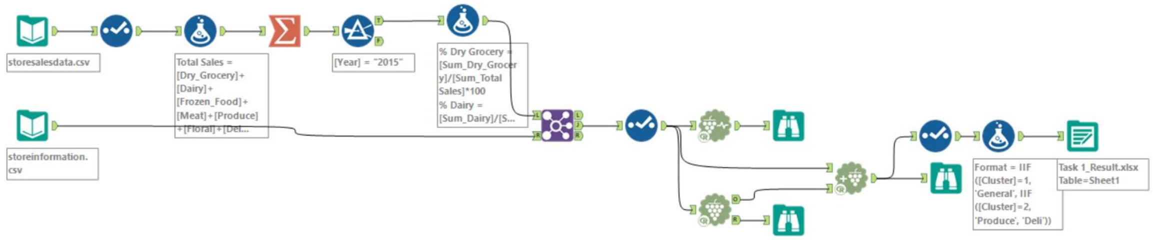
Project: Predictive Analytics Capstone

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?

   After applying k-means cluster tool (based on % category sales contribution in for 2015), the optimal number of store formats is 3 formats.



storesalesdata.csv

Total Sales = [Dry_Grocery]+ [Dairy]+ [Frozen_Food]+ [Meat]+[Produce] +[Floral]+[Del...

[Year] = "2015"

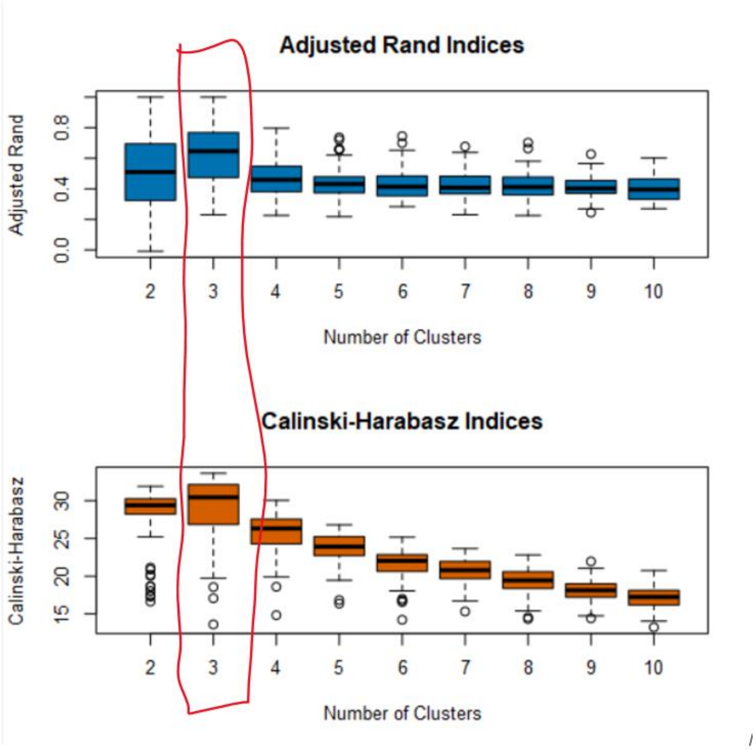
% Dry Grocery = [Sum_Dry_Grocery]/[Sum_Total Sales]*100 % Dairy = [Sum_Dairy]/[S...

storeinformation.csv

Format = IIF ([Cluster]=1, 'General', IIF ([Cluster]=2, 'Produce', 'Deli'))

Task 1_Result.xlsx Table=Sheet1

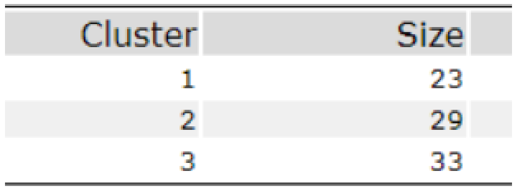
After carrying out the requisite steps in data preparation to generate data set with % category sales contribution by store for the year 2015, this dataset was fed into Cluster Diagnostics tool with k-means as clustering methodology. Analysis using both Adjusted Rand and Calinski-Harabasz Indices lead to selecting 3 clusters/formats.



2. How many stores fall into each store format?

Upon feeding the dataset to Cluster Analysis tool with number of clusters = 3 with k-means as clustering method, we can establish the number of stores that fall into each store format – 23 stores fall in cluster 1, 29 stores in cluster 2 and 33 stores in cluster 3.

: Awesome: The number of stores in each format is correct - great job!



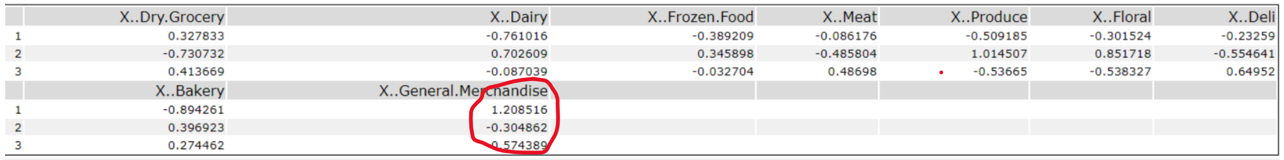
Cluster	Size
1	23
2	29
3	33

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

Based on results of clustering model, we can clearly establish that:

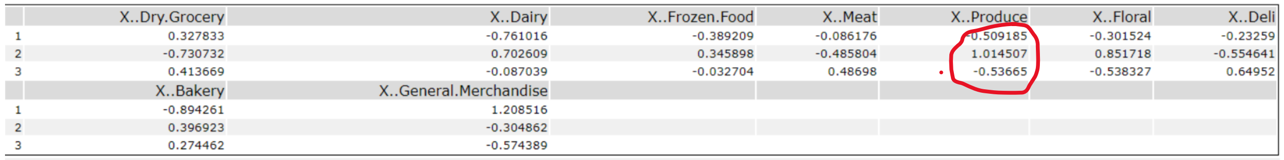
1. ***For Cluster 1 store formats, the contribution of General Merchandise is highest. We will call this cluster 'General' cluster.***

: Awesome: Excellent work providing observations about the difference among the clusters.



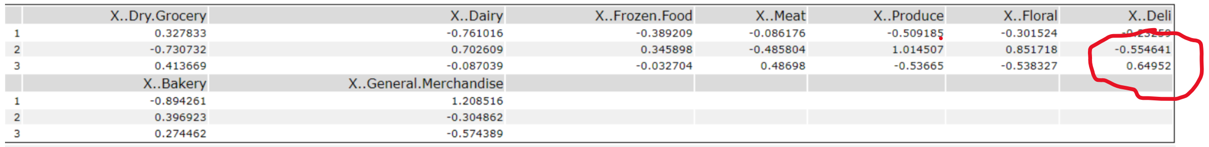
	X..Dry.Grocery	X..Dairy	X..Frozen.Food	X..Meat	X..Produce	X..Floral	X..Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952
	X..Bakery	X..General.Merchandise					
1	-0.894261	1.208516					
2	0.396923	-0.304862					
3	0.274462	-0.574389					

2. ***For Cluster 2 store formats, the contribution of Produce is highest. We will call this cluster 'Produce' cluster.***



	X..Dry.Grocery	X..Dairy	X..Frozen.Food	X..Meat	X..Produce	X..Floral	X..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
	X..Bakery	X..General.Merchandise					
1	-0.894261	1.208516					
2	0.396923	-0.304862					
3	0.274462	-0.574389					

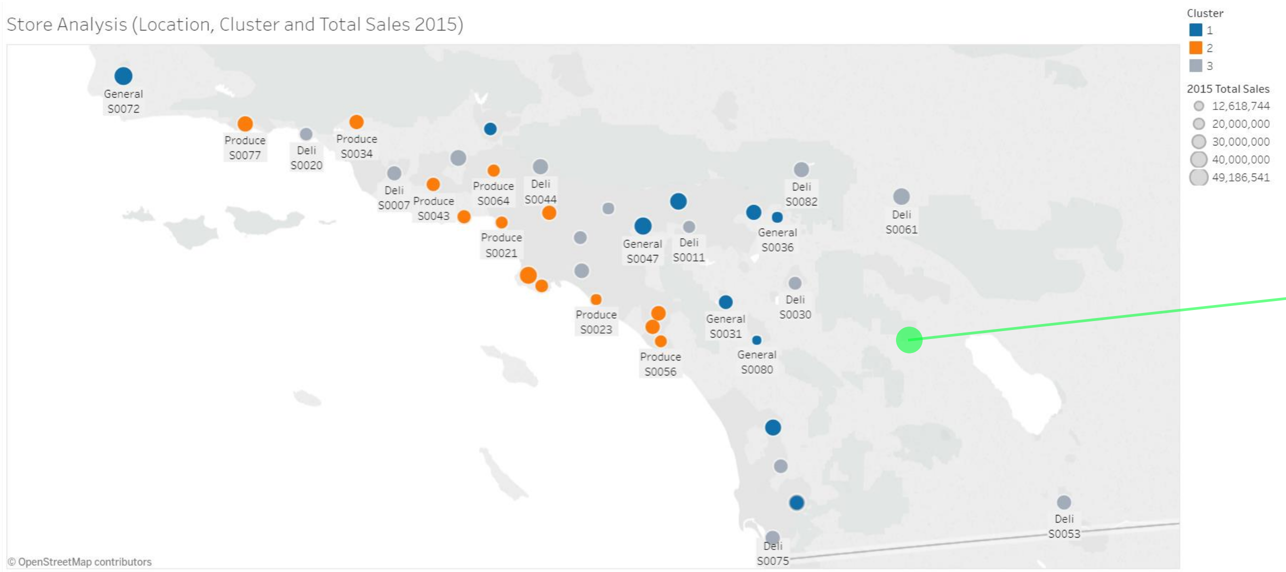
3. ***For Cluster 3 store formats, the contribution of Deli is highest. We will call this cluster 'Deli' cluster.***



	X..Dry.Grocery	X..Dairy	X..Frozen.Food	X..Meat	X..Produce	X..Floral	X..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
	X..Bakery	X..General.Merchandise					
1	-0.894261	1.208516					
2	0.396923	-0.304862					
3	0.274462	-0.574389					

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.

https://public.tableau.com/profile/karthik.subramanian#!/vizhome/Task1_Visualization_4/Sheet1?publish=yes

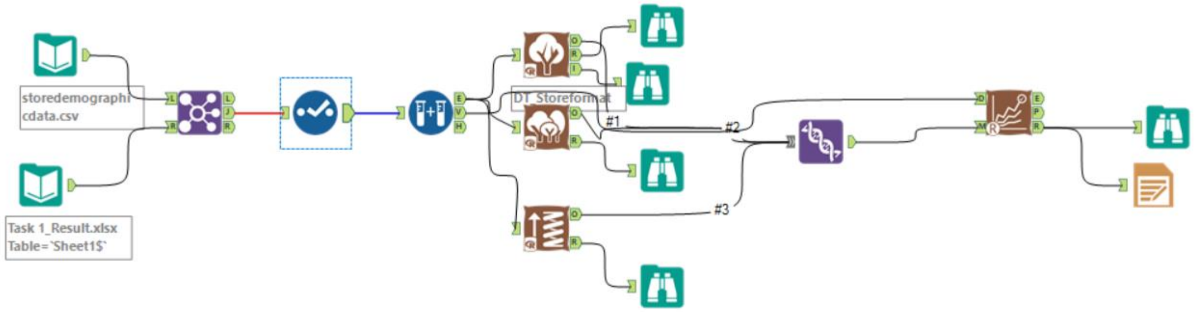


: Awesome: The map looks great. It has legends. Color is used to show the clusters and size is used to show total sales.

Task 2: Formats for New Stores

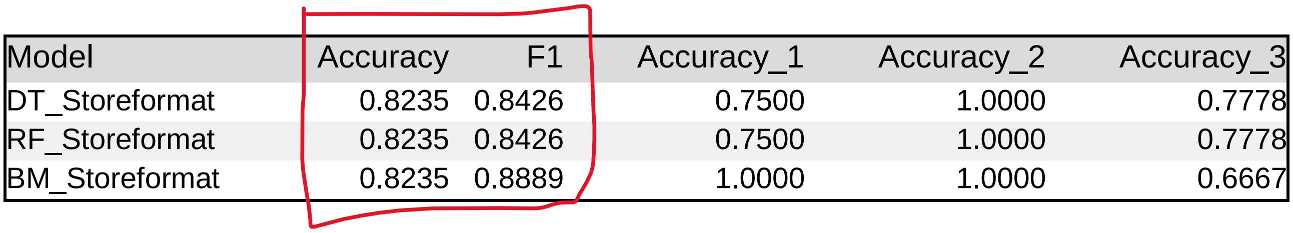
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)
   I have used Boosted Model (BM) classification methodology to predict best store format for new stores after trying out all three non-binary classification models - Decision Tree, Random Forest and Boosted Model using Alteryx workflow (below).

: Awesome: Yes, we can see that the Boosted model should be used since it has high accuracy and higher F1 score. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.



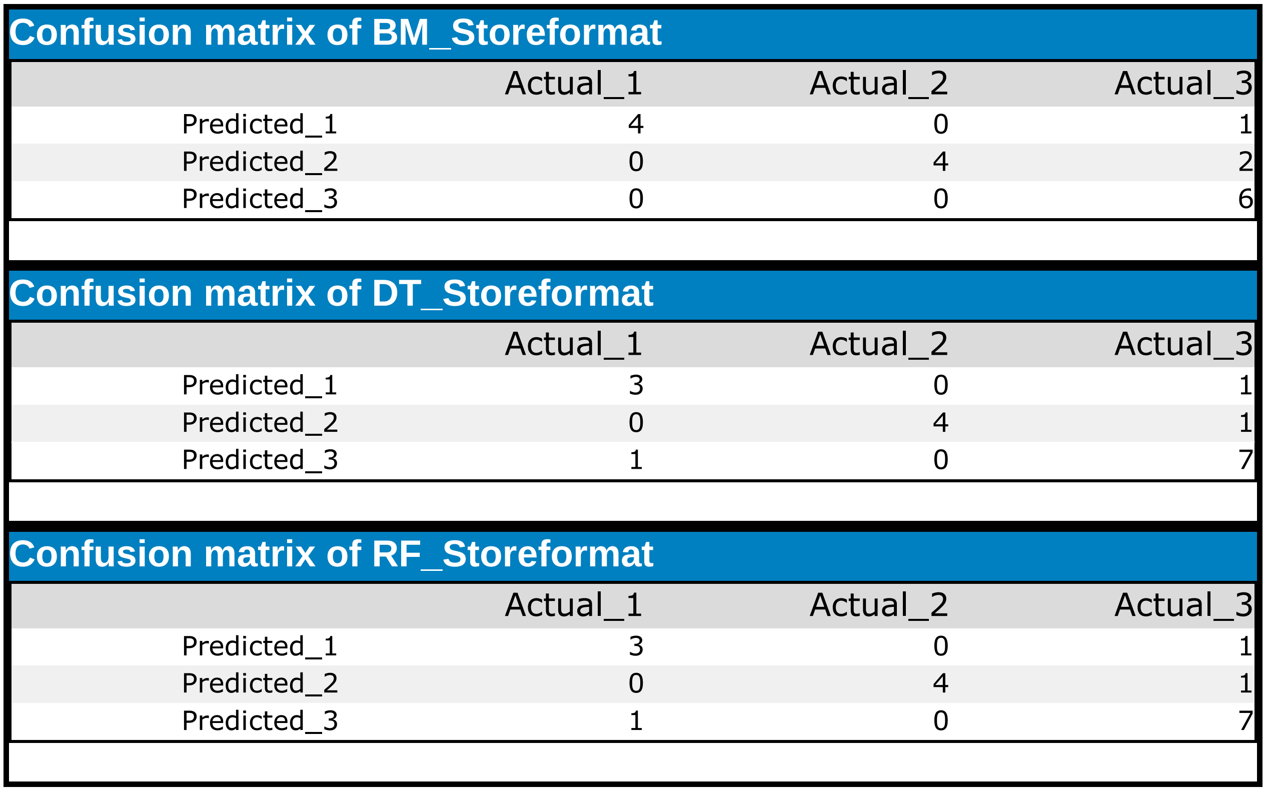
It is clear from model comparison results (below) that Boosted Model methodology has the highest F1 score even though all three non-binary classification methodologies –

Decision Tree, Random Forest and Boosted Model have similar overall accuracies.



Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Storeformat	0.8235	0.8426	0.7500	1.0000	0.7778
RF_Storeformat	0.8235	0.8426	0.7500	1.0000	0.7778
BM_Storeformat	0.8235	0.8889	1.0000	1.0000	0.6667

While Decision Tree methodology and Random Forest methodology score slightly higher on accuracy of predicting Cluster 3 (Deli), Boosted model methodology has perfect accuracy 1.0 in predicting both Cluster 1 (General) and Cluster 2 (Produce).



Confusion matrix of BM_Storeformat

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of DT_Storeformat

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

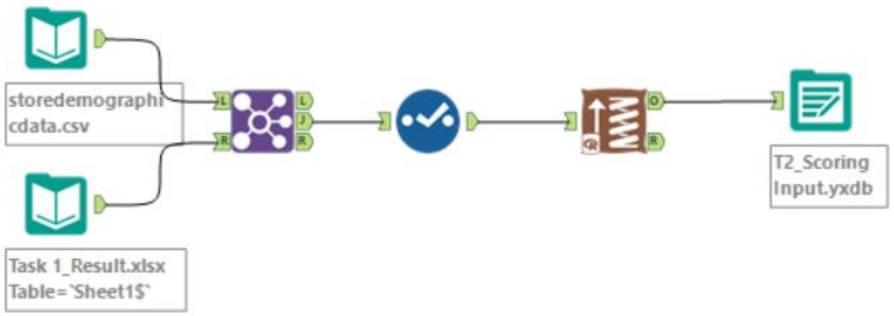
Confusion matrix of RF_Storeformat

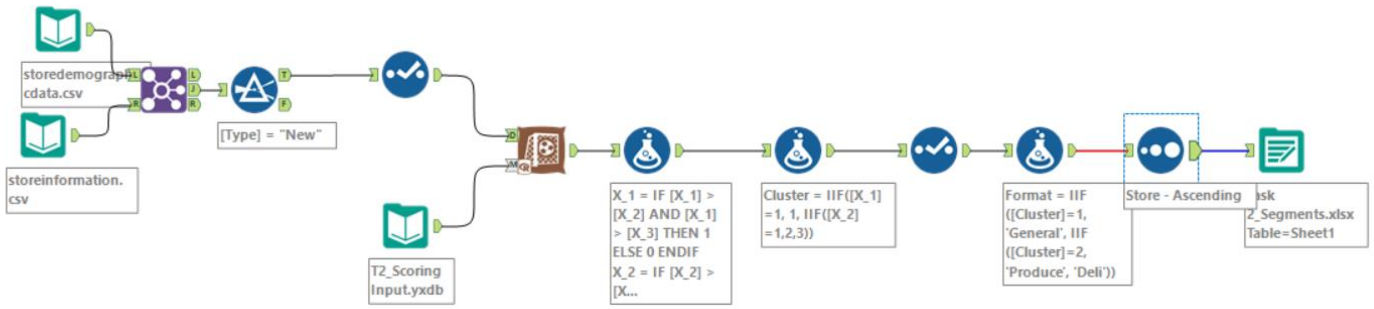
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

