GROCERIES STORE CHAIN EXPANSION AND SALES PREDICTION ANALYSIS

Task 1: Determining the Store Format for Existing Stores

The 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. I've been asked to provide analytical support to make decisions about store formats and inventory planning.

So we would need to figure out what would be the optimal number of clusters. First, we have calculated what portion of the total sales of every store represents each type of product (meat, frozen food, bakery ...) We used only data from the most recent year 2015 to come out with the segmentation. We used K-means clustering model in order to decide the number of store formats.

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.012332	0.085005	0.129167	0.198479	0.172868	0.211424	0.197457
1st Quartile	0.055047	0.28273	0.279896	0.303745	0.294079	0.281472	0.321616
Median	0.492542	0.428163	0.388131	0.353296	0.351385	0.333331	0.353529
Mean	0.406457	0.411914	0.372189	0.366041	0.367644	0.354859	0.369188
3rd Quartile	0.61678	0.50506	0.450843	0.41474	0.453322	0.409187	0.404819
Maximum	1	0.925732	0.70085	0.586379	0.5548	0.703966	0.660004

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	9.056197	9.683921	11.14097	11.15269	9.474469	8.797239	8.769803
1st Quartile	17.976426	15.402516	13.27496	12.65426	11.988572	11.311079	10.838622
Median	19.836525	16.618434	14.49044	13.49543	12.537825	12.043325	11.303199
Mean	18.604945	16.309418	14.37112	13.46494	12.624375	11.910413	11.376818
3rd Quartile	20.889876	17.734502	15.56523	14.30924	13.365637	12.535052	11.963996
Maximum	21.992647	18.908142	16.79342	16.32568	15.329887	14.179165	13.936724

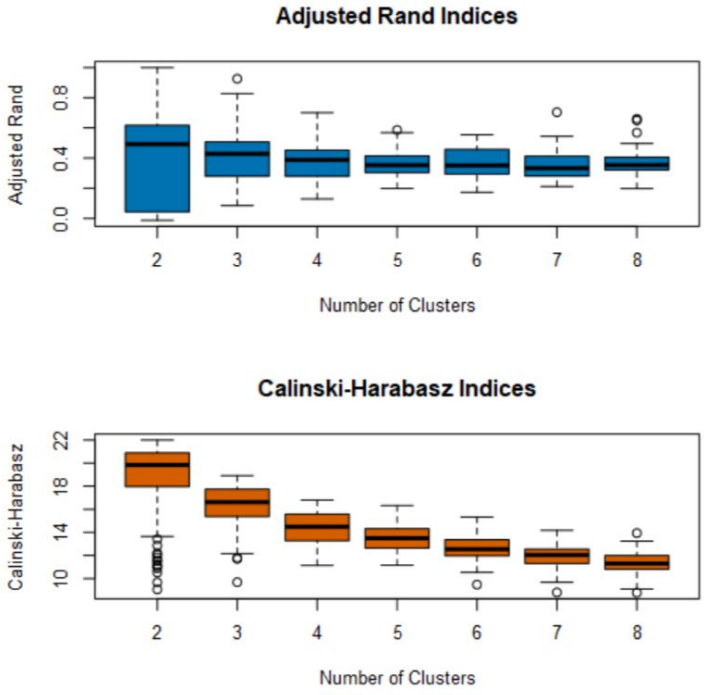


Fig. 1 AR and CH indices for the K-Means Clustering model

From the table and the plot, we can see that 3 clusters is the optimal number because it has a high median and low variability compared to using 2 or 4 clusters.

We have 23 stores in cluster1, 29 stores in cluster2 and 33 stores in cluster3.

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

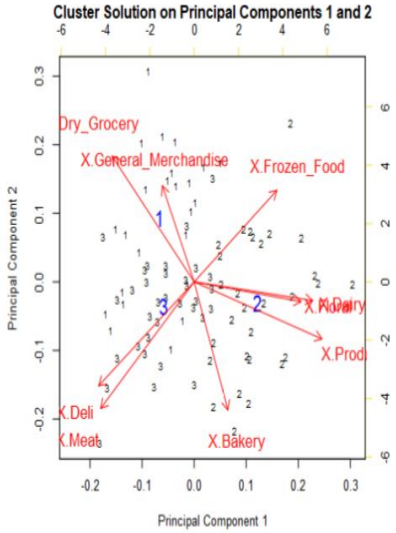


Fig 2. Stores per cluster

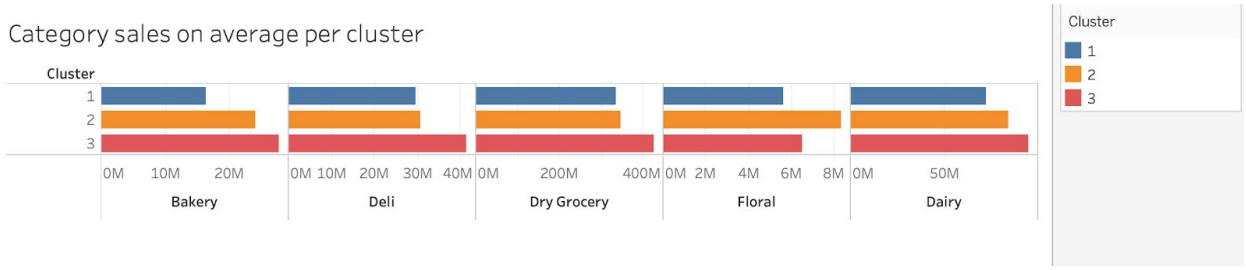


Fig 3. Category sales on average per cluster

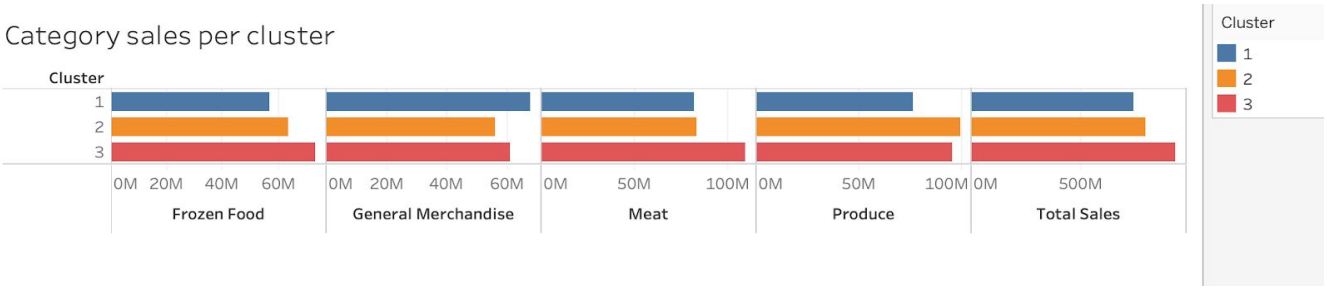


Fig 4. Category Sales and Total Sales on average per cluster

From Figures 3 and 4 we can see that:

1. Cluster 1 is oriented towards selling more General Merchandise than the other two clusters.
2. Cluster 2 is oriented towards selling more Produce and Floral products compared to the other 2 clusters.
3. Cluster 3 is oriented towards selling more Bakery, Meat, Deli, Dry Grocery, Dairy, Frozen Food. Cluster 3 also has the highest total sales.

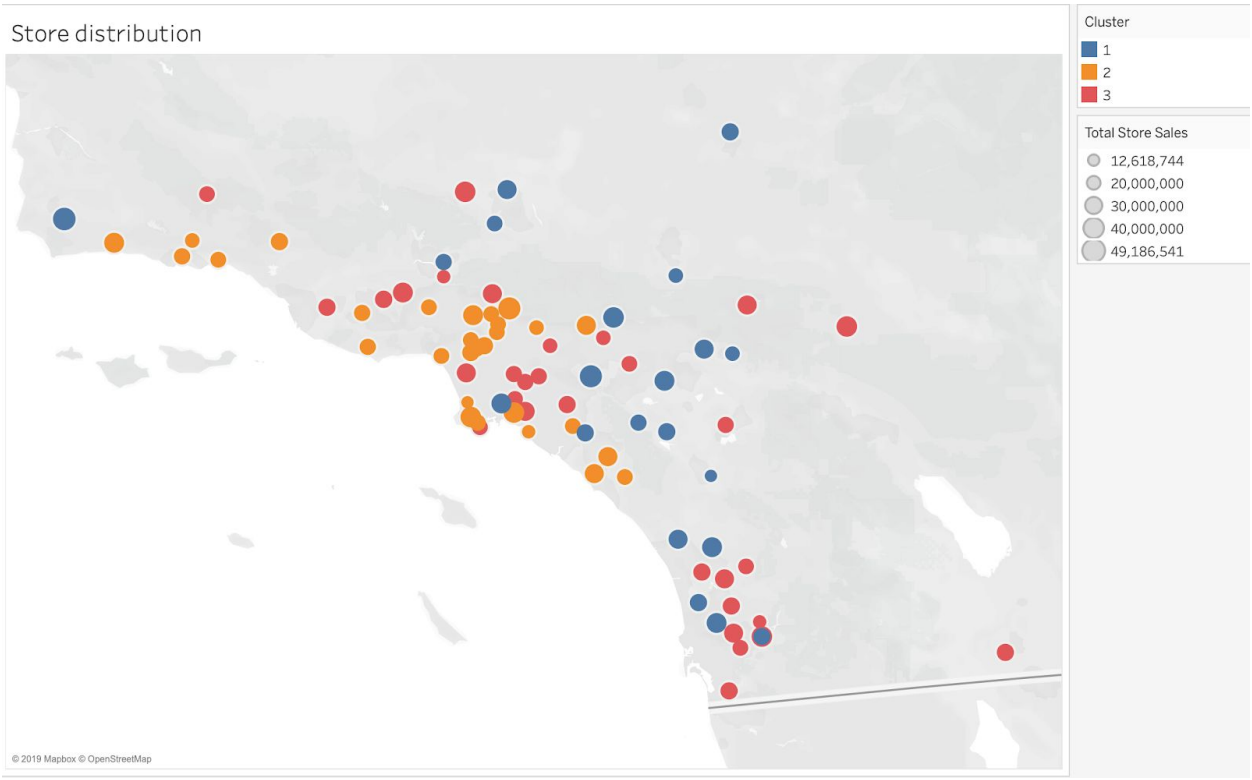


Fig 5. Store location

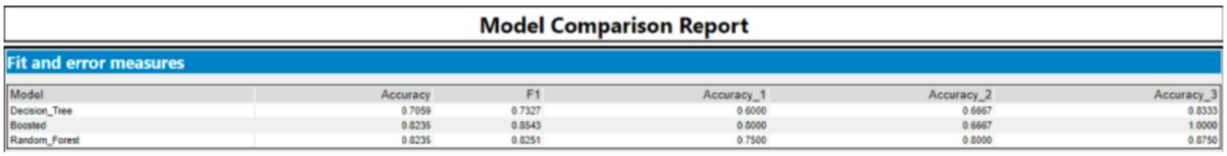
On the map, we can see the location of the stores. The color presents the cluster and the size of the dots refer to the size of the total sales. Here is a link to the map. Where we can zoom in an see the number of each store on the map.

Task 2: Store Format for New Stores

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.

Based on the demographics data for each store and based on the cluster we were able to determine the format of the 10 new stores using the boosted model. Two other

models were tested as well but the boosted model showed higher accuracy based on the F1 score.



Model Comparison Report

Fit and error measures

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

Fig 6. Model Comparison Report

Store Number	Cluster
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Fig 6. New store segmentation

Task 3: Forecasting

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast. To do that we will use ETS and ARIMA models.

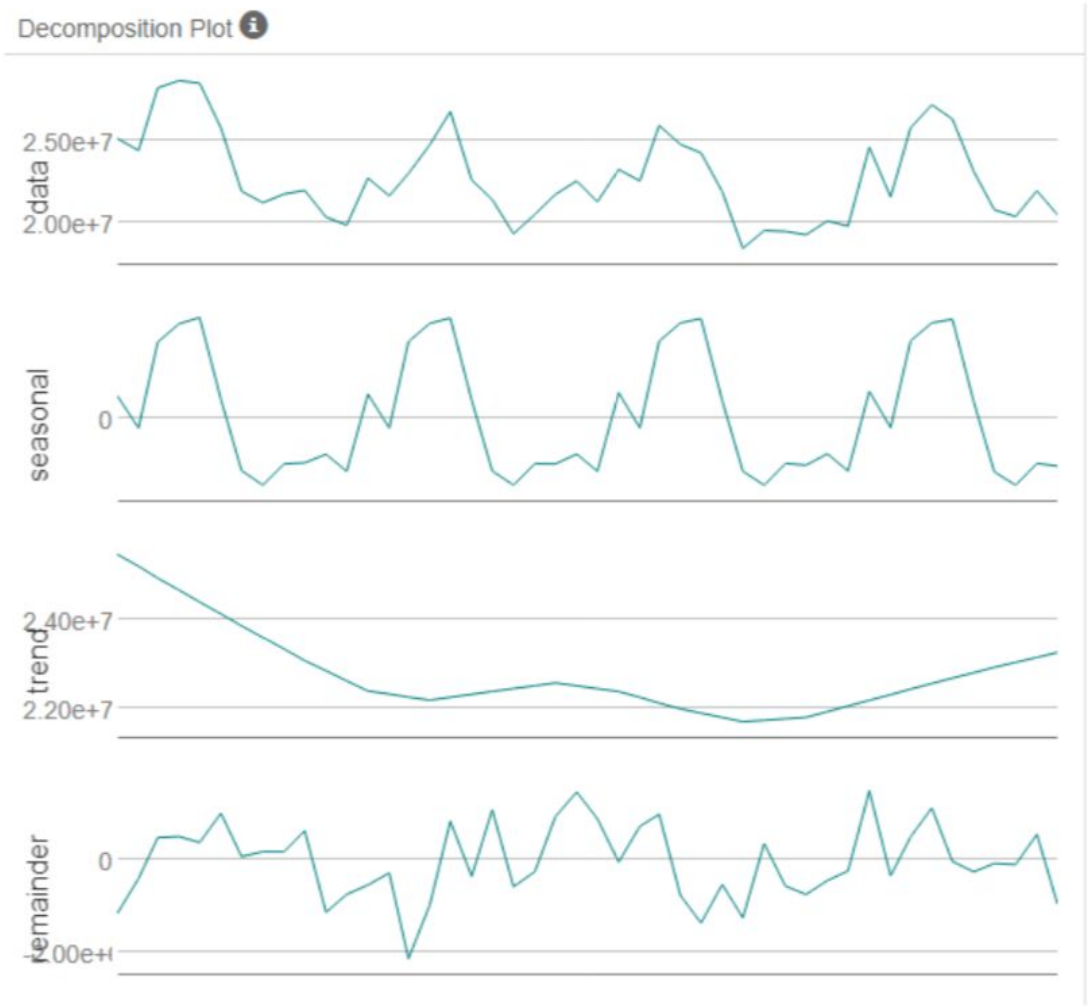


Fig 6. Time series decomposition plot

From the plot, we can see the seasonal component decreases slightly each year. Since we have change in magnitude multiplicative method will be used for that component rather than additive. There is no trend since the trend goes down then up. The error component also changes in magnitude hence multiplicative method will be used for it as well. All of that suggests that ETS(M,N,M) model will be used as the optimal model for forecasting. What further supports that is the forecasting errors against the holdout sample that we have obtained which are lower compared to other ETS models and are also lower than the optimal type of ARIMA model ARIMA(1,1,0)(1,0,0)[12].

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_M_N_M_	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA_1_0_0__1_1_0_	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463

Fig 7. Forecasting errors against the holdout sample

The errors for the ETS(M,N,M) are the lowest which suggests that it is the most accurate model and will give us the most accurate sales prediction for the stores.

Year	Month	Actual stores	New stores
2016	1	21539936,01	2587450,85
2016	2	20413770,60	2447352,89
2016	3	24325953,10	2913185,24
2016	4	22993466,35	2775745,61
2016	5	26691951,42	3150866,84
2016	6	26989964,01	3188922,00
2016	7	26948630,76	3214745,65
2016	8	24091579,35	2866348,66
2016	9	20523492,41	2538726,85
2016	10	20011748,67	2488148,29
2016	11	21177435,49	2595270,39
2016	12	20855799,11	2573396,63

Fig 8. Forecasts sales for the new and existing stores

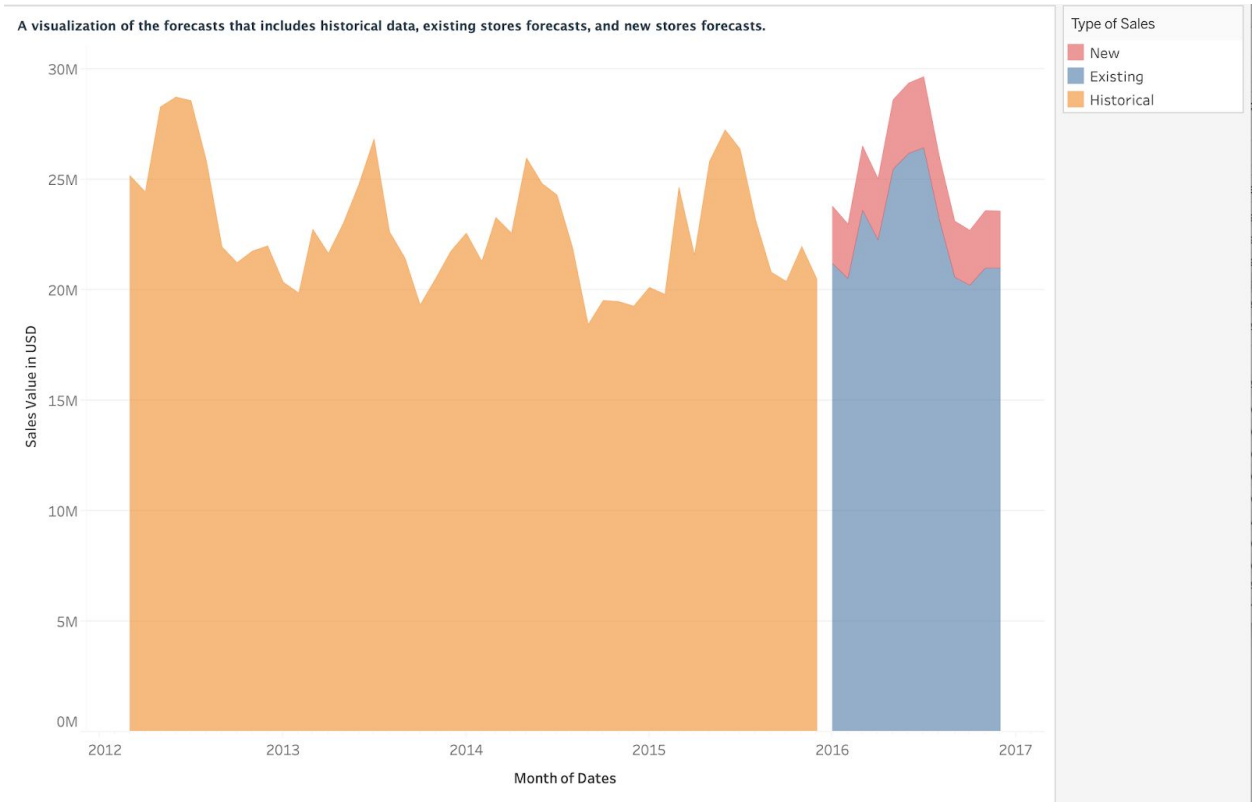


Fig 8. A visualization of the forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

The forecast of new stores represents 11% of 2016 total sales including existing stores and new ones. The growth vs last year of existing stores would be 1.6% but with the launch of the new 10 stores, the grocery chain could achieve a growth of around 13.9%.