Project 7: Combining Predictive Techniques

Task 1: Determine Store Formats for Existing Stores

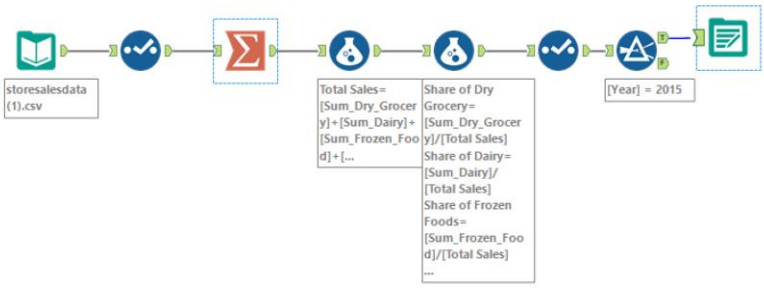
Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This 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.

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms formats and segments will be used interchangeably throughout this project. You've been asked to:

- Determine the optimal number of store formats based on sales data.
  - Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
  - Use only 2015 sales data.
  - Use a K-means clustering model.
- Segment the 85 current stores into the different store formats.

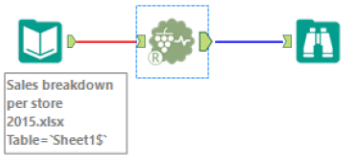
1. What is the optimal number of store formats? How did you arrive at that number?

First we arranged the data to get the percentage sales per category in 2015



Record #	Store	Year	Share of Dry Grocery	Share of Dairy	Share of Frozen Foods	Share of Meat	Share of Produce	Share of Floral	Share of Deli	Share of Bakery	Share of General Merchandises
1	S0001	2015	0.461347	0.103084	0.077199	0.107677	0.097171	0.006769	0.04355	0.035544	0.067658
2	S0002	2015	0.457528	0.106388	0.078841	0.114903	0.101259	0.007438	0.039792	0.029702	0.064149
3	S0003	2015	0.421349	0.102356	0.069027	0.114713	0.125399	0.009631	0.041761	0.036108	0.079655
4	S0004	2015	0.45456	0.097111	0.080277	0.127728	0.100409	0.00612	0.041785	0.034469	0.057542
5	S0005	2015	0.440169	0.106268	0.086339	0.101863	0.131112	0.008877	0.035389	0.022607	0.067376
6	S0006	2015	0.459835	0.098609	0.074443	0.106916	0.101782	0.008265	0.046309	0.035417	0.068424
7	S0007	2015	0.442087	0.102135	0.08917	0.115135	0.103447	0.007838	0.046298	0.030176	0.063715
8	S0008	2015	0.428457	0.108732	0.080321	0.106878	0.123696	0.007859	0.032623	0.022398	0.089036
9	S0009	2015	0.442252	0.100376	0.077051	0.118908	0.103507	0.007288	0.044204	0.025756	0.080656
10	S0010	2015	0.431445	0.114962	0.080342	0.107148	0.121629	0.010383	0.037099	0.038051	0.058941
11	S0011	2015	0.453793	0.093984	0.071882	0.121443	0.098467	0.007896	0.048717	0.032441	0.071378
12	S0012	2015	0.465999	0.097265	0.074357	0.114391	0.094315	0.006825	0.045066	0.018592	0.083191
13	S0013	2015	0.445549	0.090722	0.075352	0.114048	0.092661	0.007912	0.046277	0.034281	0.093198
14	S0014	2015	0.444373	0.097768	0.077185	0.098697	0.106849	0.01066	0.034927	0.020549	0.108992
15	S0015	2015	0.447968	0.098766	0.077869	0.09999	0.104975	0.007685	0.034977	0.025106	0.102665
16	S0016	2015	0.47724	0.091435	0.078064	0.110087	0.088469	0.006252	0.046052	0.02872	0.073681
17	S0017	2015	0.438117	0.100133	0.078777	0.097699	0.133663	0.01339	0.040857	0.028239	0.069125
18	S0018	2015	0.452848	0.098827	0.079017	0.106062	0.109022	0.009318	0.043024	0.034966	0.066918
19	S0019	2015	0.482713	0.097466	0.077881	0.111624	0.0948	0.005755	0.032458	0.019593	0.07771
20	S0020	2015	0.458778	0.107142	0.077872	0.108469	0.109104	0.008395	0.038153	0.029436	0.062651

Then we run a K-centroids Diagnostics using the K-means method to determine the number of clusters.



K-Means Cluster Assessment Report

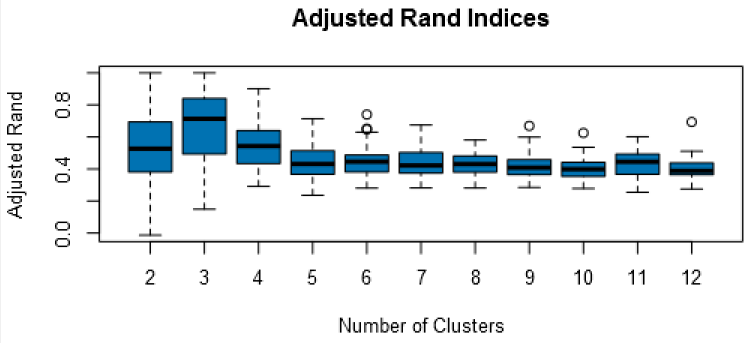
Summary Statistics

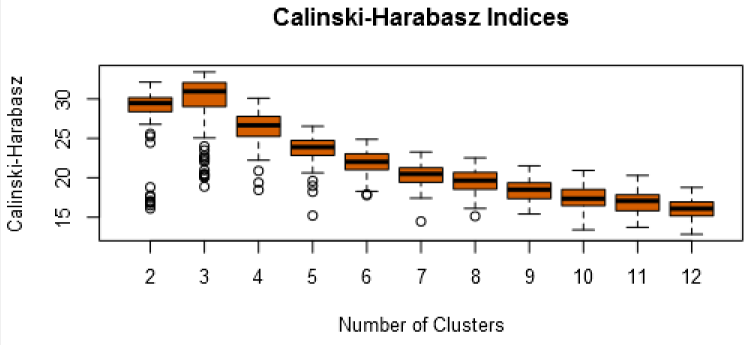
Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.01304	0.1486	0.291	0.2356	0.2802	0.282	0.2811
1st Quartile	0.3814	0.5074	0.435	0.3674	0.3838	0.3764	0.38
Median	0.5267	0.7132	0.543	0.4312	0.4452	0.4215	0.4305
Mean	0.5043	0.679	0.5409	0.4515	0.4438	0.4398	0.433
3rd Quartile	0.6942	0.8382	0.6336	0.5093	0.4866	0.5022	0.4786
Maximum	1	1	0.901	0.7141	0.7404	0.6743	0.5811
	9	10	11	12			
Minimum	0.2848	0.2781	0.2544	0.2749			
1st Quartile	0.3649	0.3543	0.3665	0.3645			
Median	0.408	0.3988	0.4446	0.3877			
Mean	0.4162	0.404	0.4331	0.3998			
3rd Quartile	0.4586	0.4412	0.4921	0.4376			
Maximum	0.6693	0.6251	0.6016	0.6938			

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	18.88	18.45	15.21	17.79	14.47	15.13
1st Quartile	28.38	29.11	25.26	22.84	21.03	19.43	18.56
Median	29.47	30.96	26.66	23.88	22.02	20.49	19.65
Mean	28.28	29.57	26.39	23.68	21.93	20.31	19.6
3rd Quartile	30.15	32.01	27.74	24.72	23.02	21.24	20.66
Maximum	32.13	33.41	30.09	26.53	24.87	23.27	22.53
	9	10	11	12			
Minimum	15.4	13.4	13.72	12.84			
1st Quartile	17.36	16.47	15.83	15.18			
Median	18.5	17.36	17.06	16.09			
Mean	18.42	17.48	16.89	16.07			
3rd Quartile	19.37	18.52	17.83	16.93			
Maximum	21.52	20.94	20.3	18.78			

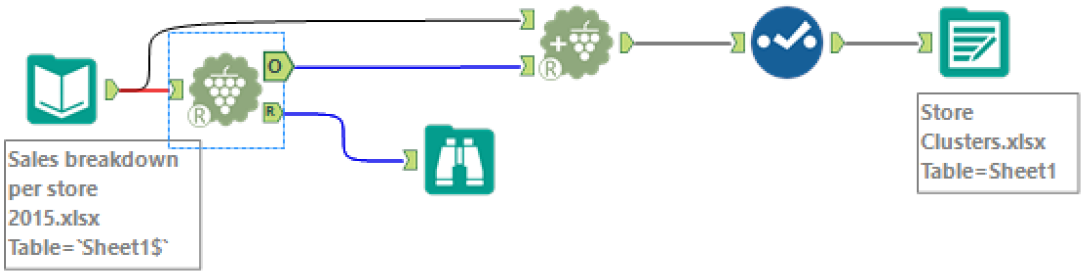




As per above Adjusted Rand Indices and Calinski-Harabasz Indices, we deduce that the optimal number of store format is three as the median of indices are higher for three and the spread of the variations is reasonable.

2. How many stores fall into each store format?

We use the K-Centroids Cluster Analysis tool to determine the number of stores that fall under each format and the summary is shown as per below.



Report

Summary Report of the K-Means Clustering Solution Clusters_results

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Share.of.Dry.Grocery + Share.of.Dairy + Share.of.Frozen.Foods + Share.of.Meat + Share.of.Produce + Share.of.Floral + Share.of.Deli + Share.of.Bakery + Share.of.General.Merchandises, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

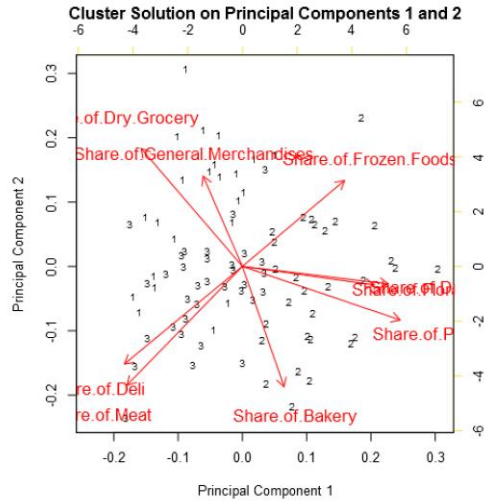
Cluster Information:

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.

	Share.of.Dry.Grocery	Share.of.Dairy	Share.of.Frozen.Foods	Share.of.Meat	Share.of.Produce	Share.of.Floral	Share.of.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

	Share.of.Bakery	Share.of.General.Merchandises
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389



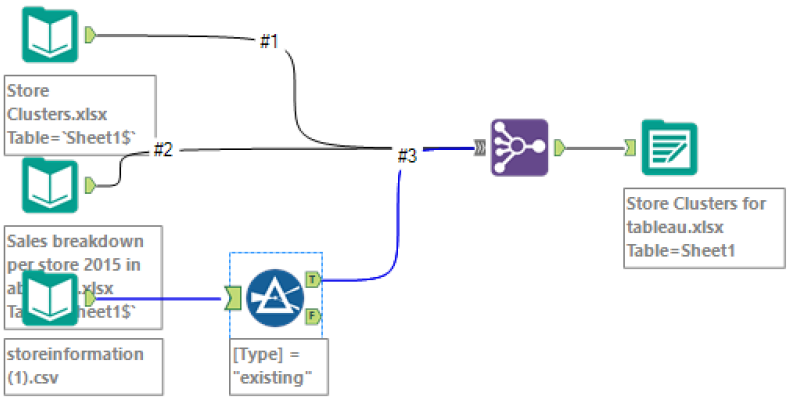
3. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**

Based on the results of the clustering model, the way to identify the clusters from one to another is to use the Append cluster tool and assign the cluster id to each store as per extract here.

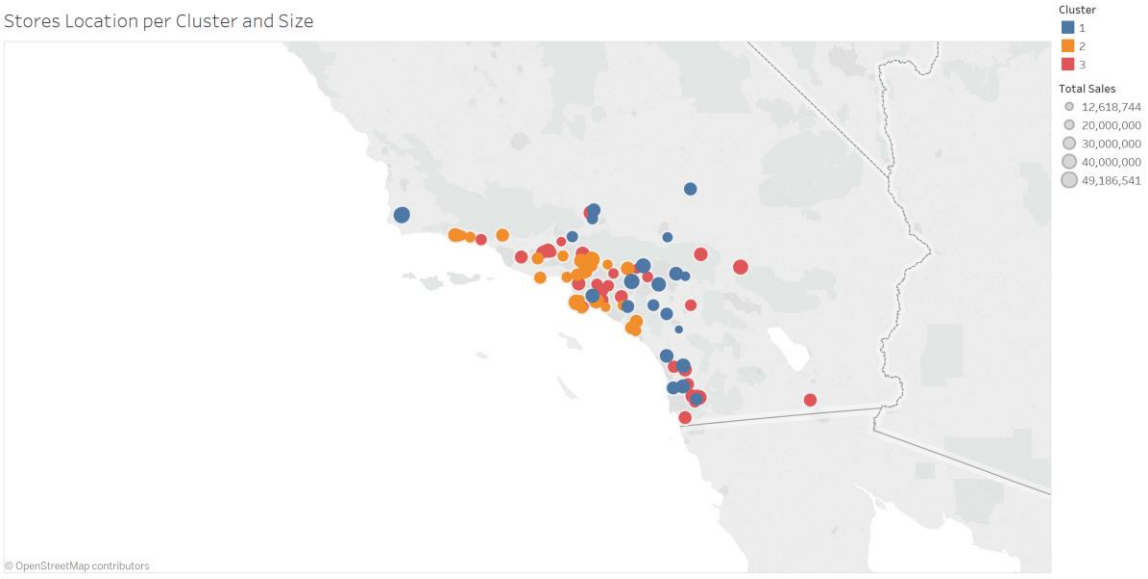
Record #	Store	Cluster
1	S0001	3
2	S0002	3
3	S0003	2
4	S0004	3
5	S0005	2

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.**

We first join all the information (cluster ID, total sales, store information) under the same excel sheet. Then we use Alteryx Public Geocoding App to obtain the Latitude and Longitude of the store.
Finally we obtain de below Tableau visualization to show the location of the stores by cluster and total sales in 2015.



https://public.tableau.com/profile/lina5384#!/vizhome/Project7-StoresLocationperClusterandSize/Sheet1?publish=yes



Task 2: Formats for New Stores

The grocery store chain is 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.

You've been asked to:

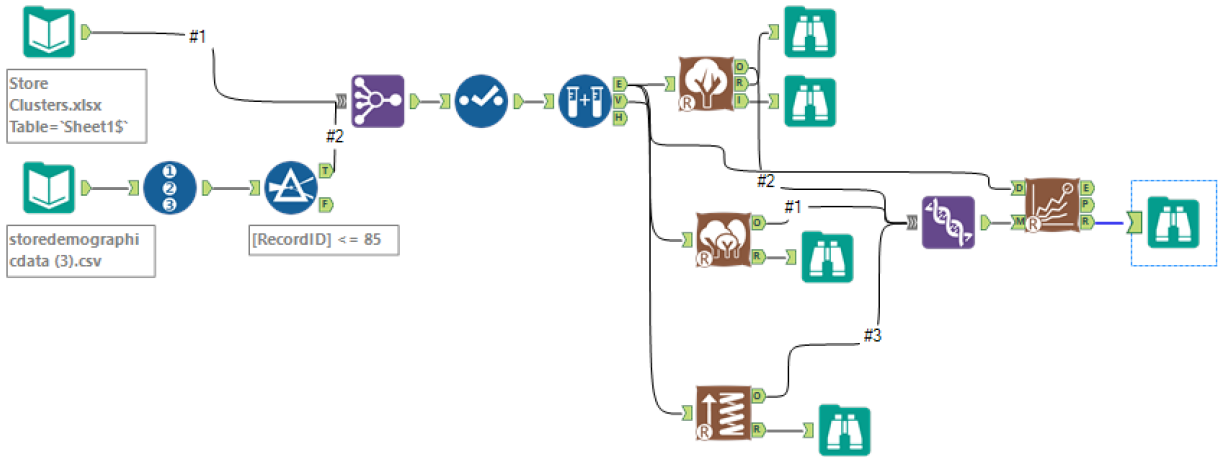
- Develop a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
- Use a 20% validation sample with Random Seed = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
- Use the model to predict the best store format for each of the 10 new stores.
- Use the StoreDemographicData.csv file, which contains the information for the area around each store.

Note: In a real world scenario, you could use PCA to reduce the number of predictor variables. However, there is no need to do so in this project. You can leave all predictor variables in the model.

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 have here a non-binary classification problem to allocate a cluster group to each new store. Therefore we run respectively a decision tree, forest and boosted model with a 20% validation sample and random seed of 3 to compare the models on the existing stores.

We note from the field summary tool that there was no missing data or outliers in the data provided.



Below are the model comparison results of the three models against the validation set.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750
Decision_Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

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

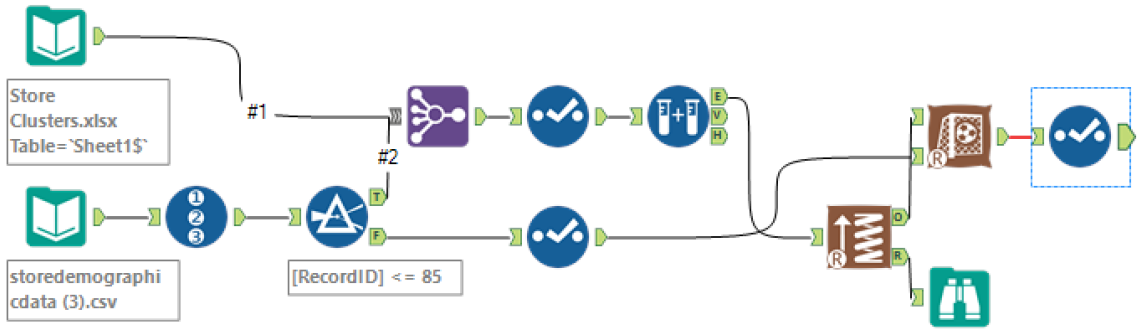
We note that

- ✓ the overall accuracy of the forest model and boosted model are the same at 82.35% and higher than the decision tree model.
- ✓ However, Boosted model's F1 score at 85.43% is higher than Forest model's F1 score at 82.51%

Therefore we choose the boosted model as the optimal model for the classification problem.

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

We run the score tool on the new stores using the boosted model and we obtain the below results



Record #	RecordID	Store	Score_1	Score_2	Score_3
1	86	S0086	0.348417	0.013522	0.638061
2	87	S0087	0.078987	0.804431	0.116582
3	88	S0088	0.486943	0.064498	0.448559
4	89	S0089	0.026597	0.935435	0.037968
5	90	S0090	0.019654	0.939601	0.040745
6	91	S0091	0.887418	0.003833	0.108749
7	92	S0092	0.028199	0.94173	0.030071
8	93	S0093	0.857561	0.005592	0.136847
9	94	S0094	0.00871	0.955864	0.035426
10	95	S0095	0.080423	0.641377	0.2782

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

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast.

Step 1: To forecast sales for existing stores you should aggregate sales across all stores by month and produce a forecast.

Step 2: To forecast sales for new stores:

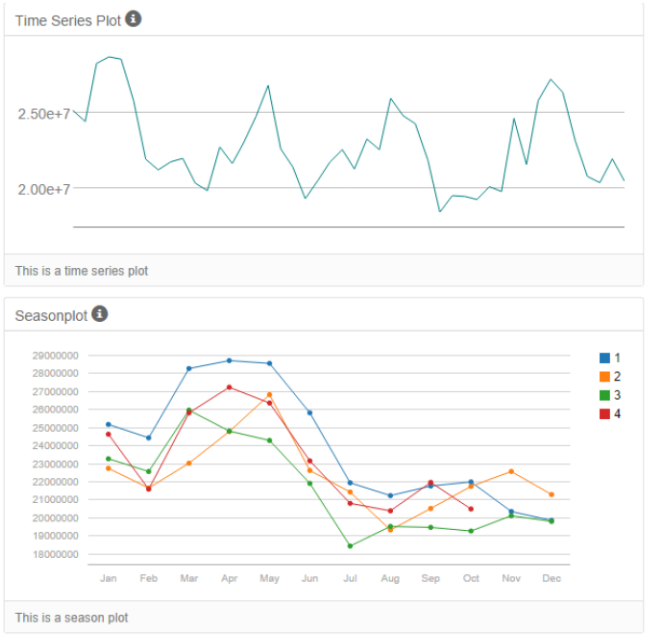
- Forecast produce sales (not total sales) for the average store (rather than the aggregate) for each segment.
- Multiply the average store sales forecast by the number of new stores in that segment.
- For example, if the forecasted average store sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
- Sum the new stores sales forecasts for each of the segments to get the forecast for all new stores.

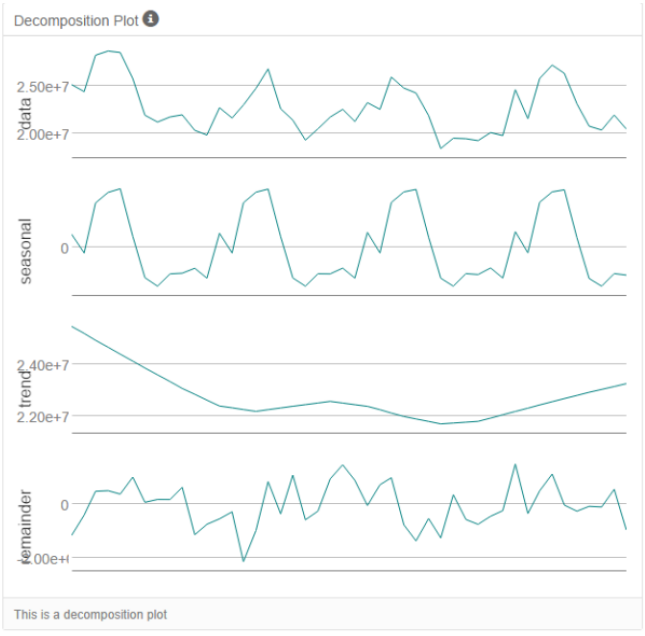
Step 3: Sum the forecasts of the existing and new stores together for the total produce sales forecast.

Forecast for the existing stores

For our forecast, we use time series model with a holdout sample of 12 months.

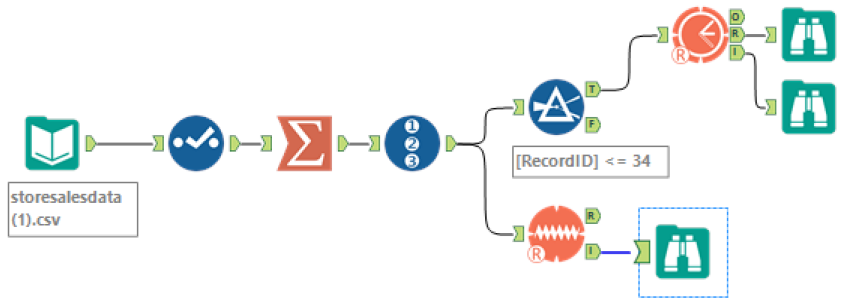
Below is the TS plot result of our historical monthly sales.





Based on the decomposition plot, we deduce the following terms for our model ETS (M,N,M)

- ✓ The error is multiplicative as the errors are growing and shrinking over time .
- ✓ The trend is none as the trend seems to decrease until July 14 and increase back in Aug 14
- ✓ The seasonality is multiplicative as the peaks change over time.



Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-241658.3191268	886787.7565482	699047.4732299	-1.1576764	3.1317204	0.3724833	0.069077

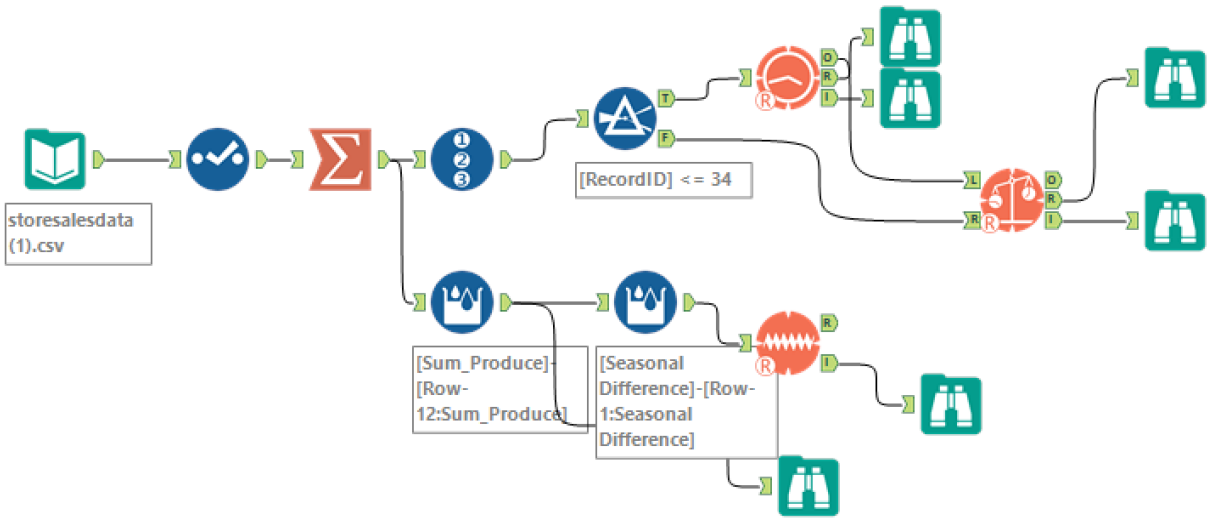
Information criteria:

AIC	AICc	BIC
1078.9536	1101.0588	1100.3226

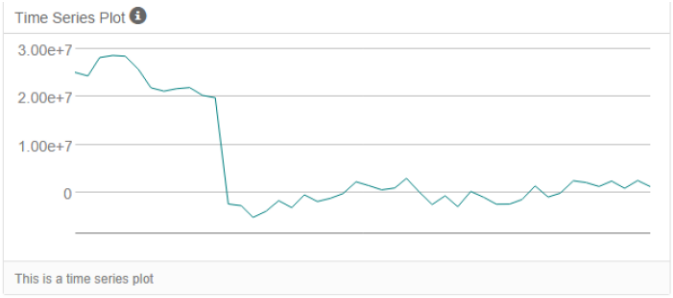
Smoothing parameters:

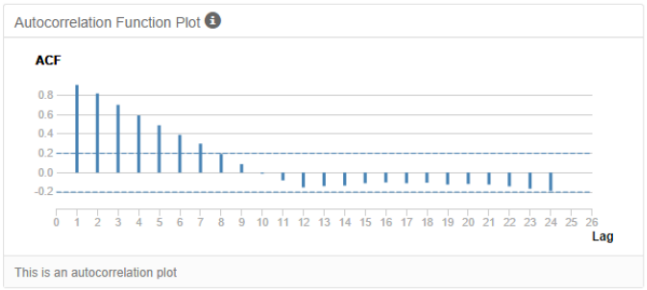
Parameter	Value
alpha	0.542014
gamma	1e-04

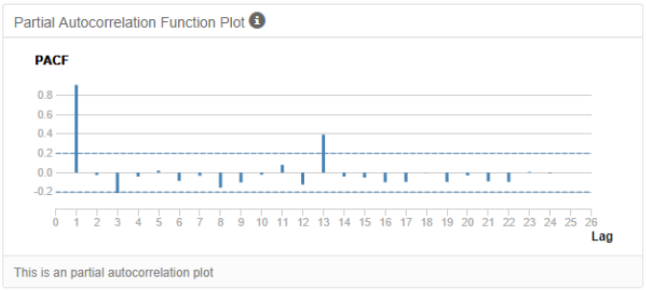
For the ARIMA model, since the dataset is seasonal we apply a seasonal difference and a seasonal first difference until the dataset is stationary.



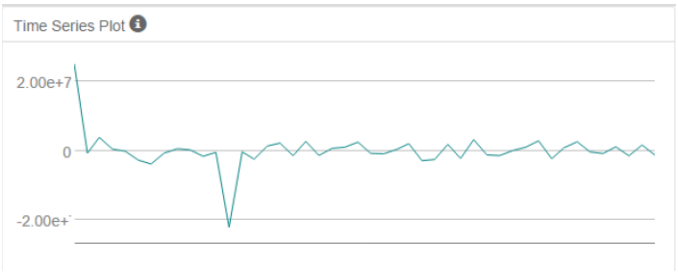
Time series plot, ACF and PACF of the seasonal difference dataset.

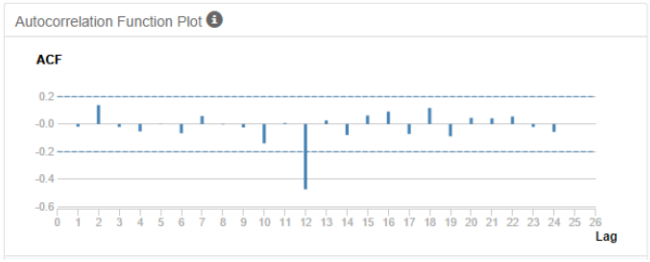


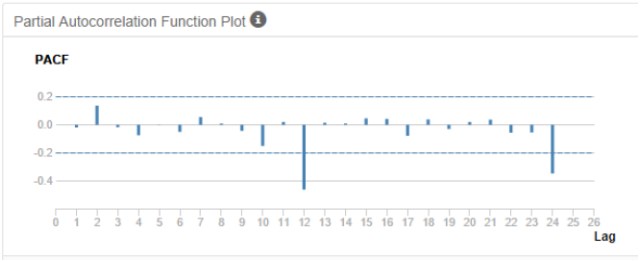




Time series plot, ACF and PACF of the seasonal first difference dataset.







We note that

- We note a negative auto-correlation at lag 1 in the ACF and PACF plot, and the partial autocorrelation drops after lag 1 and gradually with no other significant autocorrelation, which suggest a MA model therefore we choose “p=0”, “q=1”, “P=0”, “Q=0”.
- We use the seasonal difference and first seasonal difference to make our dataset stationary, therefore we choose “d=1” and “D=1”. We could have continue the differencing, but we noticed that it did not make much difference to the ACF and PACF graphs.
- We choose “m=12” as it is the number of period between each period.

Therefore we deduce the following model ARIMA (0,1,1) (0,1,0)12.

Summary of ARIMA Model ARIMA

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

Call:
Arima(Sum_Produce, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1
Value	-0.439899
Std Err	0.192927

sigma^2 estimated as 2970114829973.12: log likelihood = -331.46127

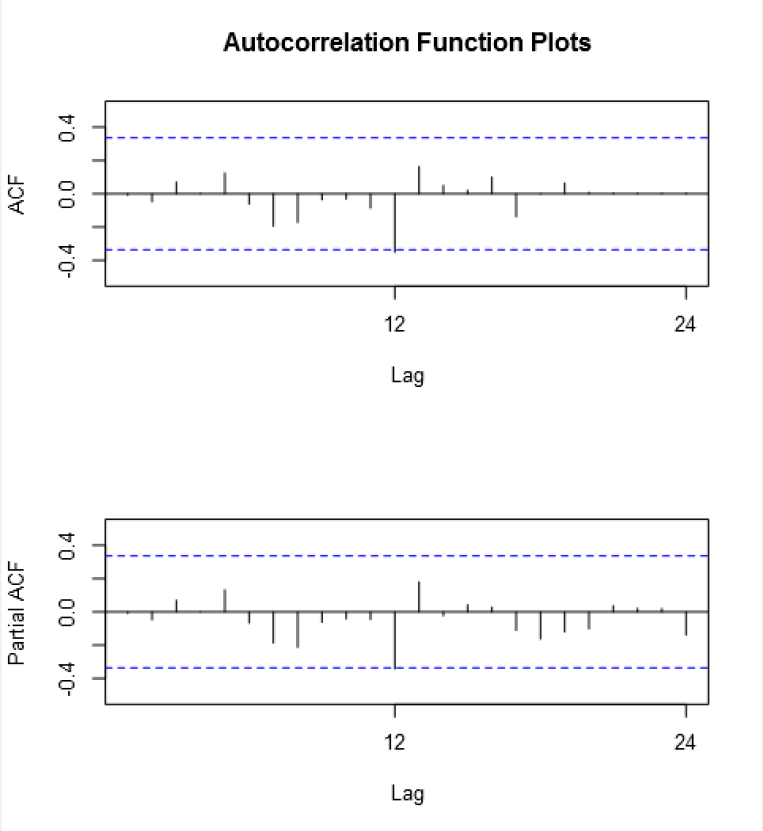
Information Criteria:

AIC	AICc	BIC
666.9225	667.5892	669.0116

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
36029.5333167	1354529.5861772	848505.992658	0.1090638	3.8058261	0.4521214	-0.0090018

Ljung-Box test of the model residuals:
Chi-squared = 4.347, df = 9, p-value = 0.910878



Forecast error measurement of both ETS and ARIMA models against the holdout sample.

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS
20088529.29	19592858.50364
19772333.34	18699732.54809
24608406.71	21034791.50129
21559729.45	20376418.33762
25792074.59	23165379.97327
27212464.15	23673094.98339
26338477.15	24203475.91264
23130626.6	21294826.71898
20774415.93	19013560.44774
20359980.58	18512113.97887
21936906.81	19125133.79398
20462899.3	19599985.21084

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	1978789	2200153	1978789	8.4769	8.4769	1.266	NA

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ARIMA
20088529.29	20747631.87903
19772333.34	19465586.61903
24608406.71	21450342.11903
21559729.45	20745161.41903
25792074.59	24146220.24903
27212464.15	22985351.92903
26338477.15	22466291.08903
23130626.6	20083162.35903
20774415.93	16610437.07903
20359980.58	17700745.44903
21936906.81	17647926.66903
20462899.3	17443558.24903

Accuracy Measures:

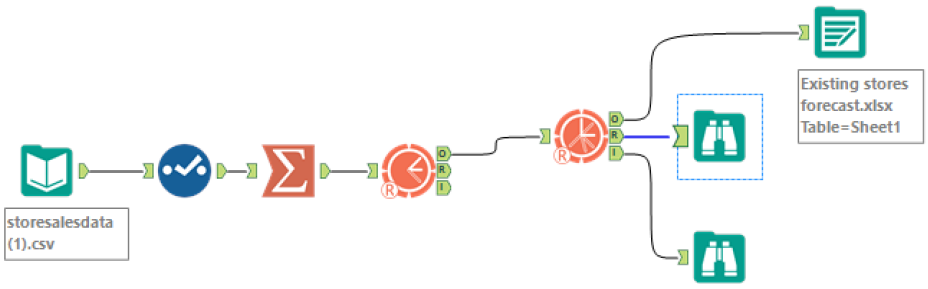
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988	NA

Comparing the on the in-sample error results and the forecast error measurement against holdout sample, we noted that :

- ✓ The in-sample errors results of ETS (M,N,M) show better results with higher AIC and lower RMSE and MASE compared to ARIMA (0,1,1) (0,1,0)12.
- ✓ For the forecast error measurement against holdout sample, the ETS model alsoshows lower error results than ARIMA model.

Therefore we choose ETS (M,N,M) for our forecast model.

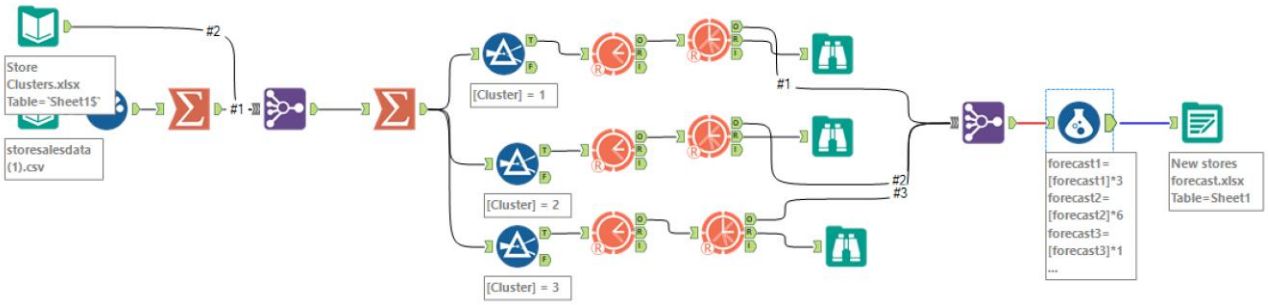
Below are the forecast for the existing stores based on ETS (M,N,M)



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21174989.40366	22840074.385973	22263729.953323	20086248.853997	19509904.421348
2016	2	20479354.577583	22316289.400956	21680461.698564	19278247.456602	18642419.75421
2016	3	23580340.680392	25927572.123553	25115112.826414	22045568.53437	21233109.237231
2016	4	22236546.234701	24649240.989504	23814122.539464	20658969.929937	19823851.479897
2016	5	25427255.457066	28396631.118655	27368825.841938	23485685.072193	22457879.795476
2016	6	26143967.404048	29399195.639379	28272446.740454	24015488.067643	22888739.168718
2016	7	26399993.267031	29879368.937501	28675034.733497	24124951.800565	22920617.596561
2016	8	23172393.880014	26386378.237661	25273905.29434	21070882.465689	19958409.522368
2016	9	20544268.638821	23528908.808356	22495819.94896	18592717.328682	17559628.469286
2016	10	20182471.085707	23241644.310338	22182756.941071	18182185.230344	17123297.861077
2016	11	20966876.352467	24271769.836858	23127830.049722	18805922.655212	17661982.868076
2016	12	20965097.001692	24391891.018402	23205757.172774	18724436.83061	17538302.984981

Forecast for the new stores

As recommended, we applied the ETS forecast model to the average sales of the total sum of monthly produce sales of each cluster, then we multiply the respective cluster monthly forecast with number of new store under the cluster as defined in task 2.



We obtain then the below aggregated forecast results

Record #	Period	Sub_Period	forecast1	forecast2	forecast3	Total Forecast New
1	2016	1	760301.225632	1605754.401712	224510.958351	2590566.585695
2	2016	2	739272.289217	1544121.219929	219741.588077	2503135.097223
3	2016	3	868600.130769	1783917.892385	257636.056356	2910154.07951
4	2016	4	816929.249288	1718234.850234	237029.092276	2772193.191798
5	2016	5	926270.352	1941934.228027	274057.895872	3142262.475899
6	2016	6	946409.212571	1976372.705043	280912.497017	3203694.414631
7	2016	7	955858.186618	1994769.01322	282808.916356	3233436.116193
8	2016	8	836496.457417	1801409.544849	246712.000887	2884618.003153
9	2016	9	741838.843637	1603994.618979	216255.220831	2562088.683447
10	2016	10	724700.737027	1568892.213186	213077.589443	2506670.539657
11	2016	11	761409.446964	1614667.763427	222073.621794	2598150.832185
12	2016	12	765642.606909	1574820.045008	225851.383712	2566314.03563

Below is a summary table combining new and existing stores forecast for 2016

Record #	Year	Month	Existing Stores	New Stores	Total Forecast
1	2016	1	21174989.40366	2590566.585695	23765555.989356
2	2016	2	20479354.577583	2503135.097223	22982489.674807
3	2016	3	23580340.680392	2910154.07951	26490494.759902
4	2016	4	22236546.234701	2772193.191798	25008739.426499
5	2016	5	25427255.457066	3142262.475899	28569517.932965
6	2016	6	26143967.404048	3203694.414631	29347661.81868
7	2016	7	26399993.267031	3233436.116193	29633429.383224
8	2016	8	23172393.880014	2884618.003153	26057011.883167
9	2016	9	20544268.638821	2562088.683447	23106357.322268
10	2016	10	20182471.085707	2506670.539657	22689141.625364
11	2016	11	20966876.352467	2598150.832185	23565027.184652
12	2016	12	20965097.001692	2566314.03563	23531411.037321

Tableau visualization

https://public.tableau.com/views/Project7-Task3/Sheet1?:embed=y&:display_count=yes&publish=yes