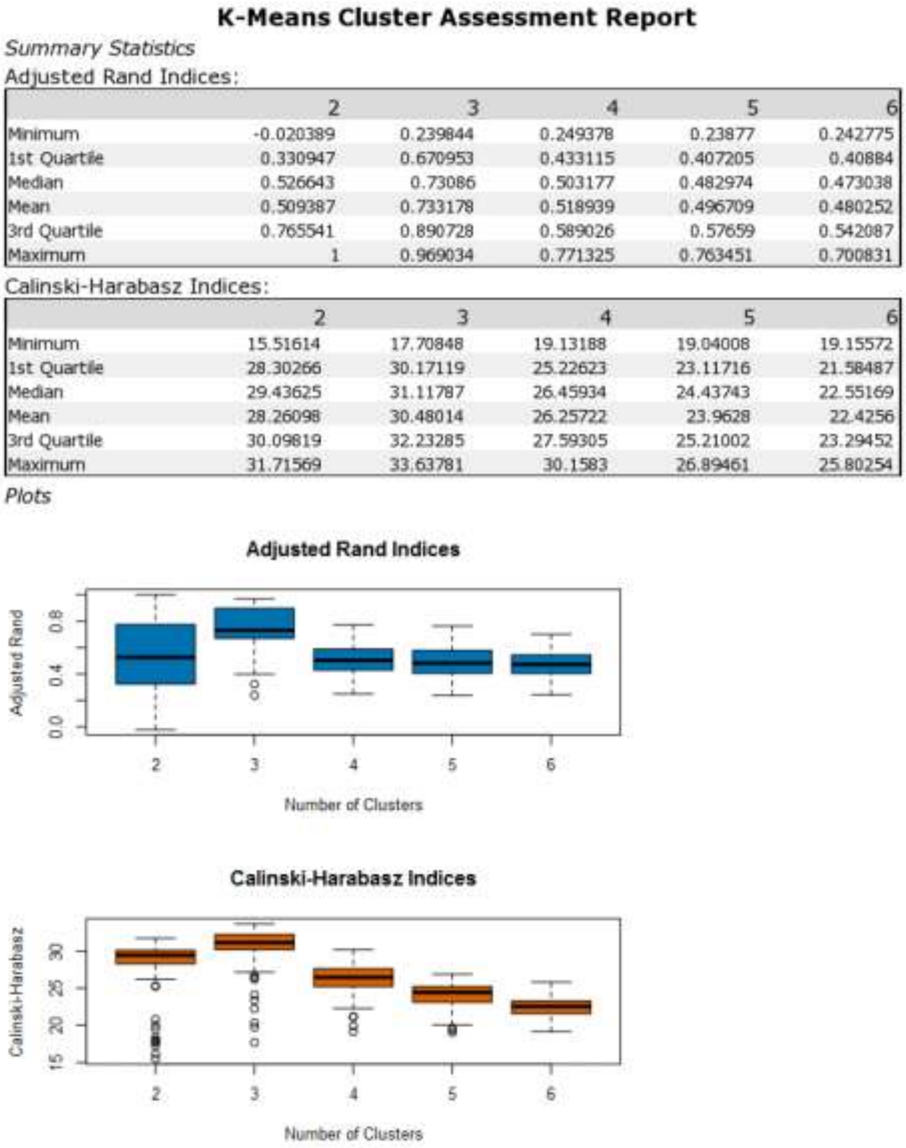
Project: Predictive Analytics Capstone

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.
   From the k-means report shown below we can see that Adjusted Rand & Calinski Harabanz indices show the highest median value when number of clusters are 3.



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

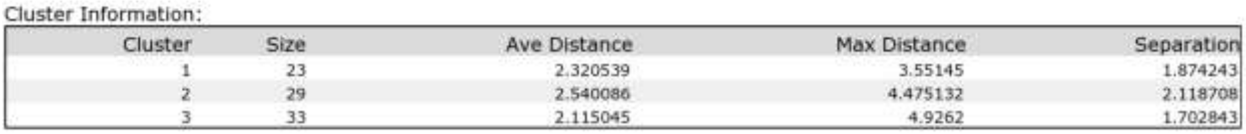
	2	3	4	5	6
Minimum	-0.020389	0.239844	0.249378	0.23877	0.242775
1st Quartile	0.330947	0.670953	0.433115	0.407205	0.40884
Median	0.526643	0.73086	0.503177	0.482974	0.473038
Mean	0.509387	0.733178	0.518939	0.496709	0.480252
3rd Quartile	0.765541	0.890728	0.589026	0.57659	0.542087
Maximum	1	0.969034	0.771325	0.763451	0.700831

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.51614	17.70848	19.13188	19.04008	19.15572
1st Quartile	28.30266	30.17119	25.22623	23.11716	21.58487
Median	29.43625	31.11787	26.45934	24.43743	22.55169
Mean	28.26098	30.48014	26.25722	23.9628	22.4256
3rd Quartile	30.09819	32.23285	27.59305	25.21002	23.29452
Maximum	31.71569	33.63781	30.1583	26.89461	25.80254

Plots

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

We can see from the report that there are 3 clusters formed namely 1, 2 & 3 and 23, 29 & 33 stores fall in each of them respectively.

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

Cluster1 has sold more General Merchandise.

Cluster 2 stores are good at selling floral products.

Whereas the totals sales of Cluster 3 stores is the highest among them.

Tableau Visualization

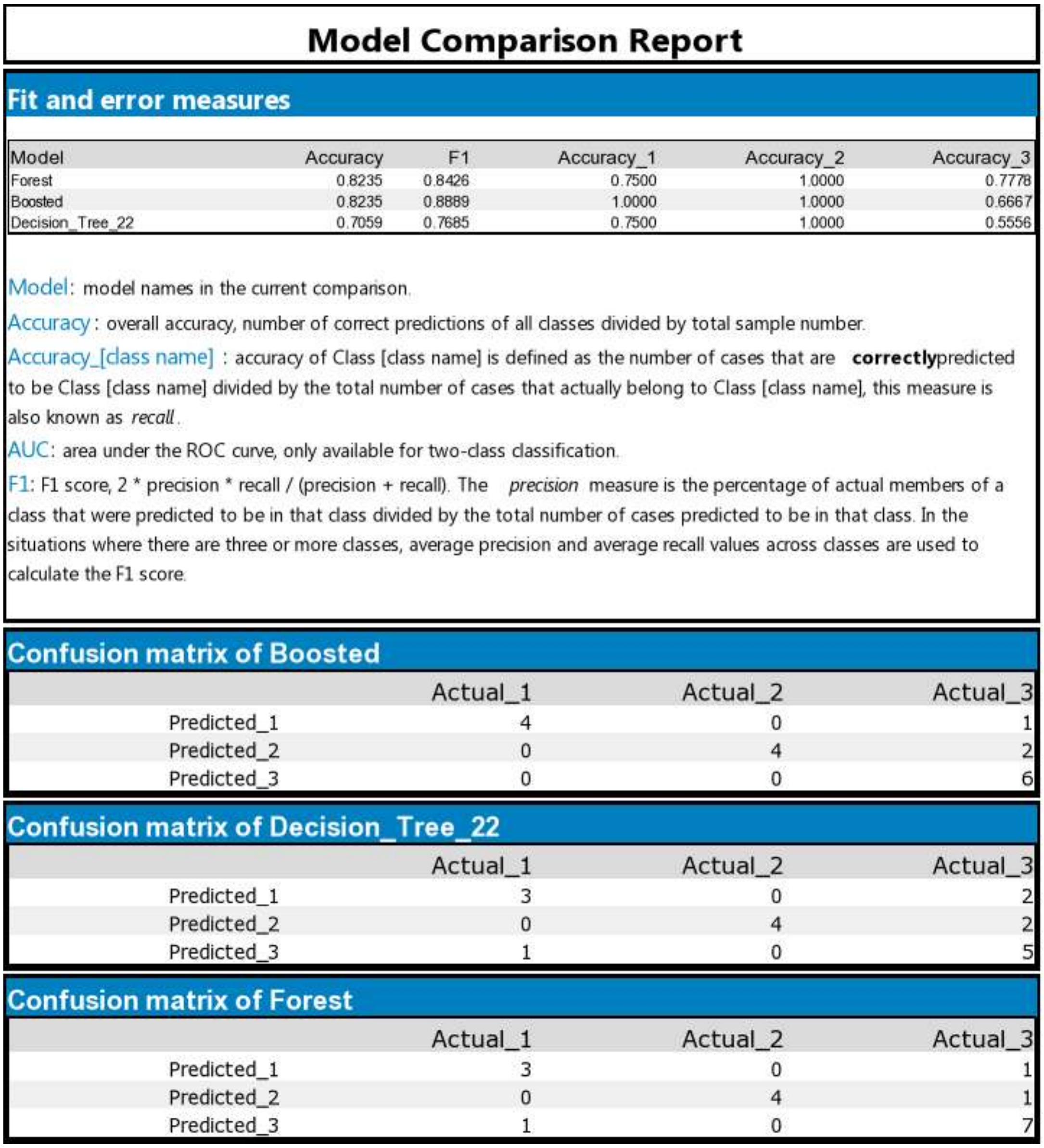
https://public.tableau.com/profile/swapnil7839#!/vizhome/Task1_192/Sheet2

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.

Tableau Visualization
https://public.tableau.com/profile/swapnil7839#!/vizhome/Task1_192/Sheet1

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.)
   Three models were used and compared namely Decision Tree, Forest and Boosted Model.
   From the report we can see that accuracy of Forest model and Boosted model is same but we choose Boosted model as our model to predict the best store format because the F1 value of boosted model is higher as compared to forest model.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Decision_Tree_22	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted

	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_22

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest

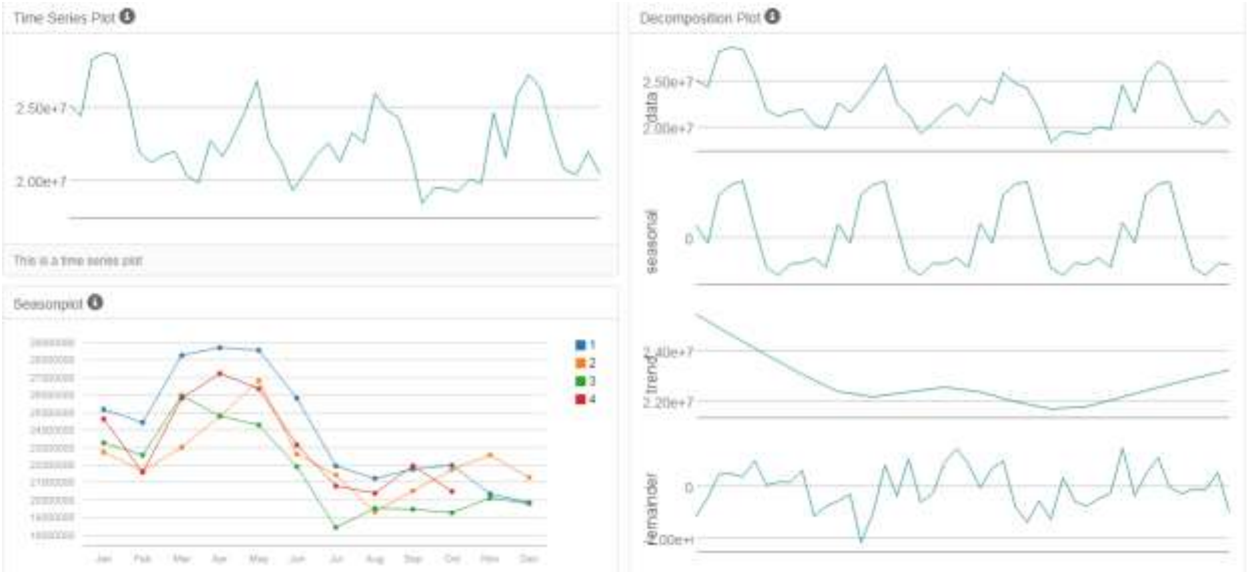
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

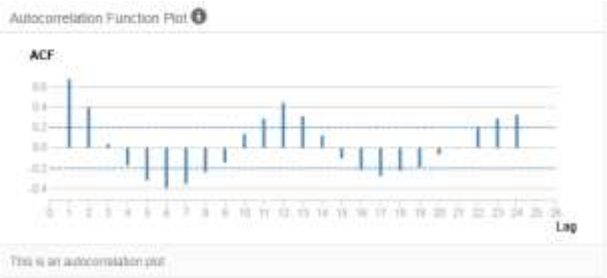
Task 3: Predicting Produce Sales

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?

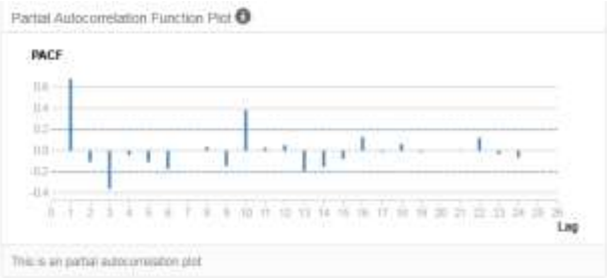


From the graph, since the seasonality is showing increasing trend we apply mulitiplicative for it.
Trend graph is not clear so we are neither applying addition nor multiplication on it and keeping it as it is.
Error graph is showing fluctuations so we apply multiplication for it.
So we apply ETS(m,n,m).

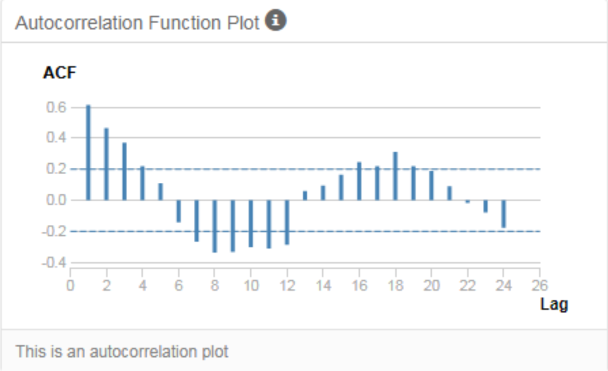
Now, for ARIMA we see that there is correlations so we have applied seasonal differencing.


Autocorrelation Function Plot

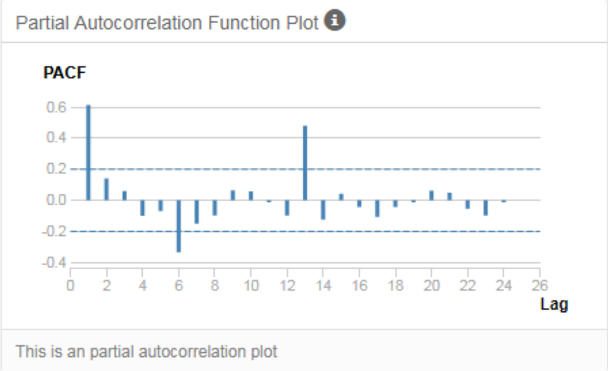
This is an autocorrelation plot


Partial Autocorrelation Function Plot

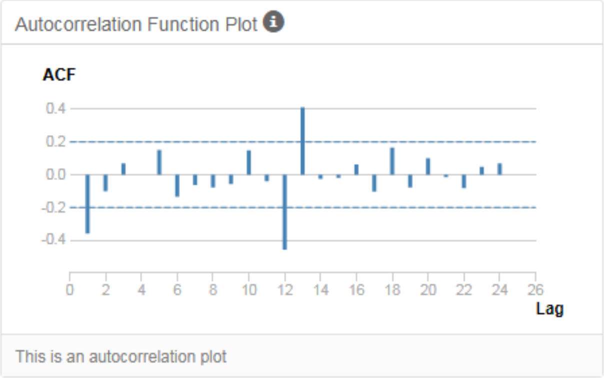
This is an partial autocorrelation plot


Autocorrelation Function Plot

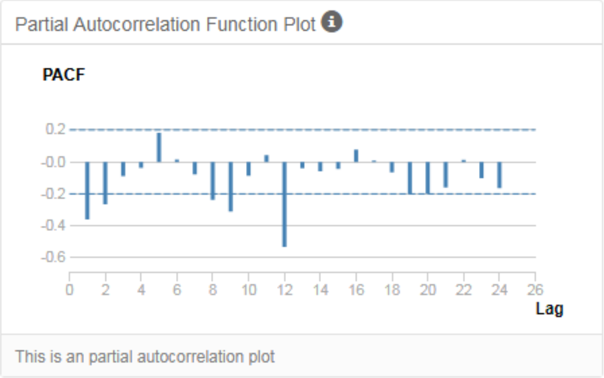
This is an autocorrelation plot


Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

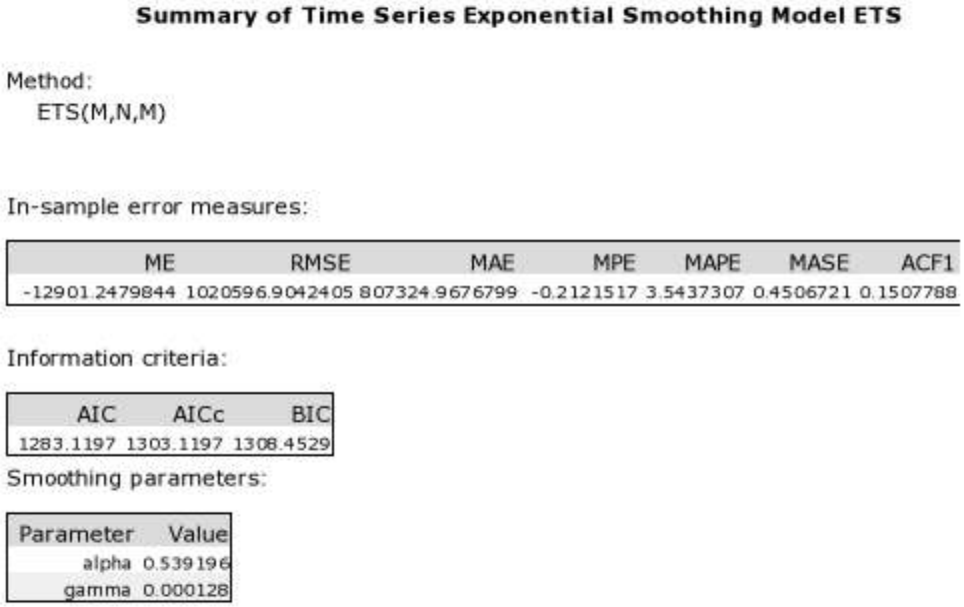

Autocorrelation Function Plot

This is an autocorrelation plot


Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

After applying seasonal differencing we get the values for ARIMA as ARIMA(0,1,2)(0,1,0)



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

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

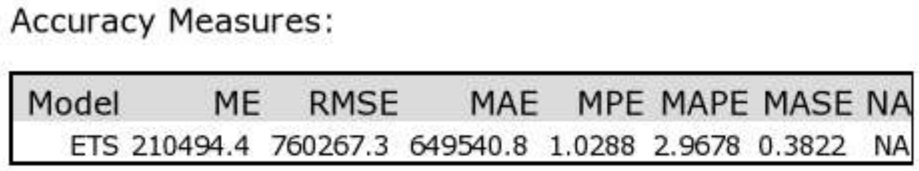
Smoothing parameters:

Parameter	Value
alpha	0.539196
gamma	0.000128

By comparing the values of ETS and ARIMA model we see that the ETS model is better at forecasting since the RMSE of ETS 1020596.91 is greater than that of ARIMA which is 1429296.30. Also AIC of ETS model 1283.11 is high than that of ARIMA which is 858.78.

So we choose ETS model for forecasting.

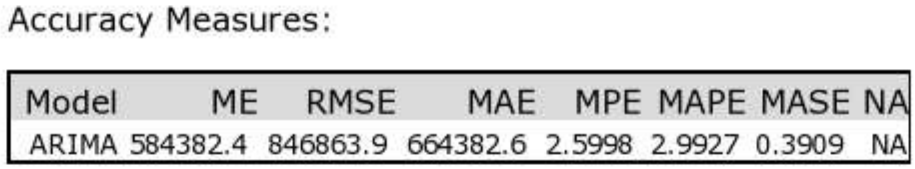
ETS forecast error measurements against the holdout sample



Accuracy Measures:

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

ARIMA forecast error measurements against the holdout sample

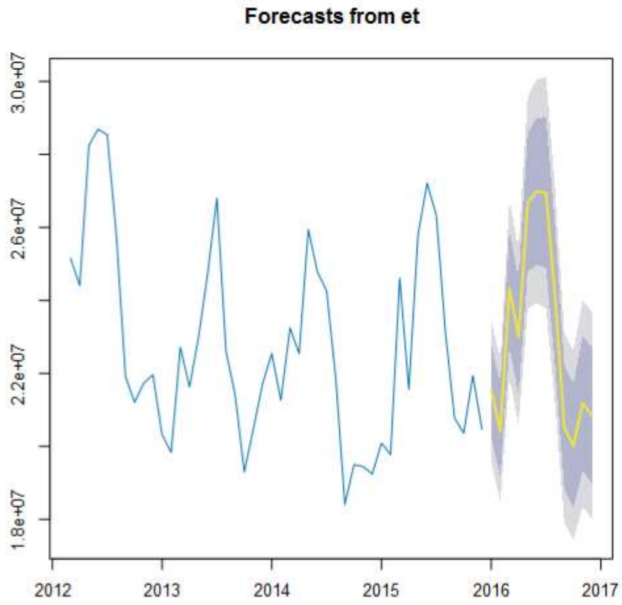


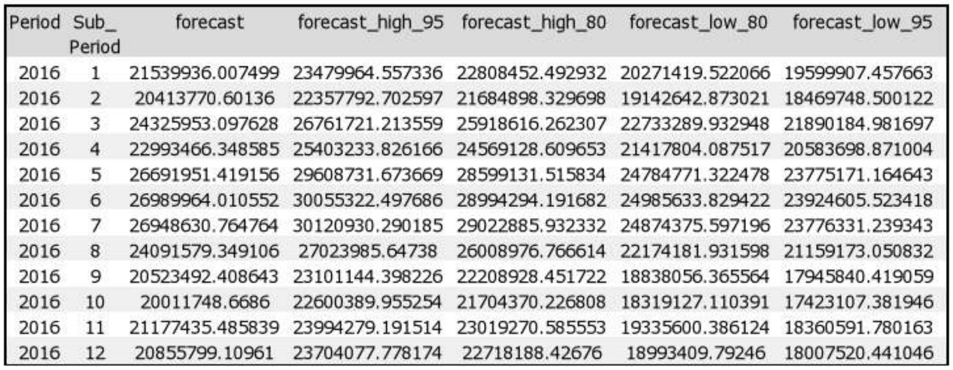
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909	NA

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.
   After choosing the efficient model forecasting i.e. ETS model we forecast the values.
   And we calculate the forecast values by 95% of larger confidence interval & 80% smaller confidence interval.

12 Period Forecast from et

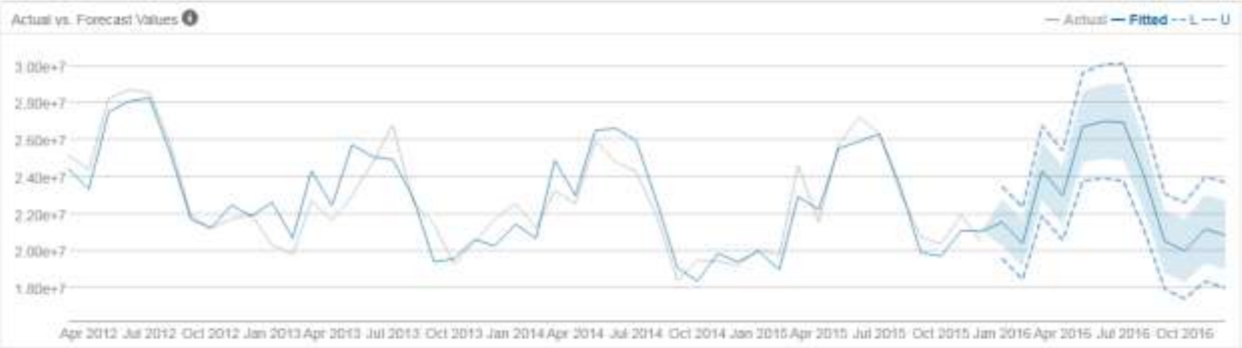




Period	Sub_ Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

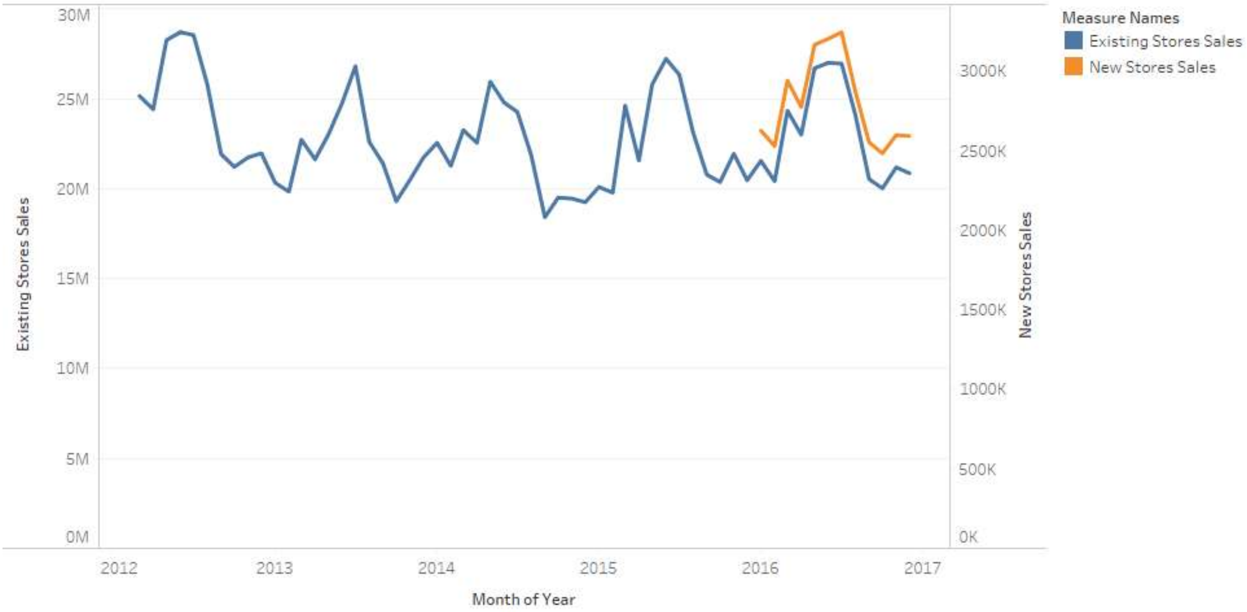
The table shows the new and existing stores sales for 12 months ranging from January 2016 to December 2016.

Year	**Month**	**New Store Sales**	**ExistingStore Sales**
2016	1	2,626,198	21,539,936
2016	2	2,529,186	20,413,771
2016	3	2,940,264	24,325,953
2016	4	2,774,135	22,993,466
2016	5	3,165,320	26,691,951
2016	6	3,203,286	26,989,964
2016	7	3,244,464	26,948,631
2016	8	2,871,488	24,091,579
2016	9	2,552,418	20,523,492
2016	10	2,482,837	20,011,749
2016	11	2,597,780	21,177,435
2016	12	2,591,815	20,855,799



Visualization of Sales Forecast

Total Produce Sales Forecast



The trends of Existing Stores Sales and New Stores Sales for Month of Year. Color shows details about Existing Stores Sales and New Stores Sales. The data is filtered on Month, Year of Year, which keeps 58 of 58 members.