

How can I identify outliers in SAS with an example?

Authored by
stats writer

July 1, 2024

RECOMMENDED CITATION

stats writer (2024). *How can I identify outliers in SAS with an example?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=164193>

The process of identifying outliers in SAS involves using statistical methods to detect data points that are significantly different from the majority of the dataset. This can be achieved by analyzing the distribution of the data and calculating measures such as mean, median, and standard deviation. For example, in a dataset of exam scores, an outlier may be a student who scored significantly higher or lower than the rest of the class.

To identify outliers in SAS, one can use various techniques such as box plots, histograms, and scatter plots. These visualizations can help to identify extreme values that fall outside the expected range of the data. Additionally, statistical tests such as Grubbs' test or Dixon's Q test can be used to detect outliers with a high degree of certainty.

For instance, in a dataset of monthly sales, a box plot may reveal a data point that is significantly higher than the rest of the sales data. Using Grubbs' test, this outlier can be confirmed and further investigated to determine if it is a valid data point or an error. Overall, identifying and handling outliers in SAS is crucial for ensuring the accuracy and reliability of data analysis results.

Identify Outliers in SAS (With Example)

An outlier is an observation that lies abnormally far away from other values in a dataset. Outliers can be problematic because they can affect the results of an analysis.

The most common way to identify outliers in a dataset is by using the interquartile range.

The interquartile range (IQR) is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) in a dataset. It measures the spread of the middle 50% of values.

We typically define an observation to be an outlier if it is 1.5 times the interquartile range greater than the third quartile (Q3) or 1.5 times the interquartile range less than the first quartile (Q1).

Outliers = Observations $> Q3 + 1.5 \times IQR$ or $< Q1 - 1.5 \times IQR$

The following example shows how to use this formula to identify outliers in a dataset in SAS.

Example: Identifying Outliers in SAS

Suppose we have the following dataset in SAS:

```
/*create dataset*/  
data original_data;  
input team $ points;  
datalines;
```

A 18

B 24

C 26

D 34

E 38

F 45

G 48

H 54

I 60

J 73

K 79

L 85

M 94

N 98

O 221

P 223

;

run;

/*view dataset*/

proc printdata=original_data;

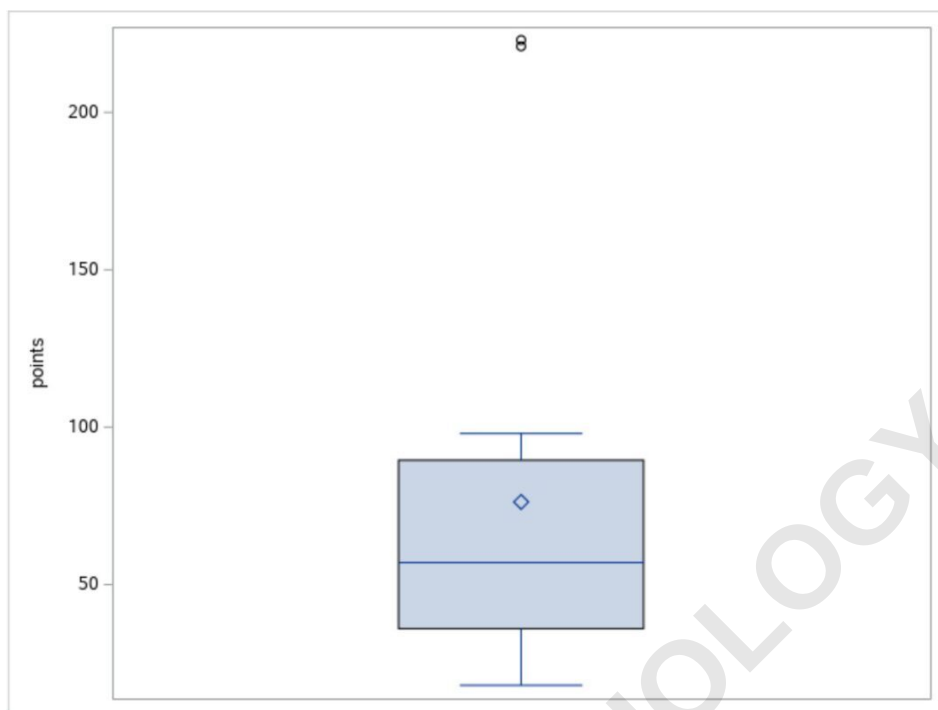
Obs	team	points
1	A	18
2	B	24
3	C	26
4	D	34
5	E	38
6	F	45
7	G	48
8	H	54
9	I	60
10	J	73
11	K	79
12	L	85
13	M	94
14	N	98
15	O	221
16	P	223

The easiest way to identify outliers in SAS is by creating a boxplot, which automatically uses the formula mentioned earlier to identify and display outliers in the dataset as tiny circles:

```
/*create boxplot to visualize distribution of points*/  
odsoutput sgplot=boxplot_data;  
proc sgplot data=original_data;  
vbox points;  
run;
```

```
/*view summary of boxplot descriptive statistics*/
```

```
proc printdata=boxplot_data;
```



Obs	BOX(points)___Y	BOX(points)_ST	points
1	18.000	MIN	18
2	36.000	Q1	24
3	57.000	MEDIAN	26
4	89.500	Q3	34
5	98.000	MAX	38
6	76.250	MEAN	45
7	62.076	STD	48
8	16.000	N	54
9	18.000	DATAMIN	60
10	223.000	DATAMAX	73
11	223.000	OUTLIER	79
12	221.000	OUTLIER	85
13	.		94
14	.		98
15	.		221
16	.		223

From the boxplot we can see that there are two tiny circles near the top of the plot. This indicates that there are two outliers.

From the table below the boxplot we can see the exact values for the two outliers: 221 and 223.

We can manually verify that these two values are outliers by using the formula from earlier:

The interquartile range is: $Q3 - Q1 = 89.5 - 36 = 53.5$.

The upper limit for outliers would be: $Q3 + 1.5 \cdot IQR = 89.5 + 1.5 \cdot 53.5 = 169.75$.

Since 221 and 223 are both greater than this value, they are classified as outliers.

We could then use the following code to remove these two outliers from the dataset if we'd like:

```
/*create new dataset with outliers removed*/  
data new_data;  
set original_data;  
if points >= 221 then delete;  
run;
```

```
/*view new dataset*/  
proc printdata=new_data;
```

Obs	team	points
1	A	18
2	B	24
3	C	26
4	D	34
5	E	38
6	F	45
7	G	48
8	H	54
9	I	60
10	J	73
11	K	79
12	L	85
13	M	94
14	N	98

Notice that the two outliers have been removed.

Additional Resources

The following tutorials explain how to perform other common tasks in SAS: