k-Means Clustering

k-Means clustering is a popular clustering algorithm that groups items into k distinct clusters so that items within the same cluster are more similar to each other than items within different clusters.

When using the k-Means script, please keep in mind the following best practices for cluster analysis:

1) Clustering works best when items are clustered ***by their nature*** – similar to how you might find books, movies or music grouped together in the public library or a well-organized store. Using an item's characteristics will result in clusters being distinguished by the most discriminating features of the items they contain.
2) The Cluster model is really ***just the beginning*** of the analytical process: It's up to the model creator to study the resulting model and understand which characteristics differentiate one cluster from another.
3) Once the cluster model is built, then business measures, such as Revenue per Customer or Transaction Frequency, can be applied. In this way, the relationship between the intrinsic nature defined by the clusters and key performance indicators can be established.

For more detailed information on k-means clustering, please consult the k-Means Wikipedia page [3].

This R Script has two functional modes:

- The user specifies the exact number of clusters.
- The user specifies a maximum number of clusters, without specifying an exact number, then the model optimally clusters the data in such a way that the Average Silhouette Width [4] is maximized.

How to Deploy to MicroStrategy:

Prerequisite: Please follow the instructions in the R Integration Pack User Guide [1] for configuring your MicroStrategy environment with R and that the R Script functions have been installed in your MicroStrategy project(s).

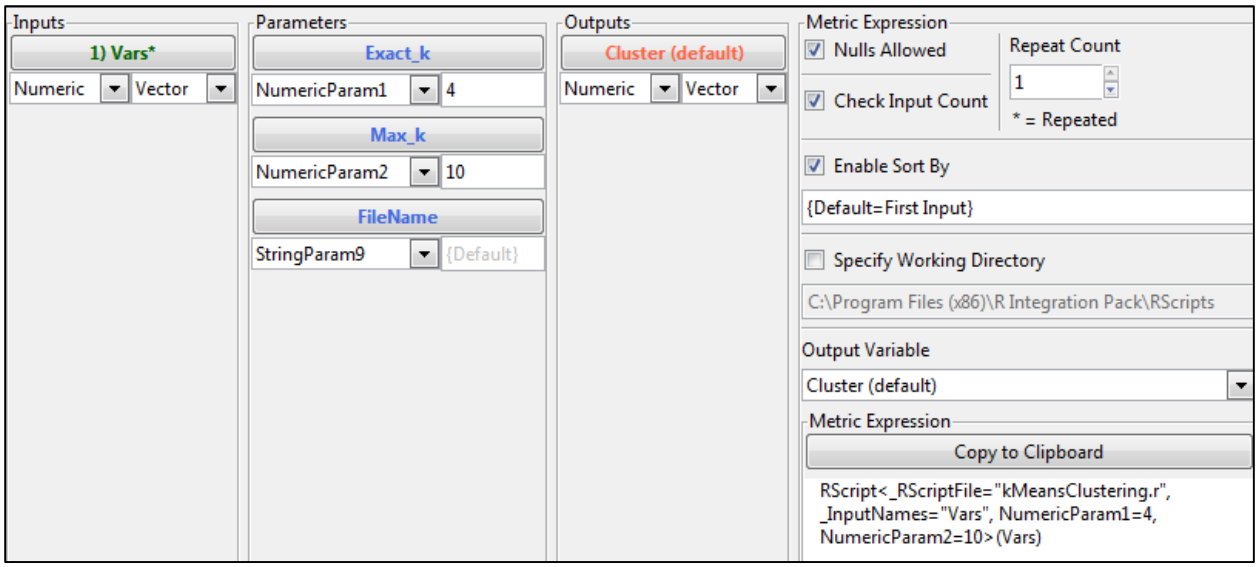
1) Download the kMeansClustering.R file from the Off-The-Shelf R Script Repository. [2].
2) From the R console, run the kMeansClustering.R script to verify the script runs correctly. For details, see the "**Running from the R Console**" section below.
3) Cut-and-paste the appropriate metric expression below in any MicroStrategy metric editor. Map the arguments to the appropriate MicroStrategy metrics. A description of each output is shown below.
4) Use the new metric in reports, dashboards and documents.

Metric Expressions:

1) **Cluster:** Returns the cluster to which the record belongs:

```
RScript<_RScriptFile="kMeansClustering.R",_InputNames="Vars",NumericParam1=4,Nu
mericParam2=10>(Vars)
```

Analytic Signature:



Inputs:

Vars: The numeric variables that are used to cluster the data. Since the Vars argument is a repeated input, it can be mapped to any number of MicroStrategy metrics. In this way, we enable our algorithm to consider any number of variables when generating predictions.

Parameters:

Exact_k: Uses NumericParam2 with a default of 4. This parameter specifies the number of clusters. If Exact_k is greater than 0, then the data will be grouped into Exact_k clusters. If not, then the number of clusters, less than or equal to the maximum number specified in Max_k, that best partitions the data will be used.

Max_k: Uses NumericParam1 with a default of 10. By using this parameter without a valid Exact_k value, the clustering algorithm will find the optimal number of clusters between 2 and the Max_k value.

FileName: Uses StringParam9 with a default of "". This parameter specifies the name of the .Rdata file used to persist relevant objects from the R environment isuch as the model and the data used to create it, allowing for additional inspection and analysis. Please note the R Script automatically appends the ".Rdata" file extension to this file name. If the default of ""is used, then no Rdata file is created.

Outputs:

Cluster: A numeric value representing the cluster to which that record was assigned.

Additional Results Generated by the R Script:

One file is stored in the working directory:

Rdata File: This file persists the state of several objects from the R environment for later inspection, analysis, and reuse, including df (a data frame containing the data read in from MicroStrategy), and model (the k-means model object).

Running from the R Console:

In addition to processing data from MicroStrategy during execution of a report or dashboard, the R script is also configured to run from the R console. Running the script for the R Console verifies that the script is functioning as expected, a good practice when initially deploying this analytic to a new system (for more details, see "Configuring dual execution modes" in [1]).

When run from the R Console, if the script is executing properly, a "Success!" message will appear in the console. If a "Success!" message does not appear, then please note the error in order to take appropriate action. For common pitfalls, please consult the **Troubleshooting** section below.

Troubleshooting:

This section covers certain situations you might encounter but it's not intended as a comprehensive list of possible errors

1) If an error occurs, the report will either fail with an error message, or return nulls as the output. In these cases, please refer to the RScriptErrors.log file generated for further guidance and the DSSErrors.log. Please consult the User Guide [1] and the R documentation for additional guidance.
2) If a non-numeric variable is passed in to the Vars argument, the report will fail with an error message indicating that a variable with unexpected type was passed in. This can be remedied by using all numeric variables as inputs.

Example:

If you're using MicroStrategy version 9.4.1 or higher and would like to try this example yourself, please download the StorePerformance.mstr file from this location and import it into your MicroStrategy environment:
http://download-codeplex.sec.s-msft.com/Download?ProjectName=rintegrationpack&DownloadId=837963

Store Data Layout:

A retail organization wants to identify poor performing stores in order to detect the store managers that are most in need of training. Traditionally, they have identified poor performers by using standard business measures such as Revenue or Revenue per Square Foot. Such measures simply assume that all stores have equal revenue potential which is not a valid assumption because of differences in the nature of each store and the communities they serve.

There are specific characteristics that would allow each store to be clustered by its nature. These characteristics can include the size and age of the store plus the population density, median income and the number of local competitors in its area. It would be unfair to focus solely on the stores with the lowest

overall performance since stores in more populous and prosperous areas would have an advantage over stores in smaller, less affluent locations or locations with higher numbers of competitors. Instead, by clustering stores by their nature, the performance of stores within each cluster can be easily compared and the lowest performers identified.

Using the Store Characteristic grid, let's build a four-cluster k-Means clustering model, by inserting a new metric named "Cluster (R)" with this expression that returns the cluster for each store and then add it to the grid.

```
RScript<NumericParam1=4, NumericParam2=10,_RScriptFile="kMeansClustering.R">([Median
Income],Population,[Store Sqft],[Store Age],Competitors)
```

This is the result:

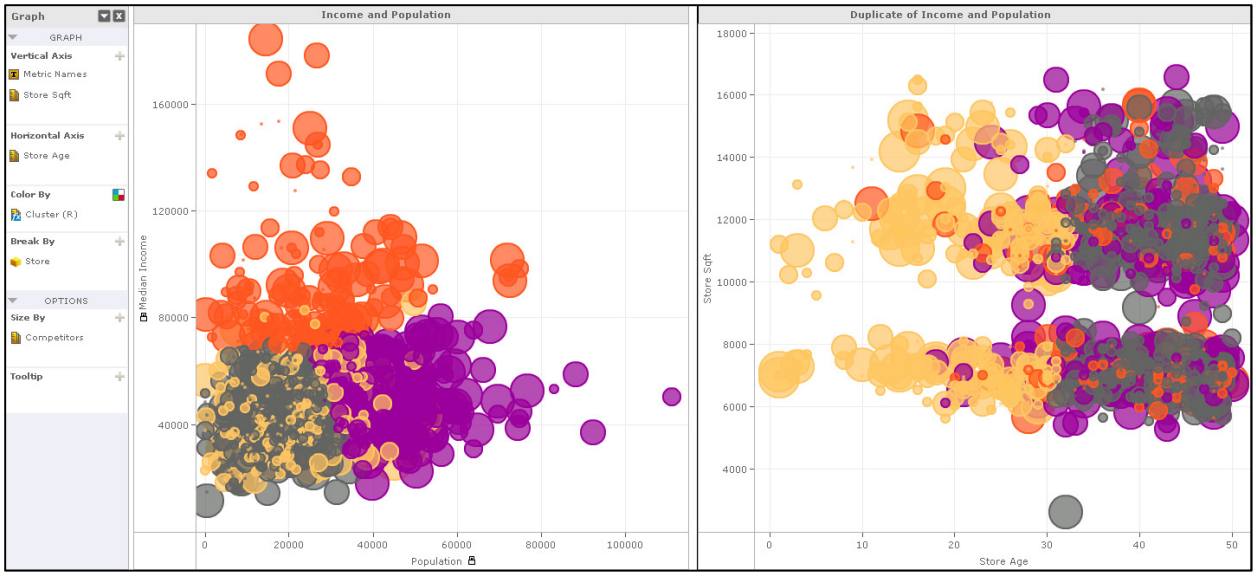
Stores Details							
Store	City	Store Sqft	Population	Competitors	Store Age	Median Income	Clusters (R)
10001	DOTHAN	7,109.0	35759	6	14	$38,007.9	3
10005	BOAZ	6,957.2	15693	4	14	$33,993.4	3
10011	BIRMINGHAM	12,605.3	19484	1	8	$54,491.3	3
10012	FORT PAYNE	11,907.0	4575	2	32	$38,898.5	4
10019	SHEFFIELD	7,500.6	9042	3	45	$37,147.4	4
10029	OPELIKA	14,256.2	22214	4	46	$37,923.5	4
10034	TALLASSEE	6,548.6	13179	4	38	$41,203.2	4
10038	ANNISTON	8,035.9	20150	3	33	$25,135.5	4
10039	HUNTSVILLE	7,114.0	21509	1	41	$60,686.3	4
10054	DECATUR	12,097.5	34434	5	40	$38,537.5	1
10055	DOTHAN	11,199.2	14470	3	31	$60,075.0	3
10056	BIRMINGHAM	11,461.6	19686	1	29	$31,883.6	3
10069	EUFAULA	7,502.2	13277	2	43	$27,106.0	4
10078	ANNISTON	10,973.1	19801	3	43	$50,954.2	4
10083	FOLEY	6,861.8	26767	3	33	$39,140.6	4
10085	DECATUR	5,428.7	30545	4	32	$54,565.1	1
10087	MOBILE	14,512.0	13967	1	33	$23,747.2	4
10092	TUSCALOOSA	6,504.0	39878	5	42	$17,806.3	1
10100	FAIRHOPE	7,495.6	27829	5	44	$54,295.8	1
10101	TALLADEGA	7,413.7	26822	2	41	$35,734.3	4
10108	PRATTVILLE	7,971.6	26954	3	14	$52,862.8	3
10109	CARROLLTON	7,415.7	3460	1	34	$27,124.4	4
10112	DEMOPOLIS	7,931.3	8395	1	26	$35,480.0	3
10113	MOBILE	11,772.1	14269	1	40	$48,565.5	4
10114	BESSEMER	6,784.0	27139	3	30	$25,590.5	3
10118	SELMA	11,060.5	24268	4	29	$33,488.1	3
10125	HALEYVILLE	12,903.6	13485	5	35	$31,198.1	4
10131	HUNTSVILLE	6,753.7	21509	4	34	$60,686.3	1
10144	MOBILE	6,415.9	33607	5	29	$49,567.9	1
10150	GREENVILLE	11,918.0	14074	5	33	$31,849.0	4
10152	MOBILE	12,334.3	17847	3	8	$55,482.6	3
10158	RUSSELLVILLE	6,706.8	11150	4	40	$31,015.6	4
10164	SYLACAUGA	10,951.5	18472	5	24	$33,340.4	3
20006	PALMER	7,070.1	25176	5	25	$73,230.0	2
20017	ANCHORAGE	16,574.8	35857	4	44	$56,888.5	1
20026	ANCHORAGE	10,947.2	35857	4	47	$56,888.5	1
30002	PHOENIX	12,828.1	25742	5	40	$21,619.9	4
30006	TUCSON	11,865.1	32666	4	32	$34,312.3	1

Clusters Layout:

Observe that this layout contains a scatterplot showing Stores by Median Income and Population sizing the stores by the number of local competitors. Add the Cluster metric that was just created to the Color By field. Set the coloring to be Custom by value with Purple for 1, Orange for 2, Tan for 3, and Gray for 4.

Now, duplicate this visualization. Replace Median Income on the y-axis with Store SqFt, and replace Population on the x-axis with Store Age.

This layout now contains two scatterplots. The leftmost scatterplot shows Stores by Median Income and Population; the rightmost scatterplot shows Stores by Age and Sq. Feet. Stores in both plots are sized by the number of local competitors. The following visualizations result:



From the left side plot, it's clear that the orange stores in Cluster 2 consist of stores in areas with higher median income and the purple Cluster 1 stores dominate areas with higher population density. Gray Cluster 4 and Tan Cluster 3 stores are mixed together as stores that are in less affluent and less populated areas.

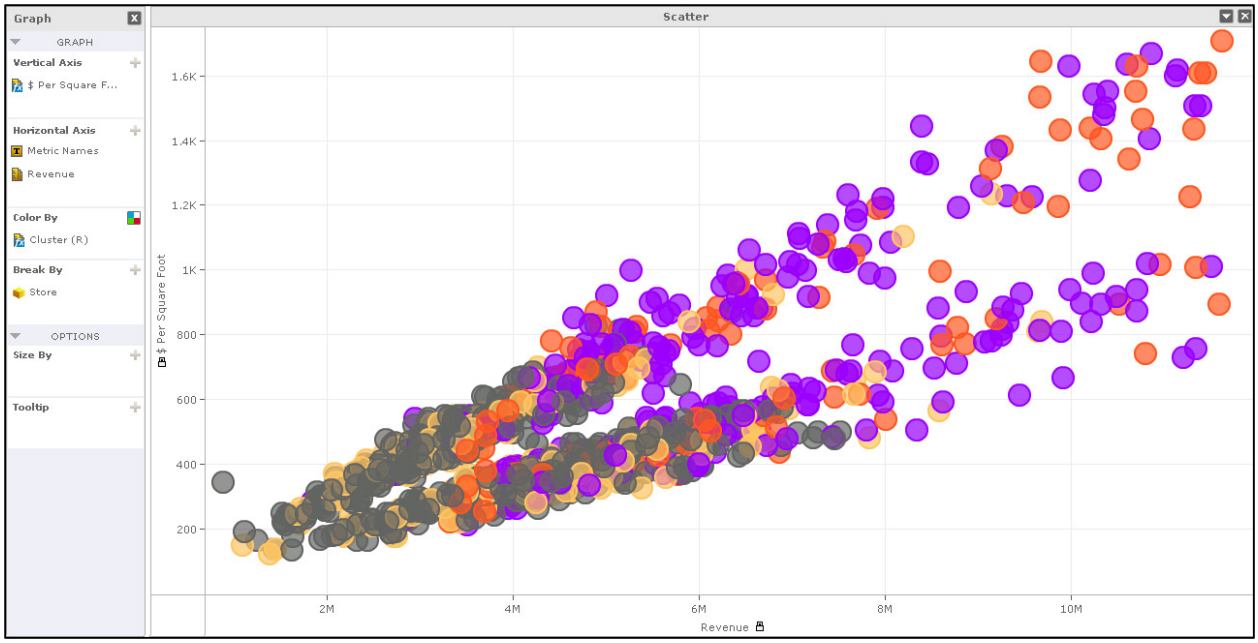
Using the right side plot, it's easy to distinguish Cluster 3 stores from Cluster 4 stores. Add the Cluster metric to the filter panel and choose where Cluster is greater than 2. Tan Cluster 3 stores are clearly lower in age than the gray Cluster 4 stores. So while both clusters 3 and 4 represent stores in relatively poorer, smaller towns, Cluster 3 stores are newer than Cluster 4.

As it turns out, the number of competitors and the size of the store were not needed to describe each cluster. The nature of the local population and the age of the store are the key characteristics identified by the k-Means algorithm to cluster stores effectively.

Now that the nature of the clusters is understood, we can apply business measures in order to identify the lowest performers in each cluster.

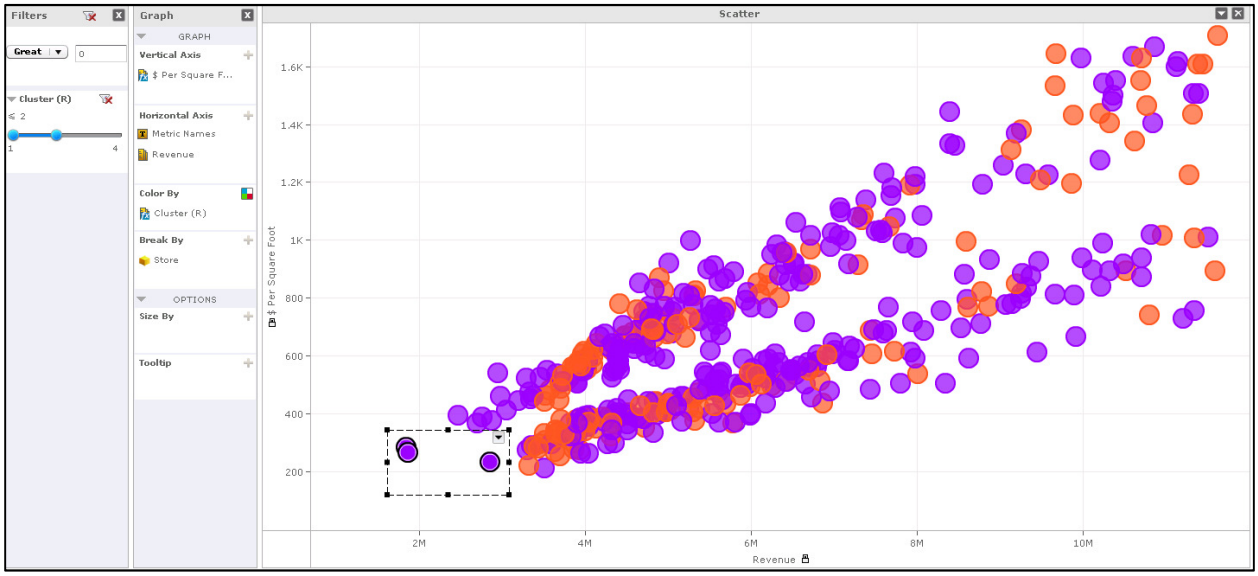
Poor Performers Layout:

On this layout, observe that there are only a handful of stores with Revenue below $2M and Revenue per Sq. Foot less than $200. Intuitively, we can identify these stores as poor performers and administer our training to the managers of these stores.



Now that we have clustered the stores, we can see which stores have lower Revenue and Revenue per Sq. Feet compared to their peers. In essence, we can use the clustering results to benchmark what the expected Revenue should be.

Add the Cluster (R) clustering metric that was created to the Color By field, and the filtering area. Set the Cluster (R) filter to Clusters 1 and 2. Observe that there are now several stores that although they have revenue greater than $2M and Revenue per Sq. Ft. greater than $200, are clearly performing poorer than their peers. These stores should be considered poor performers and their managers should also receive training. Without the clustering, we would have missed that these stores performed poorly.



Next Steps:

As noted in the introduction section, the identification of clusters is just the beginning of the analysis. The next step would be to analyze how each cluster performs based on key performance indicators such as Revenue, Profit, and Sales Volume. For users of MicroStrategy Analytics Enterprise, this process can be accomplished in either of two ways:

- Creating a data mart from the Clustering report and modeling the Cluster output as an attribute.
- Creating a custom group based on the Cluster output.

Once the Cluster output has been either modeled as an attribute or used to define a custom group, it can be linked to business performance by placing the attribute or custom group on rows and KPIs of interest on the columns. With this report, you can see which clusters are most and least beneficial to your organization.

Please see Chapters 4 and 13 of the MicroStrategy Advanced Reporting Guide for more information on how to use the outputs of a clustering model to analyze performance relative to important business measures.

References:

1) MicroStrategy R Integration Pack User Guide: https://rintegrationpack.codeplex.com/documentation
2) R Integration Pack Analytic Shelf: Off-The-Shelf R Script Repository: https://rintegrationpack.codeplex.com/wikipage?title=R%20Script%20%22Shelf%22&referringTitle=Home#
3) http://en.wikipedia.org/wiki/K-means_clustering
4) http://en.wikipedia.org/wiki/Silhouette_(clustering)

