# Project: Predictive Analytics Capstone

## 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 of stores formats is 2 because the median from both indices Rand Adjusted and Calinski-Harabasz (CH) from K-means report show the highest value for 2 clusters against other clustering values

**K-Means Informe de evaluación de clúster**

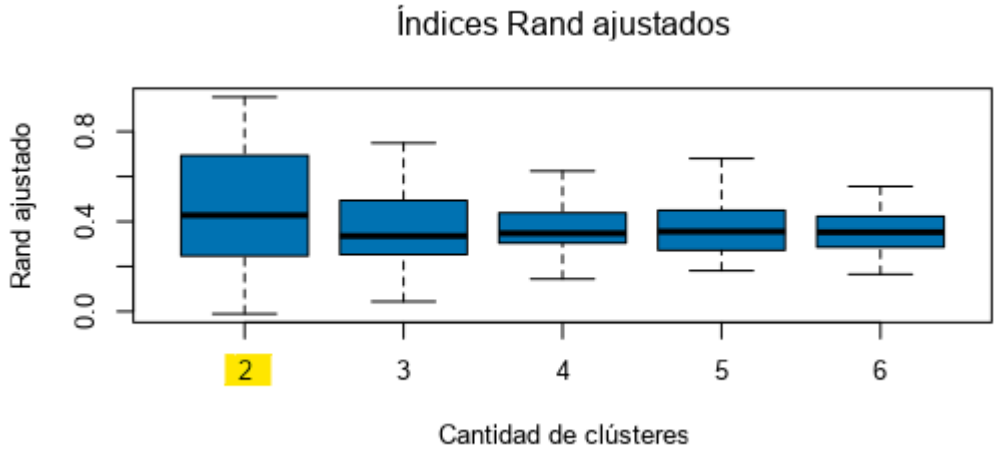
*Estadísticas de resumen*

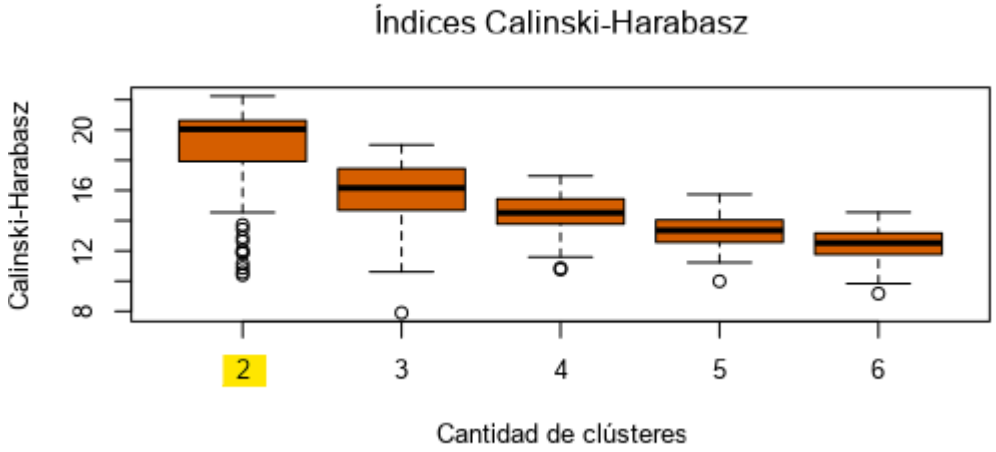
Índices Rand ajustados:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | -0.010301 | 0.044196 | 0.144519 | 0.181317 | 0.165339 |
| 1st Quartile | 0.253055 | 0.253893 | 0.304532 | 0.276847 | 0.287996 |
| Median | 0.427941 | 0.335766 | 0.347545 | 0.35514 | 0.35205 |
| Mean | 0.452115 | 0.368 | 0.365505 | 0.377865 | 0.357587 |
| 3rd Quartile | 0.693715 | 0.483006 | 0.434646 | 0.448919 | 0.42309 |
| Maximum | 0.952935 | 0.748912 | 0.624152 | 0.679767 | 0.555673 |

Índices Calinski-Harabasz:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | 10.39409 | 7.898186 | 10.7658 | 9.986445 | 9.17216 |
| 1st Quartile | 17.979 | 14.714109 | 13.78606 | 12.573044 | 11.76997 |
| Median | 20.05112 | 16.162845 | 14.51101 | 13.351895 | 12.51735 |
| Mean | 18.80087 | 15.87246 | 14.45394 | 13.31809 | 12.41103 |
| 3rd Quartile | 20.60284 | 17.437793 | 15.43122 | 14.043642 | 13.16864 |
| Maximum | 22.23741 | 19.013001 | 16.95309 | 15.737902 | 14.54993 |



2. How many stores fall into each store format?
   Perhaps the answer 1, considering Andre C's post on the discussion board: in order to successfully complete the remaining of the project we need to use 3 clusters for the project here.

