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.
   The answer is yes, the data meets the criteria of a time series type of business problem.
   - It is a numeric type of data.
   - It has enough data to make predictions.
   - It has a time-based continuous variable, where the data collected is related to this time-based variable, linked to the possible target variable.
   - The data is sequential, has equal intervals and the data has at least one data point.
2. Which records should be used as the holdout sample?
   Since there is a need to forecast 4 months ahead in time, we choose the last 4 months of the data to test the time series forecast.

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.*

   The following graphs shows the time series decomposition by using alteryx tool. This was obtained by running the workflow ahead, where the data was analyzed using the TS plot tool.

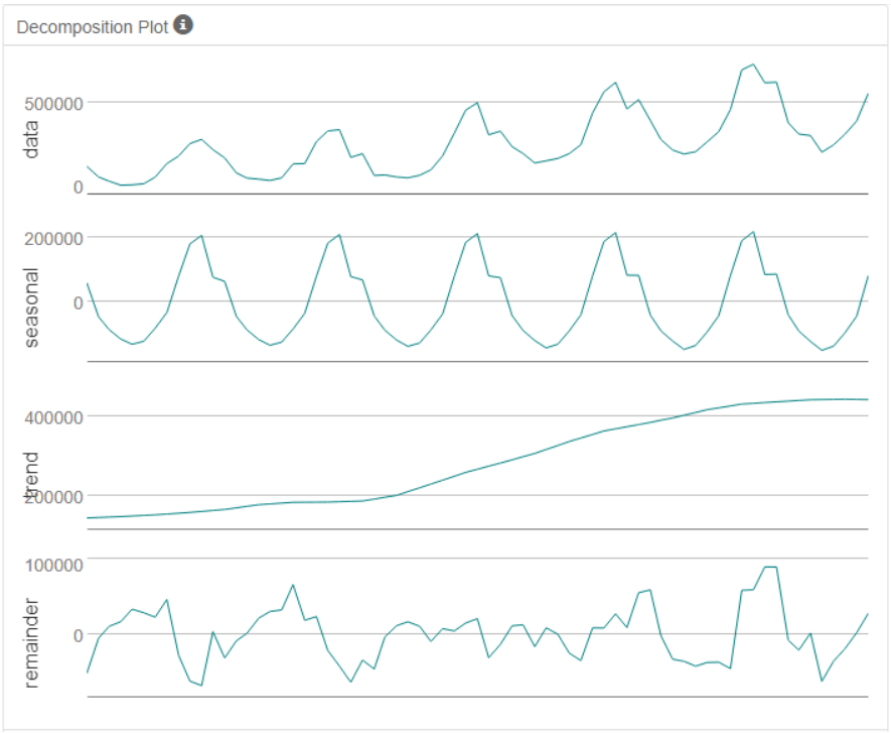

Figure 1: Time Plot Series Decomposition
Source: My Own + Alteryx Tool (2018)

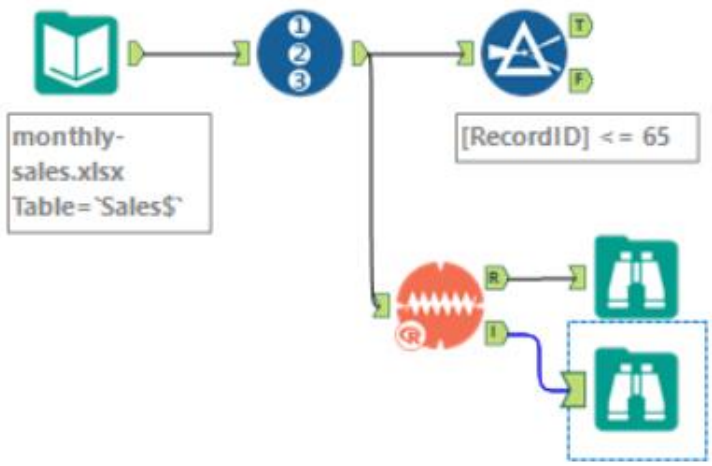

Figure 2: Alteryx Workflow to Obtain Time Series Decomposition
Source: My Own

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.*
   This is a ETS(M,A,M) case of time series forecasting.
   - The multiplicative factor in the error component is given because there is relevant variation in the error component of the time series decomposition plot.
   - The additive factor in the trend component is given because there is an ascending linear fashion in the trend plot.
   - The multiplicative factor in the seasonal component of the ETS is given because there is a slight increase in seasonal plot yet is very subtle.

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

Table 1: In-Sample Error Measures ETS

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

Source: My Own + Alteryx Tool (2018)

When looking at the MASE error indicator, proposed by Hyndman and Koehler (2006) for forecasting time series, it is stated that a MASE with a value lower than 1 is a good indicator, since it is better than a naïve forecast type of model. The value is 0.372, which is lower than 1.

When observing the RMSE indicator equal to 32992.72. RMSE is the sample standard deviation difference between predicted and observed values. The values are over 6 digits, so an RMSE with 5 digit reveals a close relationship between predicted and observed.

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.
   We will use a ARIMA (0,1,1)(0,1,0)12 model. There is a seasonal component to the data analyzed.
   - For the model to be stationary, we had to use one seasonal differencing with m=12 (month periods) and one normal differencing.
   - When stationary, the ACF plot showed a lag-1 in negative autocorrelation. This indicates a p=1. When observing ACF and PACF plots, ACF described a drop while

PACF describe a slow negative decay towards 0, but the plots are already stationary. The following graphs indicates the process of making the data stationary.

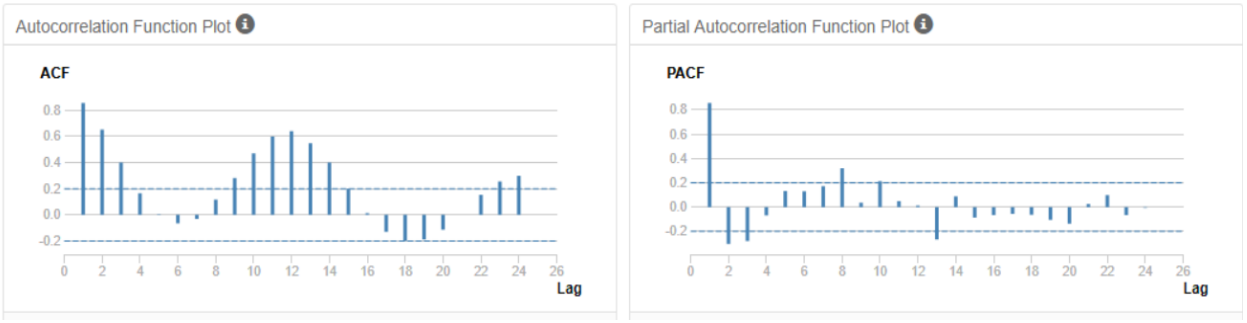

Figure 3: ACF and PACF plots Indicating Non-Stationary fashion
Source: My Own + Alteryx Tool (2018)

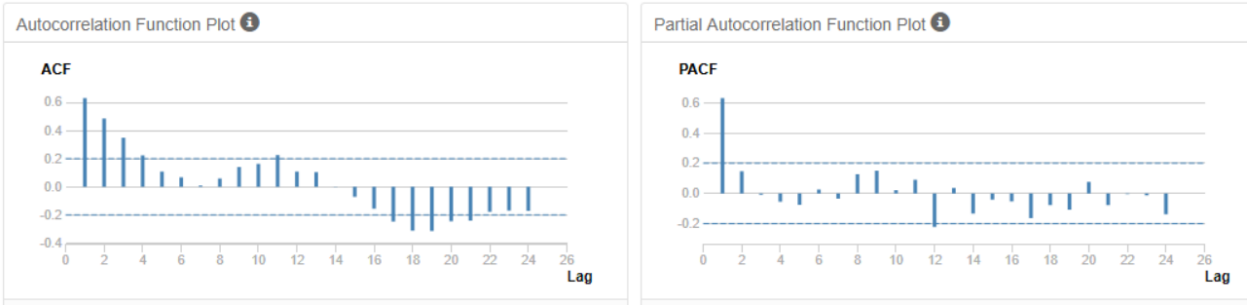

Figure 4: ACF and PACF plots after Seasonal Differencing, Indicating Non-Stationary fashion
Source: My Own + Alteryx Tool (2018)

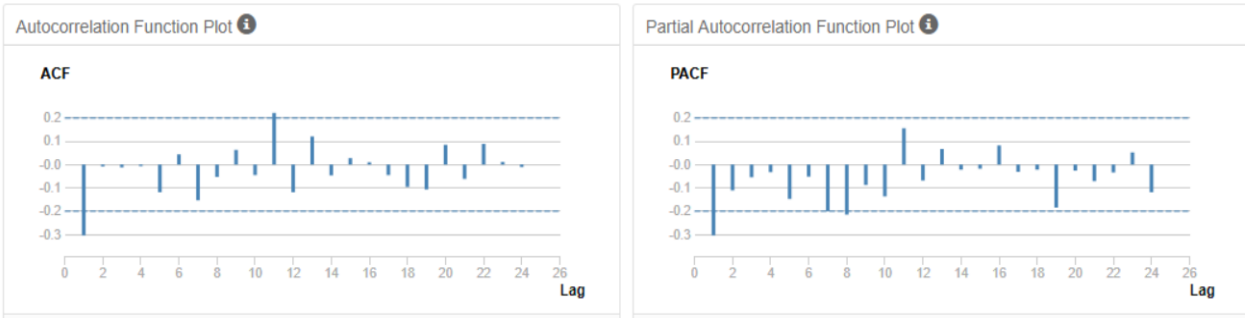

Figure 5: ACF and PACF plots after the differencing, Stationary
Source: My Own + Alteryx Tool (2018)

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

Table 2: In-Sample Measures ARIMA

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-358.1274828	36758.4027043	24996.5435416	-1.800917	9.8272386	0.3646619	0.0166958

Source: My Own + Alteryx Tool (2018)

MASE indicator is below 1, with a 0.36466 value. This indicates a good model, since it beats the Naïve Forecast Model.

RMSE is a 5-digit number, stating that the data has many values over 6 digits. It is a relatively low value, which indicates predictions will be close to the values of the holdout sample.

b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.
The graphs are in Figure 3 to Figure 5.

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.

   A good first glance when choosing a model is by looking at AIC. As coolstatblog.com (2018) says, the AIC with lower number is better. Let's take a look to these values:

   Source Link: https://coolstatsblog.com/2013/08/14/using-aic-to-test-arima-models-2/

Table 3: AIC value from ETS model.

Information criteria:

AIC	AICc	BIC
1673.431	1686.4523	1710.3956

Source: My Own + Alteryx Tool (2018)

Table 4: AIC Value from ARIMA model.

Information Criteria:

AIC	AICc	BIC
1258.5932	1259.0932	1264.447

Source: My Own + Alteryx Tool (2018)

At first glance, the ARIMA model has a lower AIC value than the ETS model.

By looking at the in-sample errors MASE and RMSE in the tables 1 and 2, ARIMA has lower values than the ETS model, which also may indicate a better model.

At last, we may see the forecasted values for each model, in order to select the best model with the accuracy measures.

Table 5: Holdout Sample actual value vs. Predicted, ETS Model

Actual and Forecast Values:

Actual	ETS_Video_Game
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Source: My own + Alteryx Tool (2018)

Table 6: Accuracy Measures ETS

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Video_Game	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA

Source: My Own + Alteryx Tool (2018)

Table 7: Holdout Sample actual value vs. predicted, ARIMA model

Actual and Forecast Values:

Actual	Video_Game_Arima
271000	263189.55788
329000	316505.01203
401000	372590.46787
553000	492977.16904

Source: My own + Alteryx Tool (2018)

Table 8: Accuracy Measures Arima

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
Video_Game_Arima	27184.45	34010.92	27184.45	6.1547	6.1547	0.4518	NA

Source: My Own + Alteryx Tool (2018)

By observing the results, it is now obvious that the ARIMA model is the best one to use, the difference between the predicted value and the actual value in the ARIMA model is smaller than the ETS.

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

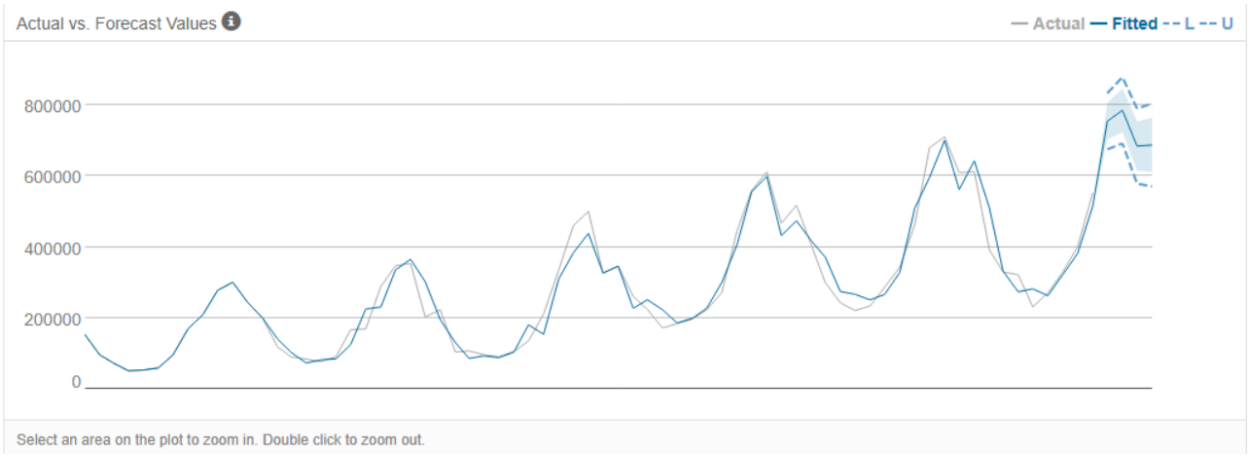

Figure 6: Forecast of Next Four Periods
Source: My Own + Alteryx Tool (2018)

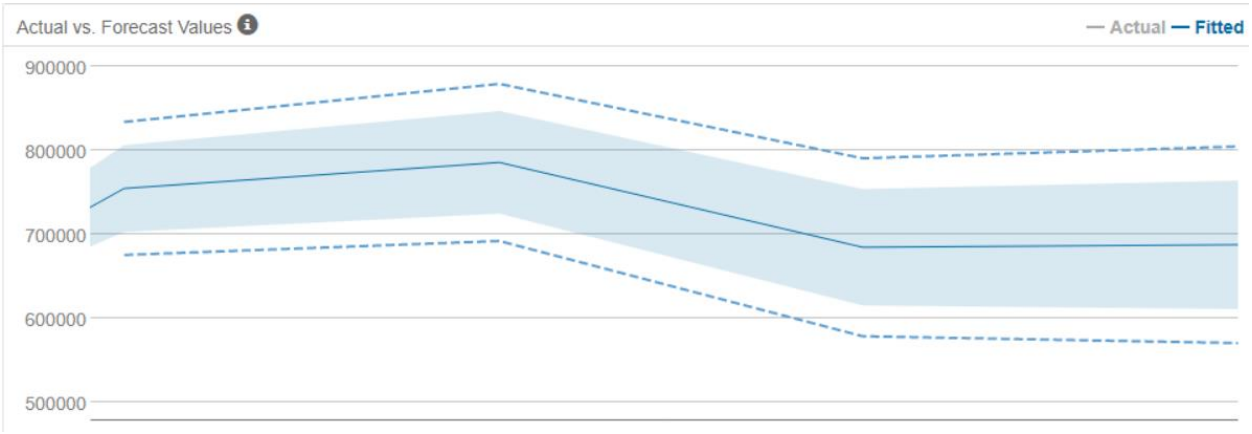

Figure 7: Zoomed of Next Four Periods
Source: My Own + Alteryx Tool (2018)

For the next four periods (Oct-2013 till Jan-2014) we have the following results:

- Oct-6 754854.46
- Nov-6 785854.46
- Dic-6 684854.46
- Jan-7 687854.46

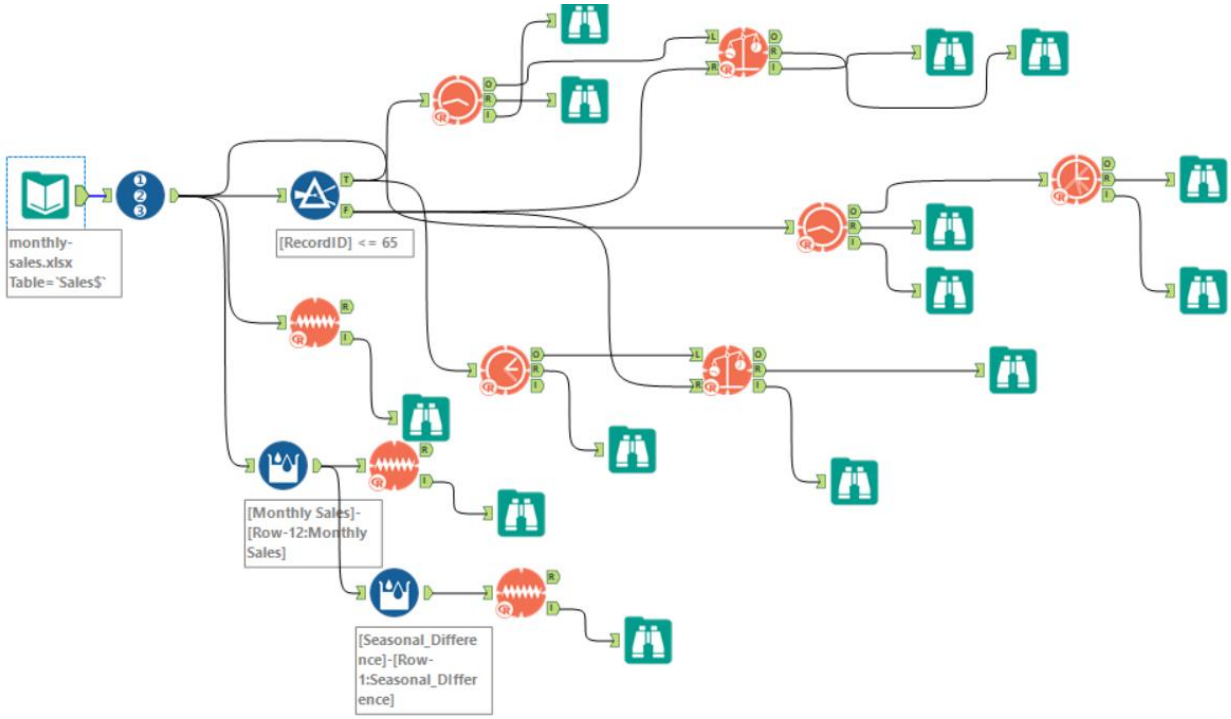

Figure 7: Alteryx Workflow
Source: My Own + Alteryx Tool (2018)

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.