Project: Forecasting Sales

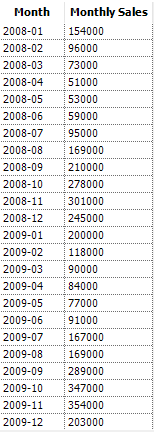
## Step 1: Plan Your Analysis

*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:*

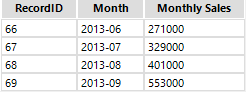
1. Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.

The dataset does meet the criteria. The data is ordered by month and year with its respective monthly sales. The data is continuous from 2008-2013. The spacing is equal between every two consecutive measurements such that after 2008-03 follows 2008-04. The dataset has no missing data point for either month or sales till the end month of 2013-09. Lastly, each time unit within the time interval has at most one data point. There is no time period with more than one data point.



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

Since I was tasked to forecast the sales for the next 4 months then it would be best for me to hold out the last 4 records of the dataset. The reason its best to use the last 4 records is because it’s feasible and does not cause disruptions. Also, the tasked forecast is followed by these last 4 records and so the model might give more importance on recent months than previous years.



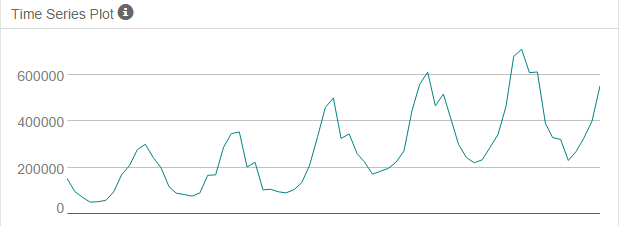
## Step 2: Determine Trend, Seasonal, and Error components

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

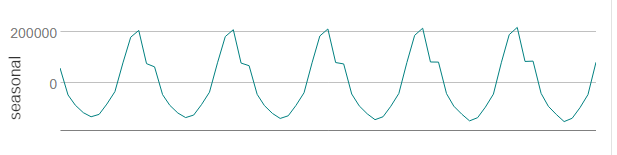
*Answer this question:*

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.

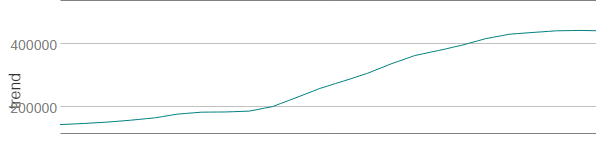
I simply made use of the TS Plot as it divides the data into the 3 components.



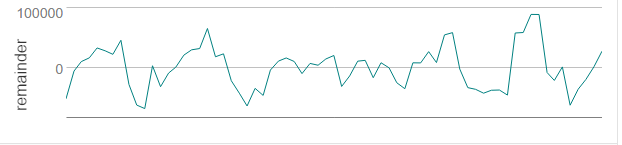
The seasonal component shows the reoccurring spikes in the sales data. These spikes suggest that any ARIMA models used in the analysis have seasonal differencing. When taking a closer look, one would see that there is slight increase in magnitude each time the sales spikes. This small increase suggests that the ETS model use a multiplicative term for seasonality.



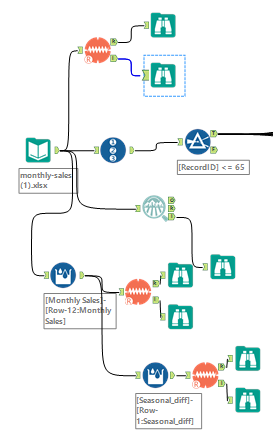
The trend component can tell us the gradual shift to higher or lower values over long period of time. In this case, we could see the general course is a linear uptrend. There is a noticeable increase in video game sales over the years however its not to the point that it is growing exponentially. This suggests the trend component of the ETS model use an additive term.

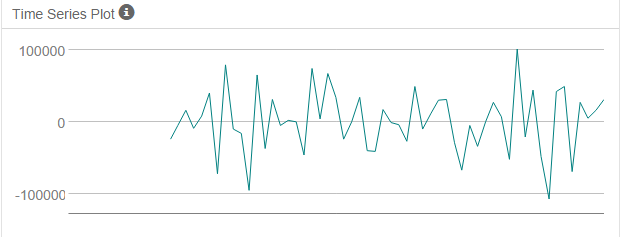


The error component is the remainder that is not accounted for in either trend or seasonality. The graph suggests that any ETS model used for the analysis will make use of a multiplicative term for error as it appears to have fluctuating peak sizes. The ARIMA models would make use of remainder to determine the lag but it would first require the data be stationary.



Since the plot was not stationary, I made a seasonal difference with the use of the multi-row tool then made the first difference. I plot the first difference and we could see that it is now stationary.





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:*

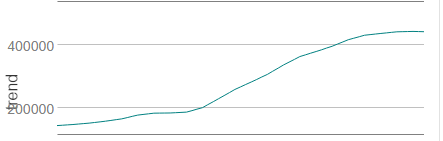
1. What are the model terms for ETS? Explain why you chose those terms.

The model terms are Error, Trend and Seasonal. There are 2 terms that I took into consideration for the ETS one is ETS(M,A,M) and ETS(M,M(d),M).

First, we would look at the plots. From the remainder plot, one can see that it fluctuates between large and small over time with varying sizes. Based from this one should most definitely make use of the Multiplicative term for error.

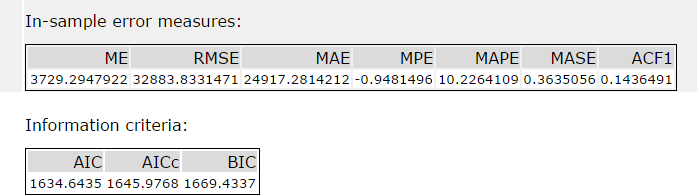
The seasonal plot at first glance appears to be constant however when looking closer, each time it peaks there is a slight increase. Therefore, we should apply multiplicative term for seasonal.

The trend plot was a bit complicated as it appears like it could be exponential but on closer inspection it isn’t. An additive term should be applied as the trend is linear.



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

ETS(M,A,M) has the following results:

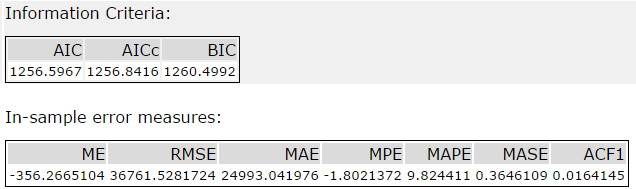


The results of ETS(M,A,M) show a lot of residue as seen from the RMSE. The ME shows the average difference of 3729.29 between the actual and forecasted which is not a huge gap. The MASE is decent as it is less than 1. The MAPE shows that there was a 10.23% error. Overall the model shows good potential.

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.

The model terms for ARIMA are Auto Regressive(p) Differencing(d) Moving Average(q) and the same goes for Seasonal ARIMA(P,D,Q). The terms I chose are ARIMA(0,1,1)(0,1,0)[12]. Since the dataset was not stationary, first difference was required to make the data stationary. The first difference also helped reduce significant lag and so there is no need for further differencing. Due to this I added only 1 point on differencing(I) for both. I added a point on MA(q) since there is a strong negative correlation at lag-1. Seasonal MA(Q) does not appear to decrease or increase. Looking at lag-12 of the ACF, it is negative (-0.118) and it reduces to negative lag (-0.011) at lag-24. The PACF on the other hand shows a small negative lag(-0.068) at lag-12 but then there is a negative increase in lag-24(-0.119). Also, lag 12 and 24 does not appear to have any significant correlation and so no MA term was given to seasonality(Q). After building the model, we could see from the resulting ACF and PACF plots that there is no more significant correlation and that means there would not be a need for adding any AR or MA terms. I then made an ARIMA with auto settings and it yield the same results.

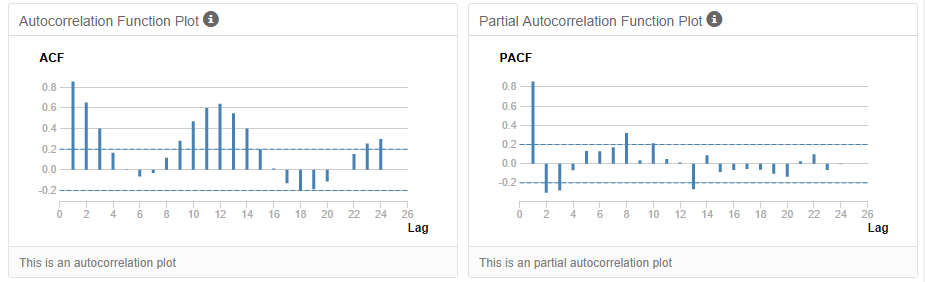
* 1. Describe the in-sample errors. Use at least RMSE and MASE when examining results



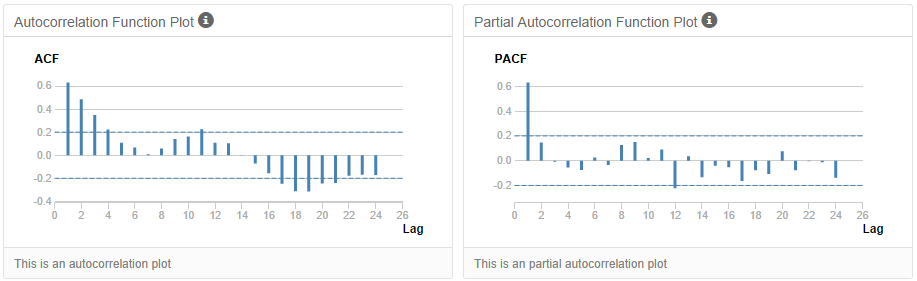
The Mean Error(ME), shows a big gap between the actual and forecasted values as it appears it is underestimating. From the RMSE one can see that there is a big difference as it has lots of residuals. The MASE is good as its value is lower than 1. The MAPE has a 9.82% error and is an improvement from the ETS. Overall the model has a good accuracy.

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

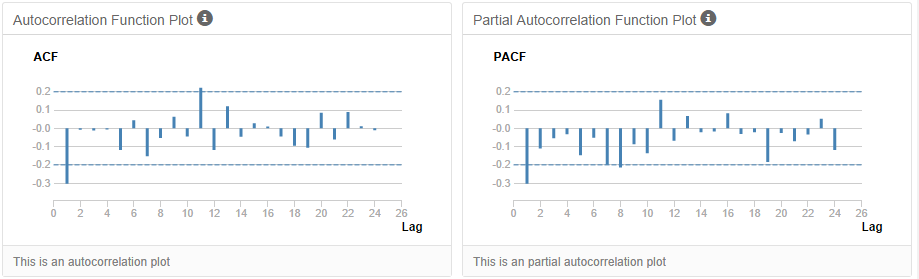
Time Series:



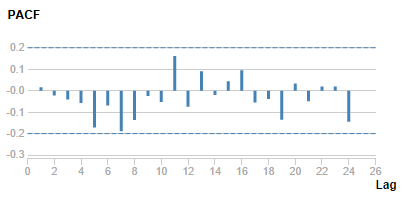
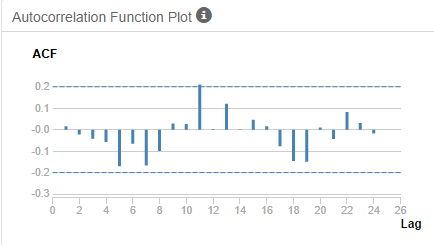
Seasonal Difference



First Difference:



Model Result:



## Step 4: Forecast

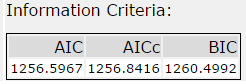
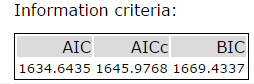
*Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)*

*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.

