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?
   Optimal number of store format = 3.
   Firstly, I filtered sales data only for 2015 then I grouped sum sales each category by store and got total sales per store after that I calculated percentage of sales for each food category.
   Secondly, I used K Centroid clustering analysis tool and use a method of K- means clustering method and z- score to standardize the fields
   I arrived at result of choosing 3 clusters on the basis of result of Adjusted rand and Calinski Harabasz indices plot .
   In adjusted Rand indices Graph 3 clusters has highest Median and well distributed spread across Y axis that shows the highest index which seems good fit to select no. of clusters among other values.
   Also, in Calinski Harbasz indices plot 3 cluster segment has highest median, highest index value and well distributed spread across y axis among other clusters.
   According to both indices 3 clusters would be good.
2. How many stores fall into each store format?
   Format 1= 25 stores
   Format 2 = 35 stores
   Format 3 =25 Stores
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   In terms of Rand indices every cluster is differently distributed.
   In cluster 1 sales of Grocery, meat and bakery sales are highest as compare to other two clusters.
   In terms of average distance, Maximum distance and Separation value every cluster is different that is shown below I the picture:

Summary Report of the K-Means Clustering Solution cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Sum_Dry_Grocery + Sum_Dairy + Sum_Frozen_Food + Sum_Meat + Sum_Produce + Sum_Floral + Sum_Deli + Sum_Bakery + Sum_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

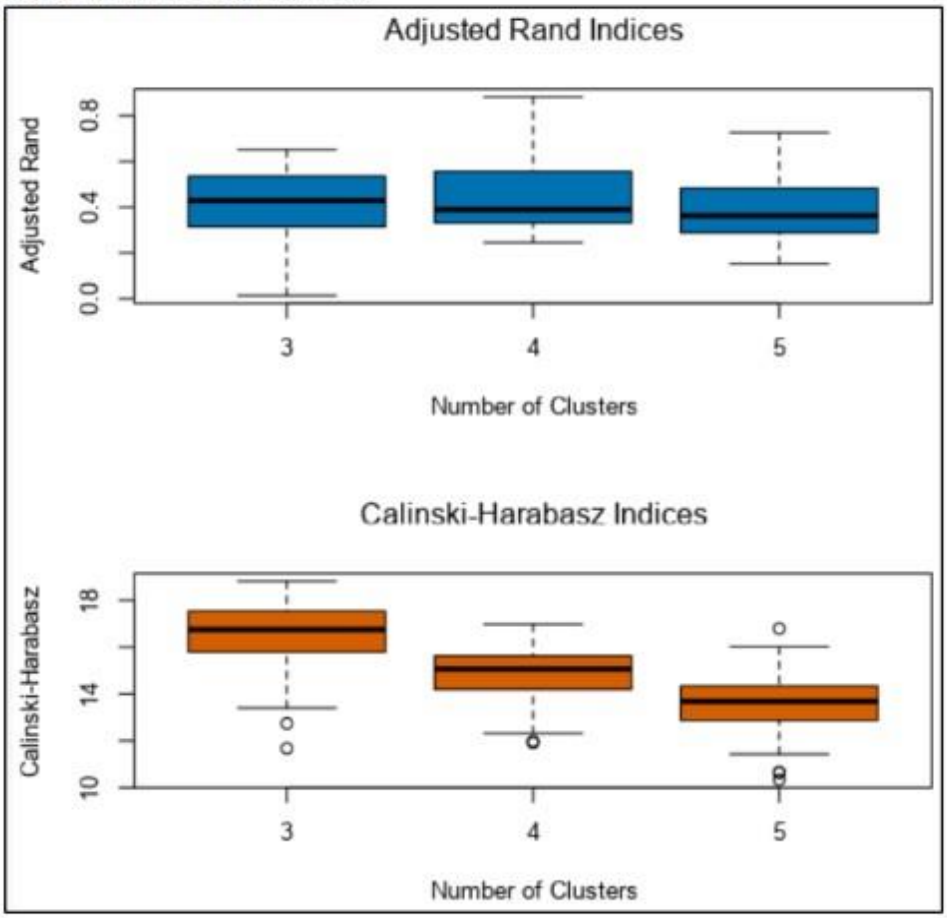
Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Produce	Sum_Floral	Sum_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

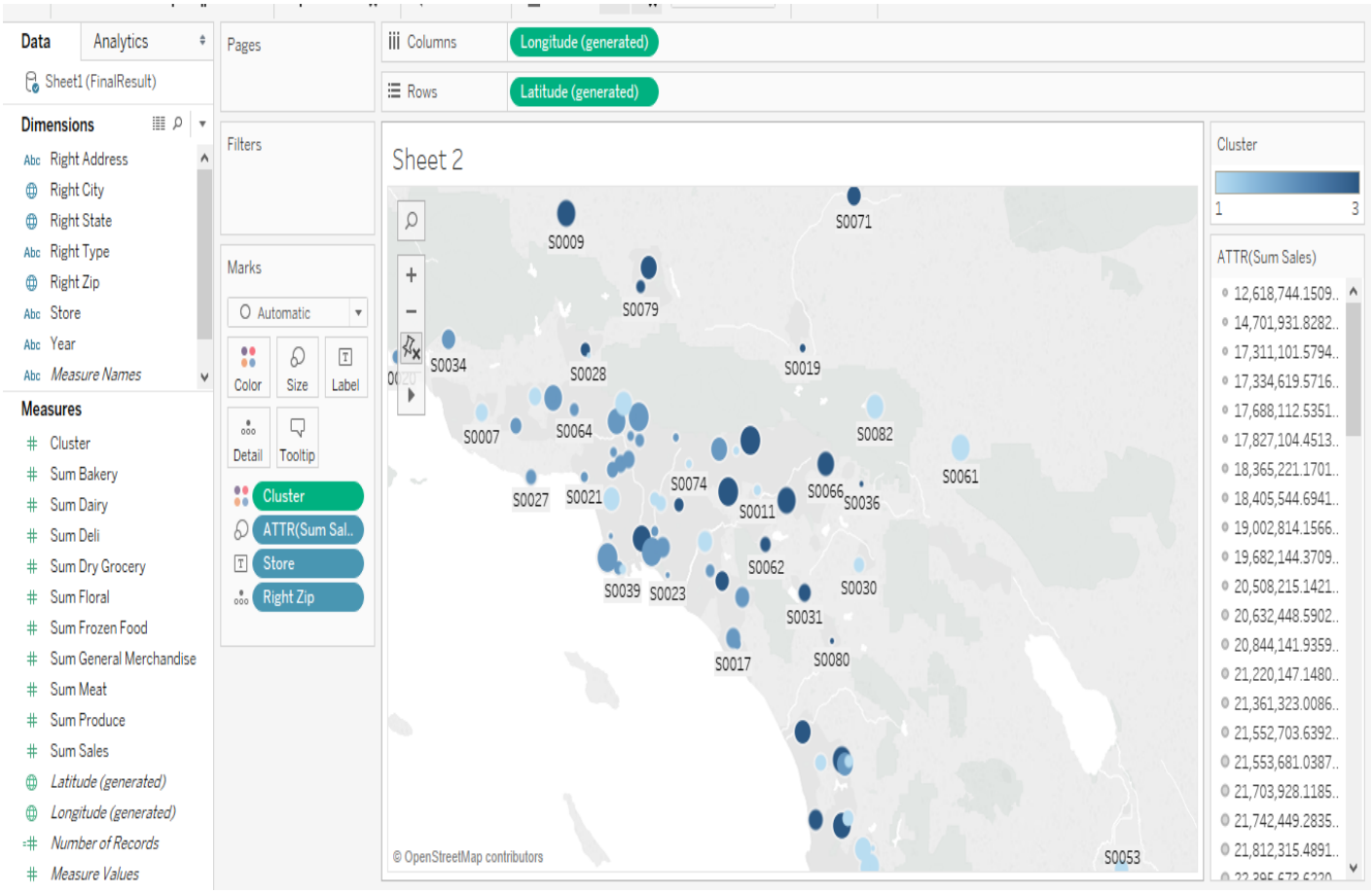
	Sum_Bakery	Sum_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Plots



6. 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

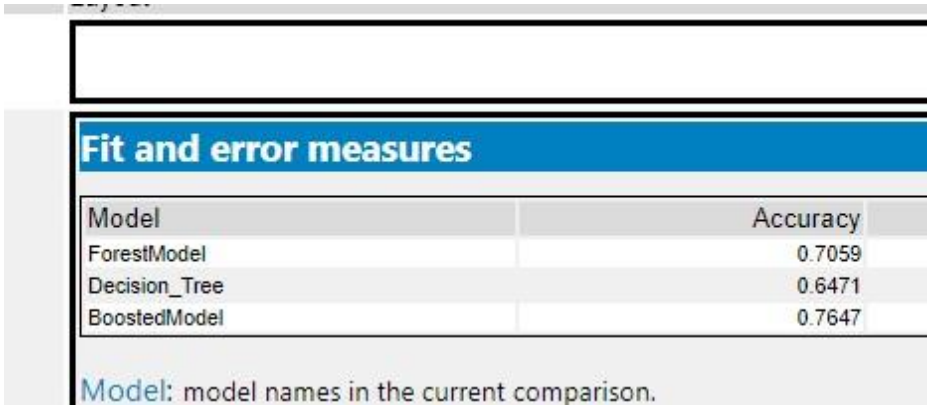
.



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 build three classification models Decision Tree, Forest Model, and Boosted model and decided to choose Boosted Model on comparing accuracy rates of all the three model. Boosted model has high accuracy rate of 76 % that is highest among others.



Fit and error measures

Model	Accuracy
ForestModel	0.7059
Decision_Tree	0.6471
BoostedModel	0.7647

Model: model names in the current comparison.

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	1
S0089	2
S0090	2
S0091	1
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?
   I decided to choose ETS(M,N,M) model on the basis of output of TS tool which shows the Decomposition Plot in which Remainder Shows the inconsistent movements between high and lows that is multiplicative , There is no Trend showing on Trend plot so It's null in this case and then comes the seasonality that's decreasing gradually and that's why multiplicative in nature.

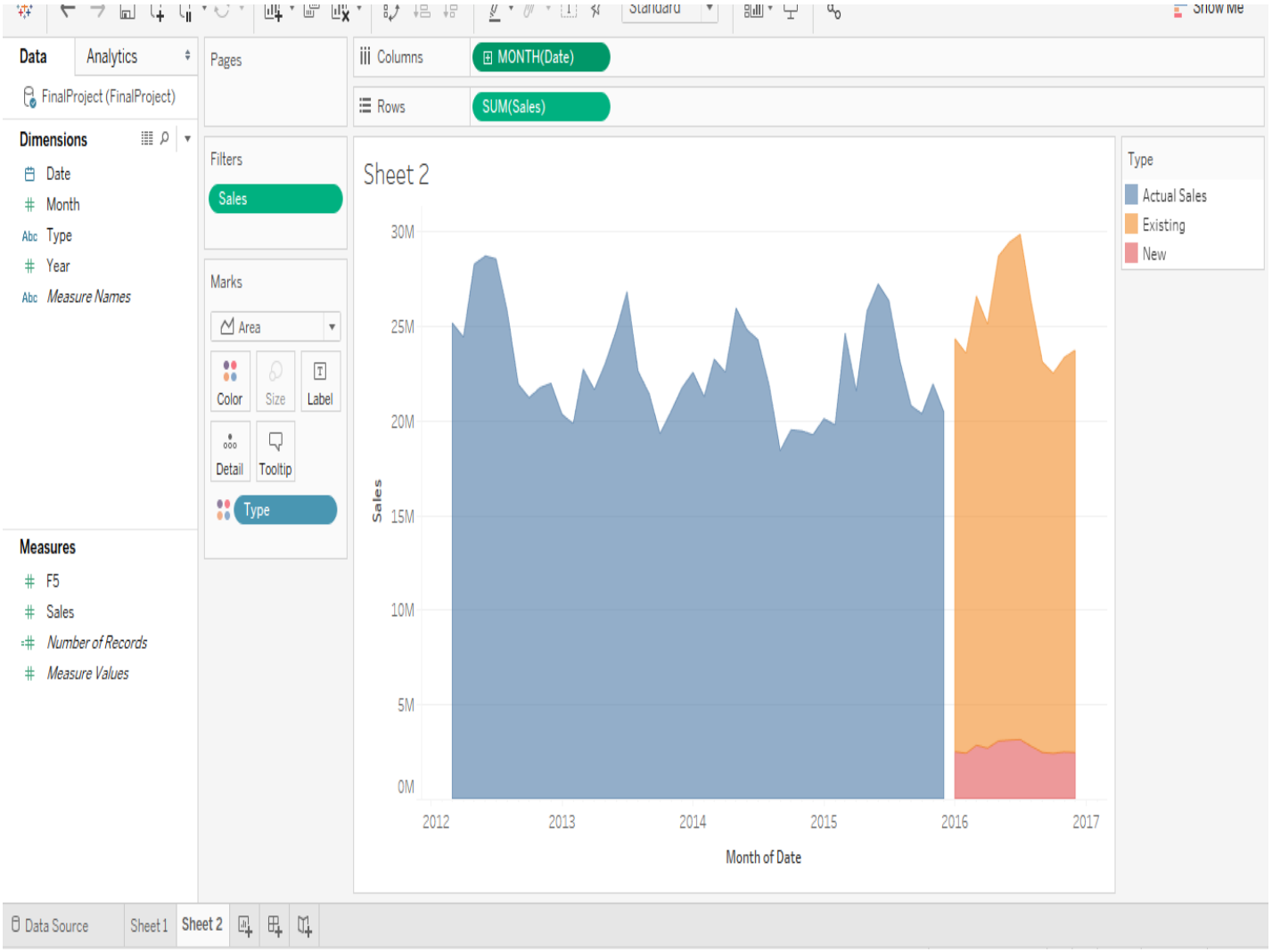
I decided to choose ETS model after comparing Both ARIMA and ETS model by Using TS compare Tool and ETS model forecast result was more closure to actual value.

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.

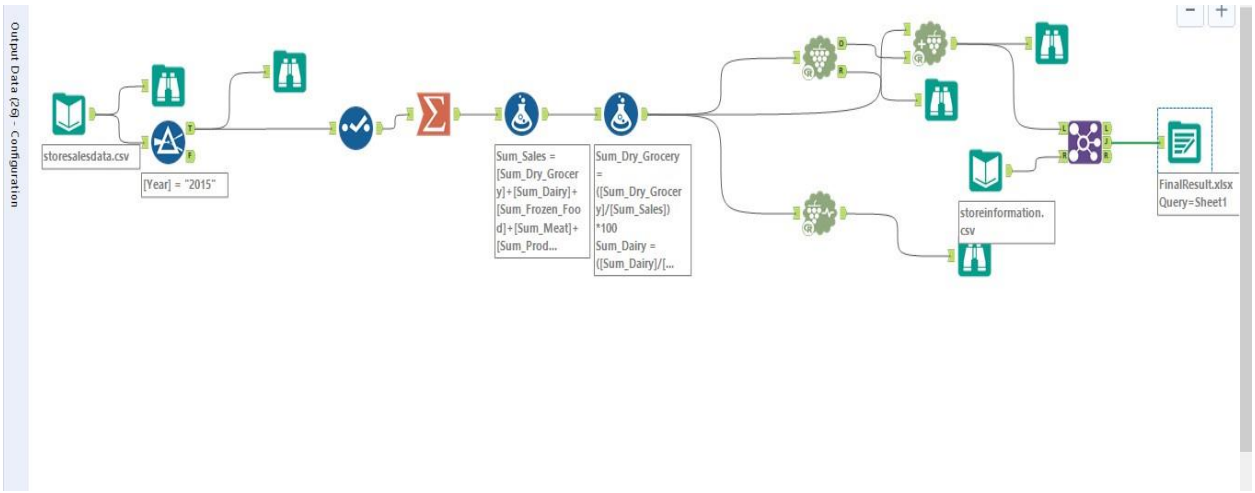
Table for Existing and new Store Forecasts Sales:

Month	New stores	Existing Stores
Jan- 2016	2491319	21829060
Feb- 2016	2408385	21146330
Mar- 2016	2833157	23735687
April- 2016	2679433	22409515
May- 2016	3054886	25621829
June- 2016	3106152	26307858
July-2016	3132699	26705093
Aug-2016	2776154	23440761
Sept- 2016	2451566	20640047
Oct- 2016	2401772	20086270
Nov-2016	2477302	20858120
Dec-2016	2452170	21255190

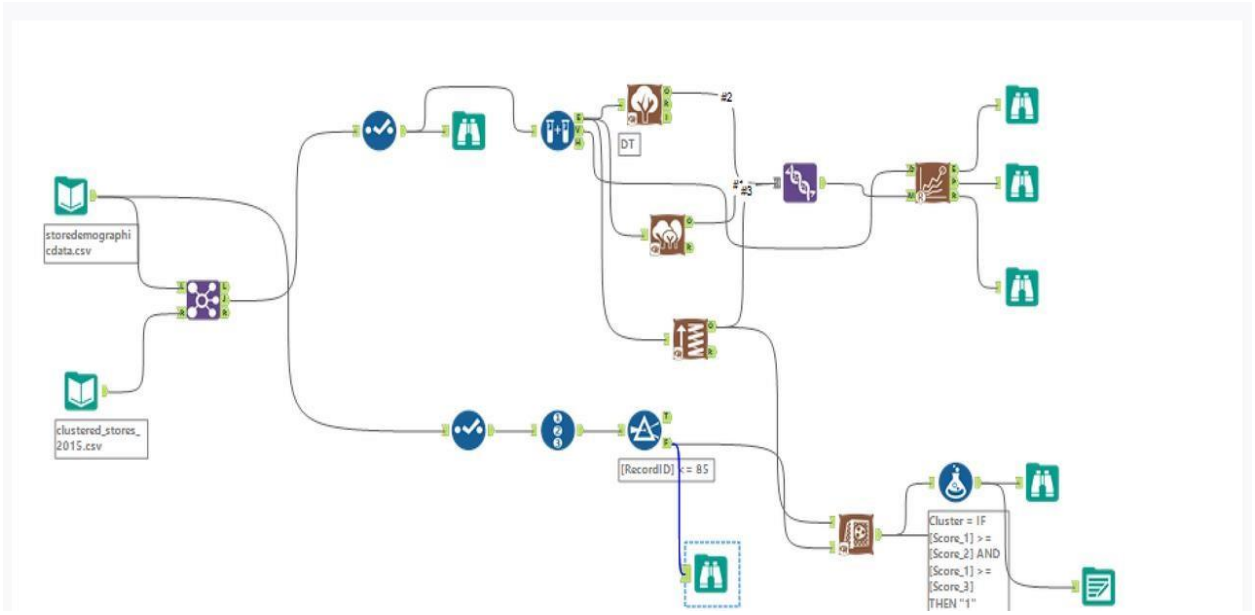
Tableau visualization for historical data vs .Forecast data:



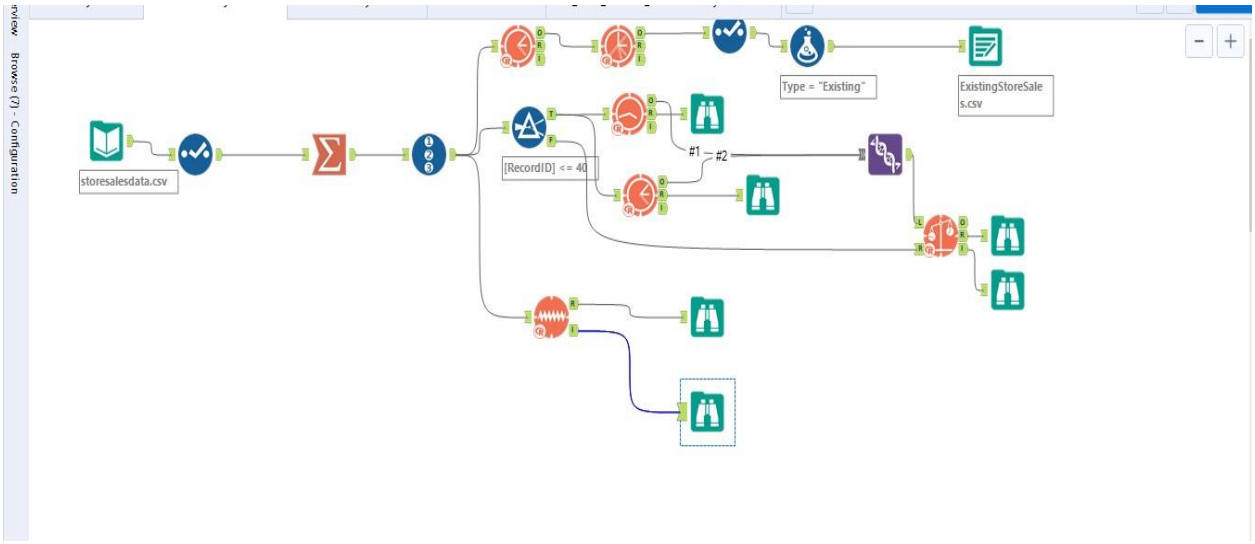
Workflow for TASK 1:



Workflow for Task 2:



Workflow for task 3 Existing stores:



Workflow for task 3 New stores: