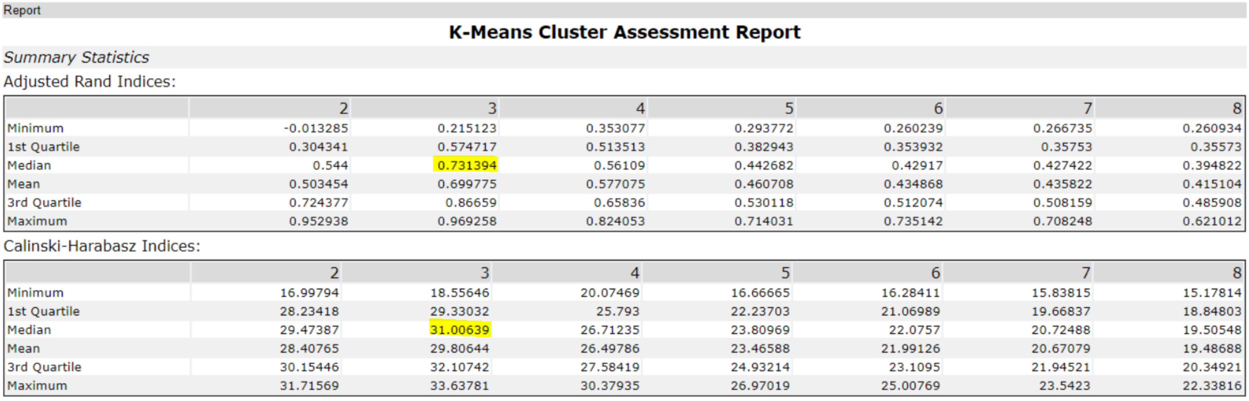
Project: Predictive Analytics Capstone

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 number of store formats is 3. This was arrived by using the K-Centroid Diagnostics tool in Alteryx. AR and CH value for each of the cluster is compared, in which Median of cluster 3 is high. AR value 0.7313 and CH Value 31.0.



Report

K-Means Cluster Assessment Report

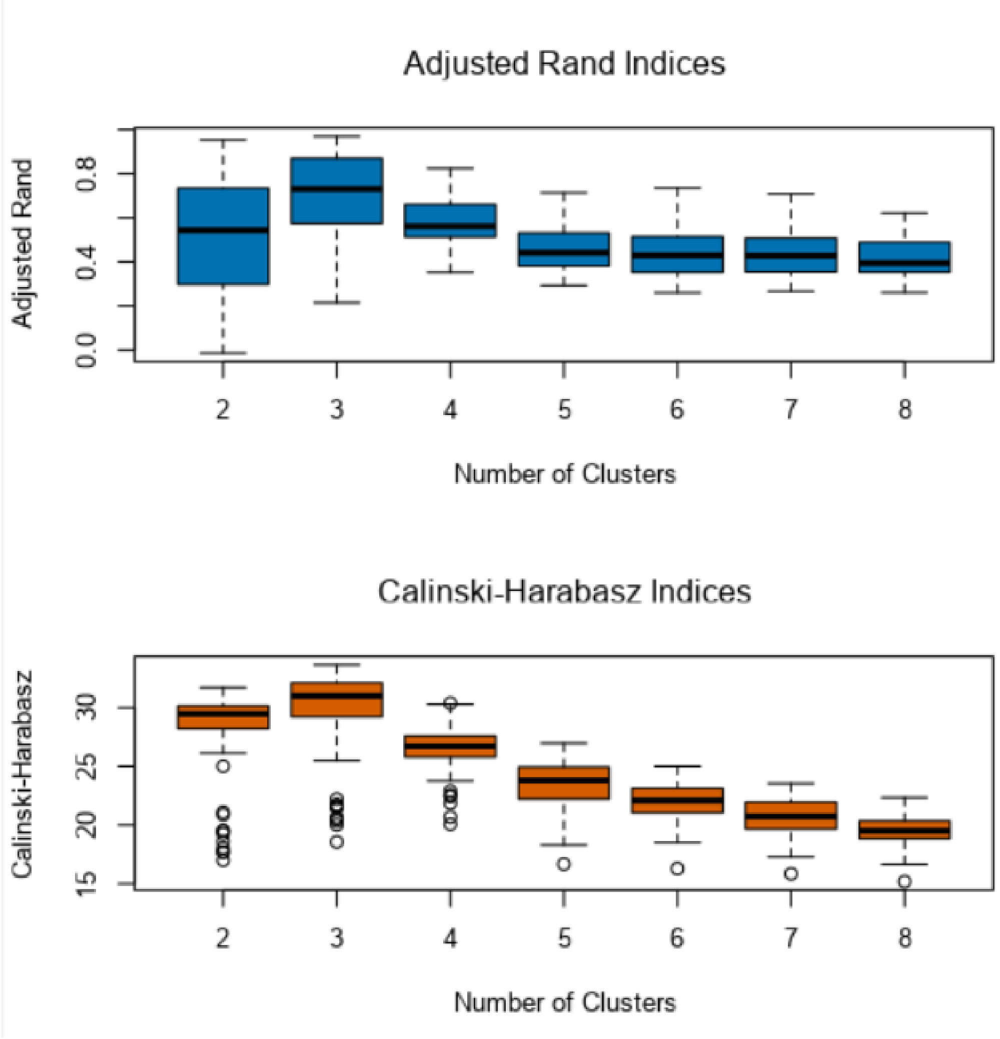
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.013285	0.215123	0.353077	0.293772	0.260239	0.266735	0.260934
1st Quartile	0.304341	0.574717	0.513513	0.382943	0.353932	0.35753	0.35573
Median	0.544	0.731394	0.56109	0.442682	0.42917	0.427422	0.394822
Mean	0.503454	0.699775	0.577075	0.460708	0.434868	0.435822	0.415104
3rd Quartile	0.724377	0.86659	0.65836	0.530118	0.512074	0.508159	0.485908
Maximum	0.952938	0.969258	0.824053	0.714031	0.735142	0.708248	0.621012

