Predictive Analytics for Business Nanodegree

Project: Predictive Analytics Capstone

Latifa M.Alyaeesh

Misk Academy & Udacity

Capstone Project Overview

The capstone project has three main tasks, each of which requires you to use skills you developed during the Nanodegree program.

Task 1: Store Format for Existing Stores

Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

Task 2: Store Format for New Stores

The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, we don’t have sales data for these new stores yet, so we’ll have to determine the format using each of the new store’s demographic data.

Task 3: Forecasting

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast.

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 best number of clusters is 3 since shows higher Median in ARI while it maintains low Variability.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.009475	0.160572	0.172381	0.204008
1st Quartile	0.353058	0.344134	0.296018	0.285294
Median	0.4926	0.498265	0.382588	0.36941
Mean	0.475777	0.487769	0.402193	0.366629
3rd Quartile	0.654984	0.614865	0.48537	0.433118
Maximum	0.952939	0.759953	0.775237	0.614868

Calinski-Harabasz Indices:

	2	3	4	5
Minimum	10.38298	10.05244	11.8645	10.77356
1st Quartile	18.7784	15.96022	14.07268	13.03449
Median	20.07012	16.90389	15.11582	13.615
Mean	19.08731	16.64035	14.79844	13.70011
3rd Quartile	20.87407	17.91537	15.72883	14.38381
Maximum	22.44228	18.93512	16.62962	16.10526

Figure 1: K-Means Cluster Assessment Report

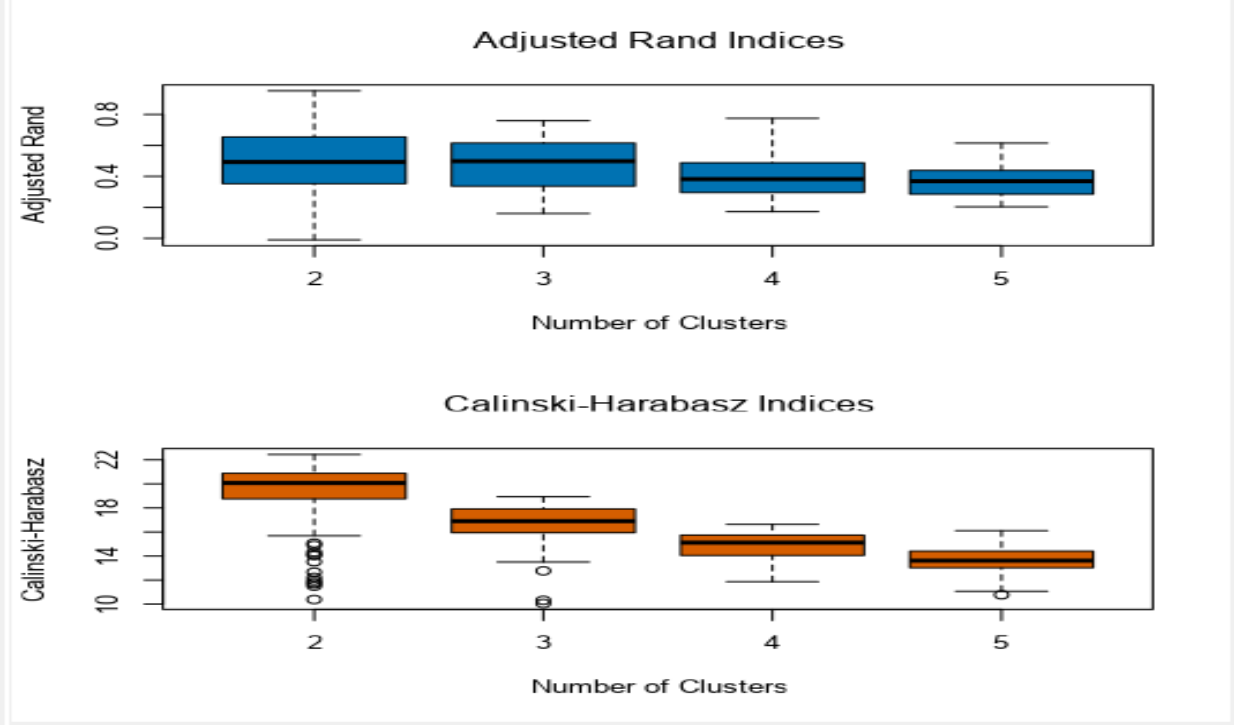

Figure 2: Plots of K-Means Cluster Assessment Report

2. **How many stores fall into each store format?**

Cluster 1 has 25 stores, cluster 2 has 35 stores and cluster 3 has 25 stores.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823872	2.191565
2	35	2.475018	4.412367	1.947297
3	25	2.289004	3.585931	1.72574

Figure 3: Cluster Information

Hint: I used Alteryx 2020.

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

Based on the result shown below, cluster 1 sells a lot of General Merchandise compared to the other two clusters. Clusters 2 sells a lot of produce compared to the other two clusters.

Summary Report of the K-Means Clustering Solution Cluster

Solution Summary

Call:

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

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823872	2.191565
2	35	2.475018	4.412367	1.947297
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.

Sum of within cluster distances: 196.35035.

	Dry.Grocery	Dairy	Frozen	Meat	Produce	Floral	Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435128	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702371	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	Bakery	General
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Figure 4: Summary Report of the K-Means Clustering Solution Cluster

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.**

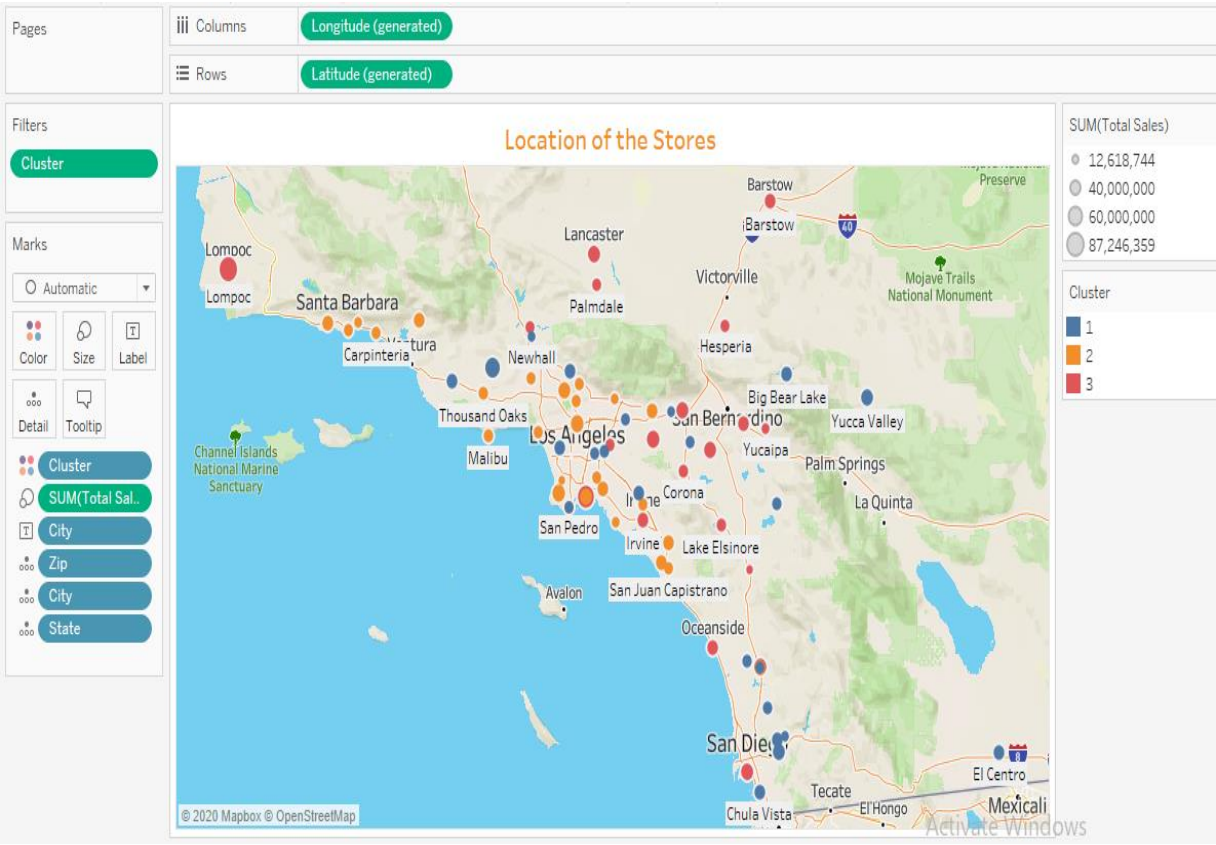

Figure 5: Location of Stores

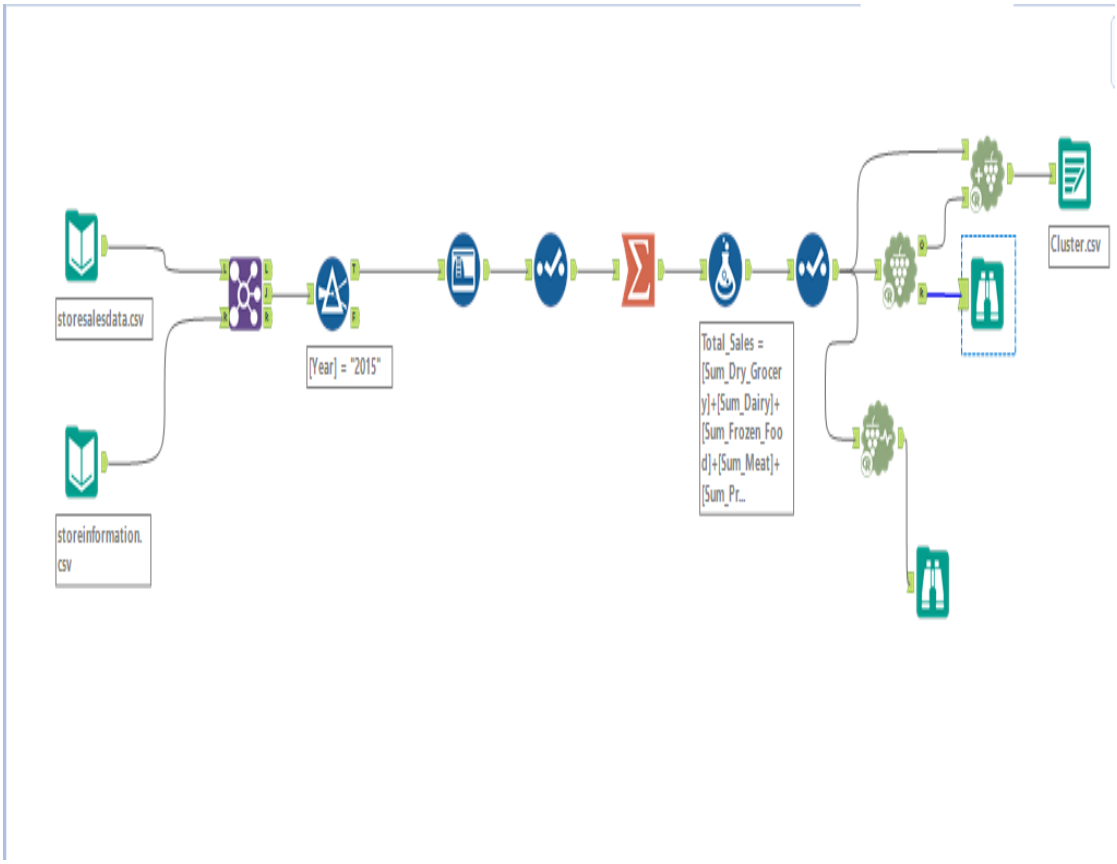

Figure 6: Altrryx workflow (Task 1)

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 model comparison report shows the comparison between Forest Model, Decision Tree and Boosted Model. All models have the accuracy but Boosted Model chosen due to higher F1 value of 0.8333.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_13	0.7059	0.7083	0.6250	1.0000	0.5000
Forest	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted	0.7647	0.8333	0.5000	1.0000	1.0000

Figure 7: Model Comparison Repot

2. **What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization**

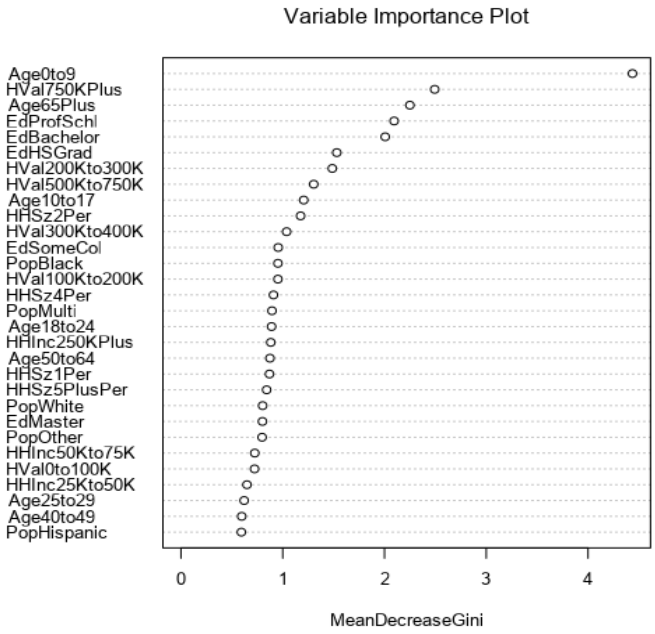

Figure 8: Variable importance plot

3- 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

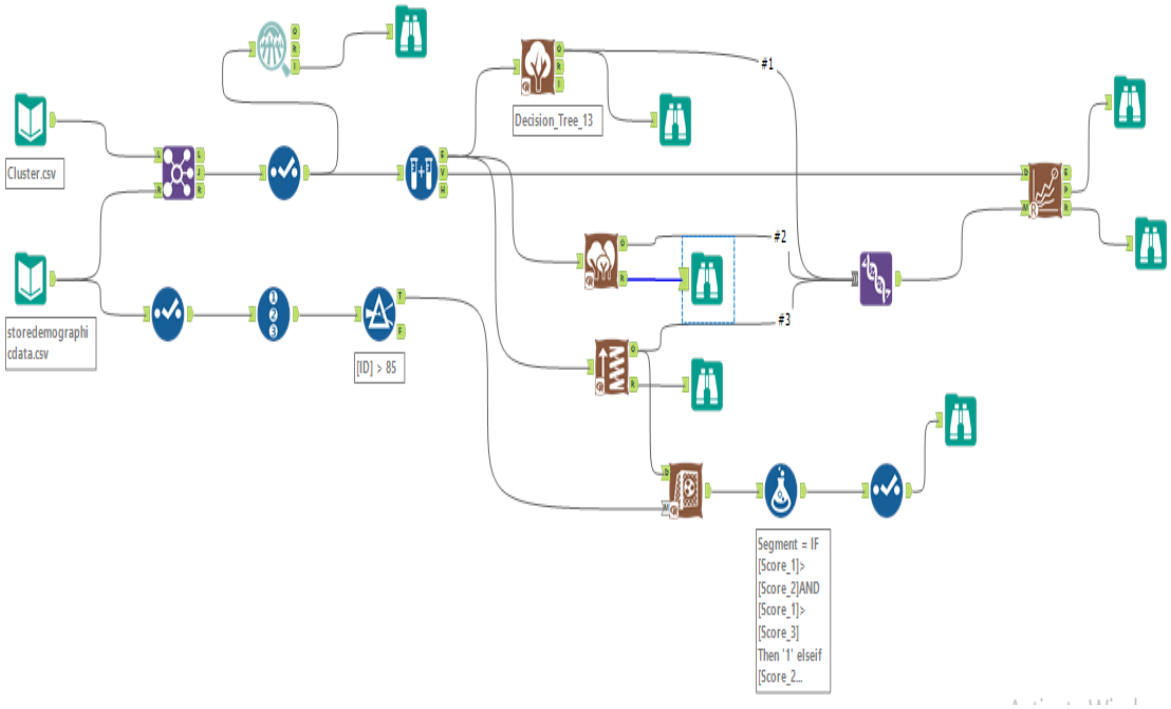

Figure 9: Alteryx Model Classification

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?

After examining the decomposition plots with ACF and PACF , the resulted ETS model is ETS (M,N,A) since the seasonality was changing through the analyzed period while there no linear or exponential trend and the error is changing in the magnitude so it would be multiplicative.

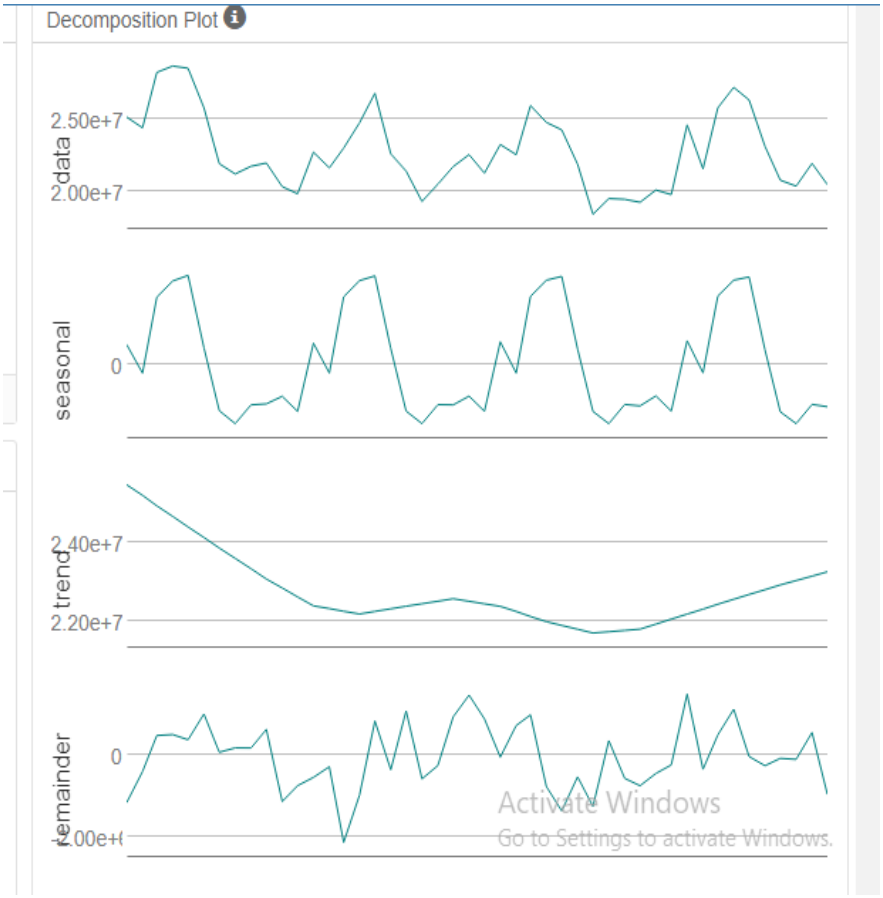

Figure 10: Decomposition Plot

TS Comparison

For the accuracy measure, the ETS model has better accuracy in all of the measures so ETS (M, N, and M) model will used.

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS	ARIMA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

Figure 11: Comparison of Time Series Model

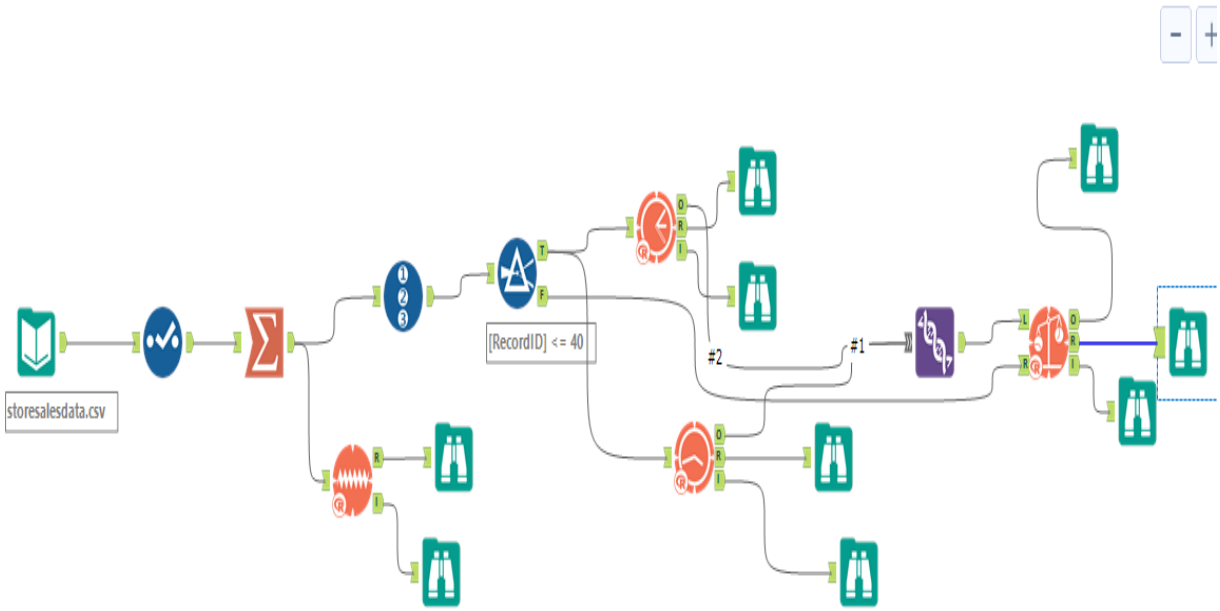

Figure12: Alteryx workflow (ETS & ARIMA)

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	New Store	Existing Store
Jan-16	2,588,357	21,745,141
Feb-16	2,498,567	21,188,192
Mar-16	2,919,067	23,671,624
Apr-16	2,797,280	22,388,691
May-16	3,163,765	22,570,588
Jun-16	3,202,813	26,293,565
Jul-16	3,228,212	26,710,714
Aug-16	2,868,915	23,472,138
Sep-16	2538,372	20,665,605
Oct-16	2,485,732	20,915,647
Nov-16	2,583,448	20,915,647
Dec-16	2,562,182	21,207,755

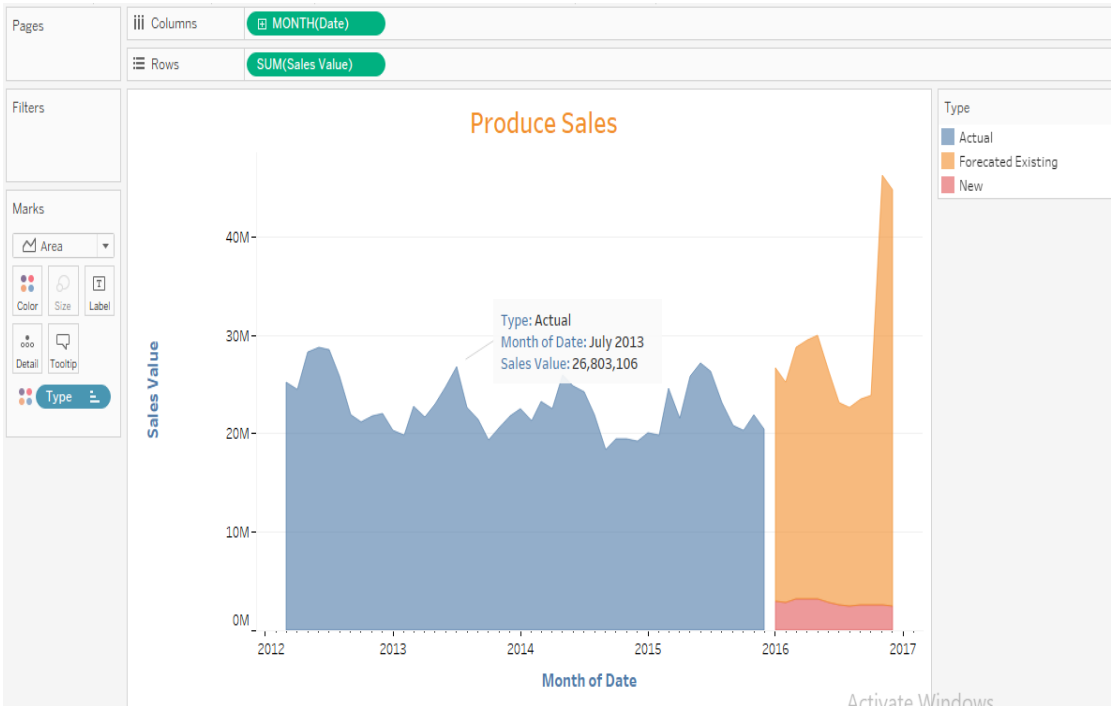

Figure 13: Total Sales Forecast

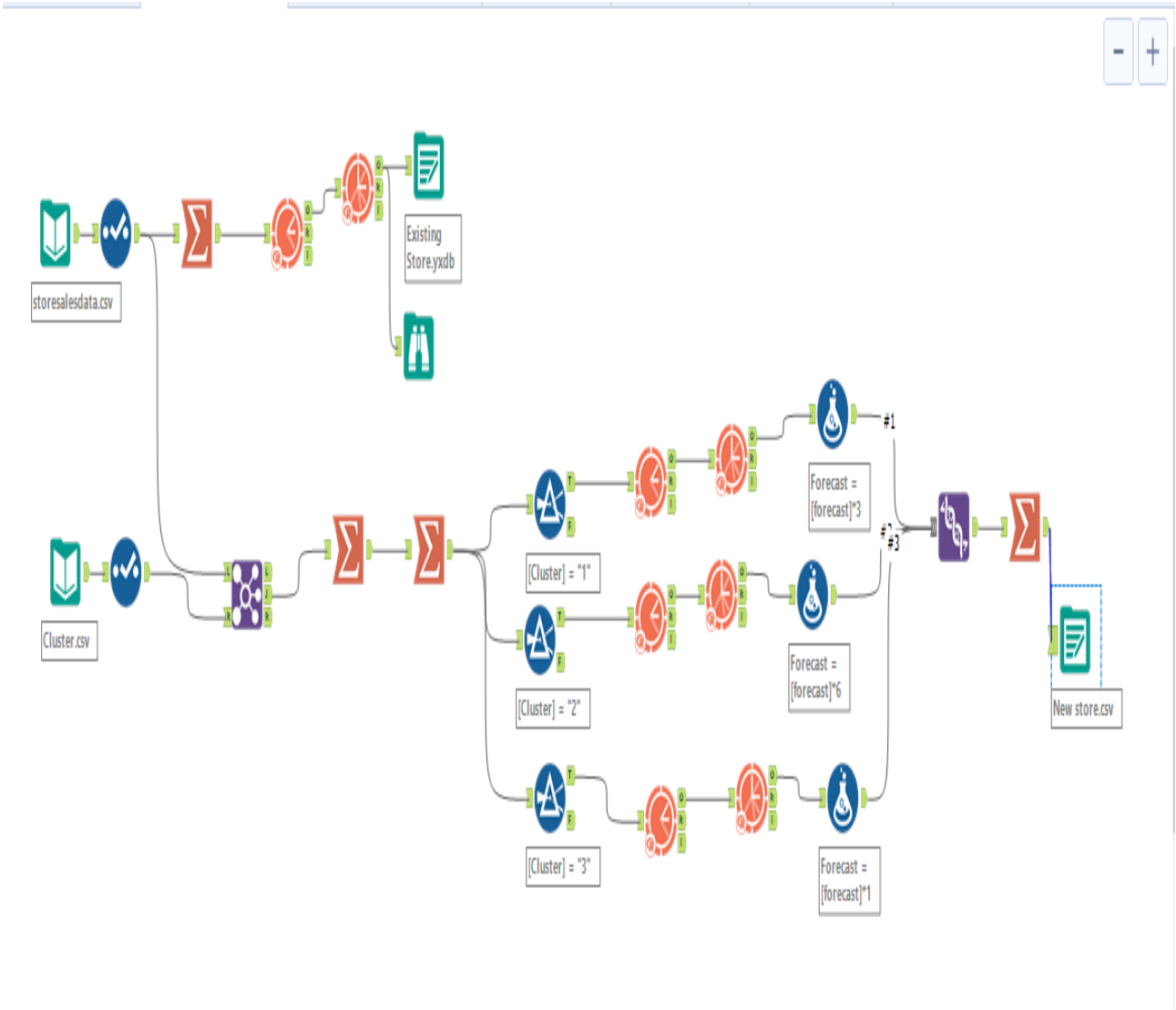

Figure 14: Alteryx workflow (Forecast)