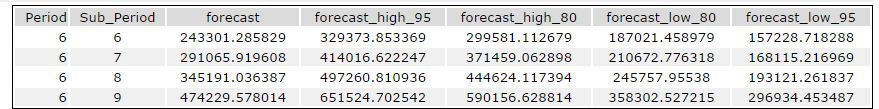
**AIC ARIMA(0,2,1)(1,1,0) 12 Method**



The best model would be model which has the most least amount or errors, least AIC , and MASE below 1.0. The best model would be ARIMA (0,2,1) (1,1,0)12 model

Forecast for the next 4 periods





Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/302/view) here. Reviewers will use this rubric to grade your project.