   I will use 3 store formats or clusters that give: 25 stores to cluster 1, 25 stores to cluster 2, and 35 stores to cluster 3

**Informe de resumen de la solución de agrupamiento en clústeres K_cluster K-Means**

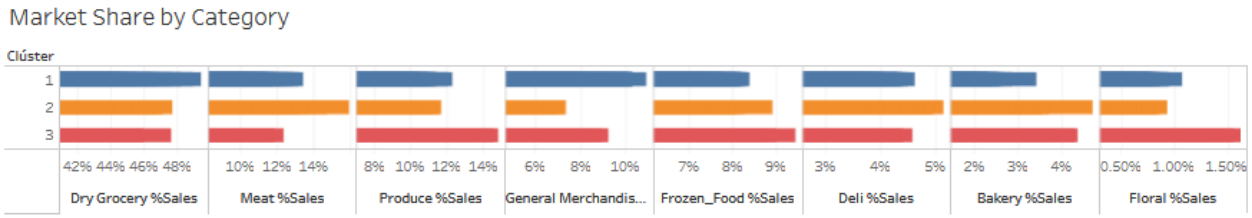
*Resumen de la solución*

Invocar:
stepFlexclust(scale(model.matrix(~-1 + Dry_Grocery_.S + Dairy_.S + Frozen_Food_.S + Meat_.S + Produce_.S + Floral_.S + Deli_.S + Bakery_.S + General_Merchandise_.S, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

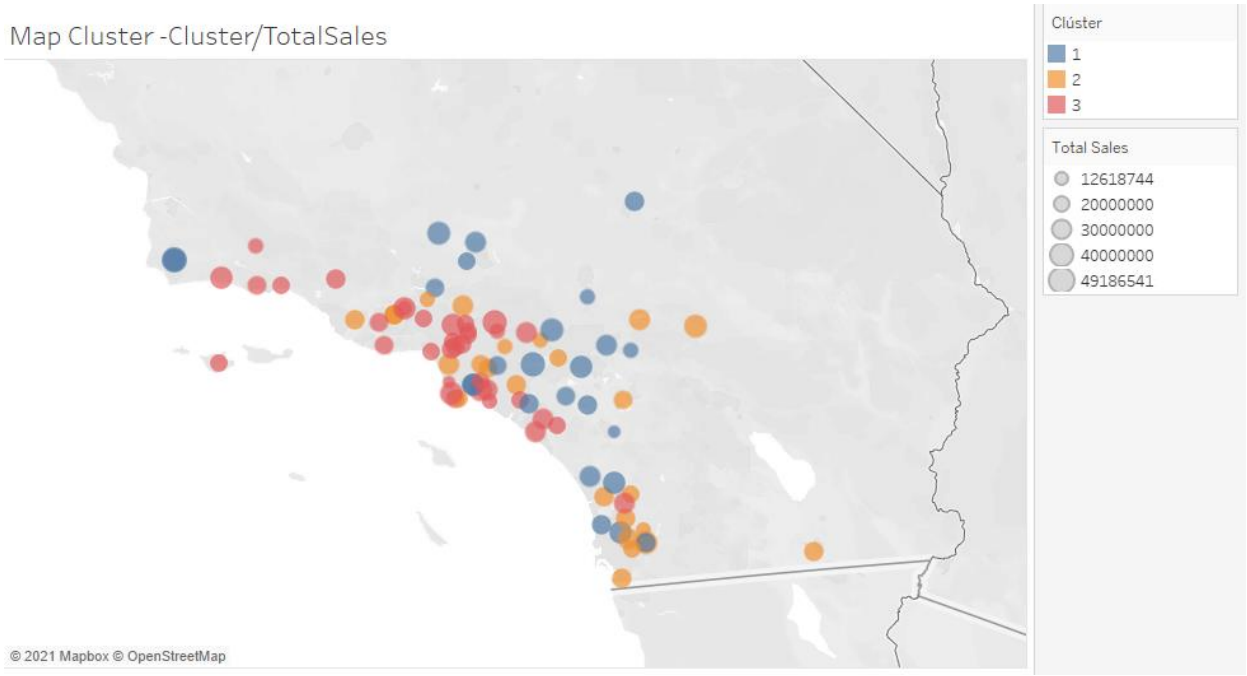
Información del clúster:

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 25 | 2.289004 | 3.585931 | 1.72574 |
| 2 | 25 | 2.099985 | 4.823871 | 2.191566 |
| 3 | 35 | 2.475018 | 4.412367 | 1.947298 |

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   For cluster 1 the highest participation by category against cluster 2 and 3 are: Dry Grocery with 49%, and General Merchandise with 11% while for cluster 2 against cluster 1 and 3 the highest participation is from Meat with a 16%, Deli and Bakery with 5% each, for cluster 3 the highest participation against cluster 1 and 2 are: Produce with 15%, Frozen Food with 9% and Floral with 1.63%as show in the following chart




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/app/profile/criverog/viz/Task1_4PredictiveAnalyticsCapstoneProject/Story1

# 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 methodology used to predict the new store format is the Boosted Model because this shows the highest accuracy (0.7647) against the accuracy of the Forest Model (0.7059) and Decision Tree (0.6471).

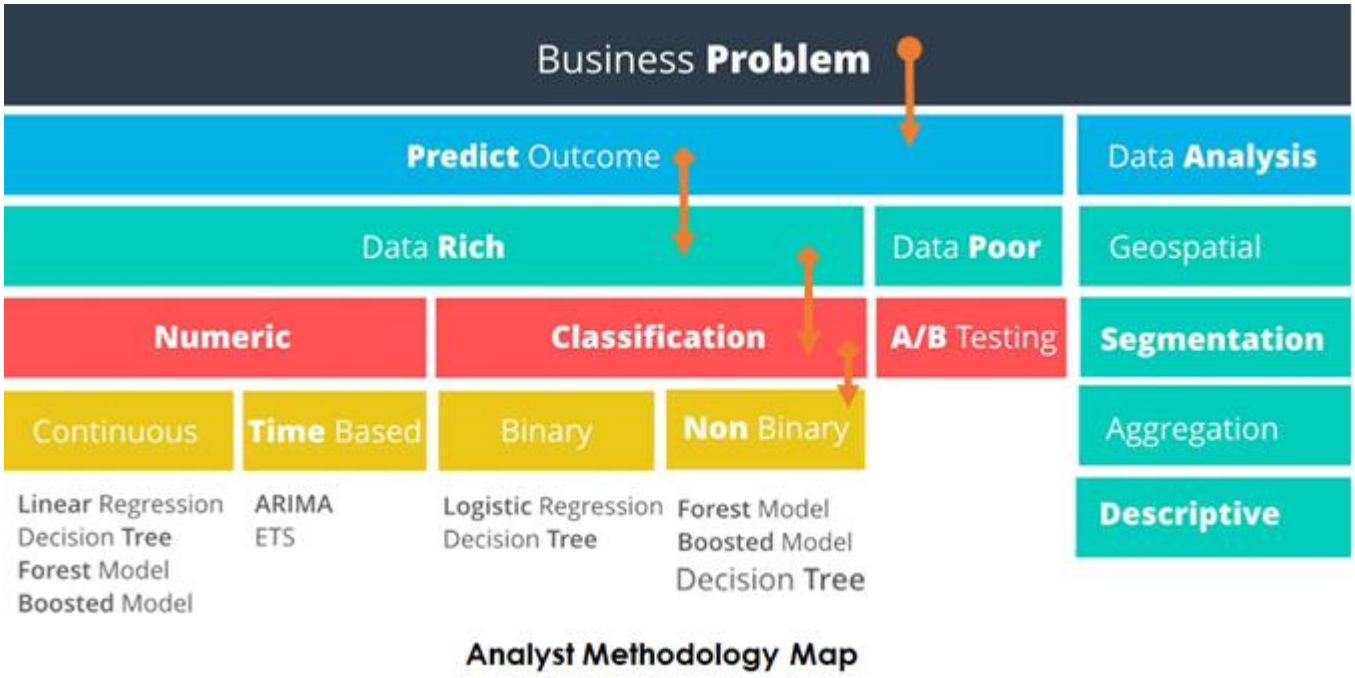
**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| DT_Cluster | 0.6471 | 0.6667 | 0.5000 | 0.5000 | 1.0000 |
| FM_Cluster | 0.7059 | 0.7917 | 1.0000 | 0.3750 | 1.0000 |
| BM_Cluster | 0.7647 | 0.8333 | 1.0000 | 0.5000 | 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.

   The Boosted Model was chosen as mentioned above because it shows the highest precision against the other two referenced models according to the Analyst's Methodology Map; considering that the business problem is to predict the format of new stores, facing a non-binary classification issue due to the nature of our data




**Analyst Methodology Map**

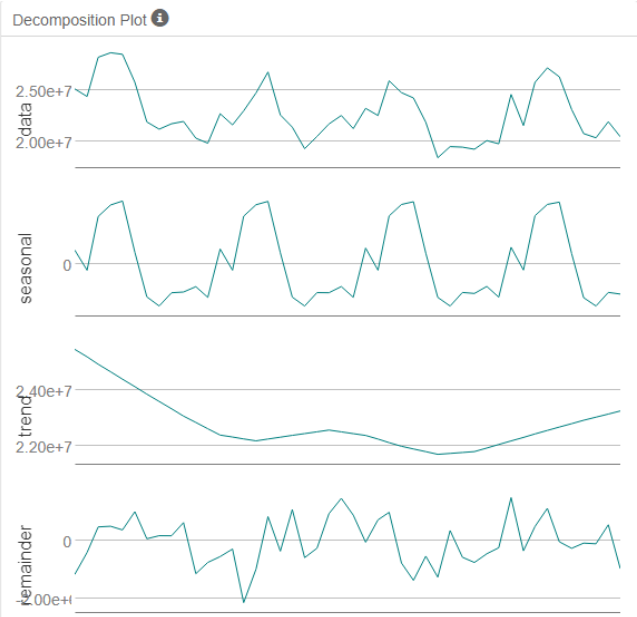
2. What format do each of the 10 new stores fall into? Please fill in the table below.
   According to the Boosted Model and using the scoring tool, the format projected for new stores is indicated in the corresponding Segment field

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

# 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?

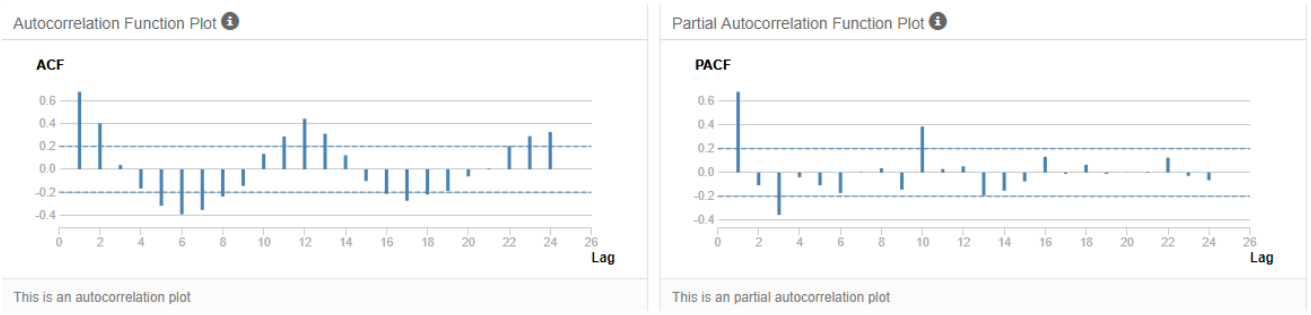
- The type of the ETS used to the forecast is **ETS (M, N, M)**, after the following consideration (based on the analysis of the TS Plot):



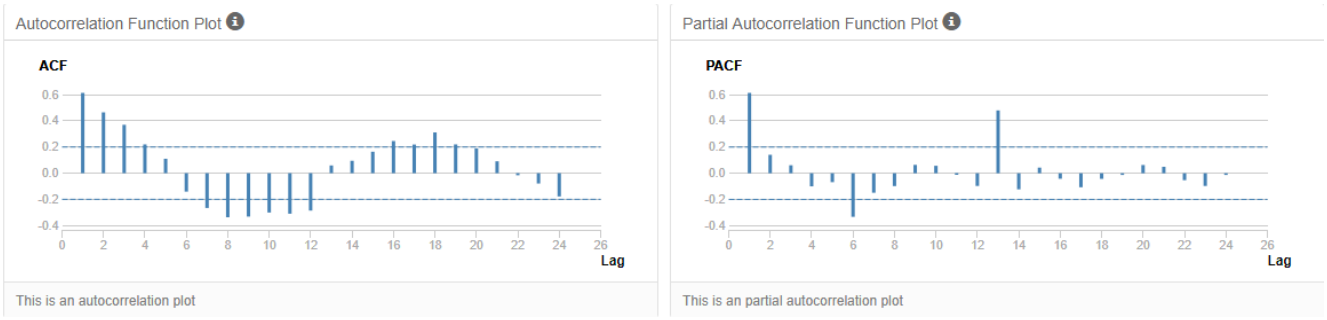

1. The **seasonal** plot shows seasonality with increasing over time, and that is way the “n” factor was considered as **M**ultiplicative
2. The **trend** plot does not show a significative tendency over time (this is showing a flat tendency), and that is way the “m” factor was considered as **N**one.
3. The **remainder or error** plot shows an irregular trend, and that is way the “a” factor was considered as **M**ultiplicative.

- The type of the ARIMA used to the forecast is **ARIMA (0, 1, 2) (0, 1, 0) [12]**, after the following consideration (based on the analysis of the ACF and PACF Plots):
  - The **ar (A**uto **R**egressive**)** term for both the **non-seasonal and seasonal** component of ARIMA model should be **0** because there is no trace of ar term in both ACF and PACF First Difference plots.
  - The **i (D**ifferencing**)** term for both the **non-seasonal and seasonal** components of ARIMA model should be **1** because the stationary condition is reached with in the first differentiation.

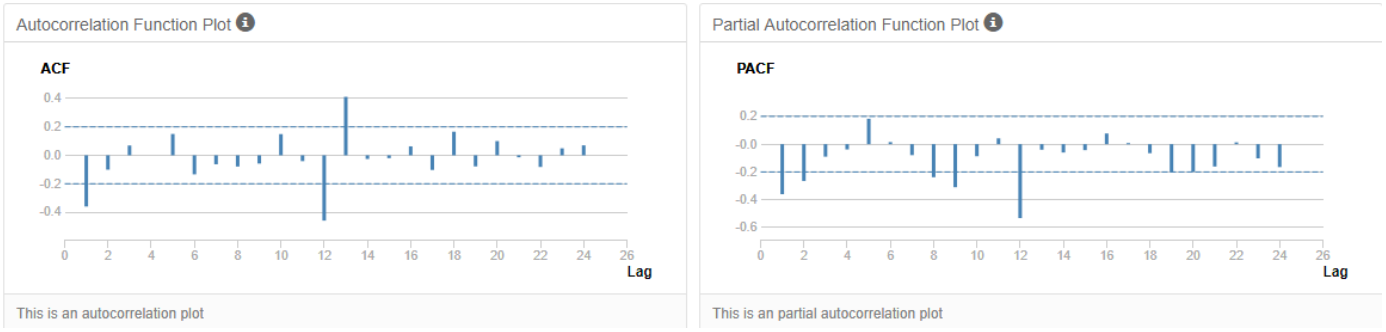
- The **ma** (**M**oving **A**verage) term for the **non-seasonal component** of the ARIMA model should be **2** because for both the ACF and PACF shown a negative correlation at Lag-1 and PACF plot show two spikes with significative correlations in Lag-1 and lag-2. To the **ma** (**M**oving **A**verage) term or the **seasonal component** of the ARIMA model should be **0** because there is no trace of ma term in the plots.
- The **m** (number of periods) of ARIMA models should be **12** because is the number of periods considered.




ACF & PACF Original Time Serie




ACF & PACF Seasonal Difference Time Serie




ACF & PACF First Difference Time Serie

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.
   First In order to get the forecast for existing and new stores; is needed select the best predictive model. That is way, using the TS Compare tool; the **ETS (M, N, M)** model shows lowest RMSE and MASE with a best accuracy against the ARIMA (0, 1, 2) (0, 1, 0) [12] model in the forecast process.

| Parameter | ETS (M, N, M) | ARIMA (0, 1, 2) (0, 1, 0) [12] | Observation |
|---|---|---|---|
| **RMSE** | 663707.2 | 846863.9 | ETS lower than ARIMA |
| **MAPE** | 2.5135 | 2.9927 | ETS lower than ARIMA |
| **MASE** | 0.3257 | 0.3909 | ETS lower than ARIMA |

Report

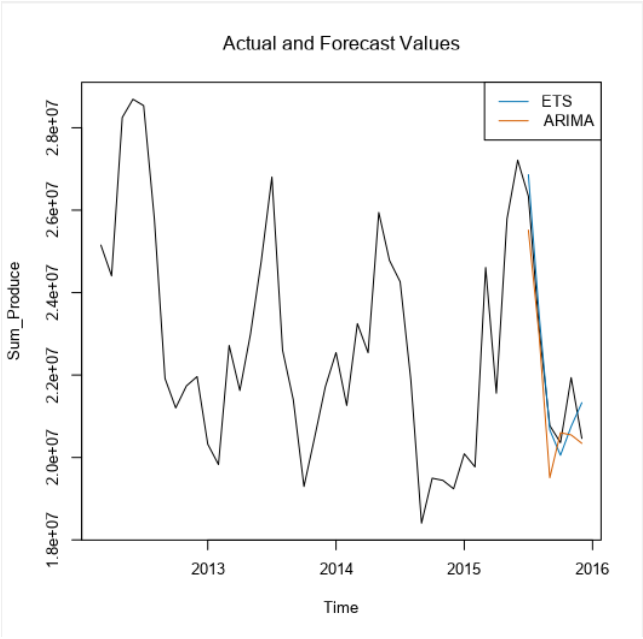
**Comparison of Time Series Models**

Actual and Forecast Values:

| Actual | ETS | ARIMA |
|---|---|---|
| 26338477.15 | 26860639.57444 | 25515002.53492 |
| 23130626.6 | 23468254.49595 | 22982398.33693 |
| 20774415.93 | 20668464.64495 | 19509673.05693 |
| 20359980.58 | 20054544.07631 | 20599981.42693 |
| 21936906.81 | 20752503.51996 | 20547162.64693 |
| 20462899.3 | 21328386.80965 | 20342794.22693 |

Accuracy Measures:

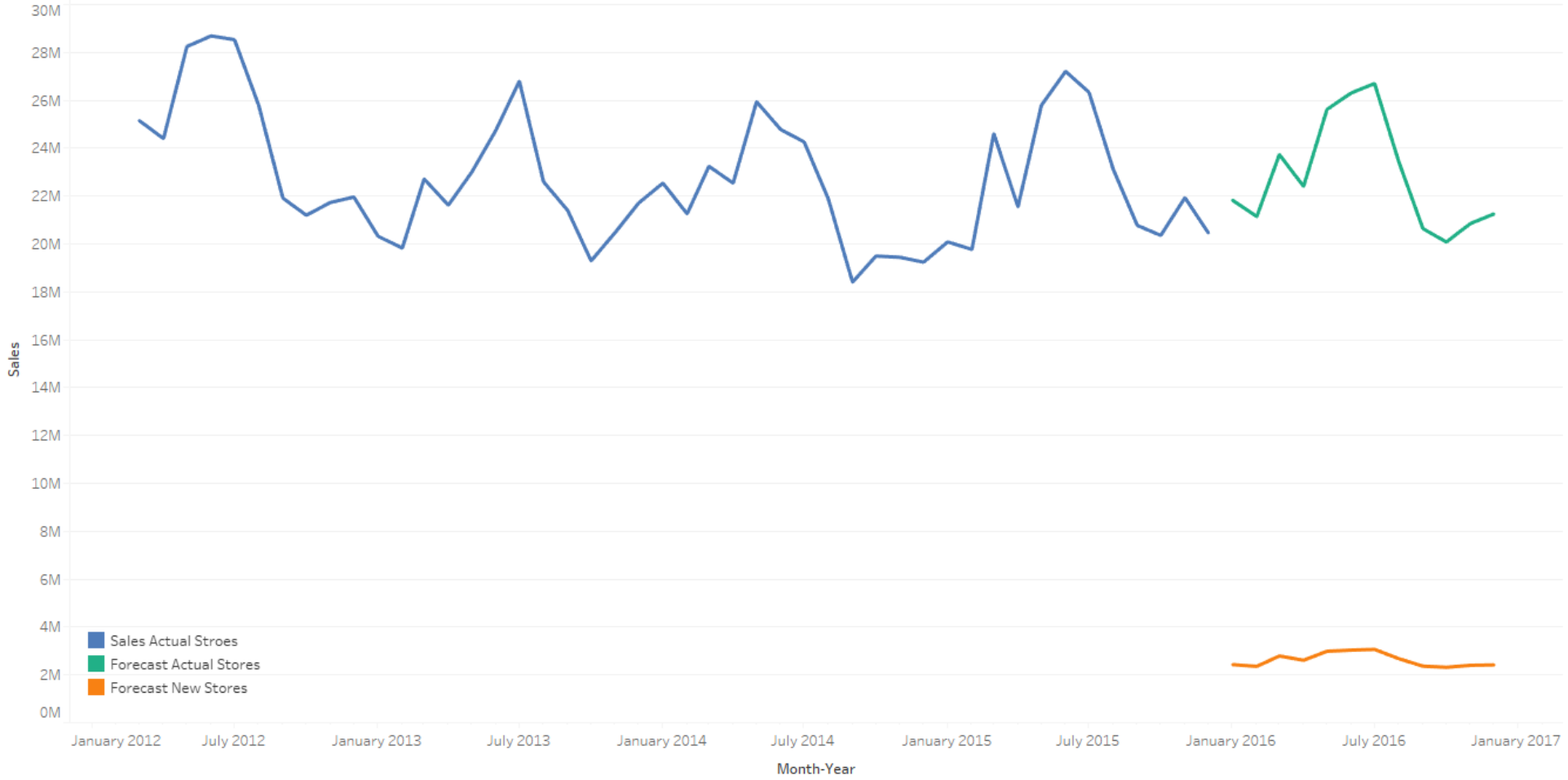
| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |
| ARIMA | 584382.36 | 846863.9 | 664382.6 | 2.5998 | 2.9927 | 0.3909 |



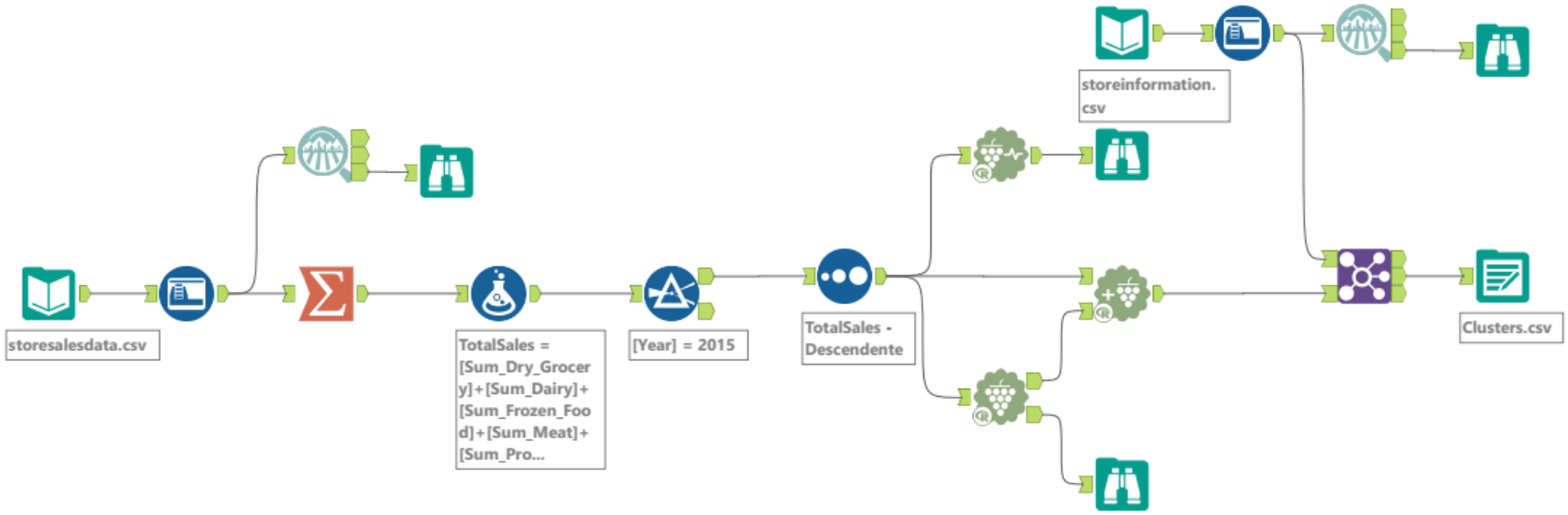

Finally, the sales forecast for existing stores a new ones for produce are:

| Month-Year | Forecast New Stores | Forecast Actual Stores |
|---|---|---|
| Jan-16 | $2,429,600.68 | $21,829,060.03 |
| Feb-16 | $2,356,690.35 | $21,146,329.63 |
| Mar-16 | $2,789,909.08 | $23,735,686.94 |
| Apr-16 | $2,613,119.70 | $22,409,515.28 |
| May-16 | $2,987,310.74 | $25,621,828.73 |
| Jun-16 | $3,036,185.79 | $26,307,858.04 |
| Jul-16 | $3,065,391.10 | $26,705,092.56 |
| Aug-16 | $2,679,384.64 | $23,440,761.33 |
| Sep-16 | $2,365,494.48 | $20,640,047.32 |
| Oct-16 | $2,318,418.48 | $20,086,270.46 |
| Nov-16 | $2,401,695.26 | $20,858,119.96 |
| Dic-16 | $2,418,318.04 | $21,255,190.24 |

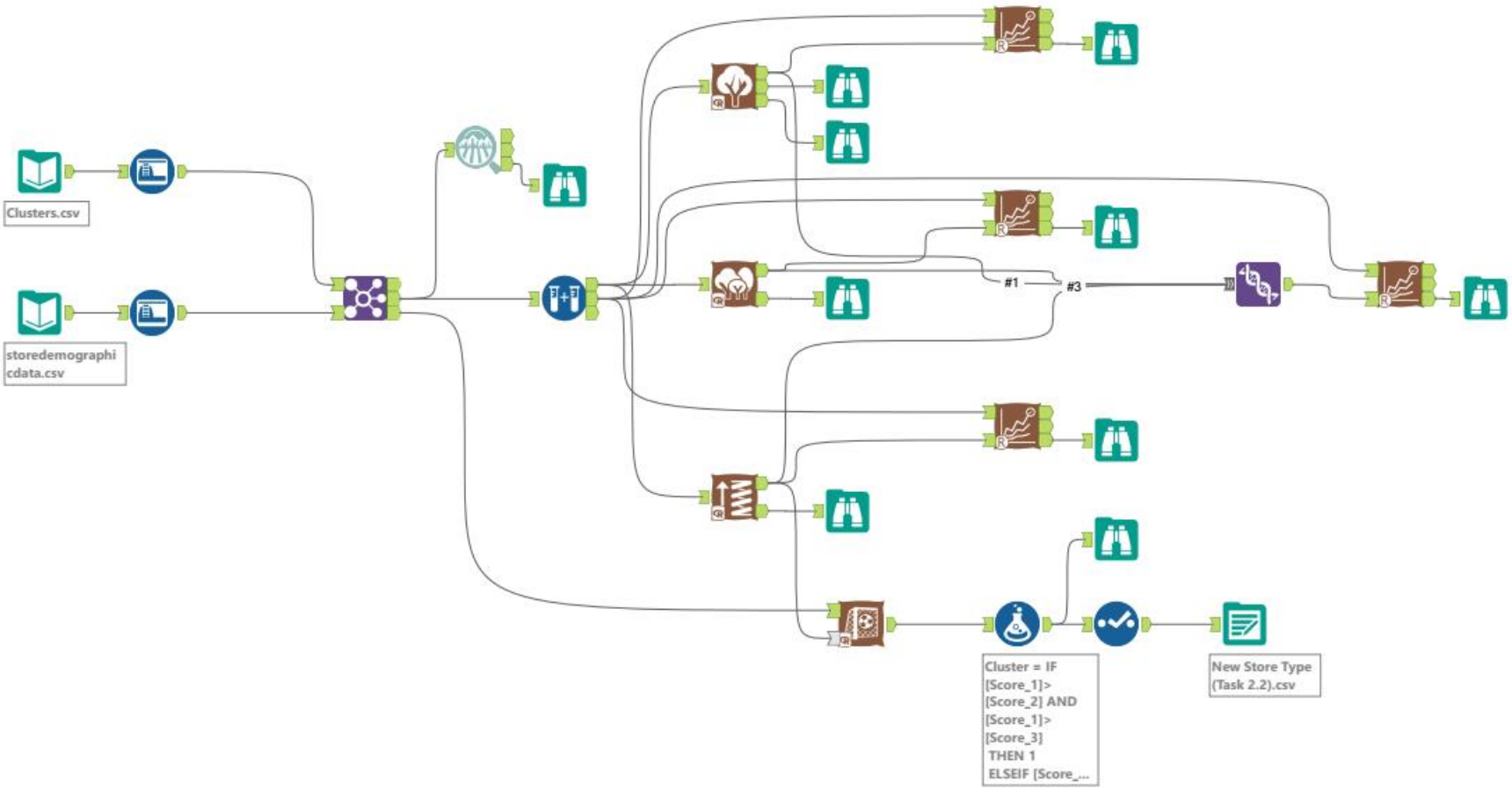
## Produce Sales Historical & Forecast (Actual and New)



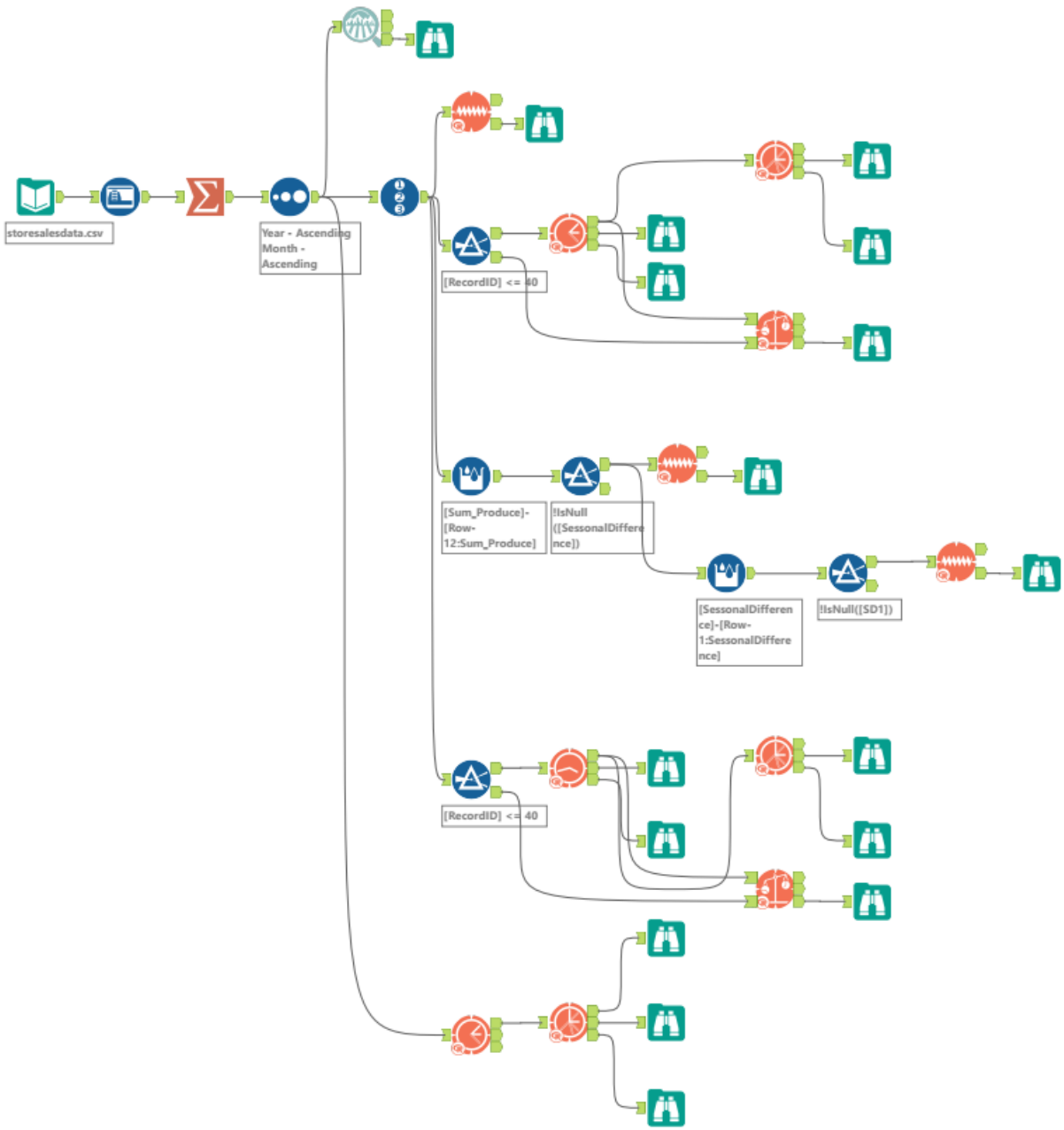
# Predictive Analytics Capstone (Task 1)



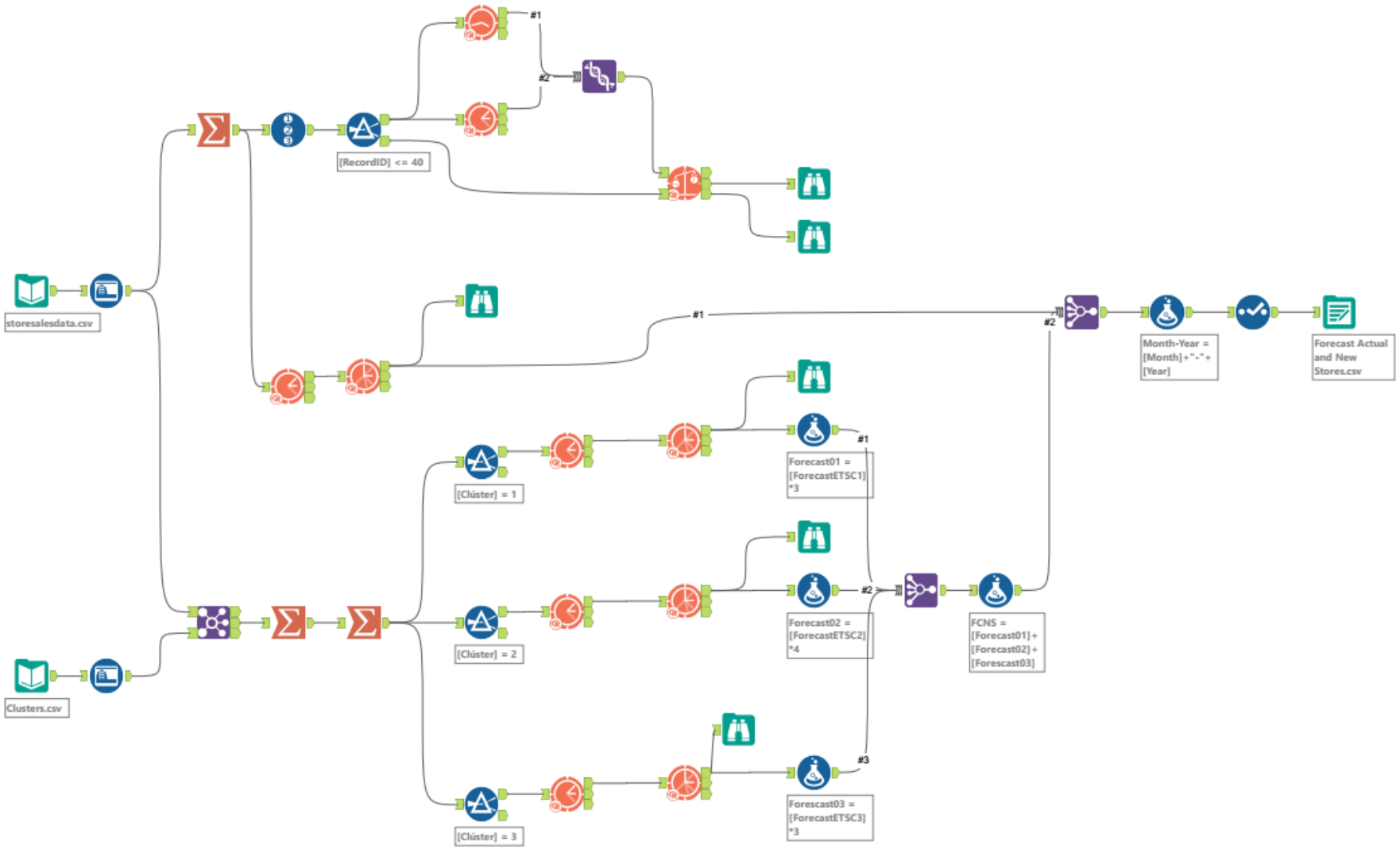
# Predictive Analytics Capstone (Task 2)



# Predictive Analytics Capstone (Task 3.1)



# Predictive Analytics Capstone (Task 3.2)



# Predictive Analytics Capstone (Task 3.2.1)