Homework 3 Boxplot Visualization 1

Rony G. Younan

2024-11-07

```r
setwd("/cloud/project")

#install reshape package to reshape the data.
install.packages("reshape")

#load the package and its dependencies
library(reshape)

#installing package needed to read .xlsx file.
install.packages("readxl")

#loading the package
library(readxl)

#Referencing the group name as Text Messages.
T.M. <- read.csv("TextMessages.csv")

# Update to the data; change from wide data to long data
long_data <- melt(T.M., id.vars= c("Participant","Group"), variable.name
                  = "Time", value.name="Value")

#seeing how many rows our data has.
nrow(long_data)
```

```
## [1] 100
```

```r
#Our Data has 100 rows.

#Installing ggplot package which will help us create the box plot.
install.packages("ggplot2")

#Loading ggplot package and its dependencies
library(ggplot2)

long_data$Group <- as.factor(long_data$Group)

is.factor(long_data$Group)
```

```
## [1] TRUE
```

```r
#Group is now a factor.

#The ggplot2(function) requires several inputs within the parenthesis:
#(1) the data set name is specified; next, the aesthetics for the x-y axis and
```
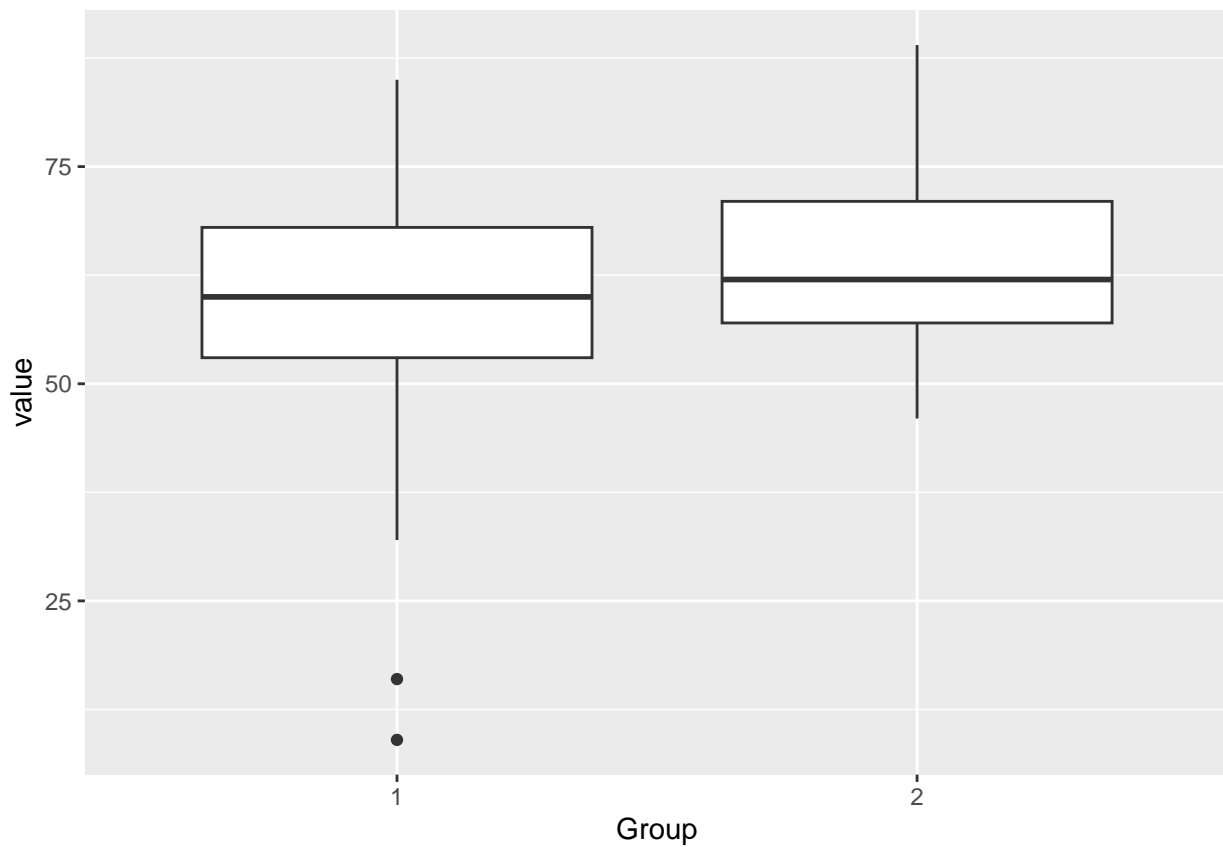
```
#the categorical variable (Group) will be specified on the x-axis and the
#continuous variable (value) will be on the y-axis, and (2) create an object
#called boxplot_T.M.ByGroup that will store this information.
boxplot_T.M.ByGroup <- ggplot(long_data,aes(Group,value))

boxplot_T.M.ByGroup
```
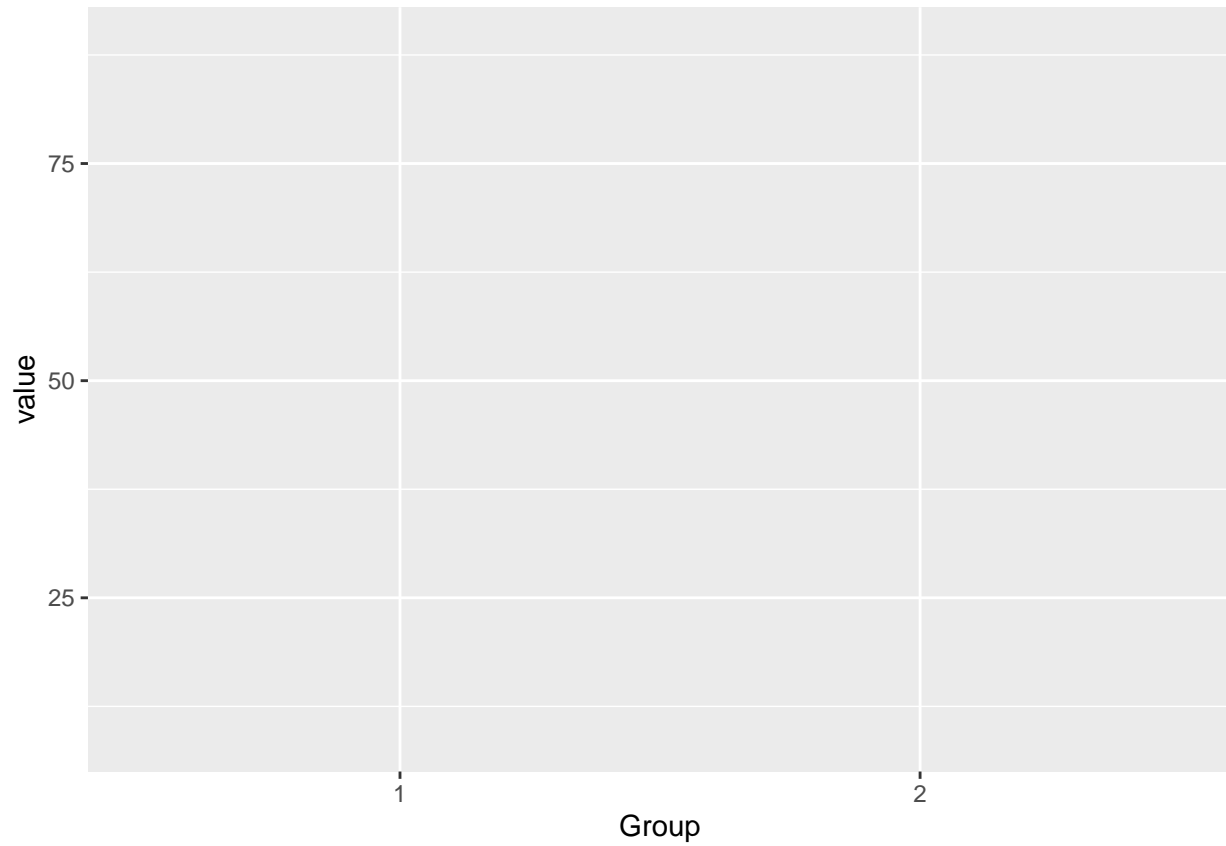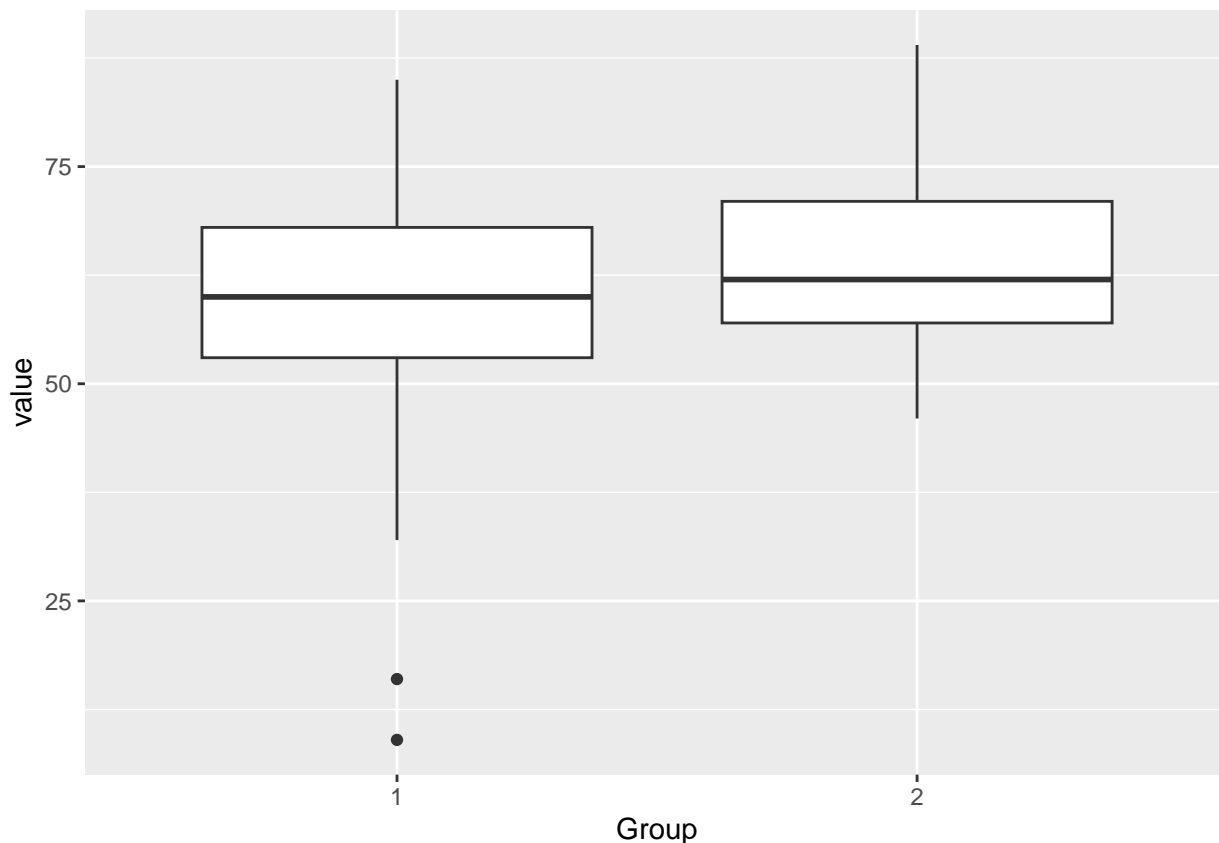


```
boxplot_T.M.ByGroup+geom_boxplot()
```

```
#The following function labs() will allow us to add user-defined labels to the
# x-y axis other than the variable names
boxplot_T.M.ByGroup+geom_boxplot()+labs(x="Group", y="value")
```

```
#We have to properly set the limits of the y-axis so that the plot is centered,
#and the plot is fully visible by first extracting the min and max values of
#the outcome variable value.
min(long_data$value)
```

```
## [1] 9
```

```
#The minimum Baseline for the entire sample is 9.
max(long_data$value)
```

```
## [1] 89
```

```
#The maximum Baseline for the entire sample is 89.

#We will now add the scale_y_continuous() function to overlay because it will
#contain new user-defined limits captured for the y-axis that will incorporate
#the min and max values extracted from our sample. Using this command, we set
#the new limits from 0 to 100 as these limits will include the min of 9 and
#the max of 89.
boxplot_T.M.ByGroup+geom_boxplot()+labs(x="Group", y="value")+
  scale_y_continuous(limits=c(0,100))
```
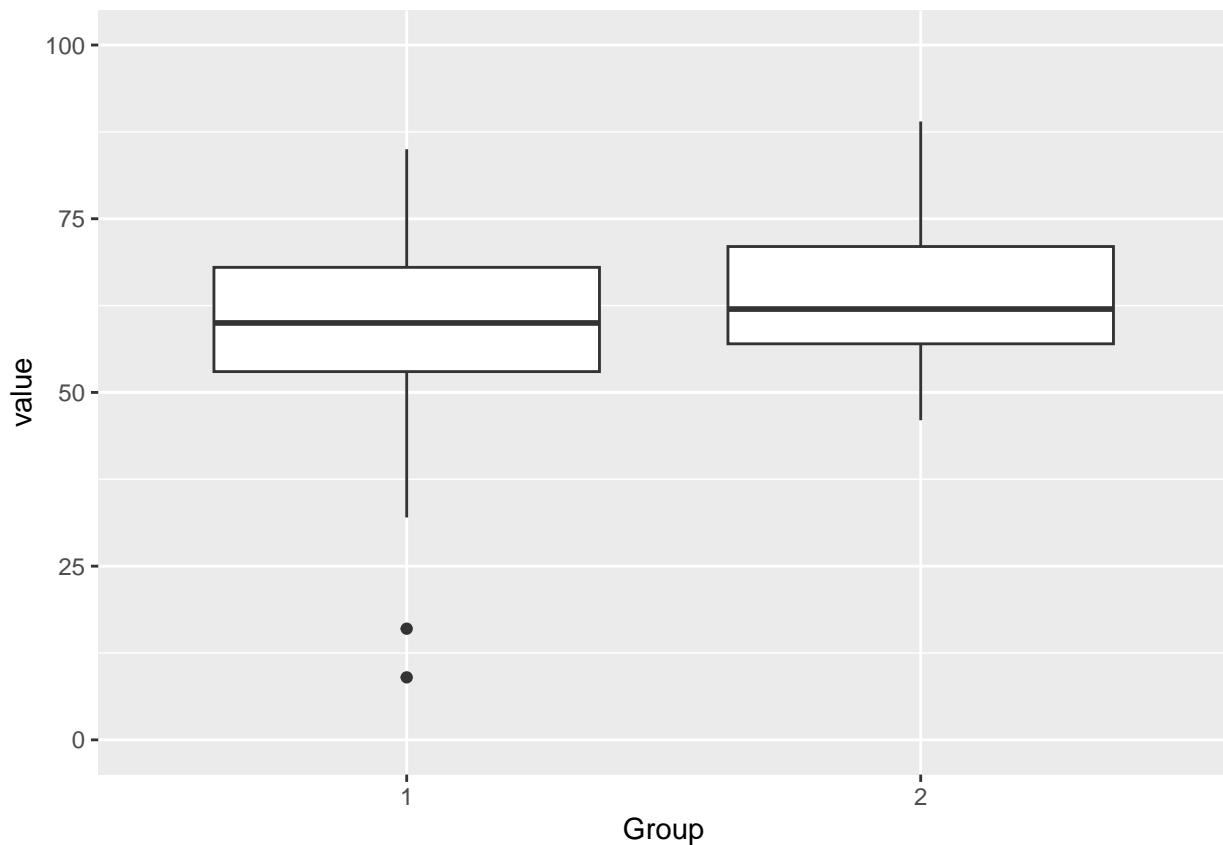
```
#Another layer will be added to add tick marks on the new user-defined limits
#on the y-axis by using the breaks() command as one of the inputs of the
#Scale_y_continuous command. Using this, we will place tick marks between
#0 and 100 increasing by 5 with every tick mark:
boxplot_T.M.ByGroup+geom_boxplot()+labs(x="Group", y="value")+
  labs(x="Group", y="value")+
  scale_y_continuous(limits=c(0,100),breaks=seq(from=0, to=100, by=5))
```
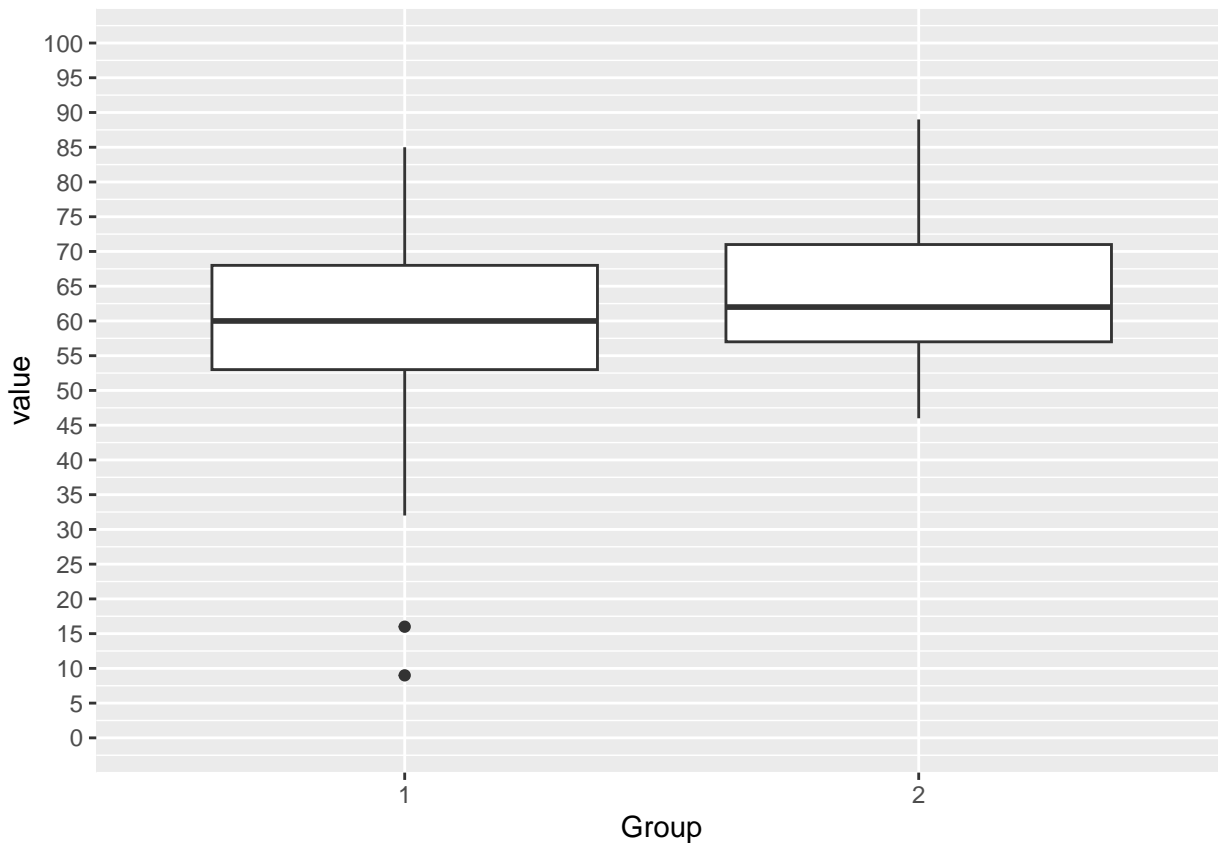
```r
#It can be seen that both group 1 and group 2 have similar median values,
#group 1 also has two outliers.

install.packages("pastecs")

#load the pastecs package and its dependencies.
library(pastecs)

colnames(long_data)
```

```
## [1] "Participant" "Group"       "variable"    "value"
```

```r
#We will output the median of text messages stratified by Group.
by(long_data$value,long_data$Group, median)
```

```
## long_data$Group: 1
## [1] 60
## ----------------------------------------------------------------
## long_data$Group: 2
## [1] 62
```

```r
#We can conclude that the median is very close for both groups with group 1
#having a median of 60 and group 2 having a median of 62.

#output the range of text messages stratified by Group by using the by() command
# again from the pastecs package
by(long_data$value,long_data$Group, range)
```

```
## long_data$Group: 1
```

```
## [1]  9 85
## ------------------------------------------------------------
## long_data$Group: 2
## [1] 46 89
```

```r
#Range is slightly different between both groups with group 1 having a min of 9
# and a max of 85, while group 2 has a min of 46 and a max of 89.

#we will now output and compare quantiles of text messages stratified by Group
#using the by() command again.
by(long_data$value,long_data$Group, quantile)
```

```
## long_data$Group: 1
##    0%   25%   50%   75%  100%
##     9    53    60    68    85
## ------------------------------------------------------------
## long_data$Group: 2
##    0%   25%   50%   75%  100%
##    46    57    62    71    89
```

```r
# The quantile are much more different around 0% because group 1 has a min of
#9 and group 2's min starts at 46.
```