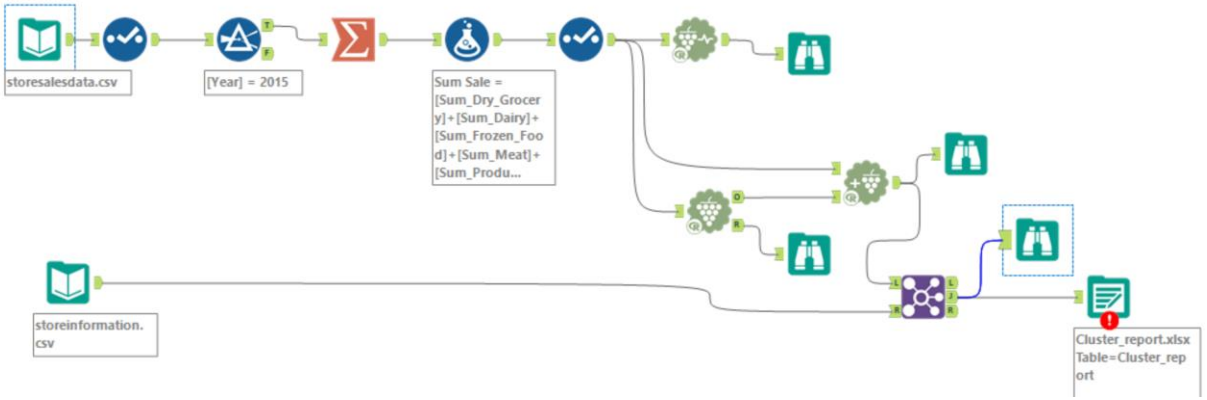
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

From the K-Means cluster assessment Report We see that the Adjusted Rand Indices and Calinski-Harabasz Indices have the highest Median Value at 3. We will utilize this as the number of store formats.



ALTERYX workflow for K-Means analysis

Report

K-Means Cluster Assessment Report

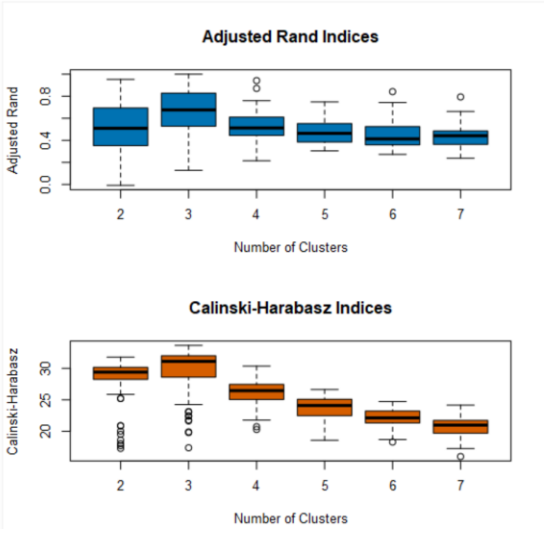
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	-0.007972	0.127969	0.214585	0.304172	0.272751	0.23775
1st Quartile	0.352204	0.53349	0.445896	0.386359	0.360333	0.365253
Median	0.509443	0.675694	0.512646	0.463469	0.412831	0.441642
Mean	0.506355	0.680038	0.532528	0.474152	0.446208	0.434426
3rd Quartile	0.684321	0.821873	0.607343	0.552362	0.520697	0.481966
Maximum	0.95293	1	0.942222	0.748402	0.841889	0.793694

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	17.281	17.38103	20.28456	18.57302	18.29328	15.98702
1st Quartile	28.24847	28.60331	25.05214	22.52428	21.3307	19.74744
Median	29.4024	31.11034	26.44701	24.10317	22.13277	20.98982
Mean	28.5332	29.86293	26.23466	23.72781	22.02897	20.75389
3rd Quartile	30.15462	31.97573	27.42205	25.07512	23.18308	21.72164
Maximum	31.78345	33.63781	30.35916	26.63063	24.72038	24.15086



K-Means Centroid from the K-Centroids Diagnostics Tool in Alteryx.

2. How many stores fall into each store format?

   The summary report shows us that we have Cluster1 - 23, Cluster2 - 29 and Cluster 3 – 33.

