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?

   - By using the K-means method for clustering, the Adjusted Rand Index and the Calinski-Harabasz Index indicate that cluster number 3 is the best solution.

     As we can see in the below figure the median of cluster number 3 is high and the maximum and minimum and interquartile range are compact which indicates a good stability of the cluster.

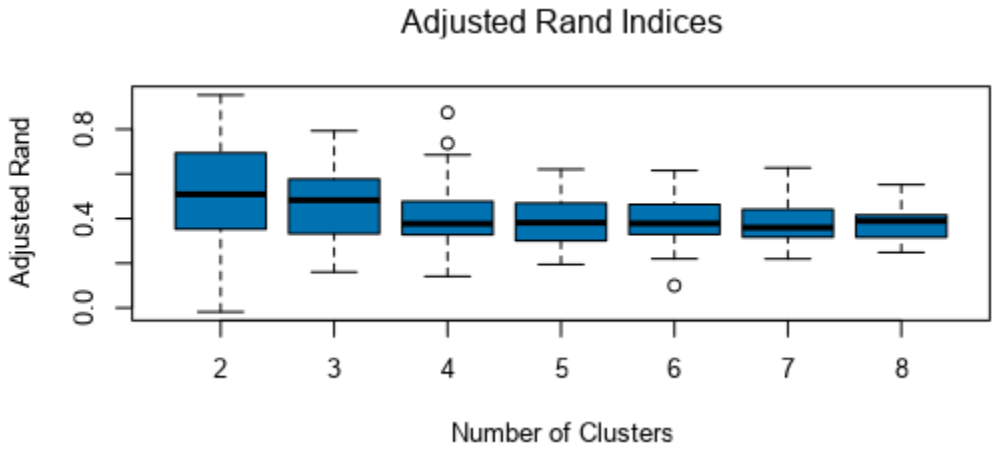

Figure 1: Adjusted Rand Indices plot

From the below Figure we can see that cluster number 3 has a high median and a compact spread.

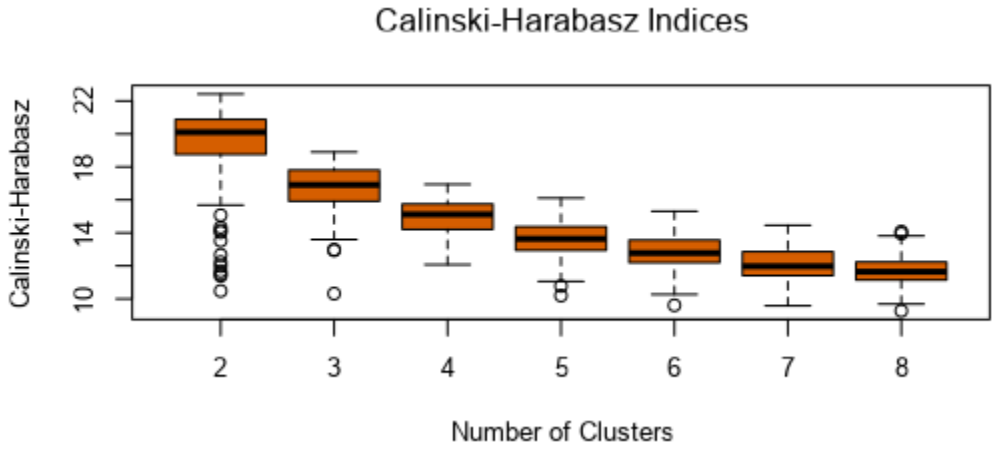

Figure 2: Calinski-Harabasz Indicies plot

2. How many stores fall into each store format?

   - The first and third clusters have 25 stores, while the second cluster contains 35 stores.

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Figure 3: Cluster information

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

   - Higher value in (Figure 4) means that the cluster is oriented towards selling more of that type of product.

     - Cluster 1 is oriented towards selling more of the following products:

       - Deli
       - Meat
       - Dry_Grocery
       - Bakery

     - Cluster 2 sells more of the following products:

       - Produce
       - Floral
       - Dairy
       - Frozen_Food
       - Bakery

     - Cluster 3 sells more of:

       - General_Merchandise
       - Dry_Grocery

	Dry_Grocery_Percent	Dairy_Percent	Frozen_Food_Percent	Meat_Percent	Produce_Percent	Floral_Percent	Deli_Percent
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	Bakery_Percent	General_Merchandise_Percent
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Figure 4: Cluster selling information

4. Tableau visualization that shows the location of the stores, where color shows clusters, and size shows total sales.

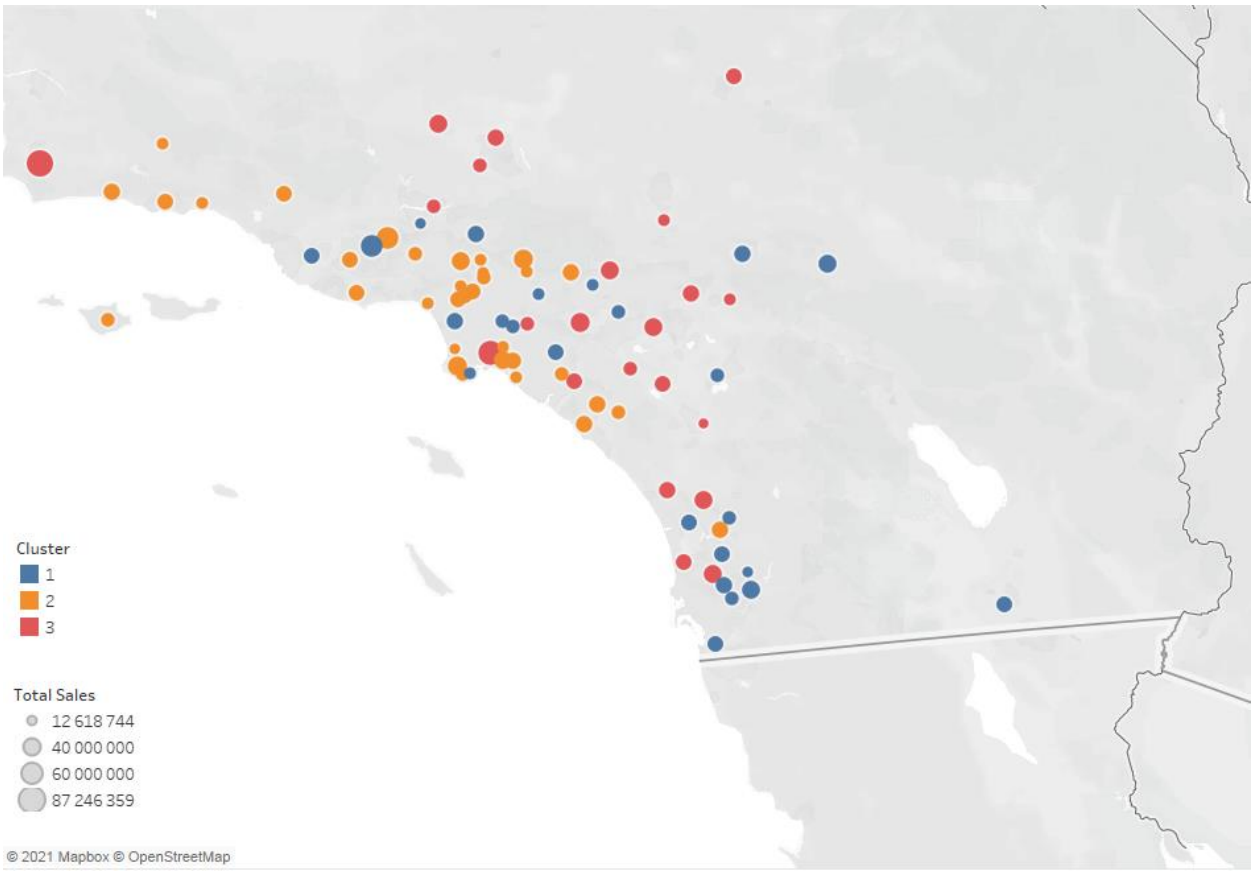

Figure 5: Existing stores map visualization

5. Alterxy Workflows

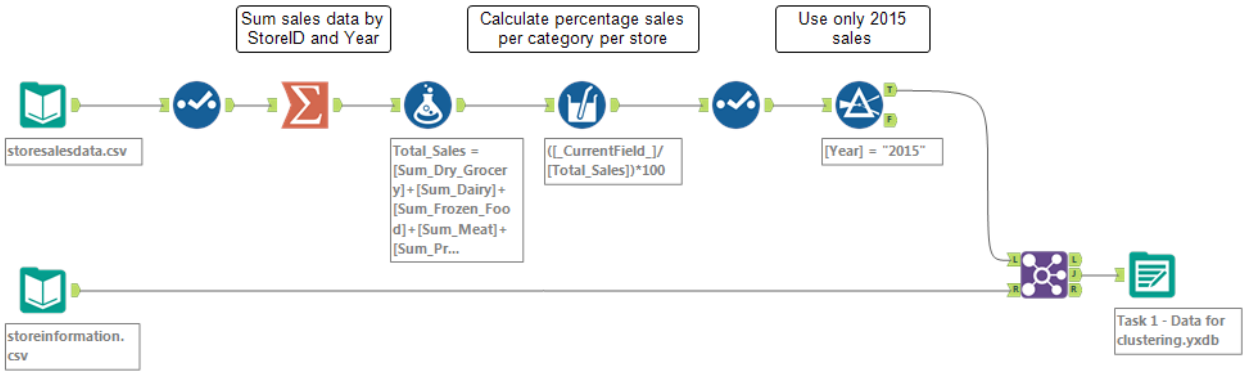

Figure 6: Preparing Data for Clustering Workflow

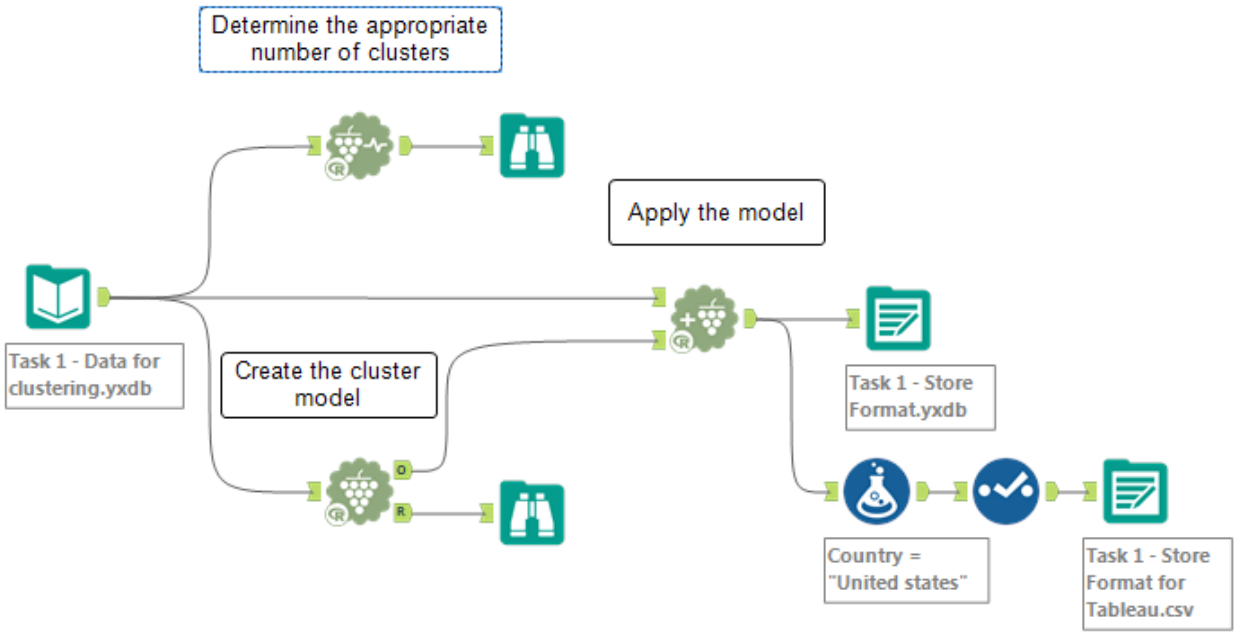

Figure 7: Determining store format for existing stores workflow

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?

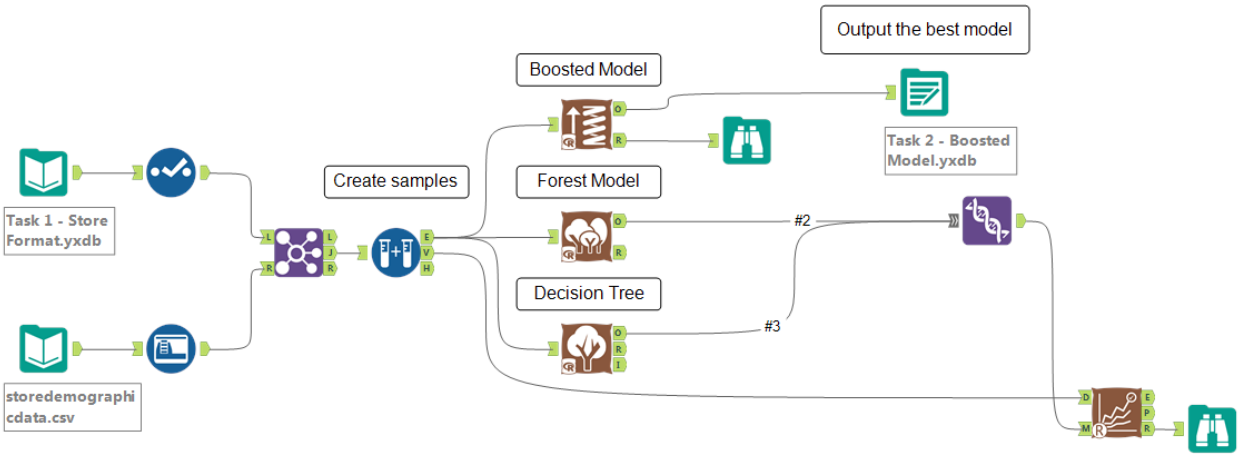

Figure 8: Training classification models workflow

- We can see in the below figure that the Boosted Model has the highest value in all measures and based on that it will be used for prediction.

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Boosted_Model	0.7647	0.8333	0.5000	1.0000	1.0000

Figure 9: Fit and error measures

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

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

Table 1: New stores segment

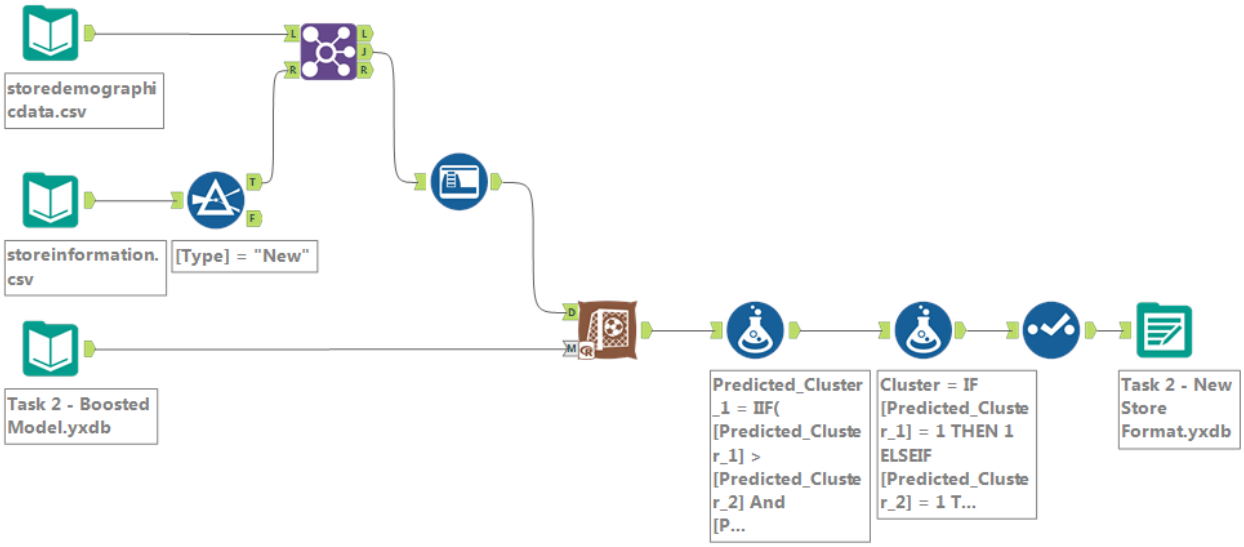

Figure 10: Predicting formats for new stores workflow

Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use for each forecast?

   - ETS model will be used to forecast since it has the lowest value in all measures (Figure 11) and it performs better than ARIMA in forecasting (Figure 12).

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

Figure 11: Comparison of accuracy measures

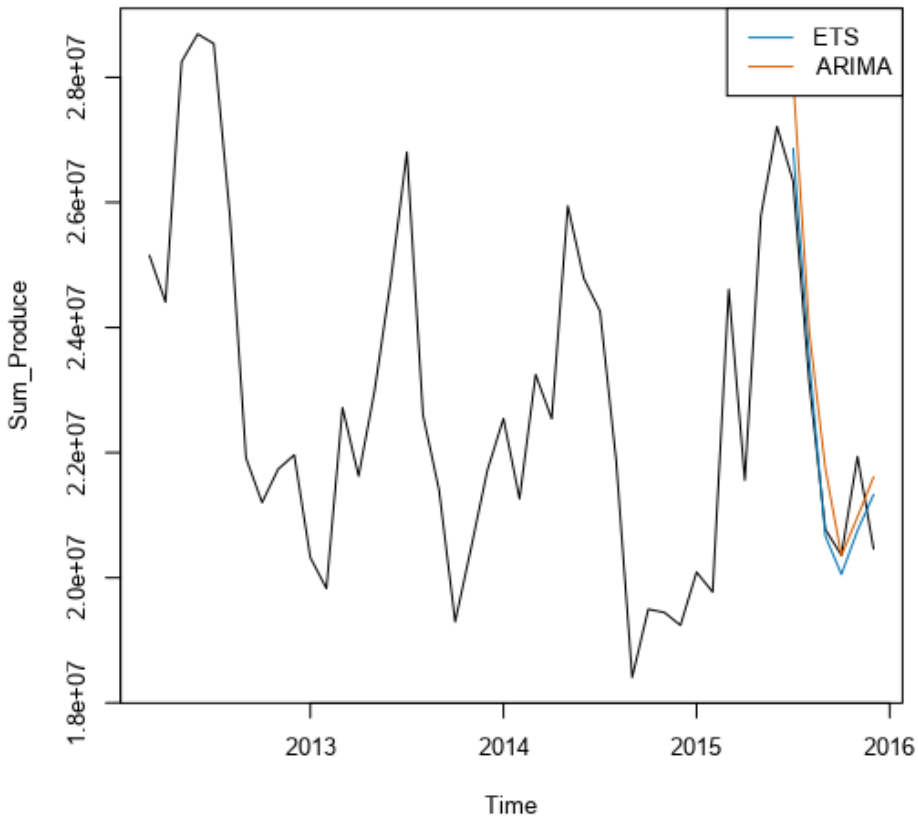

Figure 12: Actual and forecast values plot

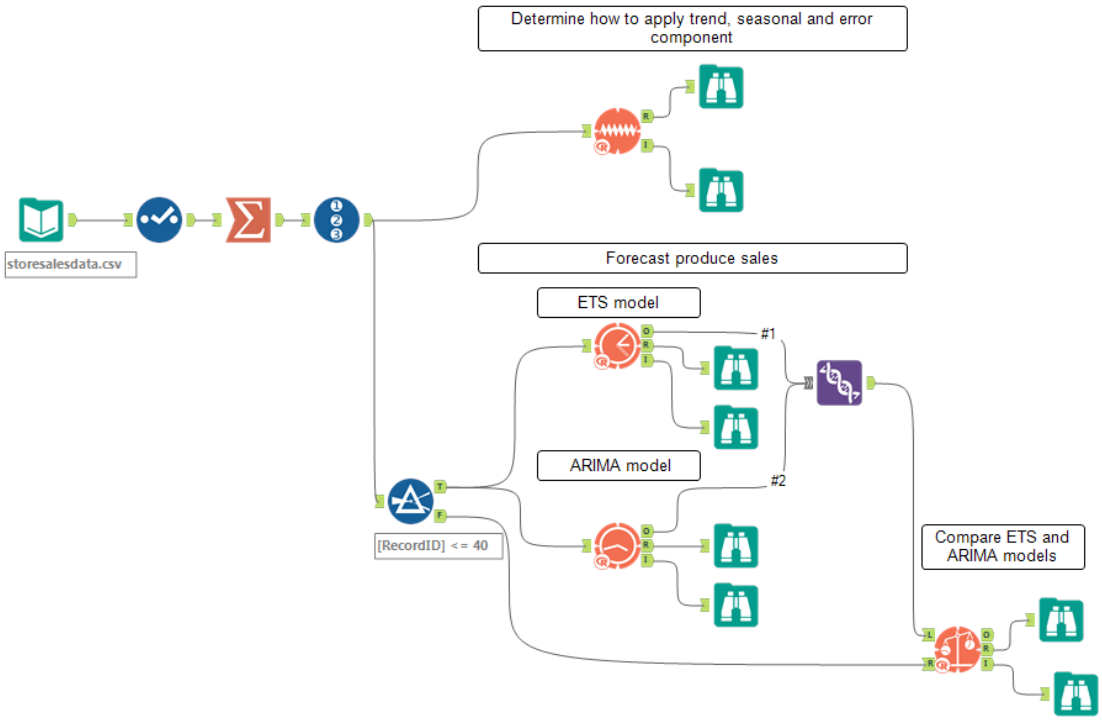

Figure 13: Comparing ETS and ARIMA models workflow

2. Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

   - From our decomposition plot (Figure 13) I have obtained the necessary information to define the best terms **ETS(M, N, M)**:

     - The trend line exhibits that there is no trend so we will use None.
     - The seasonality changes in magnitude each year, so a multiplicative method is necessary.
     - The error changes in magnitude as the series goes along so a multiplicative method will be used.

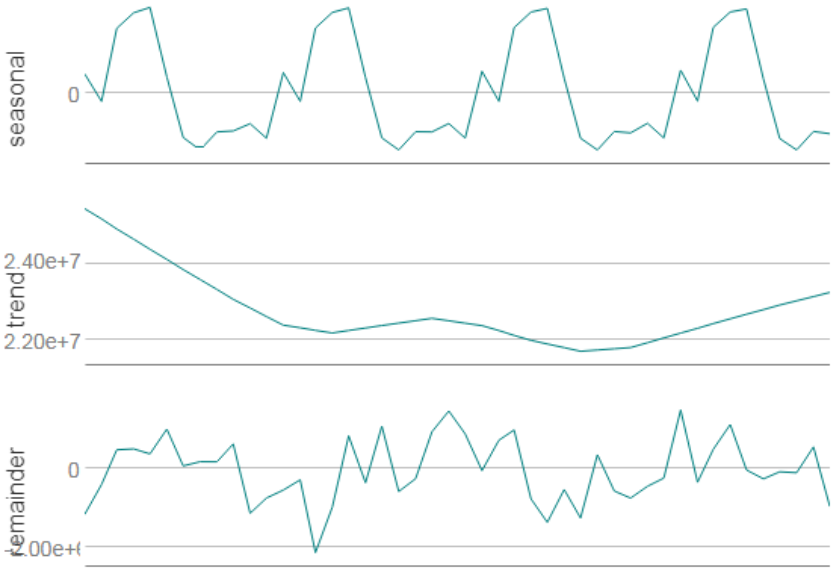

Figure 14: Decomposition plot

3. Table of forecasts for existing and new stores:

Month	New Stores	Existing Stores
Jan-16	2563357.91	21136641.78
Feb-16	2483924.73	20507039.12
Mar-16	2910944.15	23506565.98
Apr-16	2764881.87	22208405.76
May-16	3141305.87	25380147.77
Jun-16	3195054.20	25966799.47
Jul-16	3212390.95	26113792.57
Aug-16	2852385.77	22899285.77
Sep-16	2521697.19	20499583.91
Oct-16	2466750.89	19971242.82
Nov-16	2557744.59	20602665.92
Dec-16	2530510.81	21073222.08

Table 2: Forecasts for existing and new stores

Visualization of forecasts that includes historical data, existing stores forecasts, and new stores forecasts:

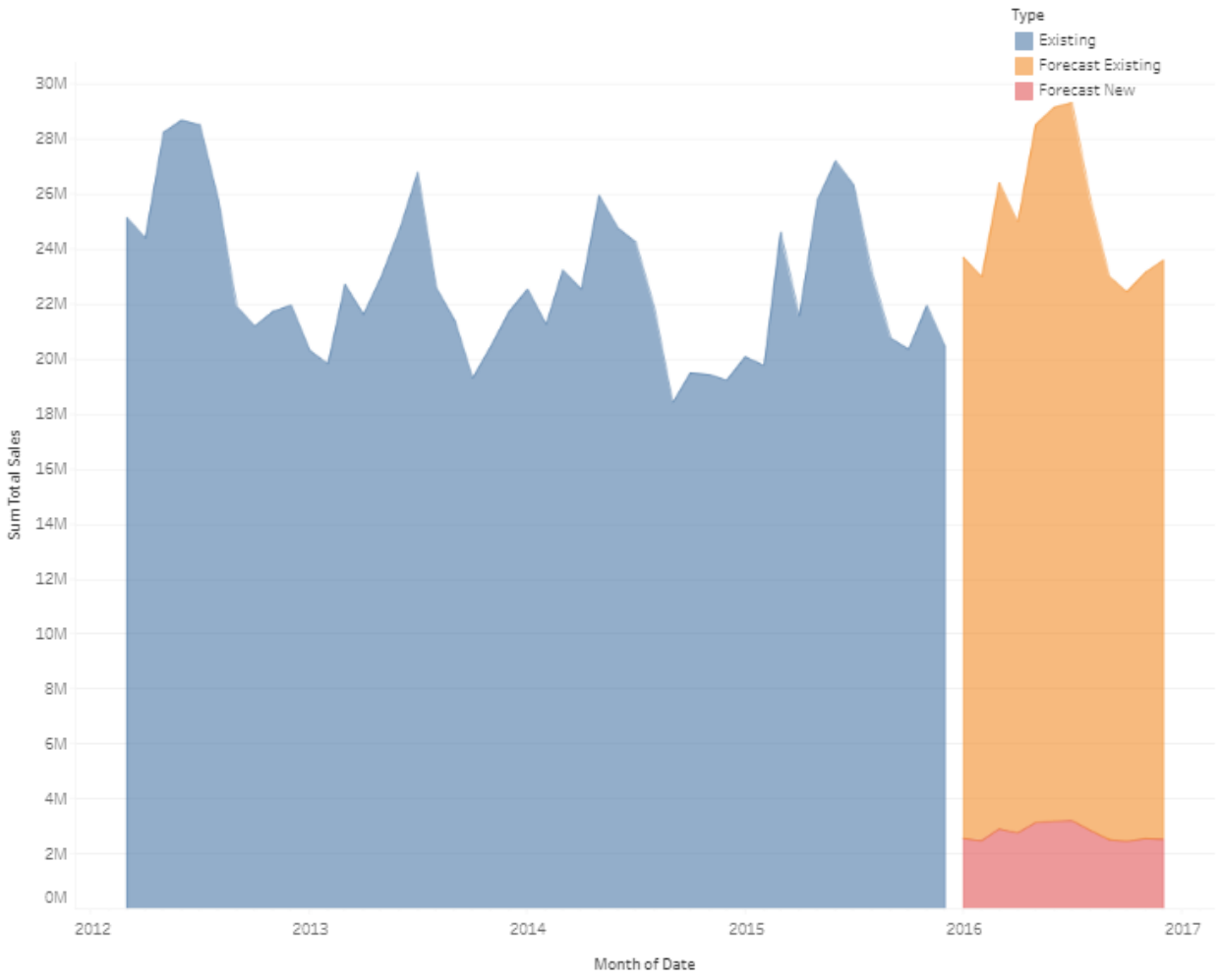

Figure 15: Visualization of forecasts

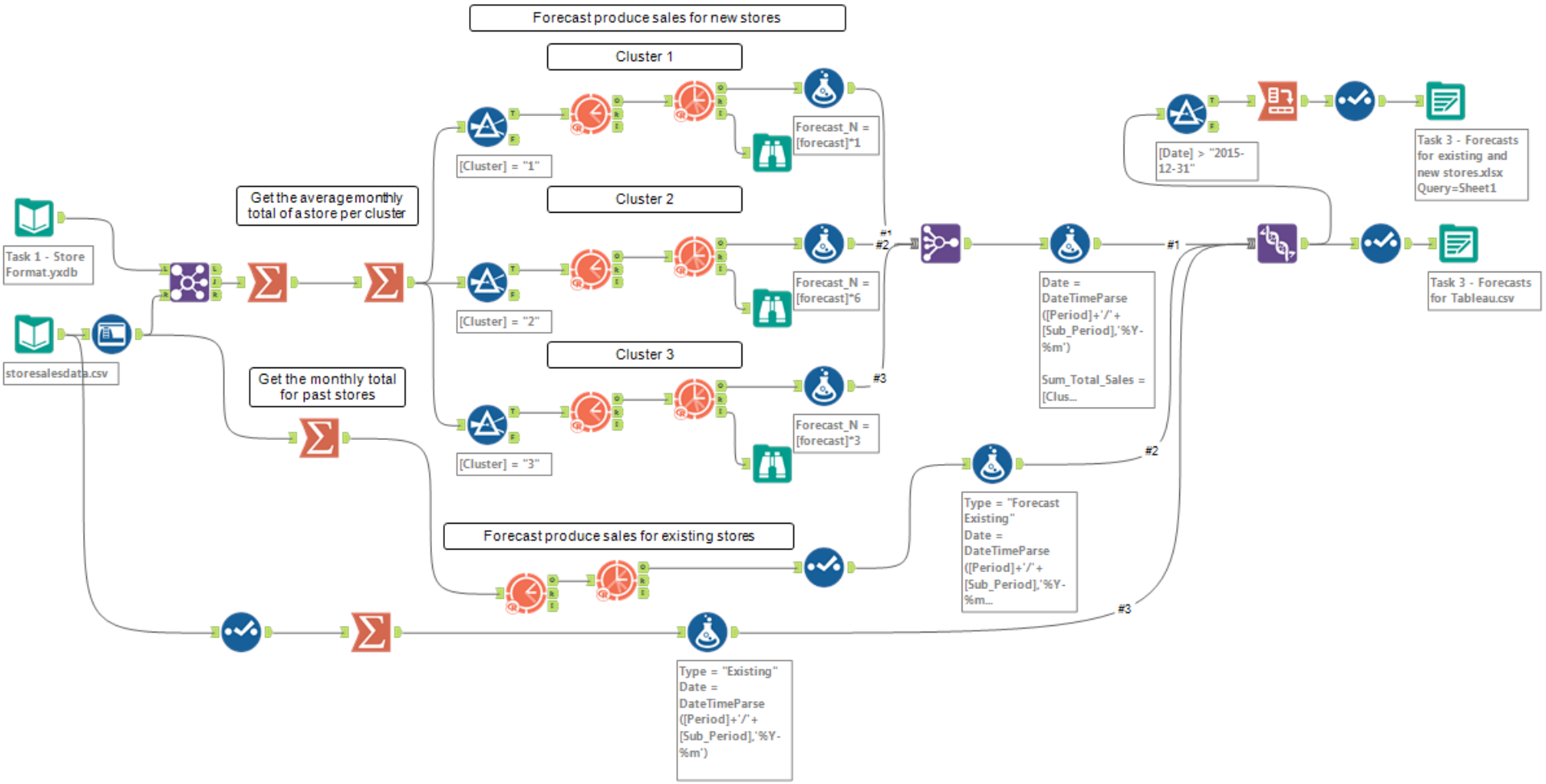

Figure 16: Predicting produce sales workflow