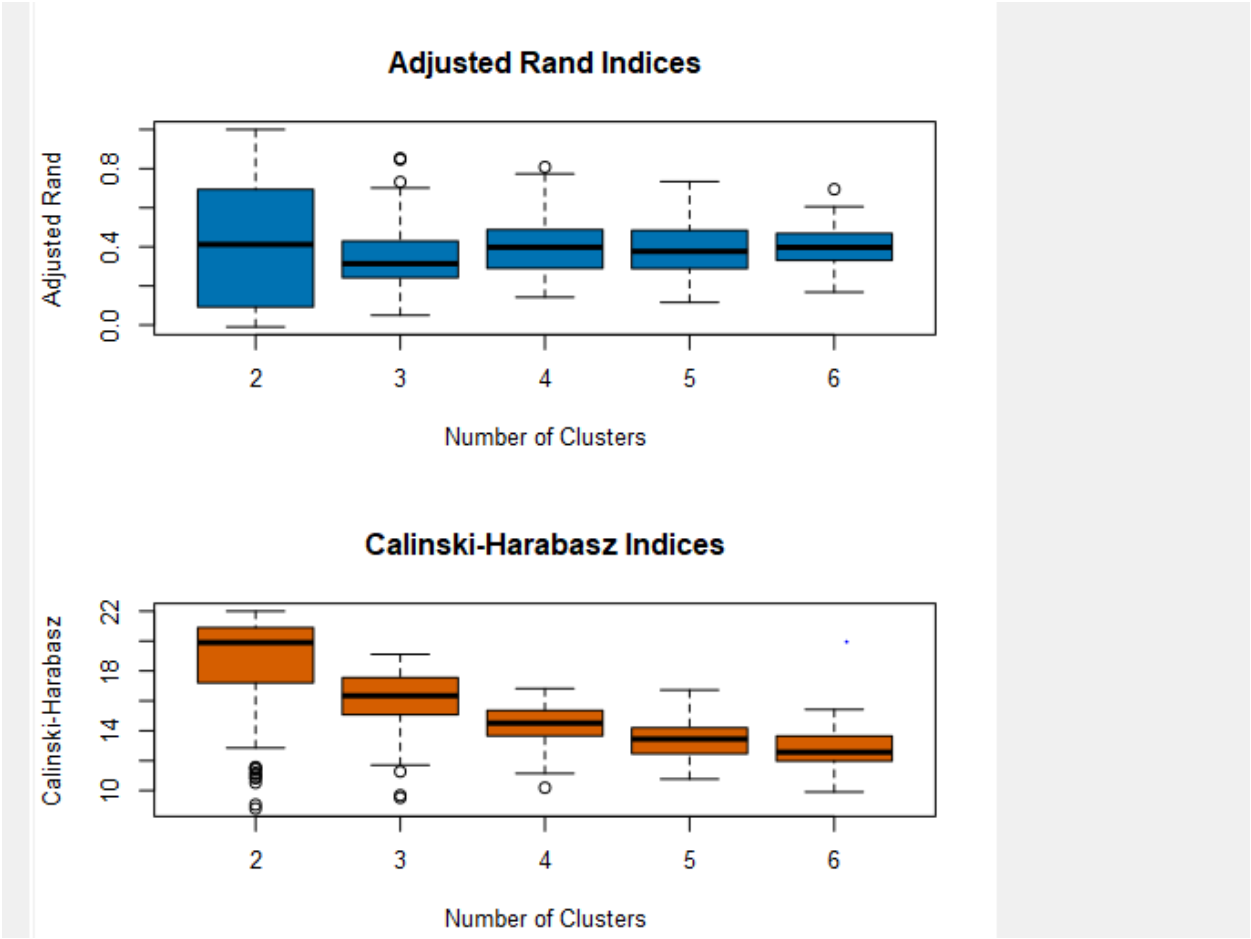
Project: Predictive Analytics Capstone

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?

- Have calculated the total sales and calculated the Percentage sales of each category as "[Sum_Dry_Grocery]/[Total_Sales]*100"
- Below is the AR index and Number of cluster representation , **Cluster 3** have high median value and equal spread between first quartile and third quartile range , So Cluster number 3 has been chosen



Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.009603	0.049806	0.141931	0.116126	0.168334
1st Quartile	0.099849	0.24398	0.290669	0.28925	0.331267
Median	0.411974	0.313552	0.397509	0.376582	0.397359
Mean	0.397078	0.366329	0.395003	0.398799	0.397648
3rd Quartile	0.684312	0.422888	0.487787	0.482849	0.466522
Maximum	1	0.85296	0.808194	0.73384	0.694723

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	8.793333	9.51835	10.1893	10.76523	9.907697
1st Quartile	17.287672	15.06886	13.67248	12.47522	11.988572
Median	19.893033	16.34166	14.514	13.44709	12.565256
Mean	18.435698	16.09378	14.40569	13.39308	12.730988
3rd Quartile	20.893269	17.54566	15.33504	14.1877	13.645308
Maximum	21.992647	19.11366	16.8206	16.72381	15.424901

2. How many stores fall into each store format?

Record | Report

Summary Report of the K-Means Clustering Solution Store_Cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Per_Grocery + Per_Dairy + Per_Frozen + Per_Meat + Per_Produce + Per_Floral + Per_Deli + Per_Bakery + Per_GenMer, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

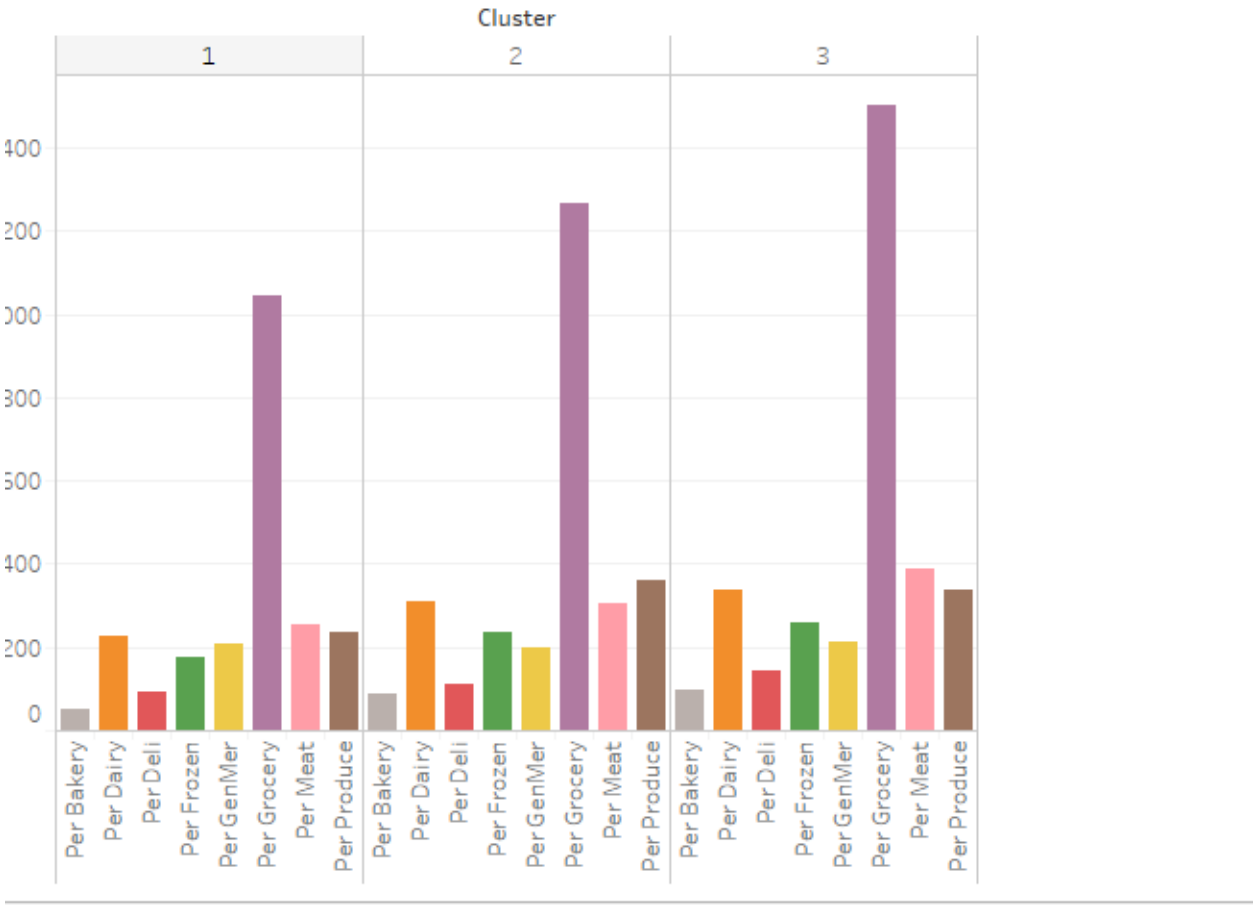
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

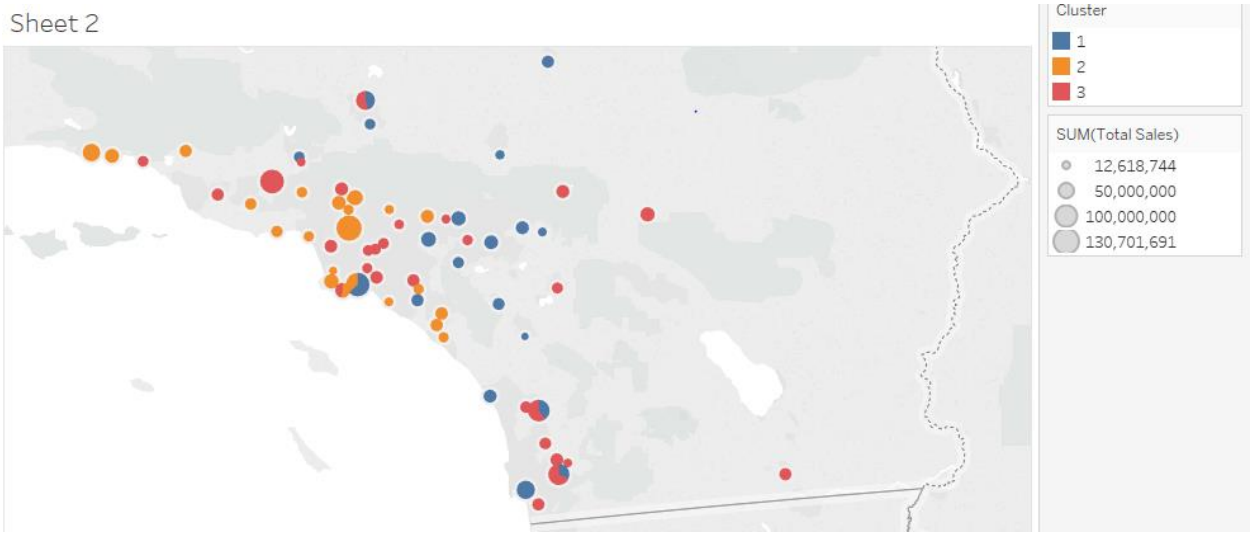
Cluster 1 has 23 stores, cluster 2 has 29 stores, Cluster 3 has 33 stores

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

Each of the cluster differ from each other in the percentage of Grocery Sales



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? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

Below is the model comparison Report, All the three models have same accuracy value but Boosted Model has higher F1 value than other two models , hence this model was chosen

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.8235	0.8426	0.7500	1.0000	0.7778
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

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	4	2
Predicted_3	0	0	6

Confusion matrix of Decision_Tree

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1

Confusion matrix of Decision_Tree

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

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

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

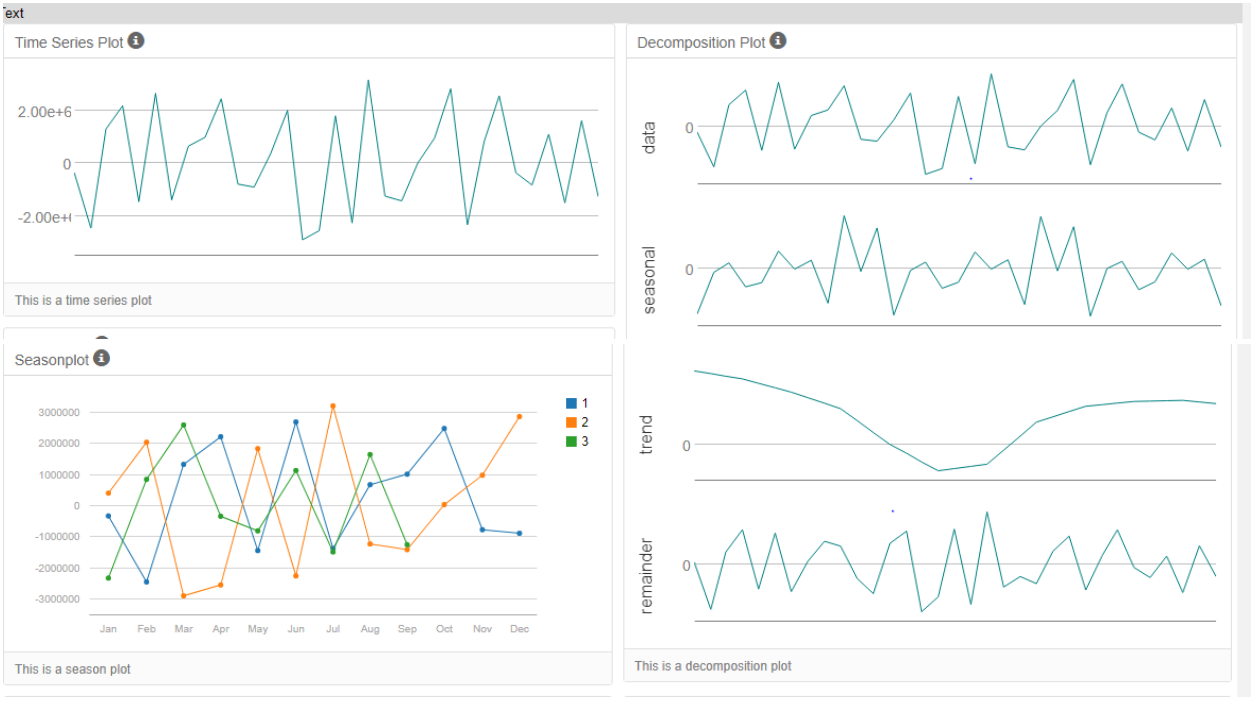
Store Number	Segment
S0086	3
S0087	2
S0088	1

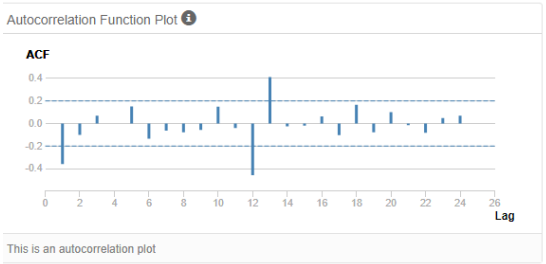
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

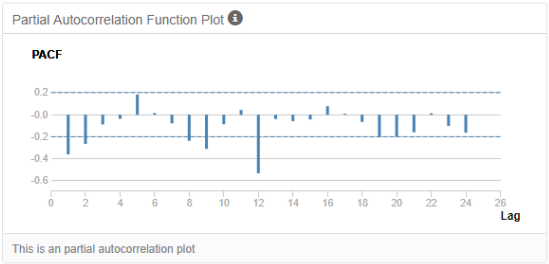
Task 3: Predicting Produce Sales

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?

Series has been made stationary by taking one seasonal difference and one difference (D= 1 , d = 1)
Below is the representation from the TS Plot
Season and Error type shows multiplicative behavior
Trend shows unequal variation, so I have kept Trend to none

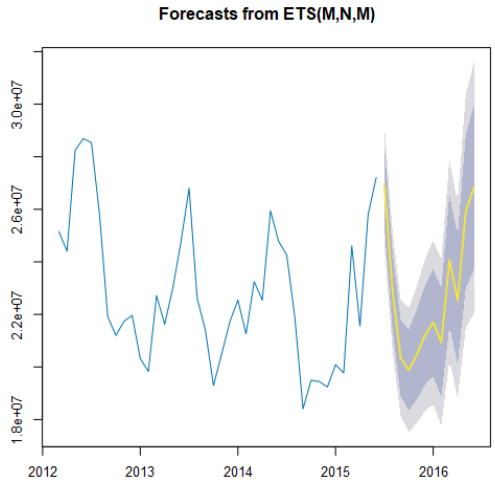






Based on the above after lag 2 it drops to 0 . (q =2)
There are no seasonal lags which are observed (so Q = 0)

ETS(M,N,M) –



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.

Summary of Time Series Exponential Smoothing Model ETS1

Method:
ETS(M,N,M)

In-sample error measures:

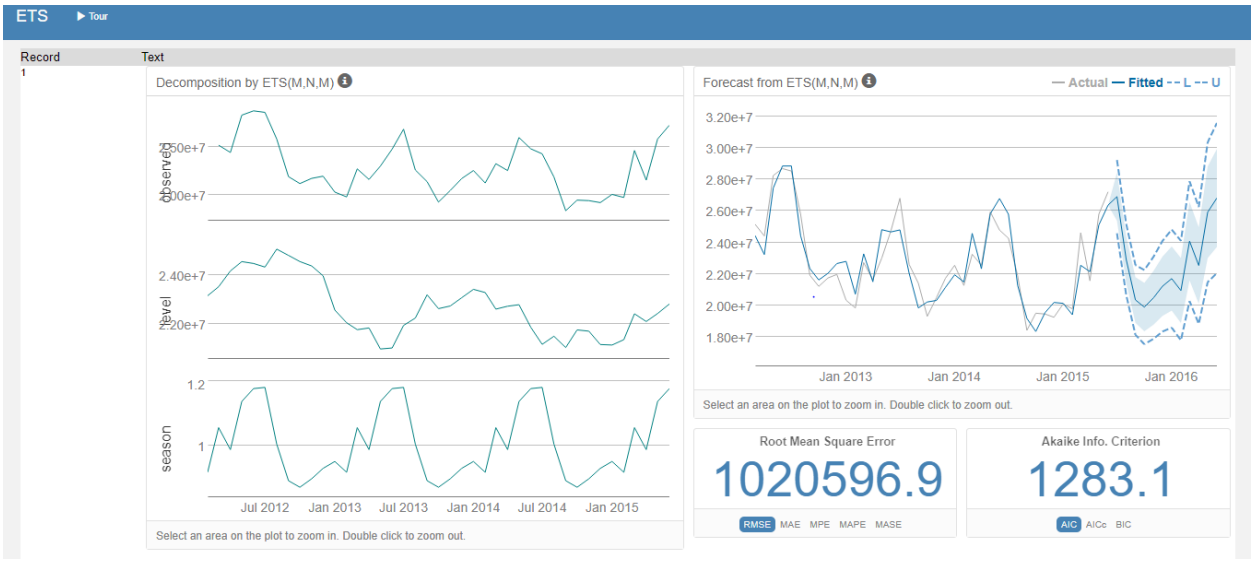
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

Smoothing parameters:

Parameter	Value
alpha	0.539196
gamma	0.000128



ARIMA(0,1,2)(0,1,0)

Report

Summary of ARIMA Model Arima_1

Method: ARIMA(0,1,2)(0,1,0)[12]

Call:
Arima(Sum_Produce, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1	ma2
Value	-0.415471	-0.054116
Std Err	0.219958	0.234438

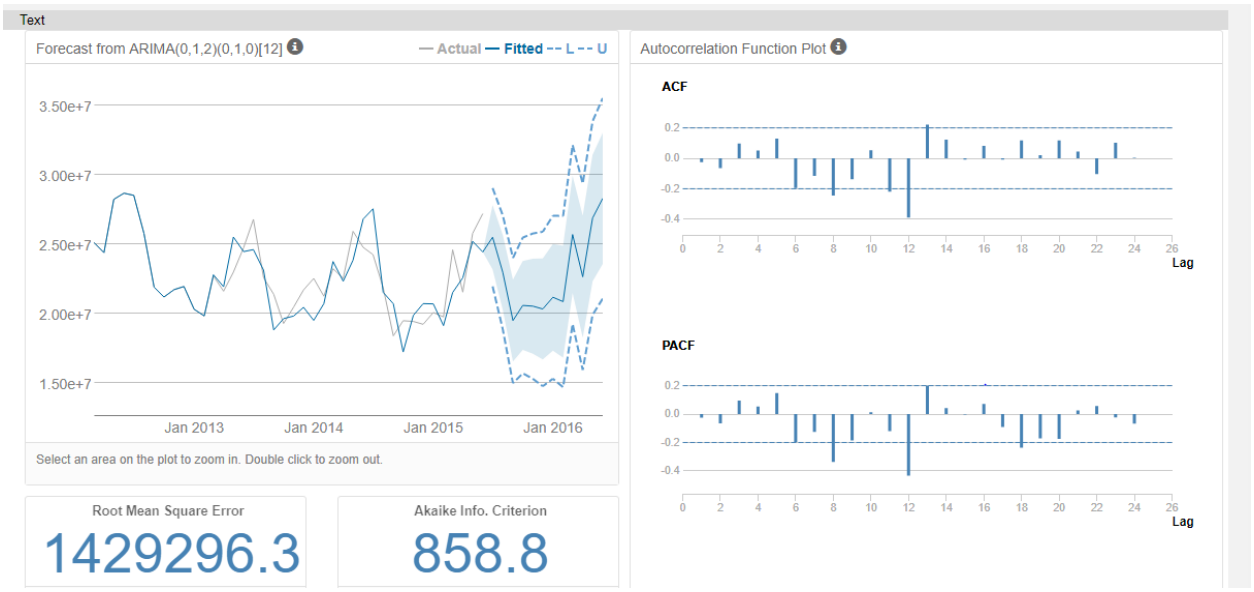
sigma^2 estimated as 3268620653560.66: log likelihood = -426.38872

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961



Validation Dataset Comparison between the two models

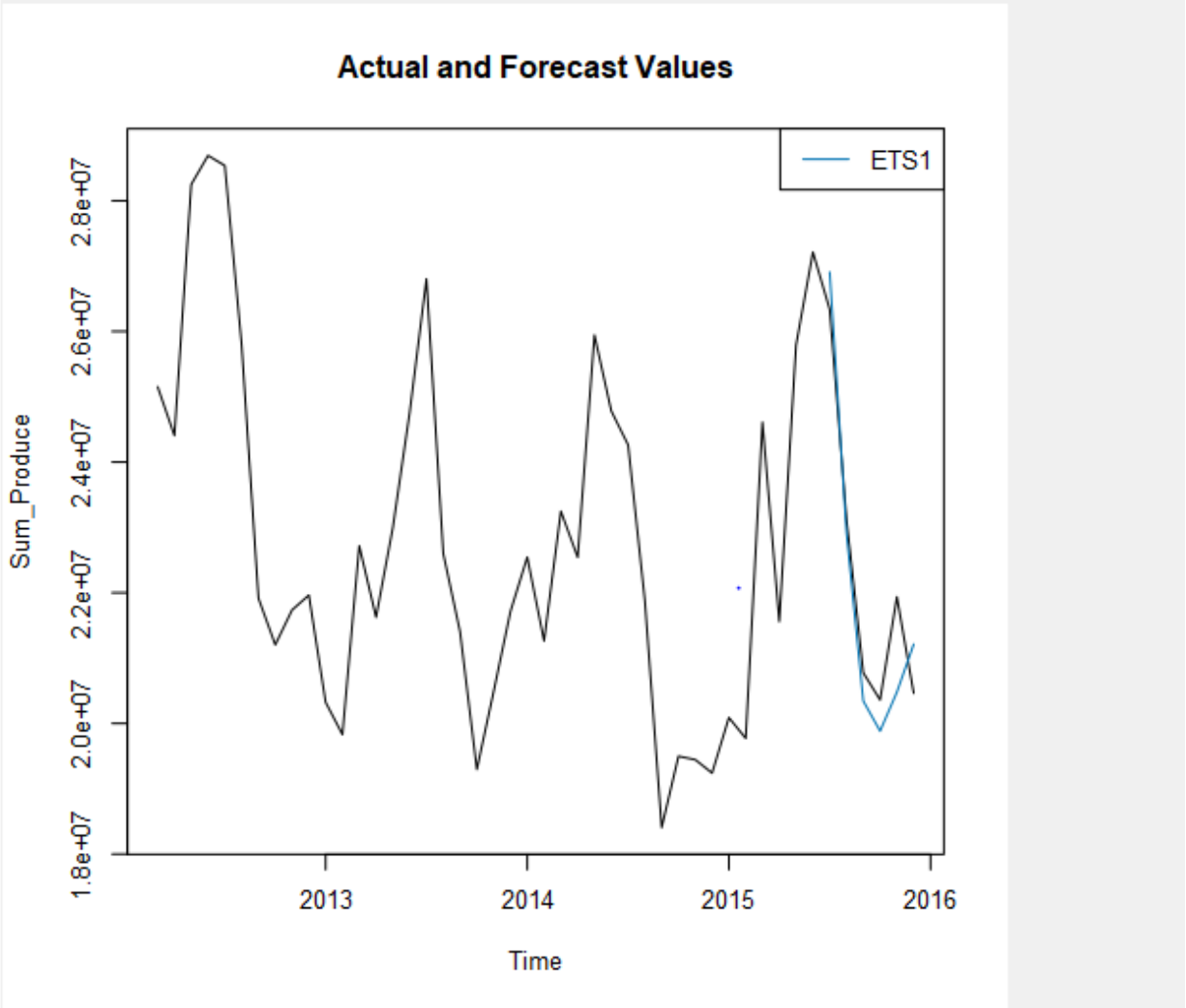
Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS1
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS1	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822


Actual and Forecast Values
ETS1
Sum_Produce
1.8e+07
2.0e+07
2.2e+07
2.4e+07
2.6e+07
2.8e+07
2013
2014
2015
2016
Time

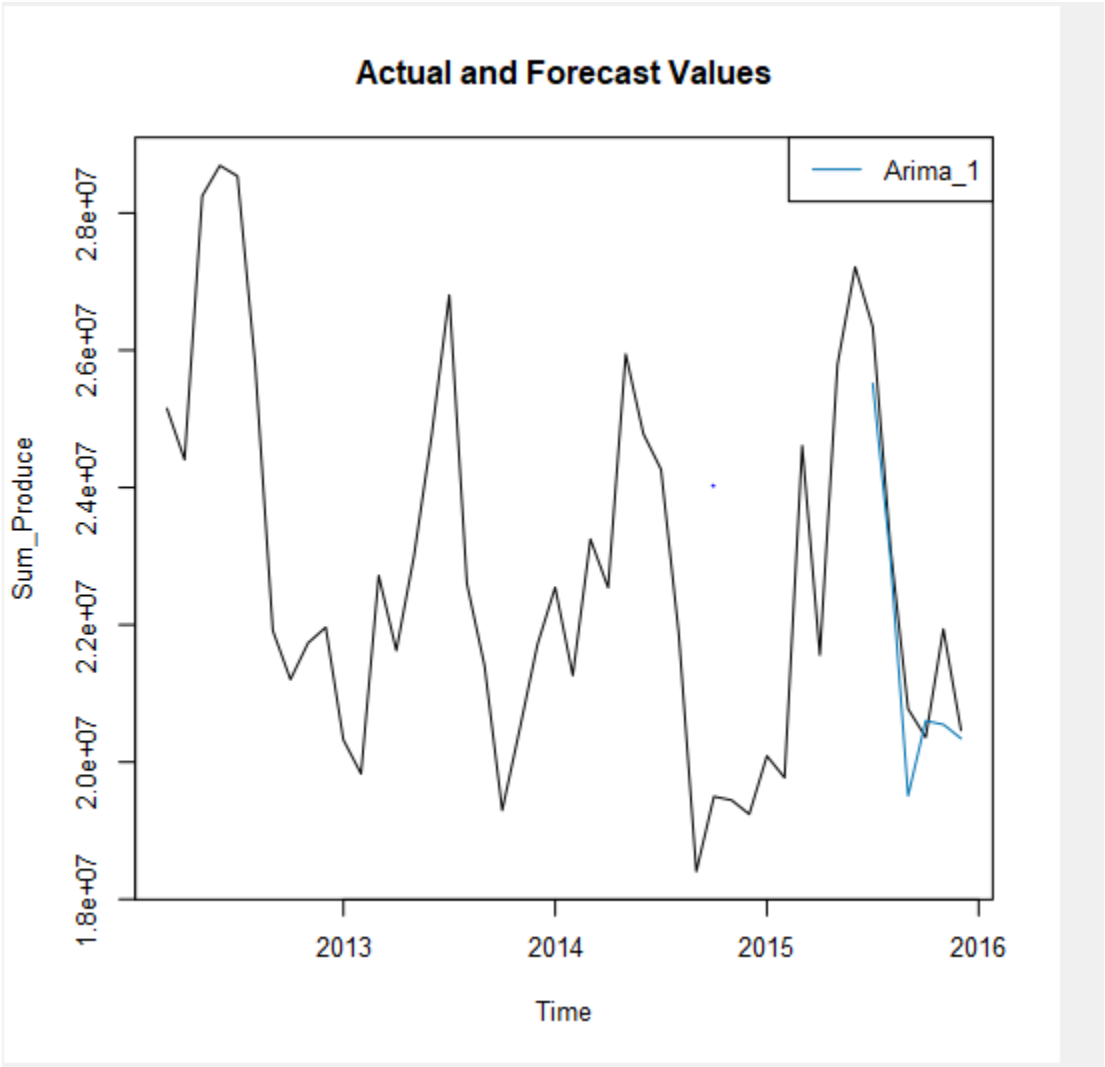
Comparison of Time Series Models

Actual and Forecast Values:

Actual	Arima_1
26338477.15	25515002.53492
23130626.6	22982398.33693
20774415.93	19509673.05693
20359980.58	20599981.42693
21936906.81	20547162.64693
20462899.3	20342794.22693

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
Arima_1	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

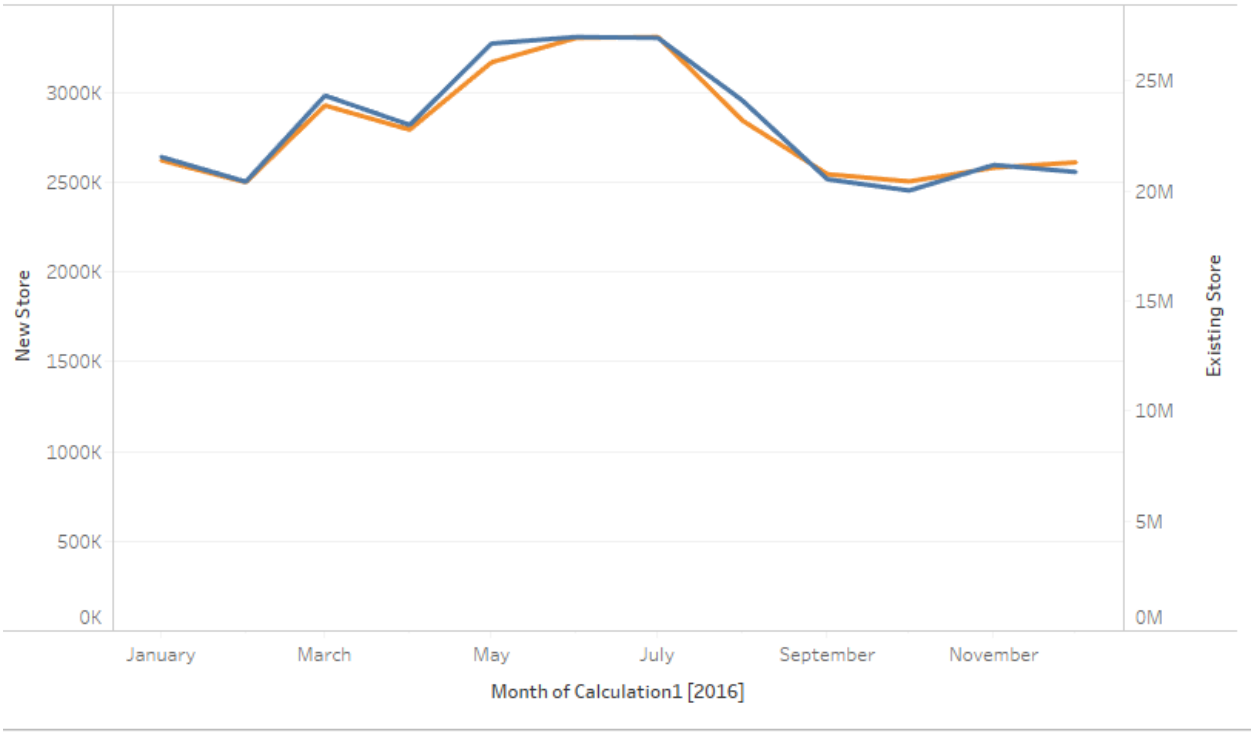


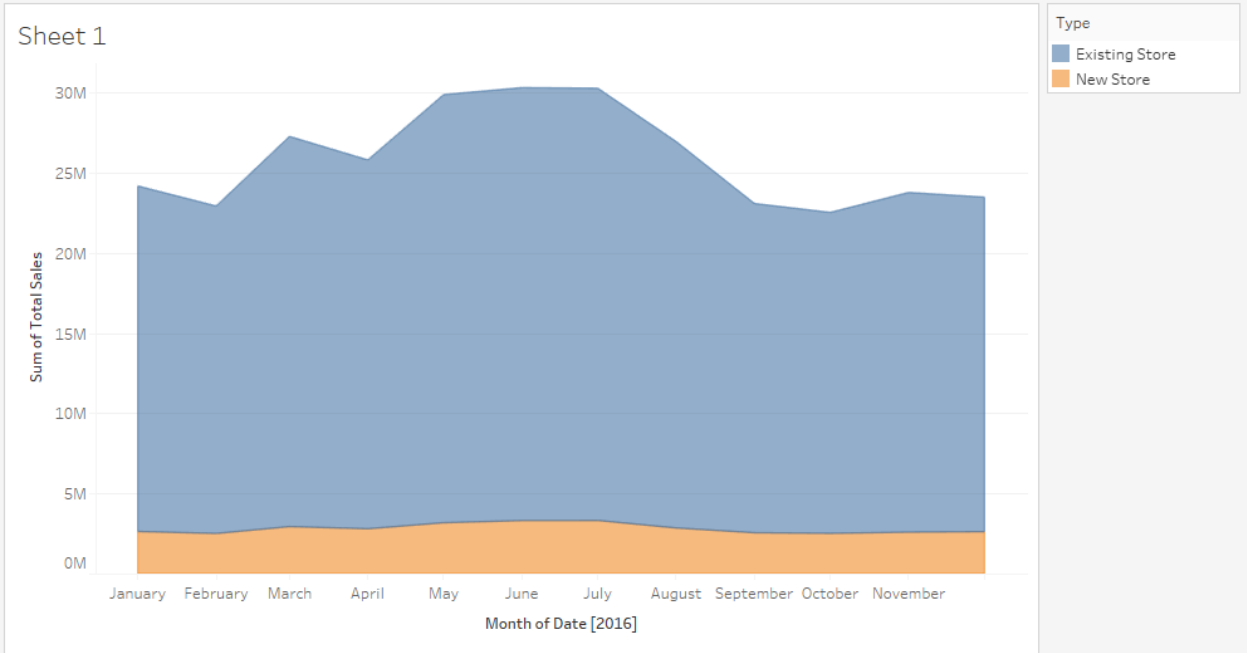
Based on the above two models , ETS model had lover ME , RMSE , MAE , MAPE values compared to ARIMA model(Even with Validation data set) . Hence ETS model is chosen

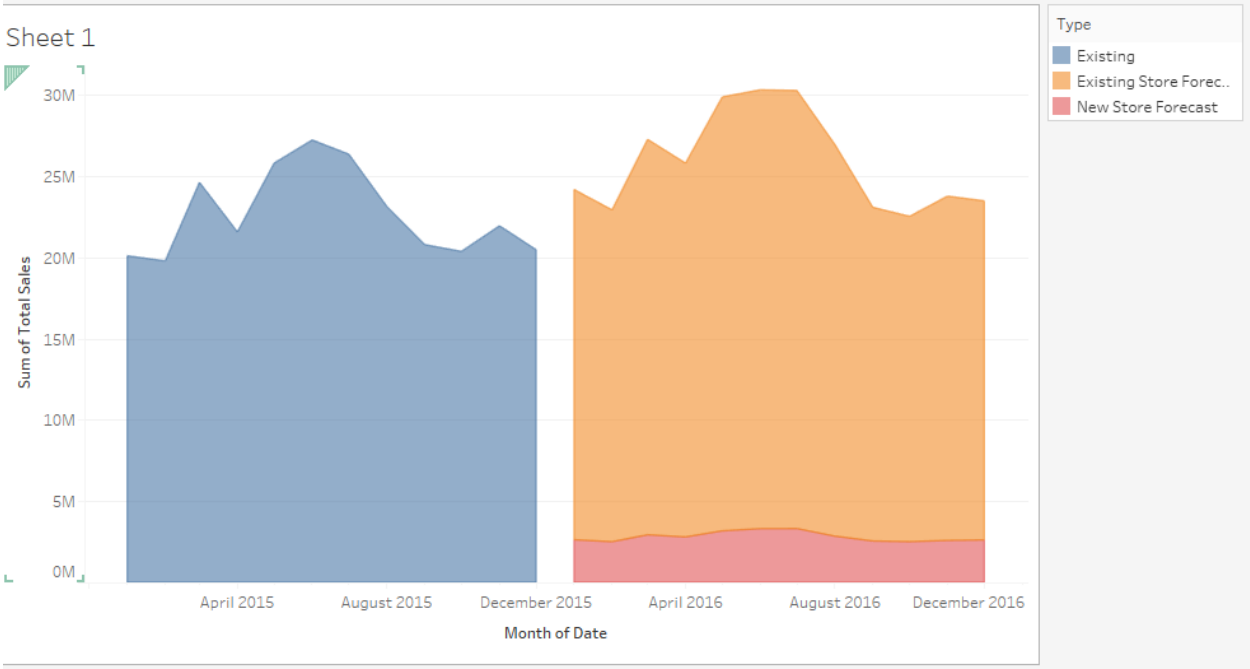
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Calculation1	Existing Stores	New Stores
1/1/2016	21,539,936.01	2,623,244.85
2/1/2016	20,413,770.60	2,500,680.44
3/1/2016	24,325,953.10	2,928,872.43
4/1/2016	22,993,466.35	2,793,733.42
5/1/2016	26,691,951.42	3,169,445.97
6/1/2016	26,989,964.01	3,305,415.33
7/1/2016	26,948,630.76	3,310,590.31
8/1/2016	24,091,579.35	2,843,645.42
9/1/2016	20,523,492.41	2,545,868.36
10/1/2016	20,011,748.67	2,505,475.79
11/1/2016	21,177,435.49	2,582,053.09
12/1/2016	20,855,799.11	2,611,623.71

Sheet 1







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.