Report

Summary Report of the K-Means Clustering Solution K_Centroid

Solution Summary

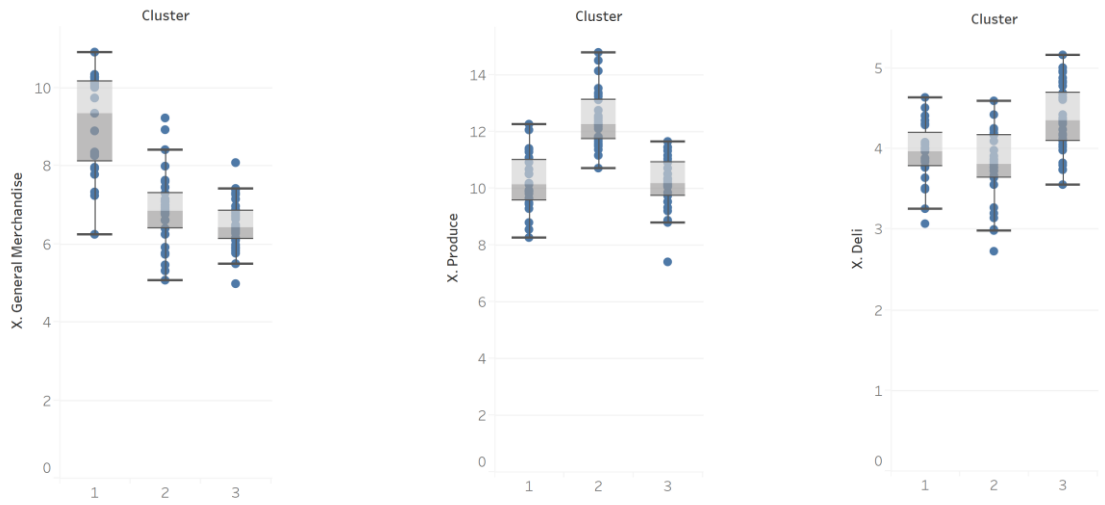
Call:

stepFlexclust(scale(model.matrix(~-1 + X._Dry_Grocery + X._Dairy + X._Frozen_Food + X._Meat + X._Produce + X._Floral + X._Deli + X._Bakery + X._General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

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

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



We can see from the Box and whiskers plot that the % of General Merchandise has a higher median for cluster 1 as compared to cluster 2 and 3. We also see that cluster 2 has a higher median compared to cluster 3.

Further analysis also shows us that % Produce has an over all greater median for cluster 2 as compared to both cluster 1 and 3.

Also, Cluster 3 has the % Deli highest median as compared to store 1 and 2.

Box and Whisker Plots

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.

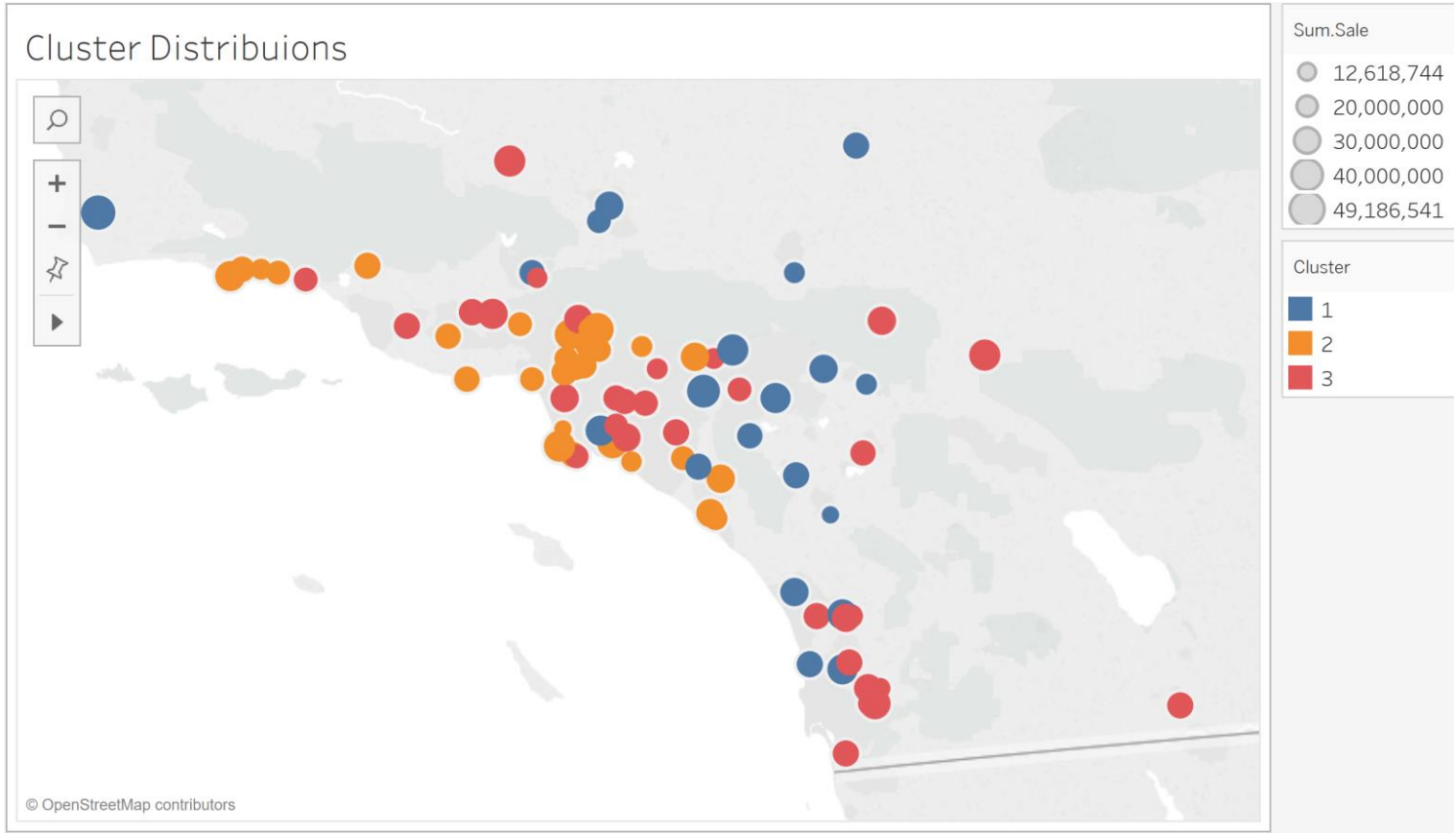


Tableau Public Link: https://public.tableau.com/profile/salman.syed2325#!/vizhome/ClusterbasedStoreLocations/Sheet1

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.)

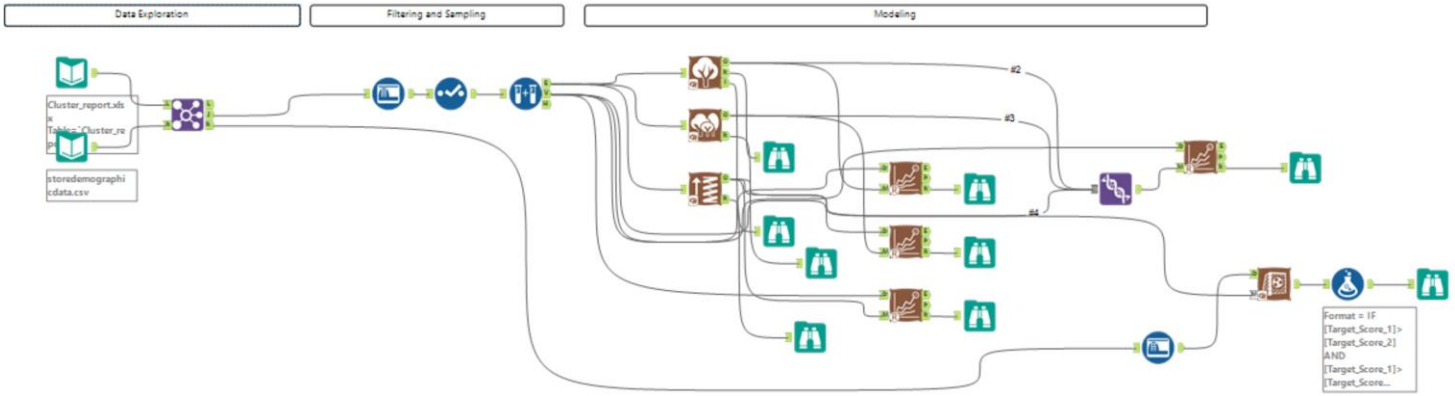
To determine the best store format, we tested out 3 different classifier models (Decision tree, Random Forest, Boosted Model). The Validation set was selected at 20% of the entire sample with Random Seed of 3. The initial data sets used are the calculated cluster report from Task 1 which contains the stores classified into 3 different clusters. This data was joined with the demographics data. From the Model comparison report the Boosted Method was used as it had the highest F1 score compared to the other two models. It also shows an overall high accuracy and no bias.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DecTree	0.7059	0.7685	0.7500	1.0000	0.5556
RandomForest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.



ALTERYX workflow for model selection.

We see from the values of the confusion matrix that there are no biases and the overall accuracy for the Boosted model surpasses the other two. We will utilize the boosted mode for our classification on the 10 new stores.

Confusion matrix of Boosted

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

Confusion matrix of DecTree

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

Confusion matrix of RandomForest

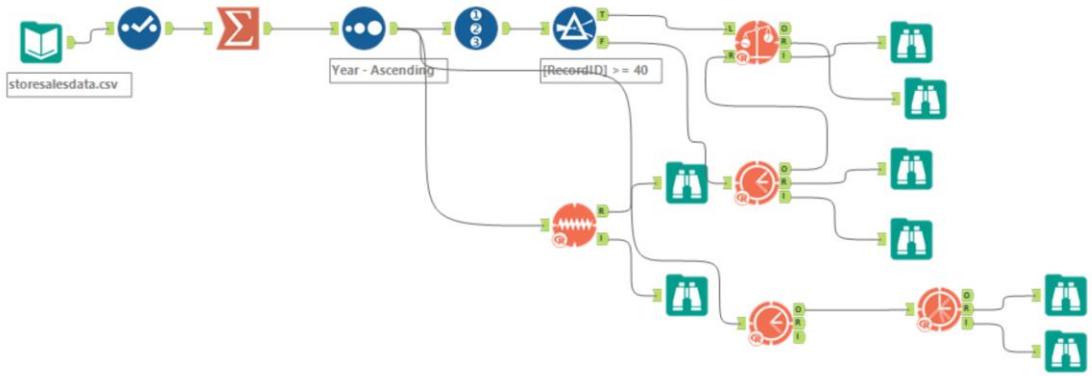
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

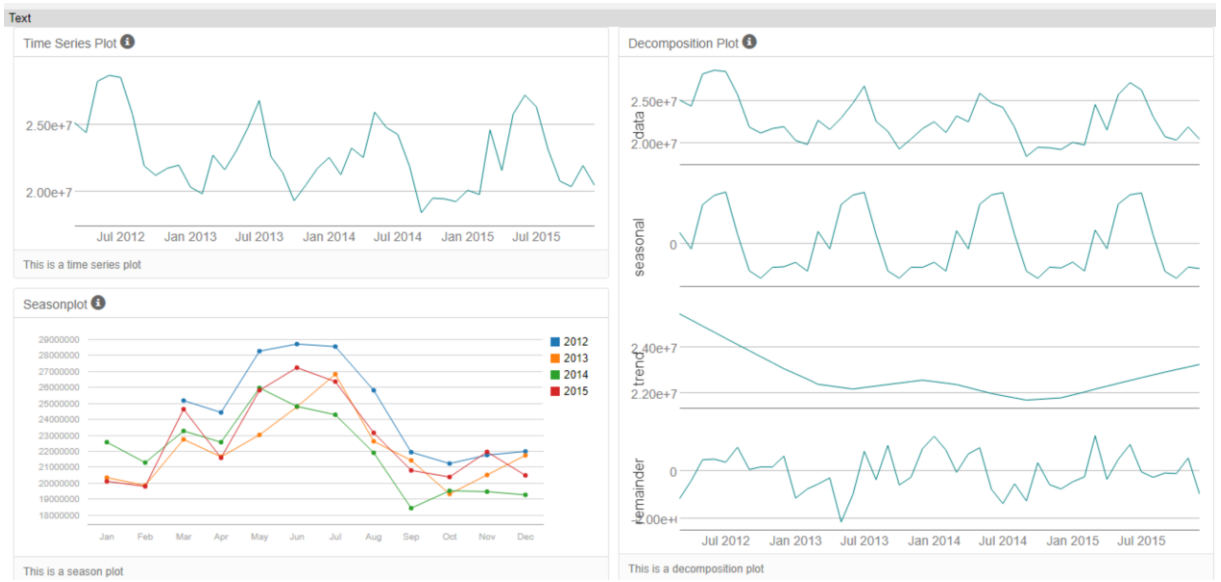
Task 3: Predicting Produce Sales

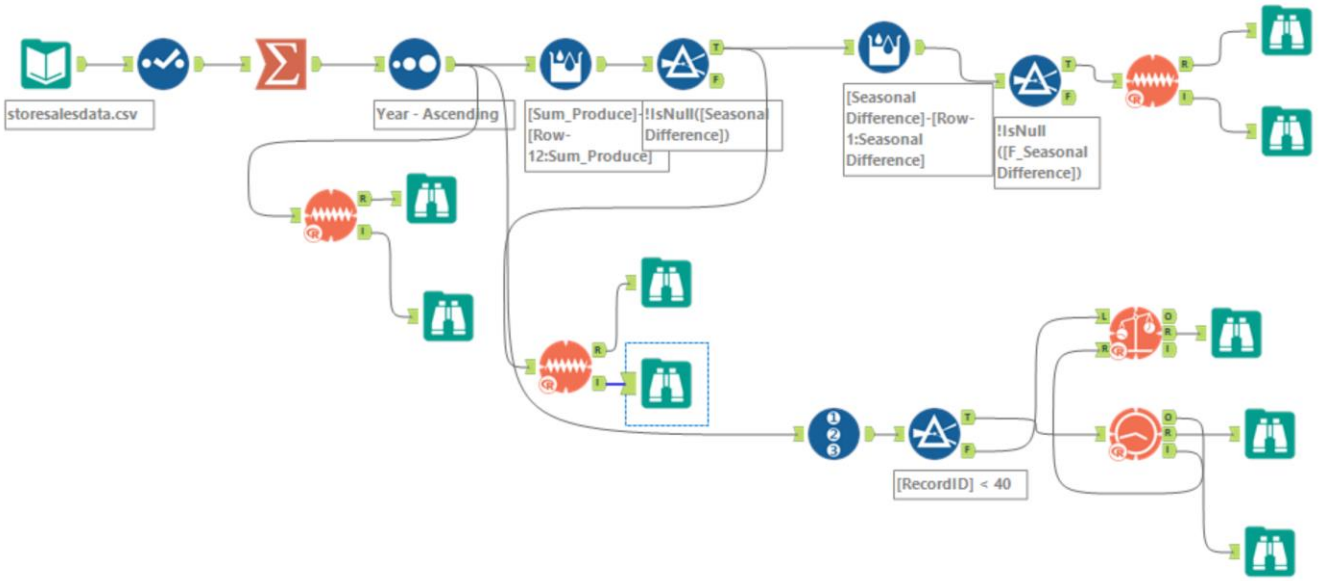
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?



ALTERYX workflow for ETS(MNM) forecasting model.

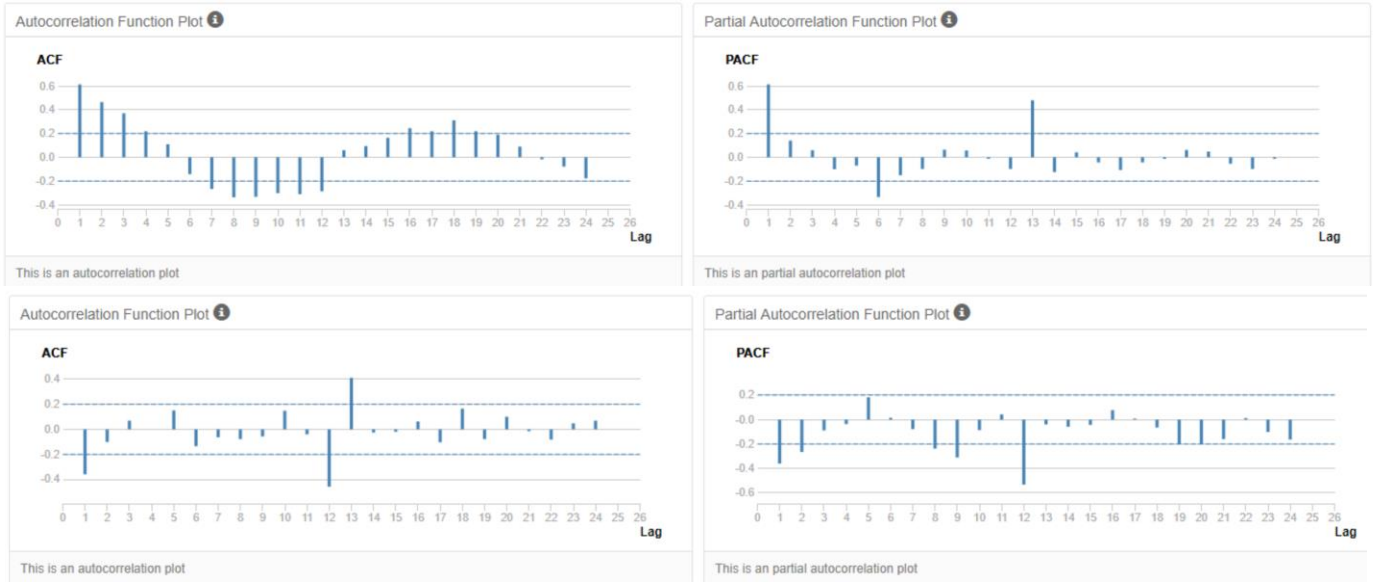
We utilize a time series plot to analyze the components of the ETS model. We see that the **ETS (MNM)** model is best suited for our analysis. The Error should be applied multiplicative as it is a randomized pattern, the trend has no pattern and will not be applied, the seasonal trend is an incremental over the season and will be applied multiplicative. In this case we have verified using an auto model.





ALTERYX workflow for ARIMA (1,0,0) (1,1,0)[12] forecasting model.

The First difference is taken for seasonality to remove any seasonal variation for our analysis purposes. The auto function recommends an ARIMA (1,0,0) (1,1,0)[12] which agrees with our assessment of the ACF and PACF plots from the previous class.



From the Summary report for both ETS (MNM) and ARIMA(1,0,0)(1,1,0)[12].

We look at the RMSE, which shows the in-sample standard deviation, and the MASE which we use for the comparison of forecasts of different models. Based on the RMSE we can see that our variance is about 1020596 units for ETS(MNM) around the mean as compared to 1042209 for ARIMA(1,0,0)(1,1,0)[12]
The MASE shows a good forecast at 0.45 with its value falling below the generic 1.0, the MASE threshold for model accuracy." We also compare the AIC for both and ETS(MNM) has a higher value compared to ARIMA(1,0,0)(1,1,0)[12].

Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

Report

Summary of ARIMA Model ARIMA

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

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

3. 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.

Aggregations for Forecasting
For existing stores, we get the monthly total for past stores before forecasting. We accomplish this using a summarize tool.

Group by Year
Group by Month
Sum Produce.

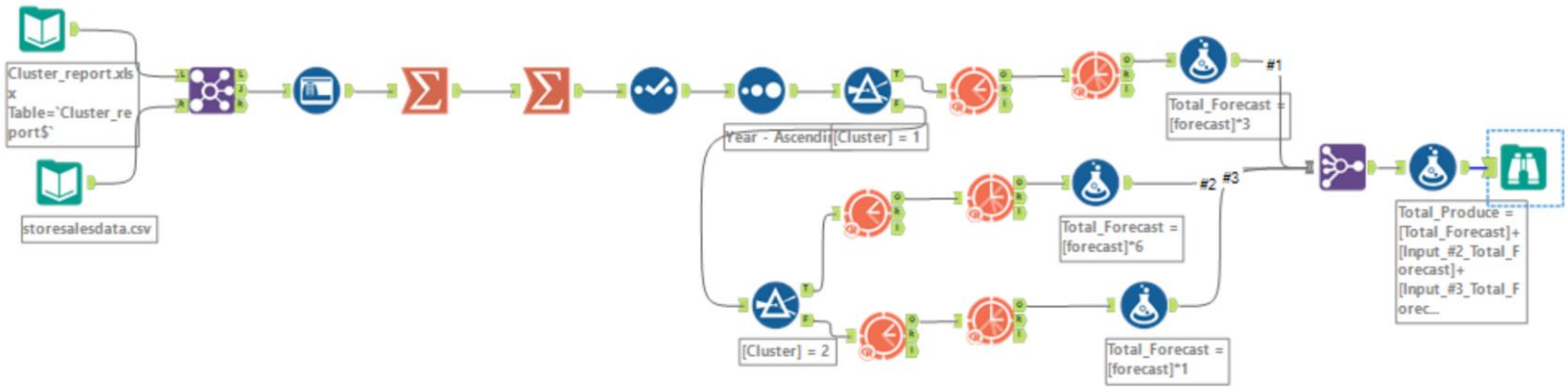
Now we determine the best forecasting model and then forecast the next 12 months for the existing stores.

For new stores we get the average monthly total of a store per cluster. We accomplish this using 2 summarize tools.

Summarize 1
Group by Store
Group by Cluster
Group by Year
Group by Month
Sum Produce.

Summarize 2
Group by Cluster
Group by Year
Group by Month
Avg Sum_Produce.

We will be forecasting for each of the clusters and then multiplying the results by the number of new stores in that cluster. Then we will be adding all of these forecasts together on the same months to get a total forecast for all the new stores.

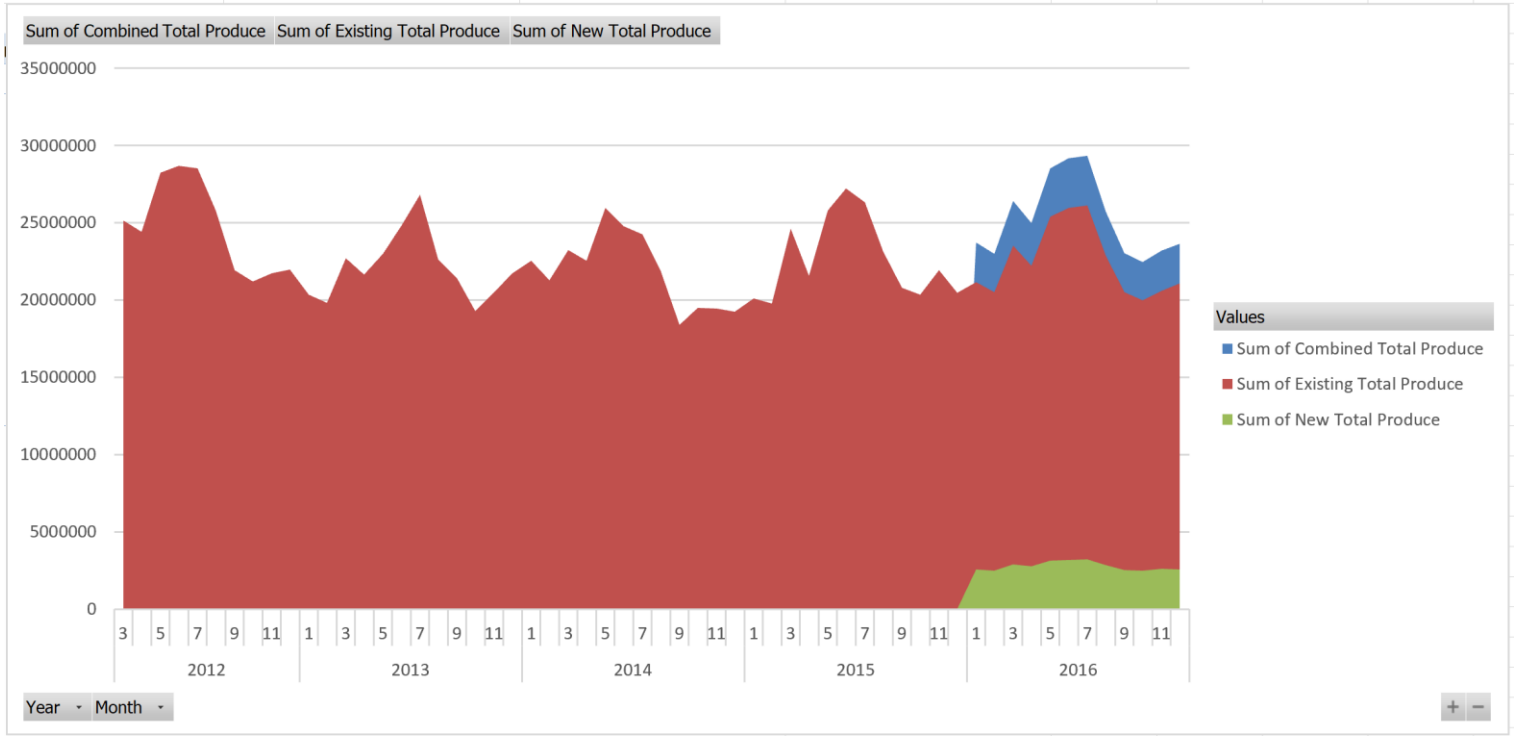


ALTERYX workflow to forecast the Cost of Produce for 10 new stores for Year 2016.

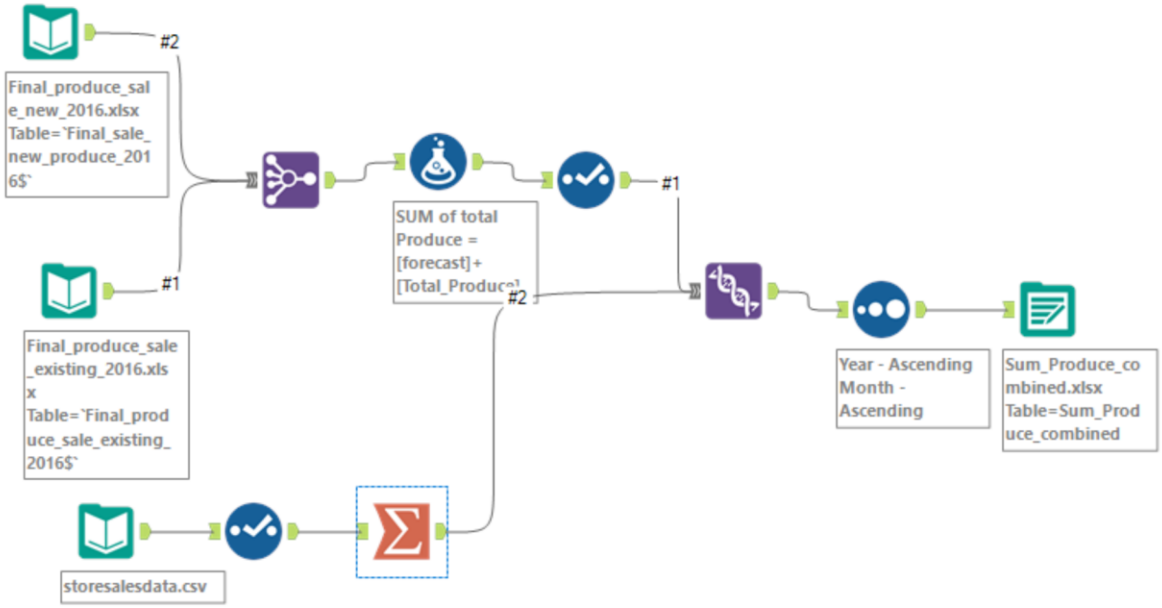
The following formula was used to calculate the Total sales of the stores.

[Total_Forecast]+[Input_#2_Total_Forecast]+[Input_#3_Total_Forecast]

Year	Month	Sum_New_stores	Sum_Existing_stores	Sum_Combined_stores
2016	1	2584384	21136208	23720592
2016	2	2470874	20506605	22977479
2016	3	2906308	23506131	26412439
2016	4	2771532	22207971	24979503
2016	5	3145849	25376698	28522547
2016	6	3183909	25963559	29147469
2016	7	3213978	26113357	29327335
2016	8	2858247	22904672	25762919
2016	9	2538174	20499151	23037325
2016	10	2483550	19970809	22454359
2016	11	2593089	20602232	23195321
2016	12	2570200	21072787	23642987



The chart shows us the combined gain on Produce as compared to current.



Alteryx Workflow to calculate sum of combined sales.

Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.