Project: Predictive Analytics Capstone

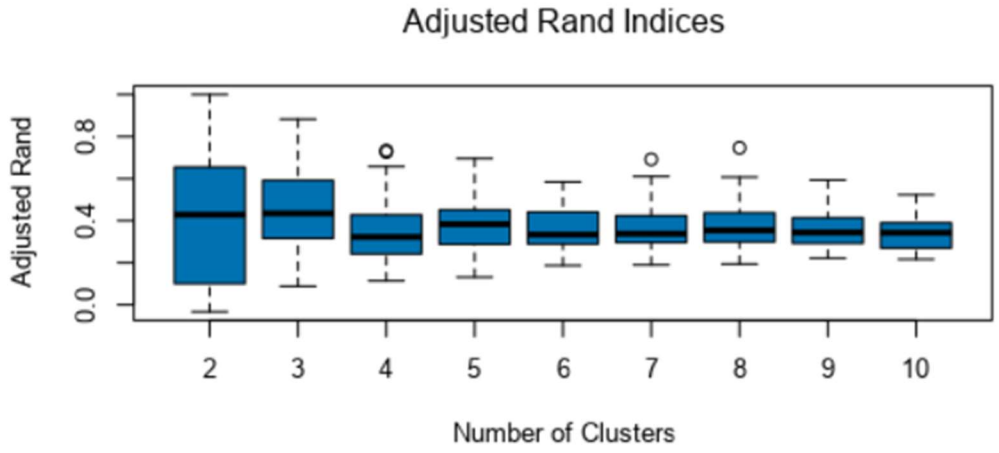
Task 1: Determine Store Formats for Existing Stores

We use the latest available sales data from 2015 to determine number of Store Formats based on a Clustering Model. We start by grouping by store to get an aggregate amount of total sales for each type grouping in a grocery store. The different types of grocery products will be used as the 'Principal Components' as the basis to cluster the Stores into an ideal number of Store Formats.

	Store	Dry_Grocery	Dairy	Frozen_Food	Meat	Produce	Floral	Deli	Bakery	General_Merchandise
1	S0001	10845787.65	2423389.38	1814872.88	2531382.34	2284388.7	159142.23	1023812.9	835599.06	1590570.68
2	S0002	7931072.94	1844188.83	1366677.52	1991807.66	1755293.15	128935.92	689786.42	514864.68	1111992.45
3	S0003	12741875.44	3095321.58	2087437.15	3468989.12	3792143.03	291259.69	1262874.27	1091930.44	2408831.27

For even more consistency, we convert the 'total sales numbers' grouping by the different types and calculate the percentage of total. With the data more standardized as a percentage, we feed it into the Alteryx tool *K-centroid Diagnostic* tool to be clustered. The tool itself offers many methods, but for our purposes we use the K-means clustering method and standardize the percentage fields further using z-score model.

The resulting report with the *Rand and CH indexes* indicates the optimal number of store formats based on the highest index median and narrowest IQR spread. This was determined to be 3 clusters.



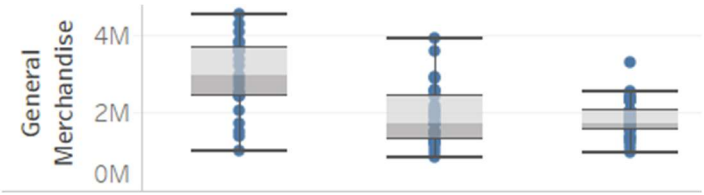
The diagnostic tool above specified the number of Clusters (representing the Store Formats) we will be using moving forward. Using *K-Centroids Cluster Analysis* tool, we now generate a report based on the same fields for K-means and z-score configurations that we used in the diagnostic tool. This results in the cluster size allocations with the most closely related stores grouped into three clusters.

Cluster	Size
1	23
2	29
3	33

Looking further into the Analysis report, we can determine some patterns that differentiate each cluster. Cluster 1, for example, is a clear differentiator when it comes to the sales of 'General Merchandise'. We can see this by looking at the results from the Analysis tool, where the value for Cluster 1 is highly positive, while the values for Cluster 2 and 3 are highly negative.

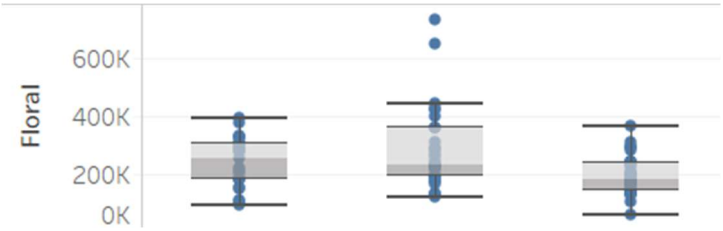
PercentageOfTotal_General_Merchandise
1.208516
-0.304862
-0.574389

This was confirmed when validating the cluster allotment with the actual sales numbers. This can be seen from the whisker plots below, where the distribution, with the bulk represented with the IQR, for 'General Merchandise' sales is significantly higher for stores in cluster one.



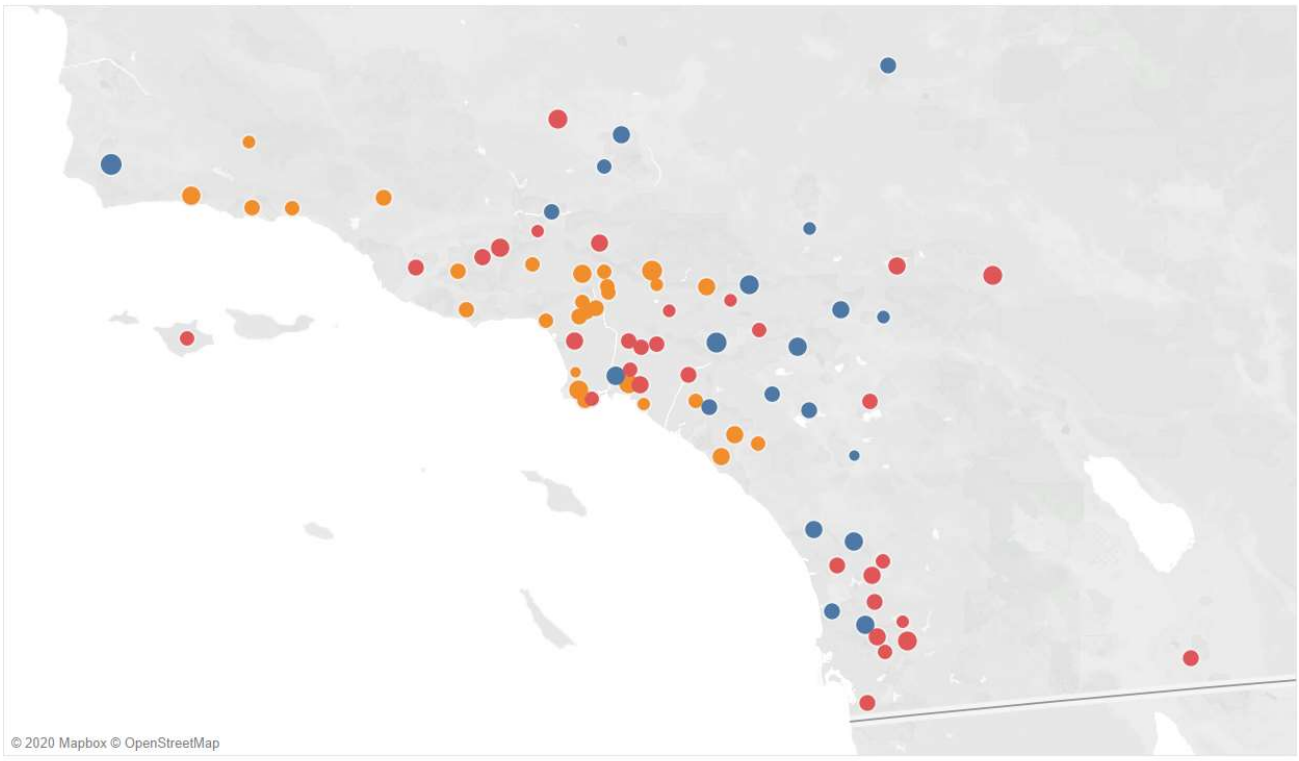
We can see other differentiator when looking at 'Floral' sales. Here the differentiator is confirmed for cluster 2 with a high positive with low negatives for the other two clusters. This can also be validated by plotting the Floral sales data as Whisker plots highlighting the higher median and IQR for cluster two.

PercentageOfTotal_Floral
-0.301524
0.851718
-0.538327



With the Store Formats identified, it is best to continue visualizing the results in Tableau. All the stores are marked on an Map overlay of California State. The level of Sales is designated by a Size Marker on the points, and the type of Cluster is marked by Color. (On next page)

Store distribution in California



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

Total Sales
12,618,744
20,000,000
30,000,000
40,000,000
49,186,541

Cluster
1
2
3

Task 2: Formats for New Stores

For the next task, we used Store Format (or the cluster identifier), determined from the previous task, as the Target variable. Three models were compared and validated using the *Model Comparison* tool to determine the best predictive model to determine the 'Store Format' designation for the new stores. The demographic data was used as predictor variables while the Store Format results required using a Non-binary classification model

In the choice to pick the best model, the number for accuracies in the predictions for Forest Model and Boosted Model are same and significantly higher than that of the Decision Tree model. We ultimately used the *Boosted* Model based on the higher F1 or precision measure which indicates the least bias amongst the False positive and False negative ratios for each respective Store Format.

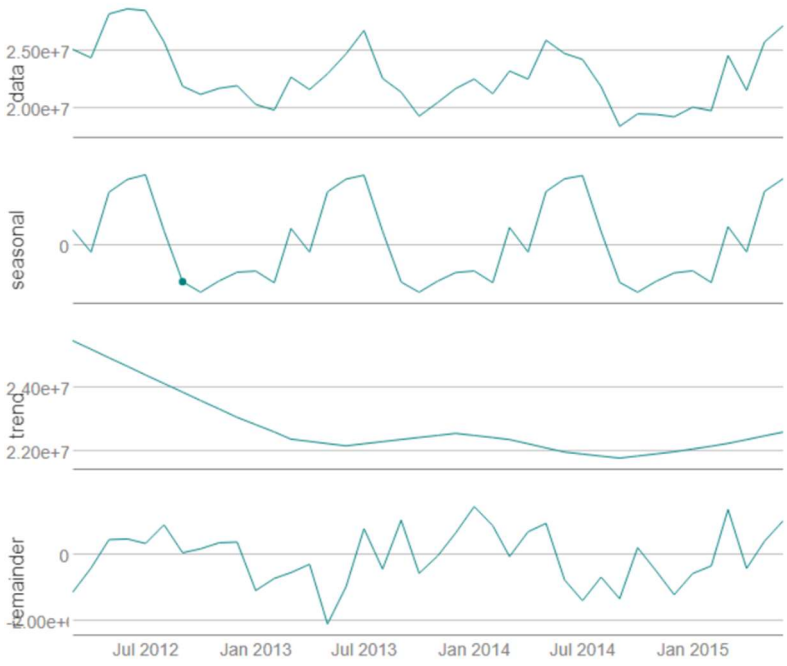
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_ClusterPredict	0.7059	0.7685	0.7500	1.0000	0.5556
FM_ClusterPredict	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_ClusterPredict	0.8235	0.8889	1.0000	1.0000	0.6667

With the model ready, the new store segments are Scored with the Model to determine their Store Format allocation.

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

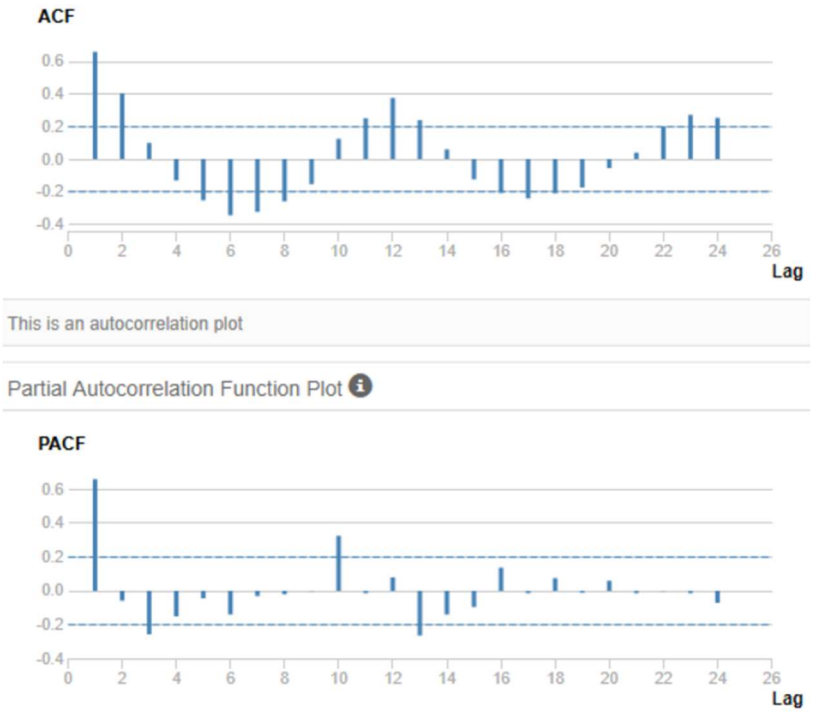
We continue to evaluate the sales data to attempt to forecast futures ‘Produce’ sales. With data available from March 2012 to the end of December 2015, we start by singularizing the Produce Sales Data, to fit with ‘Time Series Plotting’ for forecasting. We do this by aggregating by the month and year to get the total sales for that month. We can then feed the data into the *TS PLOT* Alteryx tool



The resulting report, decomposes the main plot, making it possible to determine patterns in the 'Time Series' plot. In order to identify correctly model to use with forecasting, the first step is to identify these key patterns. These can be identified in the plot in the previous page:

- There is a slight decrease seasonally, which indicates 'Multiplicative'.
- No upward or downward trend is observed, which we indicate as 'None'.
- The Remainder or error shows changing variance, which we indicate as 'Multiplicative'.

We can note that the above setup can be used with our set up for the ETS model. It is denoted as ETS (MNM). In the same report from the TS Plot tool, we continue to look at the Auto Correlation and Partial Autocorrelation plots that help us denote configurations for the other time series modelling tool called *Arima.*



There are many steps we can take to denote the different settings for a Arima model in relation to autocorrelation. Seasonality for example, can be indicated by the spikes on the lag 1 and 12. Seasonality indicates that the Arima model of the type (p,d,q)(P,D,Q). Running the Arima model on Alteryx results in Arima(1,1,0)(1,1,0)[12]

Comparisons of the models results in accuracy measures favors ETS over Arima all across the board, especially with the Root Mean Square Error (RMSE) and Mean Absolute Scale Error (MASE). RMSE which represents standard deviation between predicted and observed value, is lower for the ETS model. MASE, which is recent introduction and a reliable error measure because it is scale free, also sides favorably with ETS. You should note that both models are still valid, with MASE well below the 1 threshold.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
Arima	-604232.33	1050239.2	928412	-2.6156	4.0942	0.5463

With the ETS model chosen, we continue with two forecasts for the existing stores and the new stores using the same ETS model and the MNM configuration. The existing store forecasts was done using the same Produce total sum projections as the same data was available for those stores. The new store sales were projected based on taking an average of Sales for a Store Format based on the existing data and multiplying that by the number of clusters to determine the total sales.

Date	Forecast Existing	Forecast New
Jan-16	21829060.03	2588356.56
Feb-16	21146329.63	2498567.17
Mar-16	23735686.94	2919067.02
Apr-16	22409515.28	2797280.08
May-16	25621828.73	3163764.86
Jun-16	26307858.04	3202813.29
Jul-16	26705092.56	3228212.24
Aug-16	23440761.33	2868914.81
Sep-16	20640047.32	2538372.27
Oct-16	20086270.46	2485732.28
Nov-16	20858119.96	2583447.59
Dec-16	21255190.24	2562181.70

A complete visualization of the existing and forecasting data can be see in the next page. The Tableau visualization includes the existing data from March 2012 to the forecasted period in 2016. The forecasted period further distinguishes the ‘new stores’ from the ‘existing stores’