Project: Predictive Analytics Capstone

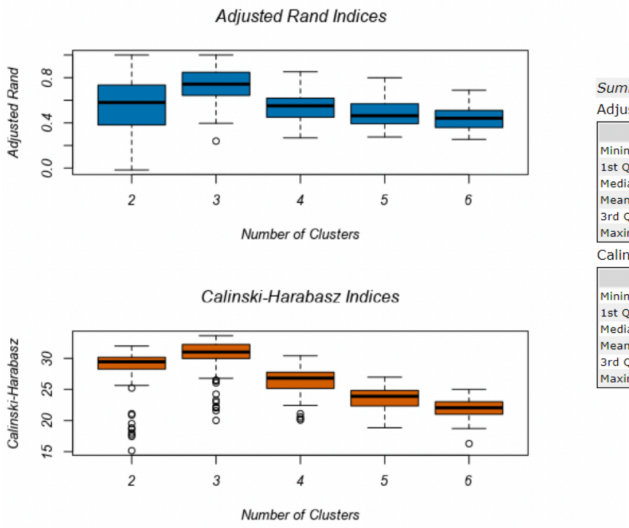
Task 1: Determine Store Formats for Existing Stores

" " "

1. *What is the optimal number of store formats? How did you arrive at that number?*
2. *How many stores fall into each store format?*
3. *Based on the results of the clustering model, what is one way that the clusters differ from one another?*
4. *Please provide a Tableau visualisation (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.*

" " "

[1] From the below box plot and the table, we came to an conclusion that the **cluster of 3 is the optimal number** of store formats. Also, considering the median values of the cluster 3 column from both the resulted tables in Adjusted Rand Indices and Calinski - Harabasz Indices, we choose a cluster of 3 is the perfect store format.



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.238908	0.26746	0.275161	0.254075
1st Quartile	0.389138	0.643526	0.451546	0.393179	0.361002
Median	0.579832	0.742946	0.550094	0.46327	0.440569
Mean	0.538248	0.716946	0.539436	0.480527	0.444128
3rd Quartile	0.734477	0.841627	0.618537	0.564177	0.507959
Maximum	1	1	0.851619	0.798934	0.689104

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.14927	20.01657	20.07469	18.84105	16.28411
1st Quartile	28.27367	30.07272	25.16346	22.35521	21.04521
Median	29.4511	31.00382	26.81884	23.89722	22.0471
Mean	28.40735	30.28555	26.35179	23.56802	21.93001
3rd Quartile	30.16162	32.23534	27.76016	24.82346	22.99673
Maximum	31.9781	33.63781	30.41396	26.97019	25.00769

[2] The number of stores falls into each store format is mentioned in the below table. 23 stores in cluster 1, 29 stores in cluster 2, 33 stores in cluster 3.

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

[3] Looking at the below resulted clustering report table, where more the positive value the more relation it has to the cluster. Hence judging from that fact, **cluster 1** has a high relation to the **general merchandise** (sale), while **cluster 2 has a higher relation to the Produce** (sale) and cluster 3 appears to have a **similar sales** unlike previous clusters.

Cluster Information:

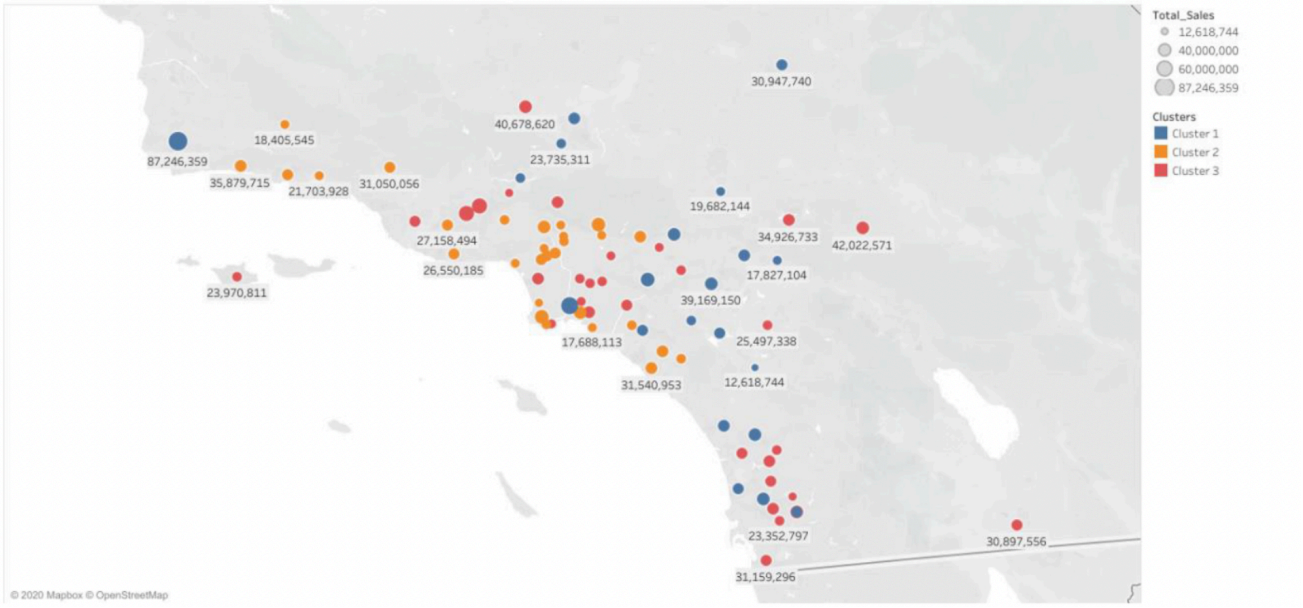
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Perc_Dry_Grocery	Perc_Diary	Perc_Sum_Frozen_Food	Perc_Sum_Meat	Perc_Sum_Produce	Perc_Sum_Floral	Perc_Sum_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Perc_Sum_Bakery	Perc_Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

[4] Tableau visualisation:



Task 2: Formats for New Stores

" " "

1. *What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)*
2. *What format do each of the 10 new stores fall into? Please fill in the table below.*

" " "

From the below model comparison report, it is clearly visible that the boosted_model and the forest model both has the highest and the same accuracy level. Since the **F1 score of the boosted model is high** comparing to the forest model, hence we chose to proceed with using the **boosted model**.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7327	0.6000	0.6667	0.8333
FM	0.8235	0.8251	0.7500	0.8000	0.8750
BM	0.8235	0.8543	0.8000	0.6667	1.0000

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of DT

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of FM

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	3

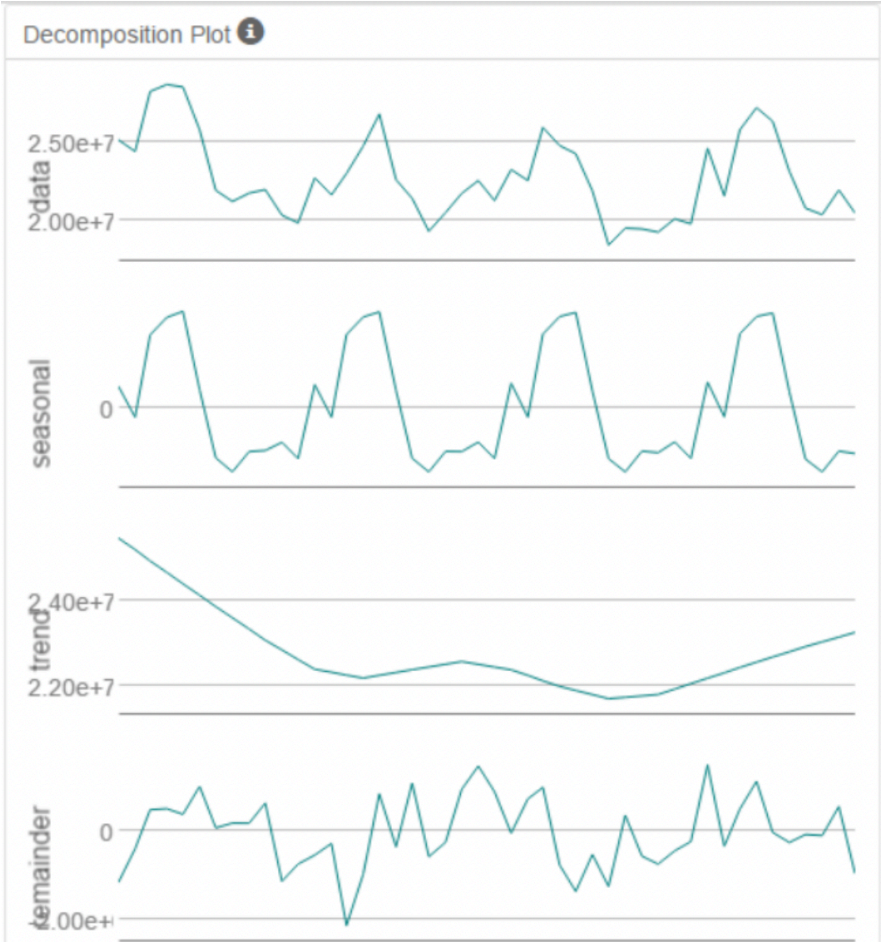
Task 3: Predicting Produce Sales

" " "

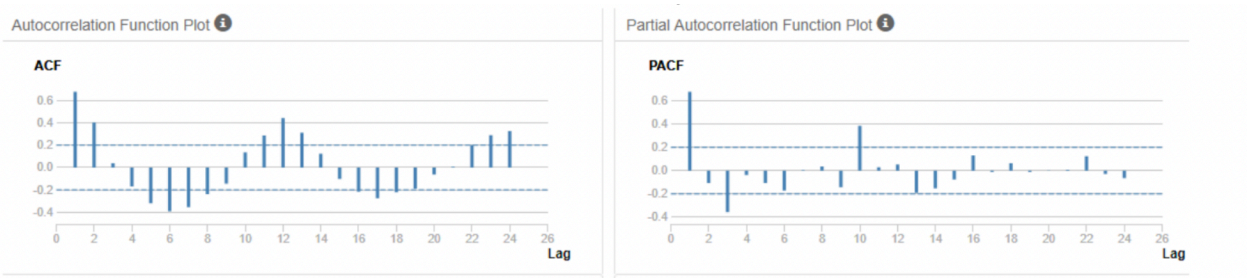
1. *What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision.*
2. *Please provide a table of your forecasts for existing and new stores. Also, provide visualisation of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.*

" " "

Here, I have used ETS (M,N,M) non dampering for ETS model. Considering the below plots, it clearly portrays that the seasonality has an increasing trend and hence I have used a **multiplicative method for seasonality**. Furthermore, the trend didn't show any signifiant pattern in it i.e. there isn't any pattern associated with trend (inconsistent) meanwhile, the error appears to be irregular, hence I've applied the **error multiplicatively**.



AFP & PAFP



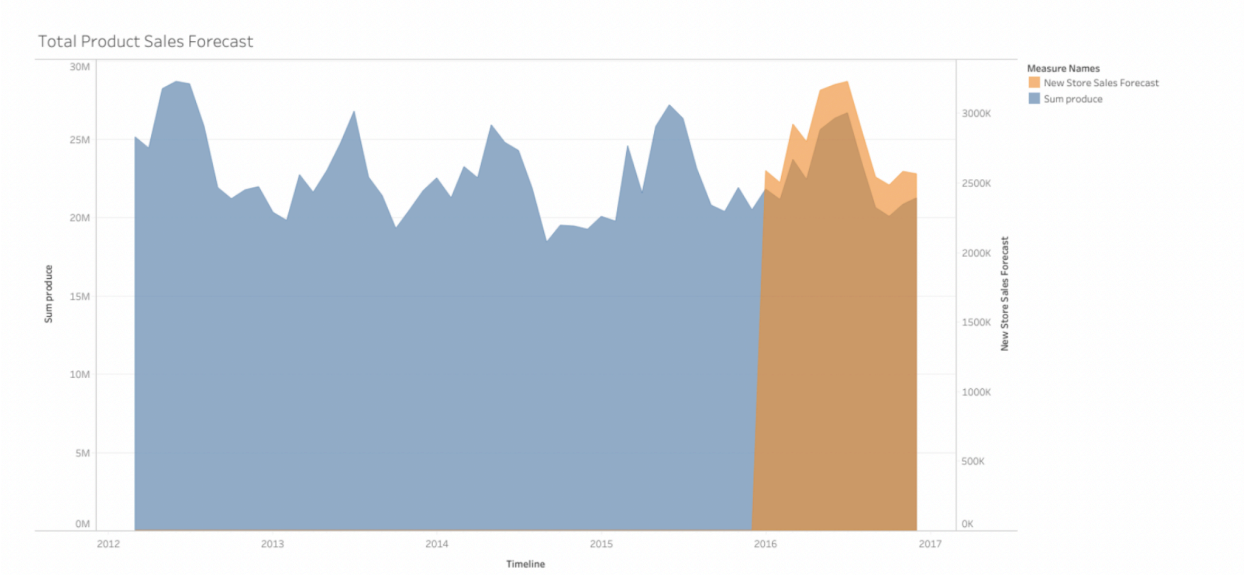
Based on the TS plot, I came to a conclusion to use the **models with parameters (0,1,2) (0,1,0).**

Once both the models are completed, we compared the performance of both the models.

Accuracy Measures:

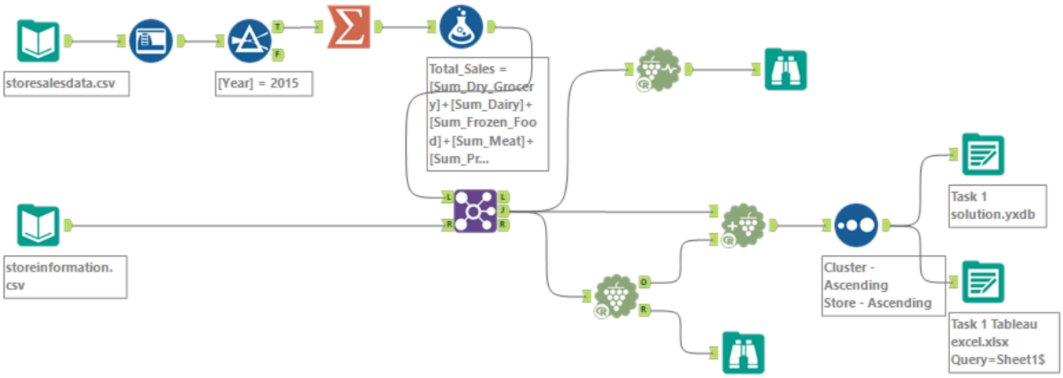
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

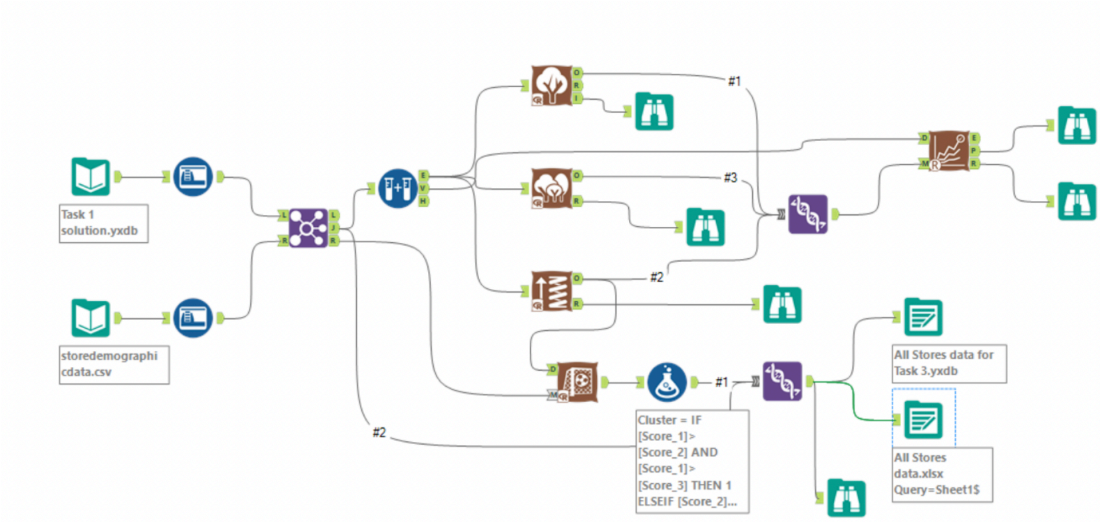
From the above comparison table which we have obtained using TS compare tool, it is very much clear that the **performance of the ETS model is higher** than the ARIMA model. Hence my answer for the question *"What type of ETS or ARIMA model did you use for each forecast?"* is ETS because of the above mentioned reasons.

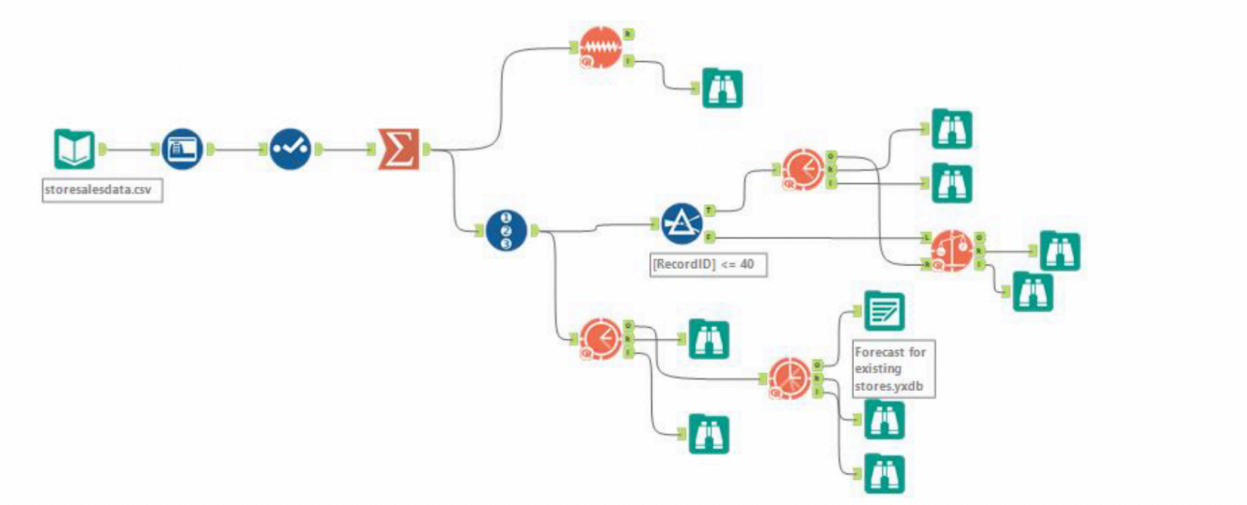


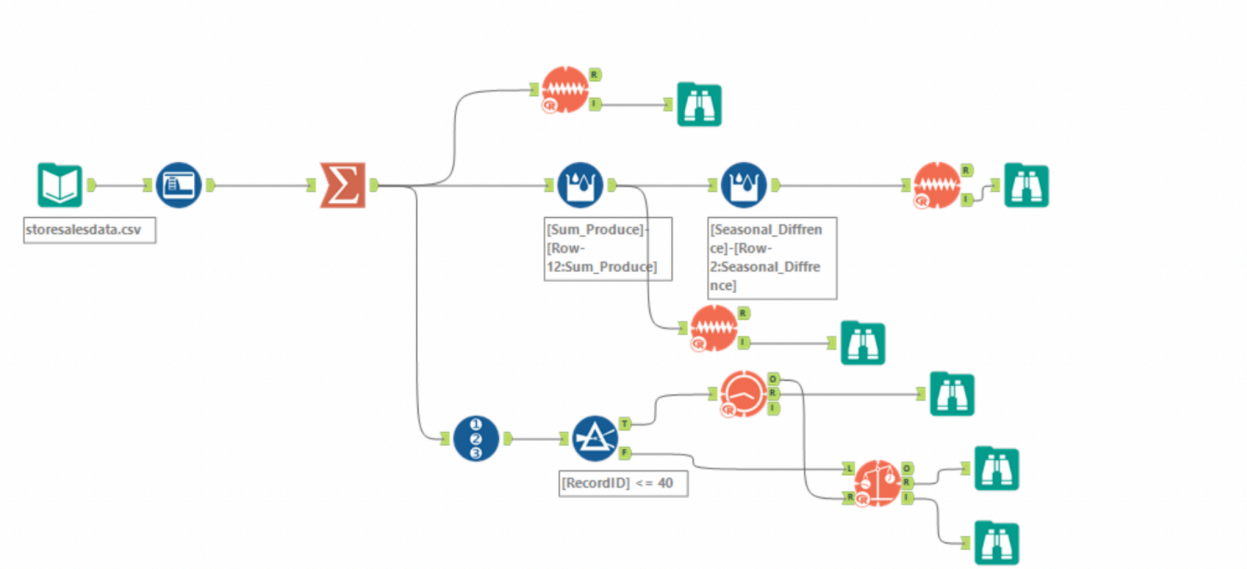
Month	**New Stores**	**Existing Stores**
Jan 2016	2,587,451	21,539,936
Feb 2016	2,477,353	20,413,771
Mar 2016	2,913,185	24,325,953
April 2016	2,775,746	22,993,466
May 2016	3,150,867	26,691,951
June 2016	3,188,922	26,989,964
July 2016	3,214,746	26,948,631
August 2016	2,866,349	24,091,579
September 2016	2,538,727	20,523,492
October 2016	2,488,148	20,011,749
November 2016	2,595,270	21,177,435
December 2016	2,573,397	20,855,799
Total annual sales	$33,370,160	$276,563,727

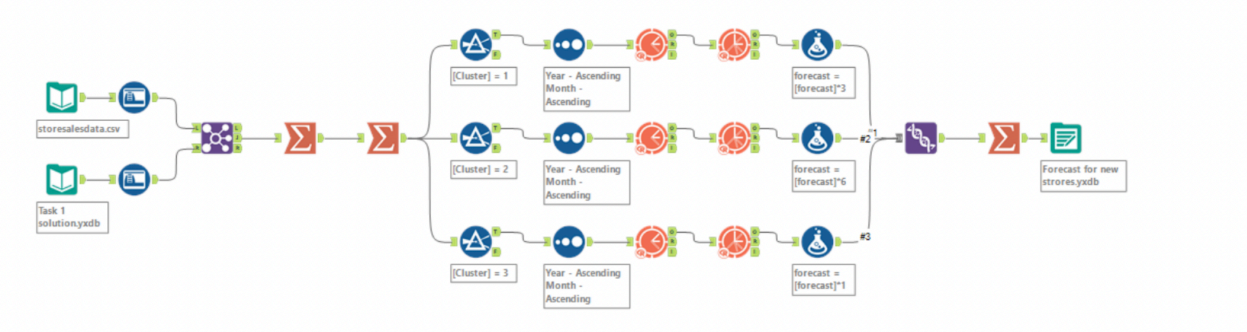
ALTERYX WORKFLOW :








storesalesdata.csv
[Sum_Produce]-[Row-12:Sum_Produce]
[Seasonal_Diffrence]-[Row-2:Seasonal_Diffrence]
[RecordID] <= 40


storesalesdata.csv
Task 1 solution.yxdb
[Cluster] = 1
[Cluster] = 2
[Cluster] = 3
Year - Ascending Month - Ascending
Year - Ascending Month - Ascending
Year - Ascending Month - Ascending
forecast = [forecast]*3
forecast = [forecast]*6
forecast = [forecast]*1
#1
#2
#3
Forecast for new strores.yxdb