Project: Forecasting Sales

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.

   A dataset should have the following four characteristics for it to be considered a Time Series.
   - It's over a continuous time interval
   - There are sequential measurements across that interval
   - There is equal spacing between every two consecutive measurements
   - Each time unit within the time interval has at most one data point

   Looking at the given dataset after cleaning(splitting Year & Month), it does have the above characteristics and hence it can be considered a time series dataset.

Record #	Year	Month	Monthly Sales
1	2008	January	154000
2	2008	February	96000
3	2008	March	73000
4	2008	April	51000
5	2008	May	53000
6	2008	June	59000
7	2008	July	95000
8	2008	August	169000
9	2008	September	210000
10	2008	October	278000
11	2008	November	301000
12	2008	December	245000
13	2009	January	200000
14	2009	February	118000
15	2009	March	90000
16	2009	April	84000
17	2009	May	77000
18	2009	June	91000
19	2009	July	167000
20	2009	August	169000
21	2009	September	289000
22	2009	October	347000
23	2009	November	354000
24	2009	December	203000
25	2010	January	223000
26	2010	February	104000
27	2010	March	107000
28	2010	April	96000
29	2010	May	91000
30	2010	June	105000
31	2010	July	135000

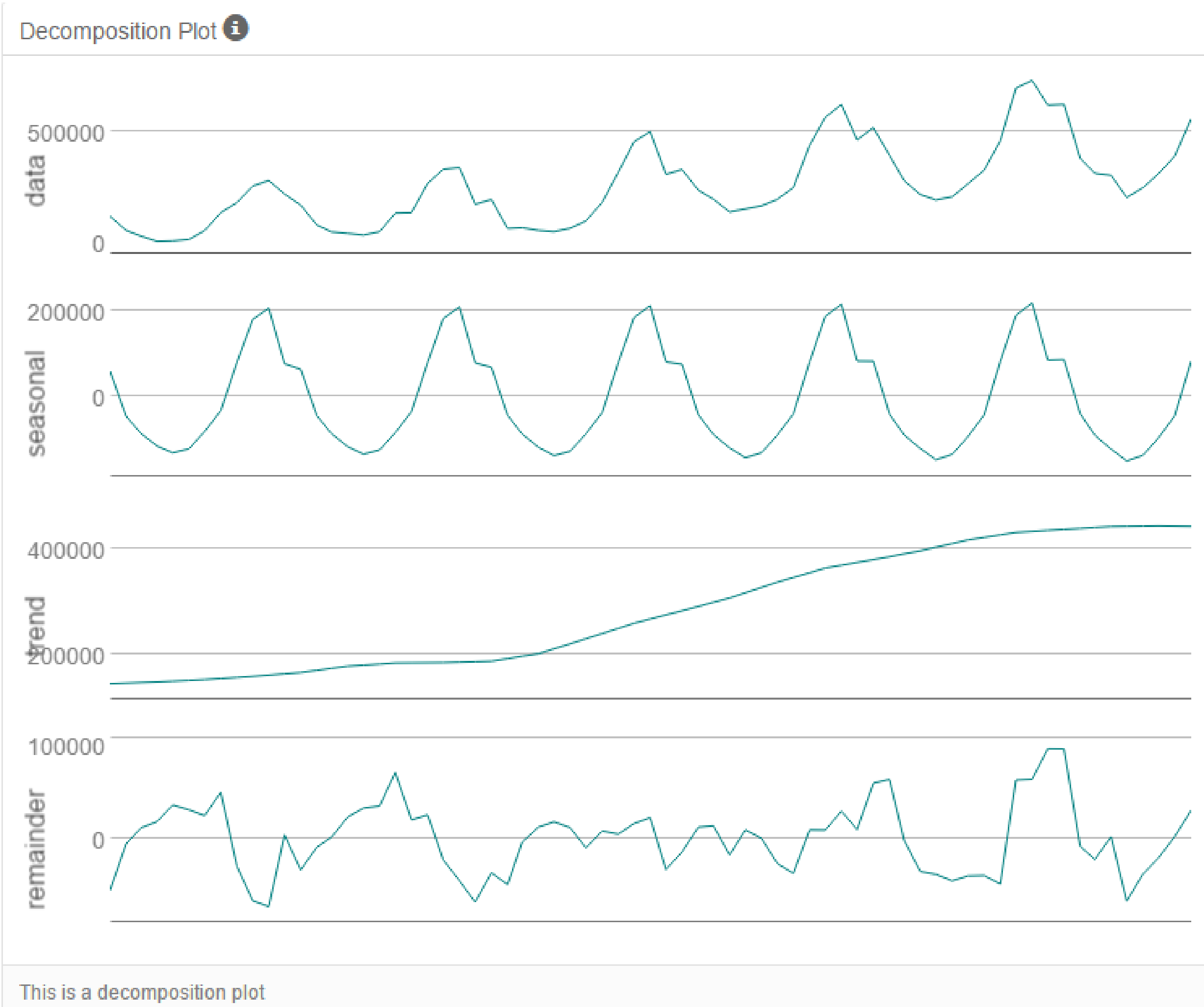
2. Which records should be used as the holdout sample?

   Holdout sample is a subset of the time series, usually the most recent data points, that you withheld and then used to check the accuracy of predictions from your model. Ideally, the size of the holdout sample should be at least the number of periods we are forecasting for. Since we need to forecast for 4 months, the holdout sample should also be the last 4 records, which in this case would from Record#66 to Record#69.

Step 2: Determine Trend, Seasonal, and Error components

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.

   Using TS Plot tool, we generate the Decomposition plot as shown below.



The first graph is the time series before being decomposed, using Year, Month (which relates to the time) and monthly sales. This shows a seasonally increasing time series plot.

Seasonality: This is the second graph. There is regular pattern, i.e., all the peaks are in November and all the valleys are in May. It shows a seasonal pattern repeating every twelve months. Also, the values of peaks and valleys are increasing throughout the years, E.g. Nov2008 is $207467.35 whereas Nov2012 is $219237.38.

Trend: The third graph shows that the monthly sales exhibits an uptrend which increases steadily with time.

Error: The remainder graph has varying altitudes of peaks which indicates that the error factors are not uniform.

Step 3: Build your Models

1. What are the model terms for ETS? Explain why you chose those terms.

   Error: It appears that Error is not uniform over time. Peaks and valleys are higher towards both ends of the graph, and smaller in the middle. So, we will apply the error multiplicatively (M).

   Trend: There is a linear upward trend, so it will be applied additively (A).

   Seasonality: At a closer look shows the peaks have a growing magnitude of sales even if it is a minor growth. So, the seasonality component will be applied multiplicatively (M).

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

   Before getting into the in-sample errors, we need to find out whether the Trend Dampening should be used or not. For which we need to compare the Damped and Un-Damped models to find out which one yields better predictions. The series starting period is Jan-2008 and we want to predict 4 periods, same as holdout.

   With Trend Dampening:

   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

   Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

   Without Trend Dampening:

   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

   Information criteria:

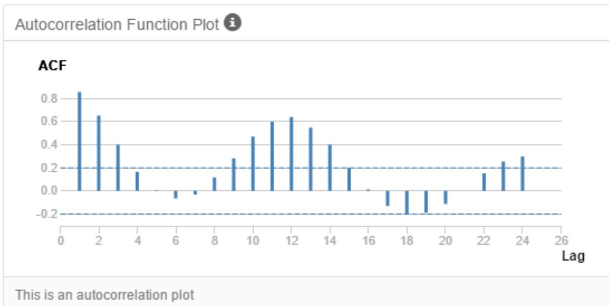
AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