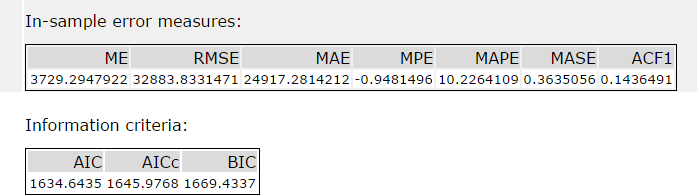
When comparing the information criteria of both models, one can see that the ARIMA model stands out because it has a lower AIC than the ETS model.

**ETS ARIMA**

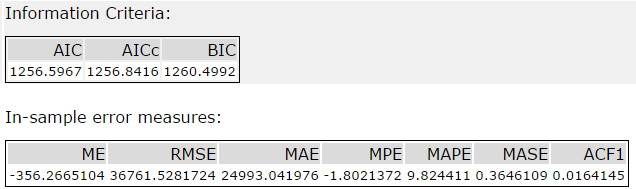


From the In-sample error measures, we could see that the ARIMA has a better ME as its value is closer to the actual(on average). Looking at the RMSE,the ETS has a lower value than the ARIMA. MPE shows a negative bias on both models but the ARIMA model still has it lower. MAPE shows the ARIMA having a lower percent error. Both ARIMA and ETS have the similar MASE but the ETS still has it slightly lower.

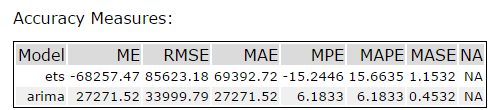
ETS Model:



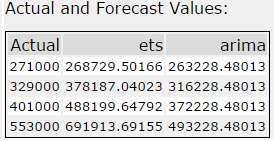
ARIMA Model:

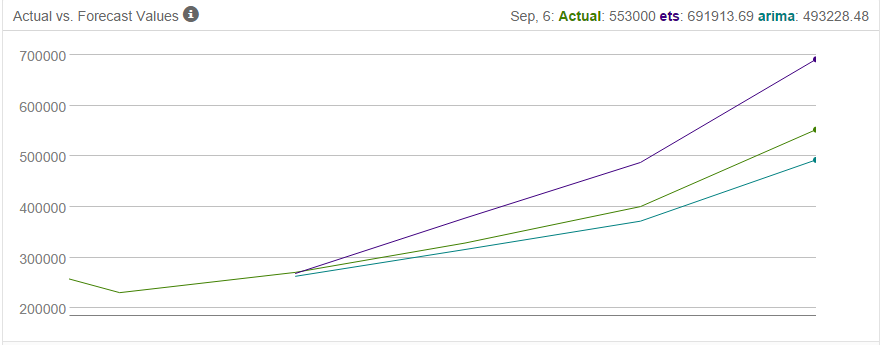


Now looking at the accuracy measures, ME shows the ETS model underestimating by -69257.47 while the ARIMA is overestimating by 27271.52. RMSE also shows a huge gap between the models, the ARIMA has less residue.The MASE shows ETS go beyond the value of 1 meaning while the ARIMA model maintains a value less than 1. MAPE confirms the difference in quality by showing the gap percent wise, as it notices a 15.66% error in ETS while only a 6.18% error in the ARIMA.



The Actual vs. Forecast Values and Chart makes it clear that ARIMA(0,1,1)(0,1,0)[12] is the better model as its estimates are closer to the actual value than ETS(M,A,M) for all 4 periods.





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

The forecast listed below is made by making use of the forecast tool with all the 69 records of data and applying the ARIMA(0,1,1)(0,1,0)[12] model.

