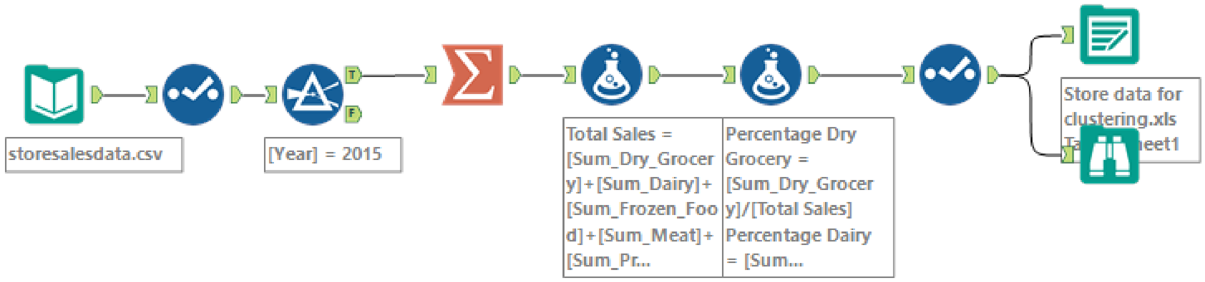
Project: Predictive Analytics Capstone

Jing Huang

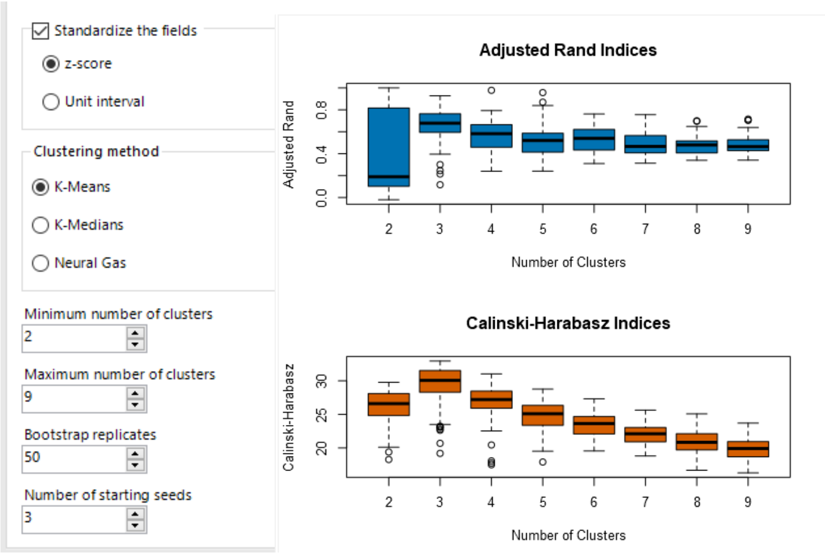
Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

   To find out what's the optimal number of store formats, we need to do clustering analysis on store sales data. As mentioned in the project we need to prepare a data set with category sales as a percentage of total store sales for 2015 sales data.
   I used below Alteryx workflow including aggregation tool, formula tool and filter tool to prepare the data set, and output the data as "Store data for clustering".


storesalesdata.csv
[Year] = 2015
T
F
Total Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
Percentage Dry Grocery = [Sum_Dry_Grocery]/[Total Sales] Percentage Dairy = [Sum...
Store data for clustering.xls
Ta eet1

   Input "Store data for clustering" data into Alteryx and connect it to a cluster diagnostic tool, set up the configuration as below. We can see from below the result report on the right that adjusted rand index and CH index both indicates that 3 clusters give the data the most compactness and distinctiveness, and is the best choice.



☑ Standardize the fields

● z-score

○ Unit interval

Clustering method

● K-Means

○ K-Medians

○ Neural Gas

Minimum number of clusters: 2

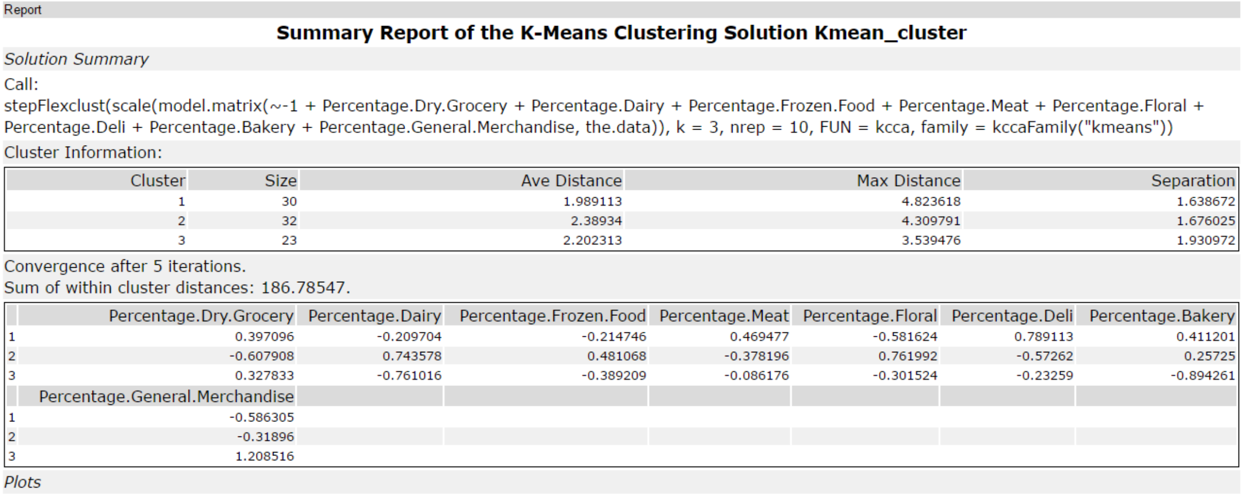
Maximum number of clusters: 9

Bootstrap replicates: 50

Number of starting seeds: 3

2. How many stores fall into each store format?

   Connect a Cluster Analysis tool to "Sales data for clustering" data, set up the configuration as 3 clusters, from the result report below we can see that cluster 1 has 30 stores, cluster 2 has 32 stores, cluster 3 has 23 stores; the most spread out cluster is cluster 2 with average distance of 2.389, and biggest cluster is cluster with Max distance of 4.8, cluster 3 located the farthest from the other 2 clusters.



Report

Summary Report of the K-Means Clustering Solution Kmean_cluster

Solution Summary

Call:

stepFlexclust(scale(model.matrix(~-1 + Percentage.Dry.Grocery + Percentage.Dairy + Percentage.Frozen.Food + Percentage.Meat + Percentage.Floral + Percentage.Deli + Percentage.Bakery + Percentage.General.Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	30	1.989113	4.823618	1.638672
2	32	2.38934	4.309791	1.676025
3	23	2.202313	3.539476	1.930972

Convergence after 5 iterations.
Sum of within cluster distances: 186.78547.

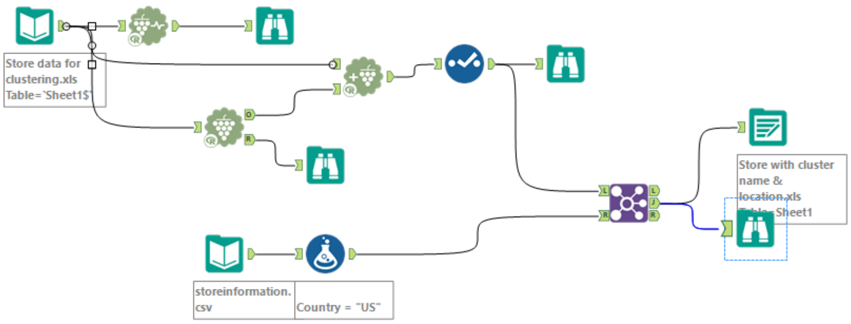
	Percentage.Dry.Grocery	Percentage.Dairy	Percentage.Frozen.Food	Percentage.Meat	Percentage.Floral	Percentage.Deli	Percentage.Bakery
1	0.397096	-0.209704	-0.214746	0.469477	-0.581624	0.789113	0.411201
2	-0.607908	0.743578	0.481068	-0.378196	0.761992	-0.57262	0.25725
3	0.327833	-0.761016	-0.389209	-0.086176	-0.301524	-0.23259	-0.894261

	Percentage.General.Merchandise
1	-0.586305
2	-0.31896
3	1.208516

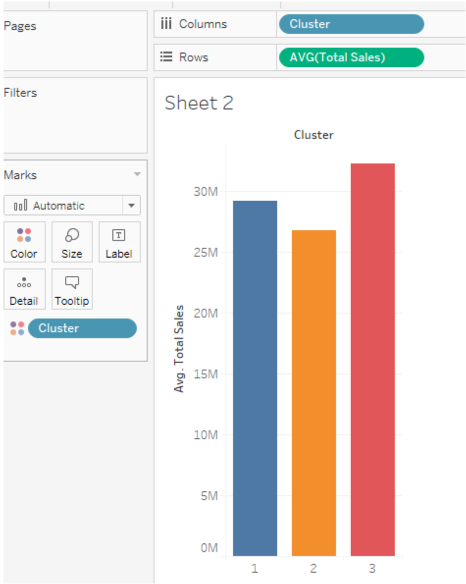
Plots

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

Join the cluster analysis result with Store information file, and added a country (value=US) column to the data, output the result as "Store with cluster name & location". (The detailed Alteryx workflow is as below)


Store data for clustering.xls Table=`Sheet1$`
D
R
L
J
R
L
R
Store with cluster name & location.xls Table=`Sheet1
storeinformation.csv
Country = "US"

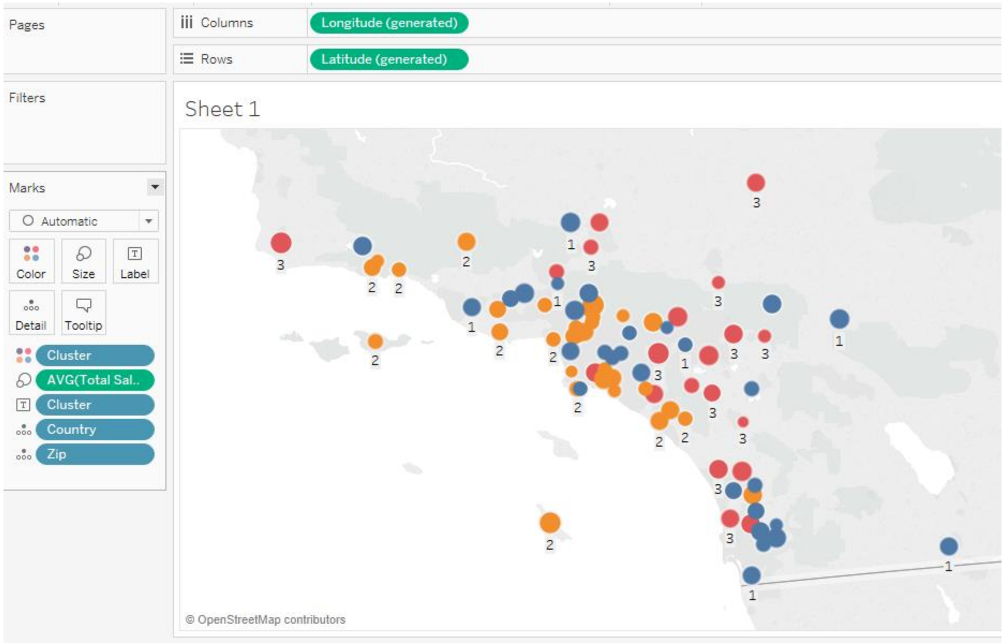
Inputted this dataset into Tableau, put average of total sales onto rows, and cluster name onto column, we can see in below chart that on average stores in cluster 3 have the most sales, cluster 2 stores have the least average sales, cluster 1 stores are in the middle.



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.

Connect dataset "Store with cluster name & location" into Tableau, created a map using latitude and longitude, put cluster onto color mark, total sales onto size, the visualization is as below:

(Tableau public file location: https://public.tableau.com/profile/angela.huang#!/vizhome/85storeswithclusters/Sheet1?publish=yes)

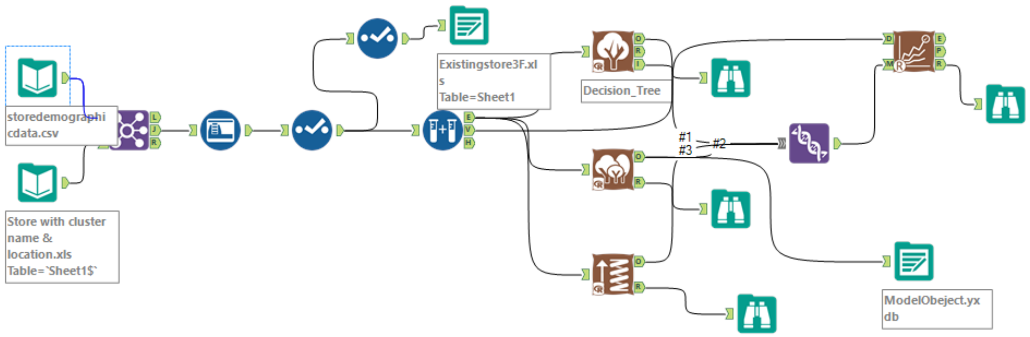


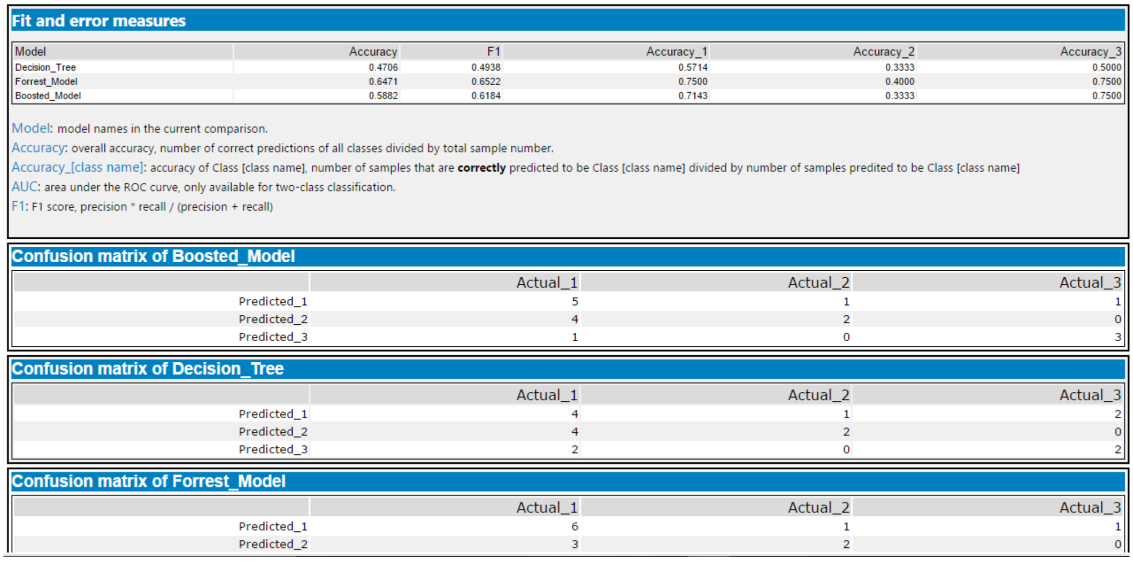
Based on above clustering result, we could arrange 3 different kinds of product selection and store layouts for the 85 existing stores.

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.)

Combine “Store with cluster name & location” data with “Storedemographic” data, use “sample tool” to create 80% of data for sampling and 20% data for validation, and connect the joined data with “Decision Tree Model”, “Forrest Model”, and “Boost Model” separately (because it’s non-binary classification analysis), multijoin their outputs, and connect to a “Model Comparison tool”, the Alteryx workflow and analysis result are as below:


storedemographicdata.csv
Store with cluster name & location.xls Table=`Sheet1$`
Existingstore3F.xls Table=Sheet1
Decision_Tree
#1
#3
#2
ModelObeject.yxdb



Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.4706	0.4938	0.5714	0.3333	0.5000
Forrest_Model	0.6471	0.6522	0.7500	0.4000	0.7500
Boosted_Model	0.5882	0.6184	0.7143	0.3333	0.7500

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	5	1	1
Predicted_2	4	2	0
Predicted_3	1	0	3

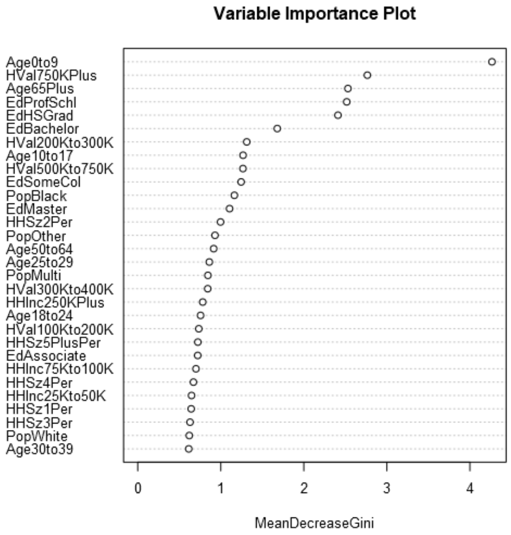
Confusion matrix of Decision_Tree

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

Confusion matrix of Forrest_Model

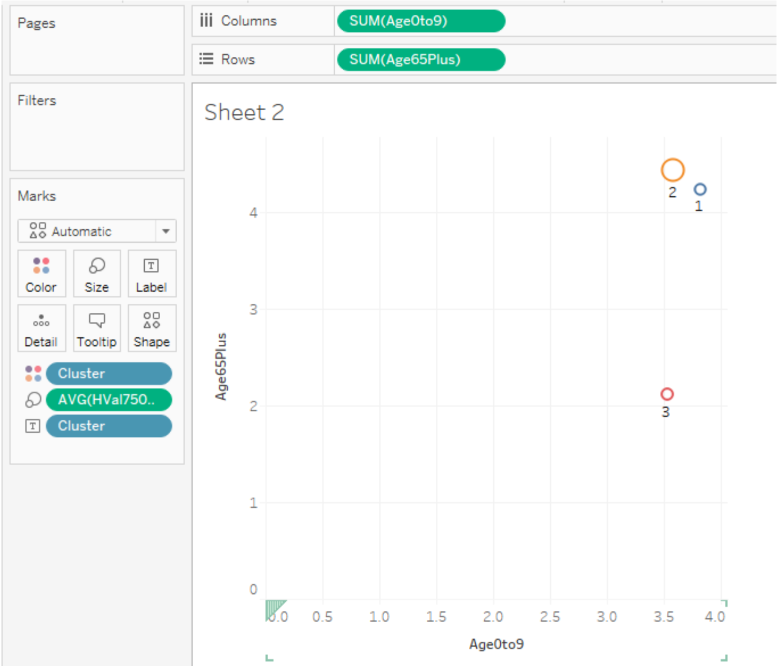
	Actual_1	Actual_2	Actual_3
Predicted_1	6	1	1
Predicted_2	3	2	0

From above classification analysis result we can see that “Forrest Model” has the best accuracy result of 64.7% among those 3 models, it has 75% prediction accuracy on predict cluster 1 and cluster 3 segments, so I decided to use Forrest Model as my prediction model to predict the cluster name for new stores, and output Forrest Model analysis result as “ModelObject.yxdb”.



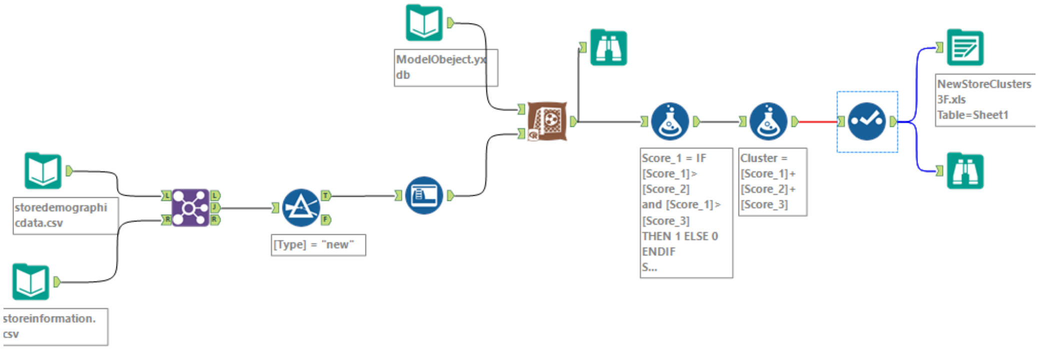
From the Forrest model analysis report above we can also see that the 3 most important prediction variables are: Age 0 to 9, HVal 750Kplus, and Age65Plus.

Connect the store demographic containing only the 3 most important factors and cluster name (save the file as “Existingstore3F.xls”) into Tableau, put Age 0-9 population percentage and Age 65+ population percentage onto column and row separately, put average income 750K+ onto size mark, and cluster name onto color and label marks. From below chart we can see that cluster3 has the most older populations and the least money (the size of income 750Kplus is smallest); cluster 2 is the richest cluster with the most Age0-9 kids at home. (Tableau Public file: https://public.tableau.com/profile/angela.huang#!/vizhome/Plot3importantfactors/Sheet1?publish=yes)



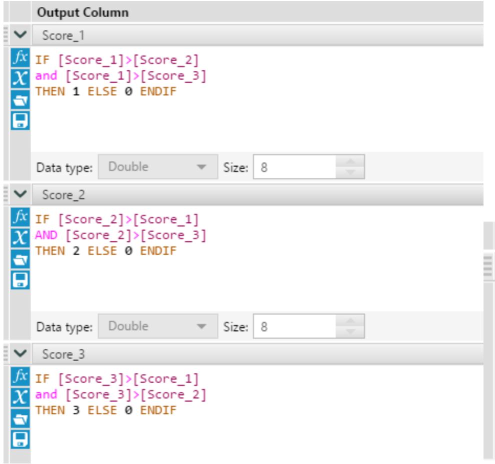
~~3.~~2. What format do each of the 10 new stores fall into? Please fill in the table below.

Connect a Score Tool to the joined store demographic data and store location data, connect another input of Score Tool to ModelObject.yxbd, running the score card, and apply 2 formula tools as below to give cluster name to each new store. (the detailed Alteryx workflow is as below):



```
Score_1 = IF
[Score_1]>
[Score_2]
and [Score_1]>
[Score_3]
THEN 1 ELSE 0
ENDIF
S...
```

```
Cluster =
[Score_1]+
[Score_2]+
[Score_3]
```



Output Column

Score_1

```
IF [Score_1]>[Score_2]
and [Score_1]>[Score_3]
THEN 1 ELSE 0 ENDIF
```

Data type: Double Size: 8

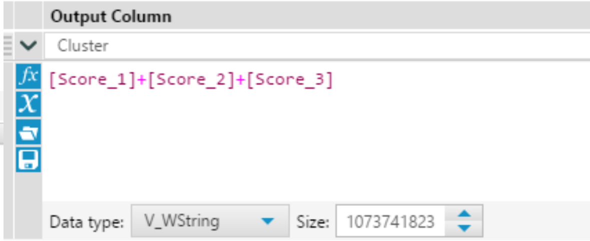
Score_2

```
IF [Score_2]>[Score_1]
AND [Score_2]>[Score_3]
THEN 2 ELSE 0 ENDIF
```

Data type: Double Size: 8

Score_3

```
IF [Score_3]>[Score_1]
and [Score_3]>[Score_2]
THEN 3 ELSE 0 ENDIF
```



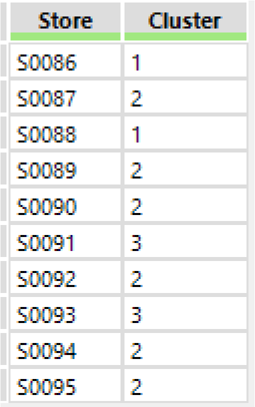
Output Column

Cluster

```
[Score_1]+[Score_2]+[Score_3]
```

Data type: V_WString Size: 1073741823

The predicted clusters for 10 new stores are as below:



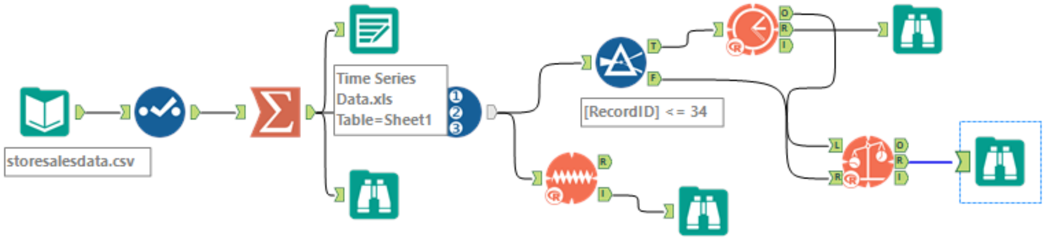
Store	Cluster
S0086	1
S0087	2
S0088	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Task 3: Predicting Produce Sales

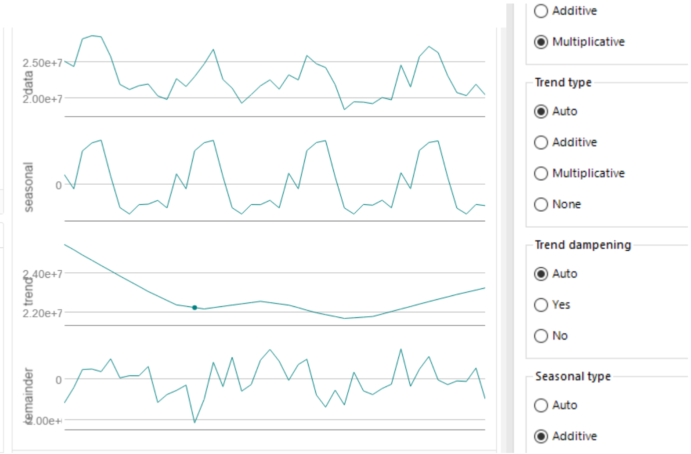
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

Forecasting Sales for Existing Stores:

1. **ETS Model:**



As show from above Alteryx workflow, use a summary tool to summarize all existing store sales by year and month, connect to a TS plot tool, the analysis result below tell us that the variance of seasonal factor is constant, so the seasonal term in ETS model is addictive, the trend is decreasing and backup upwards again, so we couldn’t decide its term, set it up as Auto, Error term’s variance is changing over time, so set up error term as multiplicative, trend damping is Auto.



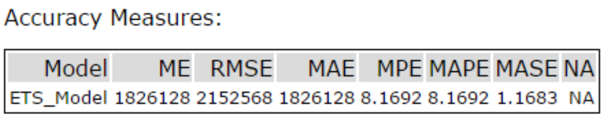
○ Additive
● Multiplicative

Trend type
● Auto
○ Additive
○ Multiplicative
○ None

Trend dampening
● Auto
○ Yes
○ No

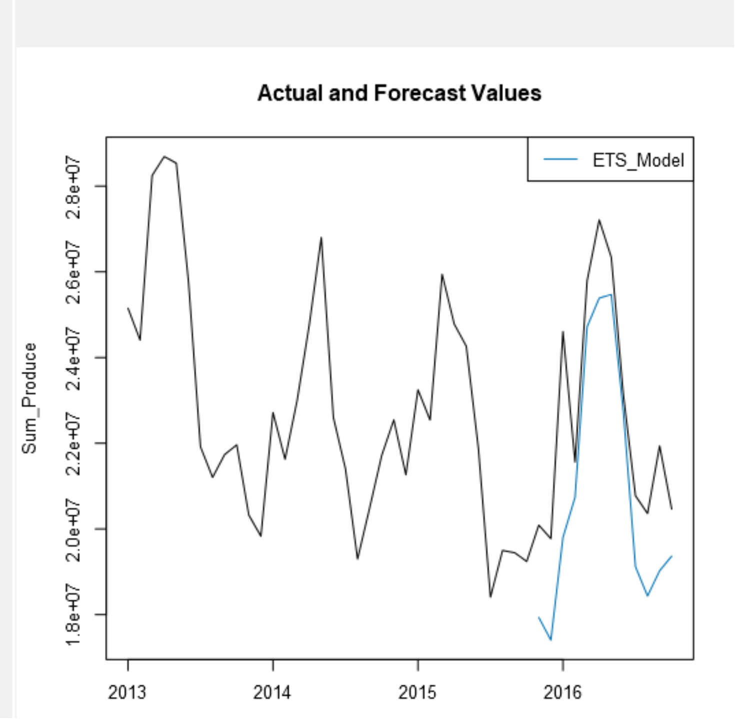
Seasonal type
○ Auto
● Additive

Filter out the most recent 12 months data for validation(cause forecasting period is 12 months), connect the summarized sales data for each month with a ETS Model Tool, set up E. T. S. figures configuration (x,0,+), connect the ETS model output to a TS Comparison tool, and using the validation dataset to validate the ETS model result, the analysis result below shows that the variance of difference between the forecast and the actual sales data are big (RMSE=2152,568), also both ME and MAE are quite big too, MASE is greater than 1 also shows it’s not good model.

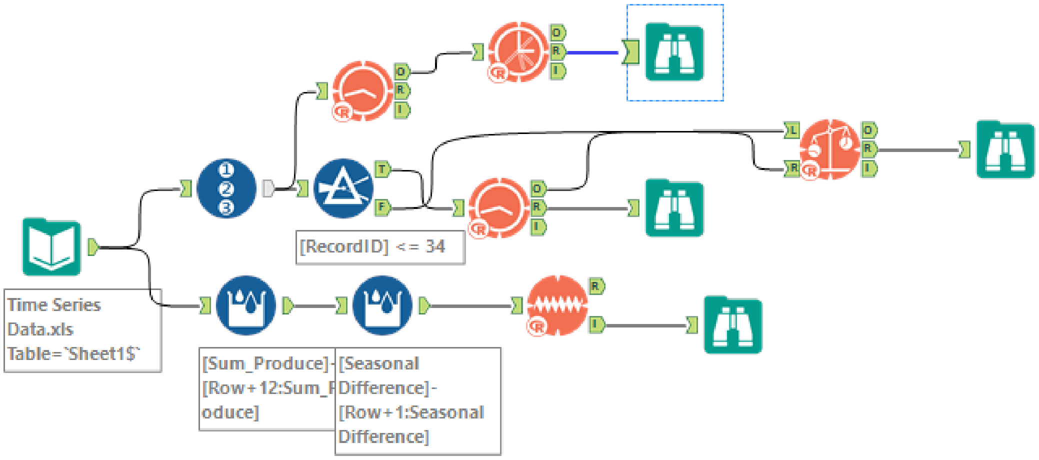


Accuracy Measures:

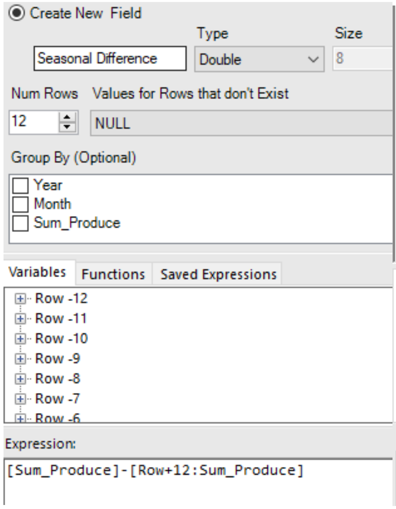
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Model	1826128	2152568	1826128	8.1692	8.1692	1.1683	NA



2. ARIMA Model:


Time Series Data.xls Table=`Sheet1$`
[RecordID] <= 34
[Sum_Produce]-[Row+12:Sum_Produce]
[Seasonal Difference]-[Row+1:Seasonal Difference]
D
R
I
T
F
L

Following above workflow, connect 2 Multi row formulas to the "Time Series Data" (saved from above ETS model process, set up the two formulas as below, connect a TS tool to the formulized data, the TS tool result of ACF and PACF plots are as below lower chart:



● Create New Field

Type: Double
Size: 8

Seasonal Difference

Num Rows: 12

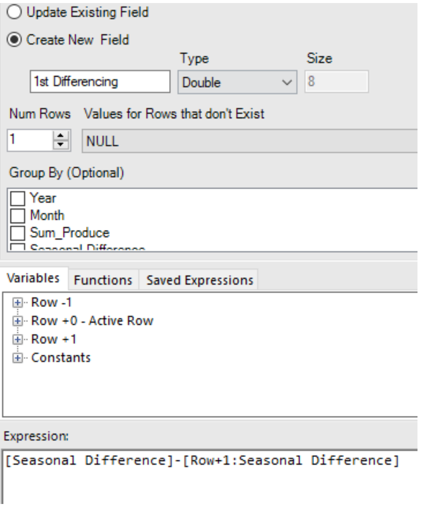
Values for Rows that don't Exist: NULL

Group By (Optional)

□ Year
□ Month
□ Sum_Produce

Expression:

```
[Sum_Produce]-[Row+12:Sum_Produce]
```



○ Update Existing Field

● Create New Field

Type: Double
Size: 8

1st Differencing

Num Rows: 1

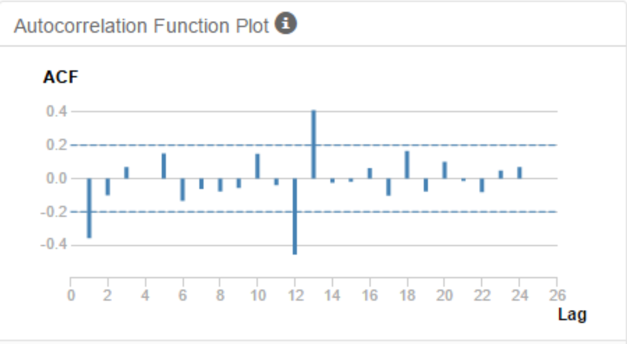
Values for Rows that don't Exist: NULL

Group By (Optional)

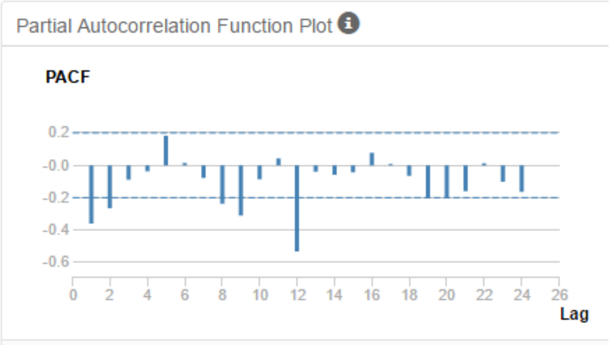
□ Year
□ Month
□ Sum_Produce
□ Seasonal Difference

Expression:

```
[Seasonal Difference]-[Row+1:Seasonal Difference]
```



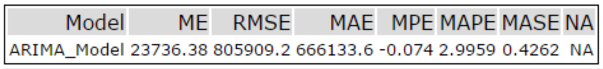
Autocorrelation Function Plot



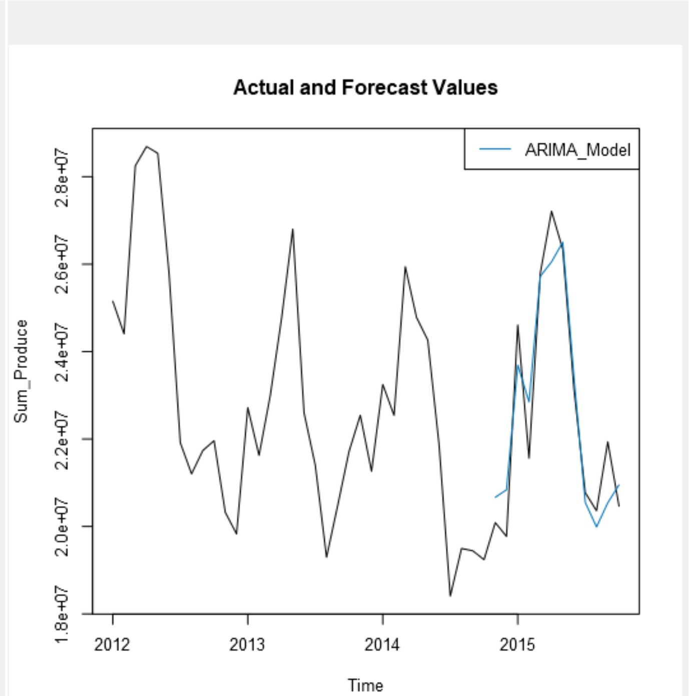
Partial Autocorrelation Function Plot

From above ACF and PACF plots we can see that the data is stationary after the 1st differencing. ACF is negative at lag1, and it's almost cut off to 0, PACF decay to 0, indicates that q=1, p=0, d=0; For seasonal P. D. Q: we can see that ACF is negative at seasonal lag-12, cutoff to 0 at lag-24, indicates P=0, Q=1, also seasonal differencing term D=1.

Set up ARIMA model figures as above and filtering out the most recent 12 months as validation dataset, connect the ARIMA model output with a TS Compare Tool to validate it with validation dataset. We can see the ARIMA model analysis result as below. RMSE figure is a lot of less than the value of ETS RMSE figure, MASE is 0.4 close to 0, and we can also see that the deviance between the predication data and the actual data are much less than the ones shows on ETS model result, so from the analysis ARIMA model is a better one to predict store sales.

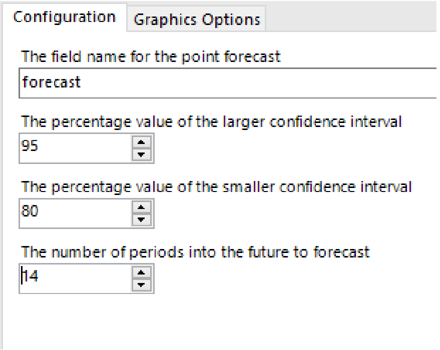


Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_Model	23736.38	805909.2	666133.6	-0.074	2.9959	0.4262	NA



3. Use ARIMA Model to Forecast Existing Store Sales:

Connect above configured ARIMA model directly with “Time Series Data”, then connect its output to a TS forecast tool to forecast 2016 existing store sales, set up its configuration as below:



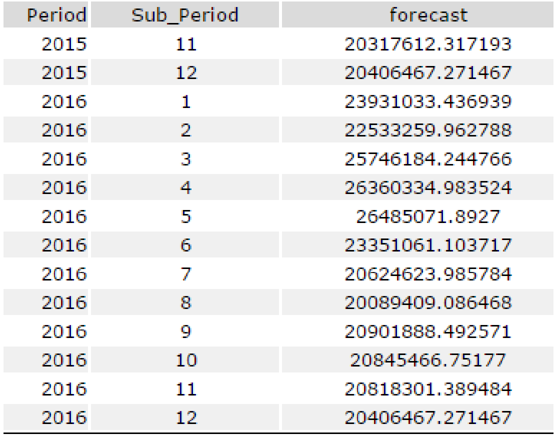
The field name for the point forecast
forecast

The percentage value of the larger confidence interval
95

The percentage value of the smaller confidence interval
80

The number of periods into the future to forecast
14

The forecast 2016 monthly store sales for existing stores as below. Also output the forecast into file “Forecast for Existing Stores” for later use.

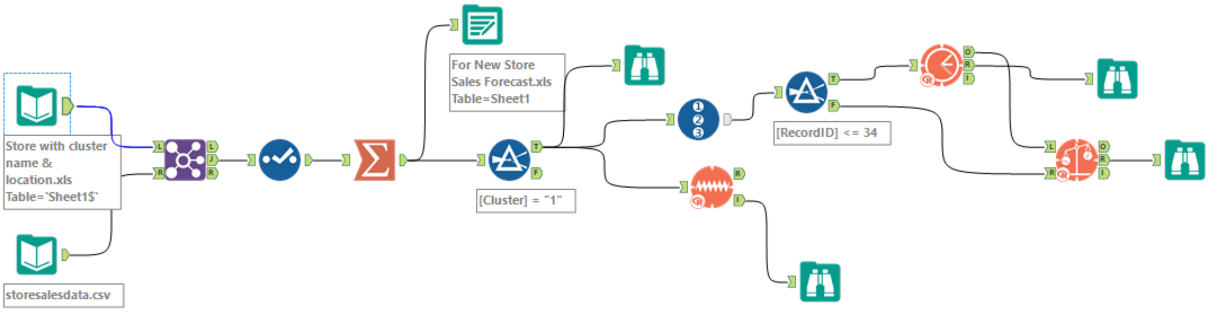


Period	Sub_Period	forecast
2015	11	20317612.317193
2015	12	20406467.271467
2016	1	23931033.436939
2016	2	22533259.962788
2016	3	25746184.244766
2016	4	26360334.983524
2016	5	26485071.8927
2016	6	23351061.103717
2016	7	20624623.985784
2016	8	20089409.086468
2016	9	20901888.492571
2016	10	20845466.75177
2016	11	20818301.389484
2016	12	20406467.271467

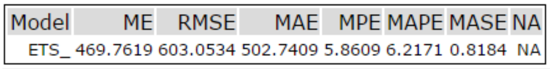
Forecasting Sales for New Stores:

(To forecast the monthly produce product average sales)

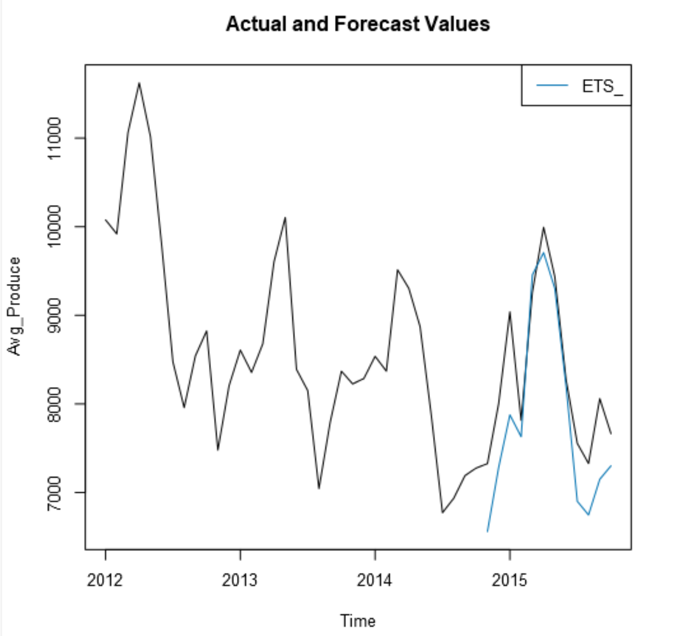
1. ETS Model:


Store with cluster name & location.xls Table=`Sheet1$`
storesalesdata.csv
For New Store Sales Forecast.xls Table=Sheet1
[Cluster] = "1"
[RecordID] <= 34
L
J
R
T
F
O
I
D

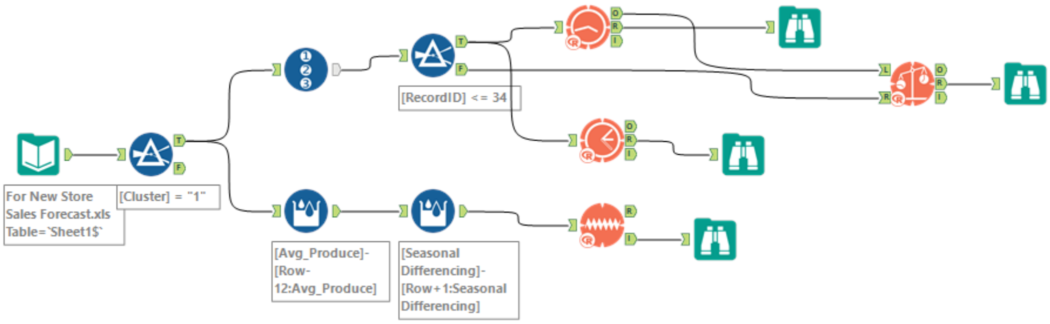
Summarize sales data into average monthly Product sales for each cluster, and apply above ETS and ARIMA model analysis on the summarized data. The TS tool analysis report on cluster 1 data is almost the same as above TS tool analysis report for monthly total product sales data, so I set up ETS model as: multiplicative, auto, addictive, then connect to a ETS model tool, also use TS compare tool to validate the result. Below report shows that doesn't like ETS model for total sales analysis, it's not a very bad model for average product sales prediction, its RMSE is quite low at 603, but MASE is quite large close to 1.



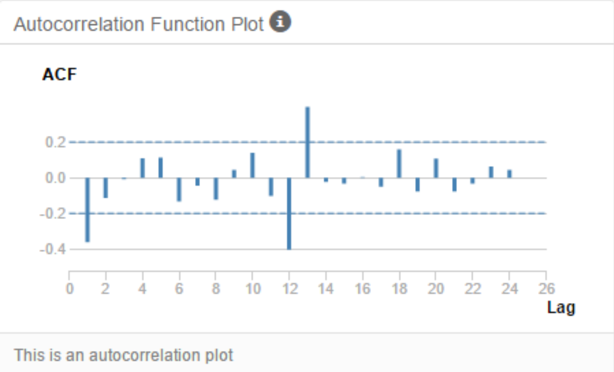
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_	469.7619	603.0534	502.7409	5.8609	6.2171	0.8184	NA



2. ARIMA model analysis:

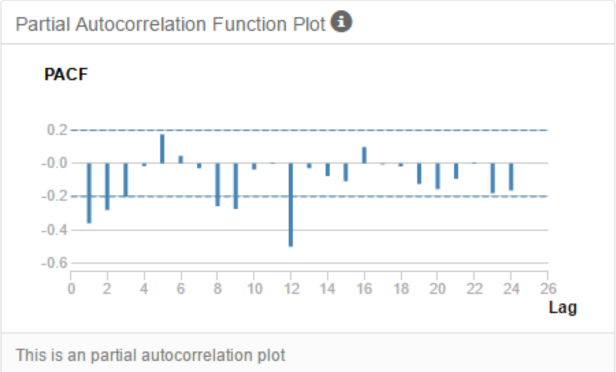

For New Store Sales Forecast.xls Table=`Sheet1$`
[Cluster] = "1"
[RecordID] <= 34
[Avg_Produce]-[Row-12:Avg_Produce]
[Seasonal Differencing]-[Row+1:Seasonal Differencing]

Following above Alteryx workflow, input “For New Store Sales Forecast” data saved from above ETS model analysis process. Connect the data to two multirow formula, and connect to another TS plot tool, according to the TS plot analysis result as below ACF and PACF plots, I set up ARIMA model as p.d.q.(0,0,1) P.D.Q(0,1,1)



Autocorrelation Function Plot

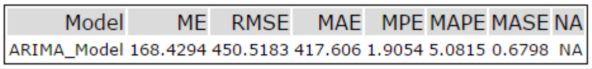
This is an autocorrelation plot



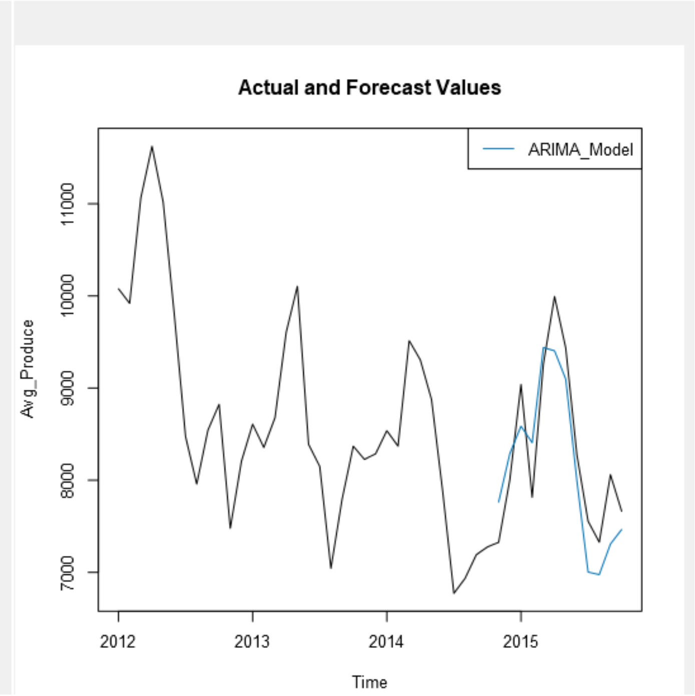
Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

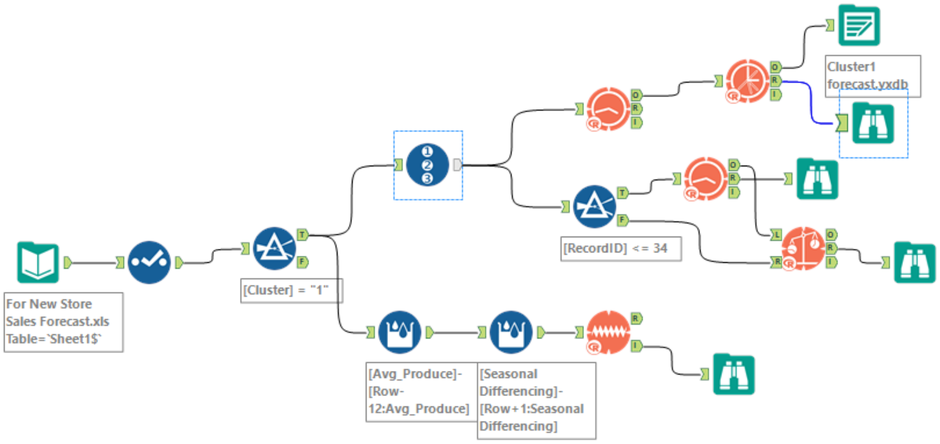
Also use TS compare tool to validate the result, from the ARIMA model analysis report below we can that its RMSE, MAE, MPE, are both lower than ETS model analysis result, its MASE figure is lower than 0.8 (ETS model result) therefore ARIMA model is a better model to forecast each cluster’s new store sales for 2016.



Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_Model	168.4294	450.5183	417.606	1.9054	5.0815	0.6798	NA

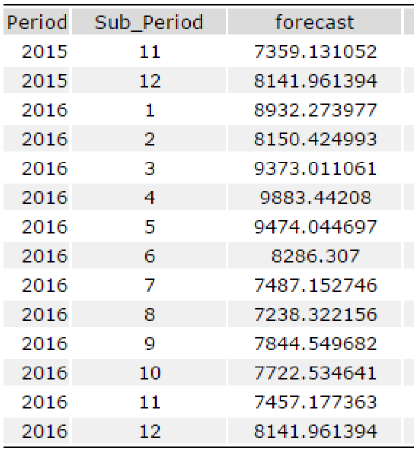


3. To Forecast Average Produce Sales for 3 Clusters:

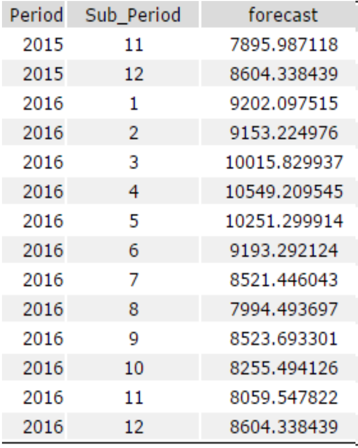

For New Store Sales Forecast.xls Table=`Sheet1$`
[Cluster] = "1"
[RecordID] <= 34
Cluster1 forecast.yxdb
[Avg_Produce]-[Row-12:Avg_Produce]
[Seasonal Differencing]-[Row+1:Seasonal Differencing]

Follow above ARIMA model analysis process, set up cluster filter as cluster 1, 2, 3 separately, the 2016 forecasts for 3 clusters are as below: (cluster 1, 2, 3 from left to the right on below graph)

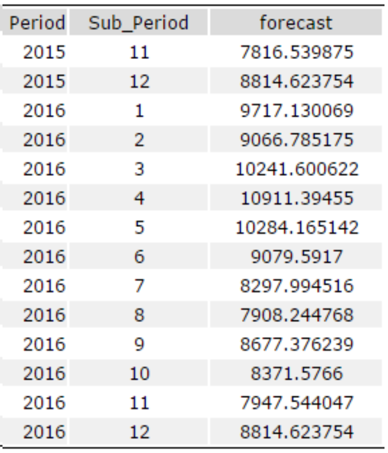
Cluster1: Cluster2: Cluster3:



Period	Sub_Period	forecast
2015	11	7359.131052
2015	12	8141.961394
2016	1	8932.273977
2016	2	8150.424993
2016	3	9373.011061
2016	4	9883.44208
2016	5	9474.044697
2016	6	8286.307
2016	7	7487.152746
2016	8	7238.322156
2016	9	7844.549682
2016	10	7722.534641
2016	11	7457.177363
2016	12	8141.961394

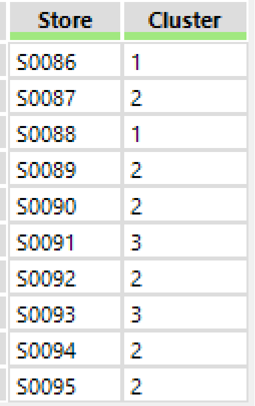


Period	Sub_Period	forecast
2015	11	7895.987118
2015	12	8604.338439
2016	1	9202.097515
2016	2	9153.224976
2016	3	10015.829937
2016	4	10549.209545
2016	5	10251.299914
2016	6	9193.292124
2016	7	8521.446043
2016	8	7994.493697
2016	9	8523.693301
2016	10	8255.494126
2016	11	8059.547822
2016	12	8604.338439



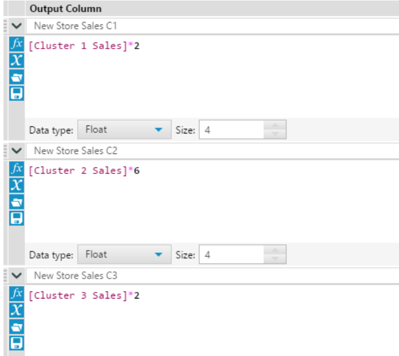
Period	Sub_Period	forecast
2015	11	7816.539875
2015	12	8814.623754
2016	1	9717.130069
2016	2	9066.785175
2016	3	10241.600622
2016	4	10911.39455
2016	5	10284.165142
2016	6	9079.5917
2016	7	8297.994516
2016	8	7908.244768
2016	9	8677.376239
2016	10	8371.5766
2016	11	7947.544047
2016	12	8814.623754

4. Use Each Cluster's Average Produce Sales to Calculate New Store Total Sales:



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

From earlier new store classification result showed as above we know that we predicate that in the 10 new stores, there will be 2 stores in cluster 1, 6 stores in cluster 2, 2 stores in cluster 3, therefore set up 2 formula tools as below, sum the sales for each cluster together.



Output Column

New Store Sales C1

```
[Cluster 1 Sales]*2
```

Data type: Float Size: 4

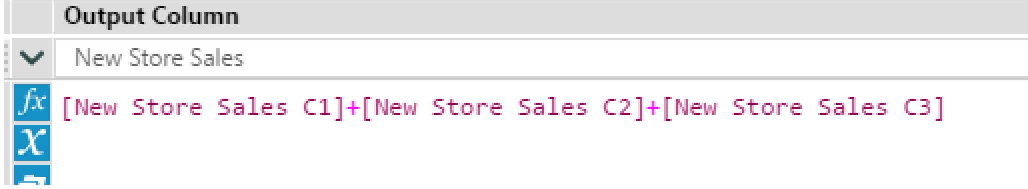
New Store Sales C2

```
[Cluster 2 Sales]*6
```

Data type: Float Size: 4

New Store Sales C3

```
[Cluster 3 Sales]*2
```

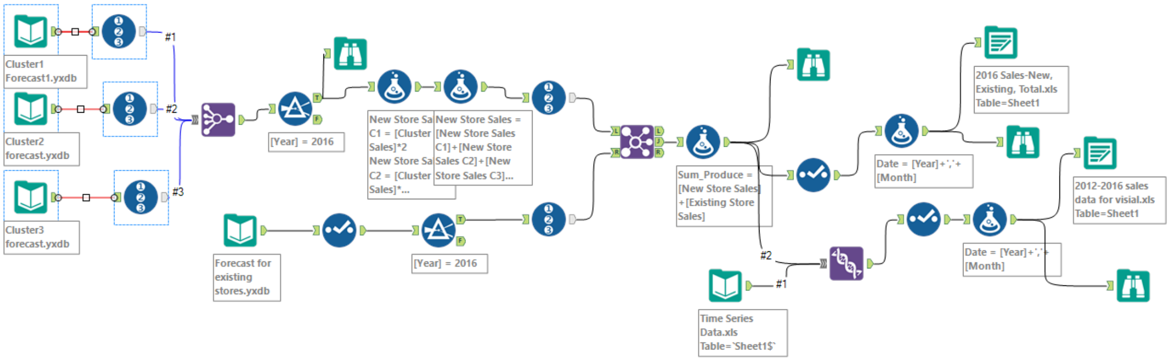


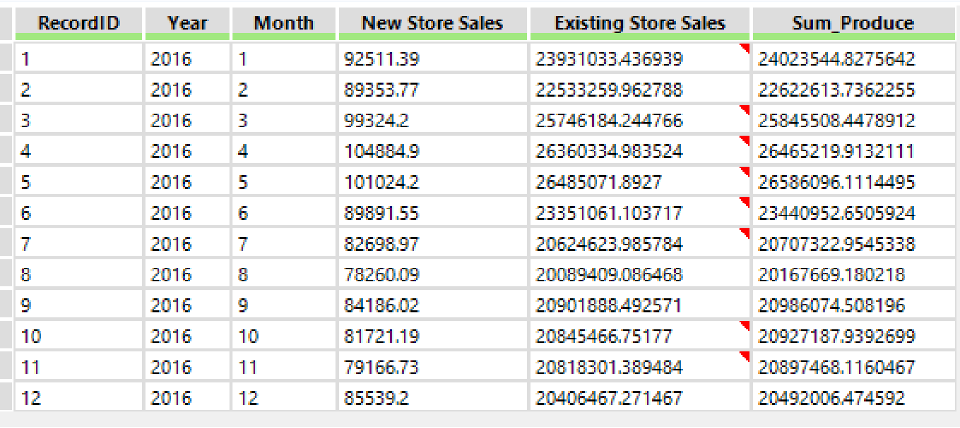
Output Column

New Store Sales

```
[New Store Sales C1]+[New Store Sales C2]+[New Store Sales C3]
```

Following below Alteryx workflow to combine existing store sales with new store sales, so the overall forecast for all stores is as below data table:


Cluster1 Forecast1.yxdb
#1
Cluster2 forecast.yxdb
#2
Cluster3 forecast.yxdb
#3
[Year] = 2016
New Store Sales C1 = [Cluster Sales]*2 New Store Sales C2 = [Cluster Sales]*...
New Store Sales = [New Store Sales C1]+[New Store Sales C2]+[New Store Sales C3]...
Forecast for existing stores.yxdb
[Year] = 2016
Sum_Produce = [New Store Sales] +[Existing Store Sales]
#2
#1
Time Series Data.xls Table='Sheet1$'
Date = [Year]+'-'+ [Month]
2016 Sales-New, Existing, Total.xls Table=Sheet1
Date = [Year]+'-'+ [Month]
2012-2016 sales data for visial.xls Table=Sheet1



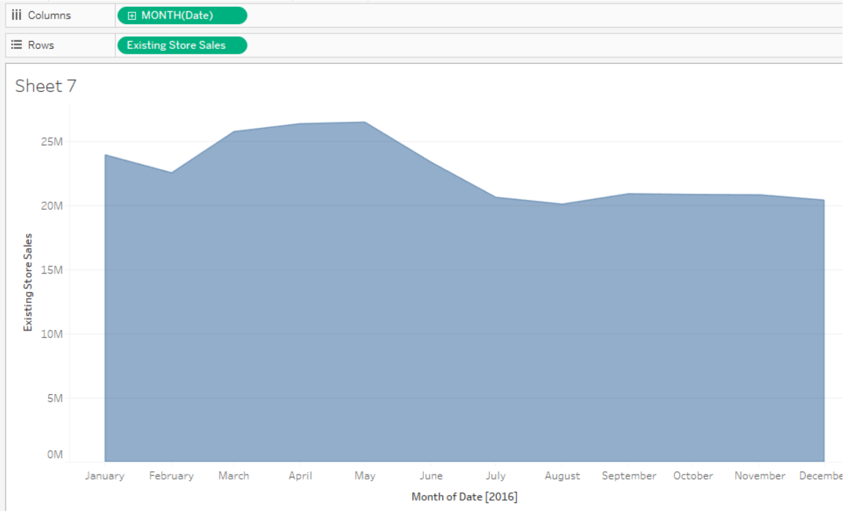
RecordID	Year	Month	New Store Sales	Existing Store Sales	Sum_Produce
1	2016	1	92511.39	23931033.436939	24023544.8275642
2	2016	2	89353.77	22533259.962788	22622613.7362255
3	2016	3	99324.2	25746184.244766	25845508.4478912
4	2016	4	104884.9	26360334.983524	26465219.9132111
5	2016	5	101024.2	26485071.8927	26586096.1114495
6	2016	6	89891.55	23351061.103717	23440952.6505924
7	2016	7	82698.97	20624623.985784	20707322.9545338
8	2016	8	78260.09	20089409.086468	20167669.180218
9	2016	9	84186.02	20901888.492571	20986074.508196
10	2016	10	81721.19	20845466.75177	20927187.9392699
11	2016	11	79166.73	20818301.389484	20897468.1160467
12	2016	12	85539.2	20406467.271467	20492006.474592

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

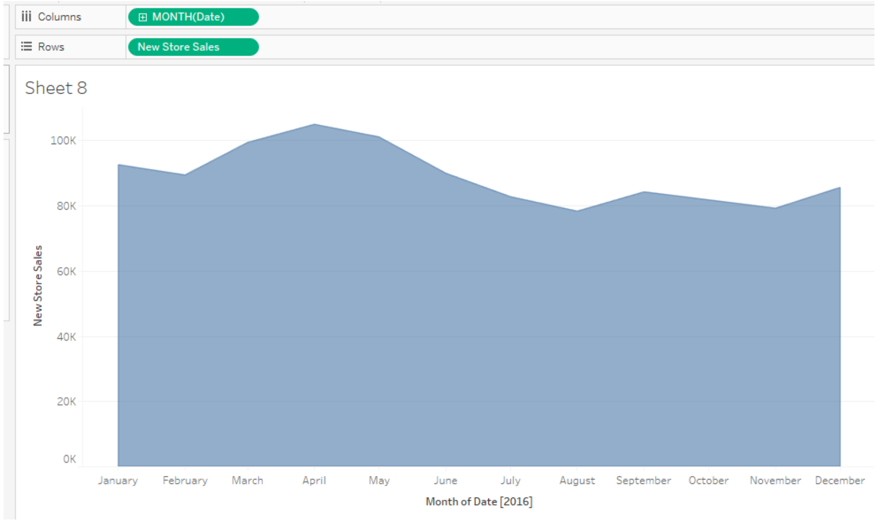
Following above Alteryx workflow, combing new store 2016 sales forecast with existing store 2016 sales forecast, output data as “2016 Store Sales-New,Exisitng,Total.xls”, also combine 2016 sales data with 2012-2015 sales data, output data as “2012-2016 sales data for visial.xls”. Input those two files into Tableau, plot sales volume against date, the visualization are as below:

(Tableau Public file location: https://public.tableau.com/profile/angela.huang#!/)

2016 existing store sales forecast plot:



2016 new store sales forecast plot:

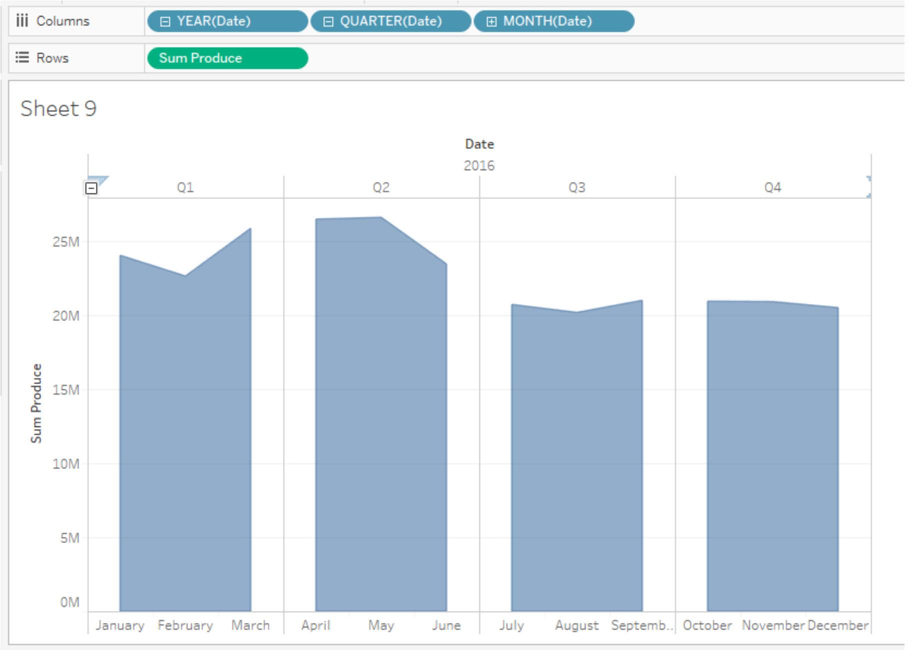


Columns: MONTH(Date)

Rows: New Store Sales

Sheet 8

2016 total store sales forecast plot:

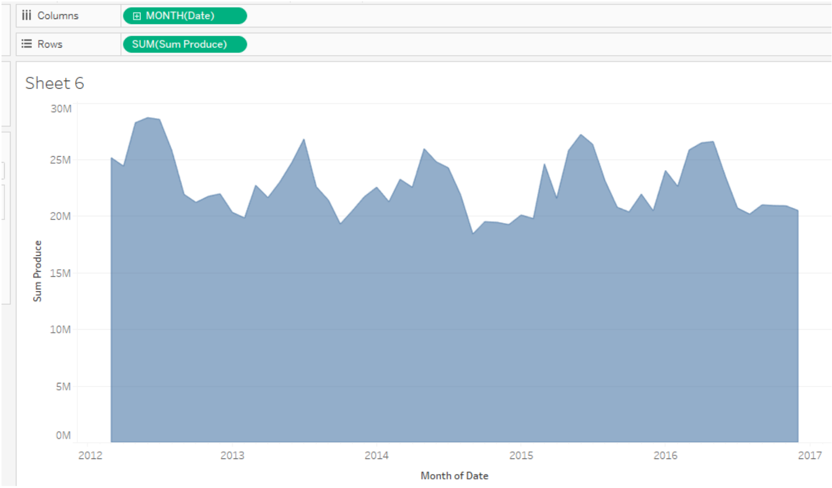


Columns: YEAR(Date) QUARTER(Date) MONTH(Date)

Rows: Sum Produce

Sheet 9

2012-2016 store sales forecast plot:



Columns MONTH(Date)

Rows SUM(Sum Produce)

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.