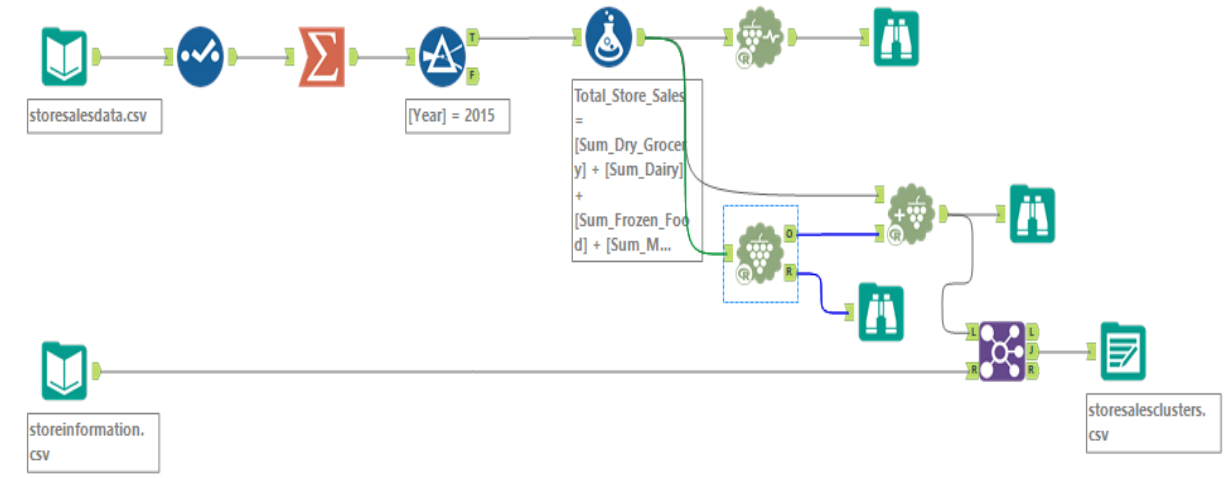
Project: Predictive Analytics Capstone

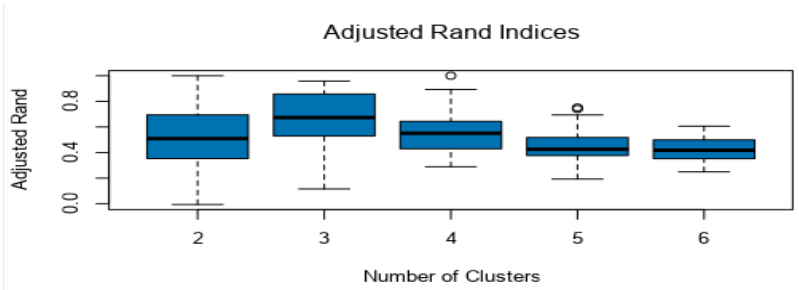
Ans Jayan
ansjayan@msn.com

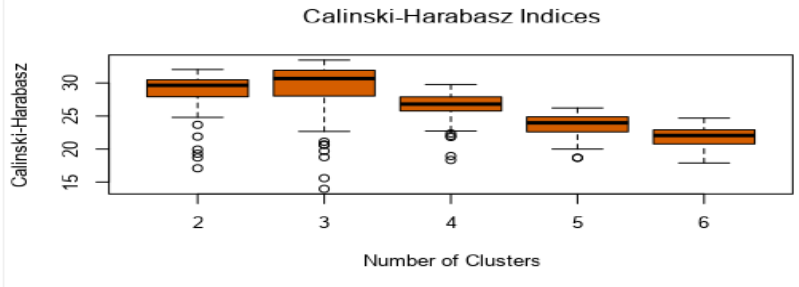
The company runs 85 grocery stores currently. All stores use the same format for selling products and company ships the same amount of product to all stores. Some stores have product surplus and some has shortage. Company is planning to open 10 more stores. So, the company want to make decisions about store formats and planning inventory.

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 format is 3. Using given data, a k-means clustering model is created.







From the adjusted rand indices, higher the index better the stability. For calinski-

harabasz indices, higher the index better the distinctness and compactness of clusters. Required is Median high and spread minimized. Cluster 3 has high median and fairly ok spread.

2. How many stores fall into each store format?

   Cluster 1: - 25
   Cluster 2: - 35
   Cluster 3: - 25

Summary Report of the K-Means Clustering Solution Cluster_Analysis

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Pct_Dry_Grocery + 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	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

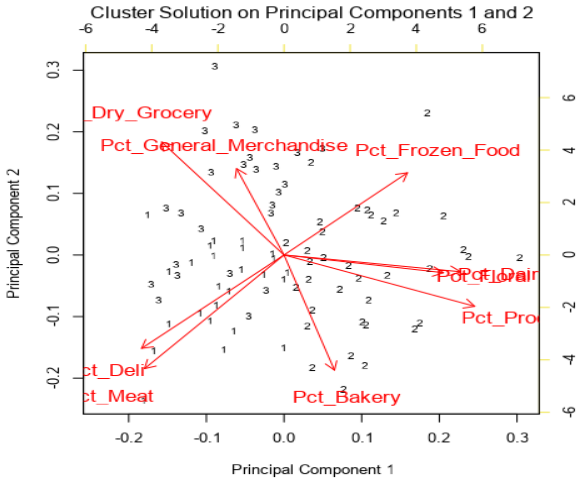
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	Pct_Dry_Grocery	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-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	-0.178481

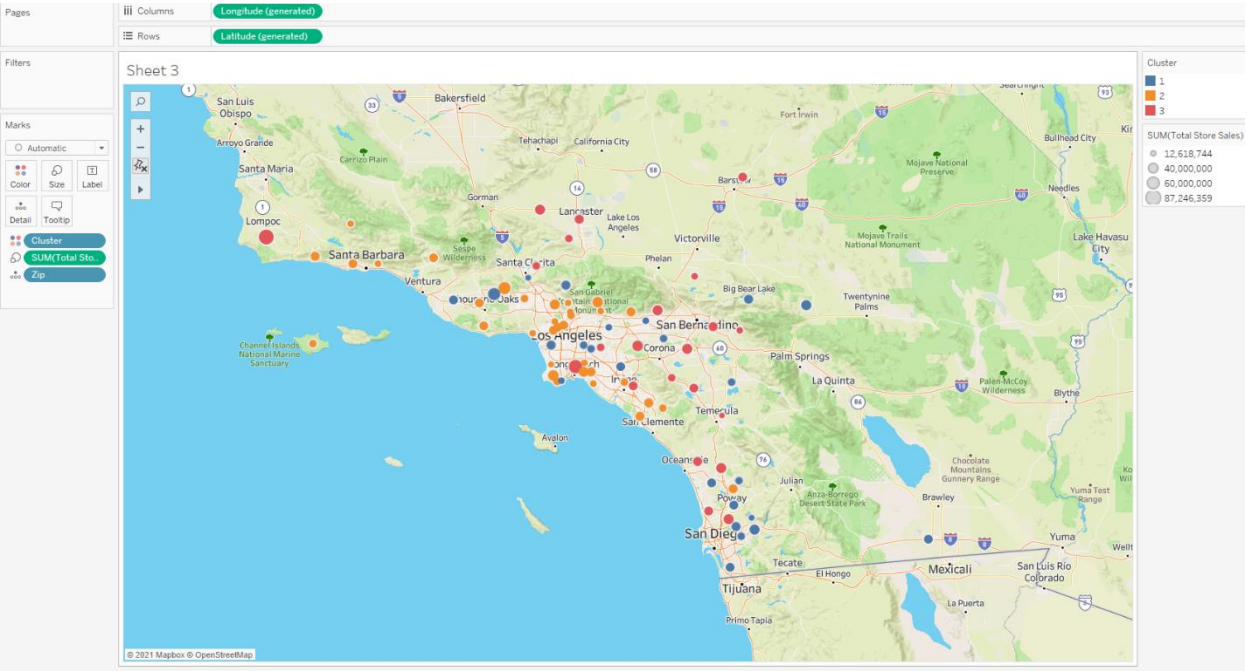
	Pct_Bakery	Pct_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Plots



The clusters differ from one another based on the percentage of sales in different categories of each store. The cluster 1 sells Meat and Deli more, cluster 2 sells Produce and Floral more and cluster 3 sells General Merchandise more.

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.

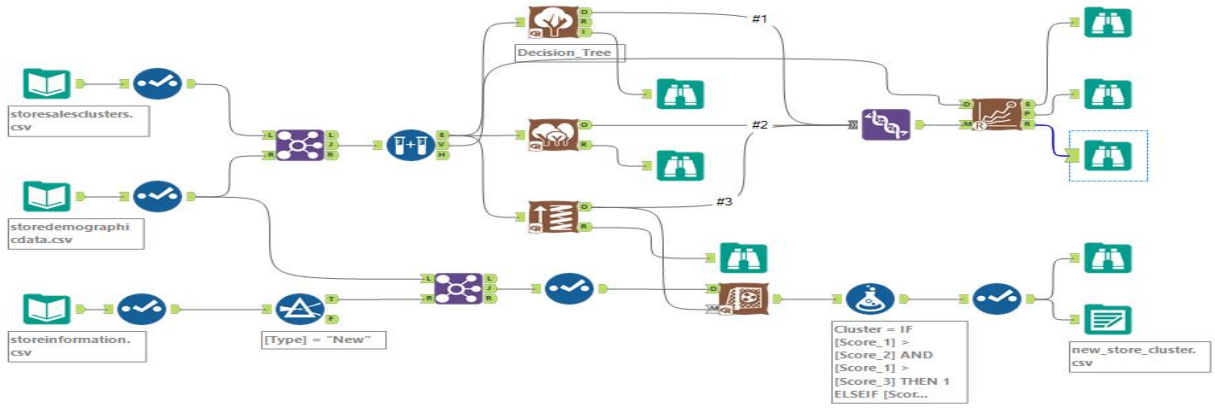


Task 2: Formats for New Stores

The company is opening 10 new stores. Using the given demographic data, we can find the store format by modeling the classification methods like random forest, boosted models or decision tree.

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.)

   I used Boosted model to predict the best store format for the new stores.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Random_Forest	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_Model	0.7059	0.7500	0.5000	1.0000	0.7500

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_Model

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Random_Forest

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

From the model comparison report we can see that both boosted model and random forest model has similar performances and better than decision tree.

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

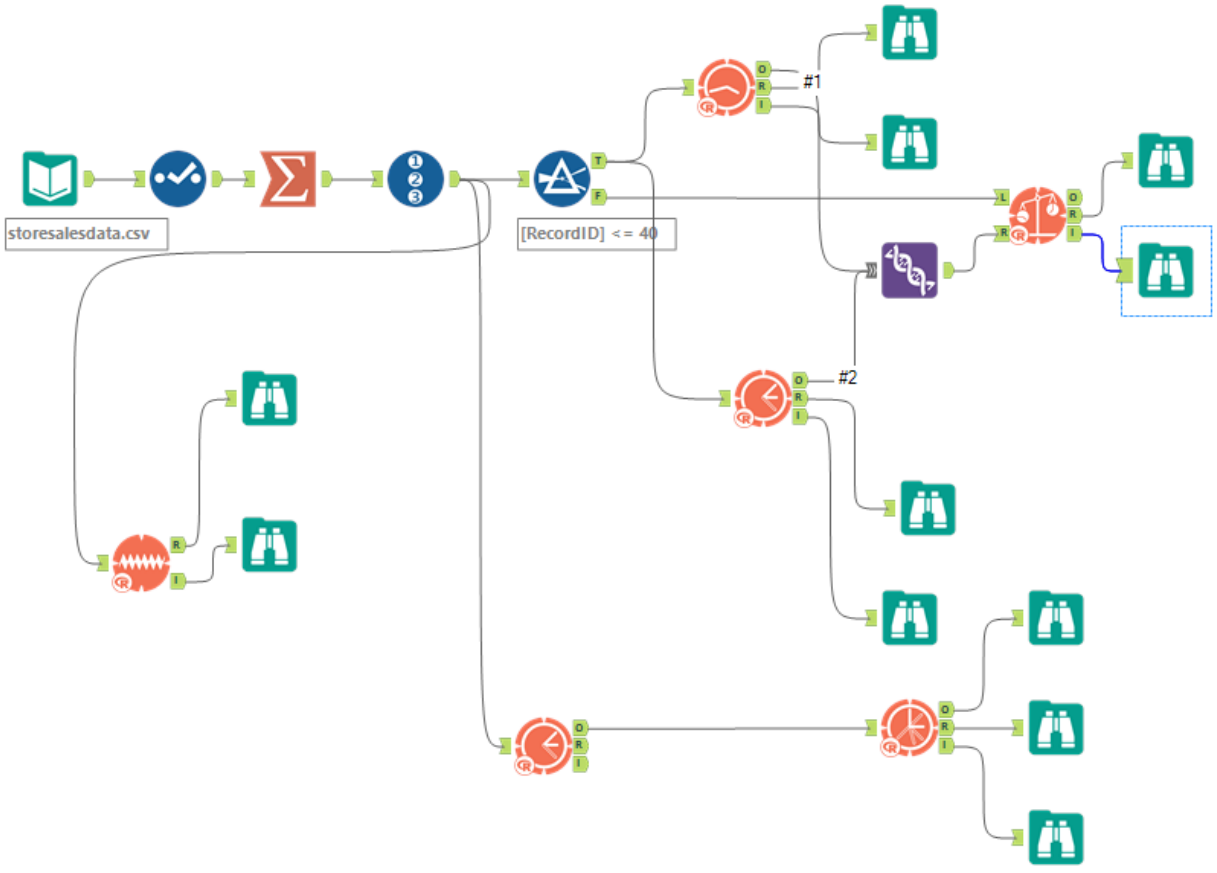
S0092	2
S0093	3
S0094	2
S0095	2

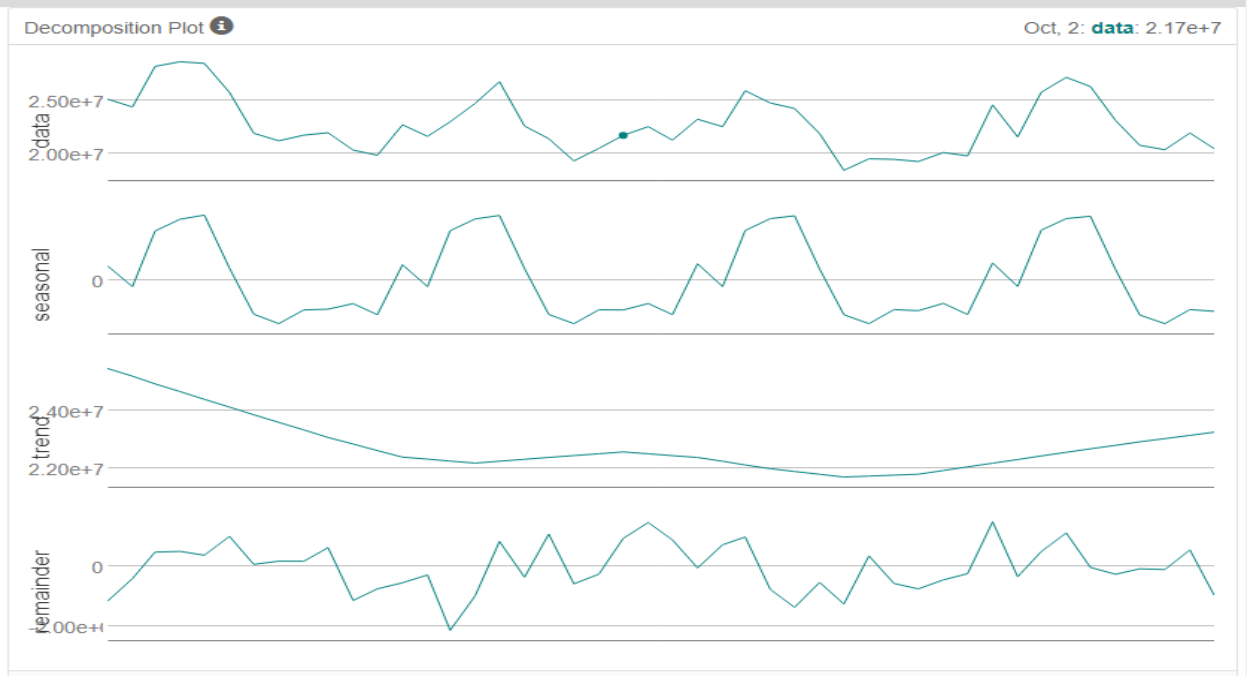
Task 3: Predicting Produce Sales

A monthly forecast of Produce sales for one year of 2016 for both existing and new stores have to be prepared.

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?

ETS (M, N, M) model is used for forecast.





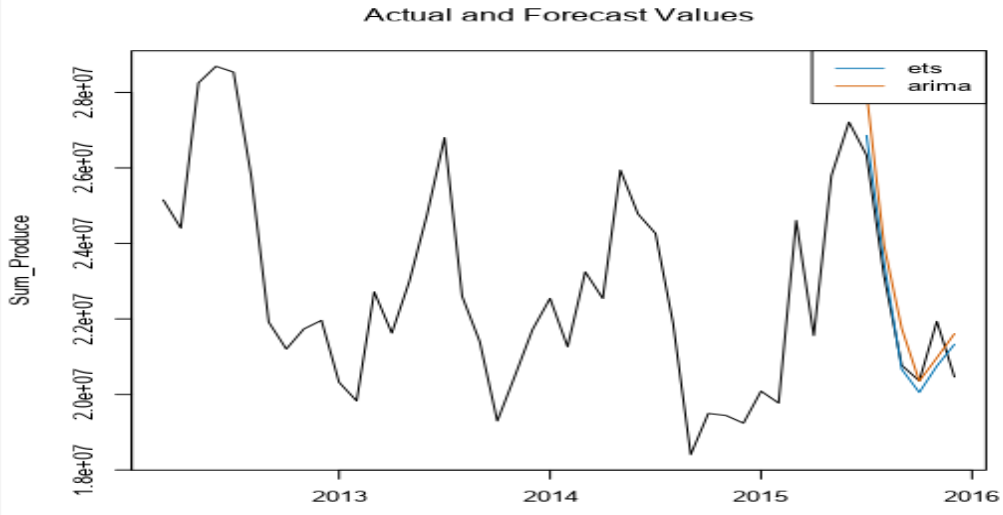
From the above decomposition plot, Error component is multiplicative, Trend is none, and Seasonal component is multiplicative.

Actual and Forecast Values:

Actual	ets	arima
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

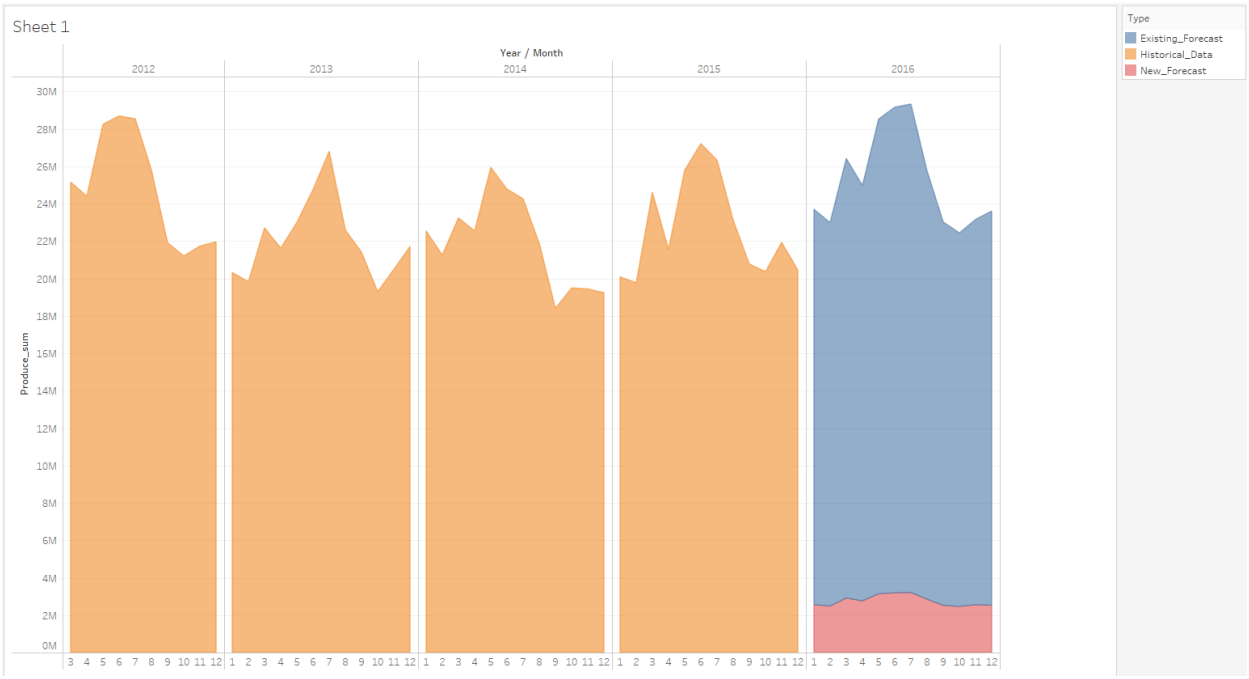
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ets	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
arima	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

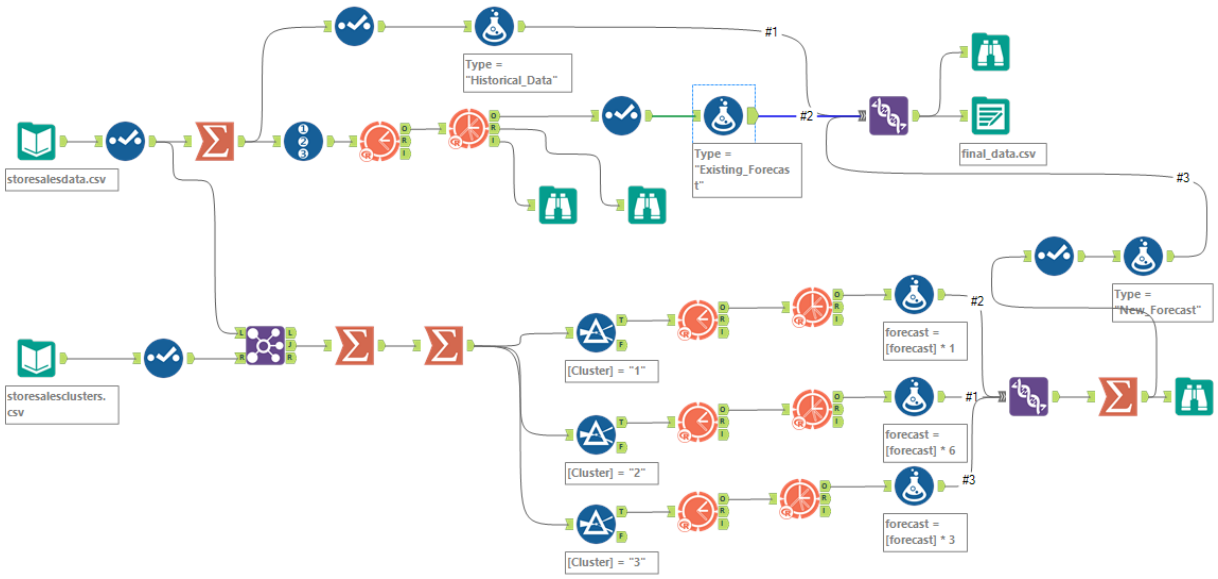


From above summary of ETS and ARIMA, the error measures are low in ETS than in ARIMA, so ETS method is used for forecasting.

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.

Year	Month	Existing	New
2016	01	21136641.78	2563357.91
2016	02	20507039.12	2483924.72
2016	03	23506565.98	2910944.14
2016	04	22208405.75	2764881.86
2016	05	25380147.77	3141305.86
2016	06	25966799.46	3195054.20
2016	07	26113792.56	3212390.95
2016	08	22899285.76	2852385.76
2016	09	20499583.90	2521697.18
2016	10	19971242.82	2466750.89
2016	11	20602665.91	2557744.58
2016	12	21073222.08	2530510.80





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.