Table of Contents

Project: Forecasting Sales

Business Issue

You recently started working for a company as a supply chain analyst that creates and sells video games. Many businesses have to be on point when it comes to ordering supplies to meet the demand of its customers. An overestimation of demand leads to bloated inventory and high costs. Underestimating demand means many valued customers won't get the products they want. Your manager has tasked you to forecast monthly sales data in order to synchronize supply with demand, aid in decision making that will help build a competitive infrastructure and measure company performance. You, the supply chain analyst, are assigned to help your manager run the numbers through a time series forecasting model. You've been asked to provide a forecast for the next 4 months of sales and report your findings.

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.

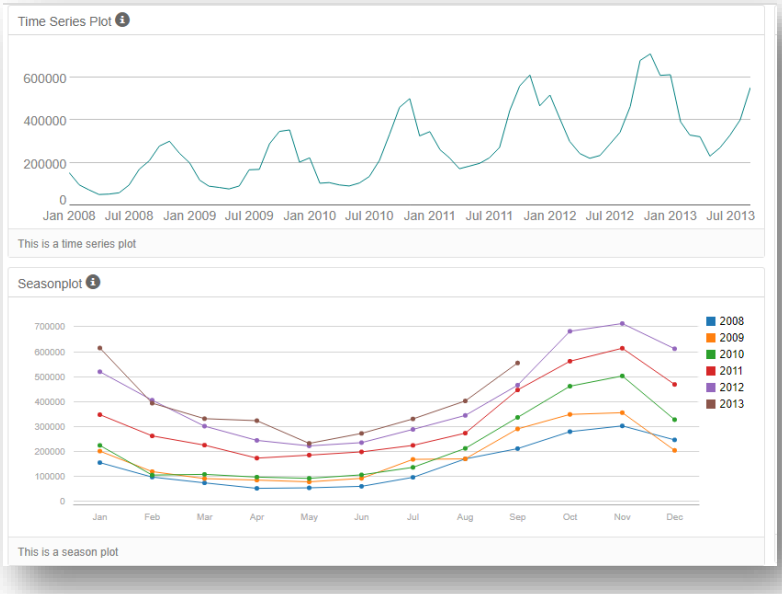
Answer the following questions to help you plan out your analysis:

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 exhibits all the 4 characteristics of a time series:
   - We have a continuous time interval where each observation is ordered chronologically. More specifically, it comprises monthly video game sales from January 2008 to September 2013. However, the dimensions month and year have been aggregated into the same column so we will need to split it into two separate columns, "Year" and "Month".
   - Measurements have been taken across sequential and equal intervals.
   - There is equal spacing between every two consecutive measurements.
   - Each time unit (month) corresponds to one data point (monthly sales).
2. *Which records should be used as the holdout sample?*
   Since the models are going to be forecasting 4 periods, we only need to use a holdout sample of 4 periods, namely Record-Id 66 to Record-Id 69. This will allow us to test the model accuracy against the 4 most recent periods in our holdout sample.

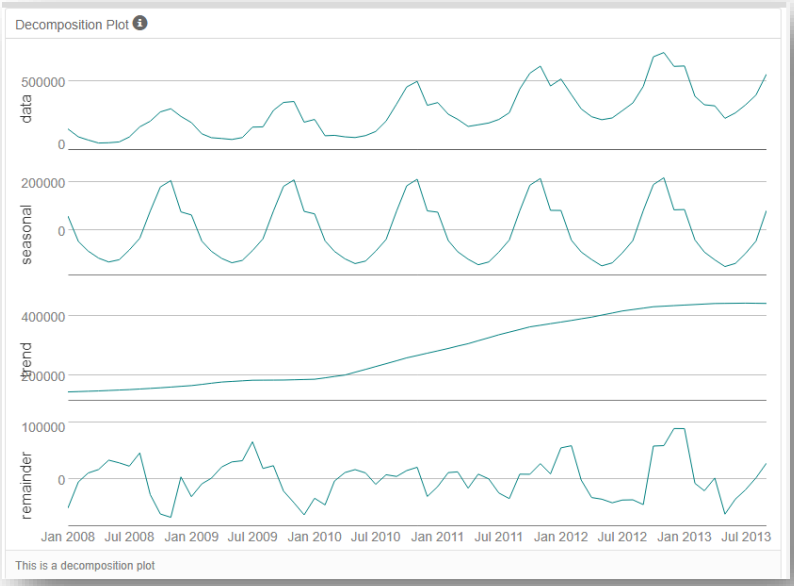
Step 2: Determine Trend, Seasonal, and Error components

Graph the data set and decompose the time series into its three main components: trend, seasonality, and error.
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.*
   To gain a better understanding of our time series we should take a closer look at the TS plot, at the seasonal plot as well as at the decomposition plot. When looking at the time series and seasonal plots, it looks like sales are subject to seasonality with clearly defined peaks and valleys:



The following decomposition plot confirms these seasonal patterns: the graph shows a repeating pattern at fixed intervals of time within a 1-year period. Another noticeable characteristic is that sales have been increasing over time, this could point to either an additive or multiplicative behavior. At first sight, it looks multiplicative since we can detect a slight increase in magnitude over time.
The trend component exhibits an upward, linear trend, suggesting an additive behavior.
Lastly, the remainder illustrates changes in variability as the time series moves along, thus, we will apply it multiplicatively.



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.
Answer these questions:

1. *What are the model terms for ETS? Explain why you chose those terms. Describe the in-sample errors. Use at least RMSE and MASE when examining results.*
   The analysis of the decomposition plot reveals an ETS(M,A,M) model. As mentioned, the error component indicates periods of higher and lower variability (multiplicative term), the trend component follows a linear pattern (additive term) while the seasonality element captures variations in magnitude (multiplicative term).
   Let's now examine some of the in-sample error measures:

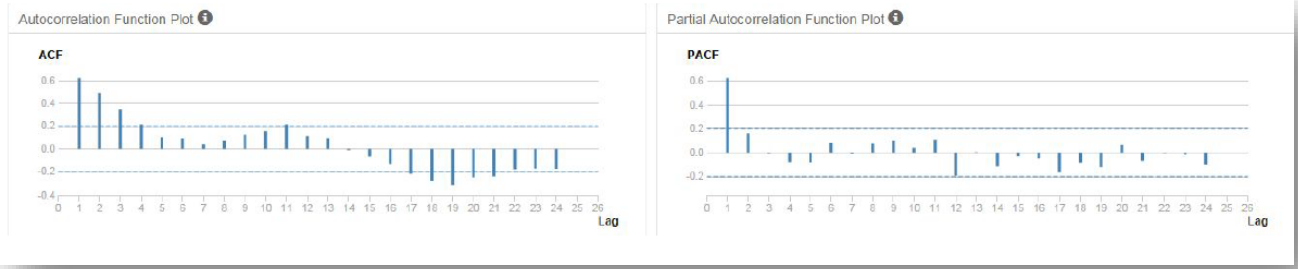
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3243.4703524	31474.3668886	24188.2167878	-0.572395	10.3052041	0.3528697	0.0087402

   The Mean Absolute Scaled Error shows the relative reduction in error as compared to a naïve mode. When this value is below the threshold of 1.00, it suggests accuracy in forecasting. The Root Mean Squared Error, which represents the sample standard deviation of the differences between forecasted values and observed values, lies at about 31000 units around the mean.
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 decomposition plot provides evidence that there are seasonal patterns in our time series, therefore, we will need to make the time series stationary in order to build our ARIMA(p,d,q)(P,D,Q) model.

After applying the seasonal difference (D1) as shown below, we notice that the ACF and PACF plot results still show high serial correlation, therefore we will need to apply an additional differencing.

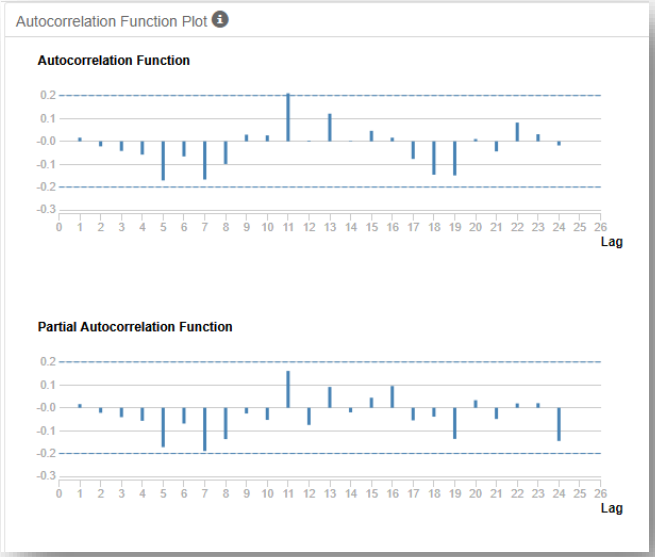


By applying a seasonal differencing and a first differencing (d1), we arrive at a stationary timeseries, as illustrated by the following ACF and PACF plots.

When selecting seasonal terms, we already know that the differencing term is D1. Both the ACF and PACF plots show that there is no significant correlation at Lag-12 and Lag-24, consequently we should account for MA0(Q) and AR0(P). As far as the non-seasonal terms of the model are concerned, we have already performed a first differencing (d1). By inspecting the ACF plot we notice a negative autocorrelation at Lag-1 which is also reflected by the PACF plot displaying a gradual decay, suggesting AR0(p) and MA1(q) terms.

Hence, we can denote our model as **ARIMA(0, 1, 1)(0, 1, 0)[12]**.

The ACF and PACF plots resulting from the above-mentioned model show no significantly correlated lags, suggesting no need to refine our model with additional AR() or MA() terms.



The in-sample errors of our ARIMA model are represented in the following table:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

When looking at the Mean Absolute Scaled Error, we can observe a value that is below the threshold of 1.00, so we may conclude that this model is accurate enough. The Root Mean Squared Error shows a variance of about 37000 units around the mean. We will compare these in-sample errors with the in-sample errors of the ETS model to identify the best fit.

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.
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 in-sample errors of both models side by side, we can observe similar results as far as the Mean Absolute Scaled Error and Root Mean Squared Error values are concerned. Additionally, the ARIMA model scores better than the ETS Model in terms of Margin Error and Mean Absolute Percentage Error – in this respect it shows a higher accuracy in forecasting values.

ARIMA Model						
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

ETS Model						
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3243.4703524	31474.3668886	24188.2167878	-0.572395	10.3052041	0.3528697	0.0087402

By testing each model accuracy against the holdout sample, we conclude that the ARIMA Model does a better job in predicting our target variable since it outperforms the ETS Model on all fronts:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
MAdM	-33,469.6095	53,828.4835	41,542.755	-6.3476	9.3266	0.6904
ARIMA	27,271.5199	33,999.7911	27,271.5199	6.1833	6.1833	0.4532

For this reason, we are going to choose the ARIMA Model to forecast sales figures.

2. *What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.*
   The forecast for our next four periods, using 95% and 80% as C.I., is displayed in the following table and graphs:

Period	Sub Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2013	10	754854.46	834046.22	806635.17	703073.75	675662.70
2013	11	785854.46	879377.75	847006.05	724702.87	692331.17
2013	12	684854.46	790787.83	754120.57	615588.35	578921.09
2014	1	687854.46	804889.29	764379.42	611329.50	570819.63