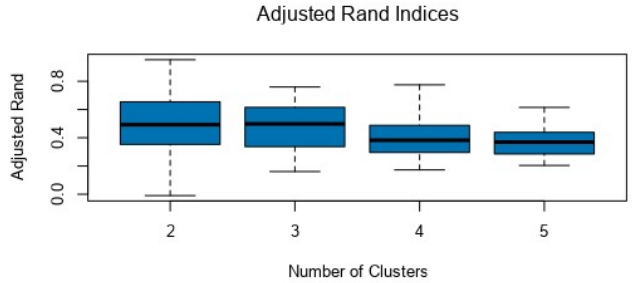
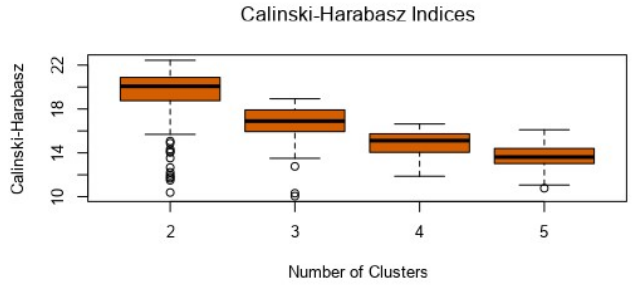
Project: Predictive Analytics Capstone

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 optional number is 3, based on the on highest median that cluster has including small variation. I have obtained it by deriving cluster analysis using Calinski-Harabasz criterion and Adjusted Rand:*





2. **How many stores fall into each store format?**

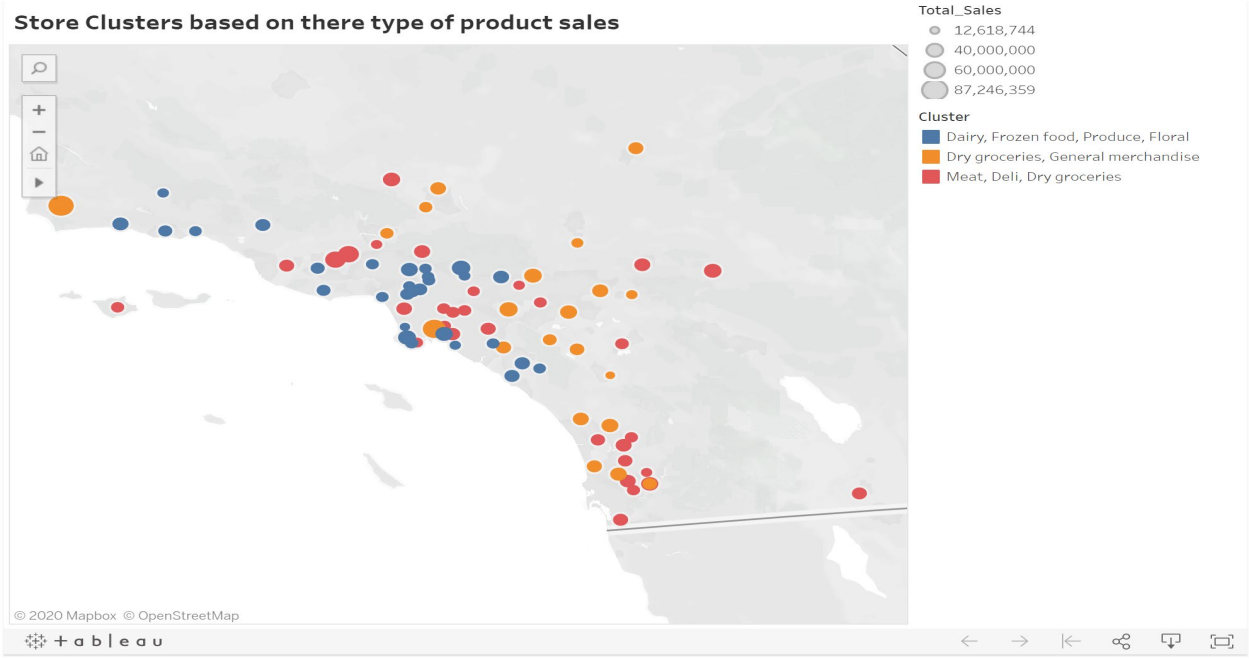
   - *The results are next – Claster 1 has 23 stores, Cluster 2 has 29 stores, and Cluster 3 has 33 stores*

3. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**
   - *At first it differs in average distance, and the cluster 3 has the best one. There is also maximum distance and separation. Using these three parameters we can chose the best one. Also we can use the cluster report to determine which cluster belongs to what segment, and determine their positive and negative values.*

	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

4. **Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show 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've used methodology map so I could be able to predict the best store format for the new store. I have chosen it because I wanted to predict the outcome and I had a plenty of data so I used Decision tree model, Forest model and Boosted model. I have found that Forest model and Boosted model have identical accuracies, but the highest F1 score is for the Boosted model 0.8543. Also the confusion matrix of Boosted model shows highest true positive values, and according to all those parameters I can be sure that Boosted model is the best one to choose.*

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7327	0.6000	0.6667	0.8333
FM	0.8235	0.8251	0.7500	0.8000	0.8750
BM	0.8235	0.8543	0.8000	0.6667	1.0000

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

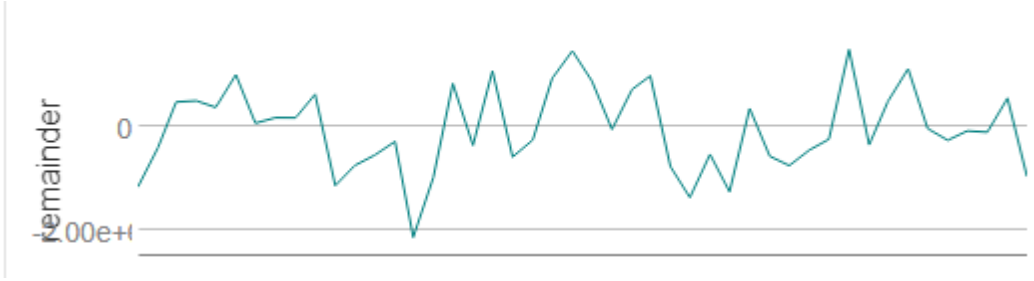
2. **What format do each of the 10 new stores fall into? Please fill in the table below.**

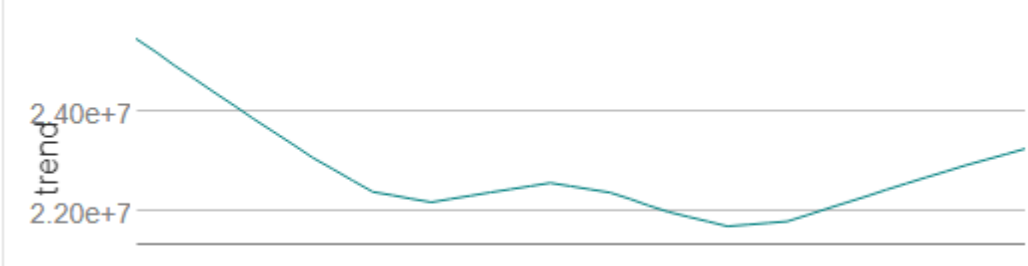
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

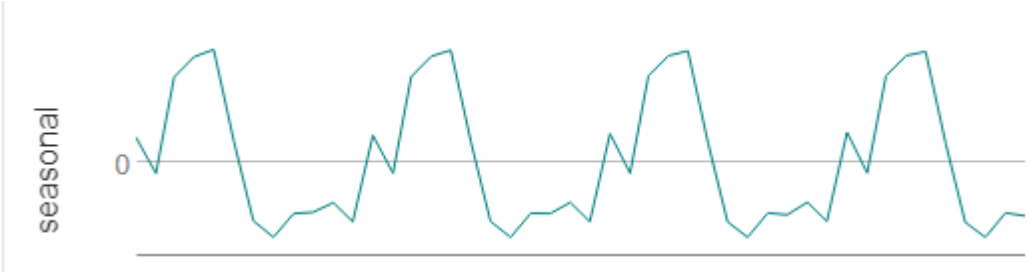
Task 3: Predicting Produce Sales

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?

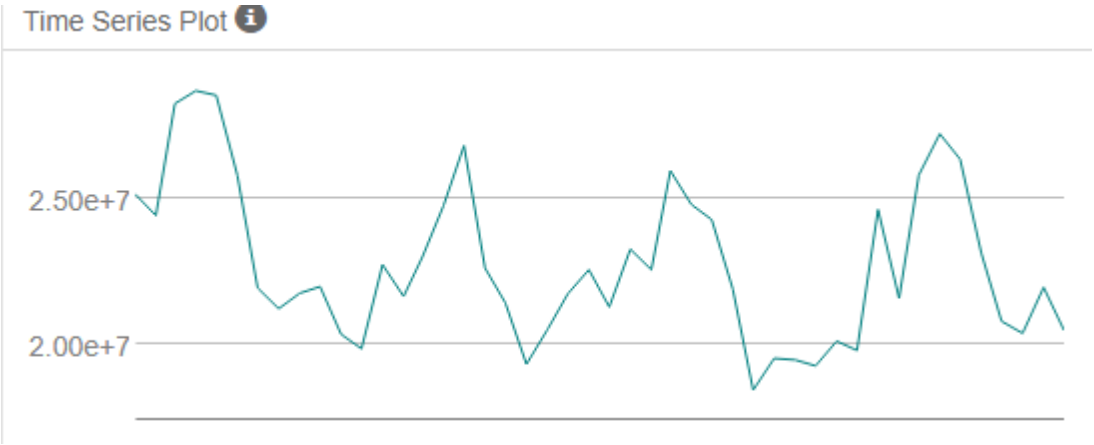
- *Using the decomposition plot of the TS plot tool, we can see that the Error is Multiplicative, the Trend is None, and the Seasonality is also Multiplicative - ETS (M,N,M).*



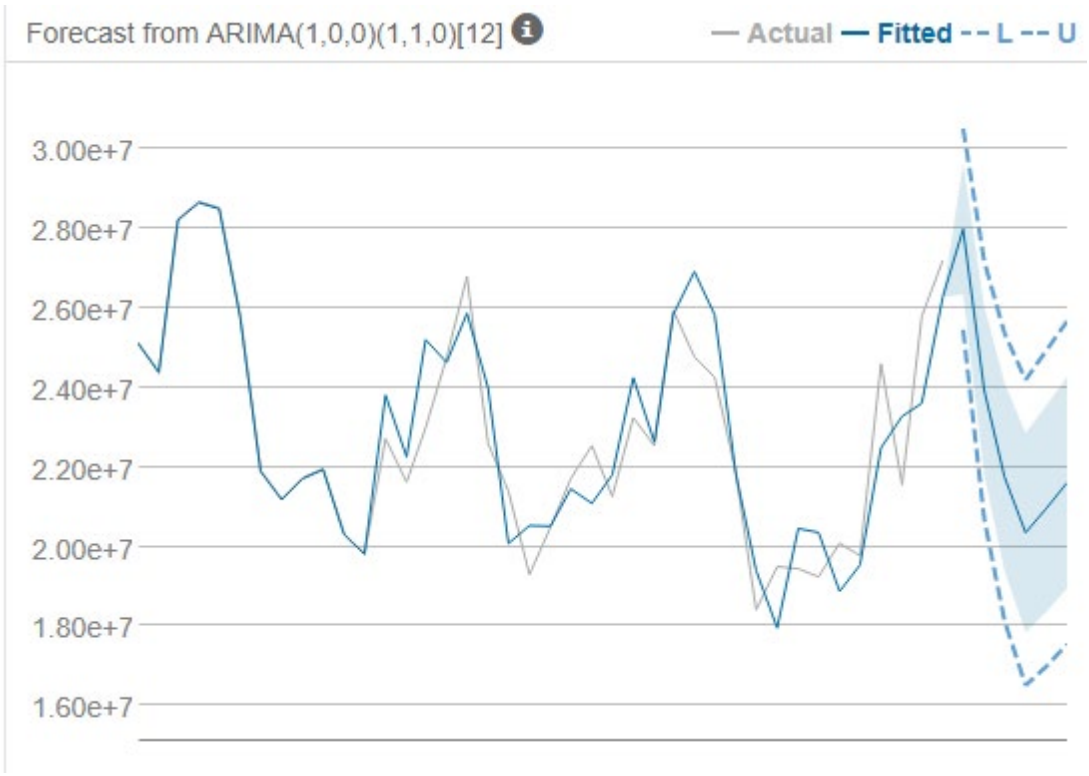




- *Contrary to ETS, applying* ARIMA (1,1,1) (0,1,0) [12]) model, and using Time series plot tool, we can see that it is not stationary, which means that it has too many variations.



- *After transforming and stabilizing the data {*ARIMA (1,0,0) (1,1,0) [12]*}, we can see that time series decreasing unto zero, which means the data was stabilized.*



Also we can see that ACF and PACF shows us the negative correlation.

- *The accuracy measures shows lower RMSE, MAPE and MASE for ETS contrary to ARIMA model that has worse accuracy measures.*

Actual and Forecast Values:

Actual	ETS
26338477.15	26918022.38381
23130626.6	23792569.05787
20774415.93	21028514.63042
20359980.58	20509999.41019
21936906.81	21121956.48609
20462899.3	21580998.03469

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	-324792.3	680122.7	596442.4	-1.4619	2.7002	0.351	NA

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	NA
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463	NA

- *Finally, we can conclude that ETS is the best model to use to predict the produce sales for the new and existing stores.*

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.

Month	New Stores	Existing Stores
Jan-2016	$2,587,451.00	$21,539,936.00
Feb-2016	$2,477,353.00	$20,413,771.00
Mar-2016	$2,913,185.00	$24,325,953.00
Apr-2016	$2,775,746.00	$22,993,466.00
May-2016	$3,150,867.00	$26,691,951.00
Jun-2016	$3,188,922.00	$26,989,964.00
Jul-2016	$3,214,746.00	$26,948,631.00
Aug-2016	$2,866,349.00	$24,091,579.00
Sep-2016	$2,538,727.00	$20,523,492.00
Oct-2016	$2,488,148.00	$20,011,749.00
Nov-2016	$2,595,270.00	$21,177,435.00
Dec-2016	$2,573,397.00	$20,855,799.00