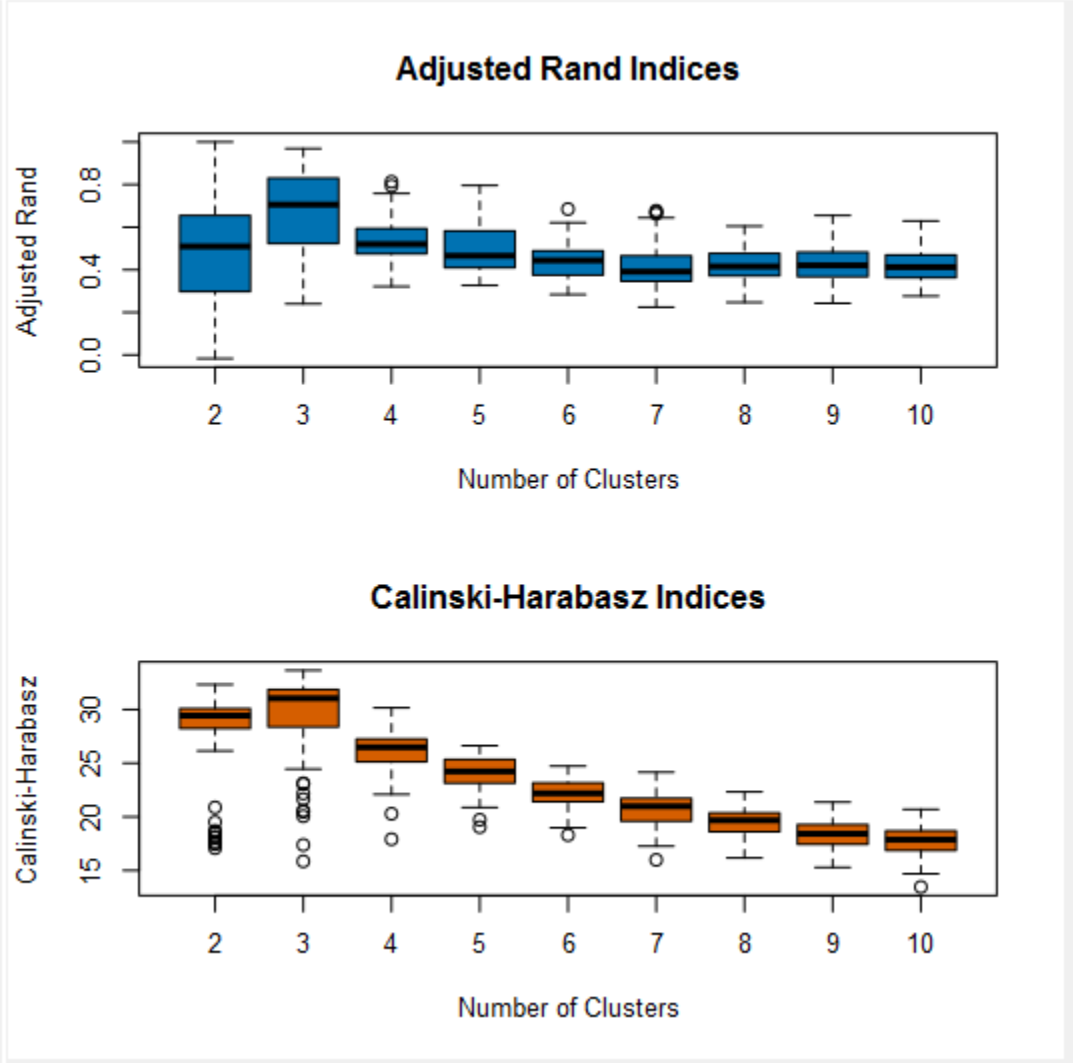
Project: Predictive Analytics Capstone

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?

Optimal number of store format is 3.
As we can inference this from below Adjusted Rand Indices and Calinski-Harabasz Indices.
Both the indices are highest at 3.



2. How many stores fall into each store format?

Store format
Cluster 1 – 23 stores
Cluster 2 – 29 stores
Clsuter 3 – 33 stores

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?

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_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

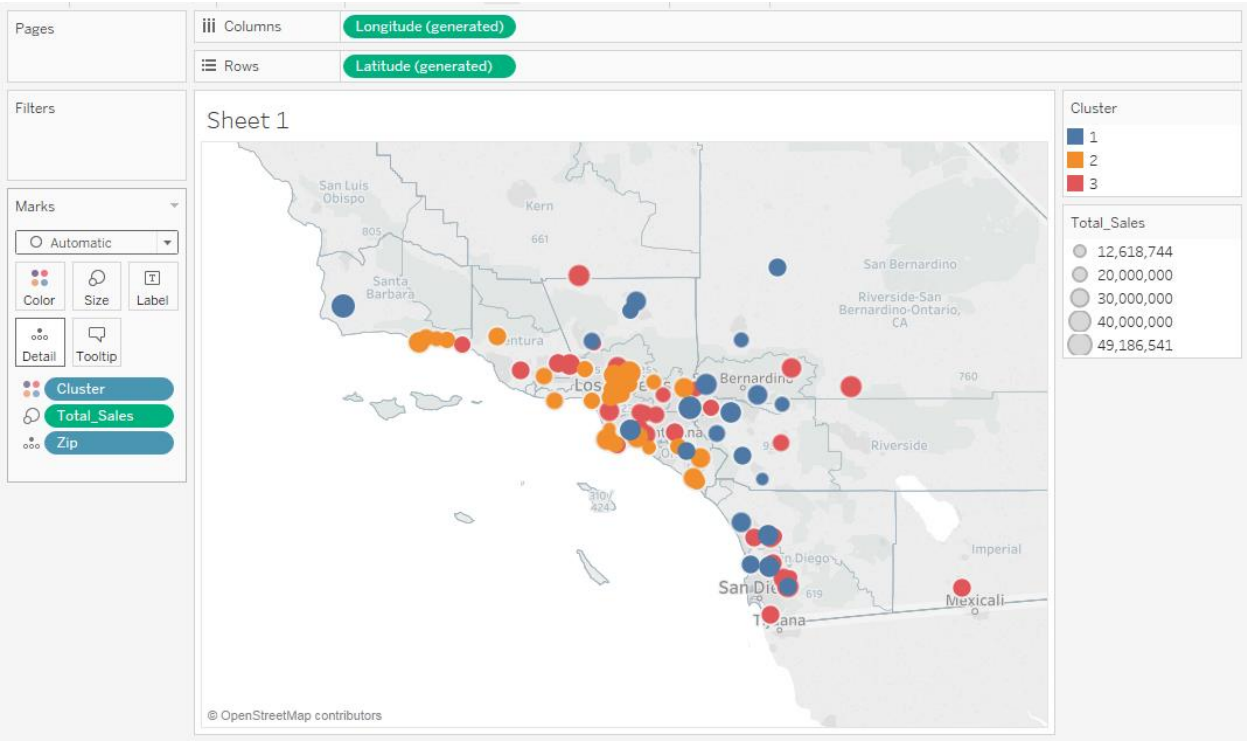
	Percent_Bakery	Percent_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

We can observe that one way cluster differ from another cluster is in '**general merchandise**' for cluster 1 it's 1.2 , cluster 2 it's -0.30 & cluster 3 it's -0.57. Which means cluster 1 differs significantly from cluster 2 & 3.

