Experiment 3.2

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

2. Objective:

- To Identify and flag data points that deviate significantly from the norm.
- To Assess the impact of outliers on statistical analysis or modeling.

3. Script:

What are outliers?

- ➢ 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.

Dataset:

The **day.csv** dataset provides information related to bike sharing and various environmental and temporal factors.

It contains the following variables:

1. instant: A unique identifier for each record.
2. dteday: The date of the record in yyyy-mm-dd format.
3. season: The season (1: spring, 2: summer, 3: fall, 4: winter).
4. yr: The year (0: 2011, 1: 2012).
5. mnth: The month (1 to 12).
6. holiday: A binary indicator of whether it is a holiday or not (1: holiday, 0: non-holiday).
7. weekday: The day of the week (0: Sunday, 1: Monday, ..., 6: Saturday).
8. workingday: A binary indicator of whether it is a working day or not (1: working day, 0: non-working day).
9. weathersit: The weather situation (1: clear, 2: mist/cloudy, 3: light rain/snow, 4: heavy rain/snow).
10. temp: The normalized temperature in Celsius.
11. atemp: The normalized feeling temperature in Celsius.
12. hum: The normalized humidity.
13. windspeed: The normalized wind speed.
14. casual: The number of casual (non-registered) bike users.
15. registered: The number of registered bike users.
16. cnt: The total count of bike rentals (casual + registered).

4. Code:

```
#Removed all the existing objects
rm(list = ls())

#Setting the working directory
setwd("D:/Documents")
getwd()

#Load the dataset
bike_data = read.csv("day.csv",header=TRUE)

# Missing Value Analysis
sum(is.na(bike_data))
summary(is.na(bike_data))
```

Output:

Prior to outlier detection, we have performed missing value analysis just to check for the presence of any NULL or missing values. For the same, we have made use of sum(is.na(data)) function.

```
> sum(is.na(bike_data))
[1] 0
> summary(is.na(bike_data))
  instant          dteday           season             yr              mnth        
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical  
 FALSE:731       FALSE:731       FALSE:731       FALSE:731       FALSE:731      
  holiday         weekday        workingday      weathersit         temp        
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical  
 FALSE:731       FALSE:731       FALSE:731       FALSE:731       FALSE:731      
   atemp            hum          windspeed         casual        registered     
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical  
 FALSE:731       FALSE:731       FALSE:731       FALSE:731       FALSE:731      
    cnt         
 Mode :logical  
 FALSE:731      
```

#From the above result, the dataset contains NO Missing Values.
#Outlier Analysis -- DETECTION

1. Outliers in the data values exists only in continuous/numeric form of data variables. Thus, we need to store all the numeric and categorical independent variables into a separate array structure.

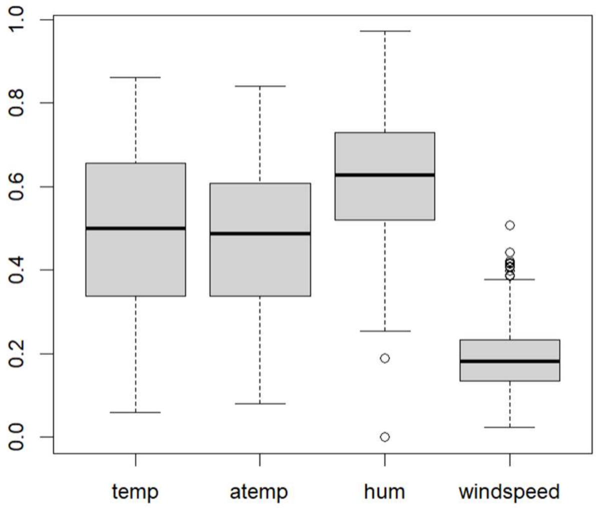
```
col = c('temp','cnt','hum','windspeed')
categorical_col = c("season","yr","mnth","holiday","weekday","workingday","weathersit")
```

2. Using BoxPlot to detect the presence of outliers in the numeric/continuous data columns.

```
boxplot(bike_data[,c('temp','atemp','hum','windspeed')])
```

Output:

From the above visualization, the data variables 'hum' and 'windspeed' contains outliers in the data values.



#Now, we will replace the outlier data values with NULL.

```
for (x in c('hum','windspeed'))
{
  value = bike_data[,x][bike_data[,x] %in% boxplot.stats(bike_data[,x])$out]
  bike_data[,x][bike_data[,x] %in% value] = NA
}
```

```
#Checking whether the outliers in the above defined columns are replaced by NULL or not
sum(is.na(bike_data$hum))
sum(is.na(bike_data$windspeed))
as.data.frame(colSums(is.na(bike_data)))
```

Output:

As a result, we have converted the 2 outlier points from the ‘hum’ column and 16 outlier points from the ‘windspeed’ column into missing(NA) values.

```
> sum(is.na(bike_data$hum))
[1] 2
> sum(is.na(bike_data$windspeed))
[1] 16
> as.data.frame(colSums(is.na(bike_data)))
           colSums(is.na(bike_data))
instant                            0
dteday                             0
season                             0
yr                                 0
mnth                               0
holiday                            0
weekday                            0
workingday                         0
weathersit                         0
temp                               0
atemp                              0
hum                                2
windspeed                         16
casual                             0
registered                         0
cnt                                0
>
```

#Removing the null values
library(tidyr)
bike_data = drop_na(bike_data)
as.data.frame(colSums(is.na(bike_data)))

Output:

At last, we treat the missing values by dropping the NULL values using drop_na() function from the 'tidyr' library.

```
> #Removing the null values
> library(tidyr)
> bike_data = drop_na(bike_data)
> as.data.frame(colSums(is.na(bike_data)))
           colSums(is.na(bike_data))
instant                            0
dteday                             0
season                             0
yr                                 0
mnth                               0
holiday                            0
weekday                            0
workingday                         0
weathersit                         0
temp                               0
atemp                              0
hum                                0
windspeed                          0
casual                             0
registered                         0
cnt                                0
```

#draw boxplot to verify whether ouliers removed or not
boxplot(bike_data[,c('temp','atemp','hum','windspeed')])

Output: