Data classification and diagrammatic representation

```
#Pie Chart

x <- c(21, 62, 10, 53)

labels <- c("London", "New York", "Singapore", "Mumbai")

pie(x,labels)

#Pie chart with title

x <- c(21, 62, 10, 53)

labels <- c("London", "New York", "Singapore", "Mumbai")

pie(x, labels, main = "City pie chart", col = rainbow(length(x)))


#Bar plot

H <- c(7,12,28,3,41)

barplot(H)


#Bar plot with labels

H <- c(7,12,28,3,41)

M <- c("Mar","Apr","May","Jun","Jul")

barplot(H,names.arg=M,xlab="Month",ylab="Revenue",col="blue",main="R
evenue chart",border="red")


#Box plot

input <- mtcars[,c('mpg','cyl')]

print(head(input))

boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders", ylab
= "Miles Per Gallon", main = "Mileage Data")


#Histogram

v <-  c(9,13,21,8,36,22,12,41,31,33,19)

hist(v,xlab = "Weight",col = "yellow",border = "blue")


#Line graph

# plot(v,type,col,xlab,ylab)
v <- c(7,12,28,3,41)
```

```
plot(v,type = "o")


#Line Graph with heading

v <- c(7,12,28,3,41)

plot(v,type = "o", col = "red", xlab = "Month", ylab = "Rain fall",
main = "Rain fall chart"


#Scatter plots

#plot(x, y, main, xlab, ylab, xlim, ylim, axes)
input <- mtcars[,c('wt','mpg')]

print(head(input))

plot(x = input$wt,y = input$mpg,

   xlab = "Weight",

   ylab = "Milage",

   xlim = c(2.5,5),

   ylim = c(15,30),

   main = "Weight vs Milage"

)


#ALL scatter plots together

pairs(~wt+mpg+disp+cyl,data = mtcars, main = "Scatterplot Matrix")


########################################################

Categorical variables in R are stored into a factor.

factor(x = character(), levels, labels = levels, ordered =
is.ordered(x))


R does not use the terms nominal, ordinal, and interval/ratio for
types of variables.

In R, nominal variables can be coded as variables
with factor or character classes.

Continuous variable/Interval/ratio data can be coded as variables
with numeric or integer classes. An L used with values to tell R to
store the data as an integer class.
```

We can code ordinal data as either numeric or factor variables, depending on how we will be summarizing, plotting, and analyzing it.

```
sex <- factor(c("male", "female", "female", "male"))

levels(sex)

nlevels(sex)
```

Nominal Categorical Variable

```
# Create a color vector

color_vector <- c('blue', 'red', 'green', 'white', 'black', 'yellow')

# Convert the vector to factor

factor_color <- factor(color_vector)

factor_color
```

Ordinal Categorical Variable

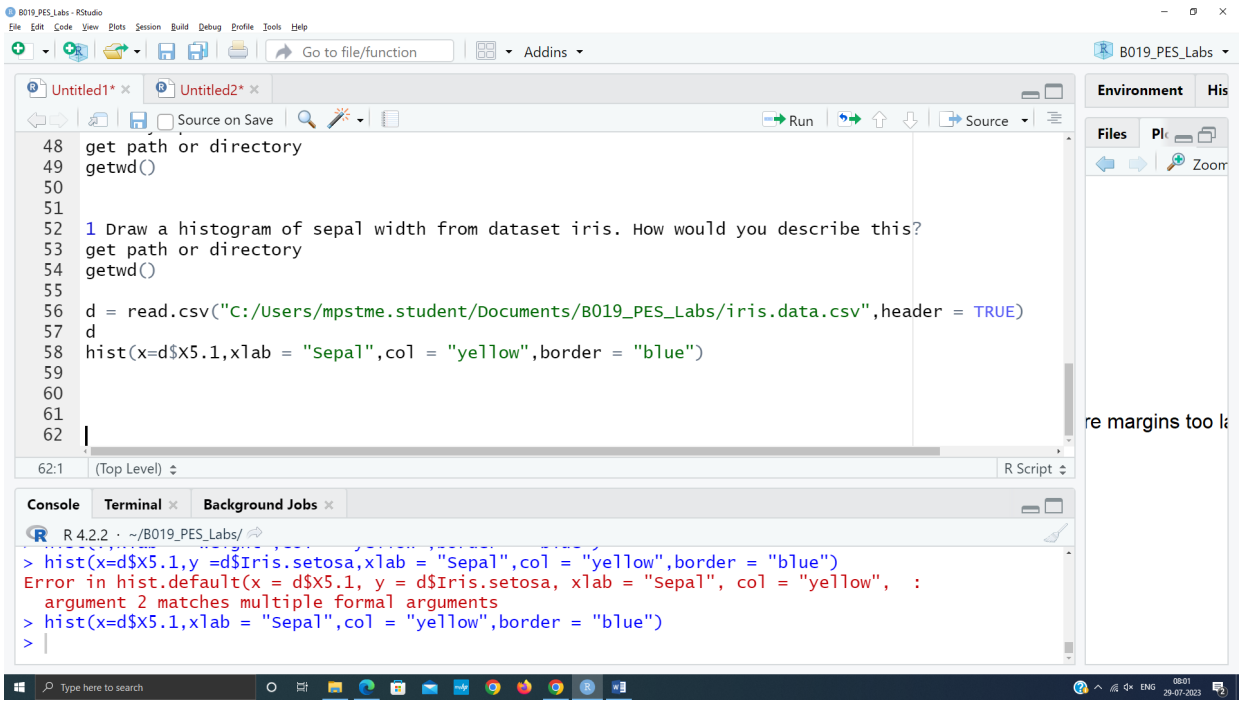
```
# Create Ordinal categorical vector

day_vector <- c('evening', 'morning', 'afternoon', 'midday', 'midnight', 'evening')

# Convert `day_vector` to a factor with ordered level

factor_day <- factor(day_vector, order = TRUE, levels =c('morning', 'midday', 'afternoon', 'evening', 'midnight'))

# Print the new variable

factor_day

## Levels: morning < midday < afternoon < evening < midnight

# Append the line to above code

# Count the number of occurence of each level

summary(factor_day)

##Continuous Data

dataset <- mtcars

class(dataset$mpg)
```

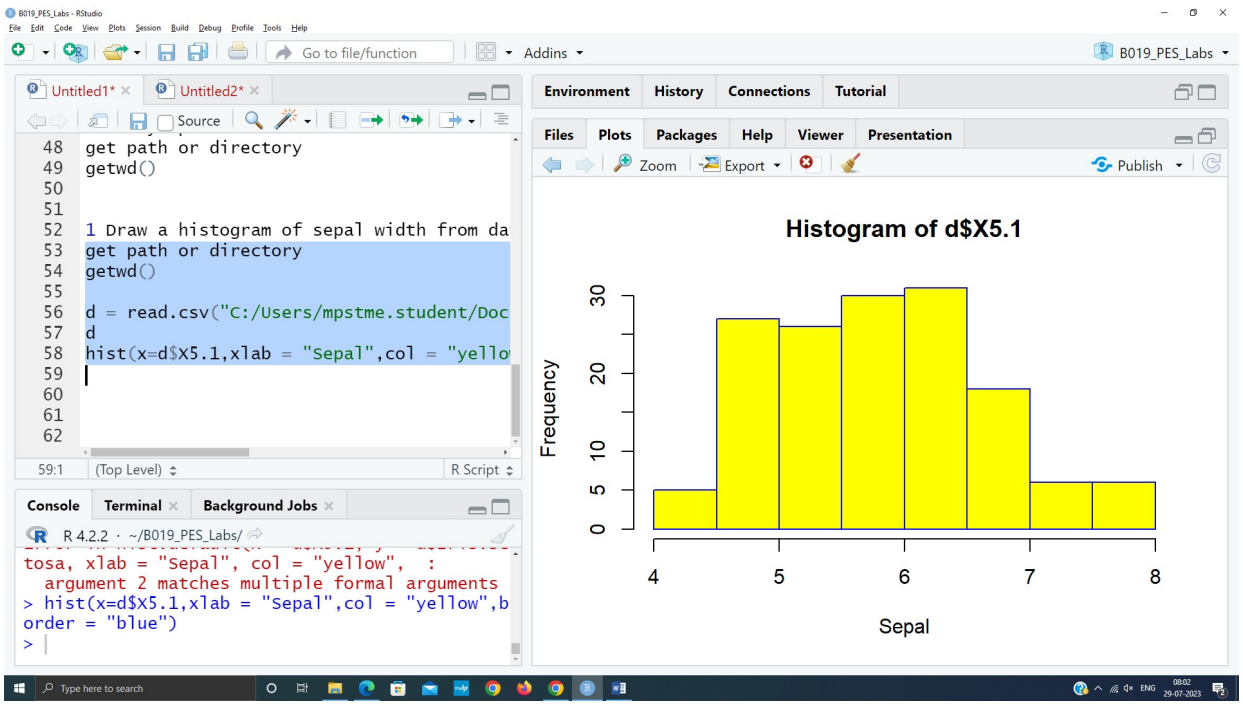
```
##Importing data from excel

dat <- read.csv( file = "data.csv", header = TRUE, sep = ",", dec =
".")


##Easy enough: at the beginning of your script, simply
add library(readxl) to the list of libraries you are loading. And
tidyverse package by using install.package(tidyverse) command

df <- read_excel("<name and extension of your file>")
```

Exercises

1 Draw a histogram of sepal width from dataset iris. How would you describe this?

```
get path or directory
getwd()

d =
read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/iris.data
.csv",header = TRUE)
d
hist(x=d$X5.1,xlab = "Sepal",col = "yellow",border = "blue")
```



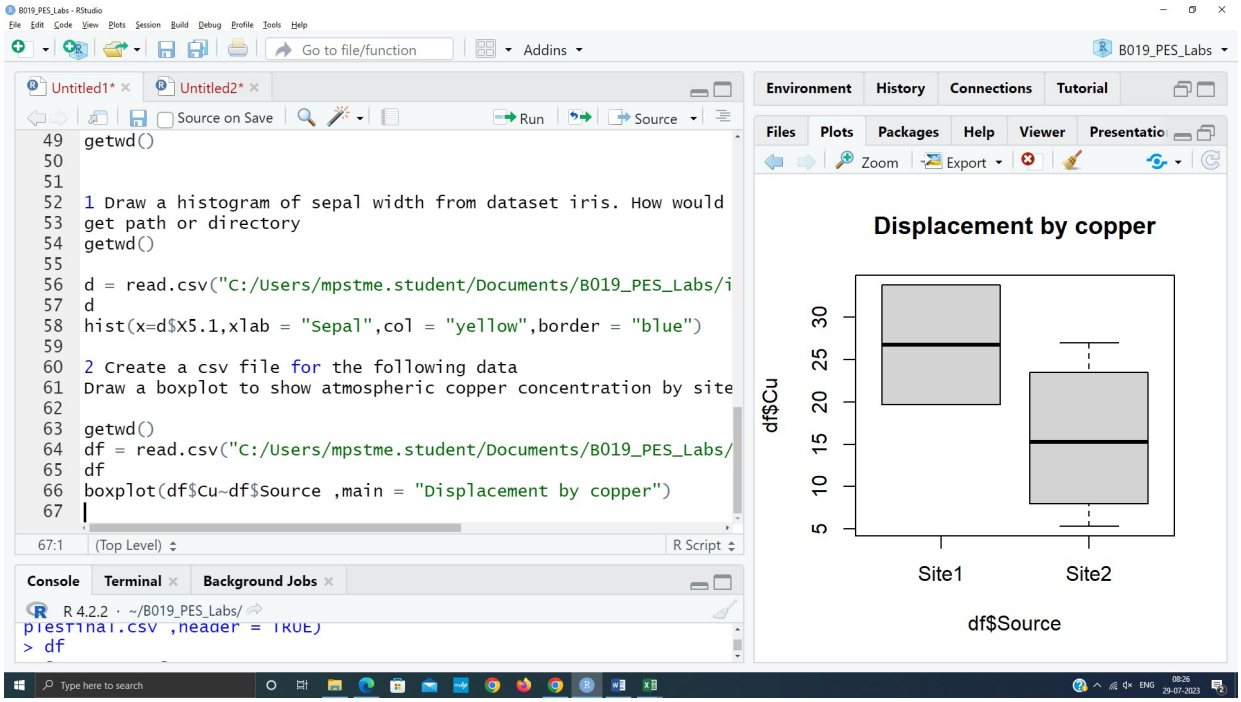


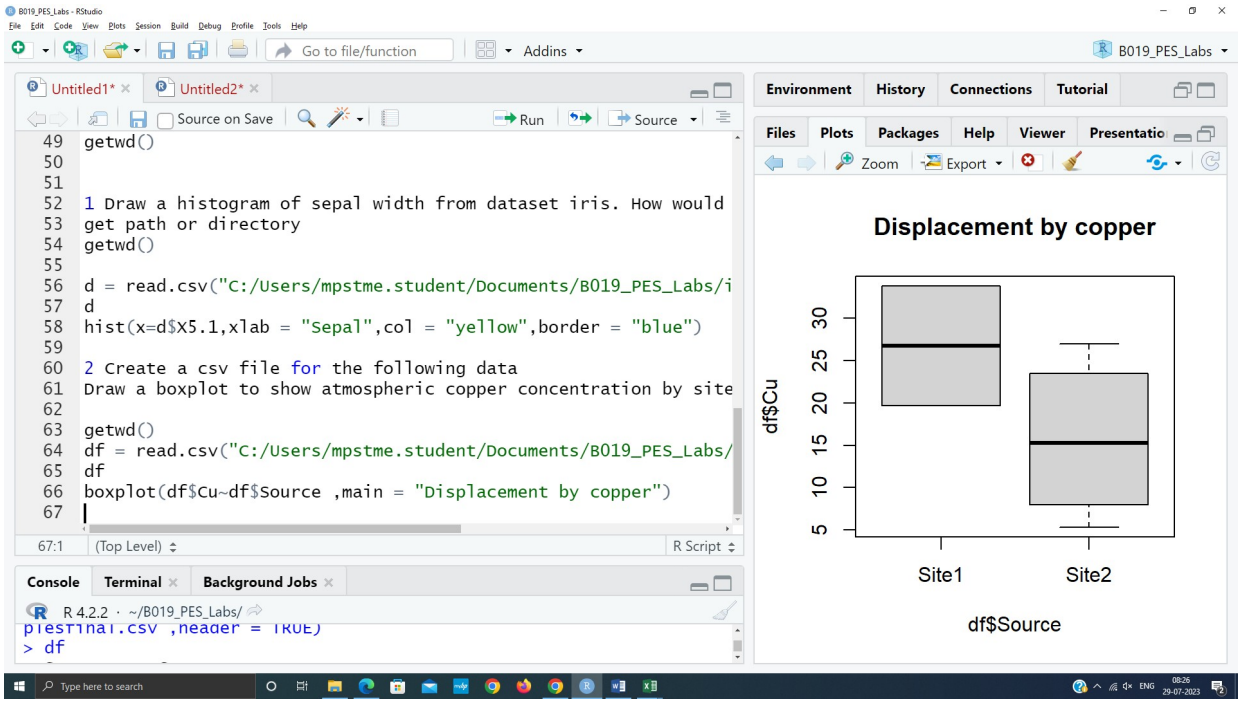
2 Create a csv file for the following data

```
##   Source     Cu
## 1  Site1 19.700
## 2  Site2 10.643
## 3  Site1 33.792
## 4  Site2  5.353
## 5  Site2 19.890
## 6  Site2 26.966
```

Draw a boxplot to show atmospheric copper concentration by sites.

```
getwd()
df =
read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/samplesfinal.csv",header = TRUE)
df
boxplot(df$Cu~df$Source ,main = "Displacement by copper")
```


getwd()
1 Draw a histogram of sepal width from dataset iris. How would
get path or directory
getwd()
d = read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/i
d
hist(x=d$X5.1,xlab = "Sepal",col = "yellow",border = "blue")
2 Create a csv file for the following data
Draw a boxplot to show atmospheric copper concentration by site
getwd()
df = read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/
df
boxplot(df$Cu~df$Source ,main = "Displacement by copper")
Displacement by copper
df$Cu
Site1
Site2
df$Source


getwd()
1 Draw a histogram of sepal width from dataset iris. How would
get path or directory
getwd()
d = read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/i
d
hist(x=d$X5.1,xlab = "Sepal",col = "yellow",border = "blue")
2 Create a csv file for the following data
Draw a boxplot to show atmospheric copper concentration by site
getwd()
df = read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/
df
boxplot(df$Cu~df$Source ,main = "Displacement by copper")
Displacement by copper
df$Cu
Site1
Site2
df$Source

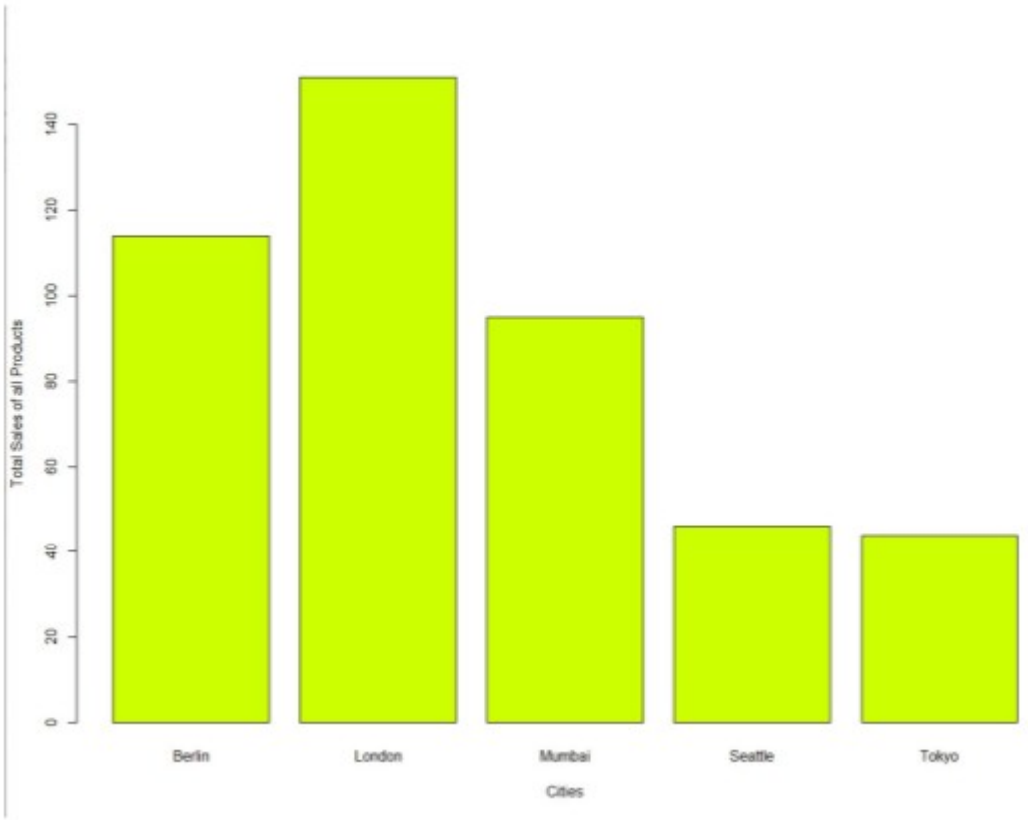
3 Create a csv file for the following data

```
##      City ProductA ProductB ProductC
## 1 Seattle       23       11       12
## 2  London       89        6       56
## 3   Tokyo       24        7       13
## 4  Berlin       36       34       44
## 5  Mumbai        3       78       14
```

Draw a bar graph to show total sales across different cities.

```
dat3 =
read.csv(file="C:/Users/mpstme.student/Documents/B019_PES_Labs
/city.csv",header=TRUE,sep=",",dec=".")

dat3

barplot(dat3$ProductA+dat3$ProductB+dat3$ProductC~dat3$City,xlab="Ci
ties",ylab="Total Sales of all

Products",col="#CCFF00")
```

Output:

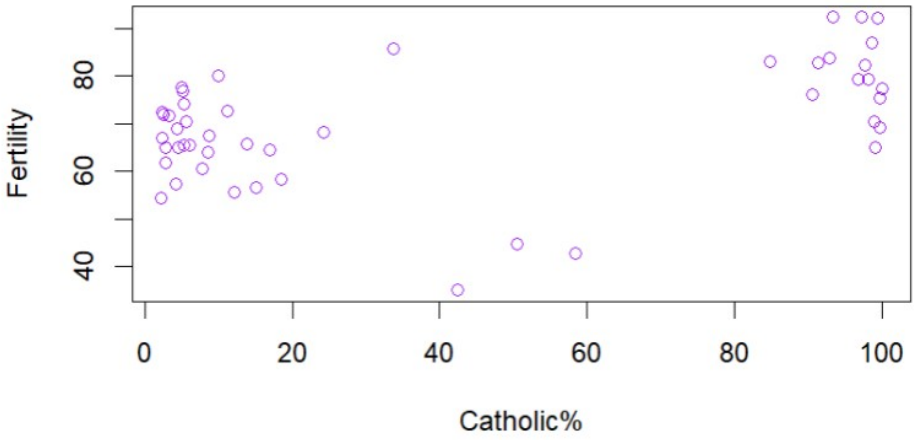


```
df =
read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/city.csv",
header = TRUE)

df

barplot(df$Sales,names.arg=df$City,xlab="sales",ylab="city",col="
blue",main="R evenue chart",border="red")
```

4 The dataset swiss contains a standardized fertility measure and various socioeconomic indicators for each of 47 French-speaking provinces of Switzerland in about 1888.

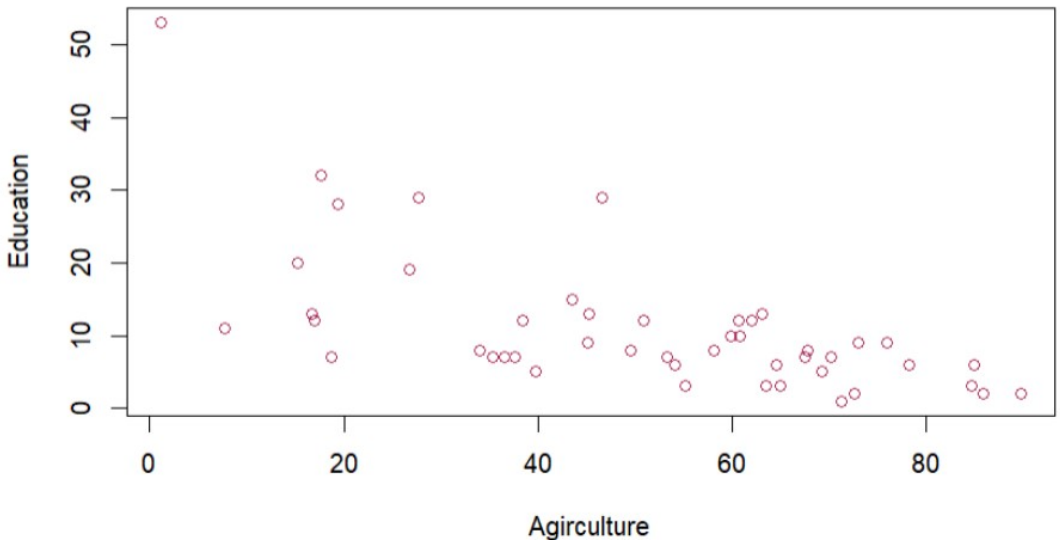
```
getwd()

df=read.csv("C:/Users/mpstme.student/Documents/B019_PES_Labs/
swiss.csv",header = TRUE)

df

plot(x = df$Catholic, y = df$Fertility,xlab = "Catholic",

ylab = "Fertility",main="scatterplot")

plot(x = df$Education, y = df$Agriculture,xlab = "Education",

ylab = "Agriculture",main="scatterplot")
```

a. Draw a scatterplot of Fertility against %Catholic. Which kind of areas have the lowest fertility rates?



.

b. Discuss the relationship between the variables Education and Agriculture.



5 Write a R program to change the first level of a factor with another level of a given factor.

```
x=c('a','b','c','d','a','b','d')
v=factor(x)
v
new_x=factor(v,levels = c('d','a','b','c'))
new_x
```

```
> x=c('a','b','c','d','a','b','d')
> v=factor(x)
> v
[1] a b c d a b d
Levels: a b c d
> new_x=factor(v,levels = c('d','a','b','c'))
> new_x
[1] a b c d a b d
Levels: d a b c
```

6. Write a R program to create an ordered factor from data of minimum 20 elements consisting of the names of months.

```
x=c('jan','feb','march','april','may','june','july','august','sept','jan','nov','dec',
    'jan','may','oct','jan','june','jan','feb','feb','feb','march','march')
v=factor(x)
v
table(v)
ordered.x=factor(x,levels = c('jan','feb','march','april','may','june','july','august'
                              'sept','oct','nov','dec'),ordered = T)
ordered.x
```

```
> x=c('jan','feb','march','april','may','june','july','august','sept','jan','nov','dec',
+     'jan','may','oct','jan','june','jan','feb','feb','feb','march','march')
> v=factor(x)
> v
 [1] jan    feb    march  april  may    june   july   august sept   jan    nov
[12] dec    jan    may    oct    jan    june   jan    feb    feb    feb    march
[23] march
Levels: april august dec feb jan july june march may nov oct sept
> table(v)
v
 april august    dec    feb    jan   july   june  march    may    nov    oct   sept
     1      1      1      4      5      1      2      3      2      1      1      1
> ordered.x=factor(x,levels = c('jan','feb','march','april','may','june','july','augus
t',
+                               'sept','oct','nov','dec'),ordered = T)
> ordered.x
 [1] jan    feb    march  april  may    june   july   august sept   jan    nov
[12] dec    jan    may    oct    jan    june   jan    feb    feb    feb    march
[23] march
12 Levels: jan < feb < march < april < may < june < july < august < sept < ... < dec
```