Project: Predictive Analytics Capstone

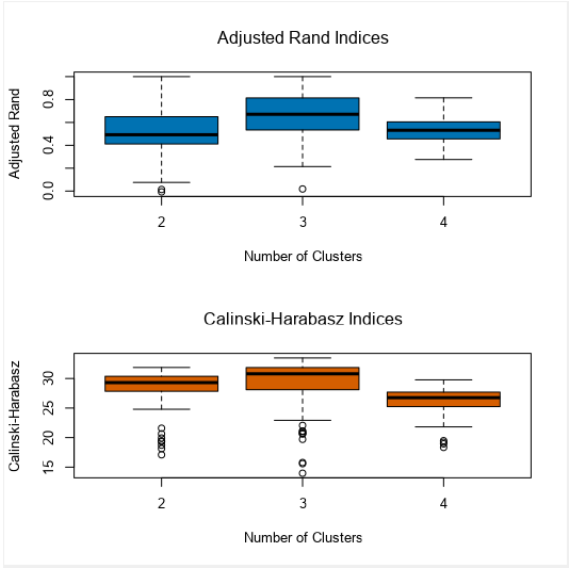
Valencia Becerra Gerardo

Task 1: Determine Store Formats for Existing Stores

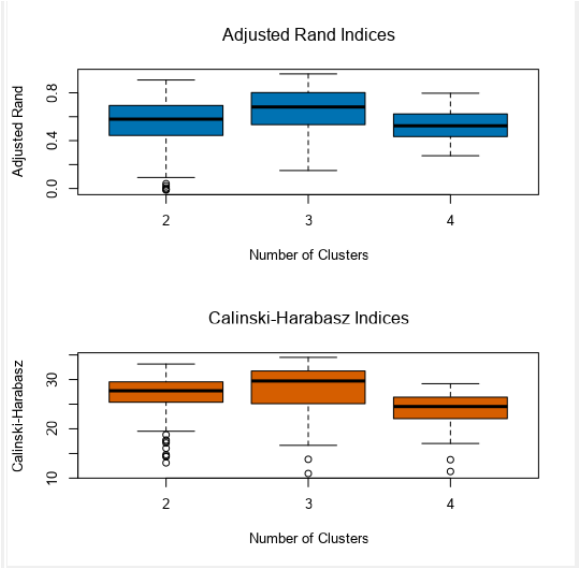
1. What is the optimal number of store formats? How did you arrive at that number?

   The optimal number of store formats is 3. This is the reported value (using 3 different methods: K-Means, K-Medians and Neural Gas) that shows higher Adjusted Rand and CH Indices when comparing to 2 and 4 clusters, it also shows a higher median.

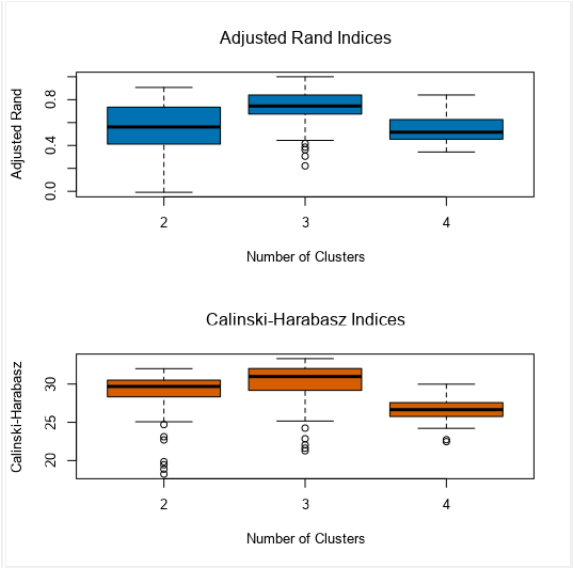
   K-Means



   K-Medians



Neural Gas



2. How many stores fall into each store format?

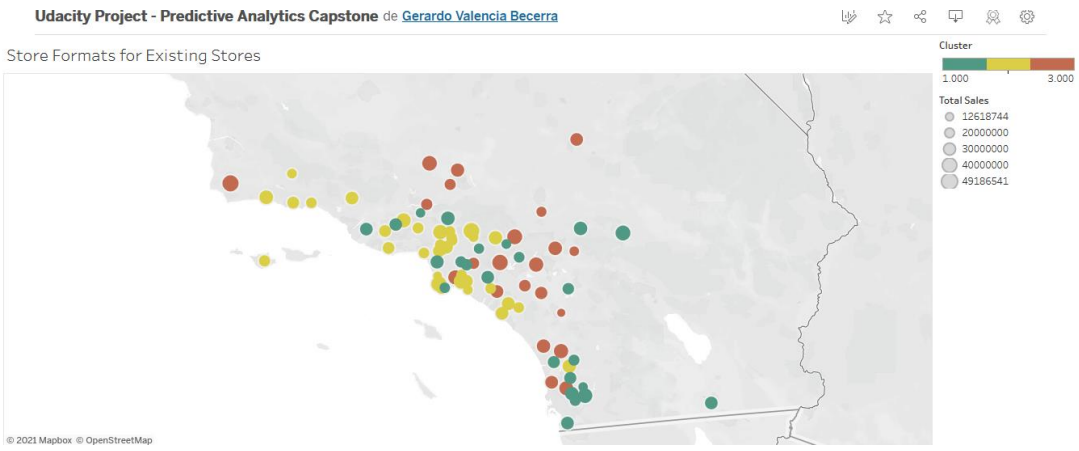
   Store Format #1 has 25 stores; Format #2 has 35 and Format #3 25 stores.

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

   In general: Cluster #1 is high in Deli and/or low in General Merchandise.
   Cluster #2 is high in Produce and/or Floral.
   Cluster #3 is high in General Merchandise and/or low in Dairy and/or Bakery.

	Percentage_Dry_Grocery	Percentage_Dairy	Percentege_Frozen_Food	Percentage_Meat	Percentage_Produce	Percentage_Floral	Percentage_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482
	Percentage_Bakery	Percentage_General_Merchandise					
1	0.428226	-0.674769					
2	0.312878	-0.329045					
3	-0.866255	1.135432					

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.



The Tableau Public file can be accessed through the following URL: https://public.tableau.com/app/profile/gerardo.valencia.becerra/viz/UdacityProject-PredictiveAnalyticsCapstone/Hoja1?publish=yes

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?

   To classify the new stores in the best format 3 different classification models were tried: Decision Tree, Random Forests and Boosted Models. I chose the Boosted Model as it has the highest accuracy.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
StoreSegment_DT	0.6471	0.6667	0.5000	1.0000	0.5000
StoreSegment_RF	0.7059	0.7500	0.5000	1.0000	0.7500
StoreSegment_Boosted	0.7647	0.8333	0.5000	1.0000	1.0000

Confusion matrix of StoreSegment_Boosted

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	2	5	0
Predicted_3	2	0	4

Confusion matrix of StoreSegment_DT

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	2
Predicted_2	3	5	0
Predicted_3	1	0	2

Confusion matrix of StoreSegment_RF

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

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	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
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?

   Two models were compared: ETS(M,N,M) method and ARIMA(1,0,0)(1,1,0)[12].

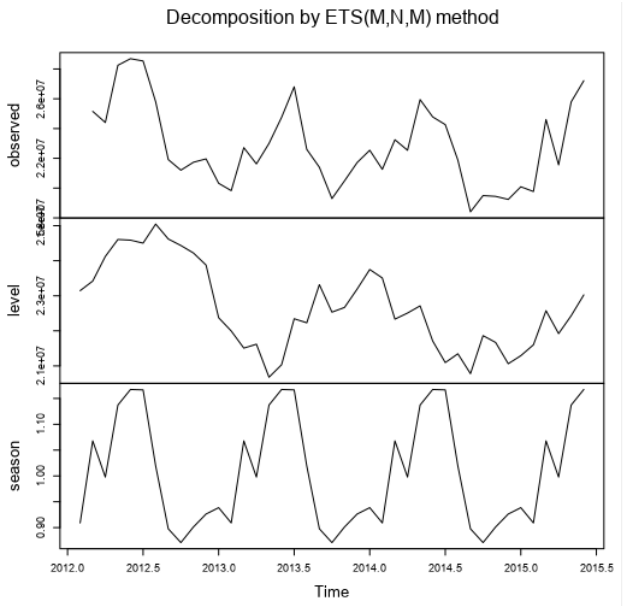
Actual and Forecast Values:

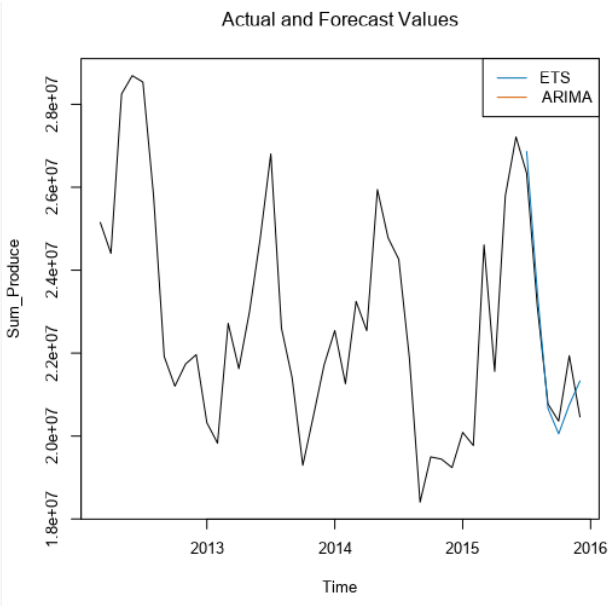
Actual	ETS	ARIMA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

The comparison shows that the ETS Model has significantly smaller error so it's the one we'll choose for our forecasting.





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-16	2,563,357.910041	21,829,060.031666
Feb-16	2,483,924.727562	21,146,329.631982
Mar-16	2,910,944.145687	23,735,686.93879
Apr-16	2,764,881.869697	22,409,515.284474
May-16	3,141,305.867305	25,621,828.725097
Jun-16	3,195,054.203804	26,307,858.040046
Jul-16	3,212,390.95409	26,705,092.556349
Aug-16	2,852,385.769198	23,440,761.329527
Sep-16	2,521,697.18679	20,640,047.319971
Oct-16	2,466,750.893696	20,086,270.462075
Nov-16	2,557,744.587714	20,858,119.95754
Dec-16	2,530,510.805133	21,255,190.244976