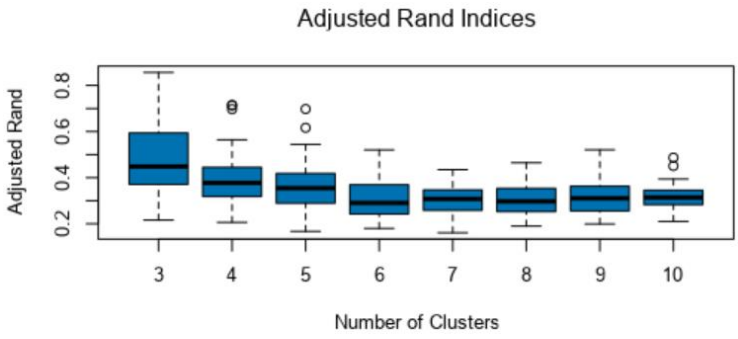
Predictive Analytics for Business Nanodegree

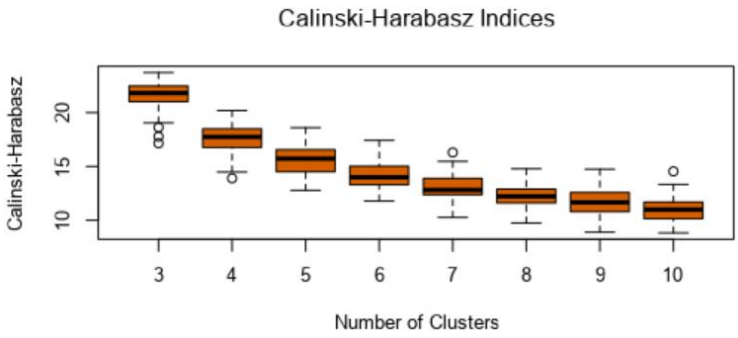
Project: Combining Predictive Techniques

Laila Hussain Alqawain

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. I arrived this number by using K-Centroids Cluster Analysis and K-Centroids Diagnostics Tools with K-Means Clustering Method.





From the Adjusted Rand Indices box plot and Calinski-Harabasz Indices box plot, we can see the Cluster 3 is the highest one of all of them.

2. How many stores fall into each store format?

Cluster	Size
1	23
2	29
3	33

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

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:

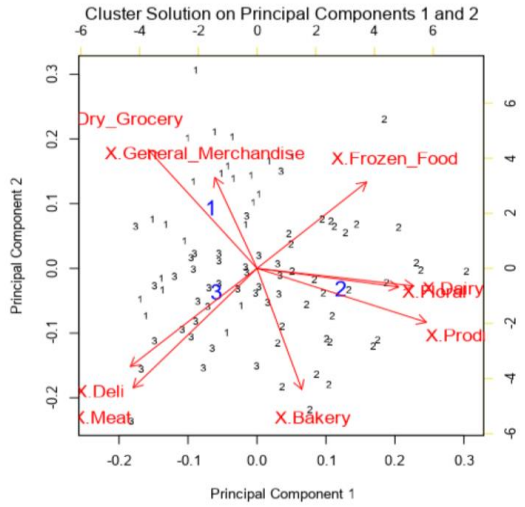
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

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

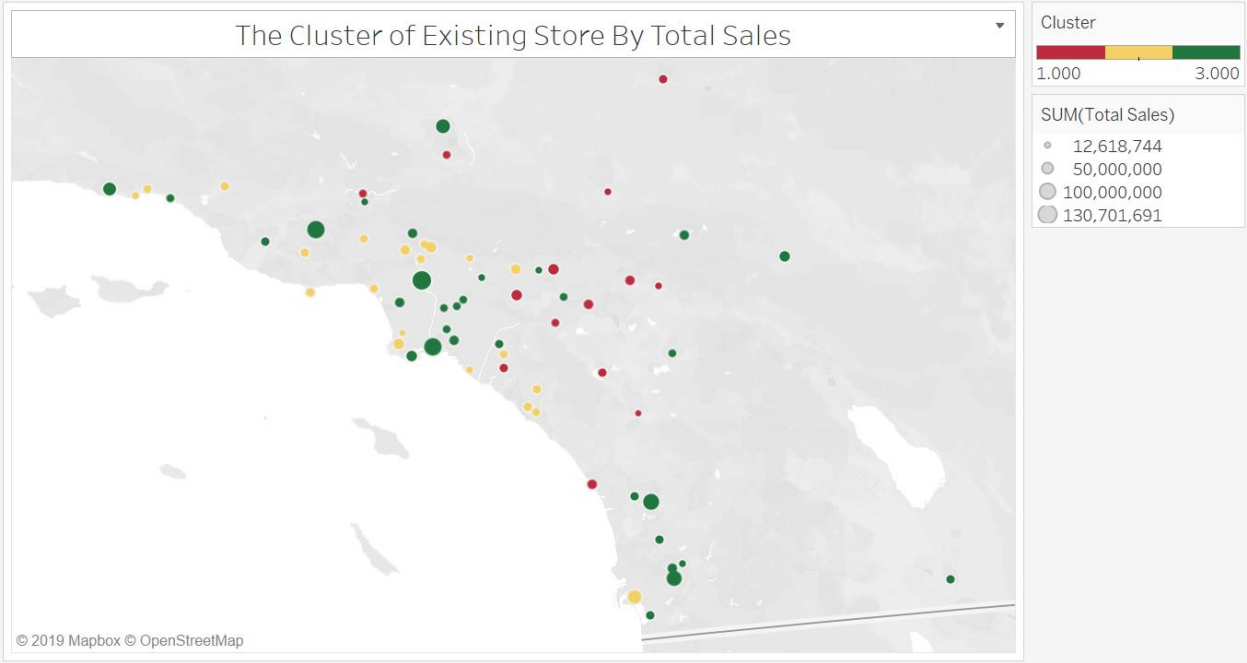
Plots



From the report of the K-Means Clustering, we can see the Cluster 1 has the lowest size which is 23 and lowest Max Distance which is 3.55. The Ave Distance and Separation for Cluster 1are between the Cluster 2 and the Cluster 3. The Cluster 2 with 29 of size and has the highest of Ave Distance and Separation. The Cluster 3 with the highest number of sizes which is 33 and Max Distance. The Cluster 3 has the lowest number of Separation which is 1.720.

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 Clustering]



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?
   used the Boosted Model to predict the best store format for the new stores after testing three models which are Decision Tree Model, Boosted Model, and Forest Model.

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

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_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

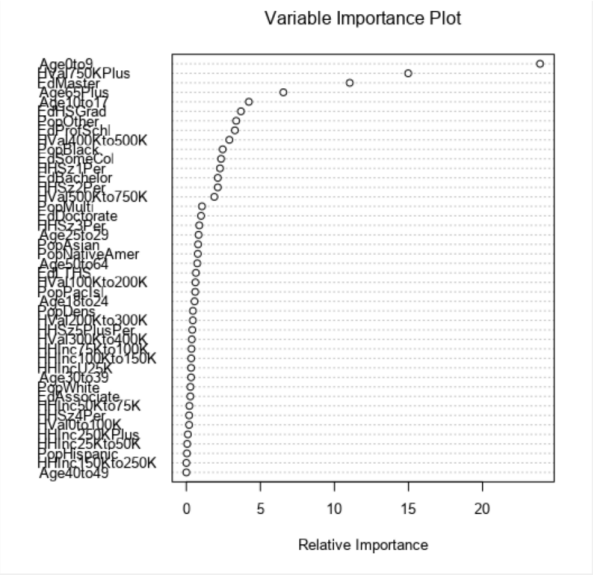
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

From Model Comparison Report, I can see the best model is the Boosted Model with 0.8235 Accuracy, 0.8889 F1, 1.0000 Accuracy_1, 1.0000 Accuracy_2, and 0.6667 Accuracy_3. Also, when I comparing between Confusion matrix of three models, I can see the Boosted Model is the best in Predicted_1 and Actual_1 which is 4, Predicted_2 and Actual_2 which is 4, and Predicted_3 and Actual_3 which is 6. So, I chose the Boosted Model to predict the best store format for the new stores.

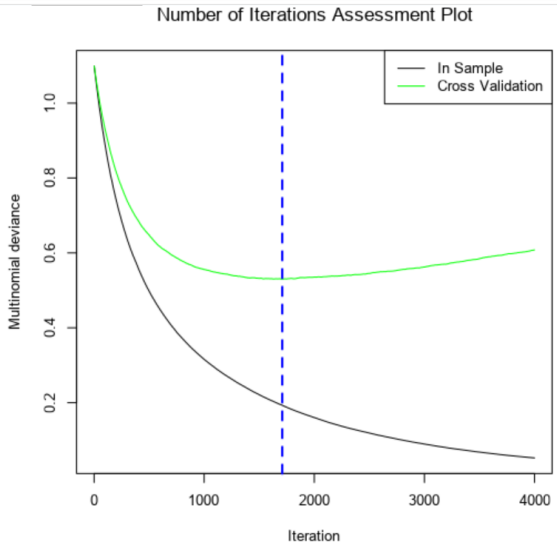
Report for Boosted Model

Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1709



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.



The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specfied assessment criteria (cross validation, the use of a test sample, or out-of-bag prediction).

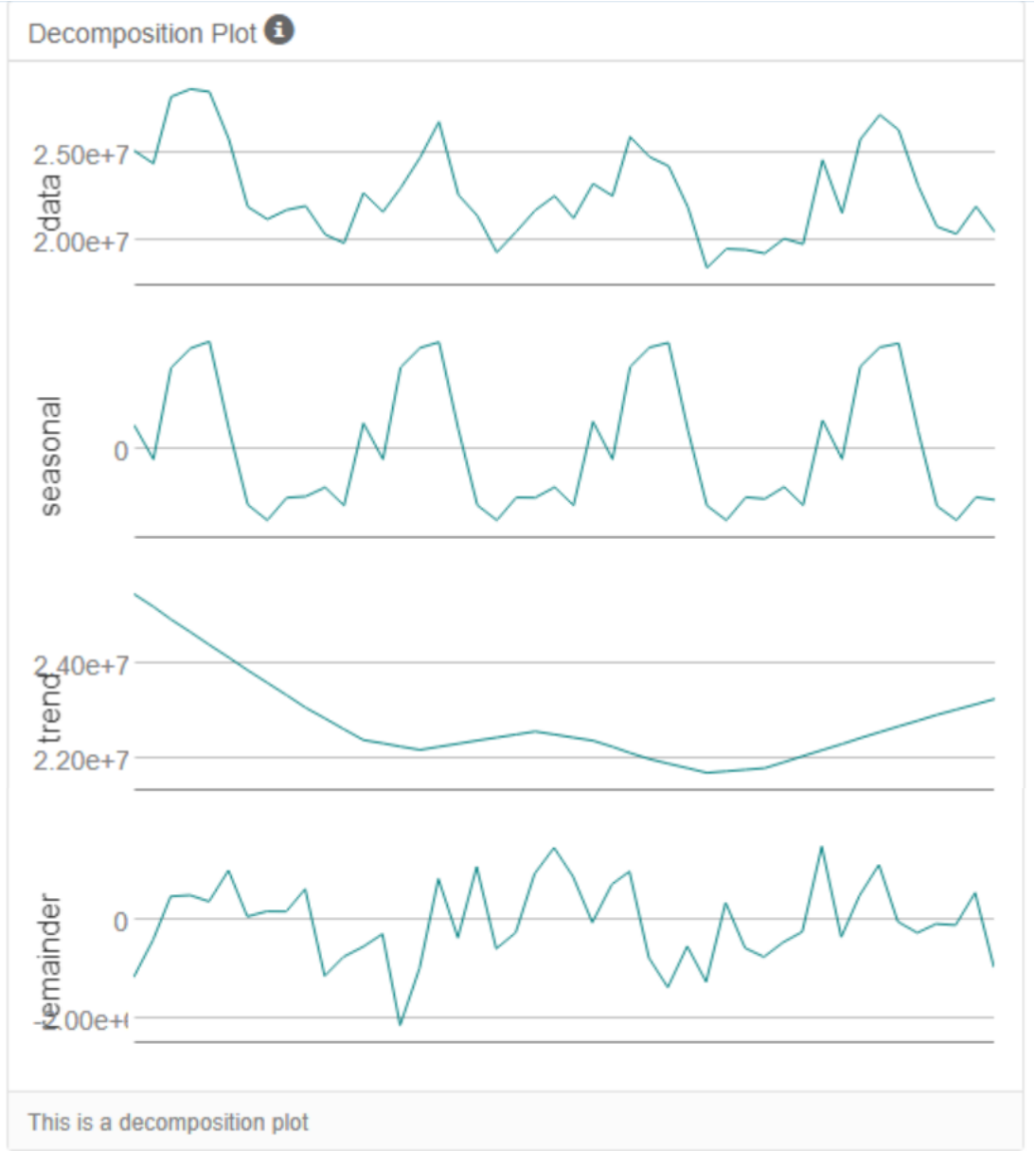
From Report for Boosted Model, the three best important variables are Age0to9, HVal750KPlus, and EdMaster.

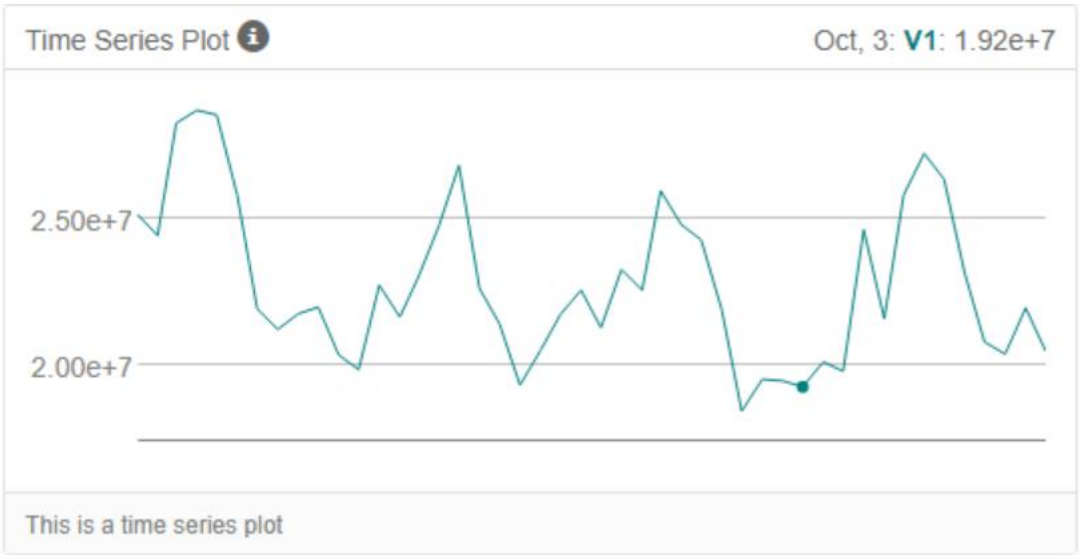
2. What format do each of the 10 new stores fall into? Please fill in the table below.

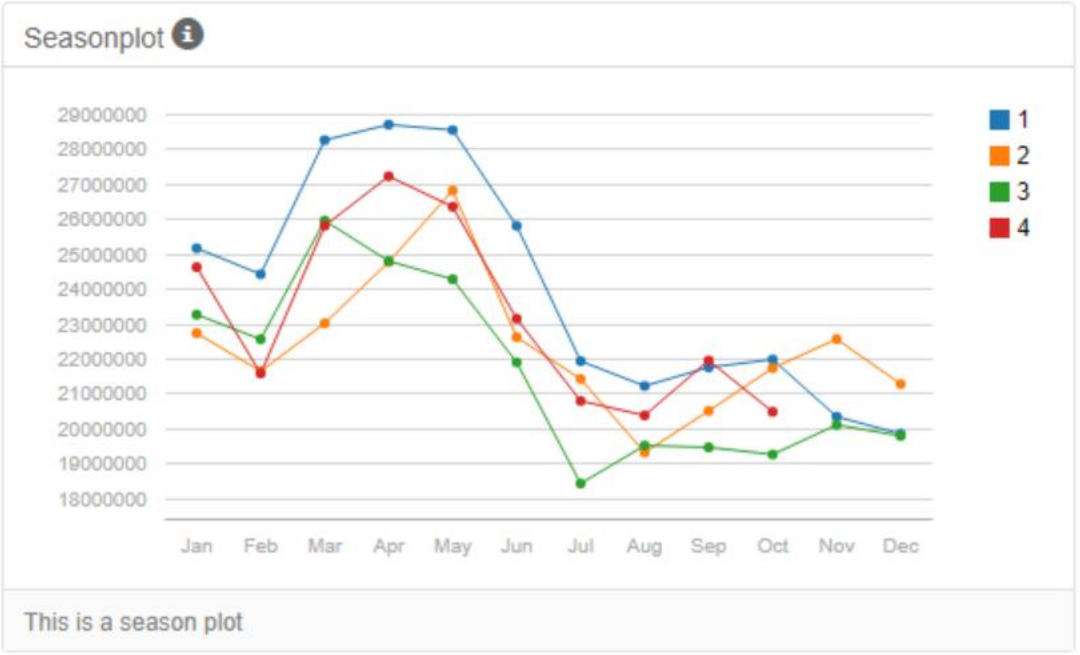
Store	Cluster
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

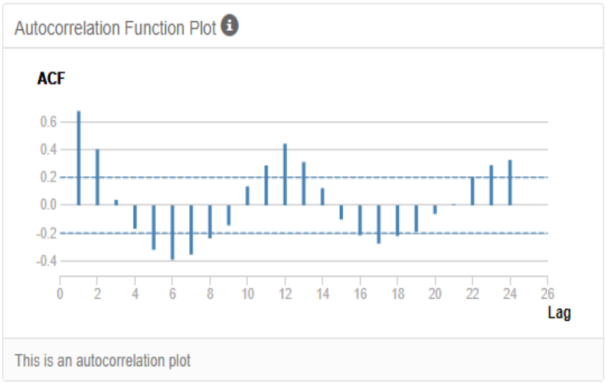
Task 3: Predicting Produce Sales

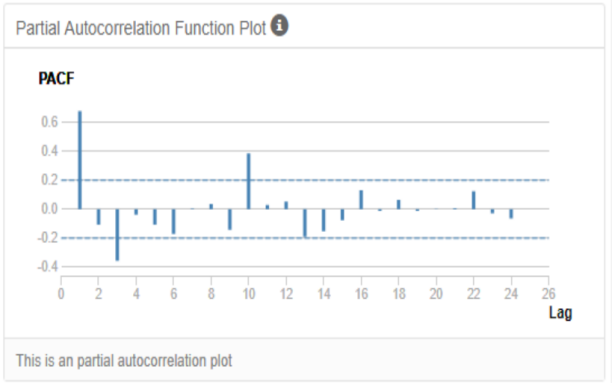
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS (a, m, n) or ARIMA (ar, i, ma) notation. How did you come to that decision?
   I used ETS model for forecast. I came to this decision after comparing between ETS and ARIMA and using TS Plot tool.











From the Decomposition plot, I see the error is multiplicative, the seasonal is also multiplicative, the trend is nonexciting. So, I chose the ETS model.

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

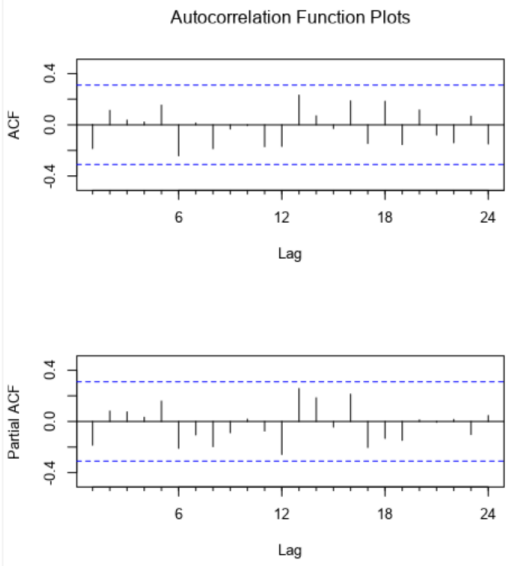
Information Criteria:

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



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:

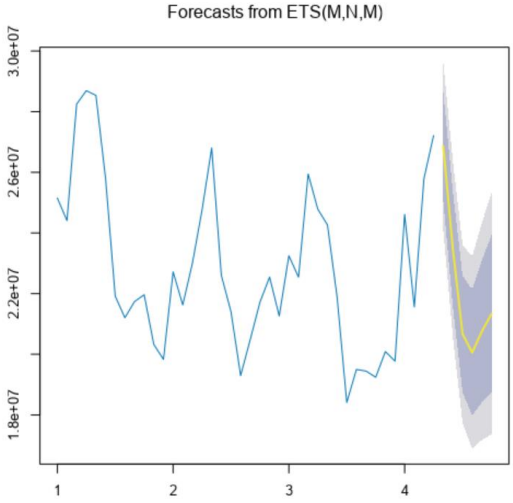
AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

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.

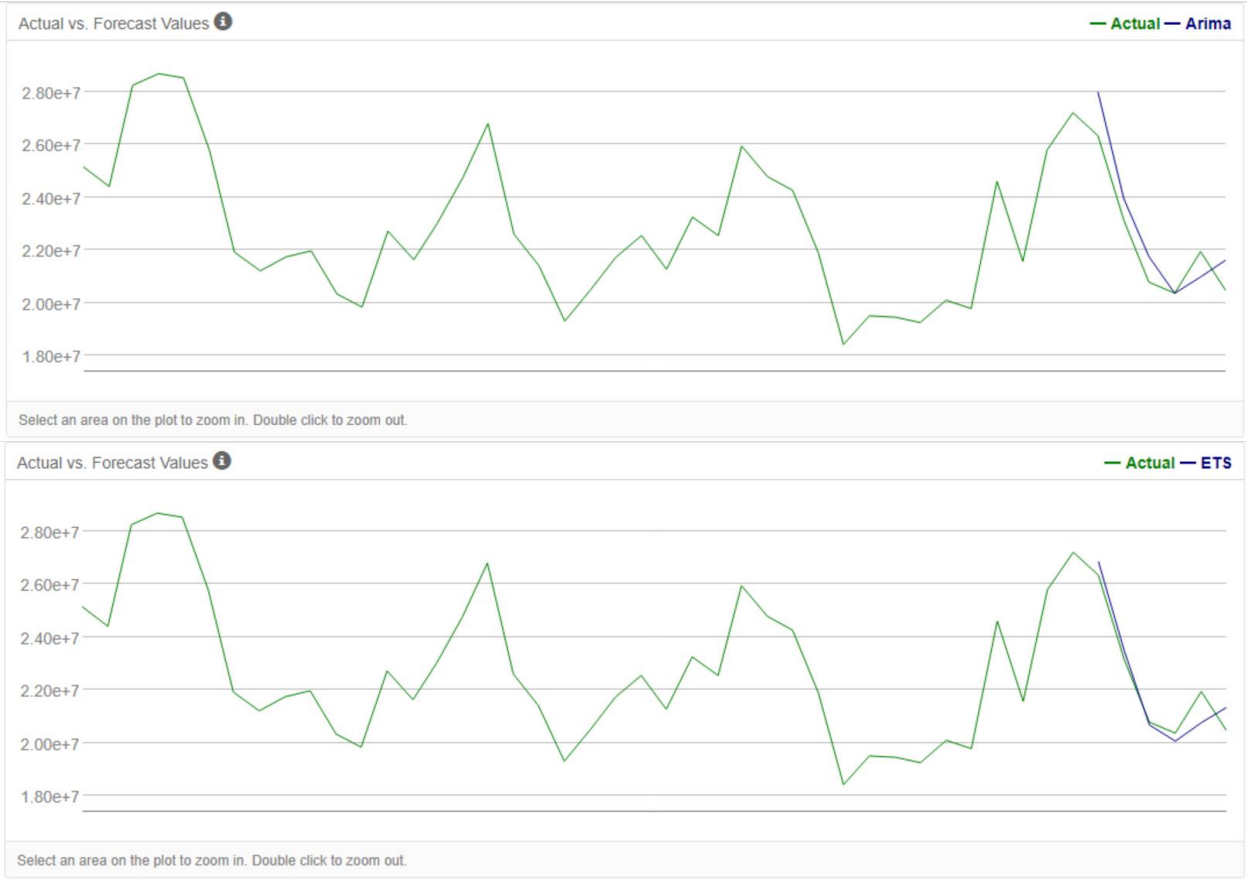
Comparison of Time Series Models

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



From the Actual vs. Forecast Values for Arima and ETS plots above, I can see the forecast values by the ETS model is most near to the actual values than the forecast values by the Arima model.

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	21829060	2603262
2016	2	21146330	2508878
2016	3	23735687	2989458
2016	4	22409515	2849287
2016	5	25621829	3224711
2016	6	26307858	3269623
2016	7	26705093	3288334
2016	8	23440761	2937302
2016	9	20640047	2606592
2016	10	20086270	2536270
2016	11	20858120	2631293
2016	12	21255190	2586562

[Visualization of Forecasting]