Project course "Introduction to Data Analysis"

Do we still need gyms?

Dataset description and variables:

A sample of 1000 people was tested with three new medicines to improve physical features

Variables:

Gender: Binary (M-F) Age: Age of people (numeric) Km Before: Running Kilometers before taking any medicine (Km) Kg Before: Weight of people before taking any medicine (Kg) Time Before: Running time before taking any medicine (Minutes) Medicine 1: Binary (Yes-No) Medicine 2: Binary (Yes-No) Medicine 3: Binary (Yes-No) Km After: Running Kilometers after taking any medicine (Km) Kg After: Weight of people after taking any medicine (Kg) Time After: Running time after taking any medicine (Minutes) SideEffects: Binary (Y-N)

Dataset link :

https://www.kaggle.com/datasets/saralattarulo/do-we-still-need-gyms (https://www.kaggle.com/datasets/saralattarulo/do-we-still-need-gyms)

kbasalim

We need some packages to download

In [ ]:
```
#Install packages ————————————————-
install.packages("RCurl")
install.packages("Hmisc")
install.packages("RColorBrewer")
install.packages("ggplot2")
install.packages("ggcorrplot")
install.packages("corrplot")
install.packages("ggcorrplot")
```

We requisition the libraries so that we can use the functions :

In [15]:
```
#Libraries ----------------------
library(skimr)
library(Hmisc)
library(RColorBrewer)
library(ggplot2)
library(dplyr)
library(ggcorrplot)
library(corrplot)
library(ggcorrplot)
```

```
Warning message:
"package 'skimr' was built under R version 3.6.3"Warning message:
"package 'Hmisc' was built under R version 3.6.3"Loading required package: lattice
Loading required package: survival
Warning message:
"package 'survival' was built under R version 3.6.3"Loading required package: Formula
Warning message:
"package 'Formula' was built under R version 3.6.3"Loading required package: ggplot2

Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, units

Warning message:
"package 'dplyr' was built under R version 3.6.3"
Attaching package: 'dplyr'

The following objects are masked from 'package:Hmisc':

    src, summarize

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
Warning message:
"package 'ggcorrplot' was built under R version 3.6.3"corrplot 0.92 loaded
```

Import the data :

In [4]:
```
#set the working directory
setwd("C:/Users/dell/Desktop/دورة R 2022/ال دورة مشروعR")
# Reading data and call it "df"
df <- read.csv("survey.csv.csv",sep = ";")
```

Discovering the dataset:

Dimensions :

In [5]:
```
# Know the dimensions of the data (row, column)
dim(df)
```

1000 · 12

look at the first and last six rows:

In [6]:
```
# Get the first sex rows
head(df)
```

A data.frame: 6 × 12

	Gender	Age	KmBefore	KgBefore	TimeBefore	Medicine1	Medicine2	Medicine3	KmAfter	KgAfter	TimeAfter	SideEffects
	<fct>	<int>	<dbl>	<dbl>	<dbl>	<fct>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<fct>
1	F	32	4.06	74.7	41.2	No	No	No	4.37	91.8	61.1	N
2	M	37	3.96	76.3	43.9	Yes	Yes	No	3.09	89.6	69.7	N
3	M	43	3.80	91.7	47.9	Yes	No	No	6.26	92.7	49.8	N
4	F	26	5.17	75.4	59.6	No	No	No	5.81	89.1	60.9	N
5	F	36	3.72	77.0	54.9	No	Yes	Yes	7.80	91.7	60.7	Y
6	M	37	5.31	93.9	50.6	No	Yes	Yes	5.67	87.8	67.9	N

In [7]:
```
# Get the last sex rows
tail(df)
```

A data.frame: 6 × 12

	Gender	Age	KmBefore	KgBefore	TimeBefore	Medicine1	Medicine2	Medicine3	KmAfter	KgAfter	TimeAfter	SideEffects
	<fct>	<int>	<dbl>	<dbl>	<dbl>	<fct>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<fct>
995	M	36	4.46	69.7	28.7	Yes	No	Yes	3.28	88.2	61.8	N
996	M	39	5.69	77.5	39.3	No	No	Yes	5.41	85.8	61.0	Y
997	F	35	4.57	93.2	50.2	Yes	Yes	No	5.86	96.6	64.4	N
998	F	36	2.72	72.3	53.8	No	Yes	No	2.63	92.8	68.9	Y
999	M	32	3.92	98.7	56.2	No	Yes	No	6.31	86.0	62.3	N
1000	M	32	3.85	79.3	52.6	No	Yes	Yes	5.21	89.1	64.4	N

Column Names:

used to rename and replace the column names of the data frame in R.

In [8]:
```
#Column Names: to check if it is need modification :
colnames(df)
```

'Gender' · 'Age' · 'KmBefore' · 'KgBefore' · 'TimeBefore' · 'Medicine1' · 'Medicine2' · 'Medicine3' · 'KmAfter' · 'KgAfter' · 'TimeAfter' · 'SideEffects'

Modify values name

The gsub() function in R can be used to replace all occurrences of certain text within a string in R.

In [9]:
```
#modify values name in the Gender column (F to Female ,M to male):
df$Gender = gsub("F","Female",df$Gender)
df$Gender = gsub("M","Male",df$Gender)
```

In [10]:
```
#modify values name in the SideEffects column (N to NO ,Y to Yes):
df$SideEffects = gsub("Y","Yes",df$SideEffects)
df$SideEffects = gsub("N","No",df$SideEffects)
```

Check if a data contains null value and if there is duplicate in DataFrame :

In [11]:
```
#checks if any of columns in the data have null values - should print False :
any((is.na(df)))
```

FALSE

In [12]:
```
#print number of duplicates in the data :
sum(duplicated(df))
```

0

Description :

skim() is an alternative to summary(),quickly providing a broad overview of a data frame.

In [16]:
```
#discovering the dataset more (type of variables and some statistic measures)
skim(df)
```

```
-- Data Summary ------------------------
                           Values
Name                       df
Number of rows             1000
Number of columns          12
_____________________
Column type frequency:
  character                2
  factor                   3
  numeric                  7
_____________________
Group variables            None

-- Variable type: character -----------------------------------------------------
# A tibble: 2 x 8
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
* <chr>             <int>         <dbl> <int> <int> <int>    <int>      <int>
1 Gender                0             1     4     6     0        2          0
2 SideEffects           0             1     2     3     0        2          0

-- Variable type: factor --------------------------------------------------------
# A tibble: 3 x 6
  skim_variable n_missing complete_rate ordered n_unique top_counts
* <chr>             <int>         <dbl> <lgl>      <int> <chr>
1 Medicine1             0             1 FALSE          2 No: 732, Yes: 268
2 Medicine2             0             1 FALSE          2 Yes: 622, No: 378
3 Medicine3             0             1 FALSE          2 No: 551, Yes: 449

-- Variable type: numeric -------------------------------------------------------
# A tibble: 7 x 11
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75
* <chr>             <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Age                   0             1 34.8  5.01  19    31    35    38
2 KmBefore              0             1  4.02 0.774  1.72  3.5   4.01  4.57
3 KgBefore              0             1 80.6  9.30  51.6  74.5  80.7  86.7
4 TimeBefore            0             1 44.9  9.99   8    38.3  44.6  51.6
5 KmAfter               0             1  5.00 1.46   0.51  3.96  5.01  5.97
6 KgAfter               0             1 90.1  2.95  81.3  88.2  90.1  92.1
```

```
7 TimeAfter             0             1 59.8  5.03  42.7  56.3  59.9  63
    p100 hist
*  <dbl> <chr>
1  50    <U+2581><U+2585><U+2587><U+2585><U+2581>
2   6.28 <U+2581><U+2585><U+2587><U+2585><U+2581>
3 109.   <U+2581><U+2583><U+2587><U+2585><U+2581>
4  74.2  <U+2581><U+2582><U+2587><U+2586><U+2581>
5  10.4  <U+2581><U+2586><U+2587><U+2583><U+2581>
6  99.5  <U+2581><U+2585><U+2587><U+2583><U+2581>
7  76.3  <U+2581><U+2583><U+2587><U+2583><U+2581>
```

In [17]:
```
#and by - Statistic Measures:
describe(df)
```

```
df

 12  Variables      1000  Observations
---------------------------------------------------------------------------
Gender
       n  missing distinct
    1000        0        2

Value      Female   Male
Frequency     447    553
Proportion  0.447  0.553
---------------------------------------------------------------------------
Age
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1000        0       31    0.996    34.79    5.674       27       28
     .25      .50      .75      .90      .95
      31       35       38       41       43

lowest : 19 20 21 23 24, highest: 46 47 48 49 50
---------------------------------------------------------------------------
KmBefore
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1000        0      312        1    4.021   0.8764    2.710    3.030
     .25      .50      .75      .90      .95
   3.500    4.015    4.570    5.011    5.290

lowest : 1.72 1.85 1.97 1.99 2.00, highest: 5.94 5.95 6.12 6.27 6.28
---------------------------------------------------------------------------
KgBefore
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1000        0      358        1    80.57    10.48    64.70    68.60
     .25      .50      .75      .90      .95
   74.50    80.70    86.73    92.30    95.20

lowest :  51.6  54.6  55.3  56.0  56.1, highest: 105.1 106.4 107.0 107.3 109.2
---------------------------------------------------------------------------
TimeBefore
       n  missing distinct     Info     Mean      Gmd      .05      .10
```

```
    1000        0      382        1    44.87    11.29    28.60    32.09
     .25      .50      .75      .90      .95
   38.30    44.55    51.62    57.71    61.90

lowest :  8.0 13.8 16.0 19.1 19.7, highest: 70.6 70.9 71.0 72.8 74.2
---------------------------------------------------------------------------
Medicine1
       n  missing distinct
    1000        0        2

Value         No   Yes
Frequency    732   268
Proportion 0.732 0.268
---------------------------------------------------------------------------
Medicine2
       n  missing distinct
    1000        0        2

Value         No   Yes
Frequency    378   622
Proportion 0.378 0.622
---------------------------------------------------------------------------
Medicine3
       n  missing distinct
    1000        0        2

Value         No   Yes
Frequency    551   449
Proportion 0.551 0.449
---------------------------------------------------------------------------
KmAfter
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1000        0      475        1    4.997    1.649    2.690    3.119
     .25      .50      .75      .90      .95
   3.960    5.010    5.970    6.861    7.333

lowest :  0.51  0.95  1.20  1.24  1.44, highest:  8.95  9.40  9.53  9.76 10.35
---------------------------------------------------------------------------
KgAfter
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1000        0      144        1    90.13    3.342    85.30    86.10
     .25      .50      .75      .90      .95
```

```
   88.18    90.10    92.10    93.90    95.10

lowest : 81.3 82.0 82.5 82.8 82.9, highest: 97.6 97.7 98.3 98.5 99.5
---------------------------------------------------------------------------
TimeAfter
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1000        0      227        1    59.76    5.676    51.40    53.40
     .25      .50      .75      .90      .95
   56.27    59.90    63.00    66.11    68.10

lowest : 42.7 42.8 43.3 45.9 46.6, highest: 73.1 73.2 74.2 74.5 76.3
---------------------------------------------------------------------------
SideEffects
       n  missing distinct
    1000        0        2

Value         No   Yes
Frequency    817   183
Proportion 0.817 0.183
---------------------------------------------------------------------------
```

Data analysis :

Now the data has been checked, and the data is ready for analysis

Knowing the numbers of men and women in the study :

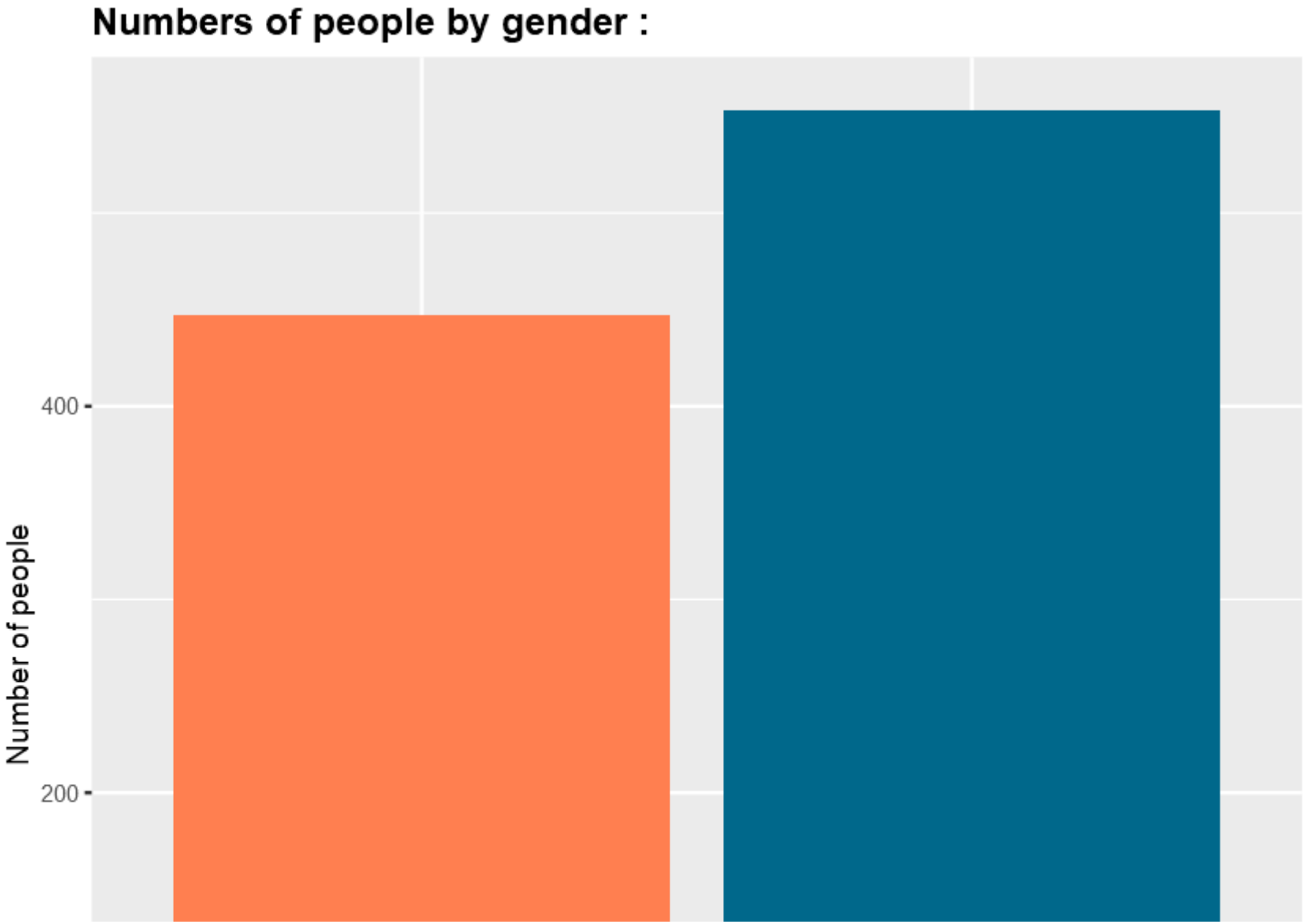
In [18]:

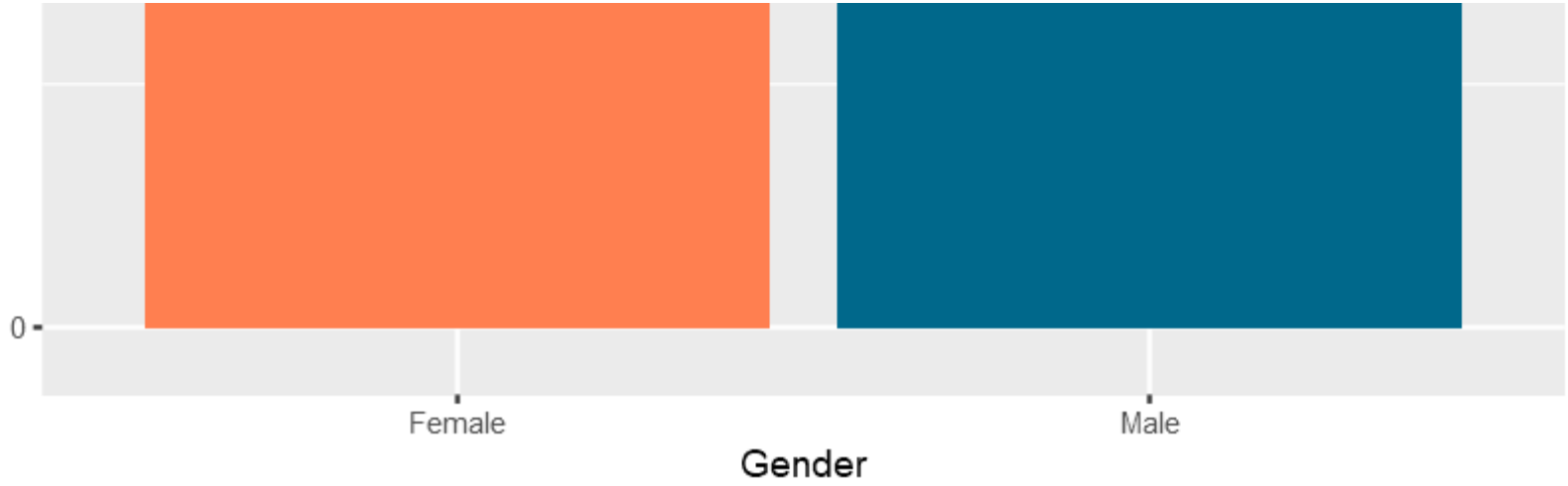
```
#The dataset contains numbers of people by gender :
df_Gender<-df%>%
  group_by(Gender)%>%
  summarize(number_of_people= n())
df_Gender
```

A tibble: 2 × 2

Gender	number_of_people
<chr>	<int>
Female	447
Male	553

In [19]:
```
#Basic barplot for Gender :
ggplot(data=df_Gender, aes(x=Gender, y=number_of_people)) +
  geom_bar(stat="identity",fill=c("coral","deepskyblue4"))+
  labs(y = "Number of people")+
  ggtitle("Numbers of people by gender :")+
  theme(plot.title = element_text(face="bold", size=14))
```





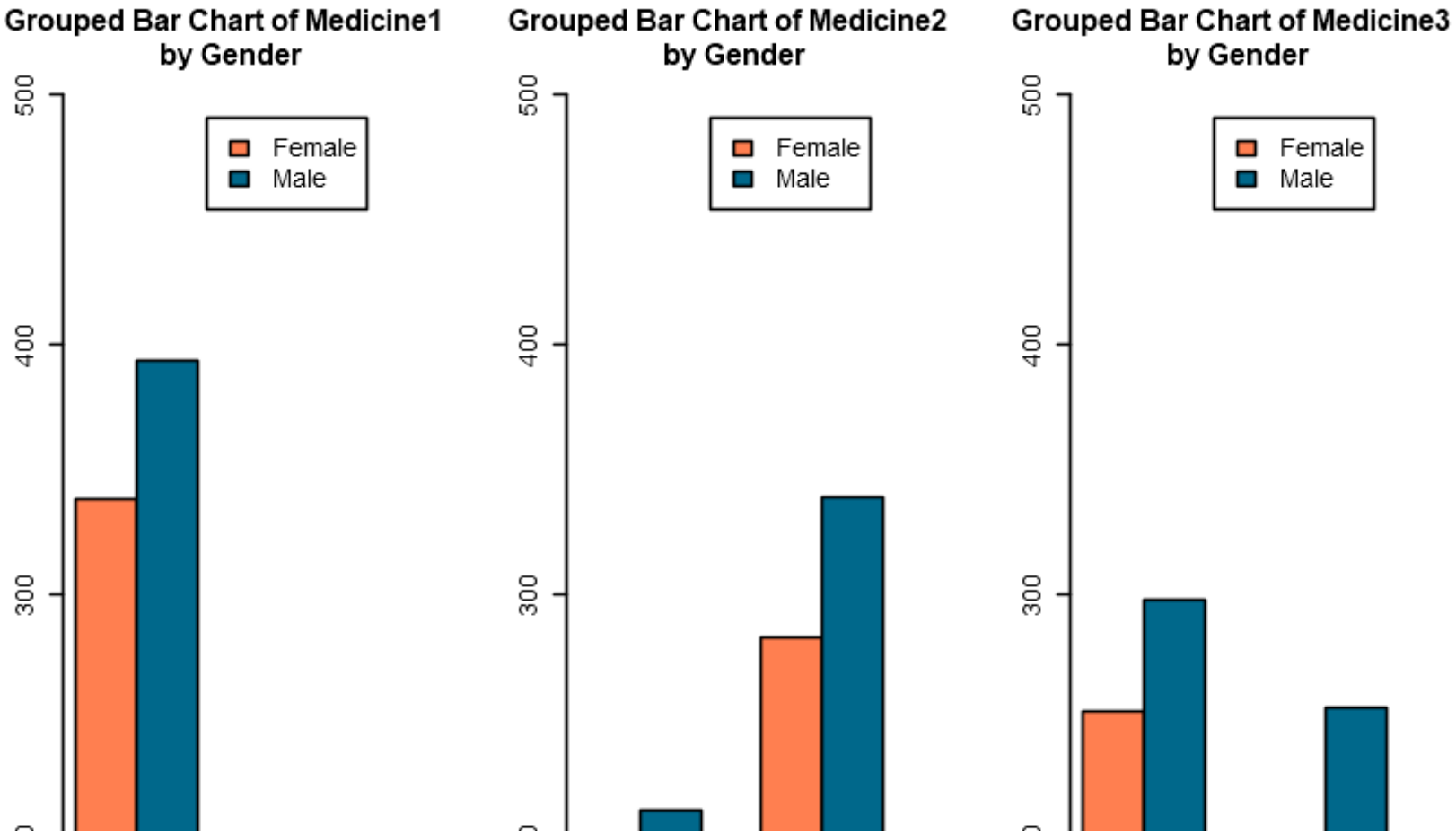
Type of Medicines by Gender:

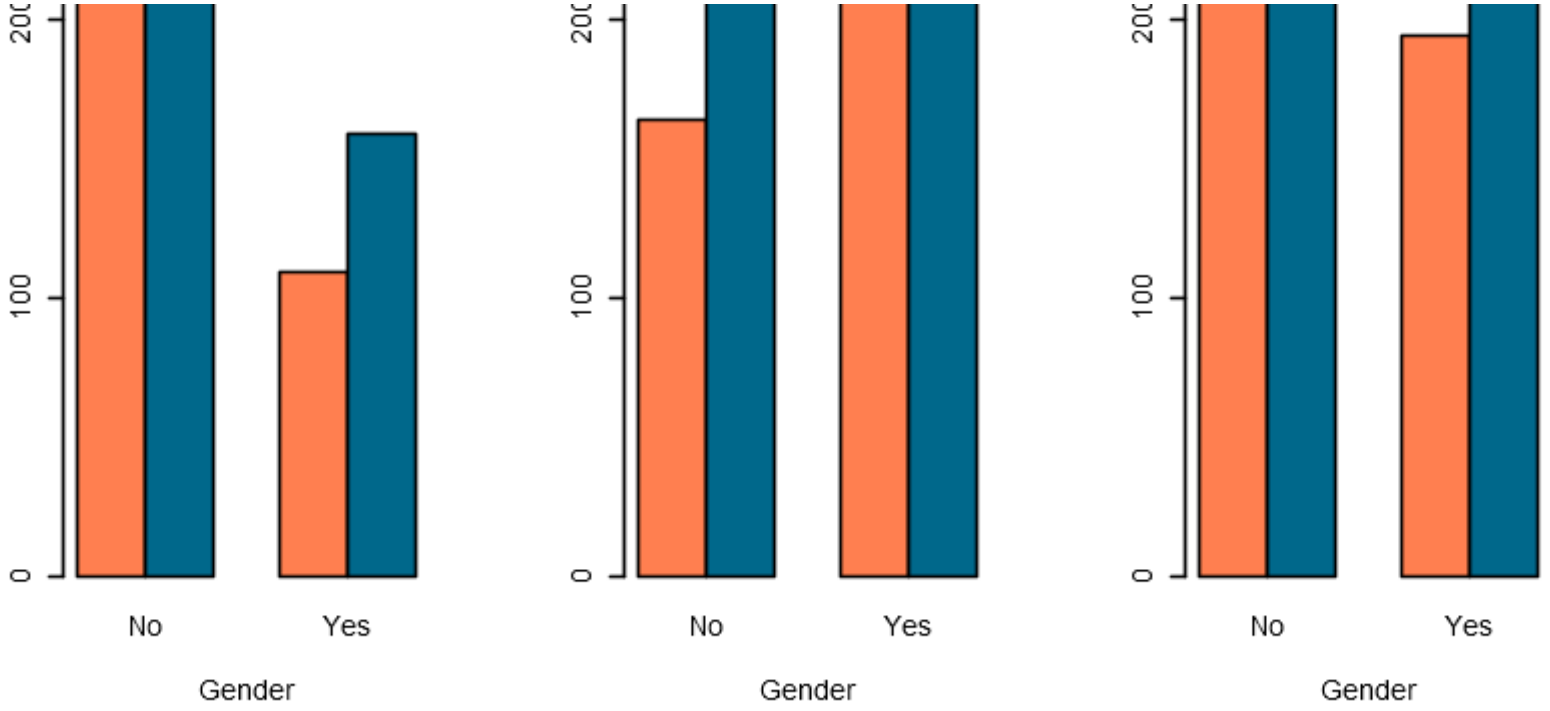
Which of the three medicines is most used according to gender?

In [20]:
```
#Number of people using each of three types of medication by gender:
#Medicine1 with other Medicine
counts1 <- table(df$Gender, df$Medicine1)
#Medicine2 with other Medicine
counts2 <- table(df$Gender, df$Medicine2)
#Medicine3 with other Medicine
counts3 <- table(df$Gender, df$Medicine3)
```

In [21]:

```
# create grouped bar chart
## Create a 1 x 3 plotting matrix
par(mfrow=c(1,3))
#Medicine1 with other Medicine:
barplot(counts1, main="Grouped Bar Chart of Medicine1 \n by Gender",
        xlab="Gender", legend = T,ylim = c(0,500),col=c("coral","deepskyblue4"), beside=TRUE)
#Medicine2 with other Medicine:
barplot(counts2, main="Grouped Bar Chart of Medicine2 \n by Gender",
        xlab="Gender", legend = T,ylim = c(0,500),col=c("coral","deepskyblue4"), beside=TRUE)
#Medicine3 with other Medicine:
barplot(counts3, main="Grouped Bar Chart of Medicine3 \n by Gender",
        xlab="Gender", legend = T,ylim = c(0,500),col=c("coral","deepskyblue4"), beside=TRUE)
```



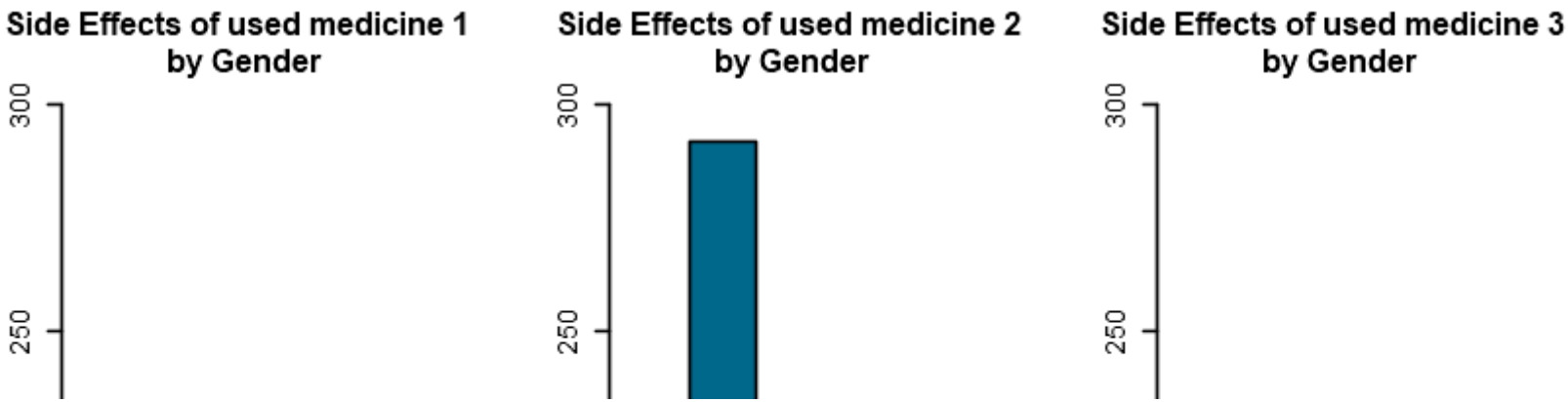


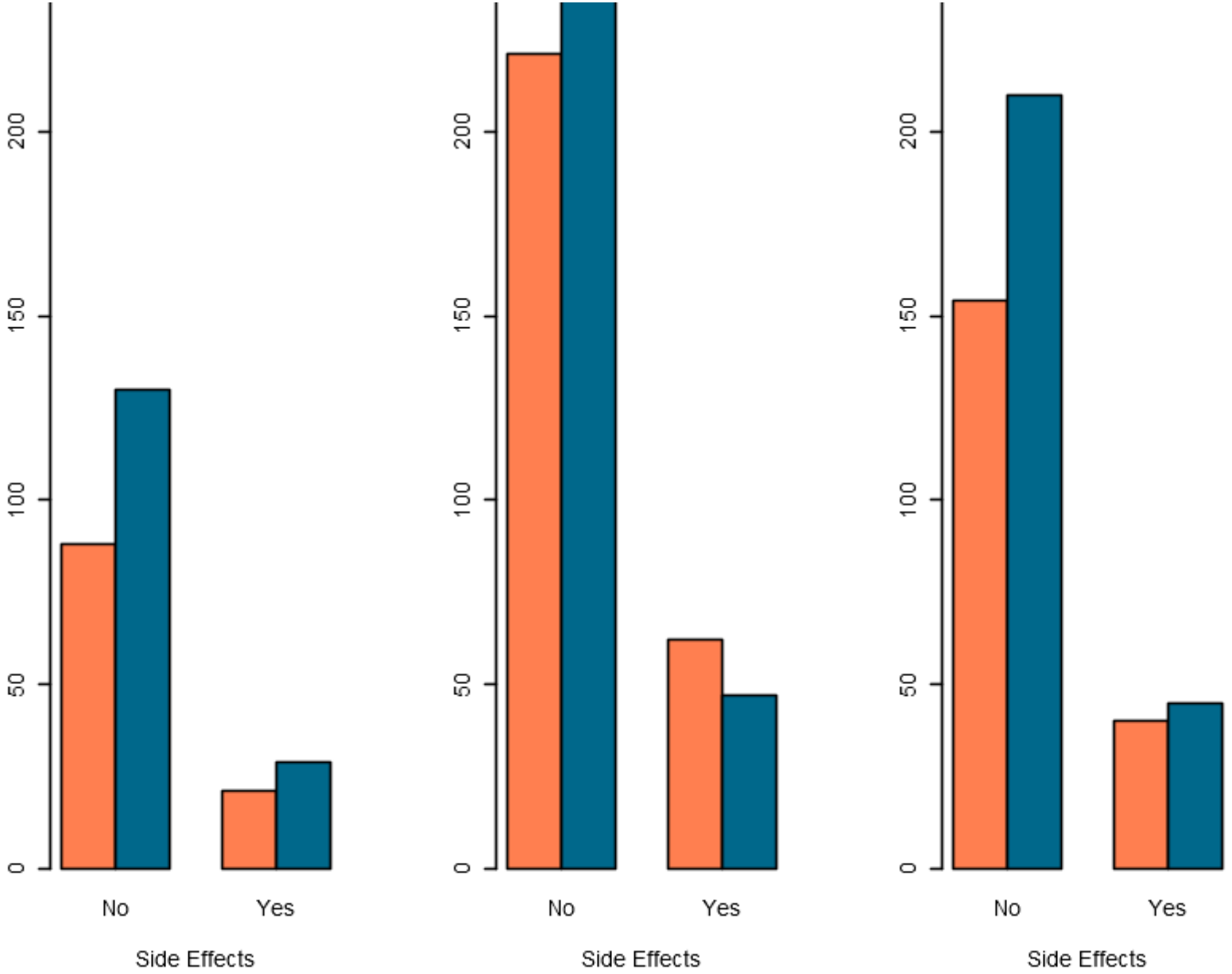
I noticed that medicine 2 is more used, I can't reveal the reason, is it the morest selling or the cheapest or is the effect useful in losing weight, I will try during the analysis of the data to reveal that if I can .

Which of the three medicines has the most Side Effects?

In [22]:

```
#Type of Medicines by Side Effects:
#Number of people using each of three types of medication by Side Effects:
df1<-df %>%
  filter(df$Medicine1 == "Yes")%>%
  select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter,Medicine2,Medicine3))


df2<-df %>%
    filter(df$Medicine2 == "Yes")%>%
    select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter,Medicine1,Medicine3))

df3<-df %>%
  filter(df$Medicine3 == "Yes")%>%
  select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter,Medicine1,Medicine2))

## Create a 1 x 3 plotting matrix
par(mfrow=c(1,3))

barplot(table(df1$Gender,df1$SideEffects),main="Side Effects of used medicine 1 \n by Gender",
          xlab="Side Effects",ylim = c(0,300),col=c("coral","deepskyblue4"), beside=TRUE)

barplot(table(df2$Gender,df2$SideEffects),main="Side Effects of used medicine 2\n by Gender",
          xlab="Side Effects",ylim = c(0,300),col=c("coral","deepskyblue4"), beside=TRUE)


barplot(table(df3$Gender,df3$SideEffects),main="Side Effects of used medicine 3 \n by Gender",
        xlab="Side Effects",ylim = c(0,300),col=c("coral","deepskyblue4"), beside=TRUE)
```





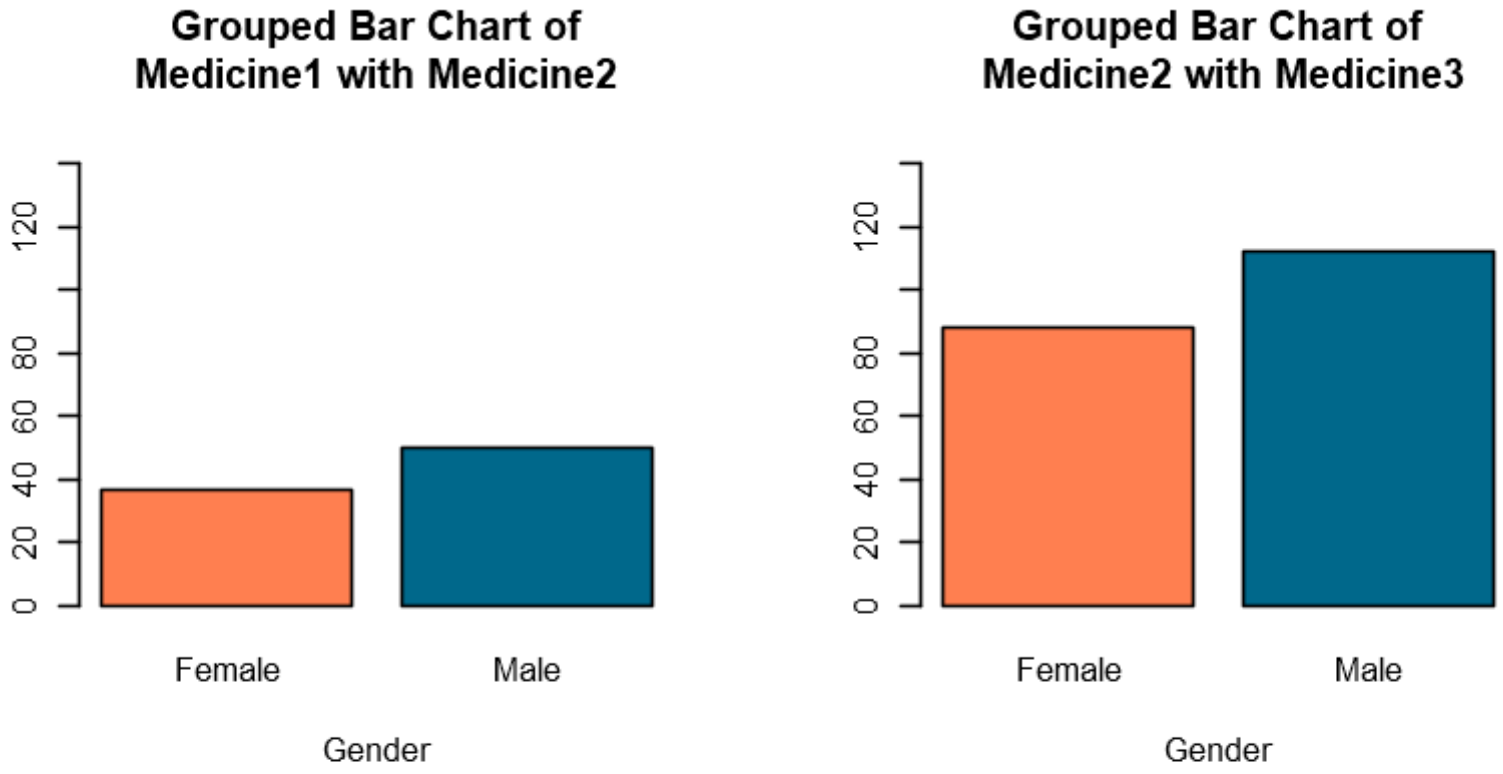
So now my focus will be on Medicine 2 because it's the most common

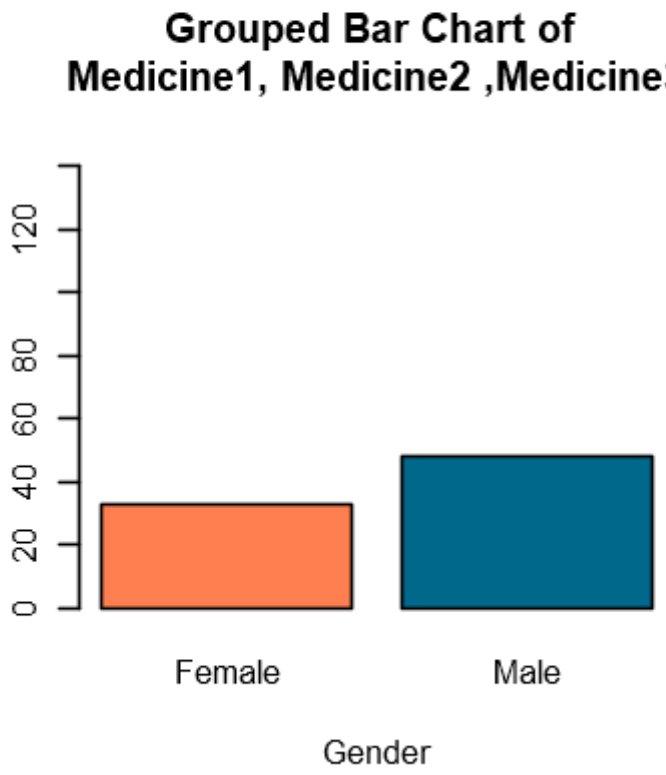
Medicine2

In [23]:

```
#I focused on the second  Medicine 2 because it is the most used from male and female:

#create of new data that contain: people whose used Medicine 1 with Medicine 2 just
df_M1_M2<- df %>%
  filter(df$Medicine1 == "Yes",df$Medicine2 == "Yes",df$Medicine3 == "No")%>%
  select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter))

#create of new data that contain Medicine : people whose used Medicine 2 with Medicine 3 just
df_M2_M3<- df %>%
  filter(df$Medicine1 == "No",df$Medicine2 == "Yes",df$Medicine3 == "Yes")%>%
  select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter))

#create of new data that contain Medicine :people whose used Medicine 1, Medicine2 and Medicine 3
df_M1_M2_M3<- df %>%
  filter(df$Medicine1 == "Yes",df$Medicine2 == "Yes",df$Medicine3 == "Yes")%>%
  select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter))

#create of new data that contain Medicine : people whose used Medicine 2 just
df_M2<- df %>%
  filter(df$Medicine1 == "No",df$Medicine2 == "Yes",df$Medicine3 == "No")%>%
  select(-c(Age,KmBefore,KgBefore,TimeBefore,KmAfter,KgAfter,TimeAfter))
```

In [24]:
```
## Create a 2 x 2 plotting matrix
par(mfrow=c(2,2))

#Now , determine which of the people who use medicine 2 highest
barplot(table(df_M1_M2$Gender),main="Grouped Bar Chart of\n Medicine1 with Medicine2 ",
        xlab="Gender",ylim = c(0,150),col=c("coral","deepskyblue4"), beside=TRUE)

barplot(table(df_M2_M3$Gender),main="Grouped Bar Chart of\n Medicine2 with Medicine3",
        xlab="Gender",ylim = c(0,150),col=c("coral","deepskyblue4"), beside=TRUE)

barplot(table(df_M1_M2_M3$Gender),main="Grouped Bar Chart of \n Medicine1, Medicine2 ,Medicine3 ",
        xlab="Gender",ylim = c(0,150),col=c("coral","deepskyblue4"), beside=TRUE)

barplot(table(df_M2$Gender),main="Grouped Bar Chart of Medicine2 just",
        xlab="Gender",ylim = c(0,150),col=c("coral","deepskyblue4"), beside=TRUE)
```





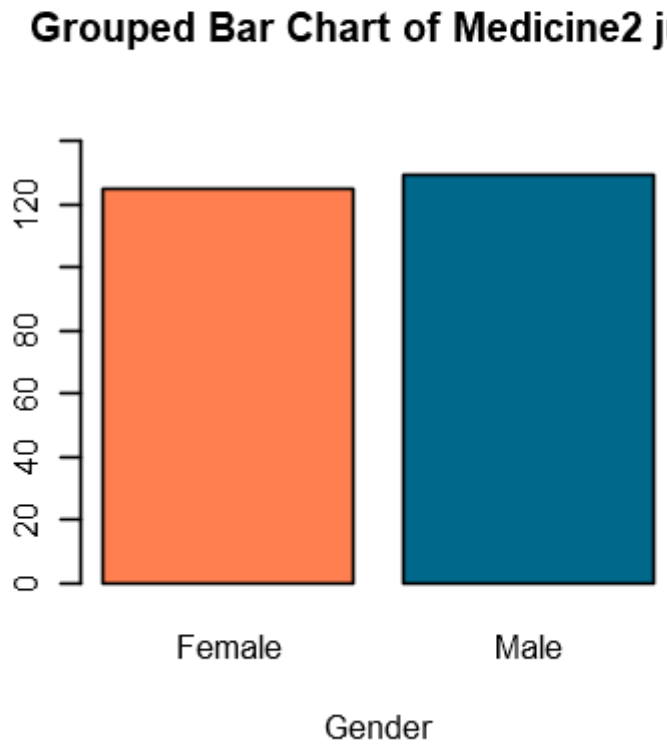
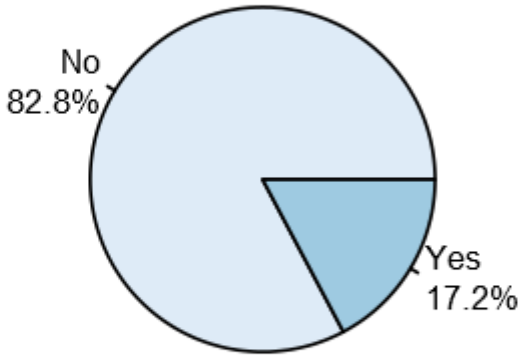


table for SideEffects with Medicine 2

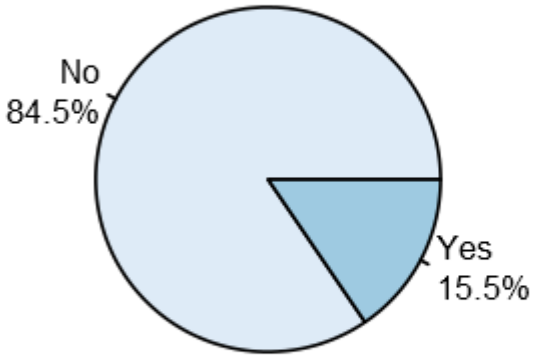
In [25]:

```
## Create a 2 x 2 plotting matrix
par(mfrow=c(2,2))

# table for SideEffects : people whose used Medicine 1 with Medicine 2 just
fs_M1_M2 <- table(df_M1_M2$SideEffects)
# calculate a percenage of each level
pct1 <- round(100*prop.table(fs_M1_M2 ), 1)
# to add labels for the plot
lbls1 <- paste(names(fs_M1_M2 ), "\n", pct1, "%", sep = "")
# plot pie chart
pie(fs_M1_M2, col = brewer.pal(3,"Blues"), labels = lbls1, main=("SideEffects M1 & M2"))


# table for SideEffects: people whose used Medicine 2 with Medicine 3 just
fs_M2_M3 <- table(df_M2_M3$SideEffects)
# calculate a percenage of each level
pct2 <- round(100*prop.table(fs_M2_M3 ), 1)
# to add labels for the plot
lbls2 <- paste(names(fs_M2_M3 ), "\n", pct2, "%", sep = "")
# plot pie chart
pie(fs_M2_M3, col = brewer.pal(3,"Blues"), labels = lbls2, main=("SideEffects M2 & M3"))


# table for SideEffects: people whose used Medicine 1, Medicine2 and Medicine 3
fs_M1_M2_M3 <- table(df_M1_M2_M3$SideEffects)
# calculate a percenage of each level
pct3 <- round(100*prop.table(fs_M1_M2_M3  ), 1)
# to add labels for the plot
lbls3 <- paste(names(fs_M1_M2_M3  ), "\n", pct3, "%", sep = "")
# plot pie chart
pie(fs_M1_M2_M3 , col = brewer.pal(3,"Blues"), labels = lbls3, main=("SideEffects M1, M2 and M3"))


# table for SideEffects : people whose used Medicine 2 just
fs_M2 <- table(df_M2$SideEffects)
# calculate a percenage of each level
pct4 <- round(100*prop.table(fs_M2  ), 1)
# to add labels for the plot
lbls4 <- paste(names(fs_M2  ), "\n", pct4, "%", sep = "")
# plot pie chart
pie(fs_M2 , col = brewer.pal(3,"Blues"), labels = lbls4, main=("SideEffects M2"))
```

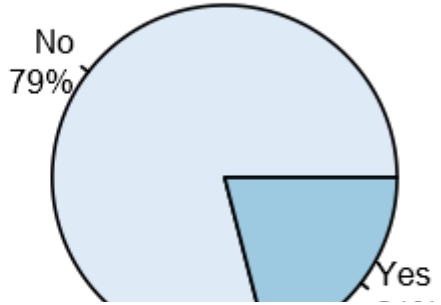
SideEffects M1 & M2



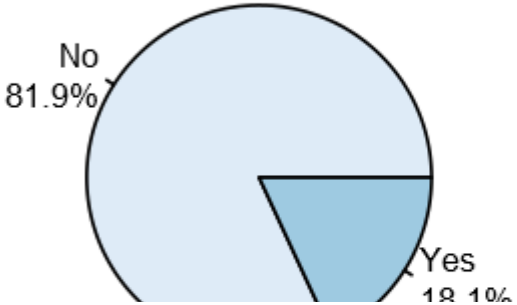
SideEffects M2 & M3

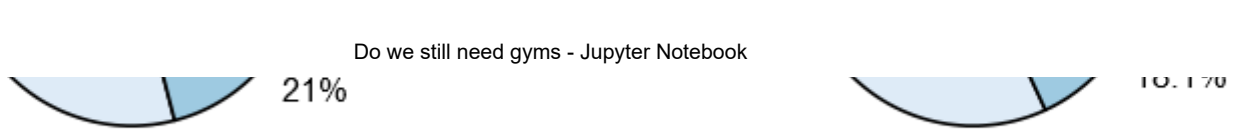


SideEffects M1, M2 and M3



SideEffects M2





All users Medicine 2 have weak side effects, which can be due to health reasons in the person

Data of Medicine2 just:

I will use the data only Medicine 2 in the future can go deeper and study all cases

In [26]:

```
df_M2_just<- df %>%
  filter(df$Medicine1 == "No",df$Medicine2 == "Yes",df$Medicine3 == "No")
#discovering the dataset:
dim(df_M2_just)
skim(df_M2_just)
#checks if any of columns in the data have null values - should print False :
any((is.na(df_M2_just)))
#print number of duplicates in the dataset :
sum(duplicated(df_M2_just))
# Get the first sex rows
head(df_M2_just)
```

254 · 12

```
-- Data Summary ------------------------
                           Values
Name                       df_M2_just
Number of rows             254
Number of columns          12
_________________________
Column type frequency:
  character                2
  factor                   3
  numeric                  7
_________________________
Group variables            None

-- Variable type: character -----------------------------------------------------
# A tibble: 2 x 8
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
* <chr>             <int>         <dbl> <int> <int> <int>    <int>      <int>
1 Gender                0             1     4     6     0        2          0
2 SideEffects           0             1     2     3     0        2          0

-- Variable type: factor --------------------------------------------------------
# A tibble: 3 x 6
  skim_variable n_missing complete_rate ordered n_unique top_counts
* <chr>             <int>         <dbl> <lgl>      <int> <chr>
1 Medicine1             0             1 FALSE          1 No: 254, Yes: 0
2 Medicine2             0             1 FALSE          1 Yes: 254, No: 0
```

```
3 Medicine3             0             1 FALSE          1 No: 254, Yes: 0

-- Variable type: numeric -------------------------------------------------------
# A tibble: 7 x 11
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75
* <chr>             <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Age                   0             1 35.4  5.12  20    32    35.5  39
2 KmBefore              0             1  4.04 0.797  1.72  3.53  4.08  4.63
3 KgBefore              0             1 80.7  9.45  54.6  74.5  81.1  87.3
4 TimeBefore            0             1 44.1  9.84  19.1  37.9  43.4  50.3
5 KmAfter               0             1  4.90 1.44   1.78  3.91  4.84  5.86
6 KgAfter               0             1 89.9  2.92  82.9  87.9  89.9  91.7
7 TimeAfter             0             1 59.4  5.17  43.3  55.7  59.5  62.8
    p100 hist
*  <dbl> <chr>
1  49    <U+2581><U+2583><U+2587><U+2586><U+2581>
2   6.12 <U+2581><U+2583><U+2587><U+2585><U+2581>
3 107    <U+2581><U+2585><U+2587><U+2585><U+2581>
4  70.9  <U+2582><U+2586><U+2587><U+2585><U+2581>
5  10.4  <U+2583><U+2587><U+2586><U+2582><U+2581>
6  99.5  <U+2582><U+2586><U+2587><U+2582><U+2581>
7  70.9  <U+2581><U+2583><U+2587><U+2587><U+2583>
```

FALSE

0

A data.frame: 6 × 12

	Gender	Age	KmBefore	KgBefore	TimeBefore	Medicine1	Medicine2	Medicine3	KmAfter	KgAfter	TimeAfter	SideEffects
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<fct>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<chr>
1	Male	32	5.12	94.1	53.6	No	Yes	No	7.07	88.2	55.5	No
2	Female	44	3.93	87.6	36.8	No	Yes	No	4.30	92.4	59.2	No
3	Female	31	3.73	71.8	61.2	No	Yes	No	3.79	94.8	60.5	No
4	Male	35	2.17	97.8	40.7	No	Yes	No	3.59	94.3	64.1	No
5	Female	34	2.96	78.9	49.3	No	Yes	No	4.85	89.4	54.1	No
6	Male	43	5.21	84.5	34.1	No	Yes	No	4.88	94.3	60.0	No

In [27]:
```
#delete the unused columns (Medicine1+Medicine2+Medicine3)--
#dataest "df_M " : The data is for the people who used " Medicine 2 "  :
df_M2_just <- df_M2_just %>%
  select(-c(`Medicine1`, `Medicine3`,`Medicine2`))
dim(df_M2_just)
```

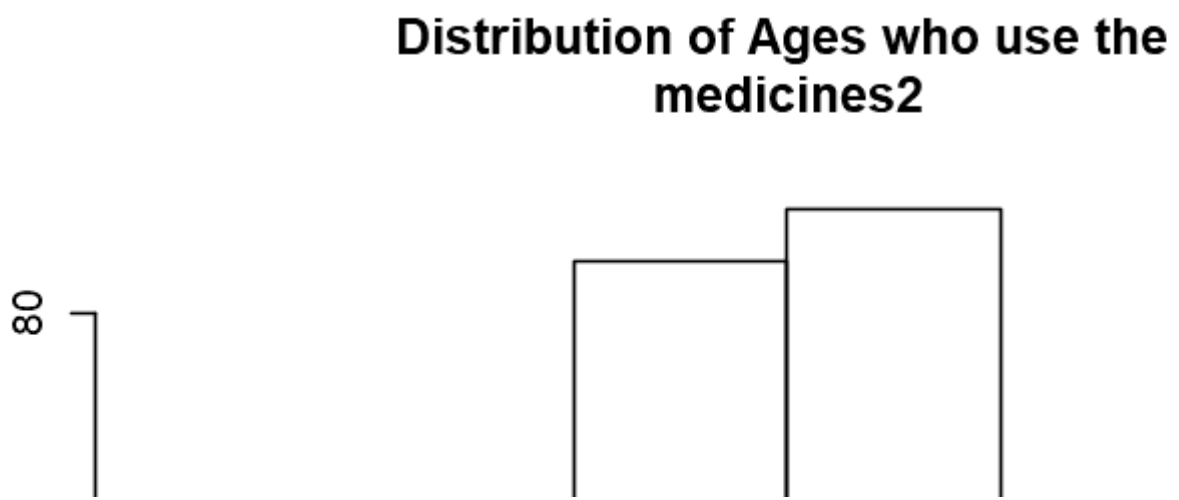
254 · 9

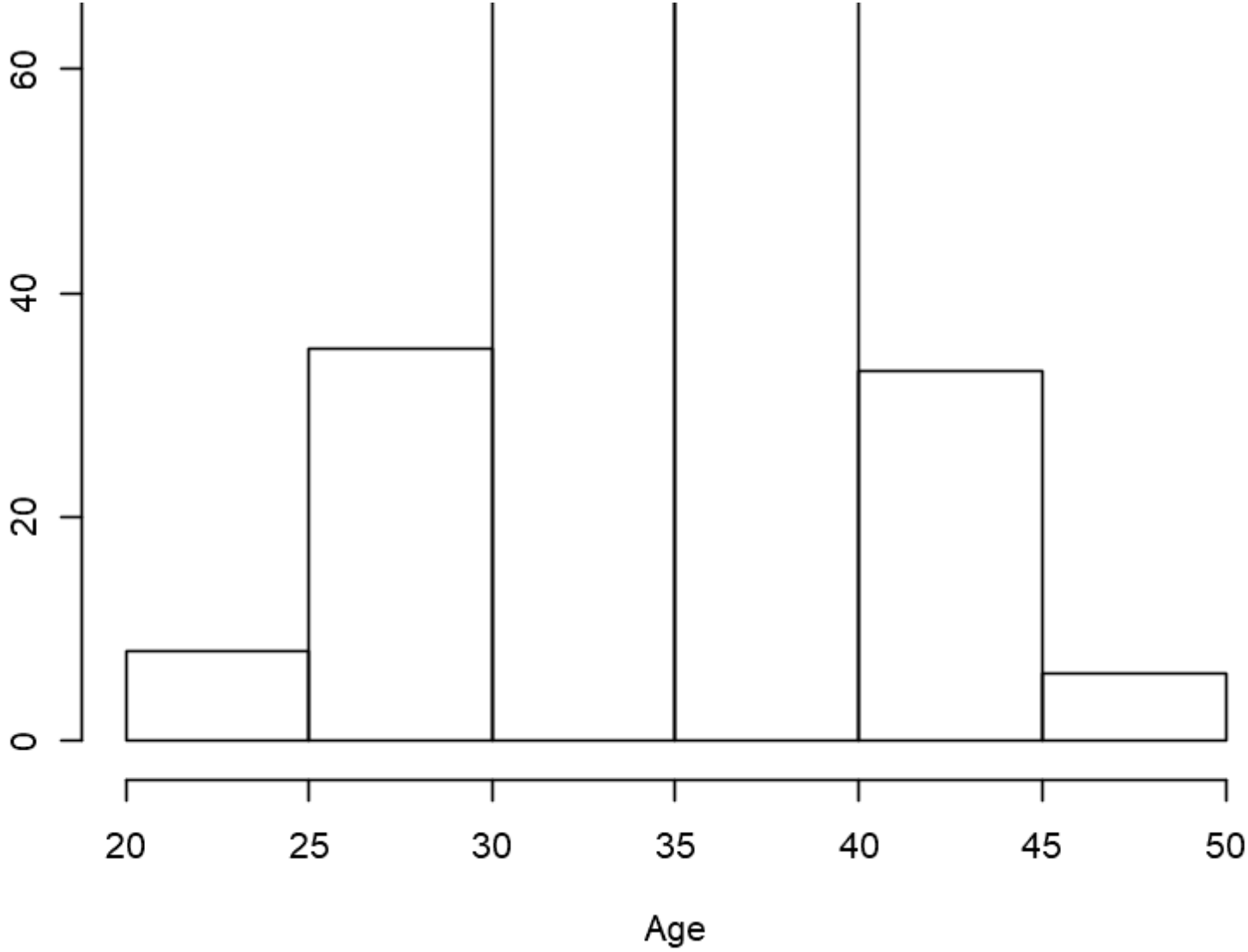
The distribution of age who use Medicine 2 :

In [28]:
```
# first, we should determine the breaks (intervals),
# and R will do that for us
breaks1 = hist(df_M2_just$Age,xlab="Age" , ylab="" , main="Distribution of Ages who use the \n medicines2 ")[[1]]
#Frequancy table :
# then assign the suitable interval for each value in the variable "age"
class1 <- cut(df_M2_just $Age, breaks1, right = F)
# then we follow the same steps as before (3rd method)
ff<-data.frame(table(class1))
mutate(ff, relFreq = round(prop.table(Freq),4)*100, cumFreq = cumsum(Freq), cumRelFreq = cumsum(relFreq))
```

A data.frame: 6 × 5

class1	Freq	relFreq	cumFreq	cumRelFreq
<fct>	<int>	<dbl>	<int>	<dbl>
[20,25)	4	1.57	4	1.57
[25,30)	29	11.42	33	12.99
[30,35)	74	29.13	107	42.12
[35,40)	91	35.83	198	77.95
[40,45)	47	18.50	245	96.45
[45,50)	9	3.54	254	99.99




0
20
40
60
20
25
30
35
40
45
50
Age

In [29]:

```
#Distribution of Male/Female Ages who use the medicines2 :
#hist for males and females separately:
par(mfrow=c(1,2))
#male:
df_M2_male<- df_M2_just %>%
  filter( df_M2_just$Gender == "Male")
head(df_M2_male)
dim(df_M2_male)
hist(df_M2_male$Age,xlab="Age", ylab="", main="Distribution of male Ages ")
#female:
df_M2_female<- df_M2_just %>%
  filter( df_M2_just$Gender == "Female")
head(df_M2_female)
dim(df_M2_female)
hist(df_M2_female$Age,xlab="Age", ylab="", main="Distribution of female Ages ")
```

A data.frame: 6 × 9

	Gender	Age	KmBefore	KgBefore	TimeBefore	KmAfter	KgAfter	TimeAfter	SideEffects
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Male	32	5.12	94.1	53.6	7.07	88.2	55.5	No
2	Male	35	2.17	97.8	40.7	3.59	94.3	64.1	No
3	Male	43	5.21	84.5	34.1	4.88	94.3	60.0	No
4	Male	42	3.55	80.6	39.5	3.36	89.3	55.7	No
5	Male	42	3.66	71.5	28.3	4.30	88.0	63.5	Yes
6	Male	39	4.08	79.5	38.5	4.39	89.8	58.5	No

129 · 9

correlation :

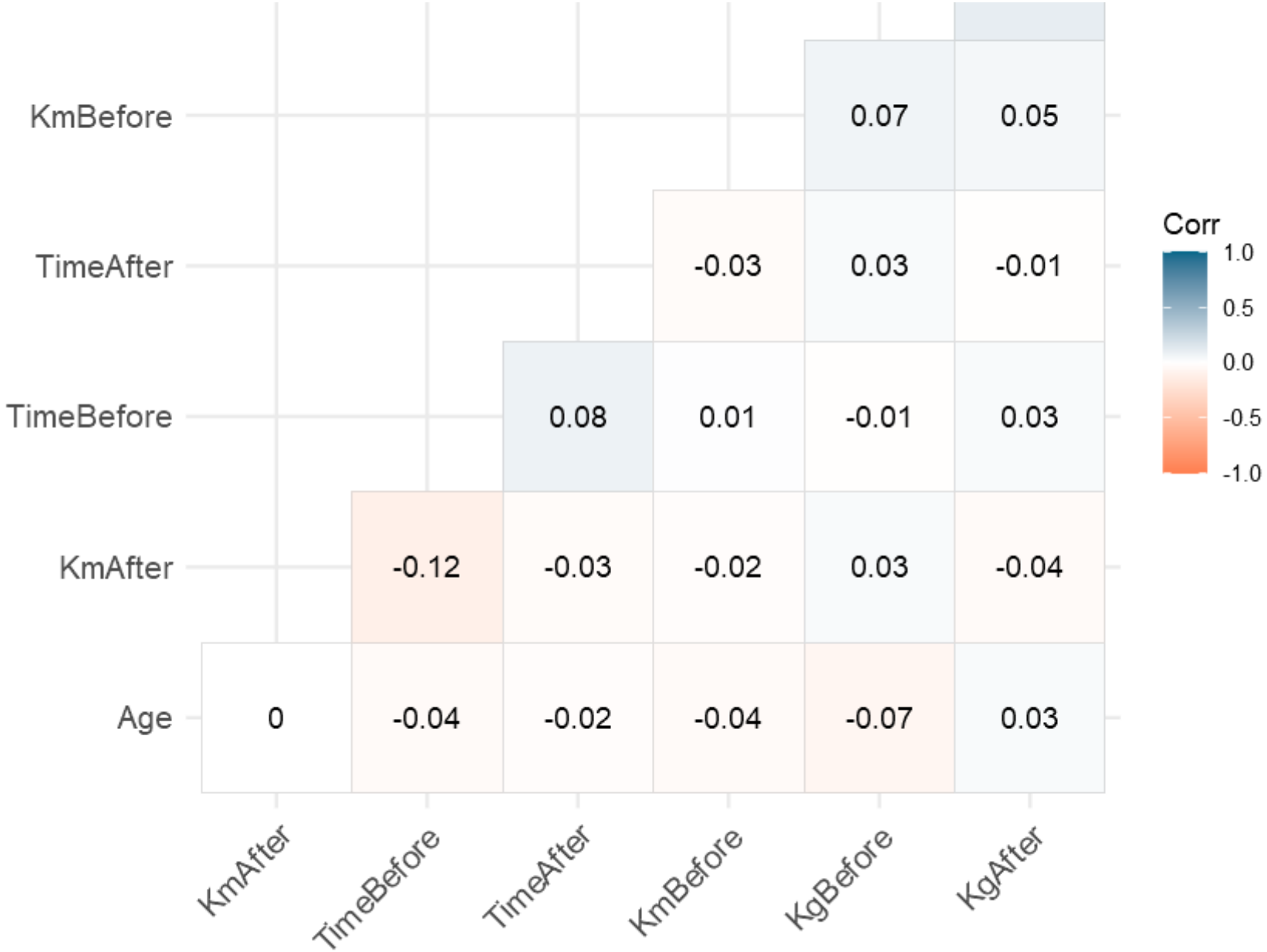
Is there a relationship between weights ,Running Kilometers and Running time before and after using medicines 2?

In [30]:
```
#Is there a relationship between weights ,Running Kilometers and Running time before and after using medicines 2?
df_M2_num<-df_M2_just %>%
  select(-c(Gender,SideEffects))

dim(df_M2_num)

res <- round(cor(df_M2_num) ,2)
res
ggcorrplot(res,
           hc.order = TRUE,
           type = "lower",
           lab = TRUE,
          colors = c("coral", "white", "deepskyblue4"))
```

254 · 7

A matrix: 7 × 7 of type dbl

	Age	KmBefore	KgBefore	TimeBefore	KmAfter	KgAfter	TimeAfter
Age	1.00	-0.04	-0.07	-0.04	0.00	0.03	-0.02
KmBefore	-0.04	1.00	0.07	0.01	-0.02	0.05	-0.03
KgBefore	-0.07	0.07	1.00	-0.01	0.03	0.11	0.03
TimeBefore	-0.04	0.01	-0.01	1.00	-0.12	0.03	0.08
KmAfter	0.00	-0.02	0.03	-0.12	1.00	-0.04	-0.03
KgAfter	0.03	0.05	0.11	0.03	-0.04	1.00	-0.01
TimeAfter	-0.02	-0.03	0.03	0.08	-0.03	-0.01	1.00


KmBefore
TimeAfter
TimeBefore
KmAfter
Age
0.07
0.05
-0.03
0.03
-0.01
0.08
0.01
-0.01
0.03
-0.12
-0.03
-0.02
0.03
-0.04
0
-0.04
-0.02
-0.04
-0.07
0.03
KmAfter
TimeBefore
TimeAfter
KmBefore
KgBefore
KgAfter
Corr
1.0
0.5
0.0
-0.5
-1.0

Find outlier

In [31]:

```
## Create a 3 x 2 plotting matrix
par(mfrow=c(3,2))

#boxplot of kg ,km ,time  by gender :(find outlier )

#boxplot of kg
boxplot( KmBefore~ Gender, data = df_M2_just,col = c("coral", "deepskyblue4"))$out
boxplot( KmAfter~ Gender, data = df_M2_just,col = c("coral", "deepskyblue4"))$out

#boxplot of km
boxplot( KgBefore~ Gender, data = df_M2_just,col = c("#FFE0B2", "#FFA726"))$out
boxplot( KgAfter~ Gender, data = df_M2_just,col = c("#FFE0B2", "#FFA726"))$out

#boxplot of time
boxplot( TimeBefore~ Gender, data = df_M2_just,col = c("#FFE0B2", "#FFA726"))$out
boxplot( TimeAfter~ Gender, data = df_M2_just,col = c("#FFE0B2", "#FFA726"))$out
```

1.72 · 1.85

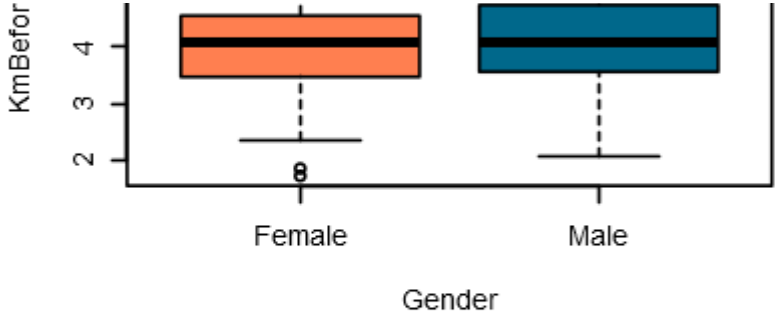
10.35

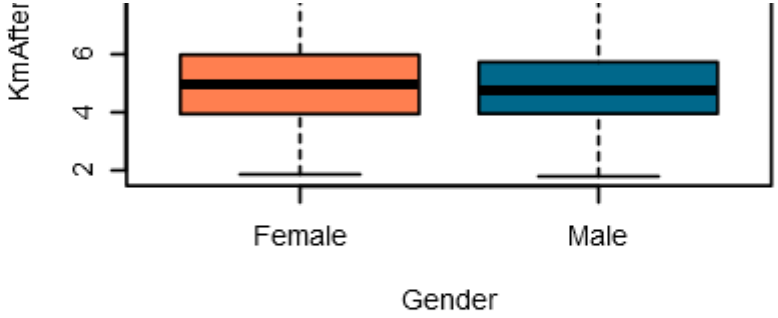
55.3 · 107 · 54.6 · 56.1

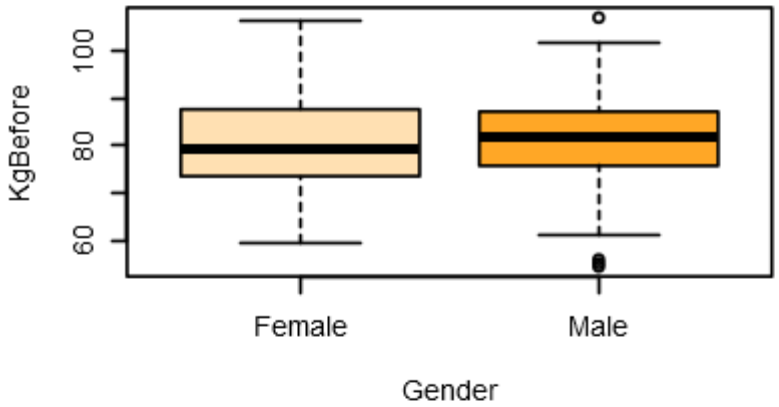
99.5 · 97.6

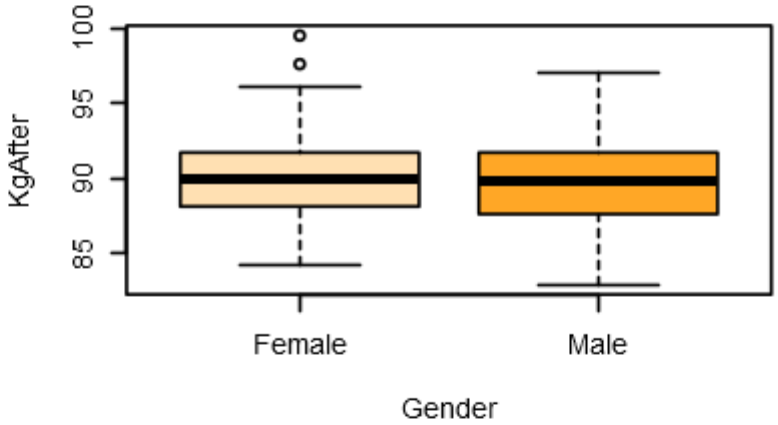
70.6 · 70.9

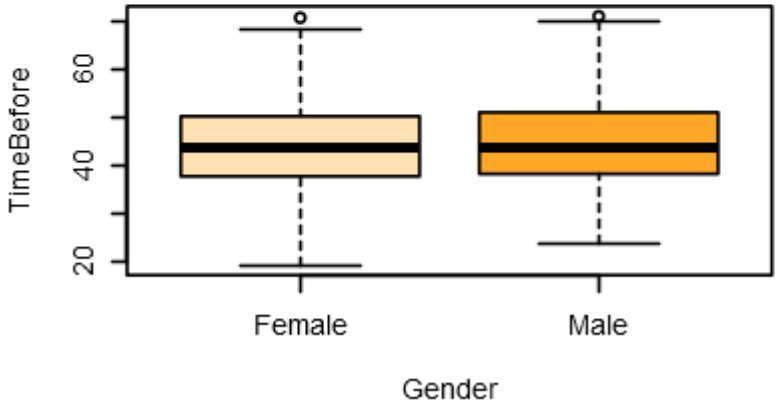
43.3

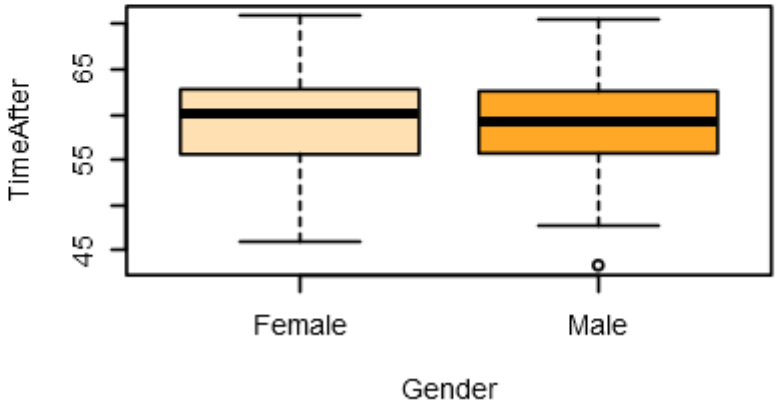

KmBefor
2
3
4
Female
Male
Gender


KmAfter
2
4
6
Female
Male
Gender


KgBefore
60
80
100
Female
Male
Gender


KgAfter
85
90
95
100
Female
Male
Gender


TimeBefore
20
40
60
Female
Male
Gender


TimeAfter
45
55
65
Female
Male
Gender

I think the values will not be affected much because they are not far from other values and also because of each person's health and physical condition

Test: the average of Weight before taking medicine 2 and after

Assumptions:

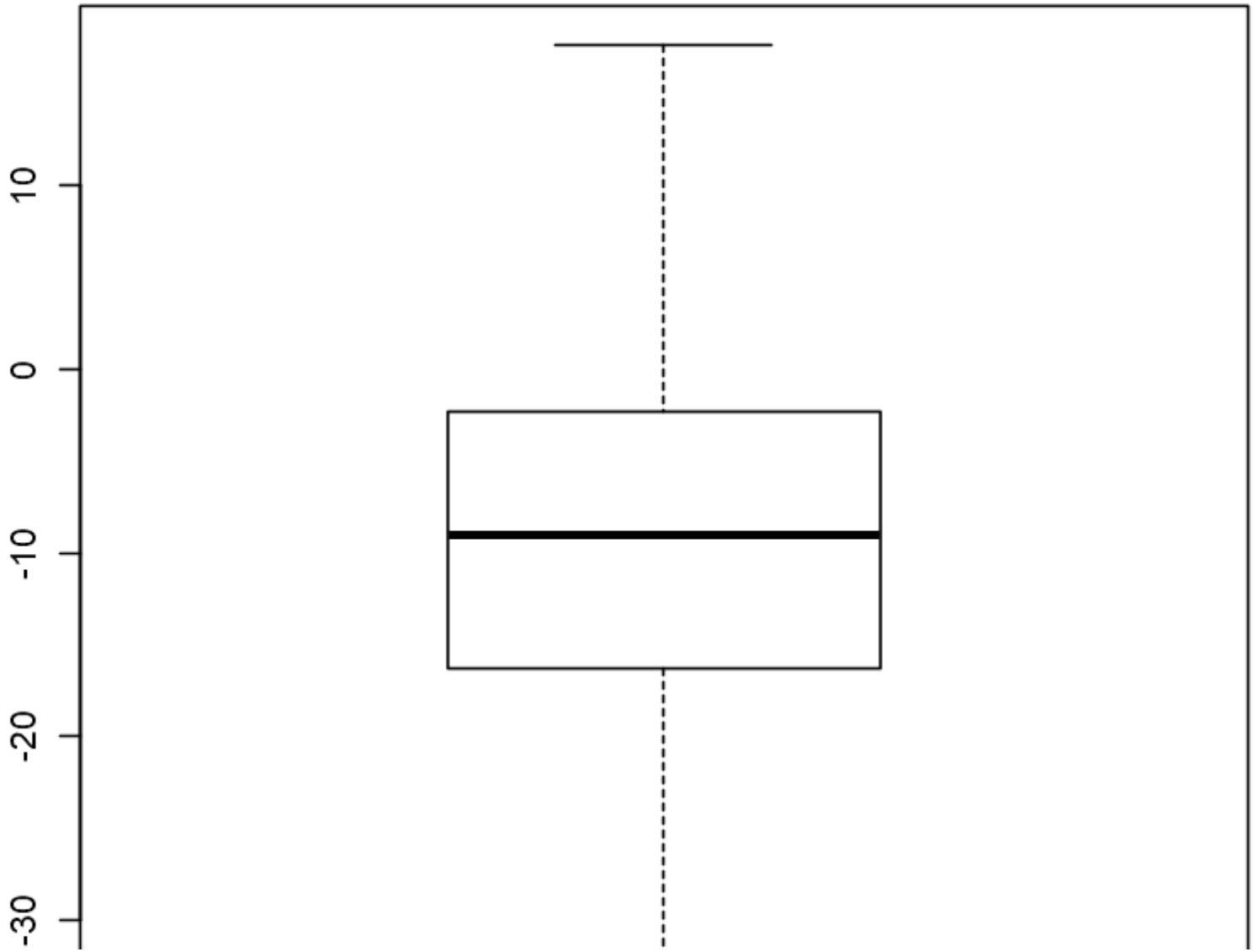
The paired samples t-test assume the following characteristics about the data: 1)the two groups are paired. 2)No significant outliers in the difference between the two related groups 3)Normality the difference of pairs follow a normal distribution. (Shapiro-Wilk test, can be used to check the of normality)

In [32]:
```
#difference:
difference=df_M2_just$KgBefore - df_M2_just$KgAfter
#Shapiro-Wilk test :
shapiro.test(difference)
```

```
        Shapiro-Wilk normality test

data:  difference
W = 0.9971, p-value = 0.928
```

H0: Data follow normal distribution vs. Ha: Data not follow normal distribution p > 0.05, We fail to reject the null hypothesis that our data are normally distributed

In [33]:
```
#boxplot:
boxplot(difference)
```



There are no outliers that affect the study

In [34]:
```
#paired t-test
t.test(df_M2_just$KgBefore,df_M2_just$KgAfter,paired=TRUE)
```

```
        Paired t-test

data:  df_M2_just$KgBefore and df_M2_just$KgAfter
t = -15.327, df = 253, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.394581  -8.027466
sample estimates:
mean of the differences
              -9.211024
```

H0: There is no difference in average weight before and after using medicine 2
vs. Ha: There is difference in average weight before and after using medicine 2 p-value is less than 0.05(significance level) , you can reject the null hypothesis.

In [35]:
```
mean(df_M2_just$KgBefore)
mean(df_M2_just$KgAfter)
```

80.6893700787402

89.9003937007874

logistic regression :

In [36]:
```
#modify values name in the SideEffects column (N to 0 ,Y to 1):
df_M2_just$SideEffects= gsub("Yes", 1 ,df_M2_just$SideEffects)
df_M2_just$SideEffects = gsub("No",0, df_M2_just$SideEffects)
```

In [37]:
```
head(df_M2_just)
tail(df_M2_just)
```

A data.frame: 6 × 9

	Gender	Age	KmBefore	KgBefore	TimeBefore	KmAfter	KgAfter	TimeAfter	SideEffects
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Male	32	5.12	94.1	53.6	7.07	88.2	55.5	0
2	Female	44	3.93	87.6	36.8	4.30	92.4	59.2	0
3	Female	31	3.73	71.8	61.2	3.79	94.8	60.5	0
4	Male	35	2.17	97.8	40.7	3.59	94.3	64.1	0
5	Female	34	2.96	78.9	49.3	4.85	89.4	54.1	0
6	Male	43	5.21	84.5	34.1	4.88	94.3	60.0	0

A data.frame: 6 × 9

	Gender	Age	KmBefore	KgBefore	TimeBefore	KmAfter	KgAfter	TimeAfter	SideEffects
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
249	Male	38	4.17	75.0	45.1	4.18	92.2	56.2	0
250	Female	39	4.78	79.2	55.6	5.76	86.8	52.9	0
251	Male	40	3.75	80.2	47.7	4.11	94.0	55.1	0
252	Male	28	5.02	97.1	48.5	4.21	86.5	60.2	0
253	Female	36	2.72	72.3	53.8	2.63	92.8	68.9	1
254	Male	32	3.92	98.7	56.2	6.31	86.0	62.3	0

In [38]:
```
str(df_M2_just)
```

```
'data.frame':   254 obs. of  9 variables:
 $ Gender     : chr  "Male" "Female" "Female" "Male" ...
 $ Age        : int  32 44 31 35 34 43 42 42 35 34 ...
 $ KmBefore   : num  5.12 3.93 3.73 2.17 2.96 5.21 3.55 3.66 5.65 4.56 ...
 $ KgBefore   : num  94.1 87.6 71.8 97.8 78.9 84.5 80.6 71.5 75.4 90.9 ...
 $ TimeBefore : num  53.6 36.8 61.2 40.7 49.3 34.1 39.5 28.3 41.4 37.6 ...
 $ KmAfter    : num  7.07 4.3 3.79 3.59 4.85 4.88 3.36 4.3 5.07 4.89 ...
 $ KgAfter    : num  88.2 92.4 94.8 94.3 89.4 94.3 89.3 88 90.9 90.7 ...
 $ TimeAfter  : num  55.5 59.2 60.5 64.1 54.1 60 55.7 63.5 68.5 65.6 ...
 $ SideEffects: chr  "0" "0" "0" "0" ...
```

The as.numeric in R is a built-in method that returns a numeric value

In [39]:
```
df_M2_just$SideEffects <- as.numeric(df_M2_just$SideEffects)
class(df_M2_just$SideEffects)
```

'numeric'

Do age and gender have a role in the side effects of people who use the Medicine 2?

In [40]:
```
#Do age and gender have a role in the side effects of people who use the Medicine 2?
model<-glm(SideEffects~Age+Gender ,family = binomial(link = "logit"),data=df_M2_just )
summary(model)
```

```
Call:
glm(formula = SideEffects ~ Age + Gender, family = binomial(link = "logit"),
    data = df_M2_just)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8881  -0.7367  -0.5214  -0.4530   2.1714

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.28305    1.18161  -1.932  0.05334 .
Age          0.03311    0.03261   1.015  0.30991
GenderMale  -0.93522    0.34517  -2.709  0.00674 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 240.31  on 253  degrees of freedom
Residual deviance: 231.73  on 251  degrees of freedom
AIC: 237.73

Number of Fisher Scoring iterations: 4
```

The model : ln(p/1-p) = -2.28305 + 0.03311 Age -0.93522 GenderMale

In this model, the increase in age for one year increases the exposure to side effects, and the increase is by 0.03311

As for the gender variable, the male participates in the decrease in exposure to side effects by a value 0.93522 if the patient is male

In [ ]:

In [ ]: