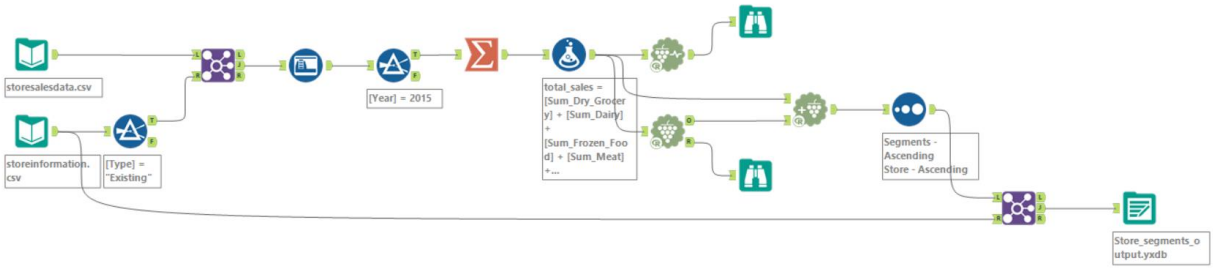
Project: Predictive Analytics Capstone

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 store formats are 3.

The optimal number of store formats is 3. This is because it has high median values within both the AR and CH index and smaller spread, showing compactness. We can get this through using the K-Centroids diagnosis tool on Alteryx.

Report

K-Means Cluster Assessment Report

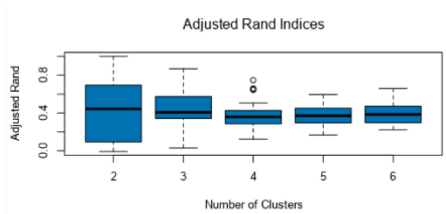
Summary Statistics

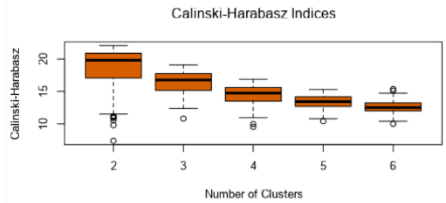
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.007639	0.029695	0.122167	0.166791	0.222111
1st Quartile	0.094172	0.343478	0.285754	0.298186	0.301965
Median	0.443213	0.406361	0.357989	0.370994	0.384296
Mean	0.405201	0.443015	0.365307	0.383051	0.389198
3rd Quartile	0.684276	0.56807	0.424442	0.450713	0.470301
Maximum	1	0.868183	0.747642	0.595251	0.659091

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	7.376319	10.80678	9.524605	10.41103	10.00938
1st Quartile	17.163364	15.15871	13.531027	12.71013	11.99892
Median	19.816152	16.75762	14.737409	13.42556	12.51619
Mean	18.520371	16.39173	14.436238	13.36015	12.61465
3rd Quartile	20.893269	17.74967	15.580417	14.17377	13.23228
Maximum	22.061691	19.089	16.865033	15.29623	15.36927





2. How many stores fall into each store format?

   Format 1: 23 Stores
   Format 2: 29 Stores
   Format 3: 33 Stores

   Using the K-Centroids Cluster Analysis tool using the same configuration as we used in K-Centroids Diagnostics tool:

Report

Summary Report of the K-Means Clustering Solution Store_Cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Pct_Dry_Grocery + Pct_Dairy + Pct_Frozen_Food + Pct_Meat + Pct_Produce + Pct_Floral + Pct_Deli + Pct_Bakery + Pct_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

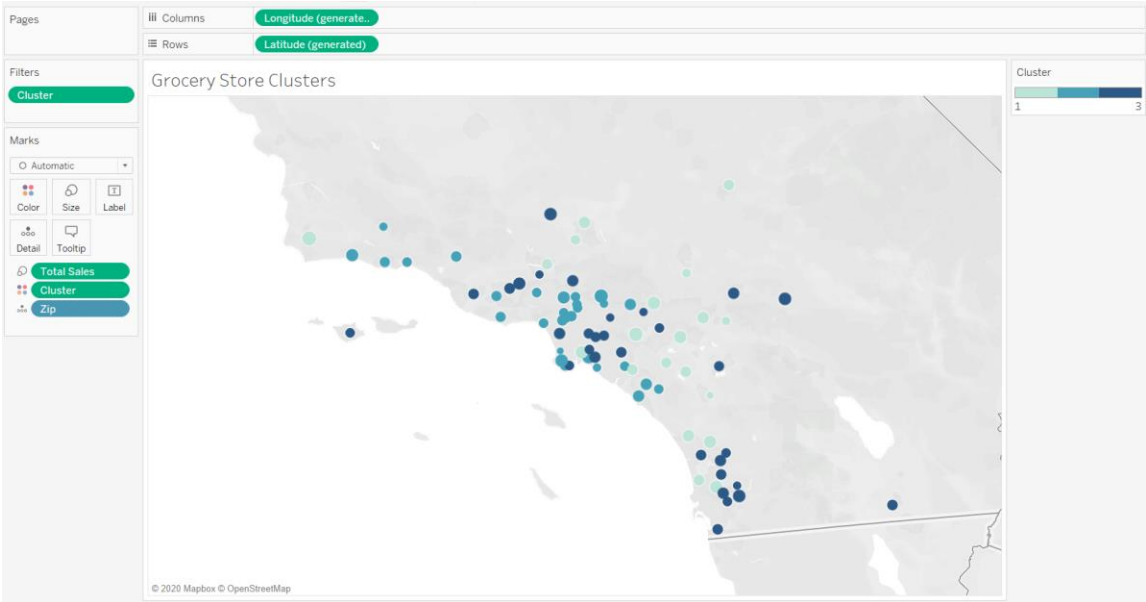
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

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

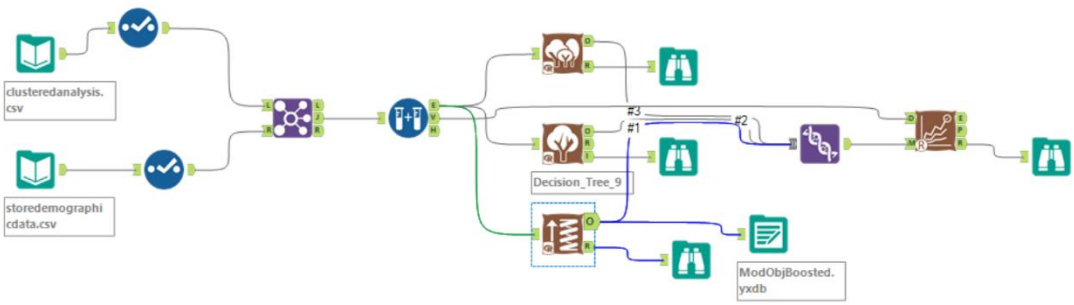
   Dairy inventory needs to be increased in case of Format/cluster 2 store as compared to stores that fall under cluster 1 & 3.

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.)



Since we must predict the clusters for the stores, so it is a classification problem and we have more than 2 classes to predict. So, we will use Decision Tree, Random Forest and Boosting. Using the workflow above, we get the statistics:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Decision Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Use the above created model to predict the clusters for the new stores:

Confusion matrix of Boosted_Model

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

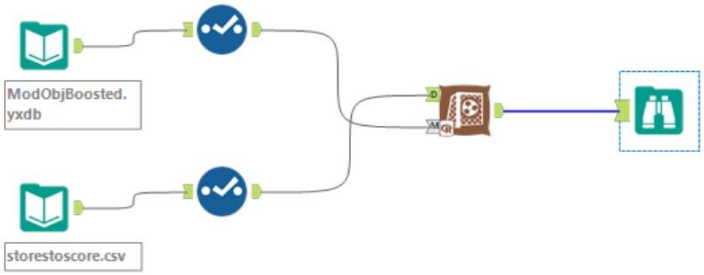
Confusion matrix of Decision Tree

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

Confusion matrix of Forest_Model

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

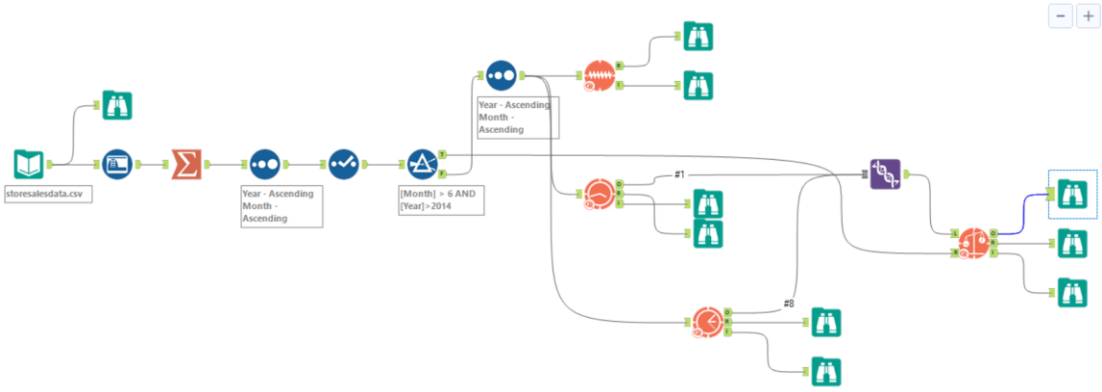
2. What format do each of the 10 new stores fall into? Please fill in the table below.



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

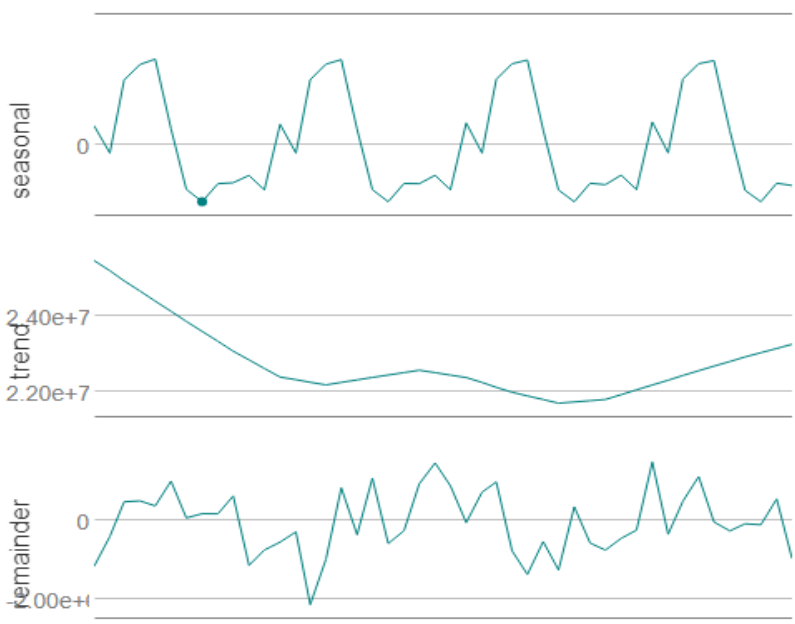
So, there are 3 stores in cluster 1, 6 stores in cluster 2 and 1 store in cluster 3.

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 existing stores, I used both the ETS and ARIMA models to find the best solution. For predicting the aggregate produce for the existing stores, I plotted the Decomposition plots to understand the trend, seasonality, and error. Looking at the three plots below, it is apparent that there exists seasonality and the error appears to decrease over time. Since the trend curve slopes upward after a period, I will not use that. So, I will have seasonality multiplicatively, trend as none, and remainder multiplicatively giving an ETS (M, N, M).



We can see in the decomposition plot above there is no trend, seasonal is multiplicative and error is multiplicative. After comparing the results against the holdout sample, the ETS performs better against the ARIMA model.

Summary of Time Series Exponential Smoothing Model ETS_MNM

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

Smoothing parameters:

Parameter	Value
alpha	0.674884
gamma	0.000203

Initial states:

State	Value
l	23146230.586012
s0	0.90906
s1	0.938619
s2	0.926304
s3	0.901291
s4	0.870972
s5	0.897637
s6	1.019225
s7	1.166556
s8	1.167388
s9	1.137259
s10	0.997793

Summary of ARIMA Model ARIMA_100_110

Method: ARIMA(1,0,0)(1,1,0)[12]

Call:
Arima(Sum_Produce, order = c(1, 0, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

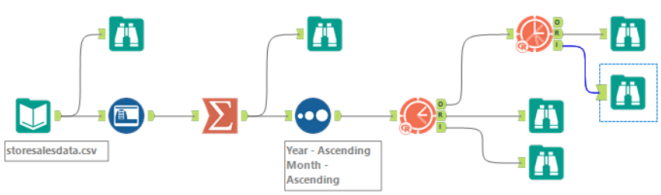
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

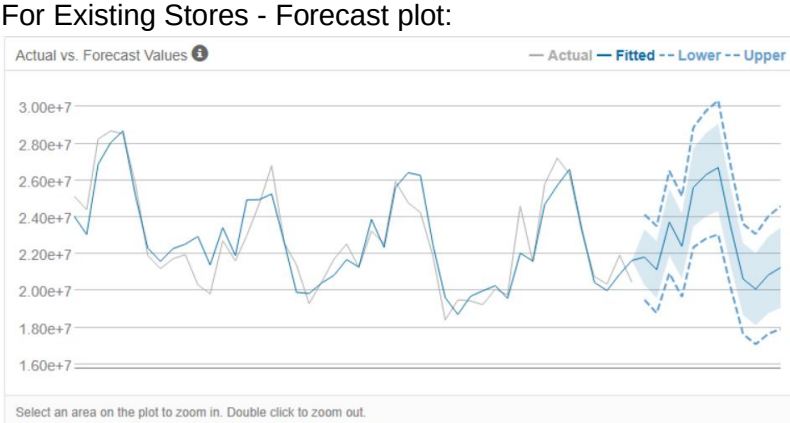
Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

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.

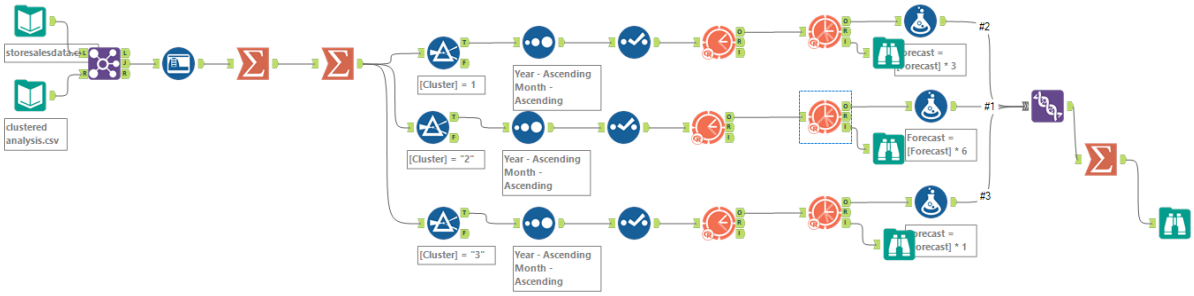
For Existing Stores - Workflow:



For Existing Stores - Forecast plot:

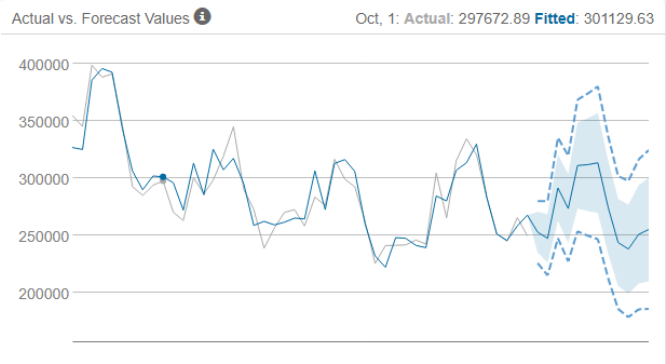


For New Stores - Workflow:

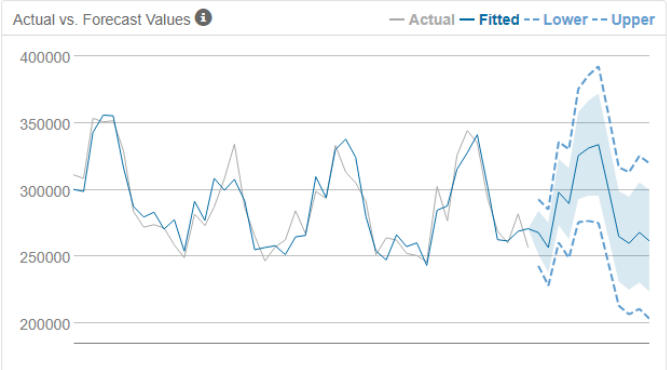


For New Stores – Individual Forecast plot:

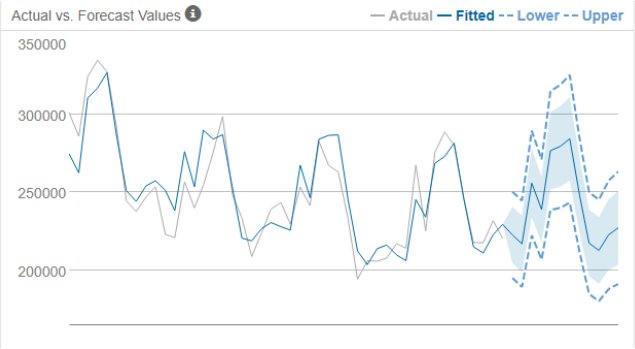
Cluster 1:



Cluster 2:



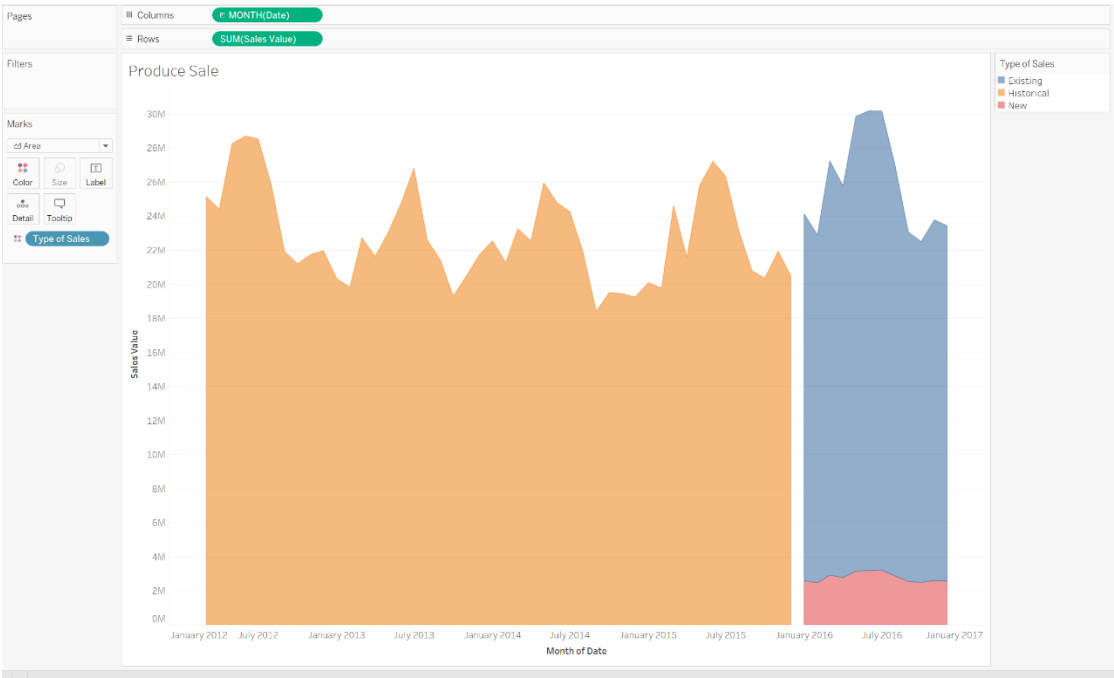
Cluster 3:


Actual vs. Forecast Values
Actual
Fitted
Lower
Upper
350000
300000
250000
200000

Forecast Table:

Period	New Stores	Existing Stores
Jan 16	2,587,451	21,539,936
Feb 16	2,477,353	20,413,771
Mar 16	2,913,185	24,325,953
Apr 16	2,775,746	22,993,466
May 16	3,150,867	26,691,951
Jun 16	3,188,922	26,989,964
July 16	3,214,746	26,948,631
Aug 16	2,866,349	24,091,579
Sep 16	2,538,727	20,523,492
Oct 16	2,488,148	20,011,749
Nov 16	2,595,270	21,177,435
Dec 16	2,573,397	20,855,799

Tableau Visualisation:



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.