Project: Predictive Analytics Capstone

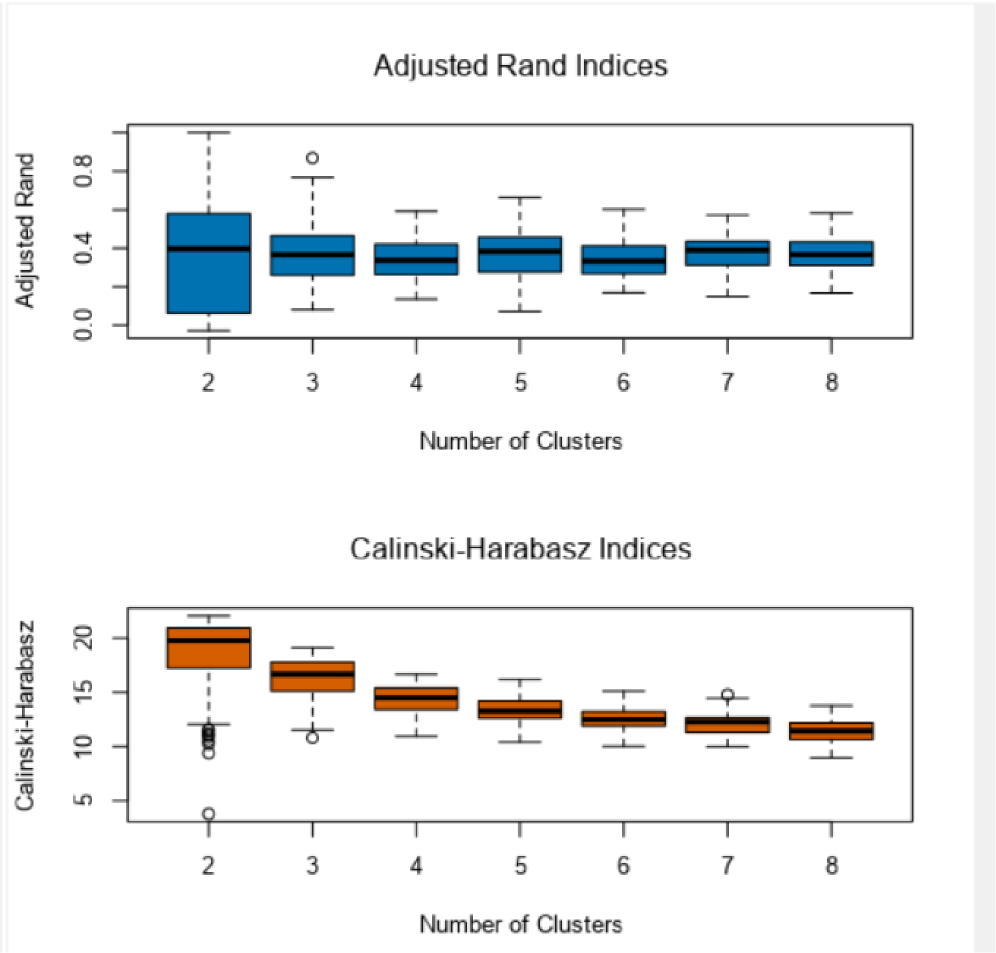
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

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 was found to be 3. Based on the 2015 sales data, it was found. The variable percentage of sales by category per store (i.e. each sales category as the percentage of total sales) was standardized by Z-Score to group
   Using the K- Centroids Diagnostic tool, the appropriate number of cluster was assessed. K-Means clustering algorithm was used to the method. Measures examined are adjusted Rand Index and Calinski-Harabasz index. The result is shown below.

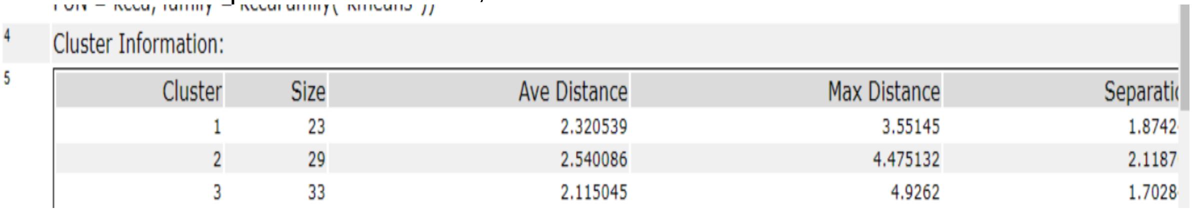
   The K-Centroids Diagnostic tool allows an assessment of the appropriate number of clusters. The clustering algorithm selected is K-Means. Two measures examined are the adjusted Rand index and the Calinski-Harabasz index. The graph below shows the ideal number of store formats:



The cluster number is based on each measure correspond to average comparing with the highest median of the solution. Therefore, the ideal storage format would be 3.

2. How many stores fall into each store format?

K-centered cluster analysis tool was used to find the cluster information. It was found that 23 stores are present in cluster 1, 29 stores in cluster 2 and 33 stores in clulster3.

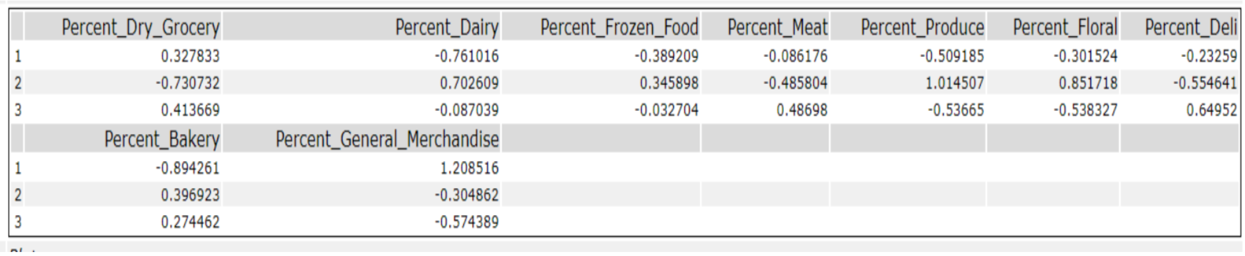


Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separati
1	23	2.320539	3.55145	1.8742
2	29	2.540086	4.475132	2.1187
3	33	2.115045	4.9262	1.7028

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

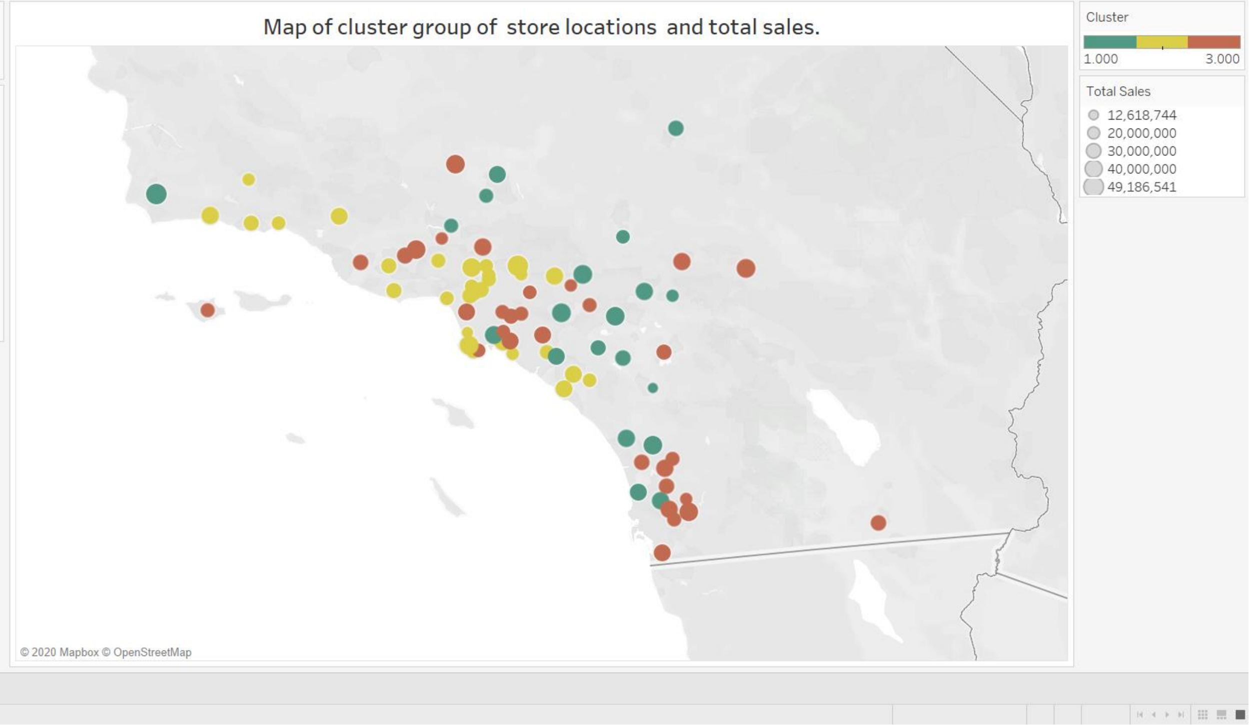
The below is the report genertated by the K-centroid cluster analysis tool.



	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 infer that cluster1 stores are generalized by high percent_general _merchandise in comparision to cluster 2 and cluster3. Similarly, cluster 2 are generally characterized by high percent_produce in comparision to cluster1 and cluster3.

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.

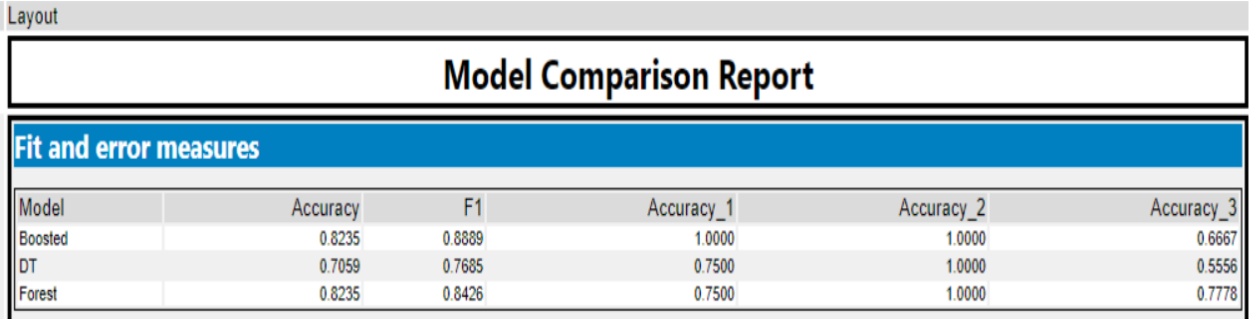


Map of cluster group of store locations and total sales.

Task 2: Formats for New Stores

10 new stores of the grocery chain is going to open in the starting of the year. Since, we don’t have any sales data, we will determine the store format using the demographic data present

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.)
   To predict the best score format for the new stores, we used the demographic data to run against boosted model, decision tree model and the random forest model. Split of 80-20 was done for training and validation data Then the model comparison tool was used to compare the model.
   The result of the model comparison tool is shown below:



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667
DT	0.7059	0.7685	0.7500	1.0000	0.5556
Forest	0.8235	0.8426	0.7500	1.0000	0.7778

Since, the boosted model and the forest model have same accuracy score, I choose to go with the boosted model as the F1 score of the boosted model was higher.

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

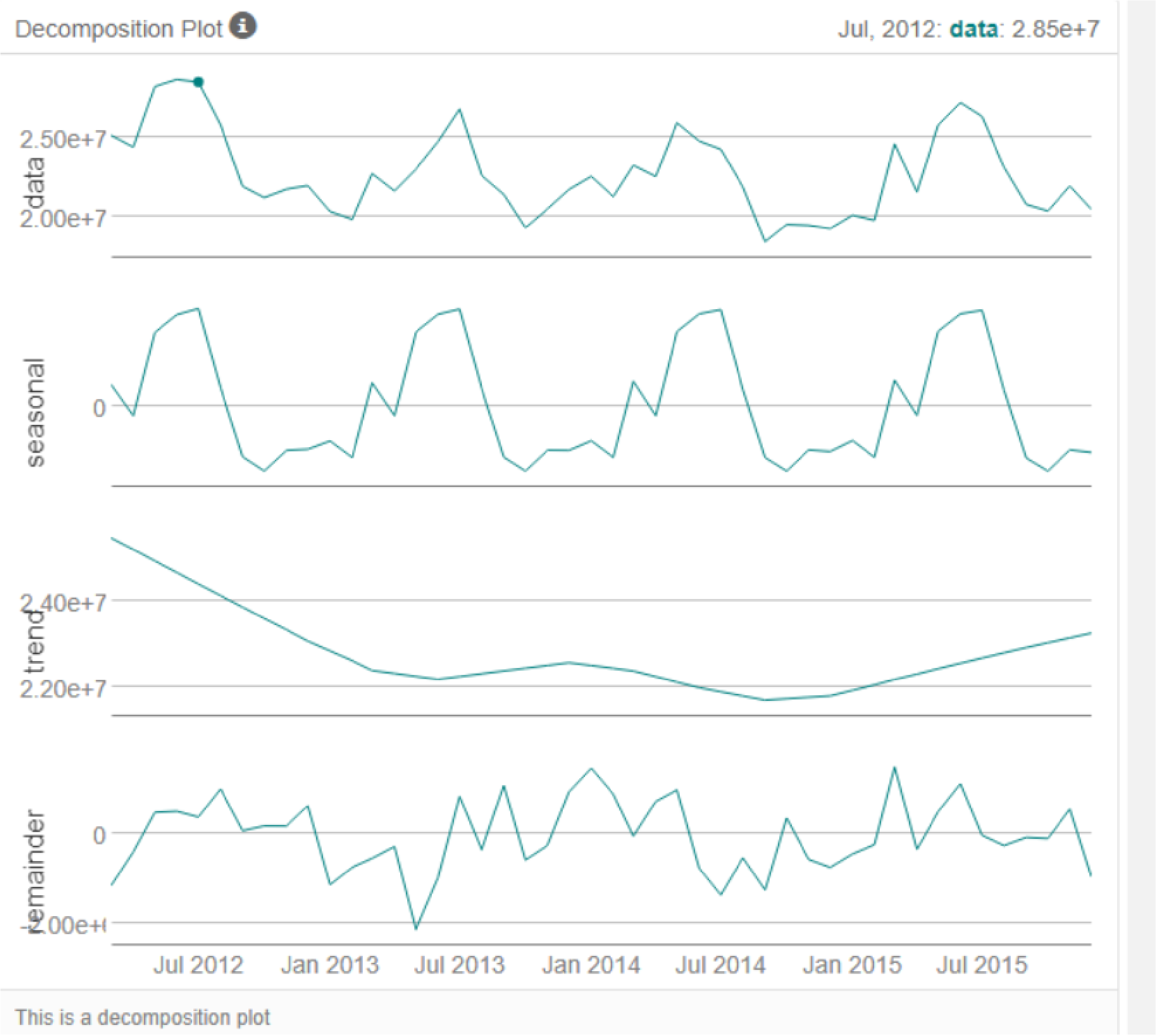
Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

Since, the fresh products have short life span and high cost, we are requested to have an accurate monthy sales forecast

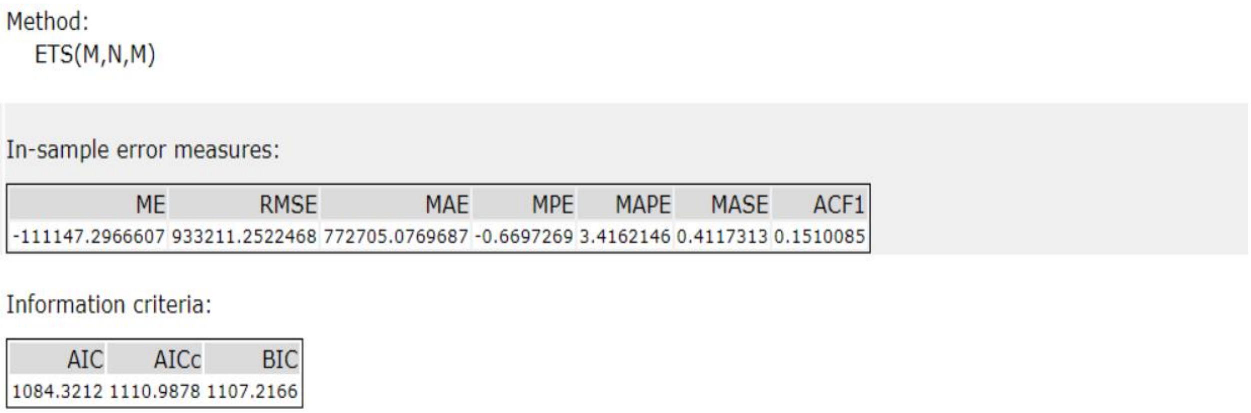
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

For the forecast, we applied and compare both the ETS and ARIMA model. By analyzing the initial time series decomposition plot which is shown below, we could further find the model parameters.
We are using sales for produce / month for all the stores aggregated.



The decomposition plot shows that there is increase in the error element, the trend is non existential and also, the seasonal trend is increasing. Hence, I have used ETS(M,N,M) model. For the ARIMA model, i had set it to auto. Both the model had a holdout period of 12.

For ETS(M,N.M) model:



Method:
ETS(M,N,M)

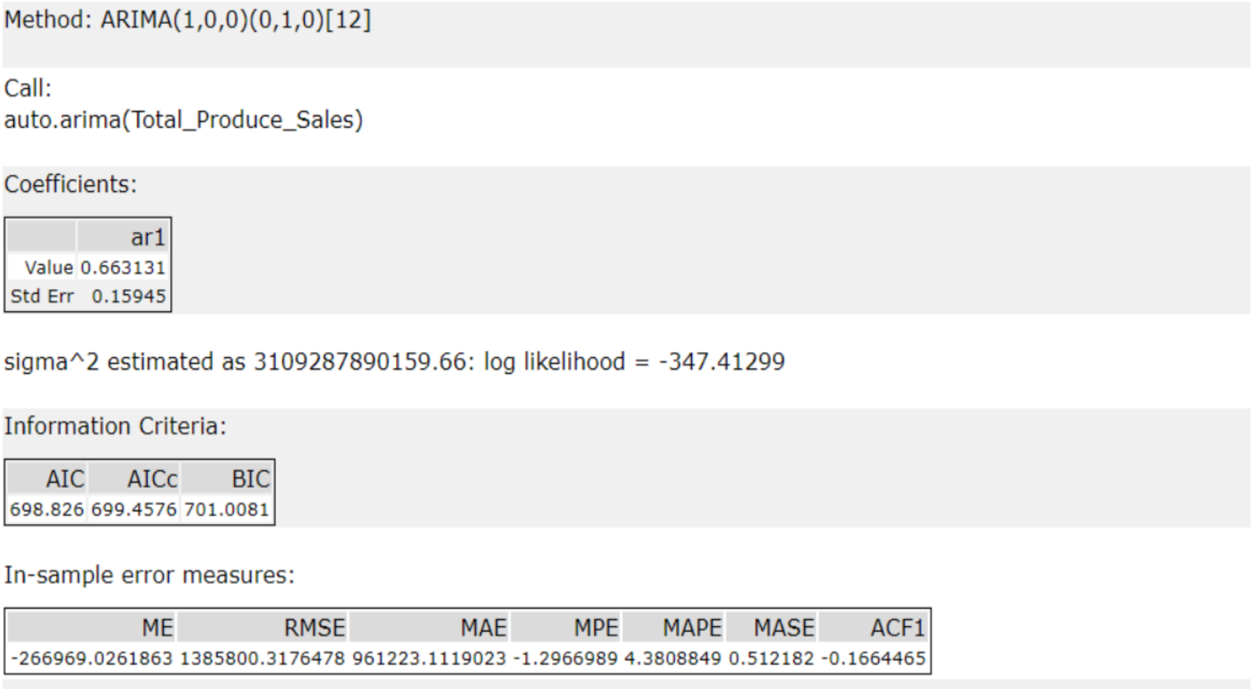
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-111147.2966607	933211.2522468	772705.0769687	-0.6697269	3.4162146	0.4117313	0.1510085

Information criteria:

AIC	AICc	BIC
1084.3212	1110.9878	1107.2166

For ARIMA(1,0,0)(0,1,0)12 selected automatically :



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

Call:
auto.arima(Total_Produce_Sales)

Coefficients:

	ar1
Value	0.663131
Std Err	0.15945

sigma^2 estimated as 3109287890159.66: log likelihood = -347.41299

Information Criteria:

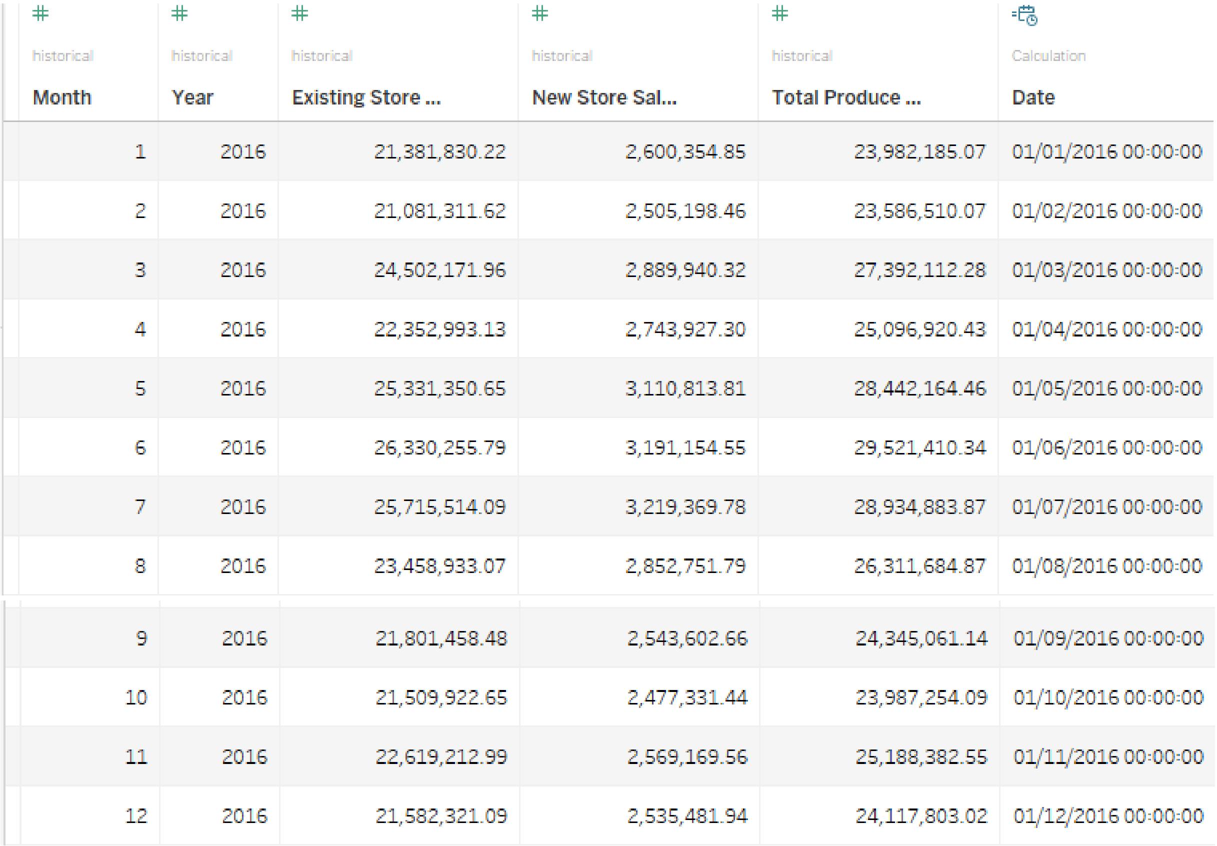
AIC	AICc	BIC
698.826	699.4576	701.0081

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-266969.0261863	1385800.3176478	961223.1119023	-1.2966989	4.3808849	0.512182	-0.1664465

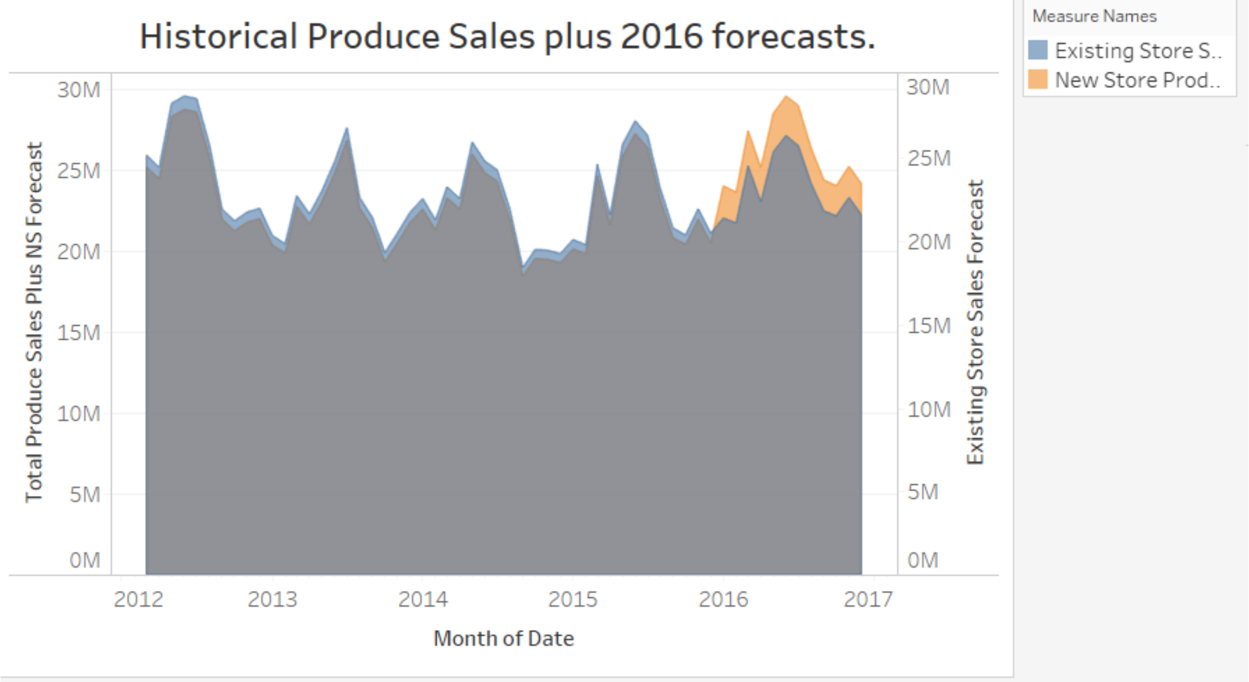
We select the ETS(M,N,M) as the forecasting model as the ETS model showed less error valued in comparisons to the ARIMA model

The table of forecast of existing and new stores is given below:



historical	historical	historical	historical	historical	Calculation
Month	**Year**	**Existing Store ...**	**New Store Sal...**	**Total Produce ...**	**Date**
1	2016	21,381,830.22	2,600,354.85	23,982,185.07	01/01/2016 00:00:00
2	2016	21,081,311.62	2,505,198.46	23,586,510.07	01/02/2016 00:00:00
3	2016	24,502,171.96	2,889,940.32	27,392,112.28	01/03/2016 00:00:00
4	2016	22,352,993.13	2,743,927.30	25,096,920.43	01/04/2016 00:00:00
5	2016	25,331,350.65	3,110,813.81	28,442,164.46	01/05/2016 00:00:00
6	2016	26,330,255.79	3,191,154.55	29,521,410.34	01/06/2016 00:00:00
7	2016	25,715,514.09	3,219,369.78	28,934,883.87	01/07/2016 00:00:00
8	2016	23,458,933.07	2,852,751.79	26,311,684.87	01/08/2016 00:00:00
9	2016	21,801,458.48	2,543,602.66	24,345,061.14	01/09/2016 00:00:00
10	2016	21,509,922.65	2,477,331.44	23,987,254.09	01/10/2016 00:00:00
11	2016	22,619,212.99	2,569,169.56	25,188,382.55	01/11/2016 00:00:00
12	2016	21,582,321.09	2,535,481.94	24,117,803.02	01/12/2016 00:00:00

Also, we have made a tabeleau plot for the store sales produce to forecast existing and the new store sales. This is shown below.


Historical Produce Sales plus 2016 forecasts.
Total Produce Sales Plus NS Forecast
30M
25M
20M
15M
10M
5M
0M
Existing Store Sales Forecast
30M
25M
20M
15M
10M
5M
0M
Month of Date
Measure Names
Existing Store S..
New Store Prod..