UDACITY ND.Business Analytics
Francihelena Uzcategui

December 29, 2018.

Project: Forecasting Sales

Step 1: Plan Your Analysis

Look at your data set and determine whether the data is appropriate to use time series models.

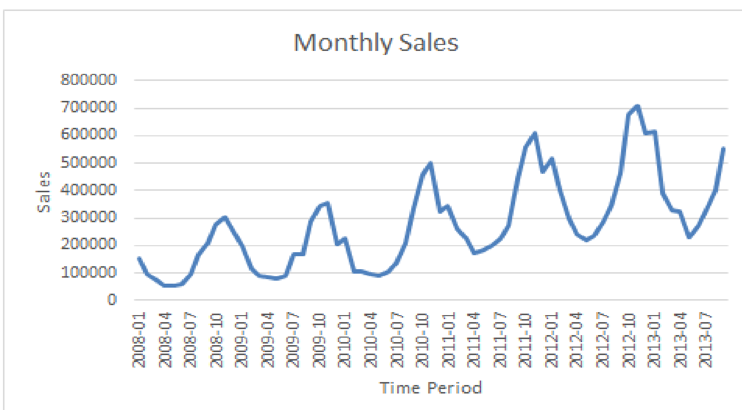
1. Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.

= In overall the dataset has a continuous time interval, due to the data goes from January 2008 to September 2009. Besides, it covers a stable and regular sales by month-to-month along five years. Consequently, there is equal spacing among every two consecutive measures, excepting in 2013 that included sales data from January to September, although I consider the criteria was met, because of lacking just the last three months in 2013. Further, there is a monthly sales value by each month along the five years.

Hence, the time plot below confirmed some features:

- There is an upper trend, for example between 2009-04 and 2010-01 is clear the changing direction such as higher and lower.
- There is a seasonality or constant pattern, like the sales that every April goes down, and every October goes up.
- The range values are similar between them, means there aren't outliers.

Thus, in my view, the dataset meets the criteria, because it has a trend and seasonality, being it reliable to use for our forecast and to design a time plot.



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

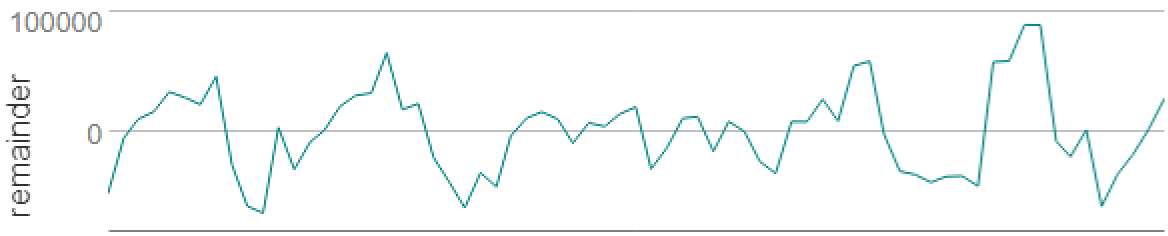
= I used a holdout sample of 4, specially the last four months of the datasets from 2013-06 to 2013-09. That's represented by <=65.

Step 2: Determine Trend, Seasonal, and Error components

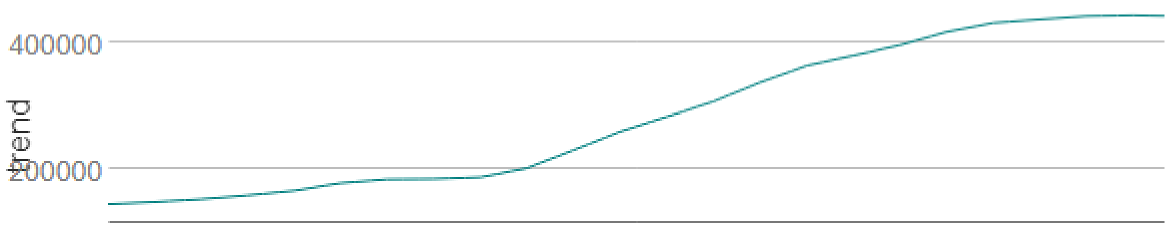
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

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.

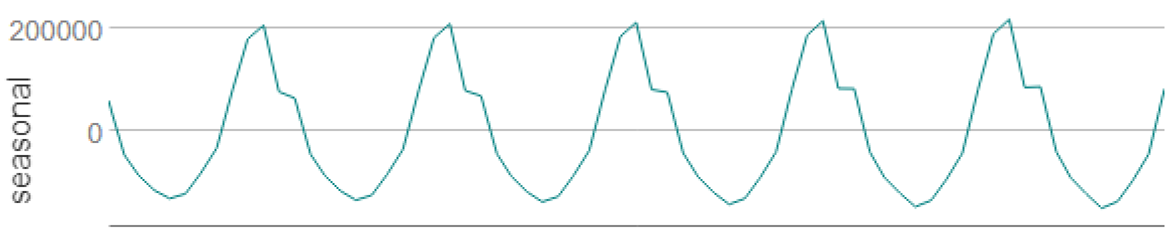
= The Error plot shows variance along the years, it is fluctuating with different sizes, this means we'll use the error multiplicatively (M.)



From this Trend Plot, we can observe the trend grows and moves uptrend, suggesting applying trend additively (A.)



Looking at the Seasonal plot there are peaks and valleys in similar periods of time, this suggests applying seasonality in a multiplicative method (M.)



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)

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 are MAM, we obtained it by the Decomposition Plot from the TS Plot tool. In details, the elements selected were Error and Seasonal terms with multiplicative patterns, contrary the Trend having an additive pattern. We choose those terms after observed the Decomposition plots movements that going up and down. Using these terms to set the ETS model, by ETS tool we obtained the next relevant measures:

Summary of Time Series Exponential Smoothing Model MAM_Video_game

Method:

ETS(M,A,M)

In-sample error measures:

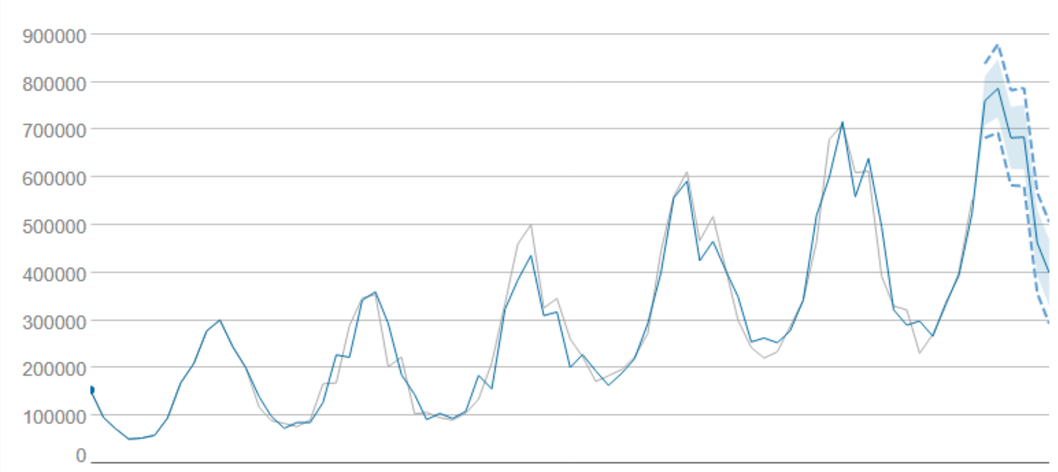
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

RMSE and MASE are important and common metrics for continuous variables.
RMSE is a scale dependent errors because is based in the set scale of the time series. RMSE's result depicts that the variance is ≈ 33000 points around deviations from the mean.
MASE is a scale-free-errors or independent scale, contrary to RMSE. A MASE significant value is less than 1, according to our result 0.3726, it's a favorable forecast of accuracy.

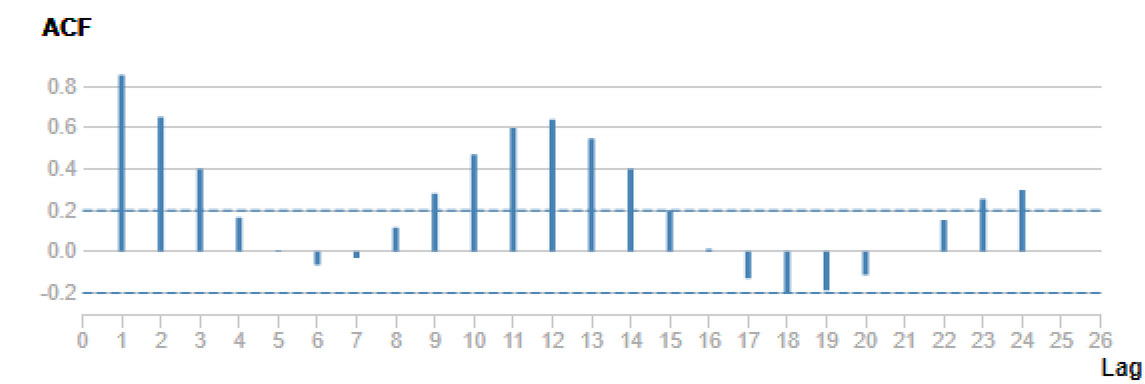
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.

= First, the ARIMA model's results from the original dataset show a time series plot non-stationary because it has a trend, as shown in the graph below.
ARIMA (1,1,1) (0,1,0) [12] depicts that there is a positive autocorrelation at period 1. Thus, it is necessary to transform the dataset to stabilized, it will be trough differencing.
ACF and PACF graphs shows the positive correlation, being suggested the AR 1 and MA 0 terms.

Forecast from ARIMA(1,1,1)(0,1,0)[12]   Jan, 1: **Actual**: 154000 **Fitted**: 153911.09

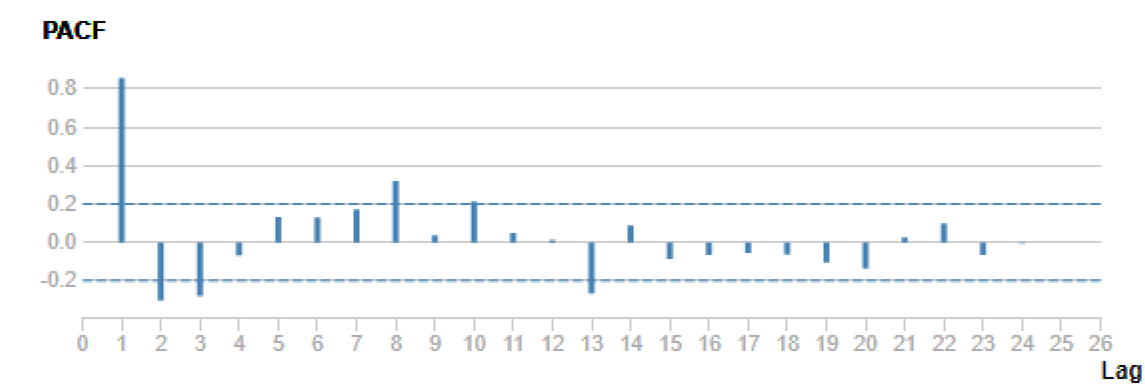


Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot

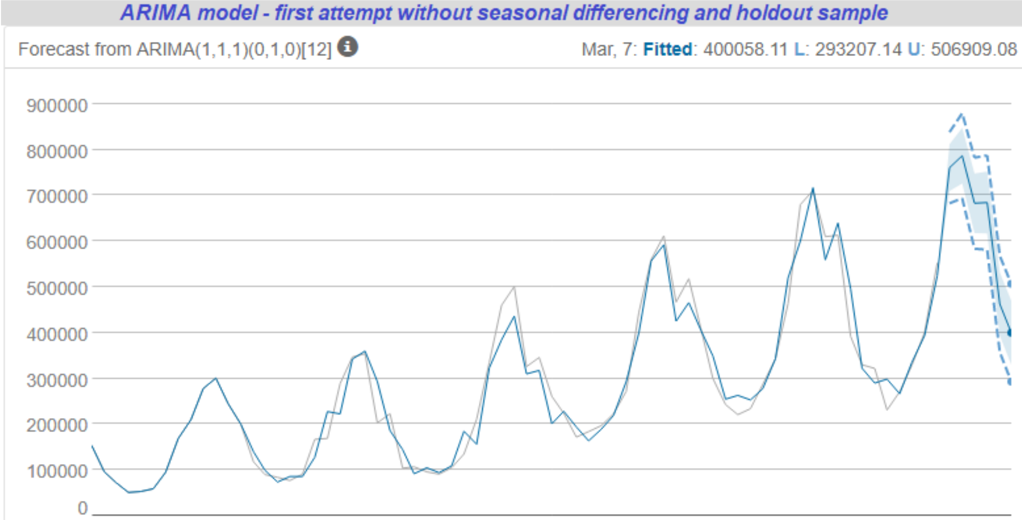


This is an partial autocorrelation plot

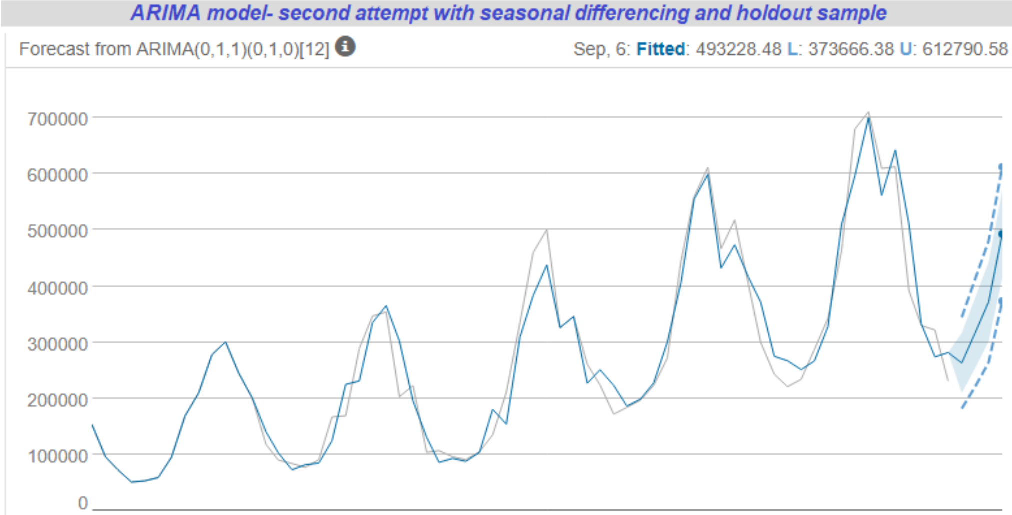
a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

= After applied the seasonal differencing and holdout sample the ARIMA's terms change respect than the first attempt. It changes to ARIMA (0,1,1) (0,1,0) [12], suggesting a negative correlation at period 1.

ARIMA model - first attempt without seasonal differencing and holdout sample



ARIMA model- second attempt with seasonal differencing and holdout sample



According to the ARIMA's summary the key measures to consider are RMSE and MASE.
RMSE's result shows that the variance is ≈ 37000 points around deviations from the mean.
Our MASE's result 0.3646, being a significant value because is less than 1, therefore it's a good forecast of accuracy.

Summary of ARIMA Model X

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

Call:
auto.arima(Monthly.Sales)

Coefficients:

	ma1
Value	-0.378032
Std Err	0.146228

sigma^2 estimated as 1722385234.94439: log likelihood = -626.29834

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

In-sample error measures:

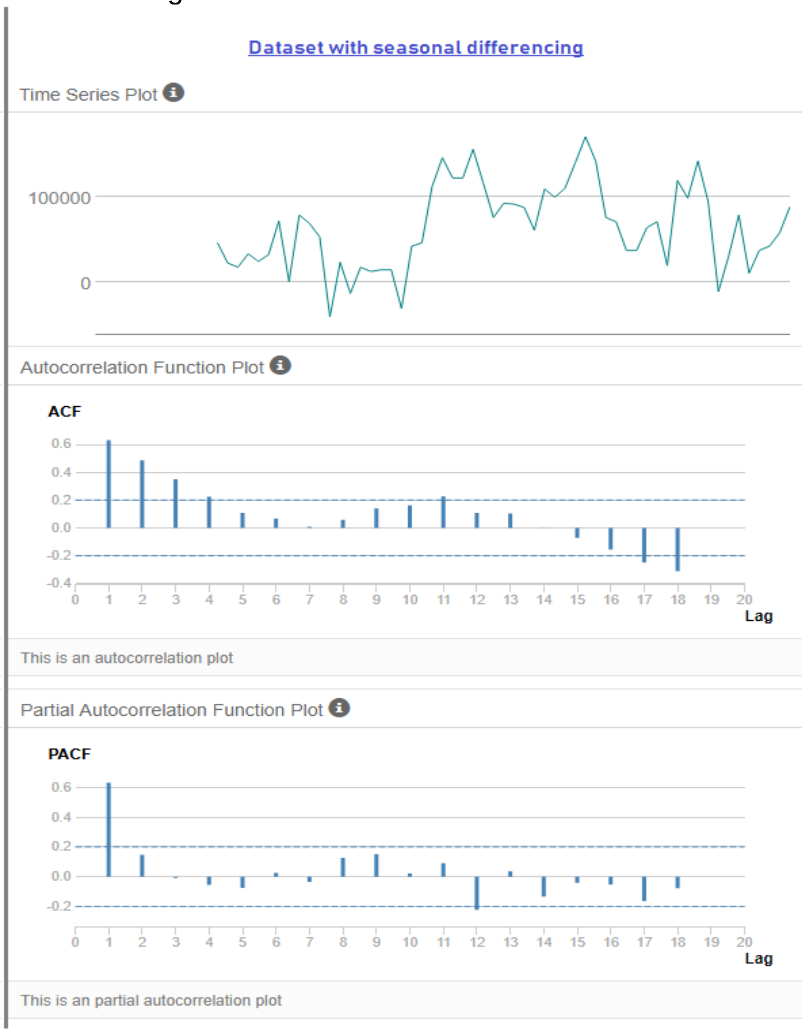
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

Ljung-Box test of the model residuals:
Chi-squared = 16.4458, df = 23, p-value = 0.83553

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

= The original dataset isn’t stationary; therefore, we'll transform it through differencing. The image at the left shows the first seasonal difference, but it was not enough, because still it fluctuates far from the zero and there is a positive correlation. Therefore, we added an extra differencing.

The image at the right describes the desired results of the extra differencing since the stationary dataset fluctuates along the zero and there is a negative correlation.

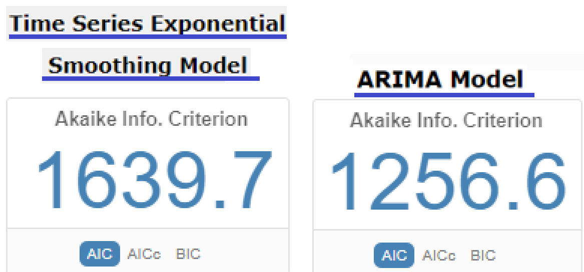


Step 4: Forecast

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)

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

= The lowest Akaike Information Criterion's model is the selecting criteria, being the ARIMA model the lowest result 1256.6, contrary to ETS 1639.7367 and ETS damped 1639.465.



ARIMA model has the best accuracy measures; specially MASE value is significant less 0.4532, being closer to 1 a good performance. RMSE is 33999.79, representing significantly fewer than ETS model result.

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS	ARIMA
271000	248063.01908	263228.48013
329000	351306.93837	316228.48013
401000	471888.58168	372228.48013
553000	679154.7895	493228.48013

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

Besides, we analyzed and compared the In-sample error measures between ETS and ARIMA model.

Mostly all the ARIMA's measures are lower than ETS, excepting the ETS's MPE that is around four times lower than the ARIMA model. Thus, lower values depict a good selecting criterion.

Highlighting the MASE measure is a key measure for selecting; therefore, we selected the ARIMA model because the MASE value is 0.3646 converse to ETS's MASE 0.3726. ARIMA's MASE value is closest to 1, being a good forecast.

Summary of Time Series Exponential Smoothing Model

In-sample error measures:

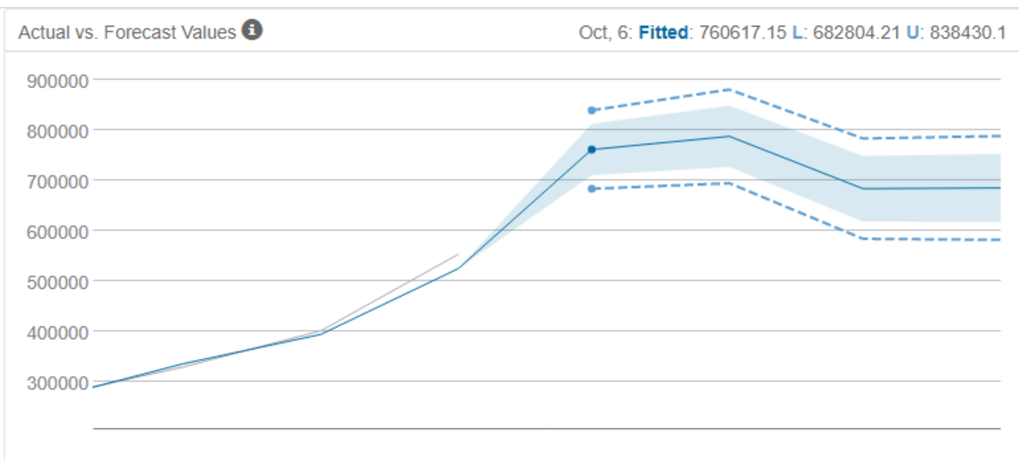
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

Summary of ARIMA Model X

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

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



ARIMA model components:

p,d,q (0,1,1)

P,D,Q (0,1,0)

	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
Oct, 6	6	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
Nov, 6	6	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
Dec, 6	6	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
Jan, 7	7	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462

Works Cited

1.1 Overview of Time Series Characteristics
https://newonlinecourses.science.psu.edu/stat510/node/47/

Another look at forecast-accuracy metrics for intermittent demand
https://robjhyndman.com/papers/foresight.pdf

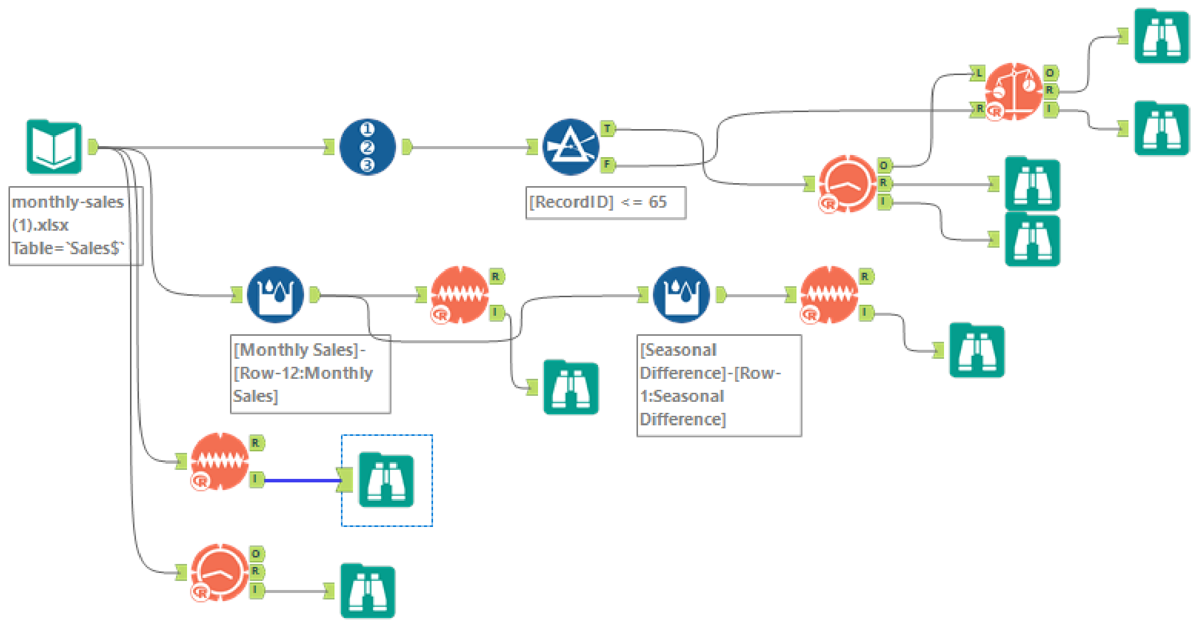
A Gentle Introduction to Exponential Smoothing for Time Series Forecasting in Python
https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/

TS PLOT error (11)
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/TS-Plot-Errors/m-p/323356

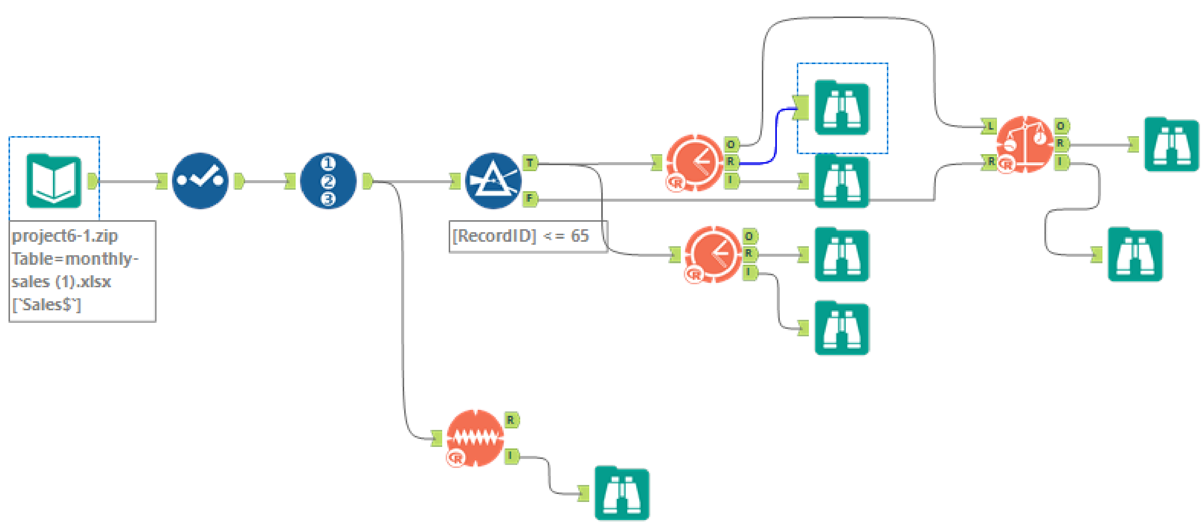
Performing Time Series Forecasting in Alteryx Designer
https://www.youtube.com/watch?v=abBzvDijEnM

Annexes

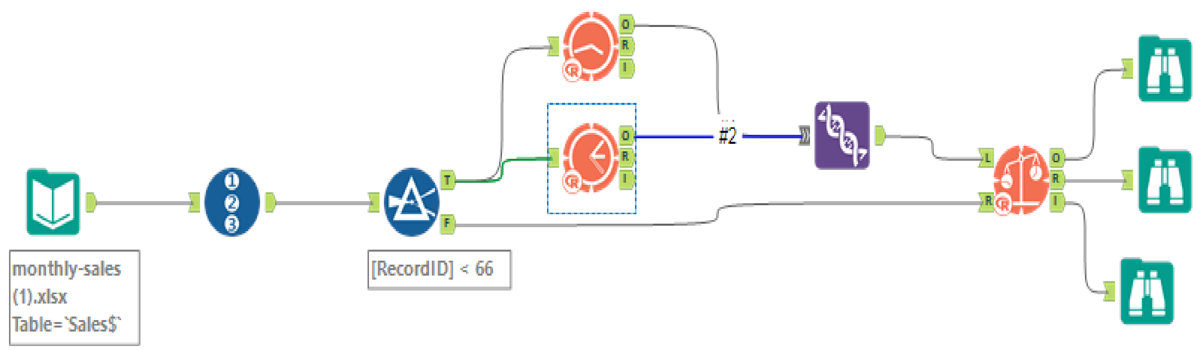
ARIMA_Sales_Video_Game.yxmd*



ETS_Sales_Video_Game.yxmd*



Comparison_ARIMA_ETS.yxmd



Forecast_model_selected_ARIMA.yxmd