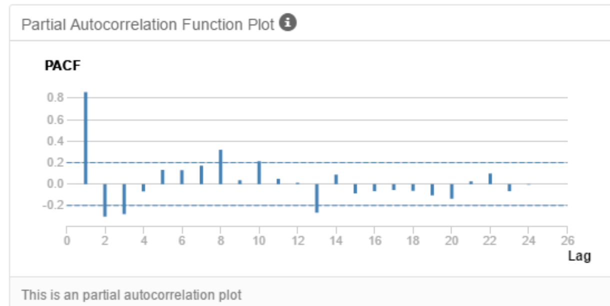
   The damped model has a lower AIC (damped 1639.465 vs undamped 1639.7367) and lower MASE (damped 0.3675478 vs un-damped 0.372685), so we will be choosing the damped ETS (M, A, M) model over the un-damped.

   From the Errors for the damped ETS (M, A, M) model (shown above), we note that MASE is much lower than 1, which generally means it is a good prediction 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.

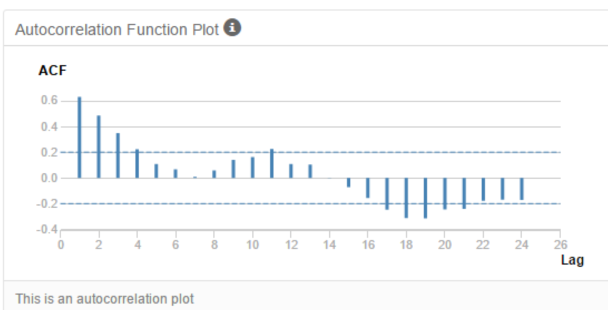
To determine the Model terms, we need to determine ACF and PACF.

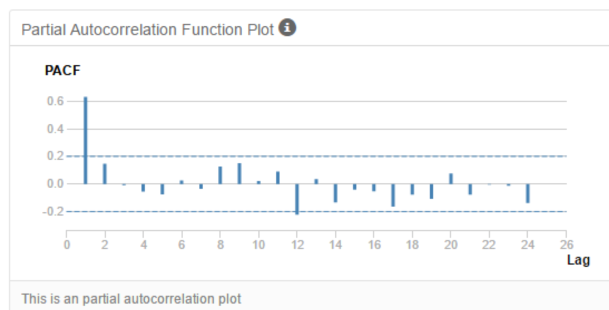




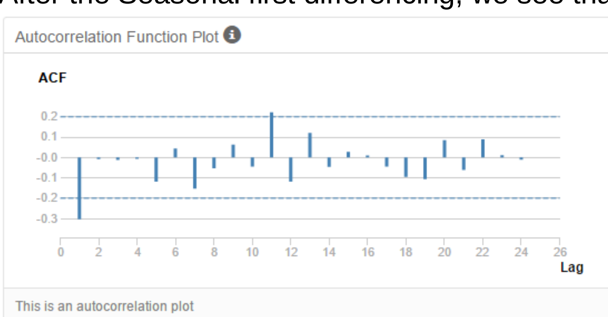
From the above graph, it can be noted that ACF is slowly decreasing towards 0 with seasonal increases at the Lags. This indicates a serial correlation, so we need to difference the series.

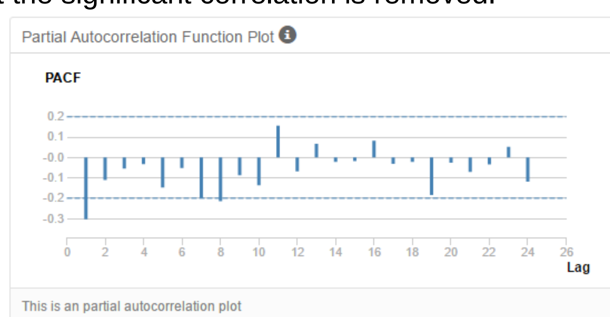
The below graph, with Seasonal difference, is still similar to the previous one. But, the correlation is lot lesser, so we will do the Seasonal first differencing.



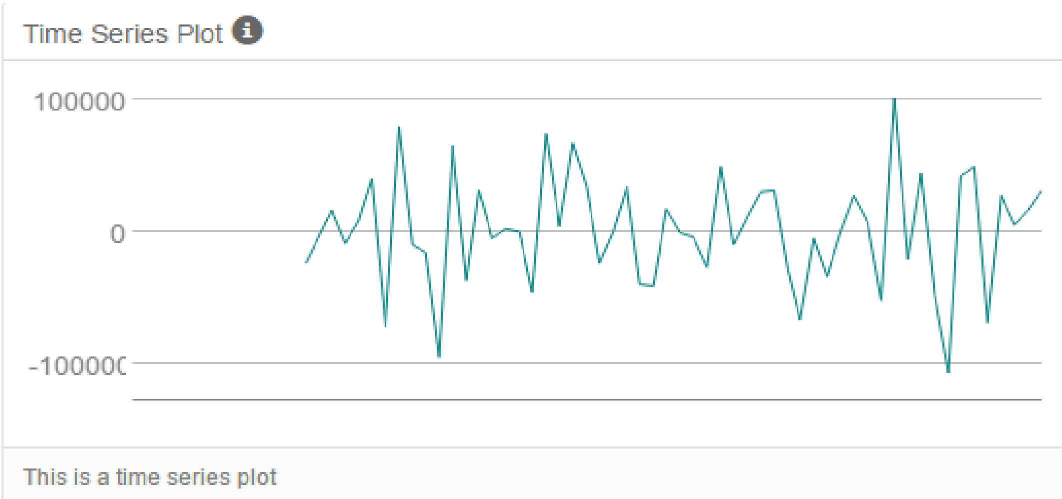


After the Seasonal first differencing, we see that the significant correlation is removed.





With the Seasonal first differencing we also get a stationary time series as shown below.



The ACF has lag-1 term and is negative so p=0 and q=1, i.e. we may add an MA term, but we may not consider adding any AR component so P=0. All seasonal lags (12, 24) do not show a spike so we may not add any MA component as well, so Q=0. But, we used seasonal differencing, so d=1 and D=1.

p=0, q=1, and d=1.
P=0, Q=0, and D=1.
M=12 as the lag repeats after 12 periods.

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

Information Criteria:

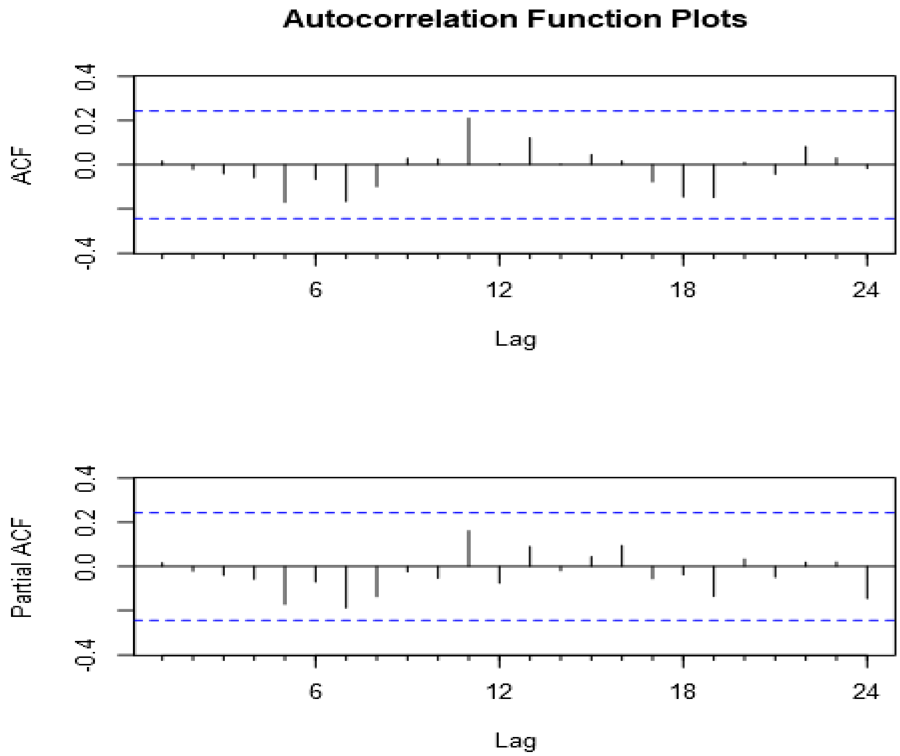
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

From the above, we see that ARIMA model's RMSE is slightly bigger than ETS model's RMSE, but ARIMA model's MASE is marginally lower than ETS model's MASE and considerably lower than 1, which indicates that ARIMA is a good model.

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



Step 4: Forecast

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

<u>ETS Model:</u>

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

Information criteria:

AIC	AICc	BIC
1639.465	1654.3346	1678.604

<u>ARIMA Model:</u>

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

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Model	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA
ARIMA_Model	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

Actual and Forecast Values:

Actual	ETS_Model	ARIMA_Model
271000	255966.17855	263228.48013
329000	350001.90227	316228.48013
401000	456886.11249	372228.48013
553000	656414.09775	493228.48013

From the above tables, we can see that

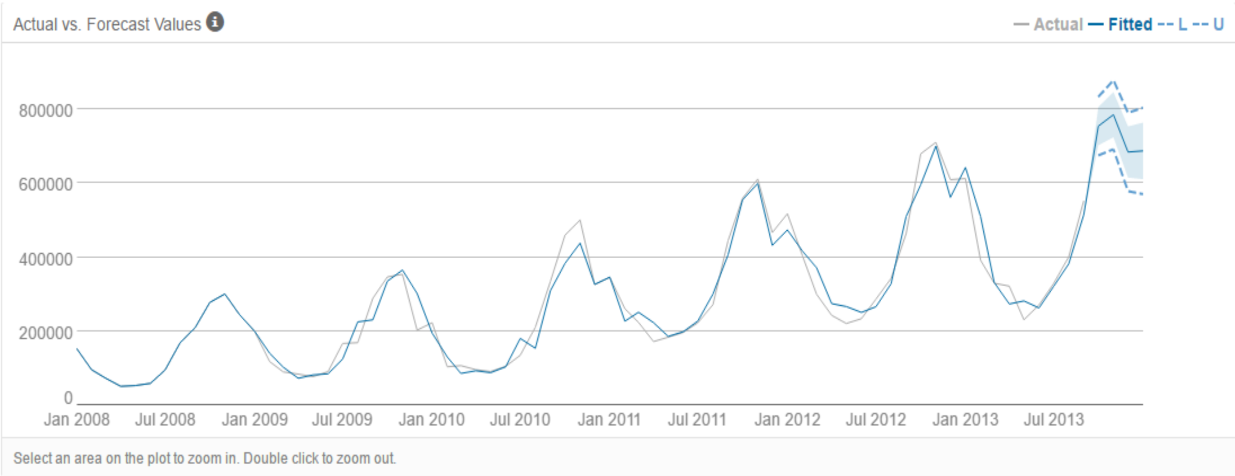
- ARIMA model has lower MASE value than ETS model.
- AIC value of ARIMA model is less than ETS model.
- *Predicted values of ARIMA model is closer to the actuals of the holdout sample than the ETS Model*.

Considering the above facts, ARIMA Model is chosen to forecast the result.

