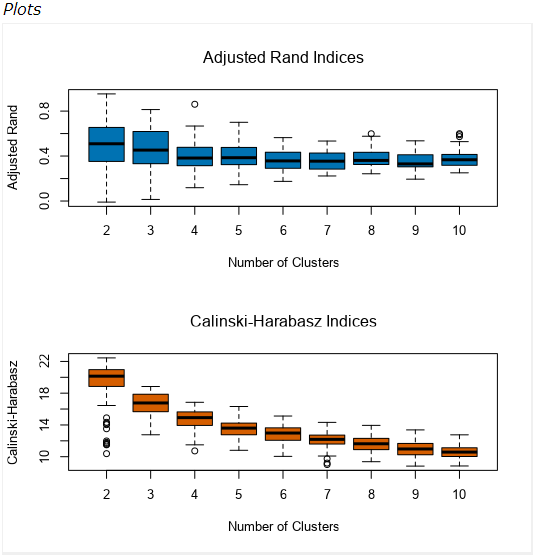
**Project: Determining the Format for the New Stores and Forecasting Produce Sales**

## **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.

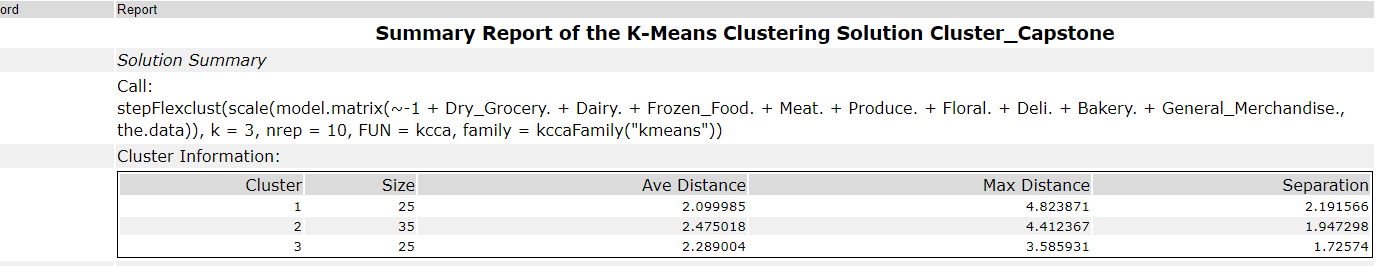


After determining the number of clusters, based on my analysis I decided how many stores fall into each store format. Each cluster has stores as follows.

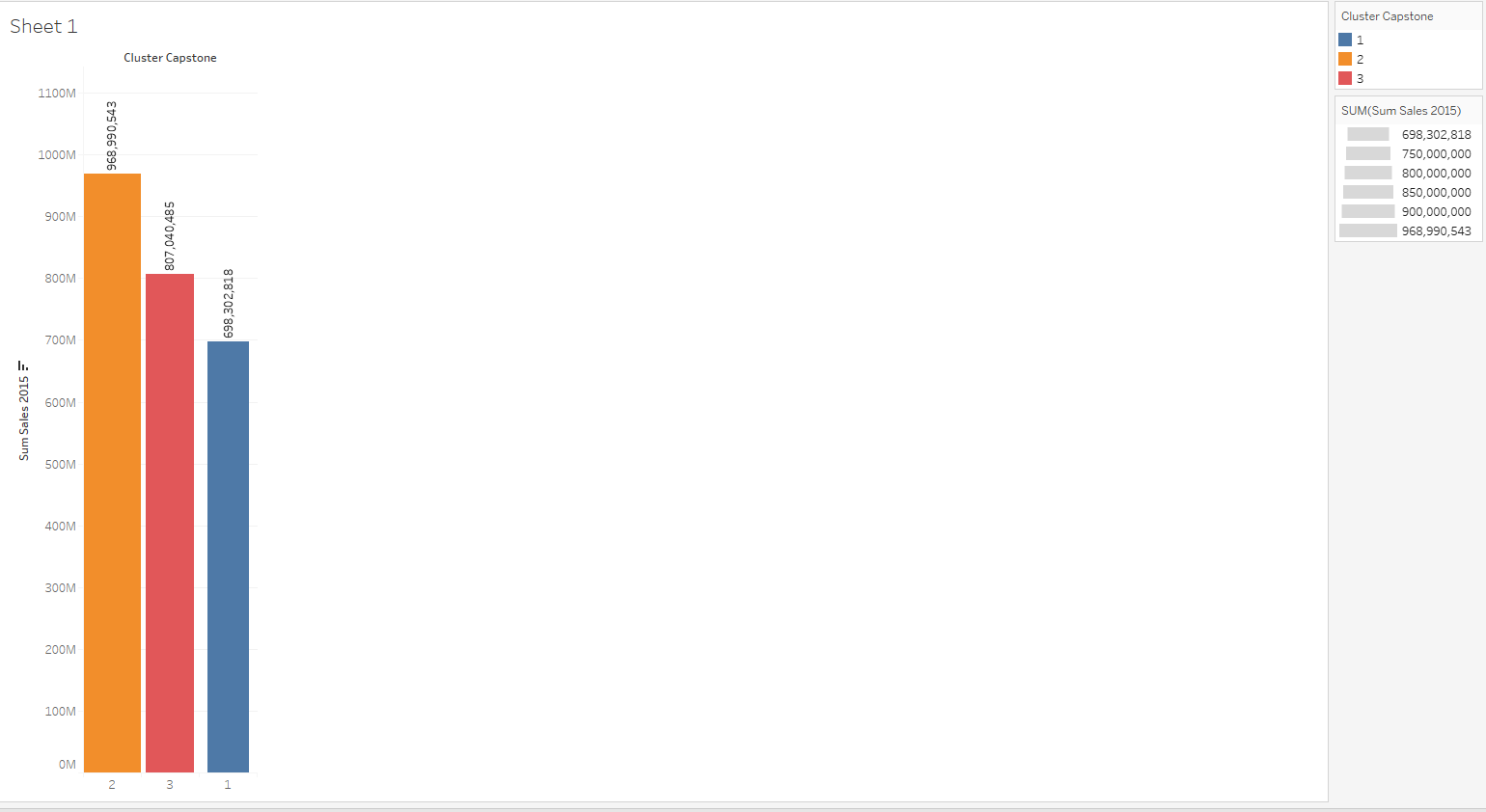
Cluster 1 = 25 Stores

Cluster 2 = 35 Stores

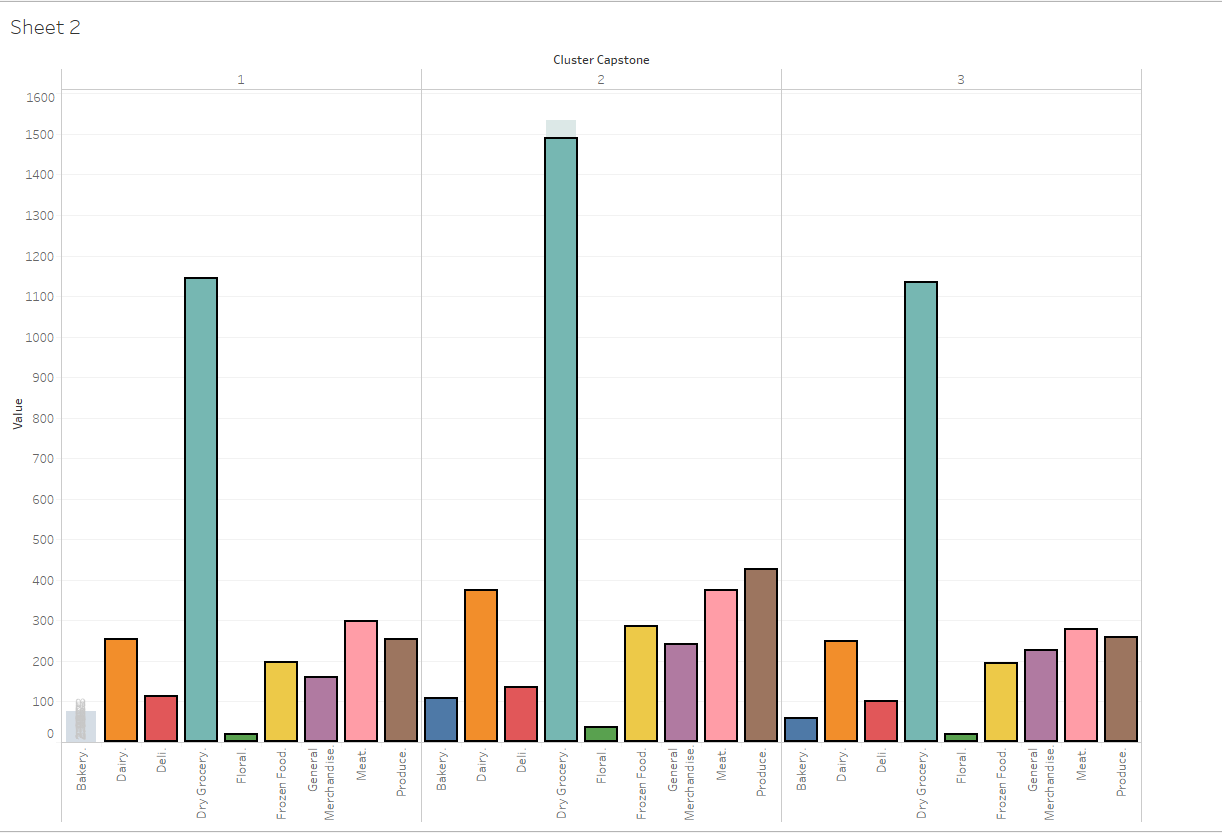
Cluster 3 = 25 Stores



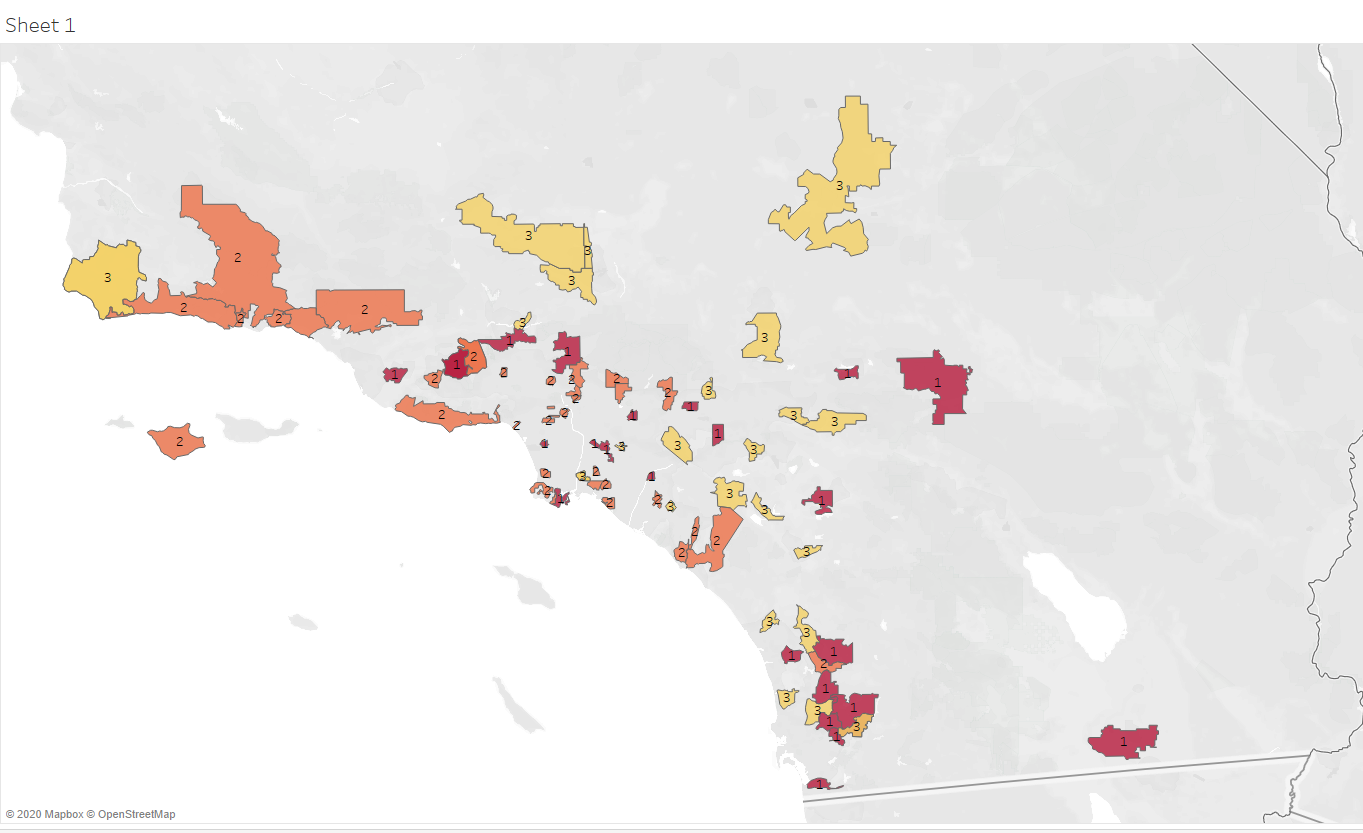
We can safely say that sales of the stores are considered in clustering process. Cluster 2 has the highest sales, then comes cluster 3 and then cluster 1. That makes sense as we used the sales numbers as our clustering values.

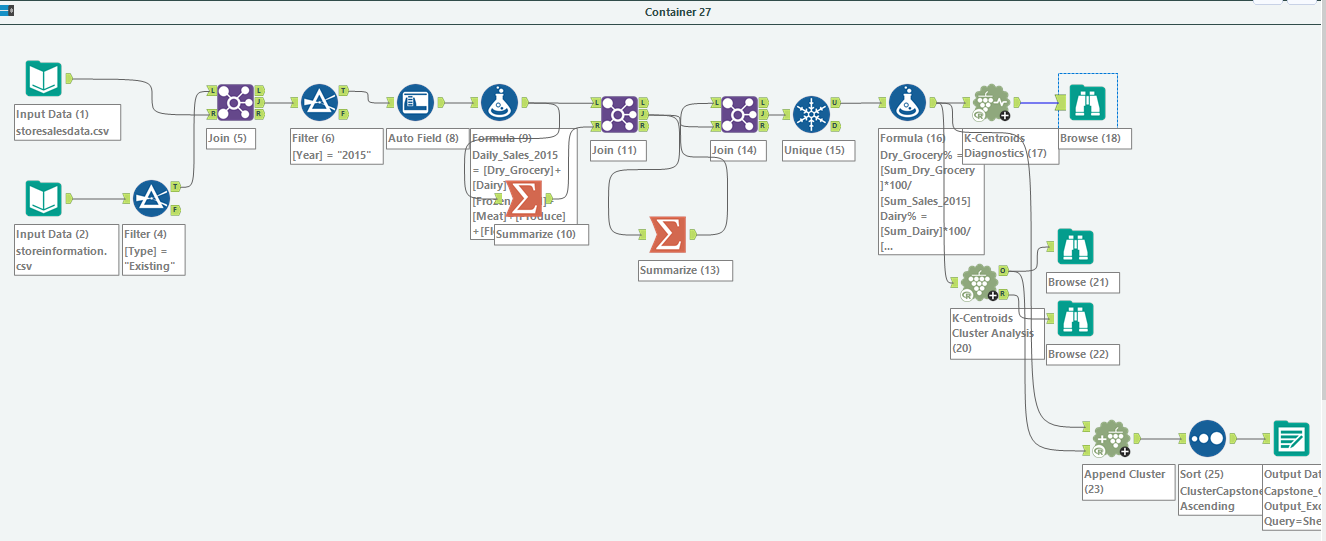


The tableau analysis indicated that; in terms of the sales by the **types of the products** there is not significant difference among the clusters. For example, as shown below, **dry grocery** is the highest selling item in all three clusters. Flora is the lowing selling product in all three clusters.



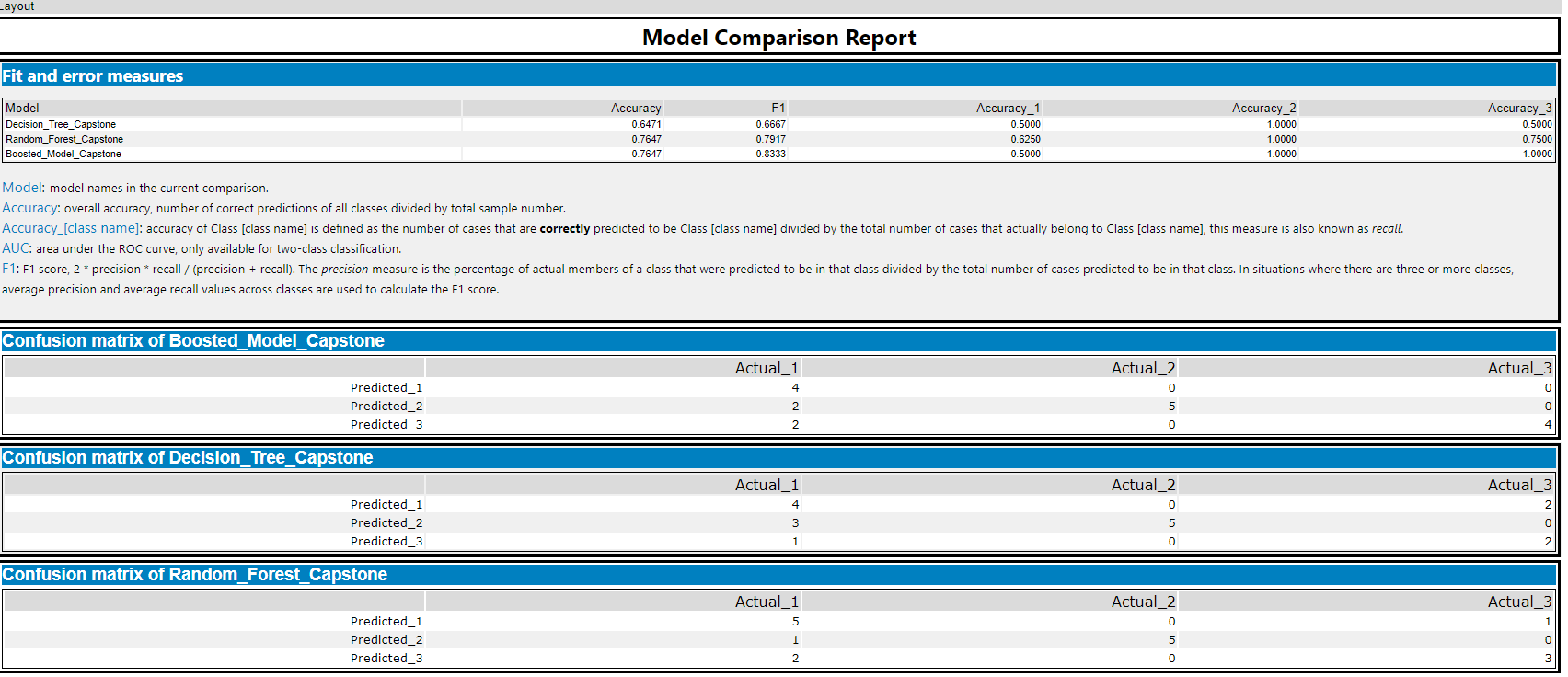
The Tableau visualization below shows the location of the stores, uses color to show cluster, and size to show total sales.

Below is the Alteryx workflow for Task 1



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

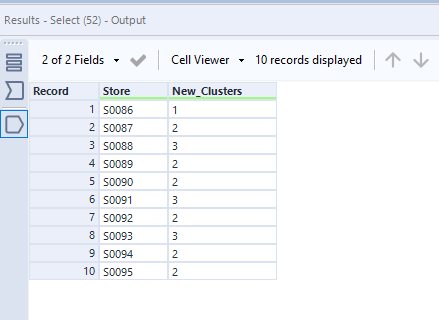
To predict the best store format for the new stores, I used classification Model as we have a categorical target variable. And since our categorical target variable has more than 2 categories I used the multinomial models.



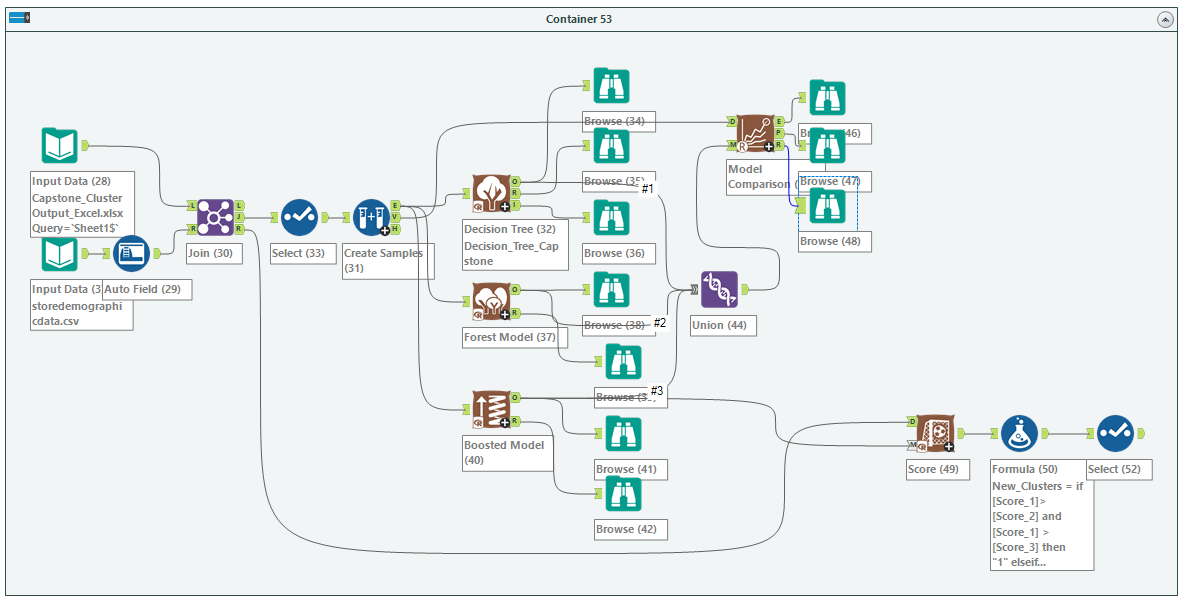
Model comparison reports above indicates that the two highest accuracy scores belong to **Boosted Model** and **Random Forest Model** both being **76.47%.** I also looked at the **F1 score** which indicates the model precision. This score ranges from 0 to 1 and higher values indicates higher precision. 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.

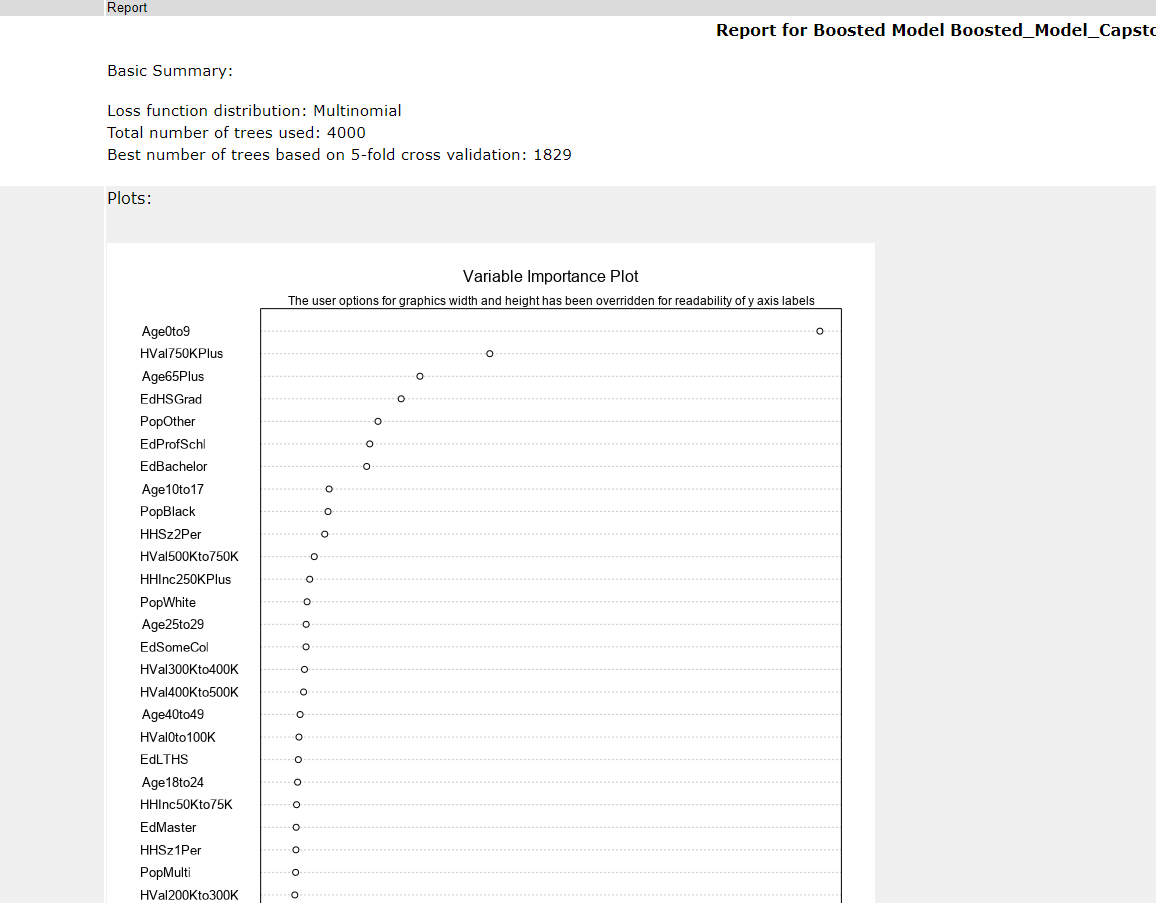
|  |  |
| --- | --- |
| Store Number | Segment |
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 3 |
| S0092 | 2 |
| S0093 | 3 |
| S0094 | 2 |
| S0095 | 2 |



Below is the Alteryx workflow for Task 2.



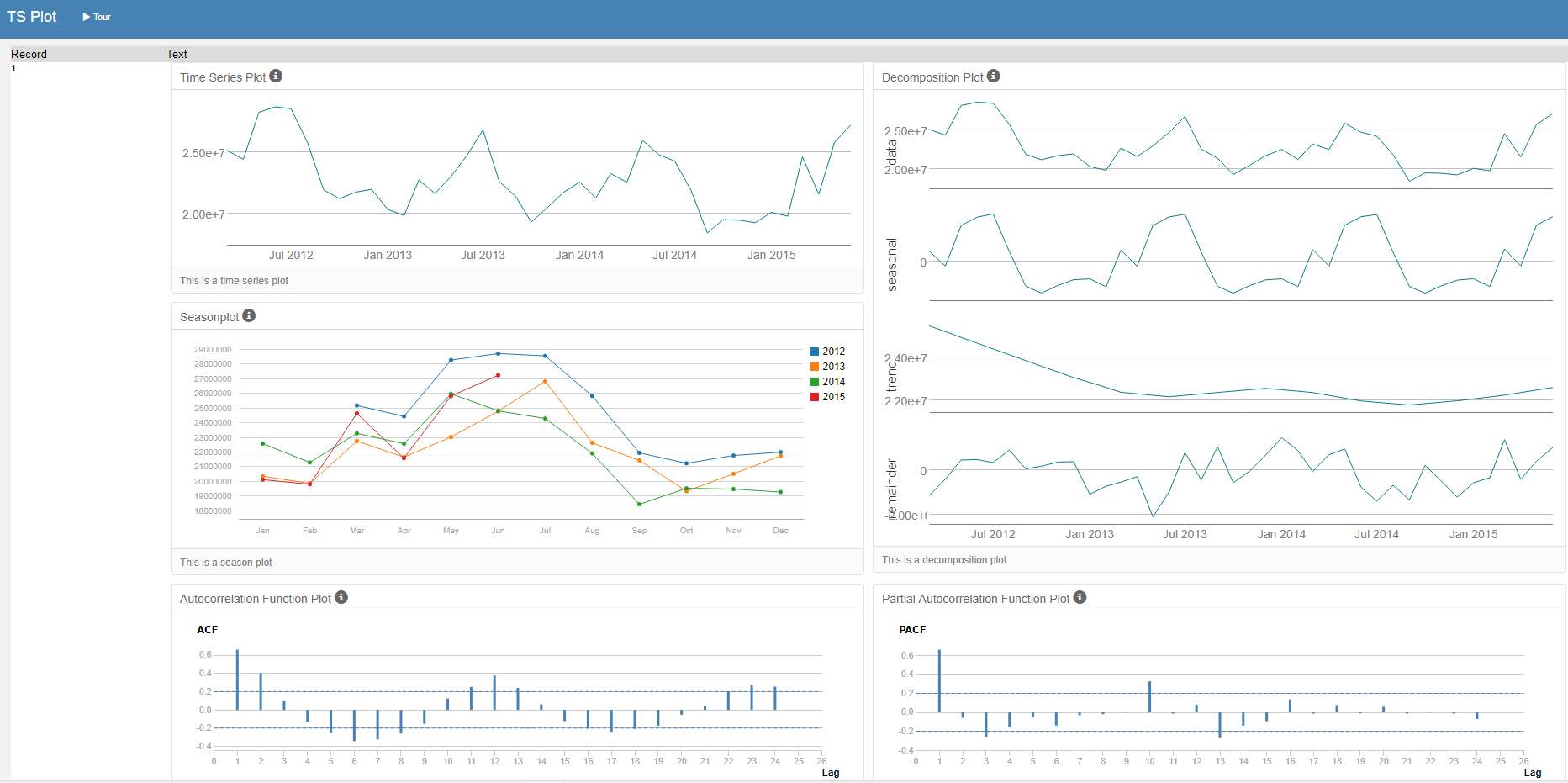
To identify the most important variables that help explain the relationship between demographic indicators and store formats I evaluated the **Variable Importance Plot below**. Accordingly, the most important three variables are “*Age 0to9*”, “*HVal750KPlus*”, “*Age65Plus*”



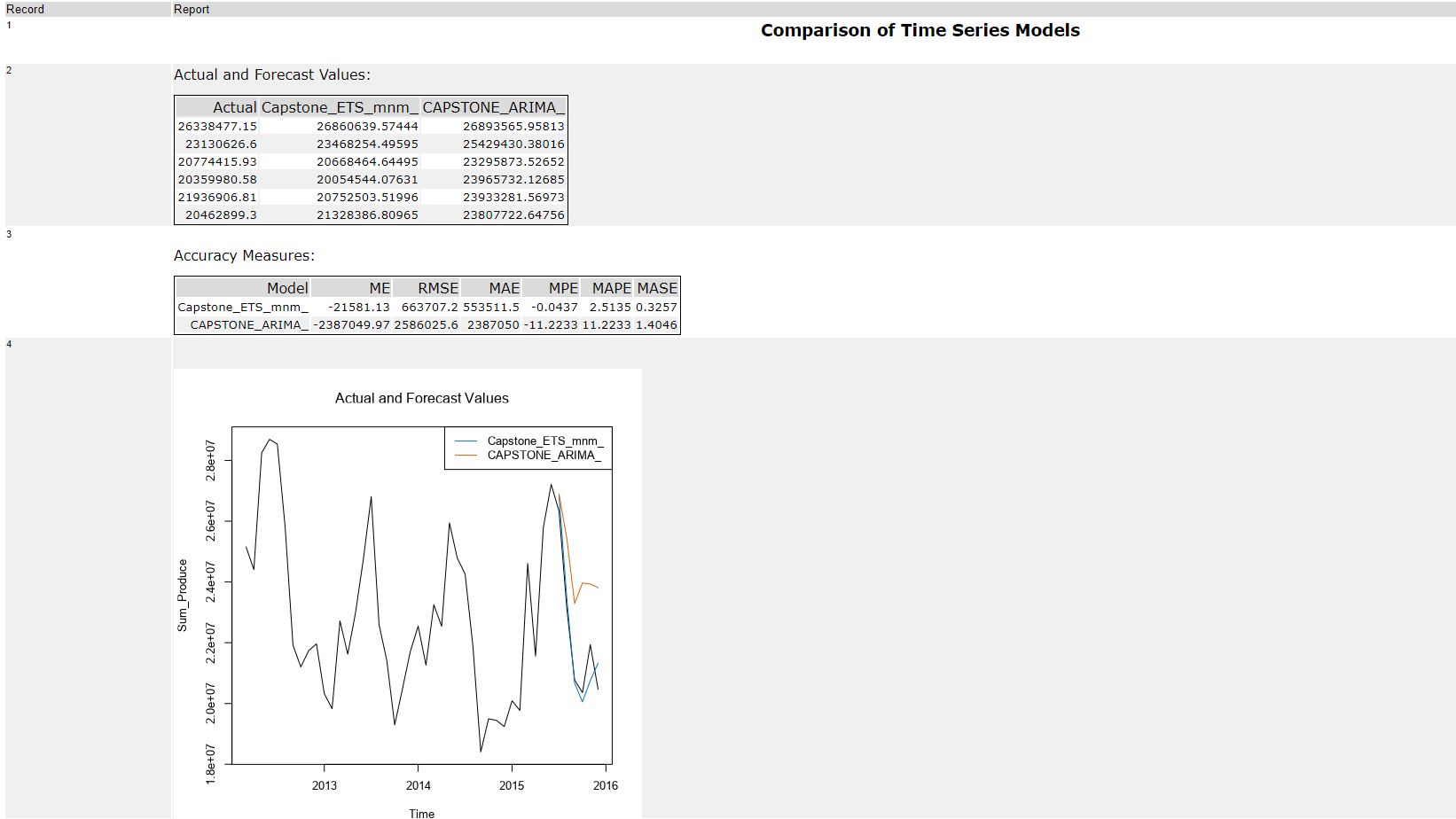
## **Task 3:** Predicting Produce Sales

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.

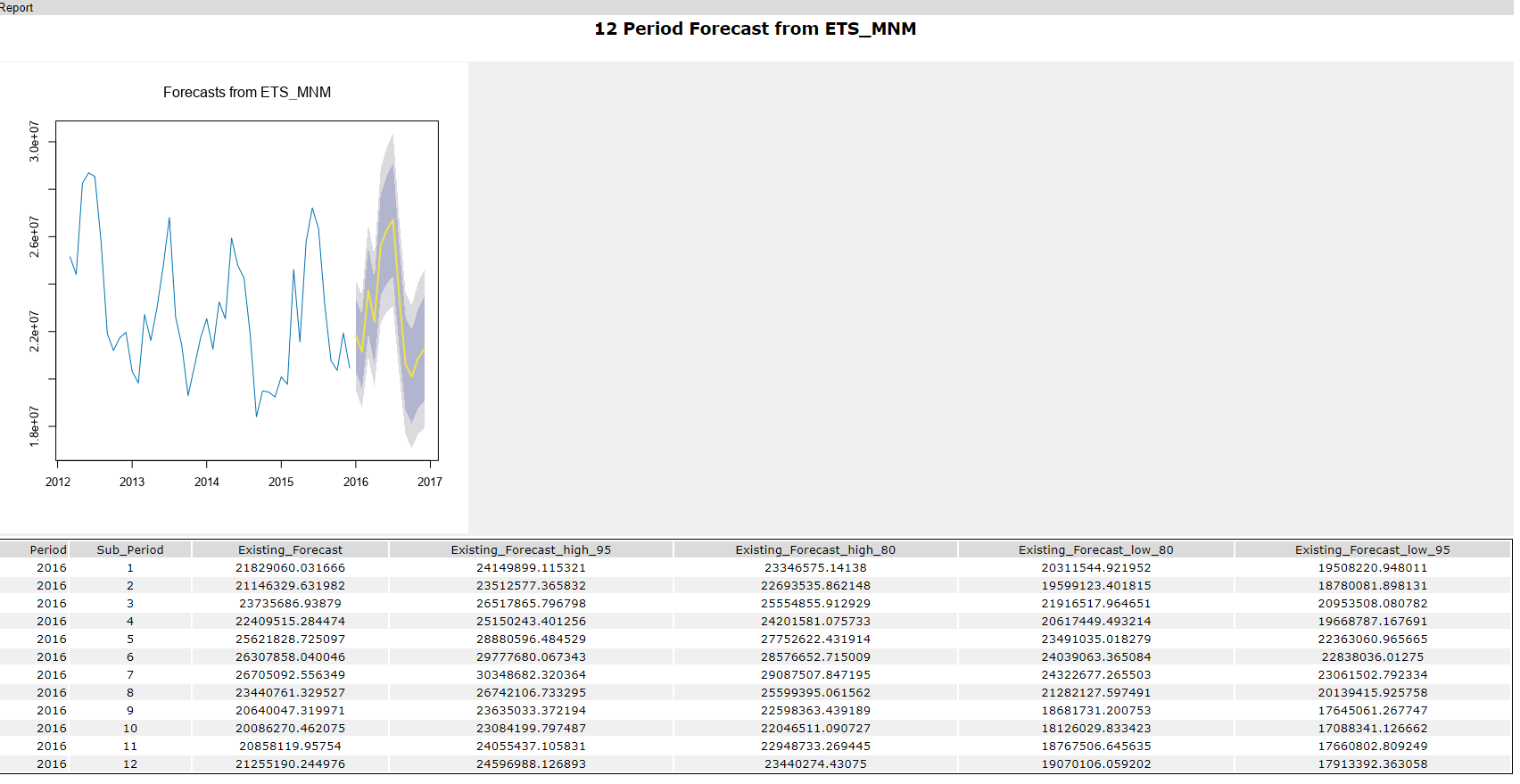


I decided to use ETS models for forecasting rather than ARIMA because when I tested the models ETS(m, n, m) model versus ARIMA, I saw that its relevant values were much better with a better forecasting performance than ARIMA indicating a better fit. Therefore, I used ETS (m, n, m) for forecasting the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



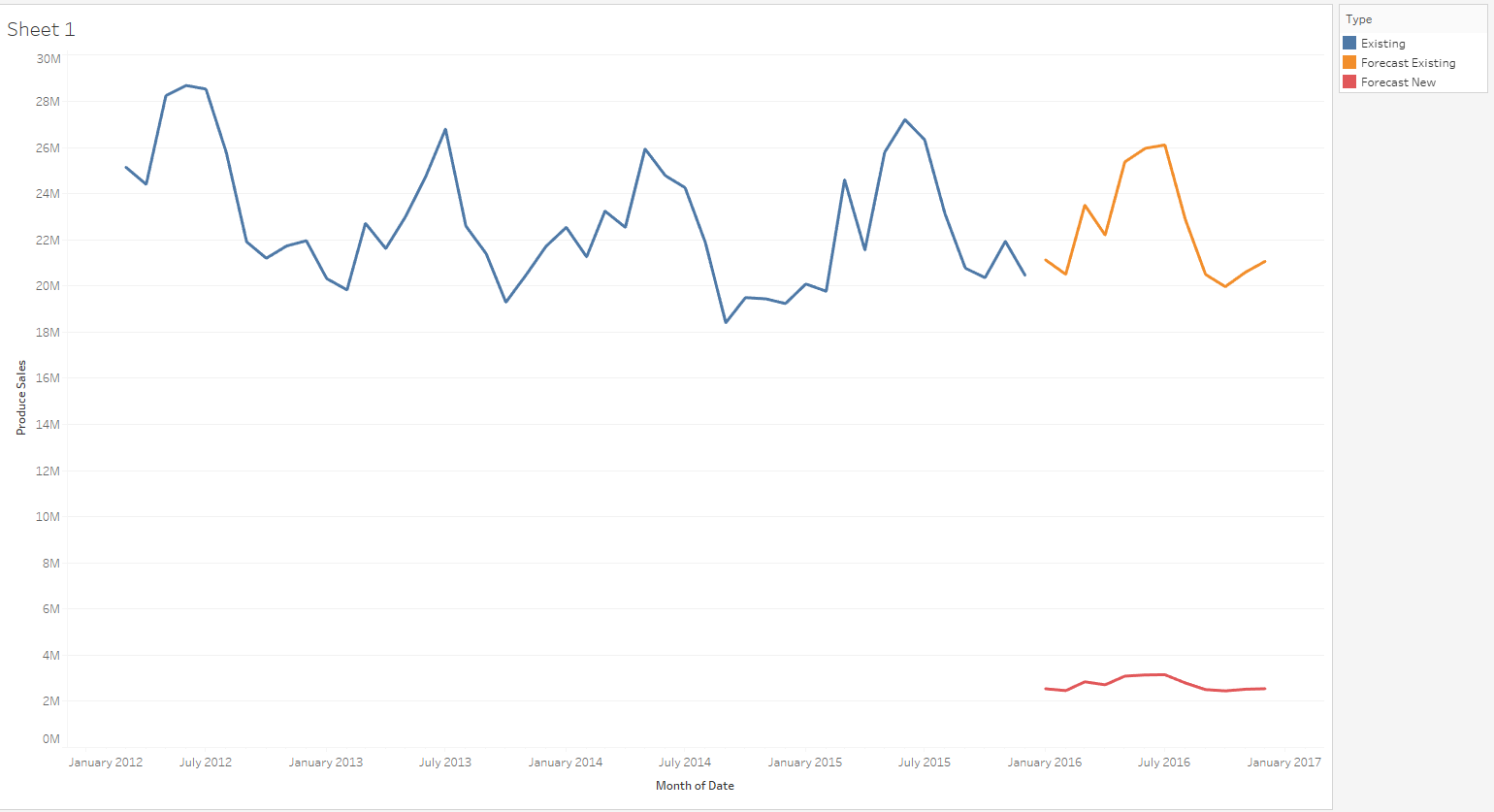
Then I created a table of my forecasts for existing and new stores. I also provided visualization the forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

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.

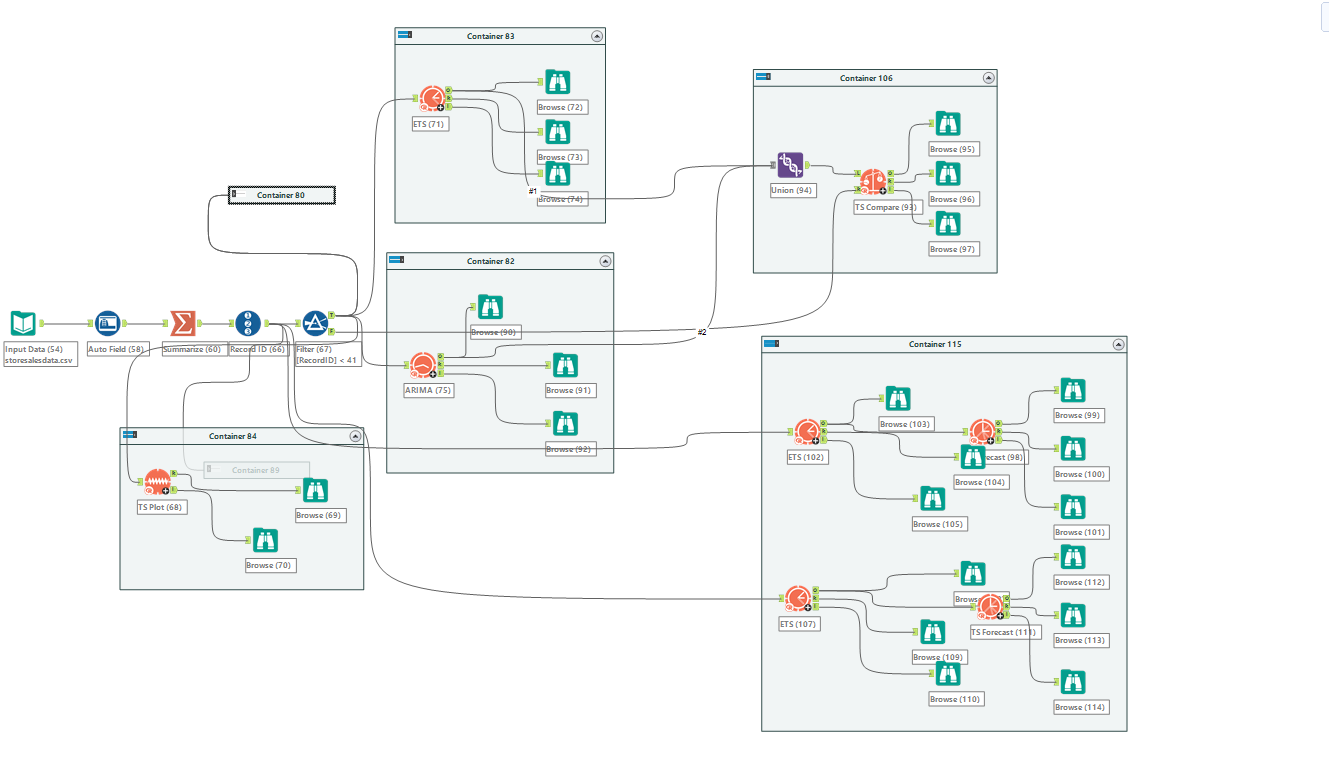


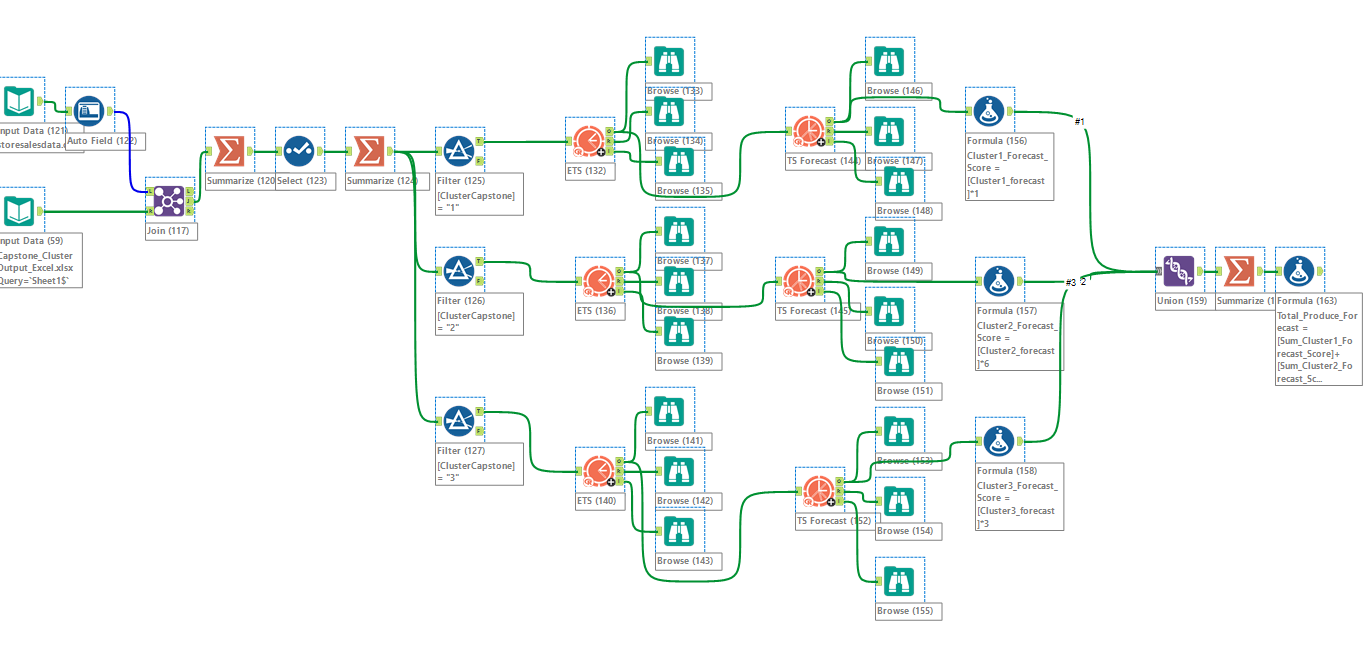
|  |  |  |
| --- | --- | --- |
| **Month** | **New Stores** | **Existing Stores** |
| Jan-16 | 2563357.91004118 | 21829060.031666 |
| Feb-16 | 2483924.72756208 | 21146329.631982 |
| Mar-16 | 2910944.1456874 | 23735686.93879 |
| Apr-16 | 2764881.86969732 | 22409515.284474 |
| May-16 | 3141305.86730493 | 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.