Laboratorio N° 01: Analisis Exploratorio de Datos con R

Objetivos

- Aplicar los conocimeintos asimilados sobre Analisis Exploratorio de Datos utilizando el lenguaje R

1. Carga el conjunto de datos

In [1]:

```
telecom <- read.csv("Telecomunicaciones.csv", sep = ";", header = TRUE, fileEncoding="l
atin1")
```

1. Revisar los seis primeros datos por defecto

In [2]:

```
head(telecom)
```

IdCliente	Género	Edad	Llamadas	Tiempo.enero	Tiempo.febrero	Monto	Espera
P50417214	Femenino	26	4	27.0	26.1	89.7	0.8
P50417215	Masculino	33	2	30.1	20.5	88.8	0.4
P50417216	Masculino	21	8	26.0	34.4	85.4	3.5
P50417217	Femenino	23	8	34.1	36.1	89.0	4.7
P50417218	Masculino	34	1	30.1	28.9	77.1	2.2
P50417219	Femenino	29	6	30.7	20.9	97.8	5.1

1. Verificar la estructura de los datos

In [3]:

```
str(telecom)
```

```
'data.frame':   120 obs. of  10 variables:
 $ IdCliente     : Factor w/ 120 levels "A80117234","A80117235",..: 71 72
73 74 75 76 77 78 79 80 ...
 $ Género        : Factor w/ 2 levels "Femenino","Masculino": 1 2 2 1 2 1
2 1 1 1 ...
 $ Edad          : int  26 33 21 23 34 29 21 40 25 38 ...
 $ Llamadas      : int  4 2 8 8 1 6 5 5 6 5 ...
 $ Tiempo.enero  : num  27 30.1 26 34.1 30.1 30.7 26.5 28.3 29.7 32.5 ...
 $ Tiempo.febrero: num  26.1 20.5 34.4 36.1 28.9 20.9 32 29.8 31.1 27.6
...
 $ Monto         : num  89.7 88.8 85.4 89 77.1 97.8 84 84.2 91.7 74.1 ...
 $ Espera        : num  0.8 0.4 3.5 4.7 2.2 5.1 2.8 5.8 4.2 0.8 ...
 $ Opinión       : Factor w/ 5 levels "Bueno","Excelente",..: 2 3 1 4 1 4
5 3 2 5 ...
 $ Empresa       : Factor w/ 4 levels "Bitele","Claros",..: 3 3 3 3 3 3 3
3 3 3 ...
```

1. Revisar un resumen de los datos

In [4]:

```
summary(telecom)
```

```
     IdCliente         Género        Edad          Llamadas
 A80117234:  1   Femenino :54   Min.   :20.00   Min.   : 0.000
 A80117235:  1   Masculino:66   1st Qu.:26.00   1st Qu.: 3.000
 A80117236:  1                  Median :31.00   Median : 5.000
 A80117237:  1                  Mean   :30.34   Mean   : 5.017
 A80117238:  1                  3rd Qu.:35.00   3rd Qu.: 7.000
 A80117239:  1                  Max.   :40.00   Max.   :13.000
 (Other)  :114
  Tiempo.enero   Tiempo.febrero      Monto            Espera
 Min.   :17.50   Min.   :17.50   Min.   : 74.10   Min.   : 0.200
 1st Qu.:31.70   1st Qu.:31.95   1st Qu.: 84.17   1st Qu.: 1.700
 Median :37.85   Median :37.60   Median : 90.70   Median : 3.650
 Mean   :38.24   Mean   :37.48   Mean   : 92.68   Mean   : 6.148
 3rd Qu.:43.62   3rd Qu.:42.25   3rd Qu.: 99.53   3rd Qu.: 7.000
 Max.   :62.20   Max.   :59.60   Max.   :119.10   Max.   :36.000

      Opinión     Empresa
 Bueno    :23   Bitele:30
 Excelente:19   Claros:30
 Muy Bueno:19   Entell:20
 Pésimo   :44   Movist:40
 Regular  :15
```

1. Revisar un resumen de los datos en columnas

In [8]:

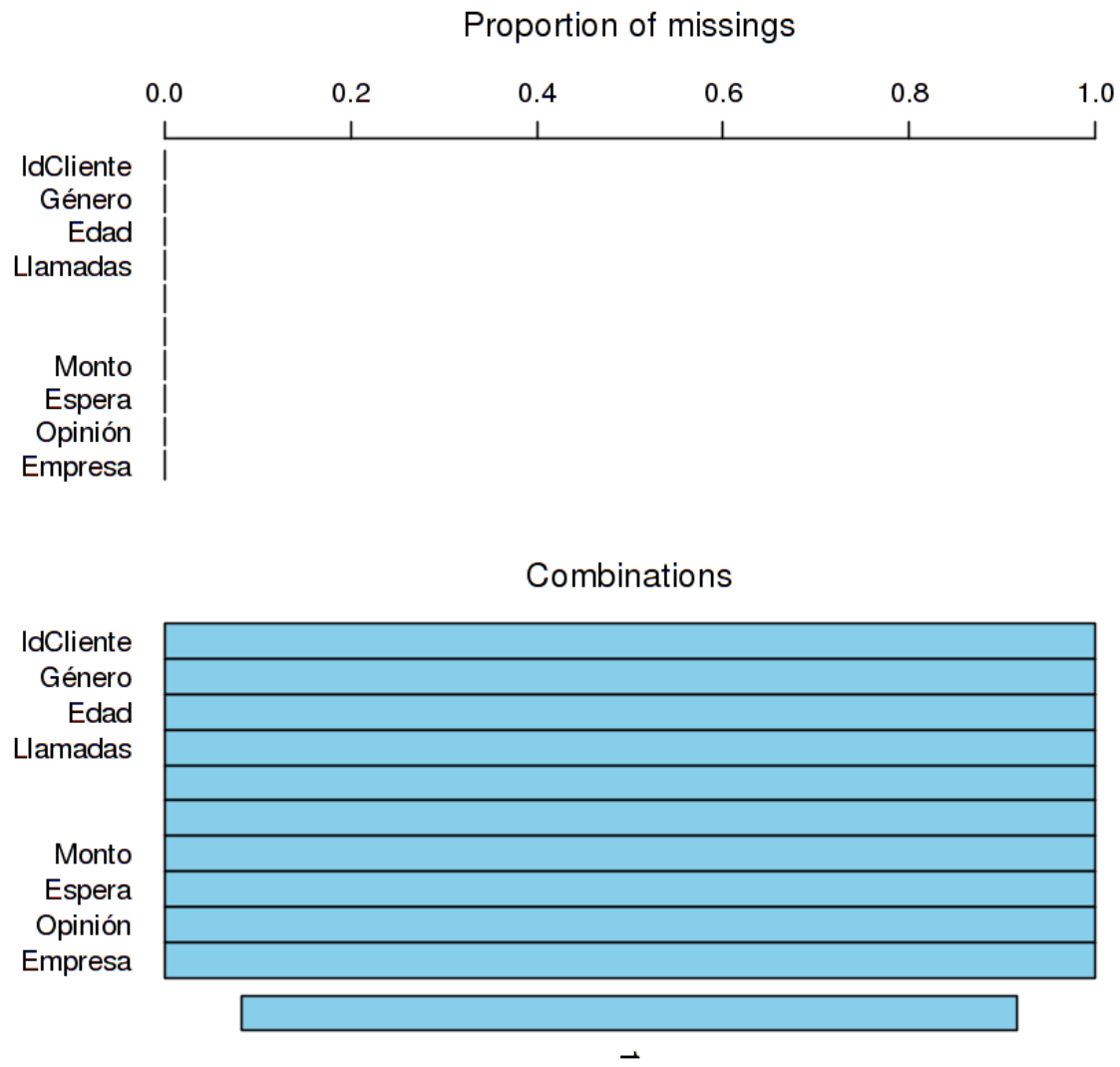
```
library(mlr)
summarizeColumns(telecom)
```

name	type	na	mean	disp	median	mad	min	max	nle
IdCliente	factor	0	NA	0.9916667	NA	NA	1.0	1.0	120
Género	factor	0	NA	0.4500000	NA	NA	54.0	66.0	2
Edad	integer	0	30.341667	5.7357728	31.00	5.93040	20.0	40.0	0
Llamadas	integer	0	5.016667	2.4218311	5.00	2.96520	0.0	13.0	0
Tiempo.enero	numeric	0	38.235833	8.7875083	37.85	9.04386	17.5	62.2	0
Tiempo.febrero	numeric	0	37.483333	8.3435713	37.60	8.00604	17.5	59.6	0
Monto	numeric	0	92.678333	10.2873846	90.70	10.30407	74.1	119.1	0
Espera	numeric	0	6.148333	6.8776582	3.65	2.96520	0.2	36.0	0
Opinión	factor	0	NA	0.6333333	NA	NA	15.0	44.0	5
Empresa	factor	0	NA	0.6666667	NA	NA	20.0	40.0	4

1. Revisar graficamente el porcentaje de nulos

In [10]:

```
library(VIM)
aggr(telecom,numbers=TRUE, plot = T)
```


Proportion of missings
0.0
0.2
0.4
0.6
0.8
1.0
IdCliente
Género
Edad
Llamadas
Monto
Espera
Opinión
Empresa
Combinations
IdCliente
Género
Edad
Llamadas
Monto
Espera
Opinión
Empresa
1

1. Visualizar la tabla y frecuencia con formato SAS

In [12]:

```
library(gmodels)
library(gdata)
CrossTable(telecom$Género, format="SAS")
```

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  120


          |  Femenino | Masculino |
          |-----------|-----------|
          |        54 |        66 |
          |     0.450 |     0.550 |
          |-----------|-----------|
```

1. Visualizar la tabla y frecuencia con formato SPSS

In [13]:

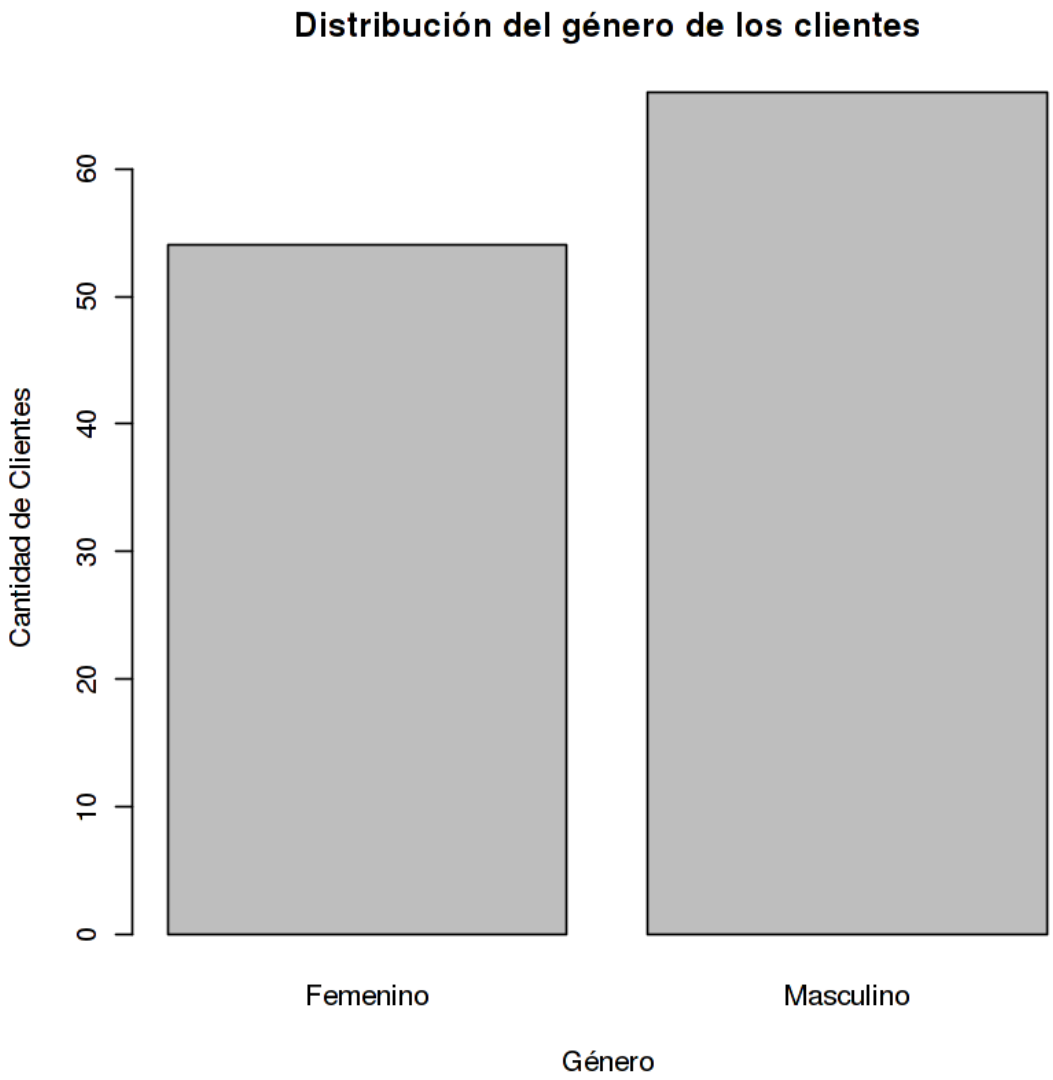
```
CrossTable(telecom$Género, format="SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  120

          |  Femenino  | Masculino  |
          |-----------|-----------|
          |       54  |       66  |
          |   45.000% |   55.000% |
          |-----------|-----------|
```

1. Visualizar gráficos de resumen de las variables: gráfico de barras

In [14]:

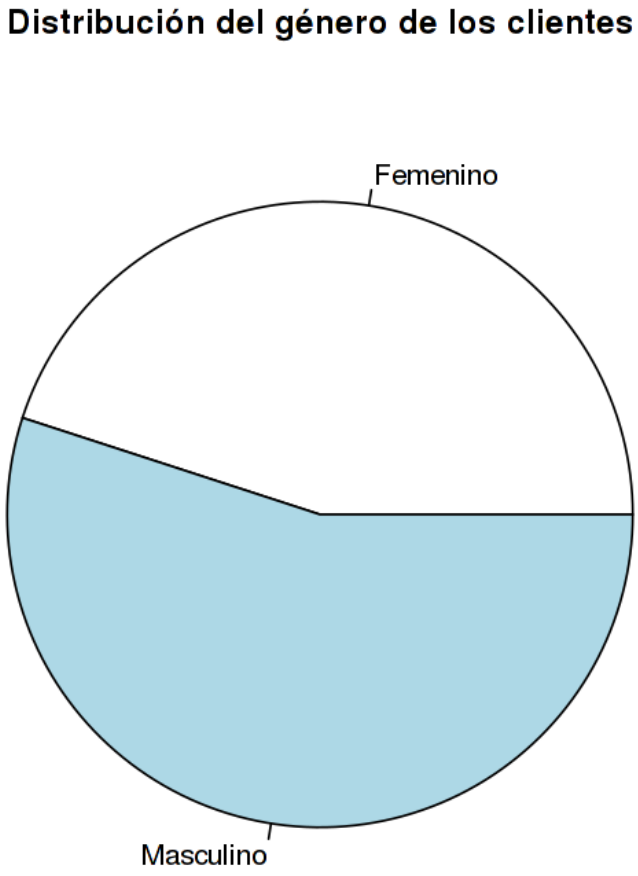
```
frec <- table(telecom$Género)
barplot(frec, main="Distribución del género de los clientes", xlab="Género", ylab="Cant
idad de Clientes")
```



1. Visualizar gráficos de resumen de las variables: pie

In [15]:

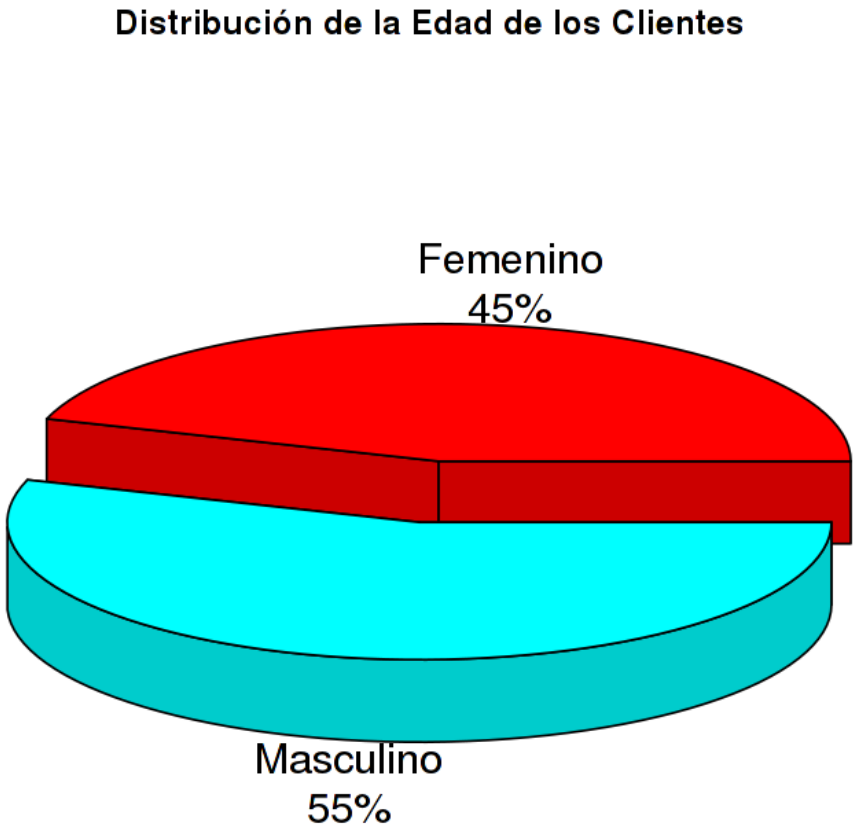
```
pie(frec, main="Distribución del género de los clientes", xlab="Género")
```



1. Visualizar gráficos de resumen de las variables: pie 3D

In [16]:

```
library(plotrix)
lbls1 <- paste(names(table(telecom$Género)), "\n", prop.table(table(telecom$Género))*10
0,"%", sep="")
pie3D(table(telecom$Género), labels = lbls1,explode=0.15, main="Distribución de la Edad
de los Clientes")
```

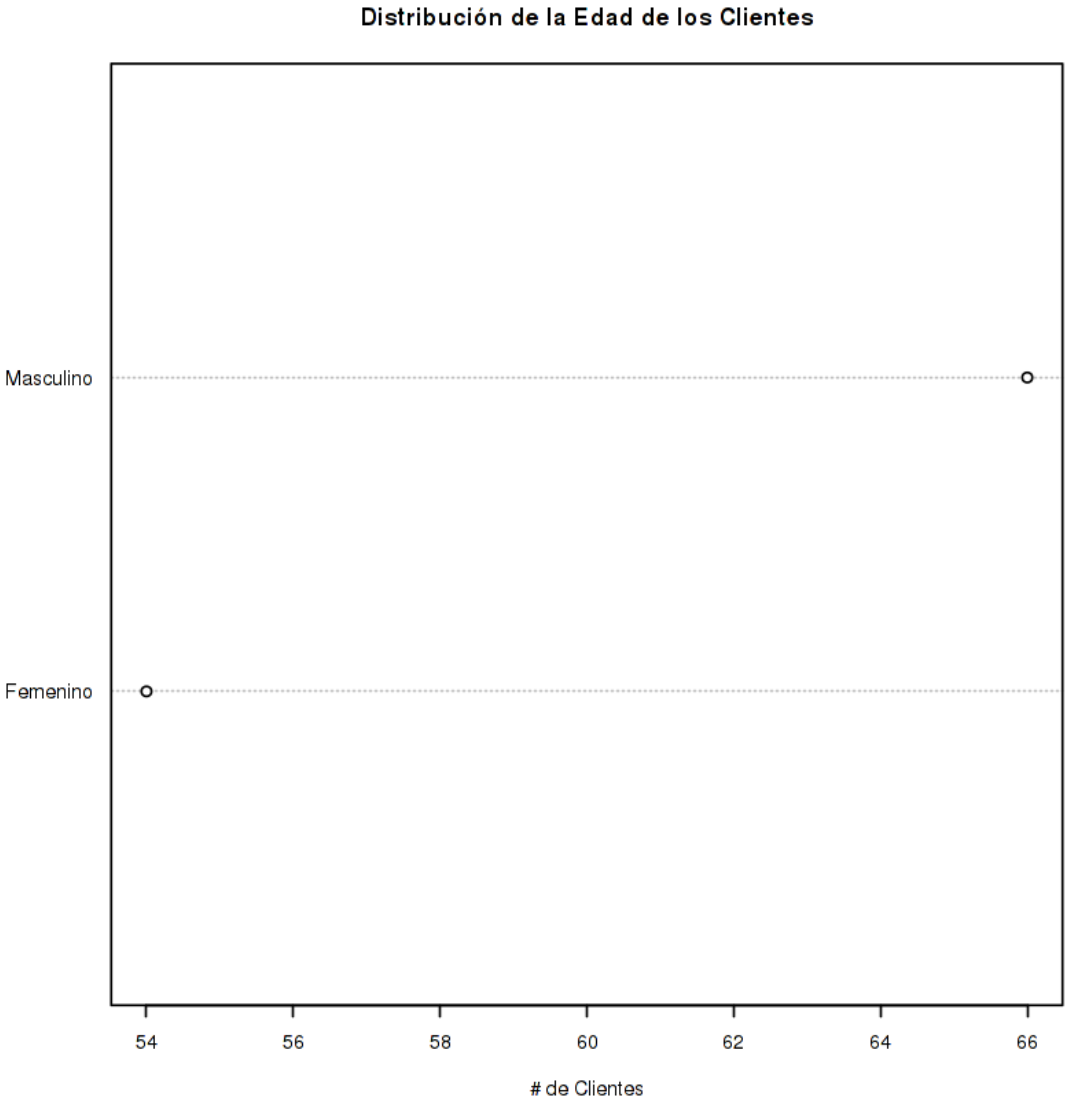


1. Visualizar gráficos de resumen de las variables: grafico de puntos

In [19]:

```
dotchart(table(telecom$Género), cex=.7, main="Distribución de la Edad de los Clientes",
xlab="# de Clientes")
```

```
Warning message in dotchart(table(telecom$Género), cex = 0.7, main = "Dist
ribución de la Edad de los Clientes", :
“'x' is neither a vector nor a matrix: using as.numeric(x)”
```



In [31]:

```
13. Generar tablas dinamicas o de contingencia
```

```
Error in parse(text = x, srcfile = src): <text>:1:5: unexpected symbol
1: 13. Generar
        ^
Traceback:
```

In [32]:

```
tablacruzada<-table(telecom$Opinión, telecom$Género)
tablacruzada
```

```
           Femenino Masculino
  Bueno             9        14
  Excelente         9        10
  Muy Bueno         8        11
  Pésimo           21        23
  Regular           7         8
```

In [21]:

```
CrossTable(telecom$Opinión, telecom$Género, format="SPSS")
```

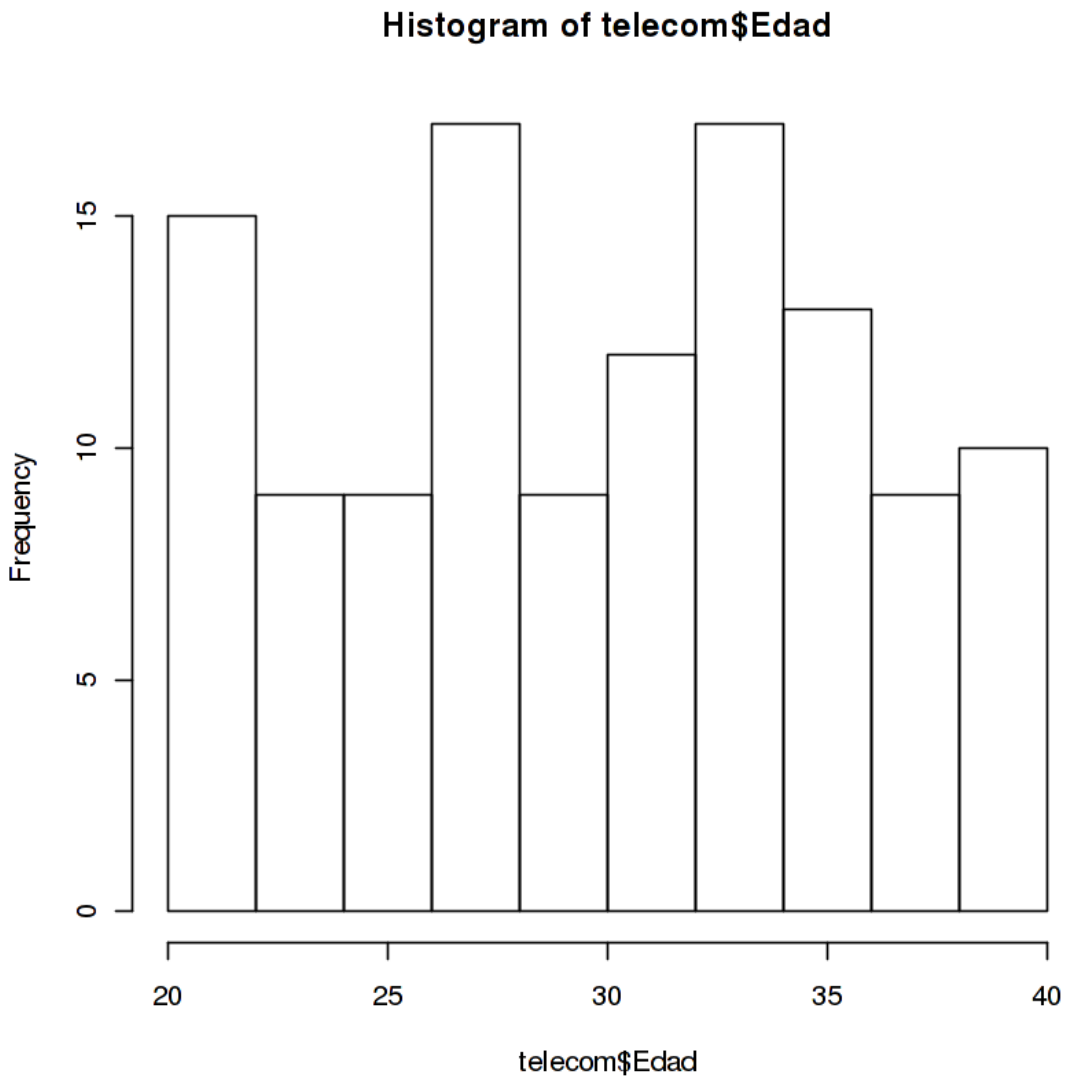
```
   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  120

                | telecom$Género
telecom$Opinión |  Femenino  | Masculino  | Row Total |
----------------|------------|------------|-----------|
          Bueno |         9  |        14  |        23  |
                |    0.176   |    0.144   |            |
                |   39.130%  |   60.870%  |   19.167%  |
                |   16.667%  |   21.212%  |            |
                |    7.500%  |   11.667%  |            |
----------------|------------|------------|-----------|
      Excelente |         9  |        10  |        19  |
                |    0.024   |    0.019   |            |
                |   47.368%  |   52.632%  |   15.833%  |
                |   16.667%  |   15.152%  |            |
                |    7.500%  |    8.333%  |            |
----------------|------------|------------|-----------|
      Muy Bueno |         8  |        11  |        19  |
                |    0.035   |    0.029   |            |
                |   42.105%  |   57.895%  |   15.833%  |
                |   14.815%  |   16.667%  |            |
                |    6.667%  |    9.167%  |            |
----------------|------------|------------|-----------|
         Pésimo |        21  |        23  |        44  |
                |    0.073   |    0.060   |            |
                |   47.727%  |   52.273%  |   36.667%  |
                |   38.889%  |   34.848%  |            |
                |   17.500%  |   19.167%  |            |
----------------|------------|------------|-----------|
        Regular |         7  |         8  |        15  |
                |    0.009   |    0.008   |            |
                |   46.667%  |   53.333%  |   12.500%  |
                |   12.963%  |   12.121%  |            |
                |    5.833%  |    6.667%  |            |
----------------|------------|------------|-----------|
   Column Total |        54  |        66  |       120  |
                |   45.000%  |   55.000%  |            |
----------------|------------|------------|-----------|
```

1. Representación de variables cuantitativas: Edad

In [22]:

```
library(agricolae)
(table.freq(hist(telecom$Edad,breaks = "Sturges")))
```

Lower	Upper	Main	Frequency	Percentage	CF	CPF
20	22	21	15	12.5	15	12.5
22	24	23	9	7.5	24	20.0
24	26	25	9	7.5	33	27.5
26	28	27	17	14.2	50	41.7
28	30	29	9	7.5	59	49.2
30	32	31	12	10.0	71	59.2
32	34	33	17	14.2	88	73.3
34	36	35	13	10.8	101	84.2
36	38	37	9	7.5	110	91.7
38	40	39	10	8.3	120	100.0

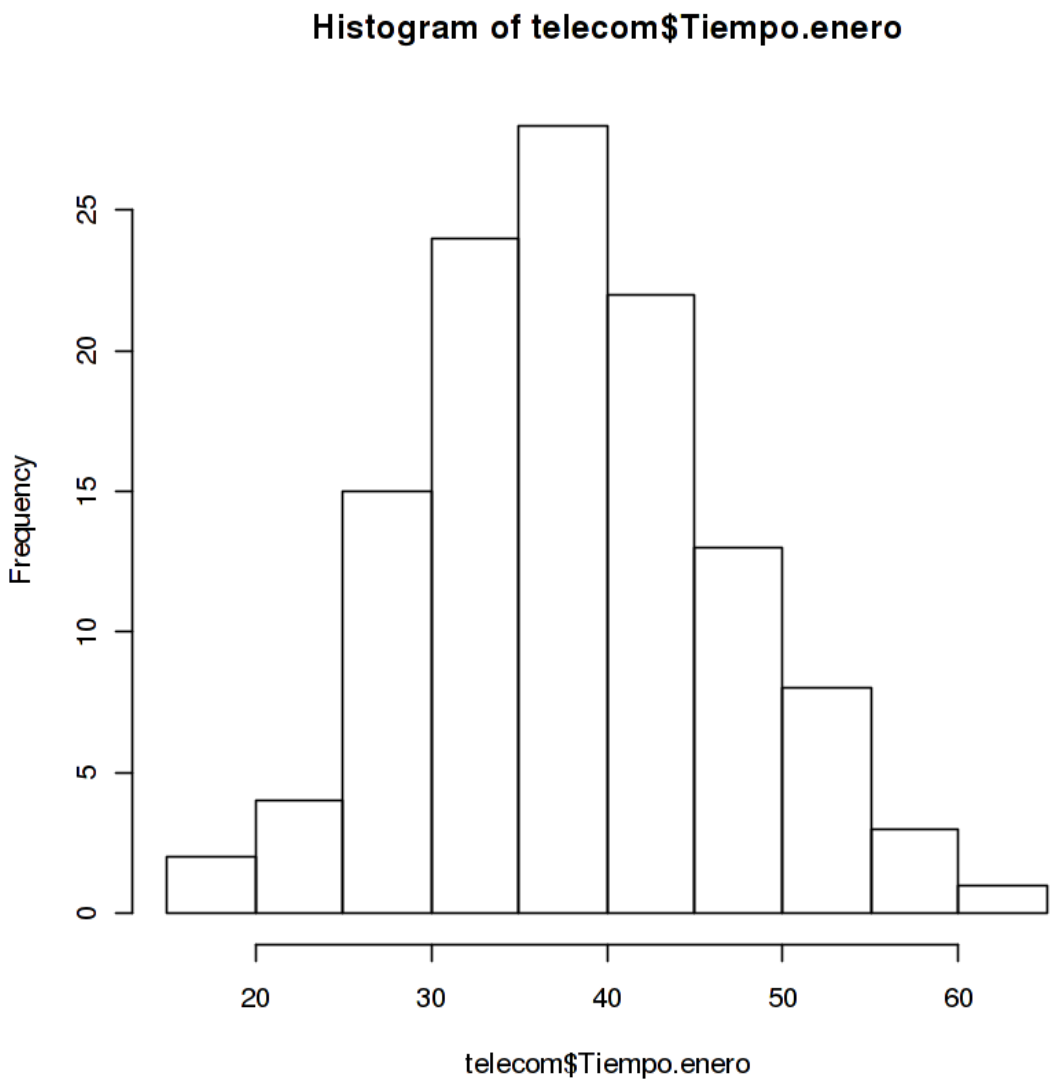


1. Representación de variables cuantitativas: Tiempo de llamadas en Enero

In [23]:

```
(table.freq(hist(telecom$Tiempo.enero,breaks = "Scott")))
```

Lower	Upper	Main	Frequency	Percentage	CF	CPF
15	20	17.5	2	1.7	2	1.7
20	25	22.5	4	3.3	6	5.0
25	30	27.5	15	12.5	21	17.5
30	35	32.5	24	20.0	45	37.5
35	40	37.5	28	23.3	73	60.8
40	45	42.5	22	18.3	95	79.2
45	50	47.5	13	10.8	108	90.0
50	55	52.5	8	6.7	116	96.7
55	60	57.5	3	2.5	119	99.2
60	65	62.5	1	0.8	120	100.0



1. Representación de variables cuantitativas: Monto

In [24]:

```
(table.freq(graph.freq(telecom$Monto,plot=FALSE)))
```

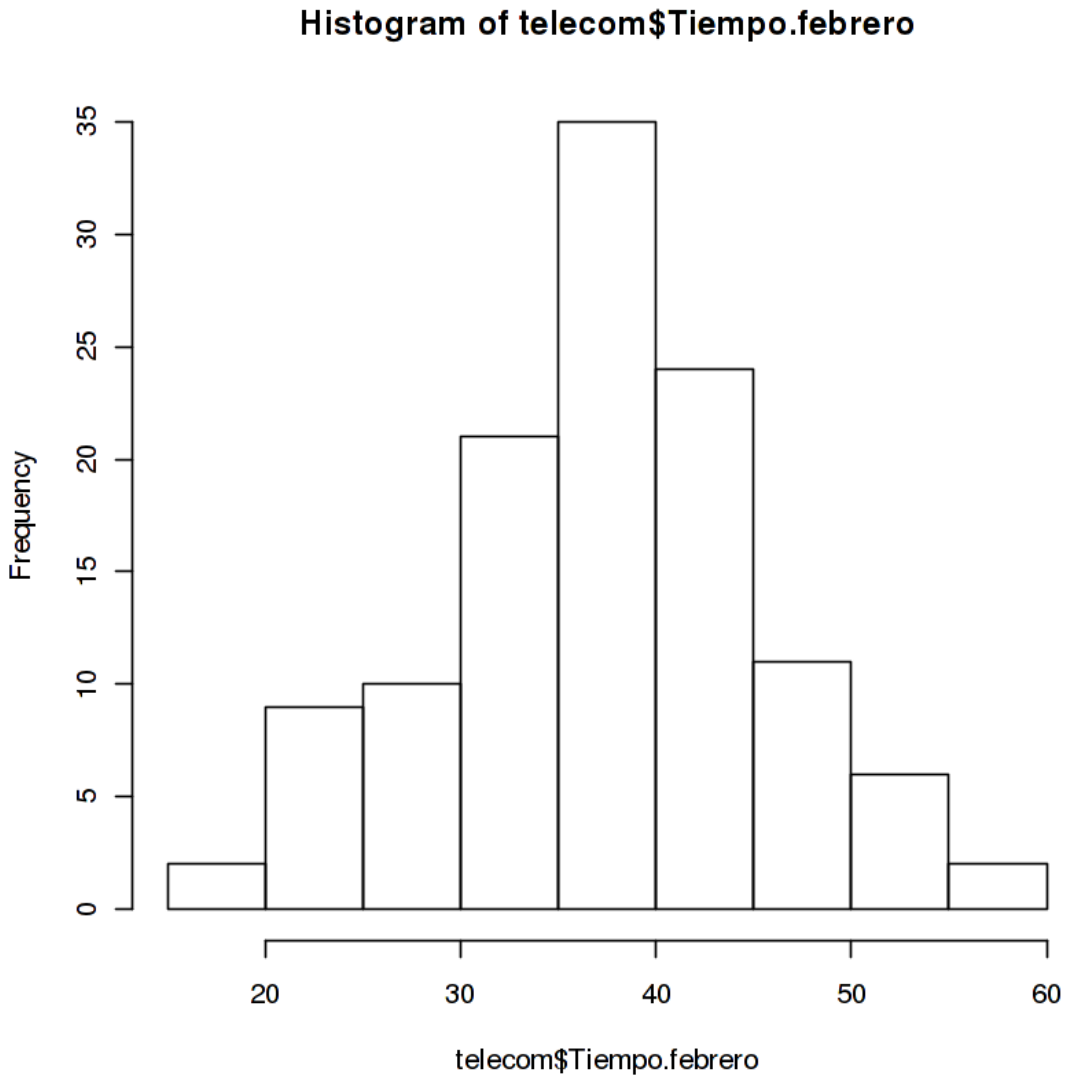
Lower	Upper	Main	Frequency	Percentage	CF	CPF
74.0	79.8	76.9	10	8.3	10	8.3
79.8	85.6	82.7	25	20.8	35	29.2
85.6	91.4	88.5	27	22.5	62	51.7
91.4	97.2	94.3	19	15.8	81	67.5
97.2	103.0	100.1	18	15.0	99	82.5
103.0	108.8	105.9	10	8.3	109	90.8
108.8	114.6	111.7	9	7.5	118	98.3
114.6	120.4	117.5	2	1.7	120	100.0

1. Representación de variables cuantitativas: Tiempo de llamadas en febrero

In [25]:

```
(table.freq(hist(telecom$Tiempo.febrero,breaks = "FD")))
```

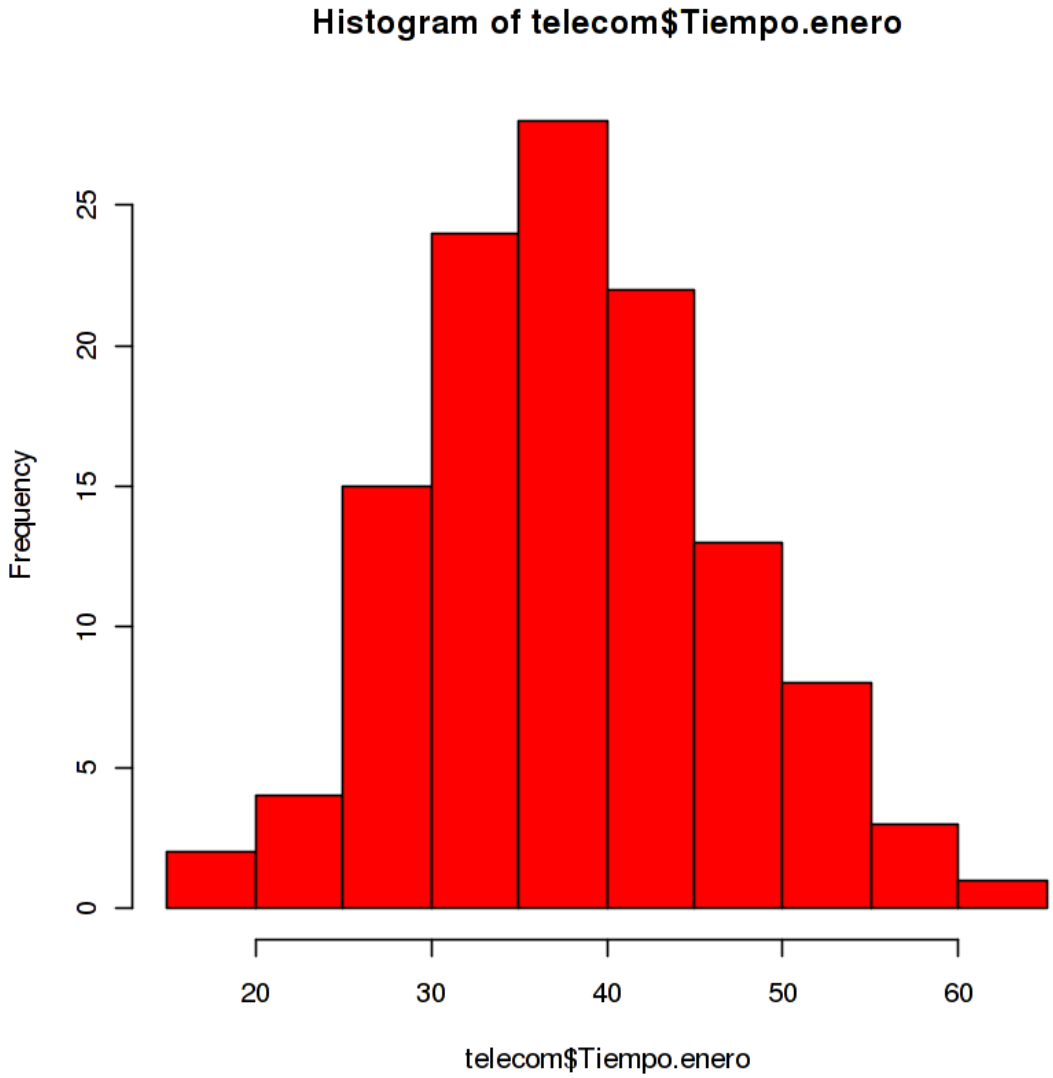
Lower	Upper	Main	Frequency	Percentage	CF	CPF
15	20	17.5	2	1.7	2	1.7
20	25	22.5	9	7.5	11	9.2
25	30	27.5	10	8.3	21	17.5
30	35	32.5	21	17.5	42	35.0
35	40	37.5	35	29.2	77	64.2
40	45	42.5	24	20.0	101	84.2
45	50	47.5	11	9.2	112	93.3
50	55	52.5	6	5.0	118	98.3
55	60	57.5	2	1.7	120	100.0



1. Representación de variables cuantitativas de forma gráfica: Histograma

In [26]:

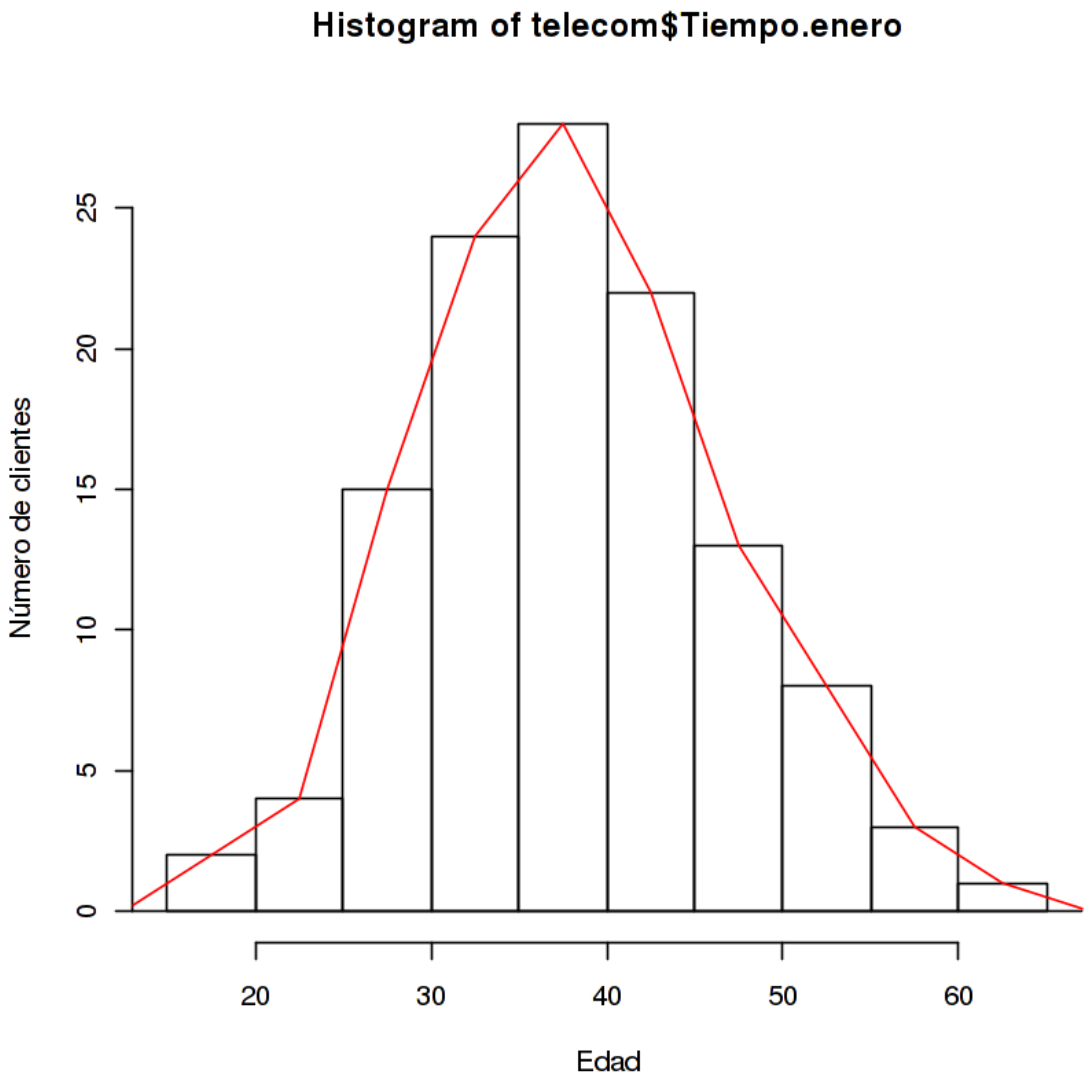
```
hist(telecom$Tiempo.enero, col = 2)
```



1. Representación de variables cuantitativas de forma gráfica: Histograma y linea de distribución

In [27]:

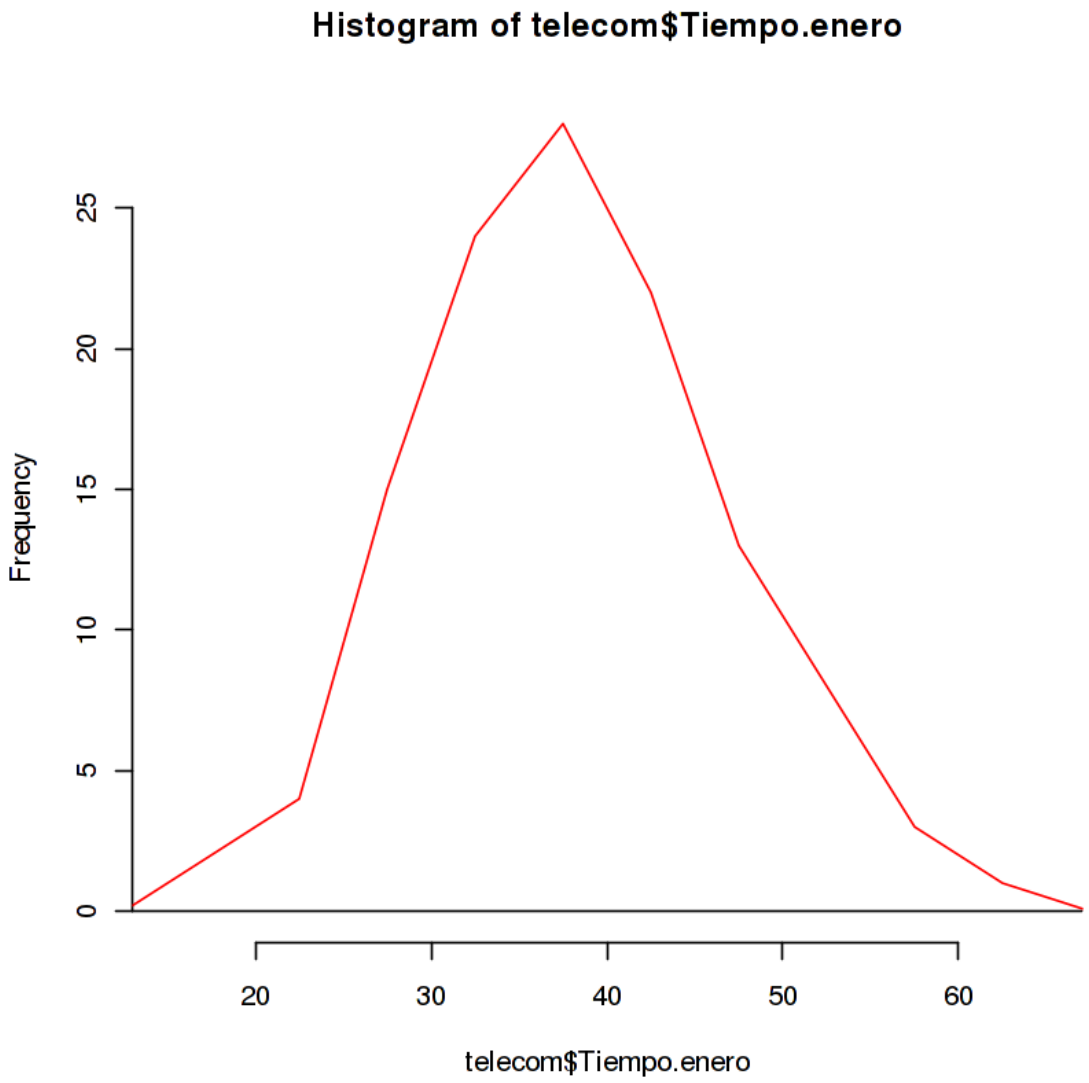
```
histograma<-hist(telecom$Tiempo.enero,breaks = "Sturges", xlab="Edad", ylab="Número de
 clientes",)
polygon.freq(histograma,frequency=1,col="red")
```



1. Representación de variables cuantitativas de forma gráfica: linea de distribución

In [28]:

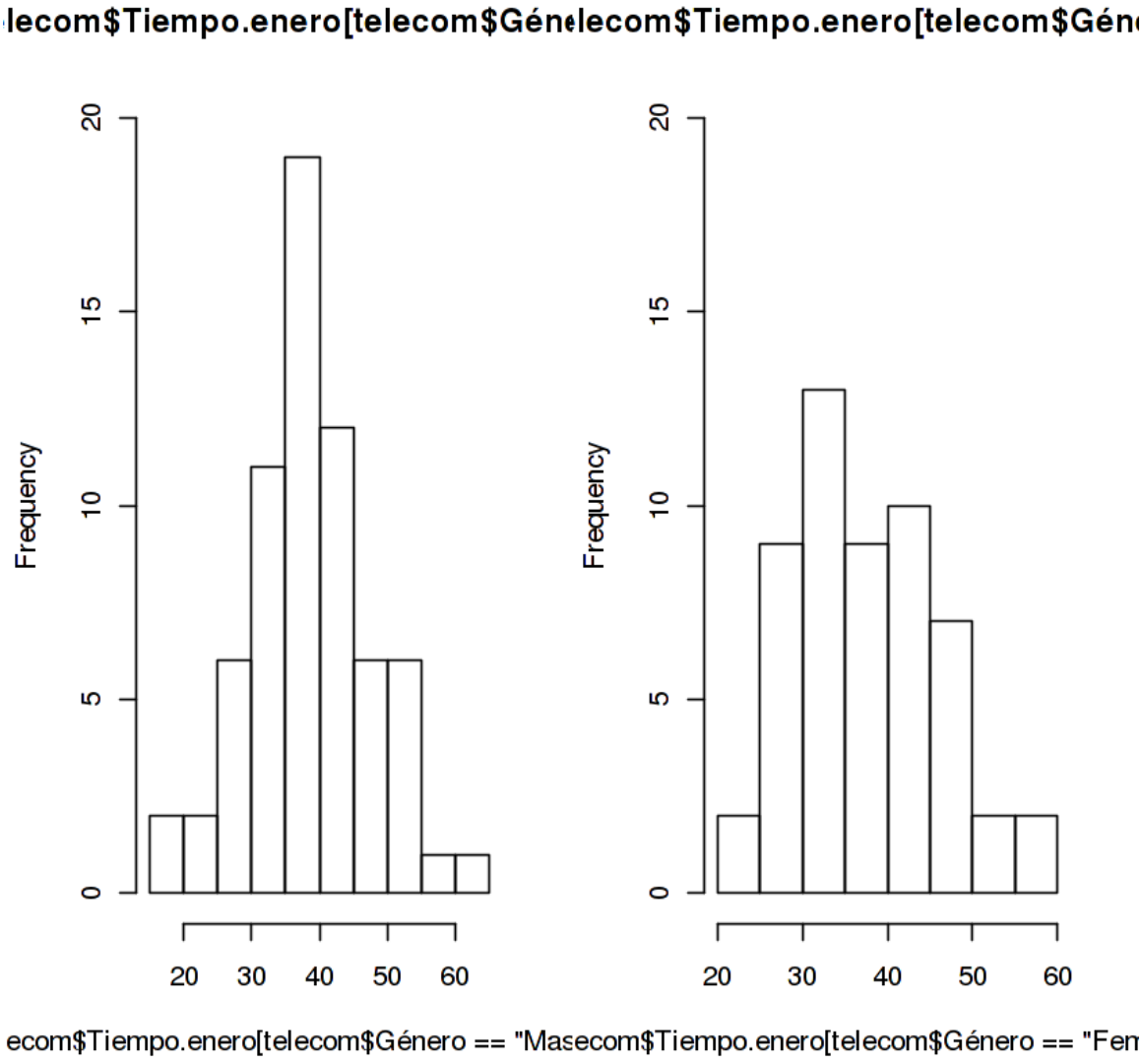
```
histograma1<-hist(telecom$Tiempo.enero,border=FALSE)
polygon.freq(histograma,frequency=1,col="red")
```



1. Representación de variables cuantitativas de forma gráfica: Comparativa de Histogramas

In [29]:

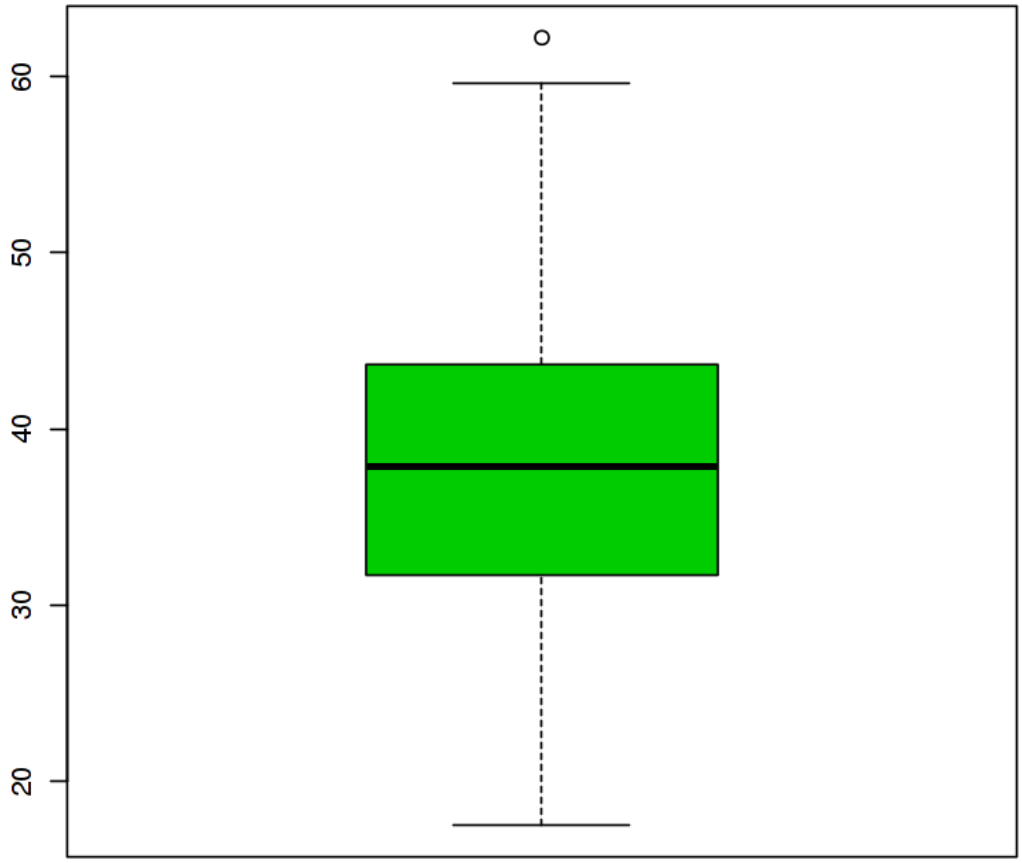
```
par(mfrow=c(1,2))
hist(telecom$Tiempo.enero[telecom$Género=="Masculino"],ylim=c(0,20))
hist(telecom$Tiempo.enero[telecom$Género=="Femenino"],ylim=c(0,20))
```



1. Representación de variables cuantitativas de forma gráfica: Gráfico de Caja

In [30]:

```
boxplot(telecom$Tiempo.enero, col = 3)
```



In [ ]:

```

```