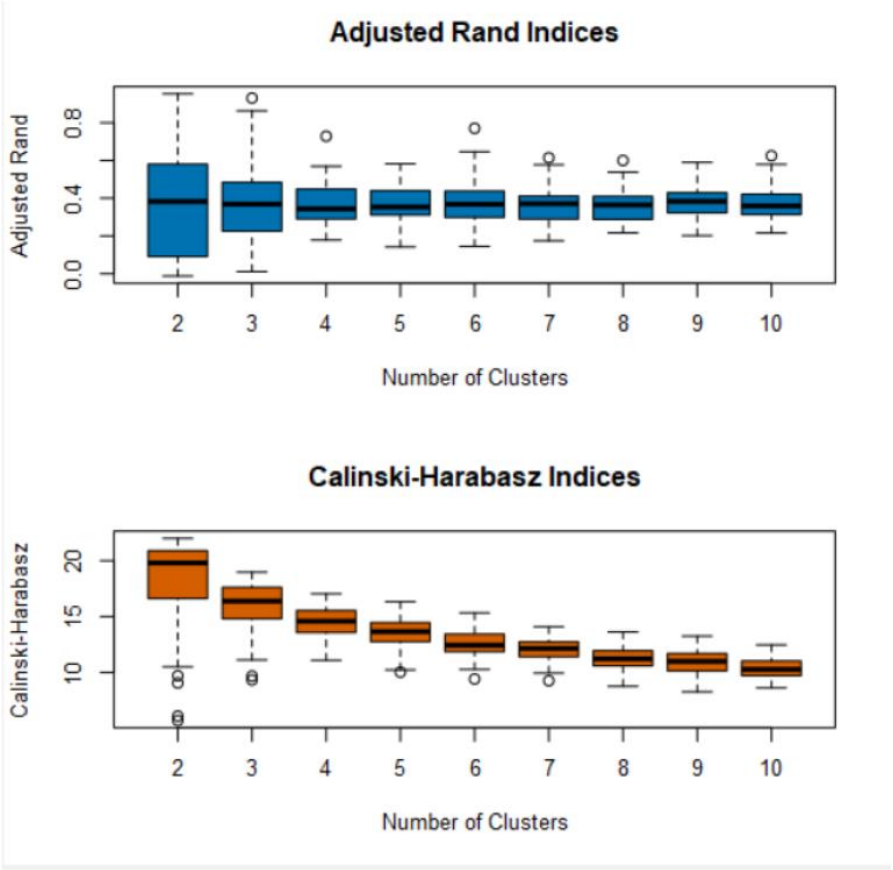
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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



According to the Adjusted Rand Indices the number of cluster with the higher stability is 2 and 3 and according to the Calinski-Harabasz Indices they are the ones with better distinctness and compactness. Therefore, they are likely to perform similarly, for the matter of this project I chose three clusters.

2. How many stores fall into each store format?

Cluster Information:

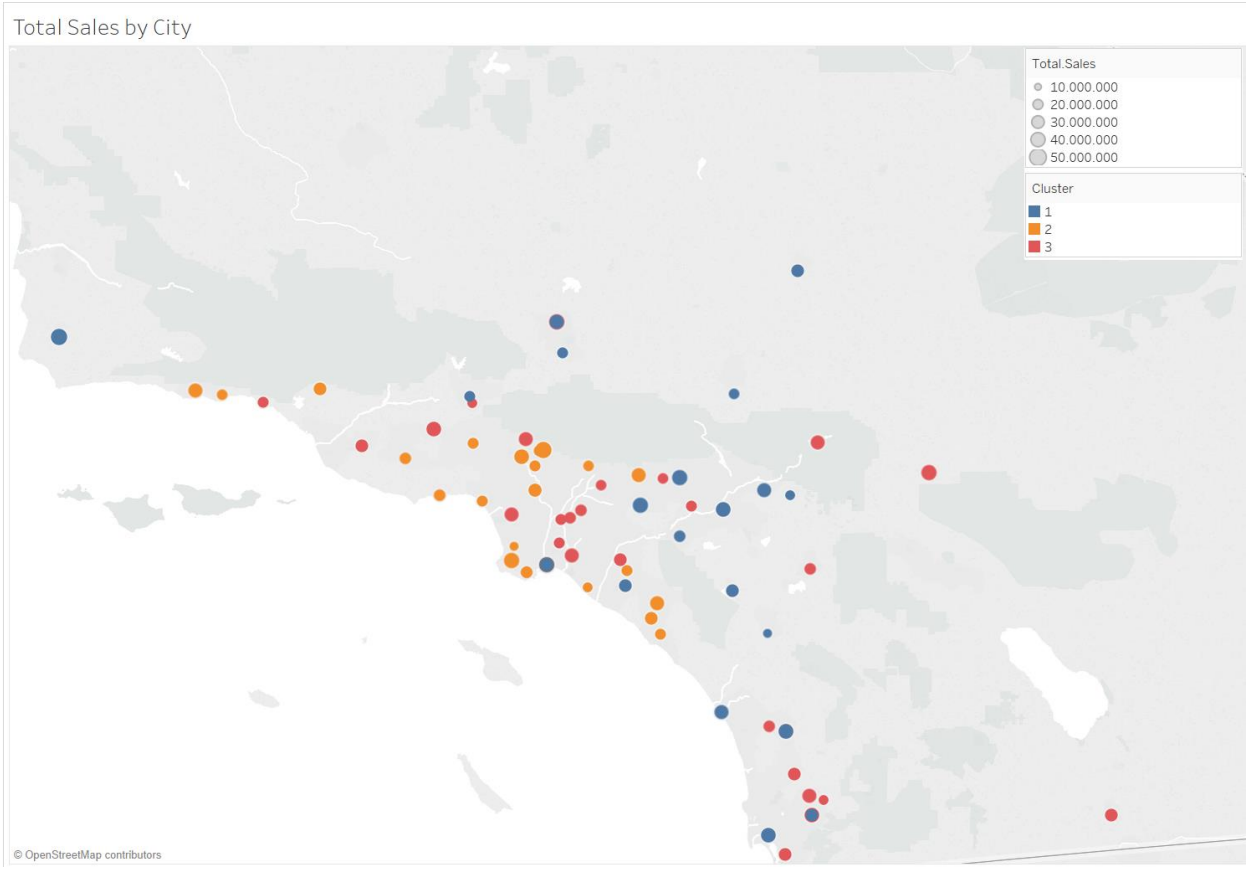
Cluster	Size
1	23
2	29
3	33

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

Cluster	General_Merchandise	Total Sales	General_Merchandise_Perc
1	67685106.61	741838363.72	0.09124
2	55824735.43	796715969.04	0.070069
3	60569632.57	935779513.54	0.064726

One way the clusters vary is the total of the General Merchandise Sales with a difference of almost seven million, even though Cluster 3 has more stores across the three and the Cluster 1 has less across the three.

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.)

Model	Accuracy	F1
RF_Model	0.8235	0.8426
DT_Model	0.7059	0.7685
Boosted_Model	0.8235	0.8889

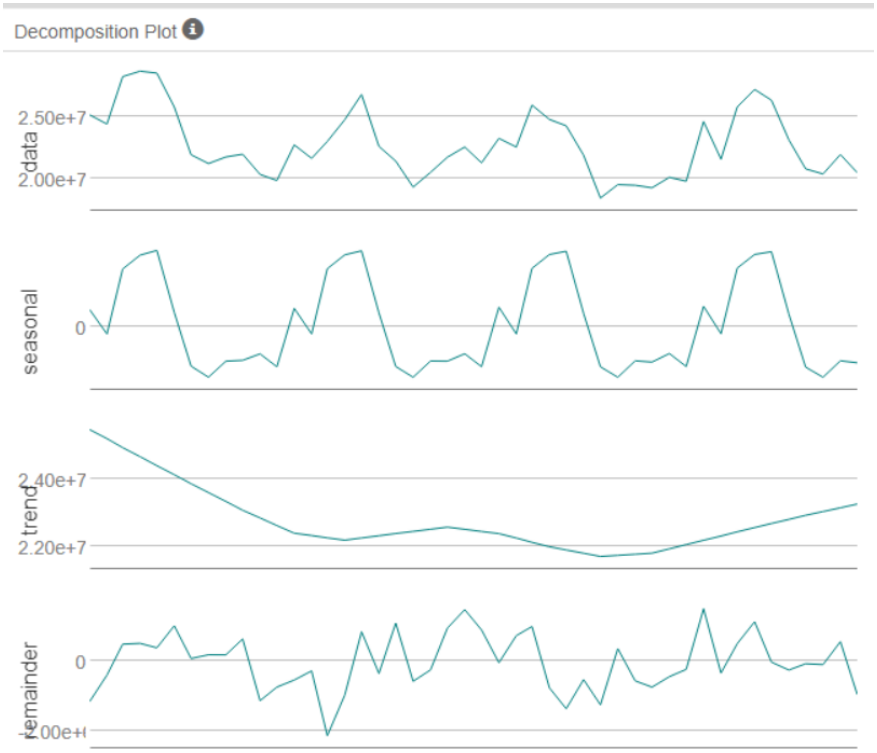
The chosen model was the Boosted Model, its accuracy is the same as the Random Forest Model, however, its F1-Score is a little higher, it means the Boosted Model generalizes better.

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

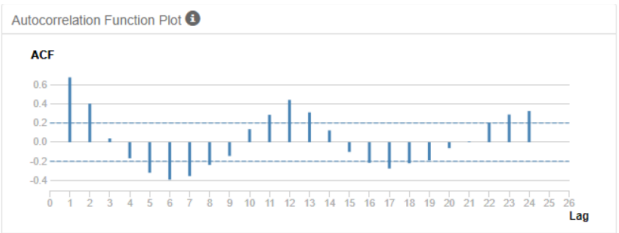
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

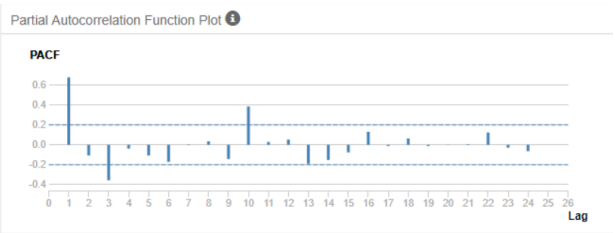
Task 3: Predicting Produce Sales

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?

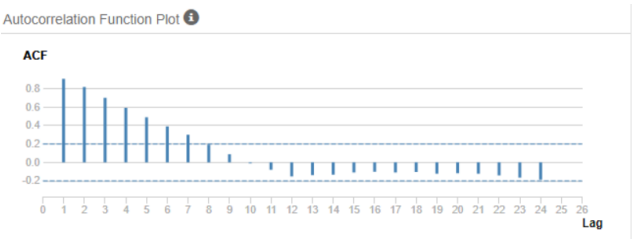


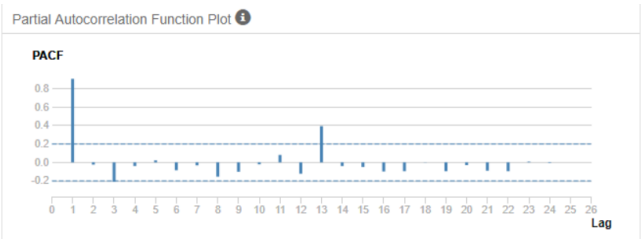
The ETS model was built from the analysis of the Decomposition Plot above and it shows that the Error is Multiplicative because the remainder plot is fluctuating between large and small errors over time, the Trend is None because there is no clear trend and the Seasonality is Multiplicative because the seasonal fluctuations tend to increase and decrease with the level of the time series. Therefore, the configuration for the model is ETS(M,N,M).



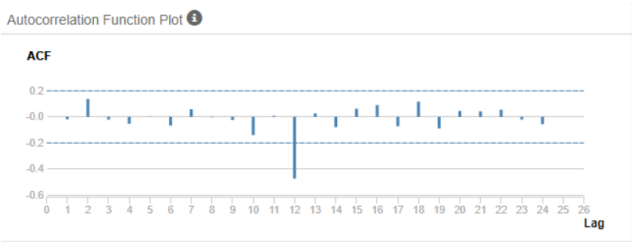


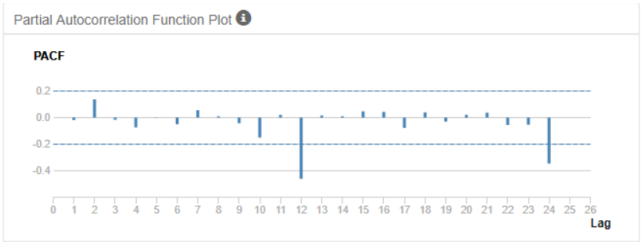
Analyze the ACF/PACF graph it is possible to see that the series is not stationary, we can verify in the ACF graph that neither the mean nor the variance is constant over time, so it is needed to take the Seasonal Differencing.





After the Seasonal Difference is taken we can see in the ACF/PACF graphs that the series is still not stationary but the seasonal component is gone, therefore we need to take the First Differencing.





After the First Difference it is possible to see that the series is now stationary except for Lag 12.

So we took the a seasonal difference and a first difference and this suggest that the lower case d and the upper case D of the ARIMA model have the value of 1. The M term is 12 because there are 12 months or periods in each season. Therefore the configuration for the model is ARIMA(0,1,0)(0,1,0)[12].

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	-718581.3	952392.2	739203.9	-3.3454	3.4394	0.435

I then compared the ETS(M,N,M) and the ARIMA(0,1,0)(0,1,0)[12] and the ETS model obtained the smallest error measures for RMSE and MASE in the validation sample, so I used this configuration to forecast the produce sales for 2016.

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.

Date	Forecast Existing Stores	Forecast New Stores
2016-1	21,539,936.01	2,584,383.53
2016-2	20,413,770.60	2,470,873.92
2016-3	24,325,953.10	2,906,307.87
2016-4	22,993,466.35	2,771,532.13
2016-5	26,691,951.42	3,145,848.57
2016-6	26,989,964.01	3,183,909.28
2016-7	26,948,630.76	3,213,977.72
2016-8	24,091,579.35	2,858,247.21
2016-9	20,523,492.41	2,538,173.64
2016-10	20,011,748.67	2,483,550.17
2016-11	21,177,435.49	2,593,089.19
2016-12	20,855,799.11	2,570,200.44