Combining Predictive Techniques

Practicing Alteryx & Tableau

Predictive Analytics for Business Nanodegree Program

UDACITY

Rosana Ferreira Soares dos Santos

Feb, 02 2021

Predictive Analytics Capstone

Combining Predictive Techniques

Introduction

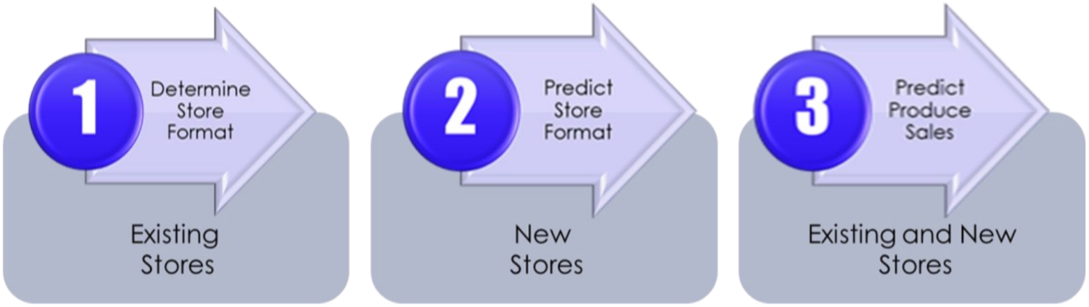
Project Rubric, https://review.udacity.com/#!/rubrics/437/view
Nanodegree, https://www.udacity.com/course/predictive-analytics-for-business-nanodegree--nd008

The Business Problem

The company where I currently work 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. I have been asked to provide analytical support to make decisions about store formats and inventory planning.

Details

I have been asked to complete three tasks, which are the following and detailed in sequence.



Three data files were provided.

- *StoreInformation.csv*
  This file contains location data for each of the stores.

- *StoreSalesData.csv*
  This file contains sales by product category for all existing stores for 2012, 2013, and 2014.

- *StoresDemographicData.csv*
  This file contains demographic data for the areas surrounding each of the existing stores and locations for new stores.

Task 1: Store Format for Existing Stores

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms "formats" and "segments" will be used interchangeably throughout this project.

I have been asked to:

1. Determine the optimal number of store formats based on sales data:
   - Sum sales data by StoreID and Year
   - Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
   - Use only 2015 sales data.
   - Use a K-means clustering model.
2. Segment the 85 current stores into the different store formats.
3. Use the StoreSalesData*.csv* and *StoreInformation.csv* files.

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. Determine the optimal number of store formats based on sales data.

I have been asked to:

1. Develop a model that predicts which segment store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
2. Use a 20% validation sample with *Random Seed* = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
3. Use the model to predict the best store format for each of the 10 new stores.
4. Use the StoreDemographicData.csv file, which contains the information for the area around each store.
5. Note: In a real world scenario, you could use PCA to reduce the number of predictor variables. However, there is no need to do so in this project. You can leave all predictor variables in the model.

Task 3: Forecasting Produce Sales

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast. I have asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, follow the steps below.

Note: Use a 6 month holdout sample for the TS Compare tool (this is because we do not have that much data so using a 12 month holdout would remove too much of the data).

1. To forecast produce sales for existing stores you should aggregate produce sales across all stores by month and create a forecast.
2. To forecast produce sales for new stores:
   - Forecast produce sales (not total sales) for the average store (rather than the aggregate) for each segment.
   - Multiply the average store produce sales forecast by the number of new stores in that segment.
   - For example, if the forecasted average store produce sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
   - Sum the new stores produce sales forecasts for each of the segments to get the forecast for all new stores.
3. Sum the forecasts of the existing and new stores together for the total produce sales forecast.

Predictive Analytics Capstone

Combining Predictive Techniques

Project Development

Project Rubric, https://review.udacity.com/#!/rubrics/437/view
Nanodegree, https://www.udacity.com/course/predictive-analytics-for-business-nanodegree--nd008

Task 1: Determine Store Formats for Existing Stores

I developed the following activities to solve task 1.

1) understood the objective of segmentation;
2) prepared data;
3) determined the number of clusters;
4) created the cluster model;
5) visualized and validated results.

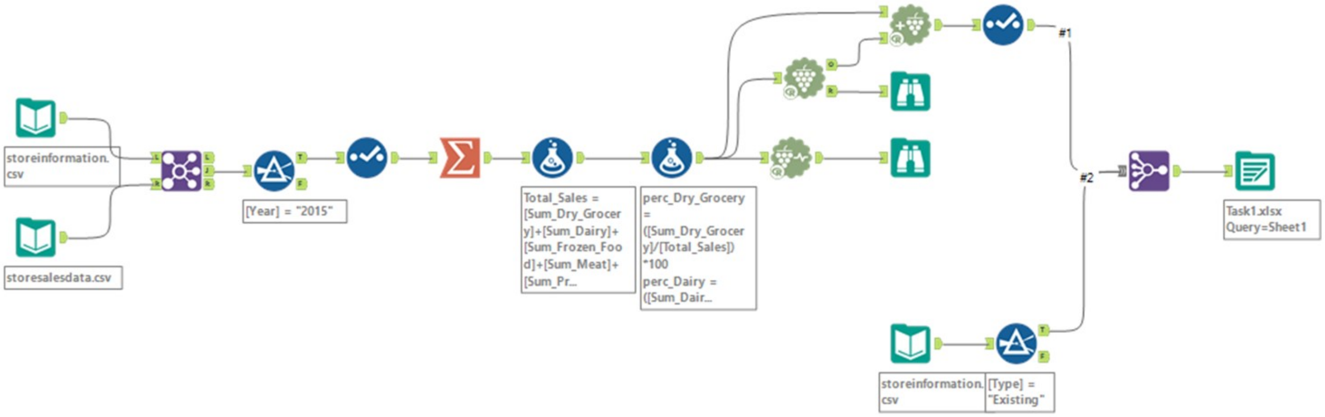

storeinformation.csv
storesalesdata.csv
[Year] = "2015"
Total_Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
perc_Dry_Grocery = ([Sum_Dry_Grocery]/[Total_Sales])*100
perc_Dairy = ([Sum_Dair...
#1
#2
storeinformation.csv
[Type] = "Existing"
Task1.xlsx
Query=Sheet1

Image 1, Determining Store Formats for Existing Stores using Alteryx

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

I concluded it would be 2 according to the results of the Alteryx workflow I developed. Although by https://knowledge.udacity.com/questions/450900 explanation and other mentor orientation that I read and understood by the time this project was conceived there was an older version of Alteryx running performing then higher median values related to cluster 3 in AR and C-H plots.

The optimal number of store formats is 3.

Understanding segmentation objective and preparing data.

The data provided had all the information needed to accomplish the analysis. First I understood the objective for the segmentation I was trying to create. Joined two data files, filtered 2015 information, changed Sales fields from VString to Double. After that I summarized data grouping it by *Store Id* and *Year* and summing by each category. Using *Formula Tool* I created a *Total_Sales* field which sum all categories. Next step was the creation of nine different columns each one receiving the sum of a category divided by *Total_Sales* multiplied by 100, in other words. By the end of this specific cycle I obtained 85 records.

Determining the number of clusters.

Proceeded using *K-Centroids Diagnostics Tool*

Customized parameters …

- Standardize the fields = enabled (z-score)
- Clustering method = K-Means
- Minimum number of clusters = 2
- Maximum number of clusters = 10
- Bootstrap replicates = 50
- Number of starting seeds = 10

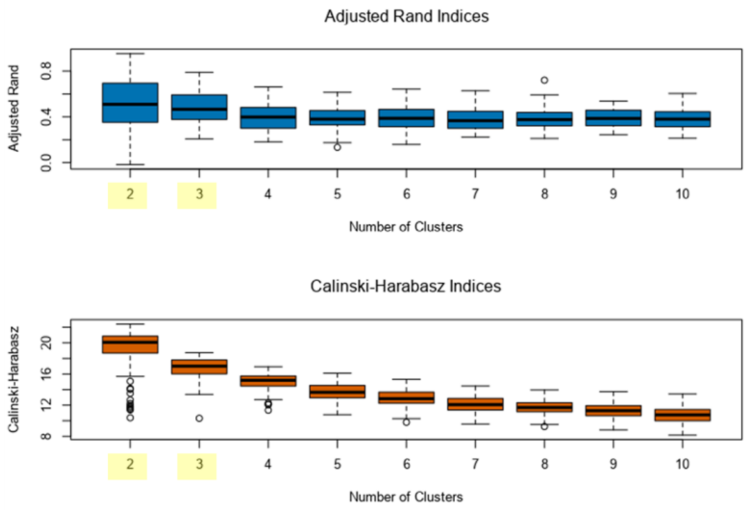
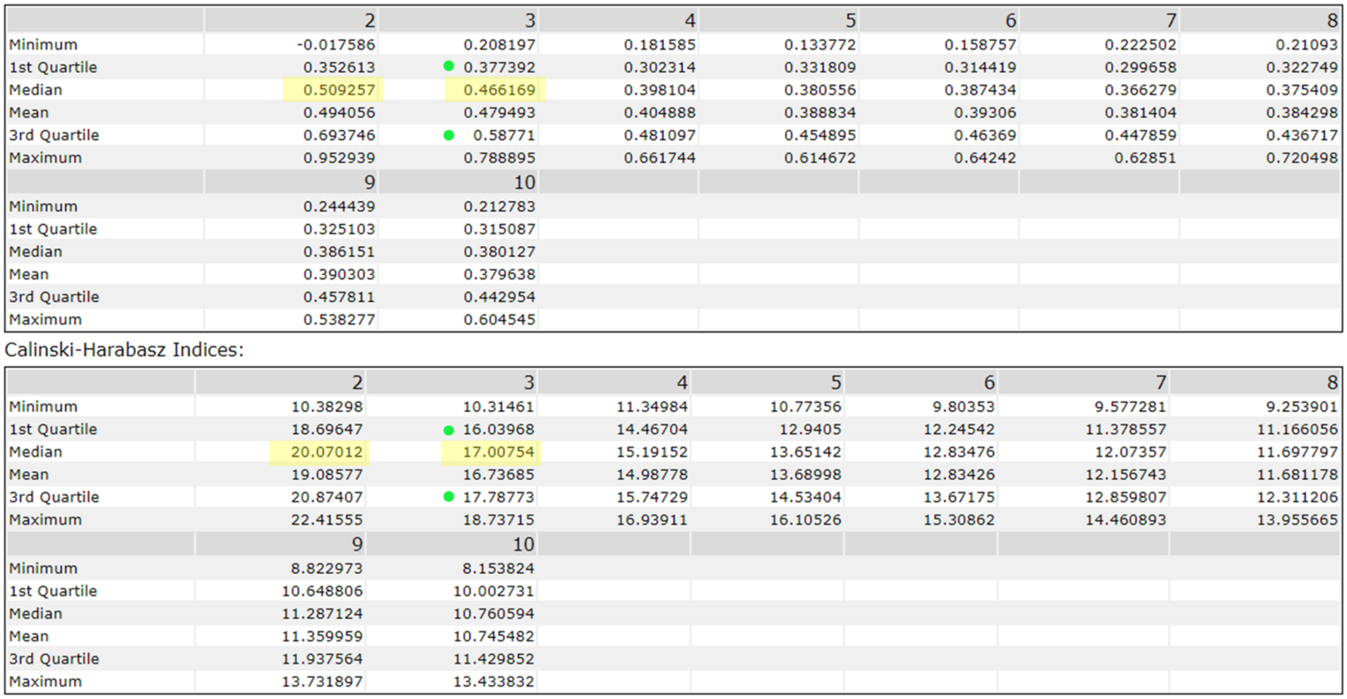


Image 2, AR and C-H Index plots

For both of the following plots AR and C-H indexes Images 1& 2) the y-axis is the value of the index and the x-axis is the number of clusters. For the AR index, the higher the index, the better the stability of the cluster. For the C-H index, the higher the index, the better the distinctness and compactness of the clusters.



	2	3	4	5	6	7	8
Minimum	-0.017586	0.208197	0.181585	0.133772	0.158757	0.222502	0.21093
1st Quartile	0.352613	● 0.377392	0.302314	0.331809	0.314419	0.299658	0.322749
Median	0.509257	0.466169	0.398104	0.380556	0.387434	0.366279	0.375409
Mean	0.494056	0.479493	0.404888	0.388834	0.39306	0.381404	0.384298
3rd Quartile	0.693746	● 0.58771	0.481097	0.454895	0.46369	0.447859	0.436717
Maximum	0.952939	0.788895	0.661744	0.614672	0.64242	0.62851	0.720498

	9	10
Minimum	0.244439	0.212783
1st Quartile	0.325103	0.315087
Median	0.386151	0.380127
Mean	0.390303	0.379638
3rd Quartile	0.457811	0.442954
Maximum	0.538277	0.604545

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.38298	10.31461	11.34984	10.77356	9.80353	9.577281	9.253901
1st Quartile	18.69647	● 16.03968	14.46704	12.9405	12.24542	11.378557	11.166056
Median	20.07012	17.00754	15.19152	13.65142	12.83476	12.07357	11.697797
Mean	19.08577	16.73685	14.98778	13.68998	12.83426	12.156743	11.681178
3rd Quartile	20.87407	● 17.78773	15.74729	14.53404	13.67175	12.859807	12.311206
Maximum	22.41555	18.73715	16.93911	16.10526	15.30862	14.460893	13.955665

	9	10
Minimum	8.822973	8.153824
1st Quartile	10.648806	10.002731
Median	11.287124	10.760594
Mean	11.359959	10.745482
3rd Quartile	11.937564	11.429852
Maximum	13.731897	13.433832

Image 3, K-Means Cluster Assessment Report

2. How many stores fall into each store format?
25 stores in cluster 1, 35 stores in cluster 2 and 25 stores in cluster 3.

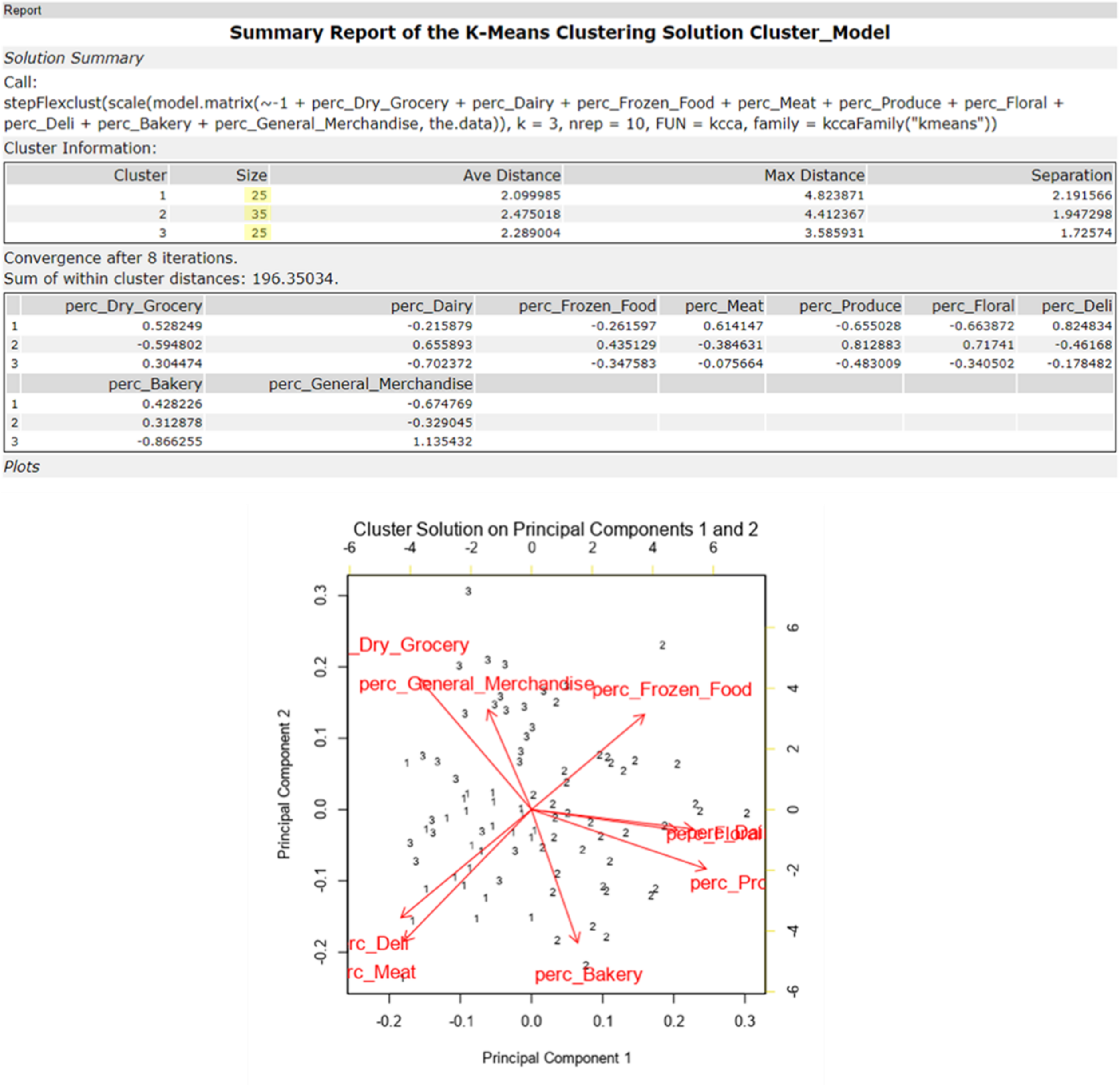
None of the clusters I determined have less than 20 stores or more than 40 stores.

Creating cluster model.

Proceeded using *K-Centroids Cluster Analysis Tool*

Customized parameters …

- Standardize the fields = enabled (z-score)
- Clustering method = K-Means
- Number of clusters = 3
- Number of starting seeds = 10



Report

Summary Report of the K-Means Clustering Solution Cluster_Model

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + perc_Dry_Grocery + perc_Dairy + perc_Frozen_Food + perc_Meat + perc_Produce + perc_Floral + perc_Deli + perc_Bakery + perc_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.

	perc_Dry_Grocery	perc_Dairy	perc_Frozen_Food	perc_Meat	perc_Produce	perc_Floral	perc_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

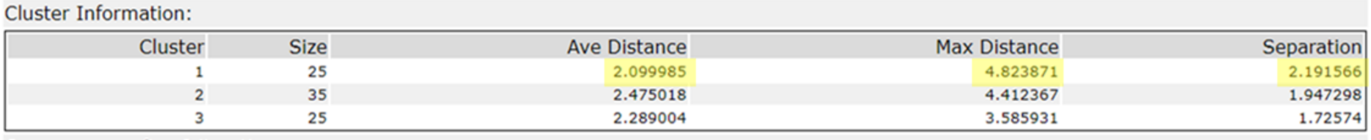
	perc_Bakery	perc_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Plots

Image 4, K-Centroids Cluster Analysis reporting

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

   The main features that distinguish clusters are *average distance*, *maximum distance* and *separation*.



