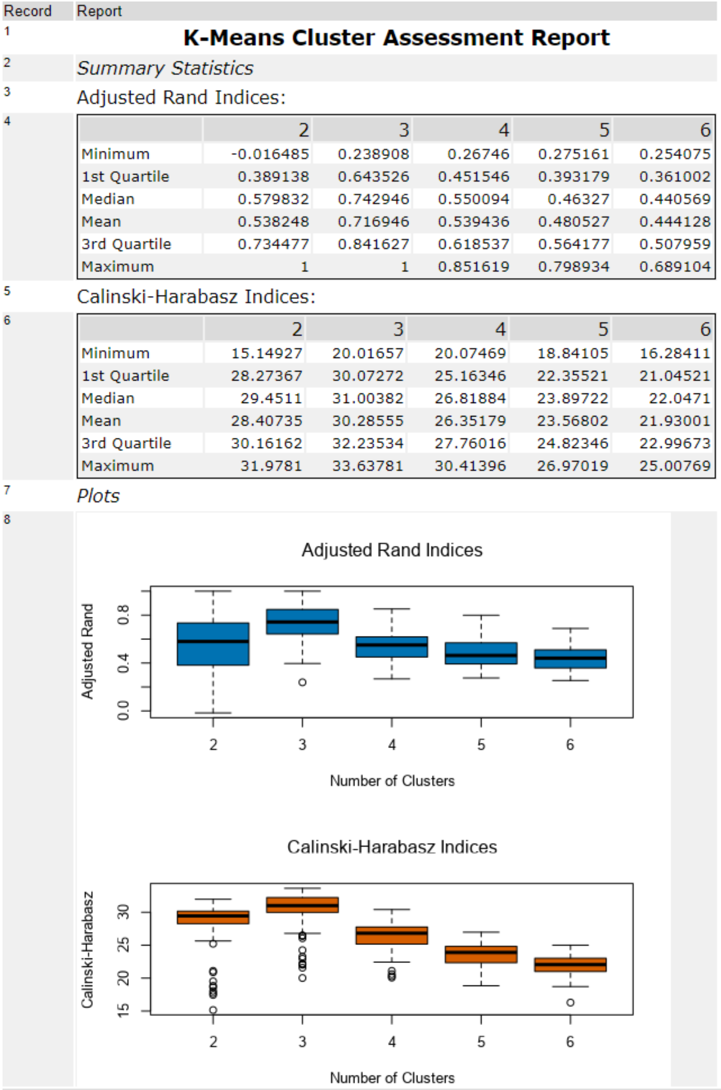
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

   After running the K-means clustering model and looking at the Adjusted Rand and Calinski-Harabasz indices, we can determine that the optimal number of store formats is 3, because it has the highest median AR and CH indices.



Record	Report
1	**K-Means Cluster Assessment Report**
2	*Summary Statistics*
3	Adjusted Rand Indices:

Record 4:

	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

5 Calinski-Harabasz Indices:

Record 6:

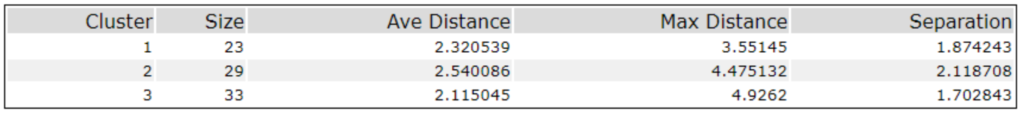
	2	3	4	5	6
Minimum	15.14927	20.01657	20.07469	18.84105	16.28411
1st Quartile	28.27367	30.07272	25.16346	22.35521	21.04521
Median	29.4511	31.00382	26.81884	23.89722	22.0471
Mean	28.40735	30.28555	26.35179	23.56802	21.93001
3rd Quartile	30.16162	32.23534	27.76016	24.82346	22.99673
Maximum	31.9781	33.63781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

7 *Plots*

8

2. How many stores fall into each store format?

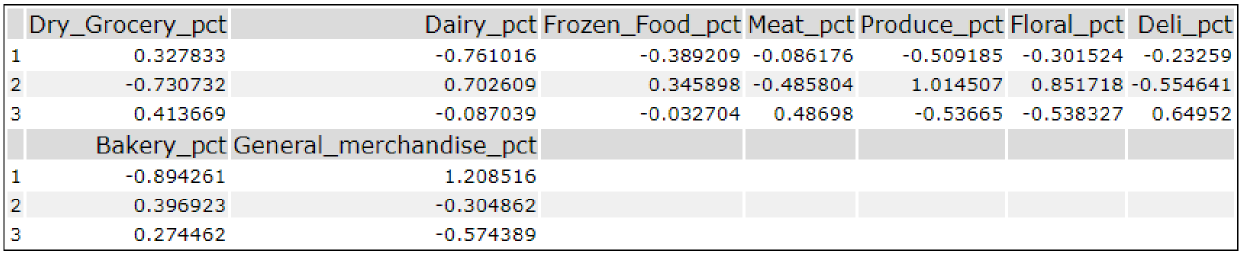
   Cluster 1 has 23 stores, Cluster 2 has 29 stores, and Cluster 3 has 33 stores



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

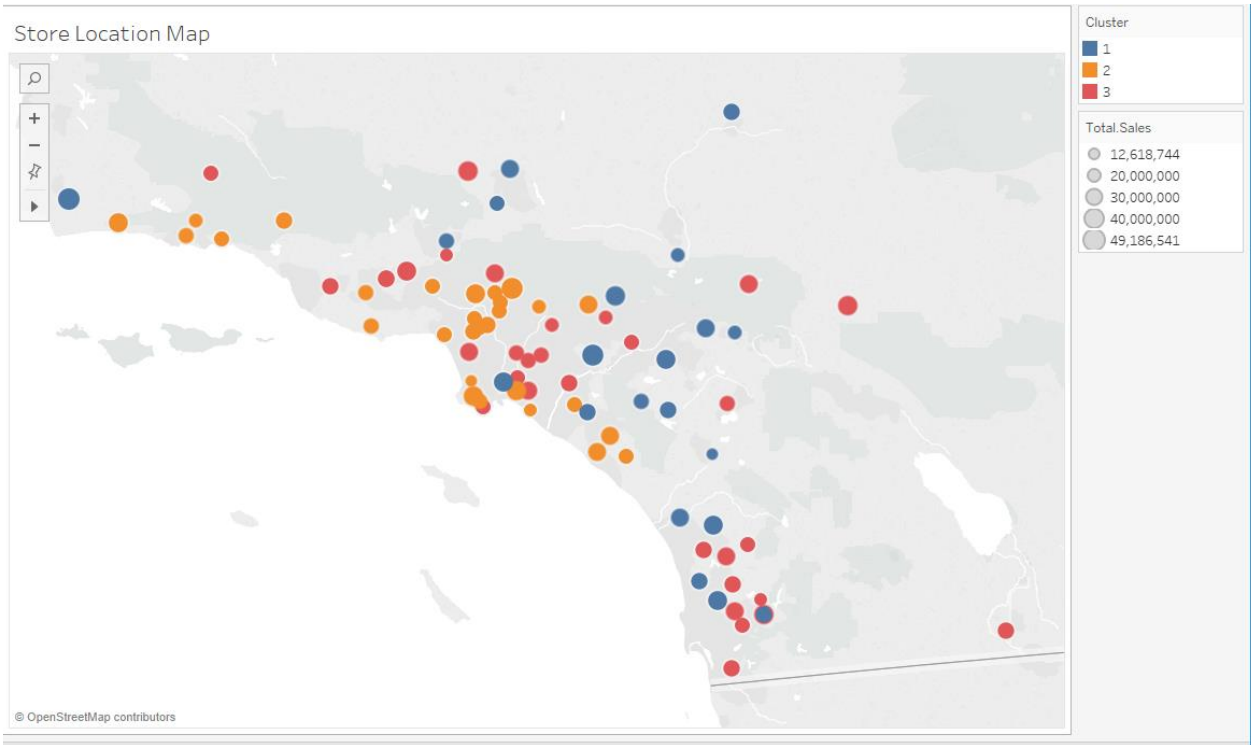
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   The results of the clustering model show that they differ in certain category sales. For example, Cluster 1 has a higher percent of General Merchandise Sales compared to Cluster 2 and Cluster 3. Cluster 2 has the highest percent of Dairy and Produce sales. Cluster 3 has a highest percent of Meat and Deli sales compared to Cluster 1 and Cluster 2.



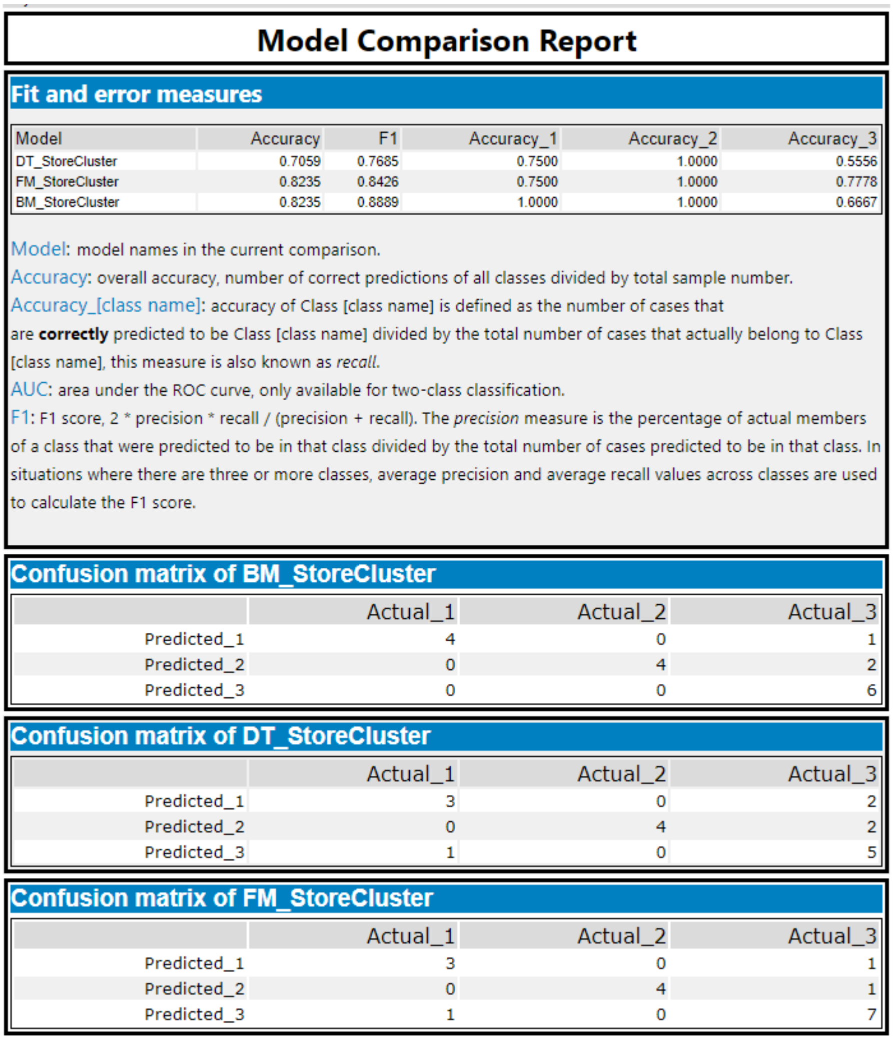
	Dry_Grocery_pct	Dairy_pct	Frozen_Food_pct	Meat_pct	Produce_pct	Floral_pct	Deli_pct	Bakery_pct	General_merchandise_pct
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259	-0.894261	1.208516
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641	0.396923	-0.304862
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952	0.274462	-0.574389

4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.



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.)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_StoreCluster	0.7059	0.7685	0.7500	1.0000	0.5556
FM_StoreCluster	0.8235	0.8426	0.7500	1.0000	0.7778
BM_StoreCluster	0.8235	0.8889	1.0000	1.0000	0.6667

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall.*

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BM_StoreCluster

	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_StoreCluster

	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_StoreCluster

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

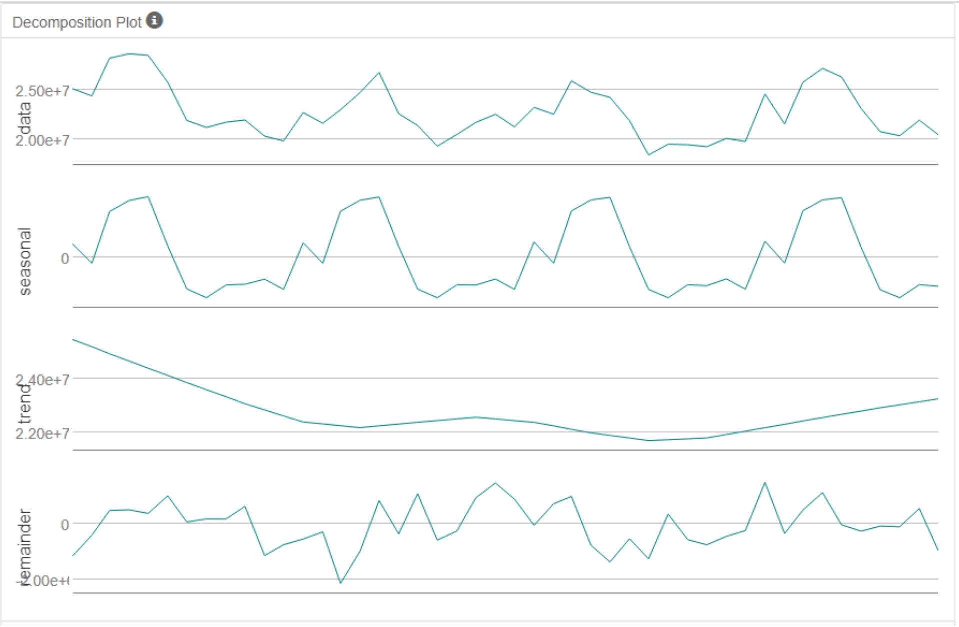
Looking at the Model Comparison Report above, the Forest Model and the Boosted Model have the highest accuracy. However, the Boosted Model has a higher F1 value, so I will choose to use the Boosted Model to predict the clusters for the new stores.

2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

Task 3: Predicting Produce Sales

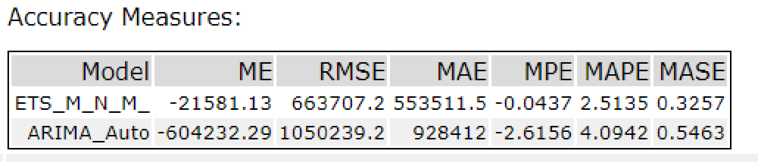
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?



By using the above Decomposition Plot we can determine the ETS model.
Error is irregular, so it will be Multiplicative. Trend is not clear, so it will be None. Seasonality is observed in the Seasonal plot, so it will be Multiplicative. The ETS model we will use is **ETS(M,N,M)**.

The ARIMA model should be **ARIMA(1,0,0)(1,1,0)[12]** as automatically determined by Aterlyx. There for we have 1 non-seasonal AR term, 1 seasonal AR term, and we used first seasonal differencing to create stationary data.

Model Comparison:



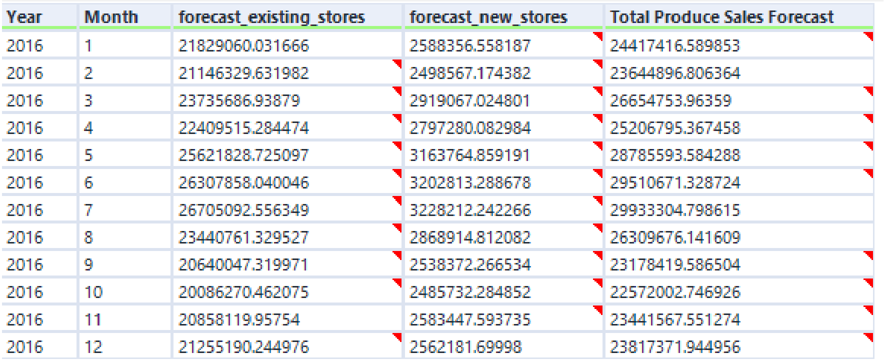
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_M_N_M_	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA_Auto	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

Looking at the Accuracy Measures table from the Model Comparison tool above, we can see that the errors for the ETS model produces lower ME, RMSE, MPE, MAPE, and MASE. So the ETS(M,N,N) model is the model that is chosen for our forecast.

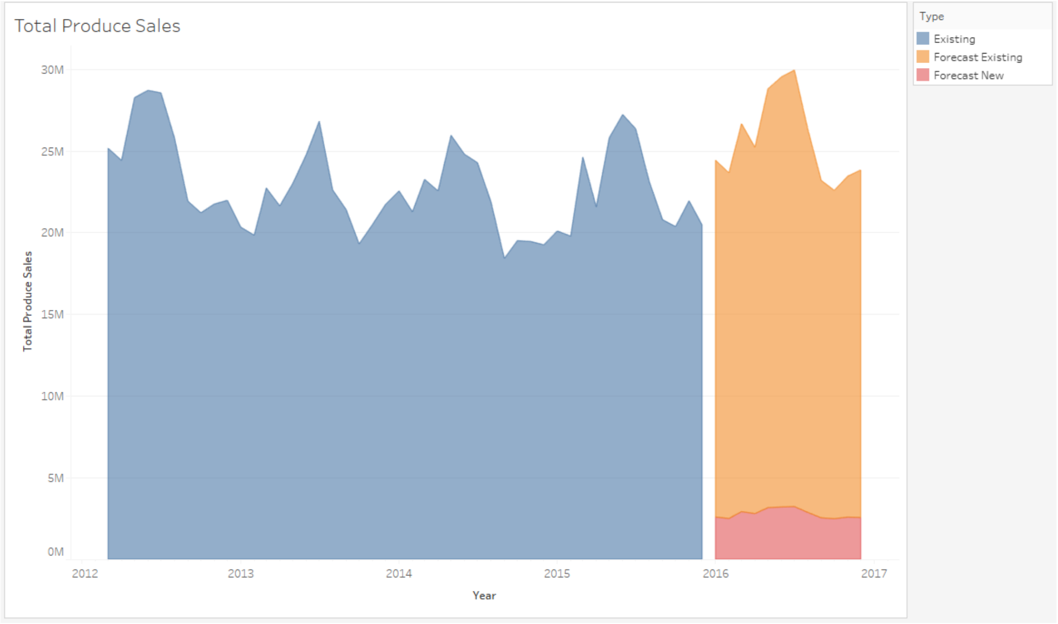
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Produce Sales Forecast

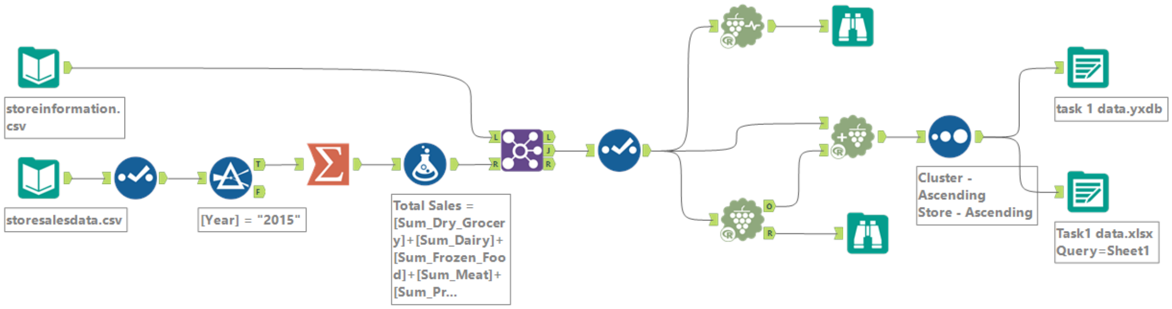


Year	Month	forecast_existing_stores	forecast_new_stores	Total Produce Sales Forecast
2016	1	21829060.031666	2588356.558187	24417416.589853
2016	2	21146329.631982	2498567.174382	23644896.806364
2016	3	23735686.93879	2919067.024801	26654753.96359
2016	4	22409515.284474	2797280.082984	25206795.367458
2016	5	25621828.725097	3163764.859191	28785593.584288
2016	6	26307858.040046	3202813.288678	29510671.328724
2016	7	26705092.556349	3228212.242266	29933304.798615
2016	8	23440761.329527	2868914.812082	26309676.141609
2016	9	20640047.319971	2538372.266534	23178419.586504
2016	10	20086270.462075	2485732.284852	22572002.746926
2016	11	20858119.95754	2583447.593735	23441567.551274
2016	12	21255190.244976	2562181.69998	23817371.944956

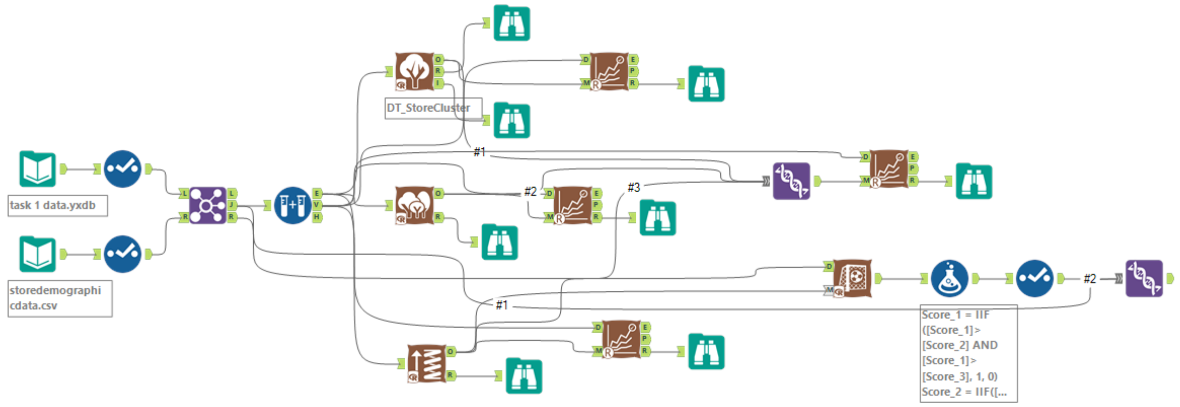
Historical and Forecast Produce Sales



Task 1 workflow

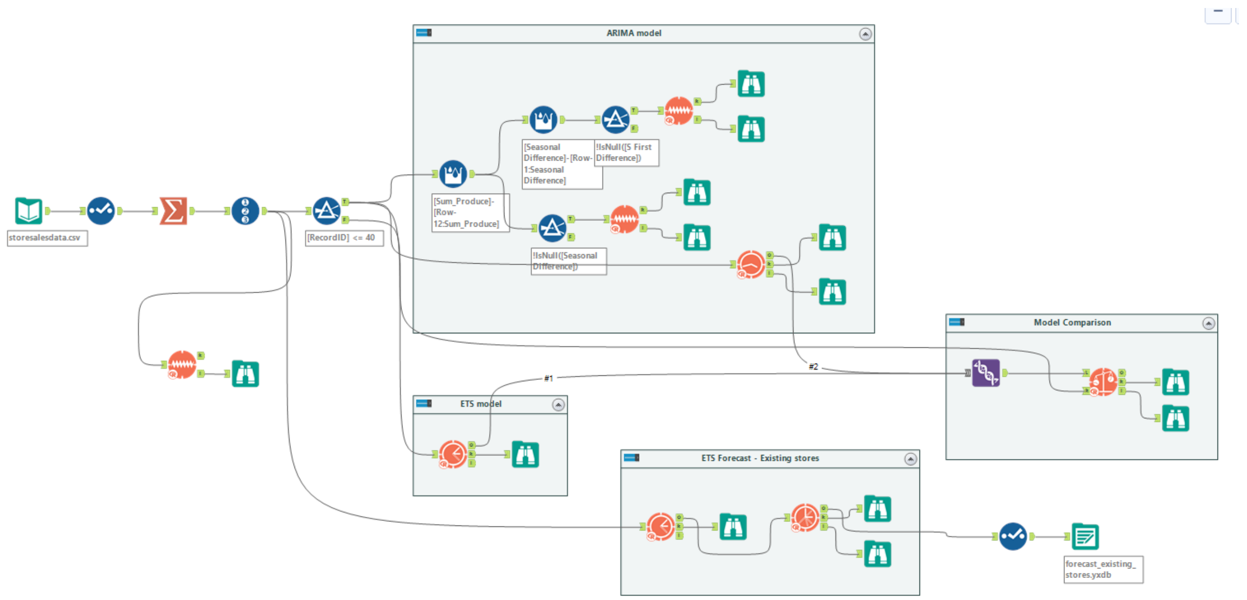


Task 2 workflow

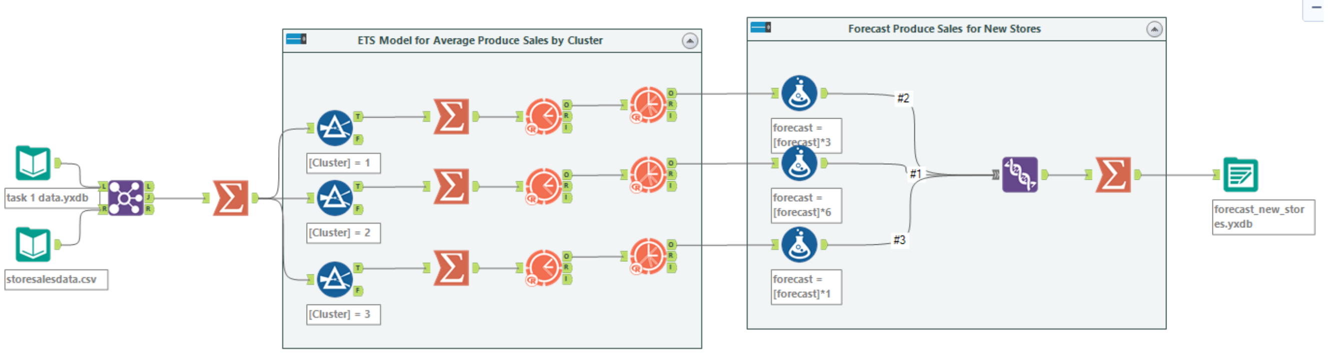

DT_StoreCluster
#1
#2
#3
task 1 data.yxdb
storedemographicdata.csv
#1
#2
Score_1 = IIF ([Score_1]> [Score_2] AND [Score_1]> [Score_3], 1, 0) Score_2 = IIF([...

Task 3 workflow

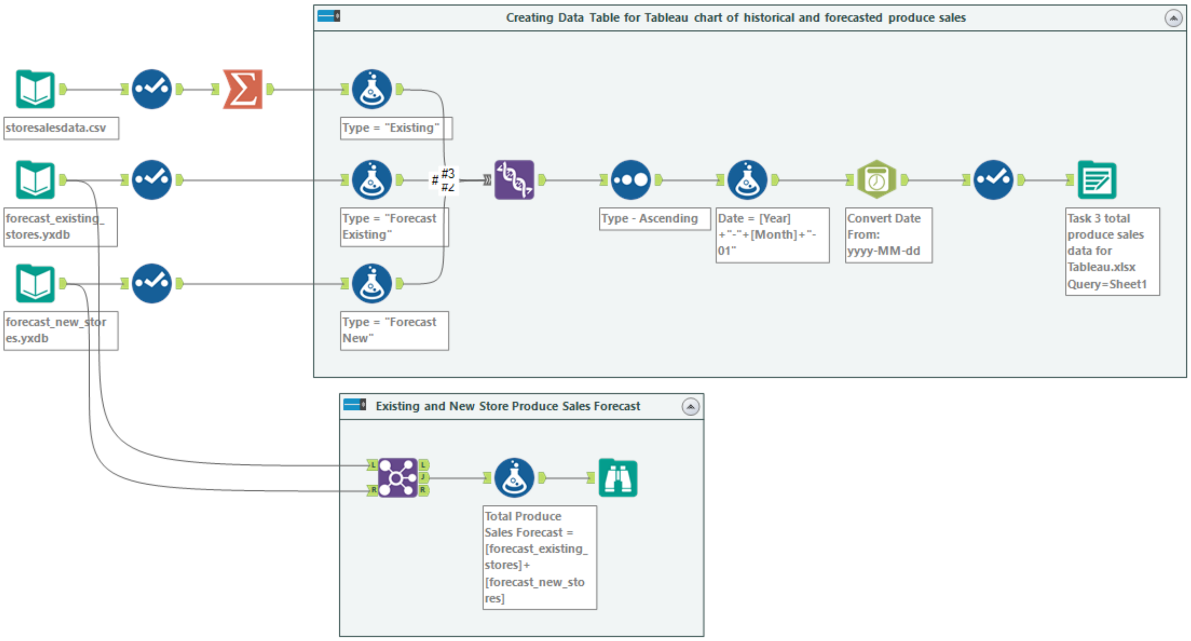
Existing stores forecast:


ARIMA model
[Seasonal Difference]-[Row-1:Seasonal Difference]
!IsNull([S First Difference])
storesalesdata.csv
[RecordID] <= 40
[Sum_Produce]-[Row-12:Sum_Produce]
!IsNull([Seasonal Difference])
Model Comparison
#2
#1
ETS model
ETS Forecast - Existing stores
forecast_existing_stores.yxdb

New stores forecast:



Creating data for Forecast table and Tableau visualization:


Creating Data Table for Tableau chart of historical and forecasted produce sales
storesalesdata.csv
Type = "Existing"
forecast_existing_stores.yxdb
Type = "Forecast Existing"
#3
#2
Type - Ascending
Date = [Year] +"-"+[Month]+"-01"
Convert Date From: yyyy-MM-dd
Task 3 total produce sales data for Tableau.xlsx Query=Sheet1
forecast_new_stores.yxdb
Type = "Forecast New"
Existing and New Store Produce Sales Forecast
L
R
L
J
R
Total Produce Sales Forecast = [forecast_existing_stores]+[forecast_new_stores]