How to Easily Identify Outliers in SAS Using PROC MEANS

Authored by

stats writer

December 1, 2025

SAS provides a comprehensive suite of powerful statistical techniques essential for identifying unusual data points, commonly known as outliers, within a given dataset. Effective outlier detection is a critical step in any robust data analysis pipeline, as these extreme values can significantly distort statistical estimates and conclusions. While simple visualization tools like **box plots** offer immediate graphical insights, SAS also provides several advanced procedures for rigorous statistical examination.

Techniques available in SAS range from calculating basic descriptive statistics using the UNIVARIATE procedure, which can highlight extreme minimum and maximum values, to employing specialized options within procedures like PROC MEANS. For instance, utilizing the `OUTLIERS` option in **PROC MEANS** allows analysts to swiftly flag observations that deviate significantly, often defined as being more than three Standard Deviations away from the calculated mean. This systematic approach ensures that potentially influential data points are meticulously identified and assessed before proceeding with modeling or interpretation.

Understanding the Statistical Definition of an Outlier

An outlier is formally defined as an observation that lies an abnormal distance from other values in a random sample from a population. The presence of these data points can be highly problematic, as they can unduly influence statistical outcomes, leading to biased parameter estimates, incorrect hypothesis testing conclusions, and ultimately, flawed decision-making based on the analysis. Therefore, mastering the methods for reliable identification and appropriate handling of these extreme values is paramount for data integrity.

The most widely accepted and mathematically robust standard method for identifying mild and extreme outliers in univariate data distributions involves calculating the limits based on the Interquartile Range (IQR). This method is preferred when the underlying distribution is not assumed to be normal or when the data may contain asymmetry, offering a non-parametric measure of spread that is resistant to the influence of the outliers themselves.

The Interquartile Range (IQR) Rule

The Interquartile Range (IQR) is a fundamental measure of statistical dispersion, calculated as the difference between the 75th percentile (the third quartile, **Q3**) and the 25th percentile (the first quartile, **Q1**) within a dataset. Essentially, the IQR captures the range spanned by the middle 50% of the data values, providing a stable measure of variability around the median. This range forms the basis for setting fences that define the boundary beyond which observations are considered atypical.

The standard statistical rule for defining an observation as an outlier relies on multiplying the IQR

by a constant factor, typically 1.5. An observation is classified as an outlier if it falls outside the range defined by these fences:

Upper Fence: Q3 + 1.5 × IQR
Lower Fence: Q1 - 1.5 × IQR

Any data point exceeding the Upper Fence or falling below the Lower Fence is statistically classified as an outlier. This widely accepted formula provides a standardized, visual, and mathematical criterion for identifying potentially influential observations.

The core formula defining these extreme observations is summarized as:

Outliers = Observations > Q3 + 1.5*IQR or < Q1 - 1.5*IQR

Example Setup: Creating the SAS Dataset

The following demonstration illustrates the practical application of the IQR rule to identify and manage outliers within a statistical dataset using the SAS statistical software environment. We begin by creating a simple dataset containing points scored by various teams, intentionally including some extreme values to serve as our target outliers.

We will utilize standard SAS data creation steps, including the `DATA`, `INPUT`, and `DATALINES` statements, to construct the initial data structure. This is followed by a `PROC PRINT` step to verify the successful creation and structure of the dataset, ensuring the data is correctly loaded for subsequent analysis procedures. Note the inclusion of two significantly higher values (221 and 223) which are expected to be flagged as outliers.

```
/*create dataset: 'original_data'*/
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 structure and contents*/
proc print data=original_data;

The result of executing the `PROC PRINT` step confirms the dataset structure:

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

Visualizing Outliers using PROC SGPLOT Box Plots

While manual calculation is possible, the most efficient and visually intuitive way to identify outliers in SAS is by generating a box plot. The **box plot** (or box-and-whisker plot) is specifically designed to visualize the distribution of data based on quartiles, and crucially, it automatically uses the 1.5 × IQR formula to determine and display outliers as distinct markers (often small circles) outside the

whiskers.

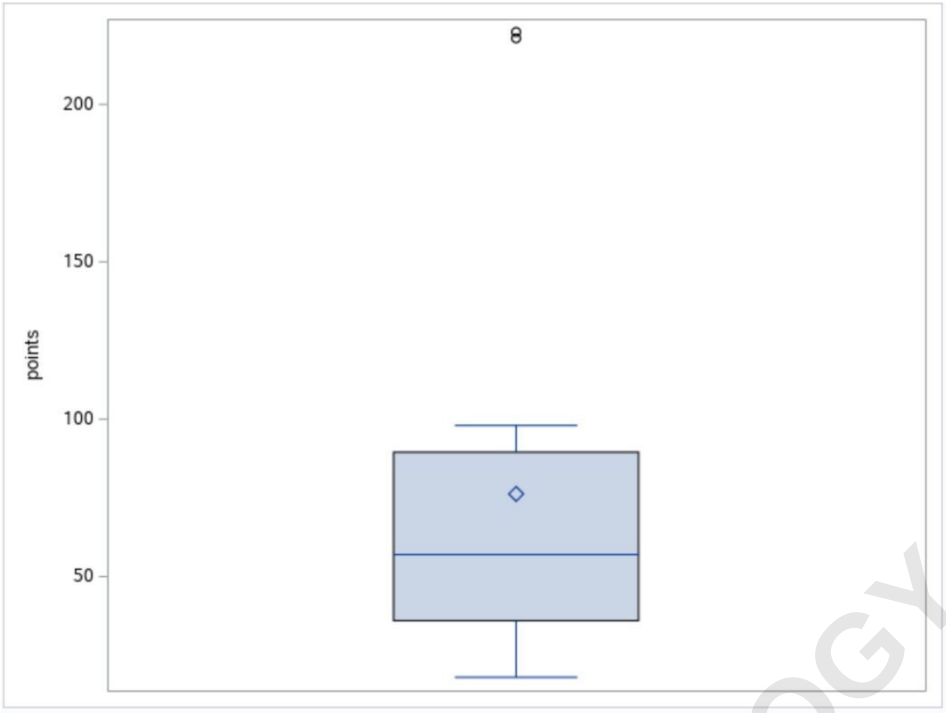
We employ the PROC SGPLOT procedure, which is part of the Statistical Graphics (SG) procedures, utilizing the `VBOX` statement to generate a vertical box plot of the `points` variable. Furthermore, we use the ODS OUTPUT statement to capture the underlying descriptive statistics calculated by the box plot procedure (such as Q1, Median, Q3, and the outlier values) into a separate SAS dataset named `boxplot_data` for later review and verification.

```
/*create boxplot to visualize distribution of points and identify outliers*/
ods output sgplot=boxplot_data;
proc sgplot data=original_data;
vbox points;
run;

/*view summary of boxplot descriptive statistics generated by ODS OUTPUT*/
proc print data=boxplot_data;
```

Interpreting the Box Plot and Statistical Output

Upon reviewing the generated graphical output, the box plot clearly displays the central distribution of the data. The rectangular box represents the Interquartile Range (IQR) (from Q1 to Q3), with the line inside indicating the median. Critically, we observe two small, isolated circular markers positioned significantly above the upper whisker of the plot.



These distinct circles represent the two observations that violate the 1.5 × IQR rule, confirming them as statistical outliers in the dataset. The summary table generated by the `ODS OUTPUT` statement and subsequently displayed via `PROC PRINT` provides the exact numerical confirmation of these extreme values, offering detailed descriptive statistics used in the plot construction.

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

Analyzing the descriptive statistics table, we can pinpoint the exact values identified as outliers: **221** and **223**. Furthermore, this table provides the essential quartile statistics (Q1, Median, Q3) necessary for manually verifying the boundaries used by the box plot visualization.

Manual Verification using the IQR Formula

To ensure complete transparency and understanding of the mechanism, we can manually apply the 1.5 × IQR rule using the quartile values obtained from the SAS output. From the statistics table, let us assume the following values:

Q1 (25th percentile) ≈ 36.0
Q3 (75th percentile) ≈ 89.5

First, we calculate the **Interquartile Range (IQR)**:

IQR = Q3 - Q1 = 89.5 - 36.0 = **53.5**.

Next, we calculate the theoretical **Upper Fence**, which defines the maximum value considered non-outlier:

Upper Limit = Q3 + 1.5 × IQR = 89.5 + (1.5 × 53.5) = 89.5 + 80.25 = **169.75**.

Since both observations **221** and **223** are substantially greater than the calculated upper fence of

169.75, they are unequivocally classified and confirmed as statistical outliers according to the 1.5 × IQR rule. This manual calculation validates the automatic detection performed by PROC SGPLOT.

Handling Outliers: Creating a Clean Dataset in SAS

Once outliers are reliably identified, the analyst must decide on the appropriate course of action, which may include retaining, transforming, or removing them, depending on the research context and the suspected cause (e.g., measurement error vs. natural variability). If the decision is made to exclude these extreme points--for instance, to prevent them from skewing regression models--this can be easily achieved in SAS using a conditional data step.

We create a new dataset, `new_data`, based on the original data, and implement a conditional `DELETE` statement. This statement efficiently removes any observation where the `points` variable meets or exceeds the minimum outlier value (221 in this case). This procedure effectively filters the dataset, resulting in a cleansed sample ready for further inferential analysis.

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

/*view the filtered dataset*/
proc print data=new_data;
```

The final output confirms that the two extreme data points corresponding to the outliers have been successfully removed from the dataset, yielding a cleaned structure suitable for analysis assuming normality or less sensitivity to variance.

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

Conclusion: Essential Techniques for Data Quality

Identifying and appropriately managing outliers is a non-negotiable step toward ensuring the quality and reliability of statistical analysis. As demonstrated, SAS provides robust and efficient tools, particularly the visual power of PROC SGPLOT combined with the statistical output capabilities of ODS OUTPUT, allowing analysts to quickly detect, verify, and address these influential observations. Whether using the IQR method through visualization or employing parametric techniques like the Standard Deviations rule in other procedures, SAS ensures that data cleansing is both accurate and streamlined.

For continued mastery of statistical analysis in SAS, consult the following related tutorials: