# Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here:

### Task 1: Store Format for Existing Stores

Your company currently 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. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

### Task 1: Determining Store Format

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. You've been asked to:

- 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.
- Segment the 85 current stores into the different store formats.
- Use the StoreSalesData.csv and StoreInformation.csv files.

## Task 1: Determine Store Formats for Existing Stores

### 1. What is the optimal number of store formats? How didyou arrive at that number?

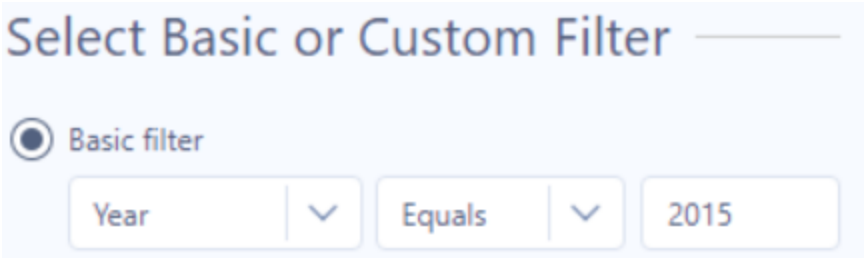
#### What is the optimal number of store formats?

Based on the findings of the Alteryx workflow I created,I came to the conclusion that it would be **two**, but after reading through the "knowledge hub", Irealized it should be **three**.

#### How did you arrive at that number?

All of the information needed to complete the analysis was contained in the data provided. First, I grasped the goal of the segmentation I was attemptingto construct.

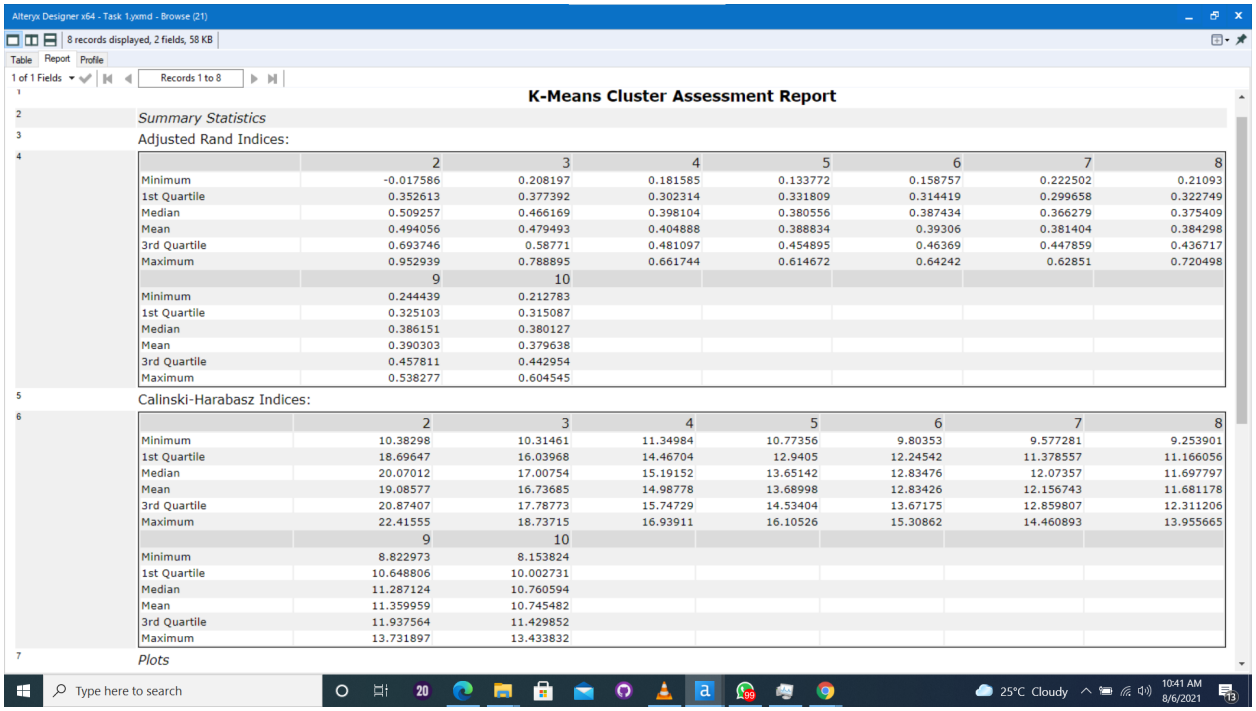
➔ I changed sales fields from "**VString**" to "**Double**" after joining two data files and filtering 2015 information.




- ➔ I summed the data by each category and grouped it by "Store Id and Year."
- ➔ I used the "**Formula Tool**" to construct a "**Total Sales**" field that totaled all categories.
- ➔ The next step was to create **nine** separate columns, each of which received the sum of a category divided by "Total Sales" multiplied by 100. By the end, I had 85 records.

| Record |
|---|
| 79 |
| 80 |
| 81 |
| 82 |
| 83 |
| 84 |
| 85 |

The y-axis in both plots "**Adjusted Rand Indices**" and "**Calinski-Harabasz Indices**" is the index value, and the x-axis is the number of clusters. The **higher** the index for the "Adjusted Rand Indices," the more **stable** the cluster. The **higher** the index for the "Calinski-Harabasz Indices," the **greater** the **distinctness and compactness** of the clusters.



Alteryx Designer x64 - Task 1.yxmd - Browse (21)

8 records displayed, 2 fields, 58 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 8

**K-Means Cluster Assessment Report**

*Summary Statistics*

Adjusted Rand Indices:

| | 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 |

*Plots*

Type here to search

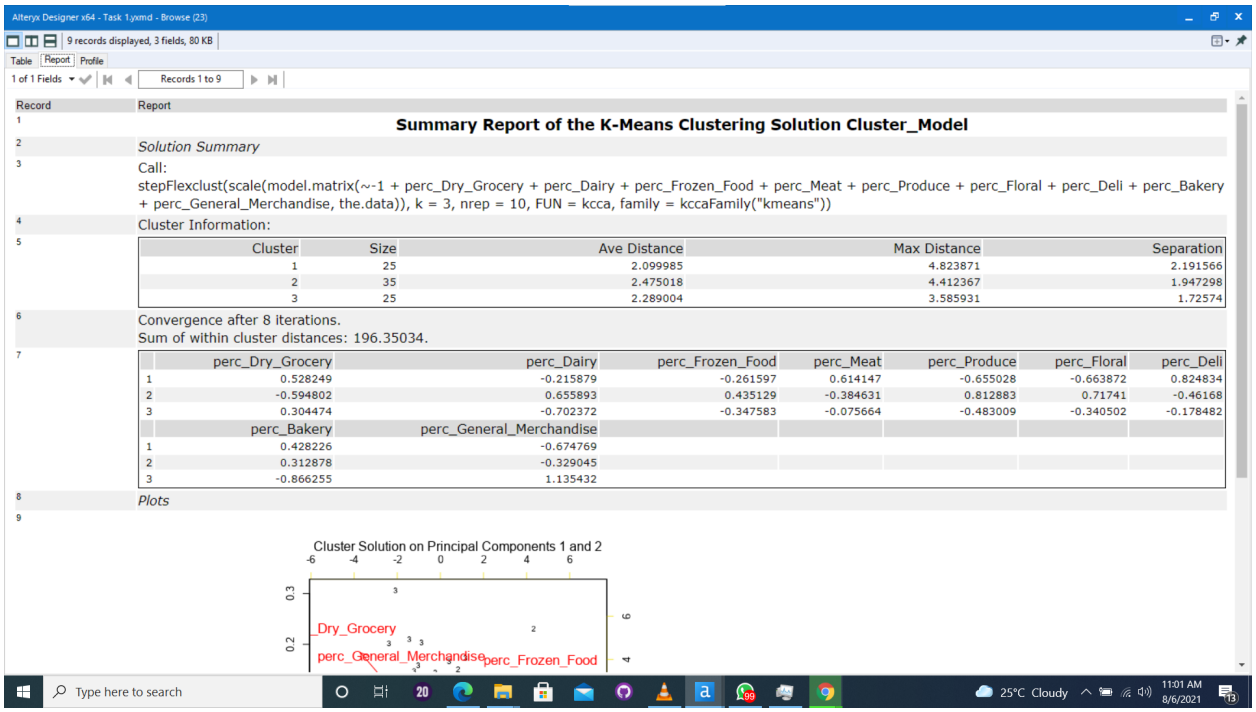
25°C Cloudy

10:41 AM 8/6/2021

## 2. How many stores fall into each store format?

- ➔ Cluster 1 has 25 stores,
- ➔ Cluster 2 has 35 stores, and
- ➔ Cluster 3 has 25 stores.

None of the clusters I identified have fewer than20 or more than 40 stores.



Alteryx Designer x64 - Task 1.yxmd - Browse (23)

9 records displayed, 3 fields, 80 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 9

| Record | Report |
|---|---|
| 1 | **Summary Report of the K-Means Clustering Solution Cluster_Model** |
| 2 | *Solution Summary* |
| 3 | Call:<br>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")) |
| 4 | 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 |

| 6 | Convergence after 8 iterations.<br>Sum of within cluster distances: 196.35034. |
|---|---|

