# 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?**
   After we summarize sales data and create new variables (percentage value of each category) per each store, we use K-Centroids Diagnostics tool in Alteryx to decide the optimal number of clusters using K-means clustering method.

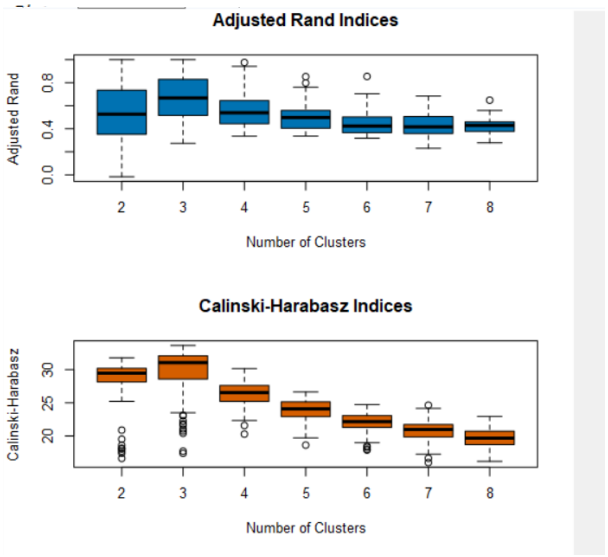
*Summary Statistics*

Adjusted Rand Indices:

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Minimum | -0.016293 | 0.27351 | 0.335359 | 0.336327 | 0.318262 | 0.230196 | 0.27786 |
| 1st Quartile | 0.352041 | 0.515917 | 0.445826 | 0.409773 | 0.366788 | 0.358895 | 0.377341 |
| Median | 0.526785 | 0.66768 | 0.538528 | 0.497192 | 0.423541 | 0.416509 | 0.428806 |
| Mean | 0.53781 | 0.664773 | 0.565975 | 0.50103 | 0.45115 | 0.432196 | 0.421514 |
| 3rd Quartile | 0.734477 | 0.826692 | 0.644691 | 0.555087 | 0.499921 | 0.502931 | 0.458601 |
| Maximum | 1 | 1 | 0.975264 | 0.852076 | 0.8539 | 0.683894 | 0.647983 |

Calinski-Harabasz Indices:

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Minimum | 16.61829 | 17.38103 | 20.28456 | 18.61989 | 17.8746 | 15.98702 | 16.16824 |
| 1st Quartile | 28.17383 | 28.57484 | 25.20913 | 22.93454 | 21.30575 | 19.85155 | 18.71365 |
| Median | 29.46587 | 31.05384 | 26.53788 | 24.086 | 22.16245 | 20.97743 | 19.6662 |
| Mean | 28.45131 | 29.70664 | 26.41806 | 23.87003 | 22.02174 | 20.77195 | 19.65973 |
| 3rd Quartile | 30.17907 | 32.08726 | 27.59305 | 25.10099 | 23.06602 | 21.72942 | 20.7099 |
| Maximum | 31.78345 | 33.63781 | 30.1583 | 26.63063 | 24.72038 | 24.63982 | 22.95166 |




As we can see from Adjusted Rand and CH indices summary table and more obviously from visualization plots, 3 clusters will show the best results (it has the highest median in both indices). As we now, the higher the index the more accurate and representative will be our clustering. So, as a number of clusters we will choose 3.

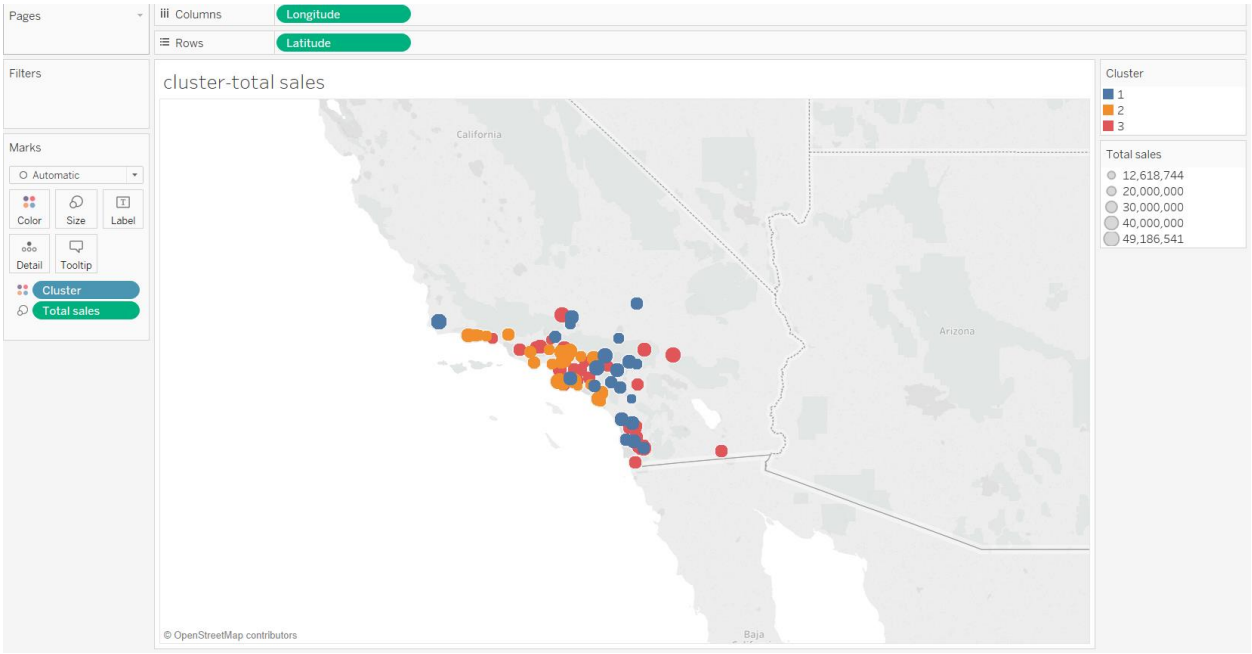
2. **How many stores fall into each store format?**
   After running the K-Centroids Cluster Analysis in Alteryx we have the following information about the 3 clusters:

Cluster Information:

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 23 | 2.320539 | 3.55145 | 1.874243 |
| 2 | 29 | 2.540086 | 4.475132 | 2.118708 |
| 3 | 33 | 2.115045 | 4.9262 | 1.702843 |

As we see, we have 23 stores in cluster 1, 29 stores in cluster 2 and 33 stores in cluster 3.

3. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**
   One of the ways the cluster differ from each other is by the variable of percentage of dry grocery in total sales.
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/azganush#!/vizhome/clustervisualization/cluster-totalsales?publish=yes

# 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.)**
   We need to decide in which segment to include the new 10 stores. We have only demographic data of the new stores, so this is the only information we can use to build the model. So we need to use the socioeconomic and demographic data about existing stores and already defined segments for the same stores to find out which variables are more correlated with each cluster/segment. We will build the models through Alteryx, using Decision Tree, Forest Model and Boosted Model tools. We will compare the results the best model from these three and then based on the analysis of the best

model will choose in which segments to include the new stores.

**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| Decision_Tree_ | 0.8235 | 0.8426 | 0.7500 | 1.0000 | 0.7778 |
| Forest_model | 0.8235 | 0.8426 | 0.7500 | 1.0000 | 0.7778 |
| Boosted_model | 0.8235 | 0.8889 | 1.0000 | 1.0000 | 0.6667 |

Comparing the accuracy measurements for three models, we find out that the overall accuracy is the same for all of them - 82%, however, when we compare accuracies for separate clusters, the boosted model has 100% accuracy for 1 and 2 clusters, and 66% for the third one, compared to decision tree and forest models that are a little higher for the 3rd cluster – 77% however, are much lower for the first two, only 75%.
So, for our analysis, we will choose the boosted model.

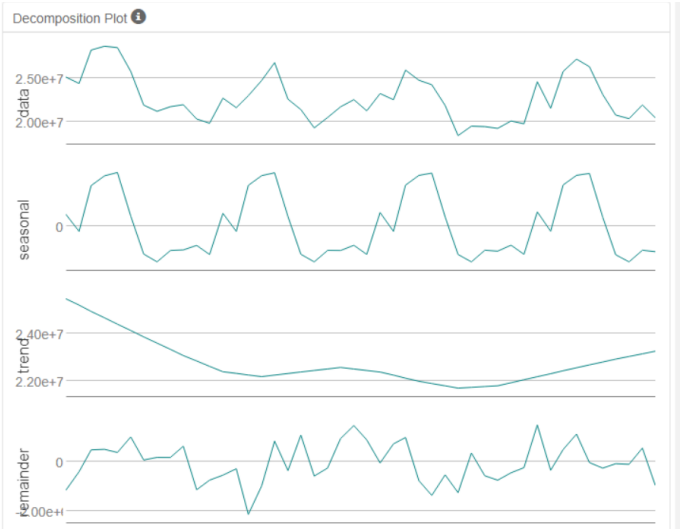
2. What format do each of the 10 new stores fall into? Please fill in the table below.

| Store Number | Segment |
|---|---|
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 1 |
| S0092 | 2 |
| 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?**

First we aggregate the produce sales data by year and month. Then we run TS Plot in Alteryx to decide the model types. First we will build ETS model.



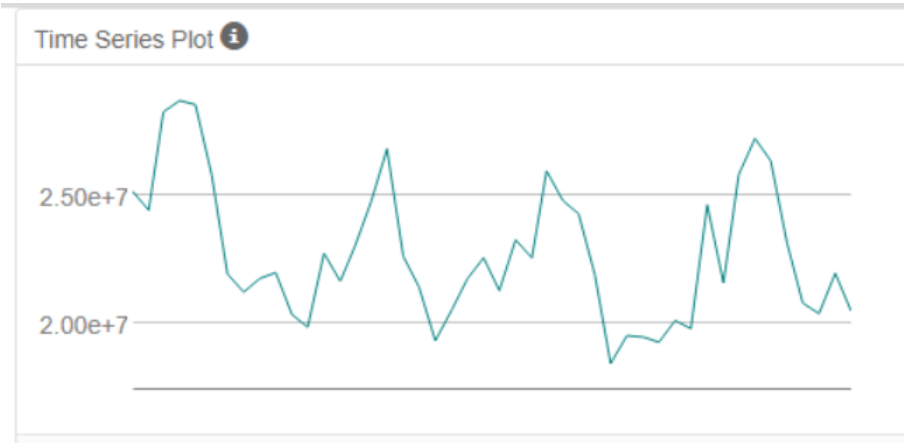

From the decomposition plot we see that the error values are randomly distributed around the mean and it has multiplicative behavior, there is no trend in this data, and there is seasonality and it again has multiplicative behavior as we can see. So we will choose ETS(m,n,m) model for our analysis.

After running ETS model we use TS Compare tool to validate the results for 6 months' holdout period. Here are the accuracy measures of the model:

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|---|---|---|---|---|---|---|---|
| ETS | 210494.4 | 760267.3 | 649540.8 | 1.0288 | 2.9678 | 0.3822 | NA |

