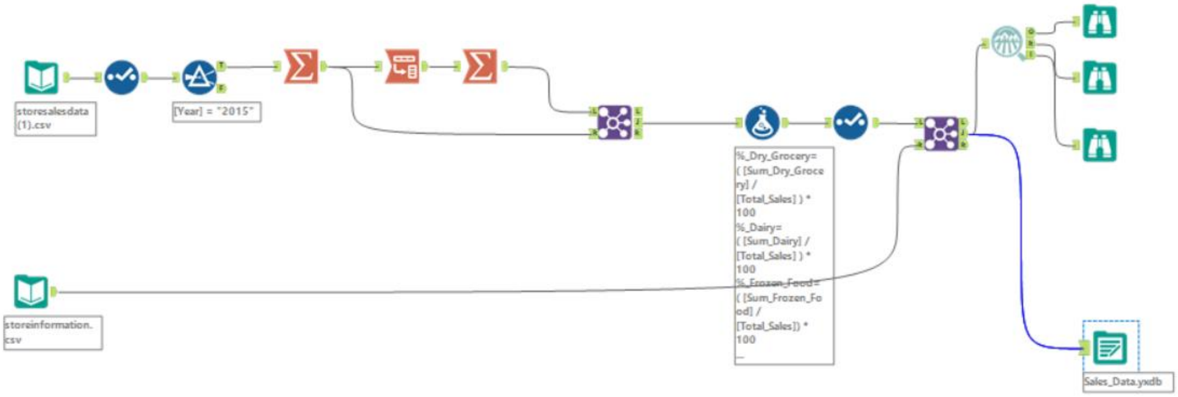
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

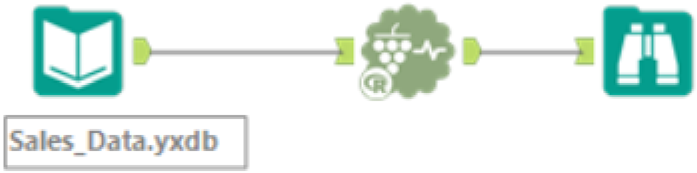
1. What is the optimal number of store formats? How did you arrive at that number?

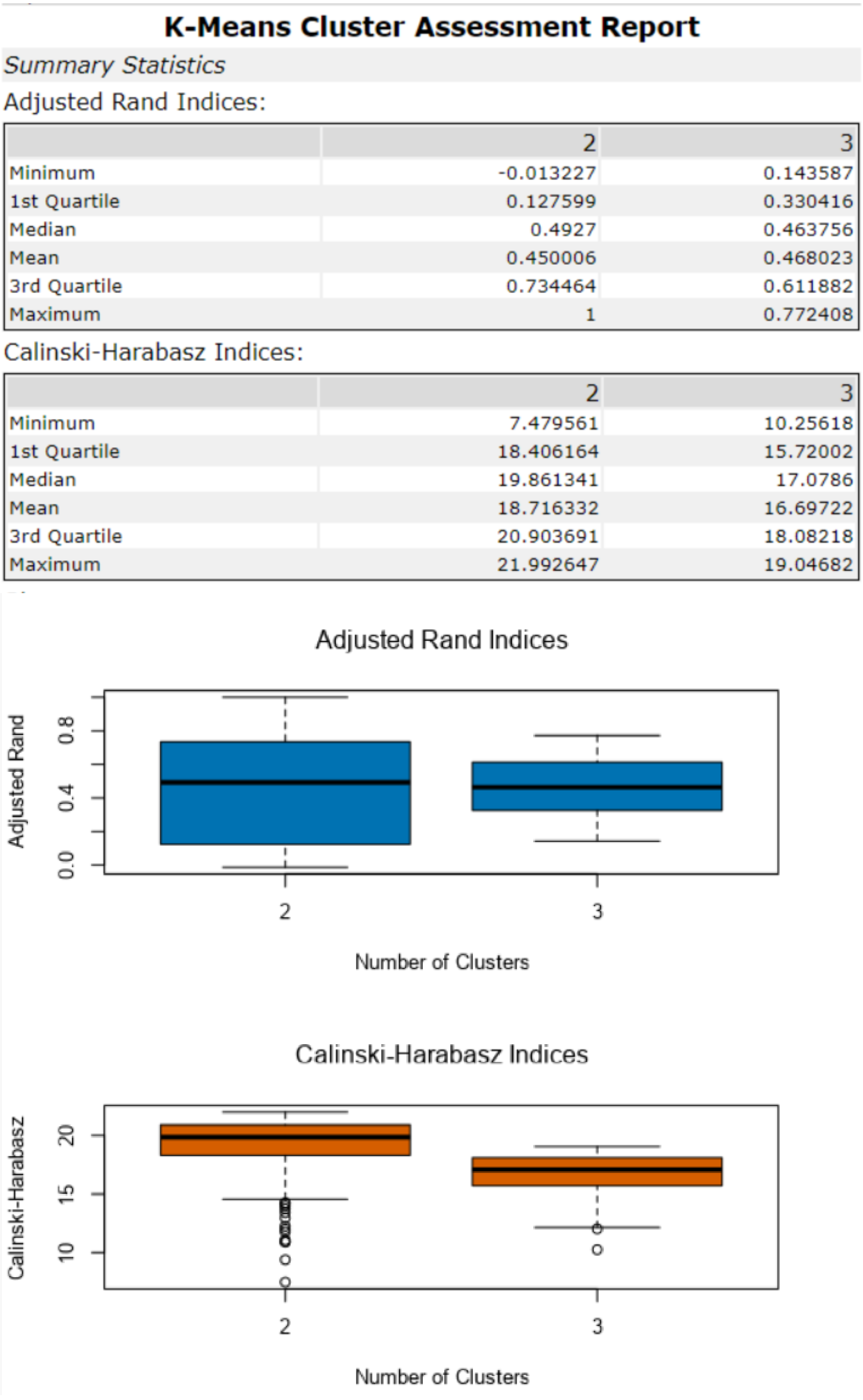
   The optimal number of store formats is 3.

- First, the data was prepared.
  Workflow:


storesalesdata (1).csv
[Year] = "2015"
%_Dry_Grocery= ([Sum_Dry_Grocery] / [Total_Sales]) * 100
%_Dairy= ([Sum_Dairy] / [Total_Sales]) * 100
%_Frozen_Food= ([Sum_Frozen_Food] / [Total_Sales]) * 100
...
storeinformation.csv
Sales_Data.yxdb

- Then used K-Means clustering model
  Workflow:





K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3
Minimum	-0.013227	0.143587
1st Quartile	0.127599	0.330416
Median	0.4927	0.463756
Mean	0.450006	0.468023
3rd Quartile	0.734464	0.611882
Maximum	1	0.772408

Calinski-Harabasz Indices:

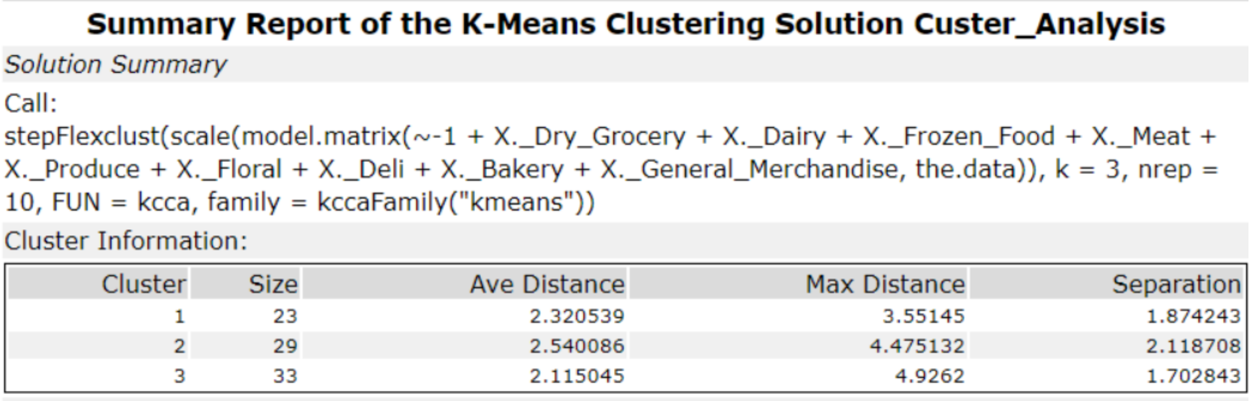
	2	3
Minimum	7.479561	10.25618
1st Quartile	18.406164	15.72002
Median	19.861341	17.0786
Mean	18.716332	16.69722
3rd Quartile	20.903691	18.08218
Maximum	21.992647	19.04682

2. How many stores fall into each store format?

Cluster 1=> 23 Stores

Cluster 2 => 29 Stores

Cluster 3 => 33 Stores



Summary Report of the K-Means Clustering Solution Custer_Analysis

Solution Summary

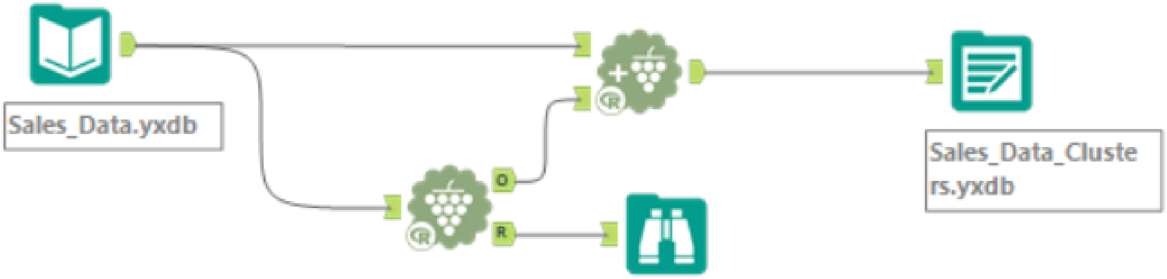
Call:

```
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"))
```

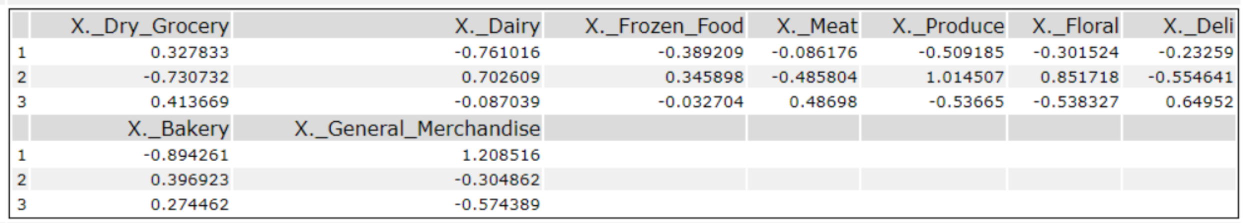
Cluster Information:

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

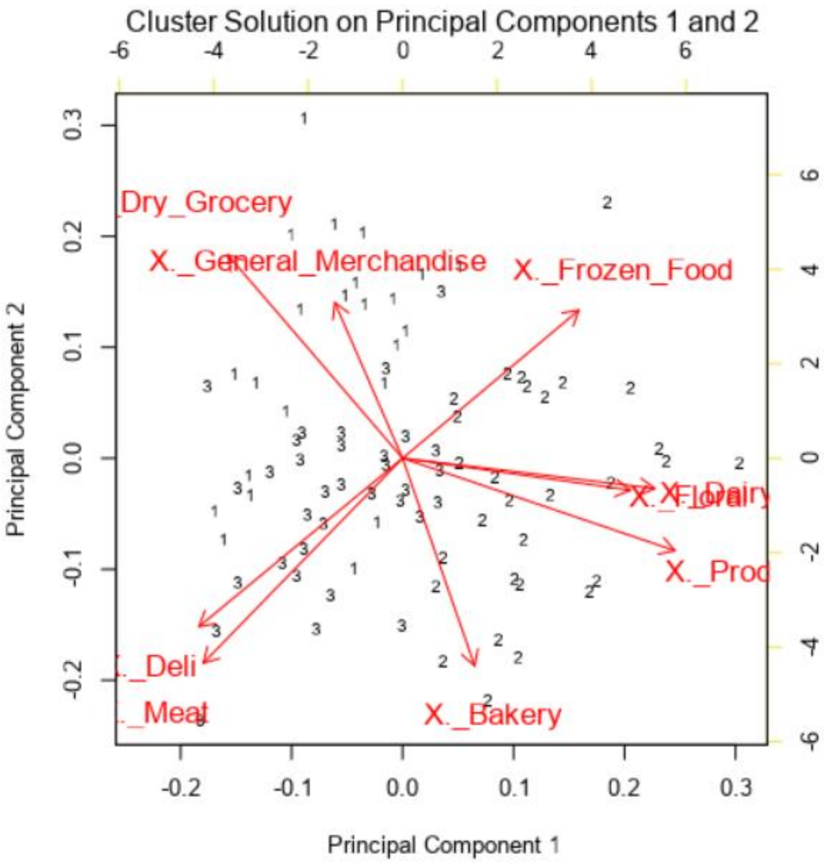
Workflow:



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

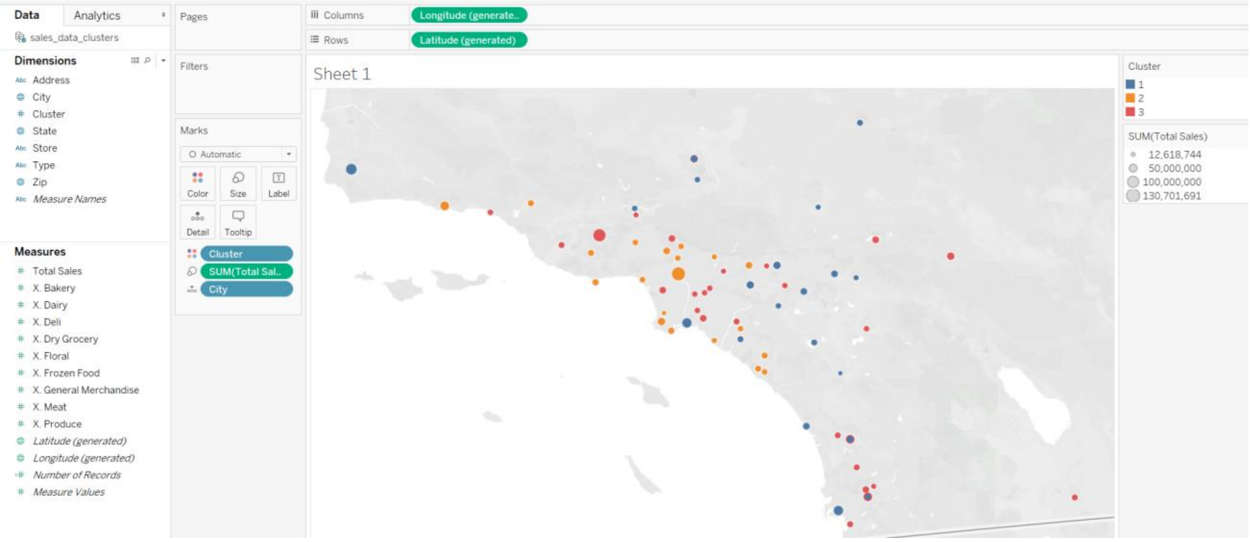


	X._Dry_Grocery	X._Dairy	X._Frozen_Food	X._Meat	X._Produce	X._Floral	X._Deli	X._Bakery	X._General_Merchandise
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259	-0.894261	1.208516
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641	0.396923	-0.304862
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952	0.274462	-0.574389



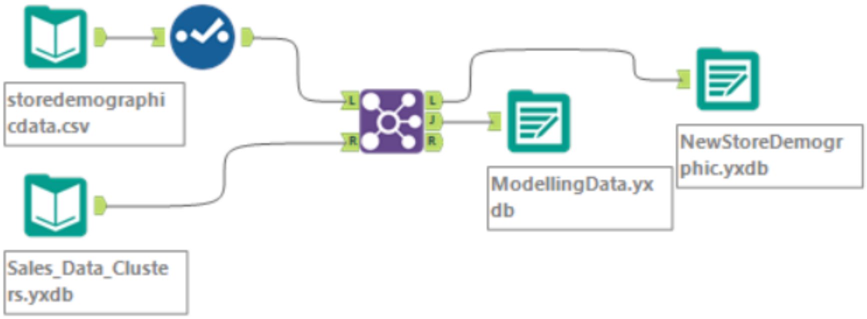
Based on general merchandise sales it can be said that I see cluster 1 is the most positive percentage of general merchandise sales on the other hand cluster 3 which is the most negative.

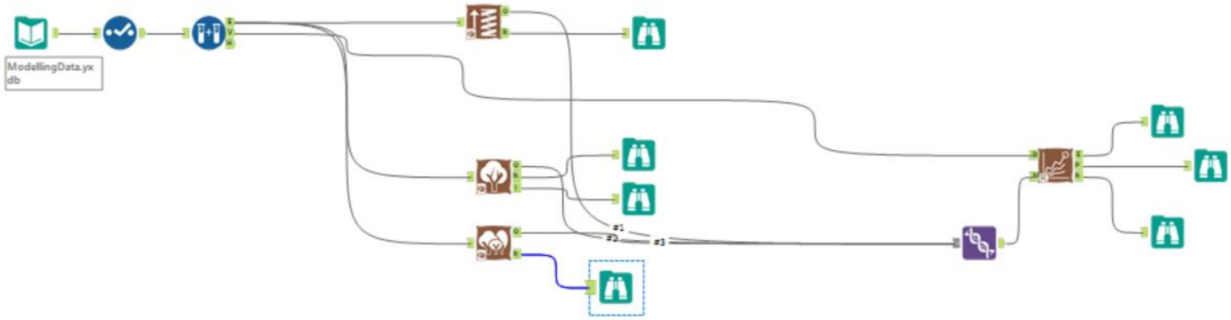
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

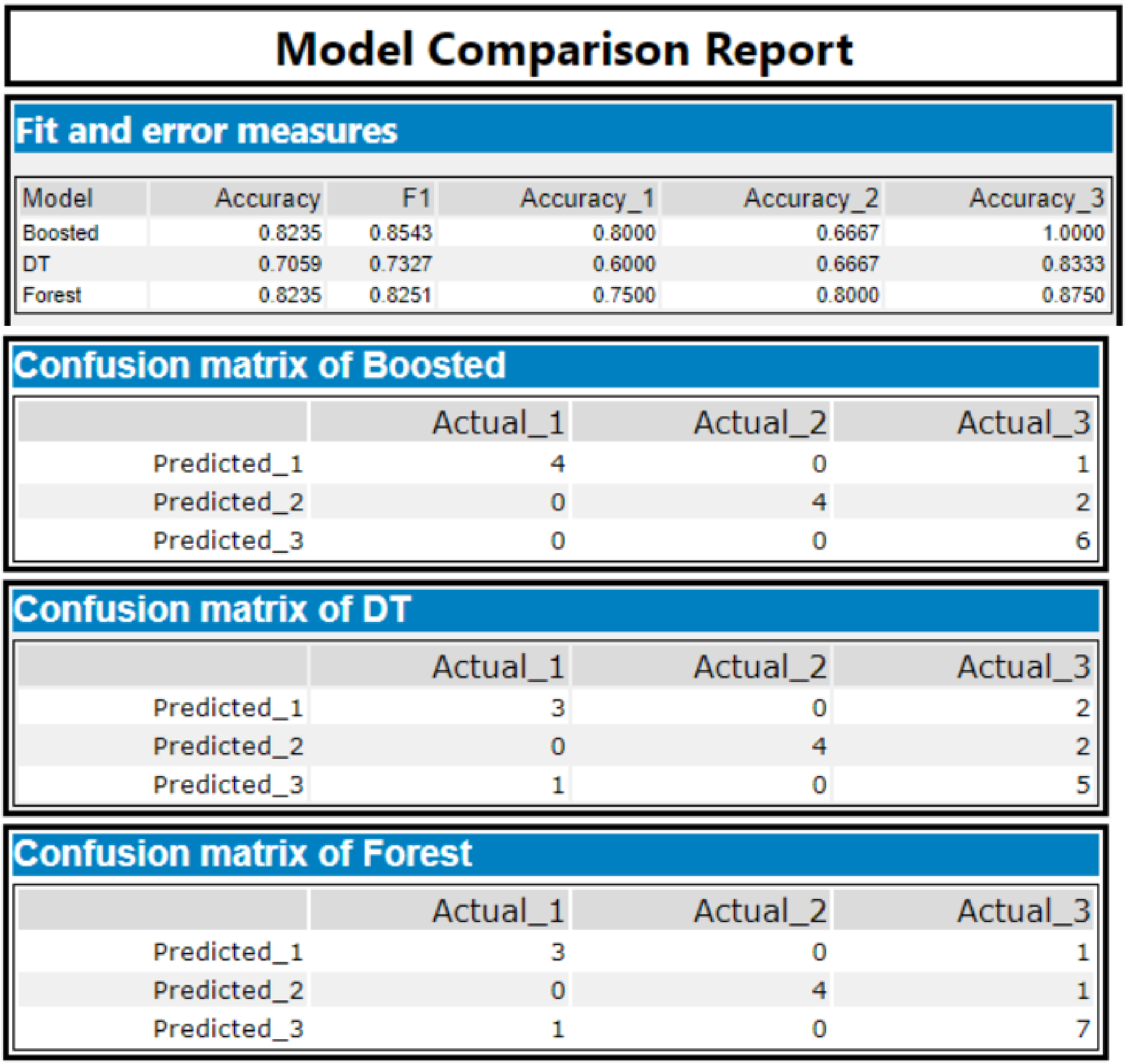


Task 2: Formats for New Stores

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.)
   Used a boosted, decision tree and random forest model. An 80/20 split of the data was used for training and validating the models.







Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000
DT	0.7059	0.7327	0.6000	0.6667	0.8333
Forest	0.8235	0.8251	0.7500	0.8000	0.8750

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 DT

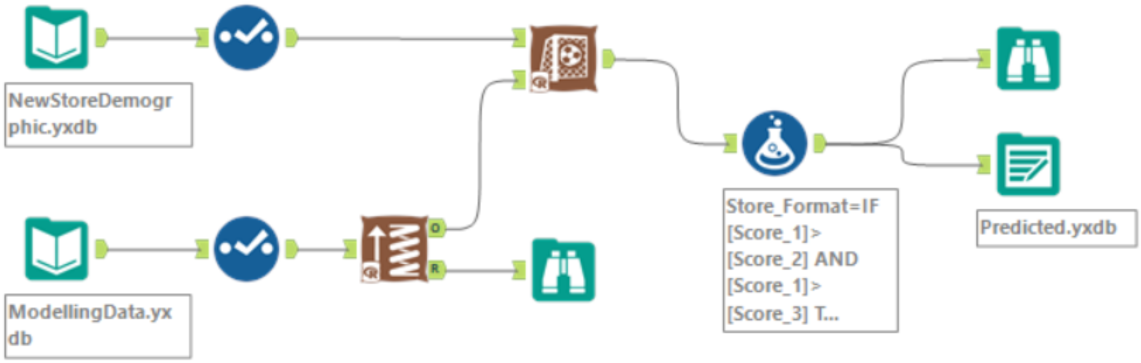
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest

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

Boosted Model is the best because it has a higher F1 score.

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

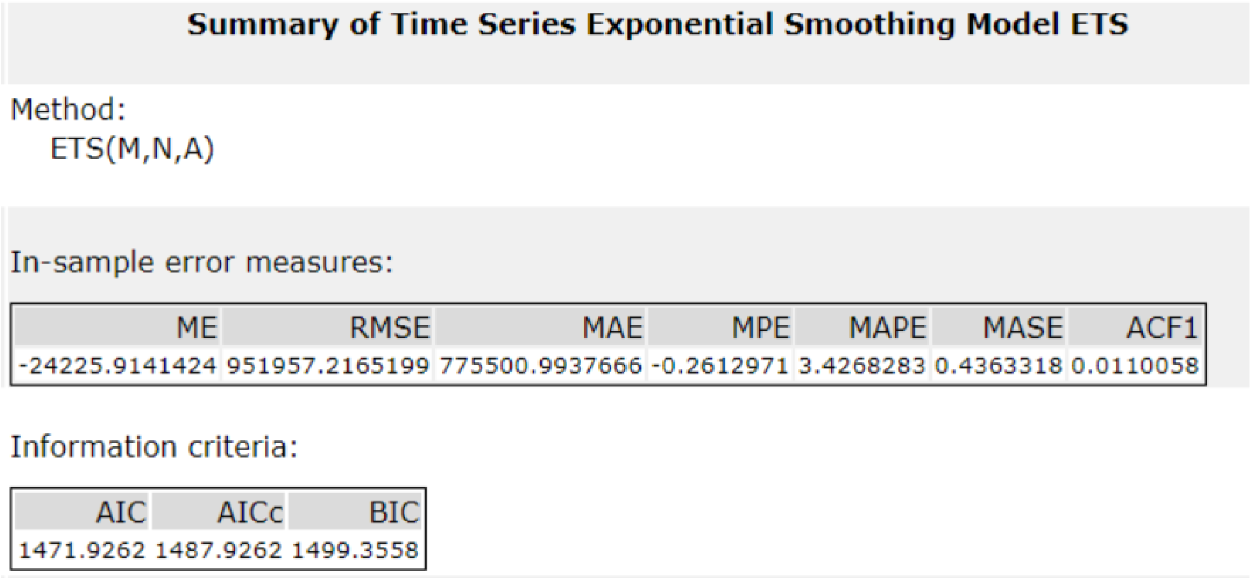

NewStoreDemogr phic.yxdb
ModellingData.yx db
O
R
Store_Format=IF [Score_1]> [Score_2] AND [Score_1]> [Score_3] T...
Predicted.yxdb

Store Number	Segment
S0086	3
S0087	2
S0088	3
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?

Both ETS and ARIMA models were run for comparison and the data used here is sales for produce only per month for all stores .



Summary of Time Series Exponential Smoothing Model ETS

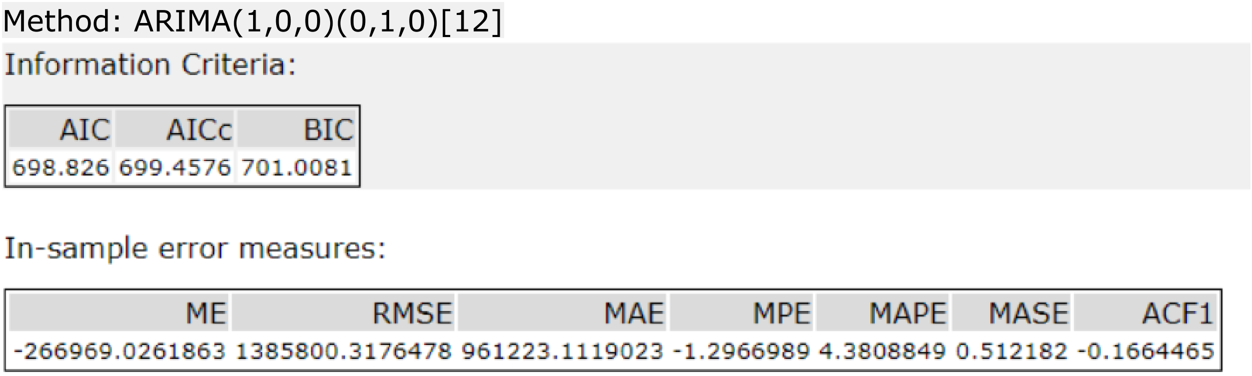
Method:
ETS(M,N,A)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-24225.9141424	951957.2165199	775500.9937666	-0.2612971	3.4268283	0.4363318	0.0110058

Information criteria:

AIC	AICc	BIC
1471.9262	1487.9262	1499.3558



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

Information Criteria:

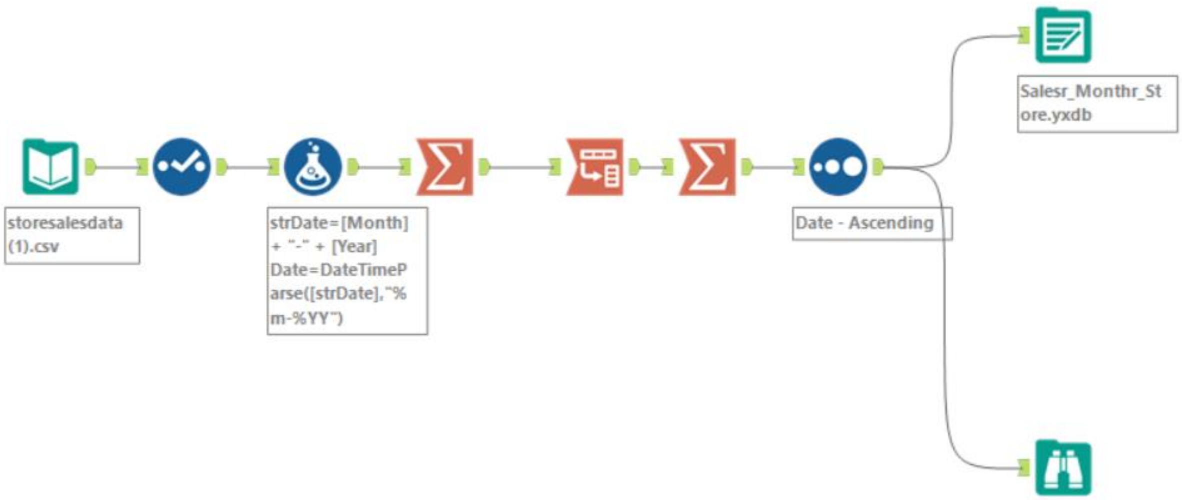
AIC	AICc	BIC
698.826	699.4576	701.0081

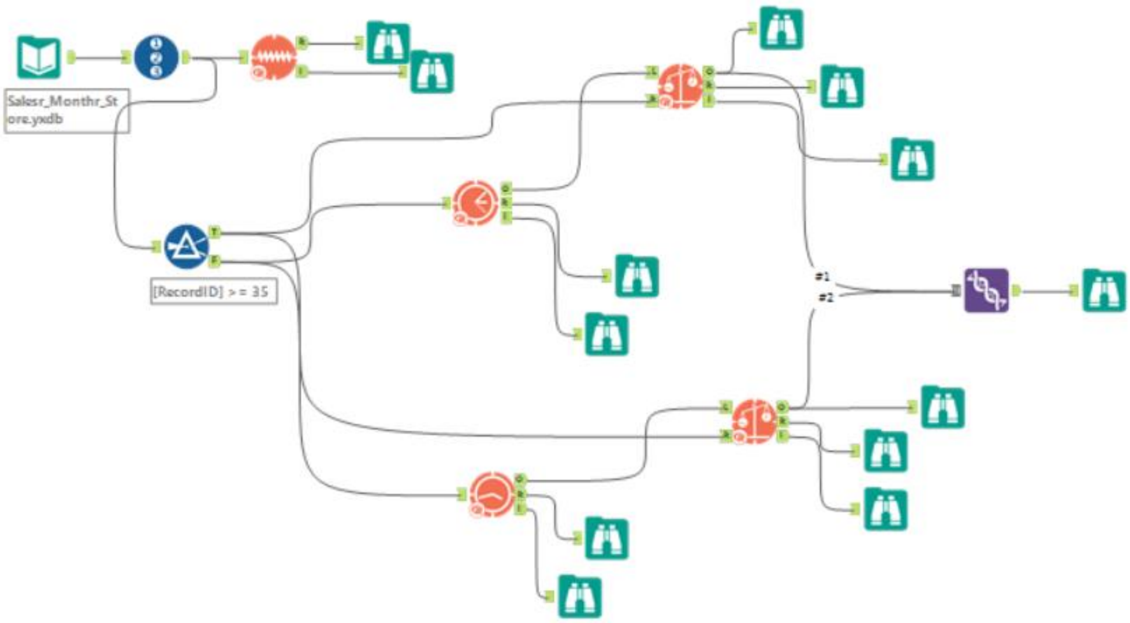
In-sample error measures:

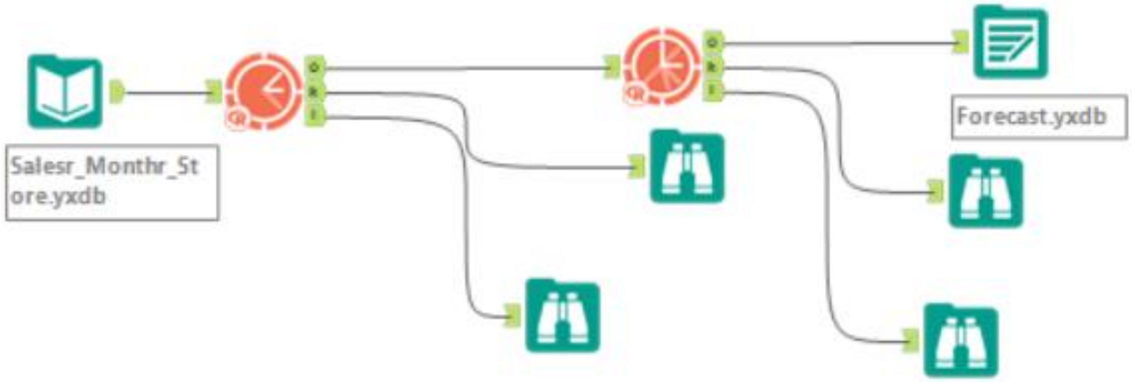
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-266969.0261863	1385800.3176478	961223.1119023	-1.2966989	4.3808849	0.512182	-0.1664465

The ETS(M,N,M) will be used for forecasting because the model having lower error values

Workflow:

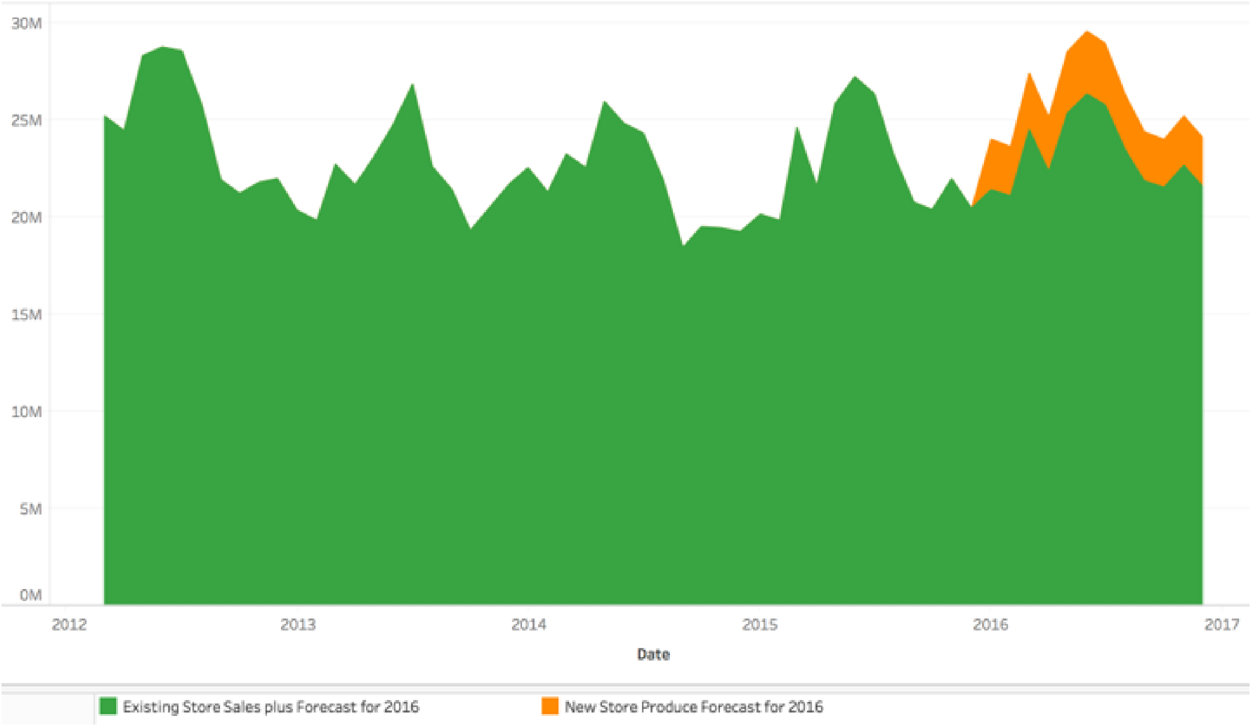

storesalesdata (1).csv
strDate=[Month] + "-" + [Year] Date=DateTimeParse([strDate],"%m-%YY")
Date - Ascending
Salesr_Monthr_Store.yxdb


Salesr_Monthr_Store.yxdb
[RecordID] >= 35
#1
#2


Salesr_Monthr_Store.yxdb
Forecast.yxdb

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.

Month-year	Existing Store Sales Forecast	New Store Sales Forecast
1-2016	21,381,830.22	2,600,354.85
2-2016	21,081,311.62	2,505,198.46
3-2016	24,502,171.96	2,889,940.32
4-2016	22,352,993.13	2,743,927.30
5-2016	25,331,350.65	3,110,813.81
6-2016	26,330,255.79	3,191,154.55
7-2016	25,715,514.09	3,219,369.78
8-2016	23,458,933.07	2,852,751.79
9-2016	21,801,458.48	2,543,602.66
10-2016	21,509,922.65	2,477,331.44
12-2016	22,619,212.99	2,569,169.56
12-2016	21,582,321.09	2,535,481.94


30M
25M
20M
15M
10M
5M
0M
Date
🟩 Existing Store Sales plus Forecast for 2016
🟧 New Store Produce Forecast for 2016