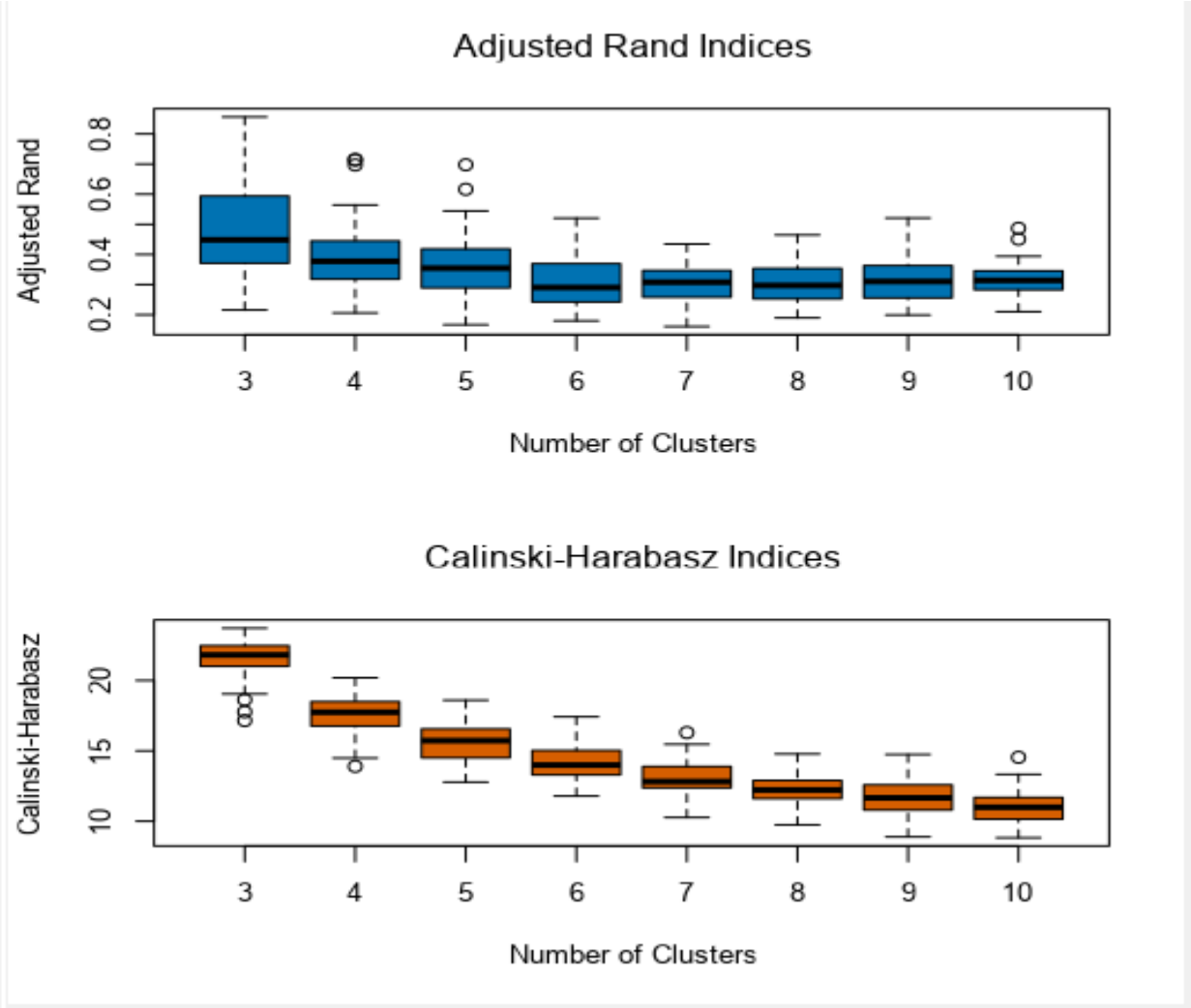
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

The optimal number of store formats is three as shown by the K-Centroids Diagnostic tool from Alteryx that shows us how the Adjusted Rand index and the Calinski-Harabasz index are both higher at three clusters .



In the first segment there are 23 stores, in the second segment there are 29 stores and in the third segment there are 33 stores.
We can see that the clusters differ by average distance and separation: the cluster number 1 has an average distance of 2,23 and a separation of 1,87, the cluster number 2 has an average distance of 2,54 and a separation of 2,12, the cluster number 3 has an average distance of 2,11 and a separation of 1,7.

Summary Report of the K-Means Clustering Solution X

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:

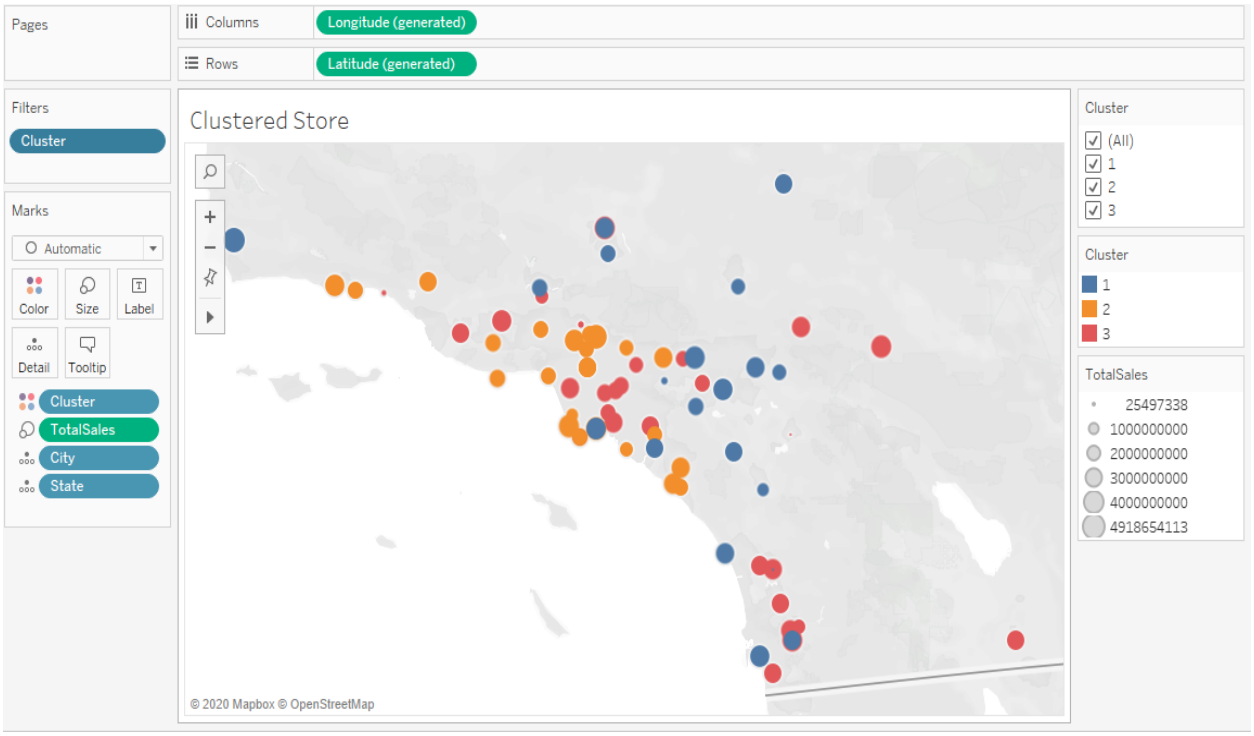
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

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

Clusters on the Map



Task 2: Formats for New Stores

I have used the Boosted model because the overall accuracy between the Bosted Model and the Forest Model was the same but it has an accuracy of 1 fore the first two segments and of 0.67 for the third segment.

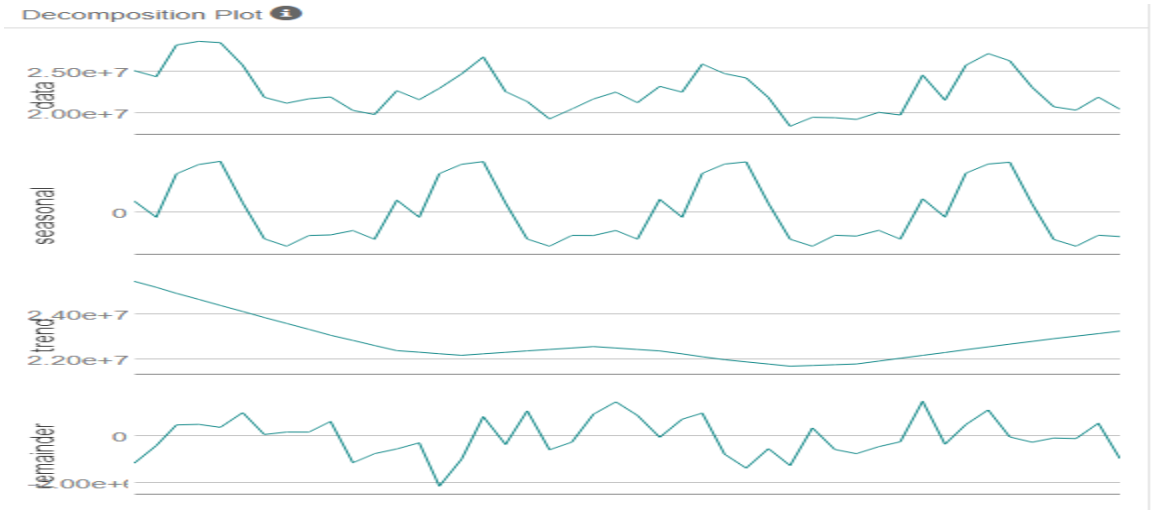
Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Model	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778

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

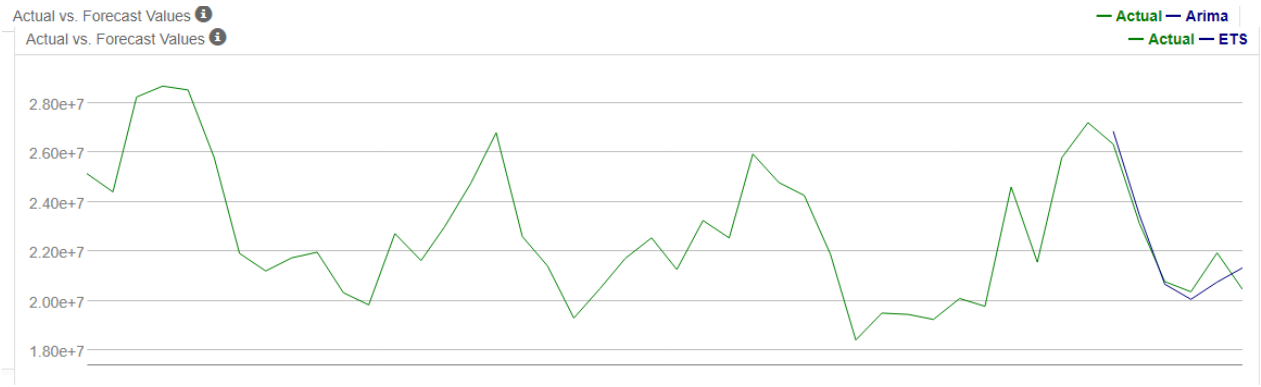
Task 3: Predicting Produce Sales

I have selected an ETS(m,n,m) after analyzing the decomposition plot that shows how the remainder is multiplicative, the trend is non existing and the seasonality is multiplicative.



This is the comparison between the ARIMA and the ETS model
ARIMA

ETS



As you can see ETS provides a better forecast.
These are the forecast error measurements against the holdout sample for the ETS model

Report

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS
26338477.15	26860639.57444
23130626.6	23468254.49595
20774415.93	20668464.64495
20359980.58	20054544.07631
21936906.81	20752503.51996
20462899.3	21328386.80965

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

and these are the forecast error measurements against the holdout sample for the ARIMA model

Report

Comparison of Time Series Models

Actual and Forecast Values:

Actual	Arima
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

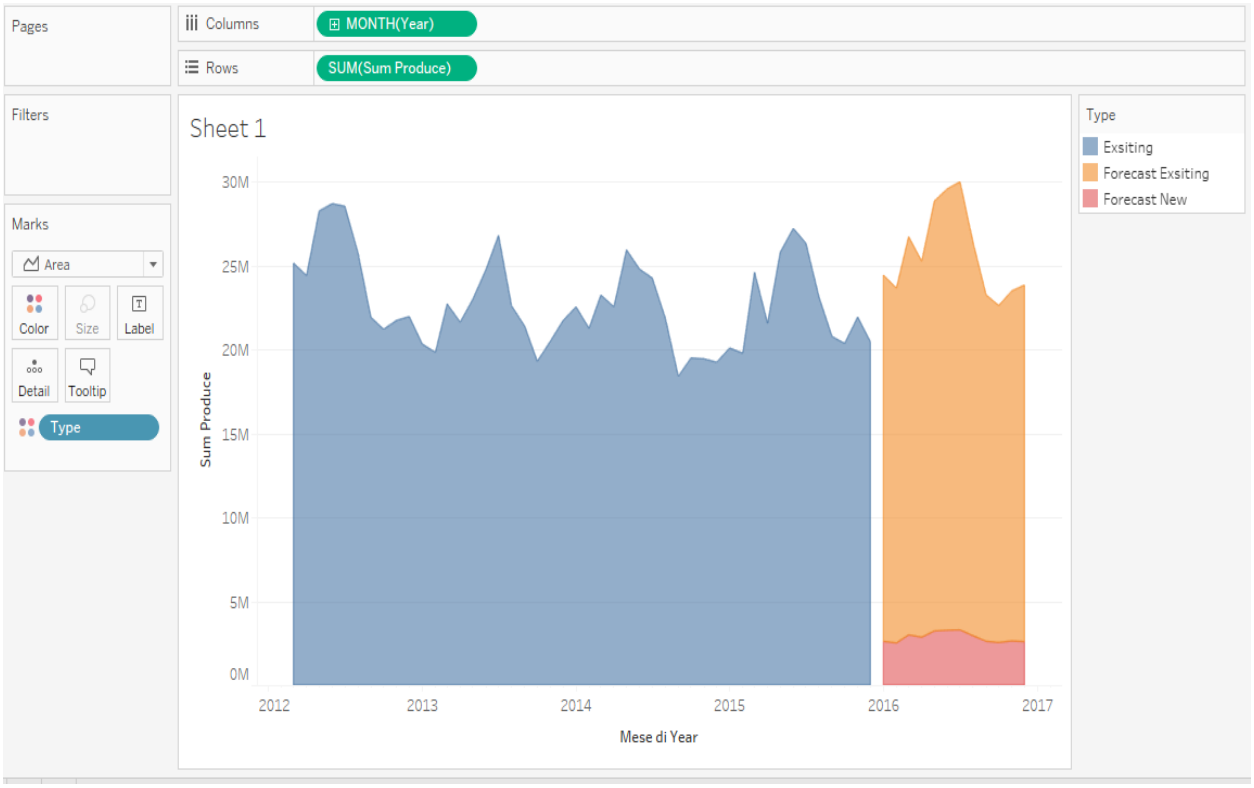
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
Arima	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

the ETS model's error rates are lower and the forecasts are more precise.

Year	Month	Forecast_Sales	New_Stores_Sales
2016	1	21829060	2603262
2016	2	21146330	2508878
2016	3	23735687	2989458
2016	4	22409515	2849287
2016	5	25621829	3224711
2016	6	26307858	3269623
2016	7	26705093	3288334
2016	8	23440761	2937302
2016	9	20640047	2606592
2016	10	20086270	2536270
2016	11	20858120	2631293
2016	12	21255190	2586562

Forecast's visualization


Pages
Columns
MONTH(Year)
Rows
SUM(Sum Produce)
Filters
Marks
Area
Color
Size
Label
Detail
Tooltip
Type
Sheet 1
30M
25M
20M
15M
10M
5M
0M
Sum Produce
2012
2013
2014
2015
2016
2017
Mese di Year
Type
Exsiting
Forecast Exsiting
Forecast New