



The Art of Data Visualization

Detecting Multivariate Data Outliers Using an Interactive Approach

Diane Peers
GCE Solutions



Agenda

- What is an outlier?
- Overview of interactive approach
- Preparing the data
- Outlier detection techniques
- Machine learning to enhance outlier detection
- Summary



What is an outlier?

Definition:

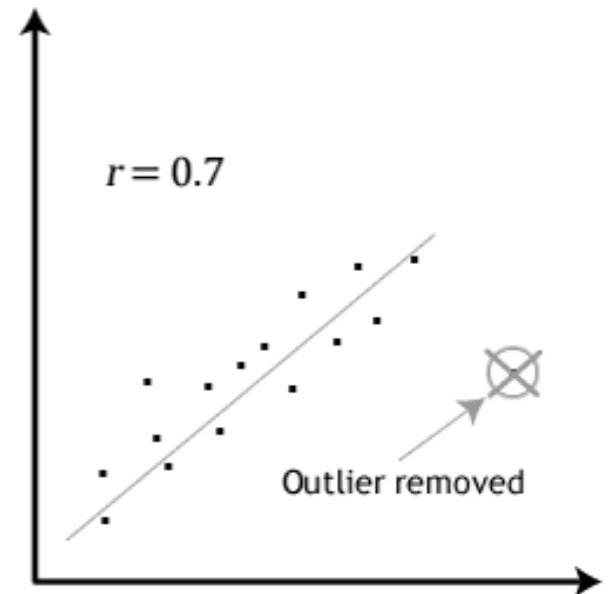
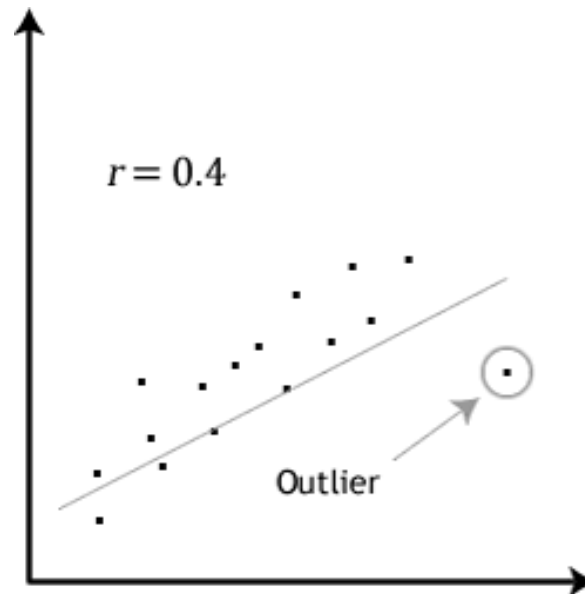
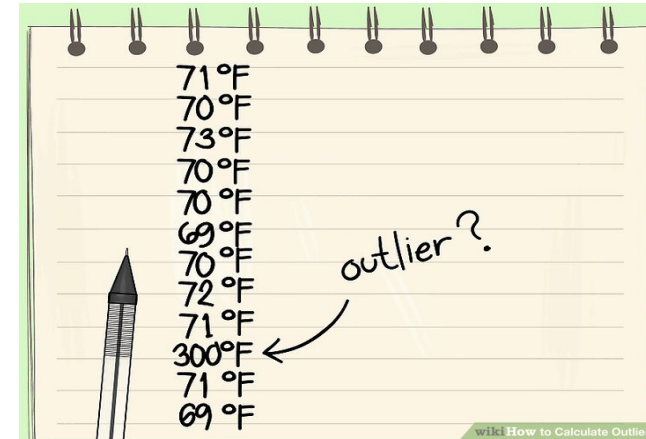
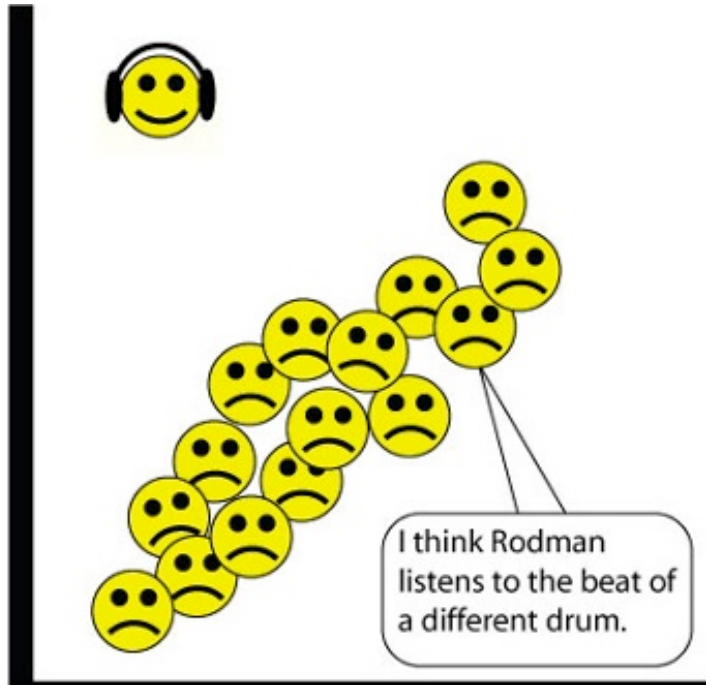
“An observation that appears to deviate markedly from other observations in the sample”

Why do we need to identify them?

- To aid data cleaning
- So they do not skew the analyses



How do you identify outliers?





Overview of interactive visualization tool

- *“One stop shop”* for outlier detection – Consistent approach

Benefits of the tool:

- Easy to use – no specialist knowledge required
- Can read a variety of data types
- Range of outlier techniques to pick from
- Checks distributional assumptions
- Visual plots of the data



Preparing the data

- Step by step process for: filtering, removing columns, subsetting and transposing the data

The screenshot shows a web browser window titled 'Outlier Application' with the address bar displaying 'localhost:11111/#/'. The navigation bar includes tabs for 'Outlier Application', 'File Uploading', 'Dataset Manipulation' (which is selected), 'Setting up the AI (Optional)', 'Outlier Detection Univariate', and 'Outlier Detection Multivariate'. The main content area is titled 'Subset the Data' and contains a 'Data Filter Control' section on the left and a dropdown menu on the right. The dropdown menu is labeled 'RACE' and lists the following options: 'American Indian or Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', and 'White'. The dropdown menu is highlighted with a red circle.



Preparing the data - transpose

Dataset before and after the transpose

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

Data Extraction result

Show 10 entries

	STUDYID	DOMAIN	USUBJID	SUBJID	VISITNUM	VISIT	VSTEST	VSTESTCD	VSPOS	VSORRES	VSORRESU	VSORNRLO	VSORNRHI	VSSTRESC	VSSTRESN	VSSTRESU	VSSTNRLO	VSSTNRHI	VSSTNRC	VSNRIND	VSSTAT	VSREASND	VSBLFL	VSDRVFL	VISITDY	VSDTC	VSDY
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																
				1001	2	Week 2	Heart Rate	HR	SUPINE	68	beats per minute																
				1001	3	Week 3	Heart Rate	HR	SUPINE	69	beats per minute																
				1001	4	Week 4	Heart Rate	HR	SUPINE	71	beats per minute																
				1002	0	Week 0	Heart Rate	HR	SUPINE	67	beats per minute																
				1002	1	Week 1	Heart Rate	HR	SUPINE	67	beats per minute																
				1002	2	Week 2	Heart Rate	HR	SUPINE	75	beats per minute																
				1002	3	Week 3	Heart Rate	HR	SUPINE	6.9	beats per minute																
				1002	4	Week 4	Heart Rate	HR	SUPINE	70	beats per minute																

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	0	64	9.9	59	18.88	
2	RF0018FA1001	1	72	128	88	21.63	37.1
3	RF0018FA1001	2	68	113	73	20.3	37.2
4	RF0018FA1001	3	69	117	77	20.67	36.9
5	RF0018FA1001	4	71	125	85	21.5	37.7
6	RF0018FA1002	0	67	111	71	20.07	36.8
7	RF0018FA1002	1	67	110	70	20.04	37.5
8	RF0018FA1002	2	75	138	98	22.76	37.8
9	RF0018FA1002	3	6.9	118	78	20.83	36.8
10	RF0018FA1002	4	70	119	79	20.95	37.3

Showing 1 to 10 of 1,501 entries

Previous 1 2 3 4 5 ... 151 Next



Selecting and filtering on specific variables

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

Data Summary

Back

Select Columns

Choose columns

☐ STUDYID
☒ DOMAIN
☒ USUBJID
☐ SUBJID
☐ RFSTDTC
☐ RFENDTC
☐ RFXSTDTC
☐ RFXENDTC
☐ RFCIDTC
☐ RFPENDTC
☐ DTHDTC
☐ DTHFL
☐ SITEID
☐ INVID
☐ INVNAM
☐ BRTHDTC
☒ AGE
☐ AGEXT
☐ AGEU
☐ SEX
☒ RACE
☐ ETHNIC
☐ SPECIES
☐ STRAIN
☐ SBSTRAIN
☐ ARMCD
☐ ARM
☐ ACTARMCD
☐ ACTARM
☐ SETCD
☒ COUNTRY
☐ DMDTC
☐ DMDY
☒ HEIGHT
☐ Weight

SEE: [DATA Statistics](#)

Display Data

Show 10 entries

Search:

	DOMAIN	USUBJID	AGE	RACE	COUNTRY	HEIGHT
3	DM	RFA00127621003	31	Black or African American	Canada	169
21	DM	RFA00127621021	38	White	Canada	182
23	DM	RFA00127621023	-65	White	Canada	156
26	DM	RFA00127621026	31	Native Hawaiian or Other Pacific Islander	Canada	177

Summary Statistics for selected variables

DOMAIN	USUBJID	AGE	
DM:1500	RFA00127621001:	1	Min. :-74.00
	RFA00127621002:	1	1st Qu.: 26.00
	RFA00127621003:	1	Median : 30.00
	RFA00127621004:	1	Mean : 29.88
	RFA00127621005:	1	3rd Qu.: 33.00
	RFA00127621006:	1	Max. :114.00
	(Other)	:1494	
		RACE	COUNTRY
American Indian or Alaska Native	:237	America:	861
Asian	:388	Canada :	251
Black or African American	:297	India :	388
Chinese	: 1		
Native Hawaiian or Other Pacific Islander:	300		
White	:277		



Outlier detection – what technique to use?

The technique used will depend upon:

- the dimensionality (number of variables used to detect outliers)
- variable data types (i.e. numeric versus categorical)
- distribution of the data

It is also valuable to compare various techniques and see if any observations are continually getting flagged as outliers



Outlier detection techniques

Univariate

- Grubbs test
- Inter-quartile range approach

Bivariate

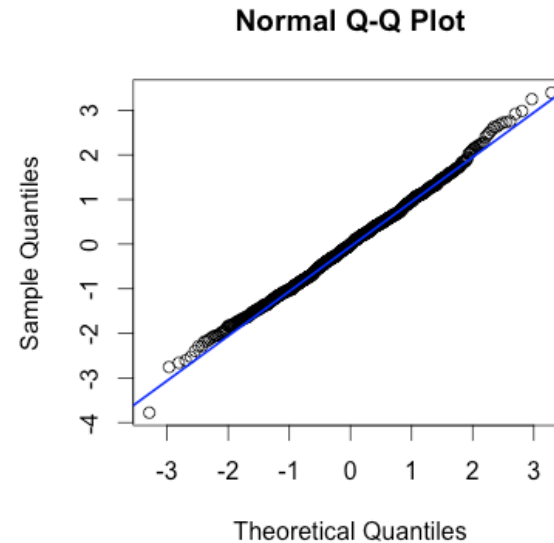
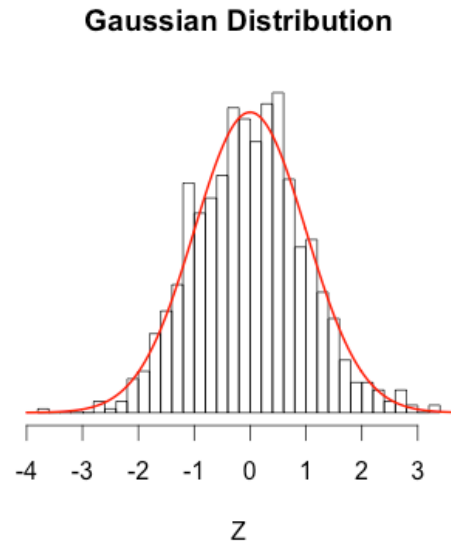
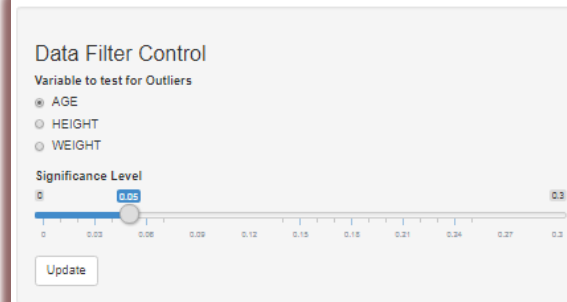
- Cook's distance
- Bag plots

Multivariate

- Mahalanobis Distance
- Outlier detection for mixed attribute data (ODMAD)



Grubb's test – normality test





Grubb's test

Data Filter Control

Variable to test for Outliers

- ☒ AGE
- ☐ HEIGHT
- ☐ Weight

Significance Level



Update

Univariate-Outliers with Grubbs Test

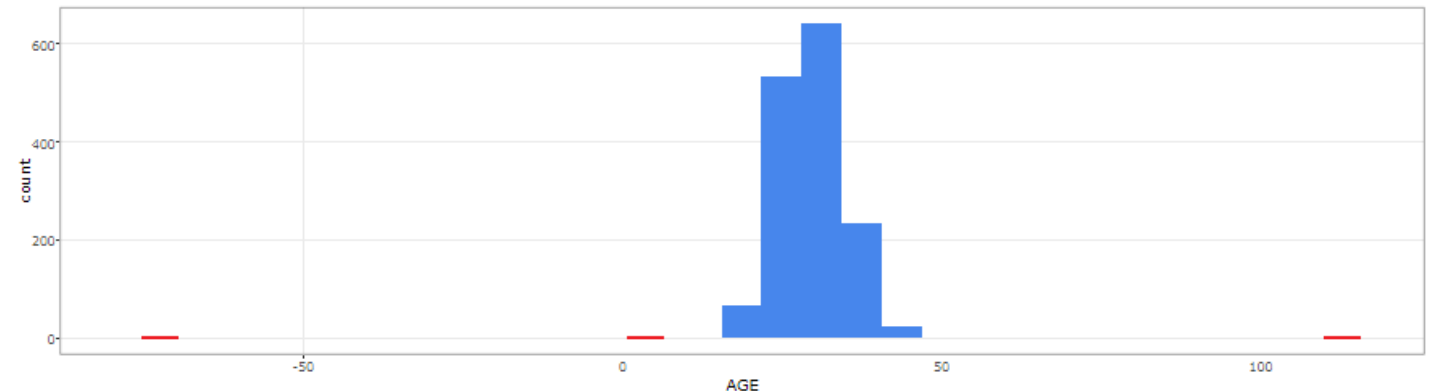
Univariate-Outliers with 1.5*IQR

Output

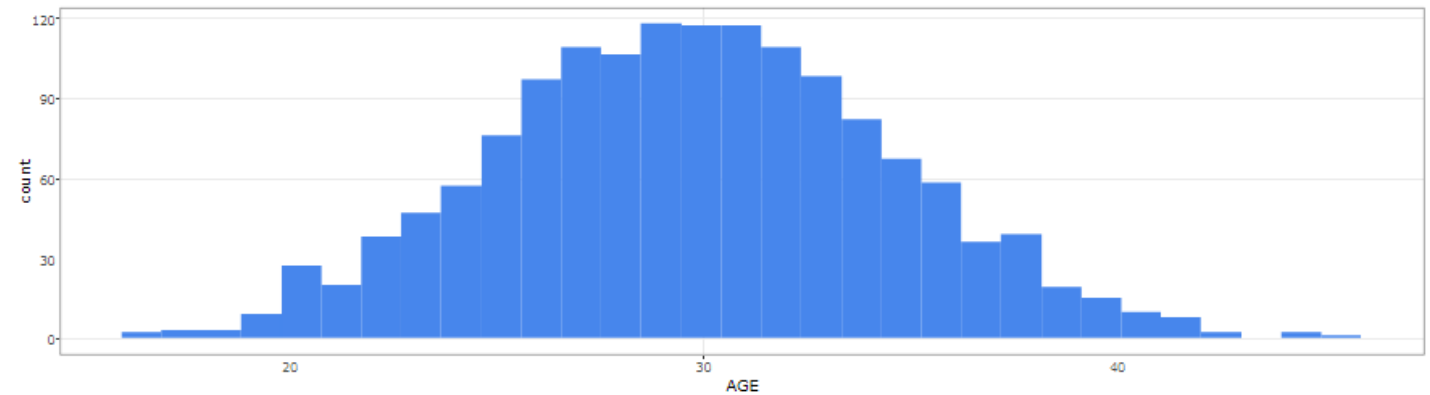
Normality Test

NOTE: This test assumes normality.

With Outliers

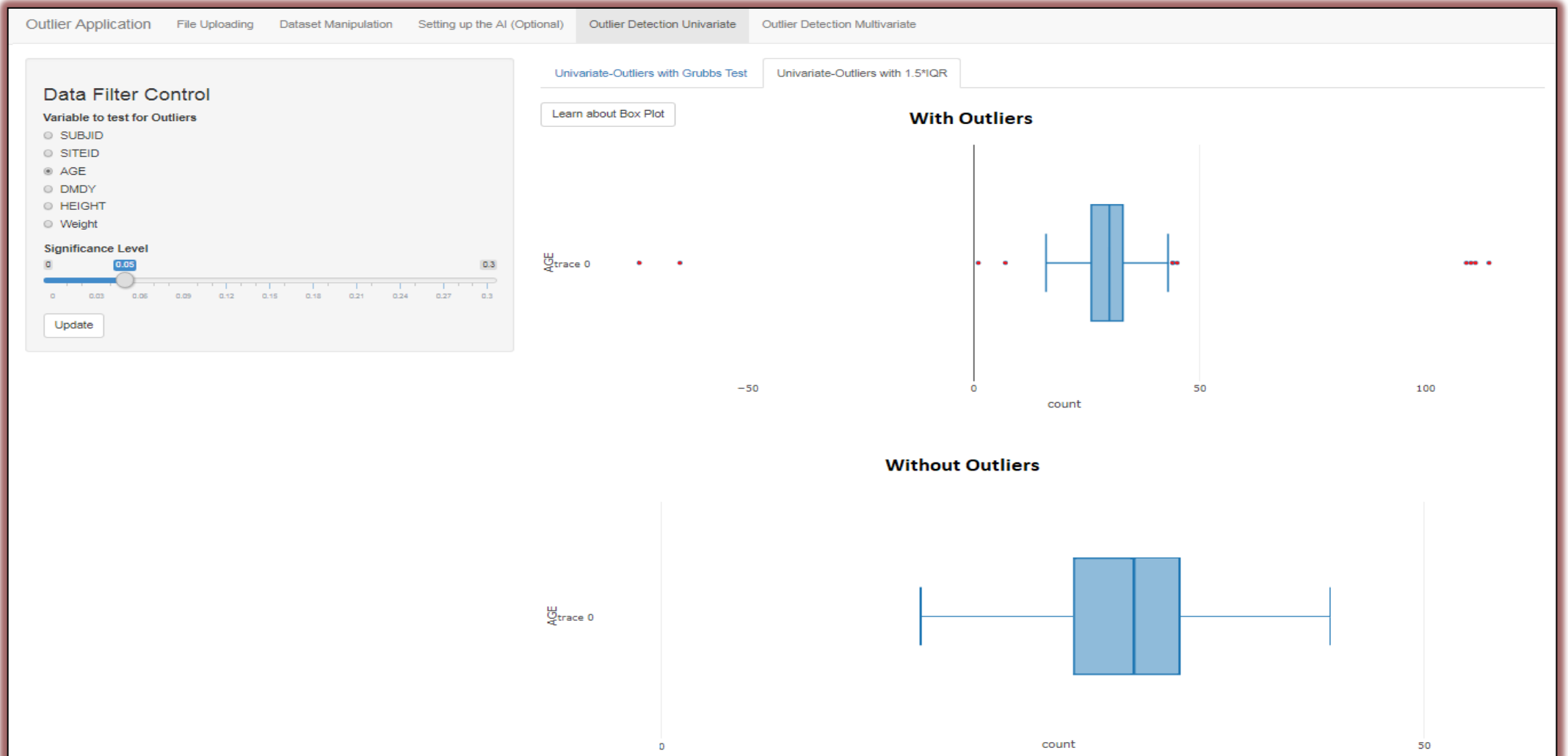


Without Outliers





Inter-Quartile range (IQR)





Comparing outliers selected

Tool lists the observations with outliers from each technique

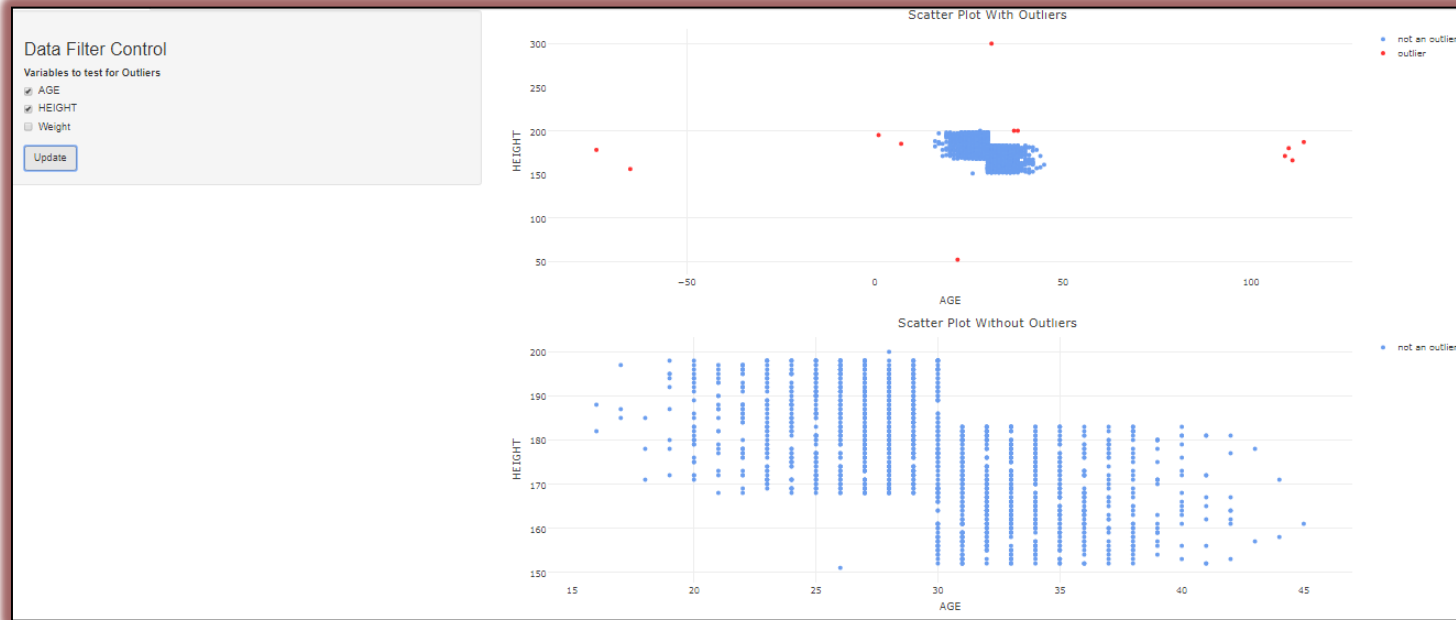
Grubbs Test				
Outlier data				
Choose columns				
☒ USUBJID	☒ AGE	☒ SEX	☒ RACE	☒ ETHNIC
Show 10 entries				
USUBJID	AGE	SEX	RACE	
1485	RFA00127629169	-74	M	Asian
23	RFA00127621023	-65	F	White
148	RFA00127621562	1	M	American
24	RFA00127621024	7	M	American
38	RFA00127621038	109	M	Black or A
108	RFA00127621108	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

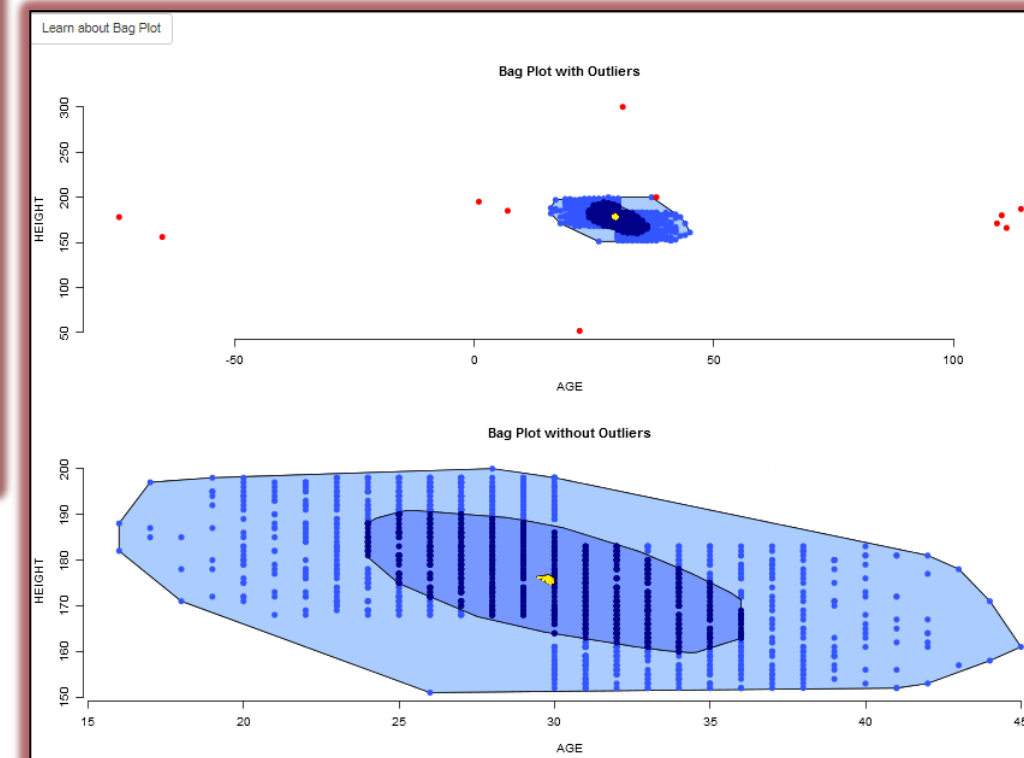


Bivariate Outlier detection techniques

Scatterplot from Cook's distance



Bag Plot





Cook's distance – outliers identified

Outlier data

Choose columns

☒ STUDYID ☒ DOMAIN ☒ USUBJID ☐ SUBJID ☐ RFSTDTC ☐ RFENDTC ☐ RFXSTDTC ☐ RFXENDTC ☐ RFICDTC ☐ RFPENDTC
☐ DTHDTC ☐ DTHFL ☐ SITEID ☐ INVID ☐ INVNAM ☐ BRTHDTC ☒ AGE ☐ AGETXT ☐ AGEU ☒ SEX ☒ RACE ☐ ETHNIC
☐ SPECIES ☐ STRAIN ☐ SBSTRAIN ☐ ARMCD ☐ ARM ☐ ACTARMCD ☐ ACTARM ☐ SETCD ☐ COUNTRY ☐ DMDTC ☐ DMDY
☒ HEIGHT ☒ WEIGHT

Show entries

Search:

	STUDYID	DOMAIN	USUBJID	AGE	SEX	RACE	HEIGHT	WEIGHT
2	RFA0012762	DM	RFA00127621002	38	F	White	200	190
10	RFA0012762	DM	RFA00127621010	111	F	White	166	221
17	RFA0012762	DM	RFA00127621017	22	M	American Indian or Alaska Native	52	209
23	RFA0012762	DM	RFA00127621023	-65	F	White	156	216
24	RFA0012762	DM	RFA00127621024	7	M	American Indian or Alaska Native	185	220
38	RFA0012762	DM	RFA00127621038	109	M	Black or African American	171	210
108	RFA0012762	DM	RFA00127621108	110	F	Black or African American	180	93
148	RFA0012762	DM	RFA00127621562	1	M	American Indian or Alaska Native	195	146
320	RFA0012762	DM	RFA00127625032	37	1	White	200	155
1477	RFA0012762	DM	RFA00127629161	114	M	Asian	187	203

Showing 1 to 10 of 12 entries

Previous 2 Next



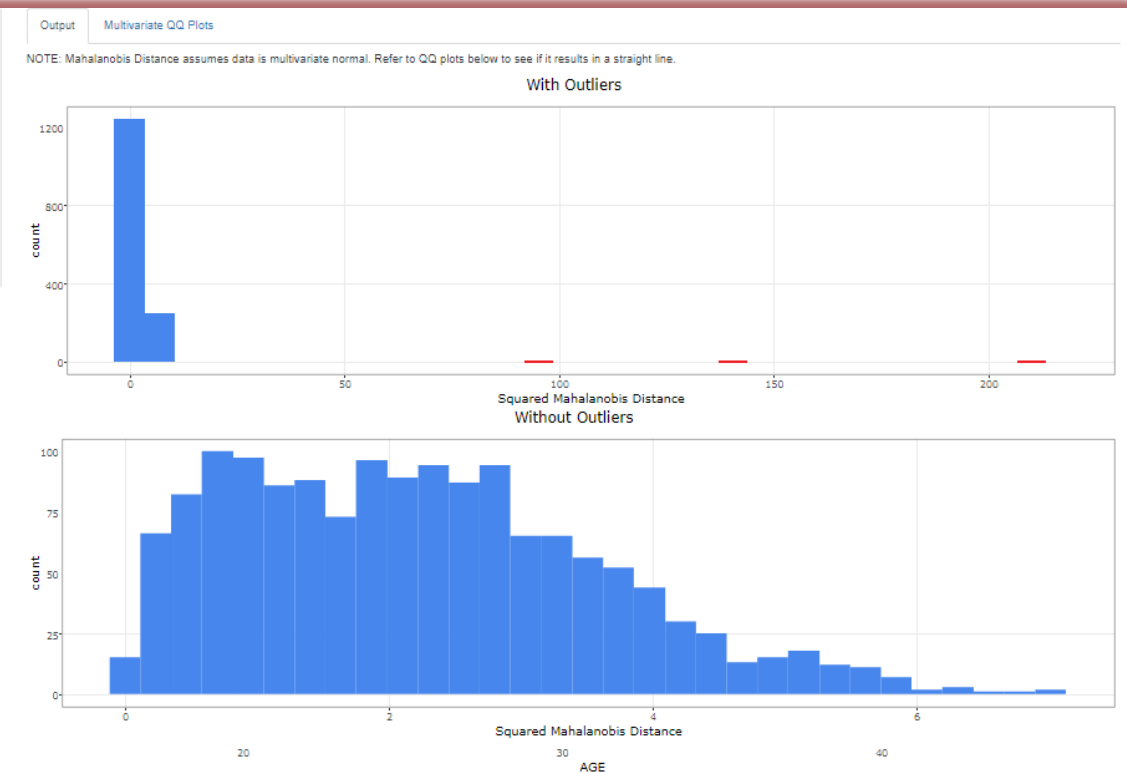
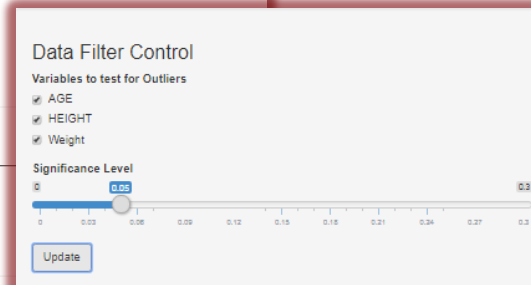
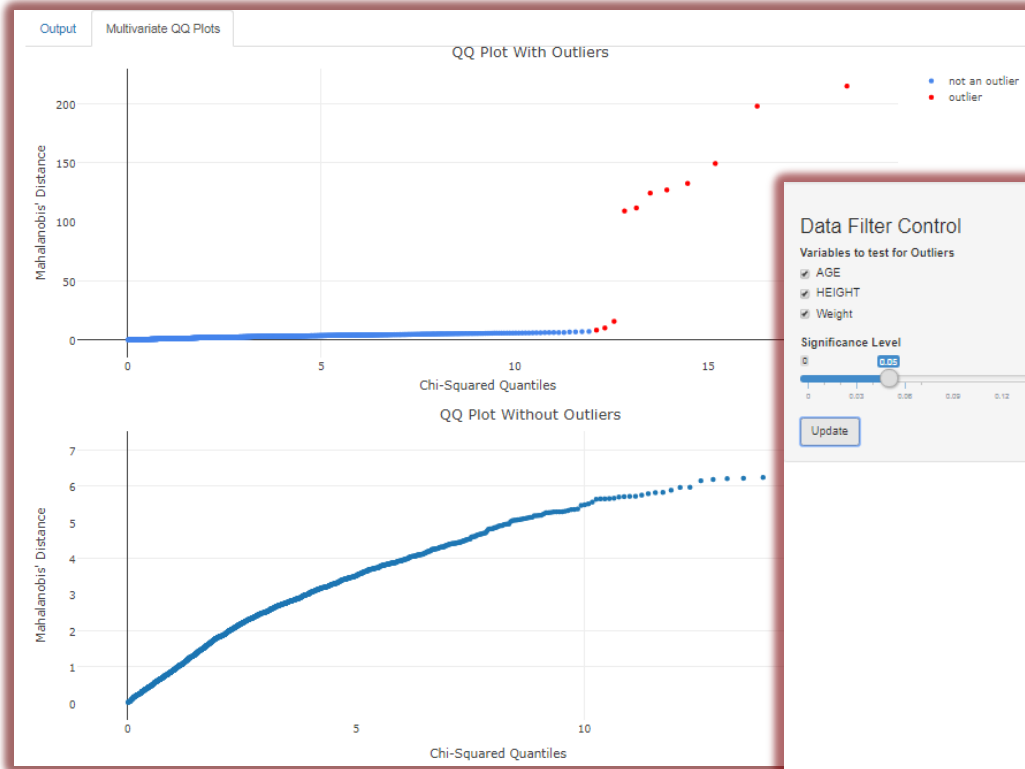
Curse of Dimensionality





Multivariate - Mahalanobis distance

Quantile-Quantile plots





Multivariate - ODMAD

Data Filter Control

Variables to test for Outliers

- ☐ USUBJID
- ☒ AGE
- ☒ SEX
- ☒ RACE
- ☐ ETHNIC
- ☐ ARM
- ☐ COUNTRY
- ☒ HEIGHT
- ☒ Weight

Numeric Outlier Detection Method

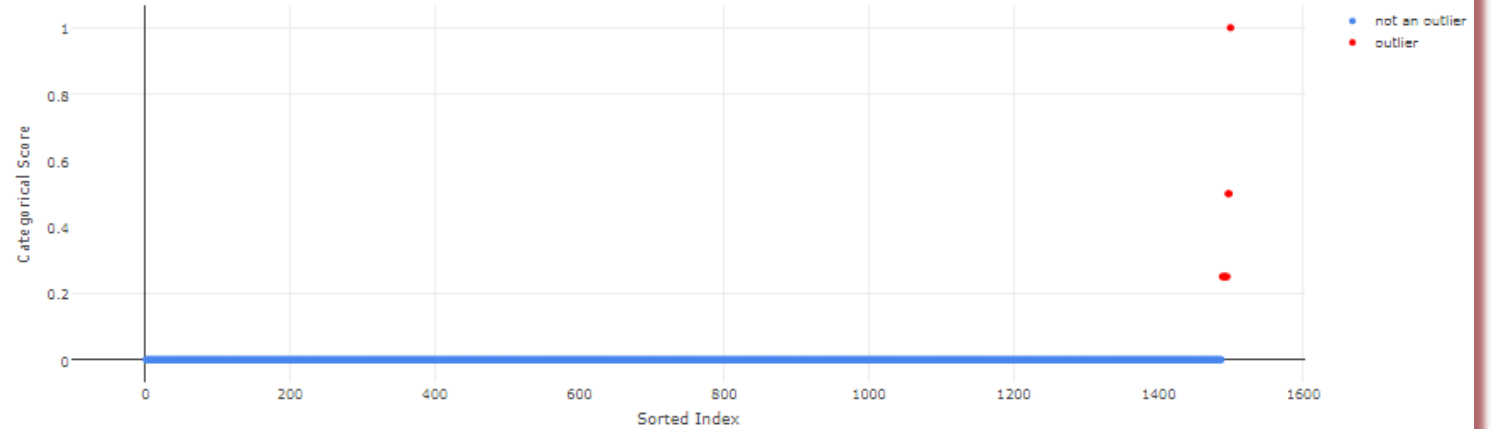
- ☒ Cosine Similarity
- ☐ Mahalanobis Distance

Update

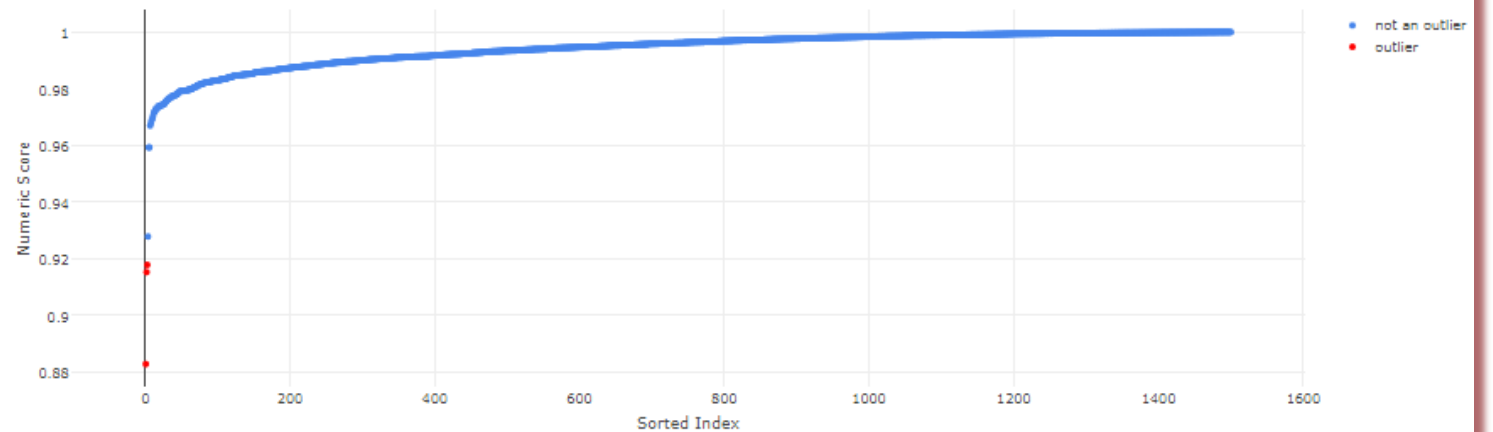
ODMAD Output

Breakdown of Outlier Observations

Categorical Scores with Outliers



Numeric Scores with Outliers





Multivariate – ODMAD - Outliers

ODMAD Output

Breakdown of Outlier Observations

Choose columns

☐ STUDYID ☐ DOMAIN ☐ USUBJID ☐ SUBJID ☐ RFSTDTC ☐ RFENDTC ☐ RFXSTDTC ☐ RFXENDTC ☐ RFICDTC ☐ RFPENDTC ☐ DTHDTC ☐ DTHFL ☐ SITEID ☐ INVID ☐ INVNAM ☐ BRTHDTC ☒ AGE ☐ AGETXT ☐ AGEU ☒ SEX ☒ RACE ☐ ETHNIC ☐ SPECIES ☐ STRAIN ☐ SBSTRAIN ☐ ARMCD ☐ ARM ☐ ACTARMCD ☐ ACTARM ☐ SETCD ☐ COUNTRY ☐ DMDTC ☐ DMDY ☒ HEIGHT ☒ WEIGHT

Show 10 entries

Search:

Row Number	Type of Flagging	Categorical/Numeric Score	Categorical Variables Caused Flagging	Categorical Values Caused Flagging	AGE	SEX	RACE	HEIGHT	WEIGHT	
1	11	numeric	0.916381137192978	Flagged under all categories	29	M	Asian	172	400	
2	17	numeric	0.841378634563727	Flagged under all categories	22	M	American Indian or Alaska Native	52	209	
3	23	numeric	0.918288458677179	RACE	White	-65	F	White	156	216
4	53	categorical	0.25	SEX	2	34	2	White	156	177
5	96	categorical	0.25	SEX	1	28	1	Asian	179	210
6	106	categorical	0.25	SEX	2	35	2	White	175	132
7	108	numeric	0.88988854476096	Flagged under all categories	110	F	Black or African American	180	93	
8	125	categorical	0.25	SEX	1	28	1	Asian	200	170
9	162	numeric	0.860842887607106	Flagged under all categories	29	M	Black or African American	179	42	
10	254	categorical	1	SEX	N	32	N	Native Hawaiian or Other Pacific Islander	176	116

Showing 1 to 10 of 21 entries

Previous123Next



Machine learning to enhance outlier detection

What happens when values are consistently identified as outliers but we know they are valid values and we don't want them to keep being flagged as outliers?

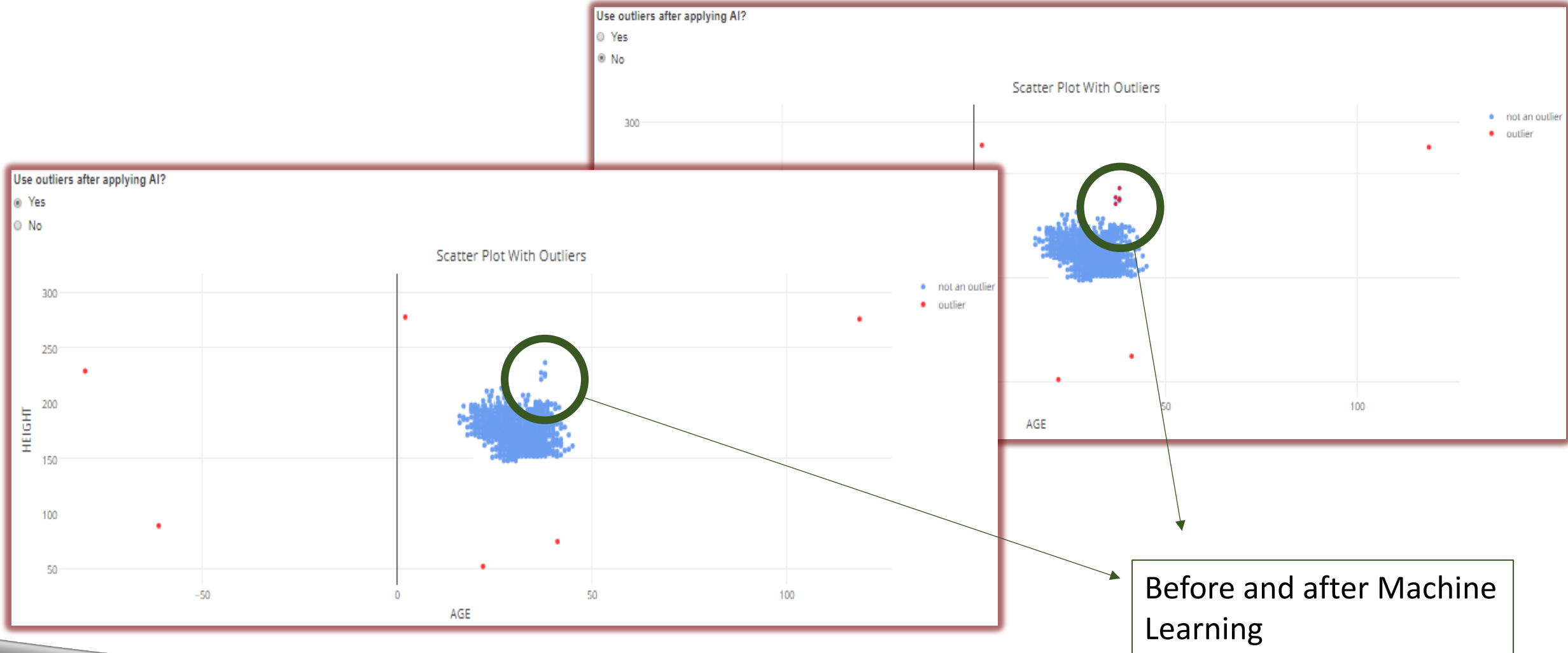
Machine learning

A predictive model that iterates randomly to find and utilize potentially hidden patterns in the underlying data, in this case outlier detection decisions

Using this technique the tool can override a flagged outlier based on the decisions previously made by the user



Machine learning to enhance outlier detection





A cautionary tale : Conclusion





Acknowledgements

This presentation was compiled with the help of Lorin Miller, co-author of the PhUSE paper

The outlier detection tool was developed by Lorin Miller and Abhay Thakur at GCE Solutions

A number of pictures used in this presentation were taken from google images <https://images.google.co.uk/>



Summary

- Tool does all steps needed to successfully identify outliers
- Easy to use
- Data visuals allow for easy outlier detection
- Can improve quality of outlier detection via machine learning
- Outliers need to be investigated

Contact details:

diane.peers@gcesolutions.com

Lorin.miller@gcesolutions.com