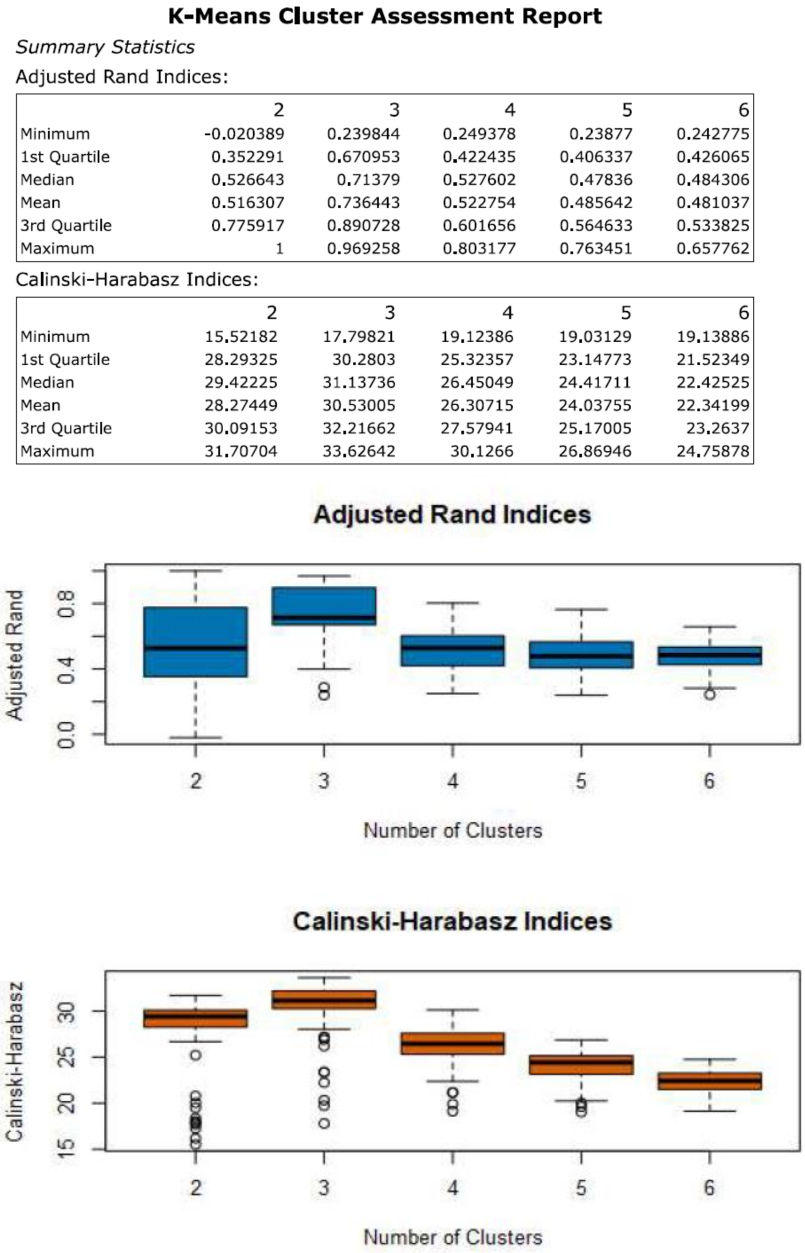
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   I took the store sales data to perform K-means clustering on a % sales by category basis. By looking at the adj. Rand indices, we choose the cluster with the highest mean i.e. number of clusters = 3.



K-Means Cluster Assessment Report

Summary Statistics

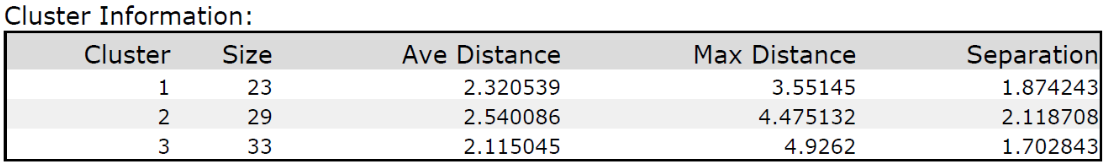
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.020389	0.239844	0.249378	0.23877	0.242775
1st Quartile	0.352291	0.670953	0.422435	0.406337	0.426065
Median	0.526643	0.71379	0.527602	0.47836	0.484306
Mean	0.516307	0.736443	0.522754	0.485642	0.481037
3rd Quartile	0.775917	0.890728	0.601656	0.564633	0.533825
Maximum	1	0.969258	0.803177	0.763451	0.657762

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.52182	17.79821	19.12386	19.03129	19.13886
1st Quartile	28.29325	30.2803	25.32357	23.14773	21.52349
Median	29.42225	31.13736	26.45049	24.41711	22.42525
Mean	28.27449	30.53005	26.30715	24.03755	22.34199
3rd Quartile	30.09153	32.21662	27.57941	25.17005	23.2637
Maximum	31.70704	33.62642	30.1266	26.86946	24.75878

2. How many stores fall into each store format?

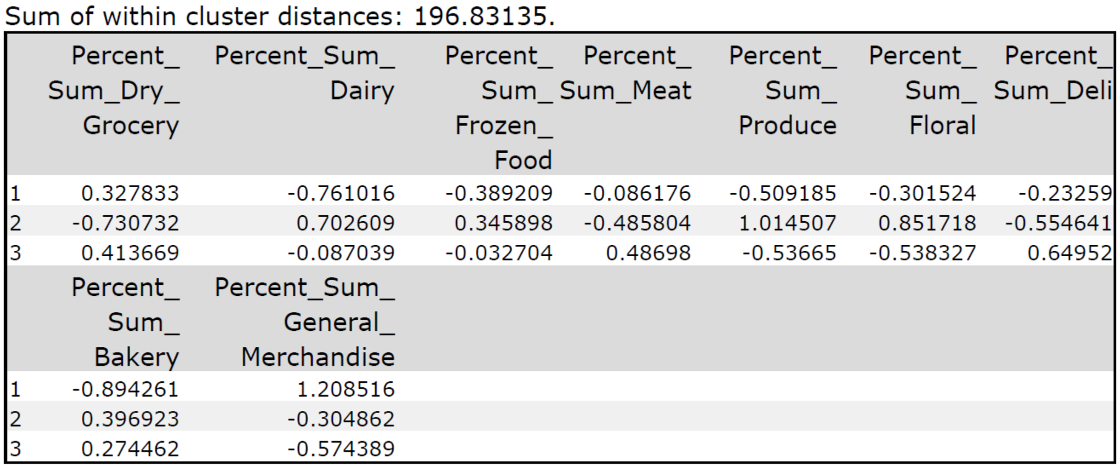


Cluster Information:

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

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

Cluster 2 is negative compared to cluster 1 and cluster 3 in terms of percentage of dry grocery. -0.730732 does not mean it is less than the other two, just that they are different.

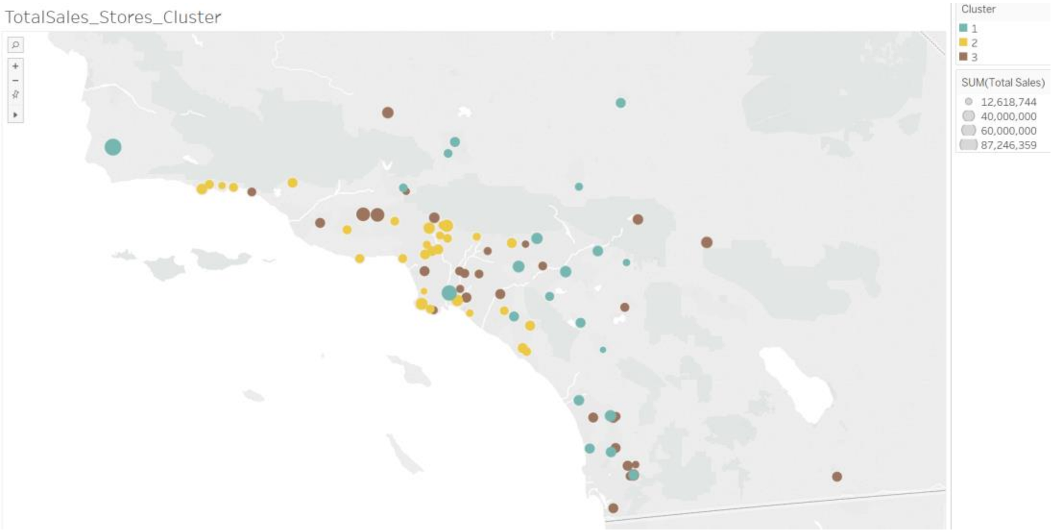


Sum of within cluster distances: 196.83135.

	Percent_Sum_Dry_Grocery	Percent_Sum_Dairy	Percent_Sum_Frozen_Food	Percent_Sum_Meat	Percent_Sum_Produce	Percent_Sum_Floral	Percent_Sum_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

	Percent_Sum_Bakery	Percent_Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

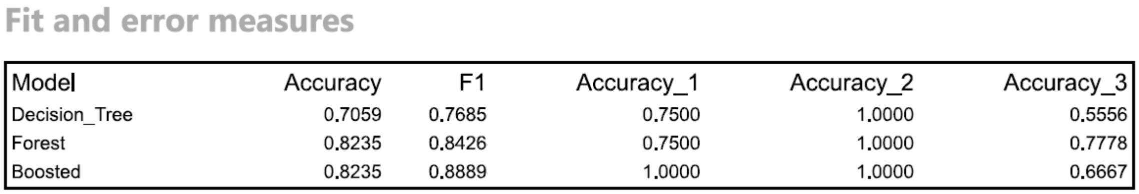
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.



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? (20% validation sample with Random Seed = 3)

   Methods used – Decision Tree, Forest and Boosted models. Demographic data was used from StoreDemographicData.csv to predict store formats for new stores. Training data – 80% data, validation data – 20%. Although, Boosted and Forest model have same accuracy, F1 score for the Boosted model is higher so it's the more ideal choice.



Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted	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

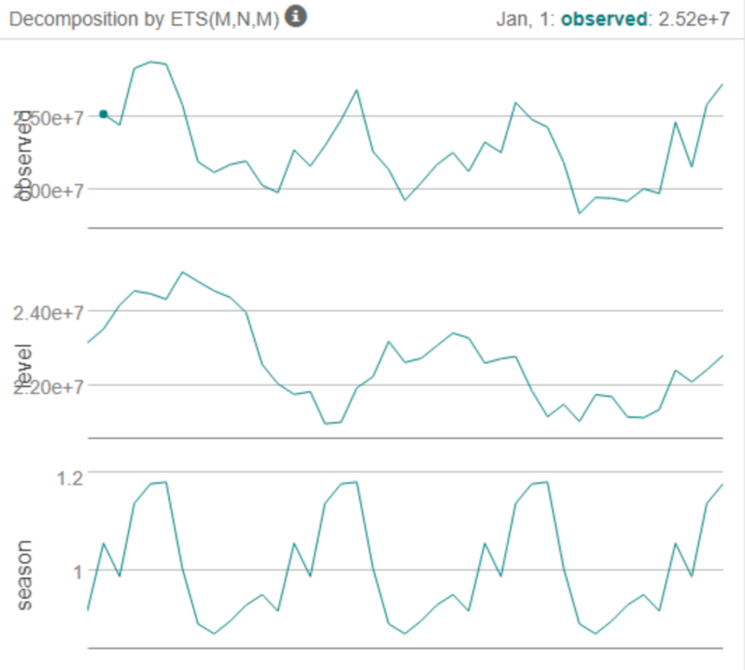
3 for cluster 1, 6 for cluster 2 and 1 for cluster 3.

Task 3: Predicting Produce Sales

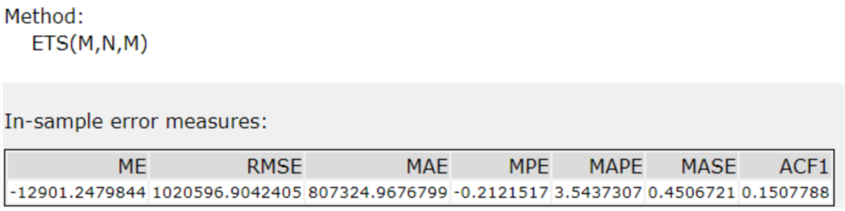
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

Holdout sample used is 6 months.

ETS modeling:-
The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. Its error is irregular and should be applied multiplicatively.



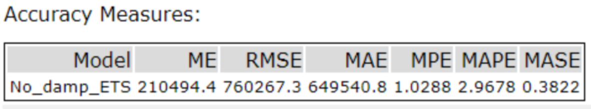
ETS model can have two configurations – with or without dampening.
Because of better in sample error and accuracy measures, ETS MNM i.e. with no dampening is chosen.
ETS w/ no dampening



Method:
ETS(M,N,M)

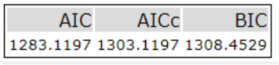
In-sample error measures:

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



Accuracy Measures:

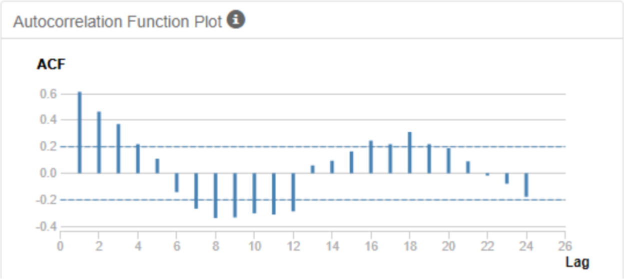
Model	ME	RMSE	MAE	MPE	MAPE	MASE
No_damp_ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

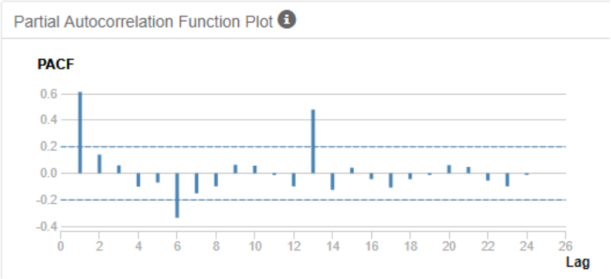


AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

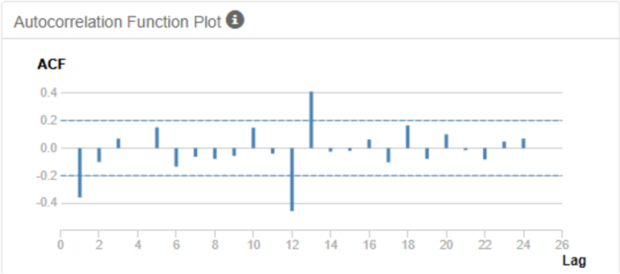
ARIMA modeling:-

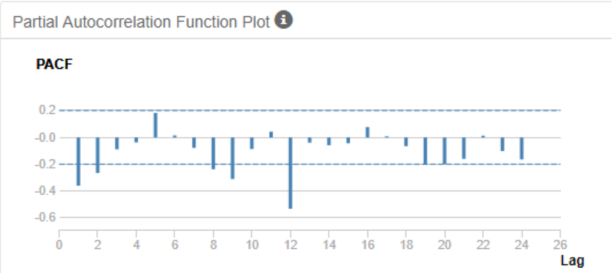
ARIMA – seasonal difference



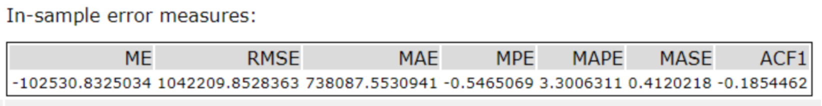


ARIMA – difference 1



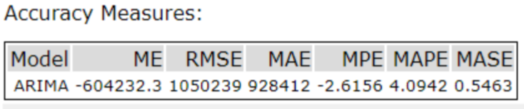


ARIMA model used is ARIMA(1,0,0)(1,1,0)[12] . Seasonal difference and seasonal first difference were performed to make the series stationary.



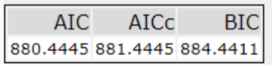
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463



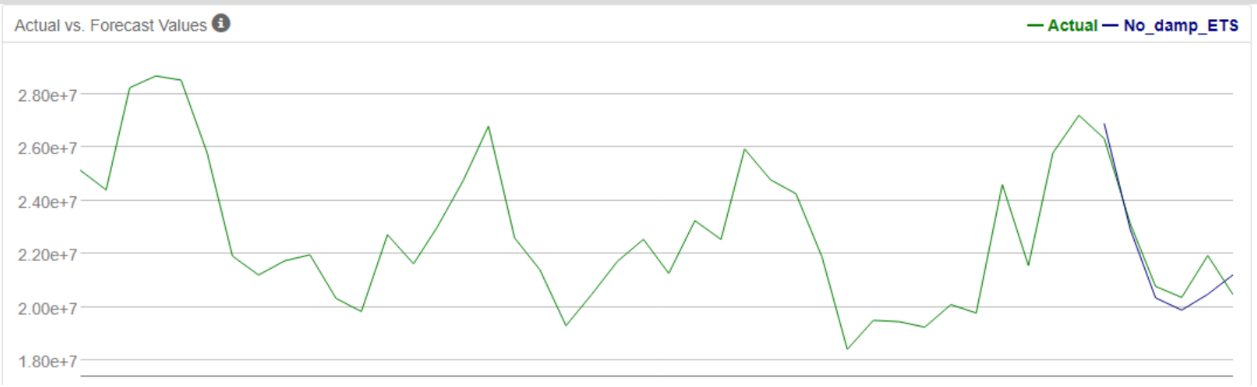
AIC	AICc	BIC
880.4445	881.4445	884.4411

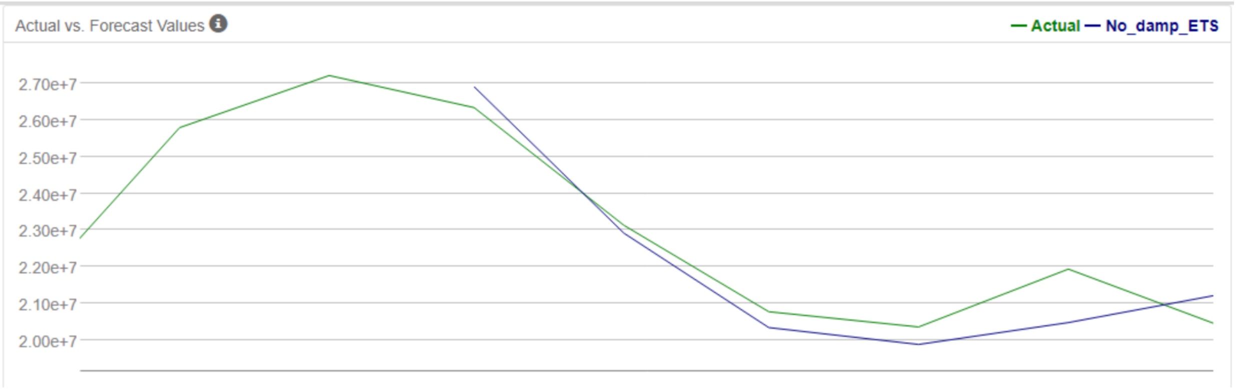
ETS model has higher accuracy when versus the ARIMA model and has lower in-sample error measures as well.

ETS model RMSE accuracy is 760267.3 vs ARIMA model RMSE accuracy of 1050239.

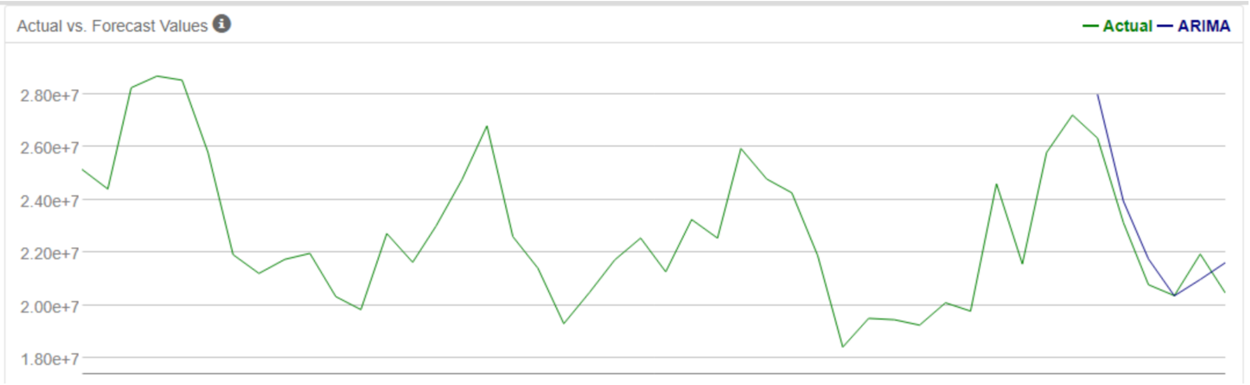
ETS model MASE accuracy is 0.3822 vs ARIMA model MASE accuracy of 0.5463.

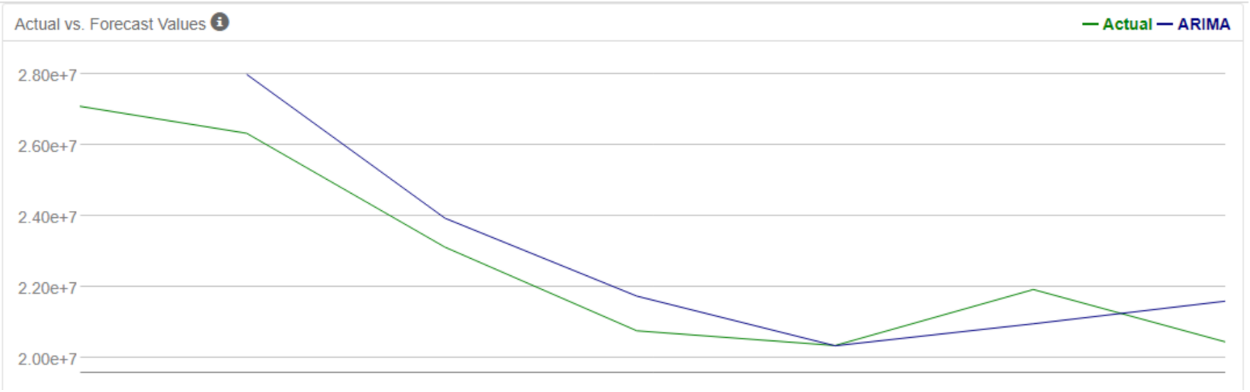
TS compare tool for ETS:

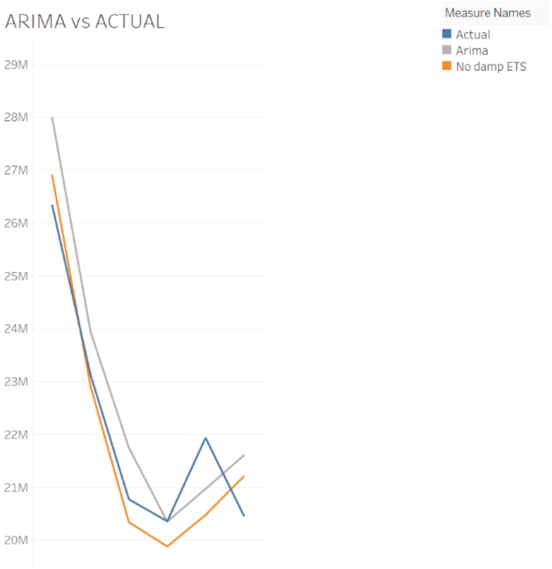




TS compare tool for ARIMA

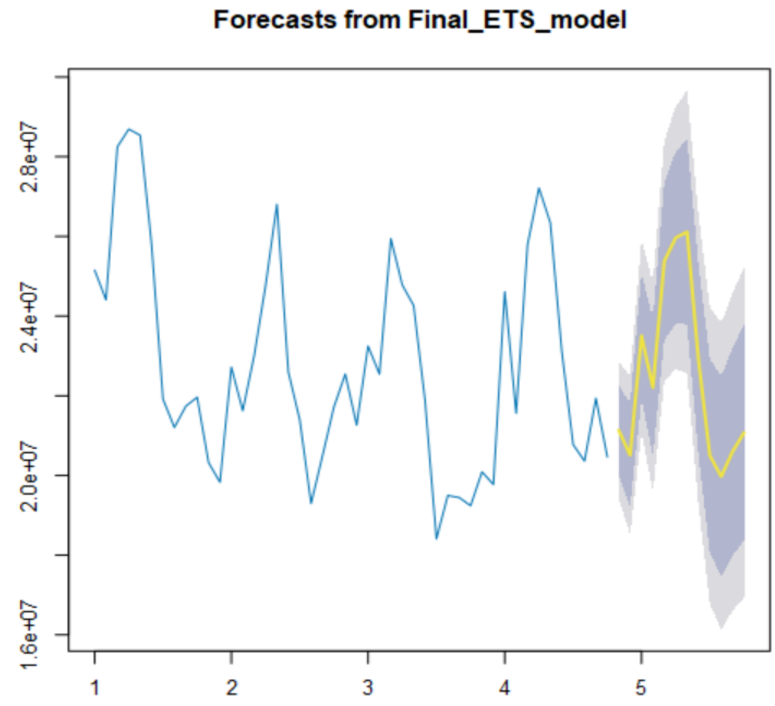


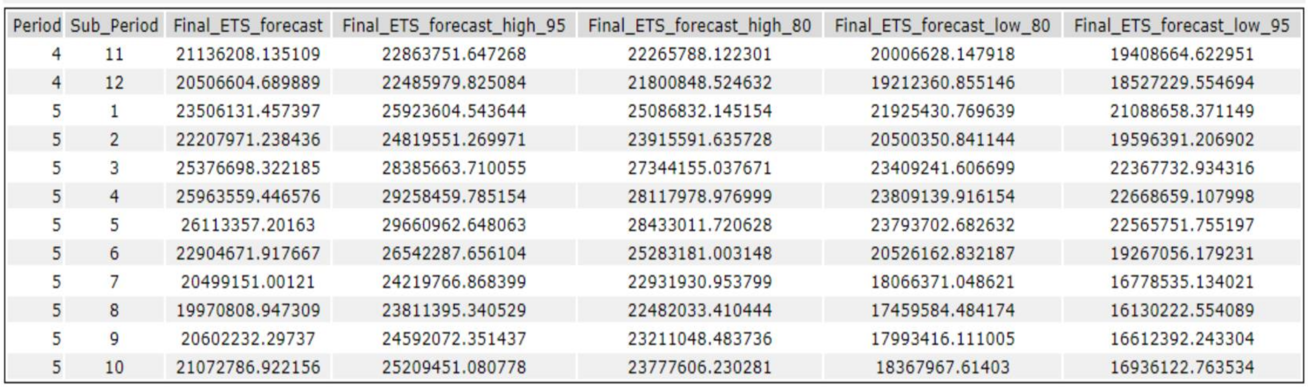




Its clear from the visualizations that ETS is actually closer to actual compared to ARIMA. So considering in-sample error measures, accuracy measures and actual vs forecast plots, I think ETS model is a better choice.

The graph and table below shows actual and forecast value with 80% & 95% confidence level interval.



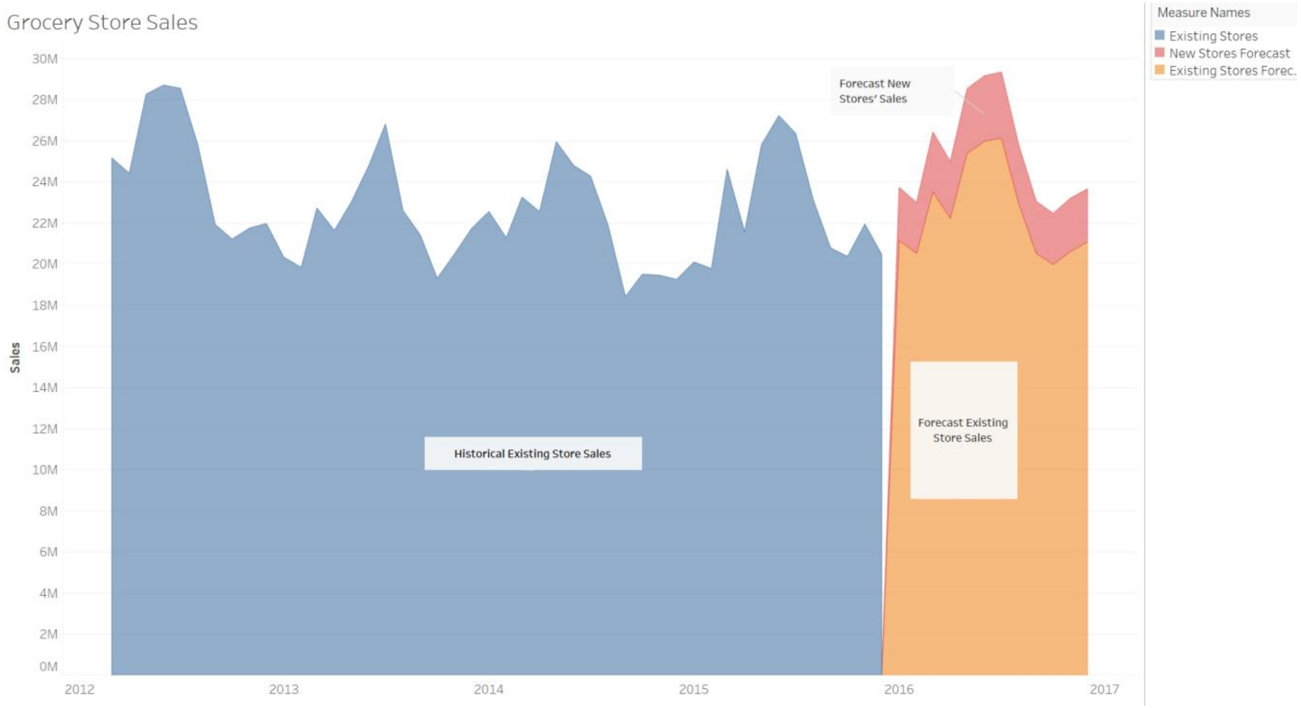


Period	Sub_Period	Final_ETS_forecast	Final_ETS_forecast_high_95	Final_ETS_forecast_high_80	Final_ETS_forecast_low_80	Final_ETS_forecast_low_95
4	11	21136208.135109	22863751.647268	22265788.122301	20006628.147918	19408664.622951
4	12	20506604.689889	22485979.825084	21800848.524632	19212360.855146	18527229.554694
5	1	23506131.457397	25923604.543644	25086832.145154	21925430.769639	21088658.371149
5	2	22207971.238436	24819551.269971	23915591.635728	20500350.841144	19596391.206902
5	3	25376698.322185	28385663.710055	27344155.037671	23409241.606699	22367732.934316
5	4	25963559.446576	29258459.785154	28117978.976999	23809139.916154	22668659.107998
5	5	26113357.20163	29660962.648063	28433011.720628	23793702.682632	22565751.755197
5	6	22904671.917667	26542287.656104	25283181.003148	20526162.832187	19267056.179231
5	7	20499151.00121	24219766.868399	22931930.953799	18066371.048621	16778535.134021
5	8	19970808.947309	23811395.340529	22482033.410444	17459584.484174	16130222.554089
5	9	20602232.29737	24592072.351437	23211048.483736	17993416.111005	16612392.243304
5	10	21072786.922156	25209451.080778	23777606.230281	18367967.61403	16936122.763534

3. 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.

Period	Sub_Period	new_stores_forecast	existing_store_forecast
4	11	$ 2,584,383.53	$ 21,136,208.00
4	12	$ 2,470,873.92	$ 20,506,605.00
5	1	$ 2,906,307.87	$ 23,506,131.00
5	2	$ 2,771,532.13	$ 22,207,971.00
5	3	$ 3,145,848.57	$ 25,376,698.00

5	4	$ 3,183,909.28	$ 25,963,559.00
5	5	$ 3,213,977.72	$ 26,113,357.00
5	6	$ 2,858,247.21	$ 22,904,672.00
5	7	$ 2,538,173.64	$ 20,499,151.00
5	8	$ 2,483,550.17	$ 19,970,809.00
5	9	$ 2,593,089.19	$ 20,602,232.00
5	10	$ 2,570,200.44	$ 21,072,787.00


Grocery Store Sales
Measure Names
Existing Stores
New Stores Forecast
Existing Stores Forec..
Forecast New Stores' Sales
Historical Existing Store Sales
Forecast Existing Store Sales
Sales
30M
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M