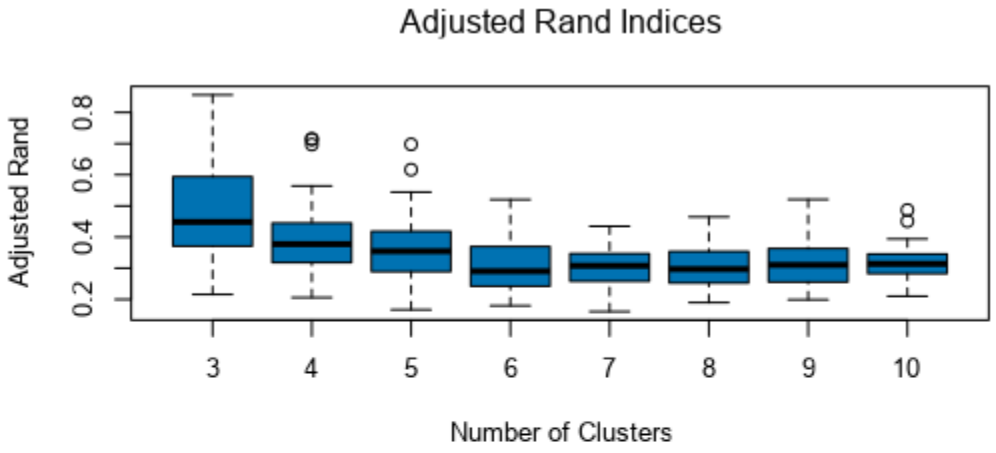
Project: Predictive Analytics Capstone

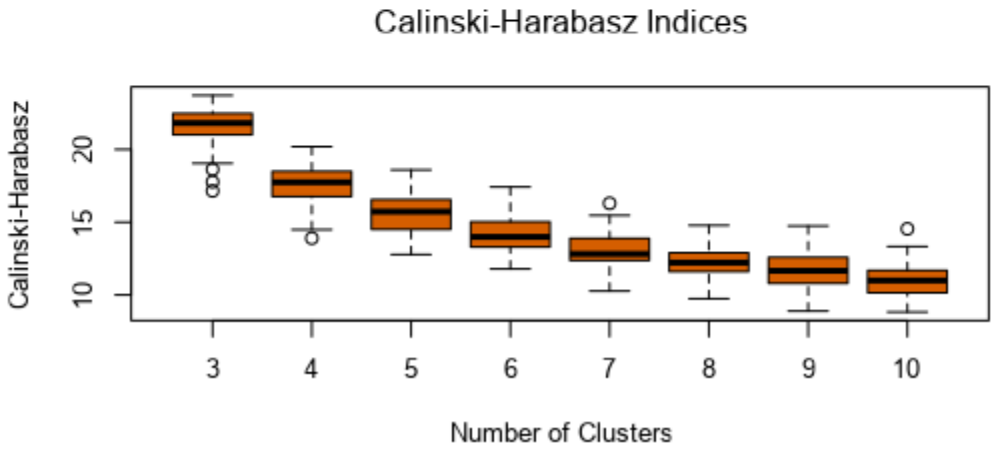
Task 1: Determine Store Formats for Existing Stores

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

   Three-store format is the ideal number. Alteryx K-Centroids Diagnostic tool was used to arrive at the value of this assumption.

   It is possible to see on the Adjusted Rand Indices box plot below and the Calinski-Harabasz Indices box plot down under that Cluster 3 is the highest of them all.





2. **How many stores fall into each store format?**

   It is possible to see the relation between cluster and size, Cluster 1 has the smallest size of 23, Cluster 2 has 29, and cluster 3 has 33.

Report

Summary Report of the K-Means Clustering Solution X

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + X.Dry_Grocery + X.Dairy + X.Frozen_Food + X.Meat + X.Produce + X.Floral + X.Deli + X.Bakery + X.General_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

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

   Based on the K Means Clustering report we can observe how clusters differ from each other.
   - Cluster 1 has the smallest size of 23 and the smallest Max Distance of 3.55.
   - Cluster 1's Ave Distance and Separation are halfway between Cluster 2 and
   - Cluster 3. Cluster 2 is the largest, with a size of 29 and the highest Ave Distance and Separation.
   - Cluster 3 has the most sizes (33), as well as the greatest distance.

Report

Summary Report of the K-Means Clustering Solution X

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + X.Dry_Grocery + X.Dairy + X.Frozen_Food + X.Meat + X.Produce + X.Floral + X.Deli + X.Bakery + X.General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

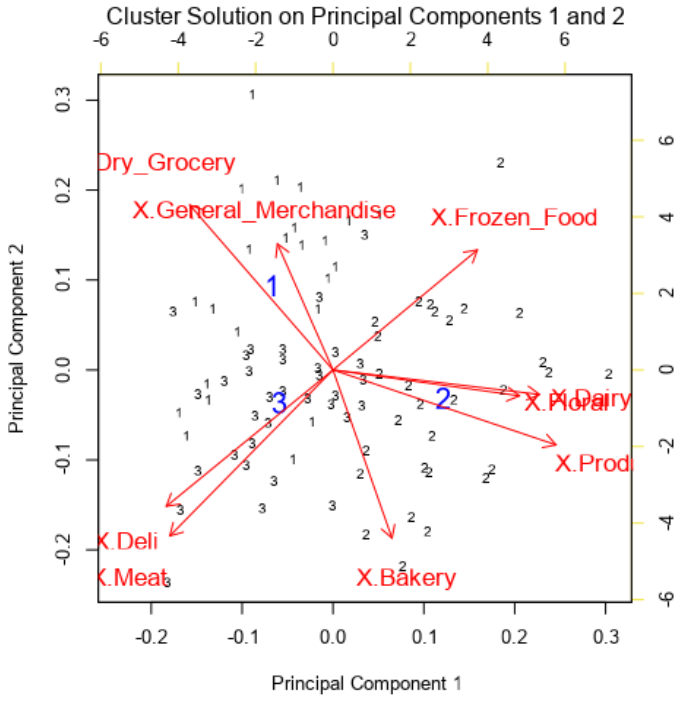
Cluster Information:

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.

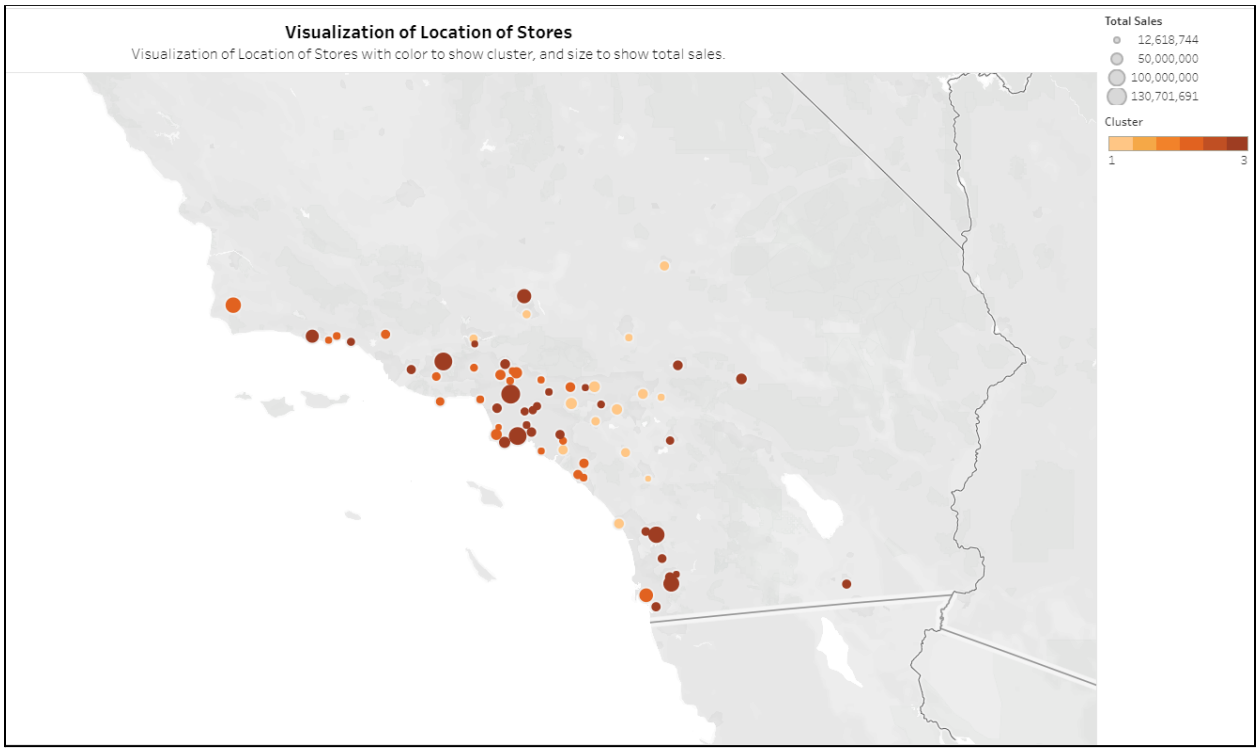
	X.Dry_Grocery	X.Dairy	X.Frozen_Food	X.Meat	X.Produce	X.Floral	X.Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	X.Bakery	X.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.**

Visualization of Location of Stores



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.)**

The methodology used to predict the best store format was by evaluating three following models

- Decision Tree Model
- Boosted Model
- Forest Model

After evaluated them, it was possible to utilized the Boosted Model to forecast the ideal shop format for the new stores.

The Boosted Model with 0.8235 Accuracy, 0.8889 F1, 1.0000 Accuracy 1, 1.0000 Accuracy 2, and 0.6667 Accuracy 3 is the best model, according to the Model Comparison Report.

Model Comparison Report

Fit and error measures

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

When I compare the Confusion matrices of the three models, I can see that the Boosted Model is the best in Predicted 1 and Actual 1 (4), Predicted 2 and Actual 2 (4), and Predicted 3 and Actual 3 (6). So, to estimate the ideal shop format for the new stores, I used the Boosted Model.

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

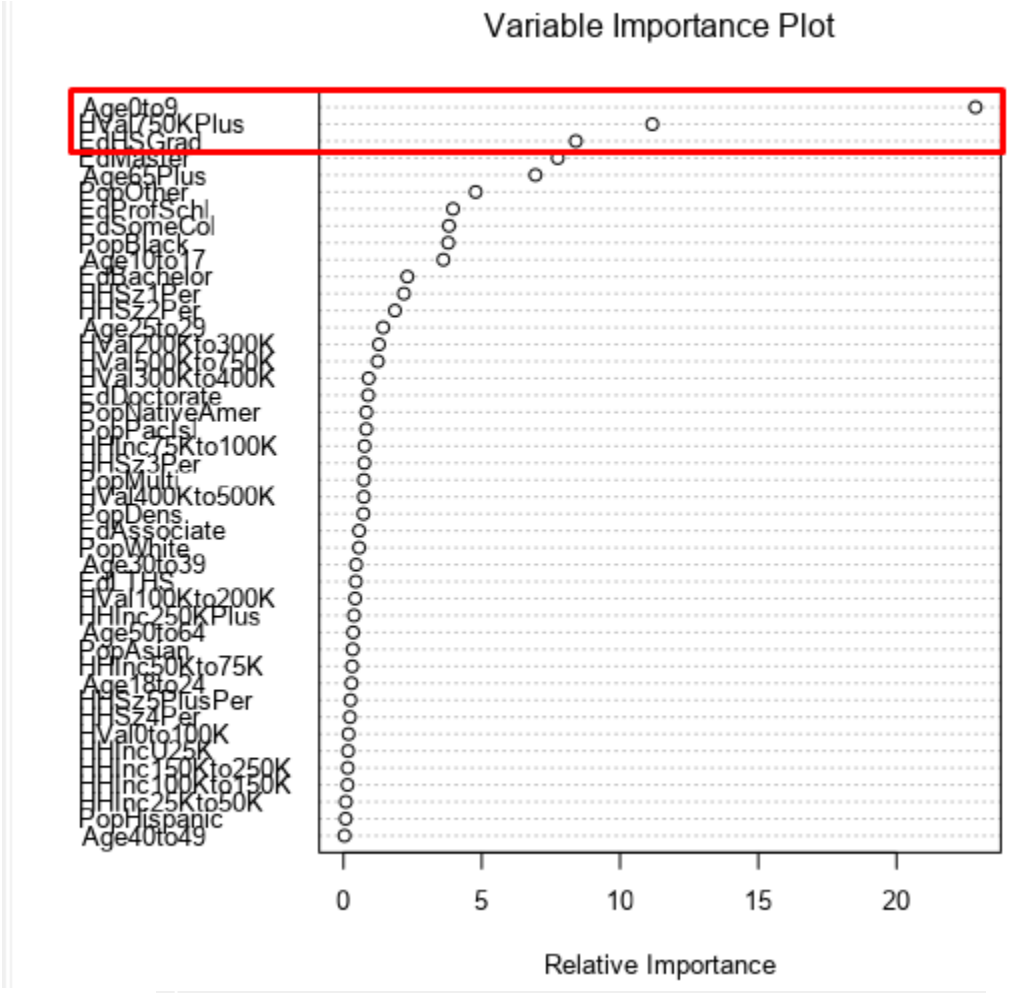
Confusion matrix of Decision_Tree_Model

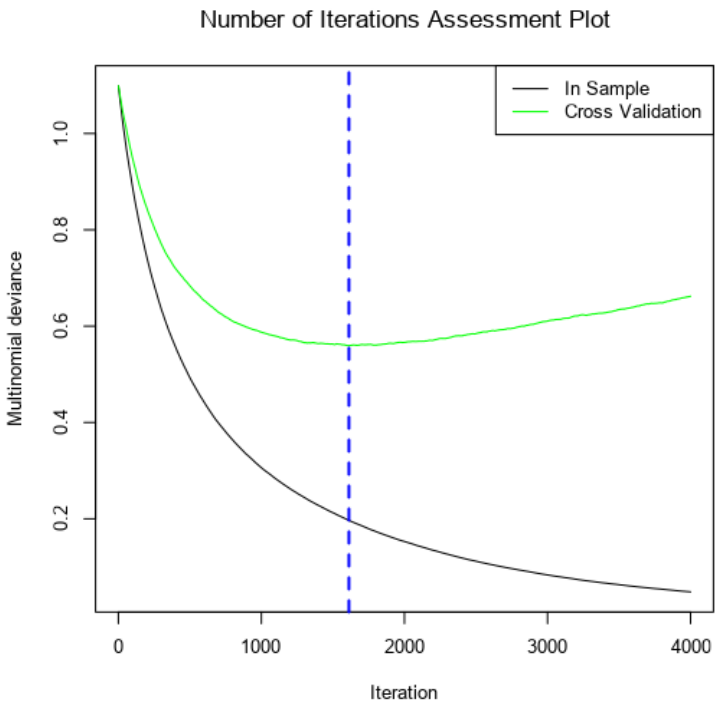
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest_Model

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

The three most essential variables, according to the Report for Boosted Model are Age0to9, HVal750KPlus, and EdMaster.





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

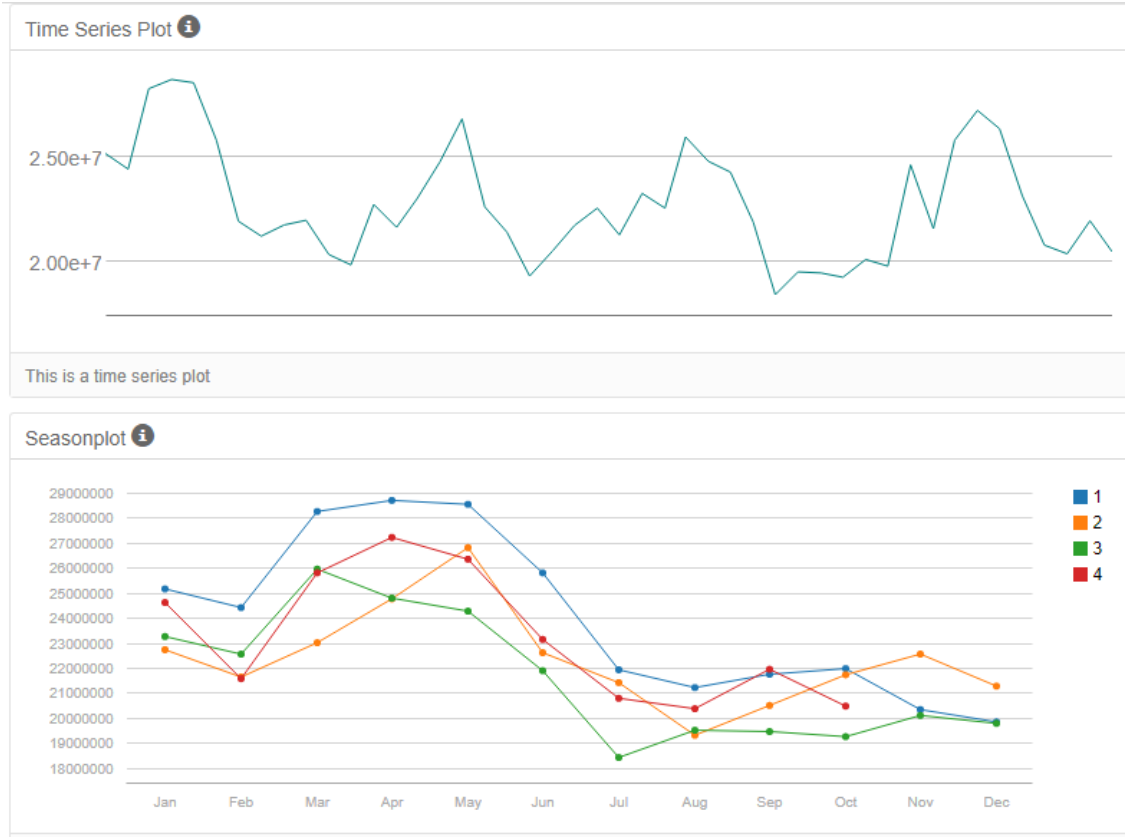
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

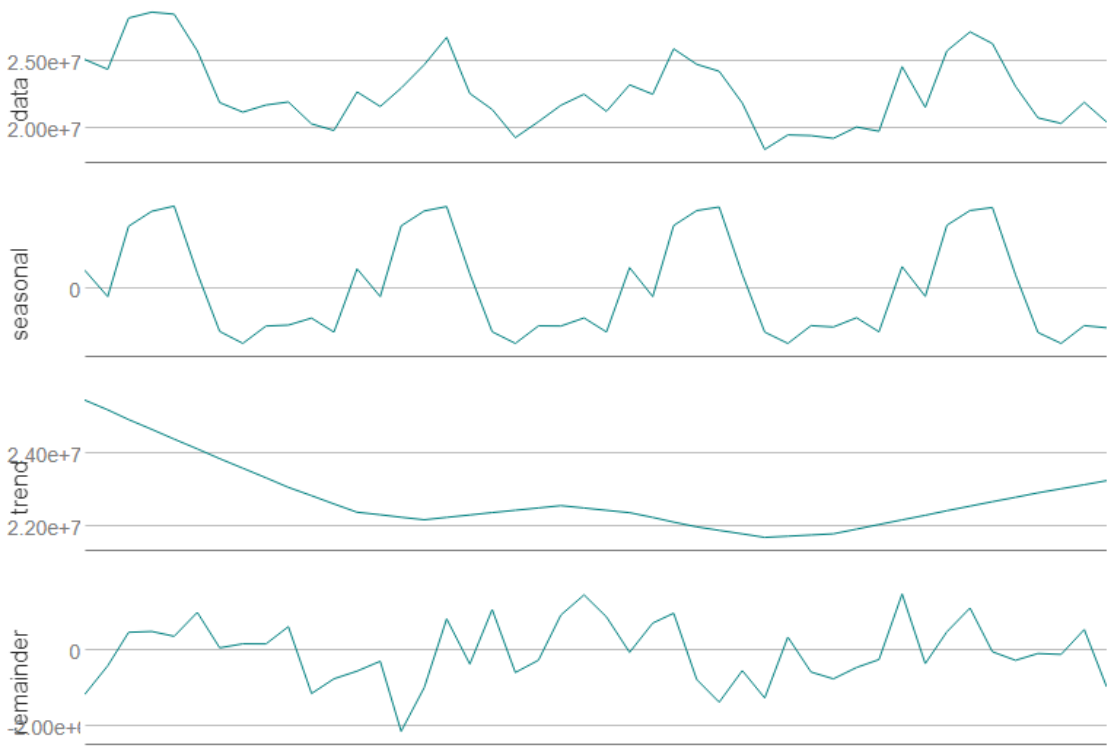
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

For forecasting, I used the ETS model. I got to this assumption to be the best option after comparing ETS with ARIMA and utilizing the TS Plot tool.

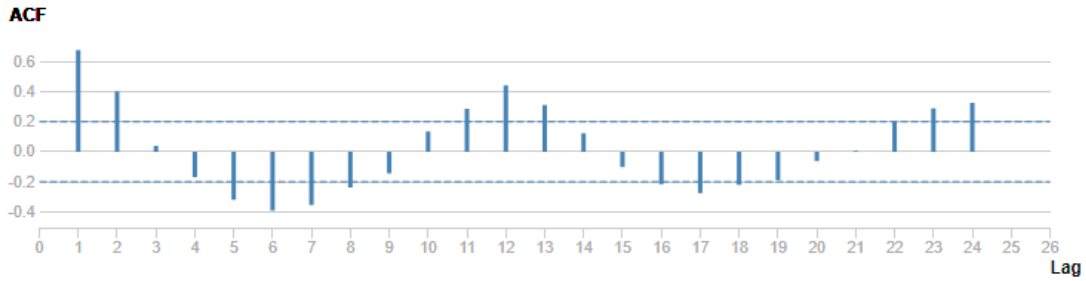
The error is multiplicative, the seasonal is also multiplicative, and the trend is nonexciting, according to the Decomposition plot. As a result, I went with the ETS model.



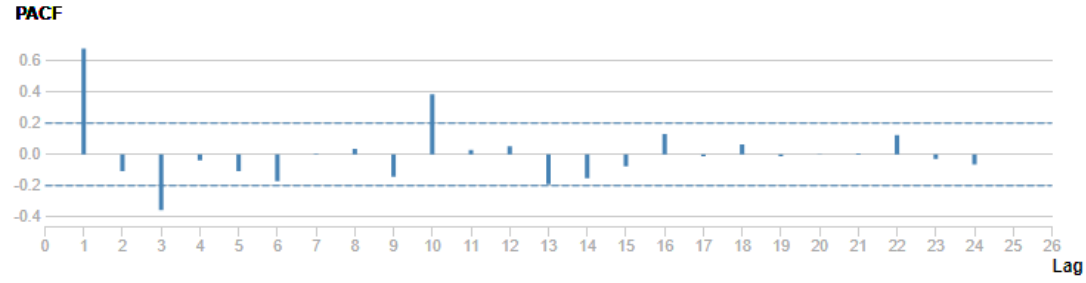
Decomposition Plot



Autocorrelation Function Plot



Partial Autocorrelation Function Plot



Summary of ARIMA Model Arima

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

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

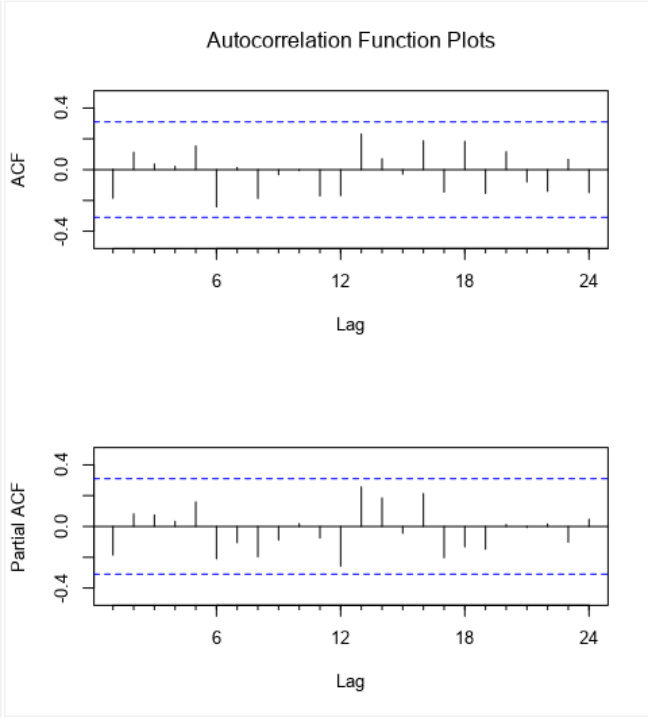
AIC	AICc	BIC
880.4445	881.4445	884.4411

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

Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

Plots



Comparison of Time Series Models

Actual and Forecast Values:

Actual	Arima
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

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

Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

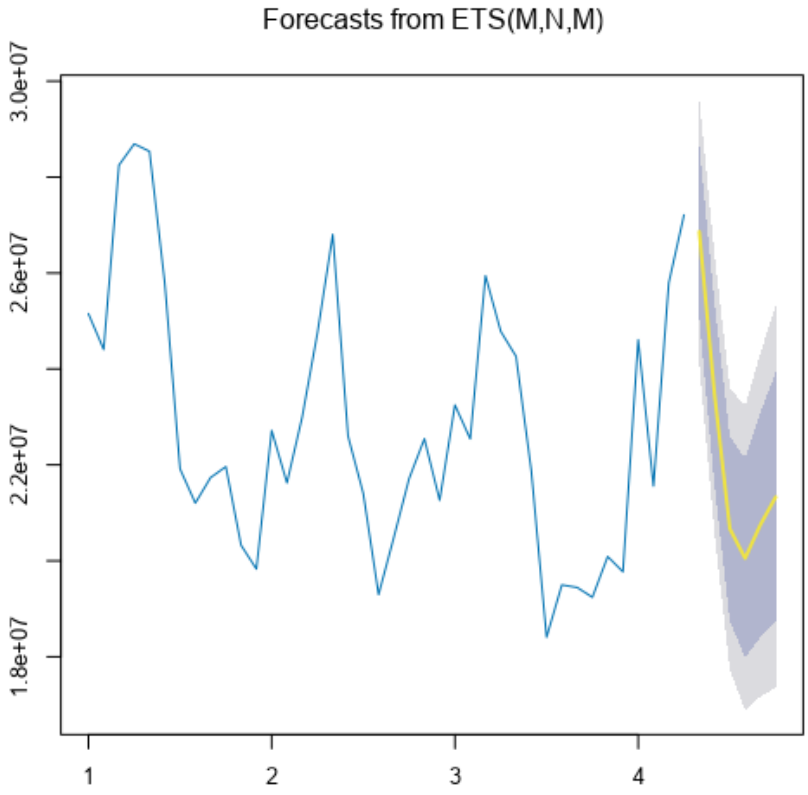
AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

Smoothing parameters:

Parameter	Value
alpha	0.674884
gamma	0.000203

Initial states:

State	Value
l	23146230.586012
s0	0.90906
s1	0.938619
s2	0.926304
s3	0.901291
s4	0.870972
s5	0.897637
s6	1.019225
s7	1.166556
s8	1.167388
s9	1.137259
s10	0.997793



The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

Comparison of Time Series Models

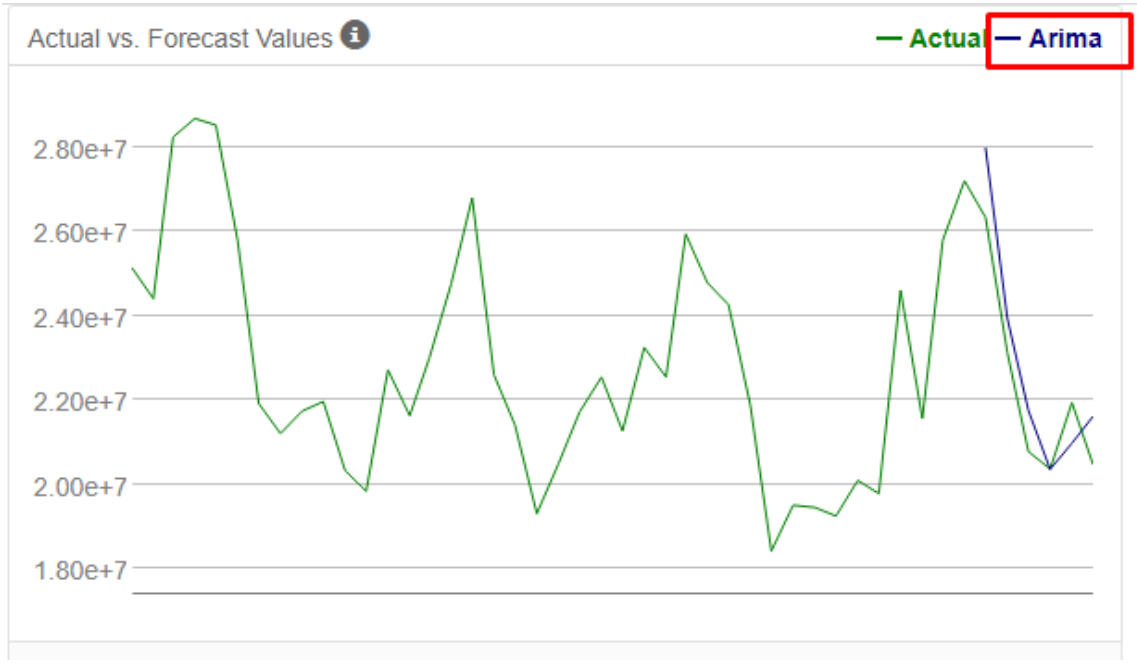
Actual and Forecast Values:

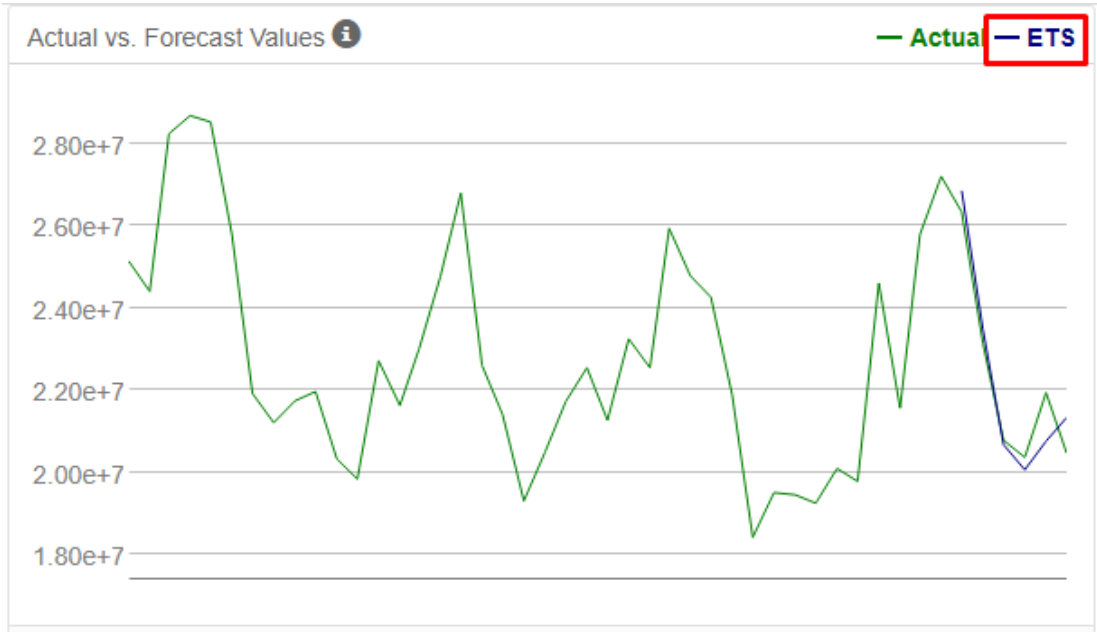
Actual	ETS
26338477.15	26860639.57444
23130626.6	23468254.49595
20774415.93	20668464.64495
20359980.58	20054544.07631
21936906.81	20752503.51996
20462899.3	21328386.80965

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

I can see from the Actual vs. Forecast Values for Arima and ETS plots that the ETS model's forecast values are the closest to the actual values than the Arima model's forecast values.





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.**

Year	Month	Forecast Integer	New Stores Sales
2016	1	21,829,060	2,603,262
2016	2	21,146,330	2,508,878
2016	3	23,735,687	2,989,458
2016	4	22,409,515	2,849,287
2016	5	25,621,829	3,224,711
2016	6	26,307,858	3,269,623
2016	7	26,705,093	3,288,334
2016	8	23,440,761	2,937,302
2016	9	20,640,047	2,606,592
2016	10	20,086,270	2,536,270
2016	11	20,858,120	2,631,293
2016	12	21,255,190	2,586,562

[Actual, Forecasting, and New Sales]