Experiment-10

Student Name: Jatin	**UID: 20BCS5951**
Branch: BE-CSE	**Section/Group:605-B**
Semester: 6th	**Date of Performance: 04/05/2023**
Subject Name: Data Mining Lab	**Subject Code: 20CSP-376**

1. **Aim:** Outlier detection using R programming.

2. **Objective:** Data points far from the dataset's other points are considered outliers. This refers to the data values dispersed among other data values and upsetting the dataset's general distribution.

   Effects of an outlier on model:

   - The format of the data appears to be skewed.
   - Modifies the mean, variance, and other statistical characteristics of the data's overall distribution.
   - Leads to the model's accuracy level being biased.

3. **Script and Output:**

   The algorithm is as follows:

   - Generates 500 normally distributed random numbers and assigned to variable **data**.
   - Adds 10 random outliers to the dataset.
   - Creates a box plot of the data variable
   - Plot shows the distribution of the data, including the outliers and it in **"Boxplot.png".**
   - Removes the outliers from the data variable.
   - Creates a box plot of the **data** variable again, but this time it shows the data after the outliers have been removed.
   - The resulting plot is saved in the file **"Boxplot1.png".**

R Script:

```
#create the data with 500 different data points using the
rnorm() function  data <- rnorm(500)

#add 10 random outliers to this data data[1:10]
<- c(46,9,15,-90,42,50,-82,74,61,-32)

# output to be present as PNG file
png(file="Boxplot.png")
# analyze the outliner in the provided data using the
boxplot boxplot(data) # saving the file dev.off()

# remove the outlier of the provided data boxplot.stats()
function in R
data <- data[!data %in% boxplot.stats(data)$out]
png(file="Boxplot1.png")
# verify if the outliner has been removed by plotting the
boxplot boxplot(data) # saving the file dev.off()
```

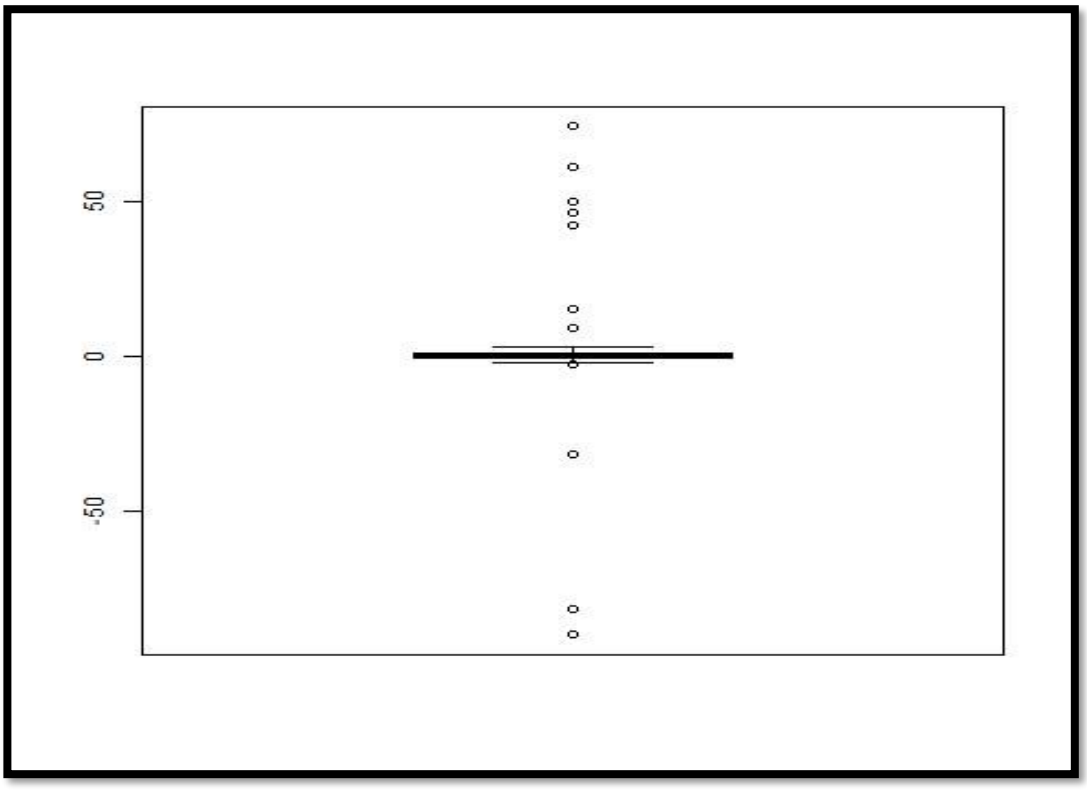
Output:

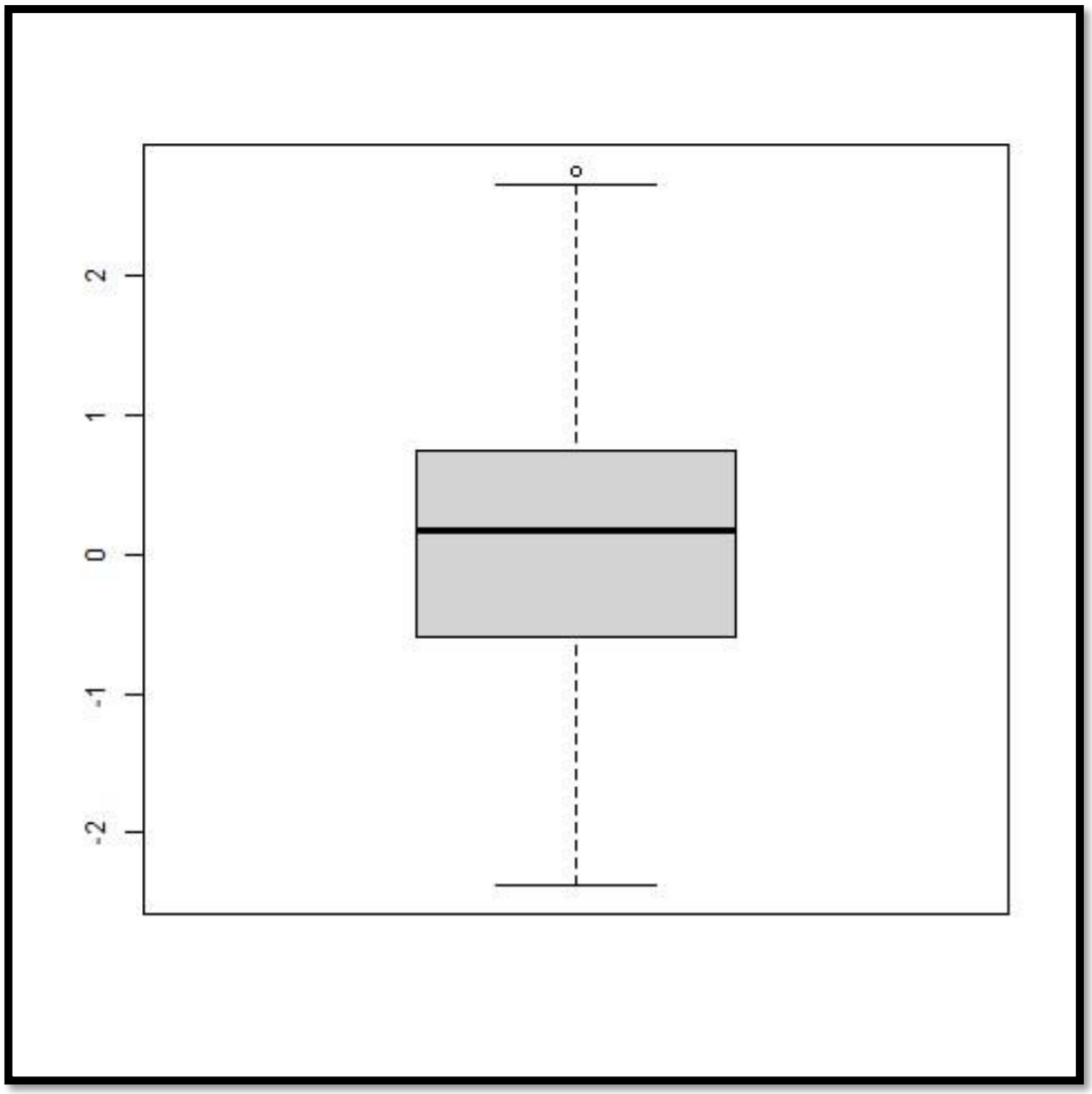
```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #create the data with 500 different data points using the rnorm() function
> data <- rnorm(500)
>
> #add 10 random outliers to this data
> data[1:10] <- c(46,9,15,-90,
+                 42,50,-82,74,61,-32)
>
> # output to be present as PNG file
> png(file="Boxplot.png")
> # analyze the outliner in the provided data using the boxplot
> boxplot(data)
> # saving the file
> dev.off()
null device
          1
> # remove the outlier of the provided data boxplot.stats() function in R
> data <- data[!data %in% boxplot.stats(data)$out]
>
> png(file="Boxplot1.png")
> # verify if the outliner has been removed by plotting the boxplot
> boxplot(data)
>
> # saving the file
> dev.off()
null device
          1
>
```




2
1
0
-1
-2