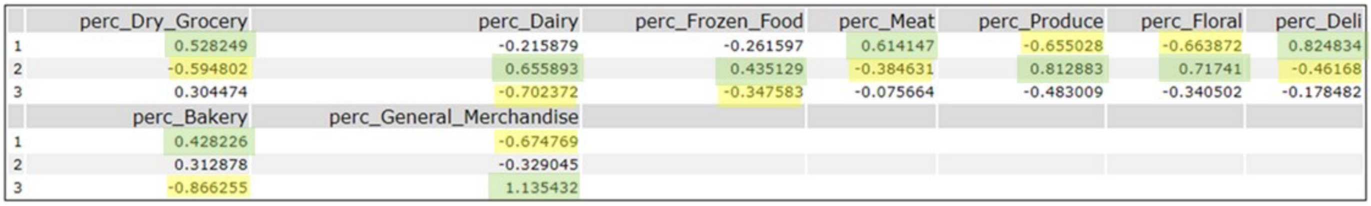
Cluster Information:

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

Image 5, Cluster differences | Avg Distange, Max Distance, Separation |

- Cluster 1 is the most compact. And cluster 2 might present more variability.
- Cluster 1 has the highest maximum distance from centroid.
- Cluster 1 is more separated from the other clusters.

I used the chart with the individual variables checking extreme values for each variable. In green the highest and in yellow the lowest values. A high positive versus a high negative might indicate that those two clusters are opposite to each other.



	perc_Dry_Grocery	perc_Dairy	perc_Frozen_Food	perc_Meat	perc_Produce	perc_Floral	perc_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

	perc_Bakery	perc_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Image 6, Cluster differences | individual variables |

Following a visual of category sales values for the three clusters generated in Tableau. It is possible to observe the nine categories are present although their distribution is different in values, evidenced in y-axis. Cluster 2 sells the most in all categories except for General Merchandise. For that category cluster 3 stands out.

https://public.tableau.com/profile/rosana7921#!/vizhome/AllClustersbycategorie/AllClusters?publish=yes

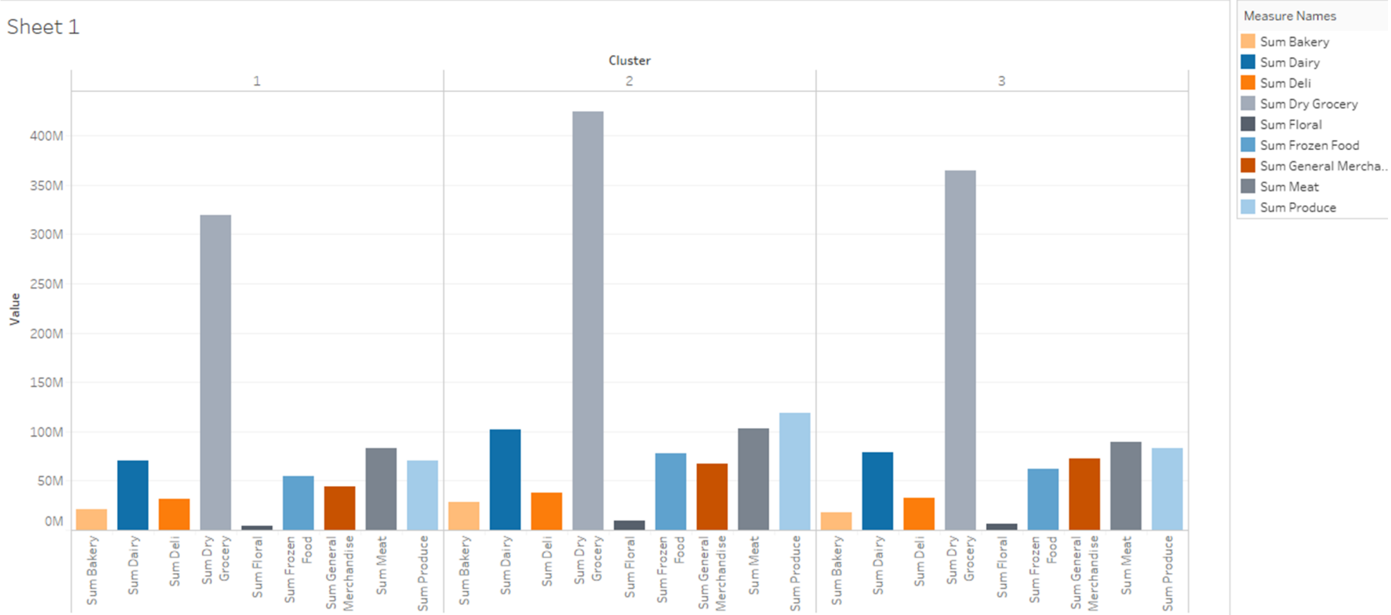


Image 7, Category Sales values per Cluster

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.

https://public.tableau.com/profile/rosana7921#!/vizhome/Task1_16120363339330/Story3?publish=yes

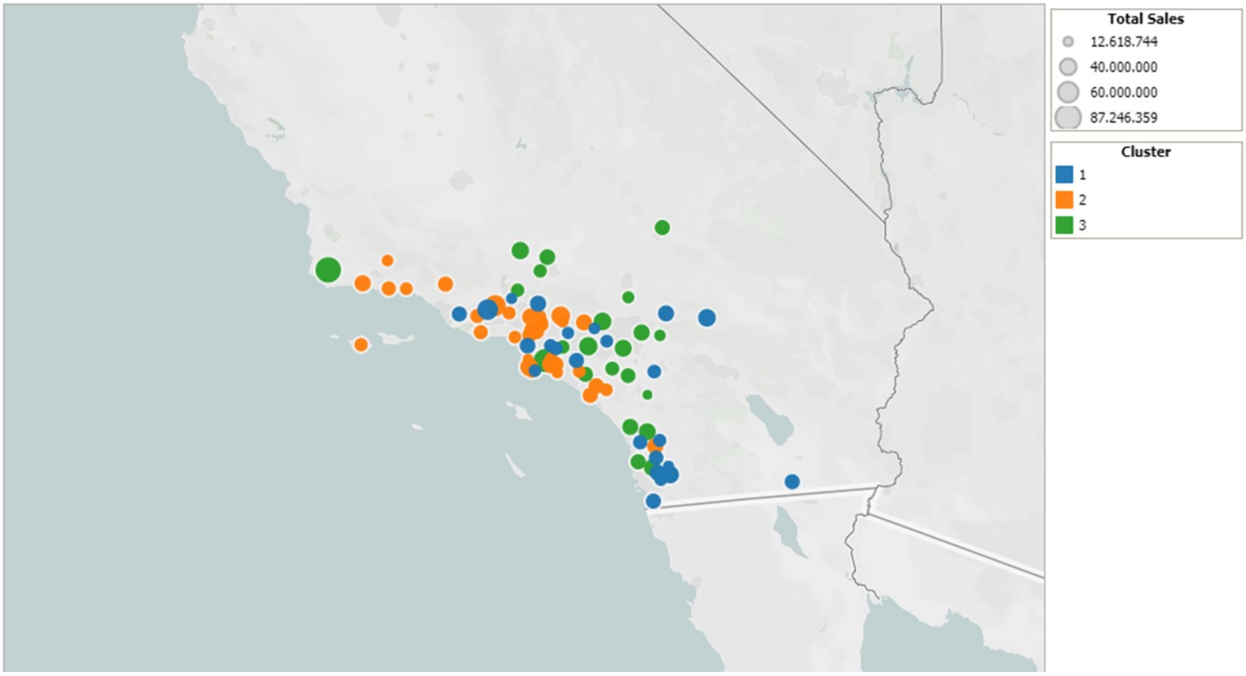


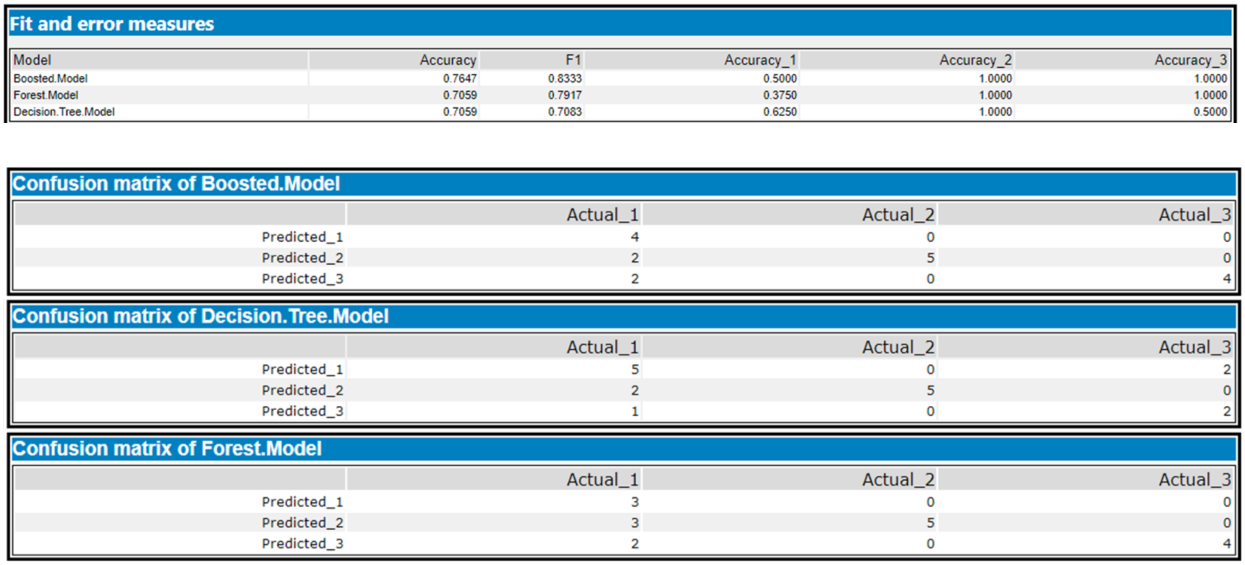
Image 8, Stores location by Total _Sales and coloured by Cluster using Tableau

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

   I considered *Cluster* as target variable and selected all demographic variables as predictors. Compared the performance of Decision Tree, Forest and Boosted models regarding Overall Accuracy, F1 Score and the accuracy of each class. Boosted model presented the best performance.

   I chose Boosted Model.



Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted.Model	0.7647	0.8333	0.5000	1.0000	1.0000
Forest.Model	0.7059	0.7917	0.3750	1.0000	1.0000
Decision.Tree.Model	0.7059	0.7083	0.6250	1.0000	0.5000

Confusion matrix of Boosted.Model

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

Confusion matrix of Decision.Tree.Model

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

Confusion matrix of Forest.Model

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

Image 9, Model Comparison Report

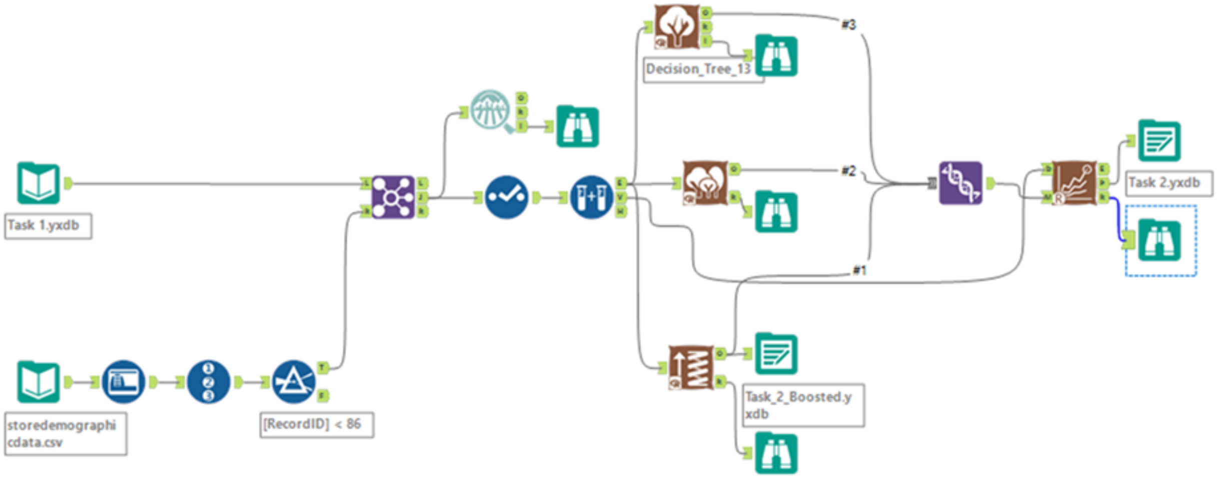


Image 10, Defining the best model using Alteryx

The three most important variables that help explain the relationship between demographic indicators and store formats are *Aget0to9*, *HVal750KPlus* and *Age65Plus*.

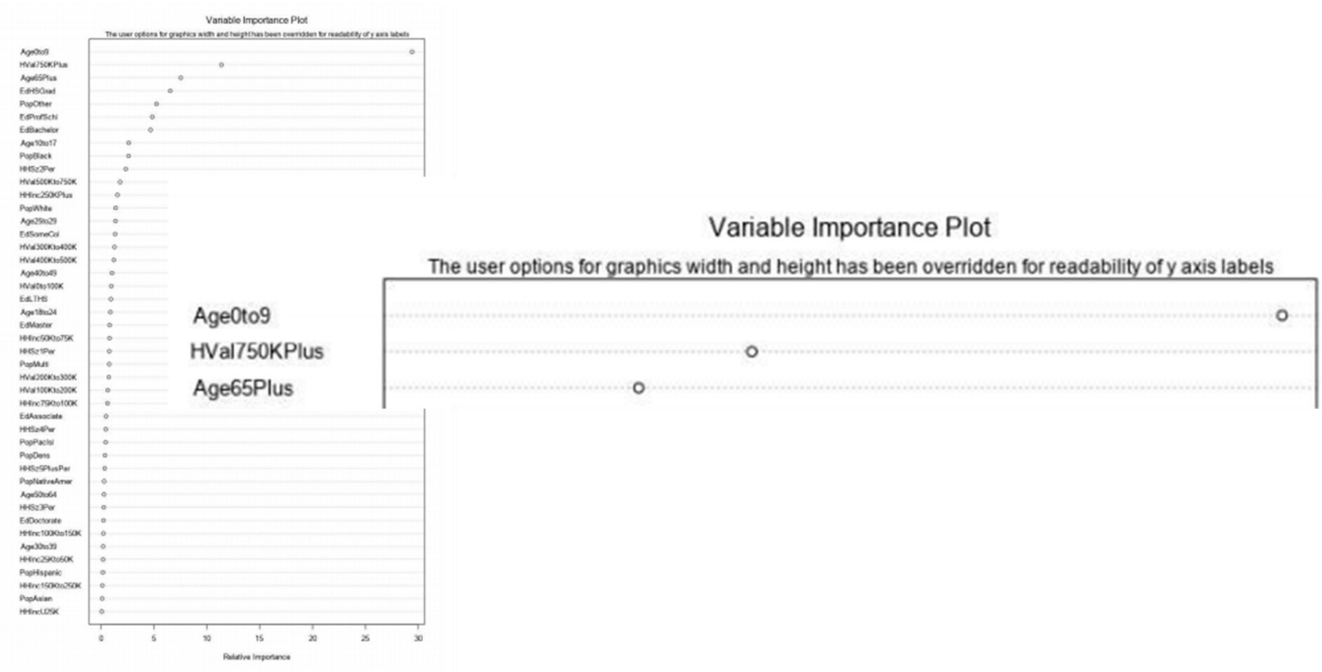


Image 11, Three most important variables

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

I applied Boosted model output, New Store data, Score tool and Formula tool resulting in

- 10% in cluster 1;
- 60% in cluster 2;

- 30% in cluster 3.

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

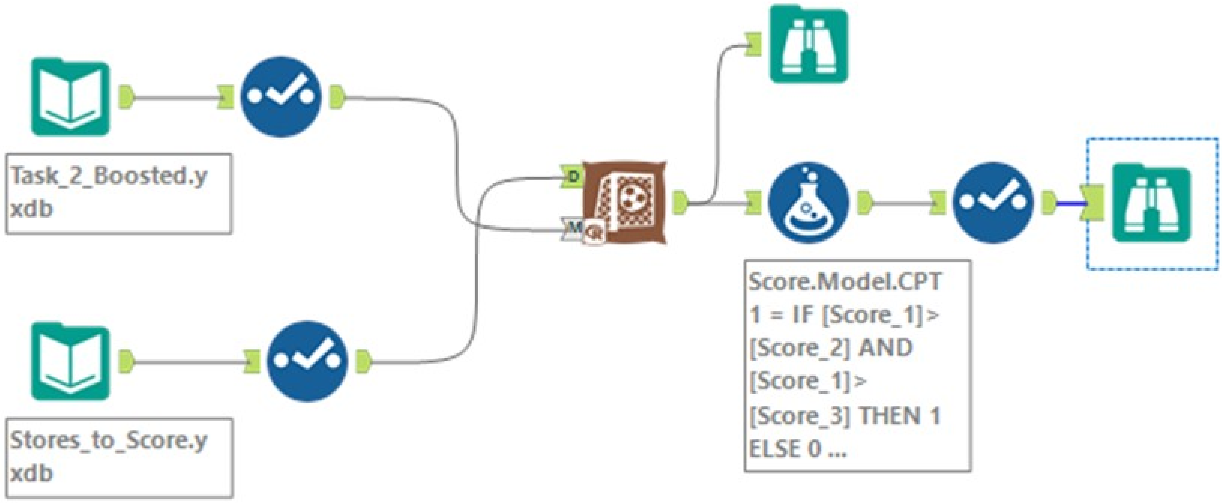


Image 12, Scoring and defining Store Formats for New Stores using Alteryx

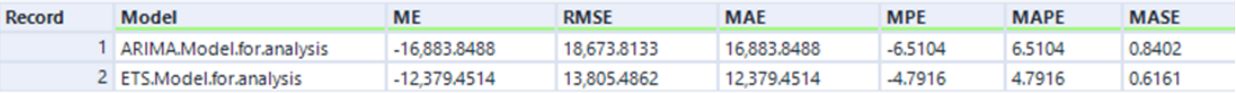
Task 3: Predicting Produce Sales

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?

   To determine the parameters to apply on ETS model building I begun visualizing a time series decomposition plot. I used Summarize tool grouping *Store Sales* data by *Year* and by *Month* and summing *Produce*. After I attached TS Plot tool and customized it for monthly target field frequency and plot type as time series decomposition plot. Based on plot it was possible to determine trend, seasonal and error components.

   - Occurs an irregular variation in magnitude over time in error component. I chose parameter **m**, multiplicative.
   - Trend plot first went downward, then upward, then continued changing. I decided by parameter none (**n**).
   - Seasonality behavior is evident with regular pattern repetition. There is a slightly growth of spikes in magnitude along the years. For this reason I considered a **m**, multiplicative method in this component.
   - Alteryx auto mode suggested ETS(m,n,m).
   - I considered default parameters for ARIMA model despite having analyzed ACF and PAC plots. From plots I concluded that data needed to be differenced to be stationary.
   - After using TS comparing I chose ETS. Compared both models using TS Compare tool. ETS performed best with less error terms.

ETS(m,n,m)



Record	Model	ME	RMSE	MAE	MPE	MAPE	MASE
1	ARIMA.Model.for.analysis	-16,883.8488	18,673.8133	16,883.8488	-6.5104	6.5104	0.8402
2	ETS.Model.for.analysis	-12,379.4514	13,805.4862	12,379.4514	-4.7916	4.7916	0.6161

Image 13, Model error comparing

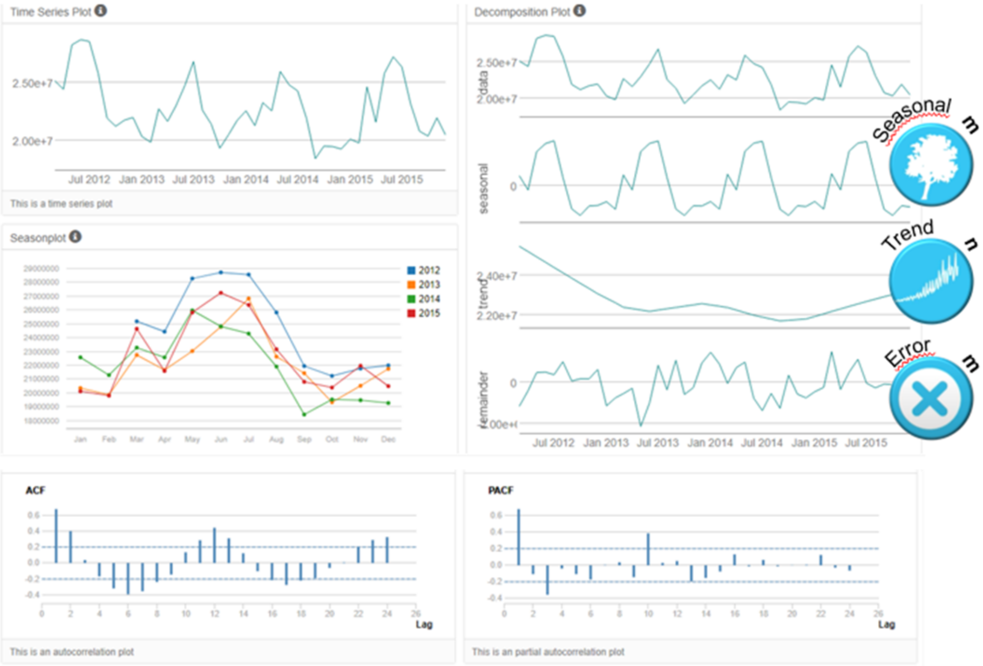
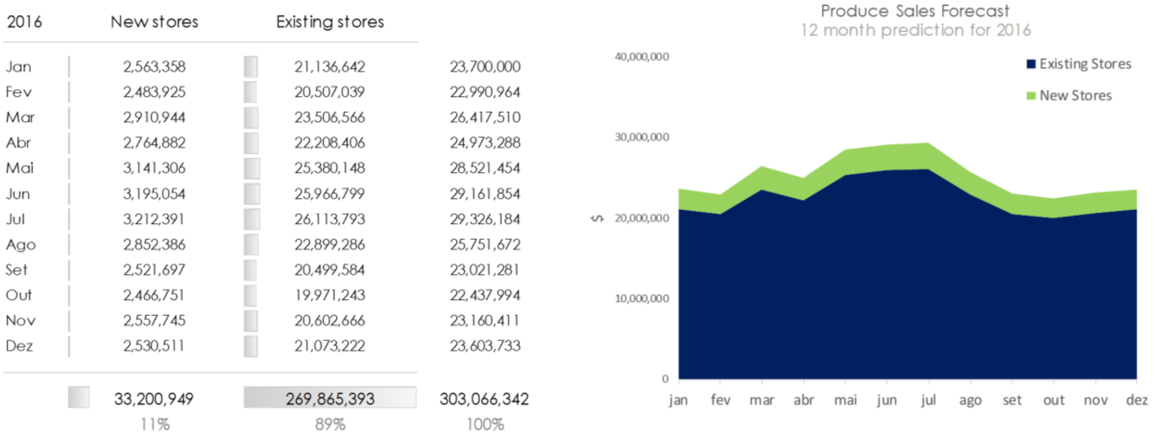

Time Series Plot
2.50e+7
2.00e+7
Jul 2012 Jan 2013 Jul 2013 Jan 2014 Jul 2014 Jan 2015 Jul 2015
This is a time series plot
Seasonplot
2012
2013
2014
2015
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Decomposition Plot
data
2.50e+7
2.00e+7
seasonal
0
Seasonal
m
trend
2.40e+7
2.20e+7
Trend
n
remainder
0
Error
m
Jul 2012 Jan 2013 Jul 2013 Jan 2014 Jul 2014 Jan 2015 Jul 2015
ACF
Lag
This is an autocorrelation plot
PACF
Lag
This is an partial autocorrelation plot

Image 14, TS plot output

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.

Based on Time Series analysis I calculated the 2016 Produce Sales forecast for existing stores and for new stores which reached 303 million with a confidence interval of 95%.

303 mi



2016	New stores	Existing stores	
Jan	2,563,358	21,136,642	23,700,000
Fev	2,483,925	20,507,039	22,990,964
Mar	2,910,944	23,506,566	26,417,510
Abr	2,764,882	22,208,406	24,973,288
Mai	3,141,306	25,380,148	28,521,454
Jun	3,195,054	25,966,799	29,161,854
Jul	3,212,391	26,113,793	29,326,184
Ago	2,852,386	22,899,286	25,751,672
Set	2,521,697	20,499,584	23,021,281
Out	2,466,751	19,971,243	22,437,994
Nov	2,557,745	20,602,666	23,160,411
Dez	2,530,511	21,073,222	23,603,733
	33,200,949	269,865,393	303,066,342
	11%	89%	100%

Image 15, 2016 Produce Sales breakdown developed in Excel

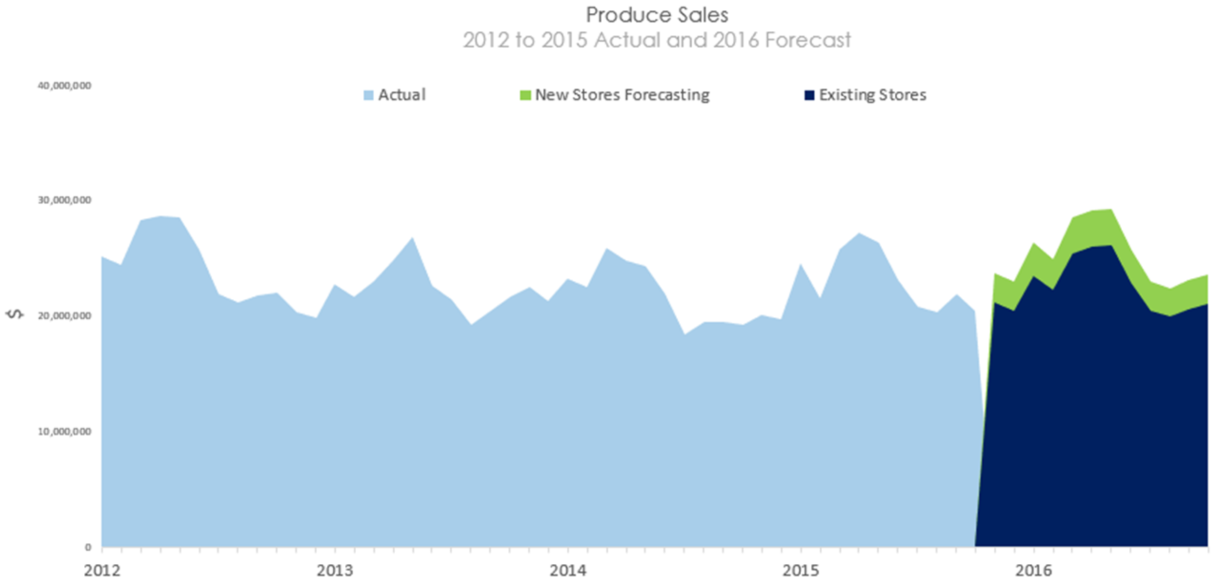
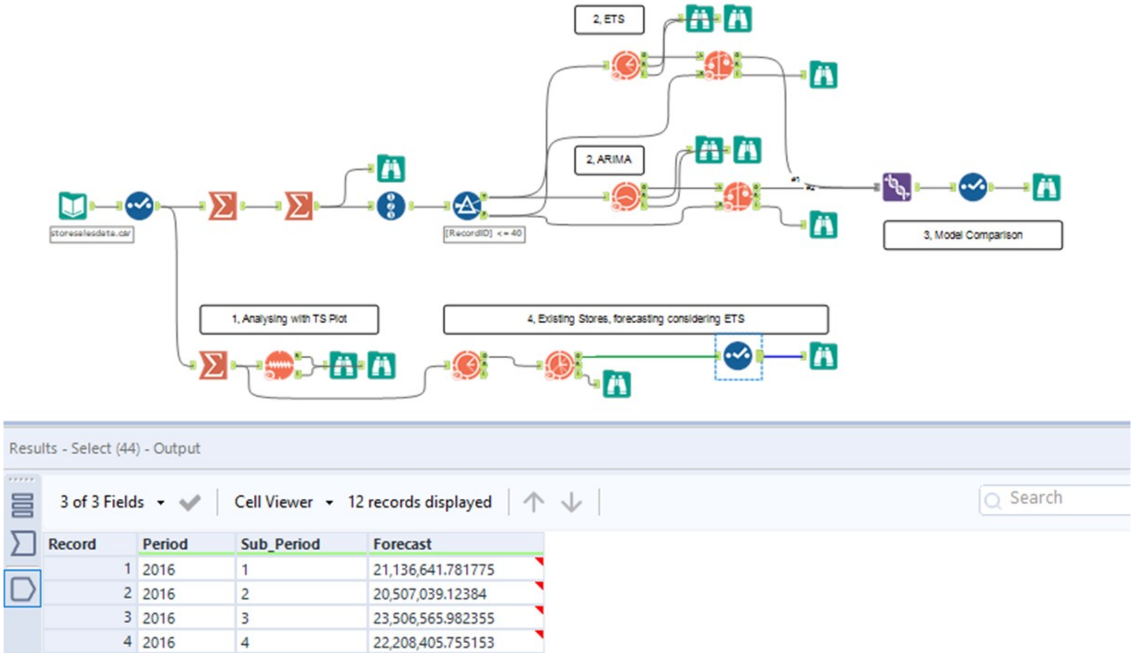


Image 16, Produce Sales historical data and forecasting for Existing and New Stores developed in Excel

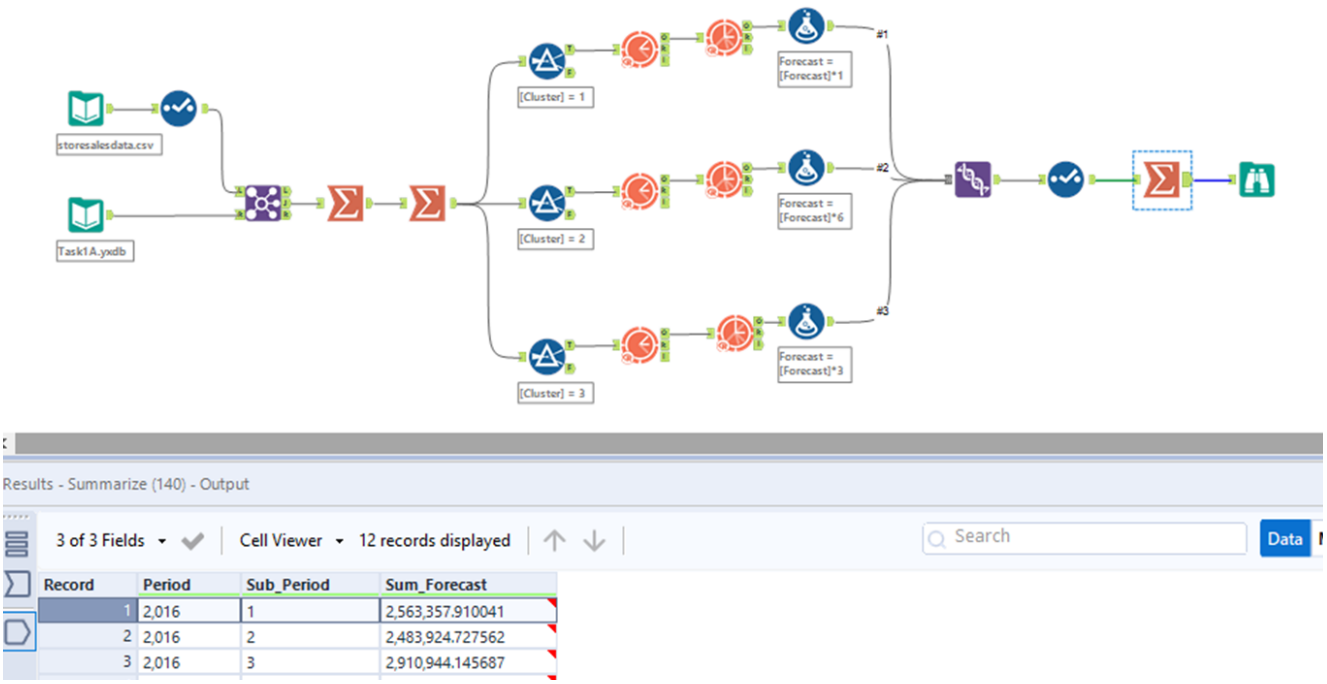
Following the first workflow I developed for task 3 resulting in existing stores Produce Sales forecasting.



Record	Period	Sub_Period	Forecast
1	2016	1	21,136,641.781775
2	2016	2	20,507,039.12384
3	2016	3	23,506,565.982355
4	2016	4	22,208,405.755153

Image 17, Existing stores Produce Sales forecasting

Following the workflow I developed resulting in new stores Produce Sales forecasting.



Results - Summarize (140) - Output

12 records displayed

Record	Period	Sub_Period	Sum_Forecast
1	2,016	1	2,563,357.910041
2	2,016	2	2,483,924.727562
3	2,016	3	2,910,944.145687

Image 18, New stores Produce Sales forecasting

Data has a better idea.