What is the process for removing outliers from multiple columns in R?

Authored by

stats writer

April 20, 2024

The process for removing outliers from multiple columns in R involves several steps. First, the data must be imported into R and organized into a data frame. Next, the outliers must be identified using statistical techniques such as the interquartile range or Z-score method. Once the outliers have been identified, they can be removed from the data frame using the subset function or by replacing them with missing values. Finally, the data can be re-analyzed and visualized to ensure that the outliers have been successfully removed. It is important to note that the specific method for removing outliers may vary depending on the nature of the data and the research question.

Remove Outliers from Multiple Columns in R

Often you may want to remove outliers from multiple columns at once in R.

One common way to define an observation as an outlier is 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).

Using this definition, we can use the following steps to create a simple function to identify outliers and then apply this function across multiple columns in an R data frame.

Step 1: Create data frame.

First, let's create a data frame in R:

df <- data.frame(index=c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10),

```
var1=c(4, 4, 5, 4, 3, 2, 8, 9, 4, 5),
var2=c(1, 2, 4, 4, 6, 9, 7, 8, 5, 29),
var3=c(9, 9, 9, 5, 5, 3, 4, 5, 11, 34))
```

Step 2: Define outlier function.

Next, let's define a function that can identify outliers and a function that can then remove outliers:

```
outliers <- function(x) {

Q1 <- quantile(x, probs=.25)
Q3 <- quantile(x, probs=.75)
iqr = Q3-Q1

upper_limit = Q3 + (iqr*1.5)
lower_limit = Q1 - (iqr*1.5)

x > upper_limit | x < lower_limit
}

remove_outliers <- function(df, cols = names(df)) {
for (col in cols) {
df <- df]),]
}
df
```

}

Step 3: Apply outlier function to data frame.

Lastly, let's apply this function across multiple columns of the data frame to remove outliers:

remove_outliers(df, c('var1', 'var2', 'var3'))

index var1 var2 var3
1 1 4 1 9
2 2 4 2 9
3 3 5 4 9
4 4 4 4 5
5 5 3 6 5
9 9 4 5 11

You can find more R tutorials here.