

# Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

**Case study:** Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

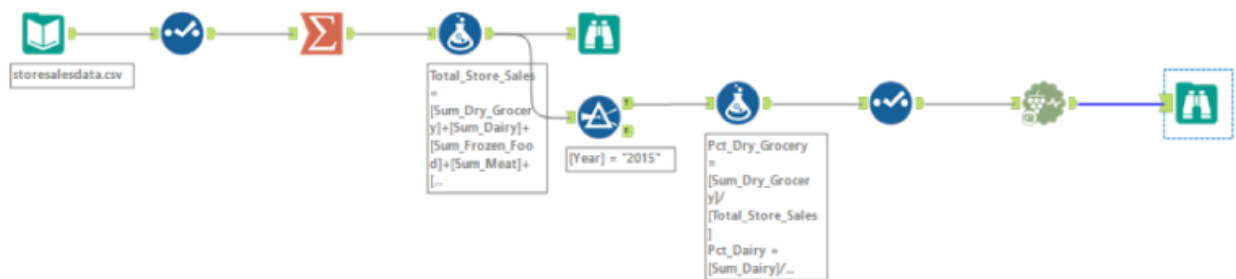
**Methodology:** Segmentation and Clustering

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

**The optimal number of store formats is 3.**

After running the K-Means clustering model and using the median and spread of the Adjusted Rand (AR) Indices and Calinski-Harabasz (CH) Indices. It revealed that 3 clusters are the optimal method because the box-whisker plots in the Adjusted Rand Indices show how tight the indices for each data point are to each other.

While it might seem that cluster 2 is the optimal number of clusters, it is 3 because the variance is too big for 2 clusters, while we see more compactness and still high median values when we have 3 clusters.



## K-Means Cluster Assessment Report

### Summary Statistics

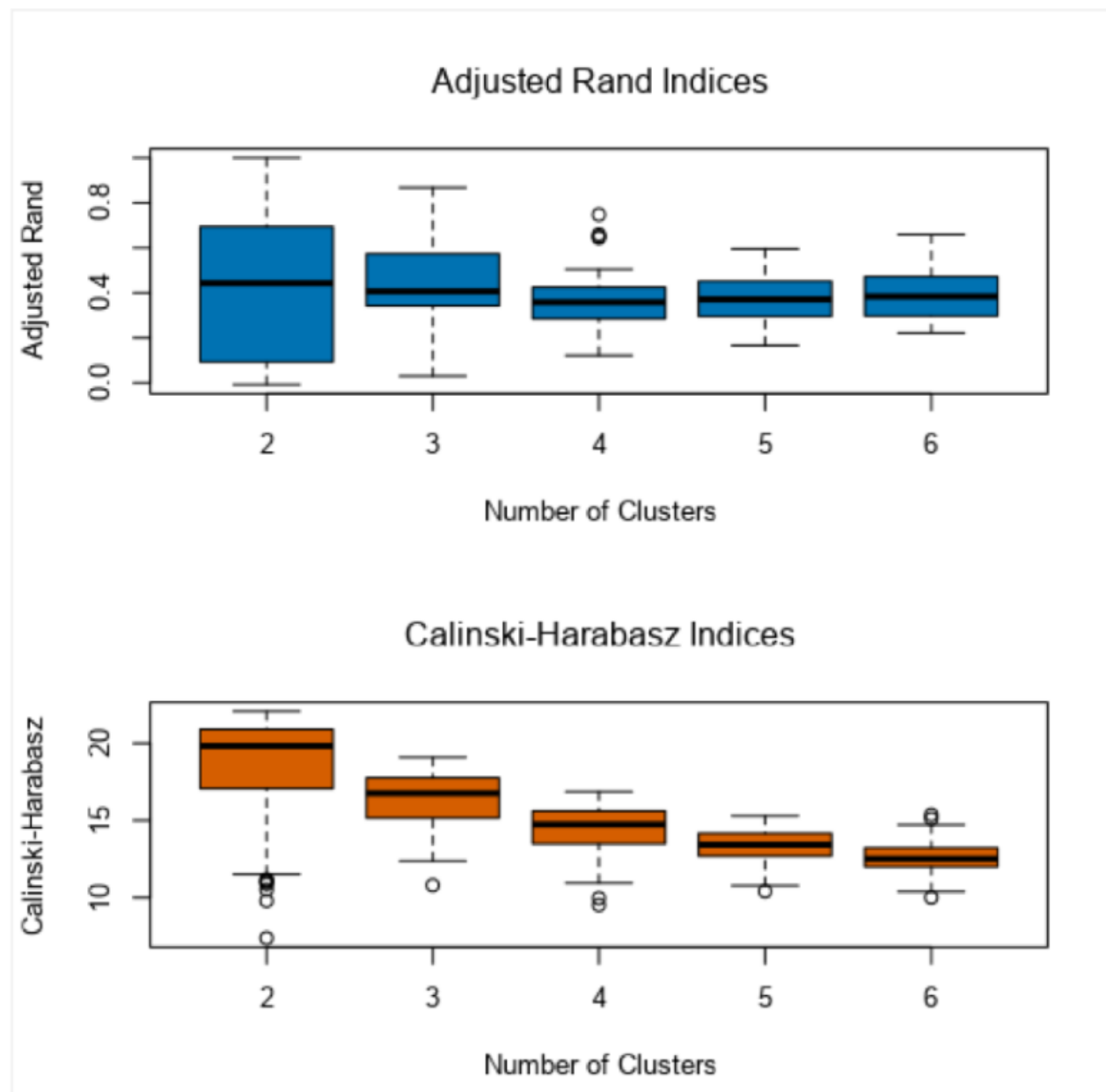
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.007639	0.029695	0.122167	0.166791	0.222111
1st Quartile	0.094172	0.343478	0.285754	0.298186	0.301965
Median	0.443213	0.406361	0.357989	0.370994	0.384296
Mean	0.405201	0.443015	0.365307	0.383051	0.389198
3rd Quartile	0.684276	0.56807	0.424442	0.450713	0.470301
Maximum	1	0.868183	0.747642	0.595251	0.659091

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	7.376319	10.80678	9.524605	10.41103	10.00938
1st Quartile	17.163364	15.15871	13.531027	12.71013	11.99892
Median	19.816152	16.75762	14.737409	13.42556	12.51619
Mean	18.520371	16.39173	14.436238	13.36015	12.61465
3rd Quartile	20.893269	17.74967	15.580417	14.17377	13.23228
Maximum	22.061691	19.089	16.865033	15.29623	15.36927

## Plots



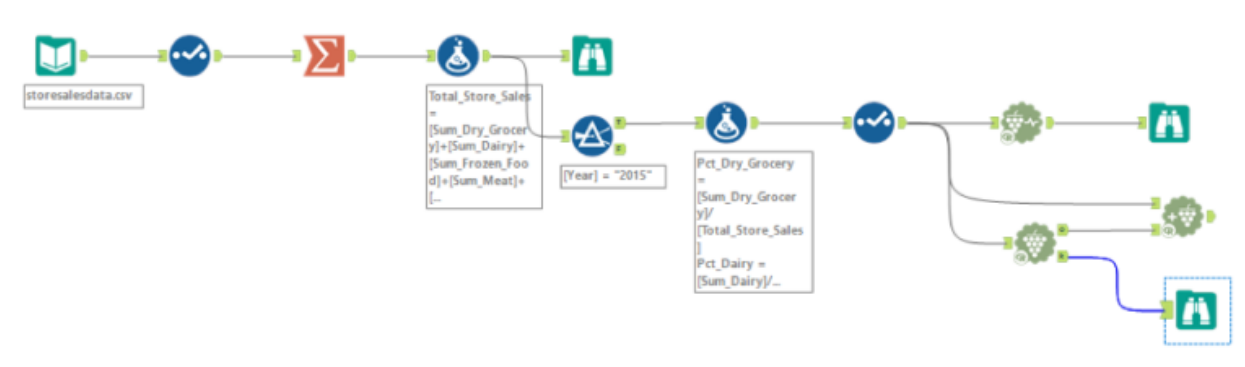
How many stores fall into each store format?

Cluster 1: **23 Stores**

Cluster 2: **29 Stores**

Cluster 3: **33 Stores**

By running the K-Centroids Cluster Analysis tool using the same configuration of the K-Centroids Diagnostics tool, I am able to get the cluster information below:



Report

Summary Report of the K-Means Clustering Solution Store\_Cluster

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	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.

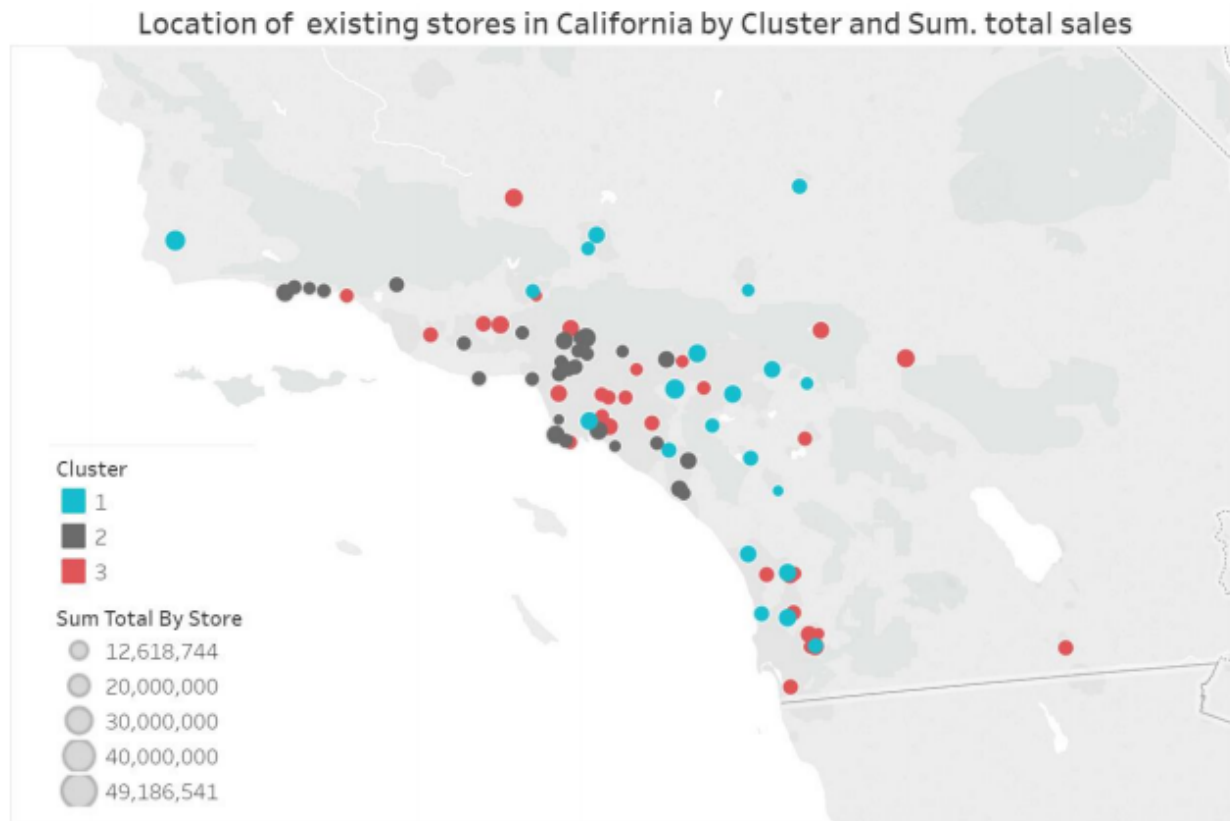
Based on the results of the clustering model, what is one way that the clusters differ from one another?

Based on the summary report of the K-Means Clustering solution, one way that the clusters differ from one another could be: considering the percentage of sales by category of each store, cluster 1 sells more in general merchandise; cluster 2 sells more in produce and floral; and cluster 3 sells more in deli and meat; etc.

	Pct_Dry_Grocery	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					

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.

Next step, I connected Tableau to the data source and created the data visualization.



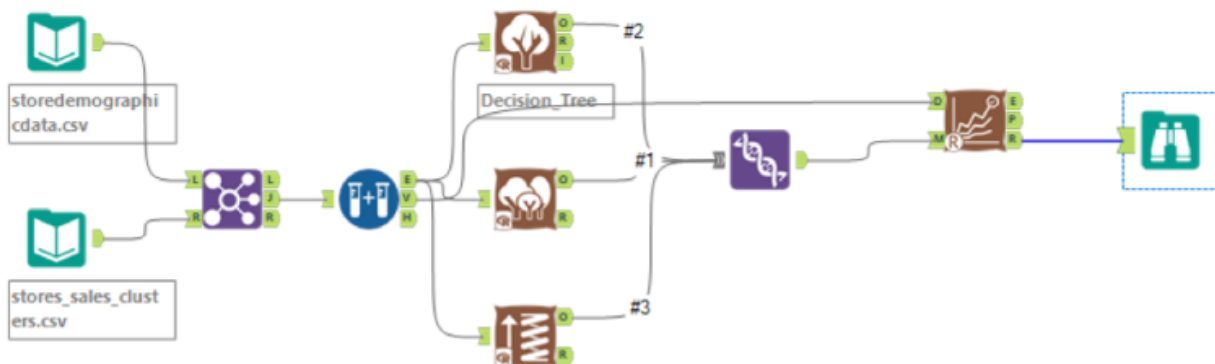
## Task 2: Formats for New Stores

**Case study:** The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, we don't have sales data for these new stores yet, so we'll have to determine the format using each of the new store's demographic data.

**Methodology:** Non-Binary Classification Models (Decision Tree, Forest Model, and Boosted Model)

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. I compared the Decision Tree, Forest Model and Boosted Model using the Model Comparison tool. Both Forest Model and Boosted Model have better accuracy than the Decision Tree. The Boosted Model is the best because it has a higher F1 score.



Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	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,  $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$ . 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	0	4	2
Predicted_3	0	0	6

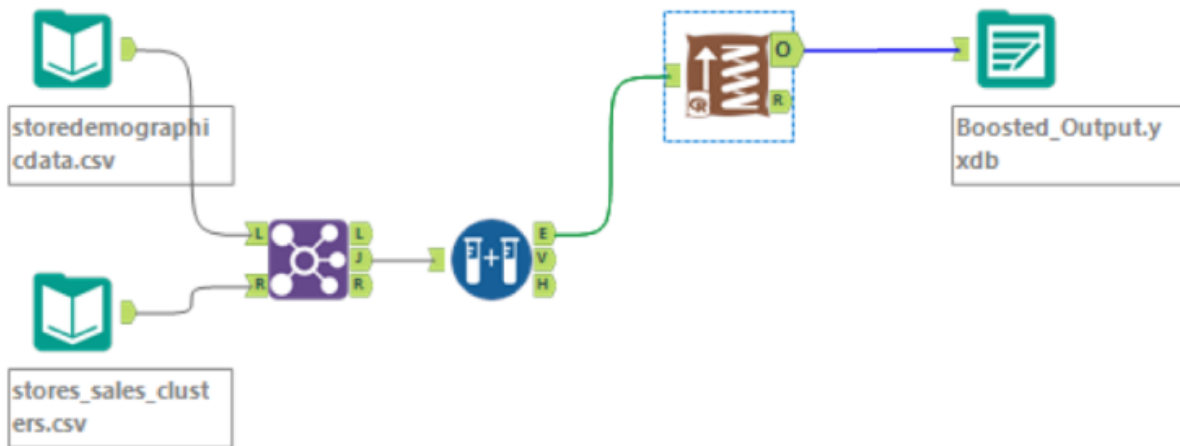
Confusion matrix of Decision Tree			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest_Model			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

I started by building the Boosted Model and output the model object so that I can use in a new workflow.



Then I created a new workflow using the score tool.

The input data sources are the data output from the Boosted Model that I mentioned above and the new stores (S0086 – S0095).

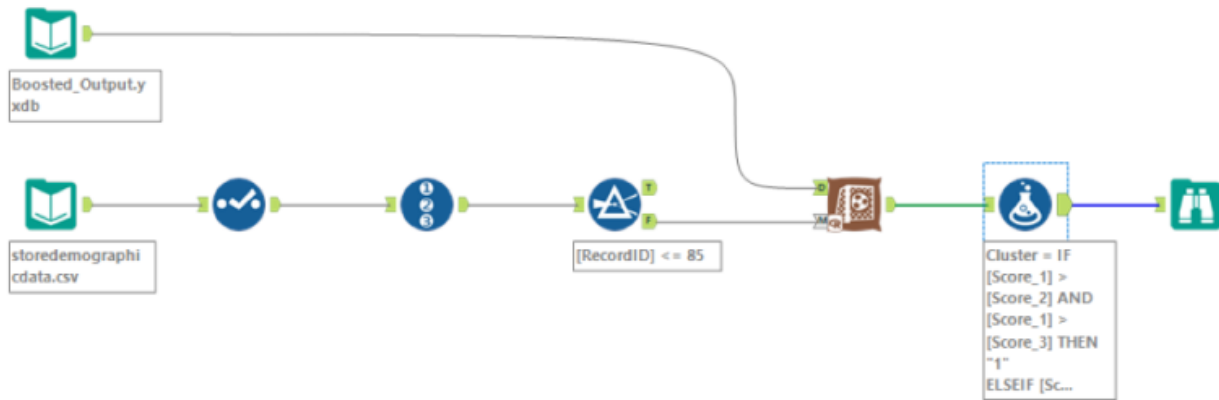
To get the new stores, I used the Record ID tool to filter out the Record ID <= 85 and the rest are the new stores.

After the Score tool, I used the Formula tool to come up with the predicted cluster for each store:

***IF [Score\_1] > [Score\_2] AND [Score\_1] > [Score\_3] THEN "1 ELSEIF [Score\_2] > [Score\_1] AND [Score\_2] > [Score\_3] THEN "2" ELSE "3" ENDIF***

Ultimately, there are 3 stores in cluster 1, 6 stores in cluster 2 and 1 store in cluster 3.





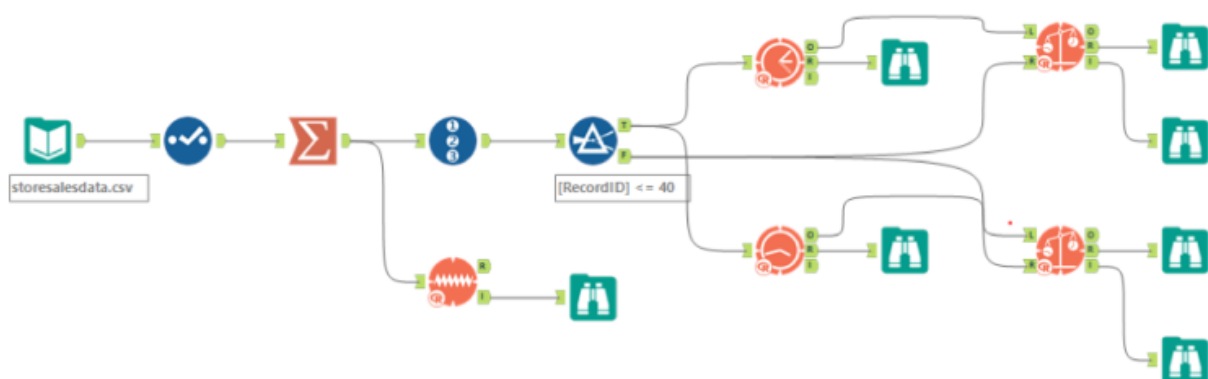
### Task 3: Predicting Produce Sales

**Case study:** You've been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, follow the steps below.

**Methodology:** Time Series Forecasting (ETS Models and ARIMA Models)

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) is the model that I chose to forecast the product sales for new and existing stores. Here my rationale for the decision:



Use "store\_sales\_data" file as the data source. Add a Summarize tool to sum the produce sales (data type: double) group by year and month.

Train ETS and ARIMA models. The optimal option for the ETS model is ETS(M,N,M) and for ARIMA is ARIMA(1,0,0)(1,1,0)[12].

Add TS Compare tool to obtain the forecast error measurements against the holdout sample for each model.

Compare the forecast error measurements against the holdout sample of ETS and ARIMA and select the model with the lower forecast error measurements. ETS(M,N,M) turns out to be the better one.

### Summary of Time Series Exponential Smoothing Model ETS

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
I	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

## 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:

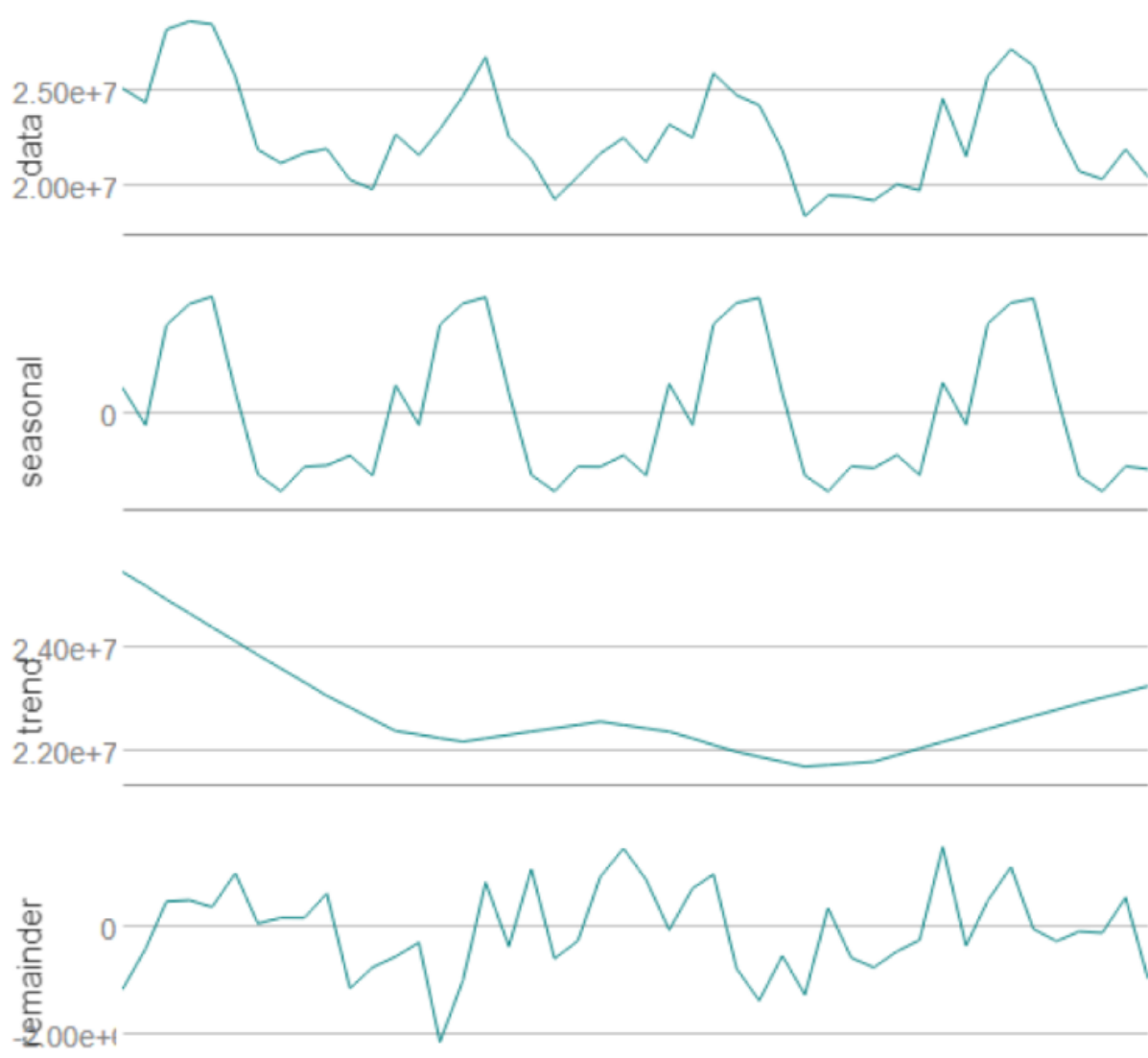
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Ljung-Box test of the model residuals:

Chi-squared = 15.0973, df = 12, p-value = 0.23616

Based on the decomposition plot obtained from the TS Plot tool, we know the error is multiplicative, the trend is none and the seasonality is multiplicative, so ETS(M,N,M) is the best model that I should choose to forecast the produce sales for the new and existing store.

Decomposition Plot 



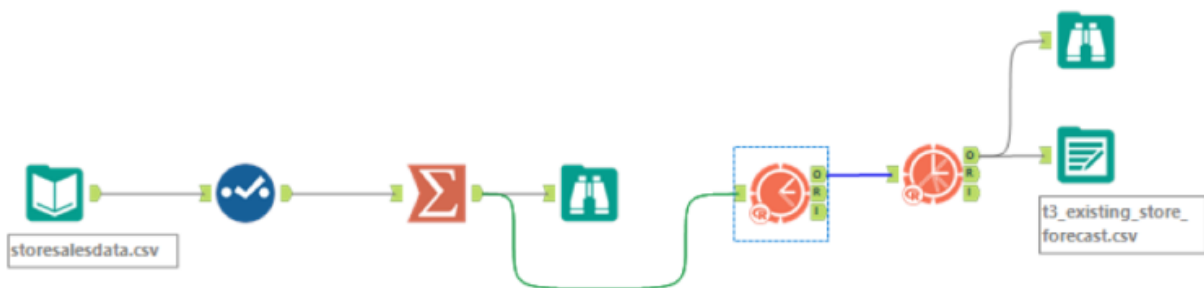
This is a decomposition plot

Month	New Stores	Existing Store
2016-01	2,588,250	21,136,642
2016-02	2,499,159	20,507,039
2016-03	2,916,908	23,506,566
2016-04	2,791,560	22,208,406
2016-05	3,156,890	25,380,148
2016-06	3,200,940	25,966,799
2016-07	3,224,858	26,113,793
2016-08	2,861,958	22,899,286

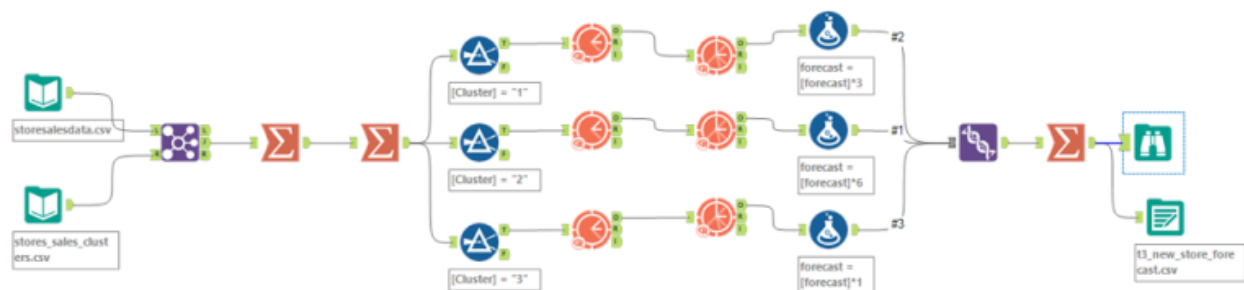
2016-09	2,534,353	20,499,584
2016-10	2,481,117	19,971,243
2016-11	2,578,336	20,602,666
2016-12	2,561,917	21,073,222

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.

Here is the workflow for existing stores forecasting:



Here is the workflow that I build for new stores forecasting:



I then created another workflow to union historical data, existing store forecasts and new store forecasts to use in Tableau.

