How to Calculate Cook’s Distance in SPSS to Identify Influential Data Points

Authored by

stats writer

January 24, 2026

Cook's Distance is a statistical measure used to identify influential data points in a regression analysis. In SPSS (Statistical Package for the Social Sciences), Cook's Distance can be calculated by first running a regression analysis and then selecting the "Plots" option. From there, the user can select "Cook's Distance" as the type of plot to be produced. This will generate a graph showing the Cook's Distance values for each observation. The larger the value, the more influential the data point is on the overall regression model. By identifying and addressing these influential points, the accuracy and reliability of the regression analysis can be improved.

Calculate Cook's Distance in SPSS

Cook's distance is used to identify influential in a regression model.

The formula for Cook's distance is:

Di = (ri2 / p*MSE) * (hii / (1-hii)2)

where:

ri is the ith residualp is the number of coefficients in the regression modelMSE is the mean squared errorhii is the ith leverage value

Cook's distance effectively measures how much all of the fitted values in the model change when the ith observation is deleted.

The larger the value for Cook's distance, the more influential a given observation.

A general rule of thumb is that any observation with a Cook's distance greater than 4/n (where *n* = total observations) is considered to be highly influential.

The following example shows how to calculate Cook's distance for a regression model in SPSS:

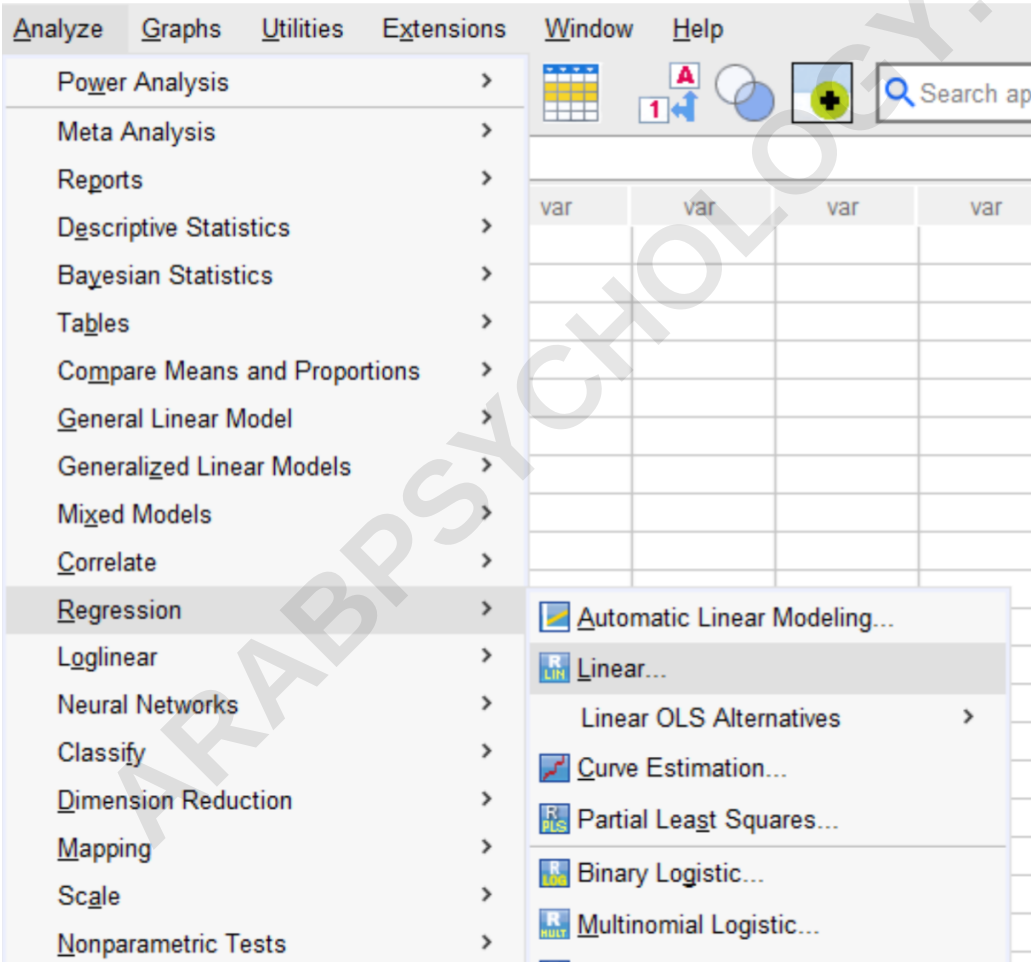
Example: How to Calculate Cook's Distance in SPSS

Suppose we have the following dataset in SPSS that contains information about total ad spend and total sales for 12 different retail stores:

	Store_ID	Ad_Spend	Sales	var
1	0001	8	41	
2	0002	12	42	
3	0003	12	39	
4	0004	13	37	
5	0005	14	35	
6	0006	16	39	
7	0007	17	45	
8	0008	22	46	
9	0009	24	39	
10	0010	26	49	
11	0011	29	55	
12	0012	30	57	
13				
14				
15				
16				
17				

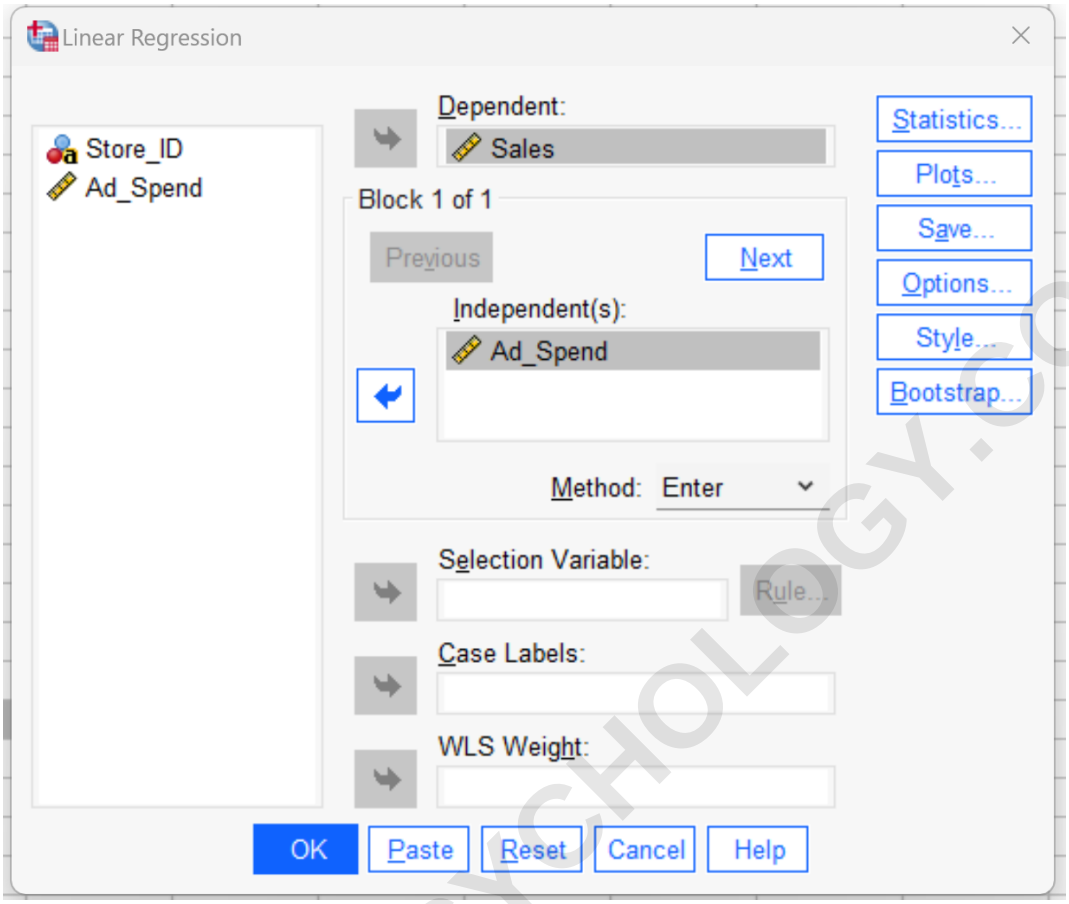
Suppose that we would like to fit a simple linear regression model to this dataset, using Ad_Spend as the predictor variable and Sales as the response variable.

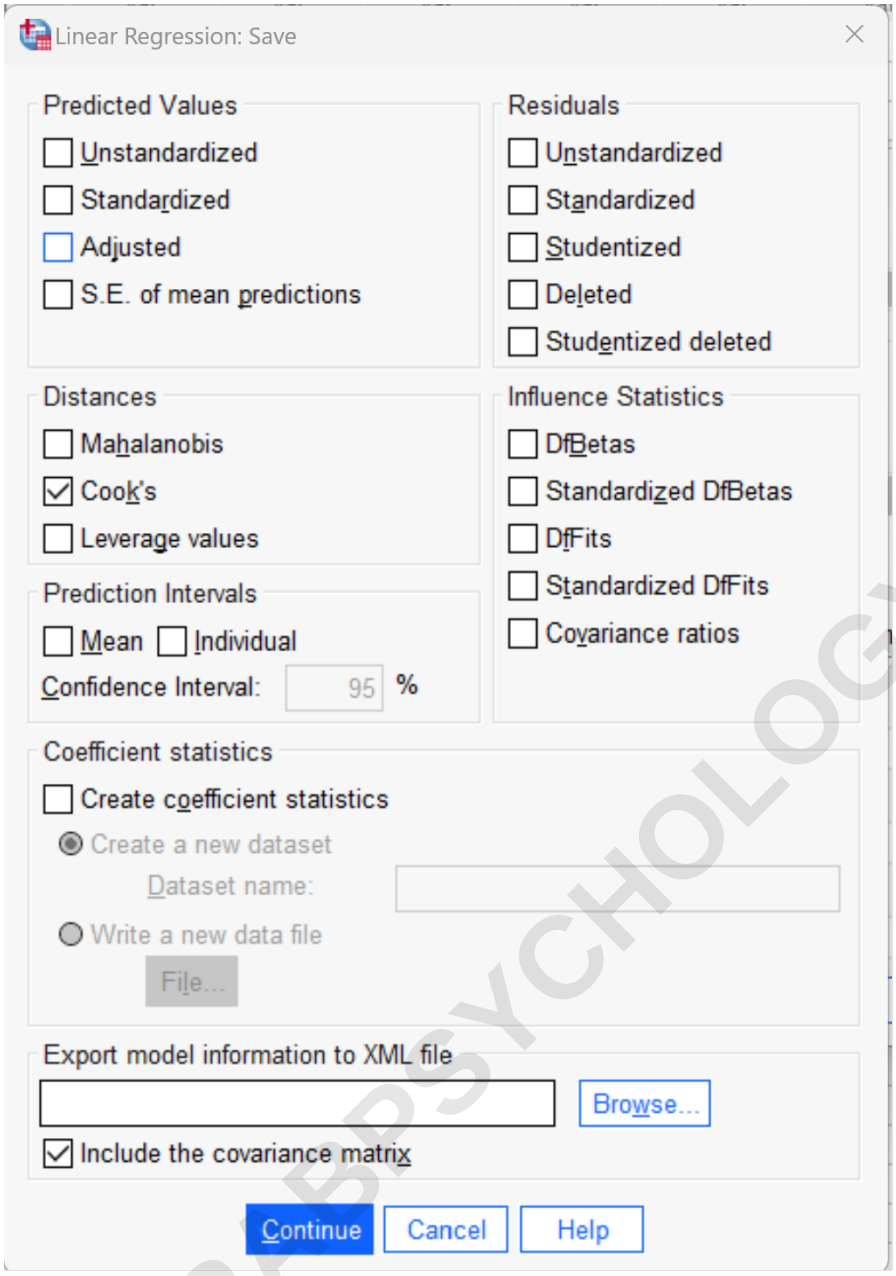
To do so, click the Analyze tab, then click Regression, then click Linear:



In the new window that appears, drag Sales to the Dependent panel and then drag Ad_Spend to the

Independent panel:





Then click Continue. Then click OK.

A new variable will be created in the Data View named COO_1 that shows Cook's distance for each observation:

	Store_ID	Ad_Spend	Sales	COO_1	var
1	0001	8	41	.36813	
2	0002	12	42	.06075	
3	0003	12	39	.00052	
4	0004	13	37	.02764	
5	0005	14	35	.10487	
6	0006	16	39	.02155	
7	0007	17	45	.01705	
8	0008	22	46	.00020	
9	0009	24	39	.34275	
10	0010	26	49	.00047	
11	0011	29	55	.15003	
12	0012	30	57	.34948	
13					
14					
15					
16					
17					
18					
19					

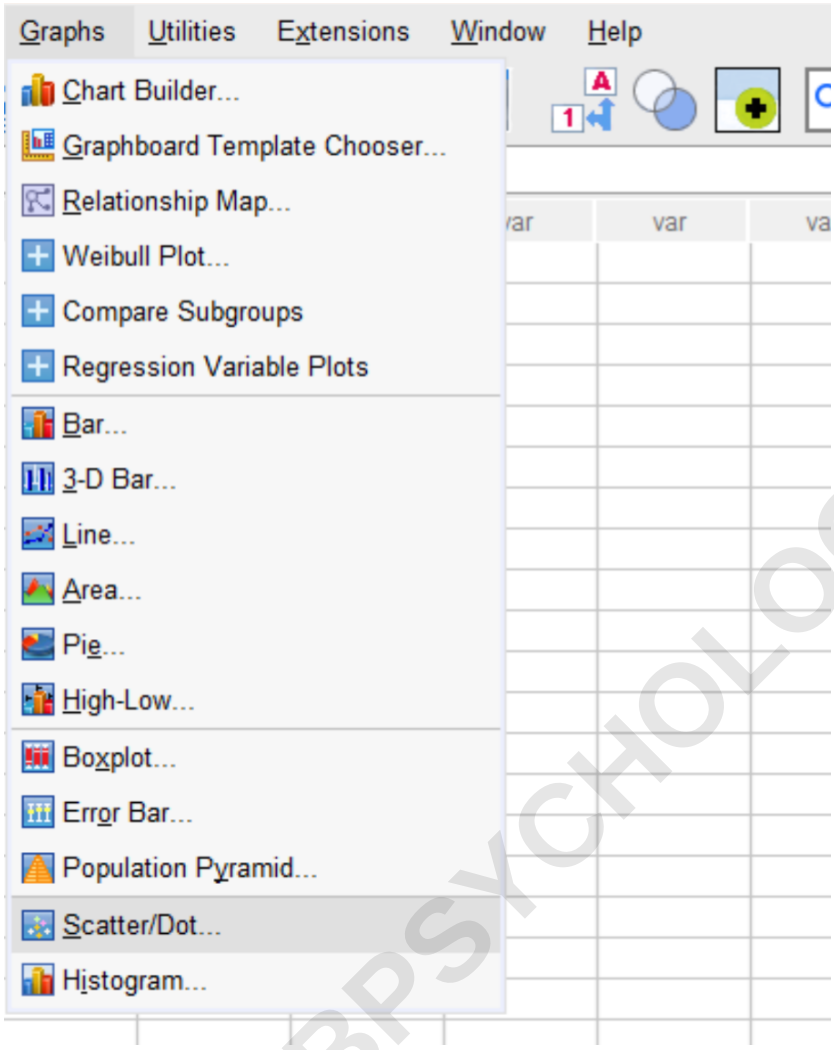
Recall that, as a rule of thumb, any observation with a Cook's distance greater than 4/n is considered to be highly influential.

In this particular dataset there are 12 observations, so any observation with a Cook's distance greater than 4/12 = 0.333 is considered to be highly influential.

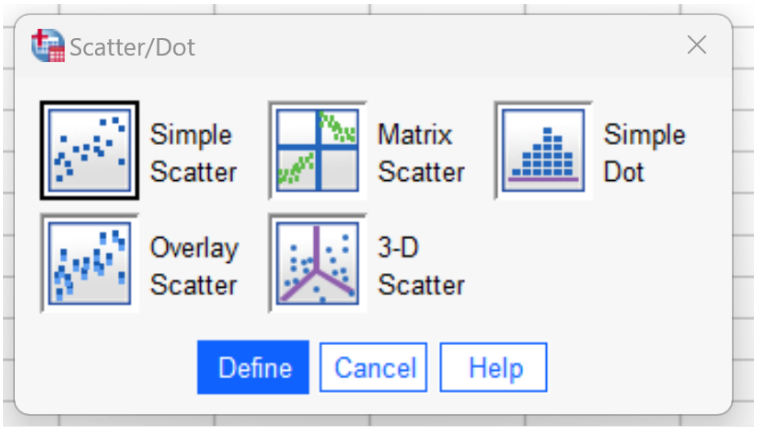
We can see that three observations in the dataset are just above this threshold.

To visualize the Cook's distance values for each

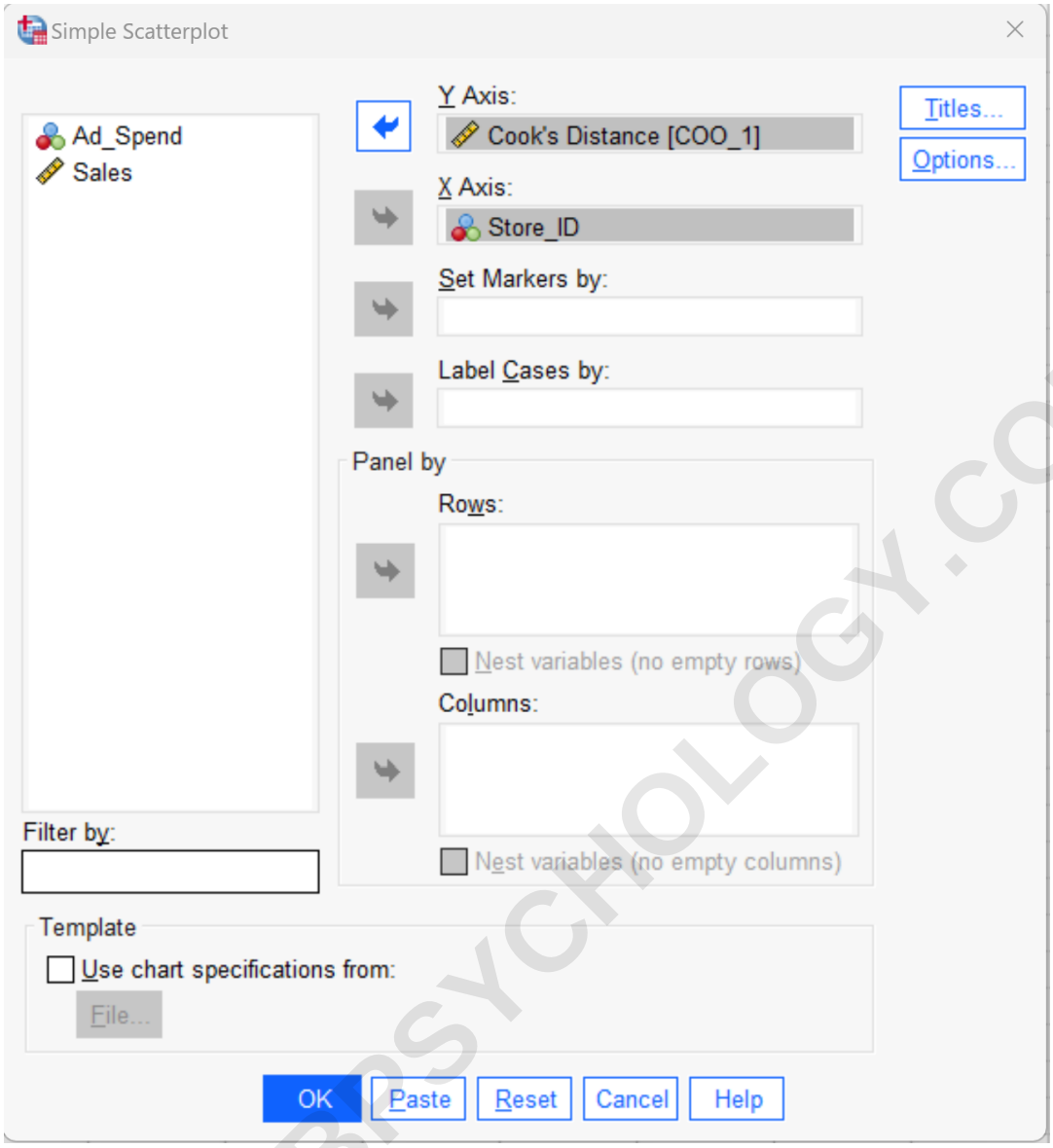
observation, click the Chart tab, then click Scatter/Dot:



Then click Simple Scatter:

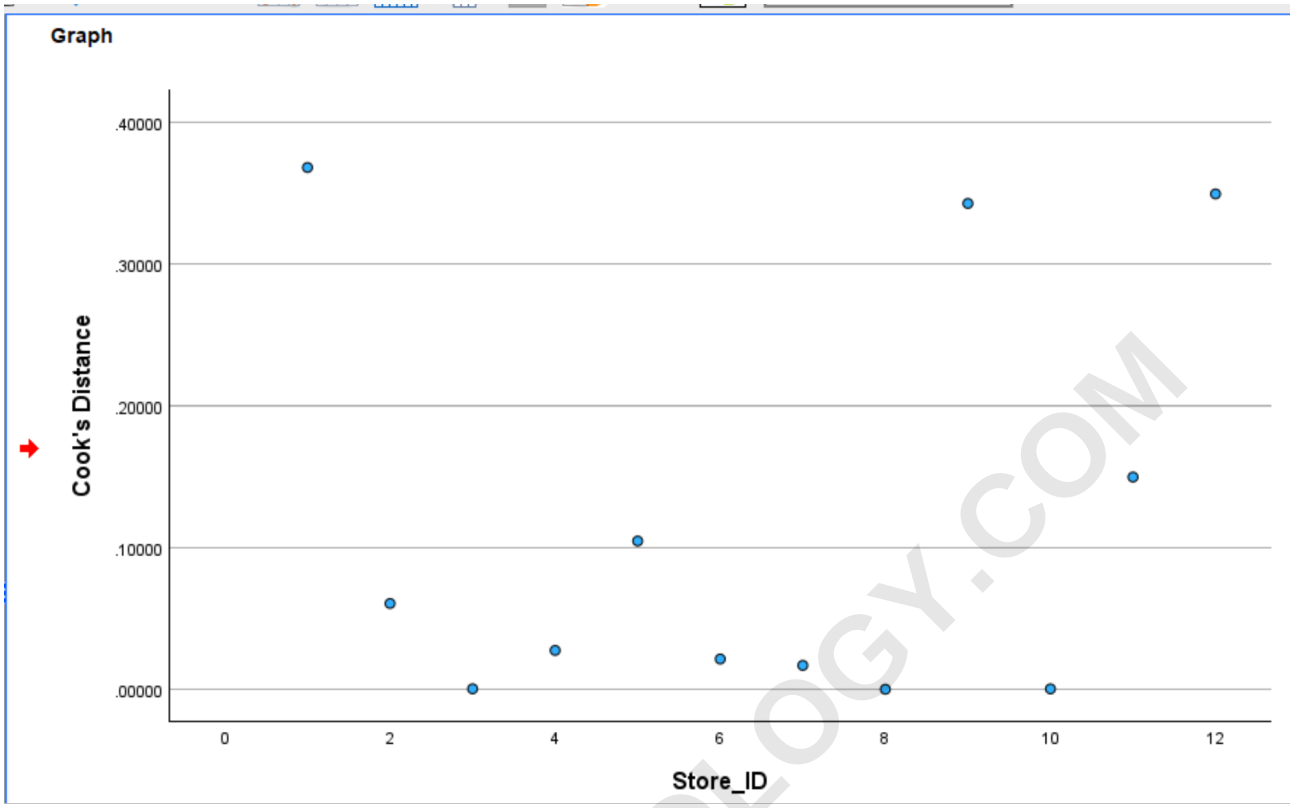


In the new window that appears, drag Store_ID to the X Axis and Cook's Distance to the Y Axis:



Then click OK.

The following scatterplot will be generated that shows the 12 store ID's along the x-axis and Cook's distance along the y-axis:



This plot helps us visualize Cook's distance for each observation and allows us to quickly spot which observations have the highest Cook's distance values.

Notes on Cook's Distance

It's important to keep in mind that Cook's Distance should be used as a way to identify potentially influential observations.

Just because an observation is influential doesn't necessarily mean that it should be deleted from the dataset. First, you should verify that the observation

isn't a result of a data entry error or some other odd occurrence.

If it turns out to be a legitimate value, you can then decide if it's appropriate to delete it, leave it, or replace it with an alternative value like the median.