PhUSE 2017

Paper DV04

# The Art of Data Visualization: Detecting Multivariate Data Outliers Using an Interactive Approach

Diane Peers, GCE Solutions, Manchester, UK
Lorin Miller, GCE Solutions, Buffalo, NY USA

## ABSTRACT

Successfully detecting outliers in multivariate data requires statistical and programming skills and can be very time consuming. Requests for outlier detection can come from different skills groups therefore it is more efficient and effective to allow users to interact directly with the data themselves. We have developed an interactive, web based data visualization application for outlier detection using R Shiny that does not require specialist knowledge to use. This application reads in various file types, manipulates and reduces datasets as needed, performs an array of different outlier detection methods, visually displays the outlier output through interactive graphs and downloadable tables, and provides tests to check any distributional assumptions. This paper visually demonstrates the functionality of the application, which includes exporting all or a subset of the outliers displayed and utilizing machine learning techniques to better predict outliers based on prior decisions made by the user.

## INTRODUCTION

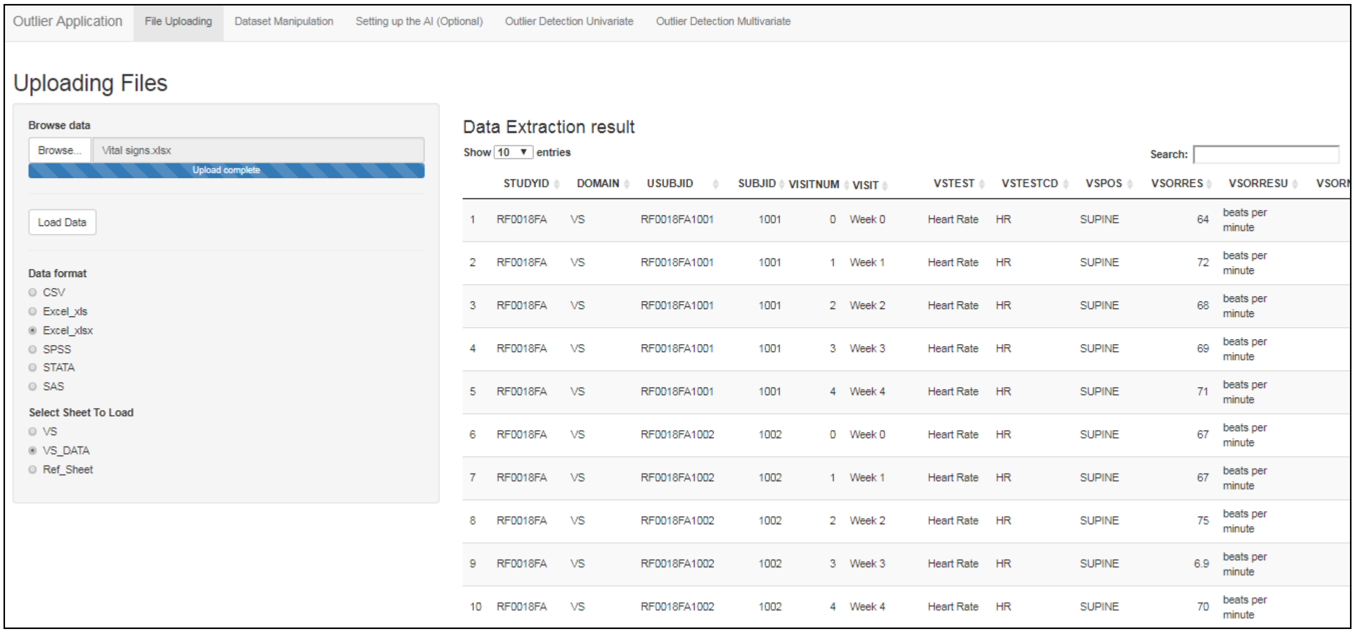
An outlier is an observation that appears to deviate markedly from other observations in the sample. Outlier detection is necessary for two key reasons: firstly, for data cleaning, so that potentially corrupt or inaccurate records can be identified and secondly, to ensure that extreme values are appropriately addressed so as not to skew the analyses. It can be a tedious task deciding which outlier detection method to use and reviewing the validity of tests through the checking of distributional assumptions. It also traditionally requires a programmer or statistician to execute the decided technique and produce any supportive visuals, which can be problematic for less technical professionals or those uncertain as to the options/methods available.

Detecting outliers is necessary but can be time and labor intensive. We have streamlined this process with an interactive tool that leads to earlier detection of outliers and hence more time left to focus on the primary analysis. Our outlier application speeds up outlier detection tasks by utilizing flexible data file types, accompanied by the ability to manipulate the dataset as necessary and apply an array of outlier techniques for ease of the user. It also checks any distributional assumptions, and provides different techniques based on the dimensionality of the data. All of the techniques provide supporting visuals before and after the outliers have been removed to further assist in the decision making.

## PREPARING THE DATA

Data is often provided in various file formats (CSV, Excel, SAS datasets, etc.), so the application provides a suite of options to read in different types of data. There is also the ability to further prepare the dataset as needed, such as filtering, removing columns and transposing the data, all of which is done in a step by step process.

For example, suppose one has a vital signs SDTM dataset (shown on the following page), that needs transposing in order to create separate variables for each vital signs measurement to allow us to then detect outliers.


Outlier Application
File Uploading
Dataset Manipulation
Setting up the AI (Optional)
Outlier Detection Univariate
Outlier Detection Multivariate

Uploading Files

Browse data

Browse... Vital signs.xlsx

Upload complete

Load Data

Data format

○ CSV
○ Excel_xls
● Excel_xlsx
○ SPSS
○ STATA
○ SAS

Select Sheet To Load

○ VS
● VS_DATA
○ Ref_Sheet

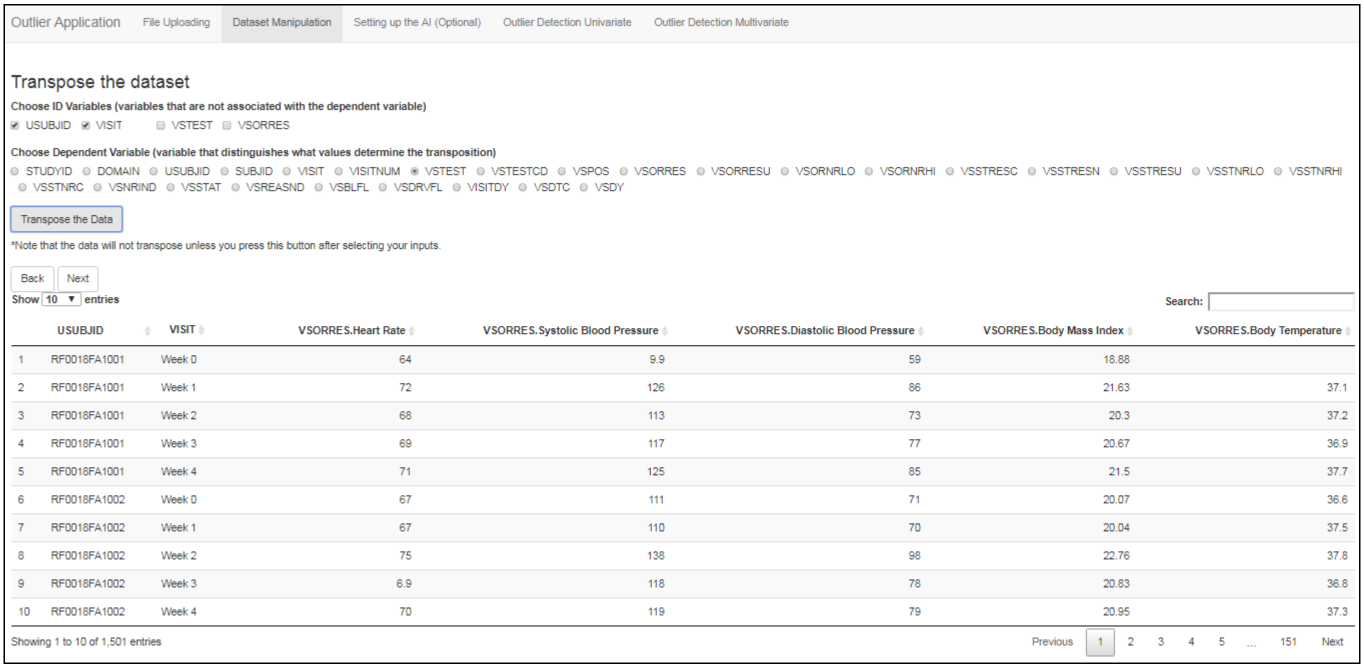
Data Extraction result

Show 10 entries

Search:

| | STUDYID | DOMAIN | USUBJID | SUBJID | VISITNUM | VISIT | VSTEST | VSTESTCD | VSPOS | VSORRES | VSORRESU | VSORI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | RF0018FA | VS | RF0018FA1001 | 1001 | 0 | Week 0 | Heart Rate | HR | SUPINE | 64 | beats per minute | |
| 2 | RF0018FA | VS | RF0018FA1001 | 1001 | 1 | Week 1 | Heart Rate | HR | SUPINE | 72 | beats per minute | |
| 3 | RF0018FA | VS | RF0018FA1001 | 1001 | 2 | Week 2 | Heart Rate | HR | SUPINE | 68 | beats per minute | |
| 4 | RF0018FA | VS | RF0018FA1001 | 1001 | 3 | Week 3 | Heart Rate | HR | SUPINE | 69 | beats per minute | |
| 5 | RF0018FA | VS | RF0018FA1001 | 1001 | 4 | Week 4 | Heart Rate | HR | SUPINE | 71 | beats per minute | |
| 6 | RF0018FA | VS | RF0018FA1002 | 1002 | 0 | Week 0 | Heart Rate | HR | SUPINE | 67 | beats per minute | |
| 7 | RF0018FA | VS | RF0018FA1002 | 1002 | 1 | Week 1 | Heart Rate | HR | SUPINE | 67 | beats per minute | |
| 8 | RF0018FA | VS | RF0018FA1002 | 1002 | 2 | Week 2 | Heart Rate | HR | SUPINE | 75 | beats per minute | |
| 9 | RF0018FA | VS | RF0018FA1002 | 1002 | 3 | Week 3 | Heart Rate | HR | SUPINE | 6.9 | beats per minute | |
| 10 | RF0018FA | VS | RF0018FA1002 | 1002 | 4 | Week 4 | Heart Rate | HR | SUPINE | 70 | beats per minute | |

We can transpose the data by setting ID variables which define what a unique row key should be. In this example, we would use the variables USUBJID and VISIT. We would then set the dependent variable i.e. the variable that determines which new columns we would like to have depending on the values in that variable, which in this case is VSTEST. Now after transposing the data, every variable is separated out as a column by the test (VSTEST) making it easier to detect outliers for each vital signs measurement, as required. Using this tool, the data preparation process does not require any programming knowledge, only an understanding of the data and familiarity with the tool/process is necessary.


Outlier Application
File Uploading
Dataset Manipulation
Setting up the AI (Optional)
Outlier Detection Univariate
Outlier Detection Multivariate

Transpose the dataset

Choose ID Variables (variables that are not associated with the dependent variable)

☑ USUBJID ☑ VISIT □ VSTEST □ VSORRES

Choose Dependent Variable (variable that distinguishes what values determine the transposition)

○ STUDYID ○ DOMAIN ○ USUBJID ○ SUBJID ○ VISIT ○ VISITNUM ● VSTEST ○ VSTESTCD ○ VSPOS ○ VSORRES ○ VSORRESU ○ VSORNRLO ○ VSORNRHI ○ VSSTRESC ○ VSSTRESN ○ VSSTRESU ○ VSSTNRLO ○ VSSTNRHI ○ VSSTNRC ○ VSNRIND ○ VSSTAT ○ VSREASND ○ VSBLFL ○ VSDRVFL ○ VISITDY ○ VSDTC ○ VSDY

Transpose the Data

*Note that the data will not transpose unless you press this button after selecting your inputs.

Back Next

Show 10 entries

Search:

| | USUBJID | VISIT | VSORRES.Heart Rate | VSORRES.Systolic Blood Pressure | VSORRES.Diastolic Blood Pressure | VSORRES.Body Mass Index | VSORRES.Body Temperature |
|---|---|---|---|---|---|---|---|
| 1 | RF0018FA1001 | Week 0 | 64 | 9.9 | 59 | 18.88 | |
| 2 | RF0018FA1001 | Week 1 | 72 | 126 | 86 | 21.63 | 37.1 |
| 3 | RF0018FA1001 | Week 2 | 68 | 113 | 73 | 20.3 | 37.2 |
| 4 | RF0018FA1001 | Week 3 | 69 | 117 | 77 | 20.67 | 36.9 |
| 5 | RF0018FA1001 | Week 4 | 71 | 125 | 85 | 21.5 | 37.7 |
| 6 | RF0018FA1002 | Week 0 | 67 | 111 | 71 | 20.07 | 36.6 |
| 7 | RF0018FA1002 | Week 1 | 67 | 110 | 70 | 20.04 | 37.5 |
| 8 | RF0018FA1002 | Week 2 | 75 | 138 | 98 | 22.76 | 37.8 |
| 9 | RF0018FA1002 | Week 3 | 6.9 | 118 | 78 | 20.83 | 36.8 |
| 10 | RF0018FA1002 | Week 4 | 70 | 119 | 79 | 20.95 | 37.3 |

Showing 1 to 10 of 1,501 entries

The display above shows the vital signs dataset after the transpose.

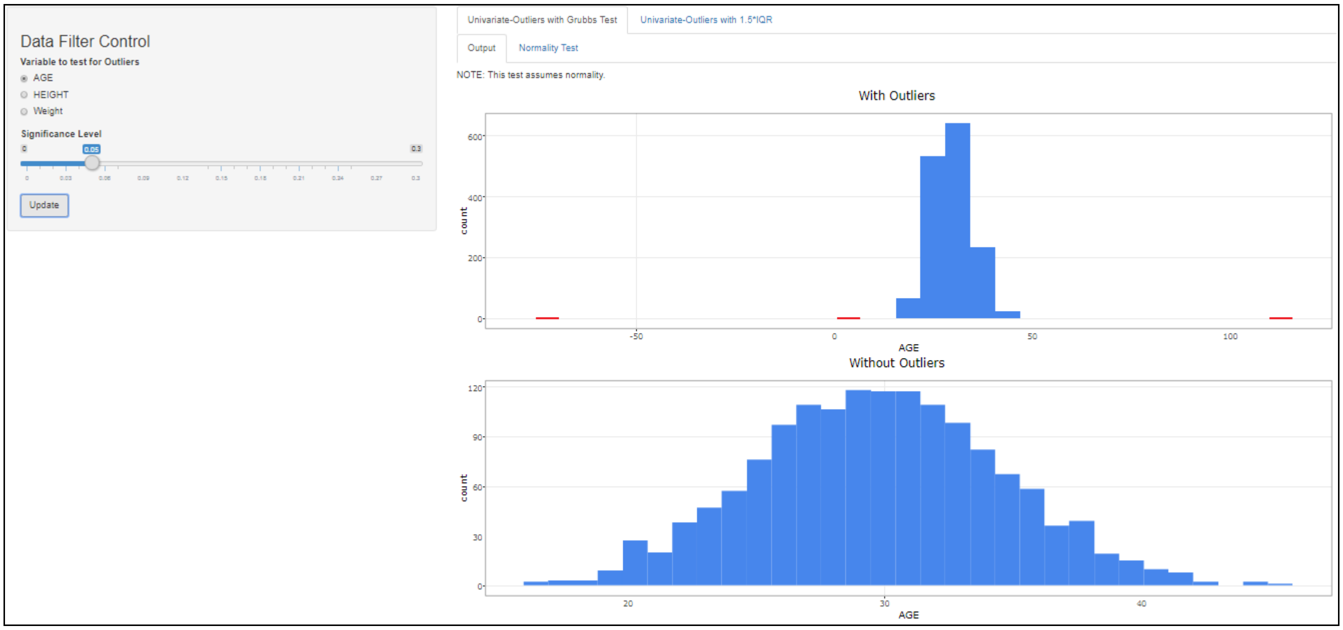
## DIFFERENT OUTLIER DETECTION TECHNIQUES

The best technique to use for outlier detection will depend on the dimensionality (number of variables used to detect outliers), variable data types (i.e. numeric versus categorical), and distribution of the data. It is also valuable to compare various techniques and see if any observations are continually getting flagged as outliers. Therefore, our application contains options to use several different techniques and to then compare the results. Additionally, each technique is provided with a high-level explanation to assist non-statisticians in understanding the theory behind each strategy.

The visuals on the various techniques should be consistent and fluid for ease of understanding. The outliers also need to be clearly visible to the user in the displays. Our application has implemented a synchronized theme of blue to represent normal observations and red to identify outlier observations. This theme is displayed on the plots produced, displaying the data with and without outliers. Additionally, most plots have "hover" abilities so that the user can move their mouse pointer over the output to easily display which observations correspond to the points on the plots. Following the identification of the outliers, a table is then produced of all the outlier observations with the variables highlighted in yellow that were used in the detection method. This table of observations with outlier values can then be exported to a dataset which could then, for example, be given to data management for investigation as potentially incorrect values.

### UNIVARIATE APPROACHES

The best ways to visualize univariate data is either by a histogram or a box plot. Our application uses two techniques to identify outliers, Grubb's Test and Inter-Quartile Range approach, and then displays the data pictorially. Grubbs test (also known as the maximum normed residual test) is used to find a single outlier in normally distributed data based on the number of standard deviations away from the mean. Once an outlier has been detected, it is expunged from the dataset and the test is iterated until no outliers are detected. Note that the significance level can be adjusted as well, and while the default is the standard 0.05 to achieve 95% confidence, this can be adjusted as needed. A normality test is applied to the data so one can check the validity of the distributional assumption, however, even on data that is not normally distributed, Grubb's test will typically still flag values that are extreme outliers. The data is then displayed with and without outliers via a histogram as shown below:



Data Filter Control

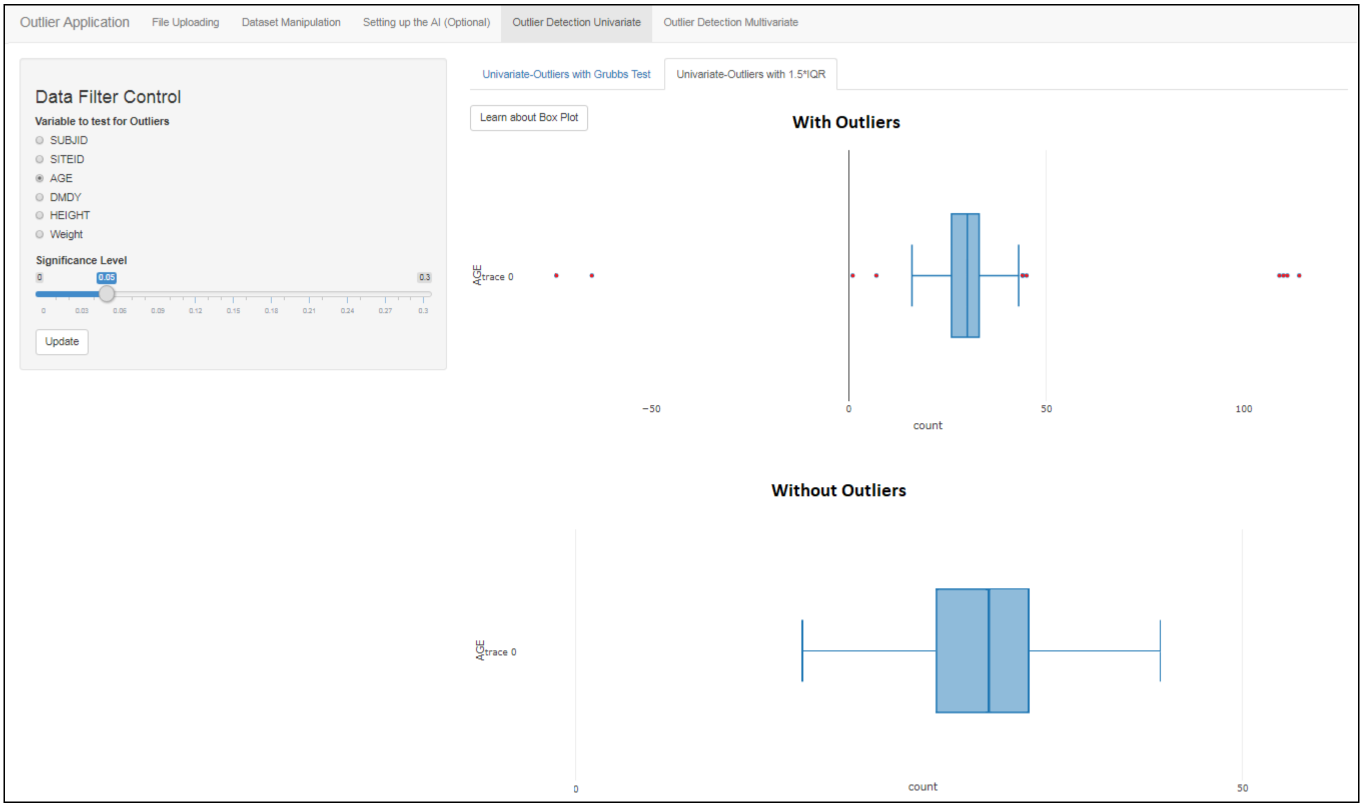
Variable to test for Outliers

● AGE
○ HEIGHT
○ Weight

Significance Level: 0.05

NOTE: This test assumes normality.

The second technique, Inter-Quartile Range (IQR) approach, uses a box plot to identify outlier values based on 1.5*IQR, which simply flags any observations that are greater or less than 1.5 times the difference of the third and first quartile from the median:


Outlier Application
File Uploading
Dataset Manipulation
Setting up the AI (Optional)
Outlier Detection Univariate
Outlier Detection Multivariate
Univariate-Outliers with Grubbs Test
Univariate-Outliers with 1.5*IQR
Learn about Box Plot
Update
With Outliers
AGE
trace 0
−50
0
50
100
count
Without Outliers
AGE
trace 0
0
count
50

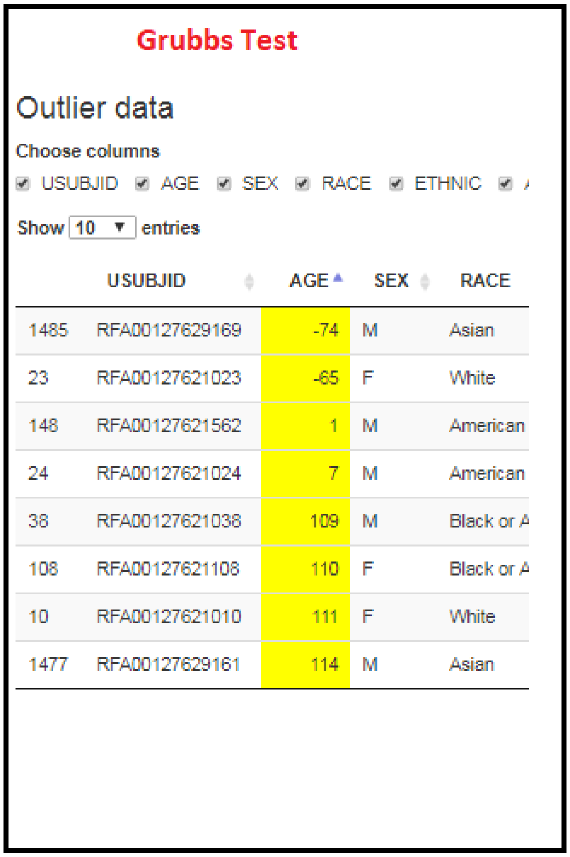
Data Filter Control

Variable to test for Outliers

○ SUBJID
○ SITEID
● AGE
○ DMDY
○ HEIGHT
○ Weight

Significance Level: 0.05 (0 – 0.3)

We can then compare outliers identified by the two different techniques:



## Grubbs Test

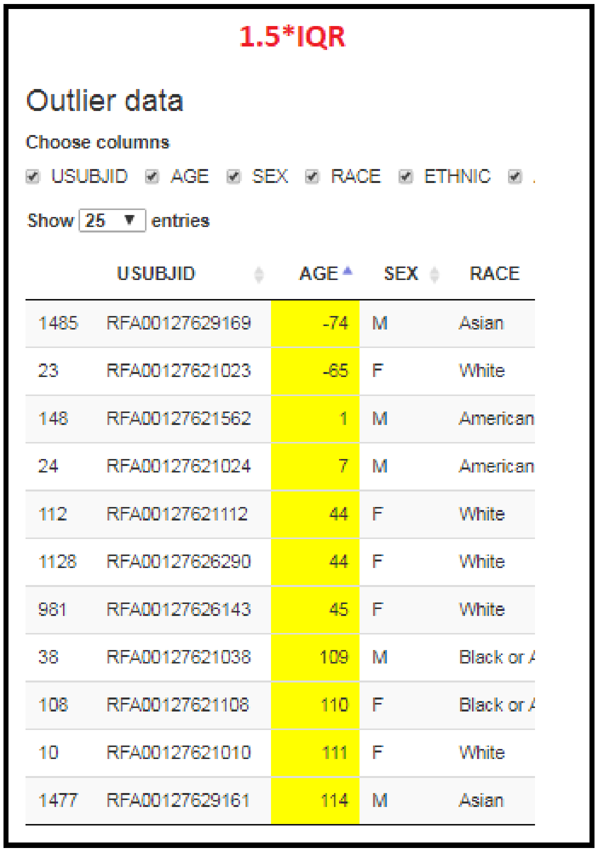
Outlier data

Choose columns

☑ USUBJID ☑ AGE ☑ SEX ☑ RACE ☑ ETHNIC ☑

Show 10 entries

| | USUBJID | AGE | SEX | RACE |
|---|---|---|---|---|
| 1485 | RFA00127629169 | -74 | M | Asian |
| 23 | RFA00127621023 | -65 | F | White |
| 148 | RFA00127621562 | 1 | M | American |
| 24 | RFA00127621024 | 7 | M | American |
| 38 | RFA00127621038 | 109 | M | Black or A |
| 108 | RFA00127621108 | 110 | F | Black or A |
| 10 | RFA00127621010 | 111 | F | White |
| 1477 | RFA00127629161 | 114 | M | Asian |



## 1.5*IQR

Outlier data

Choose columns

☑ USUBJID ☑ AGE ☑ SEX ☑ RACE ☑ ETHNIC ☑

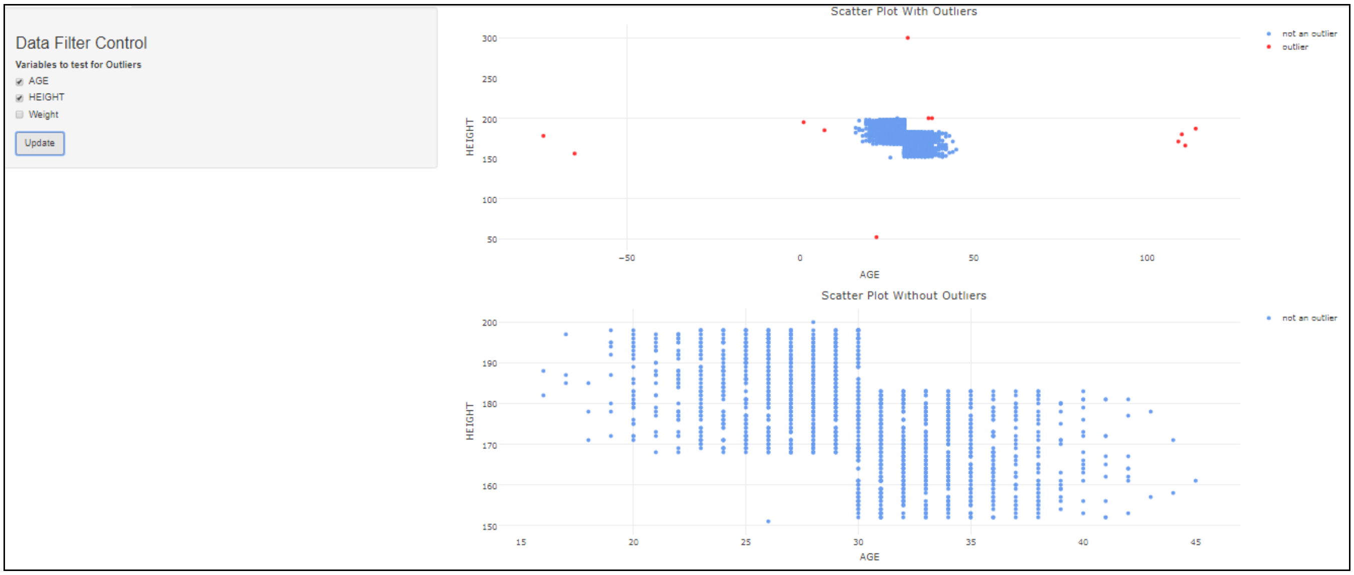
Show 25 entries

| | USUBJID | AGE | SEX | RACE |
|---|---|---|---|---|
| 1485 | RFA00127629169 | -74 | M | Asian |
| 23 | RFA00127621023 | -65 | F | White |
| 148 | RFA00127621562 | 1 | M | American |
| 24 | RFA00127621024 | 7 | M | American |
| 112 | RFA00127621112 | 44 | F | White |
| 1128 | RFA00127626290 | 44 | F | White |
| 981 | RFA00127626143 | 45 | F | White |
| 38 | RFA00127621038 | 109 | M | Black or A |
| 108 | RFA00127621108 | 110 | F | Black or A |
| 10 | RFA00127621010 | 111 | F | White |
| 1477 | RFA00127629161 | 114 | M | Asian |

As we can see from above, the Grubbs Test was less sensitive and did not detect ages 44, 44, and 45 as it did in the 1.5*IQR test. We would expect to see differences like this because of the differences in the algorithms and their settings.

### BIVARIATE APPROACHES

There are many different techniques one can use for outlier detection in the bivariate case. Bivariate data still allows ease of visualizing the data through scatter plots. Our outlier application uses both Cook's Distance and Bag Plots as outlier detection techniques. Although Cook's Distance does have a linear model assumption, it again works similar to the Grubb's test in that points that are really far away from the general area of observations will tend to get flagged as outliers, regardless of the underlying linear assumption. Similar to other techniques in the tool, it shows a scatter plot of the data both with and without outliers with all outliers colored in red as shown below for the variables age and height:



Data Filter Control

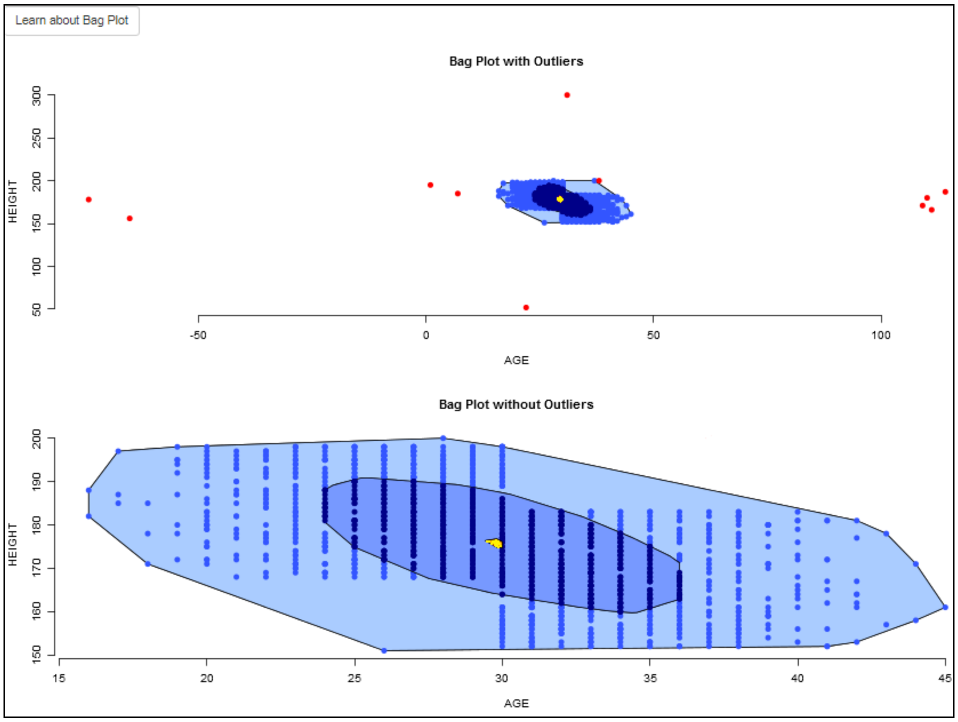
Variables to test for Outliers

☑ AGE

☑ HEIGHT

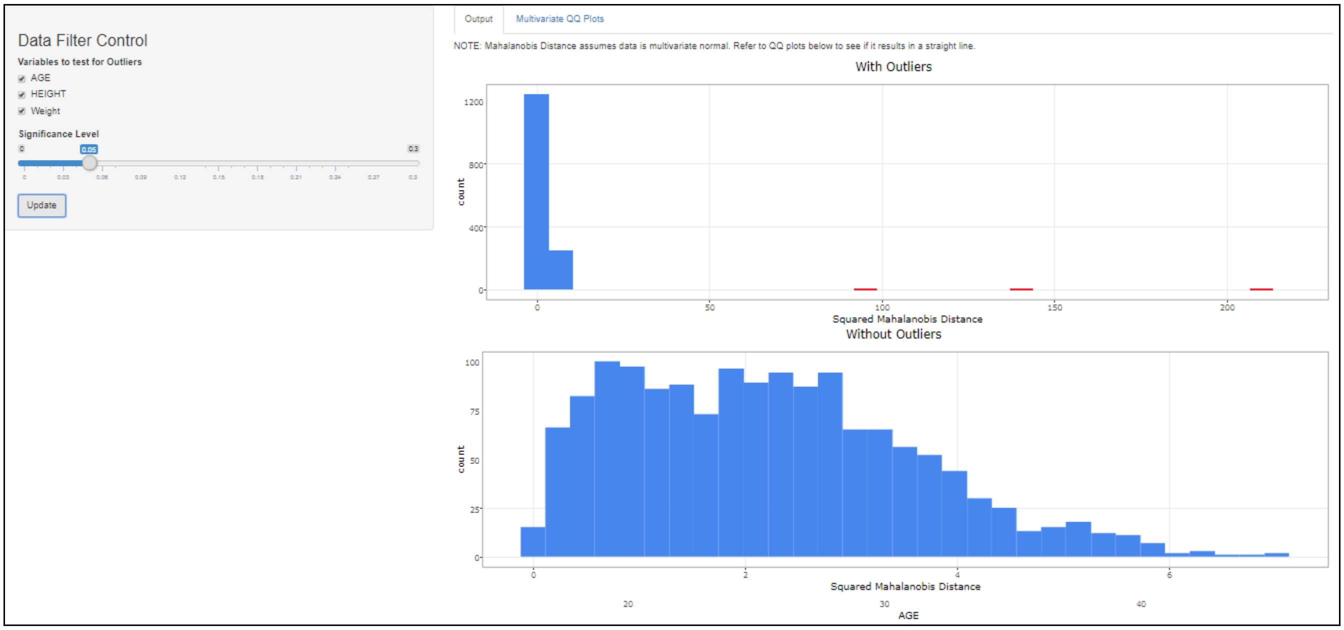
□ Weight

The Bag Plot is still similar to a scatter plot look but it places the points in polygons which can be compared to a bivariate boxplot using a calculation involving Tukey's Depth. The basic concept behind the Tukey's Depth for the bag plot generation is to find the points that can "bag" points that minimize the area of the polygons. The point in the center is the bivariate median, the dark blue polygon contains 50% of the points, and the lighter blue polygon contains the remaining 50% of points excluding outliers. The bag plot output also shows the observations both with and without outliers for consistent look and feel:



### MULTIVARIATE APPROACHES

Once moving past two variables, detecting outliers can get a lot more challenging for visualizing the data and running into the "curse of dimensionality" (for example, distributional assumptions get more challenging, computing times get longer, and data becomes more diverse). This is where an interactive application is especially helpful because it provides the visuals and distributional assumptions without having to explore too many different methods. Our application supports the Mahalanobis Distance which is similar to the Grubbs Test, but working off a multivariate normal distribution. The test finds the squared Mahalanobis distance and finds outliers now under a Chi-Squared distribution (note that Chi-Square distribution is the standard normal random variable squared):



Data Filter Control

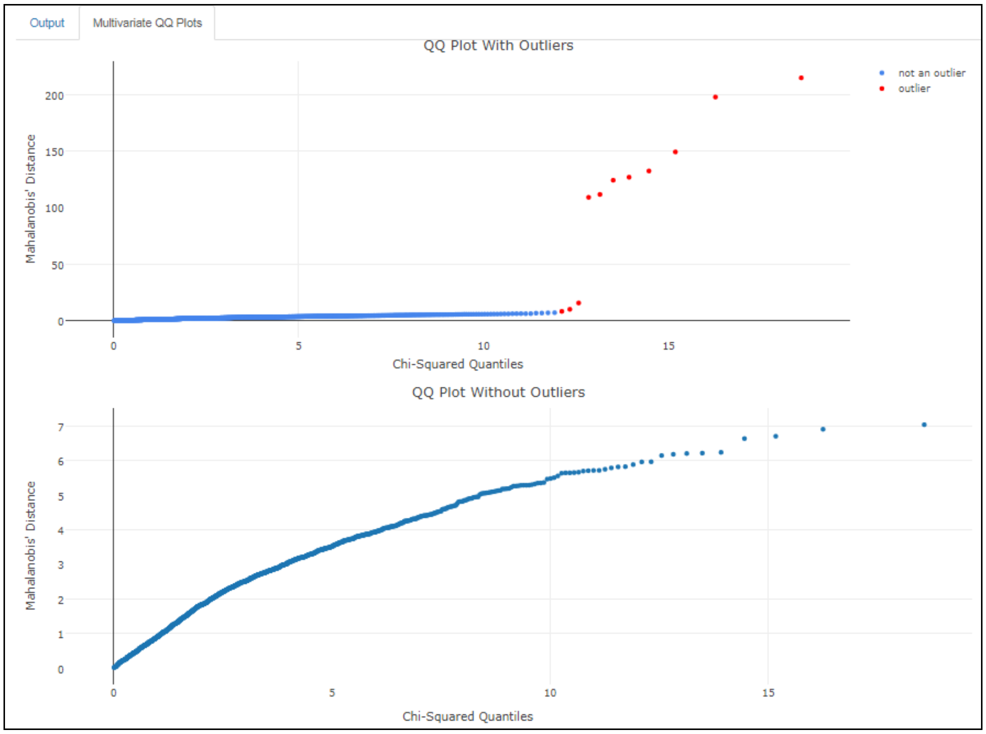
Variables to test for Outliers

☑ AGE
☑ HEIGHT
☑ Weight

Significance Level: 0.05

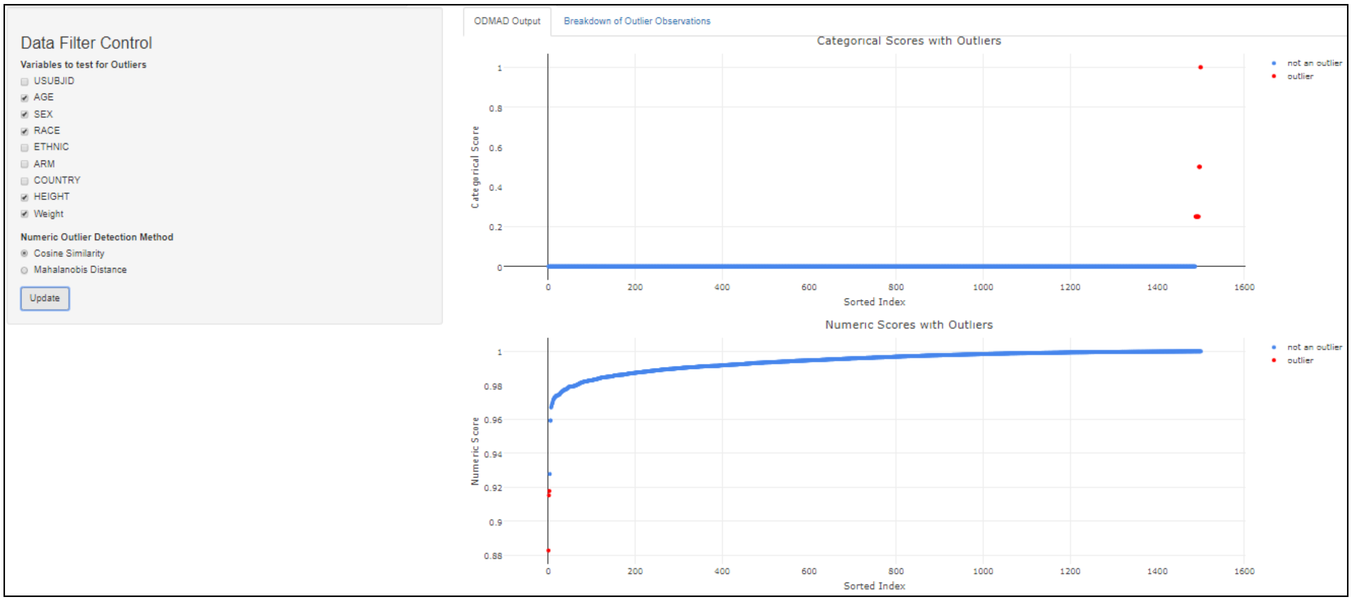
NOTE: Mahalanobis Distance assumes data is multivariate normal. Refer to QQ plots below to see if it results in a straight line.

Additionally, since it is assuming multivariate normality, the assumption is checked again through QQ (Quantile-Quantile) plots by checking if the data is in a straight line, as shown below:



Note again that checking the QQ plot both with and without outliers can be helpful since sometimes outlandish points can skew the plot.

The application also expands the capability to a newly discovered approach called ODMAD (Outlier Detection for Mixed Attribute Data), which can handle both categorical and numeric data with the flexibility to choose different algorithms based on the nature of your data. ODMAD accounts for the frequencies of the categorical variables and combines that with numeric approaches by sub setting the data according to categorical values. Different approaches may be used for the numeric portion based on the data's distribution, dimensionality, etc. For example, for datasets with a lot of numeric variables, cosine similarity may be more appropriate due to it's simple angular distance based calculation that is less sensitive to dimensionality and distributional assumptions. For other datasets that may have less numeric variables and assumed underlying distributions, other techniques such as Mahalanobis may be more appropriate. Below is a snapshot of ODMAD outputs using cosine similarity:



Data Filter Control

Variables to test for Outliers

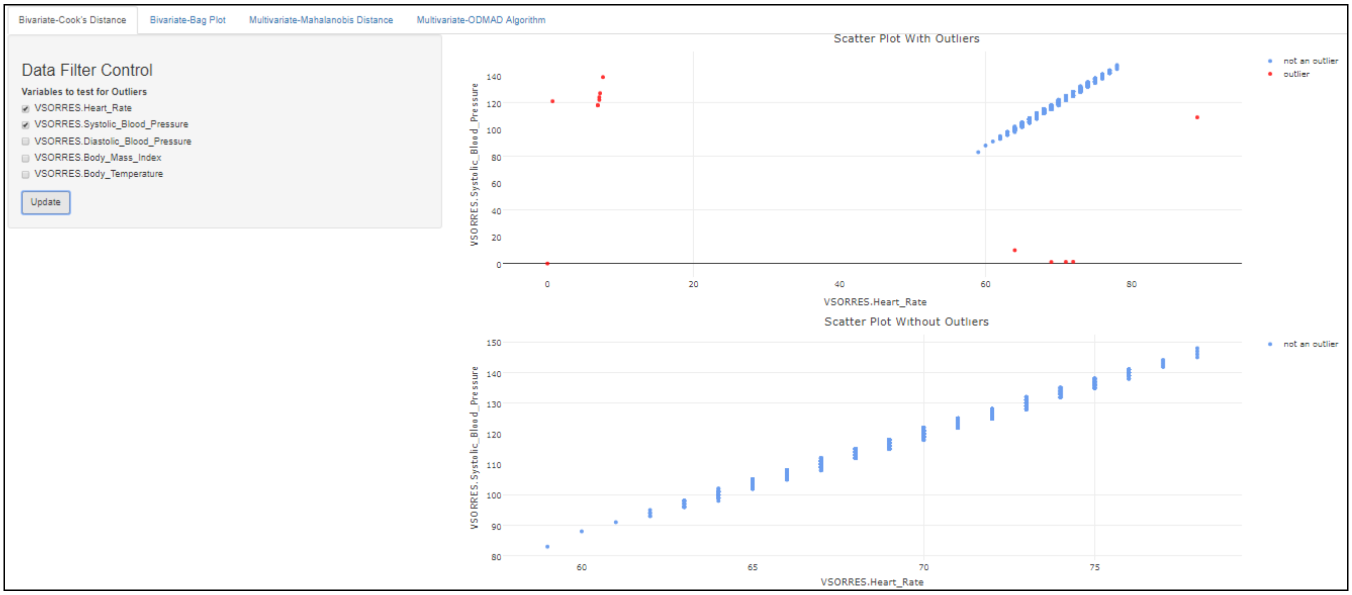
☐ USUBJID
☑ AGE
☑ SEX
☑ RACE
☐ ETHNIC
☐ ARM
☐ COUNTRY
☑ HEIGHT
☑ Weight

Numeric Outlier Detection Method

● Cosine Similarity
○ Mahalanobis Distance

## CONSISTENCY BETWEEN TECHNIQUES

As has been mentioned earlier, a key feature of our tool is that it allows comparison of various techniques in a consistent fashion. Regardless of the methods/techniques used, the results are presented with the same "look and feel". The basic idea is to maintain the same color scheme; outliers are always colored in red and normal points are always colored in blue. The outputs always show both "with outliers" and "without outliers" options, so the user can easily view "before" and "after" effects as shown in the scatterplot below using bivariate Cook's Distance approach:



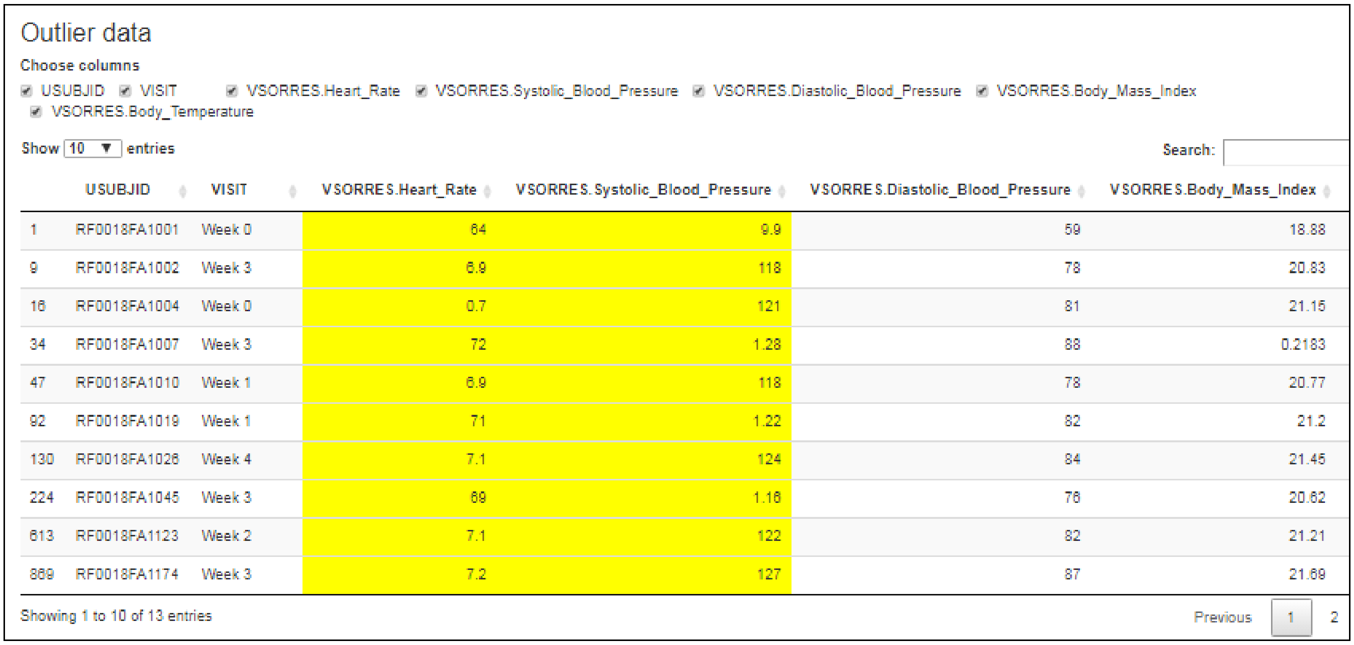
Data Filter Control

Variables to test for Outliers

☑ VSORRES.Heart_Rate
☑ VSORRES.Systolic_Blood_Pressure
☐ VSORRES.Diastolic_Blood_Pressure
☐ VSORRES.Body_Mass_Index
☐ VSORRES.Body_Temperature

Bivariate-Cook's Distance
Bivariate-Bag Plot
Multivariate-Mahalanobis Distance
Multivariate-ODMAD Algorithm
Update
Scatter Plot With Outliers
VSORRES.Systolic_Blood_Pressure
0
20
40
60
80
100
120
140
0
20
40
60
80
VSORRES.Heart_Rate
● not an outlier
● outlier
Scatter Plot Without Outliers
VSORRES.Systolic_Blood_Pressure
80
90
100
110
120
130
140
150
60
65
70
75
VSORRES.Heart_Rate
● not an outlier

The list of outliers is always shown below the outputs with the variables of interest highlighted in yellow so the user can easily identify which variables were used in the outlier detection from that given dataset and with the ability to manipulate the dataset view however they wish by dropping variables not required and specifying the number of observations to display as seen on the next page:



Outlier data

Choose columns

☑ USUBJID ☑ VISIT ☑ VSORRES.Heart_Rate ☑ VSORRES.Systolic_Blood_Pressure ☑ VSORRES.Diastolic_Blood_Pressure ☑ VSORRES.Body_Mass_Index ☑ VSORRES.Body_Temperature

Show 10 ▼ entries

Search:

| | USUBJID | VISIT | VSORRES.Heart_Rate | VSORRES.Systolic_Blood_Pressure | VSORRES.Diastolic_Blood_Pressure | VSORRES.Body_Mass_Index |
|---|---|---|---|---|---|---|
| 1 | RF0018FA1001 | Week 0 | 64 | 9.9 | 59 | 18.88 |
| 9 | RF0018FA1002 | Week 3 | 6.9 | 118 | 78 | 20.83 |
| 16 | RF0018FA1004 | Week 0 | 0.7 | 121 | 81 | 21.15 |
| 34 | RF0018FA1007 | Week 3 | 72 | 1.28 | 88 | 0.2183 |
| 47 | RF0018FA1010 | Week 1 | 6.9 | 118 | 78 | 20.77 |
| 92 | RF0018FA1019 | Week 1 | 71 | 1.22 | 82 | 21.2 |
| 130 | RF0018FA1026 | Week 4 | 7.1 | 124 | 84 | 21.45 |
| 224 | RF0018FA1045 | Week 3 | 69 | 1.16 | 76 | 20.62 |
| 613 | RF0018FA1123 | Week 2 | 7.1 | 122 | 82 | 21.21 |
| 869 | RF0018FA1174 | Week 3 | 7.2 | 127 | 87 | 21.69 |

Showing 1 to 10 of 13 entries

Previous 1 2

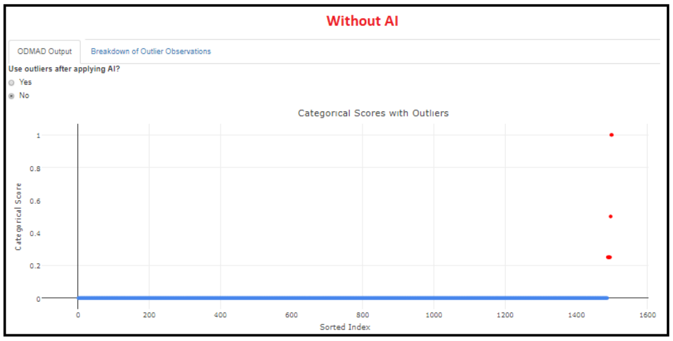
Finally, if there are any distributional assumptions required for a particular technique, there is an extra tab in the application that allows the user to view any necessary outputs to establish the validity of the assumptions. The outputs are also shown before and after outliers are removed, for the chance that there may be some points that distort the view of certain outputs, such as in a QQ plot.

This consistency allows the user to navigate around the application and once they are comfortable with the view, understanding the remaining pages is an easy task.

## MACHINE LEARNING TO ENCHANCE OUTLIER DETECTION

Statistical methods often detect points that may be considered outliers, but are in actuality fine from a human perspective. When exploring outliers, it can be frustrating to repeatedly have data flagged as outliers when you have already determined they are not. For example, suppose that a test keeps flagging a patient with the age 45 as an outlier, but after investigating your data further, you have established that there are no issues with this value. After making this decision, suppose the user returns to the application, and they do not want to see this patient flagged anymore. This is where the machine learning can assist (a predictive model that iterates randomly to find and utilize potentially hidden patterns in the underlying data, in this case outlier detection decisions). The goal is to learn over time with use of the application on how to better accommodate both statistical/computational outlier detection with contextual evaluations and judgements of the user. Simply put, it is a way to 'override' a flagged outlier based on the decisions previously made by the user. There is always an option to opt out of the machine learning element of the application, but it is a way to potentially enhance the user experience and prevent time spent weeding through previously detected but overridden "outliers".

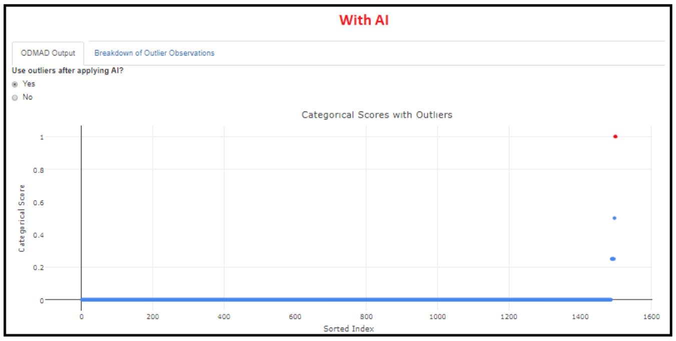
Below is an example of the outliers with and without machine learning after applying the predictions to previous "human judgement" of true outliers, using the ODMAD detection method:



Without AI

Use outliers after applying AI?

○ Yes
● No



With AI

Use outliers after applying AI?

● Yes
○ No

As one can see from the previous page, the actual computed values are still the same from the tests, but after applying the machine learning, it overrode some of the determinants of what actually is an outlier based on the user's previous judgements that those patients previously marked red (or patients similar to them), were indeed not problematic.

## CONCLUSION

This outlier application provides the ability for outlier detection for users that are not familiar with programming or do not have time to program themselves. It also provides a suite of appropriate detection techniques so the user does not need to spend time shuffling through different ideologies to find which one best fits their needs. Additionally, it provides pop-up boxes of detailed theoretical explanations of the techniques, distributional assumptions, and visuals all in one place. The machine learning abilities can also enhance efficiency, filtering out points that had already previously been determined from personal review/judgement to not be outliers. The overall goal is to gain efficiency in detecting outliers, both to aid data cleaning and to ensure extreme values are appropriately addressed prior to analysis. Thus, more time can then be spent on the important underlying statistical analysis.

## REFERENCES

1. **Koufakou, Anna**, "*Scalable And Efficient Outlier Detection In Large Distributed Data Sets With Mixed-type Attributes*" (2009). Electronic Theses and Dissertations. 3986.

## ACKNOWLEDGMENTS

The authors are indebted to Abhay Thakur for his contribution to the development of the data visualization application for outlier detection.

## CONTACT INFORMATION

Contact the author at:

Author: Diane Peers
Company: GCE Solutions
Address: 82, King St
City / Postcode: Manchester, M2 4WQ
Work Phone: +44 161 871 7483
Email: diane.peers@gcesolutions.com
Web: http://gcesolutions.com