Multivariate Outliers Tile

Introduction

This page explains what the multivariate outlier detection algorithm does and what it is used for. It also provides step-by-step instructions on how to implement it in DOMO's Magic ETL interface using a sample DataSet.

Multivariate outlier detection is an anomaly detection algorithm which aims to detect outlying or unusual observations on a set of one or more numeric columns in a DataSet. Other outlier detection algorithms available in DOMO include parametric, nonparametric, and time series outlier detection methods.

Unlike the parametric, nonparametric, and time series outlier detection methods found in DOMO, multivariate outlier detection allows the user to detect an outlying observation (or row) in more than one dimension. In some scenarios, an observation may not be an outlier with respect to a single column but may be an outlier with respect to multiple columns. This is particularly useful as the number of columns increases.

For each row of a DataSet, the Mahalanobis distance measure is computed which provides a pseudo measure of how extreme the values in that row are. This Mahalanobis distance measure is then compared against a prespecified quantile of an appropriately chosen Chi-Square distribution to determine if the row is an outlier or not. Observations in the DataSet are labeled as outliers if the distance measure is greater than the pre-specified quantile.

Configuring This Action in Magic ETL

The following example illustrates how the outlier detection algorithm can be implemented and used in Magic ETL in Domo. The sample DataSet, "Wholesale_Distributor_Sales.csv," (440 rows) is publically available from the UCI Machine Learning Repository.

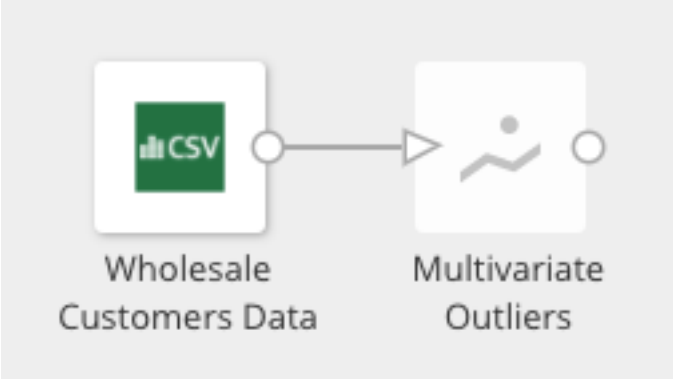
Wholesale Customers Data CONFIGURATION DETAILS DATA

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
1	2	3	12,669	9,656	7,561	214	2,674	1,338
2	2	3	7,057	9,810	9,568	1,762	3,293	1,776
3	2	3	6,353	8,808	7,684	2,405	3,516	7,844
4	1	3	13,265	1,196	4,221	6,404	507	1,788
5	2	3	22,615	5,410	7,198	3,915	1,777	5,185
6	2	3	9,413	8,259	5,126	666	1,795	1,451
7	2	3	12,126	3,199	6,975	480	3,140	545

It contains annual spending information (in monetary units, “m.u.”) on various product categories from clients of a wholesale distributor. Each row contains data for a single client. The first two columns are Region and Channel, which describe demographic information for each client. The next 6 columns contain the spending information on each of the 6 product categories: Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen. Multivariate outlier detection can be used to detect outlying, or unusual, clients based to how much was spent on the various product categories.

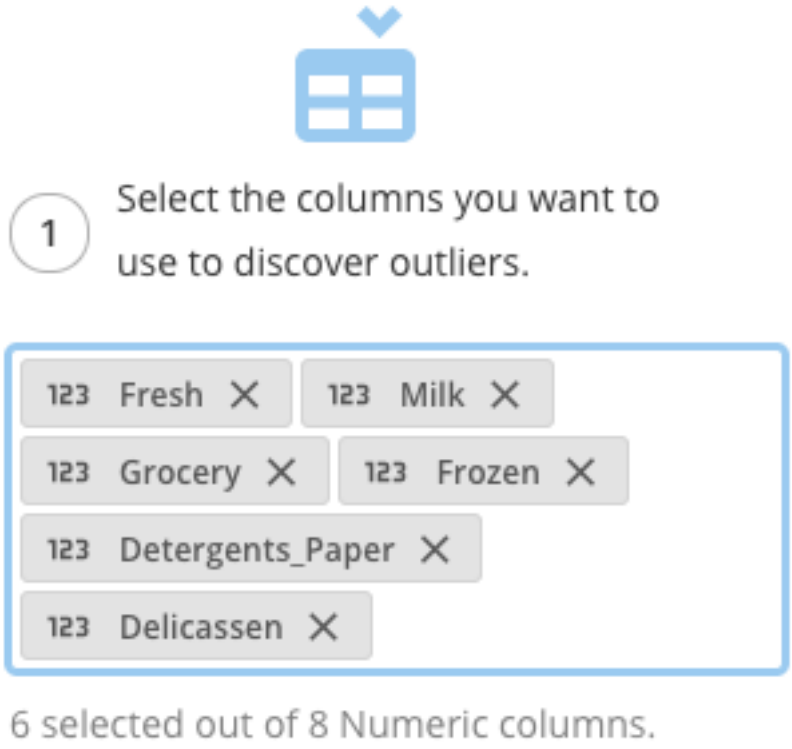
Step 1: Add tiles in Magic ETL

In Magic ETL, the first step is to load the DataSet “Wholesale_Distributor_Sales.csv,” and then connect it to the Multivaraite Outliers Data Science tile.



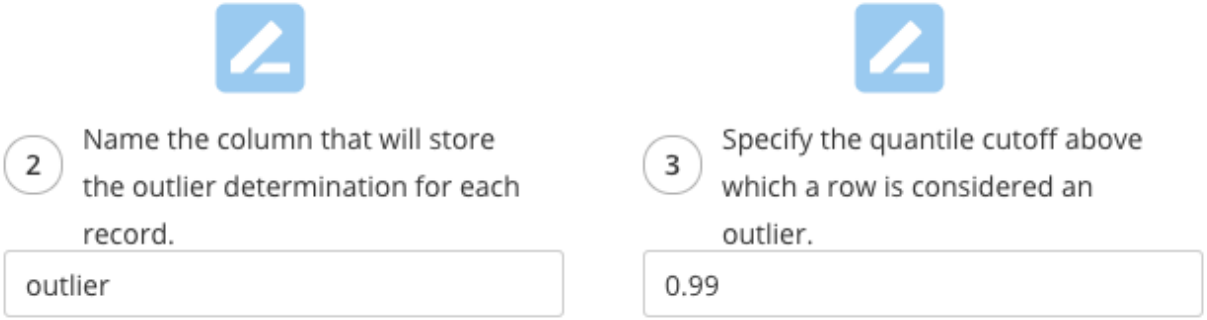
Step 2: Select columns for outlier detection

Within the Multivariate Outliers tile, one or more of the 6 product categories must be chosen as the columns for which outliers will be detected on. For this example, all of the product categories are chosen.



Step 3: Name the column and select the quantile cutoff

The quantile (a value between 0 and 1) of the Chi-square distribution that will be used as a cutoff must be selected next as well as the name of the column (default is “outlier”) that contains either TRUE (observation is an outlier) or FALSE (observation is not an outlier) values. Typically a quantile between .95 and .99 is a good starting point, with higher quantiles leading to stricter cutoffs. It is recommended that different values be explored. Note that using too small of a cutoff will label most of observations as outliers.



Step 4: Add an Output DataSet tile

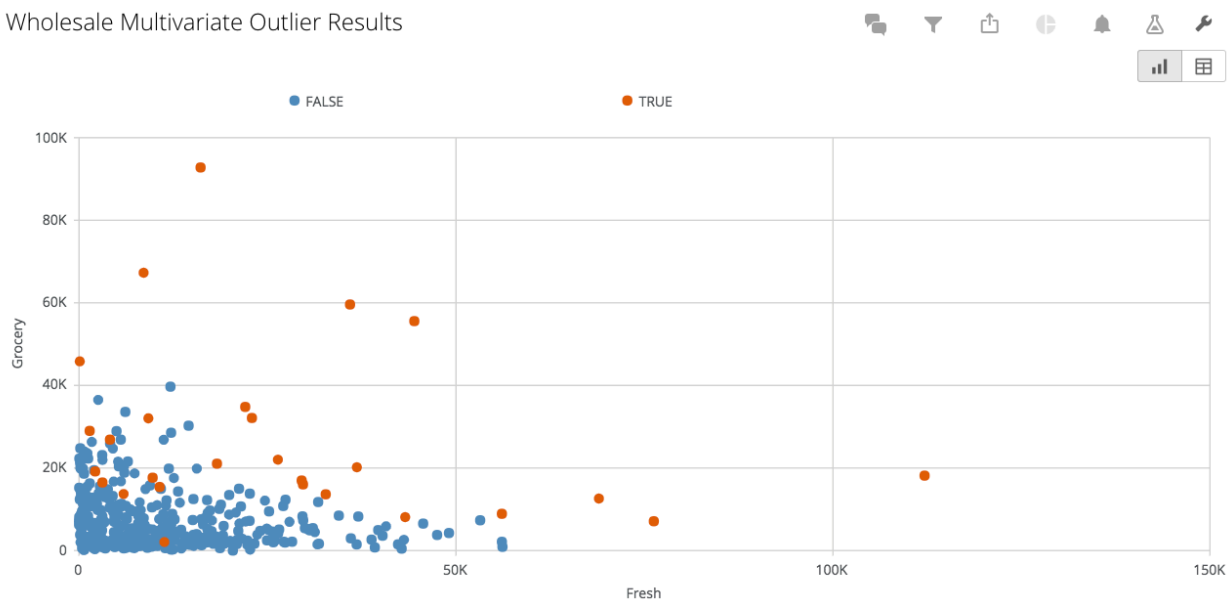
Last, an output DataSet must be connected and named.

The resulting output DataSet will include the original DataSet with an appended column containing the outlier indicators (circled below in the preview pane).

Frozen	Detergents_Paper	Delicassen	outlier
214	2674	1338	FALSE
1762	3293	1776	FALSE
2405	3516	7844	FALSE
6404	507	1788	FALSE
3915	1777	5185	FALSE
666	1795	1451	FALSE

Building a Card with This DataSet

A Scatter Plot graph is a great way to visualize the data in this DataSet.



This Scatter Plot shows client spending on Grocery (X-axis) products against Fresh (Y-axis) products. Recall that all six product categories were chosen to detect outliers on. The clients that are considered outliers are in red. Most of the outlying clients either spent a lot on Fresh or Grocery, or both. There also appeared to be some outlying clients that did not spend very much on either Fresh or Grocery (located in the bottom left corner of the plot). These clients may be high spenders in one of the other four product categories. Different scatter plots could be used to gain insight into why some of the clients were considered outliers.