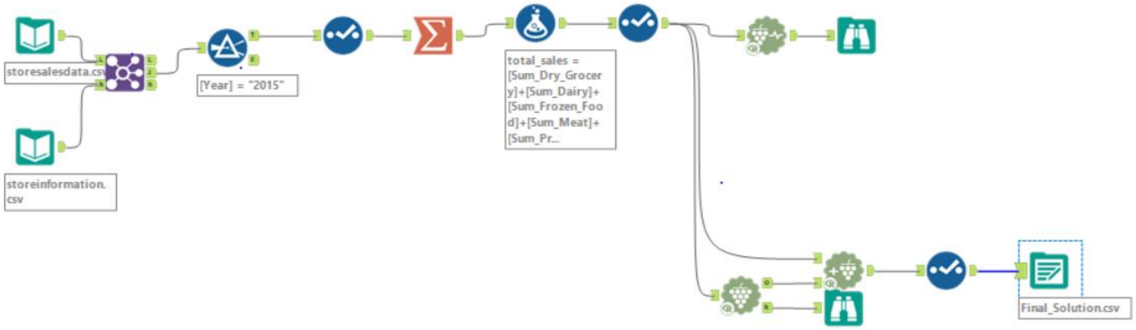
Project: Predictive Analytics Capstone

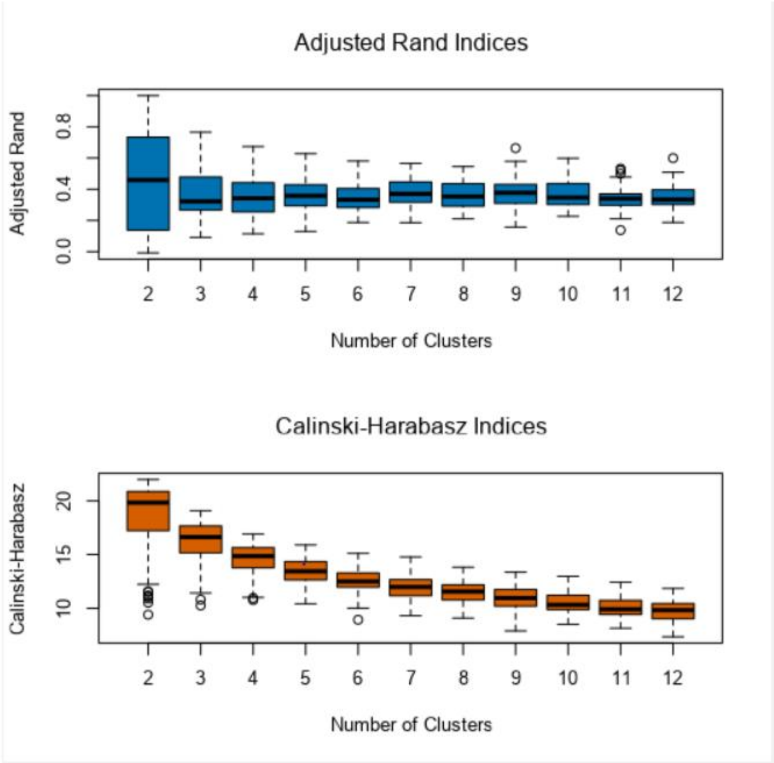
Task 1: Determine Store Formats for Existing Stores

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

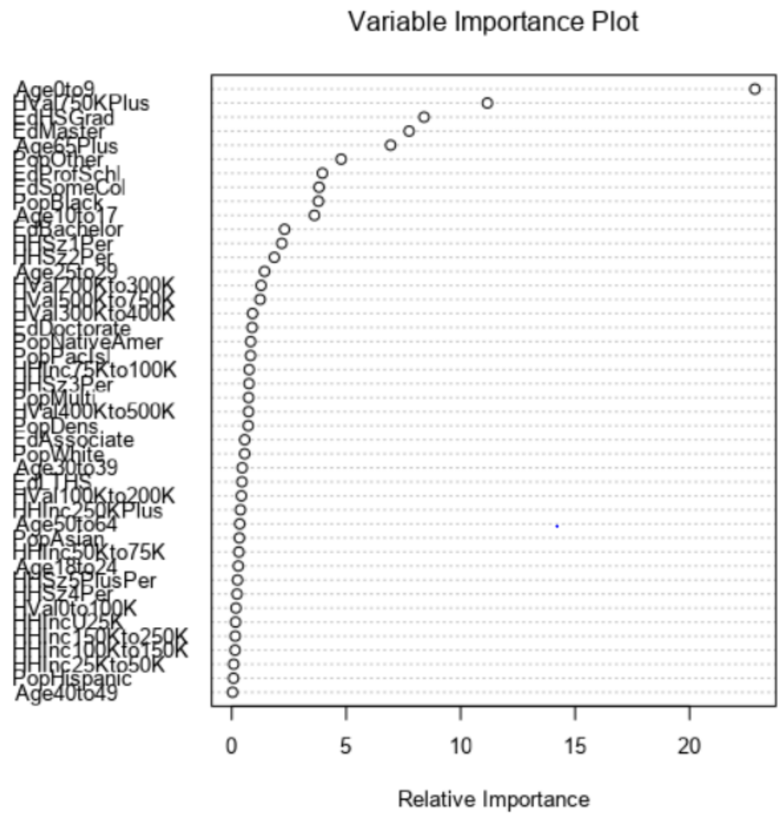
   The optimal number of store formats were 3. Using the K means (Neural Gas) method we obtained the clusters for the stores. Finally I used append cluster to store the cluster results in a file “Final_Solution.csv”


storesalesdata.csv
storeinformation.csv
L
J
R
T
F
[Year] = "2015"
total_sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
Final_Solution.csv

From the above alteryx workflow, we get 3 final clusters.

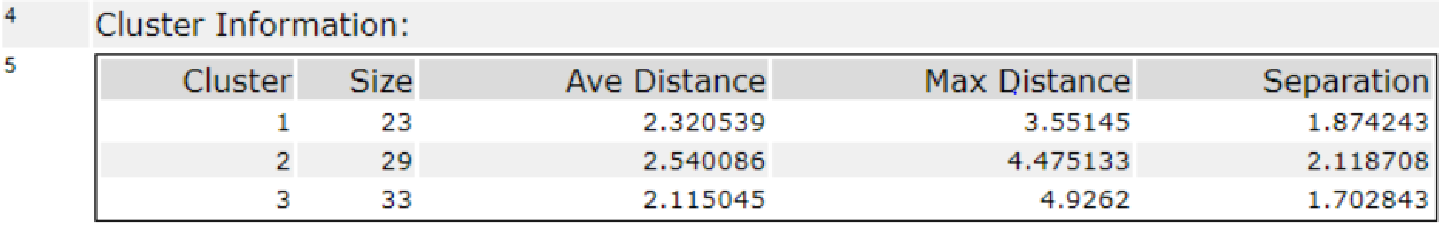


Using the Adjusted Rand Index and the CH index, we attain a cluster of 3 and the number of starting seeds at 10.



The most important variable is Age 0to9.

2. How many stores fall into each store format?



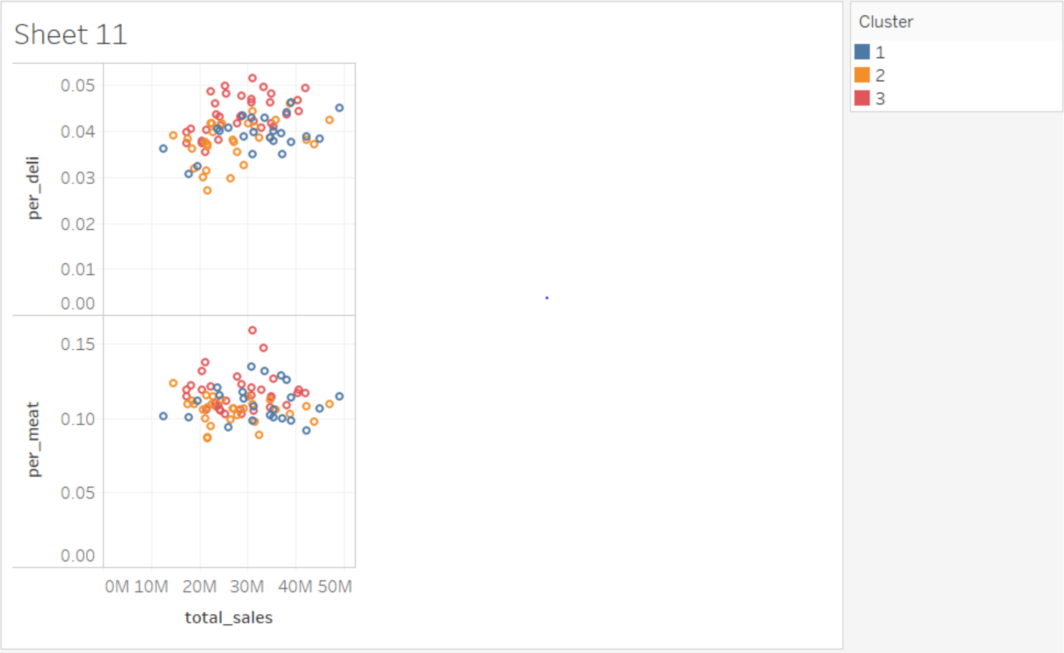
4 Cluster Information:

5

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475133	2.118708
3	33	2.115045	4.9262	1.702843

From the above table we can see that all clusters have a size of more than 20 (the required number). It also displays the intra cluster and intercluster distances.
Hence, we can proceed for validating the clusters.

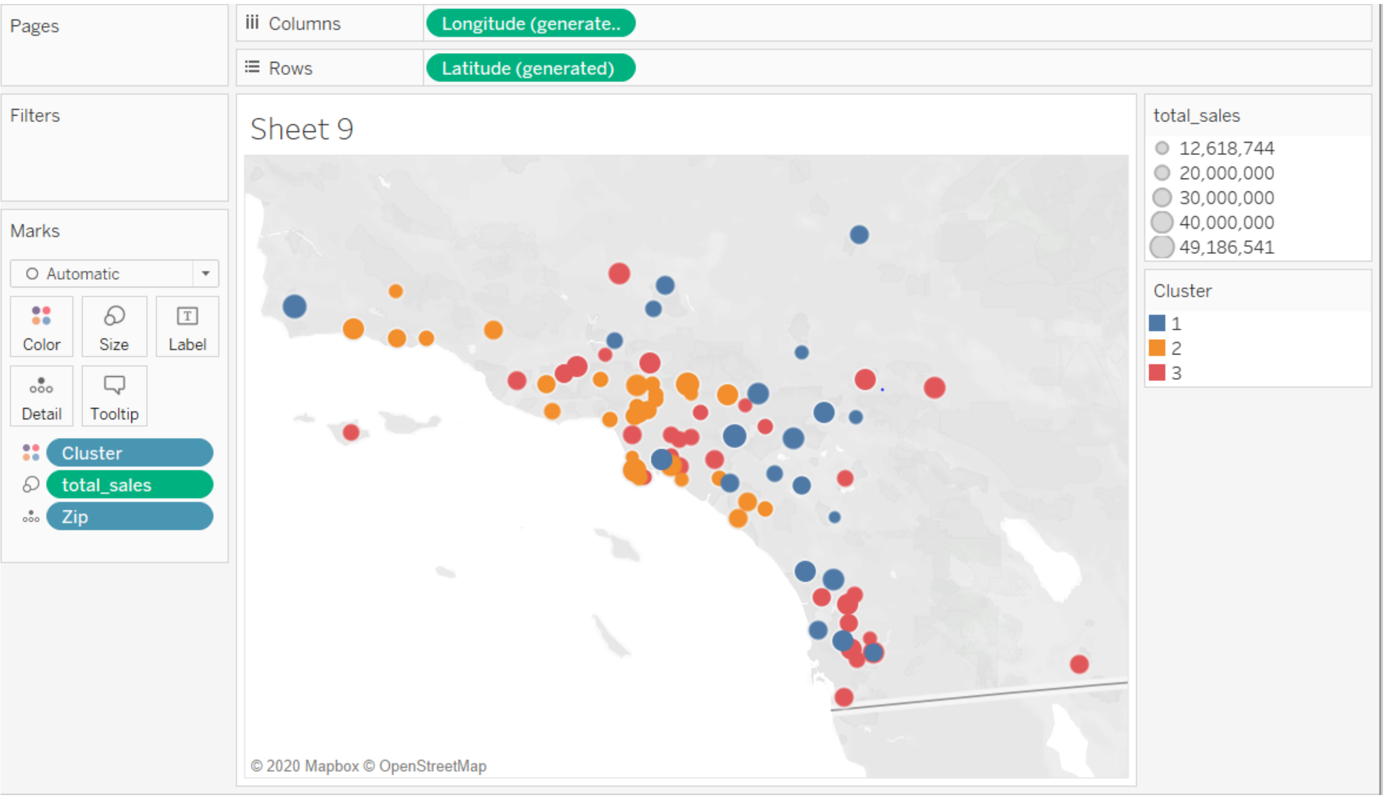
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?



From the above scatter plot, we notice that the cluster 3(red) have higher sales in deli and meat section as compared cluster 1 and 2. Hence, we can say that for stores in cluster 2 the sale of deli sandwiches and meat can is high.

Also, from the above table cluster 1 has the smallest size and cluster 3 has the largest cluster size. The intracluster range is highest for cluster 2.

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.


Pages

Columns: Longitude (generate..

Rows: Latitude (generated)

Filters

Marks

Automatic

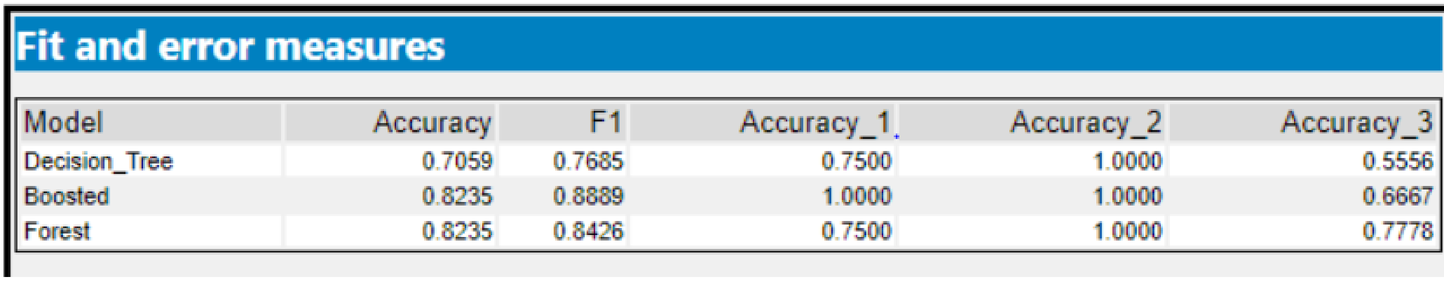
Cluster

total_sales

Zip

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 used a boosted model for the classification of clusters for the new stores.



Fit and error measures

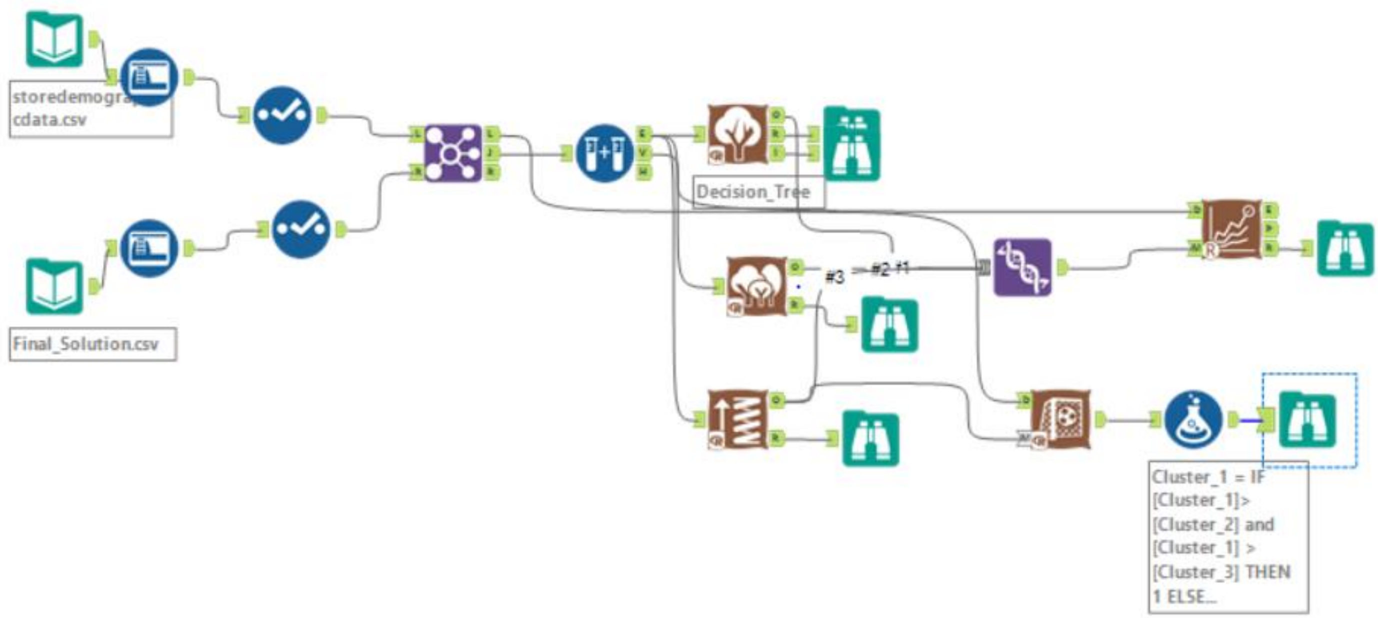
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Forest	0.8235	0.8426	0.7500	1.0000	0.7778

It is noted that the accuracy of boosted and forest model are the same(82.3). But the F1 score of the Boosted Model is higher. Hence, after using the model comparison tool to determine the best model, I used the score tool to determine the cluster segments for the new stores.

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

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

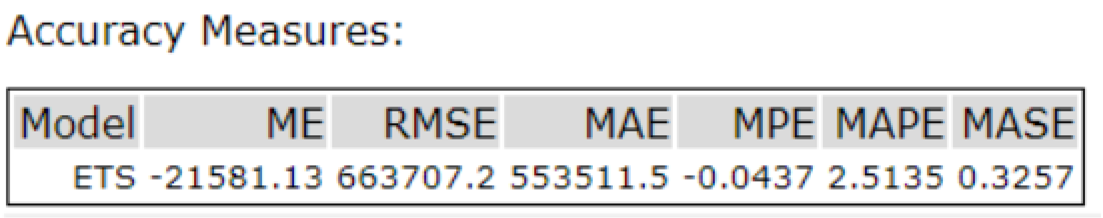
Workflow



Task 3: Predicting Produce Sales

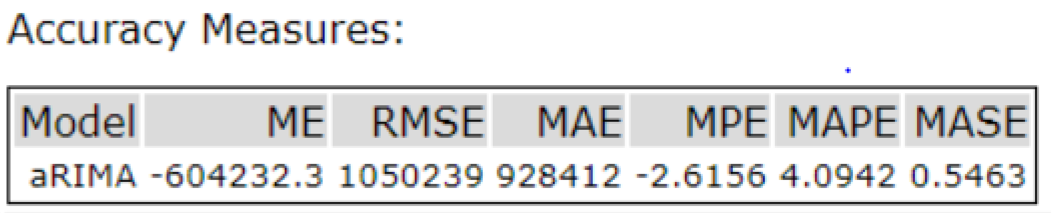
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

For this forecast, I used the ETS model. Using the TS compare tool, From the decomposition plot we observe the seasonal,trend and error components. There is no trend, seasonal is multiplicative and error is multiplicative. ETS also has the best accuracy measures, including lowest MASE,MAPE AND RMSE. Therefore, we not use the ETS model to forecase sales for existing and new stores



Accuracy Measures:

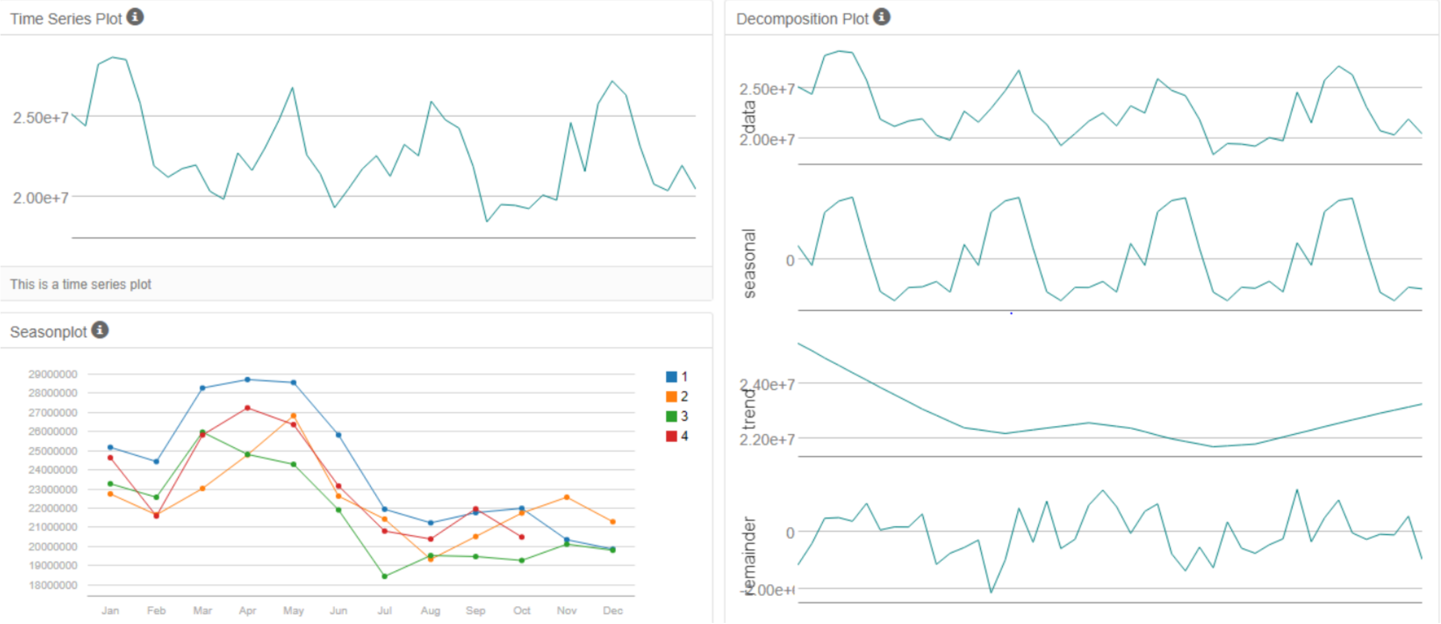
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257



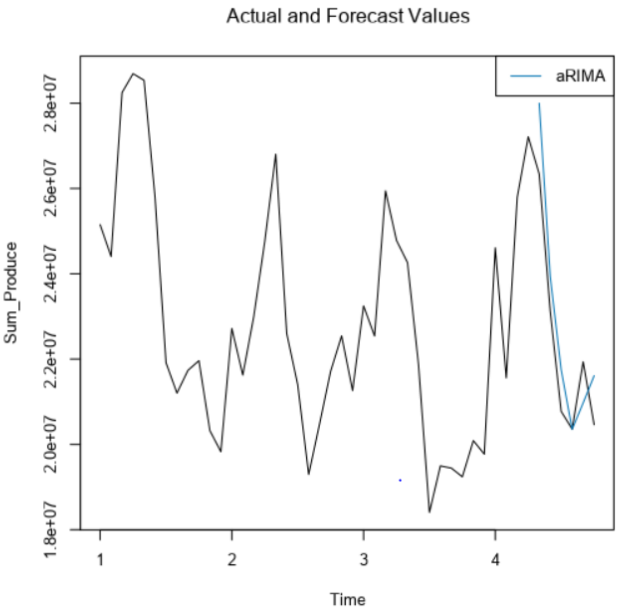
Accuracy Measures:

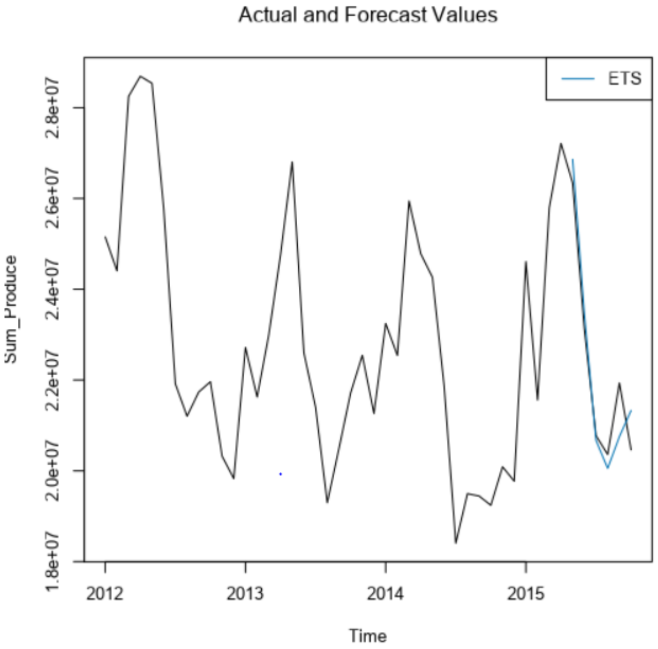
Model	ME	RMSE	MAE	MPE	MAPE	MASE
aRIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

TS COMPARE PLOT



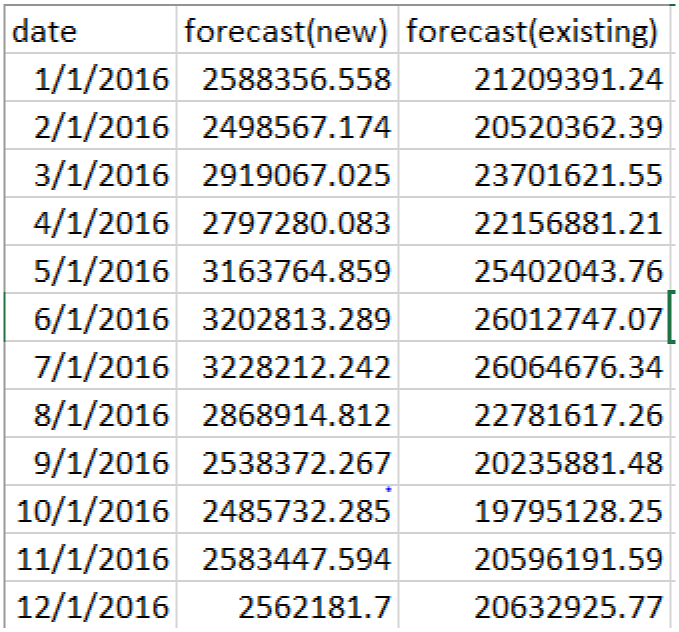
This is a time series plot





It is noted that the blue line(forecasted values) for ETS is more closer to the black line(actual values)

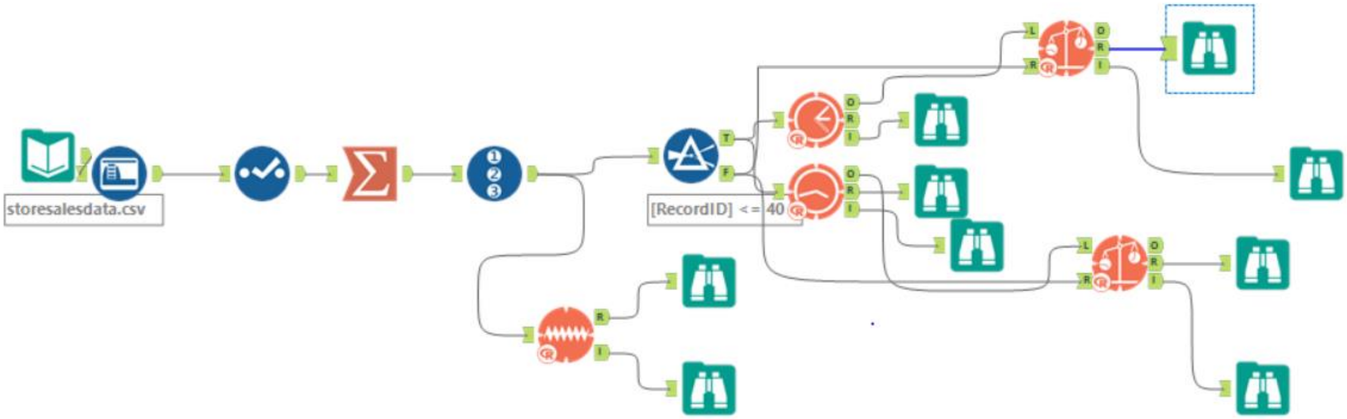
3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.



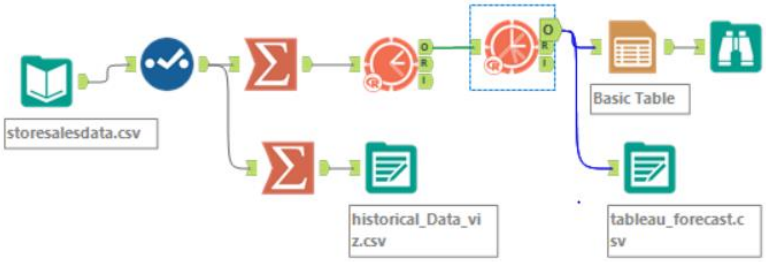
date	forecast(new)	forecast(existing)
1/1/2016	2588356.558	21209391.24
2/1/2016	2498567.174	20520362.39
3/1/2016	2919067.025	23701621.55
4/1/2016	2797280.083	22156881.21
5/1/2016	3163764.859	25402043.76
6/1/2016	3202813.289	26012747.07
7/1/2016	3228212.242	26064676.34
8/1/2016	2868914.812	22781617.26
9/1/2016	2538372.267	20235881.48
10/1/2016	2485732.285	19795128.25
11/1/2016	2583447.594	20596191.59
12/1/2016	2562181.7	20632925.77

Using the below workflows, we attained the forecasted values for the next 12 months in the year 2016 for the new and existing stores.

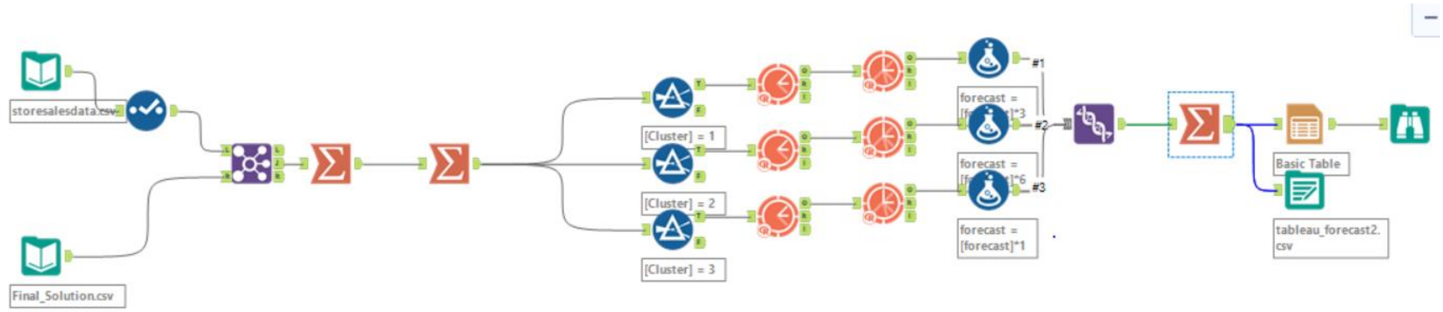
Model Comparison



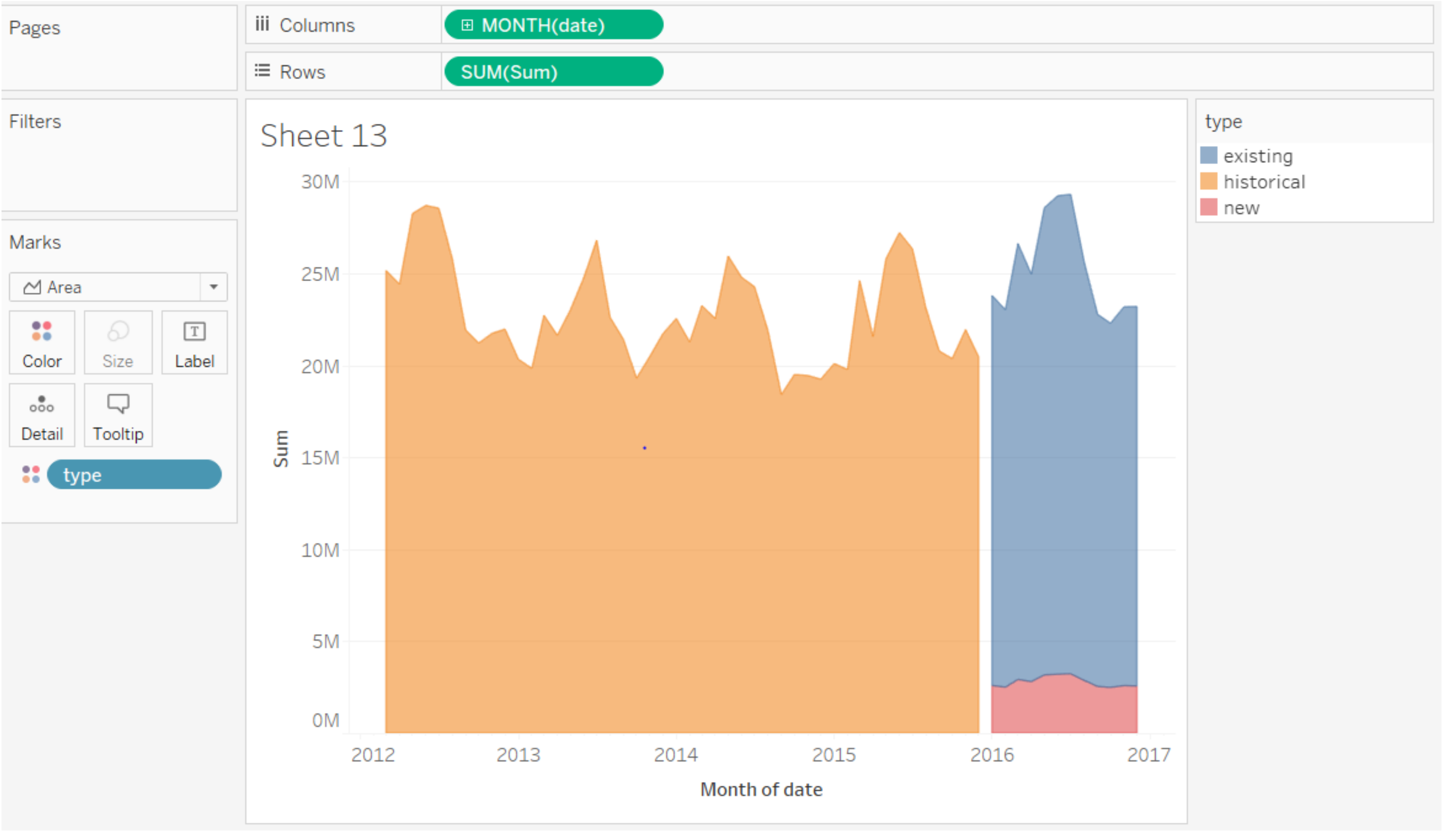
Forecast for existing stores



Forecast for new stores



Visualization


Columns: MONTH(date)

Rows: SUM(Sum)

Marks: Area

Color: type

After arranging the data from the tables for the new and existing stores we visualize the data for sales forecast according to month. Here the color in the area chart represents the historical data(2012-2015) and forecast data for the existing and new stores.

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.