Experiment 3.3

Student Nam : Mayank Kumar	**UI** :20BCS1353
Branch: BE-CSE	**Section/Group**: 20BCS_DM_705 A
Semester: 6th	**Date of Performance**:
Subject Name: Data Mining Lab	**Subject Code**: 20CSP-376

Aim: Study of Outlier detection using R programming.

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.

Script and Output:

```
#creating the data containing 500 random values


data <- rnorm(500)

print(data)

#adding 10 random outliers to this data.

data[1:10] <- c(46,9,15,-90,42,50,-82,74,61,-32)
```

#draw boxpolot and an outlier is defined as a data point that is located outside the whiskers of the box plot.

boxplot(data)

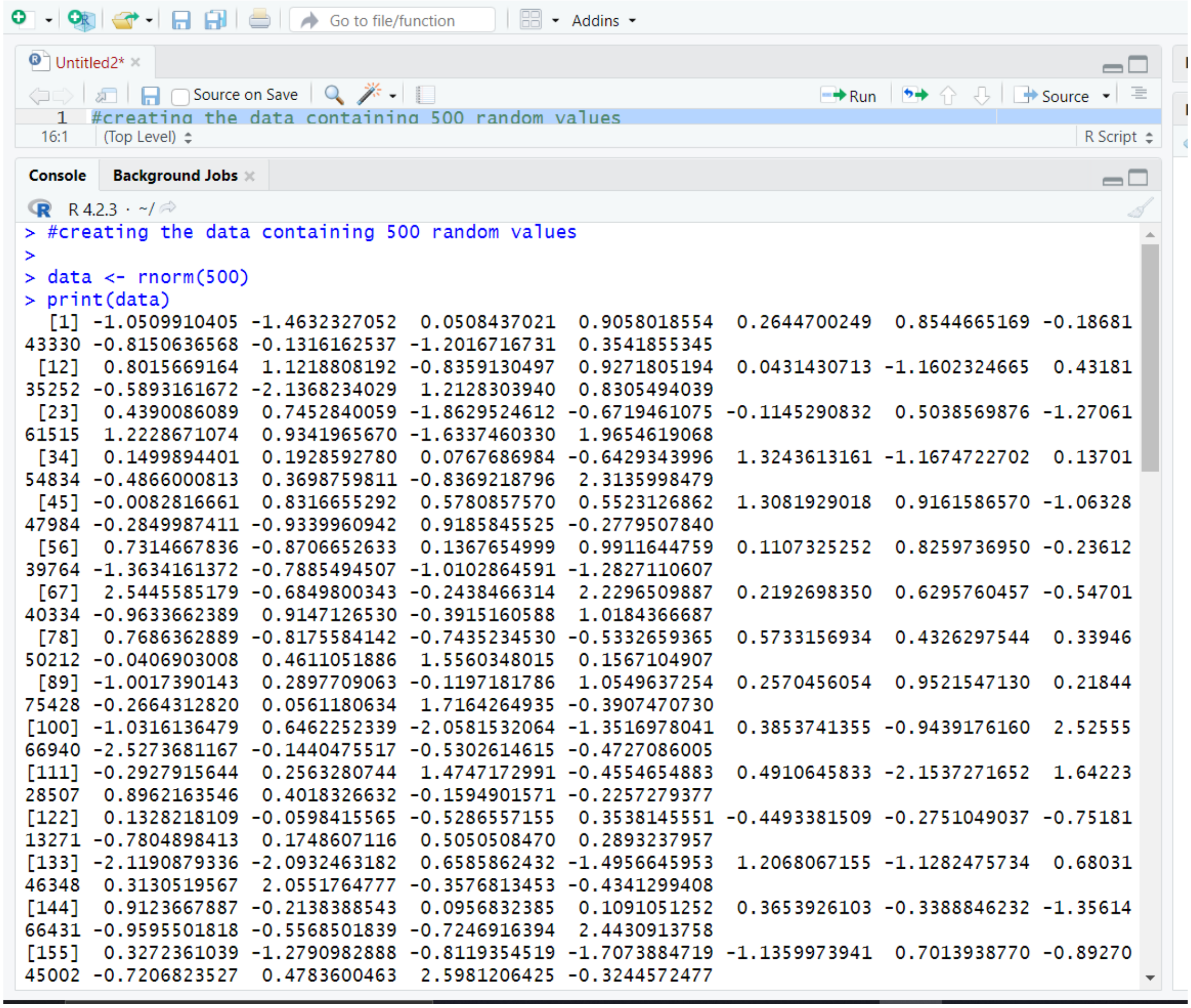
#remove the outlier of the provided data boxplot.stats() function in R

data <- data[!data %in% boxplot.stats(data)$out]

#draw boxplot to verify whether ouliers removed or not

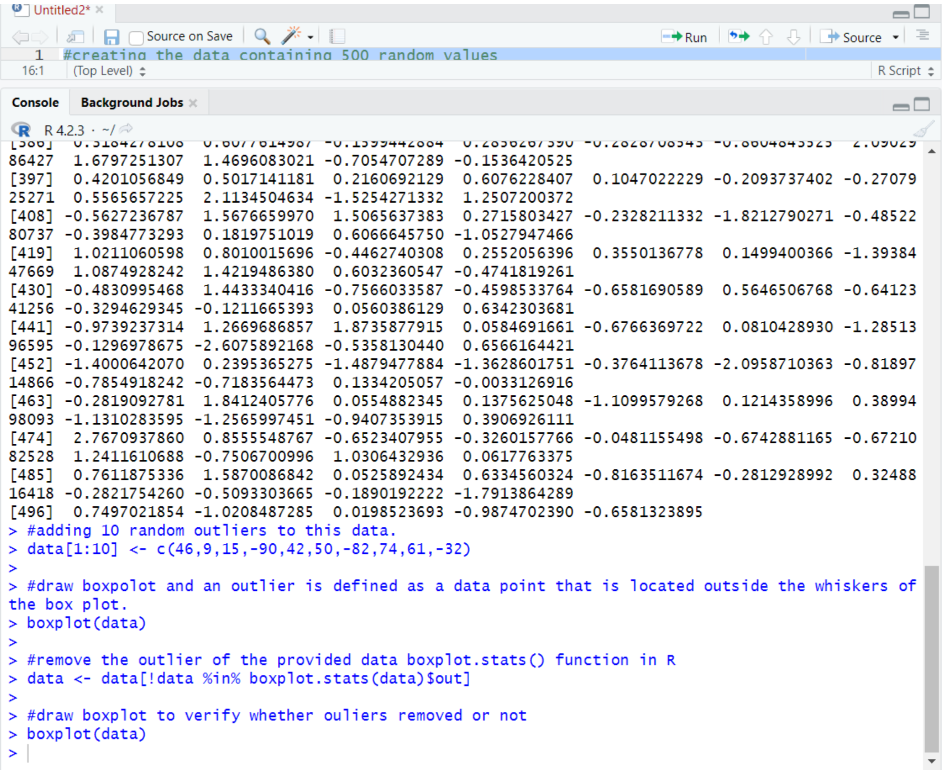
boxplot(data)

OUTPUT:



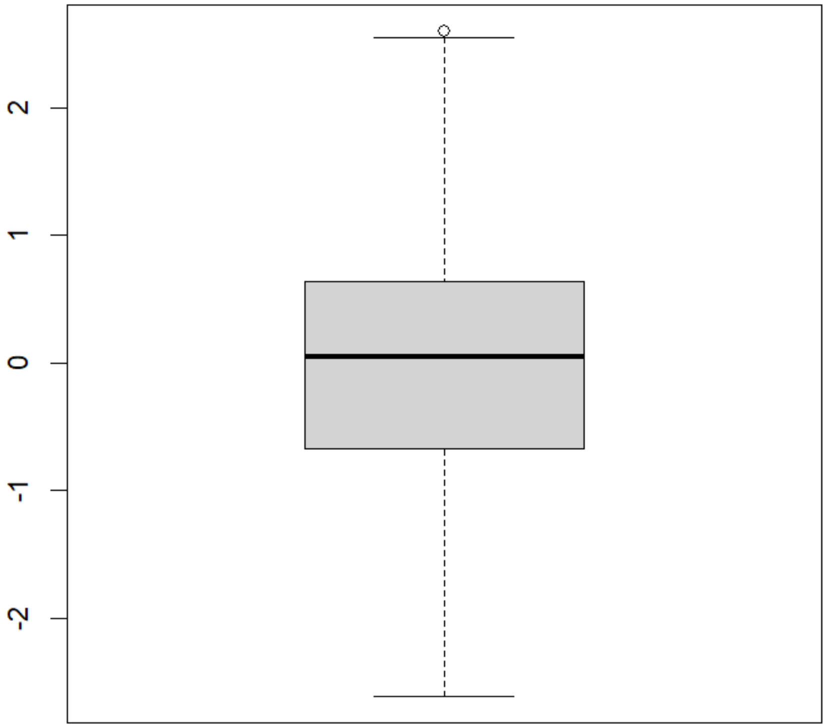
```
#creating the data containing 500 random values
```

```
> #creating the data containing 500 random values
>
> data <- rnorm(500)
> print(data)
  [1] -1.0509910405 -1.4632327052  0.0508437021  0.9058018554  0.2644700249  0.8544665169 -0.18681
43330 -0.8150636568 -0.1316162537 -1.2016716731  0.3541855345
 [12]  0.8015669164  1.1218808192 -0.8359130497  0.9271805194  0.0431430713 -1.1602324665  0.43181
35252 -0.5893161672 -2.1368234029  1.2128303940  0.8305494039
 [23]  0.4390086089  0.7452840059 -1.8629524612 -0.6719461075 -0.1145290832  0.5038569876 -1.27061
61515  1.2228671074  0.9341965670 -1.6337460330  1.9654619068
 [34]  0.1499894401  0.1928592780  0.0767686984 -0.6429343996  1.3243613161 -1.1674722702  0.13701
54834 -0.4866000813  0.3698759811 -0.8369218796  2.3135998479
 [45] -0.0082816661  0.8316655292  0.5780857570  0.5523126862  1.3081929018  0.9161586570 -1.06328
47984 -0.2849987411 -0.9339960942  0.9185845525 -0.2779507840
 [56]  0.7314667836 -0.8706652633  0.1367654999  0.9911644759  0.1107325252  0.8259736950 -0.23612
39764 -1.3634161372 -0.7885494507 -1.0102864591 -1.2827110607
 [67]  2.5445585179 -0.6849800343 -0.2438466314  2.2296509887  0.2192698350  0.6295760457 -0.54701
40334 -0.9633662389  0.9147126530 -0.3915160588  1.0184366687
 [78]  0.7686362889 -0.8175584142 -0.7435234530 -0.5332659365  0.5733156934  0.4326297544  0.33946
50212 -0.0406903008  0.4611051886  1.5560348015  0.1567104907
 [89] -1.0017390143  0.2897709063 -0.1197181786  1.0549637254  0.2570456054  0.9521547130  0.21844
75428 -0.2664312820  0.0561180634  1.7164264935 -0.3907470730
[100] -1.0316136479  0.6462252339 -2.0581532064 -1.3516978041  0.3853741355 -0.9439176160  2.52555
66940 -2.5273681167 -0.1440475517 -0.5302614615 -0.4727086005
[111] -0.2927915644  0.2563280744  1.4747172991 -0.4554654883  0.4910645833 -2.1537271652  1.64223
28507  0.8962163546  0.4018326632 -0.1594901571 -0.2257279377
[122]  0.1328218109 -0.0598415565 -0.5286557155  0.3538145551 -0.4493381509 -0.2751049037 -0.75181
13271 -0.7804898413  0.1748607116  0.5050508470  0.2893237957
[133] -2.1190879336 -2.0932463182  0.6585862432 -1.4956645953  1.2068067155 -1.1282475734  0.68031
46348  0.3130519567  2.0551764777 -0.3576813453 -0.4341299408
[144]  0.9123667887 -0.2138388543  0.0956832385  0.1091051252  0.3653926103 -0.3388846232 -1.35614
66431 -0.9595501818 -0.5568501839 -0.7246916394  2.4430913758
[155]  0.3272361039 -1.2790982888 -0.8119354519 -1.7073884719 -1.1359973941  0.7013938770 -0.89270
45002 -0.7206823527  0.4783600463  2.5981206425 -0.3244572477
```



```
1  #creating the data containing 500 random values
16:1   (Top Level)                                                    R Script
```

```
[386]  0.3184278108  0.6077614987 -0.1399442884  0.2836267390 -0.2828708345 -0.8604843325  2.09029
86427  1.6797251307  1.4696083021 -0.7054707289 -0.1536420525
[397]  0.4201056849  0.5017141181  0.2160692129  0.6076228407  0.1047022229 -0.2093737402 -0.27079
25271  0.5565657225  2.1134504634 -1.5254271332  1.2507200372
[408] -0.5627236787  1.5676659970  1.5065637383  0.2715803427 -0.2328211332 -1.8212790271 -0.48522
80737 -0.3984773293  0.1819751019  0.6066645750 -1.0527947466
[419]  1.0211060598  0.8010015696 -0.4462740308  0.2552056396  0.3550136778  0.1499400366 -1.39384
47669  1.0874928242  1.4219486380  0.6032360547 -0.4741819261
[430] -0.4830995468  1.4433340416 -0.7566033587 -0.4598533764 -0.6581690589  0.5646506768 -0.64123
41256 -0.3294629345 -0.1211665393  0.0560386129  0.6342303681
[441] -0.9739237314  1.2669686857  1.8735877915  0.0584691661 -0.6766369722  0.0810428930 -1.28513
96595 -0.1296978675 -2.6075892168 -0.5358130440  0.6566164421
[452] -1.4000642070  0.2395365275 -1.4879477884 -1.3628601751 -0.3764113678 -2.0958710363 -0.81897
14866 -0.7854918242 -0.7183564473  0.1334205057 -0.0033126916
[463] -0.2819092781  1.8412405776  0.0554882345  0.1375625048 -1.1099579268  0.1214358996  0.38994
98093 -1.1310283595 -1.2565997451 -0.9407353915  0.3906926111
[474]  2.7670937860  0.8555548767 -0.6523407955 -0.3260157766 -0.0481155498 -0.6742881165 -0.67210
82528  1.2411610688 -0.7506700996  1.0306432936  0.0617763375
[485]  0.7611875336  1.5870086842  0.0525892434  0.6334560324 -0.8163511674 -0.2812928992  0.32488
16418 -0.2821754260 -0.5093303665 -0.1890192222 -1.7913864289
[496]  0.7497021854 -1.0208487285  0.0198523693 -0.9874702390 -0.6581323895
> #adding 10 random outliers to this data.
> data[1:10] <- c(46,9,15,-90,42,50,-82,74,61,-32)
> 
> #draw boxpolot and an outlier is defined as a data point that is located outside the whiskers of
the box plot.
> boxplot(data)
> 
> #remove the outlier of the provided data boxplot.stats() function in R
> data <- data[!data %in% boxplot.stats(data)$out]
> 
> #draw boxplot to verify whether ouliers removed or not
> boxplot(data)
> 
```



Learning Outcome:

1. Learnt about Regression Analysis using R Programming.
2. Learnt about Simple Linear Regression.
3. Learnt about Multiple Linear Regression.