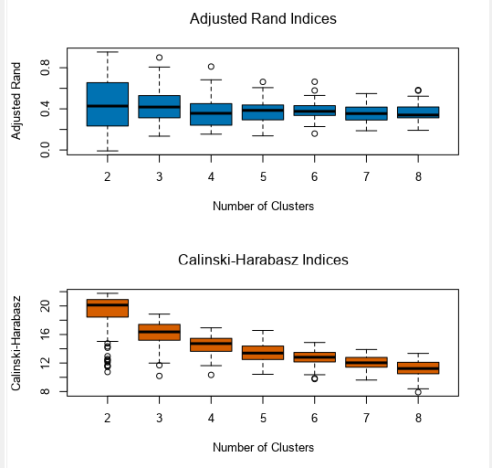
Project: Predictive Analytics Capstone

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?
   The optimal number of the store format is 3. I arrived to this conclusion via the following steps:
   1. To determine the optimal number of stores format I used the file that contained the sales data,
   2. Used only 2015 sales data,
   3. I used percentage sales per category per store for clustering diagnostics (category sales as a percentage of total store sales); The number of clusters that I chose is 3 because it shows the highest value in the Adjusted Rand Index, and although it shows the second highest median in the Calinski-Harabasz Index, the number of outliers in the first one are many,



   4. Applied K-Means clustering model.

2. How many stores fall into each store format?
   The number of stores that fall in each store format are:
   Cluster 1 = 25; Cluster 2 = 35, Cluster 3 = 25

Summary Report of the K-Means Clustering Solution Clustering

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Percent_Dry_Grocery + Percent_Dairy + Percent_Frozen_Food + Percent_Meat + Percent_Produce + Percent_Floral + Percent_Deli + Percent_Bakery + Percent_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

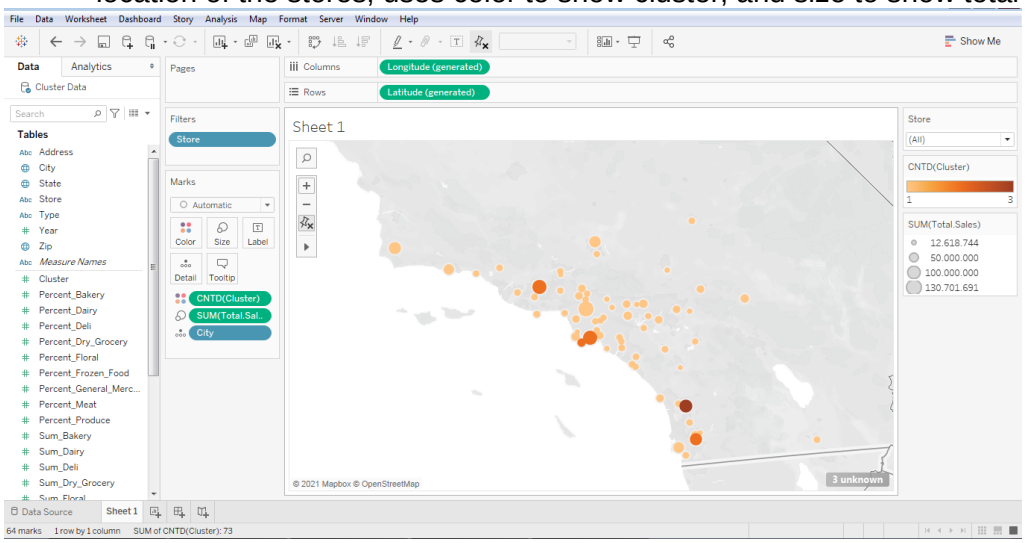
Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   A higher value means that the cluster is oriented towards selling more of that type of goods. For example (as seen in the table below) Percent_Dry_Grocery has positive values in Clusters 1 and 3, and a negative value in Cluster 2. That means that Dry Grocery sells better in the 1st and 3rd Cluster.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-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	-0.178481

	Percent_Bakery	Percent_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

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.



Task 2: Formats for New Stores

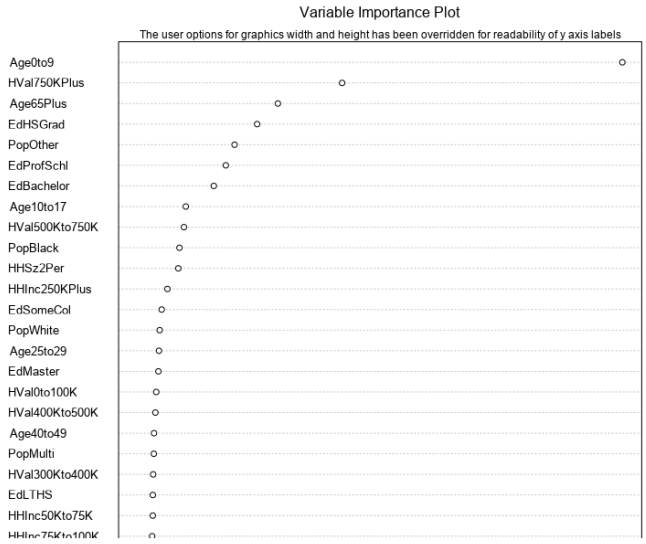
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)
   The Methodology I used to predict the best store format is Boosted Model. Choosing the Forest Model would also be another best choice, since it shows the same Accuracy and F1 score as the Boosted Model

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Boosted_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.
   The three most important variables are: Age0to9, HVal750KPlus and Age65Plus.



3. 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	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?

For the forecast of existing stores I used an ETS(M,N,M) model. The decision was based upon a comparison between the Accuracy Measures for an ETS and ARIMA Model:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Existing_Stores	-286567.1	1223861	1123843	-1.6924	5.5153	0.5982
ARIMA_Existing_Stores	-532064.6	1291405	1121404	-2.8793	5.5696	0.5969

ETS Model has lower errors. Reduced probability of error equals better accuracy.

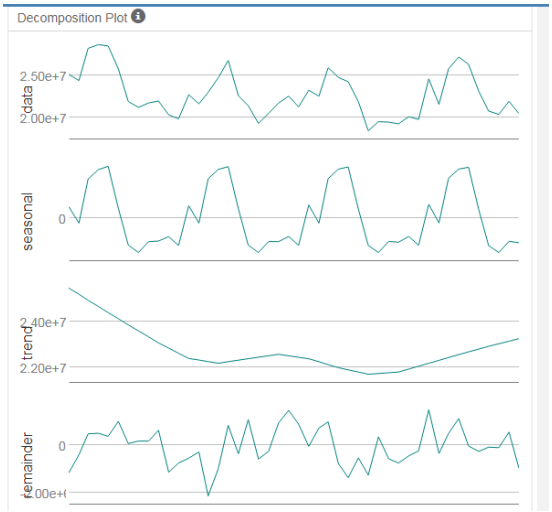
For the Forecast of new stores I used ETS (M,N,M) model as well:

Accuracy Measures:

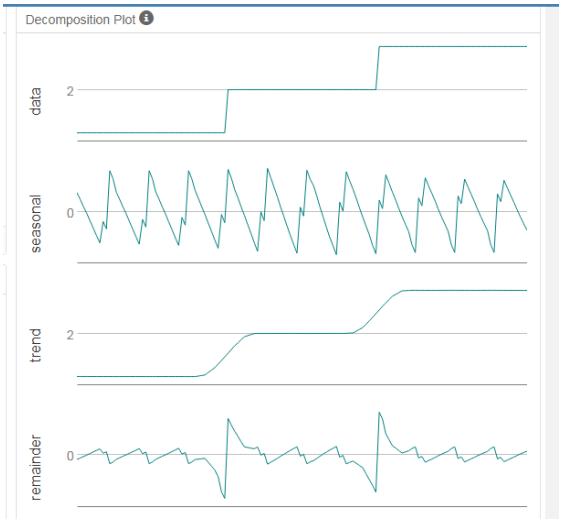
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_New	-22982.91	25784.09	22982.91	-8.9095	8.9095	1.0656
ARIMA_New	-31451.91	32766.59	31451.91	-11.9754	11.9754	1.4583

The M,N,M Configuration is used hence Seasonality changes in magnitude that is why Multiplicative (M) method**,** None (N) for the Trend component because the trend first goes down then up, or for the New Store data it goes up, than it flattens. The error component there is a variation in its magnitude over time so I use Multiplicative (M) method.

Existing Stores:



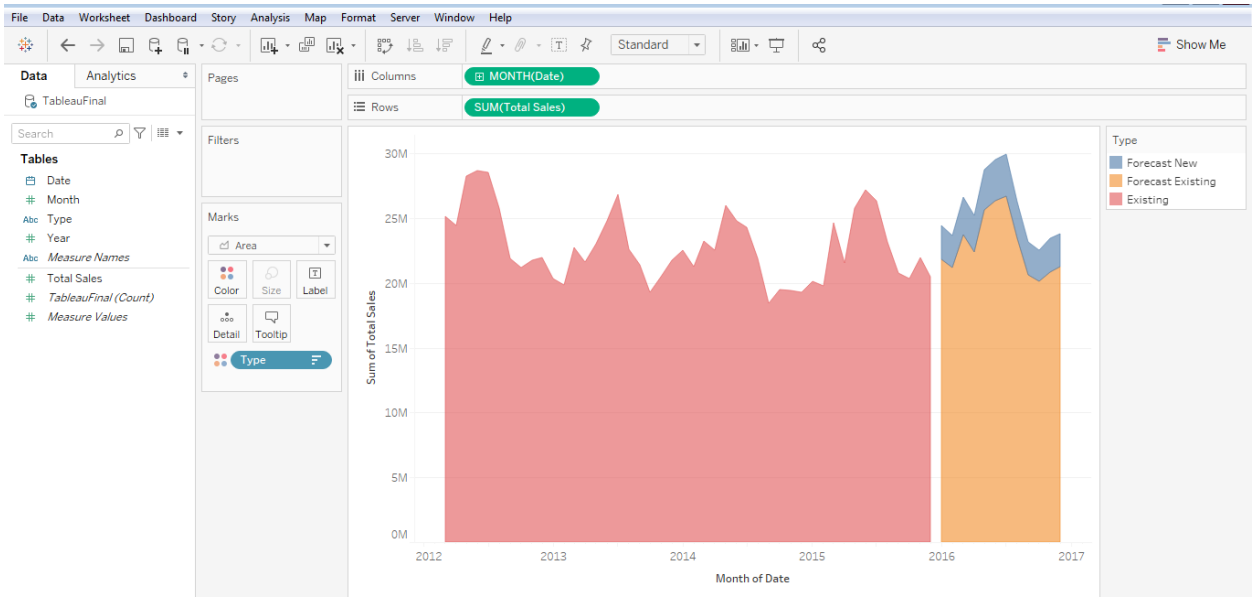
New Stores:



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.

Month	New Stores	Existing Stores
Jan-16	2563357.91	21829060.03
Feb-16	2483924.728	21146329.63
Mar-16	2910944.146	23735686.94
Apr-16	2764881.87	22409515.28

May-16	3141305.867	25621828.73
Jun-16	3195054.204	26307858.04
Jul-16	3212390.954	26705092.56
Aug-16	2852385.769	23440761.33
Sep-16	2521697.187	20640047.32
Oct-16	2466750.894	20086270.46
Nov-16	2557744.588	20858119.96
Dec-16	2530510.805	21255190.24



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.