Q1. What is the optimal number of store formats? How did you arrive at that number?

Answer: The optimal number of store formats is 3. It was arrived by running the data through the clustering model “K-means” as required by the project

1. Rand Indices

Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.0152	0.2276	0.2198	0.2392	0.2903	0.2399	0.2674
1st Quartile	0.4196	0.5498	0.4171	0.3733	0.3714	0.3754	0.3784
Median	0.562	0.7083	0.5162	0.4366	0.4184	0.4288	0.4228
Mean	0.533	0.678	0.5246	0.4563	0.4341	0.4254	0.4398
3rd Quartile	0.7656	0.8173	0.6249	0.5156	0.4768	0.4774	0.5136
Maximum	1	0.9583	0.8277	0.8215	0.7202	0.6221	0.6977
	9	10					
Minimum	0.2232	0.2398					
1st Quartile	0.3626	0.3412					
Median	0.4117	0.3743					
Mean	0.41	0.3984					
3rd Quartile	0.4501	0.4736					
Maximum	0.6294	0.5636					

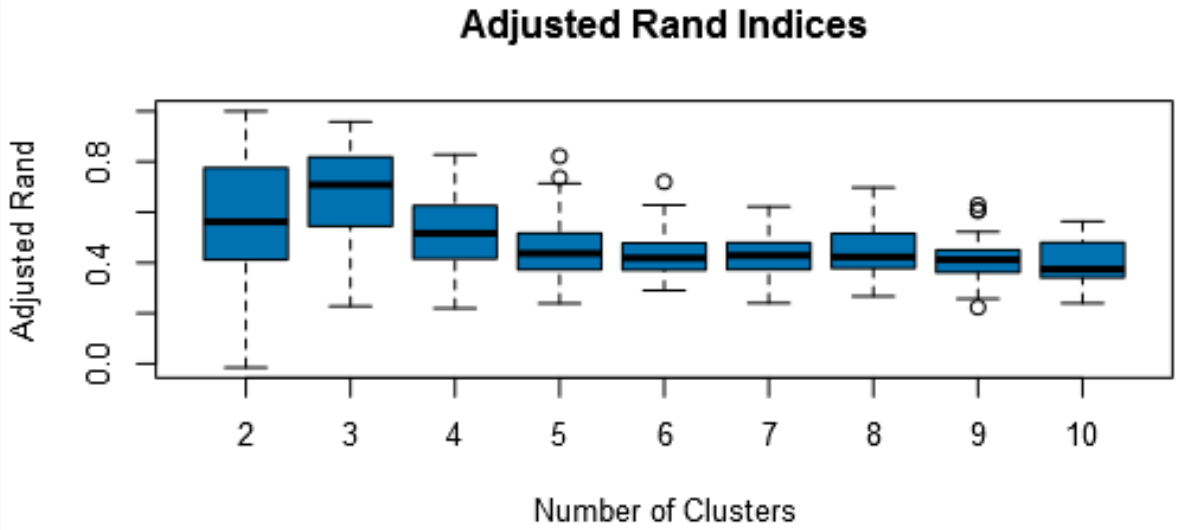
2. Calinski-Harabasz Indices

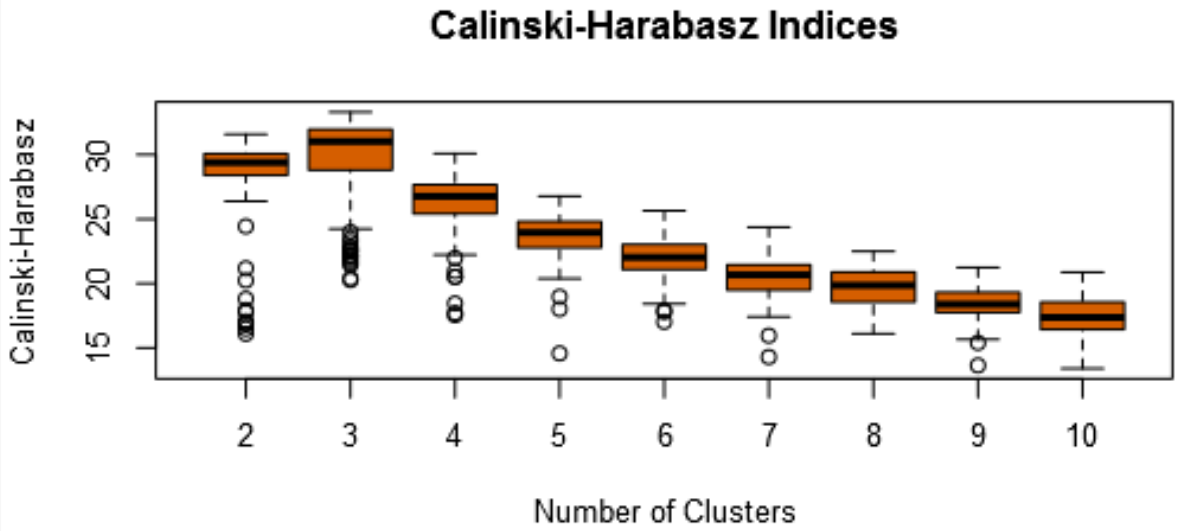
Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.27	17.55	14.58	17.03	14.3	16.11
1st Quartile	28.42	28.82	25.46	22.79	21.1	19.5	18.57
Median	29.39	31.02	26.77	23.95	22.02	20.66	19.85
Mean	28.21	29.65	26.21	23.67	21.98	20.48	19.72
3rd Quartile	30.07	31.96	27.68	24.8	23.01	21.45	20.89
Maximum	31.58	33.31	30.09	26.78	25.65	24.37	22.5
	9	10					
Minimum	13.64	13.4					
1st Quartile	17.78	16.43					
Median	18.39	17.36					
Mean	18.47	17.46					
3rd Quartile	19.31	18.55					
Maximum	21.24	20.87					

K-means plots

Plots





Based on the above images, we can also conclude the median value is the highest under the column of 3 clusters (in both indices). Thus, we can conclude, the 3 would be the optimal number of clusters. And in this case, it would be the optimal number of store formats.

Q2. How many stores fall into each store format?

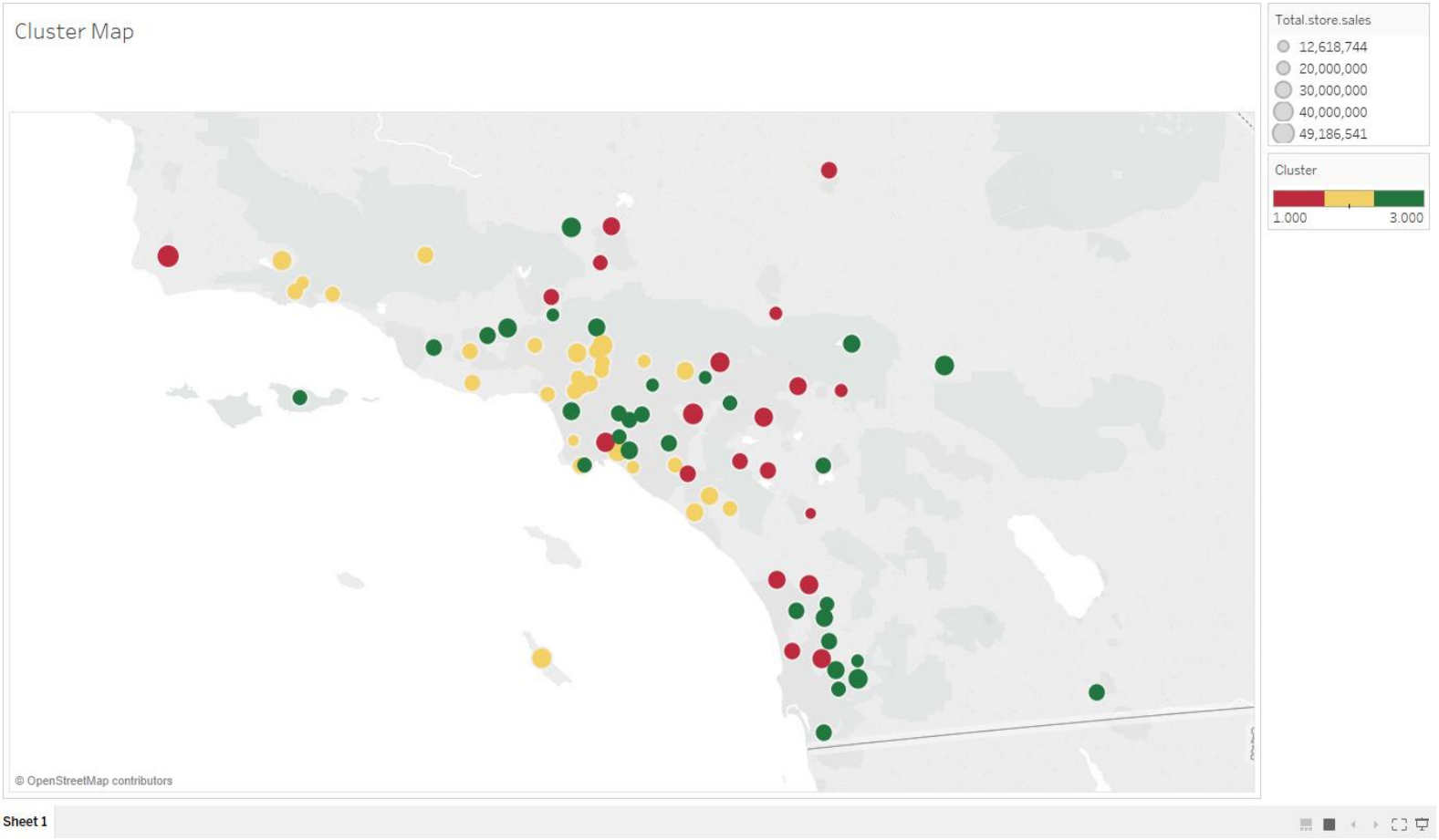
Clusters 1,2,3:

Record #	Store	Cluster
1	S0012	1
2	S0013	1
3	S0014	1
4	S0015	1
5	S0019	1
6	S0029	1
7	S0031	1
8	S0032	1
9	S0036	1
10	S0037	1
11	S0040	1
12	S0045	1
13	S0047	1
14	S0048	1
15	S0055	1
16	S0059	1
17	S0062	1
18	S0066	1
19	S0070	1
20	S0071	1
21	S0072	1
22	S0079	1
23	S0080	1

Record #	Store	Cluster
1	S0003	2
2	S0005	2
3	S0008	2
4	S0010	2
5	S0017	2
6	S0021	2
7	S0023	2
8	S0026	2
9	S0027	2
10	S0034	2
11	S0043	2
12	S0046	2
13	S0049	2
14	S0050	2
15	S0051	2
16	S0052	2
17	S0056	2
18	S0063	2
19	S0064	2
20	S0065	2
21	S0068	2
22	S0069	2
23	S0073	2
24	S0076	2
25	S0077	2
26	S0078	2
27	S0081	2
28	S0083	2
29	S0085	2

Record #	Store	Cluster
1	S0001	3
2	S0002	3
3	S0004	3
4	S0006	3
5	S0007	3
6	S0009	3
7	S0011	3
8	S0016	3
9	S0018	3
10	S0020	3
11	S0022	3
12	S0024	3
13	S0025	3
14	S0028	3
15	S0030	3
16	S0033	3
17	S0035	3
18	S0038	3
19	S0039	3
20	S0041	3
21	S0042	3
22	S0044	3
23	S0053	3
24	S0054	3
25	S0057	3
26	S0058	3
27	S0060	3
28	S0061	3
29	S0067	3
30	S0074	3
31	S0075	3
32	S0082	3
33	S0084	3

Here is how they look on a Map:



Q3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

When we analyse the results of the clustering model in Alteryx, we can observe that

For Cluster 1:

Sum_Total.store.sales	Median_Total.store.sales
741838364	33606658

For Cluster 2:

Sum_Total.store.sales	Median_Total.store.sales
796715972	26550185

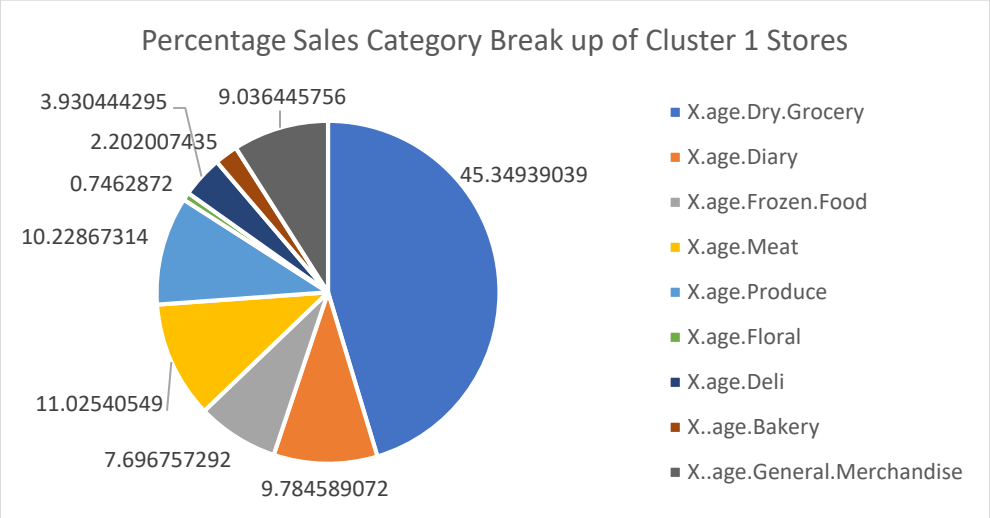
For Cluster 3:

Sum_Total.store.sales	Median_Total.store.sales
935779515	28489432

We can observe the difference in sales performances of the 3 clusters through the 'Median_Total.store.sales' value. Stores in cluster 1 clearly sell a lot more on average when compared to the other 2 clusters, whereas stores in cluster 2 have the least sales on average of all 3 clusters.

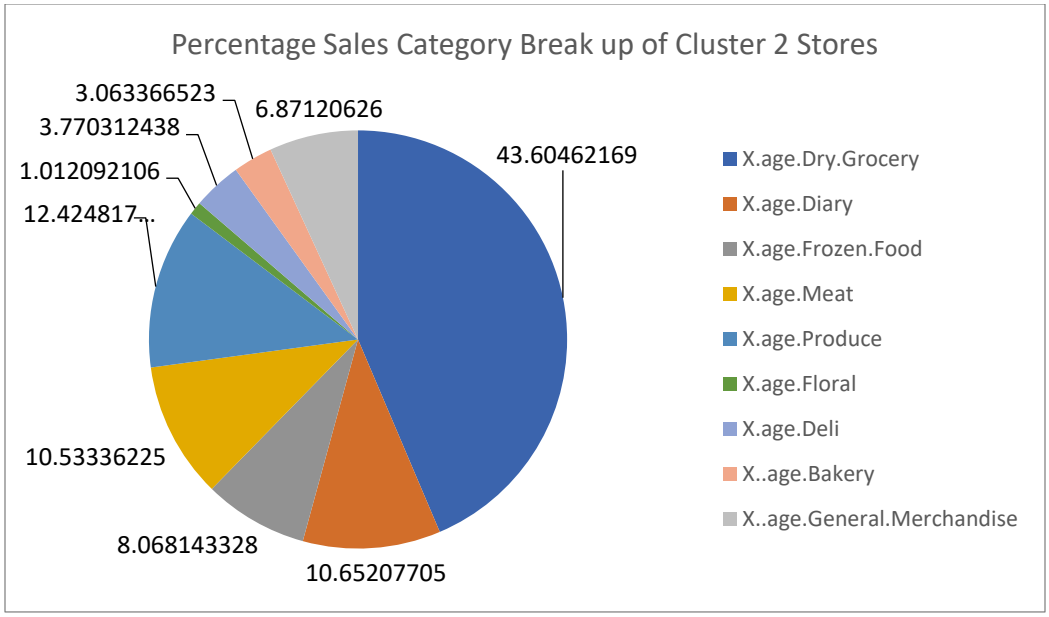
We can also further analyse the clustering data through Excel to understand how exactly does the sales break of various items vary between the 3 clusters.

Cluster 1:



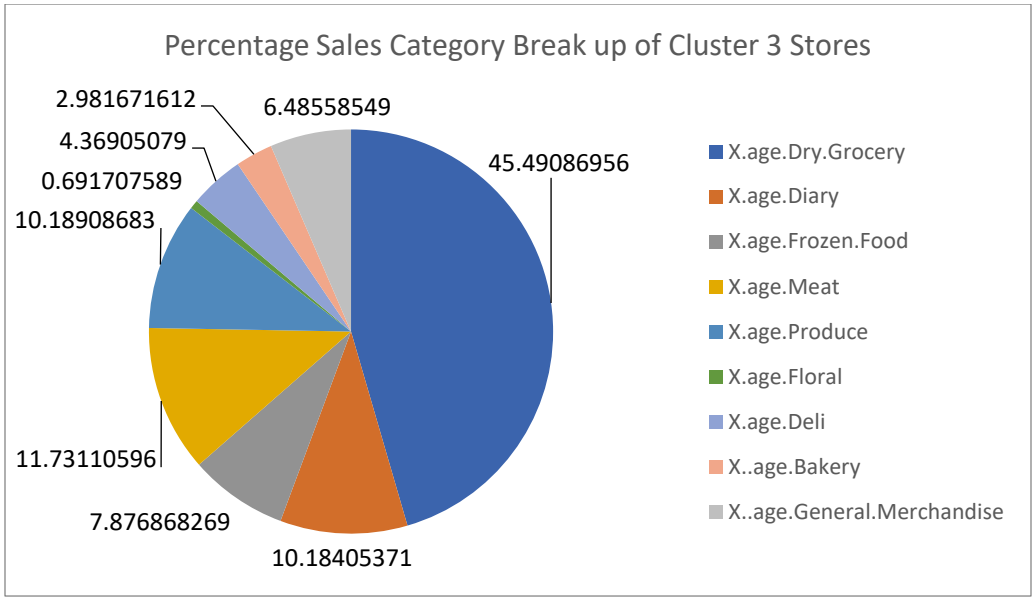
Note: Values reflected on each of the portions above is the average of that category

Cluster 2:



Note: Values reflected on each of the portions above is the average of that category

Cluster 3:



Note: Values reflected on each of the portions above is the average of that category

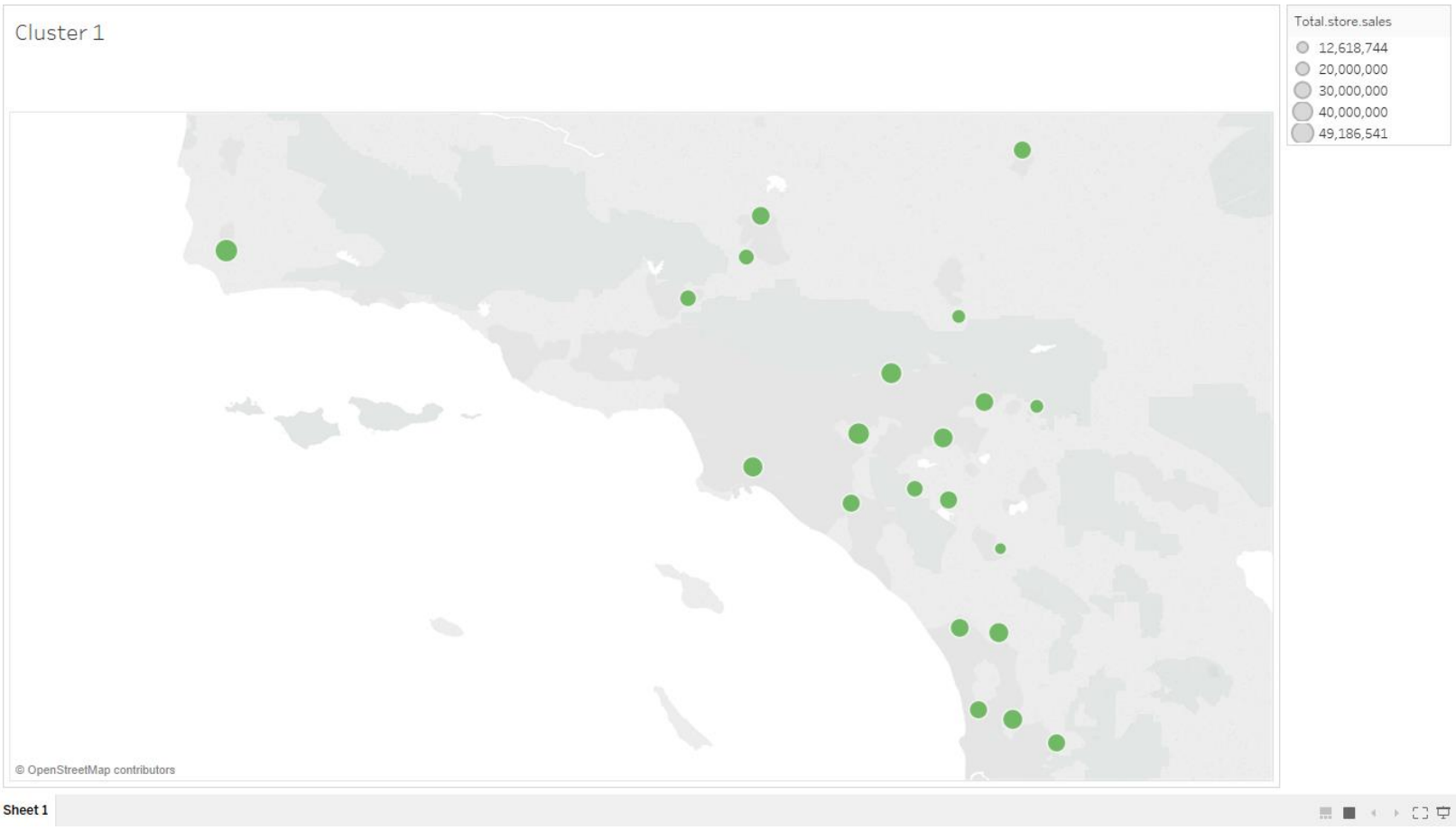
So, with the above pie charts, we can also observe, that on an average, stores in cluster 1, sell a lot more of General Merchandise than the stores of the other 2 clusters.

We can also observe that stores in cluster 2 sell more of 'produce' and stores in the other 2 clusters sell more of "Dry Grocery".

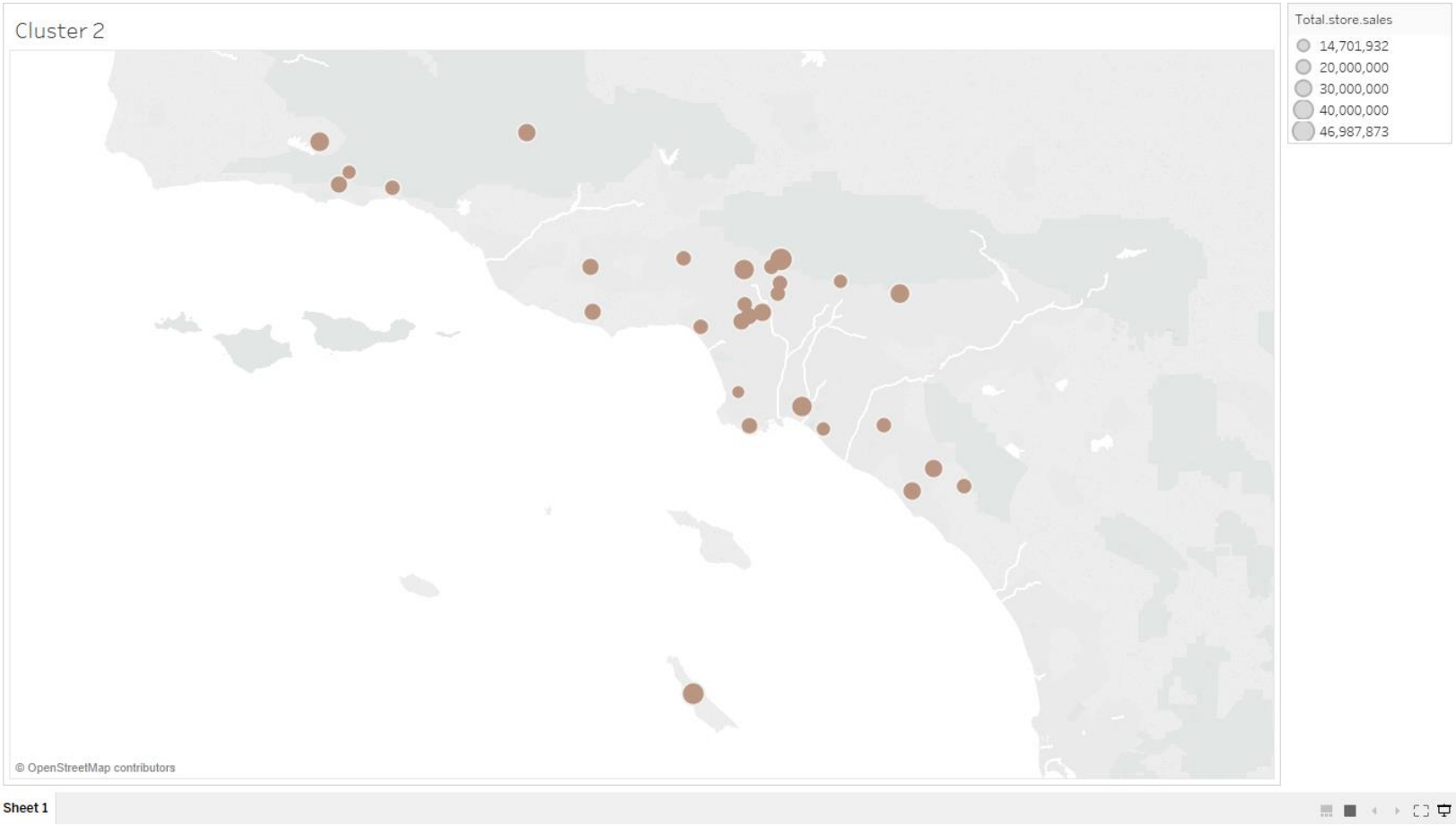
Q4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.

Answer: I've already provided a map of clusters above in this report, please also find maps of each individual cluster below:

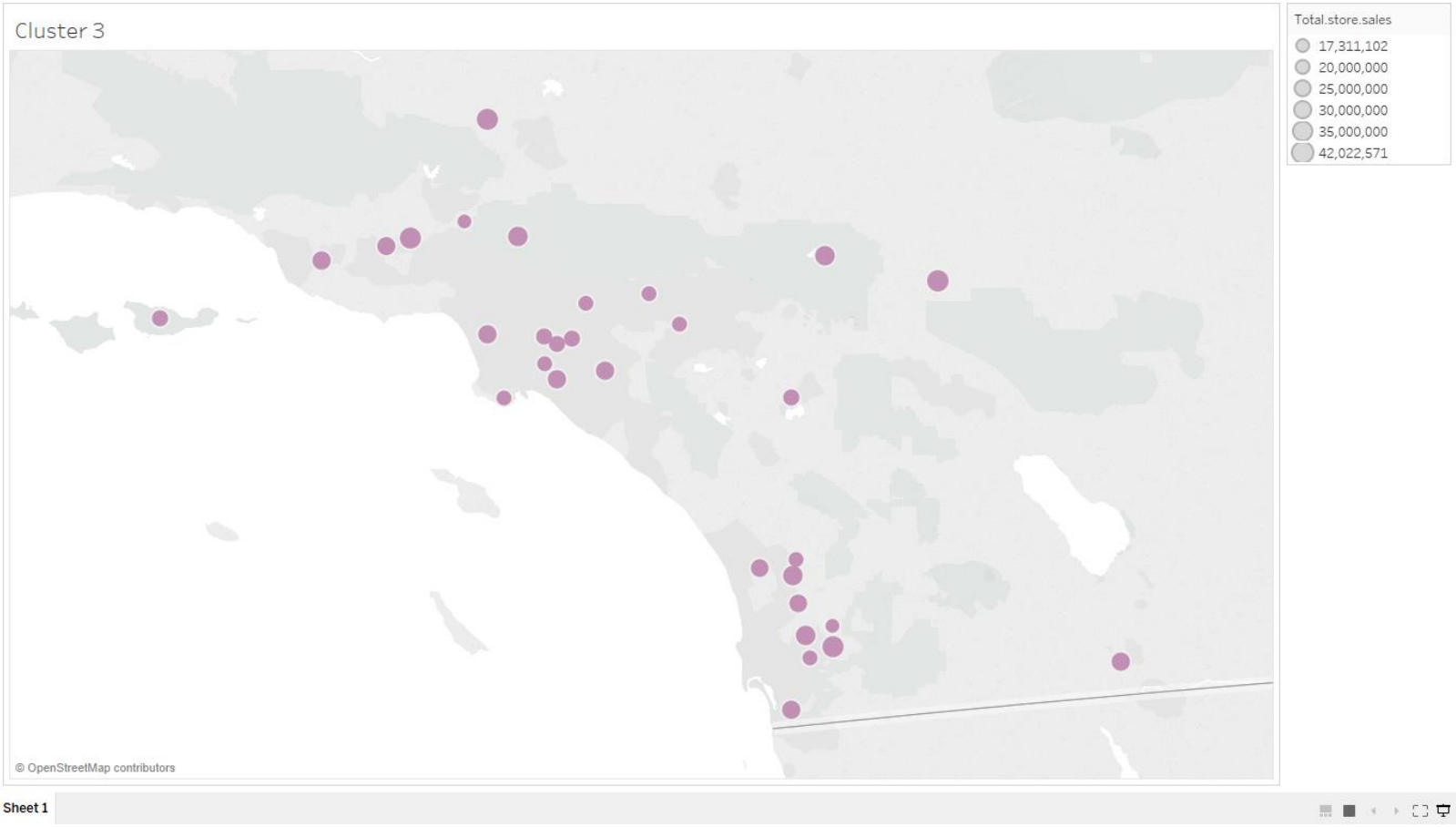
Cluster 1:


Cluster 1
Total.store.sales
12,618,744
20,000,000
30,000,000
40,000,000
49,186,541
© OpenStreetMap contributors
Sheet 1

Cluster 2:


Cluster 2
Total.store.sales
14,701,932
20,000,000
30,000,000
40,000,000
46,987,873
© OpenStreetMap contributors
Sheet 1

Cluster 3:


Cluster 3
Total.store.sales
17,311,102
20,000,000
25,000,000
30,000,000
35,000,000
42,022,571
© OpenStreetMap contributors
Sheet 1

Q1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?

I used three methods to derive an answer: Decision Tree, Boosted Model and Forest Model. Their Results/reports are as follows:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750
Decision_tree	0.7647	0.7810	0.7500	0.6667	0.8571

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of Decision_tree

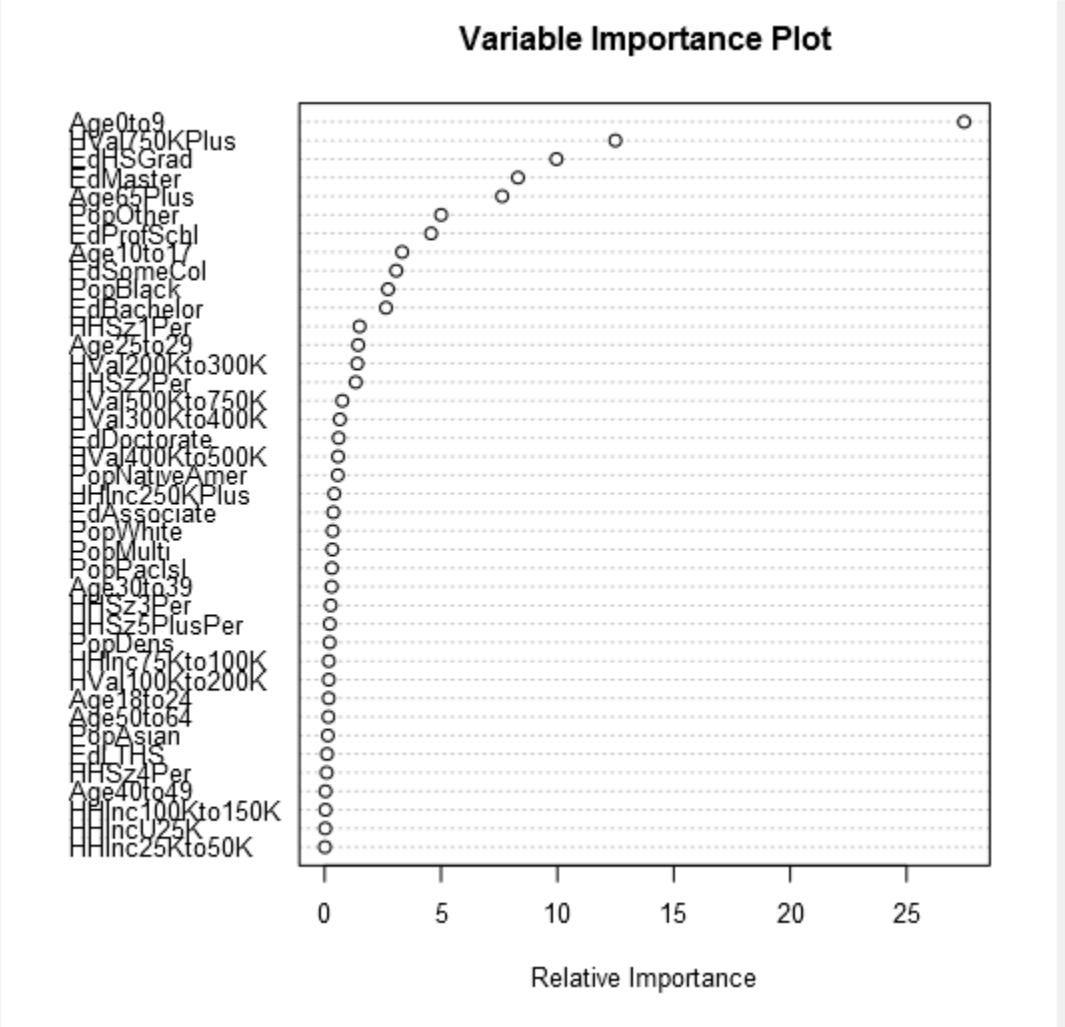
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	2
Predicted_3	1	0	6

Confusion matrix of Forest_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Based on the above reports, we can observe that the overall accuracy of two of the models (Forest and Boosted Model) are exactly the same. This is not uncommon when the datasets are small as in this case. Therefore, to pick the best model, I chose the F1 score which is the precision + recall score. Based on this score, I conclude that the Boosted Model is the best fit in this case. To summarize the Boosted Model Overall accuracy – 82.35%, F1 score: 85.43%

Q2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.



Based on the above visualization, the 3 most important demographic variables are:

1. Age0to9
2. HVal750KPlus
3. EdHSGrad

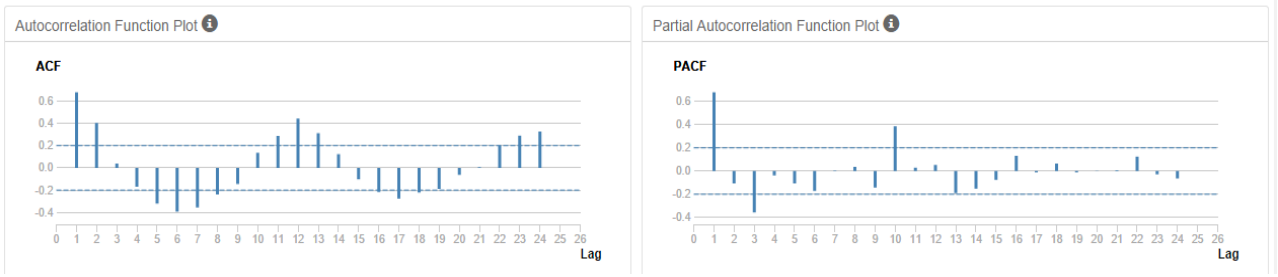
Q3. What format do each of the 10 new stores fall into? Please provide a data table.

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

TASK-3

Q1. To forecast sales for existing stores you should aggregate sales across all stores by month and produce a forecast.

To Forecast sales for existing stores, I first plotted the data with the TS-Plot tool

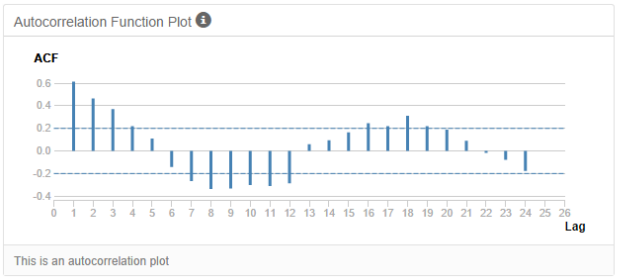


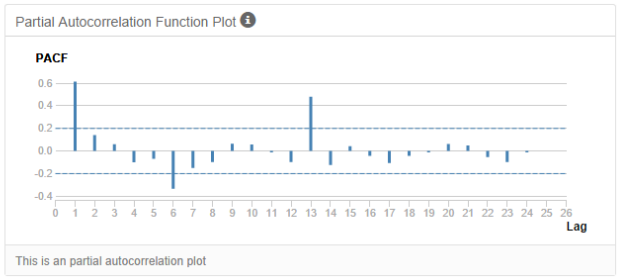
From the Decomposition Plot, we can observe:

1. The sales fluctuate up and down every month of every year in a similar pattern/ similar time intervals, thereby indicating presence of seasonality. Also visible is that the magnitude of sales is increasing, thereby suggesting a Multiplicative method for ETS model
2. There is no clear trend of data
3. The Error/Remainder is increasing in variance, indicating a multiplicative model For ETS Model.

Besides, The ACF graph shows gradually decline to zero with seasonal lag but there is high serial correlation, therefore it is required to be seasonally differenced. The PACF graph also shows high serial correlation after lag1, reaffirming the need to difference the dataset.

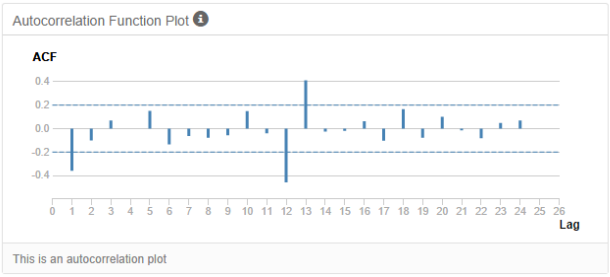
Seasonally differenced Graph

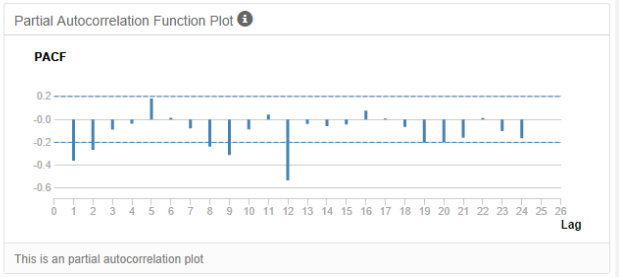




The ACF graph shows positive co-relation at lag 1 and slow decay towards 0 with seasonal lags. However, the serial co-relation of is still high. Therefore, we will need to seasonally difference it, to stationarize the time-series

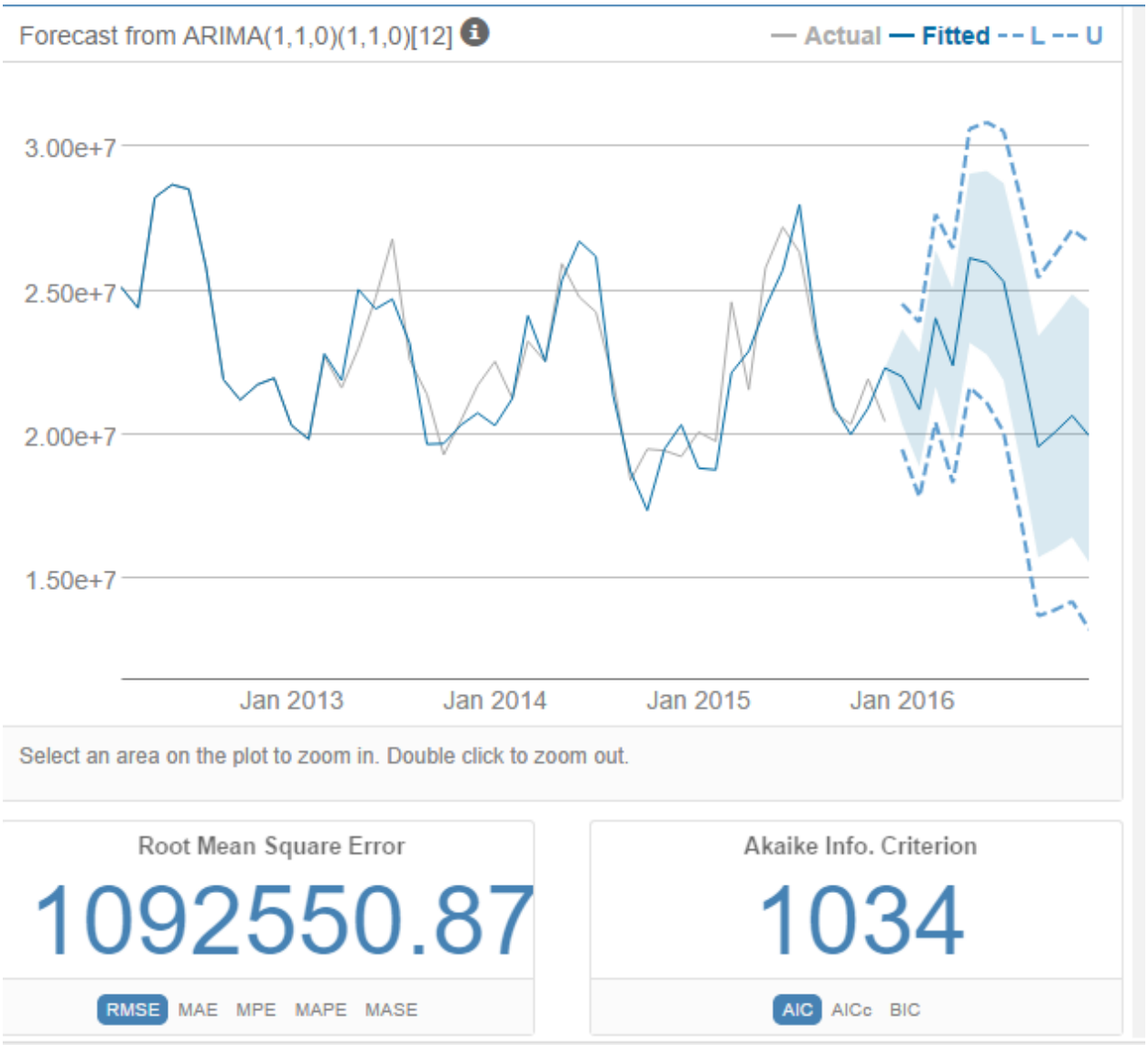
First Seasonally Differenced Graph





As we can see the serial correlation has now subsidized and the time series is now stationary

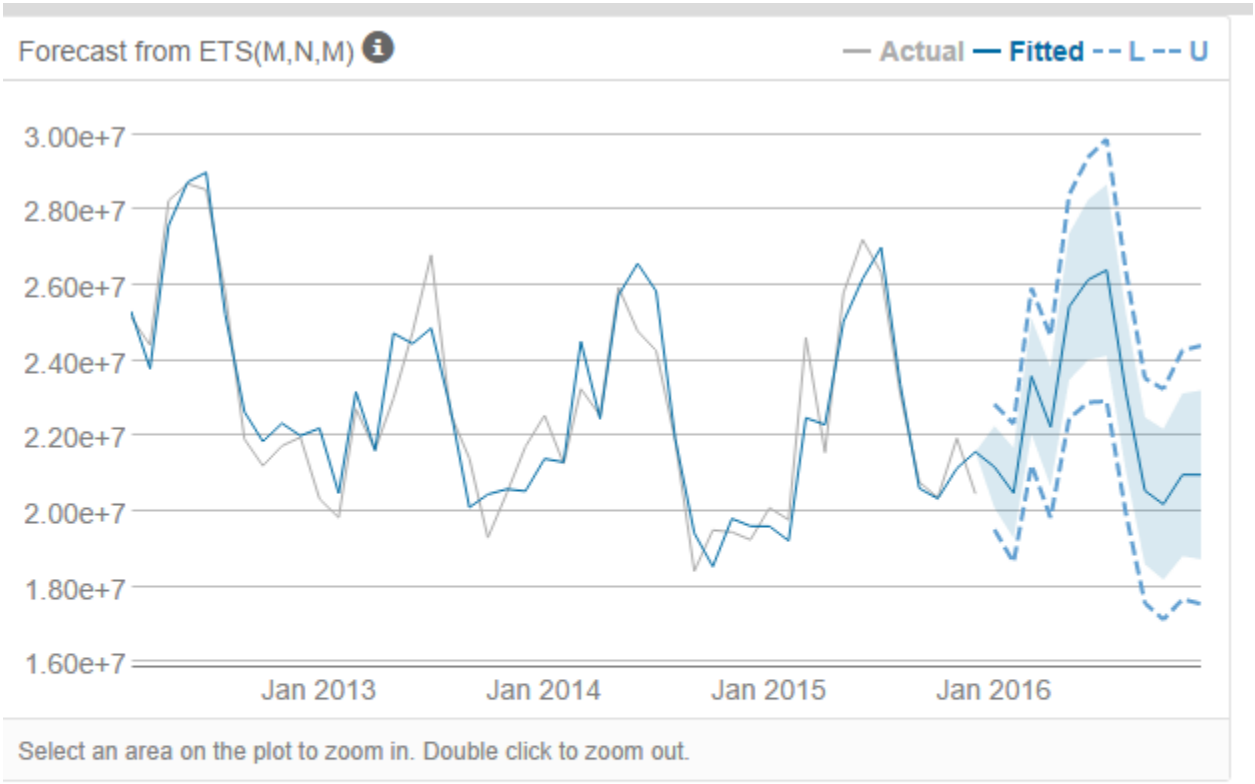
We can now use the Auto function in the ARIMA Model tool to come up with a model

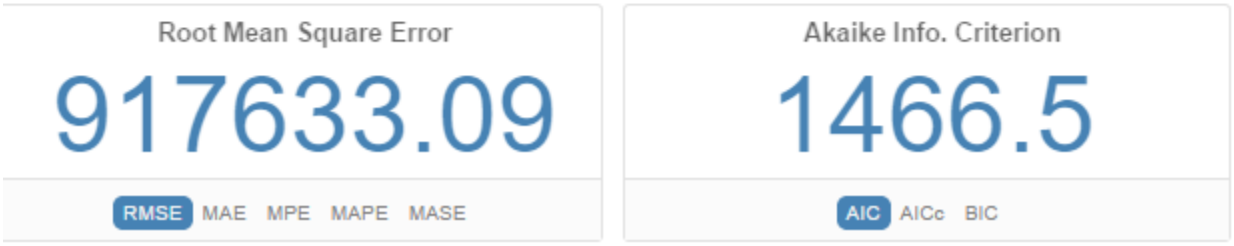


This model seems correct because when we see the Seasonally differenced graph , We see ACF is positive at Lag-1,Indicating an AR(1) Model. Besides the data was differenced and seasonally differenced once and shows a serial co-relation at 13 indicating “P” = 1. Therefore we get ARIMA (1,1,0)(1,1,0){12}

The AIC value of the model is 1034

Now as per the above decomposition plots, our ETS model should be ETS(M,N,M).The results of this model are:





The AIC Value is 1466.5

Therefore, though we see the AIC values, of ARIMA model is lesser

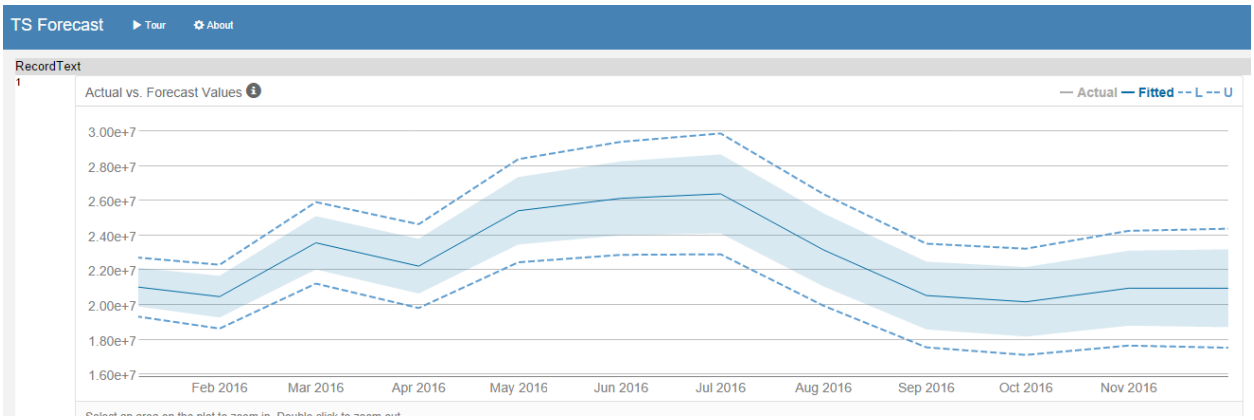
we must also observe: RMSE of ETS is lesser which means forecasts are in a narrower range

MAE value of ETS is smaller and MAPE value of ETS is too smaller, which means the ETS model better fits the data

Thus, we should use the ETS model for further forecasts.

The forecast sales of produce for existing stores in 2016 are:

Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2016	1	21174989.403659	22840074.392095	22263729.957326	20086248.849992	19509904.415223
2	2016	2	20479354.577586	22316289.407718	21680461.702986	19278247.452186	18642419.747454
3	2016	3	23580340.680459	25927572.132266	25115112.832134	22045568.528784	21233109.228653
4	2016	4	22236546.234725	24649240.998413	23814122.545298	20658969.924152	19823851.471036
5	2016	5	25427255.45708	28396631.129606	27368825.849104	23485685.065057	22457879.784554
6	2016	6	26143967.404054	29399195.651375	28272446.7483	24015488.059808	22888739.156732
7	2016	7	26399993.267031	29879368.95032	28675034.741879	24124951.792182	22920617.583741
8	2016	8	23172393.88002	26386378.249512	25273905.302091	21070882.457949	19958409.510528
9	2016	9	20544268.63882	23528908.819357	22495819.956153	18592717.321487	17559628.458283
10	2016	10	20182471.085708	23241644.321619	22182756.948447	18182185.222969	17123297.849798
11	2016	11	20966876.352467	24271769.849046	23127830.057692	18805922.647242	17661982.855887
12	2016	12	20965097.001691	24391891.031043	23205757.181039	18724436.822343	17538302.972339



Applying the same technique to calculate Produce sales for new stores

ETS Model (M, N, M) Method

Q2. Forecast produce sales for the new stores (average rather than the aggregate)

Record #	Period	Sub_Period	Sales For New Stores
1	2016	1	2467001.55059
2	2016	2	2396942.174936
3	2016	3	2797668.615104
4	2016	4	2636525.354518
5	2016	5	3005430.5559
6	2016	6	3073028.935463
7	2016	7	3104846.495517
8	2016	8	2734334.430984
9	2016	9	2420969.111842
10	2016	10	2376263.023341
11	2016	11	2467841.255045
12	2016	12	2464318.336148

Combined Table of Forecasts

Record #	Period	Sub_Period	Sales For New Stores	Sales of Existing stores
1	2016	1	2467001.55059	21174989.403659
2	2016	2	2396942.174936	20479354.577586
3	2016	3	2797668.615104	23580340.680459
4	2016	4	2636525.354518	22236546.234725
5	2016	5	3005430.5559	25427255.45708
6	2016	6	3073028.935463	26143967.404054
7	2016	7	3104846.495517	26399993.267031
8	2016	8	2734334.430984	23172393.88002
9	2016	9	2420969.111842	20544268.63882
10	2016	10	2376263.023341	20182471.085708
11	2016	11	2467841.255045	20966876.352467
12	2016	12	2464318.336148	20965097.001691

Historical and forecast plot of sales for New and Existing stores