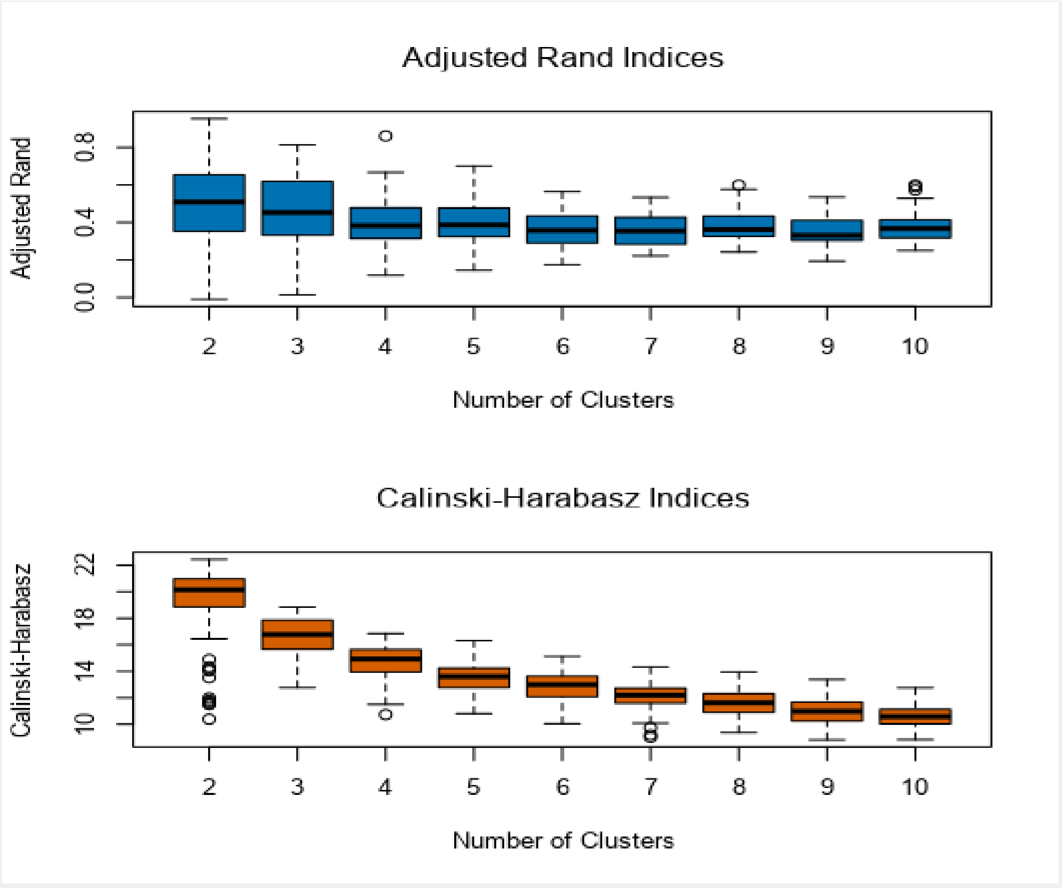
# Project: Determining the Format for the New Stores and Forecasting Produce Sales

**Summary**: In this multi-stage project, I first analyzed the optimal number of store formats for a grocery store chain and then developed a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store. In the last step I prepared a monthly forecast for produce sales for the full year for both existing and new stores.

## **Task 1:** Determining the Format

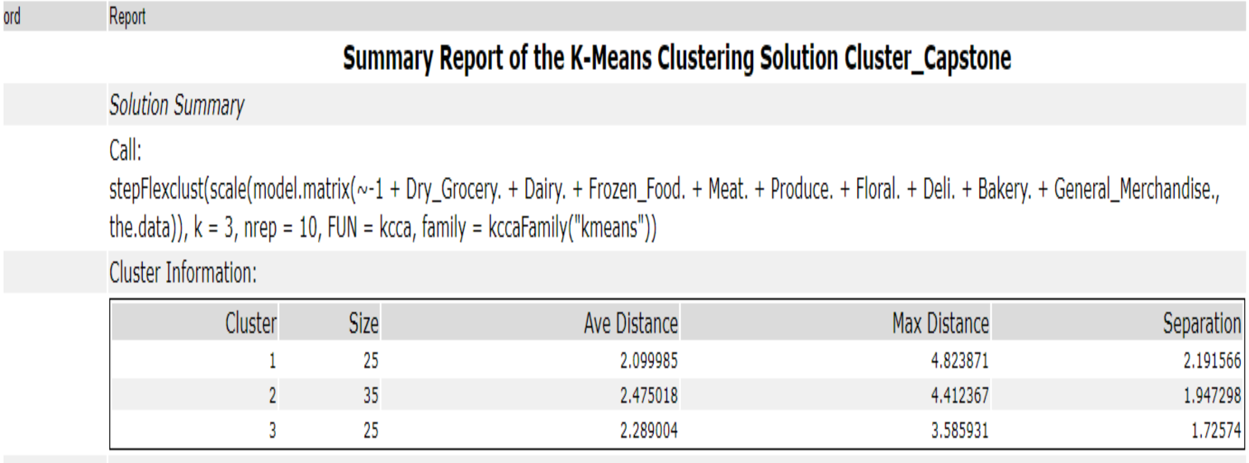
To determine the optimal number of store formats I evaluated the **K-Centroid Diagnostics** results. As can be seen from the plots below **Adjusted Rand Indices** and **Calinski-Harabasz Indices** indicate that 3 clusters will provide the best segmentation. Several factors are considered in these plots to determine the optimal number of clusters. The X axes in the plots indicate the number of clusters, higher median scores as well as compactness and less dispersion indicate better fit. Taken altogether the plots indicate number 3 as the optimal number of clusters.

*Plots*



After determining the number of clusters, based on my analysis I decided how many stores fall into each store format. Each cluster has stores as follows.

Cluster 1 = 25 Stores
Cluster 2 = 35 Stores
Cluster 3 = 25 Stores



| ord | Report |
|---|---|

**Summary Report of the K-Means Clustering Solution Cluster_Capstone**

*Solution Summary*

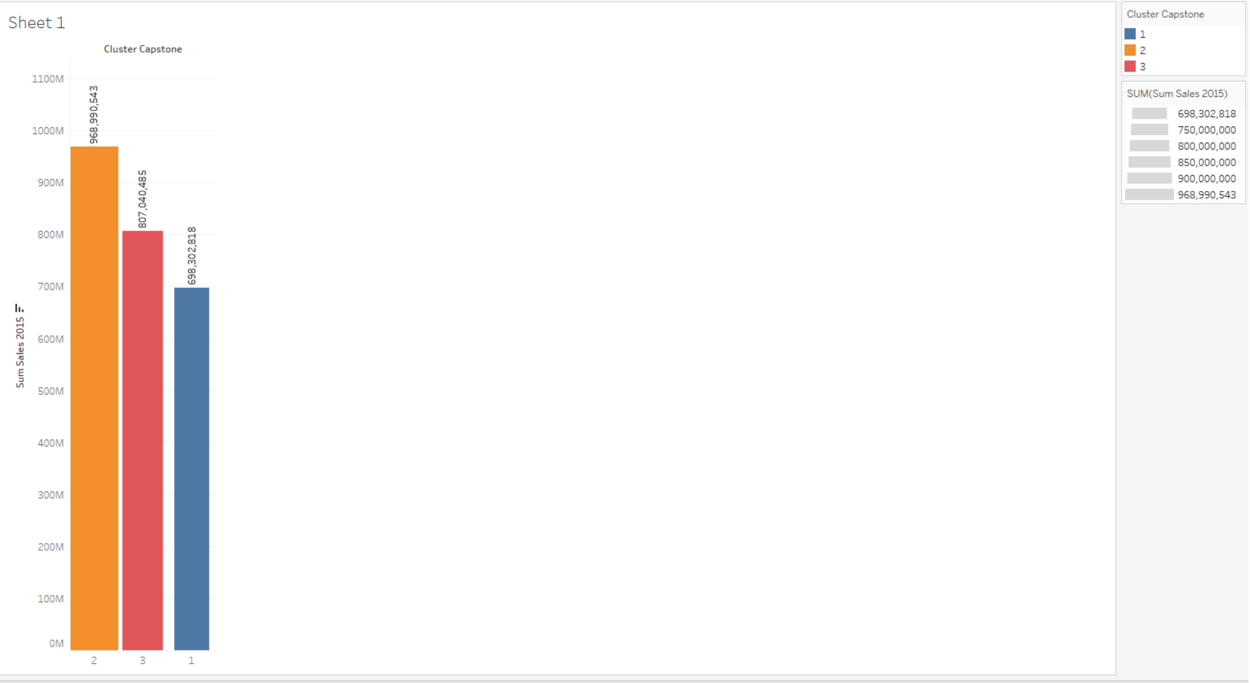
Call:

```
stepFlexclust(scale(model.matrix(~-1 + Dry_Grocery. + Dairy. + Frozen_Food. + Meat. + Produce. + Floral. + Deli. + Bakery. + General_Merchandise.,
the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))
```

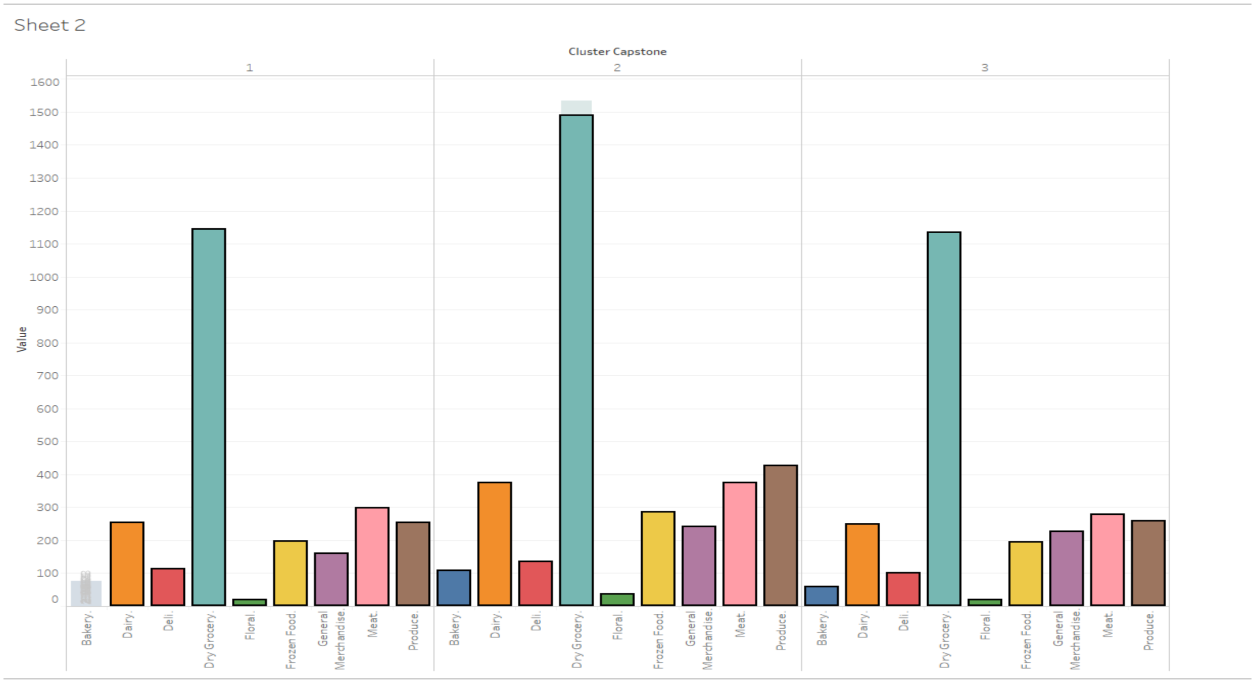
Cluster Information:

| 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 |

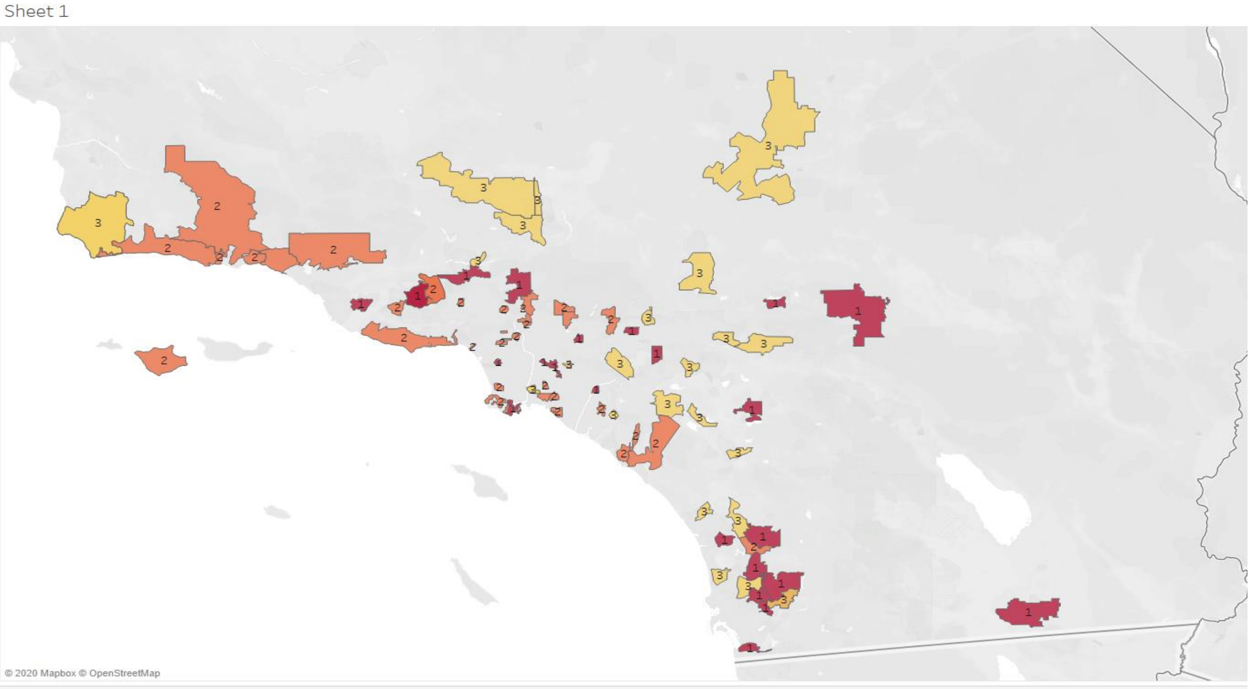
We can safely say that sales of the stores are considered in clustering process. Cluster 2 has the highest sales, then comes cluster 3 and then cluster 1. That makes sense as we used the sales numbers as our clustering values.



The tableau analysis indicated that; in terms of the sales by the **types of the products** there is not significant difference among the clusters. For example, as shown below, **dry grocery** is the highest selling item in all three clusters. Flora is the lowing selling product in all three clusters.

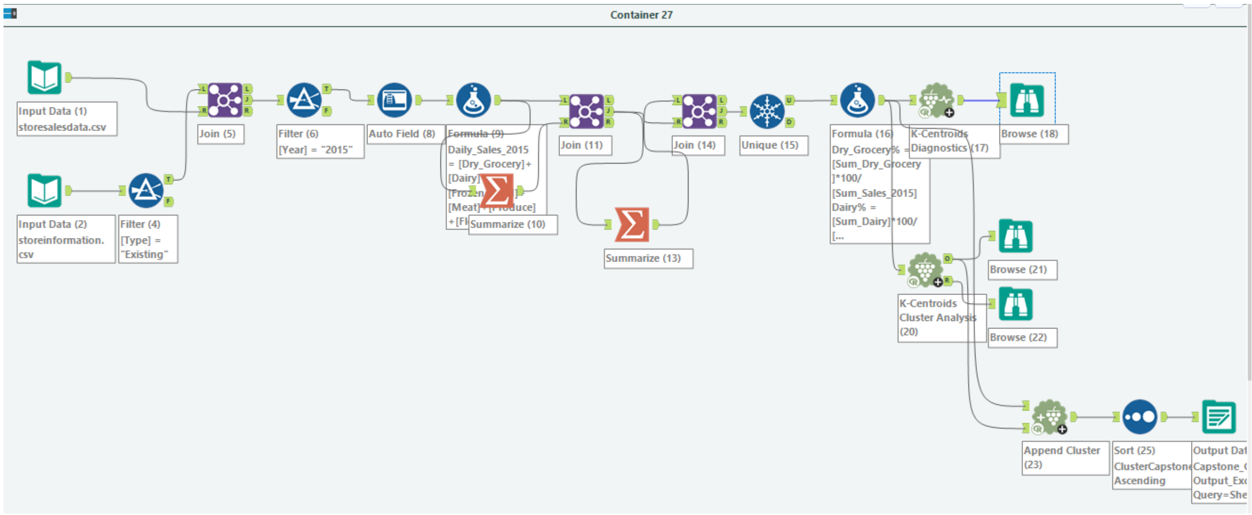

Sheet 2

The Tableau visualization below shows the location of the stores, uses color to show cluster, and size to show total sales.


Sheet 1

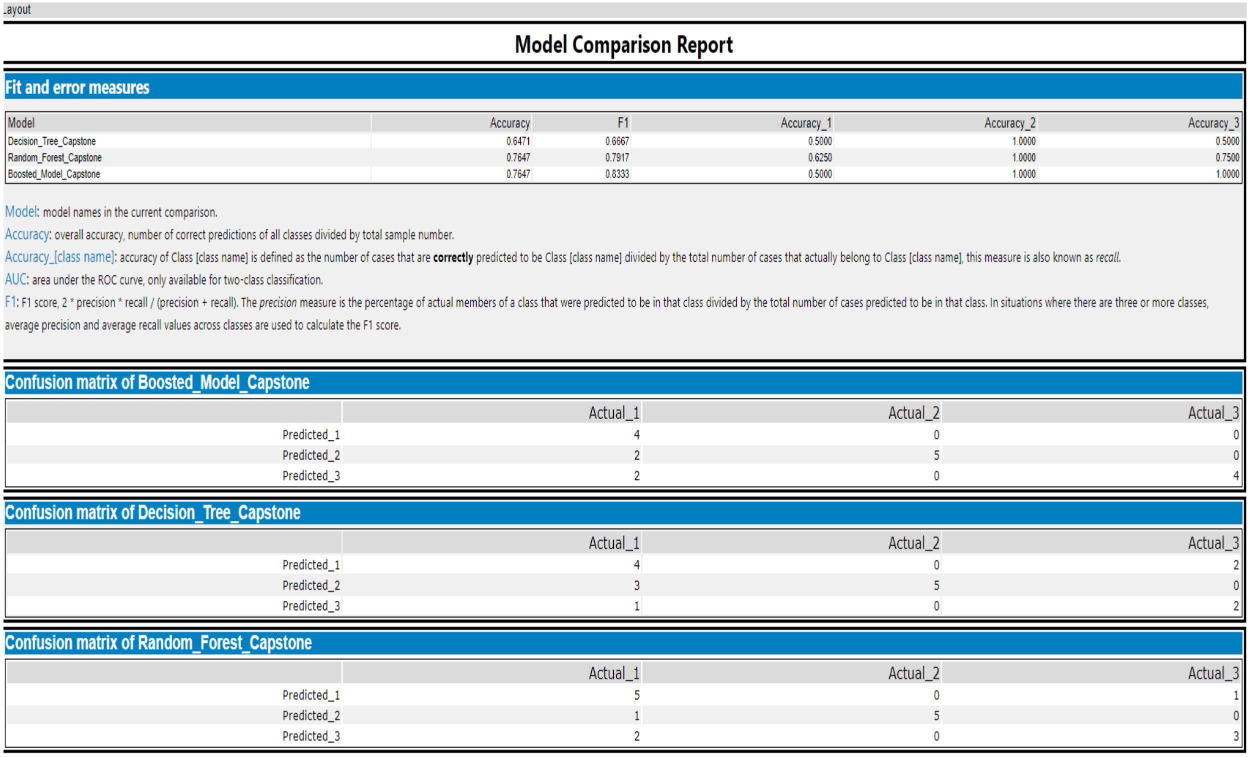
© 2020 Mapbox © OpenStreetMap

Below is the Alteryx workflow for Task 1



## **Task 2:** Formats for New Stores

To predict the best store format for the new stores, I used classification Model as we have a categorical target variable. And since our categorical target variable has more than 2 categories I used the multinomial models.


.ayout

### Model Comparison Report

#### Fit and error measures

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| Decision_Tree_Capstone | 0.6471 | 0.6667 | 0.5000 | 1.0000 | 0.5000 |
| Random_Forest_Capstone | 0.7647 | 0.7917 | 0.6250 | 1.0000 | 0.7500 |
| Boosted_Model_Capstone | 0.7647 | 0.8333 | 0.5000 | 1.0000 | 1.0000 |

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.

#### Confusion matrix of Boosted_Model_Capstone

| | 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_Capstone

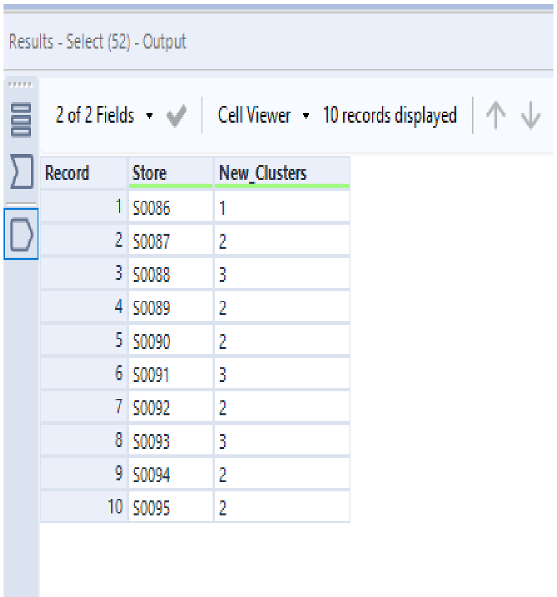
| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 4 | 0 | 2 |
| Predicted_2 | 3 | 5 | 0 |
| Predicted_3 | 1 | 0 | 2 |

#### Confusion matrix of Random_Forest_Capstone

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 5 | 0 | 1 |
| Predicted_2 | 1 | 5 | 0 |
| Predicted_3 | 2 | 0 | 3 |

Model comparison reports above indicates that the two highest accuracy scores belong to **Boosted Model** and **Random Forest Model** both being **76.47%.** I also looked at the **F1 score** which indicates the model precision. This score ranges from 0 to 1 and higher values indicates higher precision. As a result, I chose the boosted model as the best, among others.

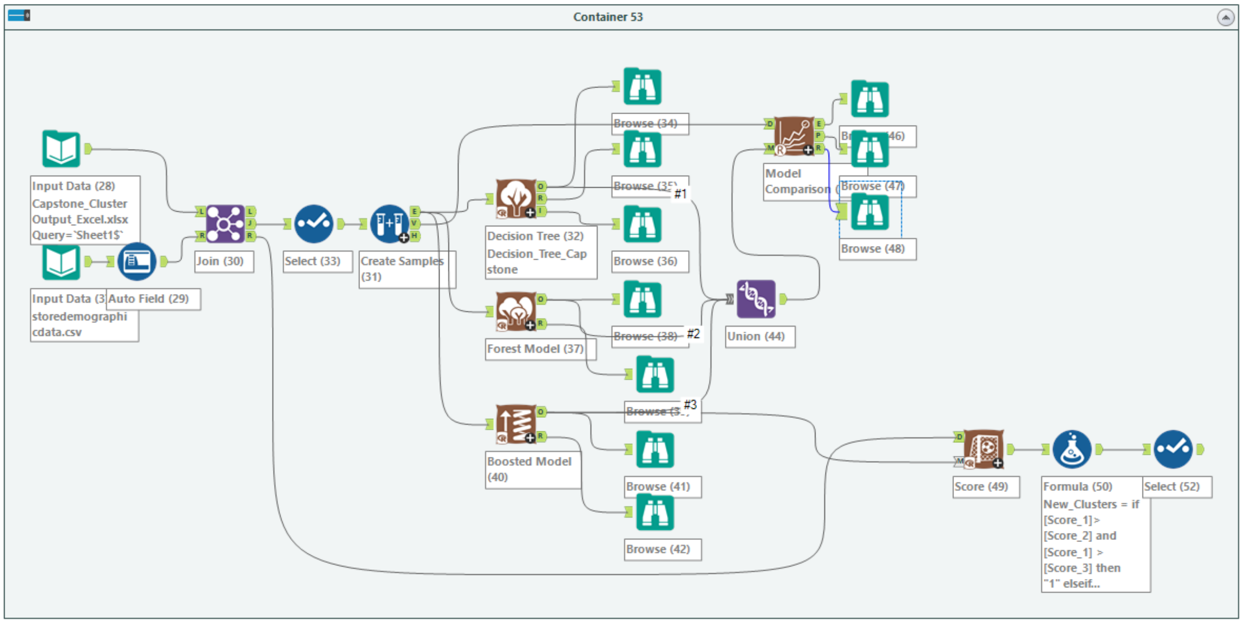
Then I applied the **Score tool** to get the scores for the new 10 stores and then I determined the clusters for the new stores using the values I obtained from the Score tool's analysis. The results below indicate what format each of the 10 new stores falls into.



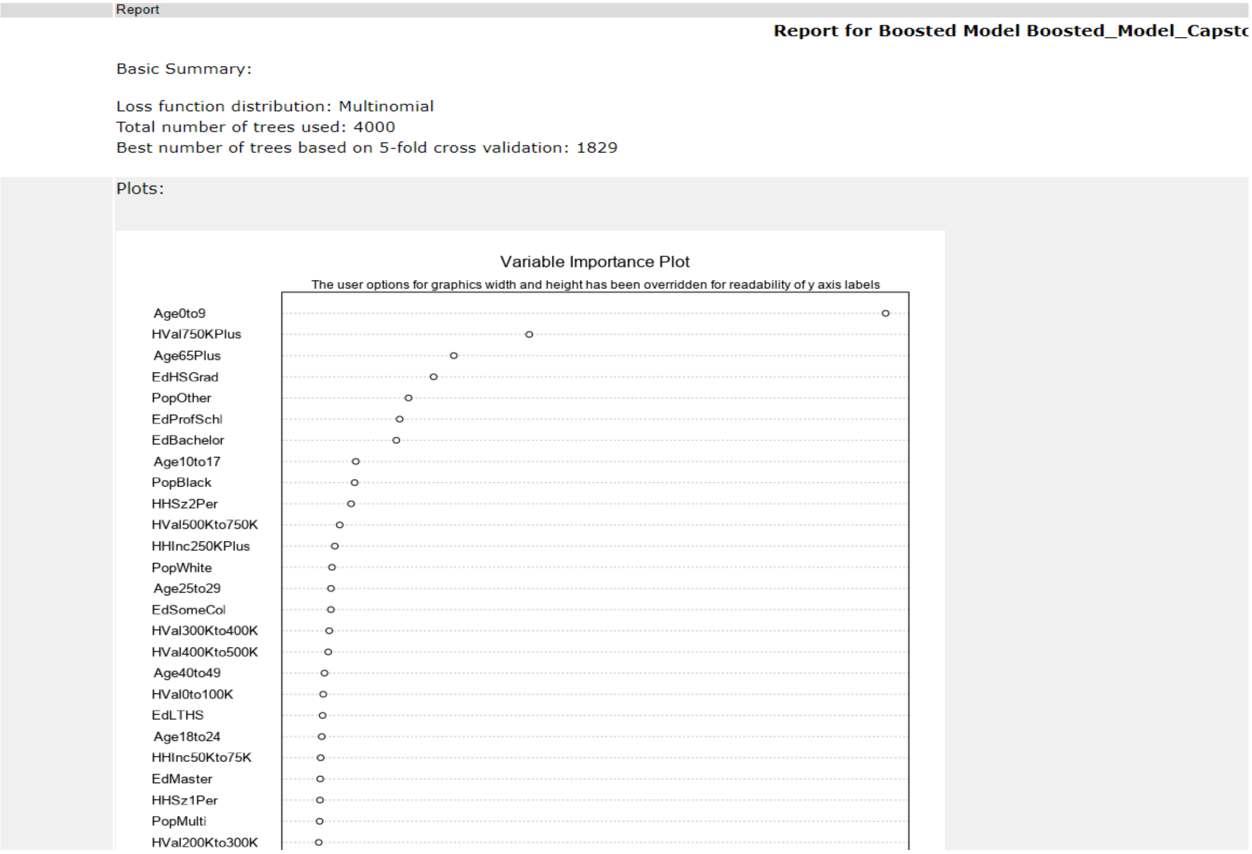
| Record | Store | New_Clusters |
|---|---|---|
| 1 | S0086 | 1 |
| 2 | S0087 | 2 |
| 3 | S0088 | 3 |
| 4 | S0089 | 2 |
| 5 | S0090 | 2 |
| 6 | S0091 | 3 |
| 7 | S0092 | 2 |
| 8 | S0093 | 3 |
| 9 | S0094 | 2 |
| 10 | S0095 | 2 |

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

Below is the Alteryx workflow for Task 2.



To identify the most important variables that help explain the relationship between demographic indicators and store formats I evaluated the **Variable Importance Plot below**. Accordingly, the most important three variables are “*Age 0to9*”, “*HVal750KPlus*”, “*Age65Plus*”



Report

Report for Boosted Model Boosted_Model_Capstc

Basic Summary:

