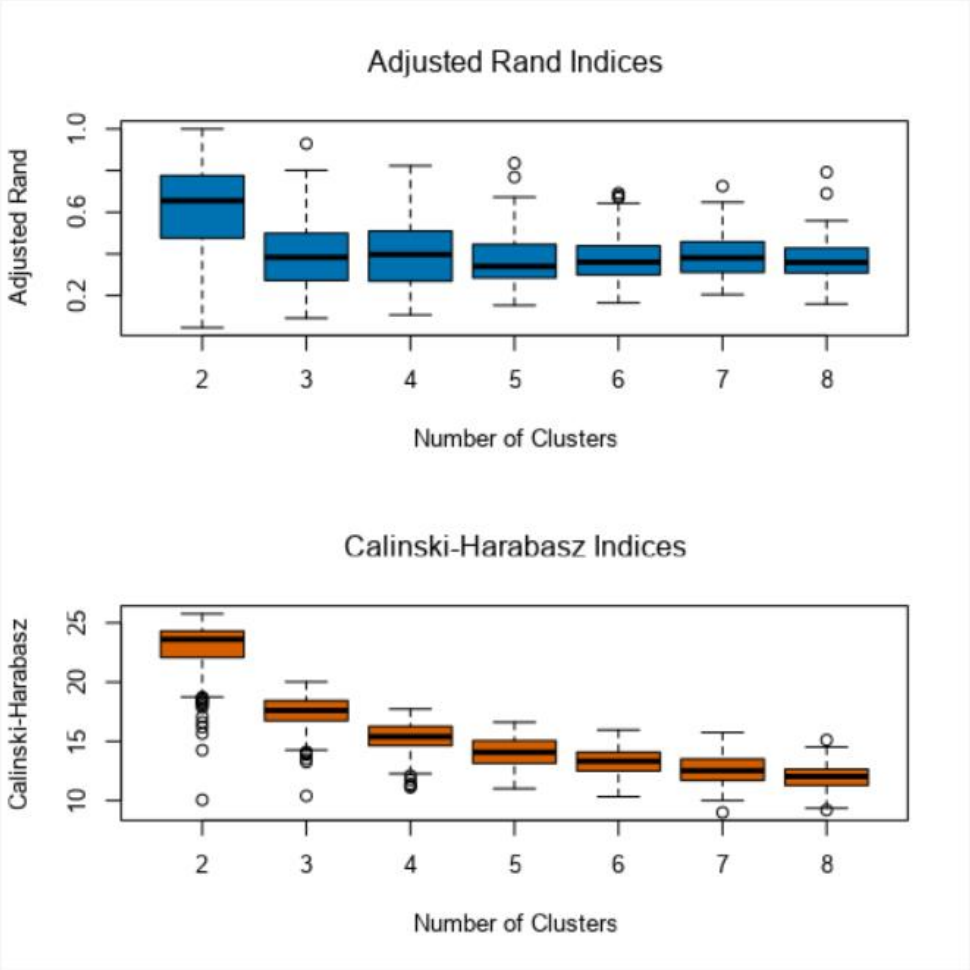
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   Detailed analysis was performed using K-Centroids diagnostic tool. The clustering method was chosen as K-Means. We get the below report for the same.

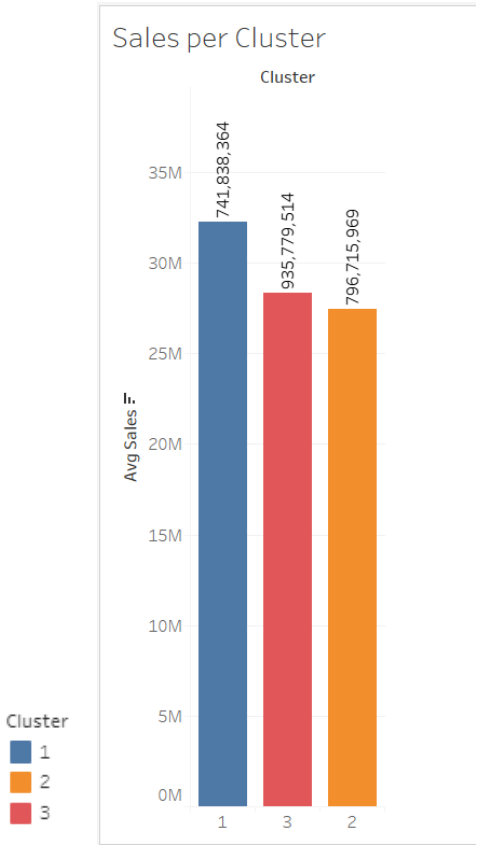
   Based on the Calinski-Harabasz Index we can see that the optimal cluster number is either 2 or 3. However in cluster number as 2 we can see there are many outliers and also the spread is more (suggesting an unstable cluster) as seen by the Adjusted Rand Indices Plot. Furthermore, choosing cluster number as 2 will result in having more than 40 stores per cluster which is contrary to the management's request. Thus we choose **3 clusters** as the optimal number of clusters.



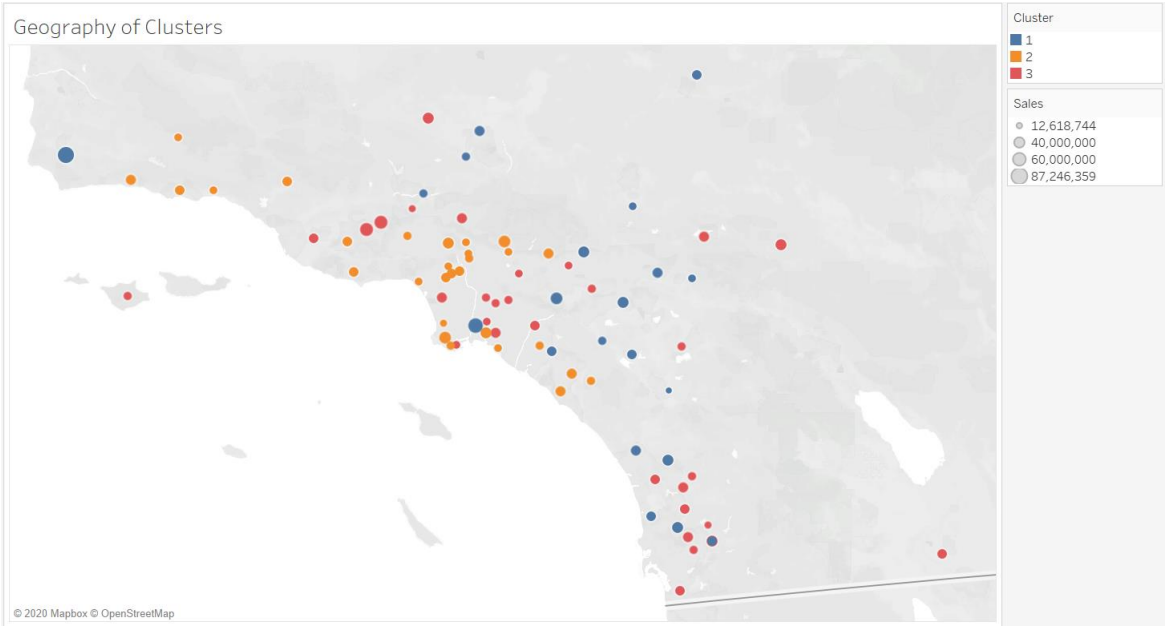
2. How many stores fall into each store format?
   Cluster 1: 23 stores
   Cluster 2: 29 stores
   Cluster 3: 33 stores

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

   As seen from the plot below the average sales in cluster 1 are the highest followed by cluster 3 and then cluster 2.



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. (https://public.tableau.com/shared/9HQ4HDDG3?:display_count=y&:origin=viz_share_link)



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.)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750
Decision_Tree	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000

As it can be seen from the table above all the models have same accuracy. Hence, the methodology chosen was **Boosted Model** as it has the highest F1 value.

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	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

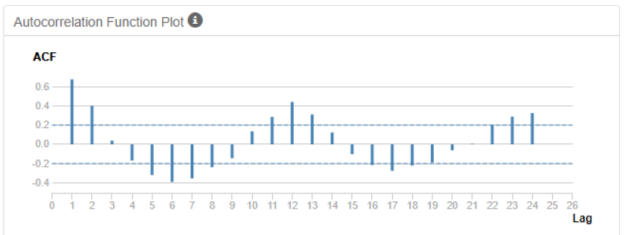
Task 3: Predicting Produce Sales

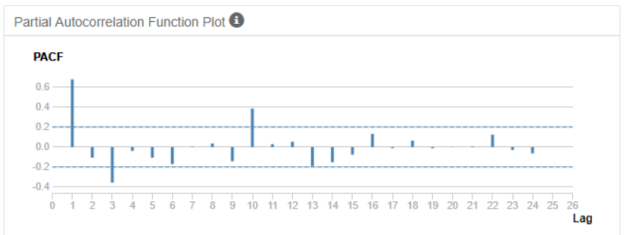
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

The decomposition plot was generated using TS-Plot.
ETS:
Error: The error is increasing in variance hence it will be Multiplicative
Trend: There is no clear trend hence it will be None
Seasonality: The sales fluctuates after regular interval thus indicating seasonality. Furthermore, the peak is growing hence it would be Multiplicative

So based on the above reasoning ETS(M,N,M) model will be used. This was verified by using ETS Auto to get the best model based on Alteryx recommendation.

ARIMA:

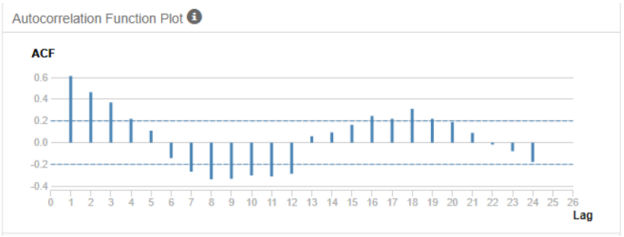


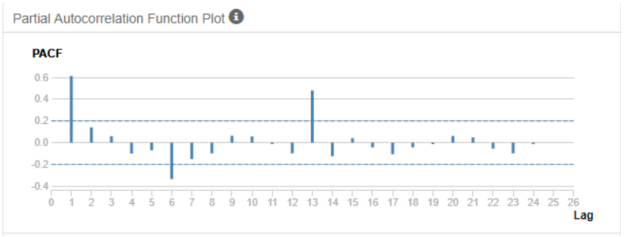


The ACF plot shows a gradual decline to zero and also a seasonal lag, but there is a

high serial correlation and hence it needs to be seasonally differenced. PACF shows high serial correlation after lag1, hence we need to difference the dataset.

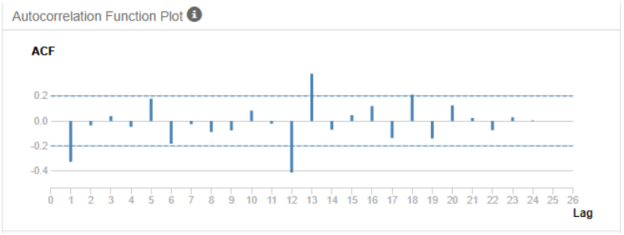
Seasonal Difference:

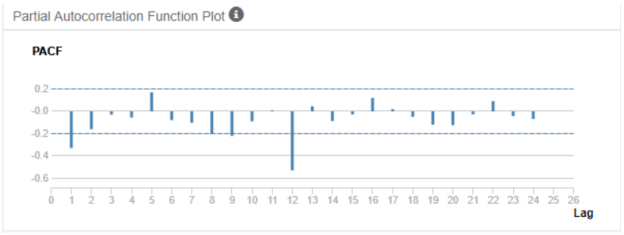




The ACF plot shows a gradual decline to zero and a positive correlation, but the serial correlation is still high. Hence we need to stationarize by differencing again.

First Seasonal Difference:





The time series is now stationary.

Non seasonal component: p=0, q=1, d=1 as ACF negative and cuts off sharply.
Seasonal component: P=0, Q=1, D=1 as ACF negative at lag 1
M = 12 as holdout sample is 12

Hence we get ARIMA(0,1,1)(0,1,1)12

From the accuracy measures we can see that ETS outperforms ARIMA and hence will be used.

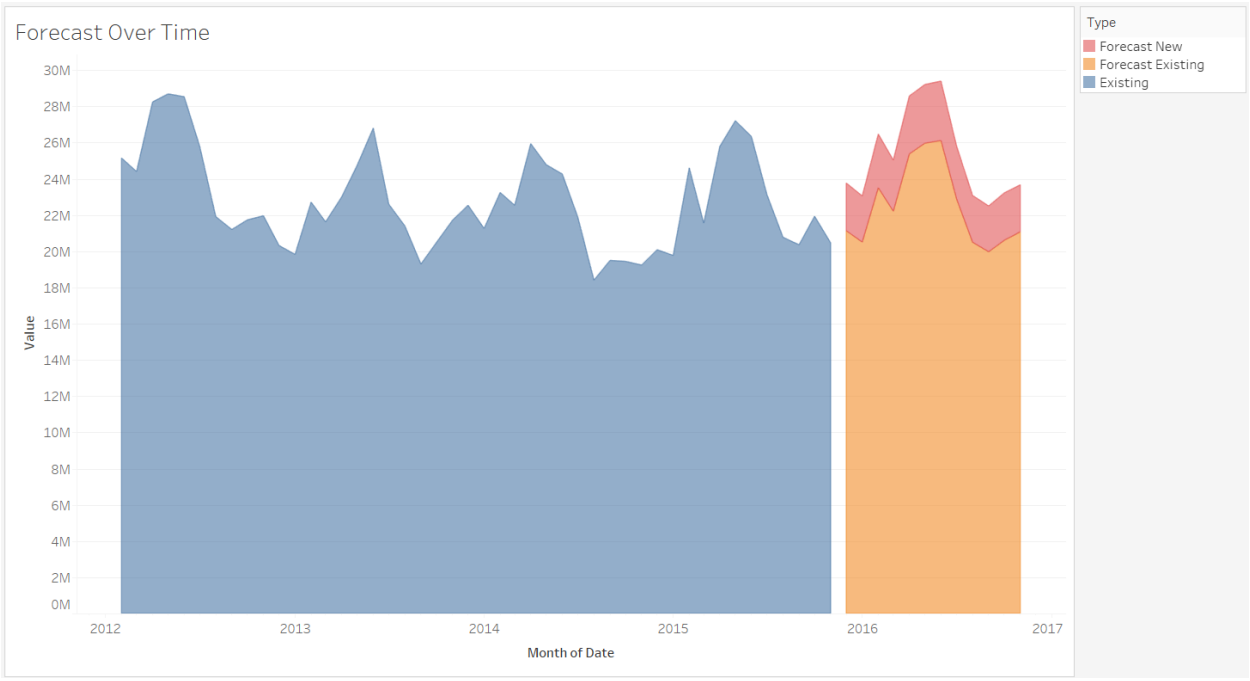
Accuracy Measures:

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

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.

Year	Month	Forecast Existing	Forecast New	Forecast Total
2016	1	21136641.78	2558028.55	23694670.33
2016	2	20507039.12	2469379.827	22976418.95
2016	3	23506565.98	2879302.753	26385868.73
2016	4	22208405.76	2751395.939	24959801.7
2016	5	25380147.77	3115792.072	28495939.84
2016	6	25966799.47	3166787.746	29133587.22
2016	7	26113792.57	3192488.639	29306281.21
2016	8	22899285.77	2828498.291	25727784.06
2016	9	20499583.91	2504011.425	23003595.34
2016	10	19971242.82	2451214.767	22422457.59
2016	11	20602665.92	2545168.699	23147834.62
2016	12	21073222.08	2533847.61	23607069.69



(https://public.tableau.com/views/GroceryStoreCapstone/Story1?:display_count=y&publish=yes&:origin=viz_share_link)