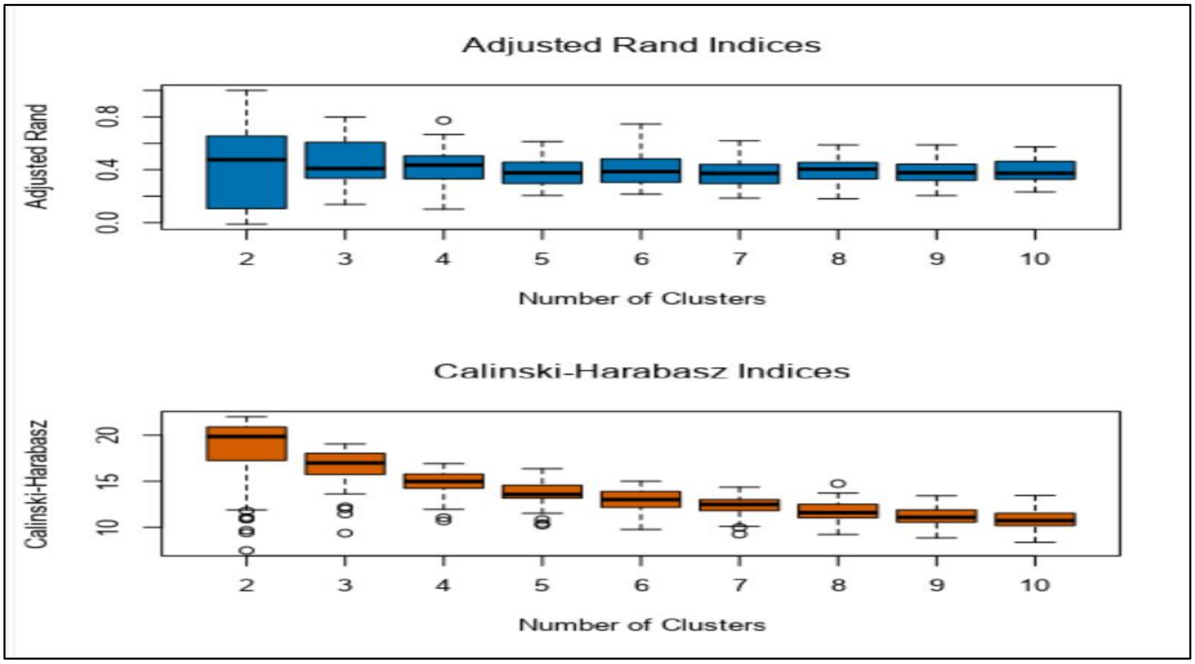
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   The optimal number of store formats is three. To arrive at this number, I analyzed the data using K-mean clustering. The variables used for clustering were the percentage of sale by category to the total sale.

   I conducted internal validation on the clustering to determine the compactness and the distinctness of the clustering.
   I used AR and CH index for this purpose. The range of clustering used for validation was 2 to 10.
   Based on the indices, cluster numbers two and three had the highest index values. But cluster 2 showed more variability with a higher IQR.
   Thus, I decided to go ahead with three as the number for store format.



AR & CH Indices

2. How many stores fall into each store format?

Cluster Number	Number of Stores
1	23
2	29
3	33

Report

Summary Report of the K-Means Clustering Solution ClusterByPctCategory

Solution Summary

Call:

```
stepFlexclust(scale(model.matrix(~-1 + Pct_Dry_Grocer + Pct_Dairy + Pct_Frozen_Food + Pct_Meat + Pct_Produce + Pct_Floral + Pct_Deli + Pct_Bakery + Pct_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

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

	Pct_Dry_Grocer	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_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

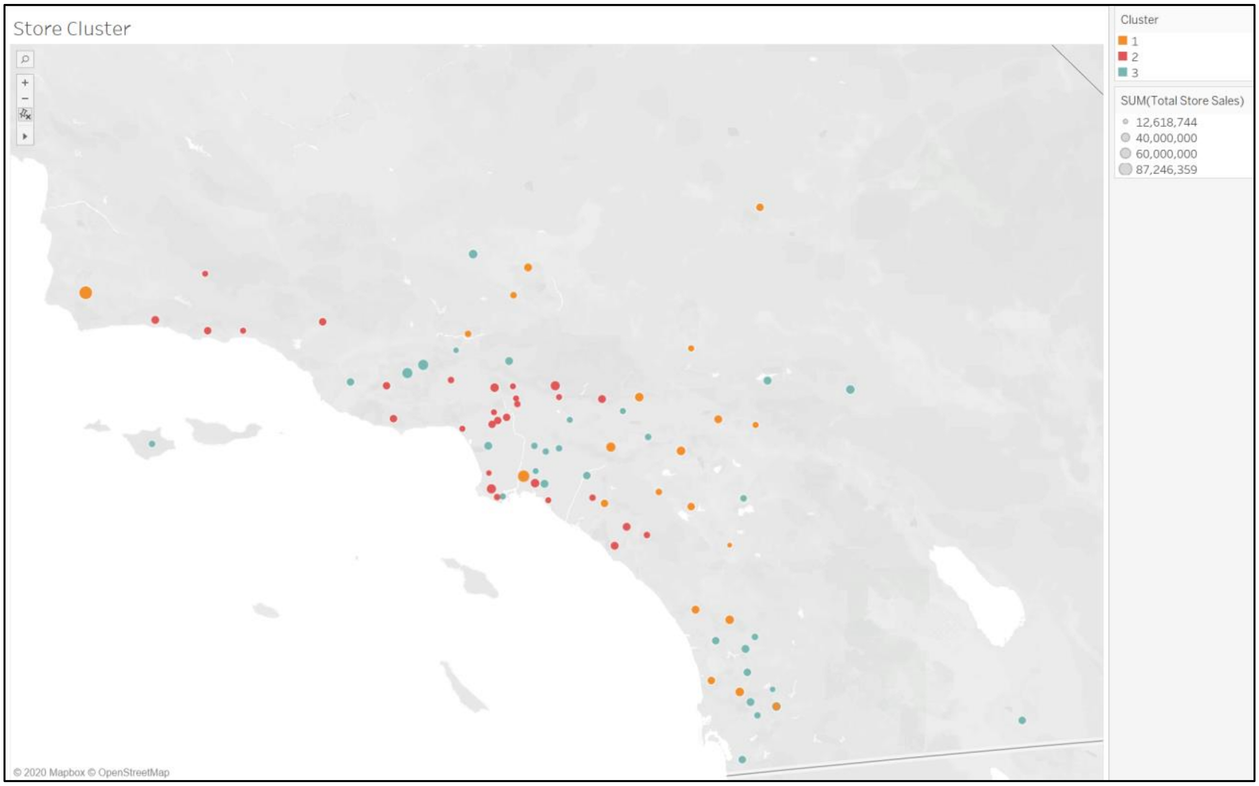
	Pct_Bakery	Pct_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

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

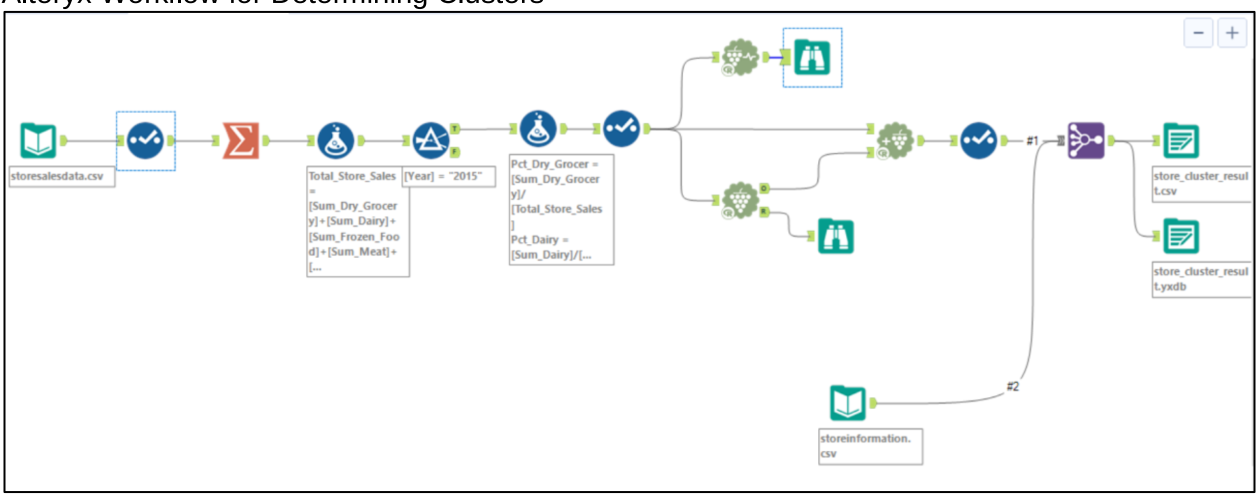
Cluster three has more in General Merchandise
Cluster two has more sale in Dairy, Frozen Food, Produce, Floral, and Bakery.
Cluster three has more sales in Dry Grocery, Meat, and Deli.

Cluster	Pct_Dry_Grocer	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_Deli	Pct_Bakery	Pct_General_Merchandise
1	0.33	-0.76	-0.39	-0.09	-0.51	-0.30	-0.23	-0.89	1.21
2	-0.73	0.70	0.35	-0.49	1.01	0.85	-0.55	0.40	-0.30
3	0.41	-0.09	-0.03	0.49	-0.54	-0.54	0.65	0.27	-0.57

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.



Alteryx Workflow for Determining Clusters



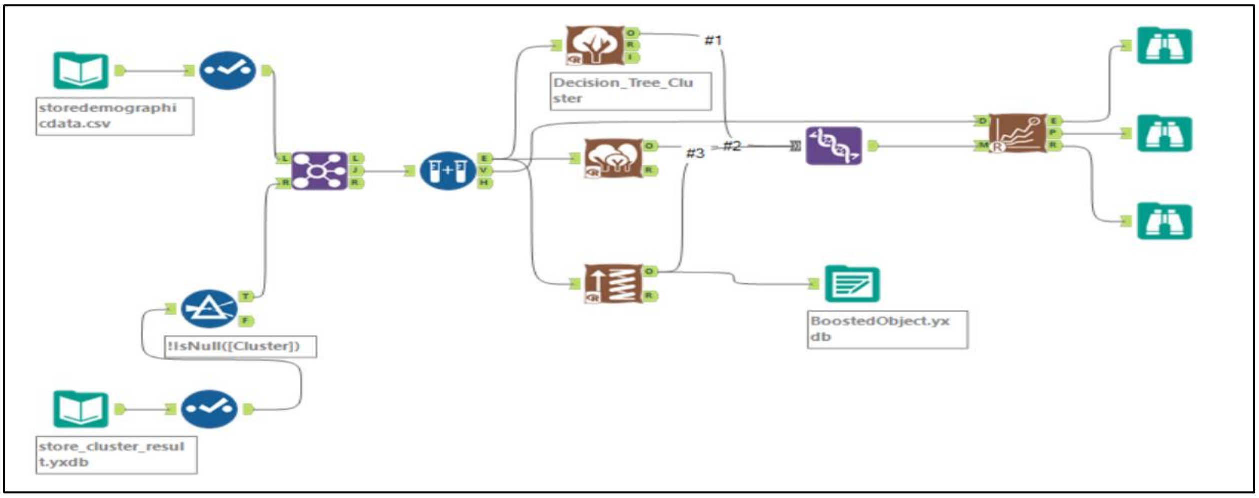
Task 2: Formats for New Stores

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.)

   We want to determine segmentations for the new stores. Hence, the methodology used to design the experiment should be a non-binary classification model. I will compare the Decision Tree, Random Forest, and Boosted Classification Model and use the model that best fits the data.

   I compared the models' output using the model comparison tool. Based on the F1 score and the confusion matrix, the Boosted Classification model does the best job out of the three models.

   We will score the model using this methodology.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Cluster	0.7059	0.7685	0.7500	1.0000	0.5556
RandomForrest_Cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Cluster	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 Boosted_Cluster

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

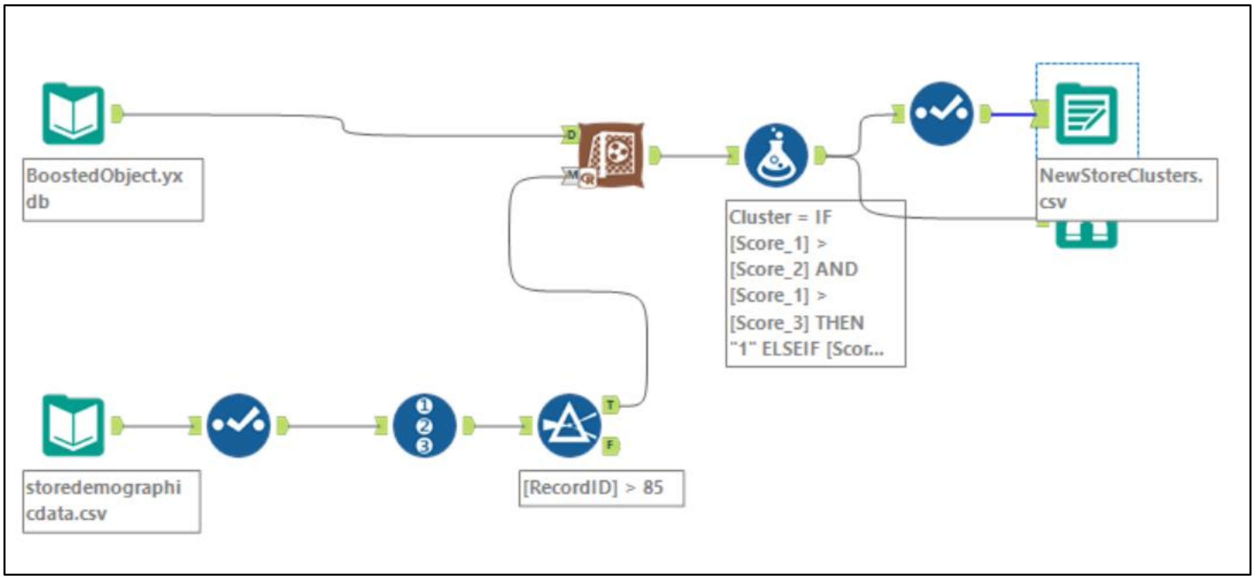
Confusion matrix of Decision_Tree_Cluster

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

Confusion matrix of RandomForrest_Cluster

	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.
Once I got the boosted object, we used the scoring tool to score the model. I assigned clusters, based on the highest score.



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?

Training ETS Model:

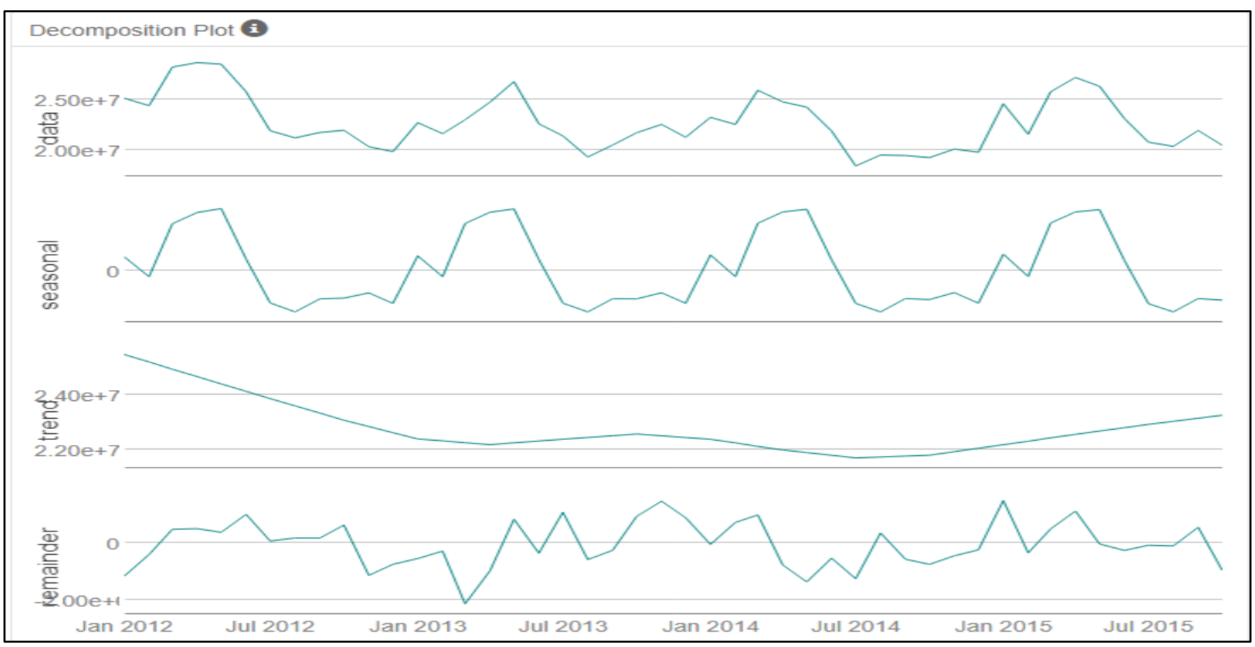
The target variable to forecast using the ETS model is Revenue of produce by year and month.

I have 46 records in total to train and validate the model. I used the first 40 records to train the model and the last 6 records to validate the model.

To examine the three components of time series error, trend, and seasonality, we build a time series decomposition plot.

Error: The error component is present and is multiplicative.
Trend: The trend component is absent.
Seasonality: The seasonal component is present. However, after initial observations of the seasonal component, I was leaning towards using it additively. But, if I let the model run on auto settings, it chooses a multiplicatively. Thus, I am going to use seasonality multiplicatively.

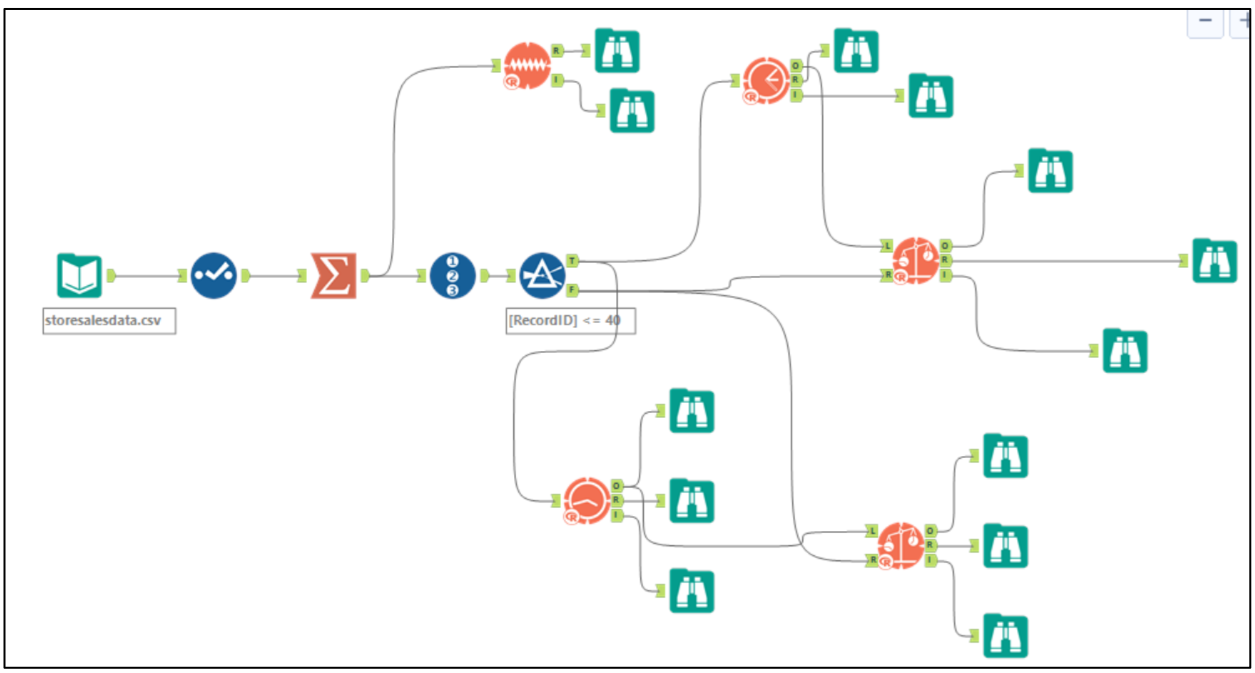
So we are going to use an **ETS(M,N,M)** model to forecast the time series.



Training ARIMA Model:
We are going to use the same training and validation sample for training the ARIMA model.

We already know that the time series has a trend and seasonal component multiplicatively. While the trend component is absent.
So, we are going to be using the Seasonal ARIMA model to forecast the time series. The seasonal ARIMA models are denoted (p,d,q)(P,D,Q)m. Where m refers to the number of periods in each season and P,D,Q refers to autoregressive, differencing and moving average term for the seasonal part of the ARIMA.

We are going to use **ARIMA (1,0,0) (1,1,0) [12]** model. So, we are using a lag of 1 for the autoregressive and one for the differencing,



Comparing the ETS(M,N,M) and ARIMA (1,0,0) (1,1,0) [12] model.

After comparing the forecasted error measurements for both the models, ETS has a lower forecast error measurement compared to the ARIMA. Thus going ahead we will be using **ETS(M,N,M) model to forecast the revenue**.

ETS(M,N,M) Result

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

Smoothing parameters:

Parameter	Value
alpha	0.674884
gamma	0.000203

Initial states:

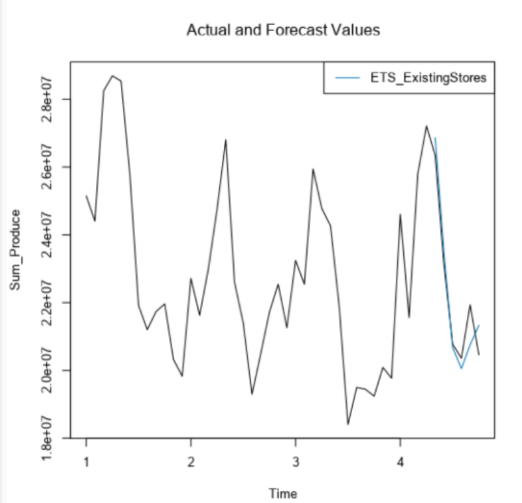
State	Value
l	23146230.586012
s0	0.90906
s1	0.938619
s2	0.926304
s3	0.901291
s4	0.870972
s5	0.897637
s6	1.019225
s7	1.166556
s8	1.167388
s9	1.137259
s10	0.997793

Actual and Forecast Values:

Actual	ETS_ExistingStores
26338477.15	26860639.57444
23130626.6	23468254.49595
20774415.93	20668464.64495
20359980.58	20054544.07631
21936906.81	20752503.51996
20462899.3	21328386.80965

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_ExistingStores	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257



ARIMA (1,0,0) (1,1,0) [12] Result

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

Call:
auto.arima(Sum_Produce, max.p = 2, max.q = 2, max.P = 1, max.Q = 1, ic = "aicc", allowdrift = TRUE)

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

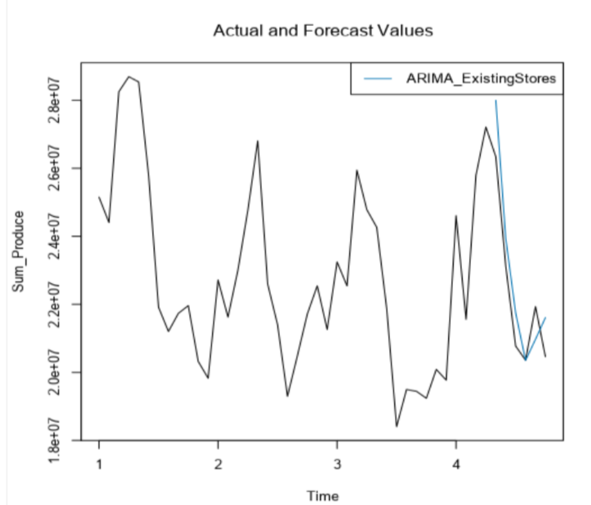
Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

Actual and Forecast Values:

Actual	ARIMA_ExistingStores
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_ExistingStores	-604232.3	1050239	928412	-2.6156	4.0942	0.5463



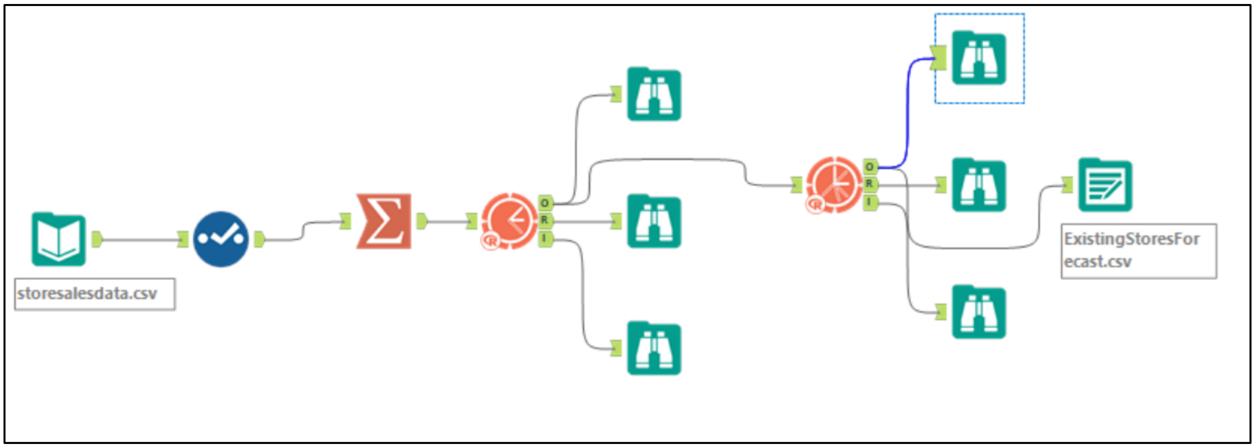
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

YearMonth	ExisitingStores	NewStores
2016-01	21,829,060.03	2,588,356.56
2016-02	21,146,329.63	2,498,567.17
2016-03	23,735,686.94	2,919,067.02
2016-04	22,409,515.28	2,797,280.08
2016-05	25,621,828.73	3,163,764.86
2016-06	26,307,858.04	3,202,813.29
2016-07	26,705,092.56	3,228,212.24
2016-08	23,440,761.33	2,868,914.81
2016-09	20,640,047.32	2,538,372.27
2016-10	20,086,270.46	2,485,732.28
2016-11	20,858,119.96	2,583,447.59
2016-12	21,255,190.24	2,562,181.70

Forecasting Revenue for Existing Stores:

I grouped the revenue for produce by Year and Month. Then used ETS(M,N,M) model to forecast the revenue for the year 2016.

Alteryx Workflow for forecasting revenue for Existing Stores.



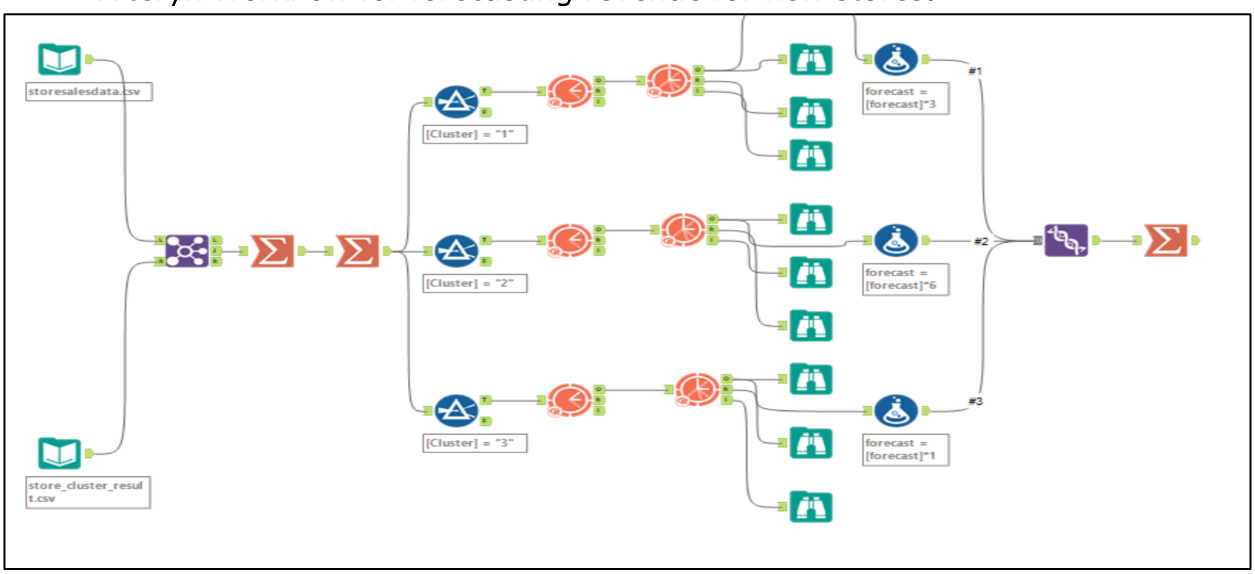
Forecasting Revenue for New Stores:

For new stores we are going forecast revenue by getting the average monthly revenue of a store per clusters.

To achieve this, we calculated the total revenue for produce grouped by store, cluster, year and month. Then calculated the average monthly revenue for a store by cluster by grouping the data by cluster, year and month.

After we got the average monthly forecasted revenue, we had to multiple the forecasted revenue with the number of stores in each cluster.

Alteryx Workflow for forecasting revenue for new stores.



Combining the data for visualization

We then combined the data from historical sales, forecasted sales for existing stores and forecasted sales for new stores for data visualization.

Alteryx Workflow for appending data

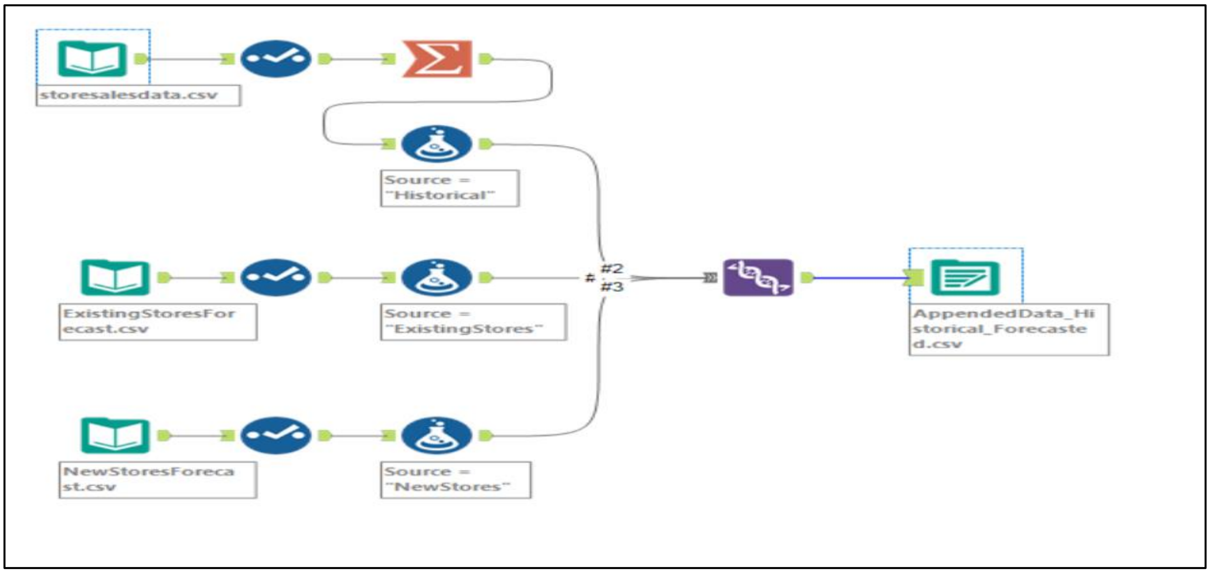


Tableau Visualization