Likewise, we can see cluster 1 significantly different from cluster 2 in sales of '**dairy items**'. Cluster 1 is -0.76, Cluster 2 is 0.70, Cluster 3 is -0.08

Another significant difference we can observe between cluster 1, 2 & 3 is in sales of '**produce**'

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

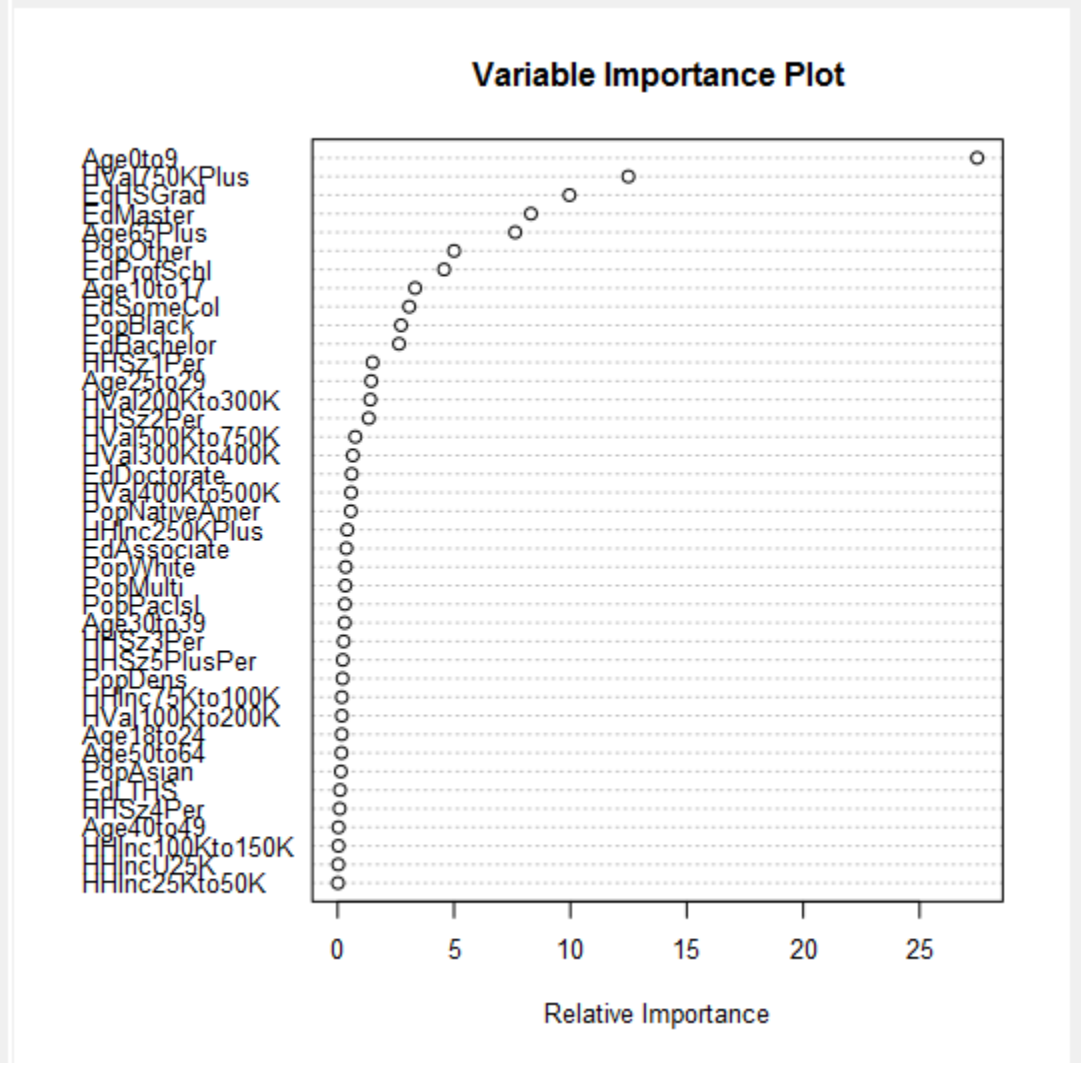
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_13	0.8235	0.8426	0.7500	1.0000	0.7778
RandomForest	0.8235	0.8426	0.7500	1.0000	0.7778
BoostedModel	0.8235	0.8889	1.0000	1.0000	0.6667

After running all the 3 models to classify best store format for new stores.
After comparing and analyzing the three best models among Decision Tree, Random Forest and Boosted Model.

Although the accuracy of all 3 models are same, **F1** score for Boosted model is higher than decision tree and random forest classification models.
Meaning boosted model is able to classify stores into clusters with greater precision.

Hence, **Boosted model is selected** for classification of new stores



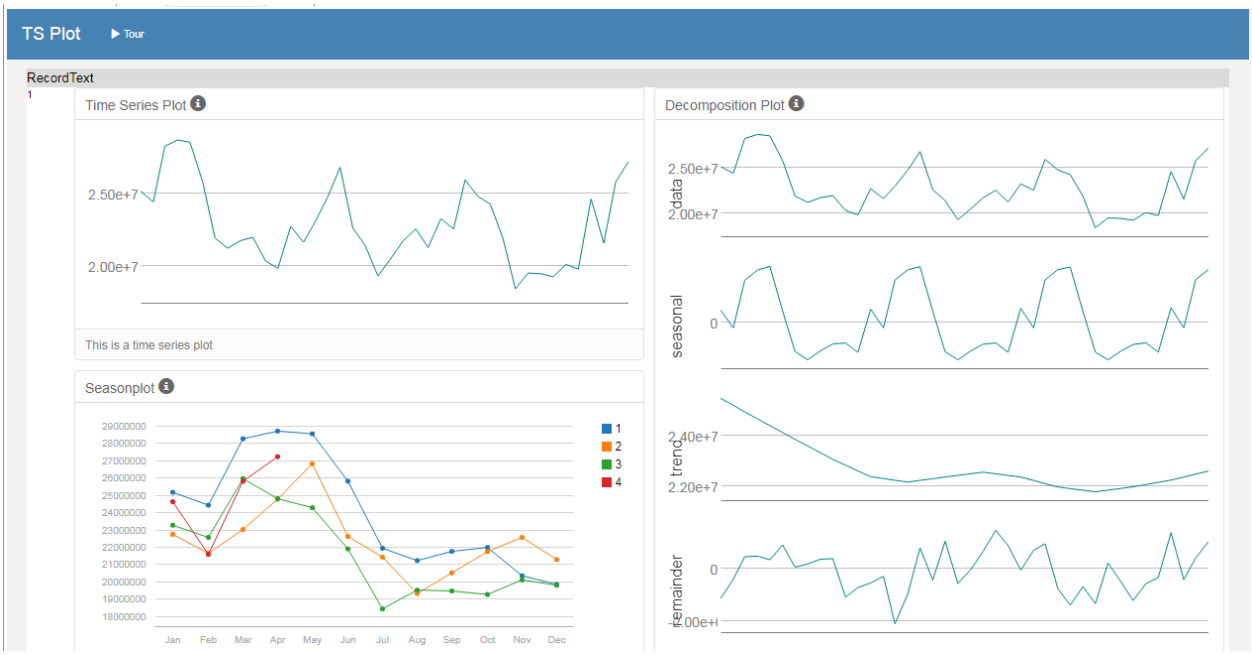
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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?

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.



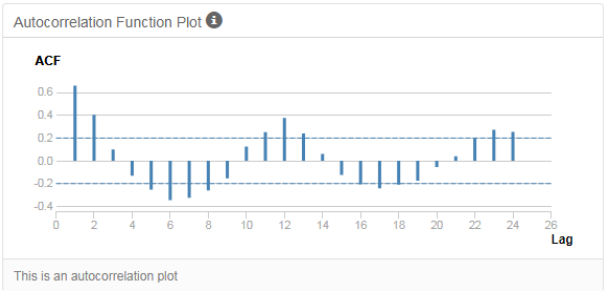
Decomposition plot shows our time series broken down into three components: trend,

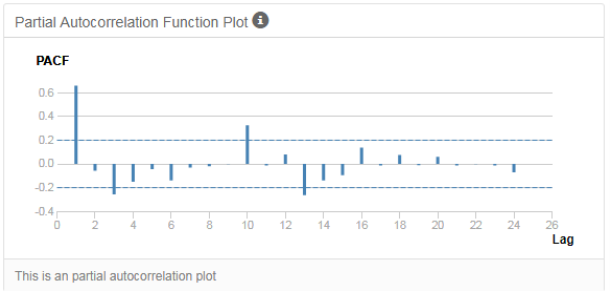
seasonality and error.

We can observer that trend is going downward with a slight curve in middle, hence there seems to be no linear or multiplicative trend we will choose none.
Seasonality graph is growing and shrinking over time hence multiplicative.
Error component is growing and shrinking over time hence it will be multiplicative

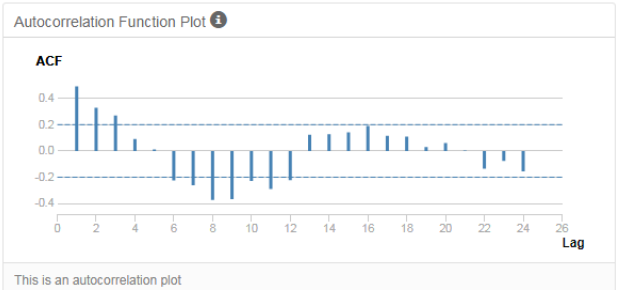
I will run ETS (M, N, M) model.

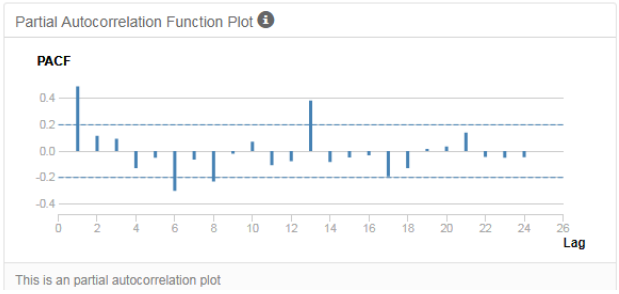
For running ARIMA model, we can look at autocorrelation plots.



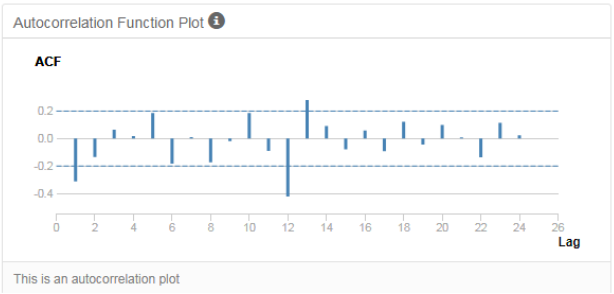


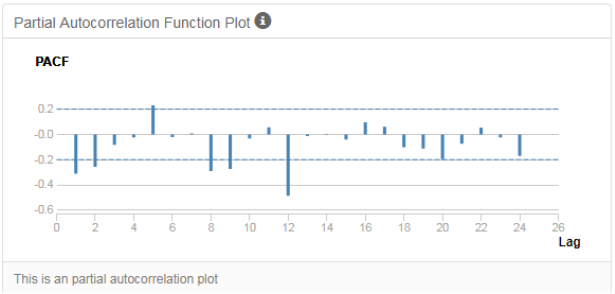
Taking Seasonal difference





First seasonal difference





After first seasonal difference, difference component would be d(1), D(1). Auto correlation plot we observe negative correlation hence moving average component would be ma(1) , MA(1).

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

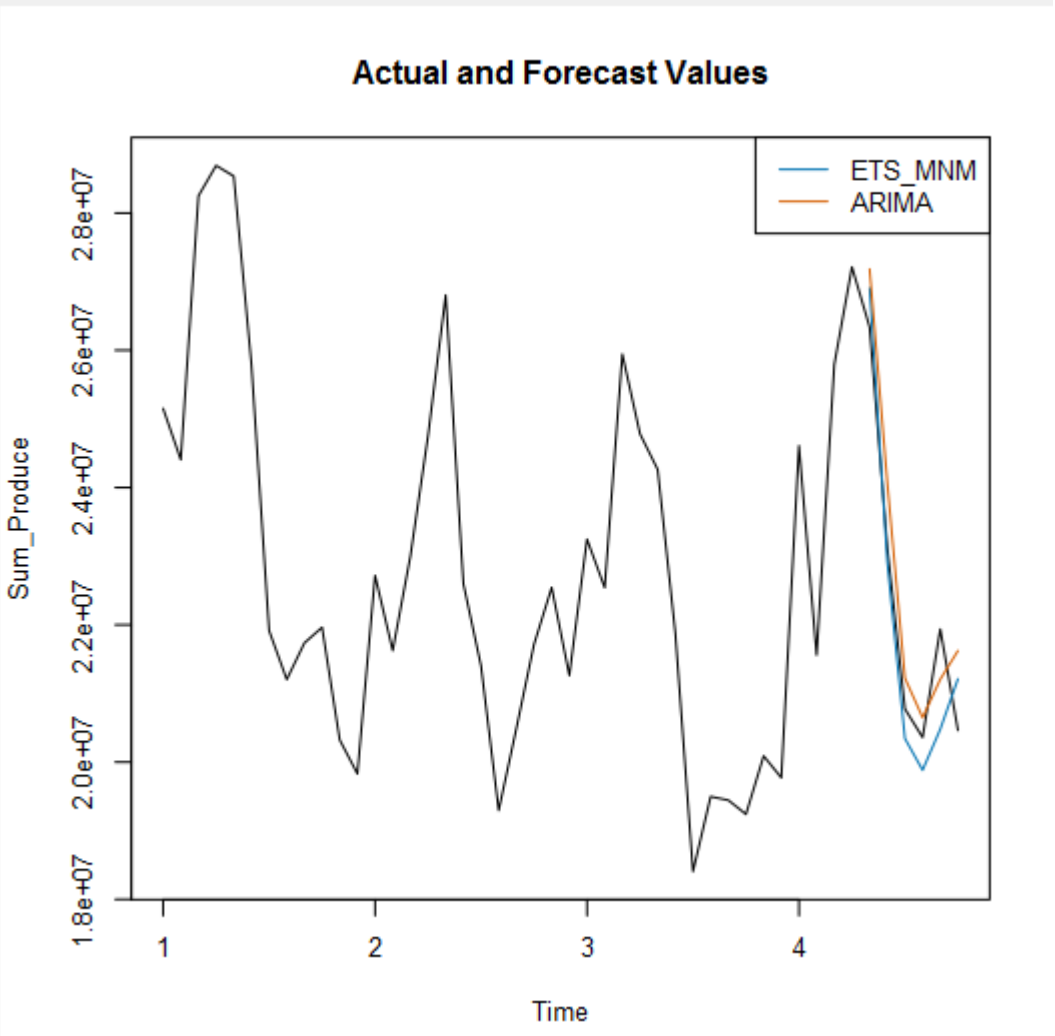
Running ETS(M, N, M) and ARIMA(0,1,1)(0,1,1)[12] and comparing the results

Actual and Forecast Values:

Actual	ETS_MNM	ARIMA
26338477.15	26907095.61191	27182961.17184
23130626.6	22916903.07434	24073582.2774
20774415.93	20342618.32222	21223756.44966
20359980.58	19883092.31778	20648299.23877
21936906.81	20479210.4317	21205988.81563
20462899.3	21211420.14021	21622151.4136

Accuracy Measures:

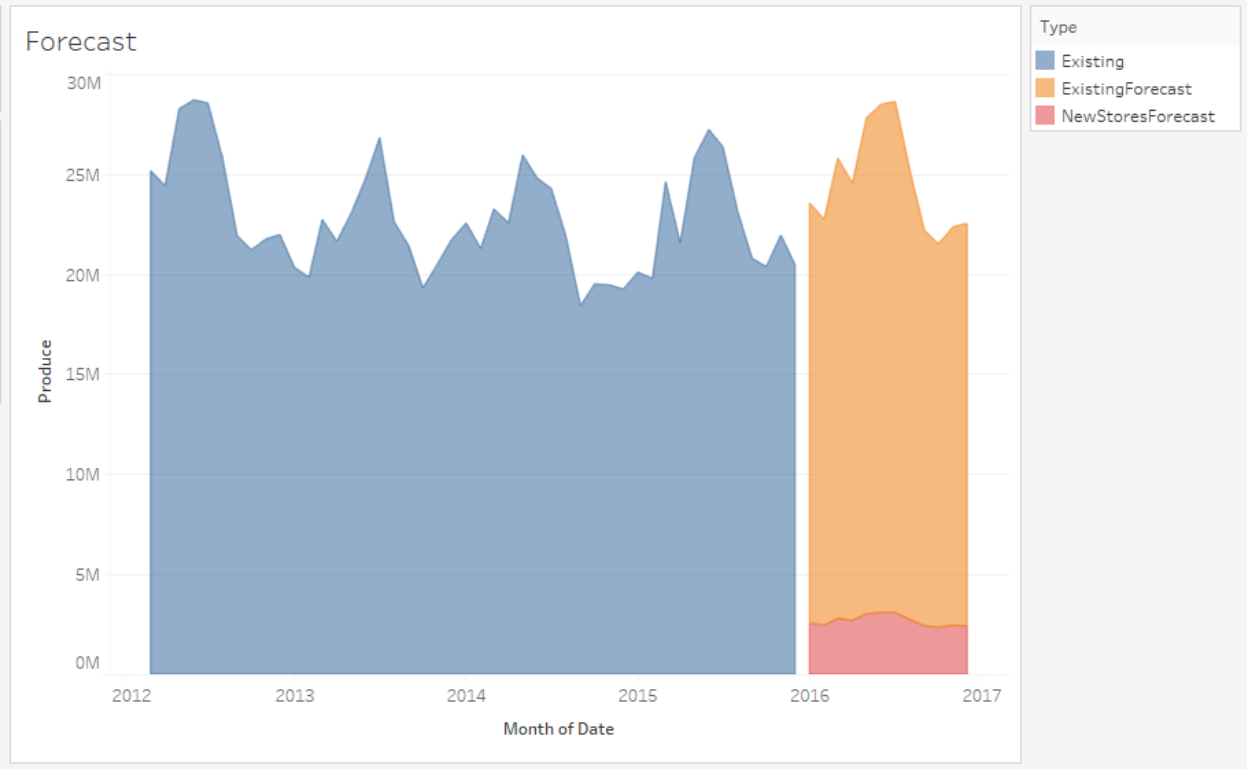
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
ARIMA	-492238.8	792197.3	735878.2	-2.1992	3.3098	0.433	NA



Both ETS and ARIMA models gives approximately same forecast
ETS(M,N,M) models seems to perform better with RMSE and MASE value less than that of ARIMA model.

Running the ETS model for forecasting sales for existing stores and new stores.
Please refer below table.

Period	New Stores	Existing Stores	Sales Forecast
Jan-16	2587450.85	20985665.43	23573116.28
Feb-16	2477352.88	20282203.23	22759556.11
Mar-16	2913185.2	22996330.77	25909515.97
Apr-16	2775745.64	21835774.09	24611519.73
May-16	3150866.82	24776343.06	27927209.88
Jun-16	3188922.01	25389437.77	28578359.78
Jul-16	3214745.66	25519682.18	28734427.84
Aug-16	2866348.66	22465594.99	25331943.65
Sep-16	2538726.82	19757858.64	22296585.46
Oct-16	2488148.3	19155636.58	21643784.88
Nov-16	2595270.37	19904834.31	22500104.68
Dec-16	2573396.66	20109948.03	22683344.69
Total	**33370159.9**	**263179309.1**	**296549469**



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.