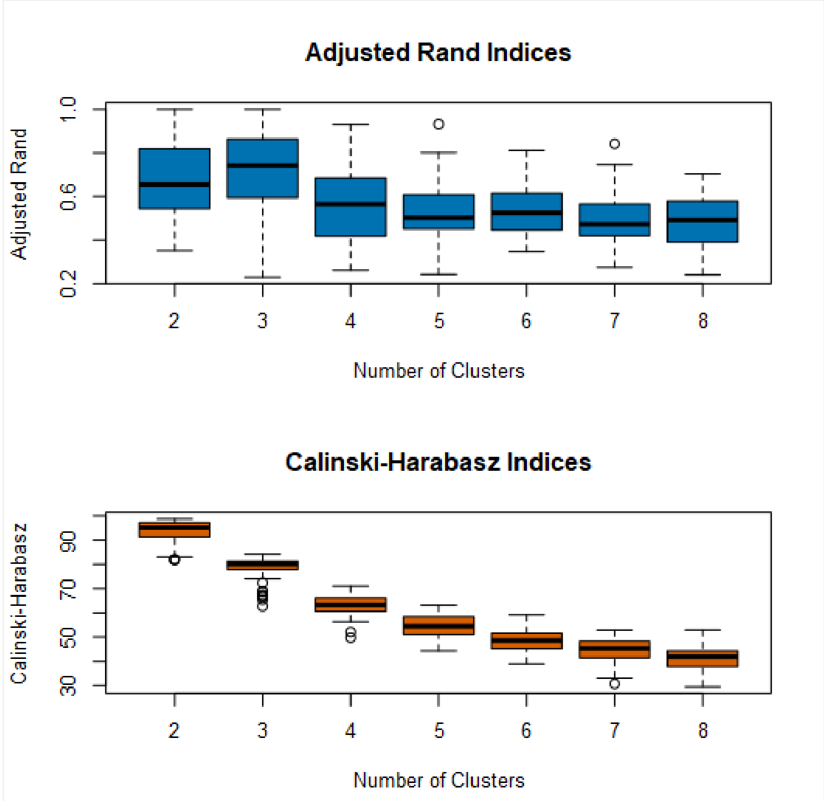
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

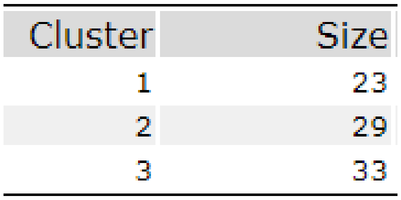
1. What is the optimal number of store formats? How did you arrive at that number?

   The optimal numbers of store formats are 3 based on the K- Means clustering method, this number is chosen based on the Adjusted Rand Indices and the CH indices below, below are shown the Box Plot Models, for the AR it show 3 cluster with the better stability, for The CH indices it show the deisticness and compactness, overall 3 formats seem to be the best solution for the data given.



2. How many stores fall into each store format?

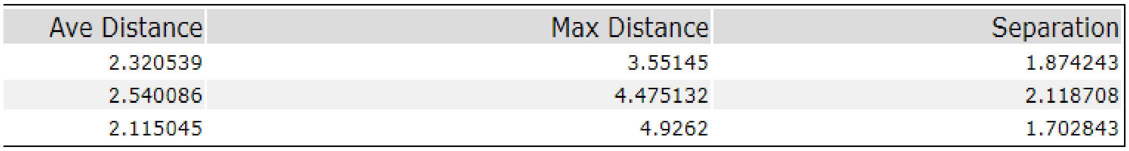
   For each store format there are different numbers of stores that fall into them, the first store format has a size of 23 stores, the second format has a sized of 29 stores, and the last 33 stores are included in the format three, overall this adds up to the 85 stores.



Cluster	Size
1	23
2	29
3	33

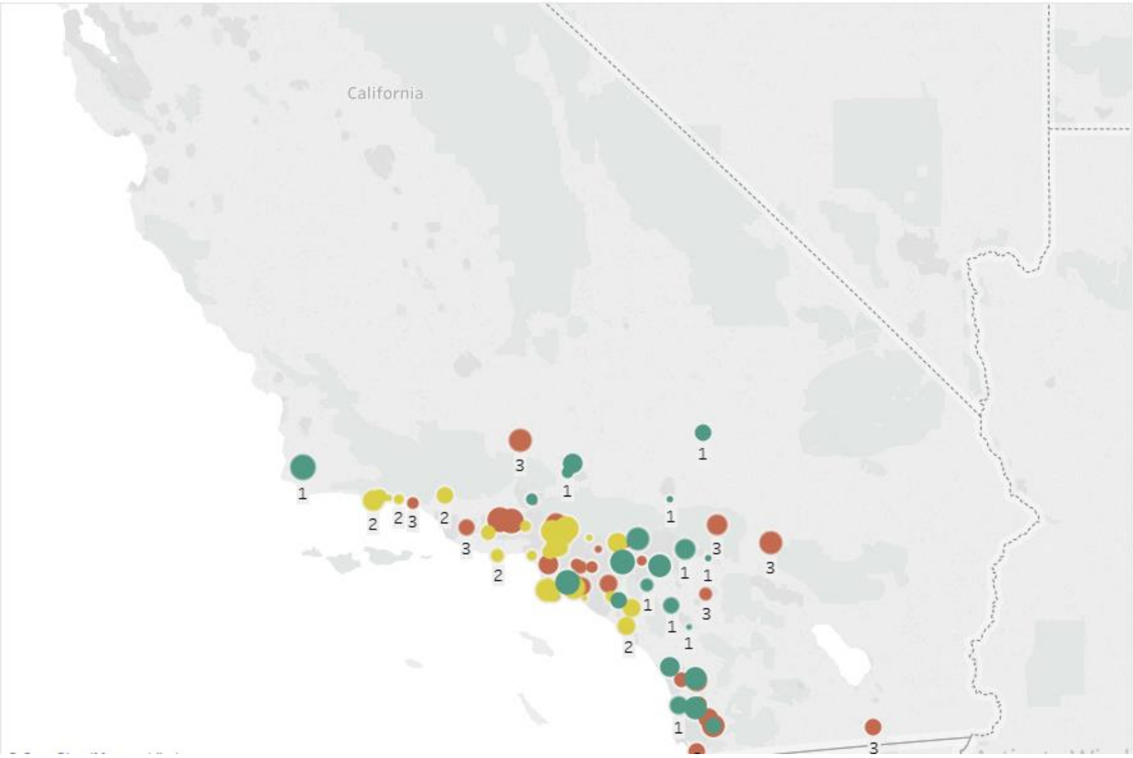
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

One way that the cluster differ form one another is the separation, for example while cluster 1 and cluster 3 have similar separation, the data for cluster 2 seem to have a bit more separation of 2.11, this showing the difference within the cluster. Finally, Cluster 1 has the highest general merchandise percentage, while cluster 2 has maximum sales in the grocery category.



Ave Distance	Max Distance	Separation
2.320539	3.55145	1.874243
2.540086	4.475132	2.118708
2.115045	4.9262	1.702843

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.



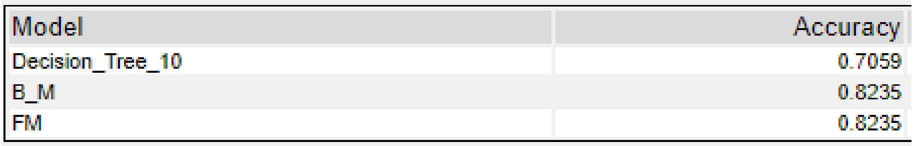
[1]

[1] https://public.tableau.com/shared/6YMR424Q6?:display_count=yes

Task 2: Formats for New Stores

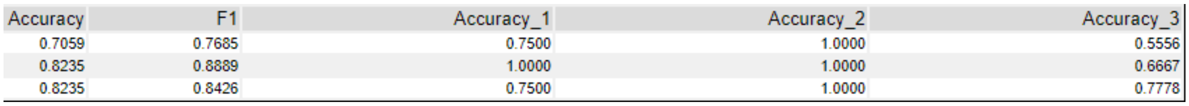
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   The methodology that was used to predict the best store format for the new stores was the Boost Model, As I used 20% of validation sample to compare for accuracy with random seed =3 to be able to see the difference between the models, the Boosted Model and forest model both showed good accuracy results.



Model	Accuracy
Decision_Tree_10	0.7059
B_M	0.8235
FM	0.8235

   the Boost Model was chosen because it shows a better precision measure (F1) than the Forest model, shown in the table



Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
0.7059	0.7685	0.7500	1.0000	0.5556
0.8235	0.8889	1.0000	1.0000	0.6667
0.8235	0.8426	0.7500	1.0000	0.7778

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

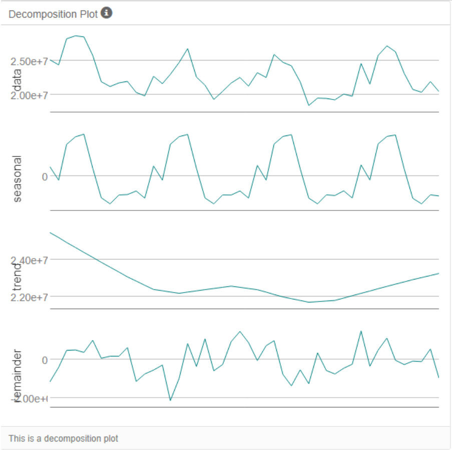
Task 3: Predicting Produce Sales

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?

For each of the forecast that were conducted I used ETS (m, n, m) and Arima (1,1,0), I came to decision of using the following model after inspecting the corresponding seasonality, trend, and the remainder as show in TS plot below. Constant and increases or decreases and the corresponding ACF and PACF plots. ETS was chosen for better correlation.

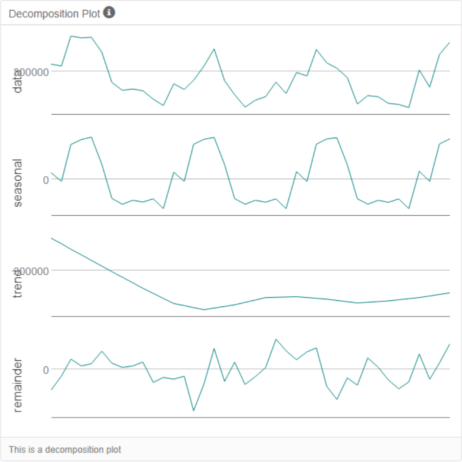
For existing Stores and new stores, the error is multiplicative as it sows in the figure because it seem to grow and shrink in time as shown on the graphic, as for the trend selected in the model was set to neutral as the trend line changes direction toward the end of the period and goes back up, finally the seasonality that was apply in the analysis was multiplicative. The mayor deference between existing and new stores were trend also show in graphs.

Graph of Error, Trend and Seasonality for existing stores



This is a decomposition plot

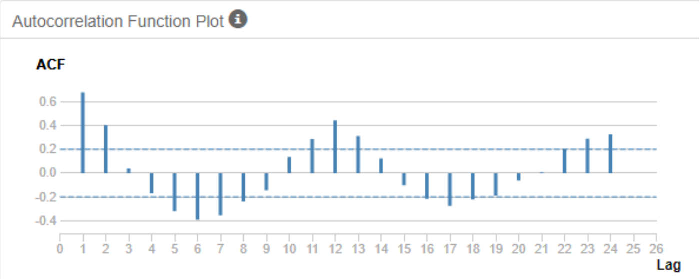
Graph of Error, Trend and Seasonality for new stores



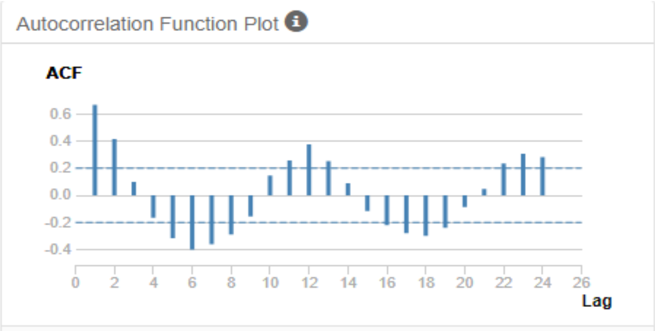
This is a decomposition plot

For the Arima model the type of model was made by analyzing the corresponding ACF and PACF plots

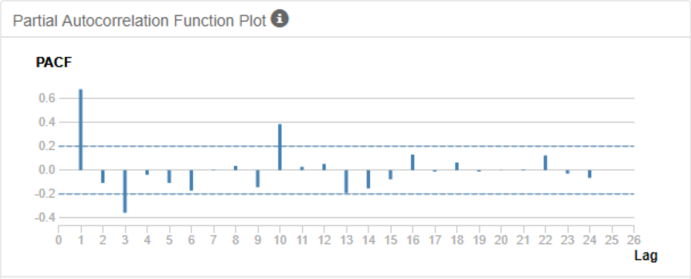
Graph of ACF for existing stores



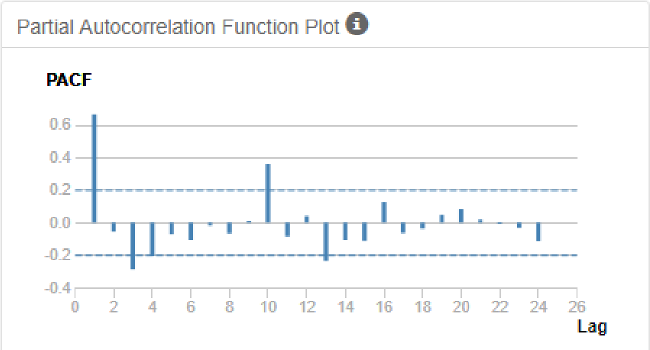
Graph of ACF for new stores



Graph of PACF for existing stores



Graph of PACF for new stores



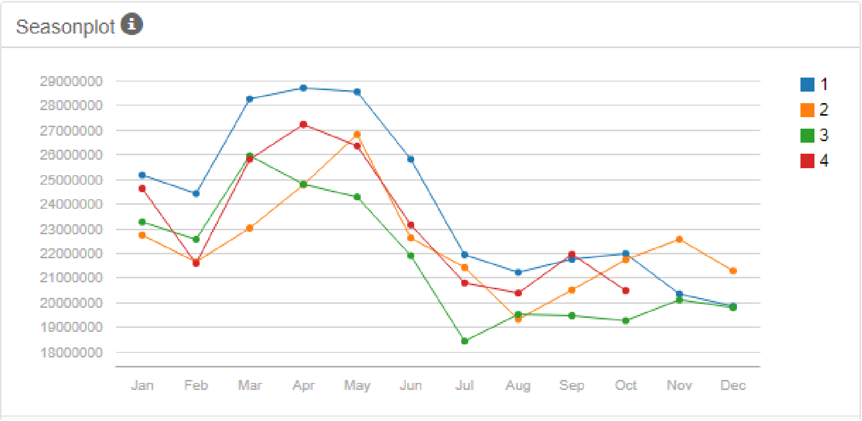
As shown the plots helps determine the best model to choose to apply in the corresponding data. As show on both cases of the ACF plot there is a tendency of decay toward 0, also in both plots there is quickly cut towards 0 indicating AR, finally the positive correlation on ACF will indicated AR as well. The PACF indicates that all autocorrelation is explain by Lag-1 indicating an AR1 model, the number of times used to make the model stationary is 1.

3. 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.

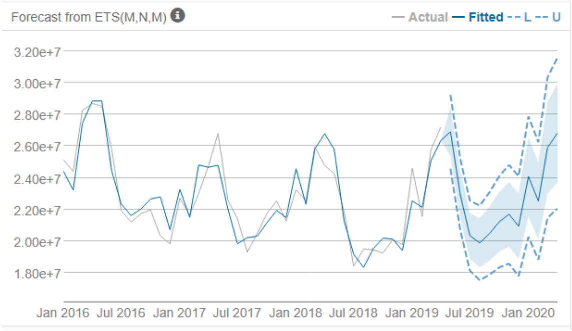
Forecast for existing and new stores

Month	New stores	Existing Stores
January-16	1,077,872	25,200,000
February-16	1,057,546	24,400,400
March-16	1,214,088	28,200,000
April-16	1,201,667	28,700,000
May-16	1,203,681	28,500,000
June-16	1,101,532	25,800,000
July-16	941,590	21,900,000
August-16	907,934	21,200,000
September-16	923,572	21,700,000
October-16	924,771	20,000,000
November-16	861,608	20,300,000
December-16	834,558	19,800,000

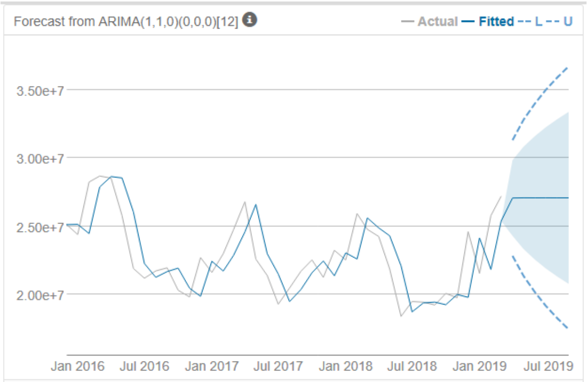
Season historical plot



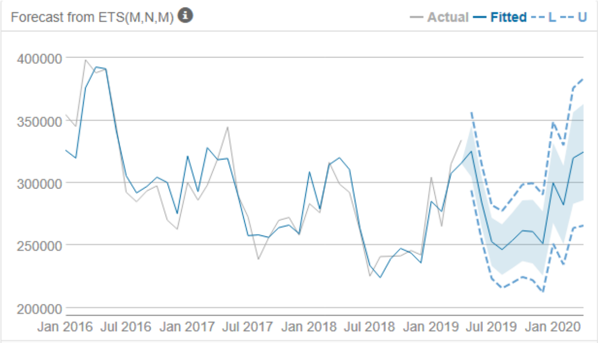
Existing Stores forecast ETS Model



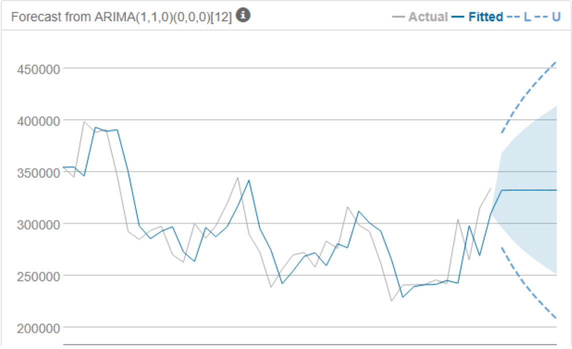
Existing Stores forecast Arima Model



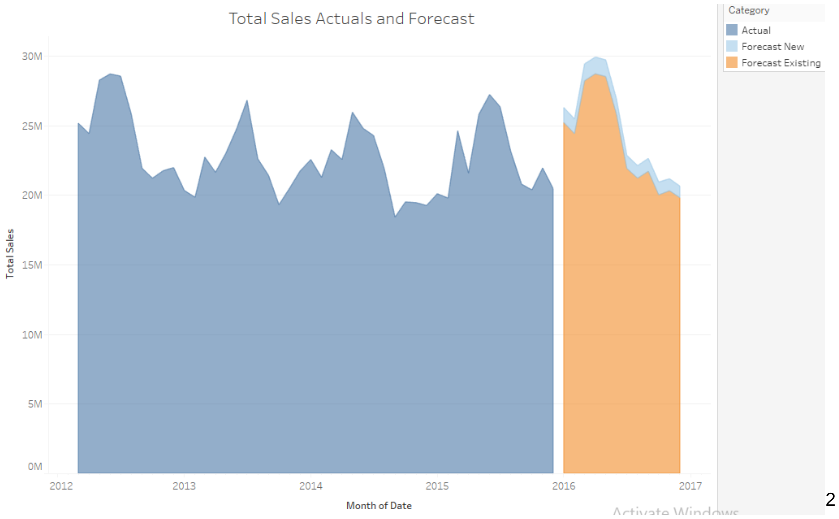
New store forecast ETS Model



New Stores forecast Arima Model



Forecast Tableau visualization



[2]

[2] https://public.tableau.com/views/Forecasttable/Sheet1?:embed=y&:display_count=yes&publish=yes