Module3_4Box_Plots_Reviewed

May 17, 2021

BOX PLOTS

0.1 Table of Contents

##

Introduction

In this notebook, we are going to explore how to create box plots in R. Box plots are a convenient way to represent the degree of dispersion (spread) and skewness in the data, and show outliers without making any assumption of the underlying statistical distribution.Let us first install the package plotly.

[2]:
```
install.packages("plotly")
library(plotly)
```

```
Installing package into ‘/resources/common/R/Library’
(as ‘lib’ is unspecified)
also installing the dependencies ‘htmltools’, ‘jsonlite’, ‘yaml’, ‘viridisLite’,
‘htmlwidgets’, ‘hexbin’, ‘purrr’

Loading required package: ggplot2

Attaching package: ‘plotly’

The following object is masked from ‘package:ggplot2’:

    last_plot

The following object is masked from ‘package:stats’:

    filter

The following object is masked from ‘package:graphics’:
```

```
    layout

The following objects are masked from 'package:SparkR':

    arrange, distinct, filter, group_by, mutate, rename, schema, select
```

##

Box Plot

First, we generate data to represent with a box plot. We will use two normal distributions with slightly different parameters to generate our samples so you can see the effect it has on the plot.

[3]:
```
#making the results reproducible
set.seed(1234)

set_a <- rnorm(200, mean=1, sd=2)
set_b <- rnorm(200, mean=0, sd=1)

#create the data frame
df <- data.frame(label = factor(rep(c("A","B"), each=200)), value = c(set_a,␣
 ↪set_b))

#output both the first and last rows
head(df)
tail(df)
```

[3]:
```
  label     value
1     A -1.414131
2     A  1.554858
3     A  3.168882
4     A -3.691395
5     A  1.858249
6     A  2.012112
```

[3]:
```
    label       value
395     B  0.52874502
396     B  0.78939440
397     B  0.45709951
398     B  0.53883312
399     B  0.01464312
400     B -0.91648914
```

As you can see, we have randomly generated two sets, labelling them respectively A and B.

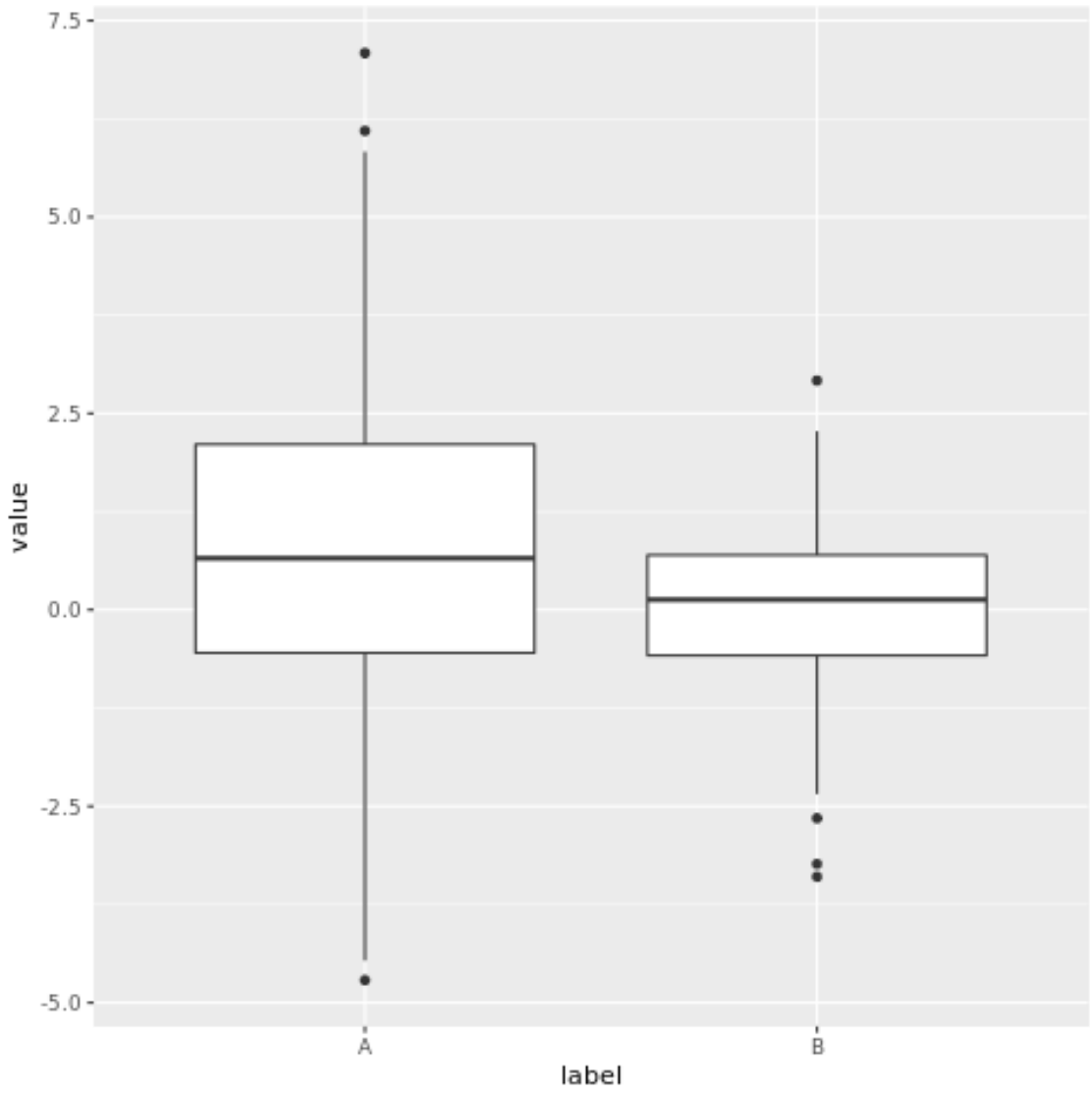
0.1.1 geom_bloxplot()

We create a plot with the ggplot function and specify the x and y-axis of the plot. Then we add the boxplot to our plot, which results in creating one boxplot for each value of x.

[4]:
```
ggplot(df, aes(x=label, y=value)) + geom_boxplot()

ggplotly()
```

[4]:



[4]:

0.1.2 Now onto a brief explanation of what a box plot is, for those who don't know

As was briefly metioned in the Introduction section, box plots are good to indicate dispertion. They do so by providing visual representations in terms of quartiles.

- The thickest line in the middle of the rectangle represents the median value (second quartile).

- The bottom and top of the rectangle represent the first and third quartiles, respectively.
- The height of the rectangle equals the IQR (interquatile range), which is the difference between the first and third quartiles.
- The superior line reaches up to the largest value that is not larger than 1.5*IQR; Values that are larger are considered outliers and represented as dots. (the inferior line is analogous).

0.1.3 About the Author:

Hi! It's Hugo Sales Correa, the authors of this notebook. We hope you found R easy to learn! There's lots more to learn about R but you're well on your way. Feel free to connect with us if you have any questions.