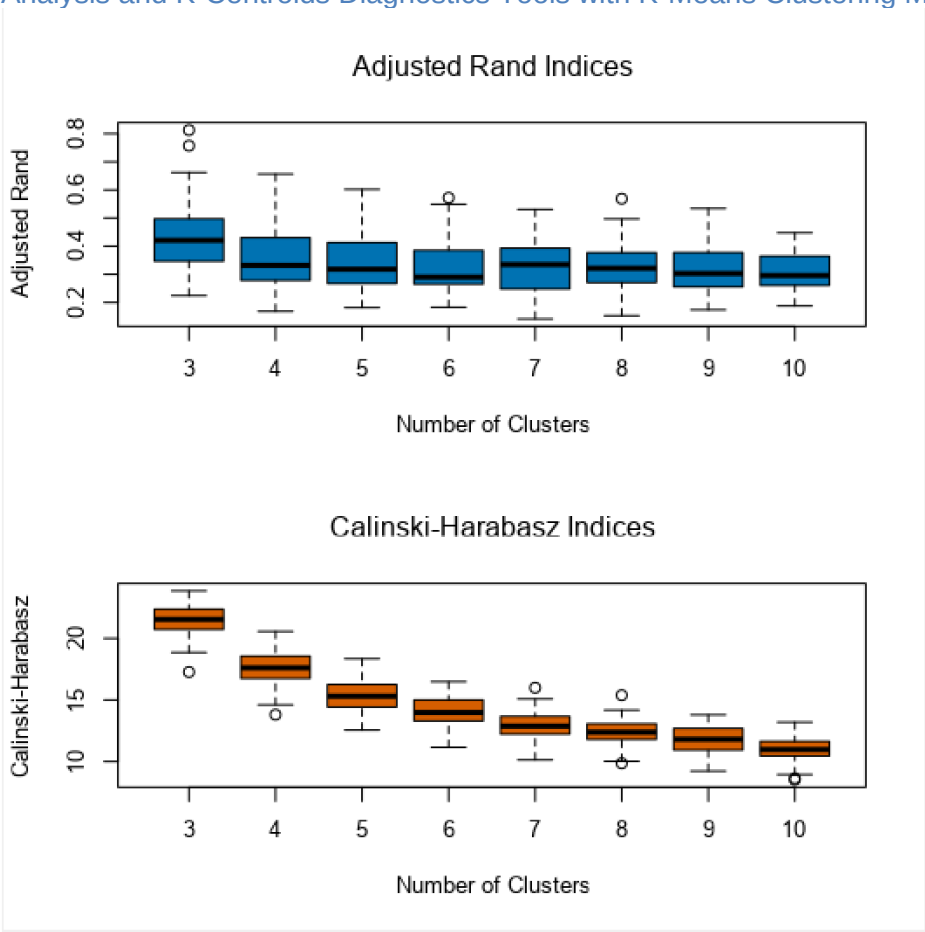
Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

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	25
2	35
3	25

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"))

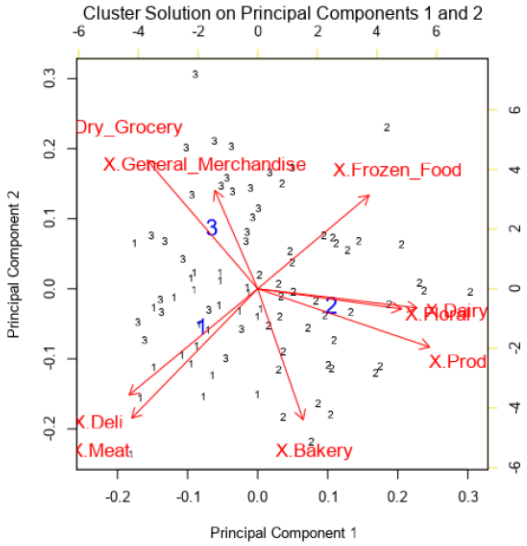
Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	X.Dry_Grocery	X.Dairy	X.Frozen_Food	X.Meat	X.Produce	X.Floral	X.Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178481

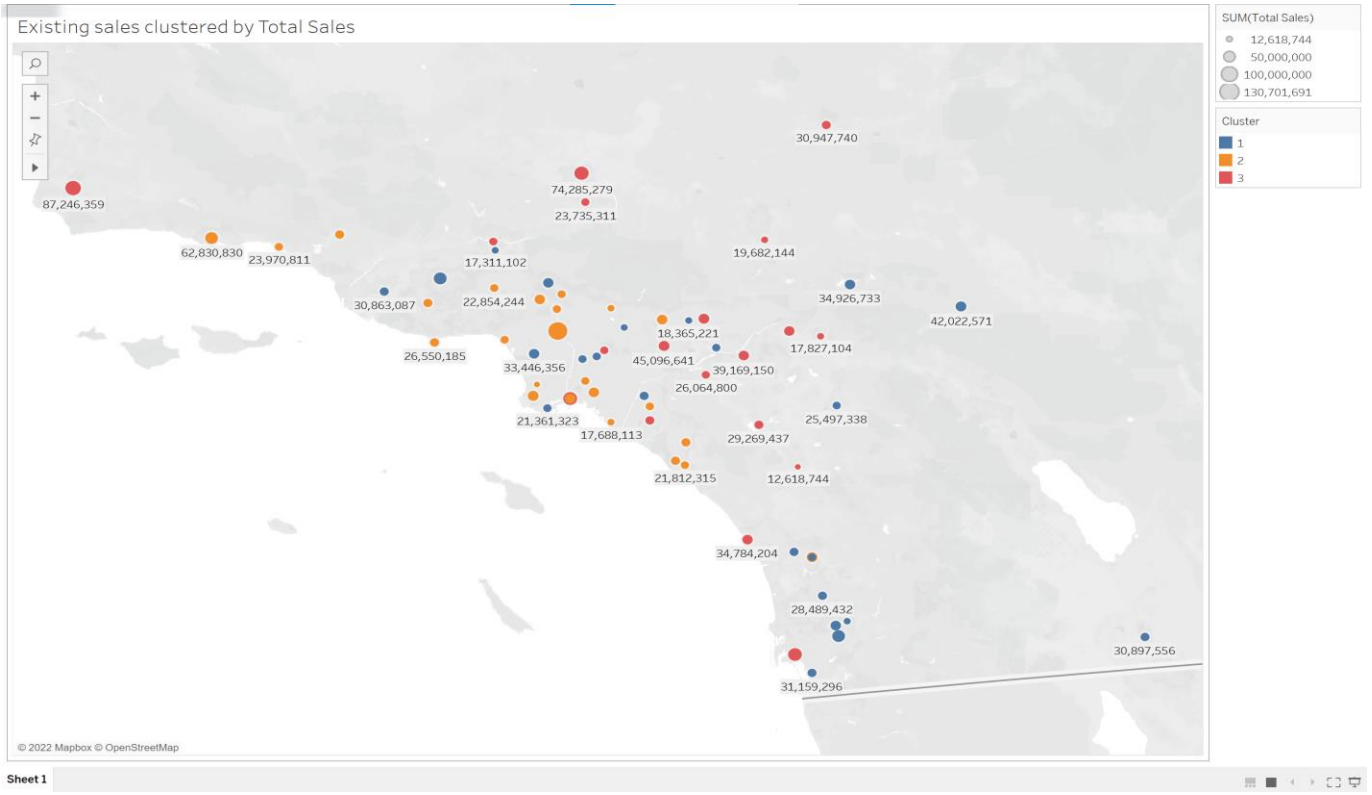
	X.Bakery	X.General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432



From the report of the K-Means Clustering, we can see the Cluster 3 has the lowest size which is 25 and lowest Max Distance which is 3.58. The Ave Distance and Separation for Cluster 3 are between the Cluster 1 and the Cluster 2. The Cluster 2 with 35 of size and has the highest of Ave Distance.The Cluster 2 has highest Separation. The Cluster 2 with the highest number of sizes which is 35. The Cluster 3 has the lowest number of Separation which is 1.725

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



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.)

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	0.6471	0.6381	0.8000	0.7143	0.4000
Forest_Model	0.8824	0.8857	0.8000	0.8571	1.0000
Boosted_Model	0.8235	0.8190	0.8000	0.8571	0.8000

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 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	6	0
Predicted_3	1	1	4

Confusion matrix of Decision_Tree_Model

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

Confusion matrix of Forest_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	1	6	0
Predicted_3	0	1	5

From the Model Comparison Report, I can see that the best model is Forest Model with accuracy of 0.8824, 0.8857 as F1, 0.8 as Accuracy_1, 0.8571 as Accuracy_2 and 1 as Accuracy_3.
Also when I compared between Confusion Matrix of three models, Boosted Model is the best in Predicted_1 and Actual_1 which is 4. Forest Model and Boosted Model are best for Predicted_2 and Actual_2 which is 6. Forest Model is best for Predicted_3 and Actual_3 which is 5. So, I choose Forest Model to predict the best store format for the new stores.

Report

Basic Summary

Call:
randomForest(formula = Cluster ~ Age0to9 + Age10to17 + Age18to24 + Age25to29 + Age30to39 + Age40to49 + Age50to64 + Age65Plus + EdLTHS + EdHSGrad + EdSomeCol + EdAssociate + EdBachelor + EdMaster + EdProfSchl + EdDoctorate + HHSz1Per + HHSz2Per + HHSz3Per + HHSz4Per + HHSz5PlusPer + HHIncU25K + HHInc25Kto50K + HHInc50Kto75K + HHInc75Kto100K + HHInc100Kto150K + HHInc150Kto250K + HHInc250KPlus + PopAsian + PopBlack + PopHispanic + PopMulti + PopNativeAmer + PopOther + PopPacIsl + PopWhite + HVal0to100K + HVal100Kto200K + HVal200Kto300K + HVal300Kto400K + HVal400Kto500K + HVal500Kto750K + HVal750KPlus + PopDens, data = the.data, ntree = 500, replace = TRUE)

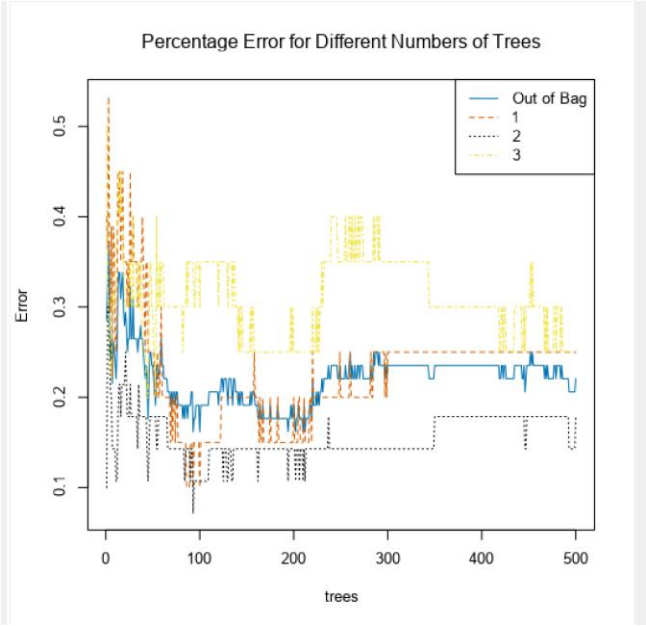
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 6

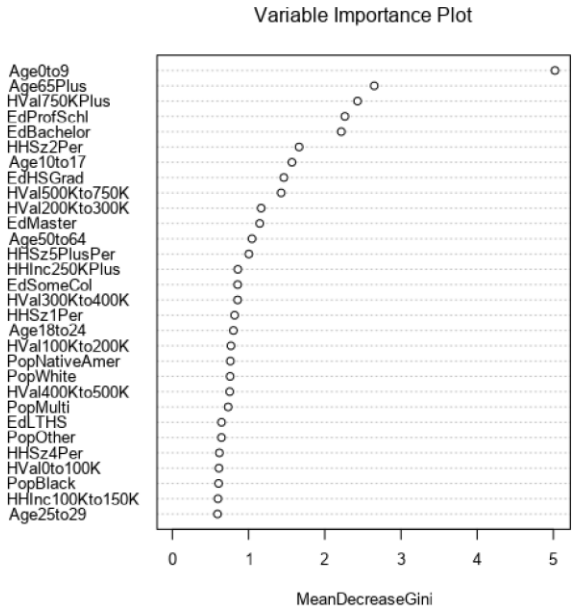
OOB estimate of the error rate: 22.1%

Confusion Matrix:

	Classification Error	1	2	3
1	0.25	15	3	2
2	0.179	5	23	0
3	0.25	4	1	15

Plots





From Report for Forest Model, the three best important variables are Age0to9, Age65Plus, and HVal750KPlus.

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

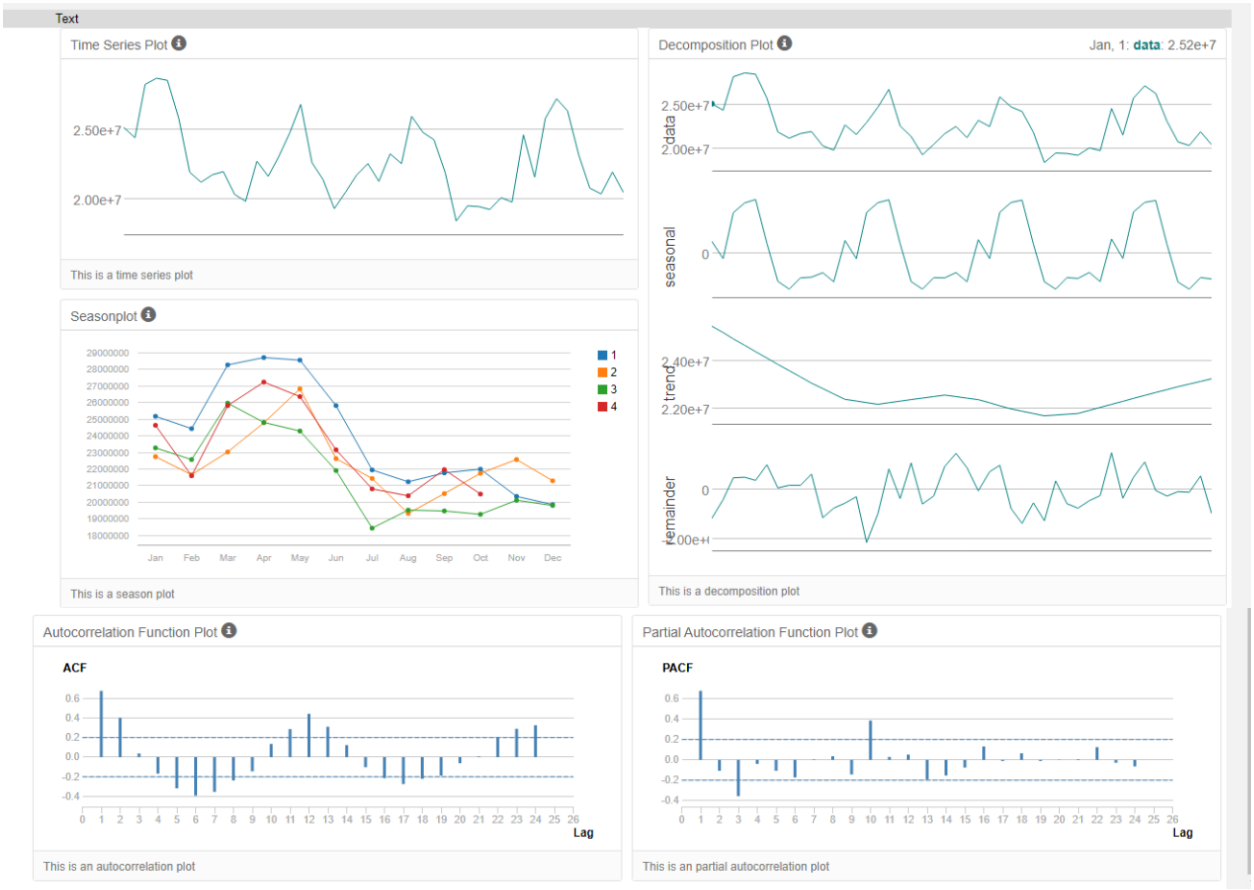
Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2

S0091	3
S0092	2
S0093	3
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?

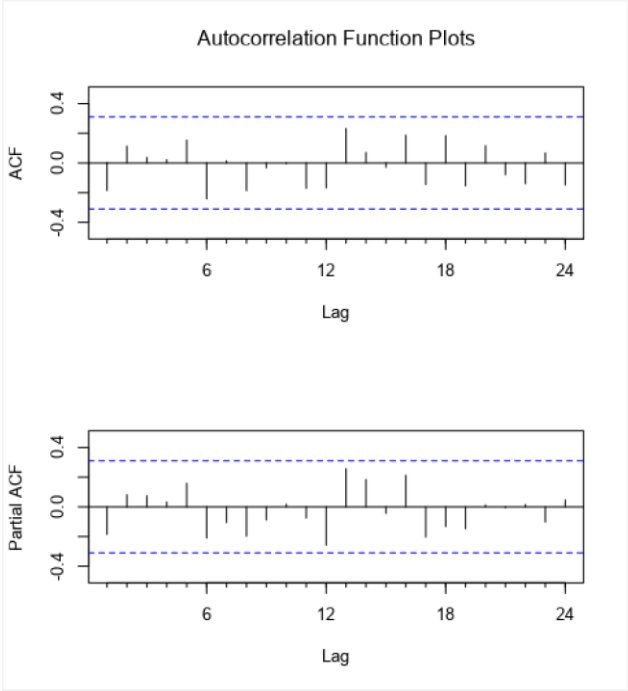
I used ETS model for forecast. I came to this decision after comparing between ETS and ARIMA and using TS Plot tool



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.

Record	Report
1	**Summary of ARIMA Model Arima**
2	Method: ARIMA(1,0,0)(1,1,0)[12]
3	Call: auto.arima(Sum_Produce)
4	Coefficients: ar1 sar1 Value 0.79852 -0.700441 Std Err 0.126448 0.140181
5	sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224
6	Information Criteria: AIC AICc BIC 880.4445 881.4445 884.4411
7	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
8	Ljung-Box test of the model residuals: Chi-squared = 15.0973, df = 12, p-value = 0.23616

Plots



Report

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:

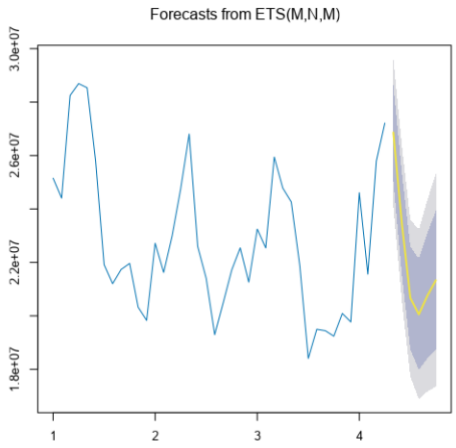
AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

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.

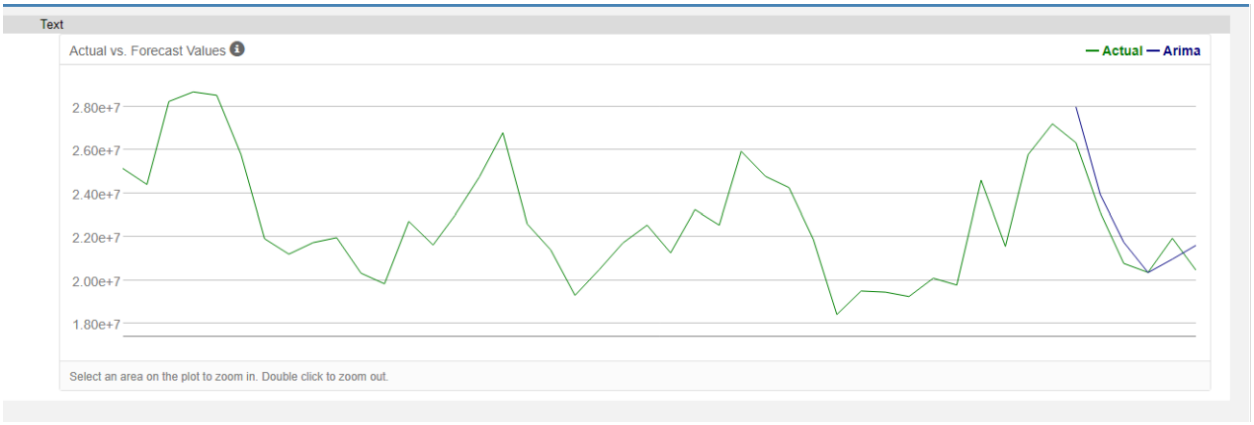
Actual and Forecast Values:

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

Text

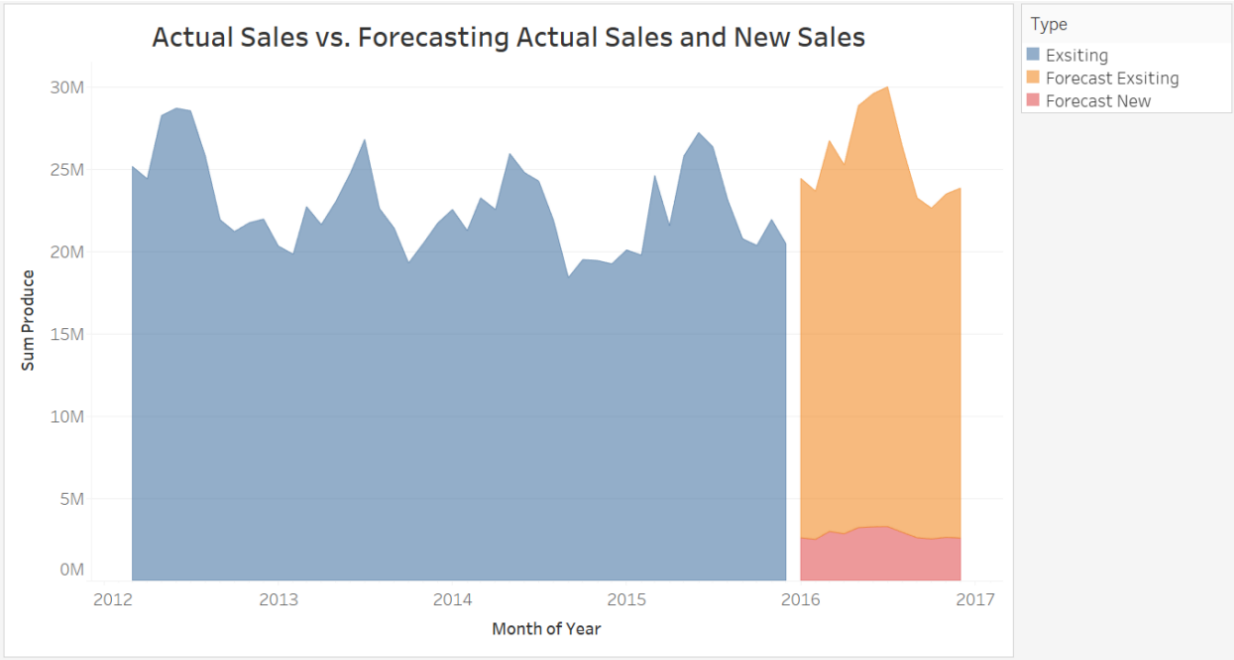


From the Actual vs. Forecast Values for Arima and ETS plots above, I can see the forecast values by the ETS model is most near to the actual values than the forecast values by the Arima model

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.

A	B	C	D
Year	Month	Forecast_Integer	New_Stores_Sales
2016	1	21829060	2493697
2016	2	21146330	2405584
2016	3	23735687	2879417
2016	4	22409515	2720393
2016	5	25621829	3089903
2016	6	26307858	3139497
2016	7	26705093	3155160
2016	8	23440761	2807733
2016	9	20640047	2482456
2016	10	20086270	2420097
2016	11	20858120	2510816
2016	12	21255190	2480120

Forecasting Visualization



Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.