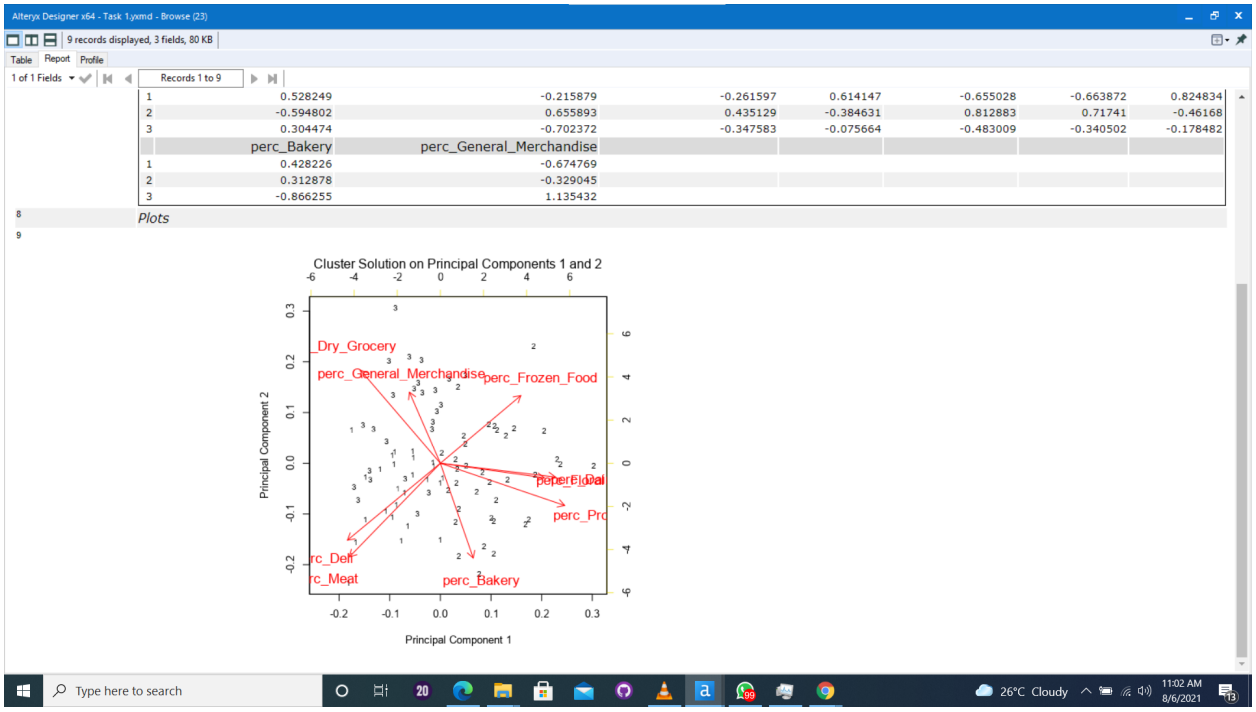
| 7 | | perc_Dry_Grocery | perc_Dairy | perc_Frozen_Food | perc_Meat | perc_Produce | perc_Floral | perc_Deli |
|---|---|---|---|---|---|---|---|---|
| | 1 | 0.528249 | -0.215879 | -0.261597 | 0.614147 | -0.655028 | -0.663872 | 0.824834 |
| | 2 | -0.594802 | 0.655893 | 0.435129 | -0.384631 | 0.812883 | 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 |

| 8 | *Plots* |
|---|---|



Type here to search | 25°C Cloudy | 11:01 AM 8/6/2021



Alteryx Designer x64 - Task 1.yxmd - Browse (23)

9 records displayed, 3 fields, 80 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 9

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 |

| 8 | *Plots* |
|---|---|



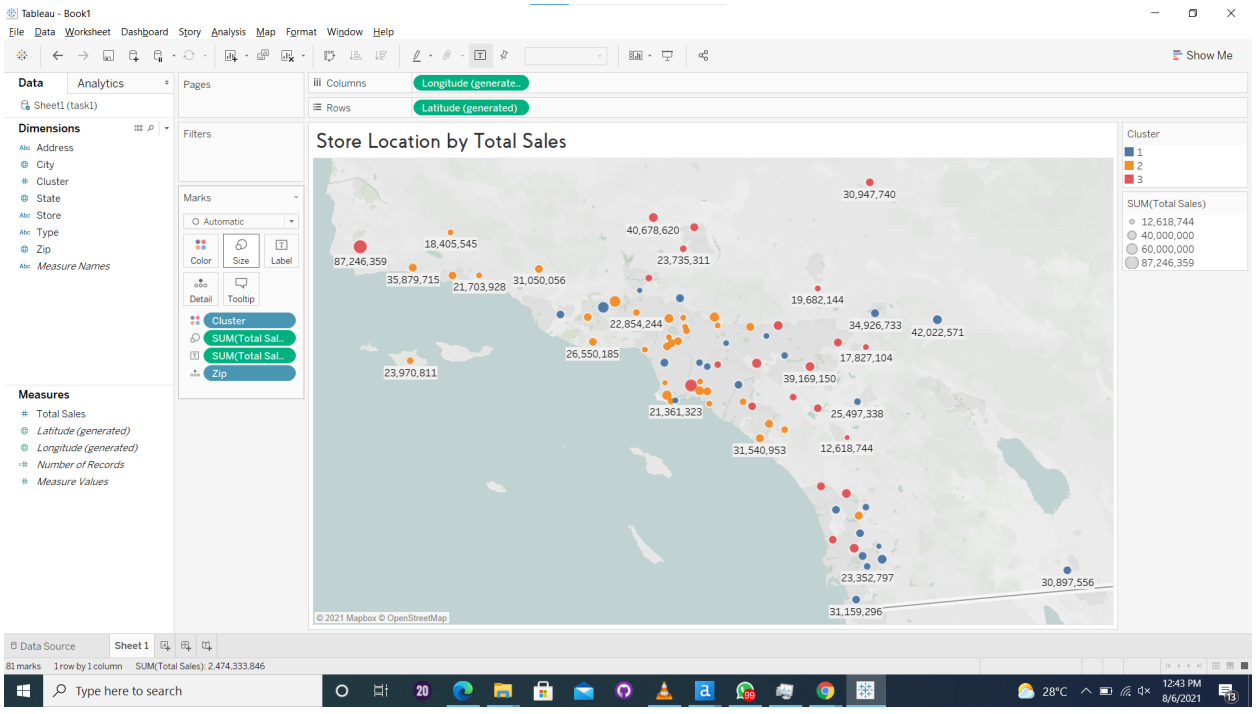
Type here to search | 26°C Cloudy | 11:02 AM 8/6/2021

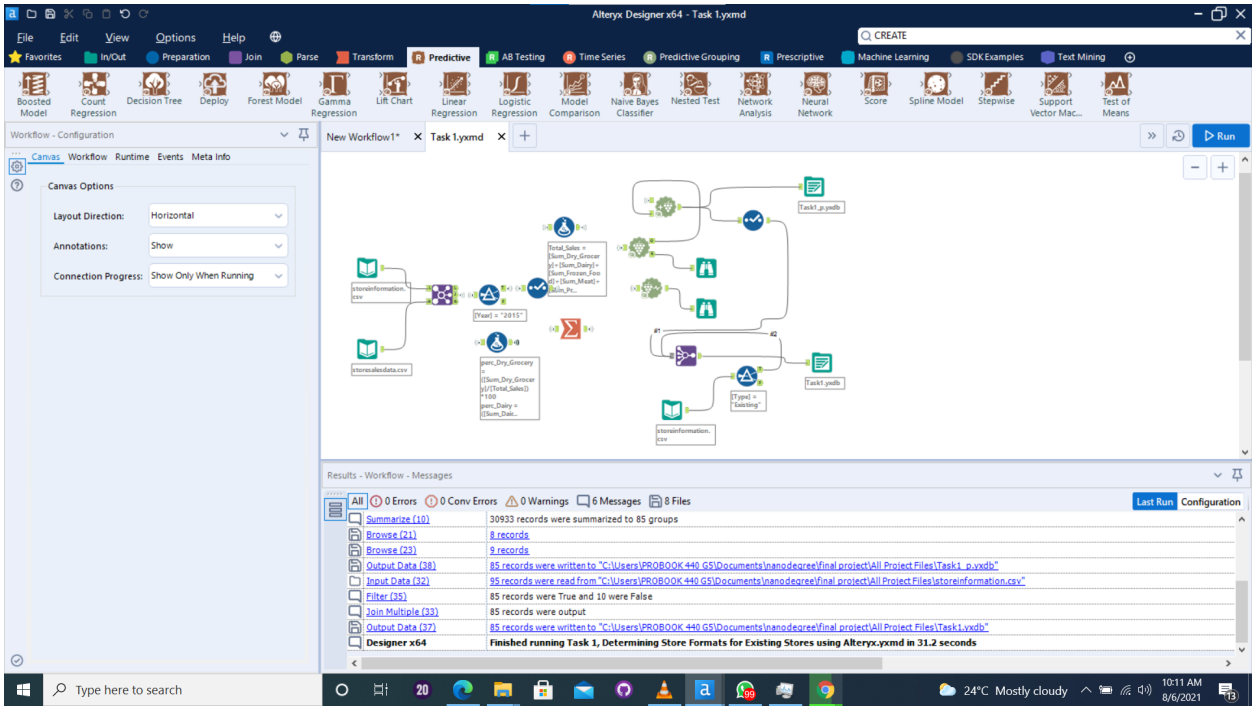
## 3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

The most compact cluster is Cluster 1. Cluster 2 could have a higher level of variation. The maximum distance from the centroid is greatest in Cluster 1.
Cluster 1 is further apart than the others.

## 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.

## Tableau Link






# 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.

Pretty sweet grocery store, right?

### Task 2: Determine the Store Format for New Stores

You've been asked to:

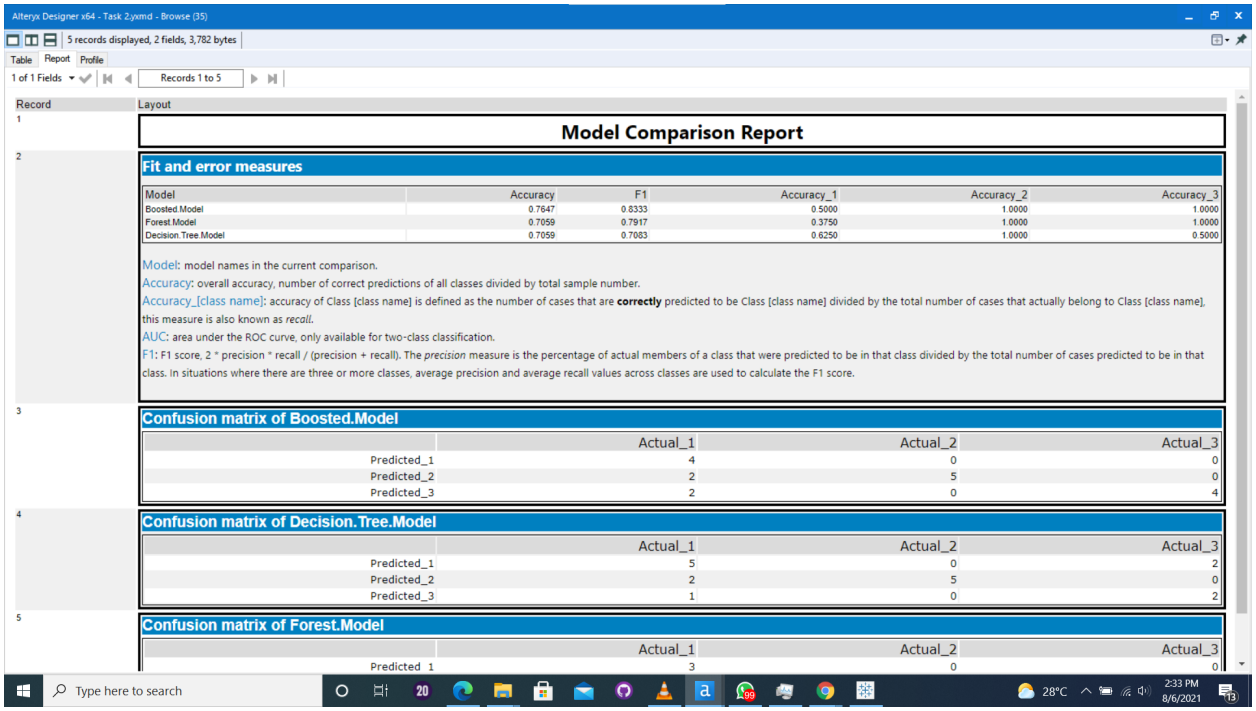
- Develop a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
- Use a 20% validation sample with *Random Seed* = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
- Use the model to predict the best store format for each of the 10 new stores.
- Use the StoreDemographicData.csv file, which contains the information for the area around each store.

- **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 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.)

**Cluster was used as the "target variable," and all demographic data were used as predictors. In terms of Overall Accuracy, "F1 Score, and the accuracy of each class," we compared the performance of "Decision Tree, Forest, and Boosted models." The"boosted model" had the best results.**



Alteryx Designer x64 - Task 2.yxmd - Browse (35)

5 records displayed, 2 fields, 3,782 bytes

Table | Report | Profile

1 of 1 Fields | Records 1 to 5

Record | Layout

**Model Comparison Report**

**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 |

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**

| | 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**

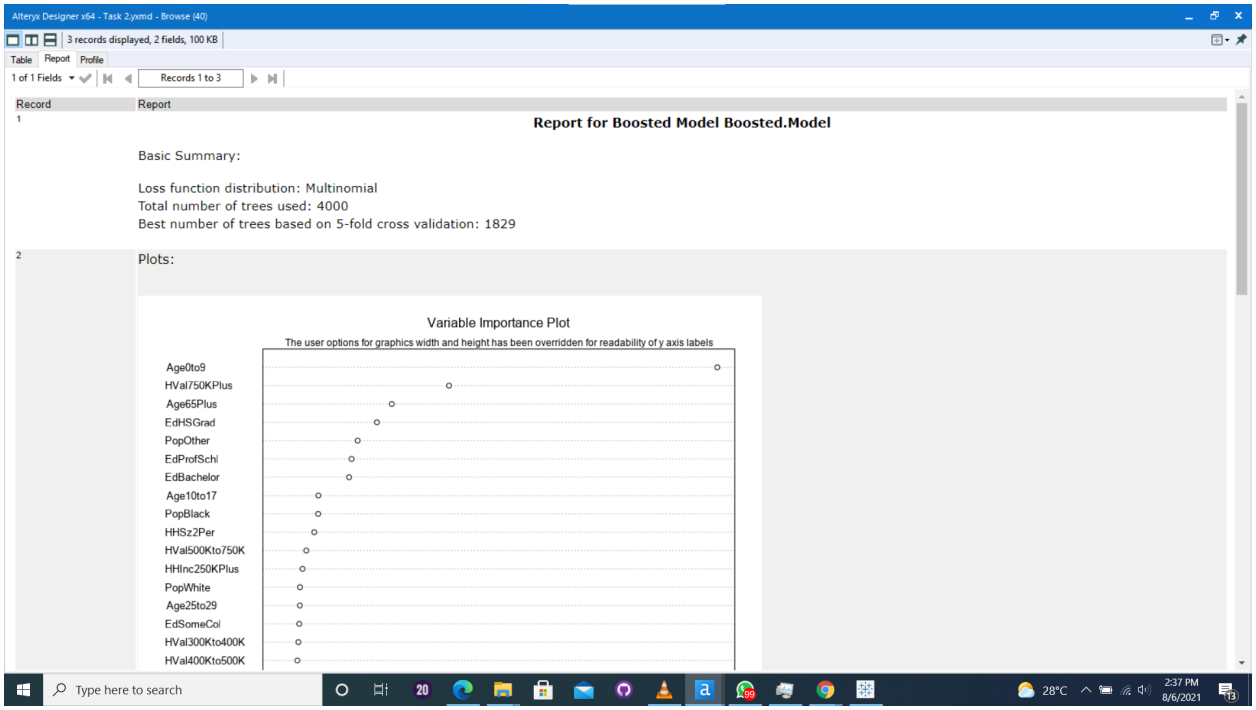
| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 5 | 0 | 2 |
| Predicted_2 | 2 | 5 | 0 |
| Predicted_3 | 1 | 0 | 2 |

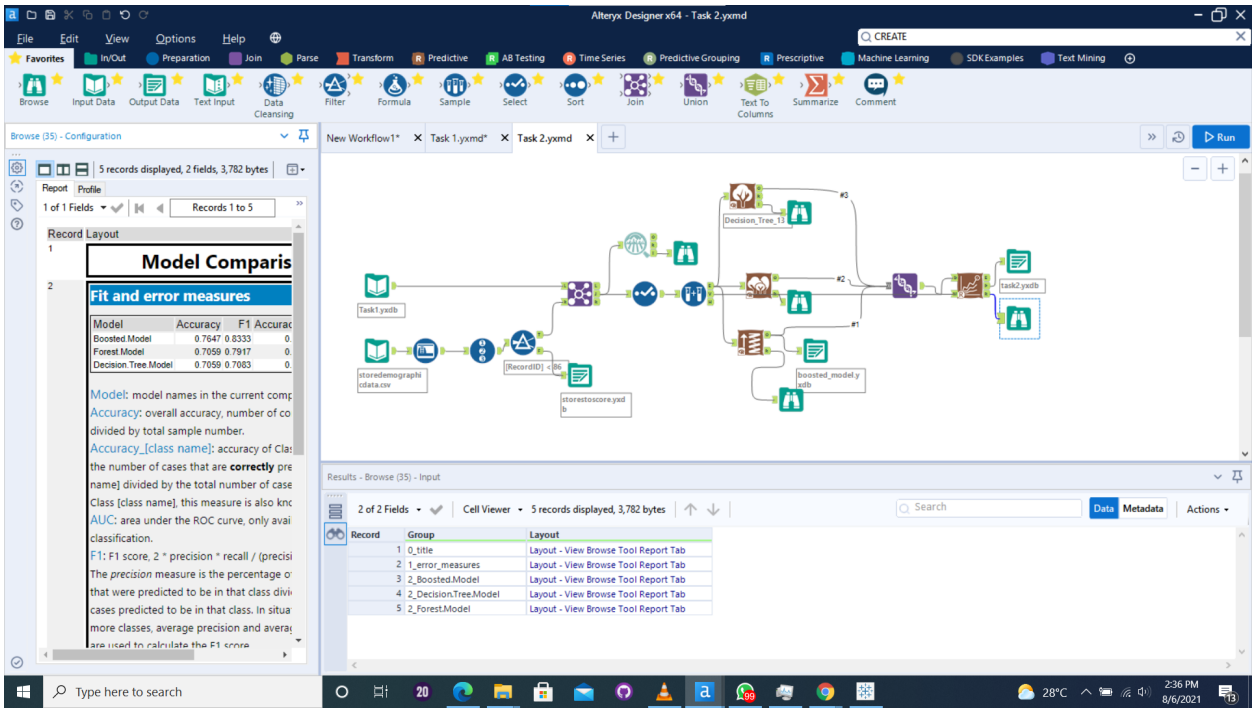
**Confusion matrix of Forest.Model**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 3 | 0 | 0 |

Type here to search | 28°C | 2:33 PM 8/6/2021

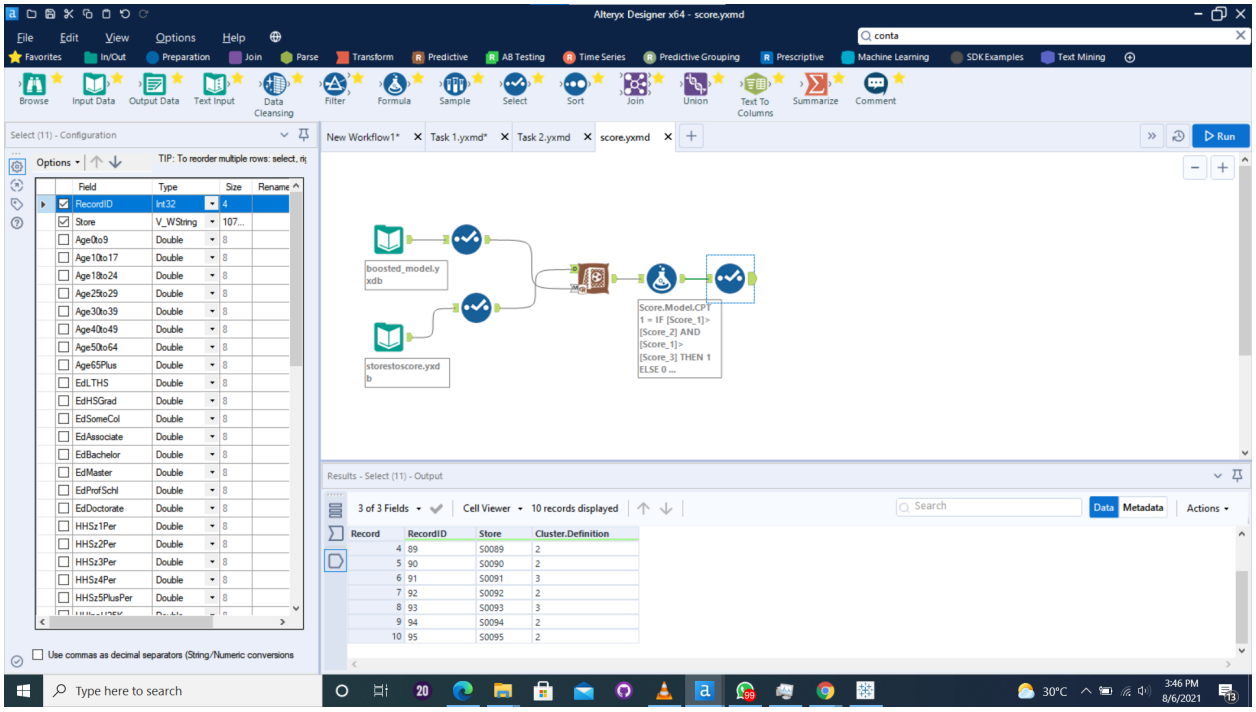
**"Aget0to9," "HVal750KPlus," and "Age65Plus" are the three most essential variables that assist explain the association between demographic indicators and store formats.**


Alteryx Designer x64 - Task 2.yxmd - Browse (40)
3 records displayed, 2 fields, 100 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 3
Record
Report
Report for Boosted Model Boosted.Model
Basic Summary:
Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829
Plots:
Variable Importance Plot
The user options for graphics width and height has been overridden for readability of y axis labels
Age0to9
HVal750KPlus
Age65Plus
EdHSGrad
PopOther
EdProfSchl
EdBachelor
Age10to17
PopBlack
HHSz2Per
HVal500Kto750K
HHInc250KPlus
PopWhite
Age25to29
EdSomeCol
HVal300Kto400K
HVal400Kto500K
Type here to search
28°C
2:37 PM
8/6/2021


Alteryx Designer x64 - Task 2.yxmd
File
Edit
View
Options
Help
CREATE
Favorites
In/Out
Preparation
Join
Parse
Transform
Predictive
AB Testing
Time Series
Predictive Grouping
Prescriptive
Machine Learning
SDK Examples
Text Mining
Browse
Input Data
Output Data
Text Input
Data Cleansing
Filter
Formula
Sample
Select
Sort
Join
Union
Text To Columns
Summarize
Comment
Browse (35) - Configuration
5 records displayed, 2 fields, 3,782 bytes
Report
Profile
1 of 1 Fields
Records 1 to 5
Record Layout
Model Comparis
Fit and error measures
Model
Accuracy
F1 Accurac
Boosted.Model
0.7647
0.8333
Forest.Model
0.7059
0.7917
Decision.Tree.Model
0.7059
0.7083
Model: model names in the current comp
Accuracy: overall accuracy, number of co
divided by total sample number.
Accuracy_[class name]: accuracy of Cla
the number of cases that are correctly pre
name] divided by the total number of case
Class [class name], this measure is also kn
AUC: area under the ROC curve, only avai
classification.
F1: F1 score, 2 * precision * recall / (precisi
The precision measure is the percentage o
that were predicted to be in that class divi
cases predicted to be in that class. In situa
more classes, average precision and avera
are used to calculate the F1 score
New Workflow1*
Task 1.yxmd*
Task 2.yxmd
Run
Task1.yxdb
storedemographicdata.csv
[RecordID] < 86
storestoscore.yxdb
Decision_Tree_13
boosted_model.yxdb
task2.yxdb
Results - Browse (35) - Input
2 of 2 Fields
Cell Viewer
5 records displayed, 3,782 bytes
Search
Data
Metadata
Actions
Record
Group
Layout
1
0_title
Layout - View Browse Tool Report Tab
2
1_error_measures
Layout - View Browse Tool Report Tab
3
2_Boosted.Model
Layout - View Browse Tool Report Tab
4
2_Decision.Tree.Model
Layout - View Browse Tool Report Tab
5
2_Forest.Model
Layout - View Browse Tool Report Tab
Type here to search
28°C
2:36 PM
8/6/2021

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

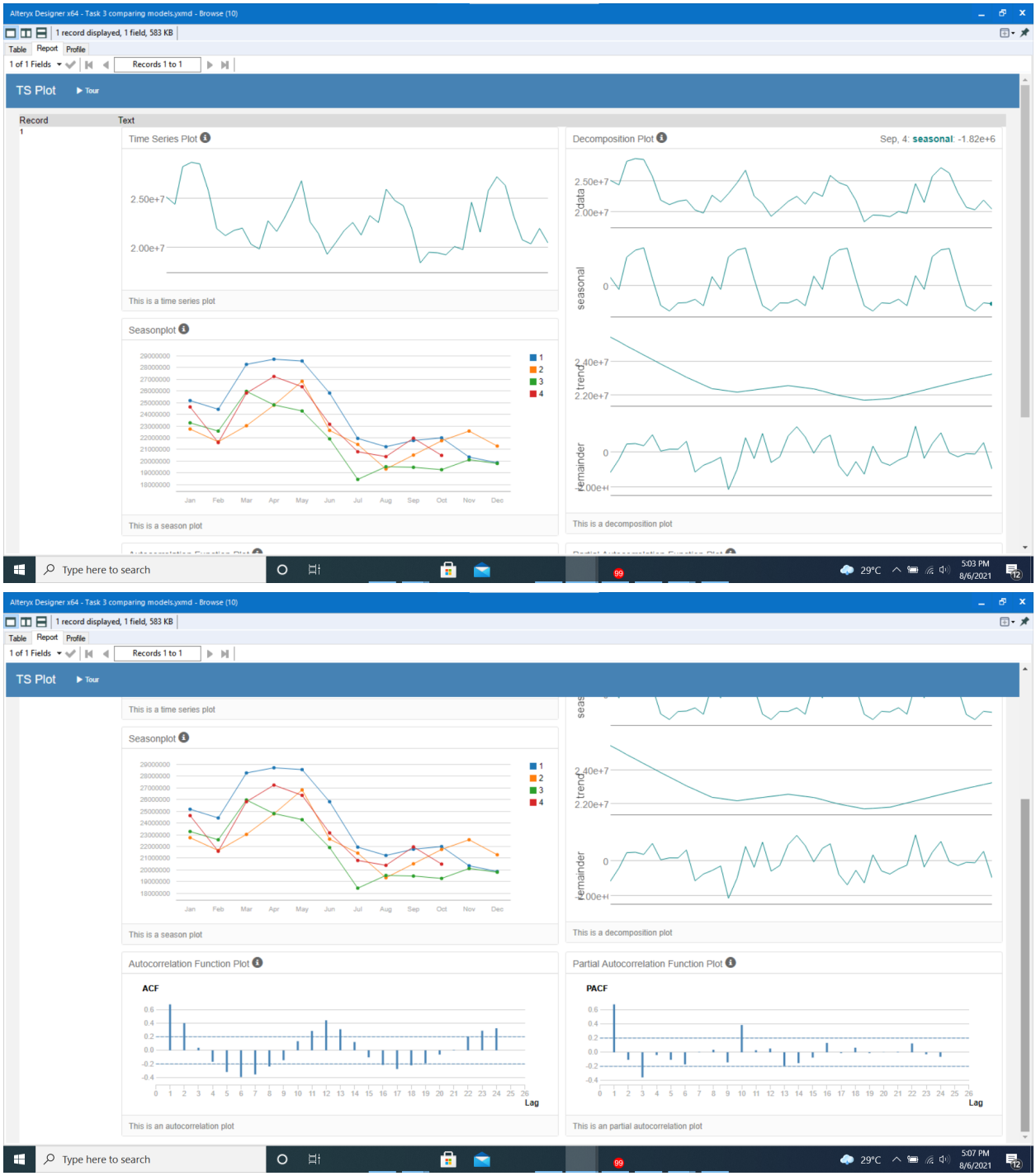
| Store Number | Segment |
|---|---|
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 3 |
| S0092 | 2 |
| S0093 | 3 |
| S0094 | 2 |
| S0095 | 2 |




# Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use foreach forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to thatdecision?

For forecasting, ETS (M.N.M) was chosen. When ETSand ARIMA are compared, ETS is found to be more accurate than ARIMA.




It was possible to determine trend, seasonality, and error components using the plot.

- ➔ In the error component, there is an irregular "variation" in "magnitude" with time. I chose the multiplicative parameter m.
- ➔ The trend plot moved downward at first, then upward, and then changed again. I chose none as a criterion (n)
- ➔ Regular pattern recurrence reveals "seasonality behavior." The magnitude of the spikes has

risen gradually over time. As a result, in this component, I considered using a m, multiplicative technique.

➔ ETS(m,n,m) was suggested by Alteryx

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |
| ARIMA | -492238.83 | 792197.3 | 735878.2 | -2.1992 | 3.3098 | 0.433 |

➔

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.

Forecast for Produce

| Month of Dat | ExistingStores | NewStores |
|---|---|---|
| January 2016 | 21,539,936 | 2,587,451 |
| February 2016 | 20,413,771 | 2,477,353 |
| March 2016 | 24,325,953 | 2,913,185 |
| April 2016 | 22,993,466 | 2,775,746 |
| May 2016 | 26,691,951 | 3,150,867 |
| June 2016 | 26,989,964 | 3,188,922 |
| July 2016 | 26,948,631 | 3,214,746 |
| August 2016 | 24,091,579 | 2,866,349 |
| September 2016 | 20,523,492 | 2,538,727 |
| October 2016 | 20,011,749 | 2,488,148 |
| November 2016 | 21,177,435 | 2,595,270 |
| December 2016 | 20,855,799 | 2,573,397 |