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

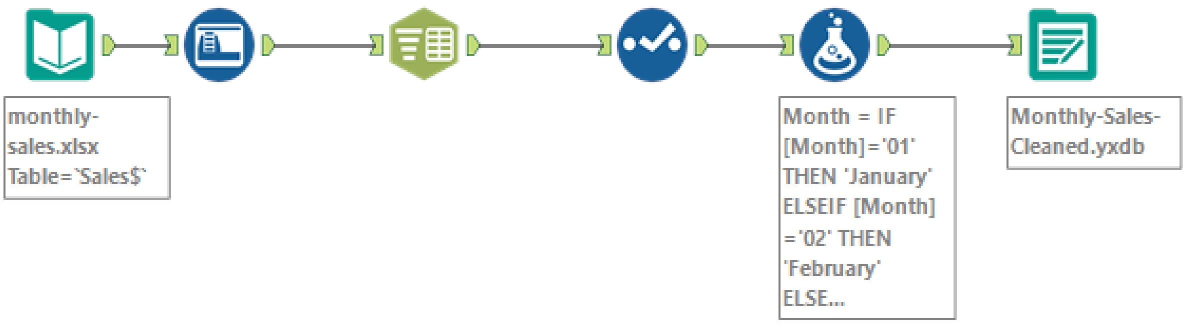
The forecasted results are as follow:

Period	Sub_Period	Final_Forecast	Final_Forecast_high_95	Final_Forecast_high_80	Final_Forecast_low_80	Final_Forecast_low_95
2013	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
2013	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
2013	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
2014	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462

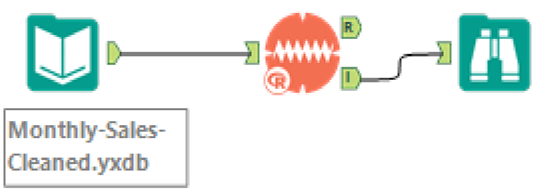


<u>Appendix:</u>

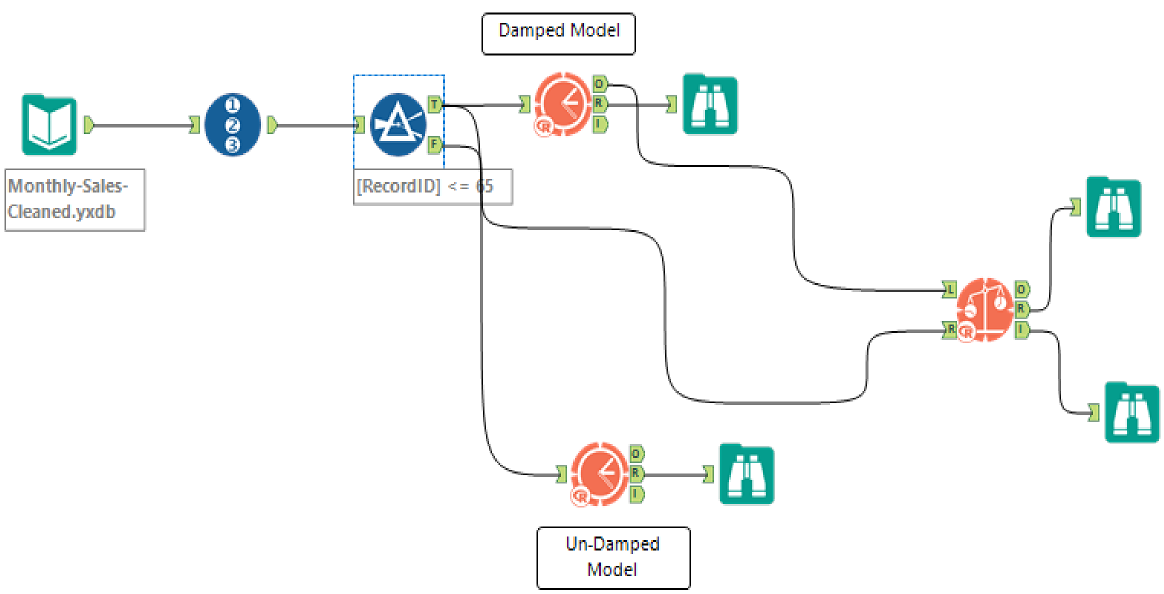
Workflow -1:



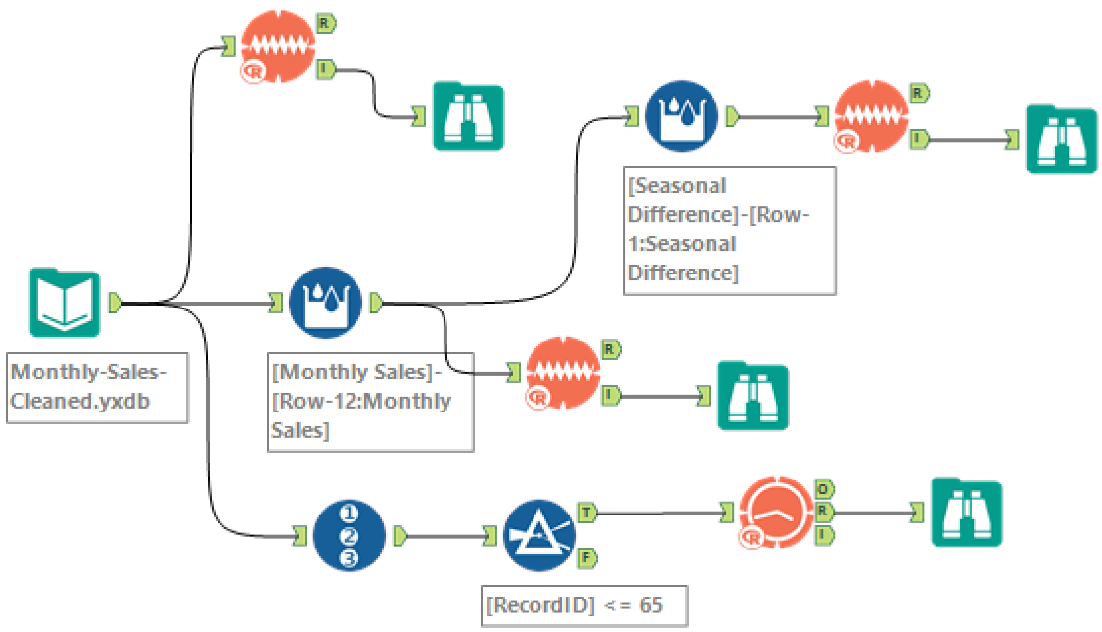
Workflow - 2:



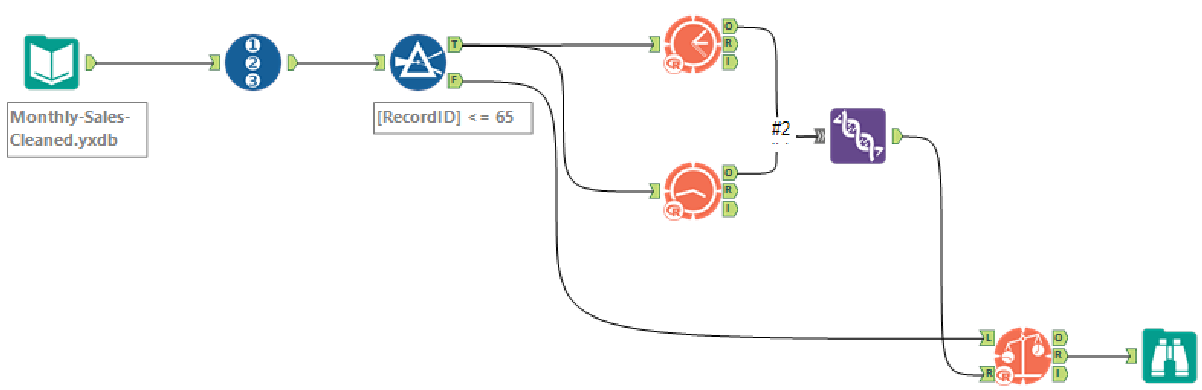
Workflow – 3:



Workflow – 4:



Workflow – 5:



Workflow – 6: