Assignment

Name : **Mohammad Wasiq**
E-mail : mohammadwasiq0786@gmail.com

Table of Contents

1 First Assignment

Exercise 1 : Company manufacturing valves are facing the problem of high torque variation in their actuator subassembly. The company wants to identify the variables impacting the torque variation in actuator subassembly. Through brainstorming the project team have identified seven variables suspected to influence the torque variation.

a. The company has two manufacturing plants located at Bangalore and Pune. The data collected from Bangalore plant is given in *Actuator_1* file and that from Pune plant is given in *Actuator_II* file. Join these files into a single dataset ?

```
library(readxl)
df1 <- read_excel("F:/ISI R Course/Assignments/Actuator_I.xlsx")

dim(df1)

## [1] 30  9

df2 <- read_excel("F:/ISI R Course/Assignments/Actuator_II.xlsx")

dim(df2)

## [1] 26  9

# Joining both the files
df <- rbind(df1 , df2)

dim(df)

## [1] 56  9

# Head of Joined Data
head(df)
```

```
## # A tibble: 6 × 9
##   SL_No Bearing_to_bearin… Driver_Shaft_le… Worm_length Star_Washer_pos…
Thickness Height  Load
##   <dbl>              <dbl>            <dbl>       <dbl> <chr>
<dbl>  <dbl> <dbl>
## 1     1               148.             128.        68.0 Bearing End
1.47   2.17  2943
## 2     2               148.             128.        68.0 Bearing End
1.50   2.19  3090
## 3     3               148.             128.        68.0 Worm End
1.52   2.2   3276
## 4     4               148.             128.        68.0 Bearing End
1.48   2.13  2992
## 5     5               148.             128.        68.0 Worm End
1.48   2.14  3041
## 6     6               148.             128.        68.0 Bearing End
1.50   2.22  3286
## # … with 1 more variable: Torque_Variation <chr>
```

b. Provide the descriptive summary of all the variables. For categorical variables provide summary using frequency table. Also provide the graphical summary of each variable using appropriate graphs like boxplot, histogram, pie chart, bar plot, etc?

```
library(psych)
round(describe(df[, -c(1 , 5 , 9)]) , 3)

##                             vars  n    mean     sd  median trimmed    mad
min     max  range
## Bearing_to_bearing_Length    1 56  147.94   0.21  147.95  147.94   0.31
147.70  148.20   0.50
## Driver_Shaft_length          2 56  127.94   0.01  127.94  127.94   0.01
127.93  127.95   0.02
## Worm_length                  3 56   68.01   0.00   68.01   68.01   0.01
68.00   68.02   0.02
## Thickness                    4 56    1.50   0.02    1.50    1.50   0.02
1.46    1.54   0.08
## Height                       5 56    2.18   0.04    2.19    2.18   0.04
2.08    2.28   0.20
## Load                         6 56 3160.32 135.30 3158.00 3164.35 159.38
2825.00 3433.00 608.00
##                            skew kurtosis    se
## Bearing_to_bearing_Length  0.00    -1.96  0.03
## Driver_Shaft_length        0.04    -1.18  0.00
## Worm_length               -0.02    -1.14  0.00
## Thickness                  0.16     0.05  0.00
## Height                    -0.33    -0.09  0.01
## Load                      -0.27    -0.52 18.08
```

1. **Bearing_to_bearing_Length**

- There are 56 observations .

- Minimum , Maximum and Range are **147.7** , **148.19** and **0.496** respectively .
- Mean and Trimmed Mean are **147.94** and **147.94** respectively .
- Median is **147.95 .**
- Standard Deviation and standard error are **0.208** and **0.028** respectively.
- Skewness and Kurtosis are **-0.003** and **-1.96** respectively .

2. **Driver_Shaft_Length**

- There are 56 observations .
- Minimum , Maximum and Range are **127.93** , **127.95** and **0.019** respectively .
- Mean and Trimmed Mean are **127.94** and **127.94** respectively .
- Median is **127.94 .**
- Standard Deviation and standard error are **0.006** and **0.001** respectively.
- Skewness and Kurtosis are **0.045** and **-1.18** respectively .

3. **Worm_Length**

- There are 56 observations .
- Minimum , Maximum and Range are **68** , **68.02** and **0.019** respectively .
- Mean and Trimmed Mean are **68.01** and **68.01** respectively .
- Median is **68.01** .
- Standard Deviation and standard error are **0.005** and **0.001** respectively.
- Skewness and Kurtosis are **-0.02** and **-1.14** respectively .

4. **Thickness**

- There are 56 observations .
- Minimum , Maximum and Range are **1.46** , **1.54** and **0.084** respectively .
- Mean and Trimmed Mean are **1.489** and **1.489** respectively .
- Median is **1.499** .
- Standard Deviation and standard error are **0.018** and **0.002** respectively.
- Skewness and Kurtosis are **0.163** and **0.048** respectively .

5. **Height**

- There are 56 observations .

- Minimum , Maximum and Range are **2.08** , **2.28** and **0.2** respectively .
- Mean and Trimmed Mean are **2.183** and **2.185** respectively .
- Median is **2.19** .
- Standard Deviation and standard error are **0.045** and **0.006** respectively.
- Skewness and Kurtosis are **-0.327** and **-0.089** respectively .

6. **Load**

- There are 56 observations .
- Minimum , Maximum and Range are **2825** , **3433** and **608** respectively .
- Mean and Trimmed Mean are **3160.32** and **3164.34** respectively .
- Median is **3158** .
- Standard Deviation and standard error are **135.3** and **18.081** respectively.
- Skewness and Kurtosis are **-0.271** and **-0.519** respectively .

Frequency Tables for Categorical Data

```
swpt <- cbind(table(df$Star_Washer_position))

swpt

##              [,1]
## Bearing End   32
## Worm End      24
```

Star_Washer_position

- There are **32** observations which are Bearing End.
- There are **24** observations which are Worm End.

```
tt <- cbind(table(df$Torque_Variation))

tt

##      [,1]
## High   28
## Low    28
```

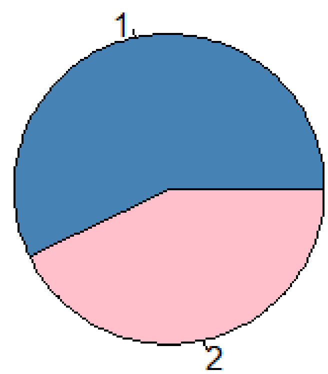
Torque_Variation

- There are **28** observations which are Hihg.
- There are **28** observations which are Low.

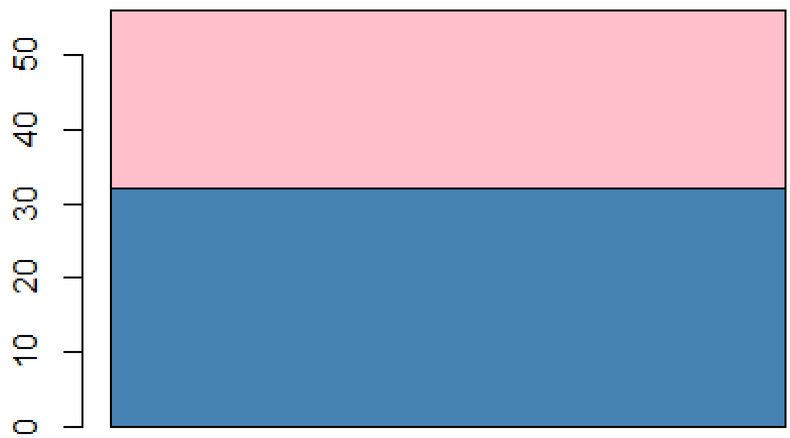
Graphics

Graphics for Categorical Data

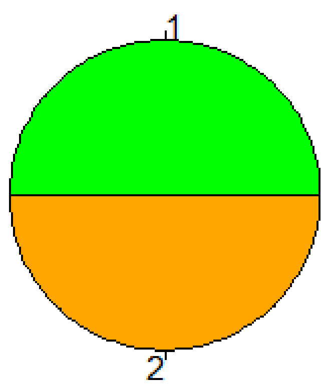
```
# Pie Chart for Star_Washer_position variable
pie(swpt , col = c("steelblue" , "pink"))
```



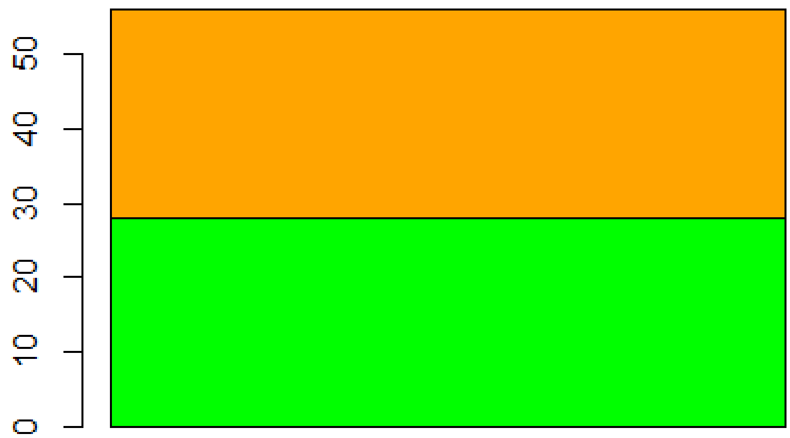

```
# Barplot for Star_Washer_position variable
barplot(swpt , col = c("steelblue" , "pink"))
```




```r
# Pie Chart for Torque Variation variable
pie(tt , col = c("green" , "orange"))
```

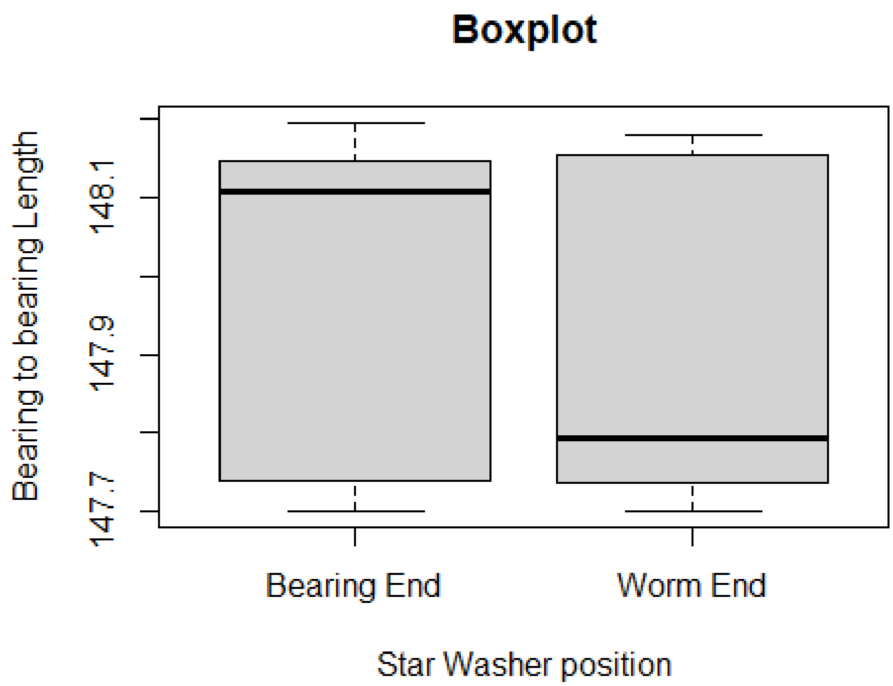


```
# Barplot for Torque Variation variable
barplot(tt , col = c("green" , "orange"))
```

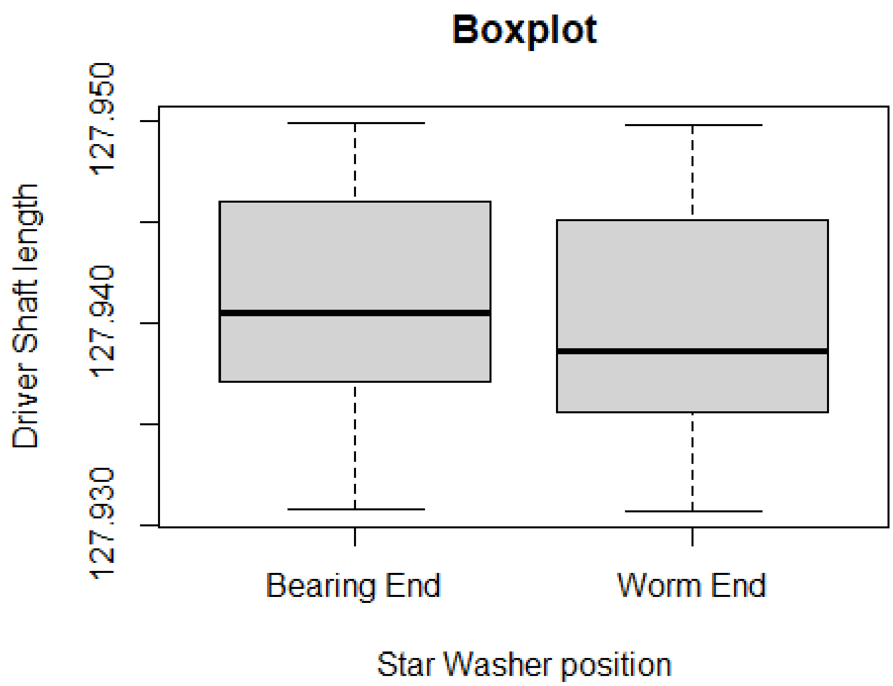


Graphics for Quantitative Data

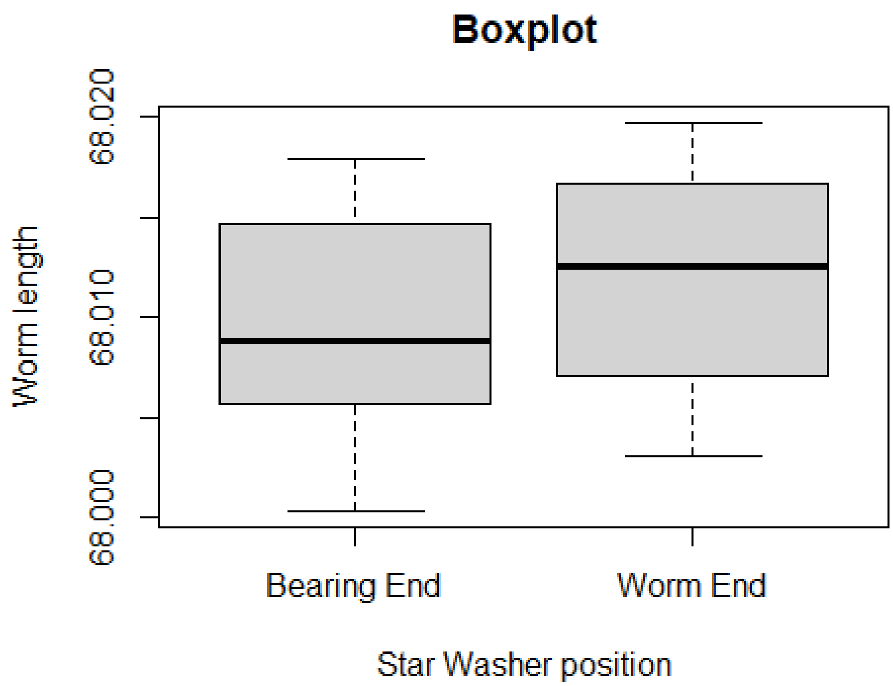
```
bl <- df$Bearing_to_bearing_Length
sl <- df$Driver_Shaft_length
wl <- df$Worm_length
wp <- df$Star_Washer_position
t <- df$Thickness
h <- df$Height
l <- df$Load
tv <- df$Torque_Variation

# Correspond to Star Washer position
# Boxplot wp and bl
boxplot(bl ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Bearing to bearing Length")
```



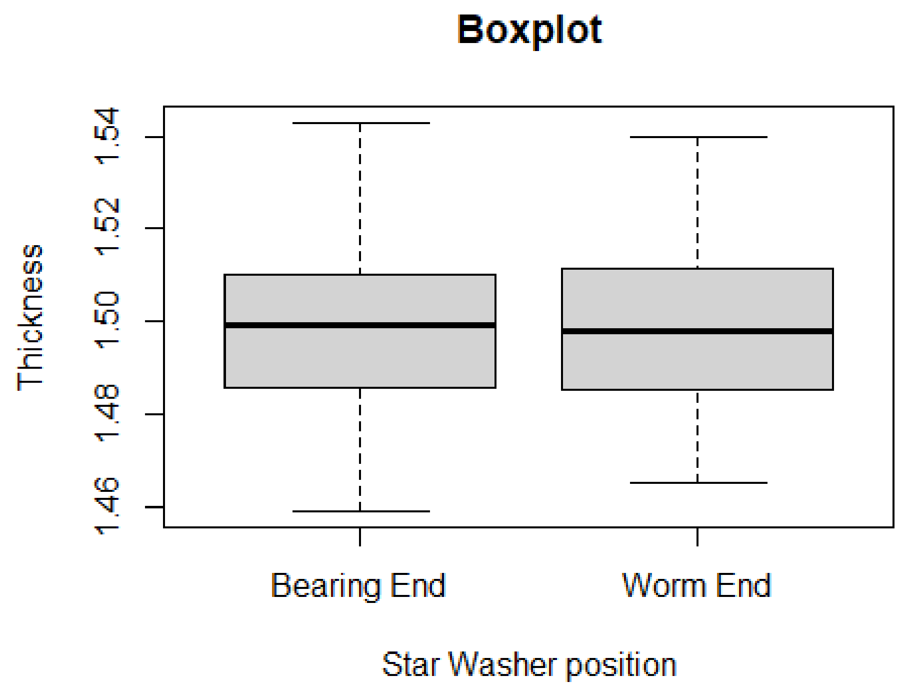

```r
# Boxplot wp and sl
boxplot(sl ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Driver Shaft length")
```



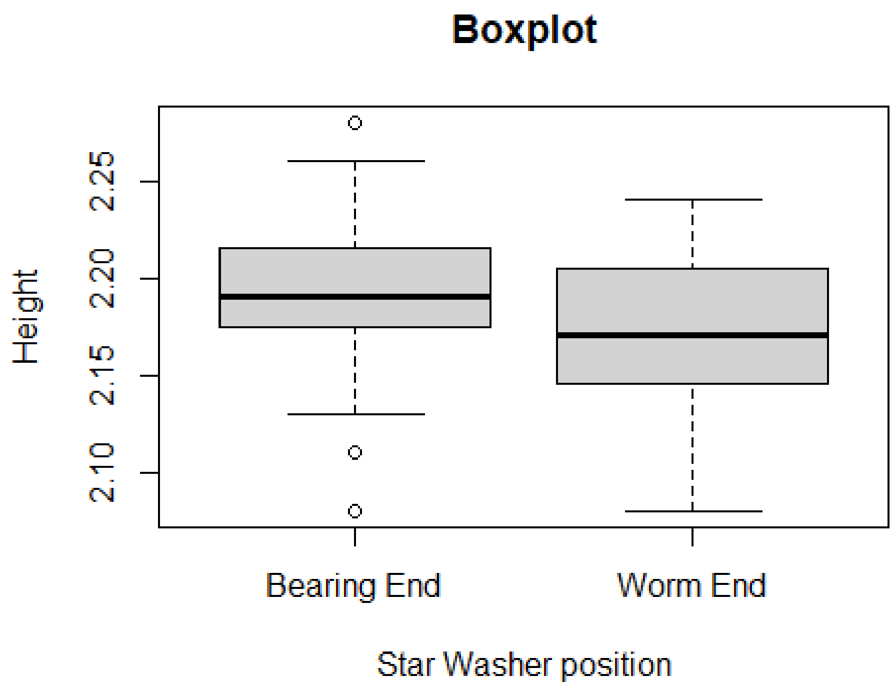

```r
# Boxplot wp and wl
boxplot(wl ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Worm length")
```



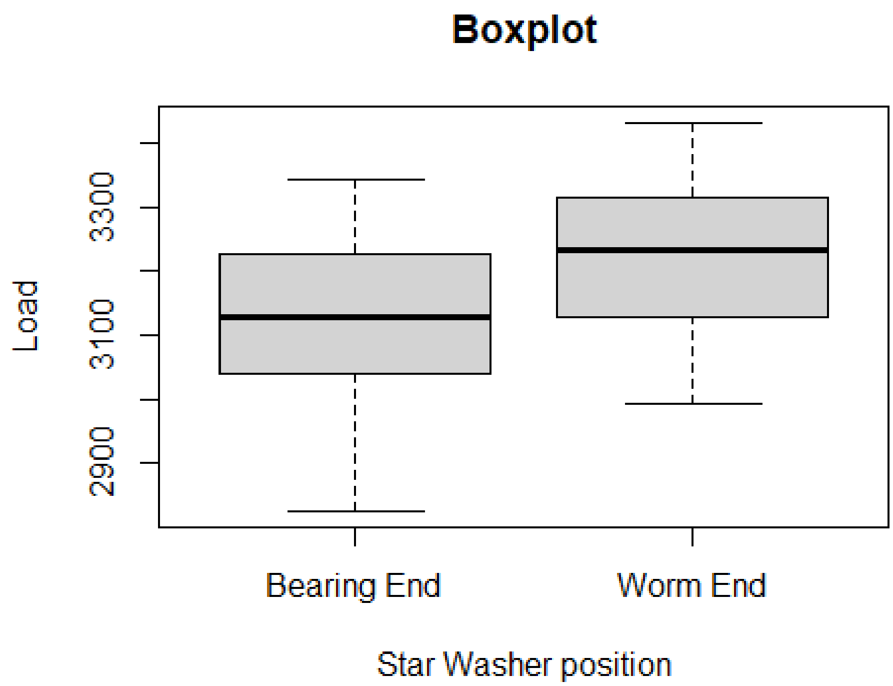

```r
# Boxplot wp and t
boxplot(t ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Thickness")
```



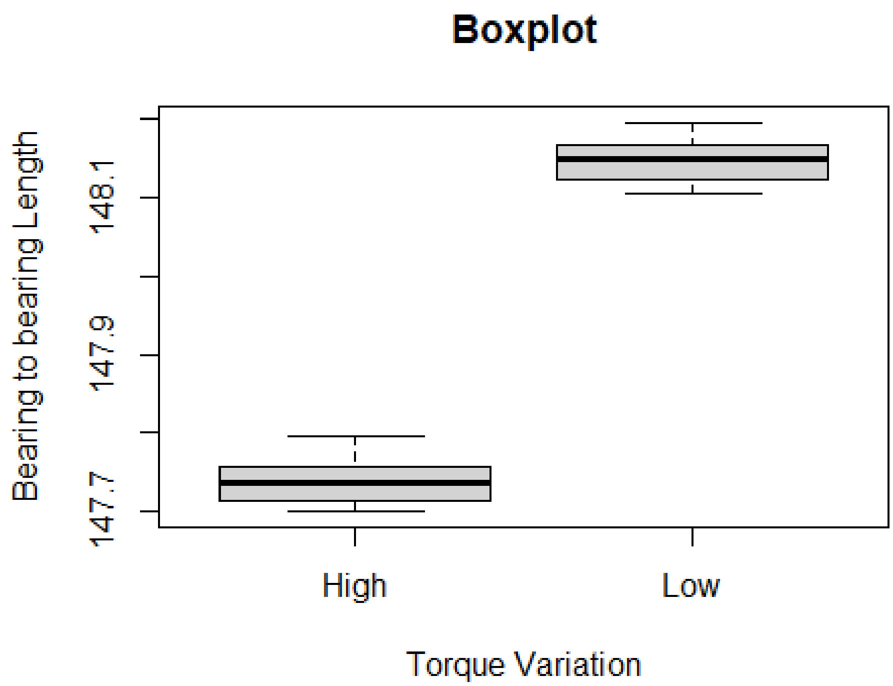

```
# Boxplot wp and h
boxplot(h ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Height")
```



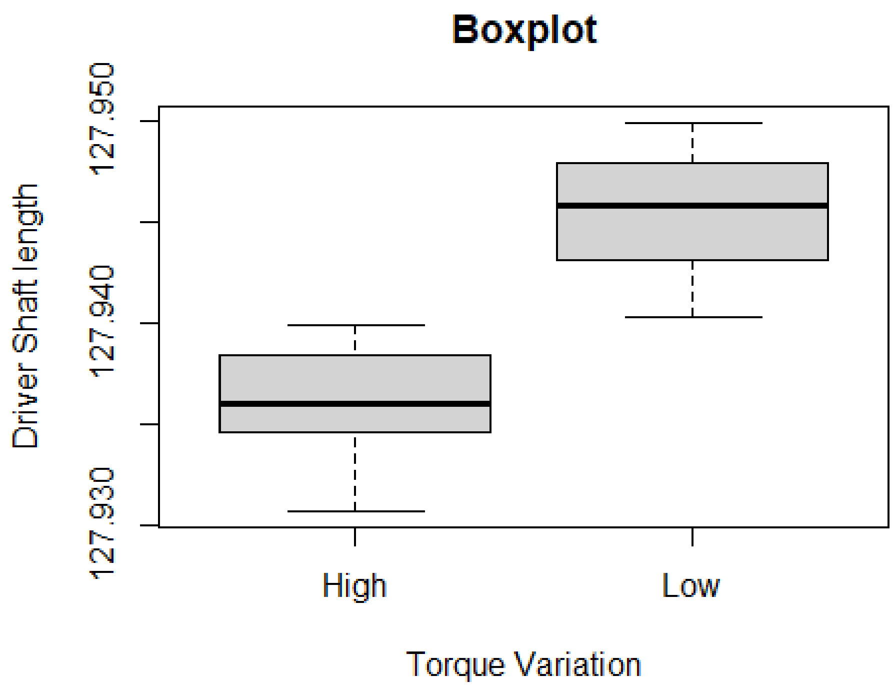

```
# Boxplot wp and l
boxplot(l ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Load")
```



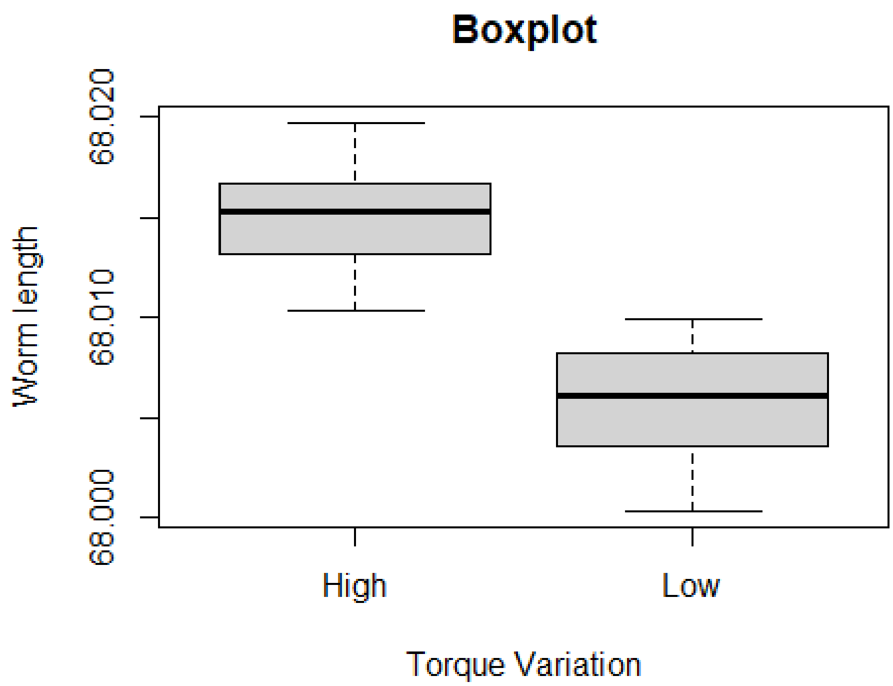

```r
# Correspond to Torque Variation
# Boxplot tv and bl
boxplot(bl ~ tv ,
        main = "Boxplot" ,
        xlab = "Torque Variation" ,
        ylab = "Bearing to bearing Length")
```



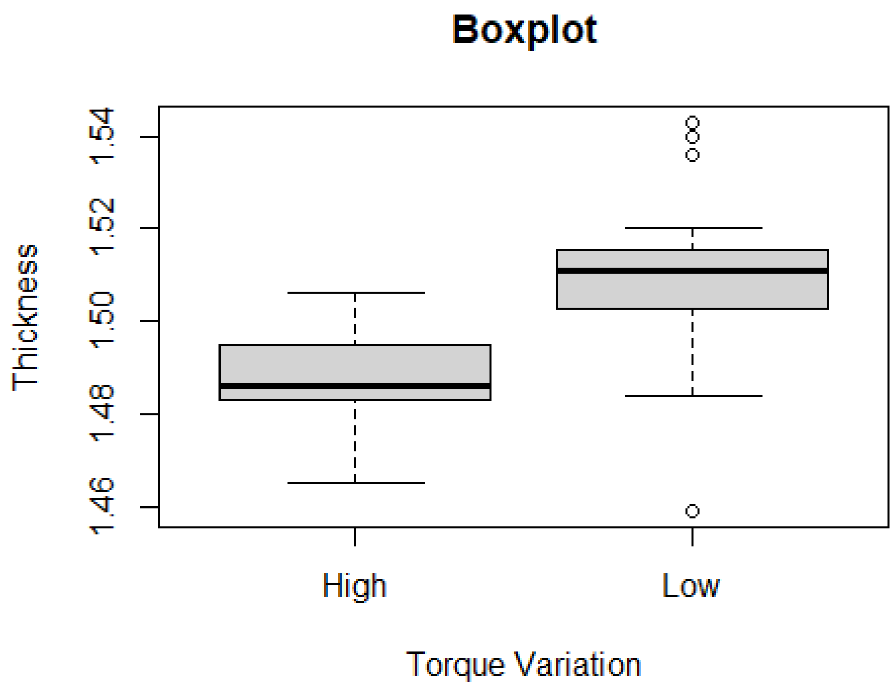

```
# Boxplot tv and sl
boxplot(sl ~ tv ,
        main = "Boxplot" ,
        xlab = "Torque Variation" ,
        ylab = "Driver Shaft length")
```



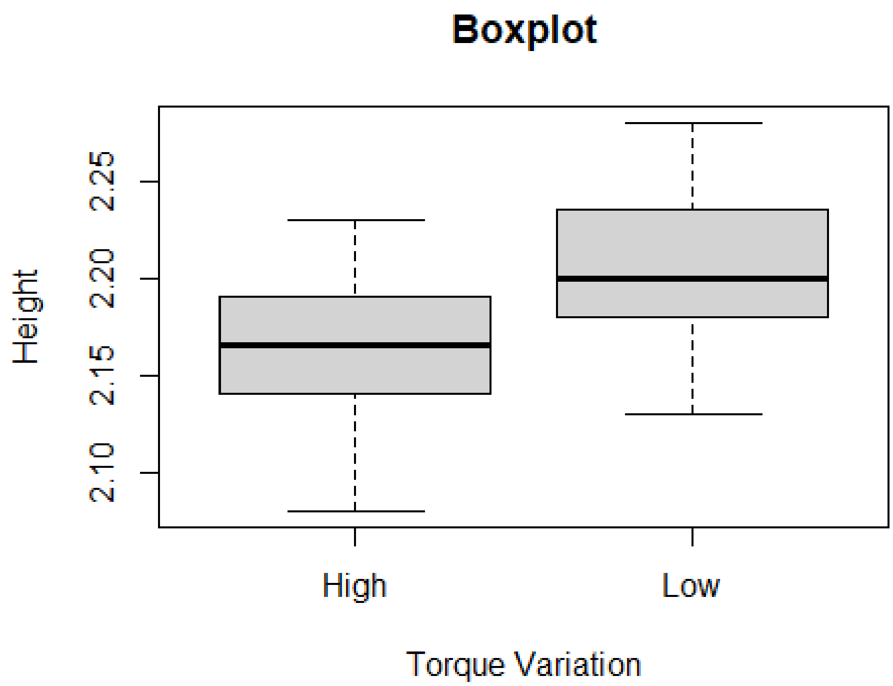

```
# Boxplot tv and wl
boxplot(wl ~ tv ,
        main = "Boxplot" ,
        xlab = "Torque Variation" ,
        ylab = "Worm length")
```



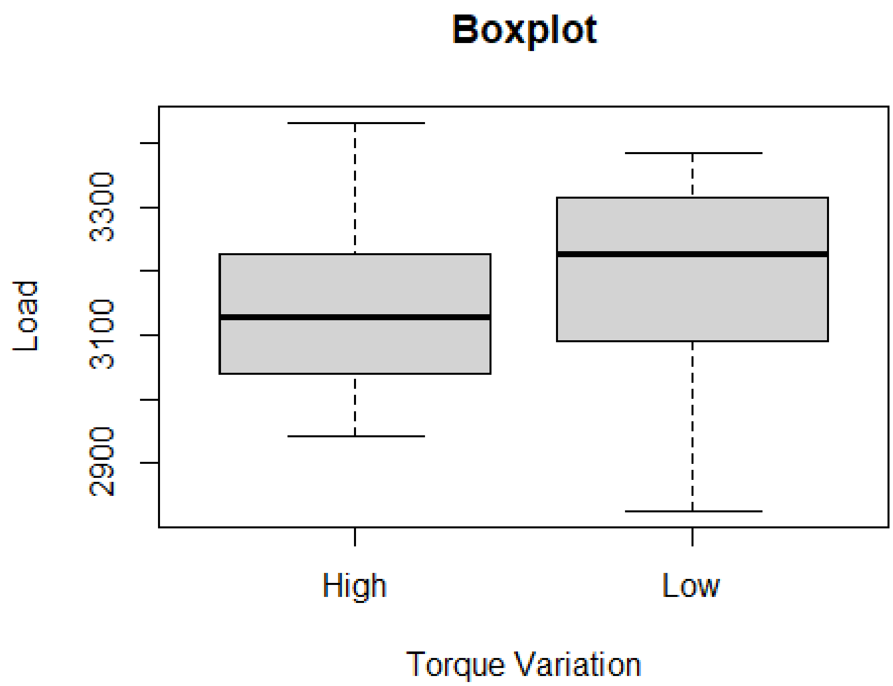

```r
# Boxplot tv and t
boxplot(t ~ tv ,
        main = "Boxplot" ,
        xlab = "Torque Variation" ,
        ylab = "Thickness")
```




```
# Boxplot tv and h
boxplot(h ~ tv ,
        main = "Boxplot" ,
        xlab = "Torque Variation" ,
        ylab = "Height")
```

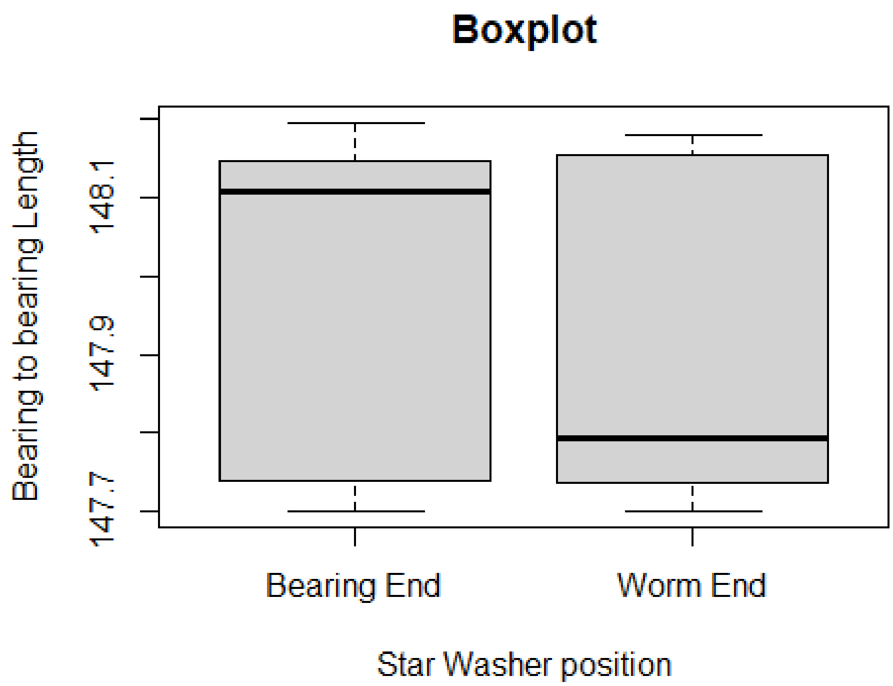



```
# Boxplot tv and l
boxplot(l ~ tv ,
        main = "Boxplot" ,
        xlab = "Torque Variation" ,
        ylab = "Load")
```



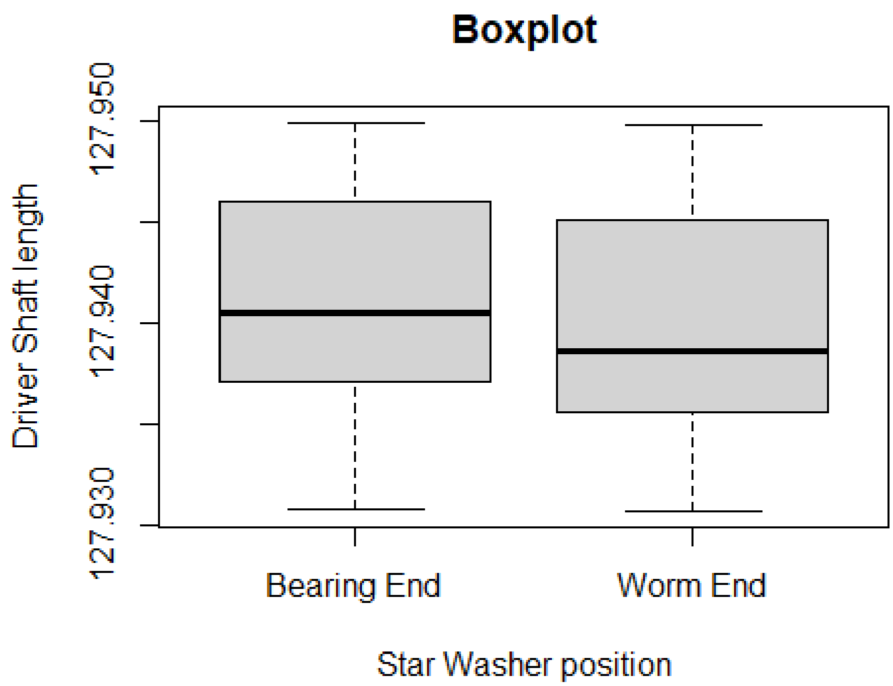
c. Identify the relationship of explanatory variables with torque variation using appropriate data visualization technique ? Give your interpretation

```
aggregate(bl ~ wp , FUN = mean)

##              wp       bl
## 1 Bearing End 147.9675
## 2    Worm End 147.9134

# Boxplot wp and bl
boxplot(bl ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Bearing to bearing Length")
```



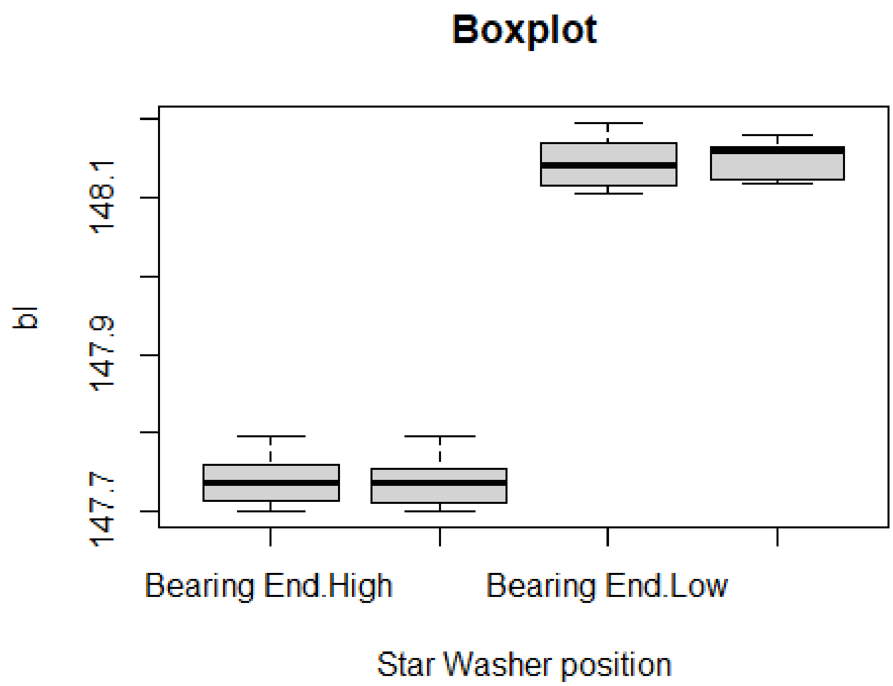
- Mean of Bearing to bearing Length where the Star Washer position is Bearing End is **147.967**
- Mean of Bearing to bearing Length where the Star Washer position is Worm End is **147.913**

```
# Boxplot wp and sl
aggregate(sl ~ wp , FUN = mean)

##            wp       sl
## 1 Bearing End 127.9414
## 2    Worm End 127.9400

boxplot(sl ~ wp ,
        main = "Boxplot" ,
        xlab = "Star Washer position" ,
        ylab = "Driver Shaft length")
```



- Mean of Driver Shaft length where the Star Washer position is Bearing End is **127.94**
- Mean of Driver Shaft length where the Star Washer position is Worm End is **127.94**

```
# Boxplot wp, tv and bl
aggregate(bl ~ wp + tv , FUN = mean)

##            wp   tv       bl
## 1 Bearing End High 147.7388
## 2    Worm End High 147.7425
## 3 Bearing End  Low 148.1455
## 4    Worm End  Low 148.1526

boxplot(bl ~ wp + tv,
        main = "Boxplot" ,
        xlab = "Star Washer position")
```



- Mean of Driver Shaft length where the Star Washer position is Bearing End and Torque Variation is High is **147.74**
- Mean of Driver Shaft length where the Star Washer position is Worm End and Torque Variation is High is **147.74**
- Mean of Driver Shaft length where the Star Washer position is Bearing End and Torque Variation is Low is **148.145**
- Mean of Driver Shaft length where the Star Washer position is Worm End and Torque Variation is Low is **148.15**

1.1 Second Assignment

Example 2 An application support process like to develop a model to estimate the time taken to resolve the tickets they receive. Through discussions, the project team has identified four explanatory variables suspected to be impacting the resolution time

a. The data is collected on the resolution time and explanatory variables. The explanatory variable data is collected from the company server and is given in Resolution_TimeI file and the resolution time data is collected from client-server and is given in Resolution_TimeII file. Kindly merge the two files.

```r
# Import require package
library(readxl)

# Read the 1st data
df1 <- read_excel("Assignments/Resolution_TimeI.xlsx")

# Dimension of 1st Data
dim(df1)

## [1] 57  4

# Head of data
head(df1)

## # A tibble: 6 × 4
##   SL_No Volume Team_Experience Domain_Expertise
##   <dbl>  <dbl>           <dbl>            <dbl>
## 1     1     69              20               15
## 2     2     84              25               20
## 3     3     72              21               15
## 4     4     79              23               18
## 5     5     20              13                9
## 6     6     NA              20               15

# Read the 2nd data
df2 <- read_excel("Assignments/Resolution_TimeII.xlsx")

# Dimension the 2nd Data
dim(df2)

## [1] 57  2

# Head of data
head(df2)

## # A tibble: 6 × 2
##   SL_No Resolution_Time
##   <dbl>           <dbl>
## 1     1             771
## 2     2             863
## 3     3             863
## 4     4             887
## 5     5             185
## 6     6             725

# Merge the data
df <- merge(df1 , df2 , by = 'SL_No')

# Dimension of merged data
dim(df)
```

```
## [1] 57  5

# Head of merged data
head(df)

##   SL_No Volume Team_Experience Domain_Expertise Resolution_Time
## 1     1     69              20               15             771
## 2     2     84              25               20             863
## 3     3     72              21               15             863
## 4     4     79              23               18             887
## 5     5     20              13                9             185
## 6     6     NA              20               15             725
```

b. Is the dataset contains missing values ? Replace the missing values with appropriate statistics.

- Yes in this data there are misssing values. So, we can replace that missing values by using appropriate statistics.

Replce the missing value of Volume by its mean

```
attach(df)
vol <- df$Volume

# Find the Mean
volume_mean <- mean(vol , na.rm = T)

volume_mean

## [1] 61.71429

# Replace Missing values by Mean
vol[is.na(vol)] = volume_mean

vol

##  [1] 69.00000 84.00000 72.00000 79.00000 20.00000 61.71429 89.00000
59.00000 81.00000 55.00000
## [11] 83.00000 85.00000 64.00000 79.00000 75.00000 88.00000 74.00000
73.00000 31.00000 68.00000
## [21] 85.00000 55.00000 59.00000 61.00000 61.00000 67.00000 62.00000
49.00000 68.00000 60.00000
## [31] 60.00000 51.00000 15.00000  5.00000 67.00000 63.00000 63.00000
70.00000 59.00000 89.00000
## [41] 61.00000 64.00000 62.00000 77.00000 69.00000 56.00000 18.00000
16.00000 68.00000 60.00000
## [51] 72.00000 66.00000 14.00000 72.00000 63.00000 49.00000 72.00000
```

Here we can see that the missing observations are replaced by mean *(61.714)* at 6th position.