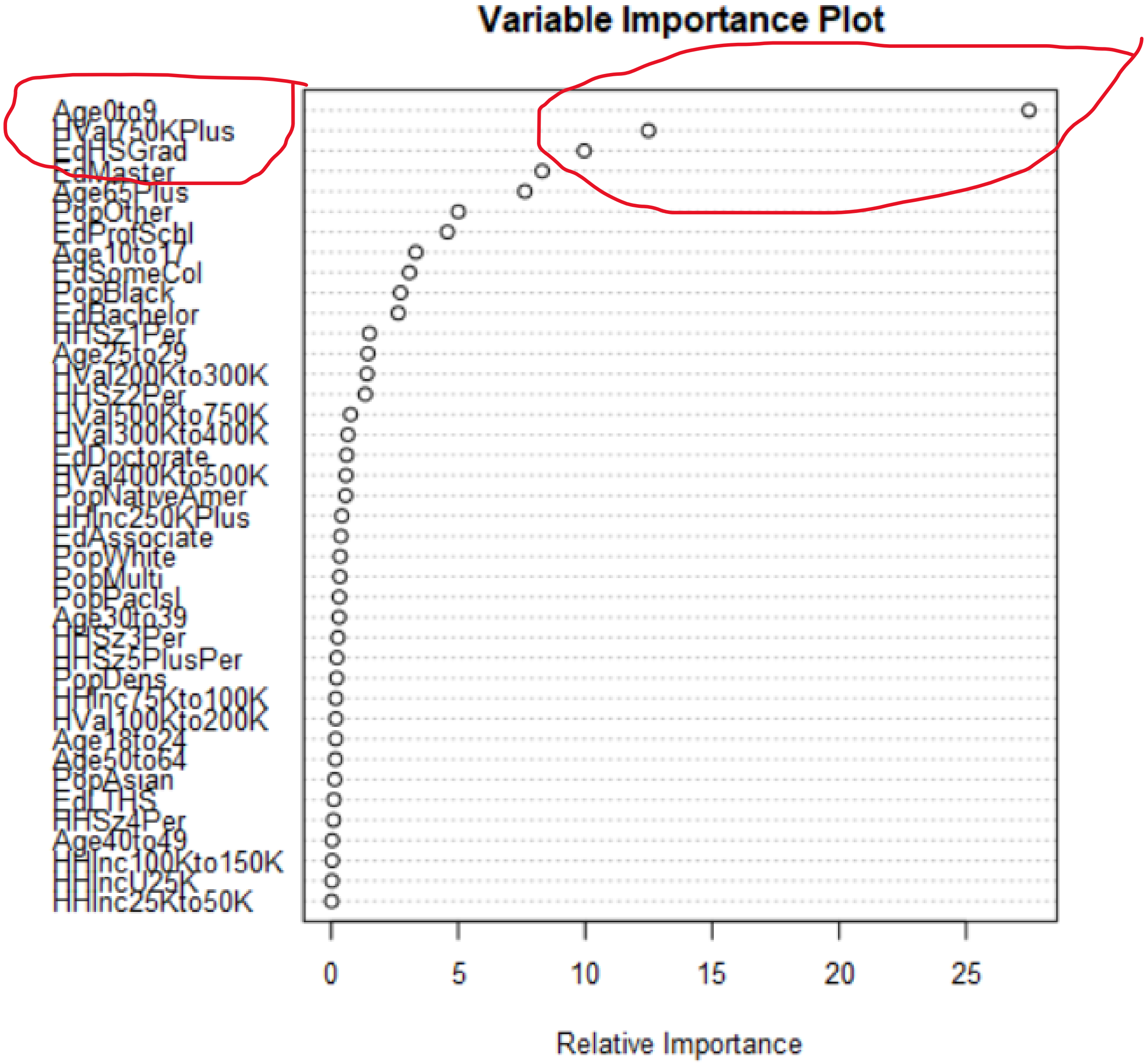
Store	City	Zip	Cluster	Format
S0086	Gilroy	95020	1	***General***
S0087	San Jose	95118	2	***Produce***
S0088	San Jose	95122	3	***Deli***
S0089	San Jose	95117	2	***Produce***
S0090	San Mateo	94403	2	***Produce***
S0091	Tracy	95304	1	***General***
S0092	El Cerrito	94530	2	***Produce***
S0093	Antioch	94531	1	***General***
S0094	Walnut Creek	94596	2	***Produce***
S0095	Pleasant Hill	94523	2	***Produce***

: Awesome: The stores are correctly segmented - great job!




storedemographicdata.csv
storeinformation.csv
[Type] = "New"
T
F
T2_Scoring Input.yxdb
X_1 = IF [X_1] > [X_2] AND [X_1] > [X_3] THEN 1 ELSE 0 ENDIF X_2 = IF [X_2] > [X...
Cluster = IIF([X_1] =1, 1, IIF([X_2] =1,2,3))
Format = IIF ([Cluster]=1, 'General', IIF ([Cluster]=2, 'Produce', 'Deli'))
Store - Ascending
ısk 2_Segments.xlsx Table=Sheet1

Three most significant variables for Boosted model are Age 0 to 9, HVal 750K Plus and EdHSGrad as illustrated in the Variable Importance Plot below:

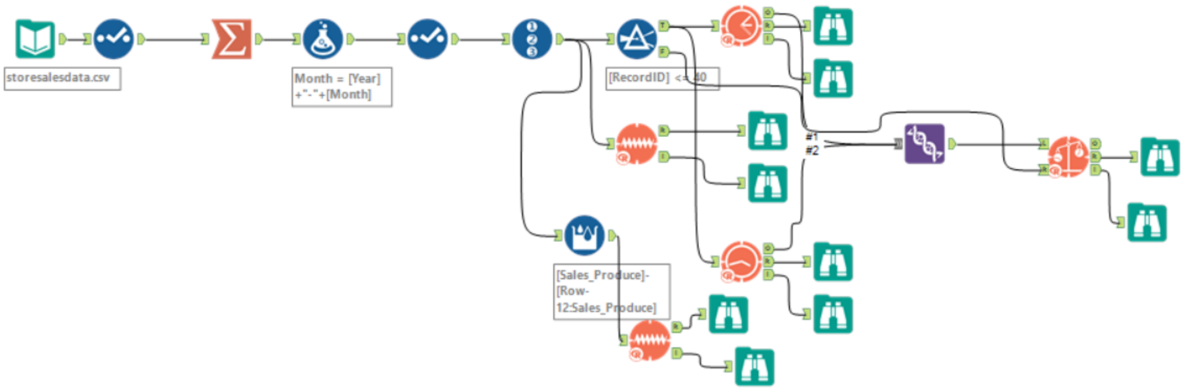


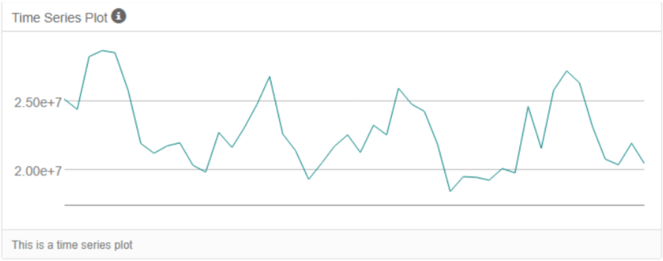
Task 3: Predicting Produce Sales

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?

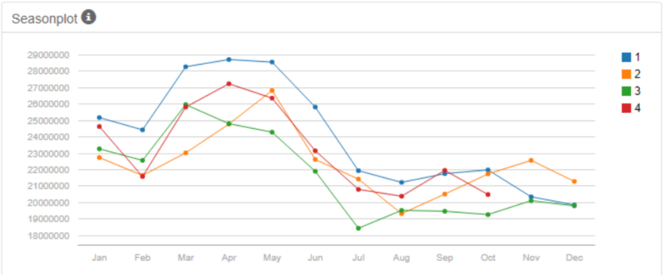
<u>Forecast Existing Stores:</u>

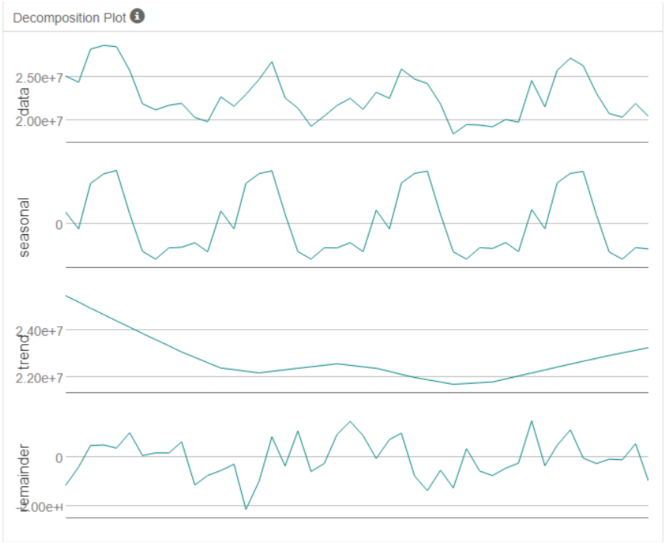
To forecast for sales of produce for existing stores, we first identify the appropriate forecast model to use through the following workflow (I have used ETS(M,N,M) and ARIMA(1,0,0)(1,1,0)[12] after evaluating relevant TS plots for comparison):

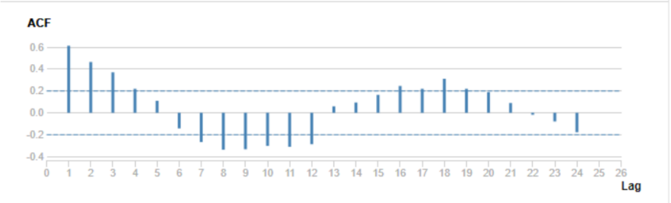

storesalesdata.csv
Month = [Year]+"-"+[Month]
[RecordID] <= 40
[Sales_Produce]-[Row-12:Sales_Produce]
#1
#2

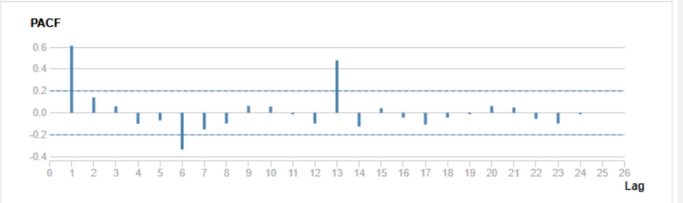


This is a time series plot

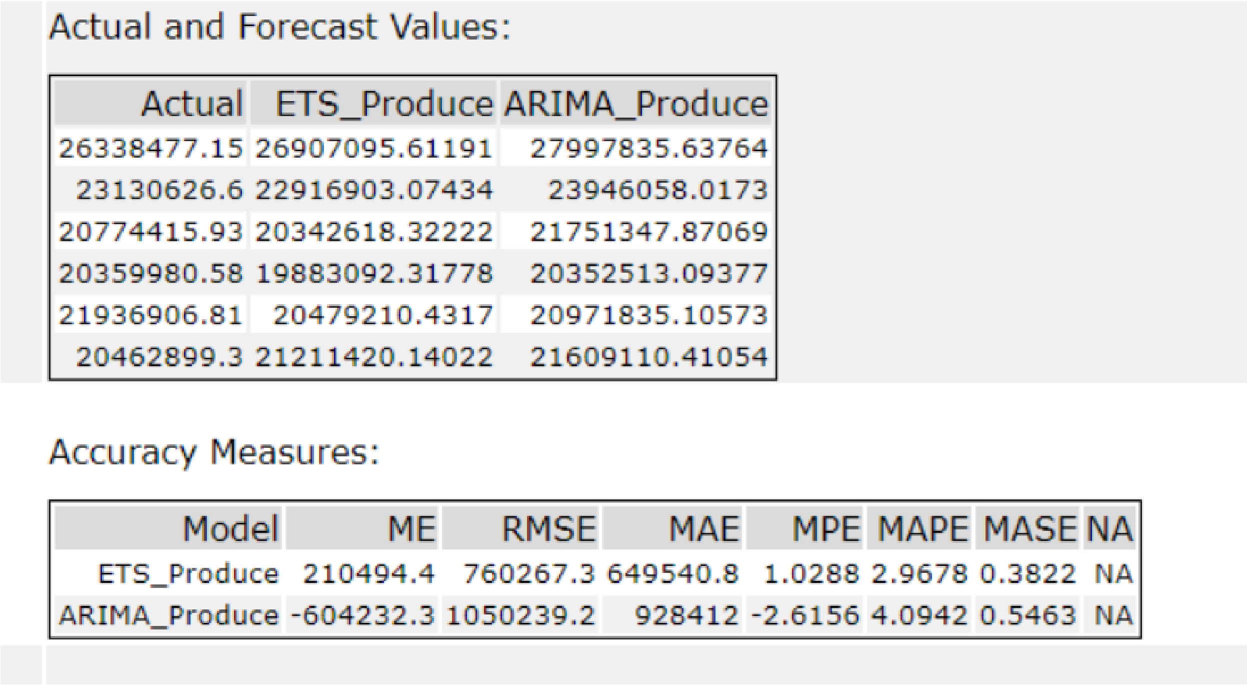








Using the TS Compare tool and by observing the forecast error measurements against the holdout sample, it is evident that ETS(M,N,M) forecast model performs better than ARIMA (1,0,0)(1,1,0)[12] forecast model for forecasting existing store sales.



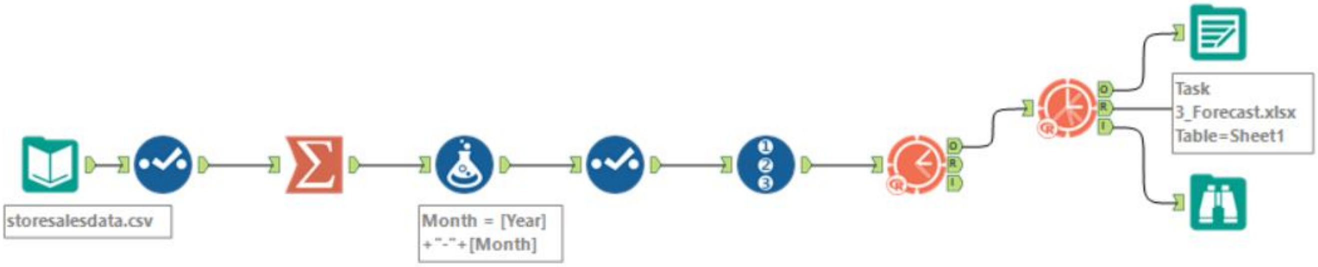
Actual and Forecast Values:

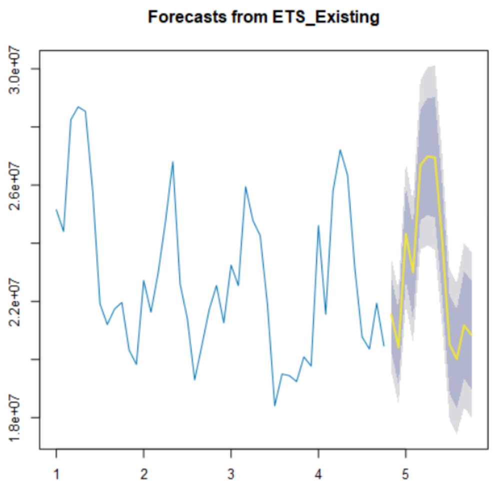
Actual	ETS_Produce	ARIMA_Produce
26338477.15	26907095.61191	27997835.63764
23130626.6	22916903.07434	23946058.0173
20774415.93	20342618.32222	21751347.87069
20359980.58	19883092.31778	20352513.09377
21936906.81	20479210.4317	20971835.10573
20462899.3	21211420.14022	21609110.41054

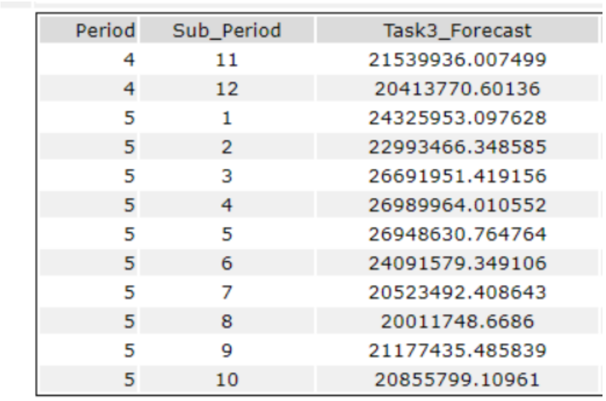
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Produce	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
ARIMA_Produce	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463	NA

Therefore, I have used ETS(M,N,M) forecast model to forecast existing store sales for all 12 months of 2016 using the following workflow:


storesalesdata.csv
Month = [Year] +"-"+[Month]
O
R
I
Task 3_Forecast.xlsx Table=Sheet1

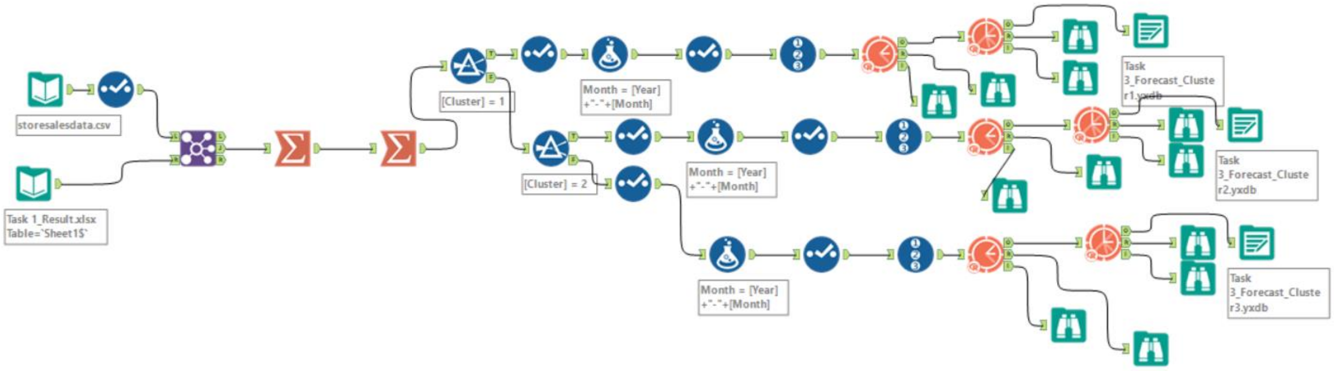




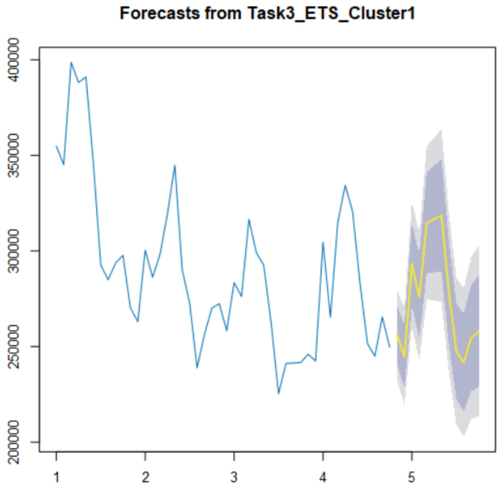
Period	Sub_Period	Task3_Forecast
4	11	21539936.007499
4	12	20413770.60136
5	1	24325953.097628
5	2	22993466.348585
5	3	26691951.419156
5	4	26989964.010552
5	5	26948630.764764
5	6	24091579.349106
5	7	20523492.408643
5	8	20011748.6686
5	9	21177435.485839
5	10	20855799.10961

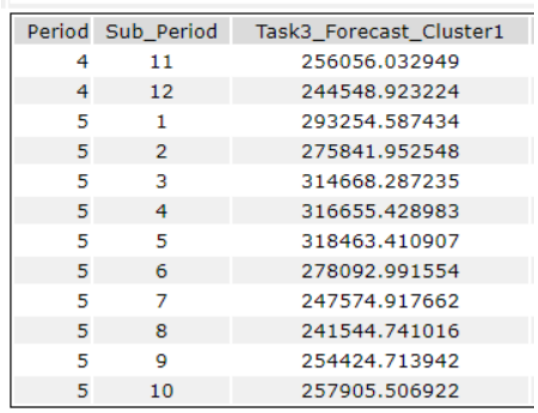
<u>Forecast New Stores:</u>

I have used ETS(M,N,M) forecast model to forecast average sales of produce per store for each of the Clusters 1, 2 and 3 for all 12 months of 2016 using the following workflow:



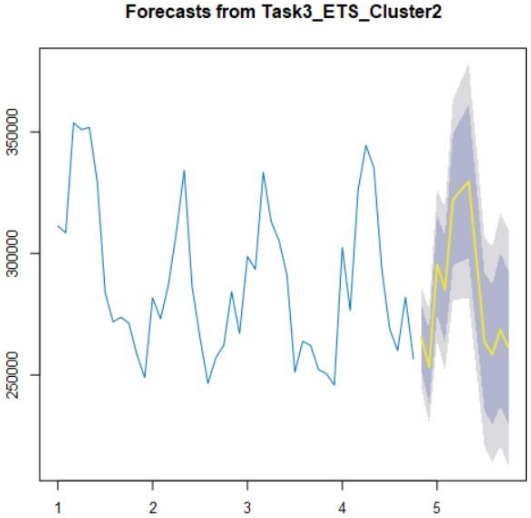
Cluster 1

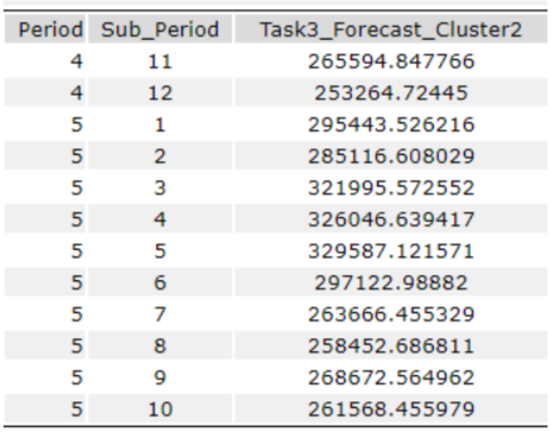




Period	Sub_Period	Task3_Forecast_Cluster1
4	11	256056.032949
4	12	244548.923224
5	1	293254.587434
5	2	275841.952548
5	3	314668.287235
5	4	316655.428983
5	5	318463.410907
5	6	278092.991554
5	7	247574.917662
5	8	241544.741016
5	9	254424.713942
5	10	257905.506922

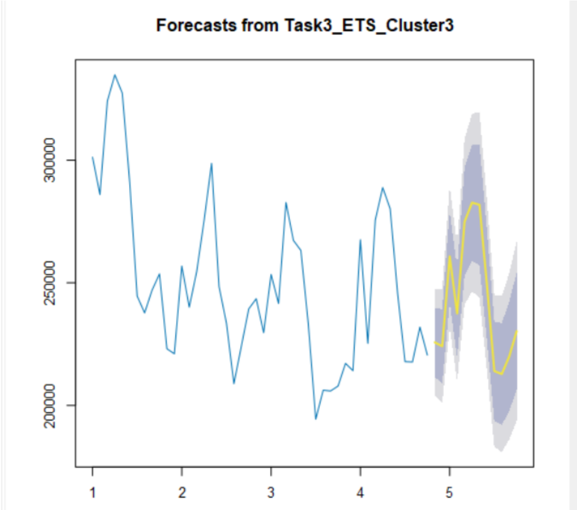
Cluster 2

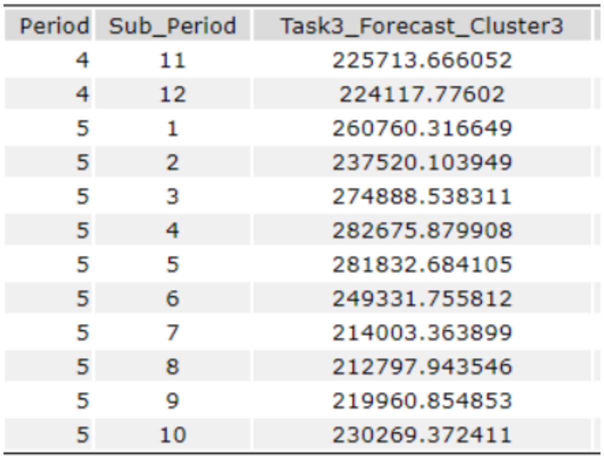




Period	Sub_Period	Task3_Forecast_Cluster2
4	11	265594.847766
4	12	253264.72445
5	1	295443.526216
5	2	285116.608029
5	3	321995.572552
5	4	326046.639417
5	5	329587.121571
5	6	297122.98882
5	7	263666.455329
5	8	258452.686811
5	9	268672.564962
5	10	261568.455979

Cluster 3





Period	Sub_Period	Task3_Forecast_Cluster3
4	11	225713.666052
4	12	224117.77602
5	1	260760.316649
5	2	237520.103949
5	3	274888.538311
5	4	282675.879908
5	5	281832.684105
5	6	249331.755812
5	7	214003.363899
5	8	212797.943546
5	9	219960.854853
5	10	230269.372411

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.

Using the forecasts of average sales of produce per store for each of the clusters 1, 2 and 3, we derive forecast of sales of produce for all new sales using the following workflow (mainly multiplying the forecast of average sales per store by each cluster with number of new stores in each cluster)

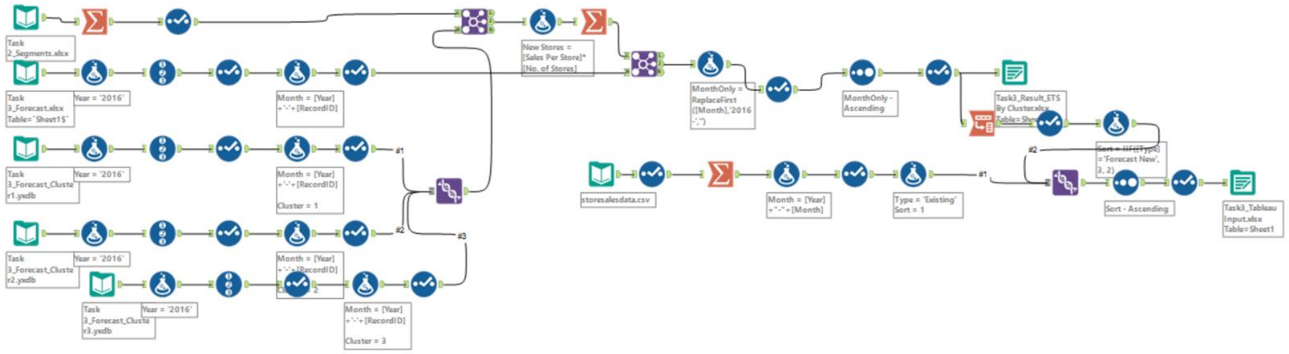


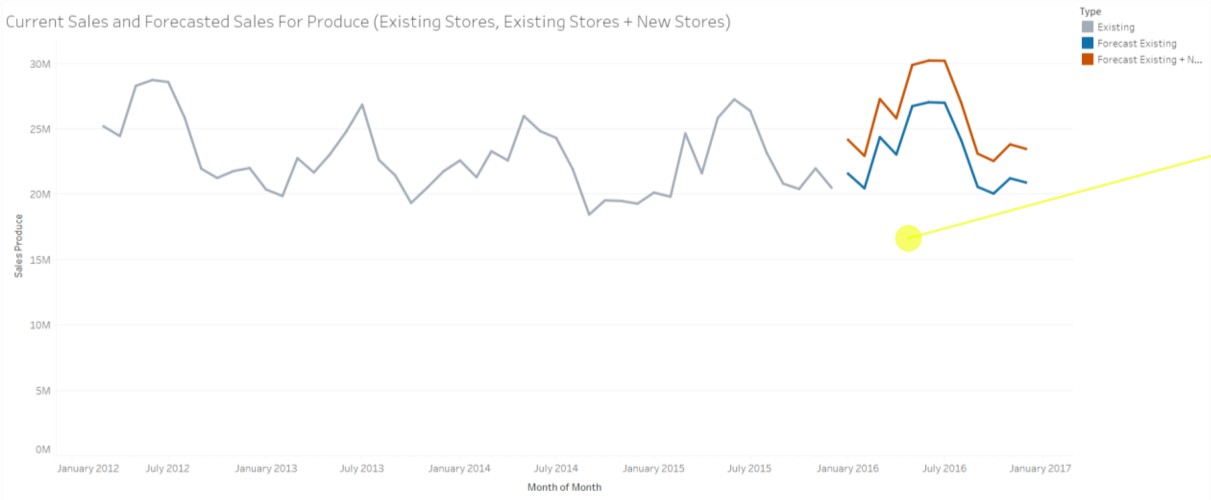
Table of forecasts for existing and new stores using ETS (M,N,M) forecast model for Existing as well as New Stores (Cluster 1, 2 and 3) provides the following results:

Month	***Forecast New Stores***	***Forecast Existing Stores***
2016-Jan	***2587451***	***21539936***
2016-Feb	***2477353***	***20413771***
2016-Mar	***2913185***	***24325953***
2016-Apr	***2775746***	***22993466***
2016-May	***3150867***	***26691951***
2016-Jun	***3188922***	***26989964***
2016-Jul	***3214746***	***26948631***
2016-Aug	***2866349***	***24091579***
2016-Sep	***2538727***	***20523492***
2016-Oct	***2488148***	***20011749***
2016-Nov	***2595270***	***21177435***
2016-Dec	***2573397***	***20855799***

: Awesome: The forecasts for the existing and new stores are within the expected range - great job! Great job plotting the results!

Visualization of forecasts that includes historical data, existing stores forecasts, and new stores forecasts is provided in the Tableau link below:

https://public.tableau.com/profile/karthik.subramanian#!/vizhome/Task3_Visualization_0/Sheet1?publish=yes



: Suggestion: Great job! Great job with the plot! We suggest presenting the plot as an area chart not as a line chart. And by the way, if you are interested how to close the gap in the plot between the actual sales and the forecasted ones you can check the example in the project review section.

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.