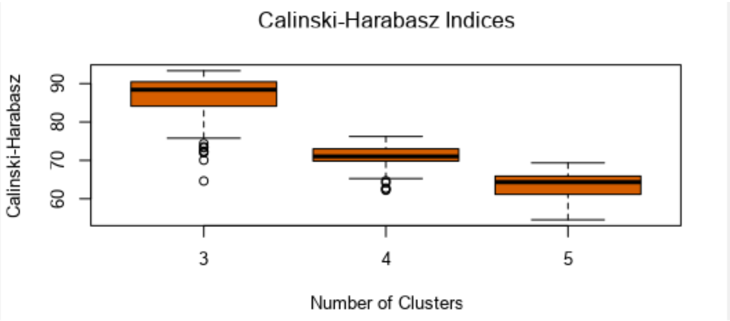
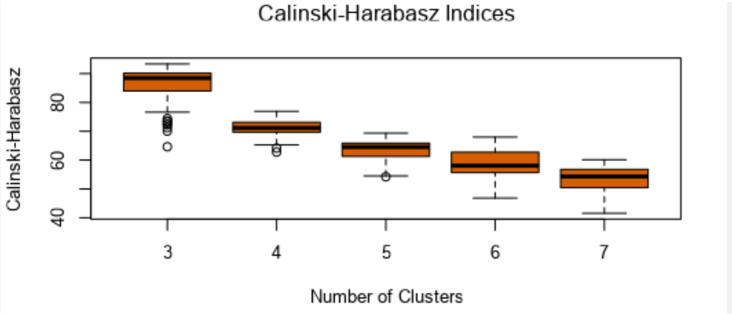
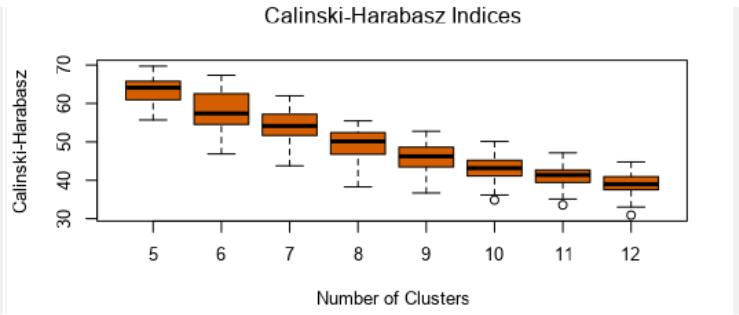
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. Before we work with new stores, we must first determine the store formats of our stores. The store format will help us categorize all stores into a group that best utilizes the supply chain of goods. The optimal number of store formats for the 85 stores is 3. To get to this decision I used the Alteryx tool K-centroid Diagnostics tool to look at how the different number of clusters perform.

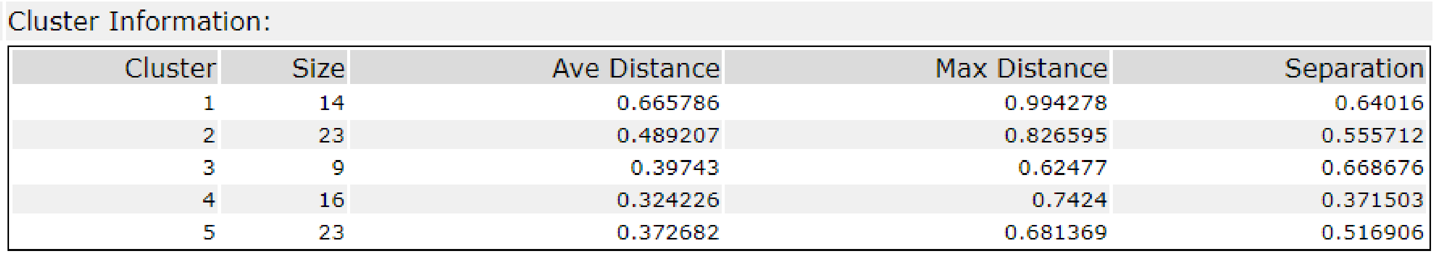






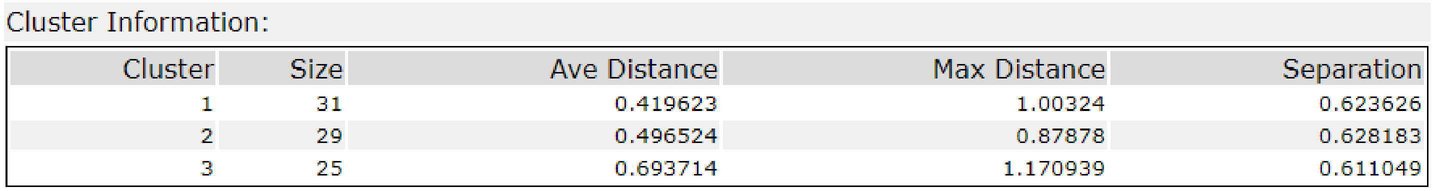
In the picture above we can see that the K-Centroid Diagnostics tool gives us box and whisker plots that shows how the different clusters are grouped. We can see that the more clusters there are (8 or 5) the more each cluster overlaps with the next. For the 3 clusters however, we can clearly see that most of the parts are separate to each other while at the same time being tightly closed and centered to itself.

2. To see how many stores are grouped in each cluster we use the K-centroid Analysis tool. This shows us the number of items in a cluster, the average distance within a cluster, and the separation between a cluster and its neighbor.



Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	14	0.665786	0.994278	0.64016
2	23	0.489207	0.826595	0.555712
3	9	0.39743	0.62477	0.668676
4	16	0.324226	0.7424	0.371503
5	23	0.372682	0.681369	0.516906

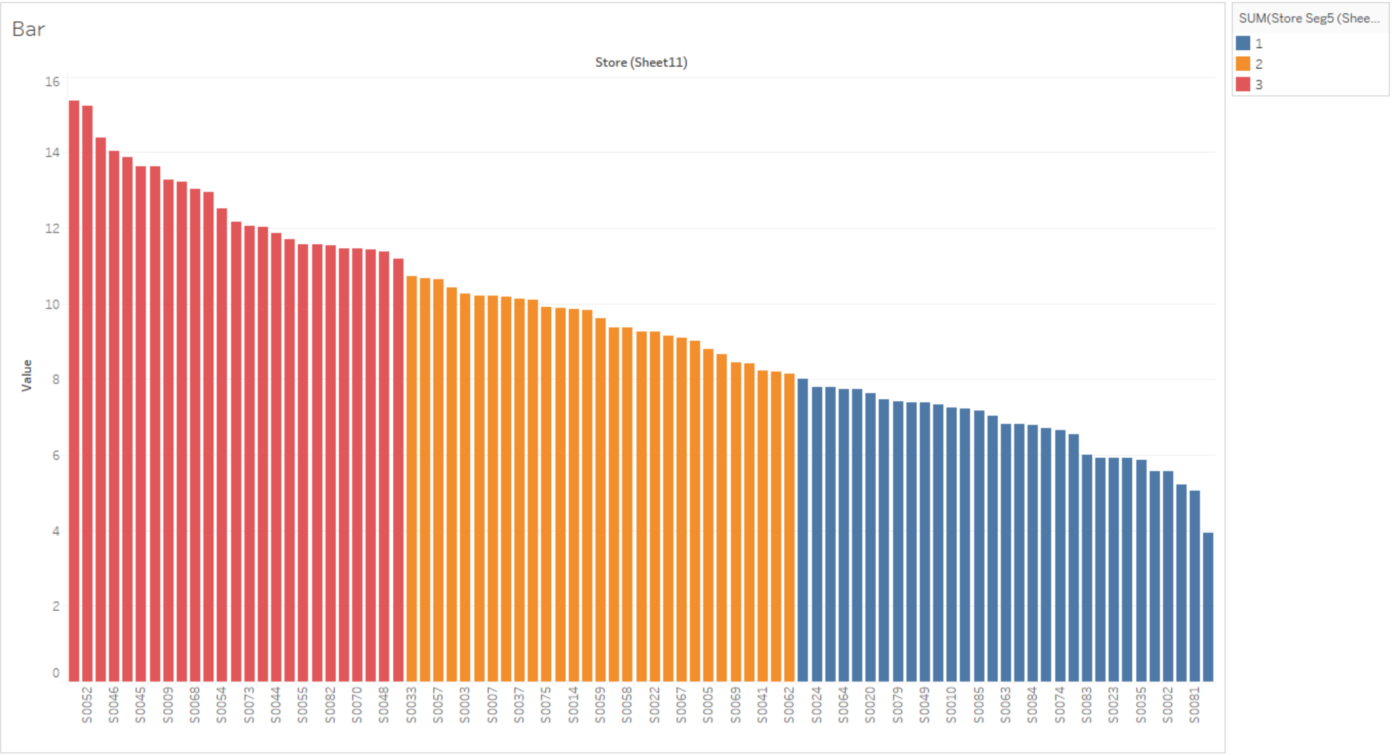


Cluster Information:

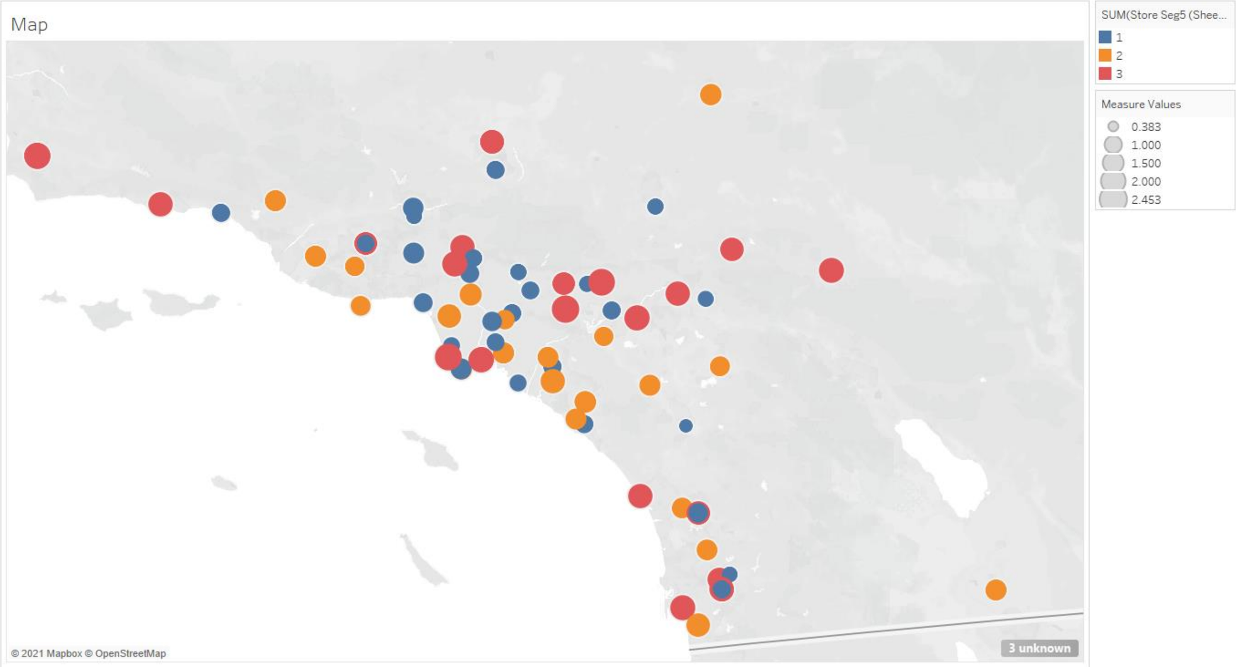
Cluster	Size	Ave Distance	Max Distance	Separation
1	31	0.419623	1.00324	0.623626
2	29	0.496524	0.87878	0.628183
3	25	0.693714	1.170939	0.611049

We can see from the photo that cluster 1 has 31 stores, cluster 2 has 29 stores, and cluster 3 has 25 stores.

3. The clusters differ from one another in their percentage sales per category per store. The clusters are characterized by high, medium, and low volume of sales in their total percentage of sales per category. Cluster 3 has high sales totals while, 2 and 1 have medium and low respectively.

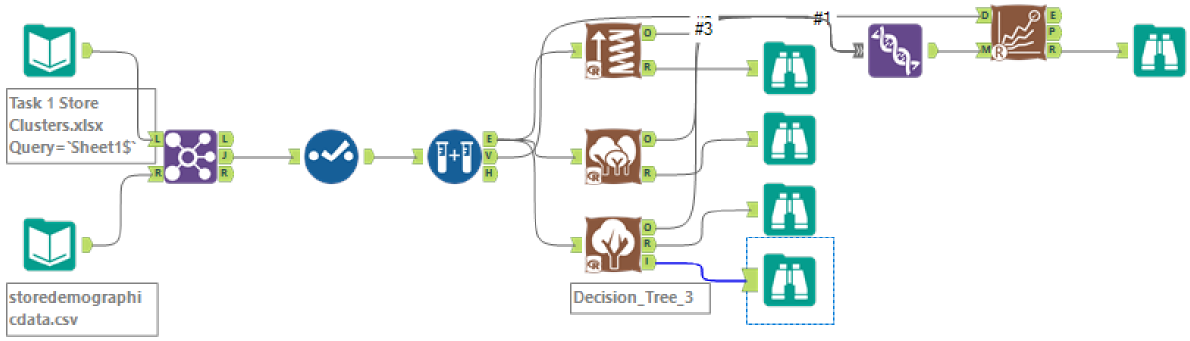


4. Putting the stores in their respective clusters then assigning their size to be equivalent to their total sales we can see that cluster 3 has the most sales followed by cluster 2 then finally cluster 1.

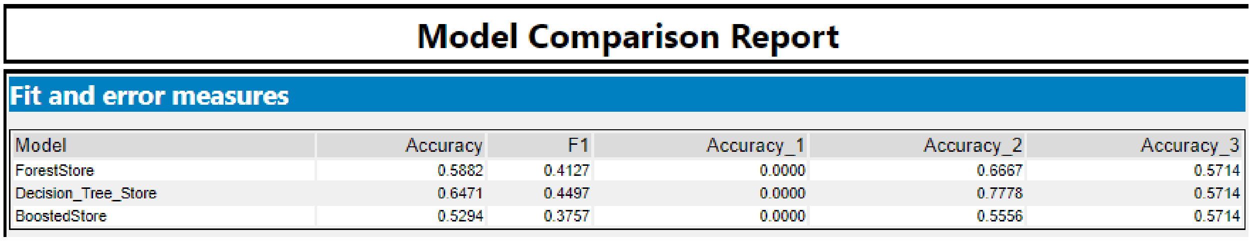


Task 2: Formats for New Stores

1. Next, we will determine the store formats for the new stores. To do this we must first use the demographics data we have about our current stores to make a predictive model. We have to compare the three different models; Boosted Model, Forest Model, and Decision Tree Model. We use the clustering data we used for determining the store formats and combine it with the demographics data to use in our models. We need to use a 20% validation sample to help compare the accuracy of the different models.



Now we use the result from the model comparison tool to make our decision. The model comparison tool gives us the accuracy of each model. As we can see, the **Decision tree** model is the best option we have with an accuracy of **0.6471**.

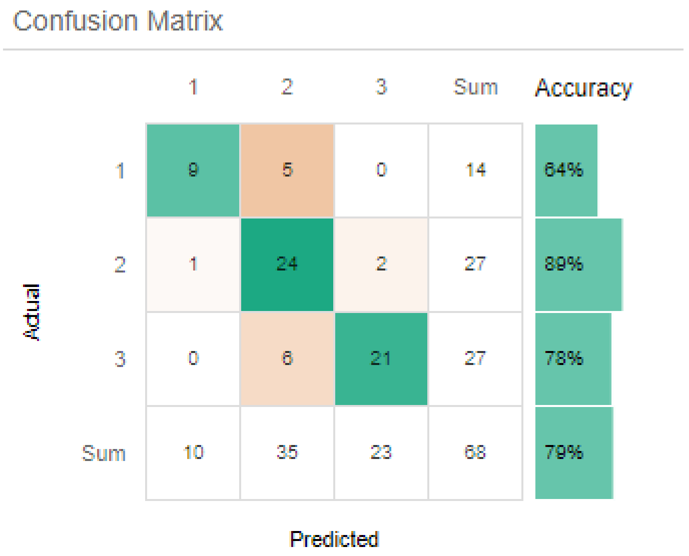


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
ForestStore	0.5882	0.4127	0.0000	0.6667	0.5714
Decision_Tree_Store	0.6471	0.4497	0.0000	0.7778	0.5714
BoostedStore	0.5294	0.3757	0.0000	0.5556	0.5714

We can further see form the confusion matrix of the decision tree model how accurate it is at predicting each store format.



Confusion Matrix

		1	2	3	Sum	Accuracy
Actual	1	9	5	0	14	64%
	2	1	24	2	27	89%
	3	0	6	21	27	78%
	Sum	10	35	23	68	79%

Predicted

2. Now that we have chosen a model, we must then score the demographic data we have about the new stores with the model to find out which format each of the new store falls in.

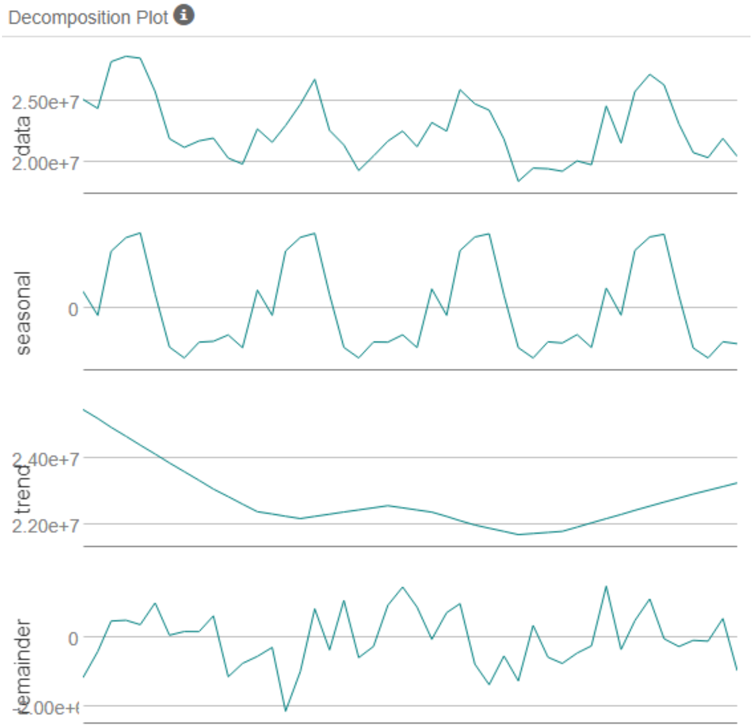
Store Number	Segment
S0086	2
S0087	2
S0088	2
S0089	2
S0090	2
S0091	1
S0092	3
S0093	2
S0094	3
S0095	2

Task 3: Predicting Produce Sales

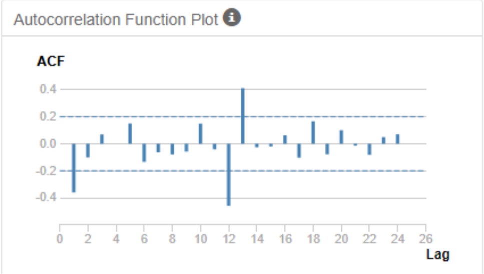
1. To predict the produce sales for the stores we need to use time series forecasting. The analysis will be split in two for the two different types of stores. First, we will aggregate the total sales of produce for each existing store. Then, we will use the average aggregated produce sales per cluster to predict the sales of the new stores.

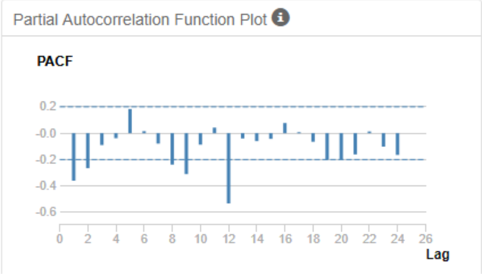
The steps taken for the prediction of the produce sales for the existing stores are as follows;

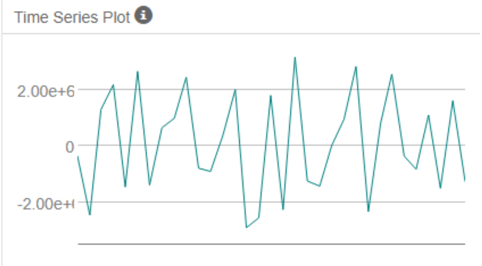
- First, we add up the produce sales grouping it by store, year, and month. We need to hold 6-month data to compare with the models later.
- For the ETS model we look for additive or multiplicative patterns in the decomposition plot


Decomposition Plot

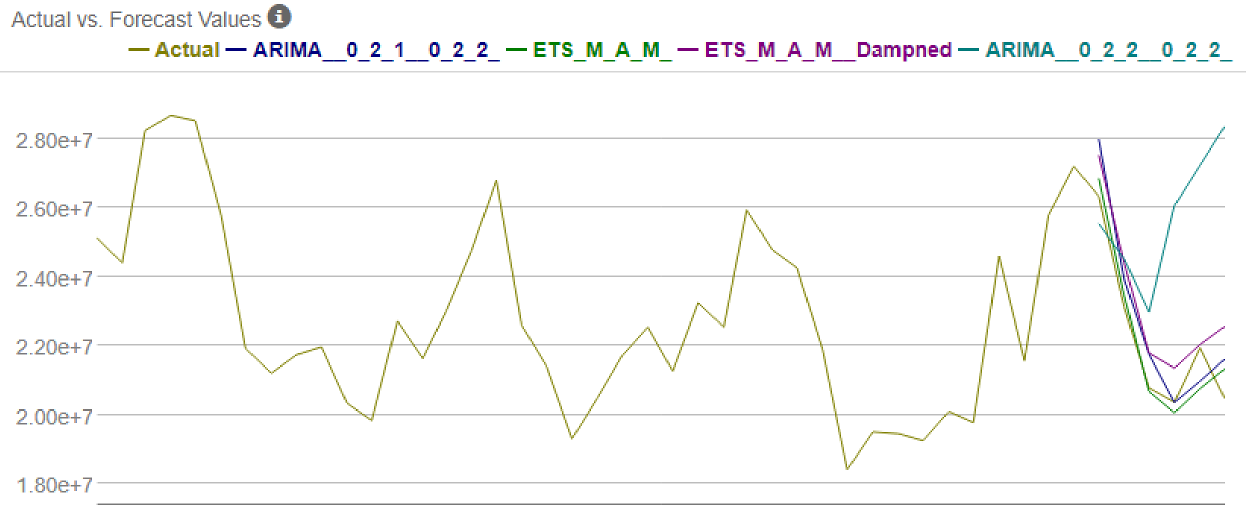
- For the ARIMA model we must first difference the seasonal patterns because the data seems to be correlated in the different lags. When differencing the data we can determine the AR, MA and I components based on the number of ACF, PACF, and the number of differencing we did to make the graph stationary.

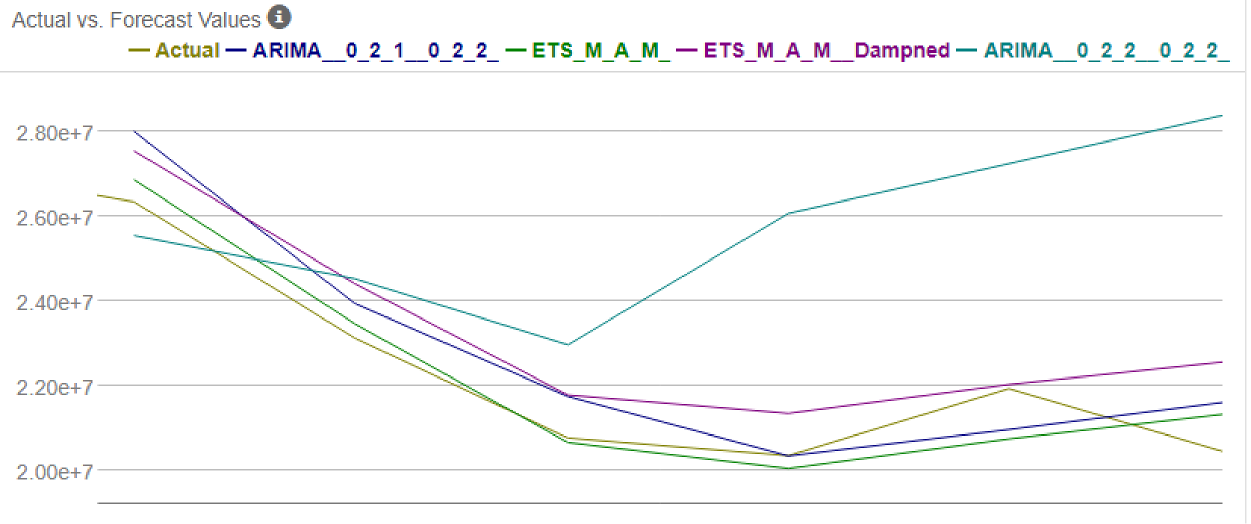

Autocorrelation Function Plot


Partial Autocorrelation Function Plot


Time Series Plot

- Then, we build the models with slightly different settings and compare them with the 6 month withheld data. Which gives us the comparison report of the models and the actual data.

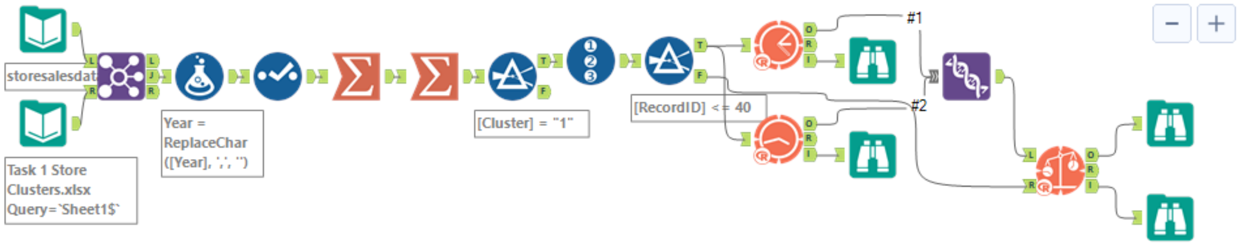




- We can see from the above diagram that the ETS MAM model is the model which closely resembles the actual data. So, we use that model to predict the 12 month produce sales.

The steps taken for the prediction of the produce sales for the new stores are as follows;

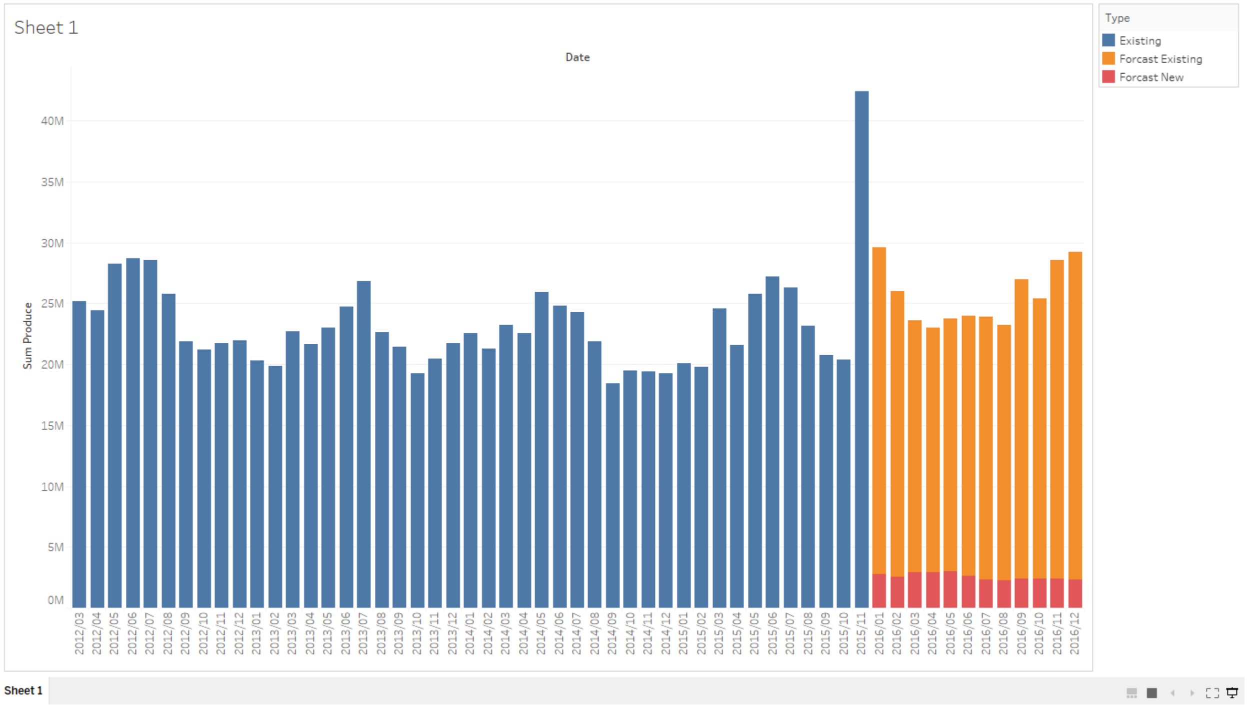
- First, we have to aggregate the sum of produce sales by store, year, month, and cluster.
- Then, we aggregate the average of produce sales by year, month, and cluster.
- Next, we split up the data and predict the outcome of each cluster separately.


storesalesdat
Task 1 Store Clusters.xlsx Query=`Sheet1$`
Year = ReplaceChar ([Year], ',', '')
[Cluster] = "1"
[RecordID] <= 40
#1
#2
L
J
R
T
F
O
I

- Afterwards, we multiply the predicted average of each cluster with the number of new stores in each cluster. Then, we merge the different clusters into one.

Month	Existing Stores	New Stores
Jan-16	26,860,639.574437	2744793.25238171
Feb-16	23,468,254.495953	2561355.91001367
Mar-16	20,668,464.644954	2912274.65081658
Apr-16	20,054,544.076312	2940851.1710502
May-16	20,752,503.519965	2998489.5097649
Jun-16	21,328,386.809651	2634188.11874266
Jul-16	21,611,877.980495	2327656.14605393
Aug-16	20,931,380.132725	2281127.65275193
Sep-16	24,588,621.430699	2367423.31143287
Oct-16	22,974,656.794772	2423056.24026646
Nov-16	26,185,910.648663	2397759.94450829
Dec-16	26,879,542.76363	2314794.72854779

- When we visualize the data, we will get something that looks like this.


Sheet 1
Date
Sum Produce
40M
35M
30M
25M
20M
15M
10M
5M
0M
2012/03
2012/04
2012/05
2012/06
2012/07
2012/08
2012/09
2012/10
2012/11
2012/12
2013/01
2013/02
2013/03
2013/04
2013/05
2013/06
2013/07
2013/08
2013/09
2013/10
2013/11
2013/12
2014/01
2014/02
2014/03
2014/04
2014/05
2014/06
2014/07
2014/08
2014/09
2014/10
2014/11
2014/12
2015/01
2015/02
2015/03
2015/04
2015/05
2015/06
2015/07
2015/08
2015/09
2015/10
2015/11
2016/01
2016/02
2016/03
2016/04
2016/05
2016/06
2016/07
2016/08
2016/09
2016/10
2016/11
2016/12
Type
Existing
Forcast Existing
Forcast New
Sheet 1