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?

Ans:
To determine the Optimal no. of store formats a K- Centroid analysis was done using K – mean clustering method using minimum and maximum clusters as 2 to 8. The following are the K-means clustering analysis results.

K-Means Cluster Assessment Report

Summary Statistics

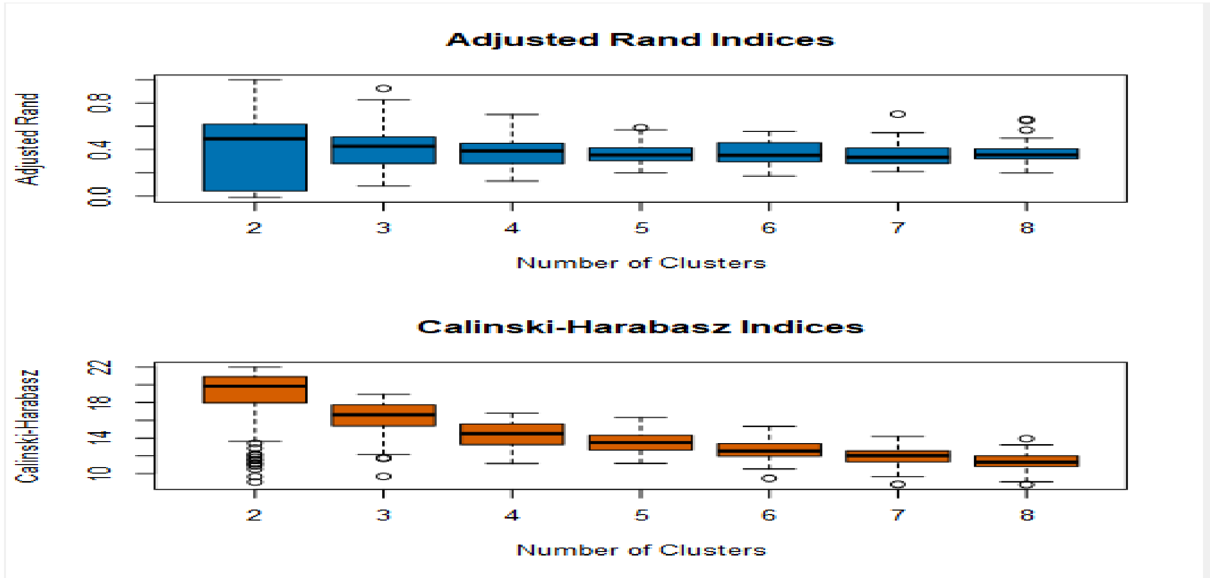
Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.012332	0.085005	0.129167	0.198479	0.172868	0.211424	0.197457
1st Quartile	0.055047	0.28273	0.279896	0.303745	0.294079	0.281472	0.321616
Median	0.492542	0.428163	0.388131	0.353296	0.351385	0.333331	0.353529
Mean	0.406457	0.411914	0.372189	0.366041	0.367644	0.354859	0.369188
3rd Quartile	0.61678	0.50506	0.450843	0.41474	0.453322	0.409187	0.404819
Maximum	1	0.925732	0.70085	0.586379	0.5548	0.703966	0.660004

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	9.056197	9.683921	11.14097	11.15269	9.474469	8.797239	8.769803
1st Quartile	17.976426	15.402516	13.27496	12.65426	11.988572	11.311079	10.838622
Median	19.836525	16.618434	14.49044	13.49543	12.537825	12.043325	11.303199
Mean	18.604945	16.309418	14.37112	13.46494	12.624375	11.910413	11.376818
3rd Quartile	20.889876	17.734502	15.56523	14.30924	13.365637	12.535052	11.963996
Maximum	21.992647	18.908142	16.79342	16.32568	15.329887	14.179165	13.936724

Plots



Basis the K means analysis, Adjusted Rand and Calinski- Harabasz Indices it is evident that the optimal no. of store format is 3 as K=3 has the tightest and compact Mean value.

2.How many stores fall into each store format?

Ans: Store formats	No. of Stores
1	23
2	29
3	33

Summary Report of the K-Means Clustering Solution Task1_Cluster_Analysis

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Per_Dry_Grocery + Per_Diary + Per_Frozen_Food + Per_Meat + Per_Produce + Per_Floral + Per_Deli + Per_Bakery + Per_Gen_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Per_Dry_Grocery	Per_Diary	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Per_Bakery	Per_Gen_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

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

Ans:

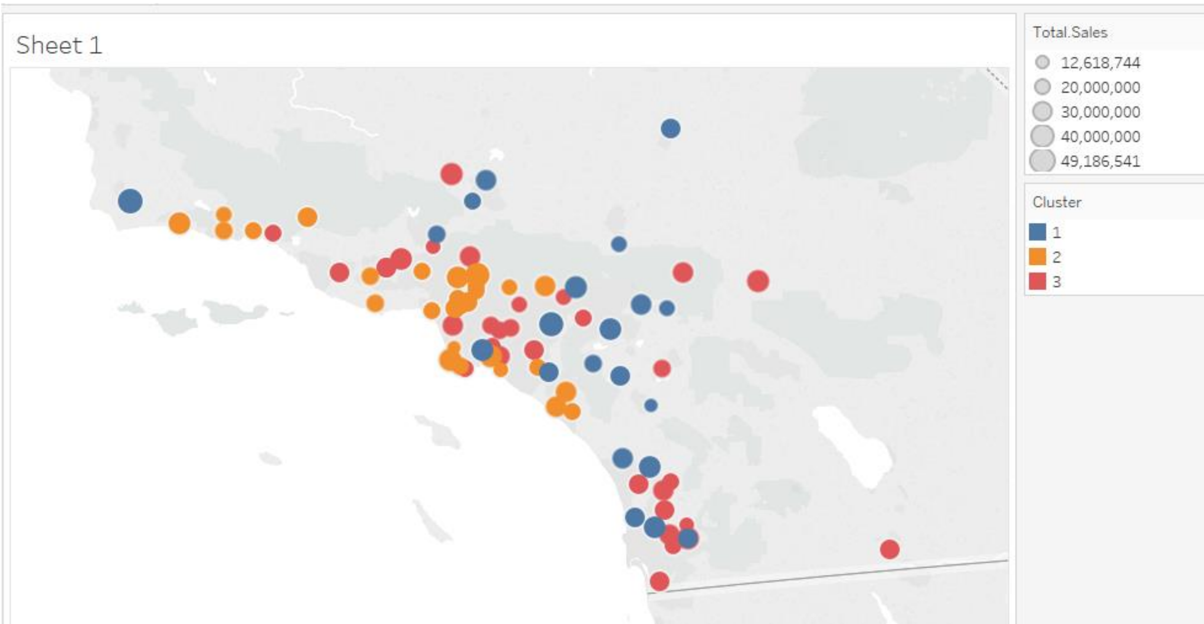
Cluster 1: It sold more of General Merchandise and have highest total sales compared to the other 2 clusters

Cluster 2: It has sold more produce than other 2

Cluster 3: These sores are more or less have the same range of sales.

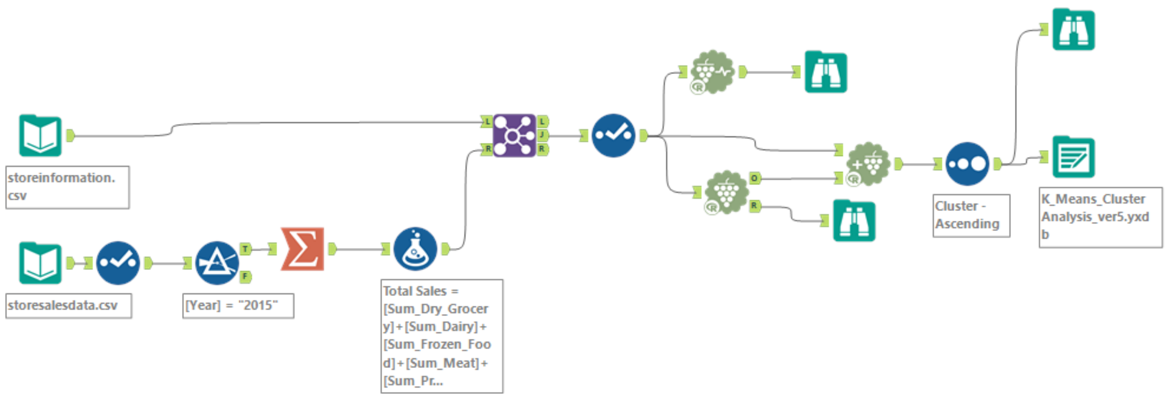
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

Ans:



https://public.tableau.com/profile/saibal.sinha#!/vizhome/Store_clusters/Sheet1?publish=yes

Alteryx Work Flow Task 1



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.)

Ans: The below figure shows the model comparison report of Decision Tree, Forest and Boosted Model. Boosted Model was chosen because of higher F1 and Accuracy values.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
ForestModel	0.8235	0.8426	0.7500	1.0000	0.7778
Decision_Tree_20	0.7059	0.7685	0.7500	1.0000	0.5556

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

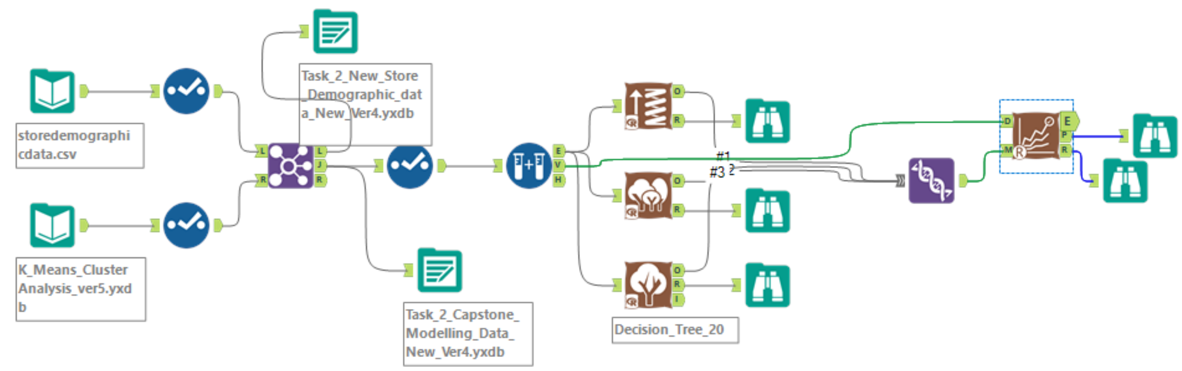
2. What format do each of the 10 new stores fall into? Please fill in the table below.

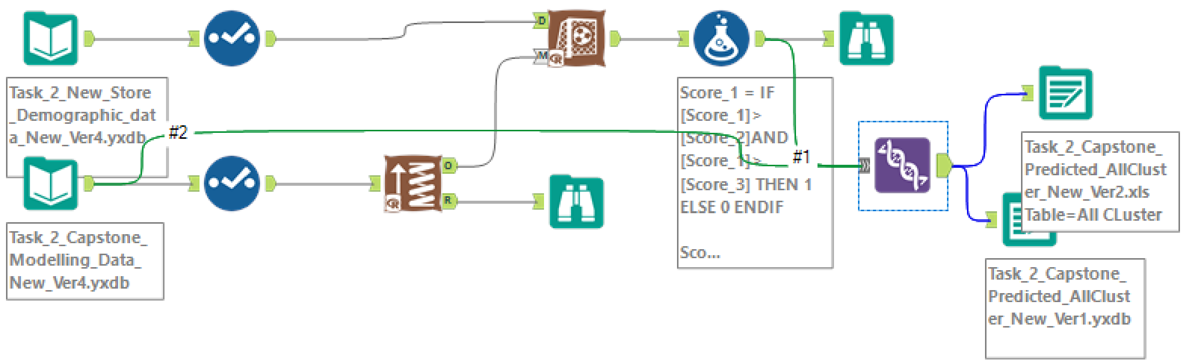
Ans:

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

Record #	Store	Score_1	Score_2	Score_3
1	S0086	1	0	0
2	S0087	0	2	0
3	S0088	0	0	3
4	S0089	0	2	0
5	S0090	0	2	0
6	S0091	1	0	0
7	S0092	0	2	0
8	S0093	1	0	0
9	S0094	0	2	0
10	S0095	0	2	0

TASK 2- Alteryx Work Flow



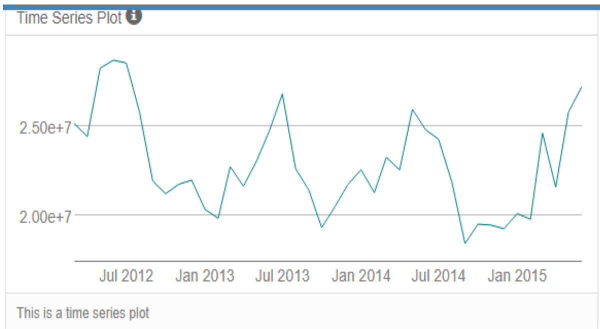


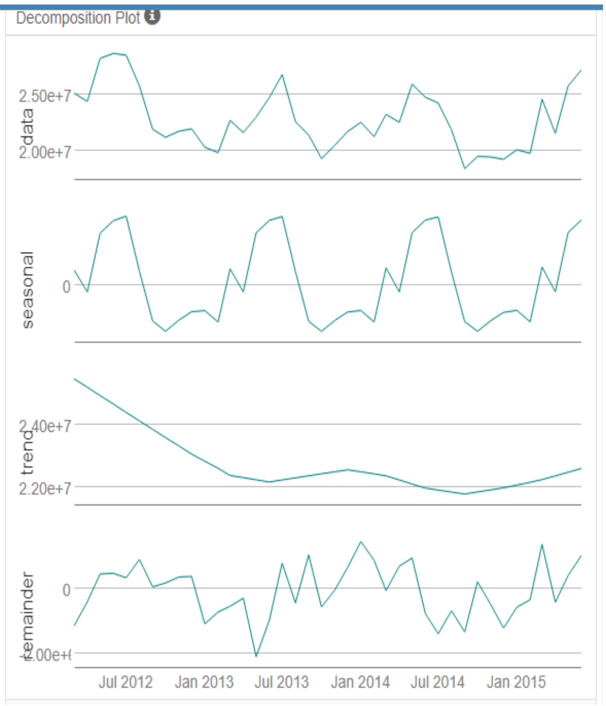
Task 3: Predicting Produce Sales

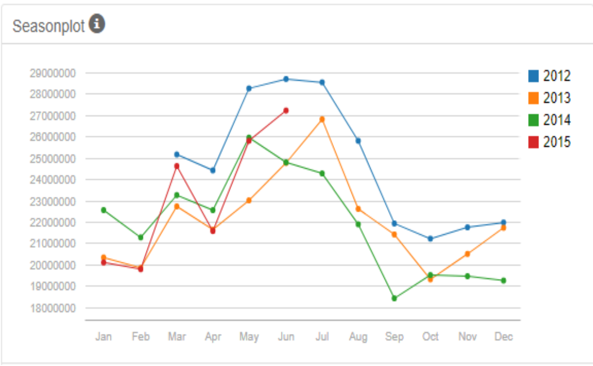
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

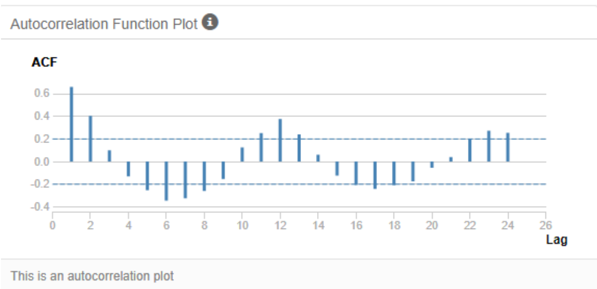
Ans:

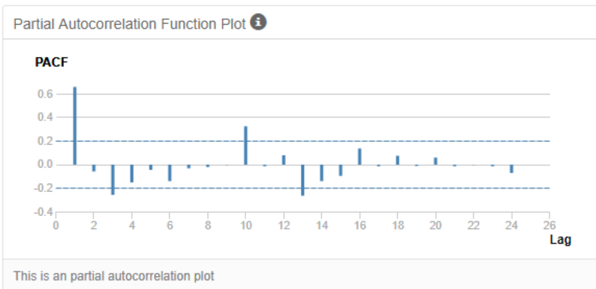
ETS (M,N,M) With Auto Trend was used since the trend was not quite clear. Looking at the decomposition plot it is found that there is a seasonality and multiplicative was applied. The error seems to be not regular and so Multiplicative was used.

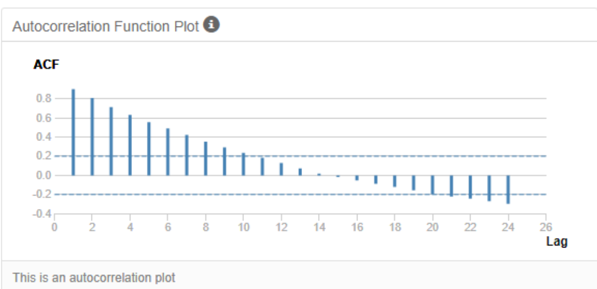


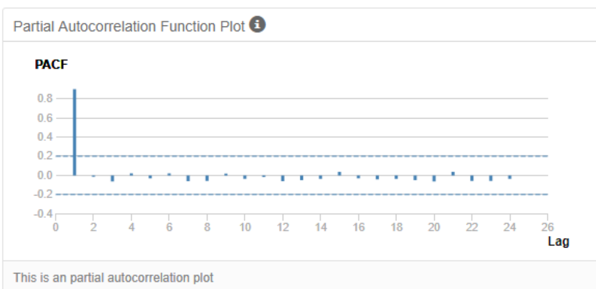


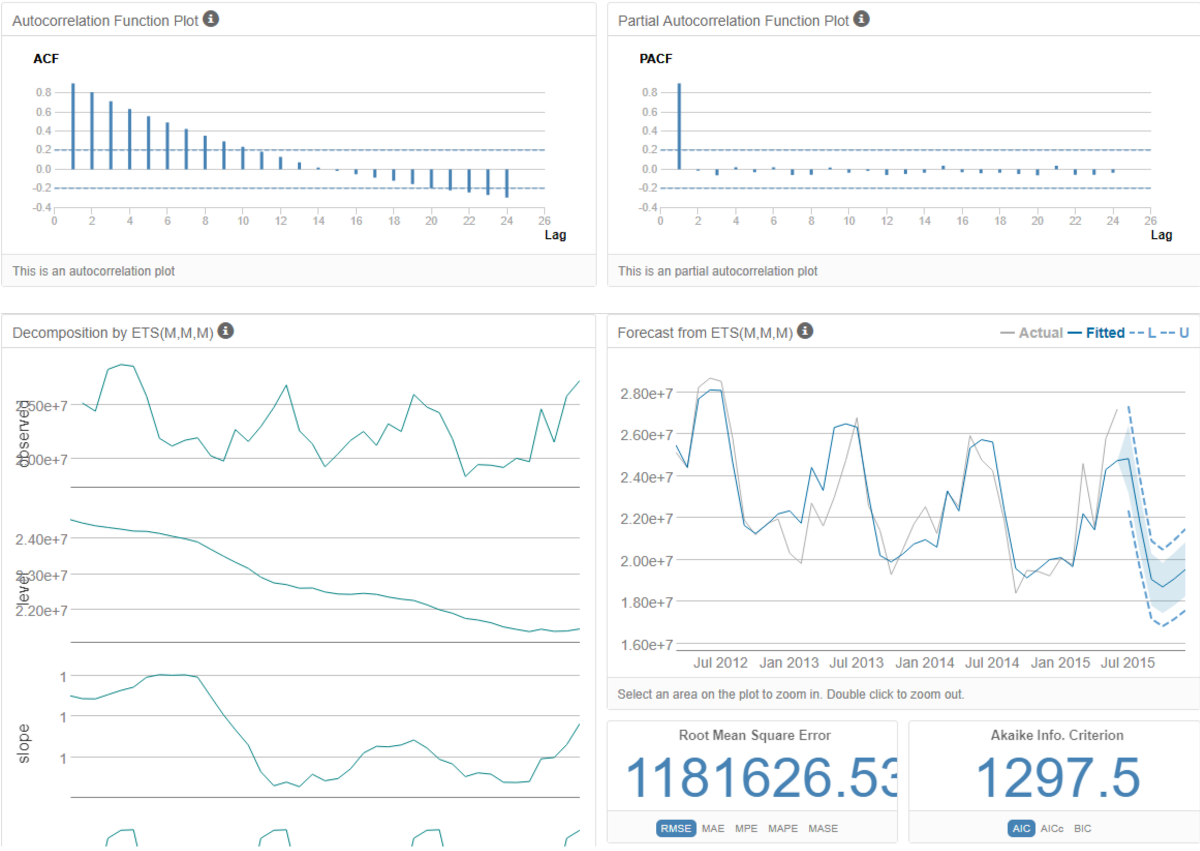




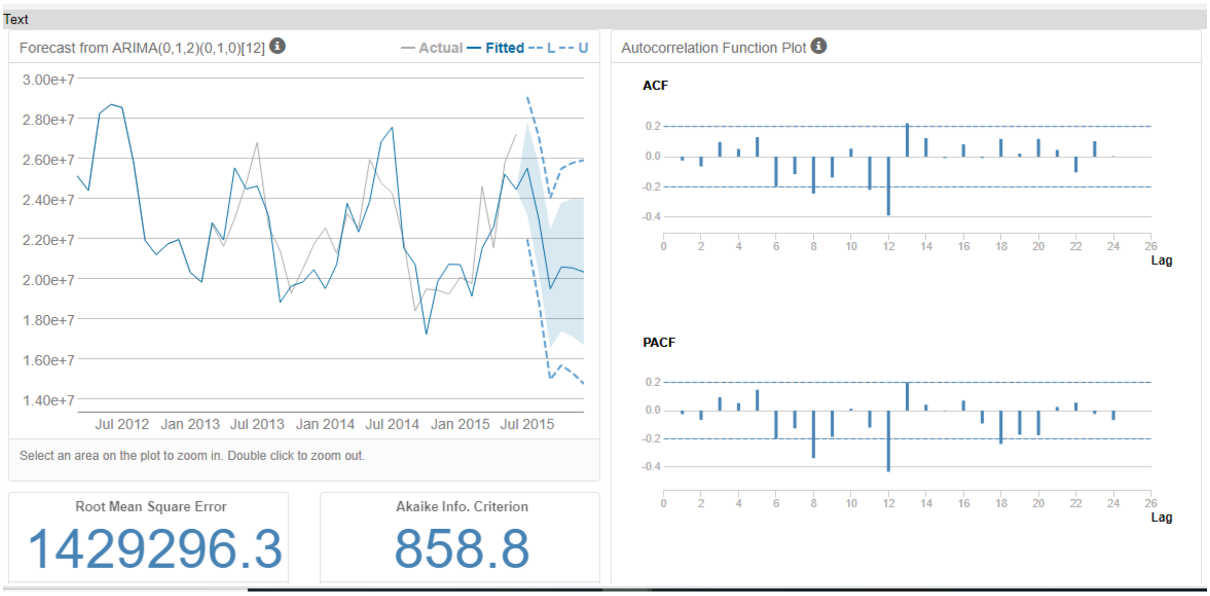








ARIMA (0,1,2) (0,1,0) was performed with seasonal difference and Seasonal first difference as there was a Lag 2.



Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Summary of ARIMA Model ARIMA

Method: ARIMA(0,1,2)(0,1,0)[12]

Call:
Arima(Sum_Produce, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 0), period = 12), include.drift = TRUE)

Coefficients:

	ma1	ma2
Value	-0.415471	-0.054116
Std Err	0.219958	0.234438

sigma^2 estimated as 3268620653560.66: log likelihood = -426.38872

Information Criteria:

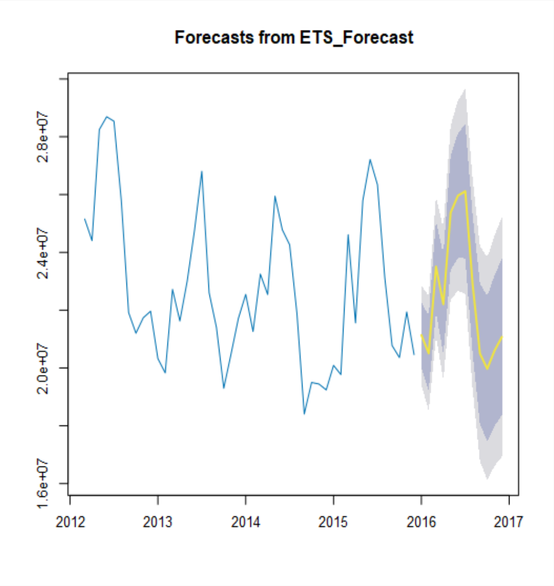
AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

The Model comparison suggests that The model accuracy of ETS is higher than that of ARIMA . The results after doing a holdout of sample of 6 months suggests the RMSE of ETS is 1020596 and that of ARIMA is 1429296. The MASE fig of ETS is 0.45 and that of ARIMA is 0.53.

12 Period Forecast from ETS_Forecast



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.

Ans: The following are the calculated table forecast for existing stores and new stores. The New store sales was performed after using ETS (M, N, M) analysis on the three different clusters. The average sales value was found after multiplying with the no. of new stores (Cluster1 – 3, Cluster 2 – 6, and Cluster 3 – 1) and adding them up for Produce New Store Sales.

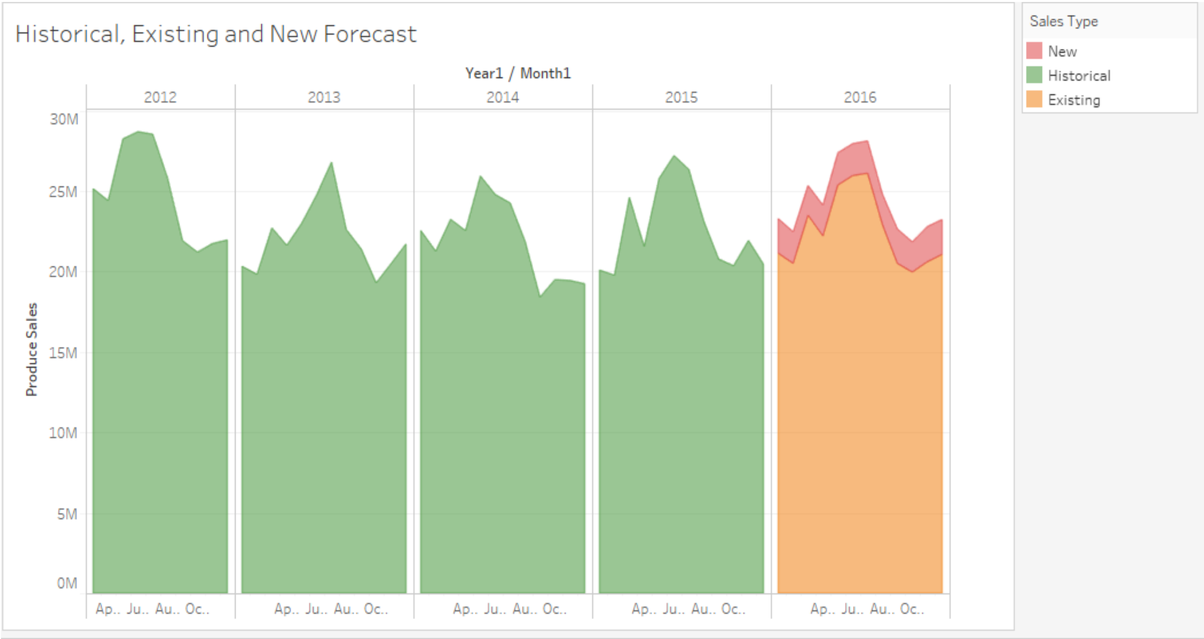
Year	Month	New Stores Sales	Existing Store Sales
2016	1	2150482	21136208
2016	2	1964811	20506605
2016	3	1845923	23506131
2016	4	1919849	22207971
2016	5	2007363	25376698
2016	6	1986007	25963559
2016	7	2007444	26113357
2016	8	1895528	22904672
2016	9	2146810	20499151
2016	10	1870057	19970809
2016	11	2187826	20602232
2016	12	2162053	21072787

Existing Store Forecast

Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2016	1	21136208.135109	22863751.647268	22265788.122301	20006628.147918	19408664.622951
2	2016	2	20506604.689889	22485979.825084	21800848.524632	19212360.855146	18527229.554694
3	2016	3	23506131.457397	25923604.543644	25086832.145154	21925430.769639	21088658.371149
4	2016	4	22207971.238436	24819551.269971	23915591.635728	20500350.841144	19596391.206902
5	2016	5	25376698.322185	28385663.710055	27344155.037671	23409241.606699	22367732.934316
6	2016	6	25963559.446576	29258459.785154	28117978.976999	23809139.916154	22668659.107998
7	2016	7	26113357.20163	29660962.648063	28433011.720628	23793702.682632	22565751.755197
8	2016	8	22904671.917667	26542287.656104	25283181.003148	20526162.832187	19267056.179231
9	2016	9	20499151.00121	24219766.868399	22931930.9538	18066371.048621	16778535.134021
10	2016	10	19970808.947309	23811395.340529	22482033.410444	17459584.484174	16130222.554089
11	2016	11	20602232.29737	24592072.351437	23211048.483736	17993416.111005	16612392.243304
12	2016	12	21072786.922156	25209451.080778	23777606.230282	18367967.61403	16936122.763534

New Store Forecast

Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	238942.418301	307662.080182	283875.790017	194009.046585	170222.75642
2	218312.34857	284888.773788	261844.333157	174780.363982	151735.923351
3	205102.601588	271027.769102	248208.751651	161996.451524	139177.434074
4	213316.546007	285231.498662	260339.20606	166293.885955	141401.593353
5	223040.299565	301588.265474	274400.053864	171680.545266	144492.333656
6	220667.451165	301571.05324	273567.473163	167767.429167	139763.84909
7	223049.355731	255795.782561	244461.093345	201637.618118	190302.928902
8	210614.259711	248101.551833	235125.882519	186102.636903	173126.96759
9	238534.469632	287322.628688	270435.332154	206633.607111	189746.310577
10	207784.112664	255172.344591	238769.612037	176798.613291	160395.880737
11	243091.811563	303728.489431	282740.004114	203443.619012	182455.133694
12	240228.146936	304890.702569	282508.719465	197947.574406	175565.591303



https://public.tableau.com/profile/saibal.sinha#!/vizhome/HistoricalExistingNewForecast-Capstone/Sheet1?publish=yes

TASK 3 – Alteryx Workflow