Replace Team_Expertise missing value by its mean

```
# Find the Mean
Team_Experience_mean <- mean(Team_Experience , na.rm = T)

Team_Experience_mean

## [1] 18.83929

# Replace Missing values by Mean
Team_Experience[is.na(Team_Experience)] = Team_Experience_mean

Team_Experience

##  [1] 20.00000 25.00000 21.00000 23.00000 13.00000 20.00000 30.00000
16.00000 24.00000 15.00000
## [11] 25.00000 26.00000 19.00000 23.00000 22.00000 27.00000 18.83929
21.00000 13.00000 19.00000
## [21] 26.00000 16.00000 16.00000 17.00000 17.00000 19.00000 18.00000
14.00000 19.00000 17.00000
## [31] 17.00000 15.00000 11.00000  9.00000 19.00000 18.00000 18.00000
21.00000 16.00000 31.00000
## [41] 18.00000 19.00000 18.00000 23.00000 20.00000 16.00000 12.00000
12.00000 19.00000 17.00000
## [51] 21.00000 19.00000 10.00000 21.00000 19.00000 14.00000 21.00000
```

Here we can see that the missing observations are replaced by mean *(18.83929)* at 17th position.

Replace Domain Expertise missing value by its mean

```
# Find the Mean
Domain_Expertise_mean <- mean(Domain_Expertise , na.rm = T)

Domain_Expertise_mean

## [1] 13.42857

# Replace Missing values by Mean
Domain_Expertise[is.na(Domain_Expertise)] = Domain_Expertise_mean

Domain_Expertise

##  [1] 15.00000 20.00000 15.00000 18.00000  9.00000 15.00000 22.00000
10.00000 19.00000  9.00000
## [11] 20.00000 13.42857 13.00000 19.00000 17.00000 21.00000 16.00000
16.00000  9.00000 14.00000
## [21] 21.00000 10.00000 10.00000 12.00000 12.00000 14.00000 12.00000
9.00000 14.00000 11.00000
## [31] 11.00000  9.00000  7.00000  5.00000 14.00000 13.00000 13.00000
15.00000 11.00000 22.00000
## [41] 12.00000 13.00000 12.00000 17.00000 15.00000 10.00000  8.00000
```

```
8.00000 15.00000 11.00000
## [51] 16.00000 13.00000  6.00000 16.00000 13.00000  9.00000 16.00000
```

Here we can see that the missing observations are replaced by mean *(13.42857)* at 12th position.

Combine all columns

```
missing_value_df <- cbind(Volume, Team_Experience, Domain_Expertise,
Resolution_Time)

mdf <- data.frame(missing_value_df) ; head(mdf)

##     Volume Team_Experience Domain_Expertise Resolution_Time
## 1 69.00000              20               15             771
## 2 84.00000              25               20             863
## 3 72.00000              21               15             863
## 4 79.00000              23               18             887
## 5 20.00000              13                9             185
## 6 61.71429              20               15             725
```

c. Prepare the descriptive summary of each variable after replenishment of missing values.

```
library(psych)

describe(mdf)

##                  vars  n   mean     sd median trimmed    mad min  max
range  skew kurtosis
## Volume              1 57  61.71  19.65     64   64.12  11.86   5   89
84 -1.25     1.17
## Team_Experience     2 57  18.84   4.58     19   18.72   2.97   9   31
22  0.32     0.18
## Domain_Expertise    3 57  13.43   4.08     13   13.31   4.45   5   22
17  0.26    -0.55
## Resolution_Time     4 57 660.30 226.63    687  679.66 179.39  57 1081
1024 -0.88     0.55
##                     se
## Volume            2.60
## Team_Experience   0.61
## Domain_Expertise  0.54
## Resolution_Time  30.02
```

1. **Volume**

- There are 57 observations .
- Minimum , Maximum and Range are **5** , **89** and **84** respectively .
- Mean and Trimmed Mean are **61.71** and **64.12** respectively .
- Median is **64** .

- Standard Deviation and standard error are **19.65** and **2.60** respectively.

- Skewness and Kurtosis are **-1.25** and **1.17** respectively .

2. **Team_Experence**

- There are 57 observations .

- Minimum , Maximum and Range are **9** , **31** and **22** respectively .

- Mean and Trimmed Mean are **18.84** and **18.72** respectively .

- Median is **19** .

- Standard Deviation and standard error are **4.58** and **0.61** respectively.

- Skewness and Kurtosis are **0.18** and **0.18** respectively .

3. **Domain_Expertise**

- There are 57 observations .

- Minimum , Maximum and Range are **5** , **22** and **17** respectively .

- Mean and Trimmed Mean are **13.43** and **13.31** respectively .

- Median is **13** .

- Standard Deviation and standard error are **4.08** and **0.54** respectively.

- Skewness and Kurtosis are **-0.55** and **0.54** respectively .

4. **Resolution_Time**

- There are 57 observations .

- Minimum , Maximum and Range are **57** , **1081** and **1024** respectively .

- Mean and Trimmed Mean are **660.30** and **679.66** respectively .

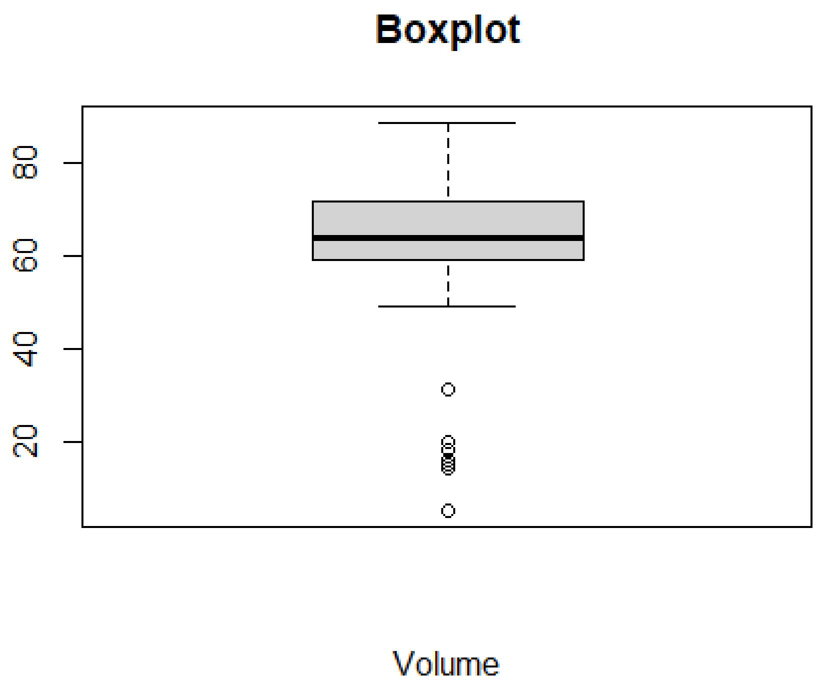
- Median is **687** .

- Standard Deviation and standard error are **226.63** and **30.02** respectively.

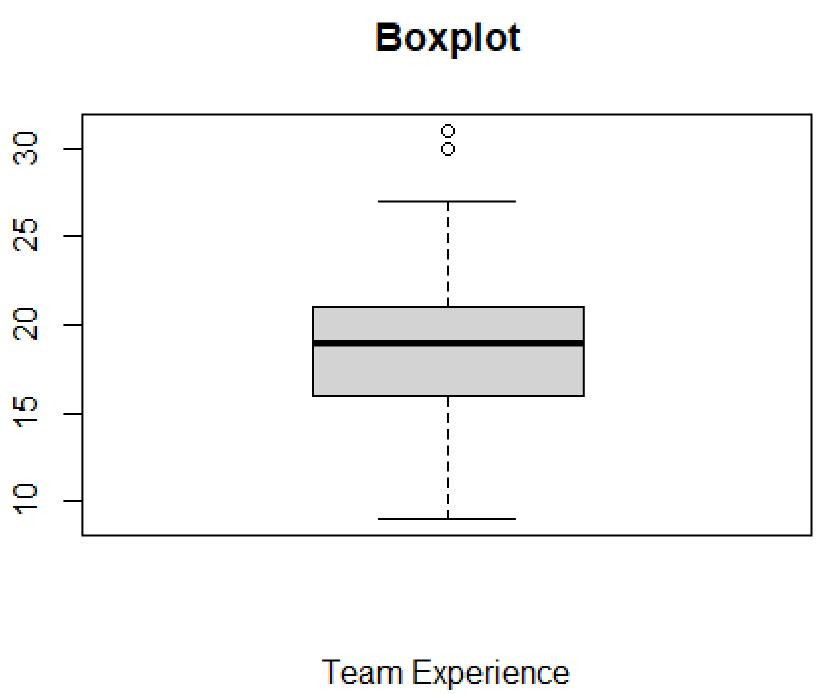
- Skewness and Kurtosis are **-0.88** and **0.55** respectively .

d. Identify the relationship of explanatory variables with resolution time using appropriate data visualization technique ? Using the graph, kindly identify the variables related to the resolution time ?

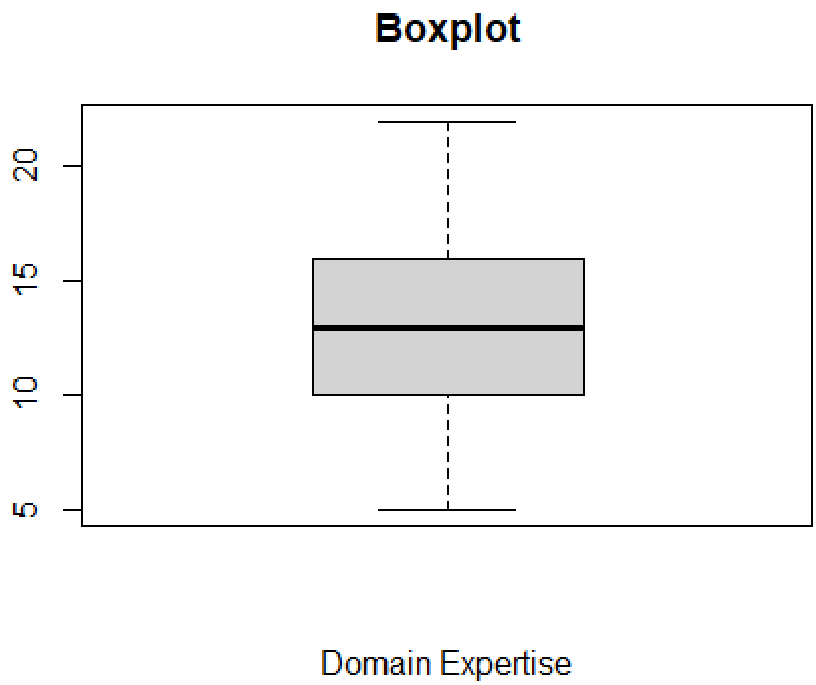
```
# Boxplot of Volume
boxplot(mdf$Volume , main = "Boxplot" , xlab = "Volume")
```



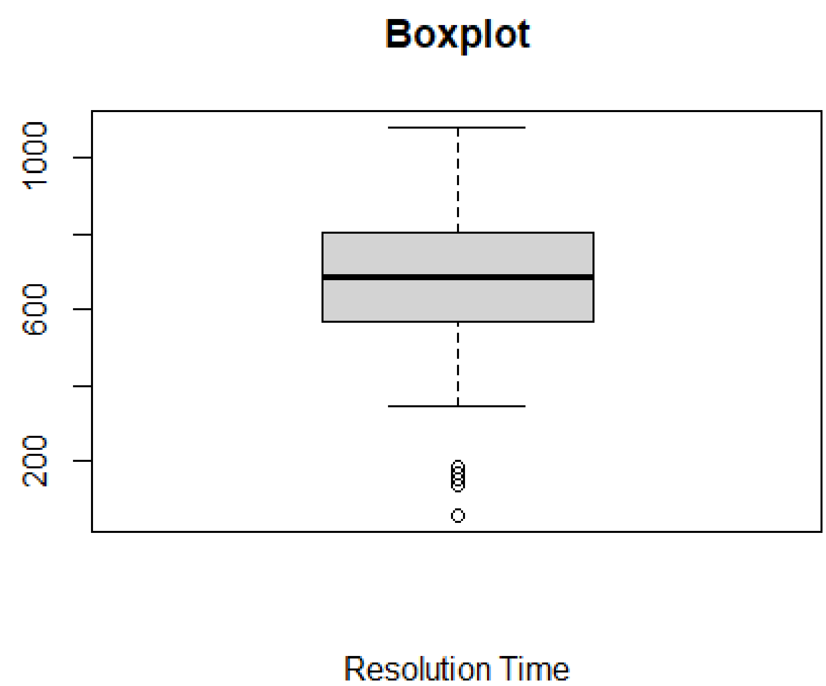

```r
# Boxplot of Team_Experience
boxplot(mdf$Team_Experience, main = "Boxplot", xlab = "Team Experience")
```



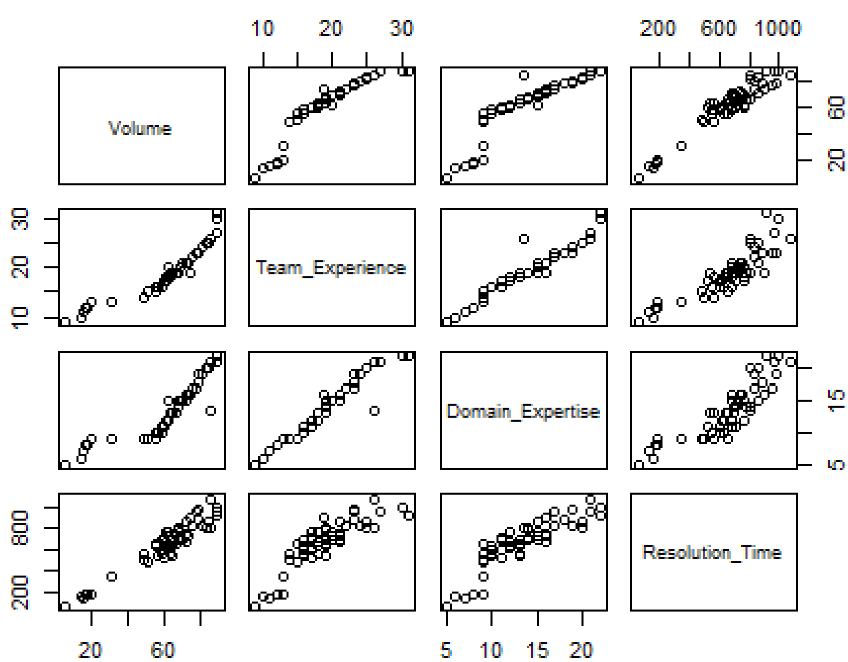
```r
# Boxplot of Domain_Expertise
boxplot(mdf$Domain_Expertise, main = "Boxplot", xlab = "Domain Expertise")
```




```r
# Boxplot of Volume Resolution_Time
boxplot(mdf$Resolution_Time, main = "Boxplot", xlab = "Resolution Time")
```



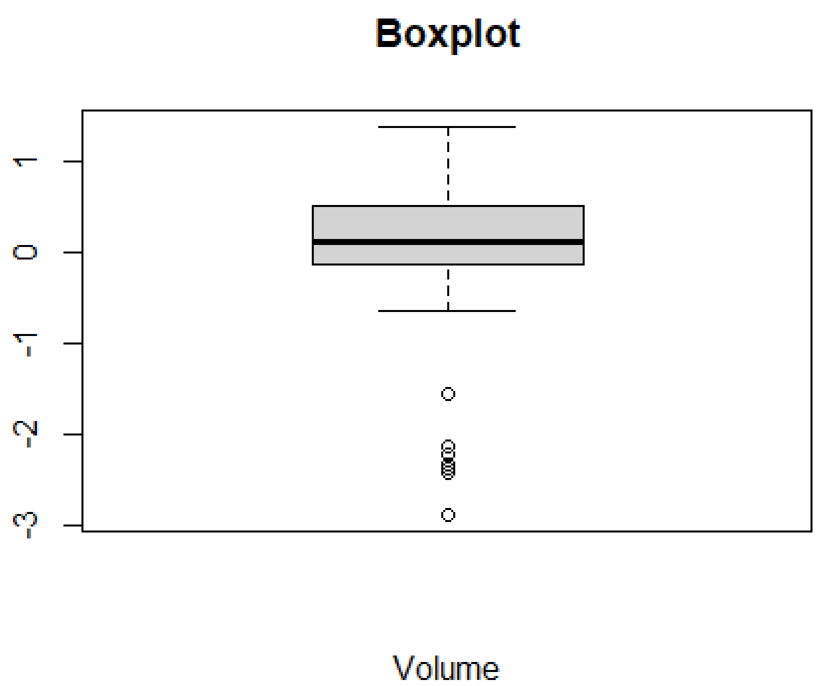

```
# Matrix Plot to find the relationship among variables
plot(mdf)
```



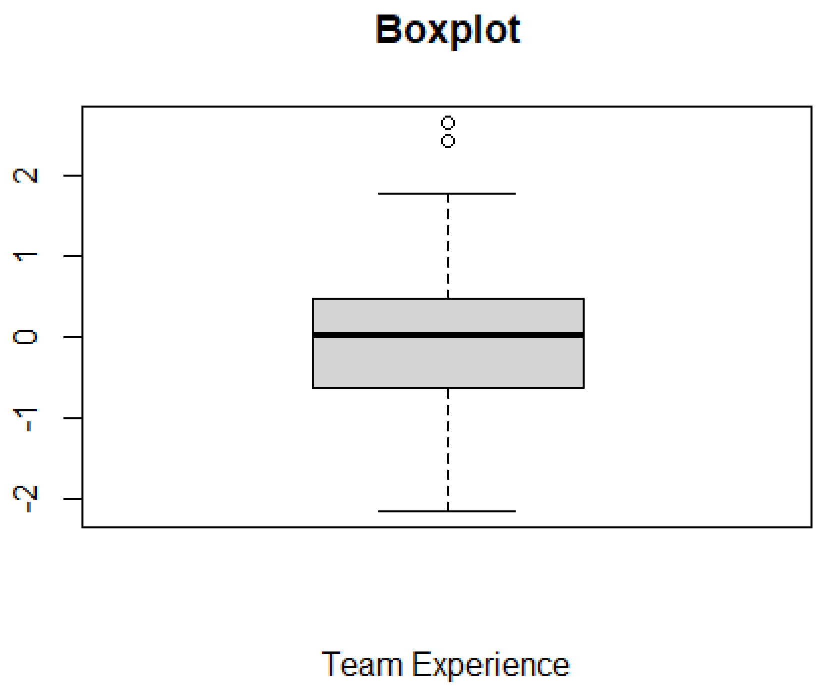
e. Scale the variables using z and Min-Max transformations. After each transformation, identify the relationship of explanatory variables with resolution time using appropriate data visualization technique ? Is there any change in the relationship of variables with resolution time after transformation ?

```r
# 1.  Using scale function
z_df <- scale(mdf)

# Transformed data head
z_df <- data.frame(z_df)
head(z_df)

##       Volume Team_Experience Domain_Expertise Resolution_Time
## 1  0.3707008       0.2535525        0.3852541       0.4884649
## 2  1.1339084       1.3457786        1.6110626       0.8944095
## 3  0.5233423       0.4719977        0.3852541       0.8944095
## 4  0.8795059       0.9088882        1.1207392       1.0003081
## 5 -2.1224439      -1.2755641       -1.0857161      -2.0972254
## 6  0.0000000       0.2535525        0.3852541       0.2854927

# 2. Using Own function
m <- apply(mdf, 2, mean)

s <- apply(mdf, 2, sd)

zt_df <- (mdf - m) / s

zt_df <- data.frame(zt_df)

#  3. Computing minimum and maximum of data
mins <- apply(mdf, 2, min)

maxs <- apply(mdf, 2, max)

# Making min-max transformation
tr_data <- scale(mdf, center = mins, scale = maxs - mins)

tr_data <- data.frame(tr_data)

# Head of Transformation
head(tr_data)

##      Volume Team_Experience Domain_Expertise Resolution_Time
## 1 0.7619048       0.5000000        0.5882353       0.6972656
## 2 0.9404762       0.7272727        0.8823529       0.7871094
## 3 0.7976190       0.5454545        0.5882353       0.7871094
## 4 0.8809524       0.6363636        0.7647059       0.8105469
## 5 0.1785714       0.1818182        0.2352941       0.1250000
## 6 0.6751701       0.5000000        0.5882353       0.6523438
```

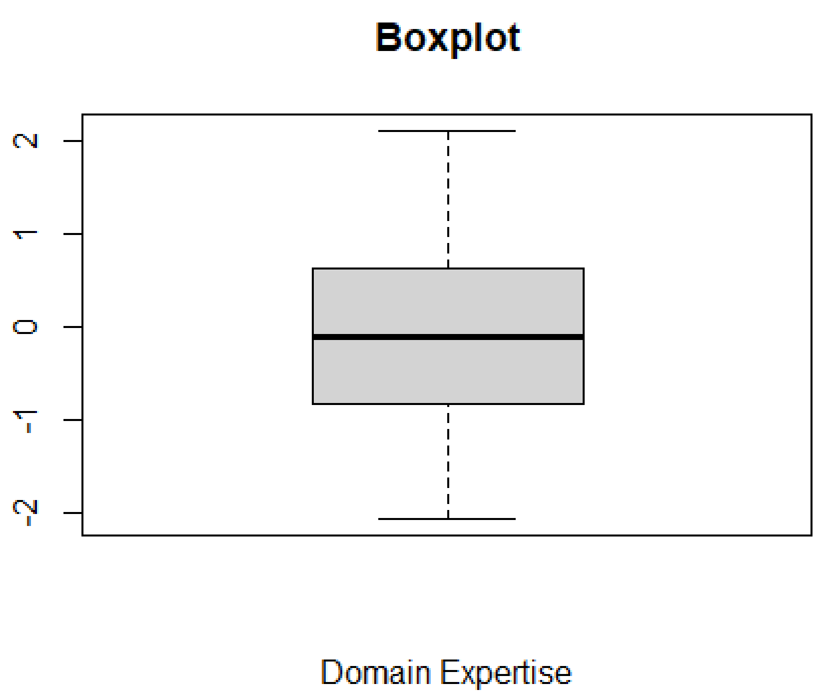
```r
# Data Vusualization

# Boxplot of Volume
boxplot(z_df$Volume , main = "Boxplot" , xlab = "Volume")
```



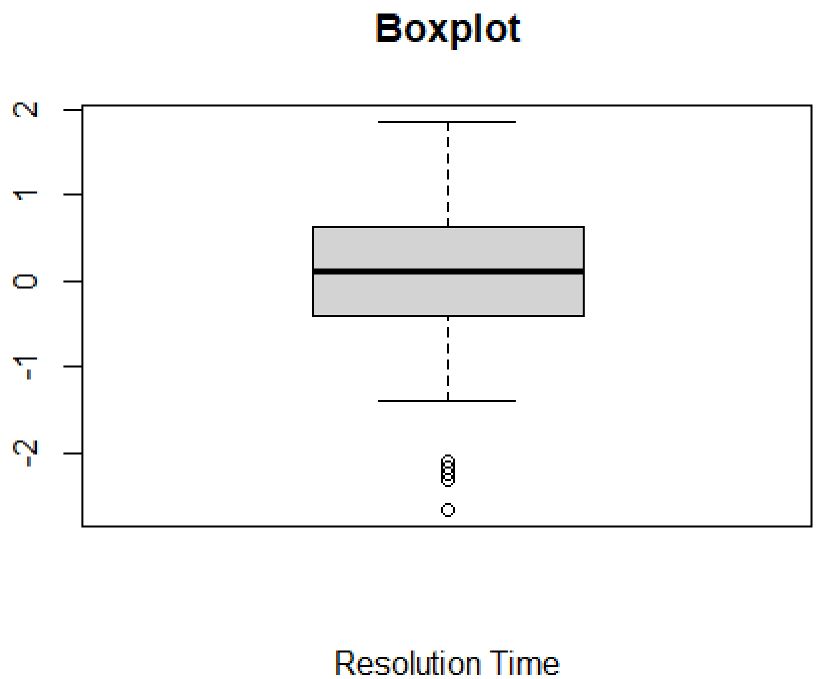

```r
# Boxplot of Team_Experience
boxplot(z_df$Team_Experience, main = "Boxplot", xlab = "Team Experience")
```



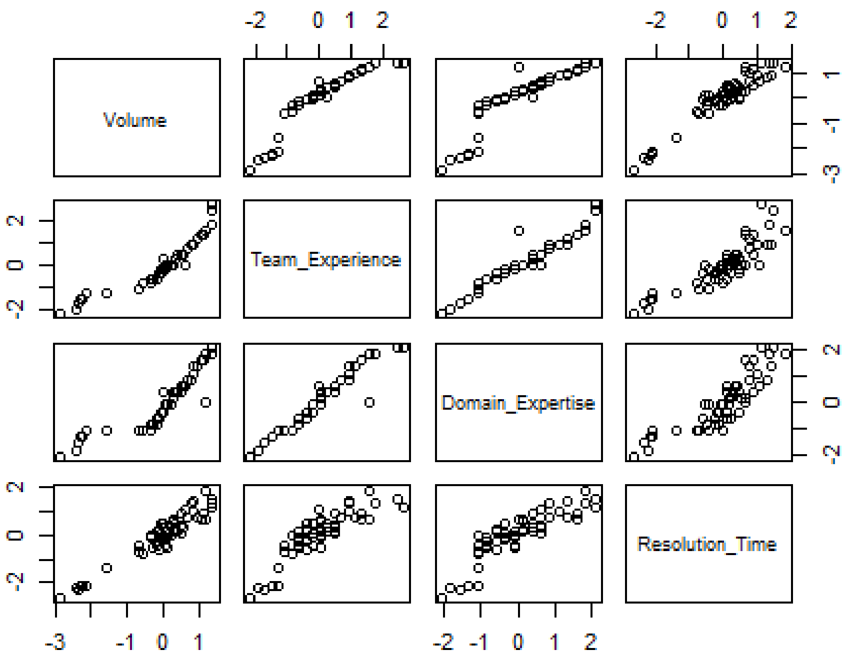

```r
# Boxplot of Domain_Expertise
boxplot(z_df$Domain_Expertise, main = "Boxplot", xlab = "Domain Expertise")
```



```r
# Boxplot of Volume Resolution_Time
boxplot(z_df$Resolution_Time, main = "Boxplot", xlab = "Resolution Time")
```




```r
plot(z_df)
```



We can easily see that the *mean* and *standard deviation* of data is reduce to *0* and *1* respectively. It means that now the data follow *Standard Normal Distribution*.