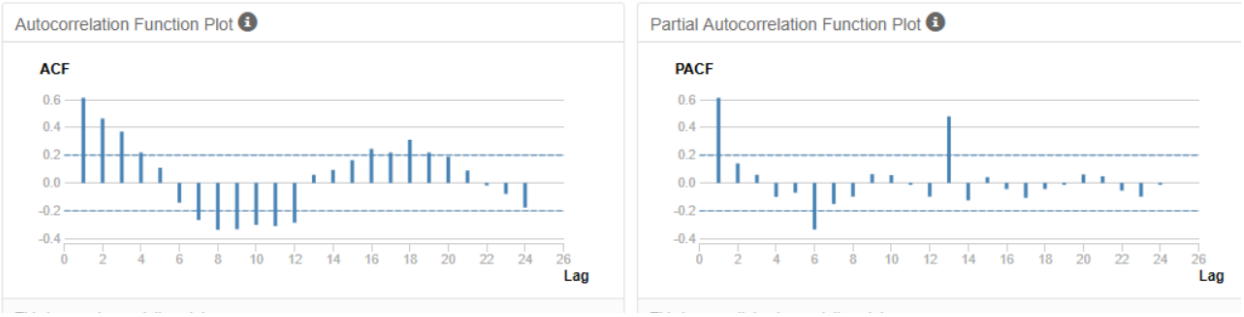
We see that MASE error is 0.3822 which is below 1 and means that the model is a good indicator of the model's accuracy. The RMSE error is quite high, however, we still need to create ARIMA model to choose the best one between them.




From the Time series plot we notice that the data is not stationary at all, so we need to make it stationary to have an accurate result in our analysis. For that reason, we will do

seasonal differencing as the model is seasonal.


After seasonal difference we see that the data is already stationary, so we can now look at ACF and PACF plots to decide model terms.




Looking at ACF we see a positive correlation at lag 1 and it slowly decays towards 0, so we have AR term, and in PACF graph we see a sharp cutoff after lag 1, so AR numeric component is 1. MA is 0, and as we did not do non-seasonal differencing, the d value is 0. As we see a negative correlation at lag 12, so we have MA term for seasonal component, and it has numeric value 1 as at lag 24 the autocorrelation is not significant. As we did one seasonal difference, the D term's value will be 1.
So our ARIMA model terms are ARIMA (1,0,0) (0,1,1) [12].
Using building ARIMA model with the following terms and using TS compare tool against the 6 months' holdout sample, we have the following model accuracy measures:

Accuracy Measures:

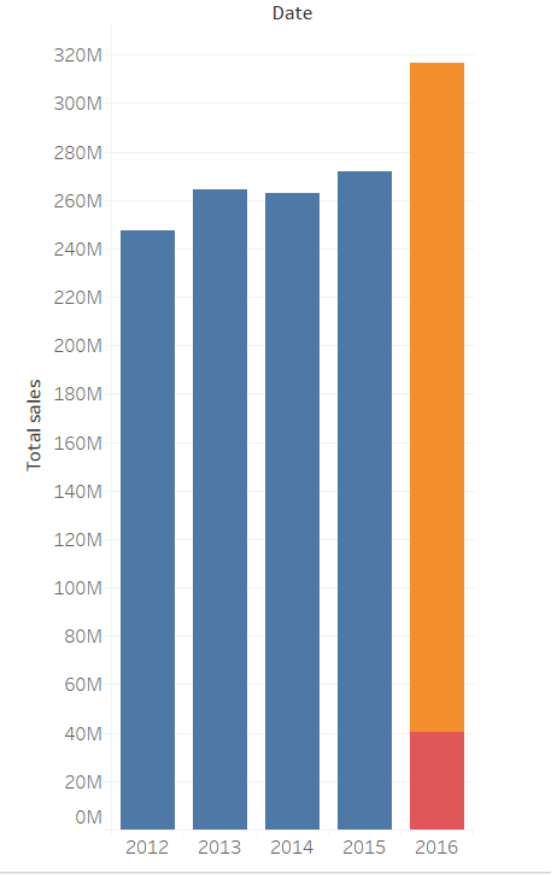
| Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|---|---|---|---|---|---|---|---|
| project_arima | -335734.3 | 791161.4 | 686064.1 | -1.4218 | 3.0188 | 0.4037 | NA |

As we see, we have better figures in ETS model, as both the MASE and RMSE errors have lower values: 0.38 and 760267. So, for our forecasting we will choose ETS model.

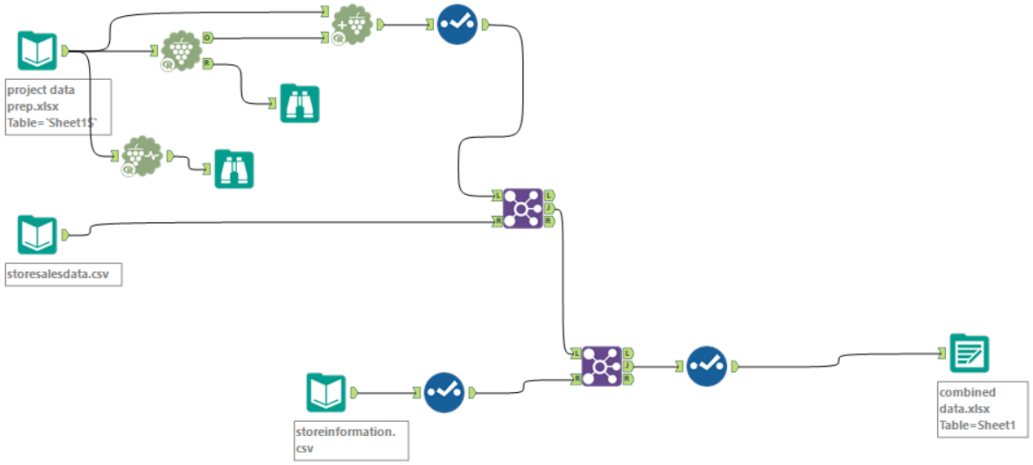
3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

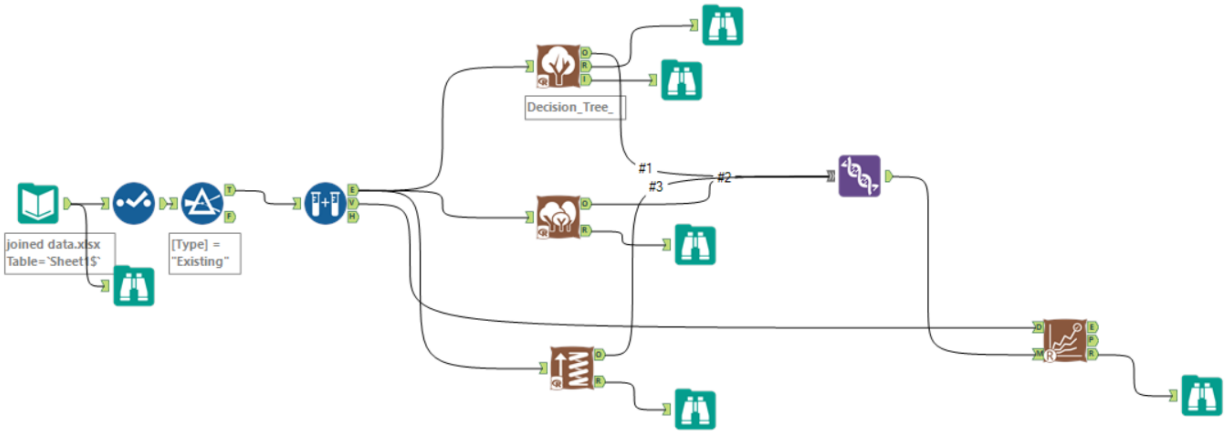
| Month | Forecasting (existing stores) | Forecasting (new stores) |
|---|---|---|
| 2016-01 | 21539936 | 2563551 |
| 2016-02 | 20413771 | 8729671 |
| 2016-03 | 24325953 | 2907666 |
| 2016-04 | 22993466 | 2760886 |
| 2016-05 | 26691951 | 3143339 |
| 2016-06 | 26989964 | 3191030 |
| 2016-07 | 26948631 | 3212766 |
| 2016-08 | 24091579 | 3464103 |
| 2016-09 | 20523492 | 2538688 |
| 2016-10 | 20011749 | 2495887 |
| 2016-11 | 21177435 | 2591871 |
| 2016-12 | 20855799 | 2551957 |

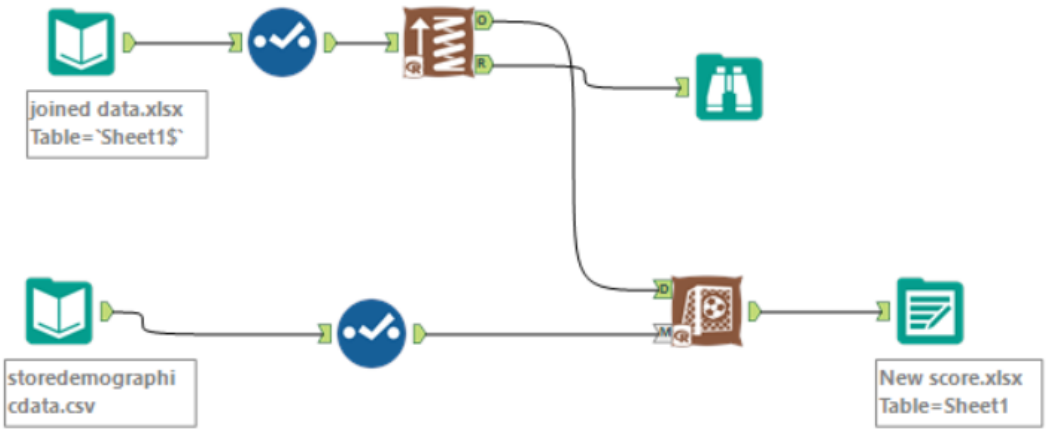
Below is the visualization of historical data together with existing and new stores sales forecasting data. As we can see, the forecasted amount of sales for 2016 combined for existing and new stores exceeds each previous year's figures. And regarding the new stores' contribution, it makes around 1/6 of 2016 total forecast.

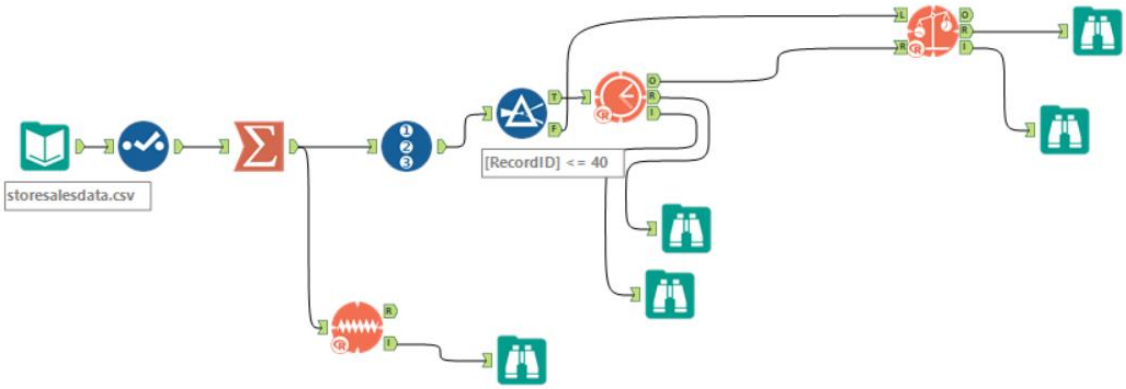


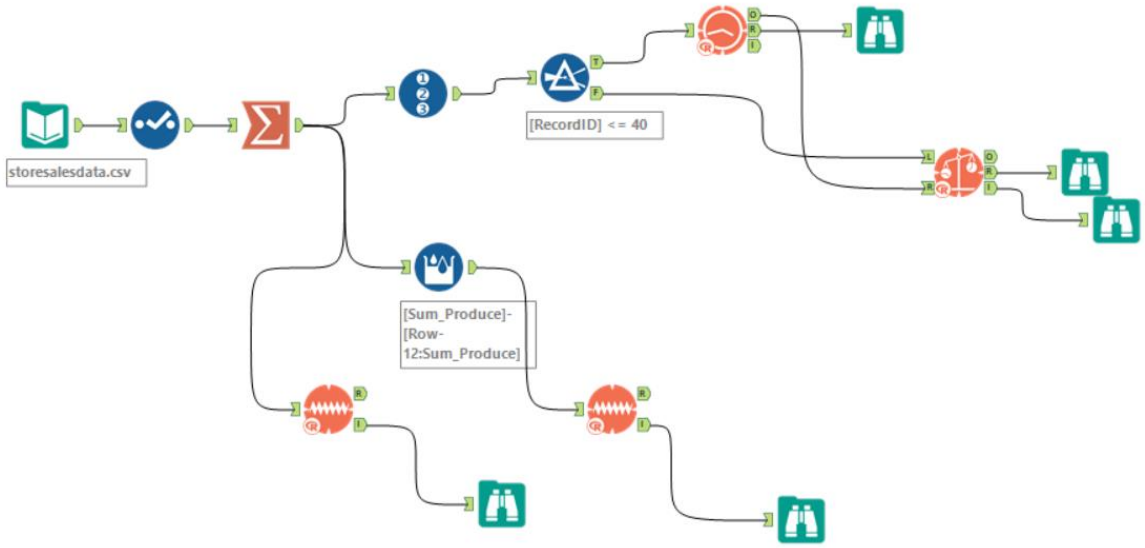
Below are Alteryx workflows used in this project:

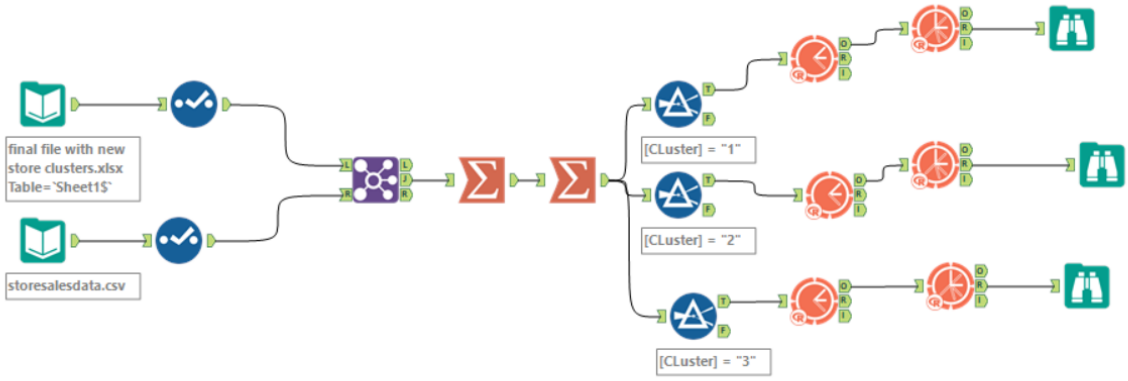



joined data.xlsx
Table=`Sheet1$`
storedemographicdata.csv
New score.xlsx
Table=Sheet1


storesalesdata.csv
[RecordID] <= 40


storesalesdata.csv
[RecordID] <= 40
[Sum_Produce]-
[Row-
12:Sum_Produce]


final file with new
store clusters.xlsx
Table=`Sheet1$`
storesalesdata.csv
[CLuster] = "1"
[CLuster] = "2"
[CLuster] = "3"