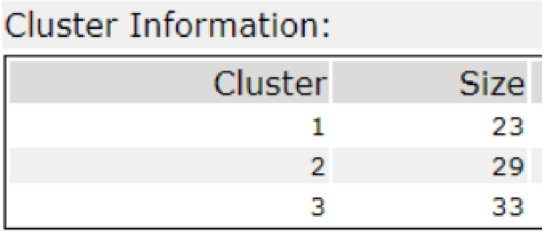
Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.99794	18.55646	20.07469	16.66665	16.28411	15.83815	15.17814
1st Quartile	28.23418	29.33032	25.793	22.23703	21.06989	19.66837	18.84803
Median	29.47387	31.00639	26.71235	23.80969	22.0757	20.72488	19.50548
Mean	28.40765	29.80644	26.49786	23.46588	21.99126	20.67079	19.48688
3rd Quartile	30.15446	32.10742	27.58419	24.93214	23.1095	21.94521	20.34921
Maximum	31.71569	33.63781	30.37935	26.97019	25.00769	23.5423	22.33816



2. How many stores fall into each store format?

   Using K-Means, cluster analysis is performed, and the distribution is as shown below.

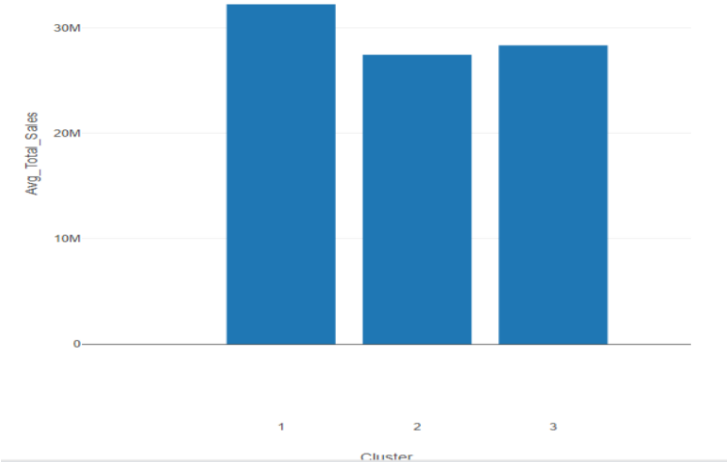


Cluster Information:

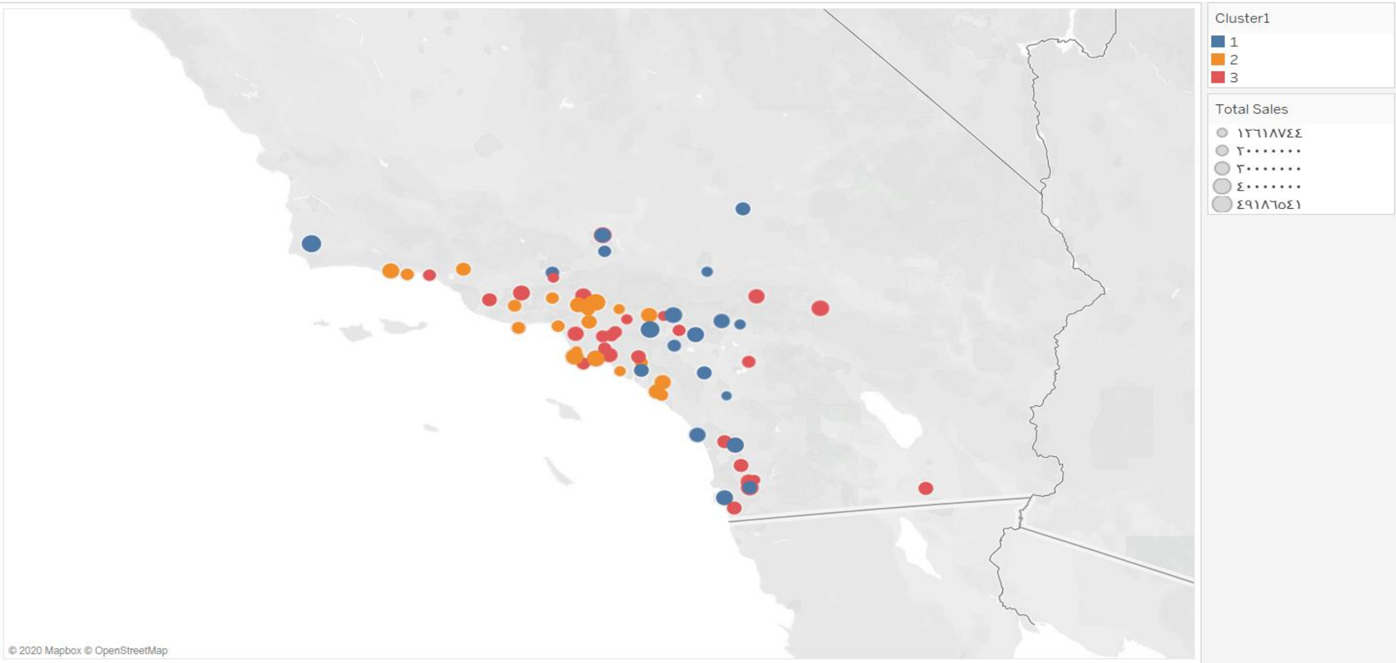
Cluster	Size
1	23
2	29
3	33

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

   Looking at the graph shown below, we can notice that Stores in Cluster 1 sell more on an average when compared to the stores in the other two clusters.



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.

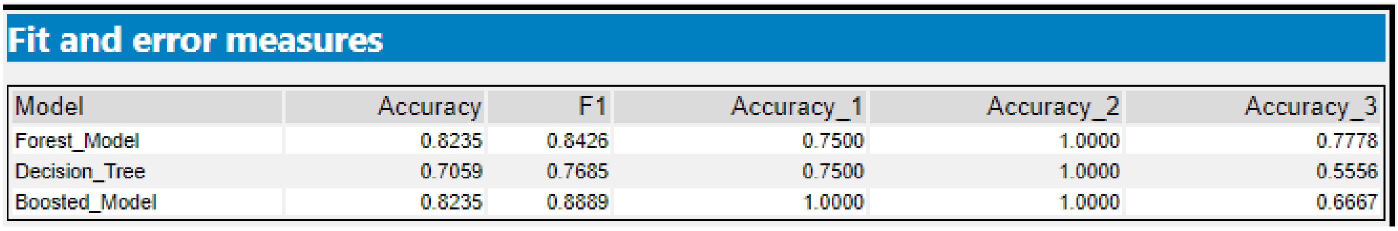


https://public.tableau.com/profile/adel.altuwaijri#!/vizhome/Storesclustersmap/store-locations

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.)

I used Decision tree, Forest, and boosted models to predict the most fit cluster for each new store and then compared the accuracy of these models using Model Comparison tool. According to resulted report, I found that accuracy and F1 measures is best with boosted model.



Fit and error measures

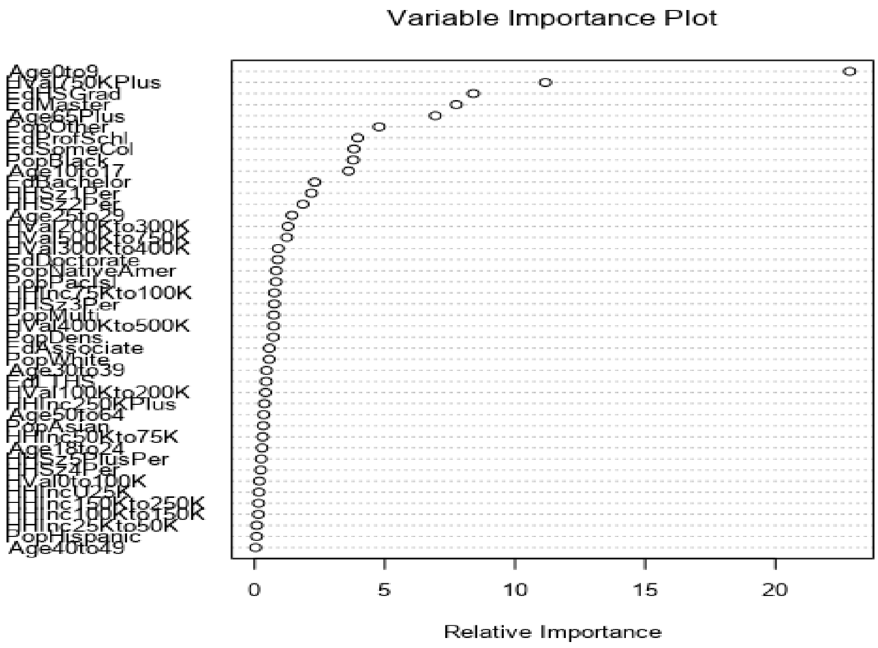
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

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

3. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

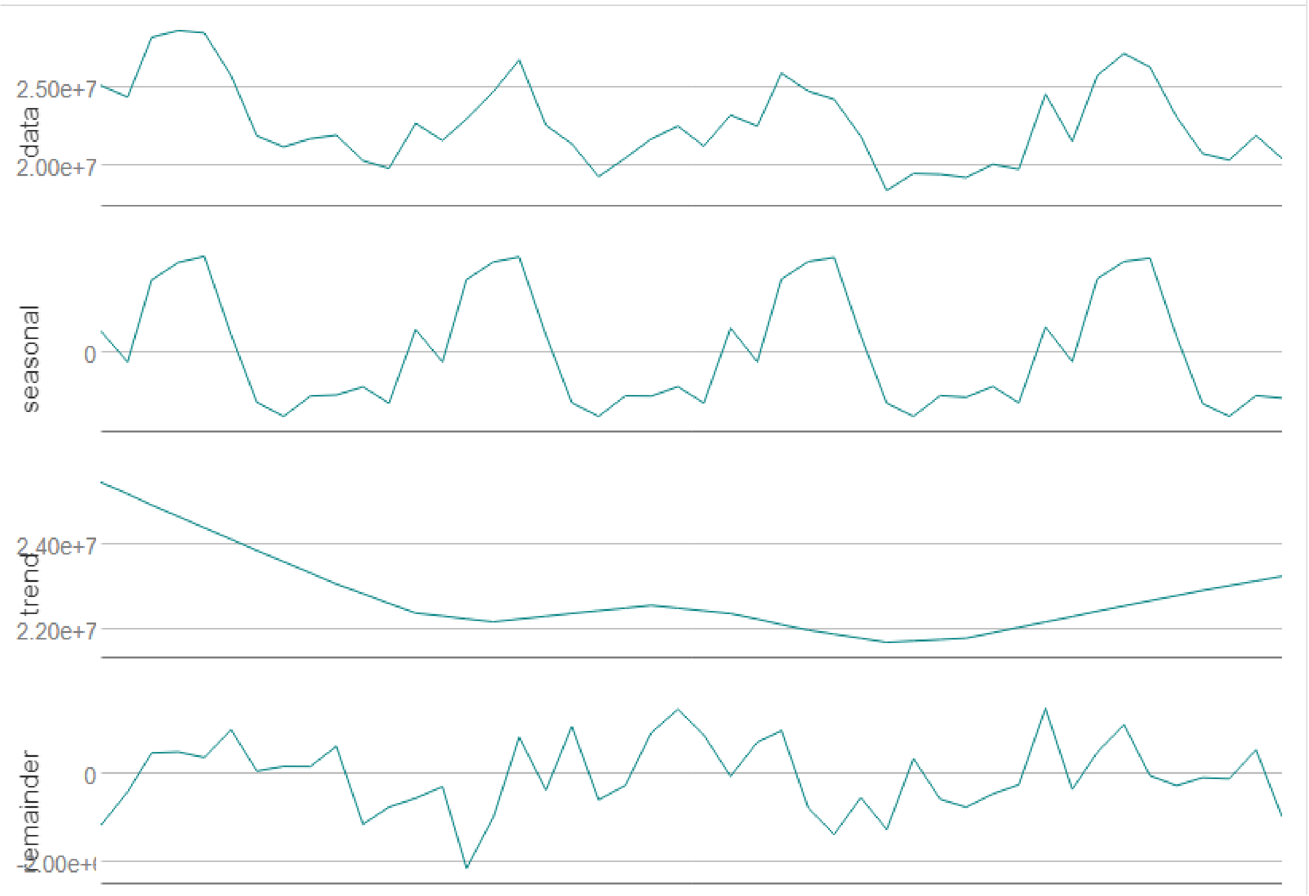
According to variable importance plot result from boosted model tool, the three most important variables are: **Age0to9**, **HVAL750kPlus**, and **EdHSGrad**.



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?

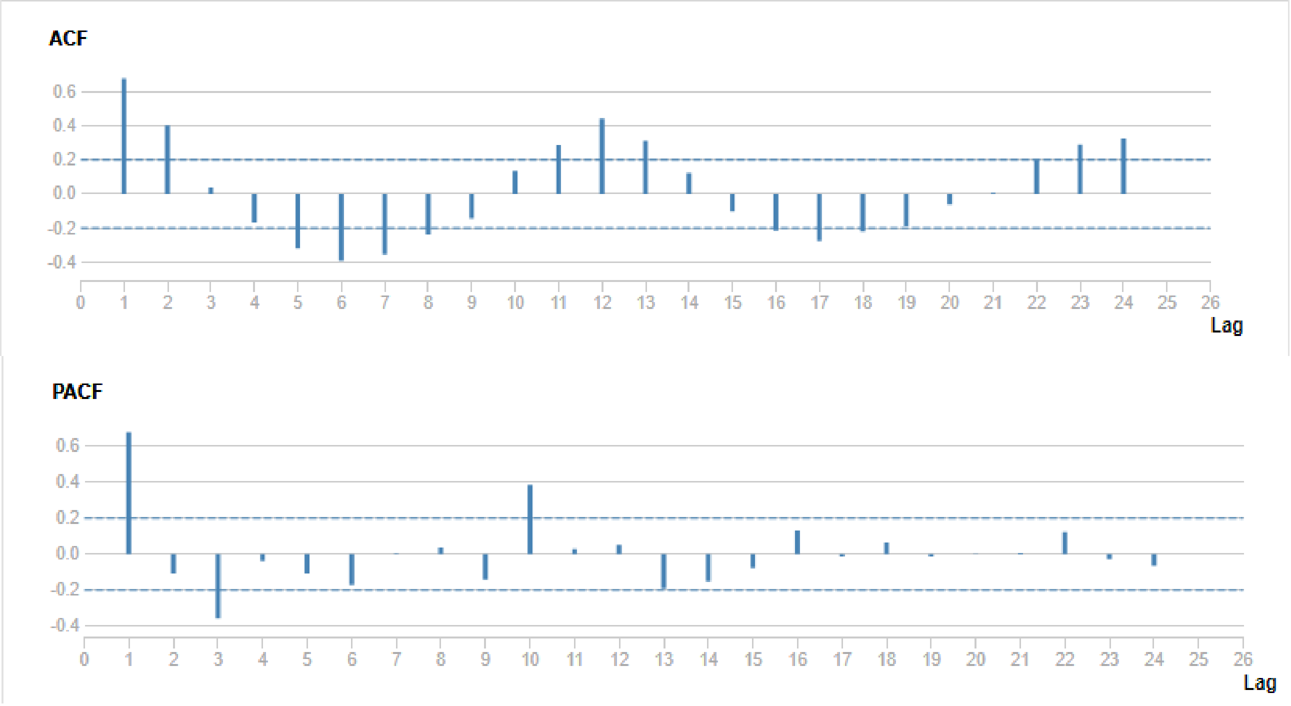
I have prepared a forecast with monthly granularity for product sales for 2016 for existing and new stores. To forecast sales for existing stores, I aggregate sales in all stores per month and produce a forecast. The time series is broken down into three time series, which is the seasonal component, the trend component and the rest. Below, I report the decomposition graph:



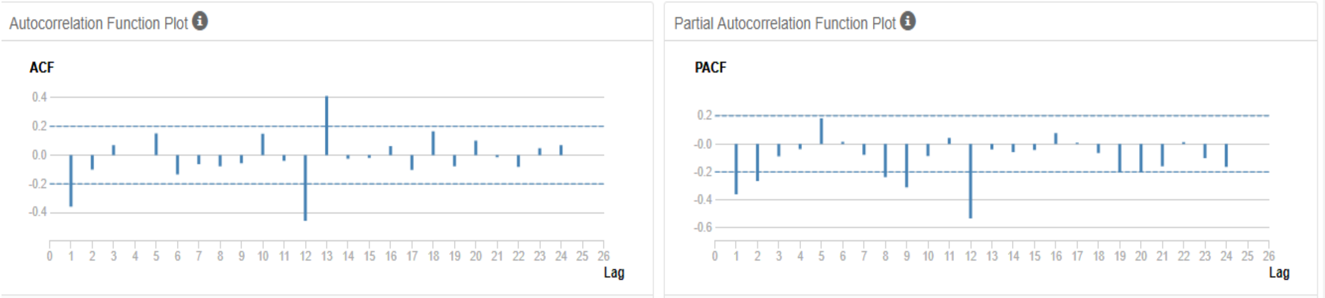
I built the ETS model, examining the seasonal component, trend component and rest component in the time series decomposition graph. Seasonality is growing slightly over time (the peaks are increasing very slowly), so i apply this multiplicatively, the series

there is no trend, the error is increasing or decreasing over time, so i apply the error multiplicatively, so i choose the ETS (M, N, M).

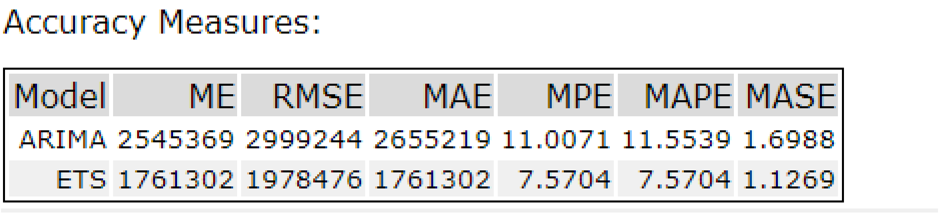
The ARIMA model requires that the series be stationary. Autocorrelation (ACF) or partial autocorrelation (PACF) charts help me determine whether autocorrelation exists:



I noted that the ACF shows an fluctuation, indicating a seasonal series, in the "lags" i can observe several seasonal periods. In the monthly data, i can observe that in the lags 12, 24, the peaks occur at intervals of 12 months and 24 months, in addition , i observed that a peak in delay 1 on an ACF graph indicates a strong correlation between each value in the series and the previous value, and then i adjust the series with the seasonal model ARIMA. Non-stationary series can be corrected by a transformation such as applying the first seasonal difference. By observing the ACF and PACF autocorrelation graphs of the first seasonal difference, i can identify the numbers of ARIMA terms needed



Observing the two negative peaks in FAC in lag 1, which indicates non-seasonal BF terms. For seasonal terms, i note that there is a negative peak at 12-month intervals. This indicates seasonal MA terms. So, the model that fits is ARIMA (0, 1, 1) (0, 1, 1) 12.

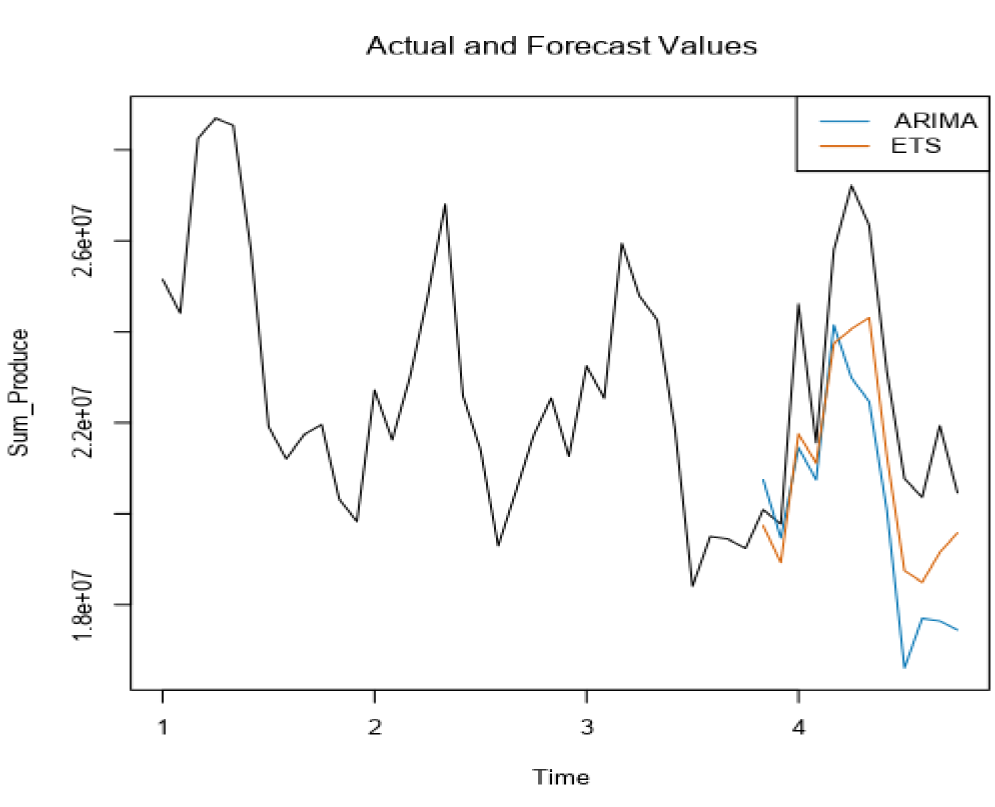


Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988
ETS	1761302	1978476	1761302	7.5704	7.5704	1.1269

From the values in the table, I concluded that the ETS model is better than the ARIMA model for this problem, given that the RMSE and MASE of the ETS model are inferior to the ARIMA model. For the ETS model, RMSE is 1983593 and MASE 1.2691 and for the ARIMA model, RMSE is 2999244 and MASE is 1.6988, here is the graph that shows all the time series values and forecast values for all the compared models.

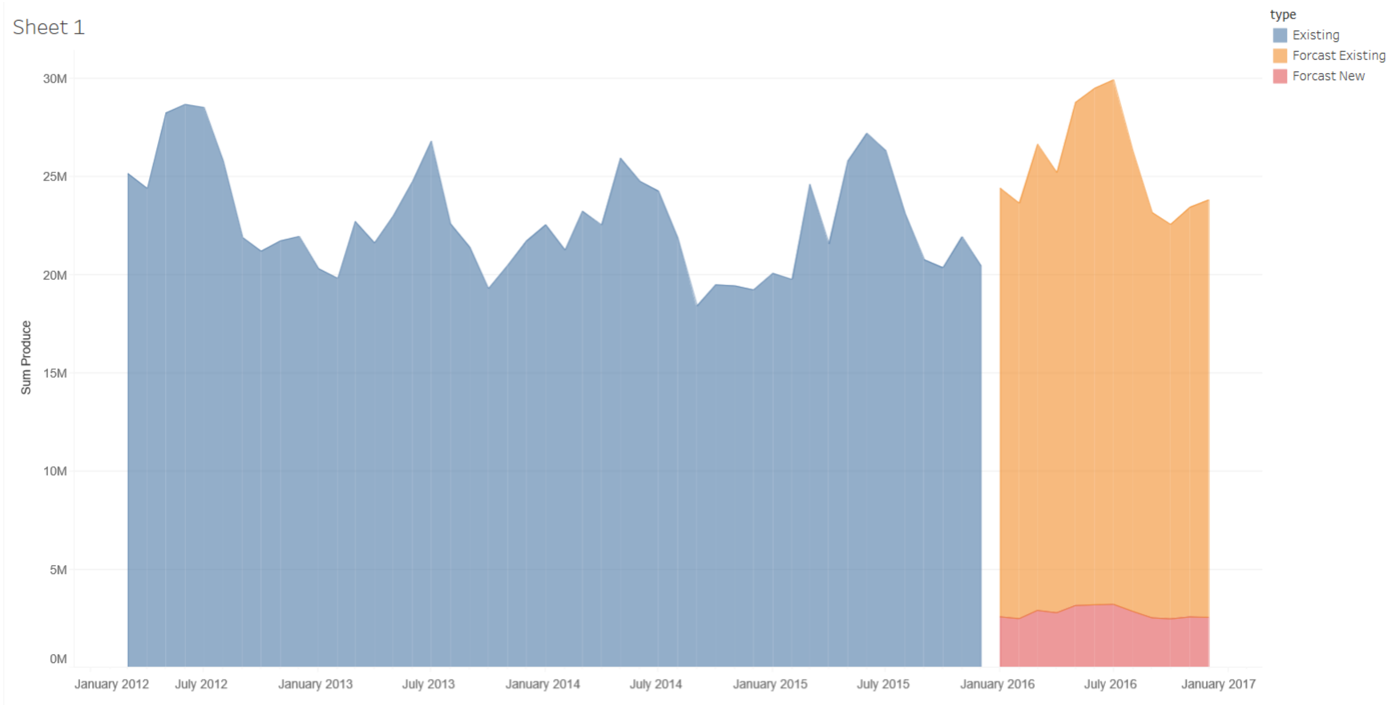
Looking at the graph below, it's obvious that ETS model behaves more accurately than the ARIMA model.



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.

Historical Sales				
Year	**Month**	**Sales**		
2012	03	25151526		
	04	24406048		
	05	28249539		
	06	28691364		
	07	28535707		
	08	25793521		
	09	21915642		
	10	21203563		
	11	21736159		
	12	21962977		
2013	01	20322684		
	02	19829621		
	03	22717070		
	04	21625385		
	05	23000152		
	06	24755406		
	07	26803106		
	08	22600217		
	09	21401266		
	10	19296578		
	11	20489773		
	12	21715707		
2014	01	22544458		
	02	21262413		
	03	23247169		

	04	22541988		
	05	25943047		
	06	24782178		
	07	24263118		
	08	21879989		
	09	18407264		
	10	19497572		
	11	19444753		
	12	19240385		
2015	01	20088529		
	02	19772333		
	03	24608407		
	04	21559729		
	05	25792075		
	06	27212464		
	07	26338477		
	08	23130627		
	09	20774416		
	10	20359981		
	11	21936907		
	12	20462899		
Forecasting Sales				
Year	**Month**		**Existing Stores**	**New Stores**
2016	1		21829060	2588357
	2		21146330	2498567
	3		23735687	2919067
	4		22409515	2797280
	5		25621829	3163765
	6		26307858	3202813
	7		26705093	3228212
	8		23440761	2868915
	9		20640047	2538372
	10		20086270	2485732
	11		20858120	2583448
	12		21255190	2562182



https://public.tableau.com/profile/adel.altuwaijri#!/vizhome/ProduceSalesAnalysis/Sheet1?publish=yes