Project: Forecasting Video Game Sales using ETS and ARIMA Model

Chandan Mishra

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.

   Four Key Characteristics of a time series data are:
   - Series over a continuous time interval
   - Sequential measurements across the interval
   - Equal spacing between every two consecutive measurements
   - Each time unit within the time interval has at most one data point

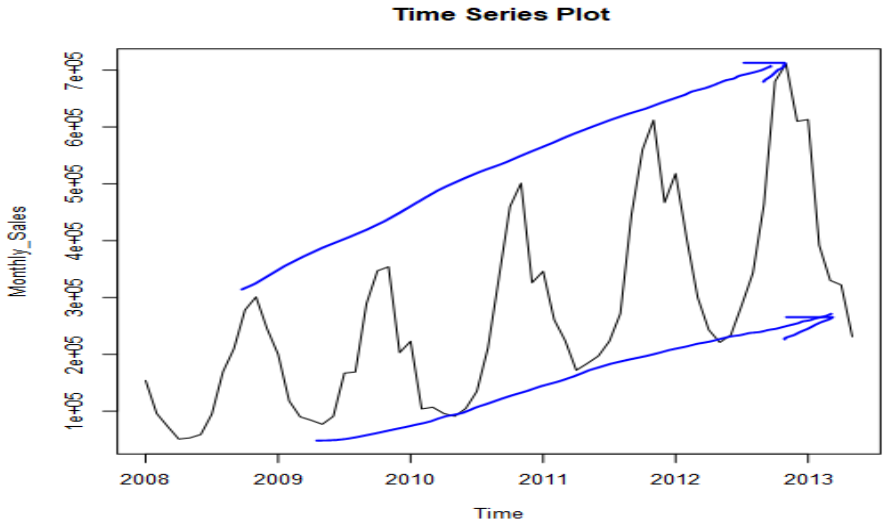
2. Which records should be used as the holdout sample?

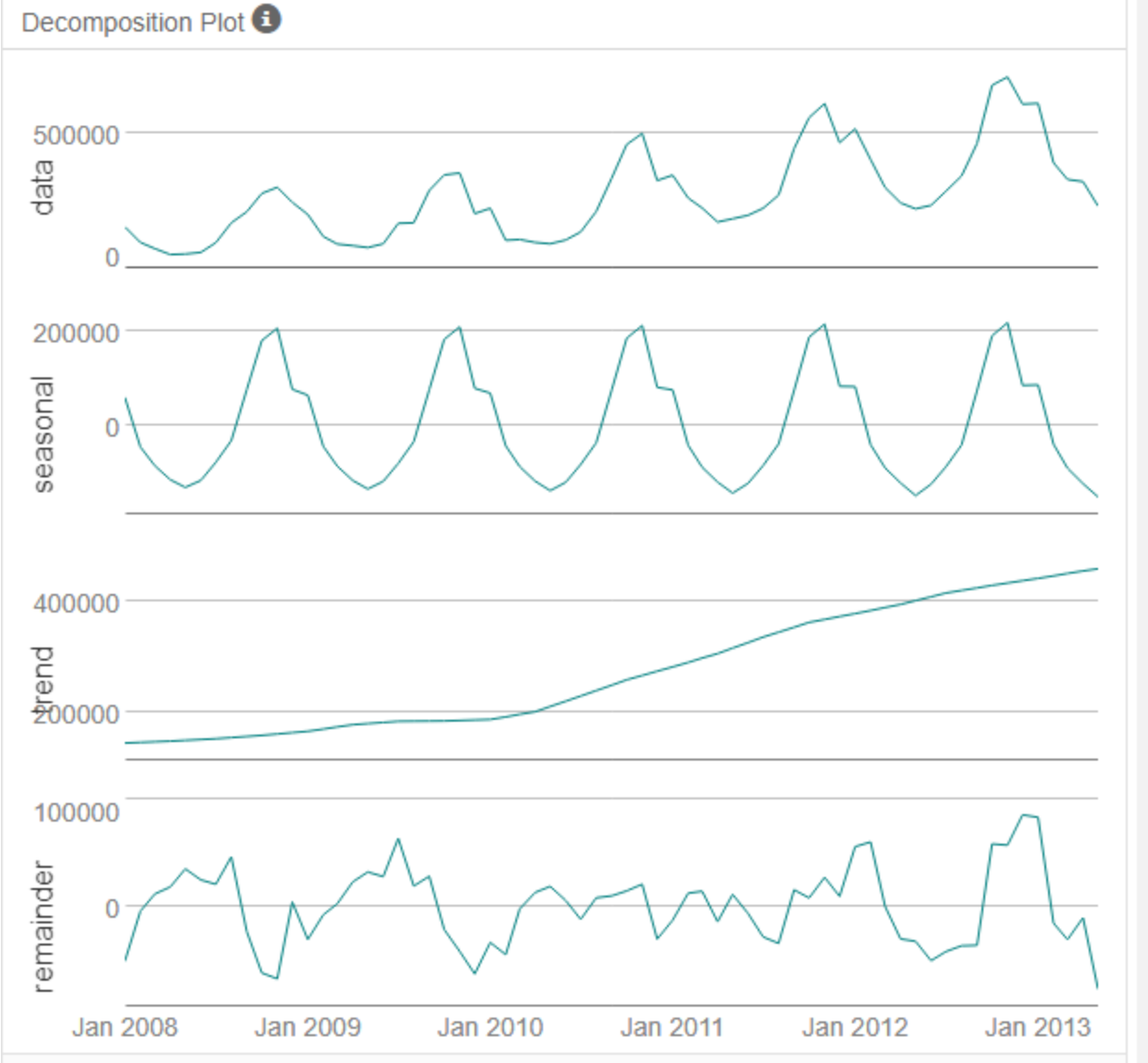
   There are a total of 69 records representing monthly sales starting from 2008-01 to 2013-09. I have filtered out last 4 records i.e. from 2013-06 to 2013-09 as a holdout sample to check the accuracy of the forecast.

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 initial time series shows an upward rising trend with spikes occurring in sales each year. This shows we have seasonality but no signs of cyclicity.





The trend seasonality and error component confirms the time series model.

Trend – Upward trending

Seasonal – Regular spikes in sales each year with magnitude of sale increasing. Since we have seasonality in our data with magnitude of sales increasing every year, we will use the following:

ARIMA model with seasonal differencing

ETS model with Multiplicative method in seasonal component

Error – Fluctuations but not consistent in magnitude between large and small errors so we will use multiplicative method in error component

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

    Trend - Trend line exhibits linear behavior, so we will use an additive method.
    Seasonality - The seasonality changes in magnitude each year so a multiplicative method is necessary.
    Error - The error changes in magnitude as the series goes along so a multiplicative method will be used.
    This leaves us with an ETS(M, A, M) model .

    Error Terms:

    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

    Two key components to look at are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models.
    We can see that our variance is about 33000 units around the mean.
    The MASE shows a strong forecast at .36 with its value falling well below the generic 1.00, the commonly accepted MASE threshold for model accuracy.

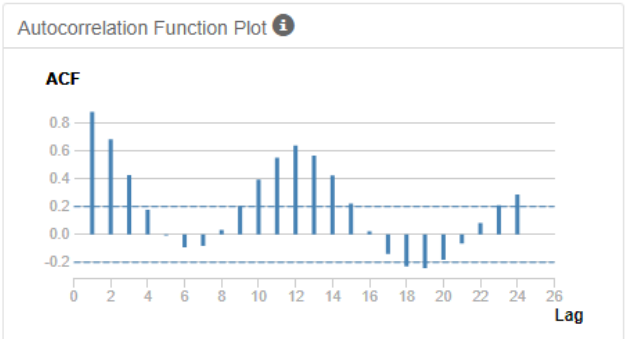
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
    a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

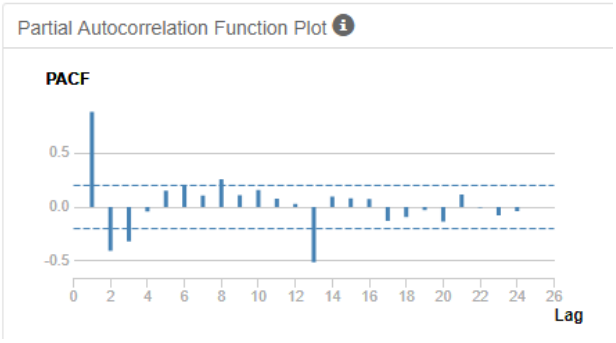
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

Since there is seasonality in the data, we will use ARIMA (p,d,q)(P,D,Q)S model for forecasting.

Time series ACF and PACF plot:

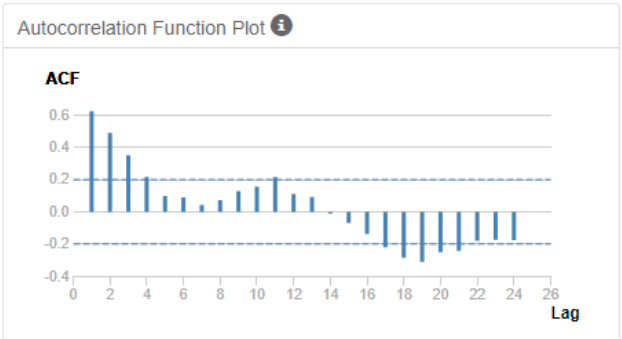
The ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high, I will need to seasonally difference the series.

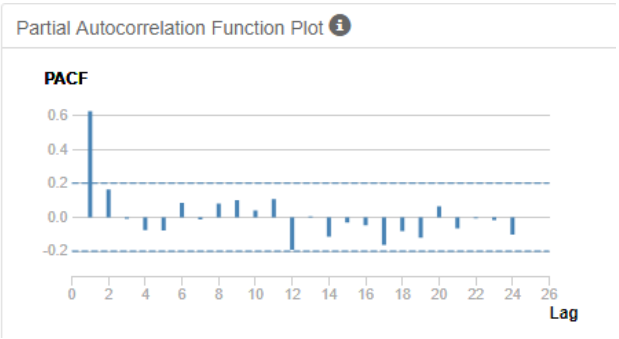




Seasonal Difference ACF and PACF plot:

The seasonal difference presents similar ACF and PACF results as the initial plots without differencing, only slightly less correlated. In order to remove correlation, we will need to difference further.

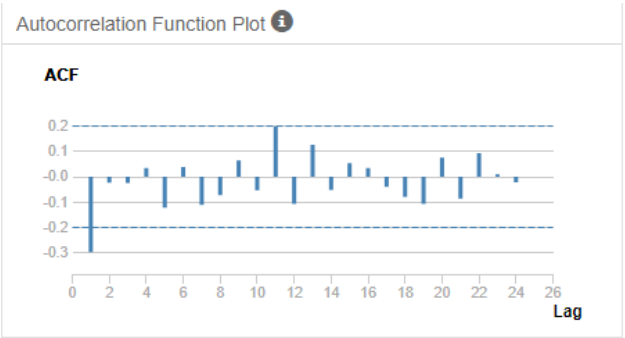


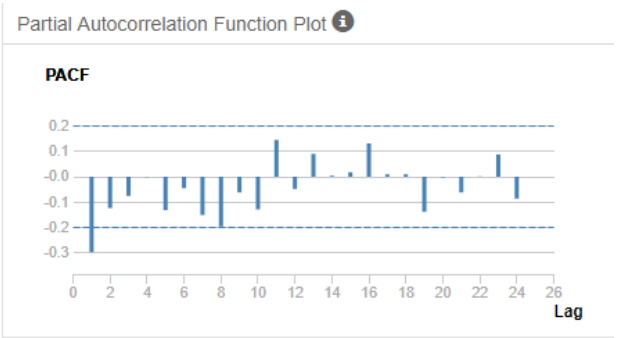


Seasonal First Difference ACF and PACF plot:

The seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be d(1) and D(1).

The ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model since there is only 1 significant lag. The seasonal lags (lag 12, 24, etc.) in the ACF and PACF do not have any significant correlation so there will be no need for seasonal autoregressive or moving average terms.





Therefore the model terms for my ARIMA model are:

ARIMA(0, 1, 1)(0, 1, 0)[12]

Error Terms:

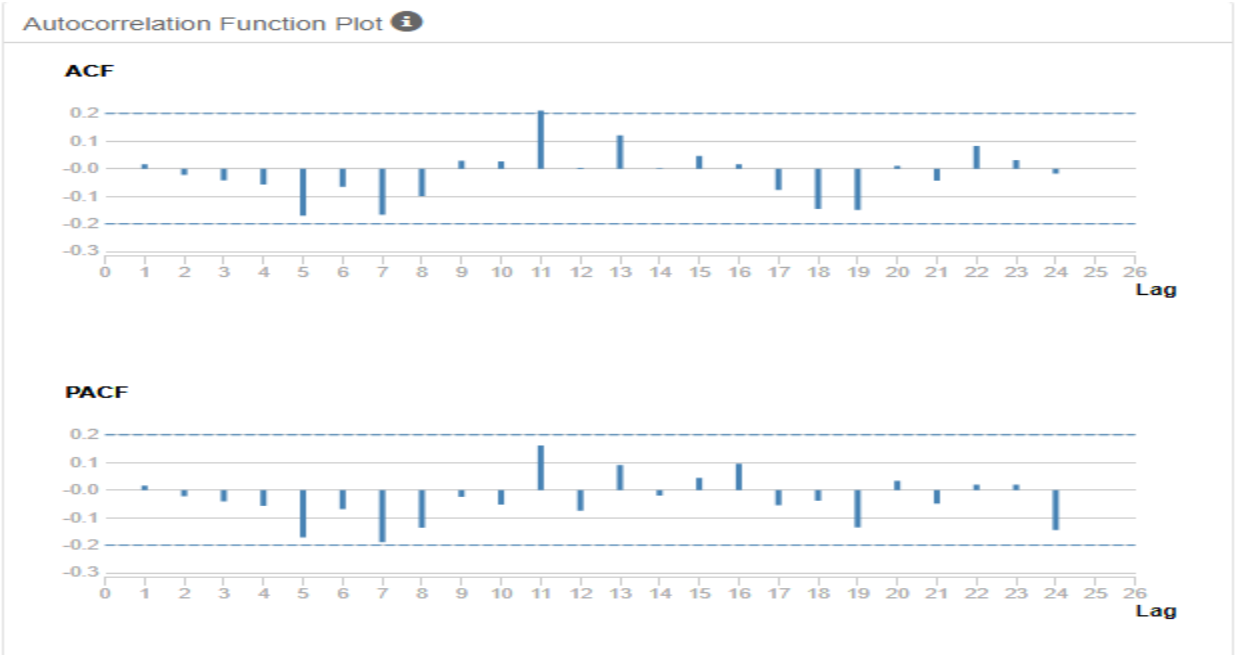
The ACF and PACF results for the ARIMA(0, 1, 1)(0, 1, 0)[12] model shows no significantly correlated lags suggesting no need for adding additional AR() or MA() terms.

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



Two key components to look at are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models. We can see that our variance is about 37000 units around the mean.
The MASE shows a strong forecast at .36 with its value falling well below the generic 1.00, the commonly accepted MASE threshold for model accuracy.

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 fitting a forecasting model we can use a series of identifiers that help us choose the best model.
When comparing the two in-sample error measures we used, the RMSE and MASE, we see very similar results. The ETS model does have a narrower standard deviation but only by a few thousand units.
Further investigation shows that the MAPE and ME of the ARIMA model are lower than the

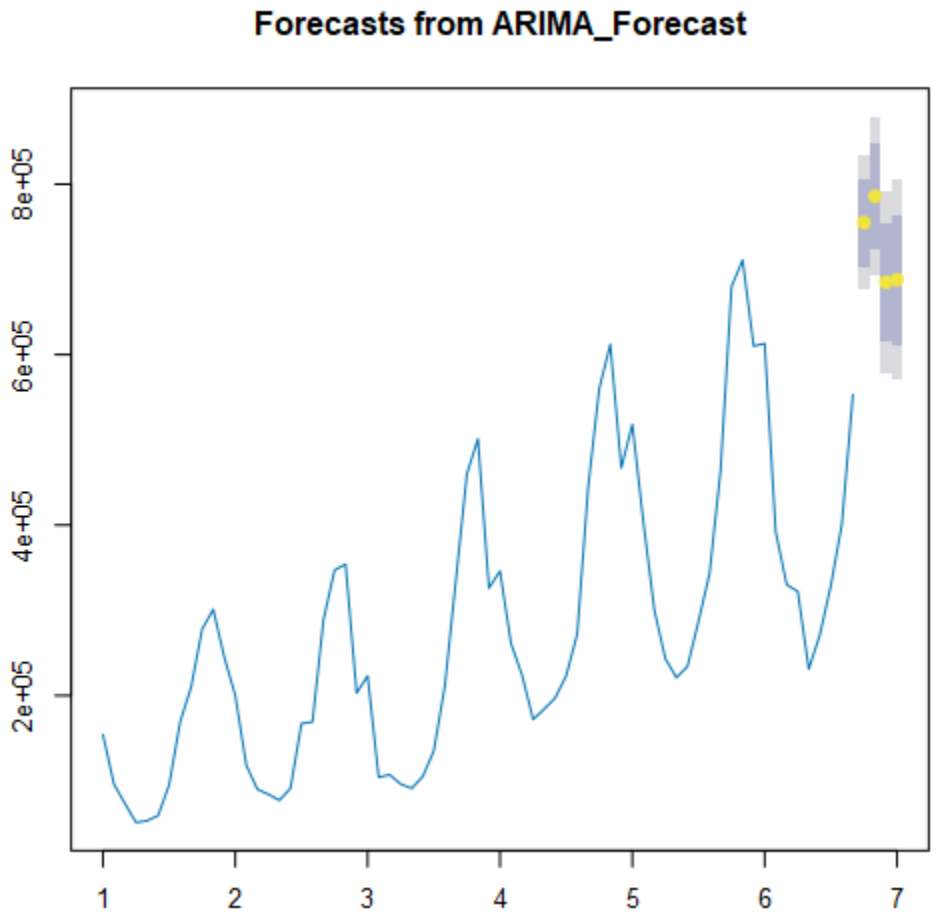
ETS. This suggests that, on average, the ARIMA model misses its forecast by a lesser amount. When looking at the model's ability to predict the holdout sample, we see that the ARIMA model has better predictive qualities in just about every metric.

Accuracy Measures:

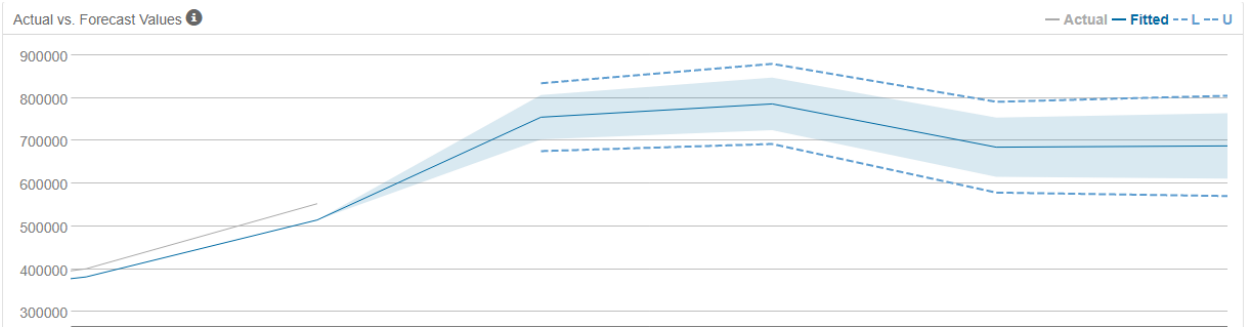
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066
ARIMA	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

For forecasting, we will use ARIMA model.

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



Period	Sub_Period	Final_Forecast	Final_Forecast_high_95	Final_Forecast_high_80	Final_Forecast_low_80	Final_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


Actual vs. Forecast Values
Actual Fitted L U
900000
800000
700000
600000
500000
400000
300000