Sri Lanka Institute of Information Technology


# Lab Submission
# 04

**IT24102798**

**Sooriyabandara U.R.G.W.K.**

**Probability and Statistics | IT2120**

B.Sc. (Hons) in Information Technology

## Exercise

1. Import the dataset (’Exercise.txt’) into R and store it in a data frame called “branch_data”.

```
5
6  # Question 01
7
8  branch_data <- read.table("Exercise.txt", header = TRUE, sep = ",")
9  attach(branch_data)
10
```

```
> # Question 01
>
> branch_data <- read.table("Exercise.txt", header = TRUE, sep = ",")
> attach(branch_data)
The following objects are masked from branch_data (pos = 3):

    Advertising_X2, Branch, Sales_X1, Years_X3

The following objects are masked from branch_data (pos = 4):

    Advertising_X2, Branch, Sales_X1, Years_X3

>
```

2. Identify the variable type and scale of measurement for each variable.

```
11
12  # Question 02
13
14  typeof(branch_data)
15  typeof(Branch)
16  typeof(Sales_X1)
17  typeof(Advertising_X2)
18  typeof(Years_X3)
19
```

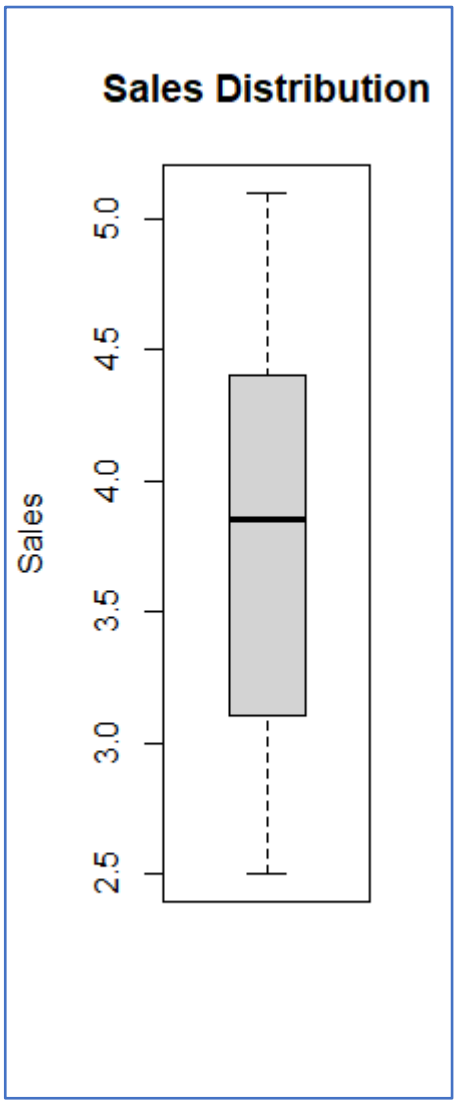
```
> # Question 02
>
> typeof(branch_data)
[1] "list"
> typeof(Branch)
[1] "integer"
> typeof(Sales_X1)
[1] "double"
> typeof(Advertising_X2)
[1] "integer"
> typeof(Years_X3)
[1] "integer"
>
```

| **Variable Name** | **Variable Type** | **Scale Of Measurement** |
|---|---|---|
| Branch | Integer | Nominal |
| Sales_X1 | Double | Ratio |
| Advertising_X2 | Integer | Ratio |
| Years_X3 | Integer | Ratio |

3. Obtain boxplot for sales and interpret the shape of the sales distribution.

```
20
21  # Question 03
22
23  boxplot(Sales_X1, main = "Sales Distribution", ylab = "Sales")
24
25
```

```
> # Question 03
>
> boxplot(Sales_X1, main = "Sales Distribution", ylab = "Sales")
>
```



4. Calculate the five number summary and IQR for advertising variable.

```
25
26  # Question 04
27
28  summary(Advertising_X2)
29  IQR(Advertising_X2)
30
```

```
> # Question 04
>
> summary(Advertising_X2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   80.0   101.2   132.5   134.8   158.8   210.0
> IQR(Advertising_X2)
[1] 57.5
>
```

5. Write an R function to find the outliers in a numeric vector and check for outliers in years variables.

```
31
32  # Question 05
33
34  find.outliers = function(numeric_vector) {
35
36    q1 <- quantile(numeric_vector)[2]
37    q3 <- quantile(numeric_vector)[4]
38    iqr <- q3 - q1
39
40    ub <- q3 + (1.5 * iqr)
41    lb <- q1 + (1.5 * iqr)
42
43    print(paste("Upper Bound =", ub))
44    print(paste("Lower Bound =", lb))
45    print(paste("Outliers:", paste( sort( numeric_vector[ (numeric_vector < lb) | (numeric_vector > ub) ]), collapse = ",")))
46
47  }
48
49  find.outliers(Years_X3)
50
```

```
> # Question 05
>
> find.outliers = function(numeric_vector) {
+
+   q1 <- quantile(numeric_vector)[2]
+   q3 <- quantile(numeric_vector)[4]
+   iqr <- q3 - q1
+
+   ub <- q3 + (1.5 * iqr)
+   lb <- q1 + (1.5 * iqr)
+
+   print(paste("Upper Bound =", ub))
+   print(paste("Lower Bound =", lb))
+   print(paste("Outliers:", paste( sort( numeric_vector[ (numeric_vector < lb) | (numeric_vector > ub) ]), collapse = ",")))
+
+ }
>
> find.outliers(Years_X3)
[1] "Upper Bound = 14.5"
[1] "Lower Bound = 10"
[1] "Outliers: 1,1,2,2,2,3,3,3,4,4,4,5,5,5,5,6,6,6,6,7,7,7,8,8,9,9"
>
```