Predictive Analytics Capstone

Task 1: Determine store formats for existing stores.

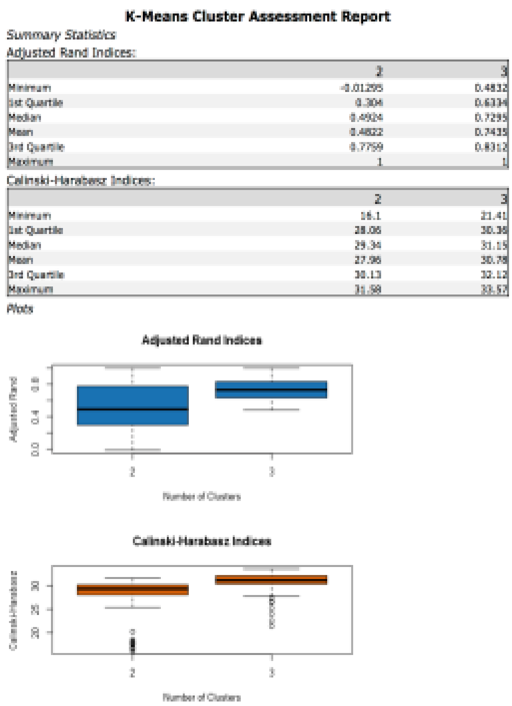
The Business Scenario: Store Format

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 the same, 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.

What is the optimal number of store formats?

In order to determine the best number of store formats to use. A k-centroids analysis was done using k-means clustering method for k=3,4,5,6.

Below are the results:


K-Means Cluster Assessment Report

Summary Statistics

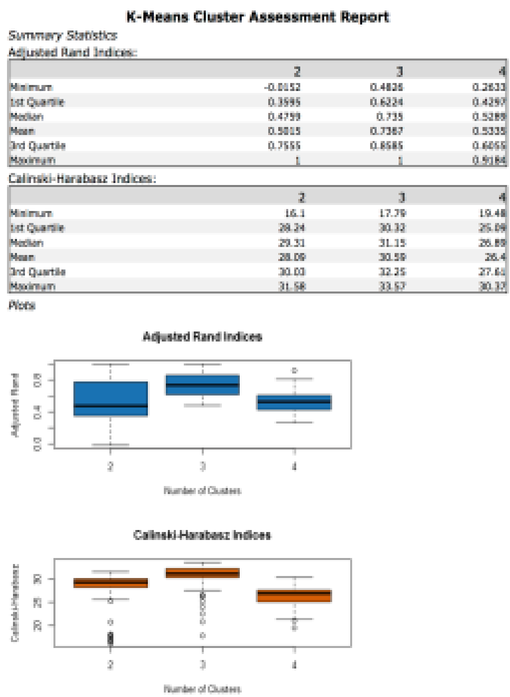
Adjusted Rand Indices:

	2	3
Minimum	-0.01295	0.4832
1st Quartile	0.304	0.6334
Median	0.4924	0.7295
Mean	0.4822	0.7435
3rd Quartile	0.7759	0.8312
Maximum	1	1

Calinski-Harabasz Indices:

	2	3
Minimum	16.1	21.41
1st Quartile	28.06	30.36
Median	29.34	31.15
Mean	27.96	30.78
3rd Quartile	30.13	32.12
Maximum	31.58	33.57

Plots


K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4
Minimum	-0.0152	0.4826	0.2633
1st Quartile	0.3595	0.6224	0.4297
Median	0.4759	0.735	0.5289
Mean	0.5015	0.7387	0.5335
3rd Quartile	0.7555	0.8585	0.6055
Maximum	1	1	0.9184

Calinski-Harabasz Indices:

	2	3	4
Minimum	16.1	17.79	19.48
1st Quartile	28.24	30.32	25.09
Median	29.31	31.15	26.89
Mean	28.09	30.59	26.4
3rd Quartile	30.03	32.25	27.61
Maximum	31.58	33.57	30.37

Plots

K-Means Cluster Assessment Report

Summary Statistics

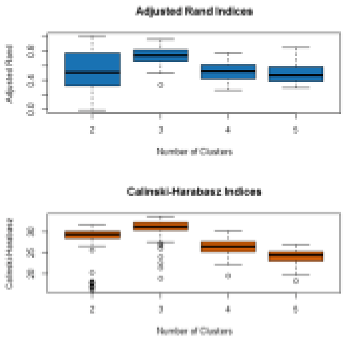
Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.0152	0.3383	0.2633	0.2967
1st Quartile	0.3384	0.6596	0.4186	0.3877
Median	0.5092	0.7422	0.5244	0.4692
Mean	0.5014	0.7361	0.528	0.4896
3rd Quartile	0.7759	0.8185	0.6115	0.5851
Maximum	1	0.9586	0.7758	0.8518

Calinski-Harabasz Indices:

	2	3	4	5
Minimum	16.1	18.05	19.48	18.23
1st Quartile	28.4	30.32	25.22	22.99
Median	29.34	31.14	26.43	24.41
Mean	28.07	30.65	26.43	23.97
3rd Quartile	30.03	32.22	27.57	25.18
Maximum	31.58	33.57	30.14	26.85

Plots



K-Means Cluster Assessment Report

Summary Statistics

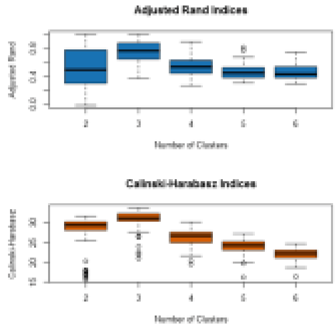
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.0152	0.3682	0.2633	0.3041	0.2923
1st Quartile	0.304	0.6459	0.4432	0.3821	0.3802
Median	0.4925	0.7816	0.533	0.4475	0.4314
Mean	0.4816	0.7503	0.5423	0.4669	0.4496
3rd Quartile	0.7555	0.8894	0.6245	0.5231	0.5332
Maximum	1	1	0.8921	0.811	0.7427

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.65	20.96	19.48	16.2	16.36
1st Quartile	28.06	30.35	24.93	23.07	21.12
Median	29.32	31.12	26.67	24.23	22.18
Mean	27.91	30.7	26.37	23.93	22.05
3rd Quartile	30.13	32.25	27.56	25.18	23.03
Maximum	31.58	33.57	30.14	27.19	24.62

Plots



From the above diagnostics, cluster 3 seems to be the best cluster and therefore I will use cluster 3 as my base to compare the number of k terms. Comparing cluster 3, between all the differing k terms, it looks like using k=3 for my analysis offers the best results with cluster 3 having the tightest range and highest mean when k=3.

The optimal number of store formats in my opinion is 3.

How many stores fall into each store format?

Store Format	Number of stores
1	23
2	29
3	33

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

Below is the summary of the cluster analysis for k=3.

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

Convergence after 12 iterations.

Sum of within cluster distances: 196.83135.

	X._Dry_ Grocery	X._Dairy	X._Frozen_ Food	X._Meat	X._ Produce	X._Floral	X._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

	X._Bakery	X._General_ Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Based on the results above, I can see cluster 1 is most positive for percentage of general merchandise sales vs cluster 3 which is the most negative. This would indicate that these two clusters are the most different in terms of the variable for percentage of general merchandise sales.

Tableau Visualization of the location of the stores, showing clusters and size due to total sales of each store.

Map of Store Locations, Cluster Group and Total Sales.

Map based on Longitude (generated) and Latitude (generated). Color shows Cluster. Size shows Total Sales. Details are shown for Zip.

Task 2: Formats for New Stores.

The Business Scenario: 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.

What methodology did you use to predict the best store format for the new stores?

In order to predict the store formats for the new stores, demographic data from StoreDemographicData.csv was used. All the variables were kept as predictor variables and run through a boosted, decision tree and random forest model. An 80/20 split of the data was used for training and validating the models.

Below is a comparison of the models:

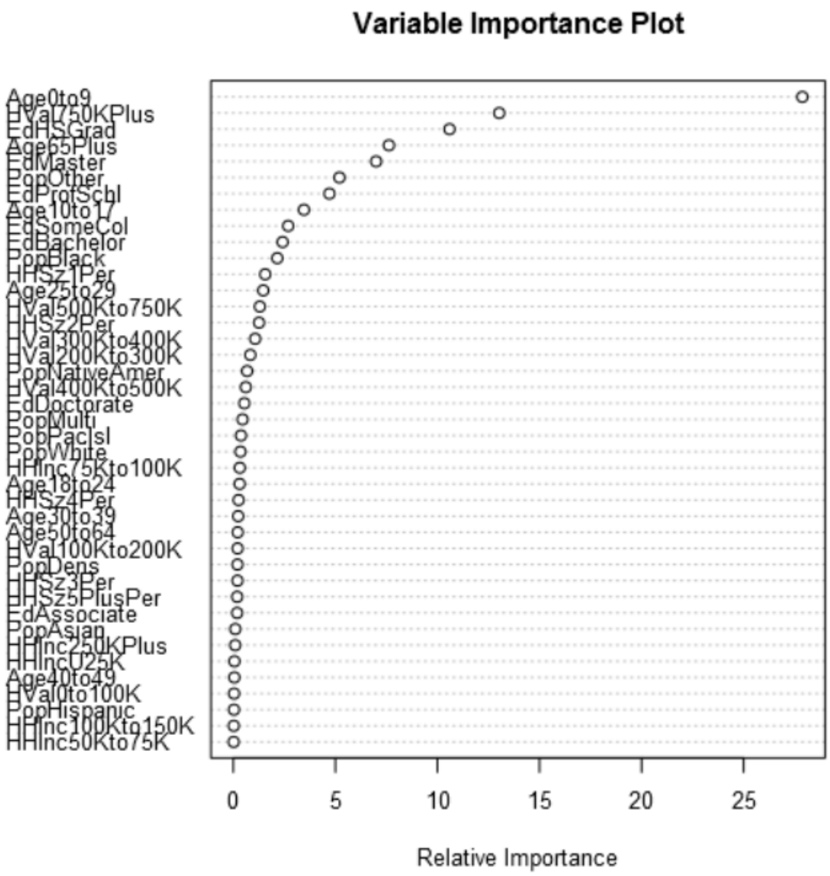
Fit and error measures

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

Based on the above results, I have decided to use the Boosted, even though the forest model and boosted model both have the same accuracy score, I have used the higher F1 score of the Boosted model as my deciding factor.

What are the three most important variables that help explain the relationship between demographic indicators and store formats?

Below is the variable importance plot for the Boosted model chosen for the final prediction.



Based on the above plot, the 3 most important variables for the Boosted model are:

Variable
Age0to9
HVal750KPlus
EdHSGrad

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

The model gave the new store predictions below:

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

Task 3: Predicting Product Sales.

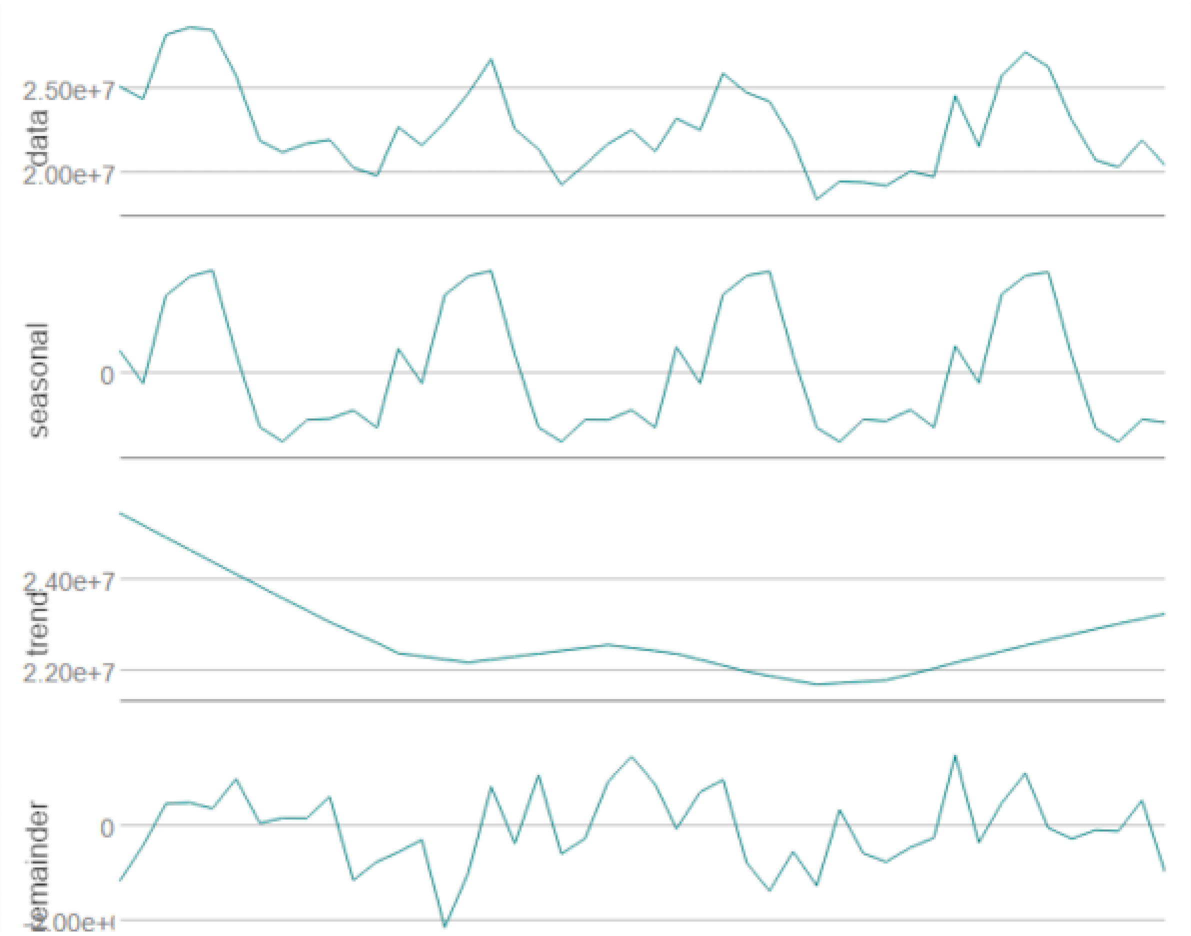
The Business Scenario: Forecasting

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

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

Both ETS and ARIMA models were run for comparison. Analysis of the initial time series decomposition plots below allowed further analysis of model parameters to be established.

The data used here is sales for produce only per month for all stores aggregated.



From the above decomposition plots, I can see that the Error element is increasing, Trend element is non-existent and the Seasonal element is also increasing, therefore an ETS(M,N,M) will be used. As for the ARIMA model, I have set the model to calculate the elements automatically.

For comparison, a holdout period of 12 periods was used to validate the ETS and ARIMA model.

Below is the ETS(M,N,M) in-sample summary.

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-241658.3191269	886787.7565481	699047.4732303	-1.1576764	3.1317204	0.3724833	0.069077

AIC	AICc	BIC
1078.9536	1101.0588	1100.3226

The ARIMA(1,0,0)(0,1,0)12 model in-sample summary.

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 due to the model having lower error values compared to the ARIMA model.

Provide a tableau dashboard that includes a table and plot of the 3 monthly forecasts; one for existing, one for new and one for all stores.

Below is a table of sales forecasts for existing stores, new stores and both new and existing stores combined.

Month	Year	Existing Store Sales Forecast	New Store Sales Forecast	Combined Store Sales Forecast
1	2016	21,381,830.22	2,600,354.85	23,982,185.07
2	2016	21,081,311.62	2,505,198.46	23,586,510.07
3	2016	24,502,171.96	2,889,940.32	27,392,112.28
4	2016	22,352,993.13	2,743,927.30	25,096,920.43
5	2016	25,331,350.65	3,110,813.81	28,442,164.46
6	2016	26,330,255.79	3,191,154.55	29,521,410.34
7	2016	25,715,514.09	3,219,369.78	28,934,883.87
8	2016	23,458,933.07	2,852,751.79	26,311,684.87
9	2016	21,801,458.48	2,543,602.66	24,345,061.14
10	2016	21,509,922.65	2,477,331.44	23,987,254.09
11	2016	22,619,212.99	2,569,169.56	25,188,382.55
12	2016	21,582,321.09	2,535,481.94	24,117,803.02

Below is a tableau plot of the sales forecasts for existing, new and the combined store sales for produce only.

Historical Produce Sales plus 2016 forecasts.