

How can I label outliers in boxplots using ggplot2?

Authored by
stats writer

June 27, 2024

RECOMMENDED CITATION

stats writer (2024). *How can I label outliers in boxplots using ggplot2?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=154761>

The process of labeling outliers in boxplots using ggplot2 involves utilizing the "stat_boxplot" function to create a boxplot, followed by adding the "geom_text" layer to the plot to label the outliers. This can be done by specifying the dataset and the variables to be used, and then using the "outliers" argument to specify the outlier values. The "position_dodge" function can also be used to adjust the position of the labels. This method allows for easy identification and labeling of the outliers in the boxplot, providing a clear representation of the data.

Label Outliers in Boxplots in ggplot2

This tutorial provides a step-by-step example of how to label in boxplots in ggplot2.

Step 1: Create the Data Frame

First, let's create the following data frame that contains information on points scored by 60 different basketball players on three different teams:

```
#make this example reproducible
```

```
set.seed(1)
```

```
#create data frame
```

```
df <- data.frame(team=rep(c('A', 'B', 'C'), each=20),
```

```
player=rep(LETTERS, times=3),
```

```
points=round(rnorm(n=60, mean=30, sd=10), 2))
```

```
#view head of data frame
```

```
head(df)
```

team player points

1 A A 23.74

2 A B 31.84

3 A C 21.64

4 A D 45.95

5 A E 33.30

6 A F 21.80

Note: We used the function to ensure that this example is reproducible.

Step 2: Define a Function to Identify Outliers

In ggplot2, an observation is defined as an outlier if it meets one of the following two requirements:

The observation is 1.5 times the interquartile range less than the first quartile (Q1)
The observation is 1.5 times the interquartile range greater than the third quartile (Q3).

We can create the following function in R to label observations as outliers if they meet one of these two requirements:

```
find_outlier <- function(x) {
```

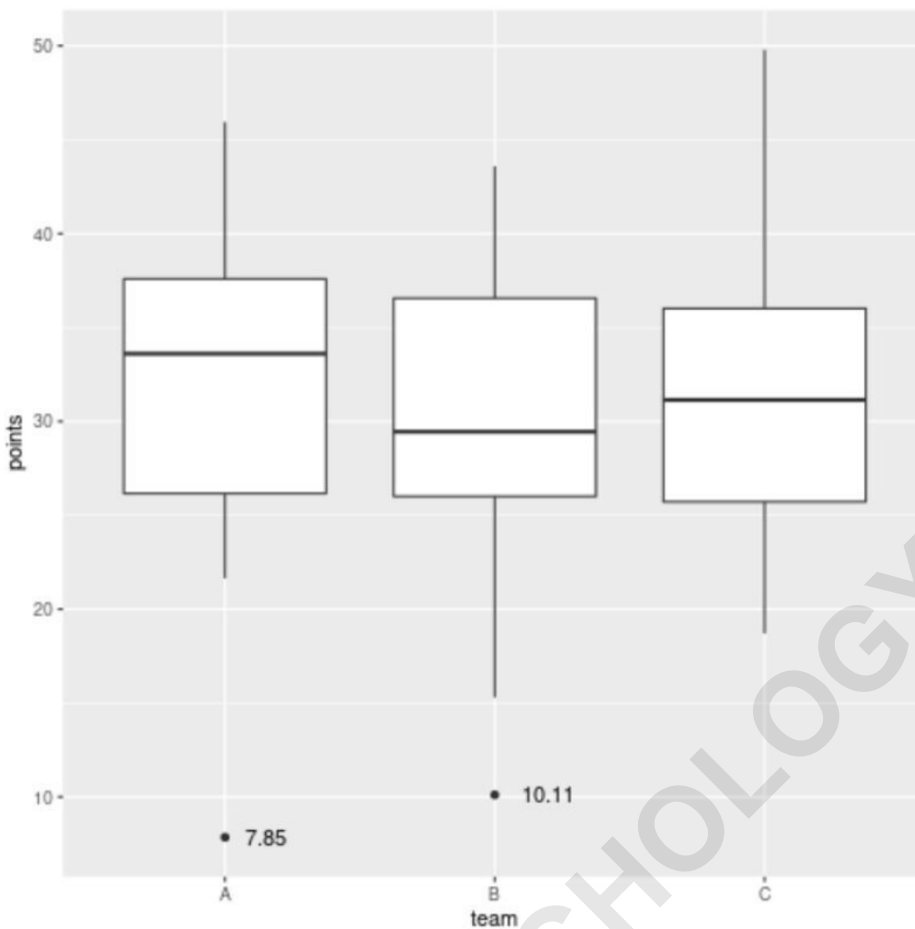
```
return(x < quantile(x, .25) - 1.5*IQR(x) | x > quantile(x,  
.75) + 1.5*IQR(x))  
}
```

Related:

Step 3: Label Outliers in Boxplots in ggplot2

Next, we can use the following code to label outliers in boxplots in ggplot2:

```
library(ggplot2)  
library(dplyr)  
  
#add new column to data frame that indicates if each  
observation is an outlier  
df <- df %>%  
group_by(team) %>%  
mutate(outlier = ifelse(find_outlier(points), points, NA))  
  
#create box plot of points by team and label outliers  
ggplot(df, aes(x=team, y=points)) +  
geom_boxplot() +  
geom_text(aes(label=outlier), na.rm=TRUE, hjust=-.5)
```



Notice that two outliers are labeled in the plot.

The first outlier is a player on team A who scored 7.85 points and the other outlier is a player on team B who scored 10.11 points.

For example, we could swap out points for player in the mutate() function to instead label the outliers based on the player name:

library(ggplot2)

```
library(dplyr)
```

```
#add new column to data frame that indicates if each  
observation is an outlier
```

```
df <- df %>%
```

```
group_by(team) %>%
```

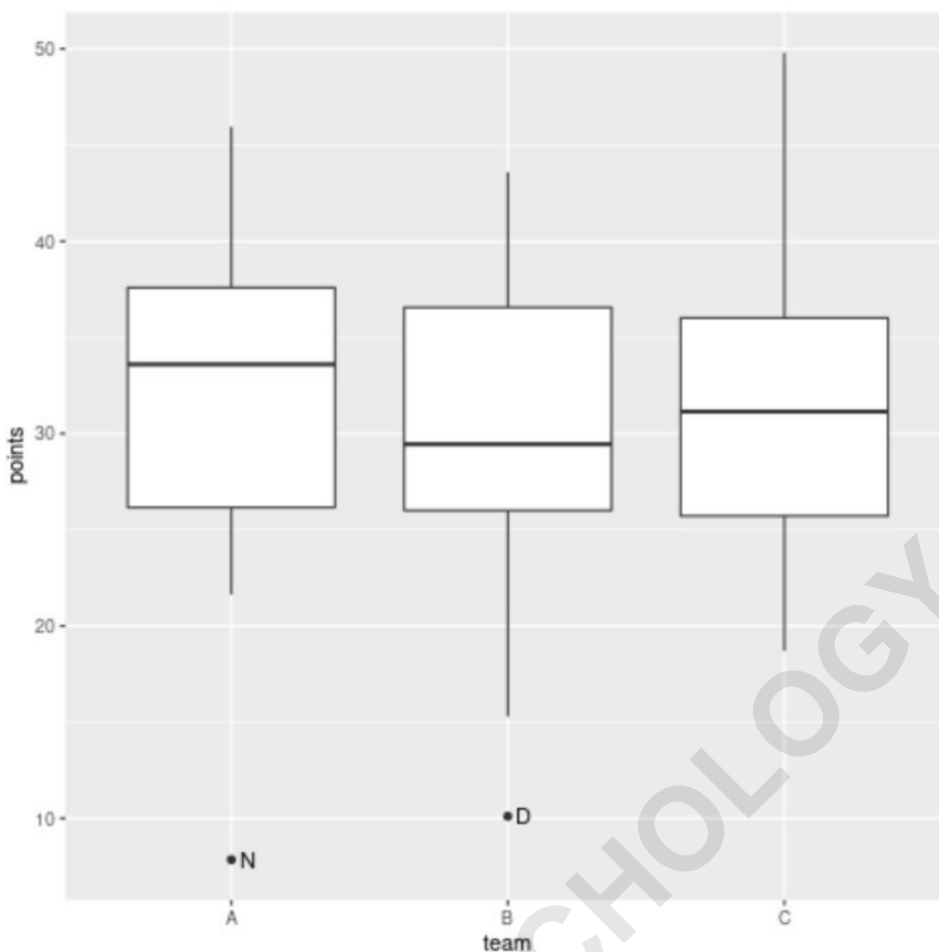
```
mutate(outlier = ifelse(find_outlier(points), player, NA))
```

```
#create box plot of points by team and label outliers
```

```
ggplot(df, aes(x=team, y=points)) +
```

```
geom_boxplot() +
```

```
geom_text(aes(label=outlier), na.rm=TRUE, hjust=-.5)
```



The outlier on team A now has a label of N and the outlier on team B now has a label of D, since these represent the player names who have outlier values for points.

Note: The argument in `geom_text()` is used to push the label horizontally to the right so that it doesn't overlap the dot in the plot.

The following tutorials explain how to perform other

common tasks in ggplot2:

ARABPSYCHOLOGY.COM