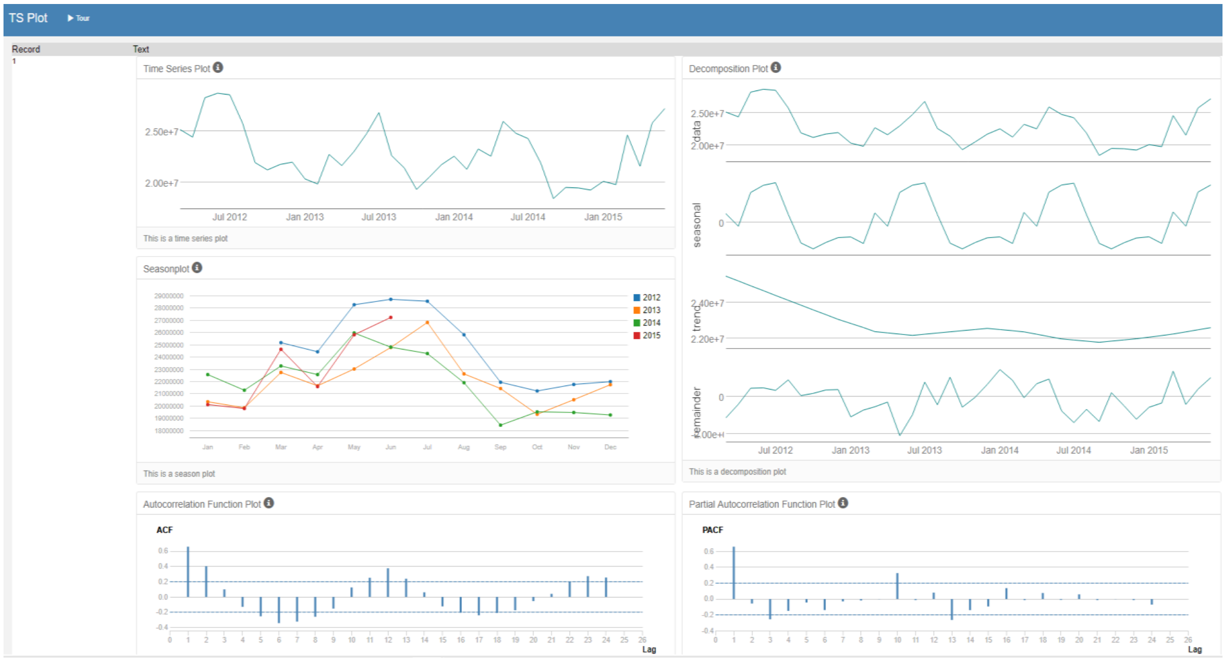
Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829

Plots:

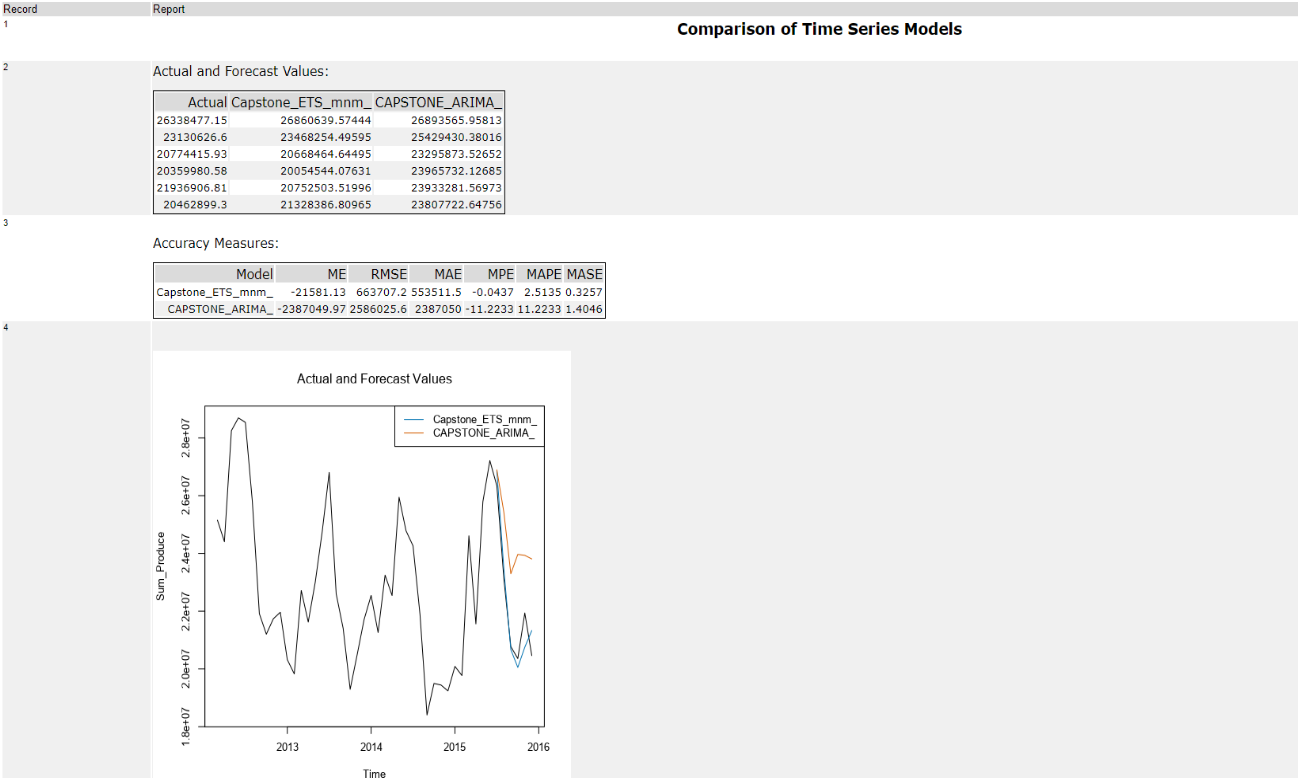
## **Task 3:** Predicting Produce Sales

To accomplish the Task 3, I first determine what type of ETS or ARIMA model I needed to use for each forecast.

For time series analysis, after initial data investigation, the first business of order is to check the **decomposition plots** below. After evaluating those plots above I concluded that errors fluctuate in magnitude indicating a need for multiplicative method. There does not seem to be a meaningful trend indicating neither of the methods is needed. As for seasonality, there is seasonality which seems to increase over time indicating the need for multiplicative method. As a result, I decided to use ETS (m, n, m) as the best model.



I decided to use ETS models for forecasting rather than ARIMA because when I tested the models ETS(m, n, m) model versus ARIMA, I saw that its relevant values were much better with a better forecasting performance than ARIMA indicating a better fit. Therefore, I used ETS (m, n, m) for forecasting the produce sales throughout 2016. Please refer to the model comparison tool’s results below.



Record Report

1

**Comparison of Time Series Models**

2

Actual and Forecast Values:

| Actual | Capstone_ETS_mnm_ | CAPSTONE_ARIMA_ |
|---|---|---|
| 26338477.15 | 26860639.57444 | 26893565.95813 |
| 23130626.6 | 23468254.49595 | 25429430.38016 |
| 20774415.93 | 20668464.64495 | 23295873.52652 |
| 20359980.58 | 20054544.07631 | 23965732.12685 |
| 21936906.81 | 20752503.51996 | 23933281.56973 |
| 20462899.3 | 21328386.80965 | 23807722.64756 |

3

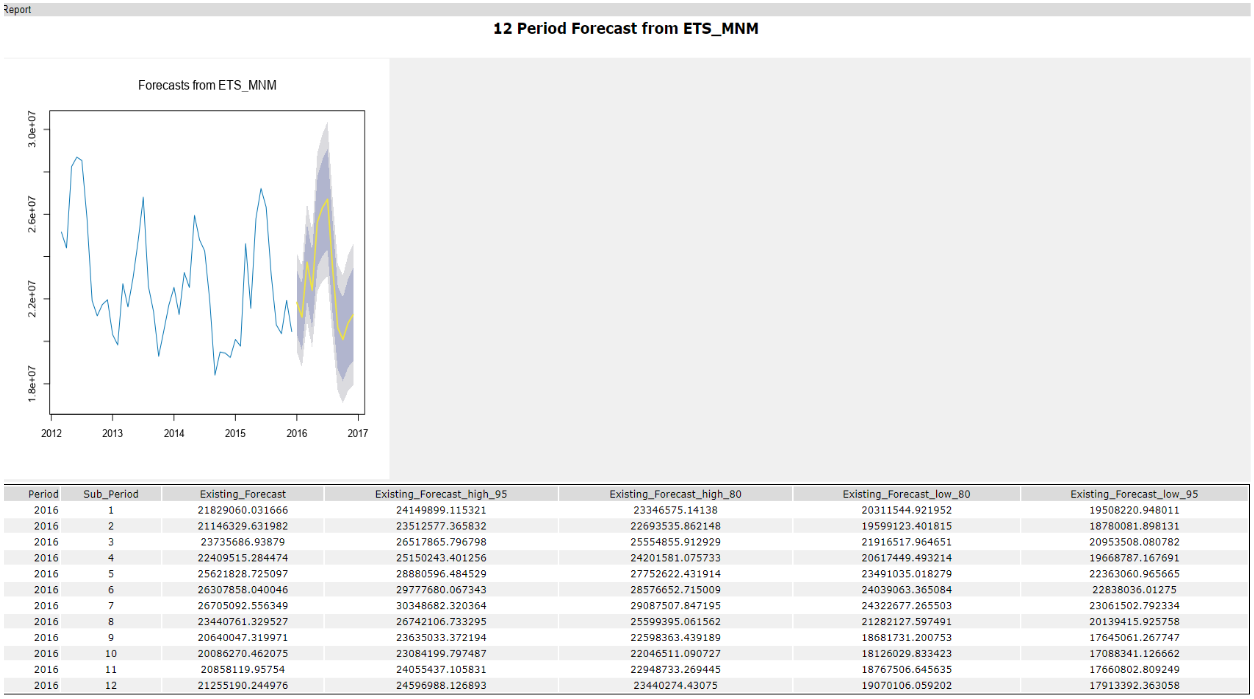
Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| Capstone_ETS_mnm_ | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |
| CAPSTONE_ARIMA_ | -2387049.97 | 2586025.6 | 2387050 | -11.2233 | 11.2233 | 1.4046 |

4

Then I created a table of my forecasts for existing and new stores. I also provided visualization the forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Below are the Forecasting results from Alteryx and the table indicating the existing and new stores' forecasting results. At the bottom you see the Tableau visualization of my forecasts that include historical data, existing stores forecasts, and new stores forecasts.

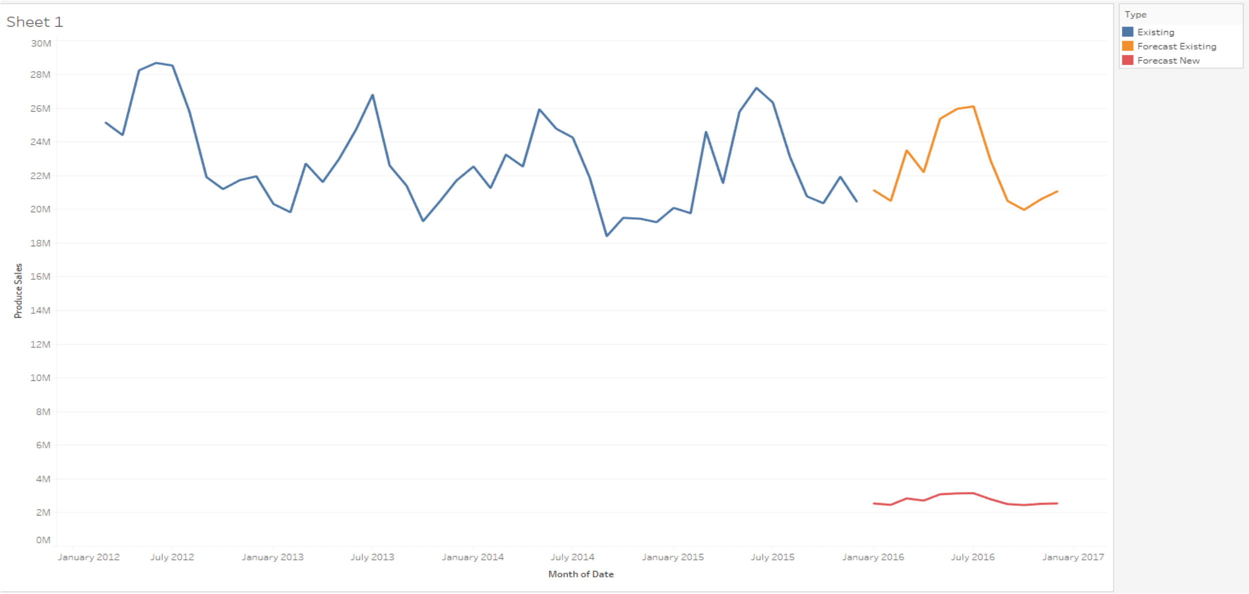


**12 Period Forecast from ETS_MNM**

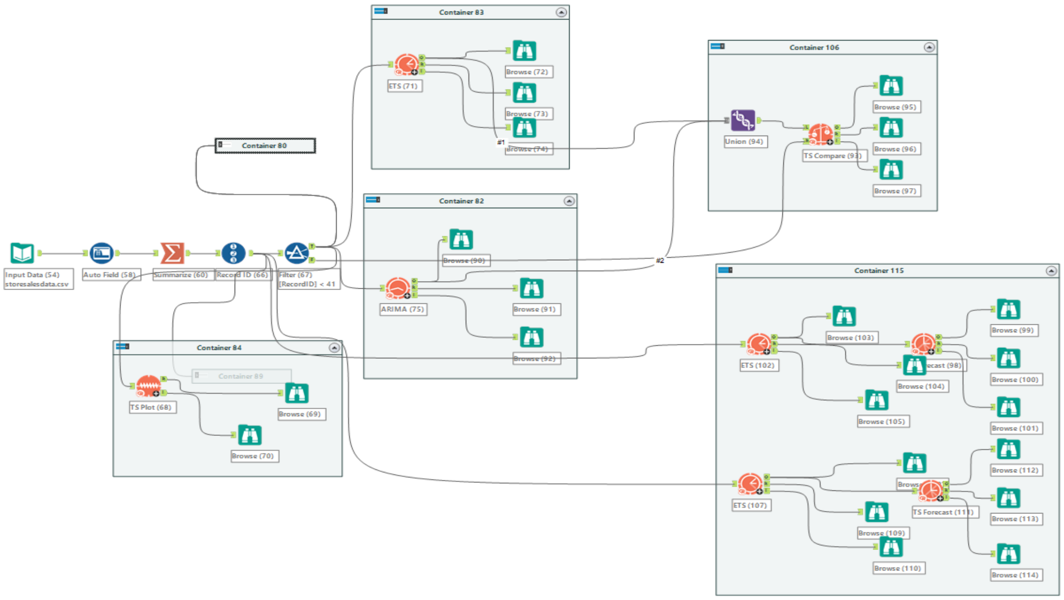
| Period | Sub_Period | Existing_Forecast | Existing_Forecast_high_95 | Existing_Forecast_high_80 | Existing_Forecast_low_80 | Existing_Forecast_low_95 |
|---|---|---|---|---|---|---|
| 2016 | 1 | 21829060.031666 | 24149899.115321 | 23346575.14138 | 20311544.921952 | 19508220.948011 |
| 2016 | 2 | 21146329.631982 | 23512577.365832 | 22693535.862148 | 19599123.401815 | 18780081.898131 |
| 2016 | 3 | 23735686.93879 | 26517865.796798 | 25554855.912929 | 21916517.964651 | 20953508.080782 |
| 2016 | 4 | 22409515.284474 | 25150243.401256 | 24201581.075733 | 20617449.493214 | 19668787.167691 |
| 2016 | 5 | 25621828.725097 | 28880596.484529 | 27752622.431914 | 23491035.018279 | 22363060.965665 |
| 2016 | 6 | 26307858.040046 | 29777680.067343 | 28576652.715009 | 24039063.365084 | 22838036.01275 |
| 2016 | 7 | 26705092.556349 | 30348682.320364 | 29087507.847195 | 24322677.265503 | 23061502.792334 |
| 2016 | 8 | 23440761.329527 | 26742106.733295 | 25599395.061562 | 21282127.597491 | 20139415.925758 |
| 2016 | 9 | 20640047.319971 | 23635033.372194 | 22598363.439189 | 18681731.200753 | 17645061.267747 |
| 2016 | 10 | 20086270.462075 | 23084199.797487 | 22046511.090727 | 18126029.833423 | 17088341.126662 |
| 2016 | 11 | 20858119.95754 | 24055437.105831 | 22948733.269445 | 18767506.645635 | 17660802.809249 |
| 2016 | 12 | 21255190.244976 | 24596988.126893 | 23440274.43075 | 19070106.059202 | 17913392.363058 |

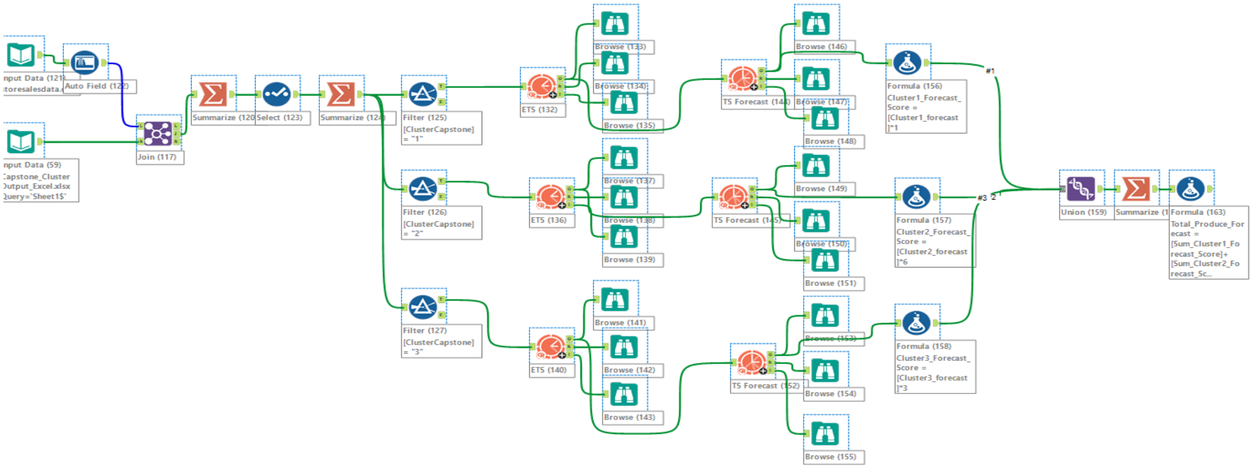
| Month | New Stores | Existing Stores |
|---|---|---|
| Jan-16 | 2563357.91004118 | 21829060.031666 |
| Feb-16 | 2483924.72756208 | 21146329.631982 |
| Mar-16 | 2910944.1456874 | 23735686.93879 |
| Apr-16 | 2764881.86969732 | 22409515.284474 |
| May-16 | 3141305.86730493 | 25621828.725097 |
| Jun-16 | 3195054.20380398 | 26307858.040046 |
| Jul-16 | 3212390.95408986 | 26705092.556349 |
| Aug-16 | 2852385.7691978 | 23440761.329527 |
| Sep-16 | 2521697.18679037 | 20640047.319971 |
| Oct-16 | 2466750.89369629 | 20086270.462075 |
| Nov-16 | 2557744.58771366 | 20858119.95754 |
| Dec-16 | 2530510.80513342 | 21255190.244976 |

Below is the visualzation of the existing and nes stores sales forecast.



Below are the Alteryx workflows for the whole analyses.





Kadir AKYUZ, Ph.D.