**Clustering**

Only use the percentage of total sales for each category as the variables in the K-Centroids Cluster Analysis tool. Do not use Total Sales, or any other variables.

**Classification**

The Cluster variable is a categorical variable. Cluster 3 is not 2 more of something than cluster 1. Therefore we must assign it the V\_String data type so that our classification models will treat it as categorical.

For the predictor variables for the classification models, use all of the variables from the demographics data set except for the Store field. Do not use any other variables as predictors in the classification models.

**Aggregations for Forecasting**

For existing stores we are trying to get the monthly total for past stores before forecasting. We can accomplish this using a summarize tool.

* Group by Year
* Group by Month
* Sum Produce.

Now we need to determine the best forecasting model and then forecast the next 12 months for the existing stores.

For new stores we are trying to get the average monthly total of a store per cluster. We can accomplish this using 2 summarize tools.

* Group by Store
* Group by Cluster
* Group by Year
* Group by Month
* Sum Produce.

And then

* Group by Cluster
* Group by Year
* Group by Month
* Avg Sum\_Produce.

We will be forecasting for each of the clusters and then multiplying the results by the number of new stores in that cluster. Then we will be adding all of these forecasts together on the same months to get a total forecast for all the new stores.

## 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 Submission

1. What is the optimal number of store formats? How did you arrive at that number?
2. How many stores fall into each store format?
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.

## 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.

### 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 Submission

* What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?
* What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.
* What format do each of the 10 new stores fall into? Please provide a data table.

## Task 3: Forecasting

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast.

### Task 3: Forecasting Produce Sales

You’ve been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, follow the steps below.

**Note:** Use a 6 month holdout sample for the TS Compare tool (this is because we do not have that much data so using a 12 month holdout would remove too much of the data)

**Step 1:** To forecast produce sales for existing stores you should aggregate produce sales across all stores by month and create a forecast.

**Step 2:** To forecast produce sales for new stores:

* Forecast **produce sales (not total sales)** for the average store (rather than the aggregate) for each segment.
* Multiply the average store produce sales forecast by the number of new stores in that segment.
* For example, if the forecasted average store produce sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
* Sum the new stores produce sales forecasts for each of the segments to get the forecast for all new stores.

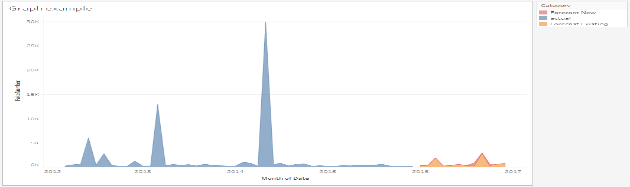
**Step 3:** Sum the forecasts of the existing and new stores together for the total produce sales forecast.

### Task 3 Submission

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?
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.

This is the table we are expecting to see filled out in this Task

This is an example of the type of graph we are expecting to see, this does not reflect the actual numbers you will get in your analysis.



Here is an example sheet (with fake data) of how to put your data together so that you can use it in

**Combining Predictive Techniques**

Overall

| CRITERIA | MEETS SPECIFICATIONS |
| --- | --- |
| Write up | The write up is written clearly, in complete sentences, and without major typos. |
| Visualizations | Several visualizations are included. All visualizations are clearly labeled and help answer the related questions. |

Task 1

| CRITERIA | MEETS SPECIFICATIONS |
| --- | --- |
| Number of store formats | Accurately identifies the correct number of formats and provides justification using the Adjusted Rand and CH indices. |
| Stores per format | Identifies the correct number of stores that fall into each store format. |
| Cluster differences | Provides one observation about the differences among clusters, and uses the results of the clusters to provide justification. |
| Cluster Map | Includes a map that shows the location of the stores, uses color to show cluster, and size to show total sales. A legend is used for both color and size. |

Task 2

| CRITERIA | MEETS SPECIFICATIONS |
| --- | --- |
| Model Selection | States the type of classification model used and adequately justifies the choice using at least one model comparison method. |
| New Store Formats | Includes a table that correctly identifies the format for each of the 10 new stores. |

Task 3

| CRITERIA | MEETS SPECIFICATIONS |
| --- | --- |
| Forecast Visualization | A table with the correct 12 month forecasts for existing and new stores is provided. A visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts is provided. |
| Model selection | Compares and identifies the best ETS or ARIMA model to use for forecasting. Justifies the decision by showing the plot and shows forecast error measurements against the holdout sample. |