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

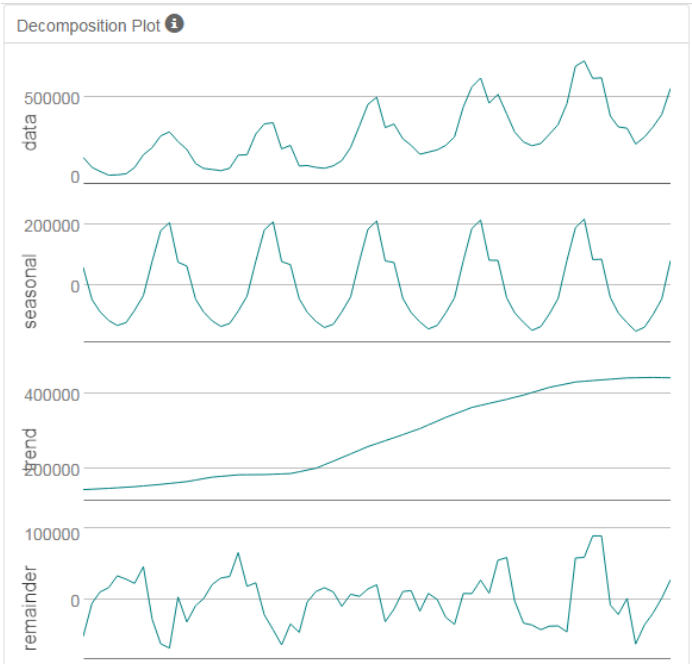
In this case, we want to forecast the sales of video game for the next 4 months, therefore we will take the 4 months of latest sales which is : Year 2013 month 6,7,8,9 to predict sales of year 2013 month 10,11,12 & year 2014 month 1

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)
Linear trend (A)
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.
    a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

    The model term for ETS is ETS (M,A,M)
    Because it has increasing error, linear trend, and increasing seasonality

ETS (M,A,M) Result

Summary of Time Series Exponential Smoothing Model ETS_M_A_M_

Method:
ETS(M,A,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056	0.1436491

AIC	AICc	BIC
1634.6435	1645.9768	1669.4337

ETS (M,A,M) Dampen Result

Summary of Time Series Exponential Smoothing Model ETS_M_A_M__Dampen

Method:
ETS(M,Ad,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5572.6821018	33302.042717	25725.4553044	0.1900065	10.54361	0.3752957	0.100576

AIC	AICc	BIC
1636.5328	1649.554	1673.4973

Based on the ETS result we can see that ETS model without dampen actually gives better model than using trend dampen with overall less errors. For example : RMSE (no-dampen 32,883 < dampen 33,302), MAPE (no-dampen 10.22% < dampen 10.54%), and MASE (no dampen 0.3635 < 0.3752)

ETS (M,A,M) TS Comparison Result

Record	Report
1	**Comparison of Time Series Models**
2	Actual and Forecast Values:

Actual	ETS_M_A_M__
271000	268729.50166
329000	378187.04023
401000	488199.64792
553000	691913.69155

3	Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_M_A_M__	-68257.47	85623.18	69392.72	-15.2446	15.6635	1.1532	NA

4	

When we test the ETS model without dampen against its holdout sample, the error was bit higher than the model. For example its Mean Absolute Percentage Error(MAPE) increase from before 10.22 to 15.66% and its MASE increase from 0.3635 to 1.1532 when tested against its holdout sample. While the model is very good at first, it's not good enough when we tested it against the holdout sample since the MASE is still above 1.0. Therefore, we still need to compare it with ARIMA 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.

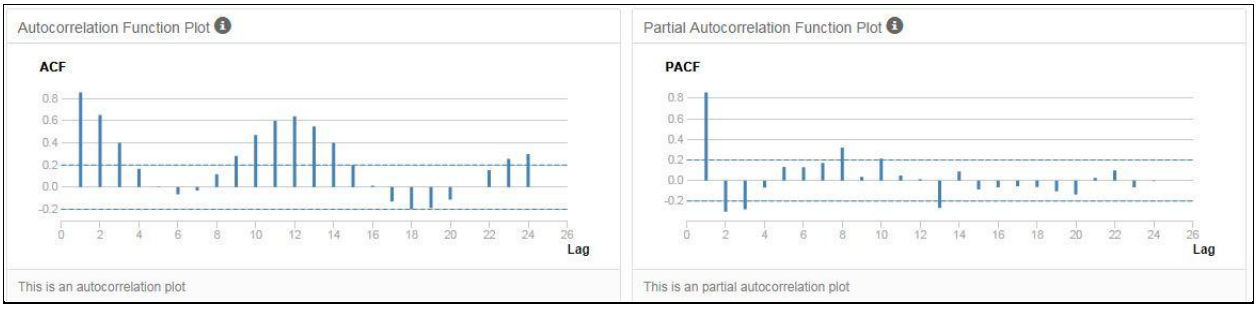
Since the data set contains seasonality, we need to use seasonality ARIMA Model ARIMA(p,d,q) (P,D,Q) m

p is the number of autoregressive

d is the degree of differencing,
q is number of moving average term

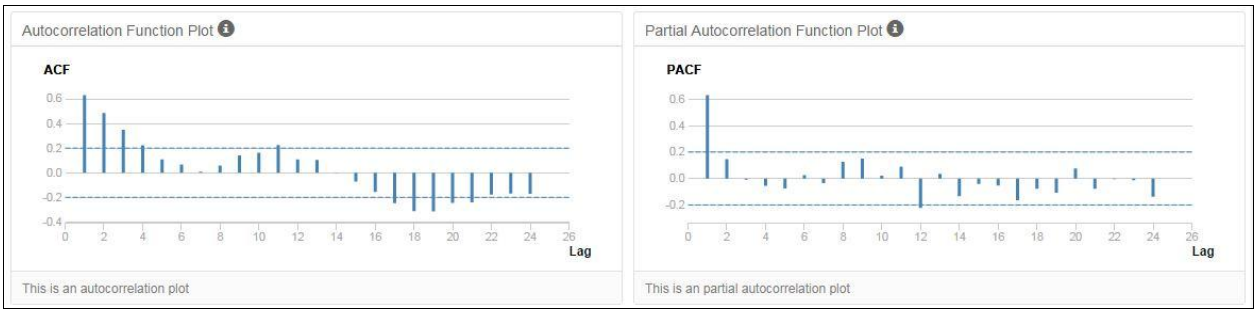
PDQ is similar with pdq but refer to the seasonality differencing
P is the number of autoregressive
D is the degree of differencing,
Q is number of moving average term

Time Series ACF & PACF



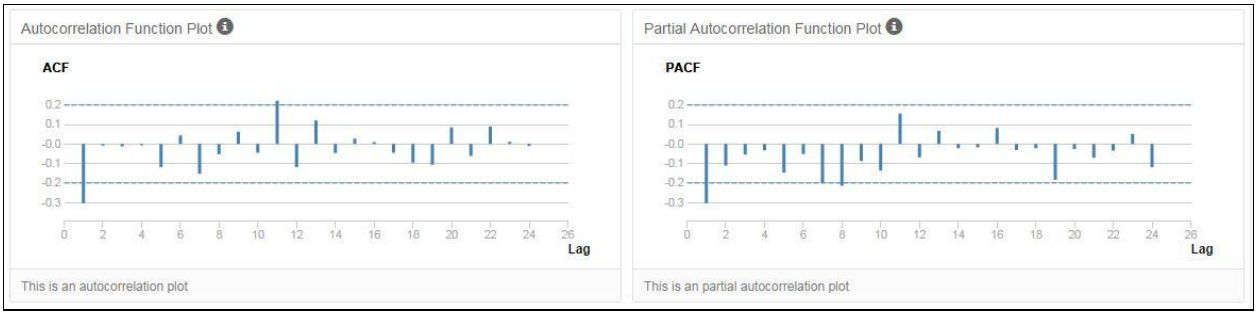
We should be able to see that the ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high we will need to seasonally difference the series

Seasonal Difference ACF & PACF



We can see that 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 & PACF



we can see see that 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. So This means we have found ARIMA (0,1,1) (P,1,Q)

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. This

means the P & Q would be zero. And since the know that the forecast is monthly, we found that m would be 12.

so the suggested model would be **ARIMA (0,1,1) (0,1,0) 12**

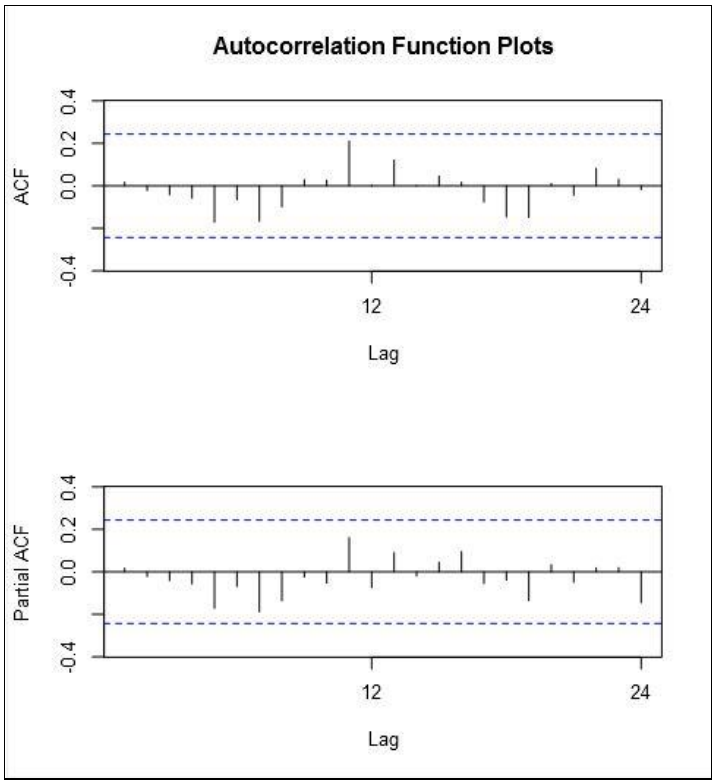
ARIMA (0,1,1) (0,1,0) 12 result

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



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

The ARIMA model shows good result with Mean Error Average Percentage Error (MPE) = -1.802% and Mean Absolute Scale Error (MASE) well below 1.0 which is 0.3646 , which show very good model, but we still need to test it against hold out sample to test its validity using TS comparison technique.

ARIMA (0,1,1) (0,1,0) 12 TS comparison result

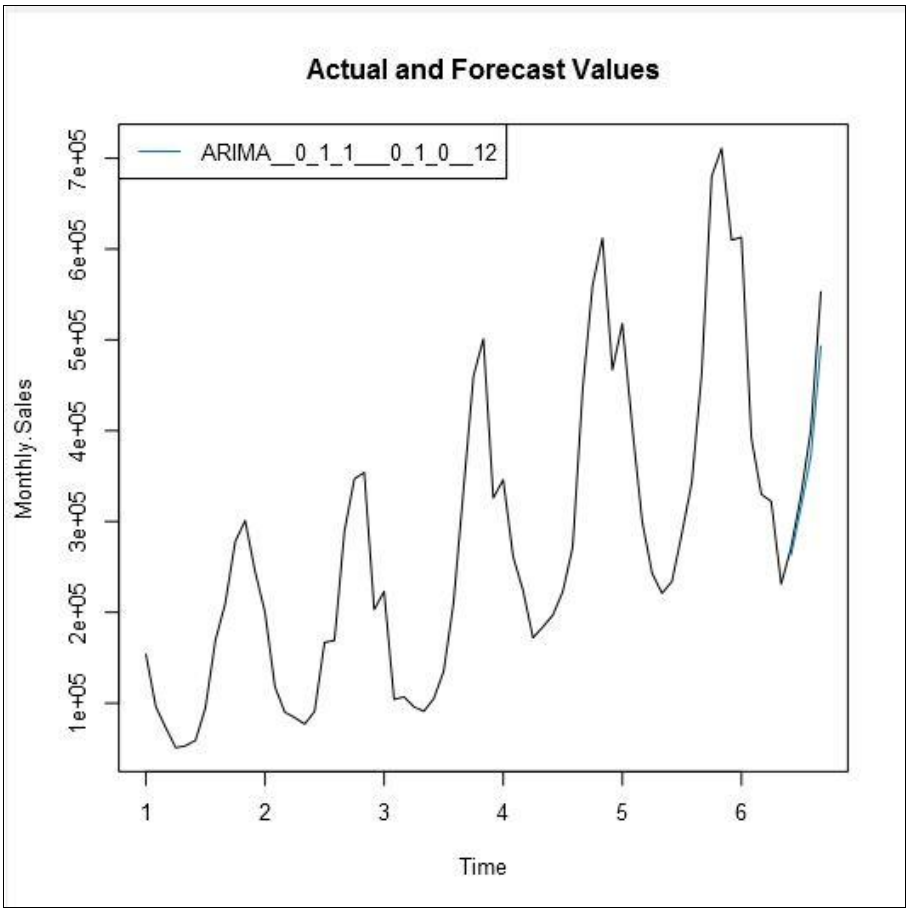
Comparison of Time Series Models

Actual and Forecast Values:

Actual	X
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

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



ARIMA (0,1,1) (0,1,0) 12 TS comparison result still shows very good result with Average Percentage Error (MPE) = 6.138% and Mean Absolute Scale Error (MASE) well below 1.0 which is 0.4532 when the model tested with the holdout sample

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.

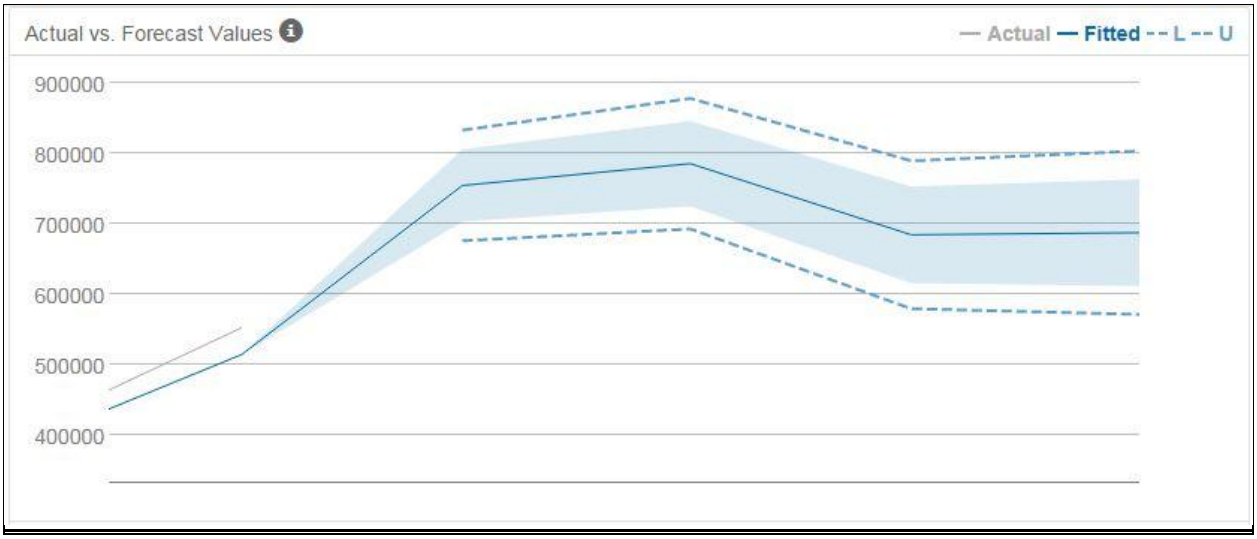
TS-COMPARISON against holdout sample for
ARIMA (0,1,1) (0,1,0) 12 and ETS (M,A,M) Dampen model

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_M_A_M__	-68257.47	85623.17	69392.71	-15.2446	15.6635	1.1532
ARIMA(0,1,1) (0,1,0) 12	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

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

- Average Error (ME) = 27271.52
- Std. Dev. of Mean (RMSE) = 33999.79
- Average Absolute Value (MAE) = 27271.52
- Average Percentage Error (MPE) = 6.1833 %
- Mean Absolute Percentage Error (MASE) = 6.1833%
- Mean Absolute Scale Error (MASE) = 0.4532

Forecast for the next 4 periods



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
6	10	754854.460048	833335.856133	806170.686679	703538.233418	676373.063963
6	11	785854.460048	878538.837645	846457.517118	725251.402978	693170.082452
6	12	684854.460048	789837.592834	753499.24089	616209.679206	579871.327263
7	1	687854.460048	803839.469806	763692.981576	612015.938521	571869.450291

The 4 months forecast would be period 6 sub period 10,11,12, and period 6 sub period 1 with 95% and 80% confidence interval. ☺

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.