Parag Sasturkar

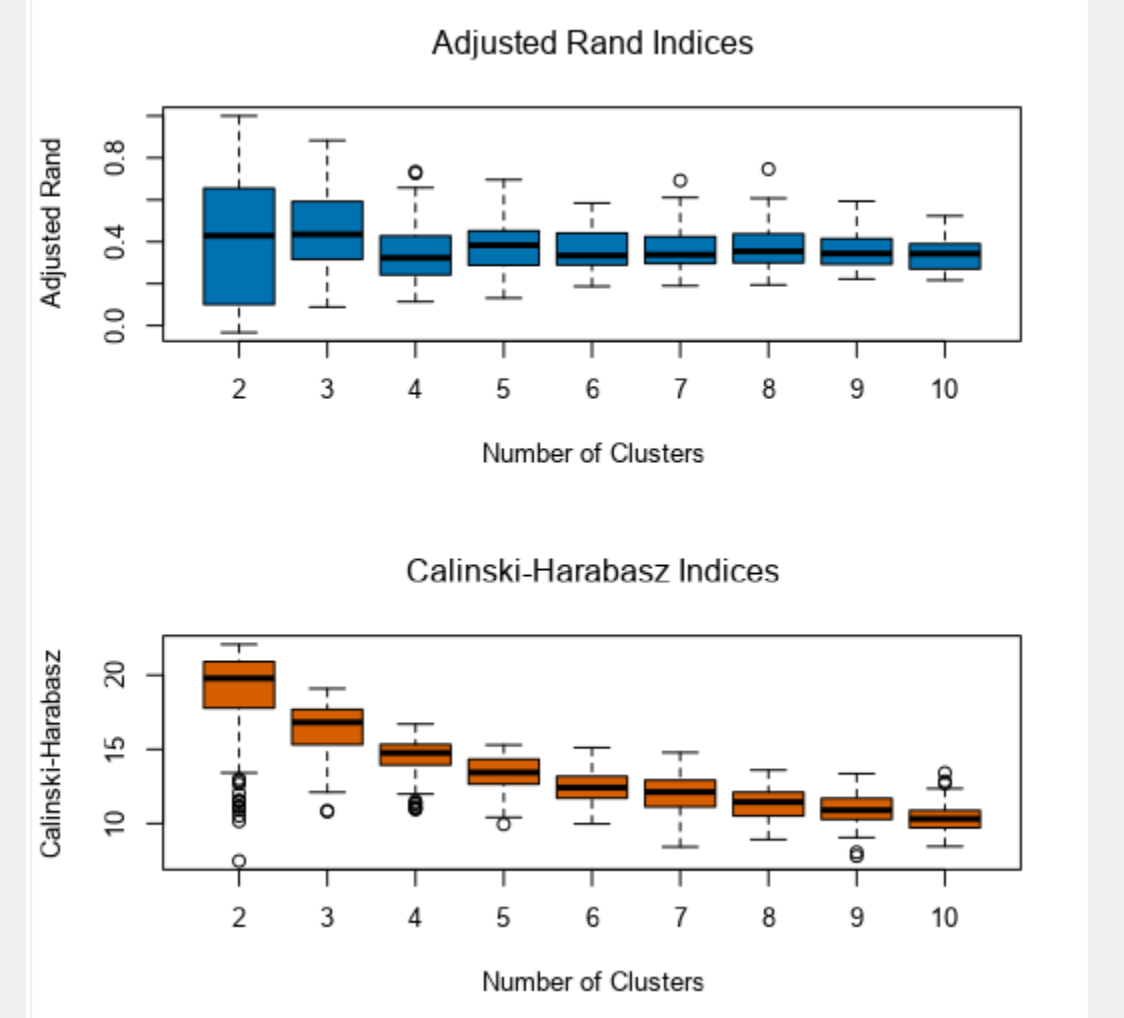
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   Optimal number of store formats is 2.

   By looking at below Adjusted Rand Indices and CH indices, I decided to choose 2 clusters because box plots for 2 clusters have the highest mean.



2. How many stores fall into each store format?
   Answer:

   Cluster 1 has 41 stores, whether Cluster 2 has 44 stores.

Cluster Information:

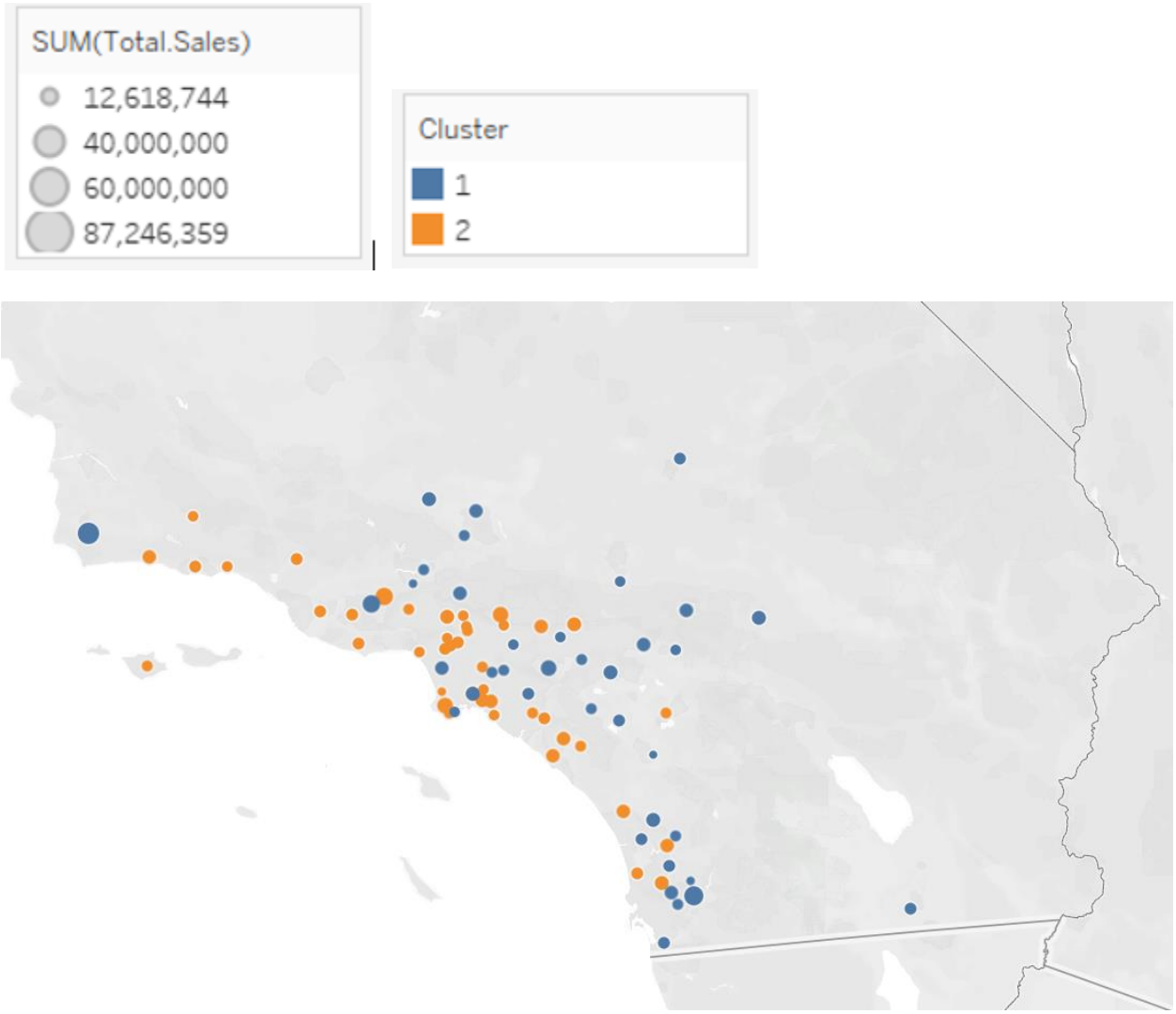
Cluster	Size	Ave Distance	Max Distance	Separation
1	41	2.449907	4.98911	2.170835
2	44	2.565327	4.530116	1.982013

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

   Based on the below image, I can say that the stores in cluster 1 are exactly opposite to the stores in cluster 2. For example, stores in cluster 1 have high sales for dry grocery and stores in cluster 2 have low sales for dry grocery and vice versa.

	X..dry.grocery	X..sum.dairy	X..frozen.food	X..meat	X..produce	X..floral	X..deli
1	0.577839	-0.488323	-0.386461	0.462019	-0.66908	-0.667953	0.404109
2	-0.538441	0.455028	0.360111	-0.430517	0.623461	0.622411	-0.376556
	X..bakery	X..general.merchandise					
1	-0.20217	0.010938					
2	0.188386	-0.010193					

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.
   Answer:

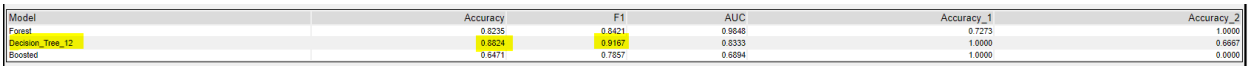


Parag Sasturkar

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.)
   Answer:

   I built three models – Decision Tree, Boosted Model, and Random Forest using 80% of the data and then compared the model results using the model comparison tool. Below are the details –



Model	Accuracy	F1	AUC	Accuracy_1	Accuracy_2
Forest	0.8235	0.8421	0.9848	0.7273	1.0000
Decision_Tree_12	0.8824	0.9167	0.8333	1.0000	0.6667
Boosted	0.6471	0.7857	0.6894	1.0000	0.0000

   Based on the model accuracy and F1 score of the models, I chose Decision Tree (Accuracy = 0.8824 and F1 score = 0.9167) as my champion model to predict the store format of the new store.

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

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

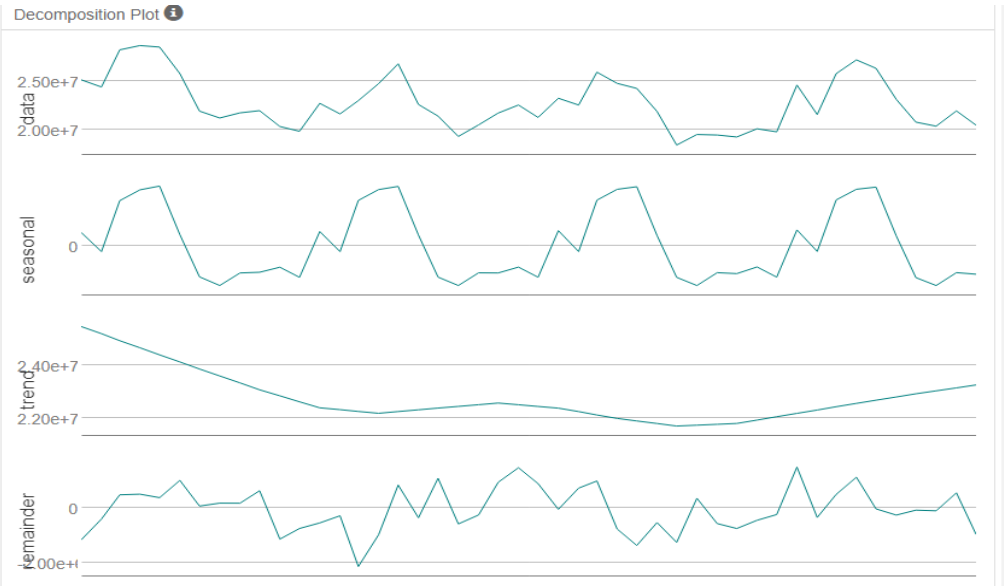
Task 3: Predicting Produce Sales

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?
   Answer:

   I built 3 models –

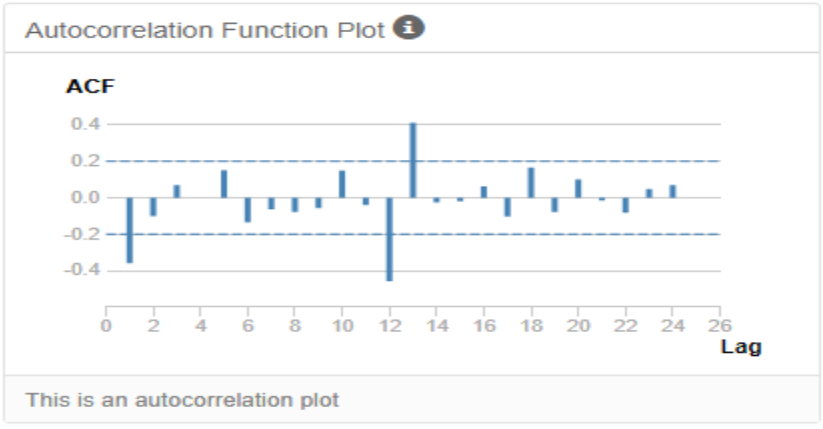
   ETS(A,A,A)
   ARIMA(0,1,1)(0,1,1)12
   ARIMA(0,1,1)(0,1,2)12

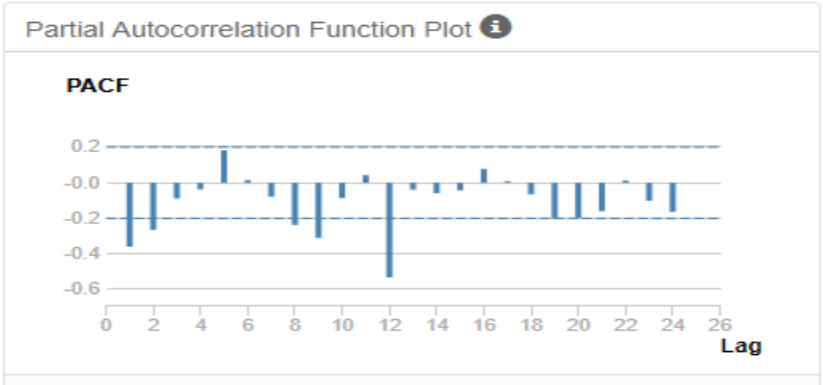
   I saw that the original data had all three components –
   Seasonal, Trend, and Error



So, I decided to take a seasonal difference and then a first difference to make the data stationary.

Once I had the stationary data, based on acf and pacf plots below, I decided to build ARIMA(0,1,1)(0,1,1)12 model because 1 lag and seasonal lag 12 had negative autocorrelations.





I also built an additional model ARIMA(0,1,1)(0,1,2)12 to account for strong seasonality in the data.

Finally, compared their AIC and RMSE values and chose ARIMA(0,1,1)(0,1,2)12 As my champion model to forecast the results.

ARIMA(0,1,1)(0,1,2)12

Information Criteria:

AIC	AICc	BIC
849.858	851.6762	855.0414

In-sample error measures:

ME	RMSE
126793.0202234	763923.4295347

ARIMA(0,1,1)(0,1,1)12

Information Criteria:

AIC	AICc	BIC
849.8292	850.8727	853.7167

In-sample error measures:

ME	RMSE
150815.8641194	935292.1712234

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.
Answer:

	Period	Sub_Period	Final New Forecast	Final Existing Forecast
1	2016	1	2482154.753384	20818603.751657
2	2016	2	2379446.053018	19998263.715615
3	2016	3	2722335.003175	22941700.427315
4	2016	4	2602490.695063	21811154.429781
5	2016	5	2926332.545815	24608787.545045
6	2016	6	2948903.97455	24846848.730104
7	2016	7	3017376.200741	25368471.082913
8	2016	8	2663498.930238	22303364.088422
9	2016	9	2320273.874618	19406302.670956
10	2016	10	2248760.352154	18821163.186906
11	2016	11	2311573.447693	19409504.24398
12	2016	12	2339234.407607	19665164.506506

Parag Sasturkar