Q1# Use R built-in data frame named cars. Find the summary of data using two different syntaxes. Find the histogram, rug plot, and fit the normal density to see how much it is deviated from normal (all are shown in one graph for each variable). Interpret your results with graphical representation. Make sure results should brief and clear. Make the boxplot of the data set which show the distribution of each variable.

Solution:

Method 1.

```
car_data_vector <- data("cars")

car_data_vector <- cars

#car_data_vector

summary(car_data_vector)
```

Result:

```
     speed           dist
 Min.   : 4.0   Min.   :  2.00
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
```

Method 2.

```
summary_x <- c(min(car_data_vector$speed),quantile(car_data_vector$speed,0.25),

median(car_data_vector$speed),mean(car_data_vector$speed)

     ,quantile(car_data_vector$speed,0.75),max(car_data_vector$speed))

#summary_x

summary_x1=c(min(car_data_vector$dist),quantile(car_data_vector$dist,0.25),median(car_data_vector
$dist),mean(car_data_vector$dist),

     quantile(car_data_vector$dist,0.75),max(car_data_vector$dist))

#summary_x1
```

Result:

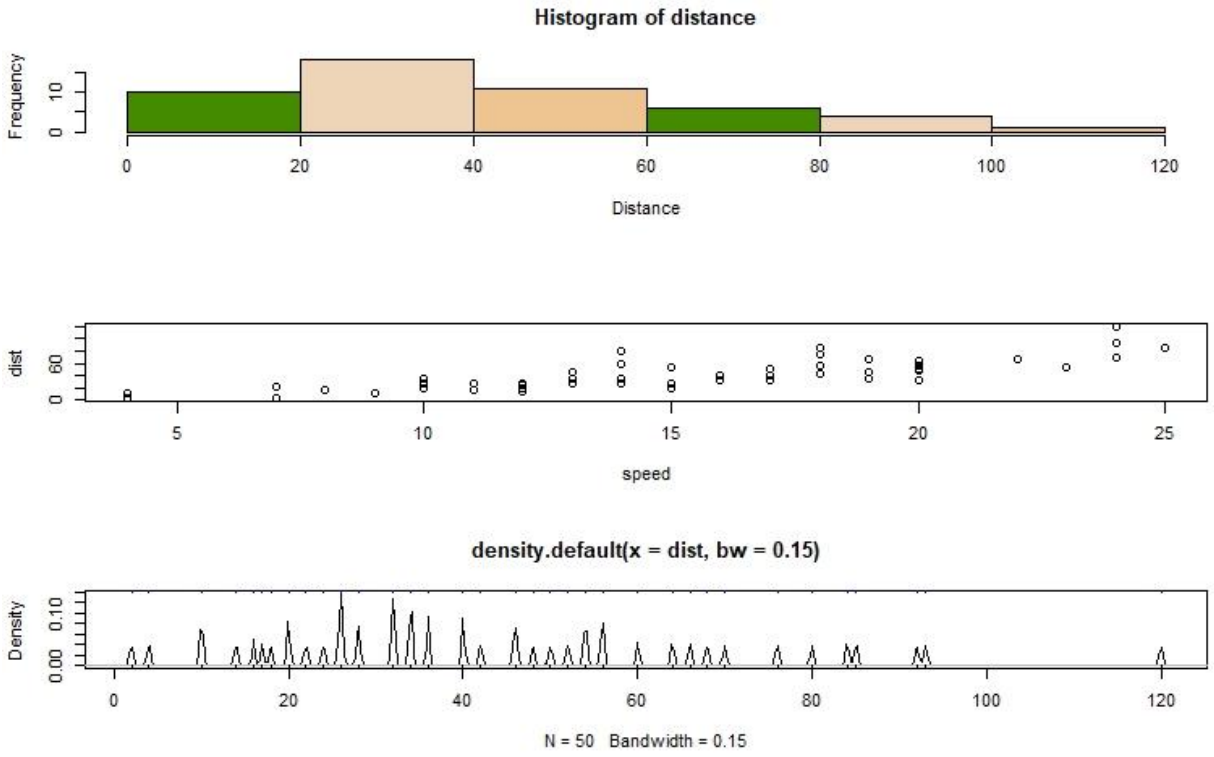
```
>summary_x
      25%             75%
 4.0 12.0 15.0 15.4 19.0 25.0
```

```
>summary_x1
          25%                     75%
  2.00  26.00  36.00  42.98  56.00 12
```

R_code:

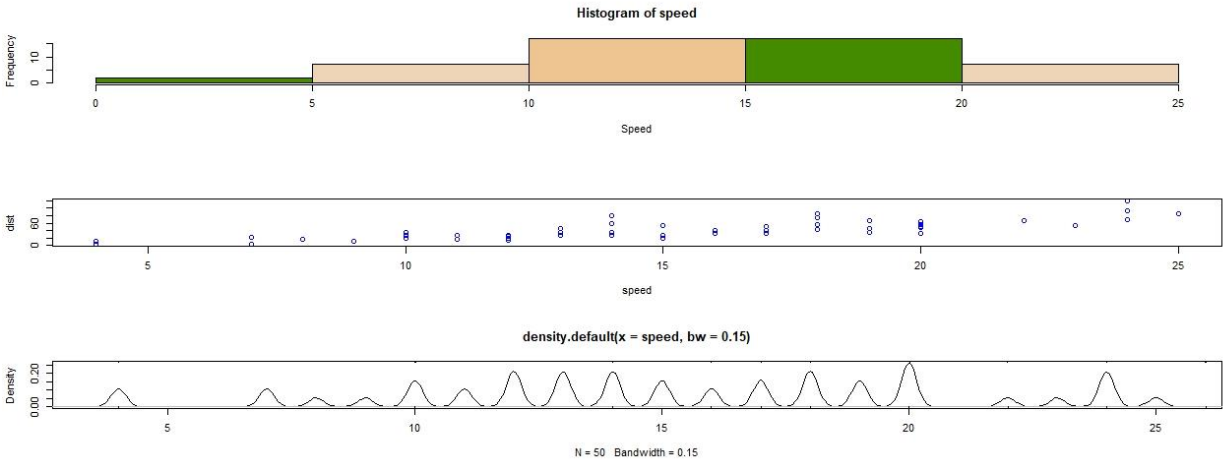
```
par(mfrow=c(3,1))
hist(car_data_vector$dist,col=c("chartreuse4","bisque2","burlywood2"),main="Histogram of distance",xlab="Distan
ce")
plot(car_data_vector,col="black")
require(stats)
with(car_data_vector,{
  plot(density(dist,bw=0.15))
  rug(dist)
  rug(jitter(dist,amount=0.01),side=3,col="black")
})
```

Result:



R_code:

```
par(mfrow=c(3,1))
hist(car_data_vector$speed,col=c("red","yellow","green"),main="Histogram  of speed",xlab="Speed")
plot(car_data_vector,col="blue")
require(stats)
with(car_data_vector,{
  plot(density(speed,bw=0.15))
  rug(dist)
  rug(jitter(dist,amount=0.01),side=3,col="blue")
})
```

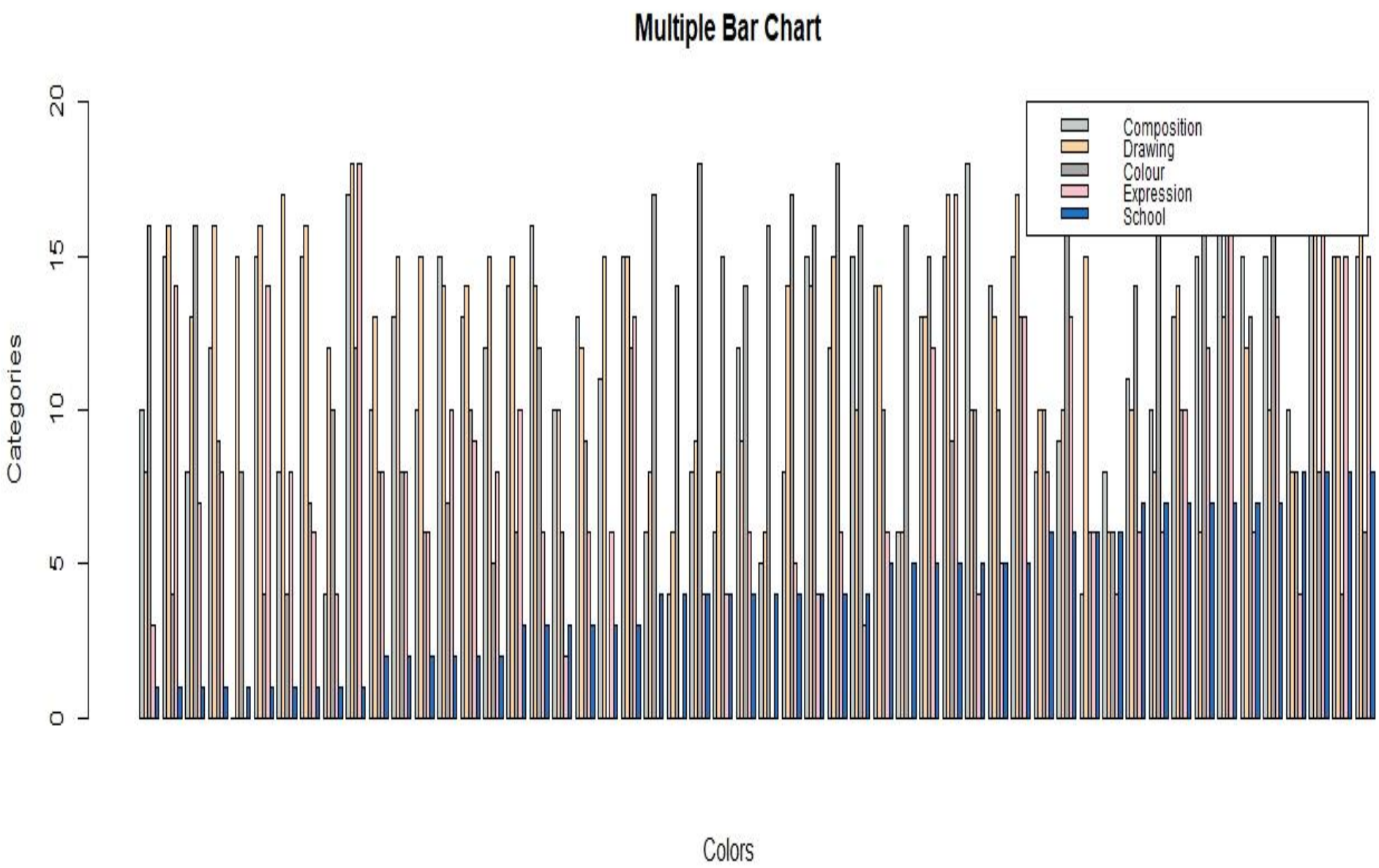


Q2# Using qualitative or categorical R built-in data frame named painters. Explore it (Graphically and with numbers) using suitable measures with R?

Solution:

Part A:

```
color_data=painters$Colour

color_data.freq <- table(color_data)

expression_data <- painters$Expression

expression_data.freq <- table(expression_data)

school_data <- painters$School

school_data.freq <- table(school_data)


colorx <- c("Da Udine","Da Vinci","Del Piombo","Del Sarto","Fr. Penni","Guilio Romano","Michelangelo","Perin del Vaga","Perugino","Raphael","F. Zucarro")

Colors <- c ("azure3","burlywood1","darkgray", "pink" , "dodgerblue3")

data <- rbind(composition , drawing , color_data, expression_data, school_data)

barplot(data,main="Multiple Bar Chart",col=Colors,beside=TRUE,ylab = "Categories",xlab = "Colors",ylim = c(0,20),xlim=c(0,310))

legend("topright",c("Composition","Drawing","Colour","Expression", "School"),fill = Colors,cex = 0.75)
```


Multiple Bar Chart
Categories
0
5
10
15
20
Colors
Composition
Drawing
Colour
Expression
School

Part B:

Relative Frequencies:

```
library(MASS)
A)
drawing_data <- painters$Drawing
drawing_data.freq <- table(drawing_data)
drawing_data.relfreq <- drawing_data.freq / nrow(painters)
drawing_data.relfreq
B)
color_data <- painters$Colour
color_data.freq <- table(color_data)
color_data.relfreq <- color_data.freq / nrow(painters)
color.relfreq
C)
expression_data <- painters$Expression
expression_data.freq <- table(expression_data)
expression_data.relfreq <- expression_data.freq / nrow(painters)
expression_data.relfreq
D)
composition <- painters$Composition
composition.freq <- table(composition)
composition.relfreq <- composition.freq / nrow(painters)
composition.relfreq
E)
school_data <- painters$School
school_data.freq <- table(school_data)
school_data.relfreq <- school_data.freq / nrow(painters)
school_data.relfreq
```

```
F
#1
hist(composition,main="Histogram of compositions",col=c("cornsilk2","darkgray"

,"deepskyblue2","darkslateblue","purple"),xlab="Composition",ylab="Values")
#2
hist(drawing_data,main="Histogram of drawing",col=c("cornsilk2","darkgray"

,"deepskyblue2","darkslateblue","purple"),xlab="Drawing",ylab="Values")
#3
hist(color_data,main="Histogram of Colour",col=c("cornsilk2","darkgray"
                                    ,"deepskyblue2","darkslateblue","purple"),xlab="Color",ylab="Values")
```

#4

```
hist(expression_data,main="Histogram of Expression",col=c("cornsilk2","darkgray"

,"deepskyblue2","darkslateblue","purple"),xlab="Expression",ylab="Values")
```

Result:

A)

6	8	9	10	12	13	14	15	16	17
0.092 59259	0.092 59259	0.037 03704	0.129 62963	0.055 55556	0.092 59259	0.129 62963	0.185 18519	0.092 59259	0.074 07407

18
0.01851 852

B)

0	4	5	6	7	8	9	10	12	13
0.018 51852	0.074 07407	0.018 51852	0.111 11111	0.037 03704	0.092 59259	0.055 55556	0.129 62963	0.055 5556	0.037 03704

14	15	16	17	18
0.05555 556	0.03703 704	0.14814 815	0.09259 259	0.03703 704

C)

0	2	3	4	5	6	7	8	9	10
0.092 59259	0.018 51852	0.037 03704	0.129 62963	0.037 03704	0.222 22222	0.018 51852	0.111 11111	0.018 51852	0.055 55556

12	13	14	15	16	17	18
0.03703 704	0.07407 407	0.03703 704	0.03703 704	0.01851 852	0.03703 704	0.01851 852

D)

0	4	5	6	8	9	10	11	12	13
0.018 51852	0.055 55556	0.018 51852	0.055 55556	0.111 11111	0.018 51852	0.111 11111	0.037 03704	0.074 07407	0.092 59259

14	15	16	17	18

0.05555 556	0.25925 926	0.03703 704	0.01851 852	0.03703 704

E)

A	B	C	D	E	F	G	H
0.185 18519	0.111 11111	0.111 11111	0.185 18519	0.129 62963	0.074 07407	0.129 62963	0.074 07407

Mean

> mean(painters$Composition)

[1] 11.55556

> mean(painters$Colour)

[1] 10.94444

> mean(painters$Expression)

[1] 7.666667

> mean(painters$Drawing)

[1] 12.46296

Median

> median(painters$Composition)

[1] 12.5

> median(painters$Expression)

[1] 6

> median(painters$Colour)

[1] 10

> median(painters$Drawing)

[1] 13.5

Std. Dev

> sd(painters$Composition)

[1] 4.087102

> sd(painters$Expression)

[1] 4.797798

> sd(painters$Colour)

[1] 4.651706

> sd(painters$Drawing)

[1] 3.457084

Variance

> var(painters$Composition)

[1] 16.7044

> var(painters$Expression)

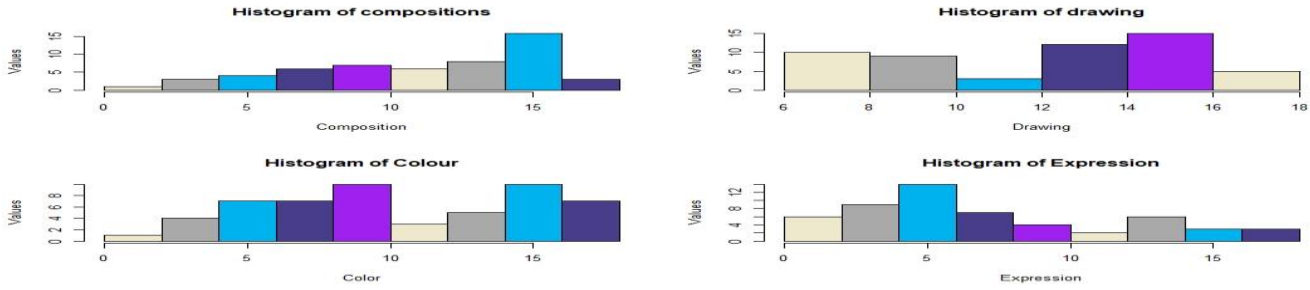
[1] 23.01887

> var(painters$Colour)

[1] 21.63836

> var(painters$Drawing)

[1] 11.95143



Q3# A class of students played computer game which tested how quickly they reached to a visual instruction to press a particular key. The computer measured their reaction times in tenths of a second and stored a record of the sex and reaction times of each student. Finally it displayed the following summary statistics for the whole class.(R-based activity**)**

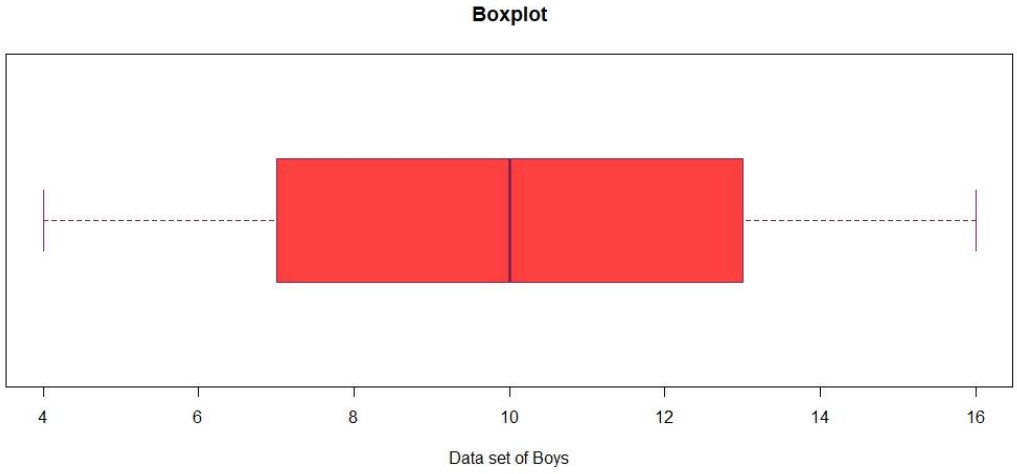
a. Draw two box plots suitable for comparing the reaction times of the boys and girls.
b. Write a brief comparison of the performance of boys and girls in the game

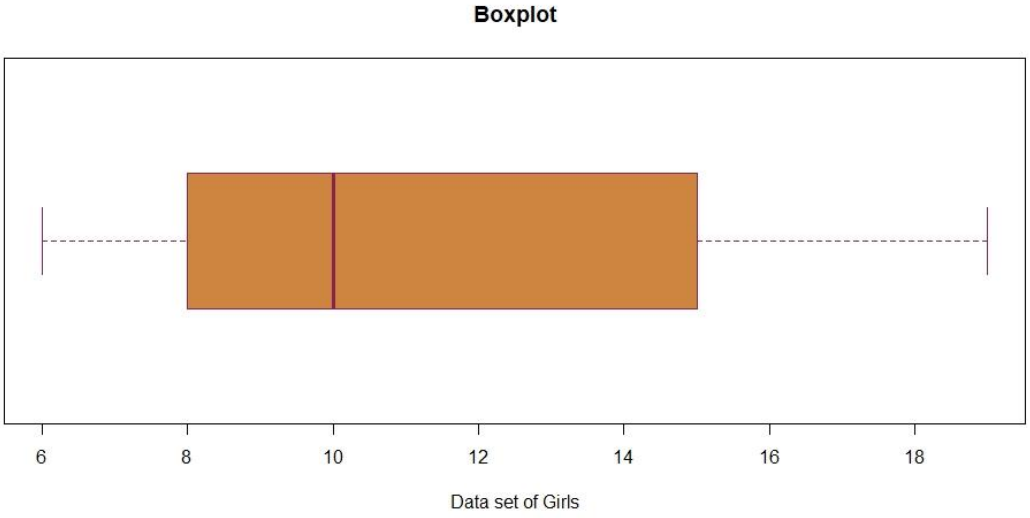
Solution:

```
#Girls
median_g <- 10
lowerQ_g <- 8
upperQ_g <- 15
min_g <- 6
max_g <- 19
```

```
#Boys
median_b <- 10
lowerQ_b <- 7
upperQ_b <- 13
min_b <- 4
max_b <- 16
```

```
data_g_vector <- c(median_g,lowerQ_g,upperQ_g,min_g,max_g)
data_b_vector <- c(median_b,lowerQ_b,upperQ_b,min_b,max_b)
boxplot( data_g_vector, main = "Boxplot",
    col = "tan3",
    names=c("Girls" ),
    border = "violetred4",xlab="Data set of Girls",
    horizontal = TRUE,
    notch = FALSE)
boxplot( data_b_vector, main = "Boxplot",
    col = "brown1",
    names=c("Boys" ),
    border = "violetred4",xlab="Data set of Boys",
    horizontal = TRUE,
    notch = FALSE)
```





(b).Comparison of performance of Boys and Grisl

1. No outliers in the data set.

2. Lowest measured reaction times in case of Boys is 4 and Girls is 6.

3. About 25% of measured reaction times of Boys is more than 13 and of Girls is more than 15.

4. About 75% of measured reaction times of Boys is more than 7 and of Girls is more than 8.

5. Maximum measured reaction times of Boys is 16 and of Girls is 19.

6. Median of measured reaction times of both Boys and Girls is 10.

Q4# In the manufacturing of a certain scientific instrument great importance is attached to the life of a particular critical component. This component is obtained in bulk from two sources, A and B, and in the course of inspection, the lives of 1000 of the components from each source are determined. The following frequency tables are obtained:-

1. Find Median and two quartiles for each group.
2. Find mean and Standard deviation for each source and compare them.
3. Which source do you think providing better quality of components and why? Note: answer this part
by considering results of mean and standard deviation.
4**. Calculate absolute and relative measure of skewness for each group and comment on life of the**
components for each source.
Note: Every calculation is carried out using R-lang.

Solution:

A.

(a)

```
Vector_components <- c(40,96,364,372,85,43)

colnames <- "No. of Components"

rownames <- c("[1000-1020]", "(1020-1040]", "(1040-1060]", "(1060-1080]",

        "(1080-1100]", "(1100-1120]")

result_1 = matrix(Vector_components, dimnames=list(rownames,
colnames),nrow=length(Vector_components))

data.frame(result_1, "cumsum"=cumsum(result_1))
```

```
        No.of.Components cumsum
[1000-1020]               40     40
(1020-1040]               96    136
(1040-1060]              364    500
(1060-1080]              372    872
(1080-1100]               85    957
(1100-1120]               43   1000
```

```
#Media A

medianA=1040 + (20/364)*(500 - 136)

medianA
```

Result:

```
[1] 1060
```

(b)

```
Vector_components <- c(339,136,25,20,130,350)

colnames <- "No. of Components"

rownames <- c("[1030-1040]", "(1040-1050]", "(1050-1060]", "(1060-1070]",

        "(1070-1080]", "(1080-1090]")

result_1 <- matrix(Vector_components, dimnames=list(rownames, colnames),
nrow=length(Vector_components))

data.frame(result_1, "cumsum"=cumsum(result_1))
```

```
        No.of.Components cumsum
[1030-1040]              339    339
(1040-1050]              136    475
(1050-1060]               25    500
(1060-1070]               20    520
(1070-1080]              130    650
(1080-1090]              350   1000
```

#MedianB

```
medianB=1050 + (10/25)*(500-475)

medianB
```

Result:

```
[1] 1060
```

#Quartiles

```
 quart_1_Data_A=1040+(20/364)*(250 - 136)
 quart_1_Data_A
[1] 1046.264

 quart_3_Data_A=1060+(20/372)*(750 - 500)
 quart_3_Data_A
[1] 1073.441

 quart_1_Data_B=1030+ (10/339) * (250-0)
 quart_1_Data_B
[1] 1037.375

 quart_3_Data_B=1080 + (10/350)*(750 - 650)
 quart_3_Data_B
[1] 1082.857
```

B.

meanA=sum of all(f.x)/sum of all(f).

> meanA=1059900/1000

> meanA

[1] 1059.9

meanA=sum of all(f.x)/sum of all(f).

> meanB=1060160/1000

> meanB

[1] 1060.16

> SD_A= sqrt(444000/1000)

> SD_A

[1] 21.07131

> SD_B= sqrt(491567.84 / 1000)

> SD_B

[1] 22.17133

C.

i. Comparing A with B source A has **provided good Quality** of Data
ii. A and B both have almost same **mean**.
iii. A has lower **Standard Deviation** than B.
iv. Considering above statement, we can say that A is closer to **average.**
v. A is **less spread** than Source B.

Note: Here A and B refers to Source A and Source B respectively

D.

Note:

If Skewness is 0. The shape of graph will be **Symmetric** and the growth rate of life of Component will be Normal.

Coefficient of Skewness for Source A:

> Sk_A=(3*(meanA - medianA)) / (SD_A)

> Sk_A

[1] -0.01423737

Shape of Graph is Symmetric

Coefficient of Skewness for Source B:

> Sk_B=(3*(meanB - medianB)) / (SD_B)

> Sk_B

[1] 0.02164958

Shape of Graph is Symmetric

Q# 5 The following table shows data about the time taken in seconds to the nearest second, for completing each one of a series of 75 similar chemical experiments.

a. State the type of the graph appropriate for illustrating.
b. Calculations using the data in the table give the estimates as follows mean time of the experiments 69.64s, and standard deviation 6.37s. Explain why these are estimates rather than precise values.
c. Estimate the median and interquartile range of the times taken for completing experiments.

d. It was subsequently revealed that the four experiments in the 50-60 class had actually taken 57, 59, 59, and 60 seconds respectively. State, without further calculation, what effect (if any) there would be on the estimates of the median, interquartile range and the mean if this information were taken into account.
e. Does the graph exhibit positive skewness, negative skewness or no skewness and how do you measure it?

Solution:

a).

Histogram **would be the most suitable graph for given scenario.**

b).

These mean and standard deviation are estimates ,because the time interval for completing each one of series of 75 chemical experiments is different for different number of experiments.Moreover ,class width is also different for some classes as in the 1st 2 classes ,class width is 11 but in the remaining classes ,width is 5.So,rather than saying exact mean and standard deviation ,we would say it estimate.

c).

Median**= l + h/f (n/2 – c.f)**

= 65.5 + 5/26 (37.5 - 17)

= 69.44

Inter Quartile Range (IQR) **= $Q_3 - Q_1$**

Q_1 = l + h/f (n/4 – c.f)

= 65.5 + 5/26 (18.75 - 17)

= 65.83

Q_3 = l + h/f (3n/4 – c.f)

= 70.5 + 5/22 (56.25 - 43)

= 73.51

So, IQR= 73.51 – 65.83

= 7.68

d).

This information will not effect mean. Median or IQR as it is the measure of location and the values will not affect the location as long as they are in the range of given class.

e).

By using Karl's measure of Skewness:

S_k = (Mean – Mode) / Std.Dev.

- **Mean = 69.64**
- **Mode = 69.32**
- **Std.Dev = 6.365**

So, S_k = (69.64 – 69.32) / 6.365

= 0.05

Since, S_k = 0, the graph has no Skewness .