Project: Forecasting Sales

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

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 data does meet the criteria of a time series dataset.
   - The data covers a continuous time interval.
   - Of sequential measurements across that interval.
   - It is equally spaced between every 2 consecutive measures.
   - Each time unit within the interval has one data point.

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

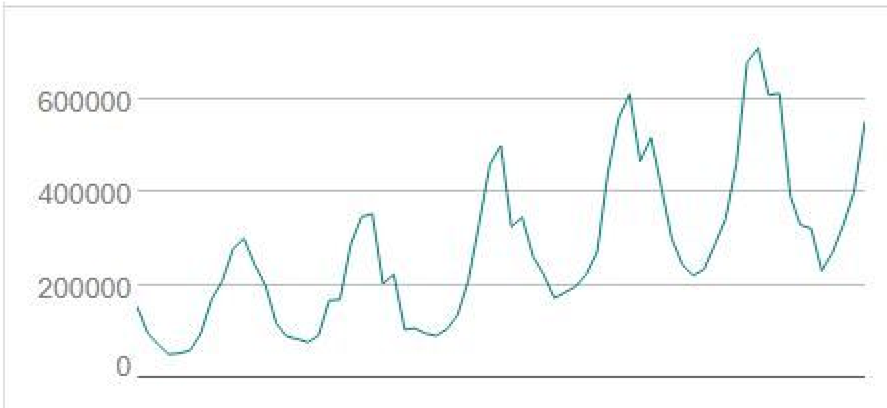
   We are forecasting for the next 4 months, so the holdout sample should be the last four months of data.

Step 2: Determine Trend, Seasonal, and Error components

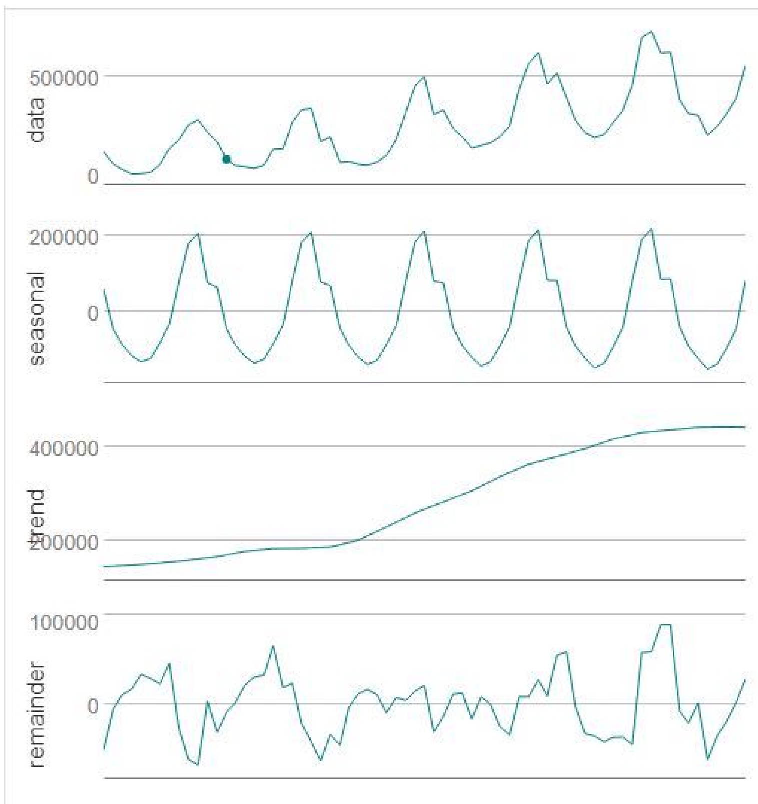
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error.

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 findings of the time series plot shows an upward rising trend with a regularly occurring spike in sales each year reported at the end of the year. This pattern shows that we have seasonality in our time series. There are no patterns within the series suggesting cyclicity.



The decomposition plot shows our time series broken down into its three components: trend, seasonal and the error. Each of these components makes up our time series and helps us confirm what we saw in our initial time series plot.



The trend of the time series is increasing.

The error (or remainder) is fluctuating between large and small errors over time.

The seasonal portion shows that the regularly occurring spike in sales each year changes in magnitude, ever so slightly. Having seasonality suggests that any ARIMA models used for analysis will need seasonal differencing.

Looking at the seasonal peaks and valleys, we can seasonality increases in volume each seasonal period.

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.

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 terms for the ETS model are :

Our trend line exhibits linear behavior so we will use an additive method.
The seasonality changes in magnitude each year so a multiplicative method is necessary
Error - fluctuates between large and small errors so a multiplicative model should be used

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

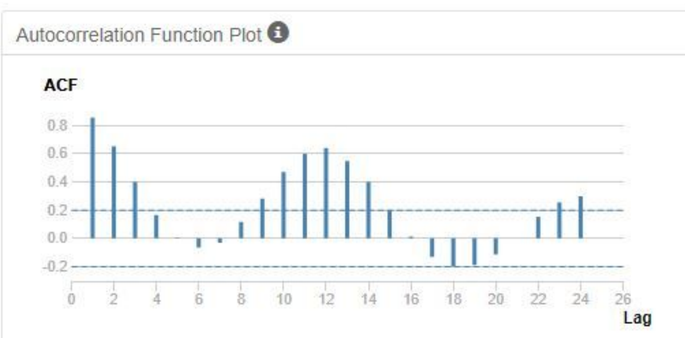
Two key components to look at are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models. We can see that our variance is about 33000 units around the mean.
The MASE shows a fairly strong forecast at .36 with its value falling well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

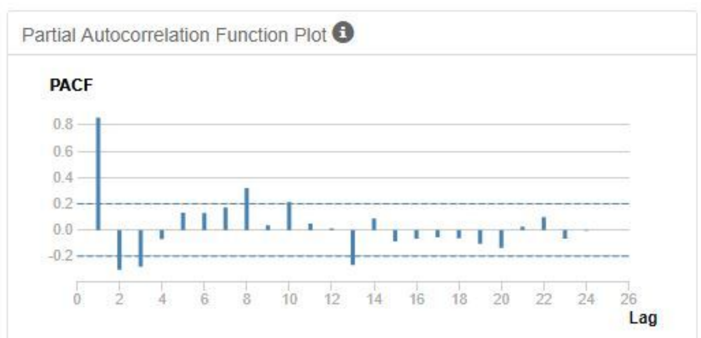
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Autocorrelation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
    a. Describe the in-sample errors. 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 are seasonal components found in the time series I will use an ARIMA(p, d, q)(P, D, Q)S model for forecasting.

<u>Time Series ACF and PACF:</u>

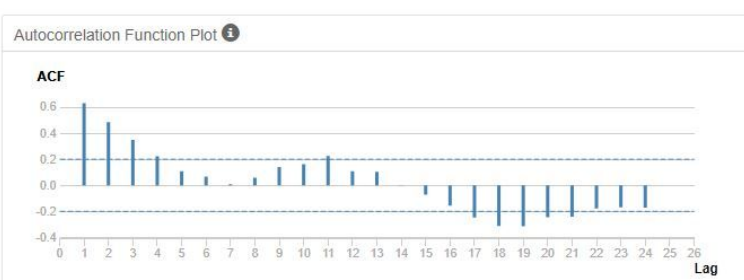
The ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high I will need to seasonally difference the series.

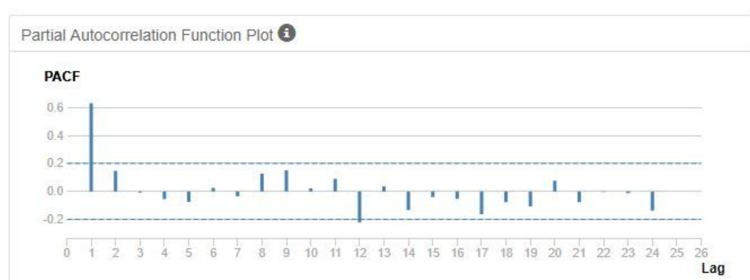




<u>Seasonal Difference ACF and PACF:</u>

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.

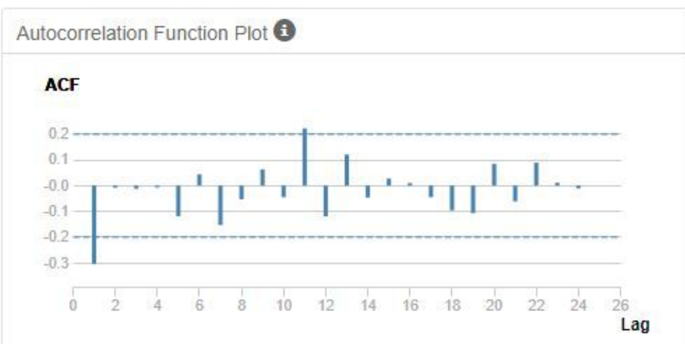


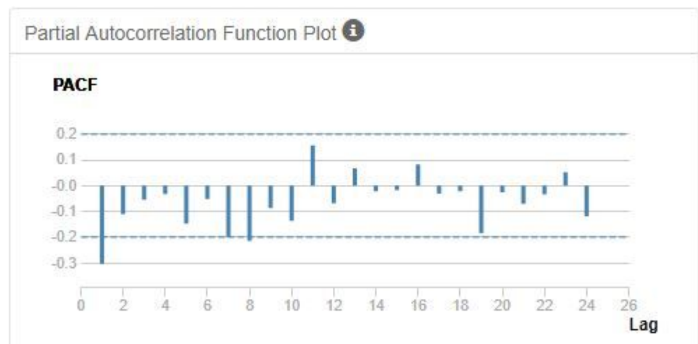


Seasonal First Difference:

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.

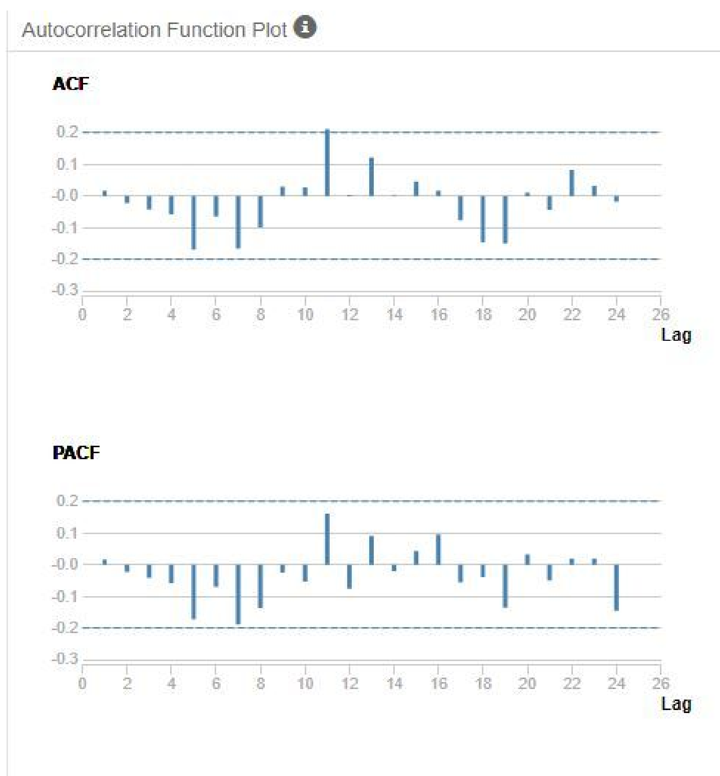




Therefore the model terms for my ARIMA model are: **ARIMA(0, 1, 1)(0, 1, 0)[12]**

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.



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 fairly 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.

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.

   The MAPE and ME measures 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	NA
ARIMA	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA
MNA_Model	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA

   For our forecast, we will use the ARIMA model.

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

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	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

Shown graphically here: