Project 5: Combine Predictive Techniques

1. Determine Store Formats for Existing Stores

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

Table 1 - K-Means Cluster Assessment Report: Adjusted Rand Indices

	2	3	4	5	6	7	8
Minimum	-0.016293	0.27351	0.335359	0.336327	0.318262	0.230196	0.27786
1st Quartile	0.352041	0.515917	0.445826	0.409773	0.366788	0.358895	0.377341
Median	0.526785	0.66768	0.538528	0.497192	0.423541	0.416509	0.428806
Mean	0.53781	0.664773	0.565975	0.50103	0.45115	0.432196	0.421514
3rd Quartile	0.734477	0.826692	0.644691	0.555087	0.499921	0.502931	0.458601
Maximum	1	1	0.975264	0.852076	0.8539	0.683894	0.647983

Table 2 - K-Means Cluster Assessment Report: Calinski-Harabasz (CH) Indices

	2	3	4	5	6	7	8
Minimum	16.61829	17.38103	20.28456	18.61989	17.8746	15.98702	16.16824
1st Quartile	28.17383	28.57484	25.20913	22.93454	21.30575	19.85155	18.71365
Median	29.46587	31.05384	26.53788	24.086	22.16245	20.97743	19.6662
Mean	28.45131	29.70664	26.41806	23.87003	22.02174	20.77195	19.65973
3rd Quartile	30.17907	32.08726	27.59305	25.10099	23.06602	21.72942	20.7099
Maximum	31.78345	33.63781	30.1583	26.63063	24.72038	24.63982	22.95166

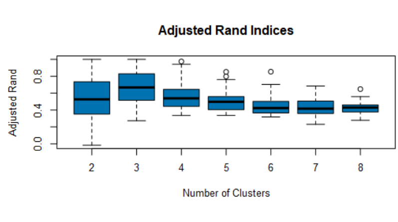


Figure 1 - K-Means Cluster Assessment Report: Adjusted Rand Indices (Box and Whisker Plots)

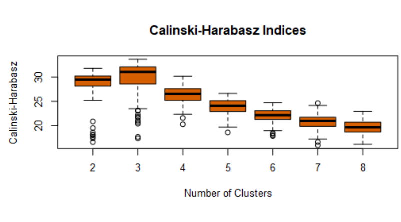


Figure 2 - K-Means Cluster Assessment Report: Calinski-Harabasz (CH) Indices (Box and Whisker Plots)

The optimal number of stores formats is 3, when both the indices registered the highest median value using the Adjusted Rand and the Calinski Harabasz Index.

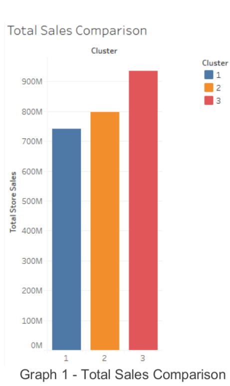
1.2. How many stores fall into each store format?

Table 3 - Report of the number of existing stores fall into each cluster

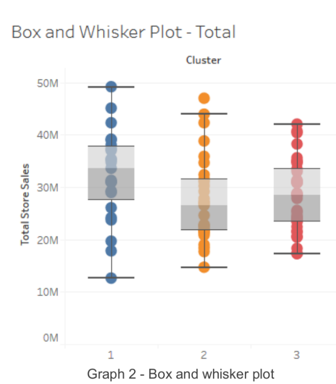
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

There are 23 stores on the Cluster 1, Cluster 2 has 29 stores while Cluster 3 has 33 stores.

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



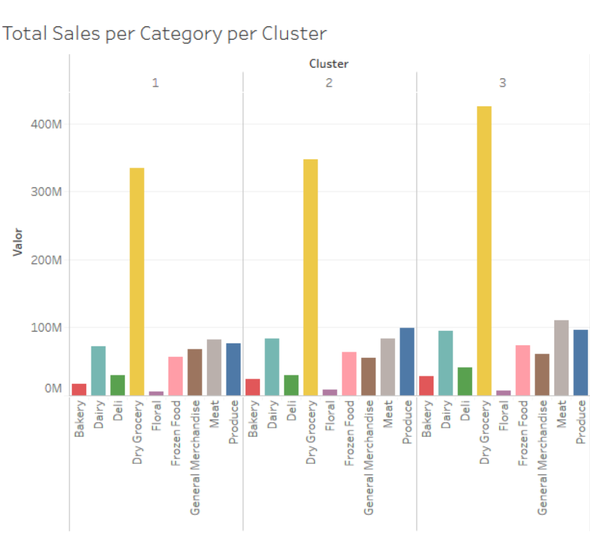
Graph 1 - Total Sales Comparison



Graph 2 - Box and whisker plot

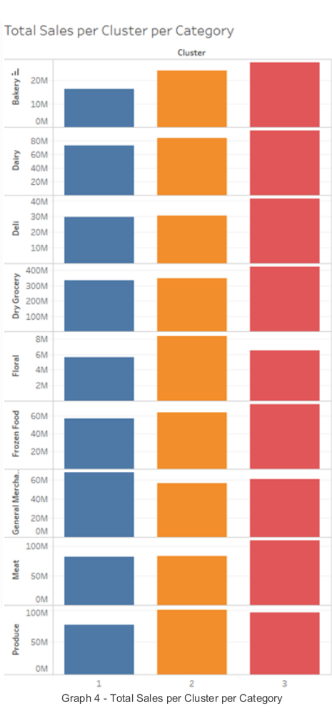
According to Graph 1, we could see that the Cluster 3 has the most total sales of them and has the most concentrated sales values, according to Graph 2, on the other hand, Cluster 1 has the lowest total sales and has the most sparse sales values.

Cluster 2 is on the middle ground between Cluster 1 and 3, both in total sales value and in spacing.



Graph 3 – Total Sales per Category per Cluster

After analyzing above Graph 3, is remarkable the difference between the dry grocery category and the other values. One of the main factors of Cluster 3 has the highest sales values is because of its dry grocery category.



Graph 4 - Total Sales per Cluster per Category

Graph 5 – Box and Whisker plot - Categories

Graph 6 – Circle views - Categories

After analyzing Graph 4, Graph 5 and Graph 6, we could see that even though the dry grocery is the largest source of revenue for all clusters, we can see some specificities of each cluster, like the following:

Cluster 1:
Top 1 – General Merchandising
Top 2 – None
Top 3 – Bakery, Diary, Deli, Dry Grocery, Floral, Frozen Food, Meat, Produce

Cluster 2:
Top 1 – Floral, Produce
Top 2 – Bakery, Diary, Deli, Dry Grocery, Frozen Food, Meat
Top 3 – General Merchandising

Cluster 3:
Top 1 – Bakery, Diary, Deli, Dry Grocery, Frozen Food, Meat
Top 2 – Floral, General Merchandising, Produce
Top 3 – None

- The first cluster has significant higher General Merchandising sales than other categories and the source of revenue of this cluster comes very diluted from each category, besides being the cluster with less effectiveness in sales. This cluster could be named as "Just Things";
- The second cluster has the most balanced incomes and the main characteristic of its consumer is their likeability of buying Produce and Floral things and not likely to buying General Merchandising. This cluster could be named as "Organics";
- The third cluster is the golden goose of this retail company! Cluster 3 stores are the most similar in terms of sales due to more compact range. The main characteristic of its consumer is their likeability of buying things with a high value-added product (and consequently high revenues and margins) as Meats, Bakery and Frozen Food. This cluster could be named as "Bon Vivant"

1.4. Please provide a Tableau visualization that shows the location of the stores, uses color to show cluster, and size to show total sales.

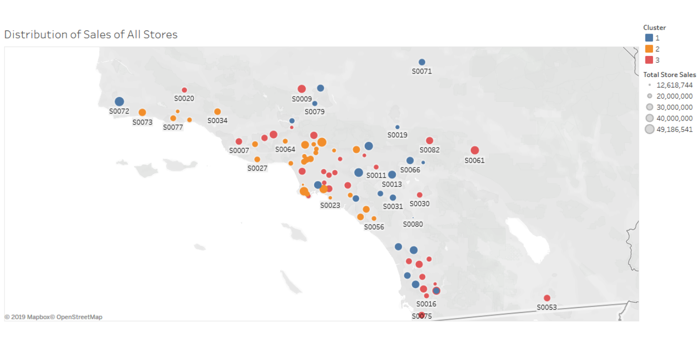


Figure 3 - Location of the stores

1.5. Workflow

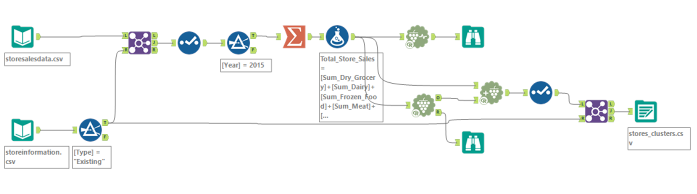


Figure 4 - Task 1 Workflow – Calculating number of cluster based on K-mean clustering model

2. Formats for New Stores

2.1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?

The following Model Comparison Report shows comparison matrix between Decision Tree, Forest Model and Boosted Model. Boosted Model is chosen despite having same accuracy as Forest Model due to higher F1 value.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Cluster	0.7059	0.7685	0.7500	1.0000	0.5556
FM_Cluster	0.8235	0.8426	0.7500	1.0000	0.7778
BM_Cluster	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 * 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 BM_Cluster

	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_Cluster

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

Confusion matrix of FM_Cluster

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

Figure 5 - Model Comparison Report

2.2. What format do each of the 10 new stores fall into?

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

2.3. Workflow

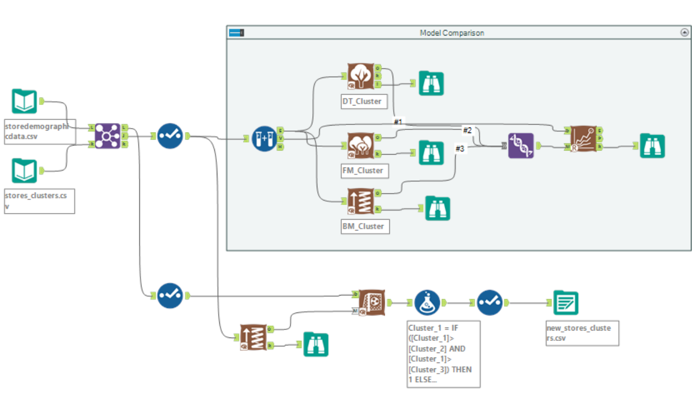


Figure 6 - Workflow used to assign cluster to new stores

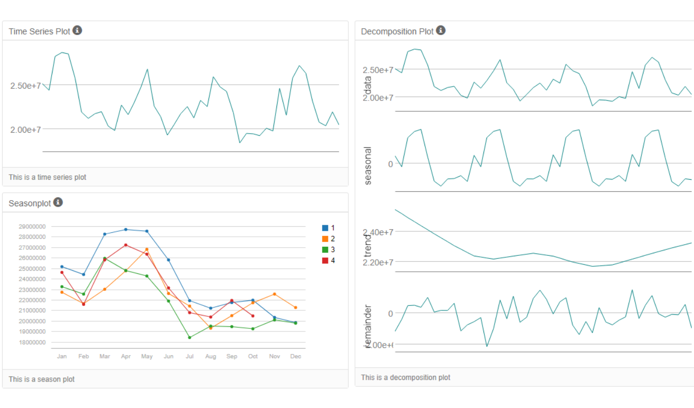
3. Predicting Produce Sales

3.1. What type of ETS or ARIMA model did you use for each forecast? How did you come to that decision?

To decide whether we'll use ETS or ARIMA model, first we will check how the Time Series behaves:

Looking for seasonality, we could see that it shows increasing trend and should be multiplicatively. The trend plot doesn't show any trending, and nothing should be applied. Its error is irregular and should be applied multiplicatively.

ETS(M,N,M) with no dampening should be used for ETS model.



Graph 7 - Time Series Plot with Decomposition Plot, without differencing

Because of the seasonality on these series, we need to differentiate our Time Series in order to Stationarize the series, as following.

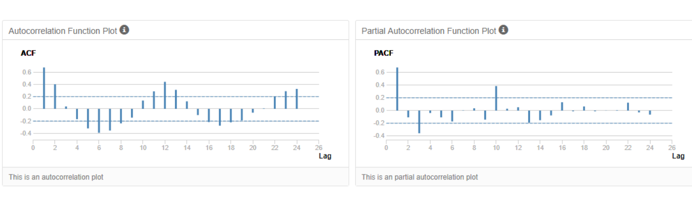


Figure 7 - ACF and PACF plot of non-seasonal component of the ARIMA model with one differencing

First, we need to look at the seasonal differencing component, to allow us to account for the value as observed in the same season one year earlier, as figure below.

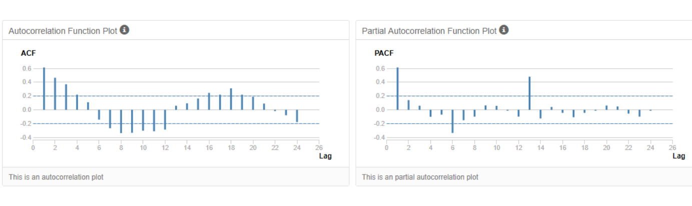


Figure 8 - ACF and PACF plot of the seasonal component of ARIMA

So, looks like we have to take the first seasonal difference to correct for seasonality before the dataset stationary

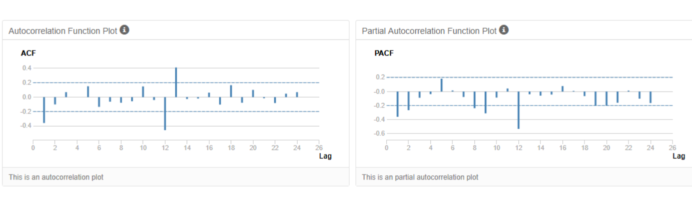


Figure 9 - ACF and PACF plot after taking the first differencing of the seasonal component of the ARIMA

After plotting the first seasonal difference, we can see that the series has stationarized. We can see this through our ACF and PACF plots, the serial correlational has now disappeared.

For the ARIMA model, the set **ARIMA(0,1,2)(0,1,0)** was chosen, seasonal difference and seasonal first difference were performed. There is a lag-2. The parameters determined for the ARIMA are based on ACF and PACF plots (above)

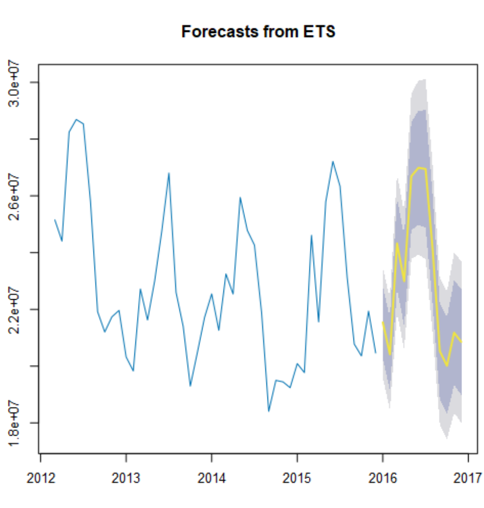
Table 4 - Accuracy between ETS and ARIMA models

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

Based on above Table 4 results, which was obtained from running the two time-series models against the holdout sample of 6 months data, the **ETS model's accuracy is higher when compared to ARIMA model**. ETS model has lower RMSE value and lower MASE value.

The following Graph 8 and table below shows actual and forecast value with 80% & 95% confidence level interval.



Graph 8 - Forecasts from ETS Model - Actual and forecast value with 80% & 95% confidence level interval

Table 5 - Forecasts from ETS Model - Actual and forecast value with 80% & 95% confidence level interval

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

3.2. Workflow

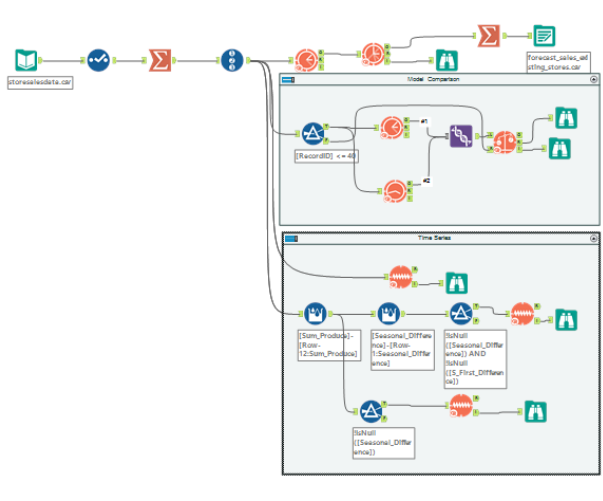


Figure 10 – Workflow used to forecast the sale value for the average store in 2016

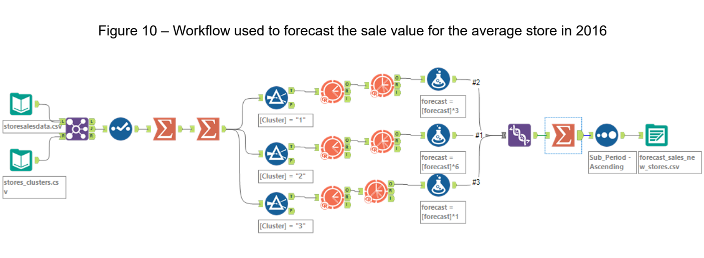


Figure 11 - Workflow used to forecast the sum produce sale for each cluster

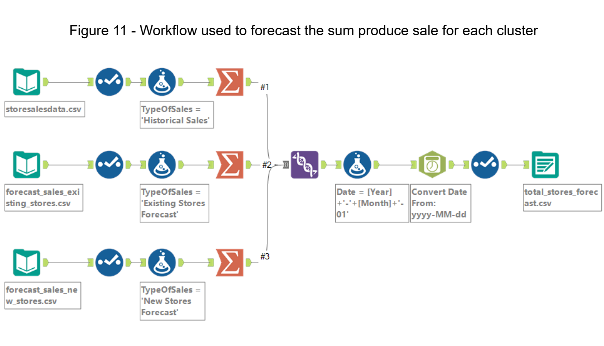


Figure 12 - Workflow used to generate the dataset for Tableau forecast plot

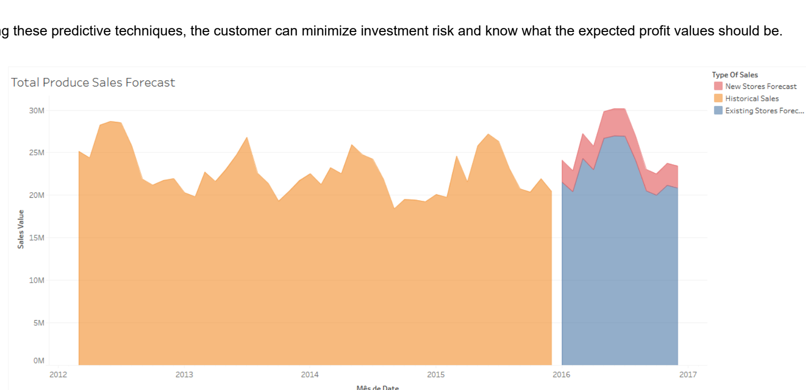
3.3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Looking at the following table, we can see the forecast sales for existing stores and new stores. New store sales was obtained by using **ETS(M,N,M)** analysis with all the 3 individual cluster to obtain the average sales per store. The average sales value (x3 cluster 1, x6 cluster 2, x1 cluster 3) are added up produce New Store Sales.

Table 6 - Forecasted sales for the next 12 months for both existing and new stores

Year	Month	New Store Sales	Existing Store Sales
2016	1	$ 2,587,450.85	$ 21,539,936.01
2016	2	$ 2,477,352.89	$ 20,413,770.60
2016	3	$ 2,913,185.24	$ 24,325,953.10
2016	4	$ 2,775,745.61	$ 22,993,466.35
2016	5	$ 3,150,866.84	$ 26,691,951.42
2016	6	$ 3,188,922.00	$ 26,989,964.01
2016	7	$ 3,214,745.65	$ 26,948,630.76
2016	8	$ 2,866,348.66	$ 24,091,579.35
2016	9	$ 2,538,726.85	$ 20,523,492.41
2016	10	$ 2,488,148.29	$ 20,011,748.67
2016	11	$ 2,595,270.39	$ 21,177,435.49
2016	12	$ 2,573,396.63	$ 20,855,799.11
Total		**$ 33,370,159.89**	**$ 276,563,727.27**

Using these predictive techniques, the customer can minimize investment risk and know what the expected profit values should be.



Graph 9 - Historical and forecast sales for existing stores and new stores over the period from Mar-12 to Dec-16