Mohammed Alrashidan

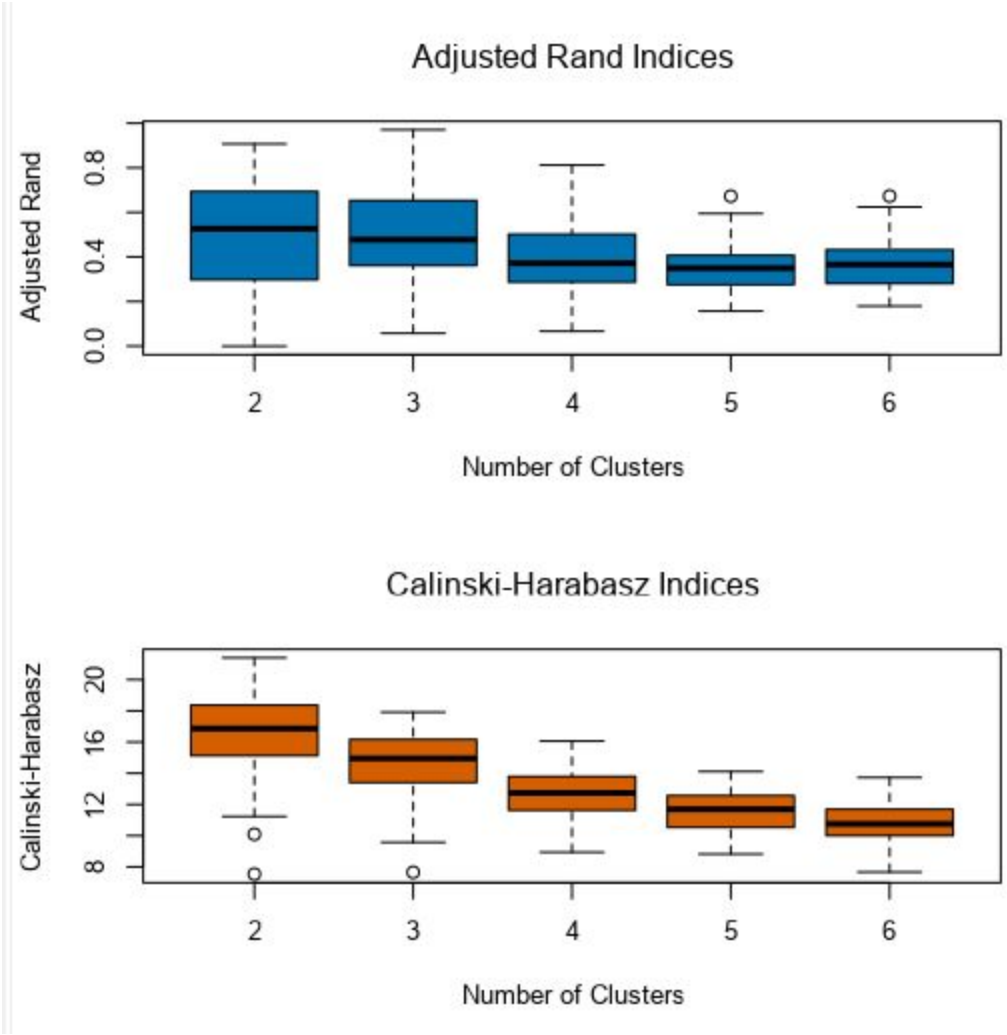
Predictive Analytics - Capstone

Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   The optimal number is 3 for the store formats. I have chosen to perform K-medians with min 2 of the number of clusters and max of 6 of the cluster. as this shows very high medians with smaller spread and compactness. Both Adjusted Rand and Calinski-Harabasz show the clusters below.



2. How many stores fall into each store format?

After calculating the sum of all sales and filtered to only year “2015”. I linked it with K-centroids Analysis to see the size (number of stores) of the 3 clusters After choosing 3 clusters. We can see from the report below:

- Cluster 1: 23 stores.
- Cluster 2: 29 stores
- Cluster 3: 33 stores

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

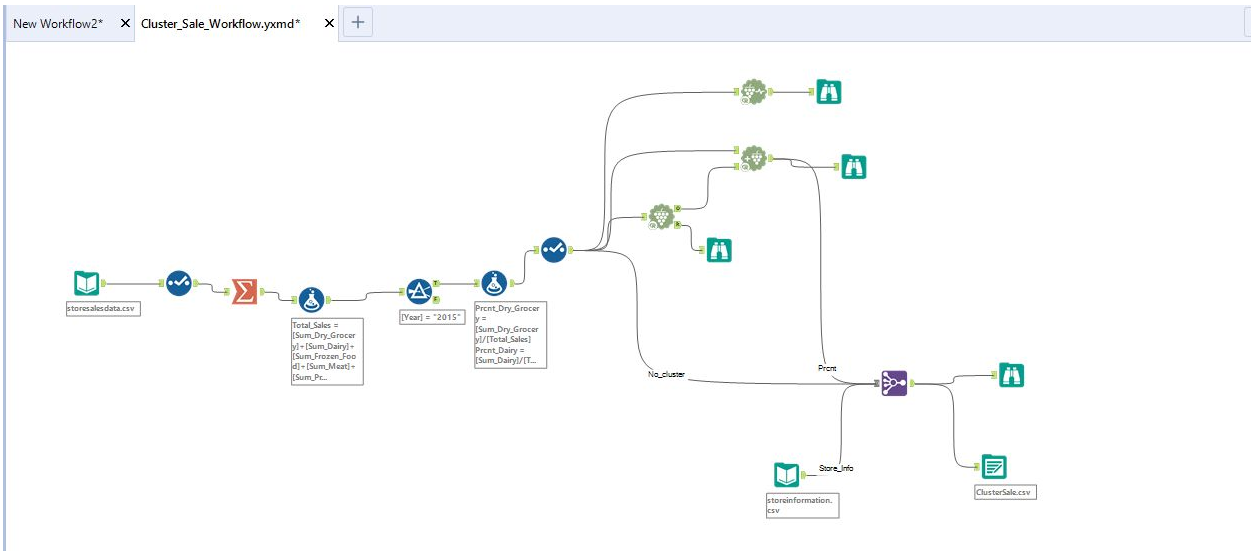
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Cluster 1 and 2 in Dairy show a significant opposite in distance where 1 is -0.76 and 2 is 0.7. This shows that each cluster varies cluster model within distance. that means Dairy in cluster 2 sells more than Dairy in both cluster 1 and 3.

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

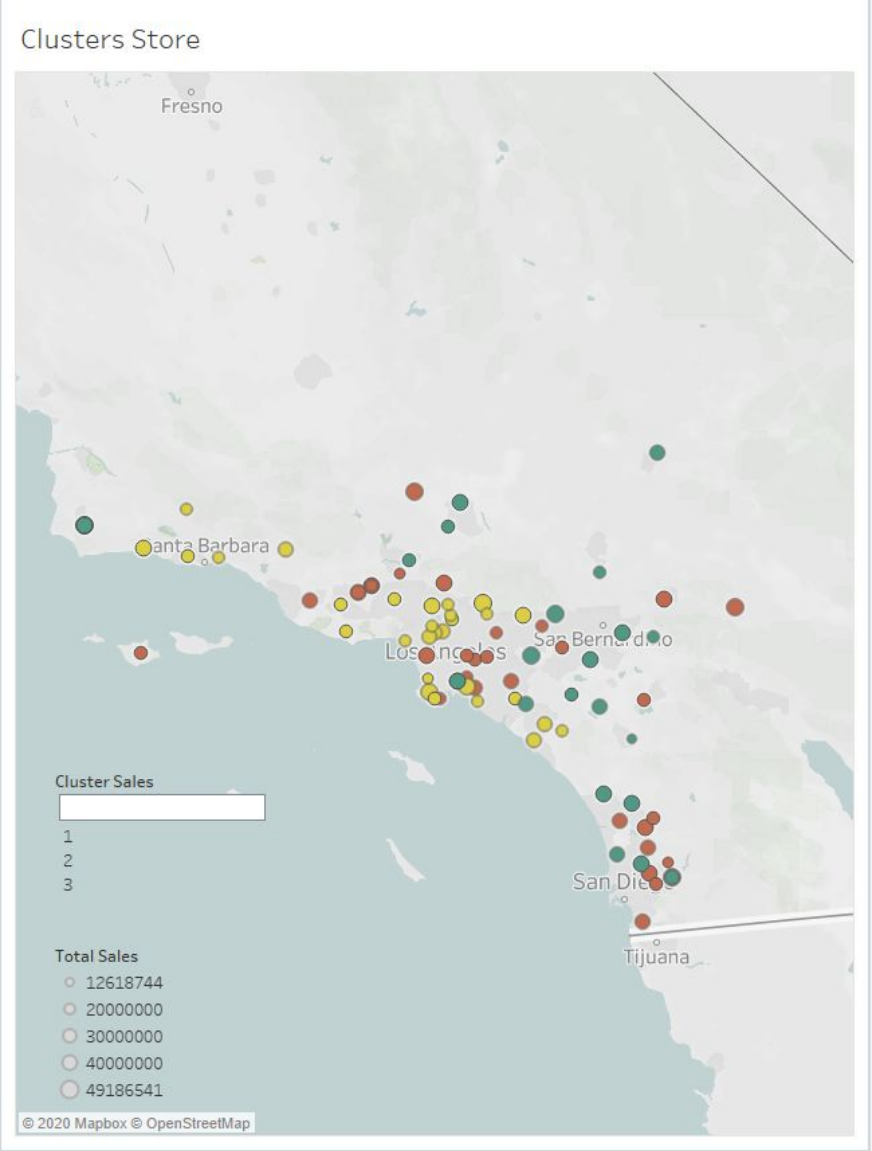
	Prcnt_Dry_Grocery	Prcnt_Dairy	Prcnt_Frozen_Food	Prcnt_Meat	Prcnt_Produce	Prcnt_Floral	Prcnt_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Prcnt_Bakery	Prcnt_General_Marchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389



4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show clusters, and size to show total sales.

The Tableau visualization shows the store location by Zip Code, and provides the 3 clusters distributed based on total sales.



Task 2: Formats for New Stores

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   I have tested all three models (Decision Tree, Forest Model, Boosted Model) to make the right decision for which model I will use to predict the best new store.
   The report below shows that Boosted has the highest F1 with 0.88 and accuracy of 0.82. Forest Model seem to have same accuracy as Boosted Model, however, BM outperform in the F1.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7685	0.7500	1.0000	0.5556
FM	0.8235	0.8426	0.7500	1.0000	0.7778
BM	0.8235	0.8889	1.0000	1.0000	0.6667

Confusion matrix of BM

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

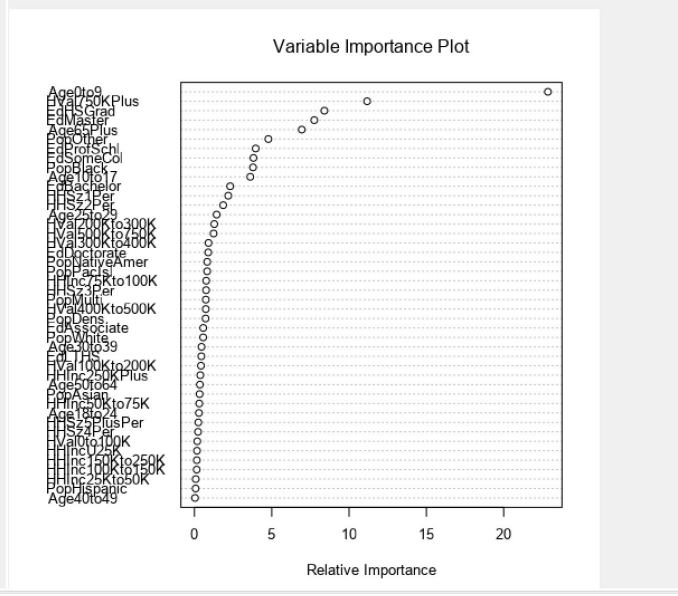
Confusion matrix of DT

	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

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

The Boosted Model shows a very positive pattern in variable importance plots.



2. What format do each of the 10 new stores fall into? Please fill in the table below.

The segment Number data was done by adding the Score, then creating the formula to predict each cluster for each store available in the dataset created.

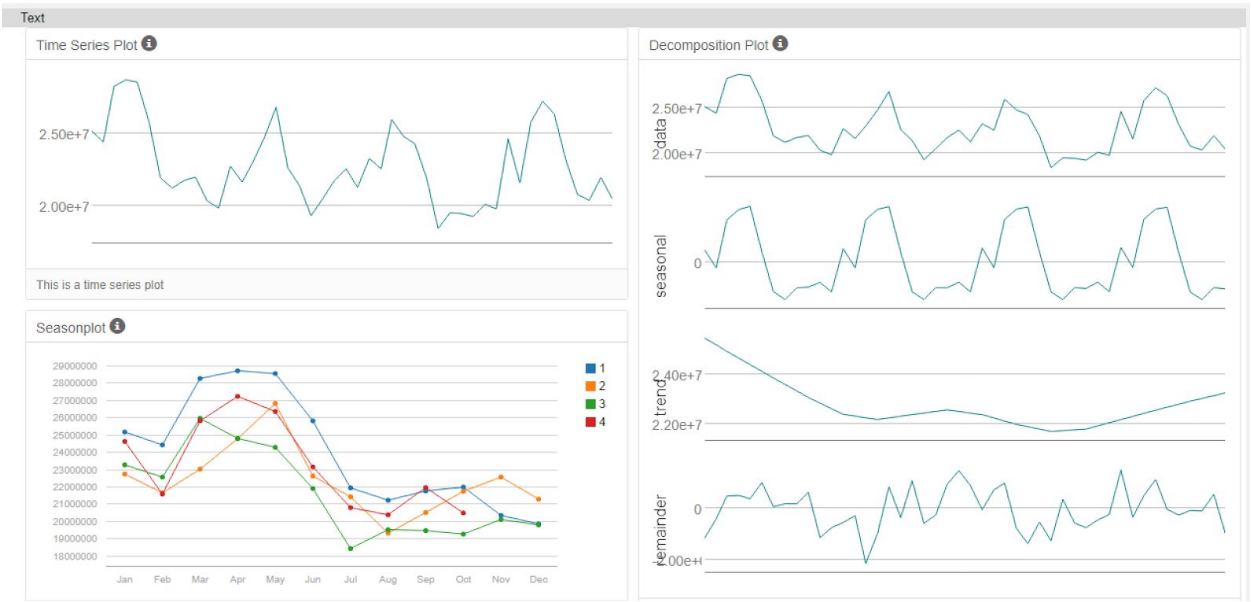
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

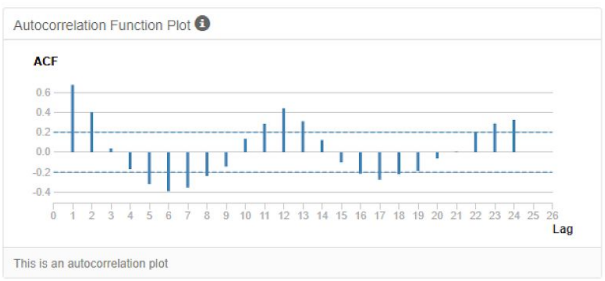
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

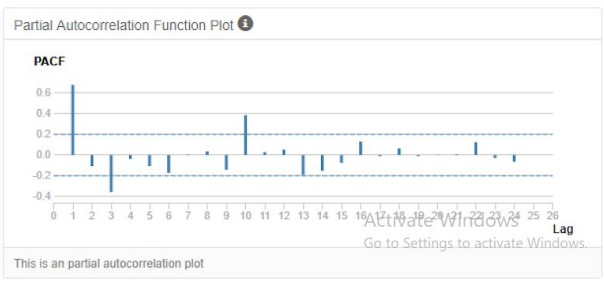
In order to begin predicting the Produce Sales I have done a couple of things. First, I added the file, storessalesdata.csv and then added a summarize tool to group the Year, Month and Sum of Produce. Then I have tested TS plot, ETS and ARIMA. The images show my results in the workflow.

In the graph below, the Seasonal show no trend and should be applied as multiplicatively.



I have used ARIMA(012)(010) to perform both seasonal difference and seasonal first difference by taking the difference from sum_produce. Autocorrelation Function Plot shows lag-2 from the graph below.



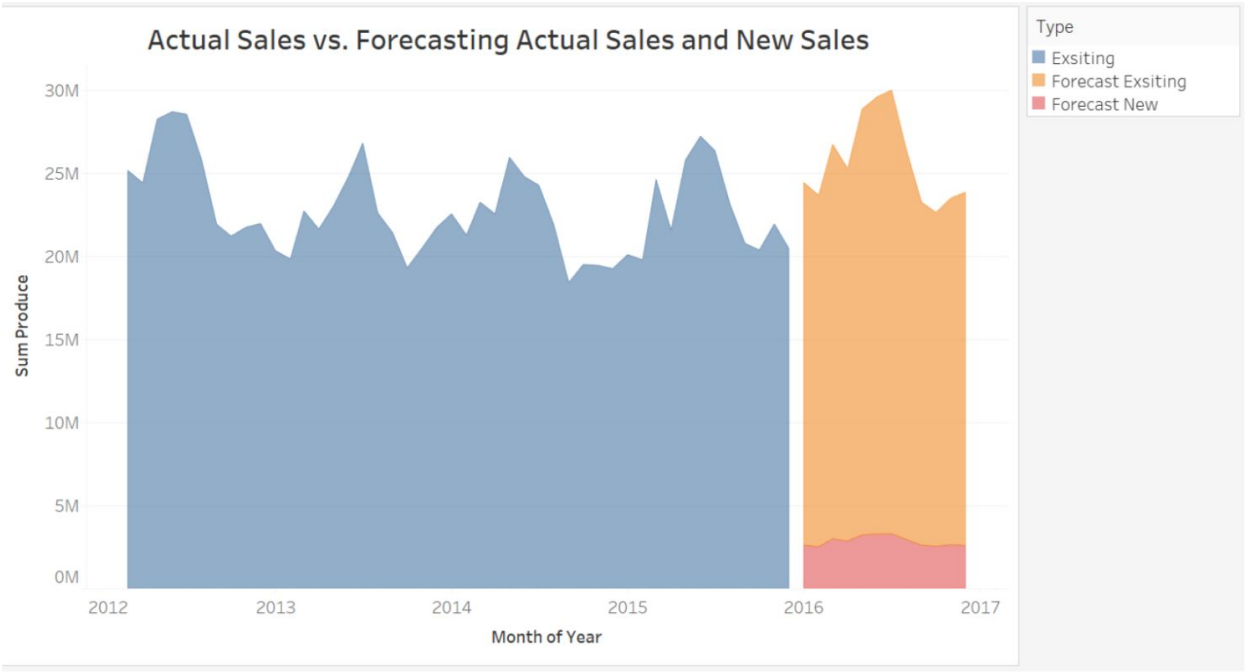


2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

This is after I joined my workflow to get the Existing and New Store for 2016 for 12 months

Year	Month	New Store	Existing Store
2016	1	2588356.558	21829060.03
2016	2	2498567.174	21146329.63
2016	3	2919067.025	23735686.94
2016	4	2797280.083	22409515.28
2016	5	3163764.859	25621828.73
2016	6	3202813.289	26307858.04
2016	7	3228212.242	26705092.56
2016	8	2868914.812	23440761.33
2016	9	2538372.267	20640047.32
2016	10	2485732.285	20086270.46
2016	11	2583447.594	20858119.96
2016	12	2562181.7	21255190.24

The Tableau shows the historical(Actual) Sales vs Forecasting Sales and new Sales. This represents all the workflows from preparing data to clustering to perform time series Model in order to finalize the output and predict the New Sales for all the stores.


Actual Sales vs. Forecasting Actual Sales and New Sales
Sum Produce
30M
25M
20M
15M
10M
5M
0M
2012
2013
2014
2015
2016
2017
Month of Year
Type
Exsiting
Forecast Exsiting
Forecast New