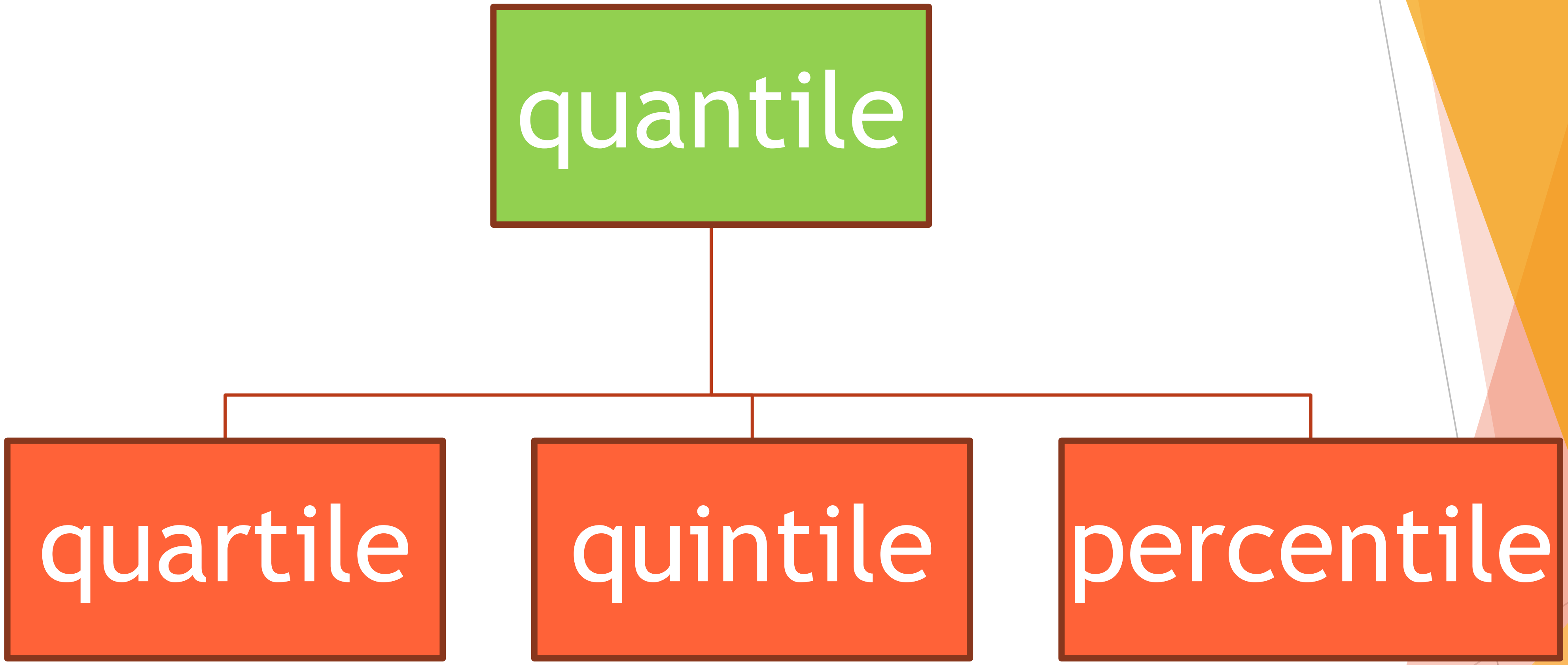
A quantile is a cut point, or line of division, that splits a probability distribution into continuous intervals with equal probabilities.



quantile

quartile

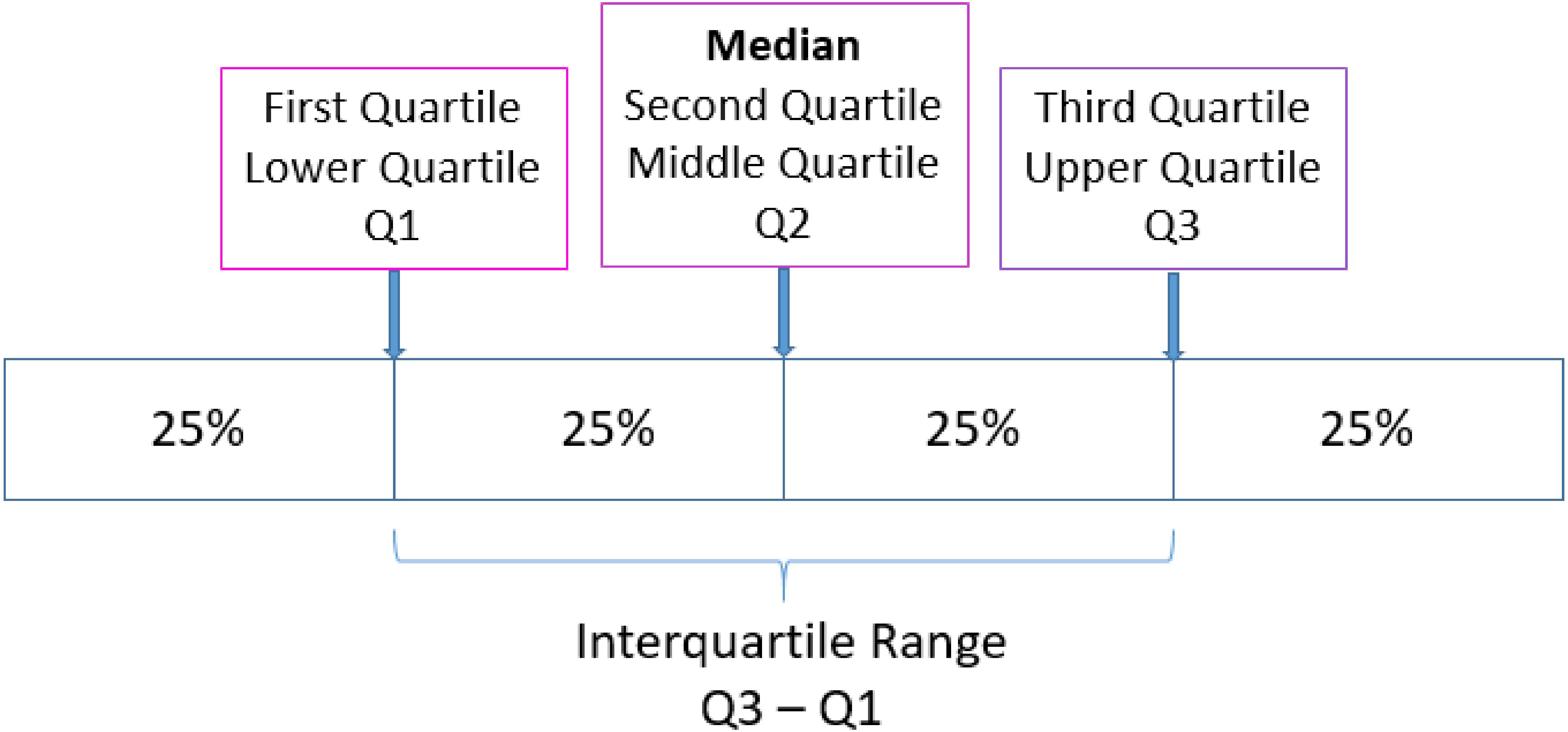
quintile

percentile

Quartile

- We sort the data in ascending order.
- The first quartile, or lower quartile, is the value that cuts off the first 25% of the ordered data
- The second quartile, or median, is the value that cuts off the first 50%.
- The third quartile, or upper quartile, is the value that cuts off the first 75%.

Median and Quartiles



Median
Second Quartile
Middle Quartile
Q2

First Quartile
Lower Quartile
Q1

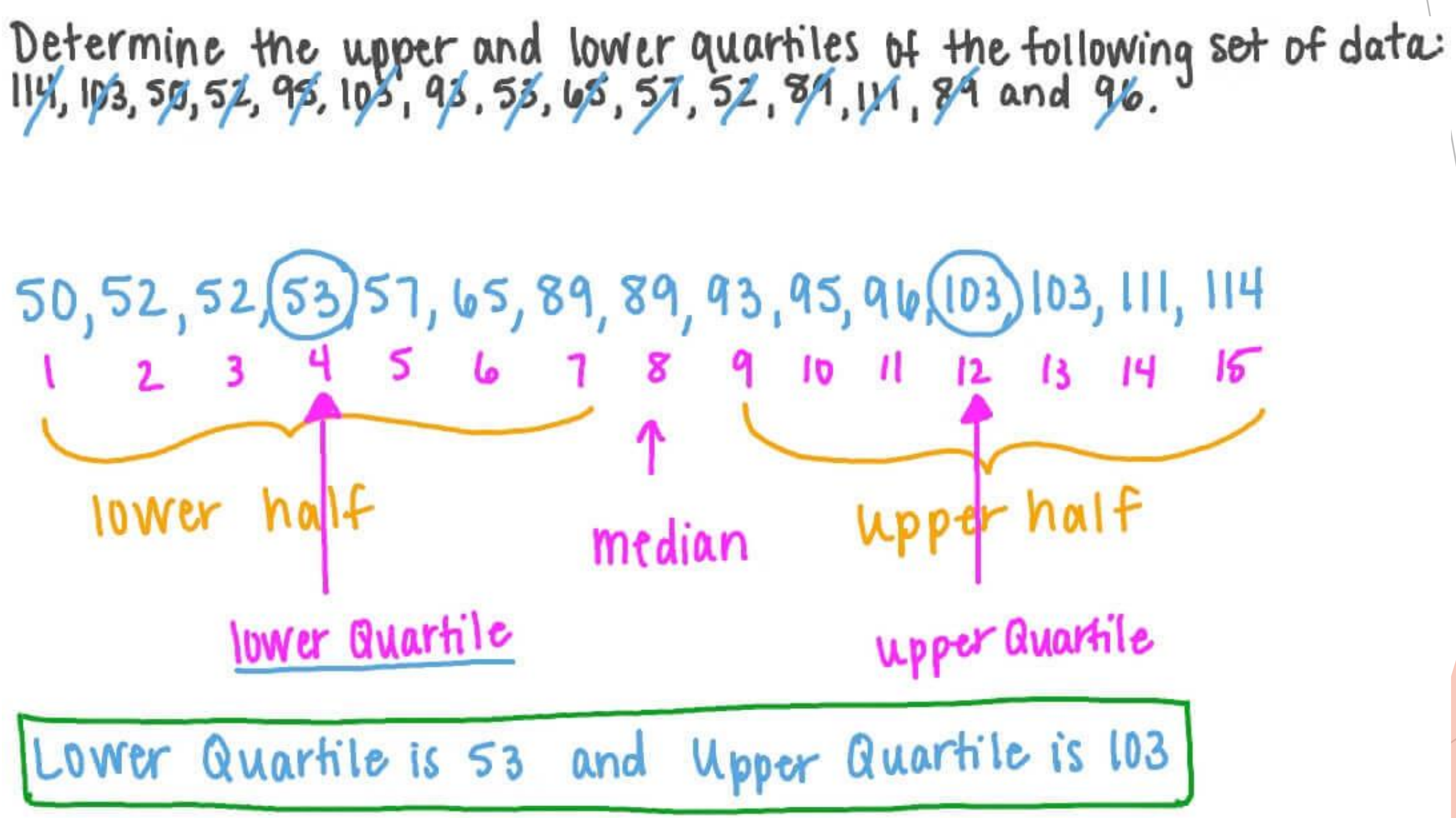
Third Quartile
Upper Quartile
Q3

25%	25%	25%	25%

Interquartile Range
Q3 – Q1

IQR 50% of data

Quartile



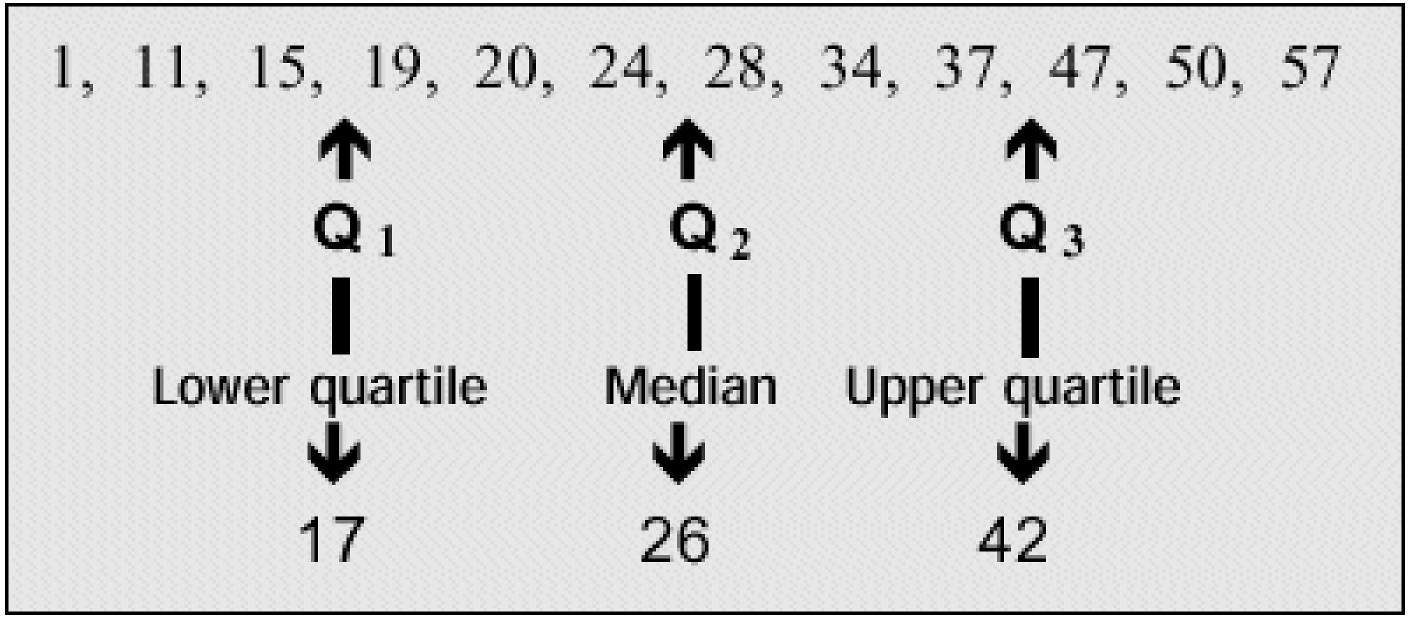
Determine the upper and lower quartiles of the following set of data:
114, 103, 50, 52, 95, 103, 93, 53, 65, 57, 52, 89, 111, 89 and 96.

50, 52, 52, 53, 57, 65, 89, 89, 93, 95, 96, 103, 103, 111, 114

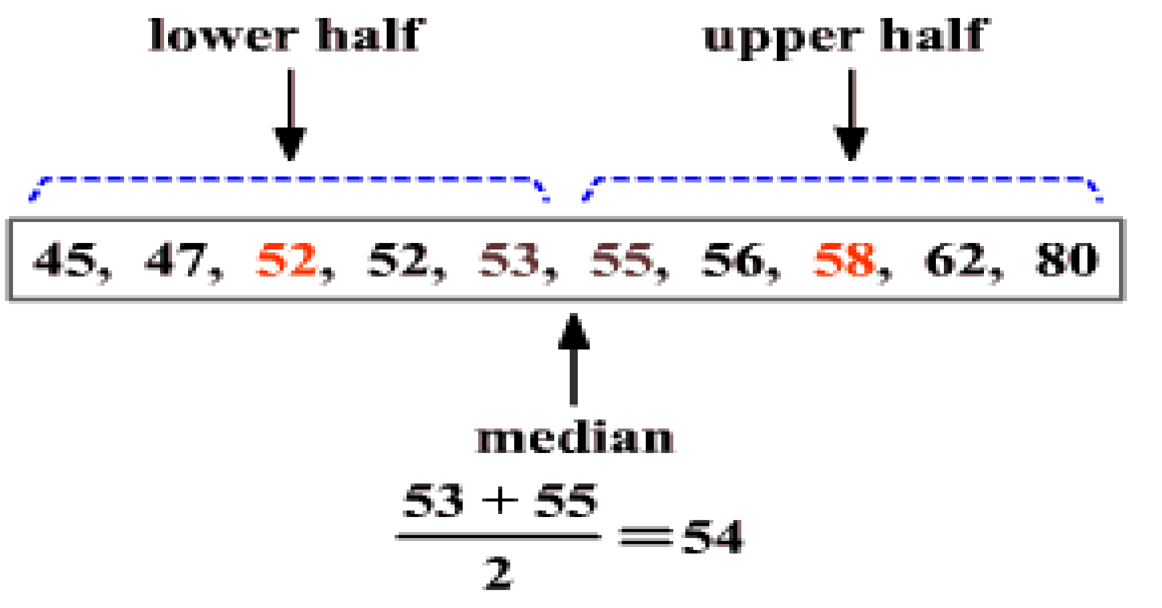
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

lower half
median
upper half
lower Quartile
upper Quartile

Lower Quartile is 53 and Upper Quartile is 103



1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57



45, 47, 52, 52, 53, 55, 56, 58, 62, 80

$$\frac{53 + 55}{2} = 54$$

Quartil in R

```
edad=mydata$age
quantile(edad)
```

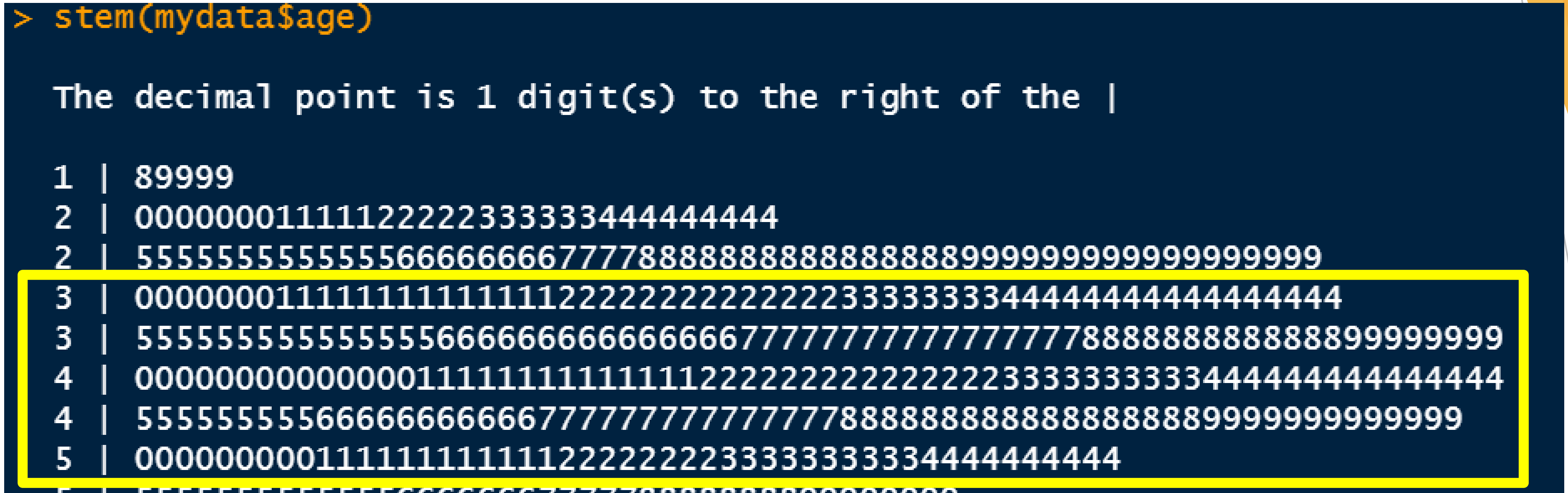
```
•     0%   25%   50%   75% 100%
•     18    32    41    51    73
```

Interquartile

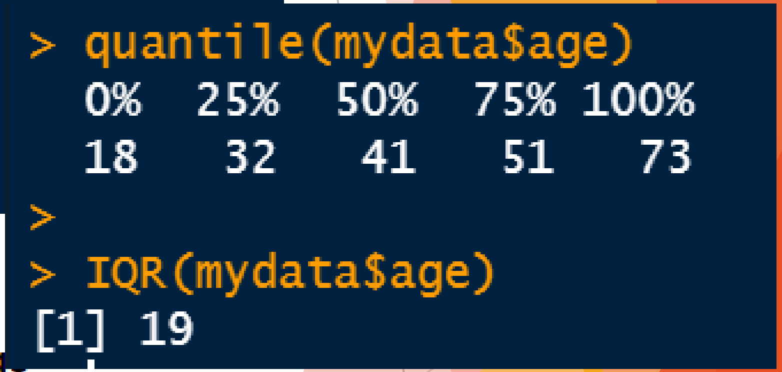
The interquartile range of an observation variable is the difference between its upper and lower quartiles. It is a measure of how far apart the middle portion of data spreads in value.

Interquartile Range = Upper Quartile - Lower Quartile

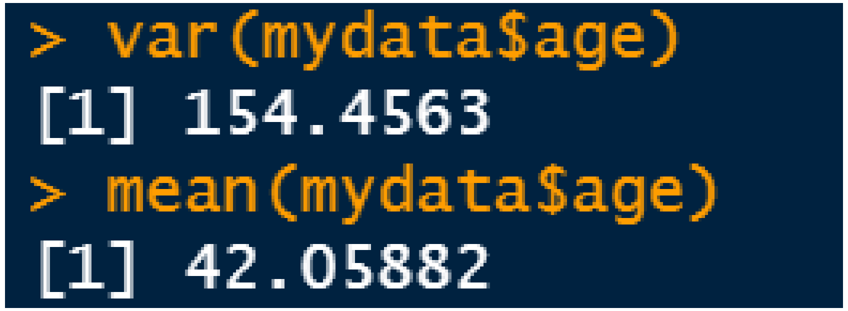
```
IQR(dat1)
[1] 28.6
```



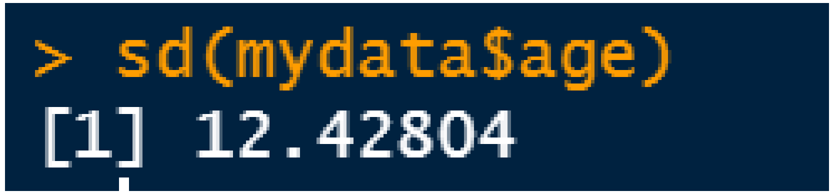
```
> stem(mydata$age)

  The decimal point is 1 digit(s) to the right of the |

  1 | 89999
  2 | 00000001111122222333334444444444
  2 | 55555555555556666666677778888888888888888999999999999999999
  3 | 000000011111111111111222222222222223333333344444444444444444
  3 | 55555555555555566666666666666677777777777777777888888888888899999999
  4 | 00000000000000111111111111112222222222222223333333333444444444444444
  4 | 555555555666666666667777777777777778888888888888888889999999999999
  5 | 0000000001111111111112222222233333333334444444444
  5 | 55555555555556666666777778888888899999999
  6 | 000000011111111112233333333334444
  6 | 5556666666678
  7 | 0001123
```



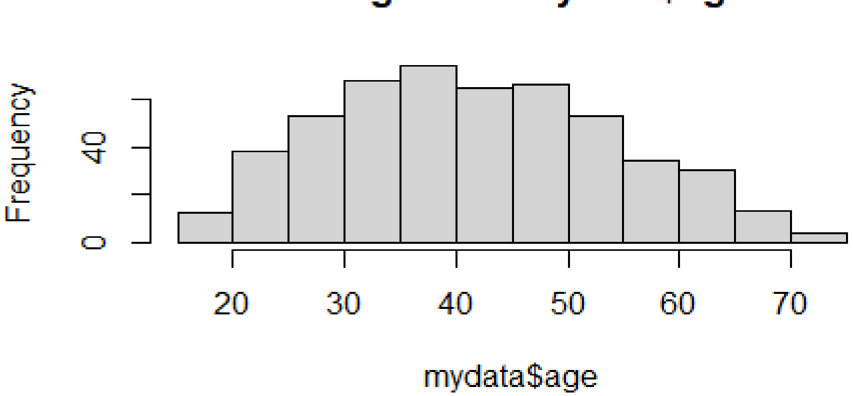
```
> quantile(mydata$age)
  0%  25%  50%  75% 100% 
  18   32   41   51   73 
> 
> IQR(mydata$age)
[1] 19
```

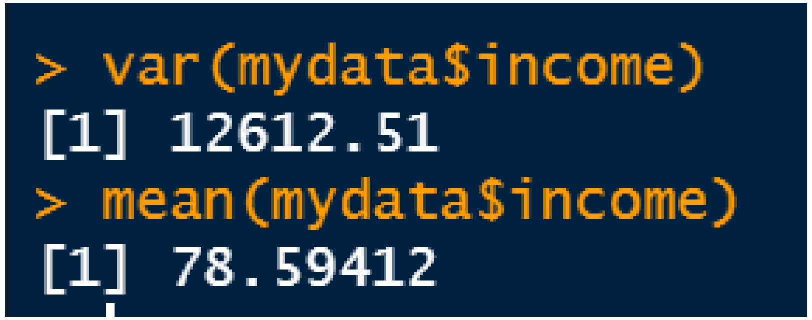


```
> var(mydata$age)
[1] 154.4563
> mean(mydata$age)
[1] 42.05882
```

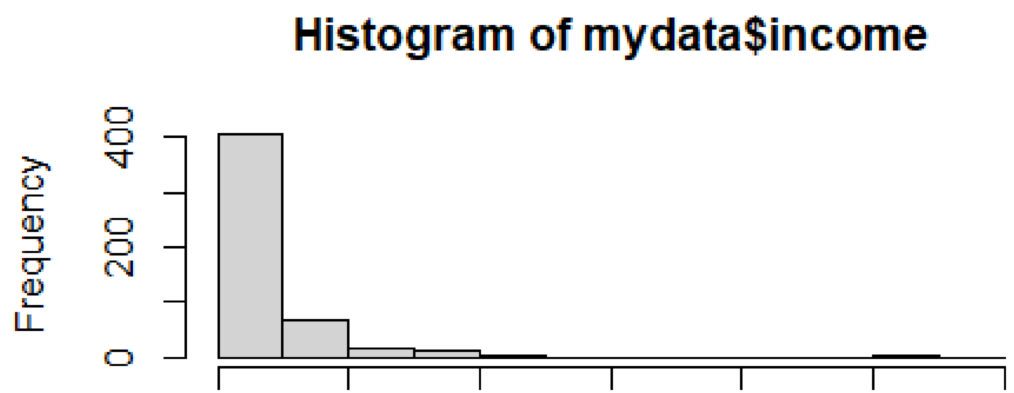


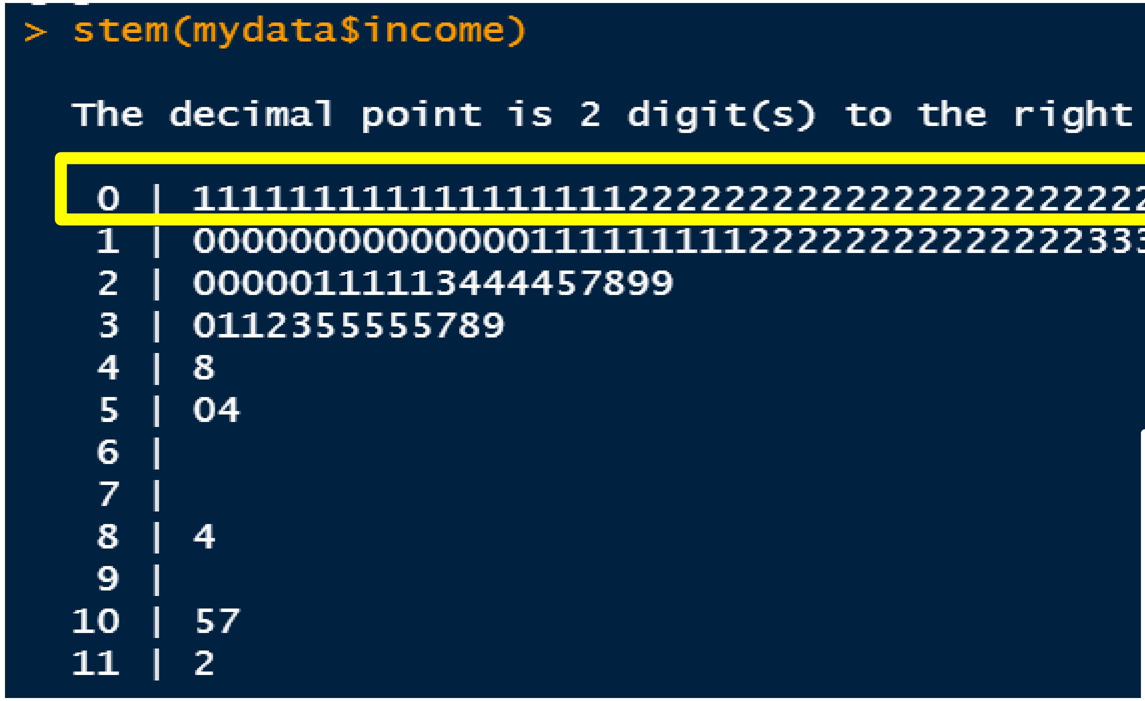
```
> sd(mydata$age)
[1] 12.42804
```



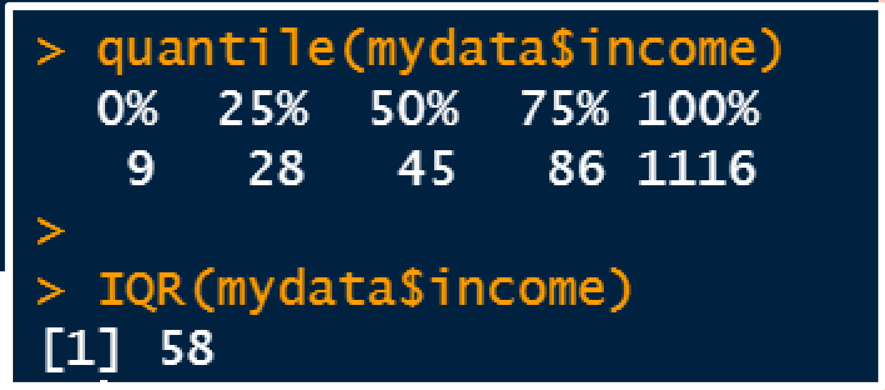


```
> var(mydata$income)
[1] 12612.51
> mean(mydata$income)
[1] 78.59412
```

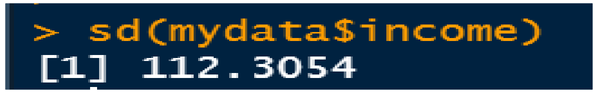




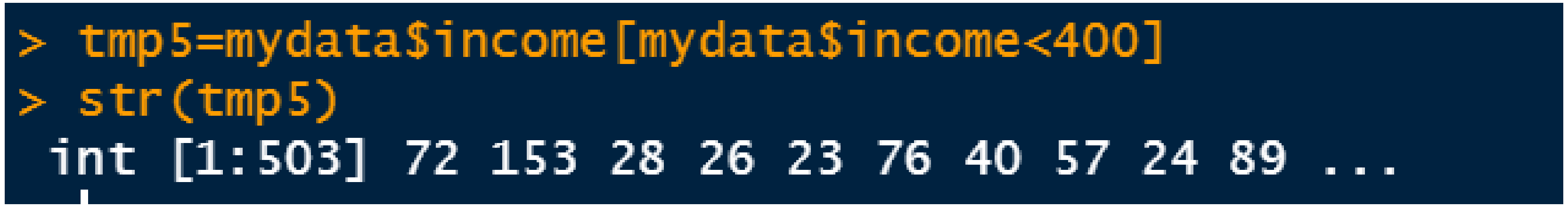
```
> stem(mydata$income)

  The decimal point is 2 digit(s) to the right of the |

   0 | 11111111111111111122222222222222222222222222222222222222222222222222+320
   1 | 0000000000000011111111122222222222222233333334444444444444445555555567
   2 | 00000111113444457899
   3 | 0112355555789
   4 | 8
   5 | 04
   6 | 
   7 | 
   8 | 4
   9 | 
  10 | 57
  11 | 2
```



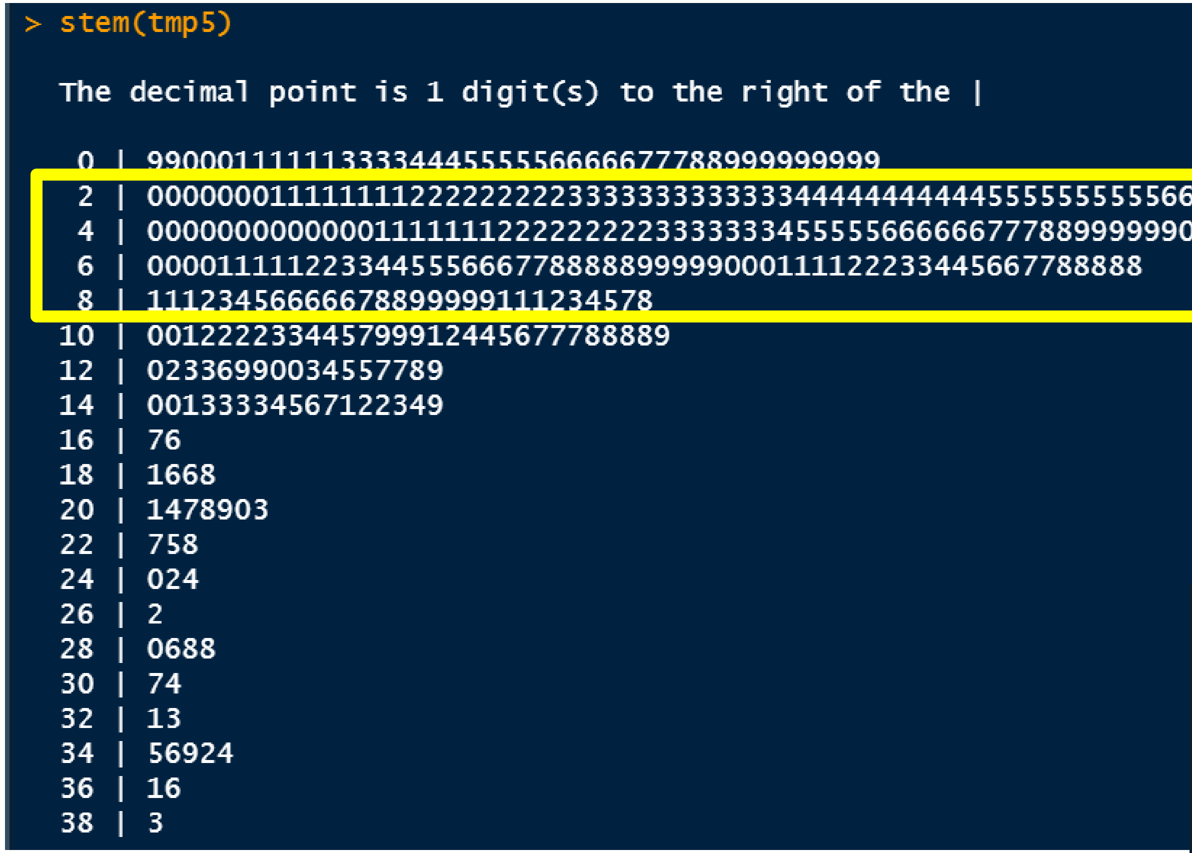
```
> quantile(mydata$income)
  0%  25%  50%  75% 100% 
   9   28   45   86 1116 
> 
> IQR(mydata$income)
[1] 58
```



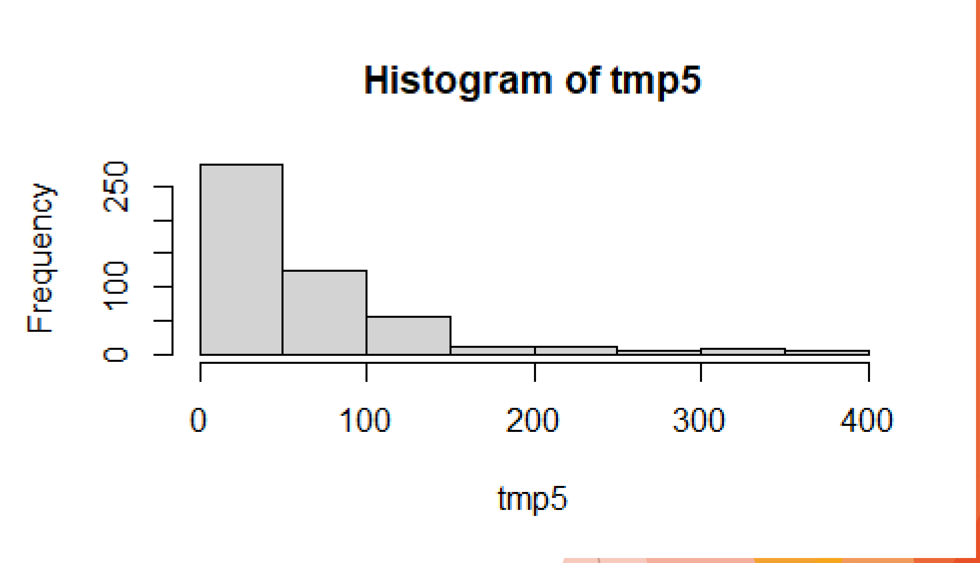
```
> sd(mydata$income)
[1] 112.3054
```

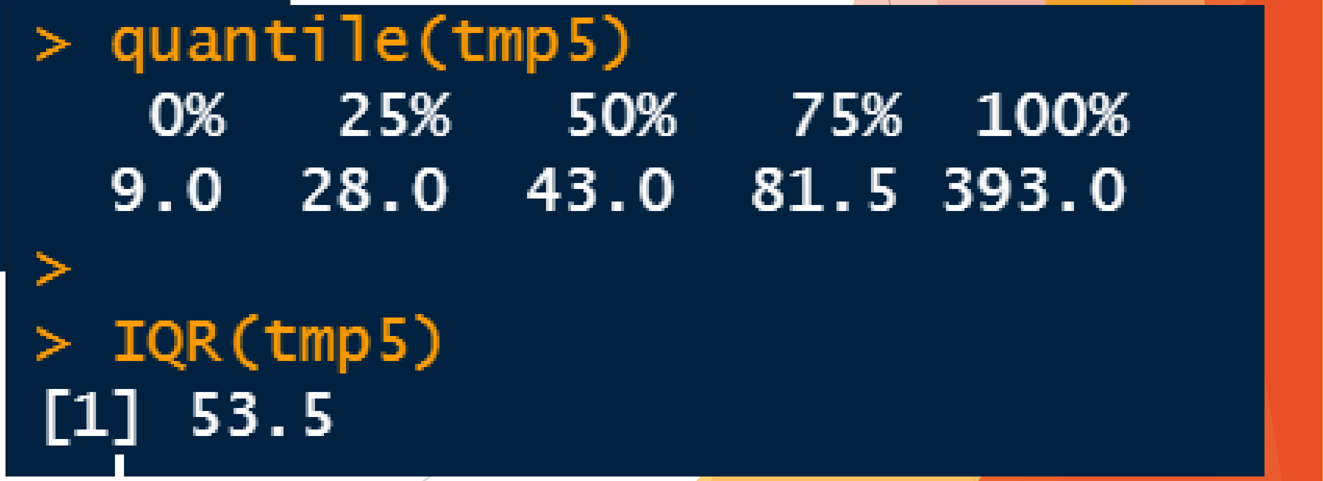


```
> tmp5=mydata$income[mydata$income<400]
> str(tmp5)
 int [1:503] 72 153 28 26 23 76 40 57 24 89 ...
```

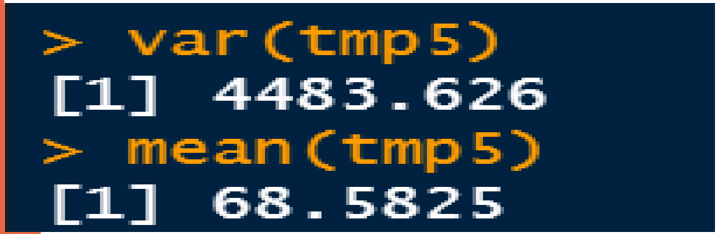


```
> stem(tmp5)

  The decimal point is 1 digit(s) to the right of the |

   0 | 99000111111333344455555666667778899999999 9
   2 | 00000011111111222222222333333333333334444444444455555555556666666666+94
   4 | 0000000000001111111222222222333333345555666666777889999990000000011+21
   6 | 000111112233445556667788888999990001111222334456677888 88
   8 | 111234566666788999991112345 78
  10 | 0012222334457999124456777 88889
  12 | 02336990034557789
  14 | 0013333456712 2349
  16 | 76
  18 | 1668
  20 | 1478903
  22 | 758
  24 | 024
  26 | 2
  28 | 0688
  30 | 74
  32 | 13
  34 | 56924
  36 | 16
  38 | 3
```

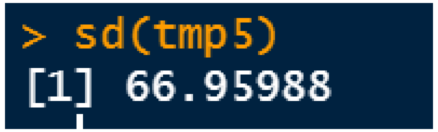




```
> quantile(tmp5)
   0%    25%    50%    75%   100% 
  9.0   28.0   43.0   81.5  393.0 
>
> IQR(tmp5)
[1] 53.5
```

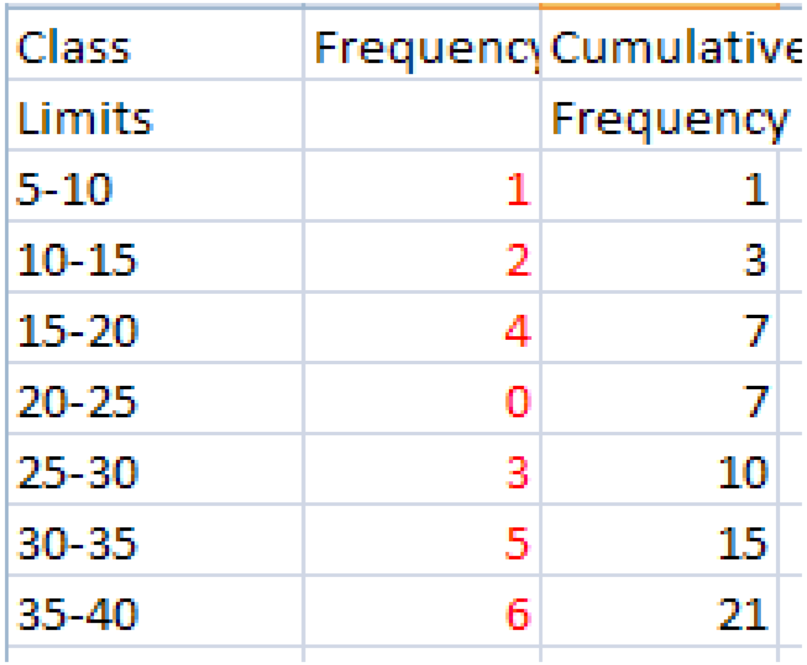


```
> var(tmp5)
[1] 4483.626
> mean(tmp5)
[1] 68.5825
```



```
> sd(tmp5)
[1] 66.95988
```

Calculate quantiles in grouped data



Class	Frequency	Cumulative
Limits		Frequency
5-10	1	1
10-15	2	3
15-20	4	7
20-25	0	7
25-30	3	10
30-35	5	15
35-40	6	21

$$Q = L_i + \frac{PN - F_{i-1}}{f_i} * A$$

Quantile

$$Q = L_i + \frac{PN - F_{i-1}}{f_i} * A$$

f	frequency
fr	relative frequency
F	accumulate frequency
Fr	accumulate relative frequency
[Li , Ls)	interval
A	interval size Ls-Li
N	number of data
P	percentage

Interval	f	fr	F	Fr
[0,5)	3	0.005882	3	0.005882
[5,10)	43	0.084314	46	0.090196
[10,15)	101	0.198039	147	0.288235
[15,20)	79	0.154902	226	0.443137
[20,25)	54	0.105882	280	0.54902
[25,30)	40	0.078431	320	0.627451
[30,35)	33	0.064706	353	0.692157
[35,40)	23	0.045098	376	0.737255
[40,45)	17	0.033333	393	0.770588
[45,50)	14	0.027451	407	0.798039
[50,55)	12	0.023529	419	0.821569
[55,60)	12	0.023529	431	0.845098
[60,65)	11	0.021569	442	0.866667
[65,70)	14	0.027451	456	0.894118
[70,75)	24	0.047059	480	0.941176
[75,80)	20	0.039216	500	0.980392
[80,85)	5	0.009804	505	0.990196
[85,90)	1	0.001961	506	0.992157
[90,95)	1	0.001961	507	0.994118
[95,100)	3	0.005882	510	1

$$Q = L_i + \frac{PN - F_{i-1}}{f_i} * A$$

First quartil
P= 0.25

Find the interval que includes P

$$Q1 = 10 + \frac{0.25*510-46}{101} *5$$

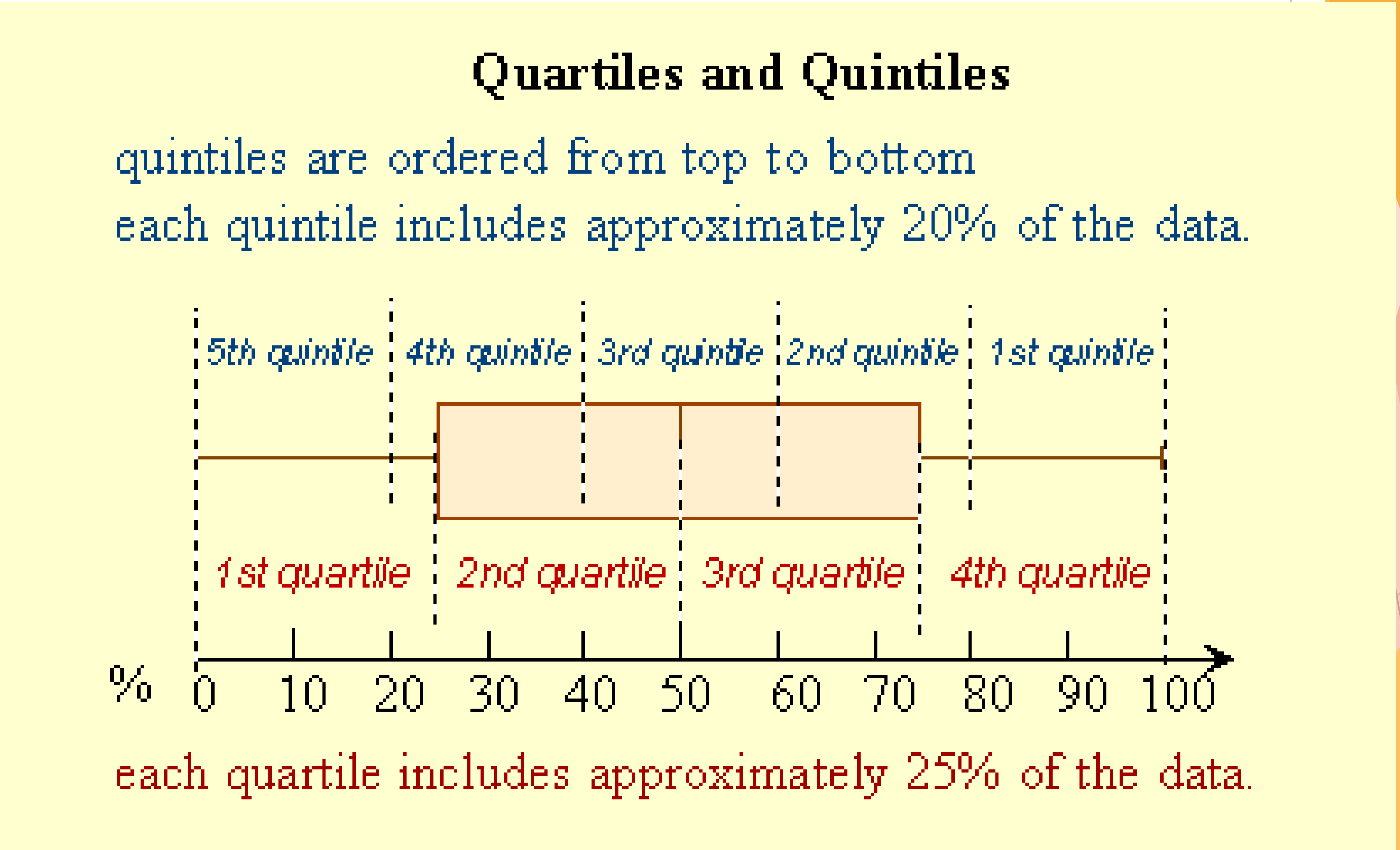
$$Q1 = 14.03$$

Quartil in R

```
edad=mydata$age
quantile(edad)
```

```
•    0%  25%  50%  75% 100%
•    18   32   41   51   73
```

Quintil



Quartiles and Quintiles

quintiles are ordered from top to bottom

each quintile includes approximately 20% of the data.

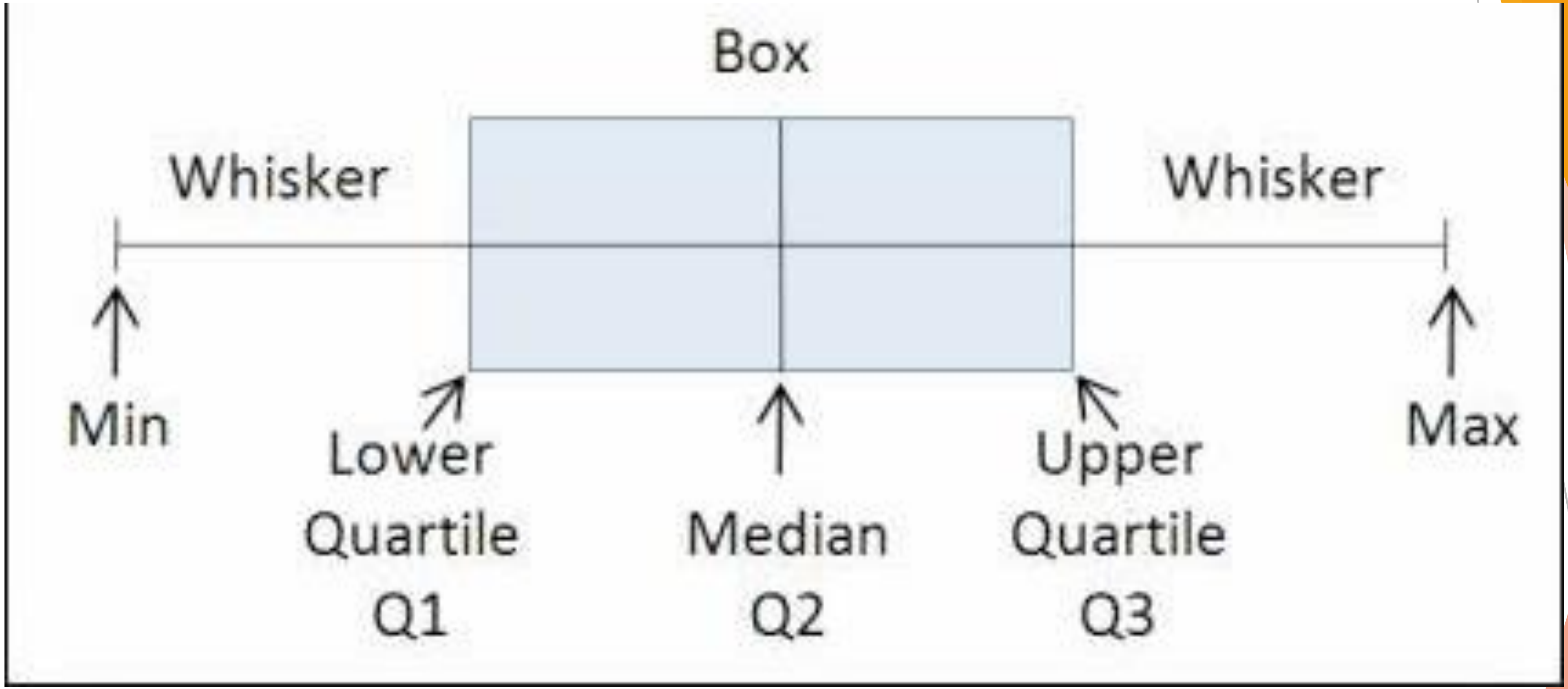
each quartile includes approximately 25% of the data.

Percentile

We have data sorted in ascending order.
The nth percentile is the value that cuts off the first n percent of the data values.

```
quantile(edad, c(0.10,0.30,0.8))
```

- 10% 30% 80%
- 26 34 54


Box
Whisker
Whisker
Min
Lower
Quartile
Q1
Median
Q2
Upper
Quartile
Q3
Max

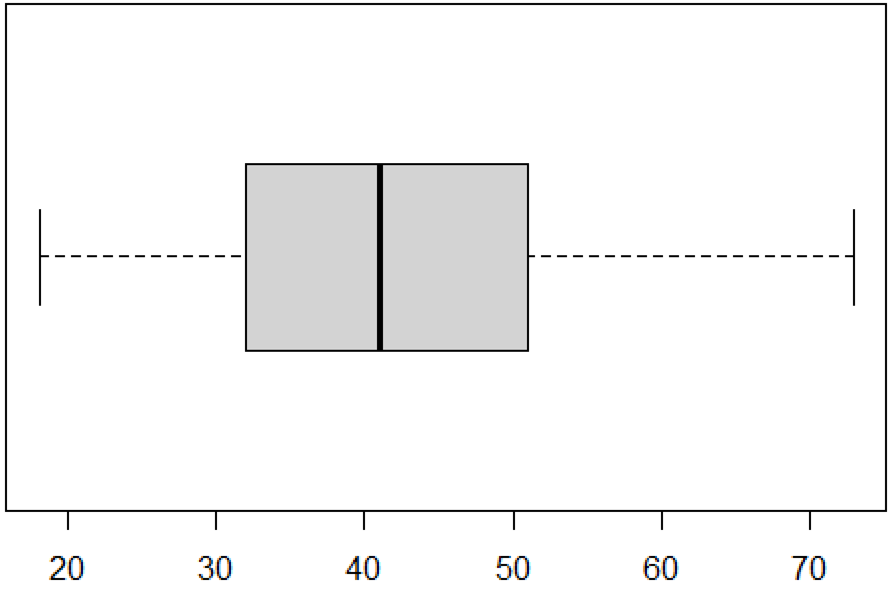
Boxplot

Box plot

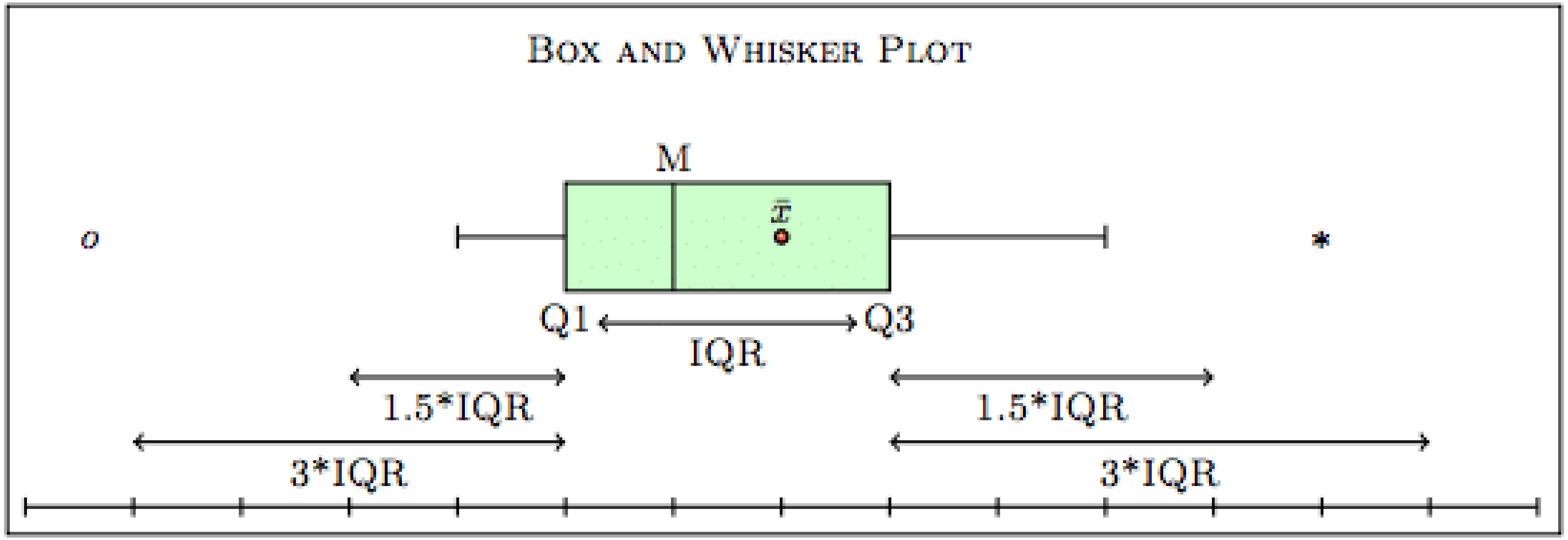
The box plot of an observation variable is a graphical representation based on its quartiles as well as its smallest and largest values. It is a simple yet effective visual representation of data distribution.

```
boxplot(edad,horizontal = TRUE)
```

- quantile(edad)
- 0% 25% 50% 75% 100%
- 18 32 41 51 73

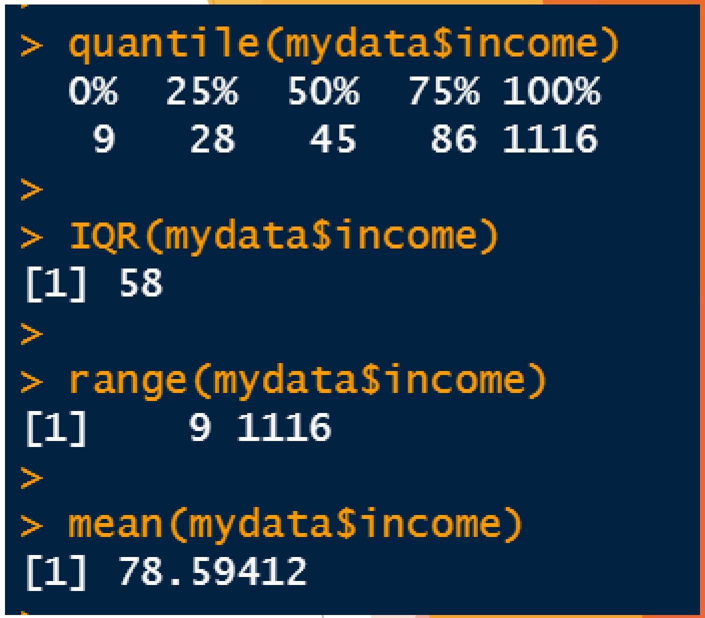


Outliers

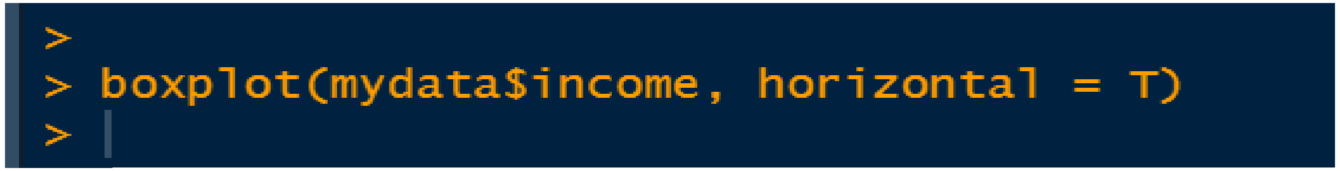


Low outliers: val< Q1- 1.5* IQR

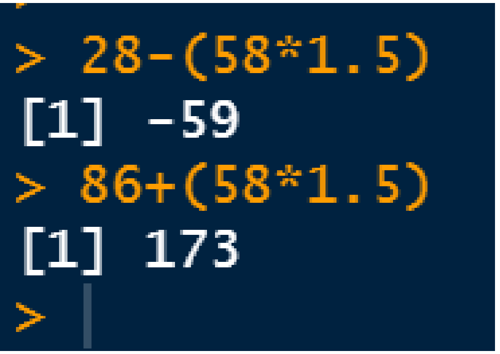
High outliers: val> Q3+1.5*IQR



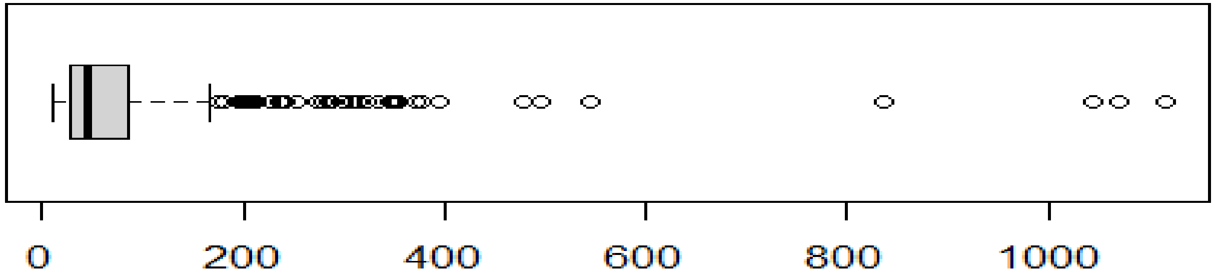
```
> quantile(mydata$income)
   0%   25%   50%   75% 100%
    9    28    45    86 1116
>
> IQR(mydata$income)
[1] 58
>
> range(mydata$income)
[1]    9 1116
>
> mean(mydata$income)
[1] 78.59412
```

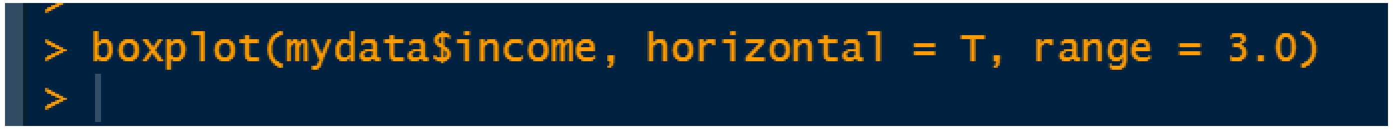


```
>
> boxplot(mydata$income, horizontal = T)
> 
```

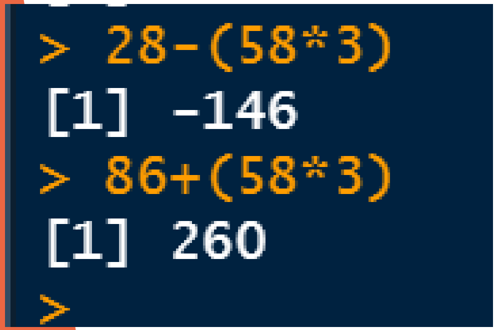


```
> 28-(58*1.5)
[1] -59
> 86+(58*1.5)
[1] 173
> 
```

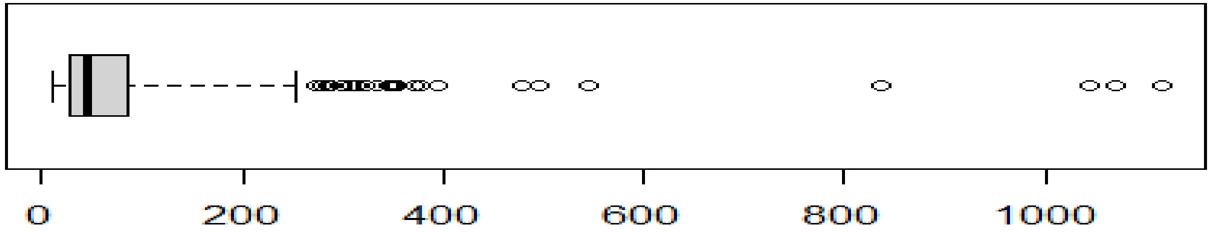


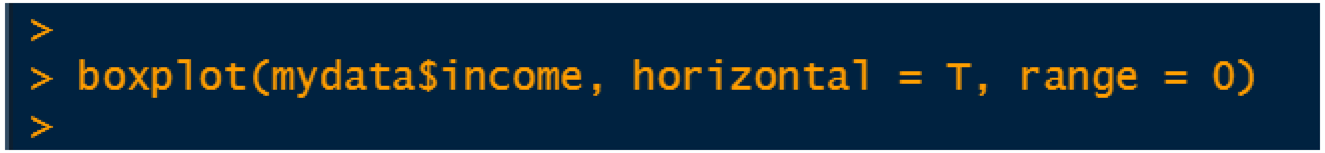


```
> boxplot(mydata$income, horizontal = T, range = 3.0)
> 
```

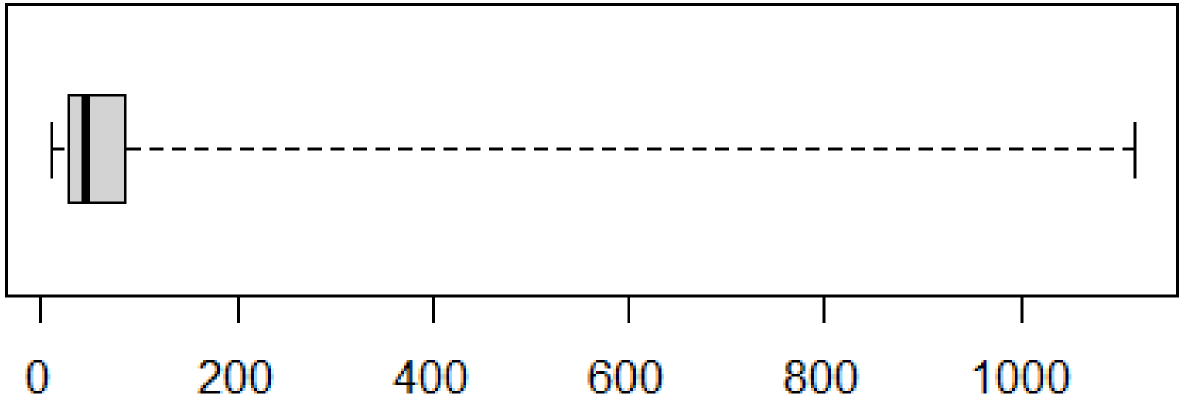


```
> 28-(58*3)
[1] -146
> 86+(58*3)
[1] 260
>
```

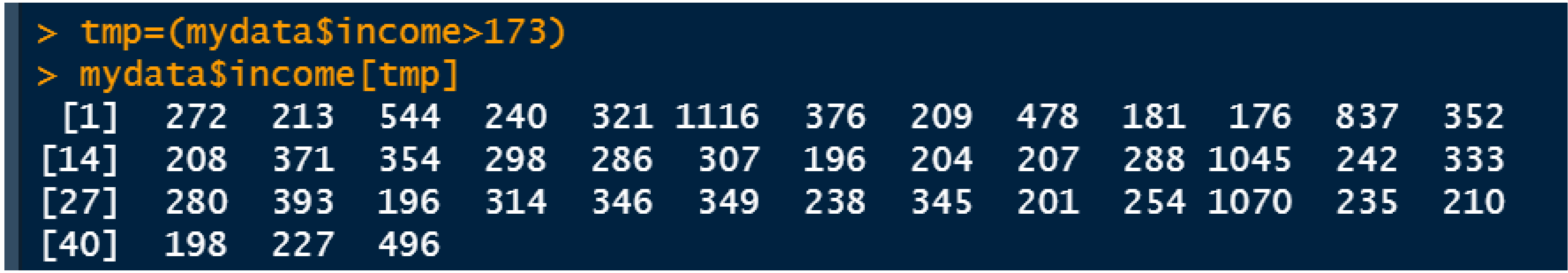




```
>
> boxplot(mydata$income, horizontal = T, range = 0)
>
```

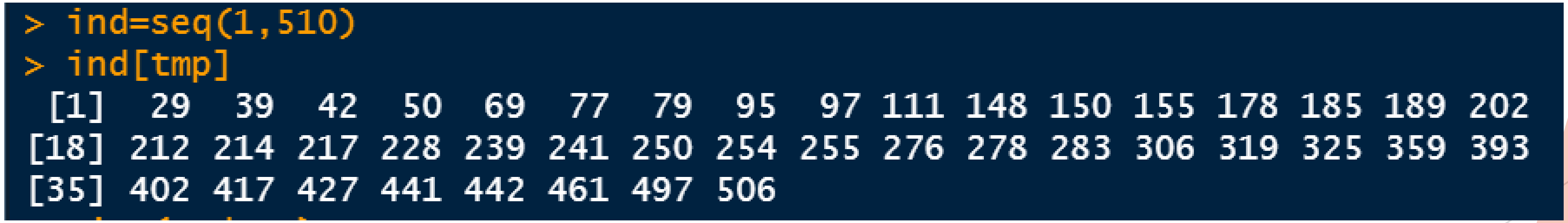


Valores mayores a 173

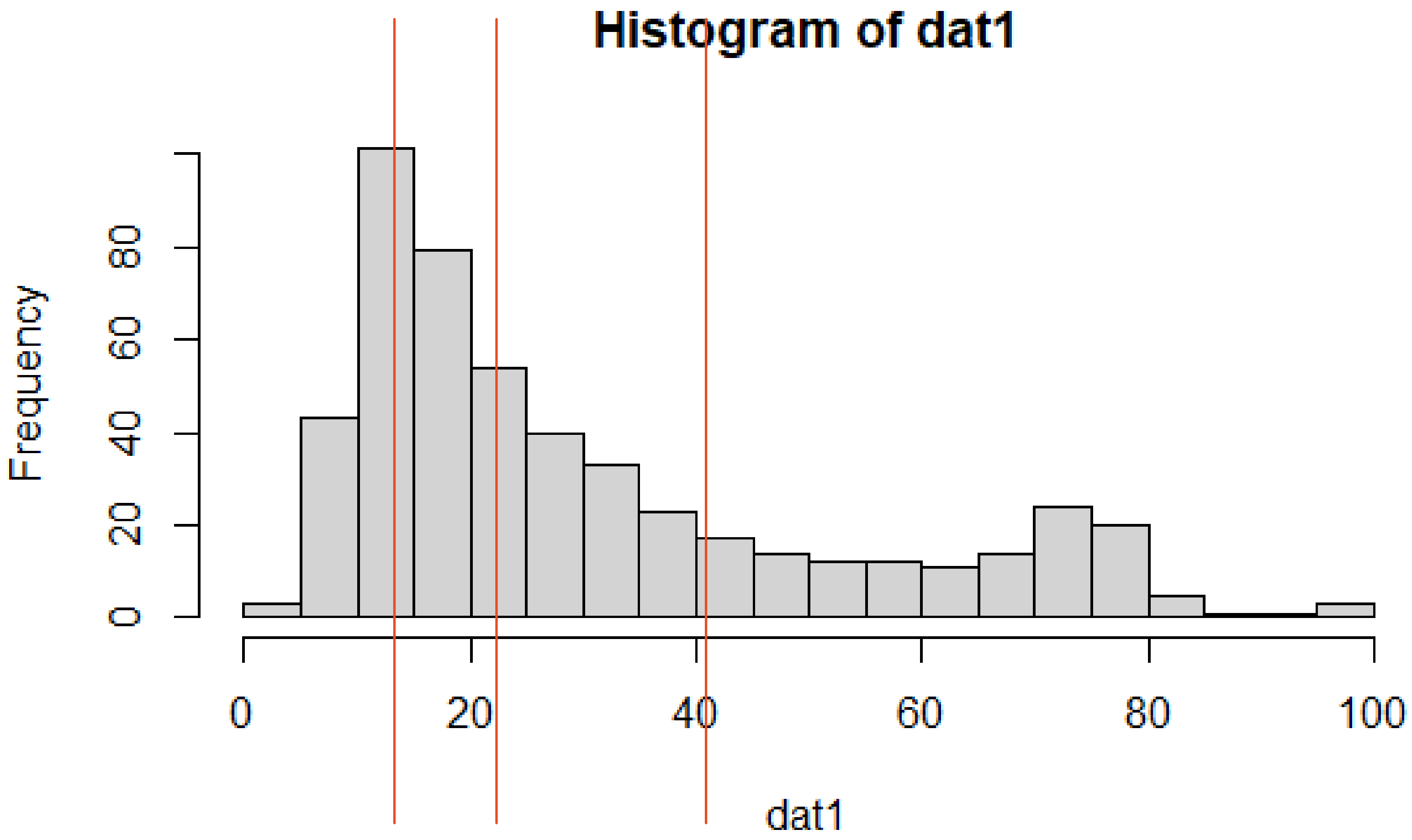


```
> tmp=(mydata$income>173)
> mydata$income[tmp]
 [1]  272  213  544  240  321 1116  376  209  478  181  176  837  352
[14]  208  371  354  298  286  307  196  204  207  288 1045  242  333
[27]  280  393  196  314  346  349  238  345  201  254 1070  235  210
[40]  198  227  496
```

Indices de valores mayores a 173



```
> ind=seq(1,510)
> ind[tmp]
 [1]  29  39  42  50  69  77  79  95  97 111 148 150 155 178 185 189 202
[18] 212 214 217 228 239 241 250 254 255 276 278 283 306 319 325 359 393
[35] 402 417 427 441 442 461 497 506
```


Histogram of dat1
Frequency
0
20
40
60
80
0
20
40
60
80
100
dat1