ASSIGNMENT 2: EXPLORATORY DATA ANALYSIS

Author: Juan Francisco H.dez F.dez

Part One.

Loading the dataset

```
datos = read.table(file = "auto-mpg.data")
```

Data cleaning and preparation

```
str(datos)
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ V1: num  18 15 18 16 17 15 14 14 14 15 ...
##  $ V2: int  8 8 8 8 8 8 8 8 8 8 ...
##  $ V3: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ V4: chr  "130.0" "165.0" "150.0" "150.0" ...
##  $ V5: num  3504 3693 3436 3433 3449 ...
##  $ V6: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ V7: int  70 70 70 70 70 70 70 70 70 70 ...
##  $ V8: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ V9: chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst" ...
```

```
names(datos) = c( "mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "model year", "orig
in", "car name")
```

We have relabelled the data, as it was too unintuitive to read the information contained therein as it was labelled V1-V9. This way we now know what each of the integers represents.

```
datos$horsepower <- as.numeric(as.character(datos$horsepower))
```

```
## Warning: NAs durch Umwandlung erzeugt
```

Because the horsepower variable was stored as a string, and we may need to operate on it later, let's convert it to an integer.

```
str(datos)
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model year  : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ car name    : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst" ...
```

```
datos <- na.omit(datos)
```

For the same reason as mentioned above, we have omited the possible NA contained in the columns of our dataset, since if it were necessary to operate with this data, the results we would obtain would be erroneous.

```
summary(datos)
```

```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140  
##   acceleration     model year        origin        car name        
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392        
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character  
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character  
##  Mean   :15.54   Mean   :75.98   Mean   :1.577                     
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000                     
##  Max.   :24.80   Max.   :82.00   Max.   :3.000                     
```

Here is a summary of our dataset, after we have cleaned and prepared the data. We are now ready to work with the data.

Part two.

Let's answer some questions related to the data

1.What year is the oldest car?

```
order(datos$`model year`)
```

```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##  [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##  [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##  [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##  [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
## [109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
## [127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
## [145] 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
## [163] 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
## [181] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
## [199] 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
## [217] 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
## [235] 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
## [253] 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
## [271] 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
## [289] 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
## [307] 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324
## [325] 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342
## [343] 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360
## [361] 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378
## [379] 379 380 381 382 383 384 385 386 387 388 389 390 391 392
```

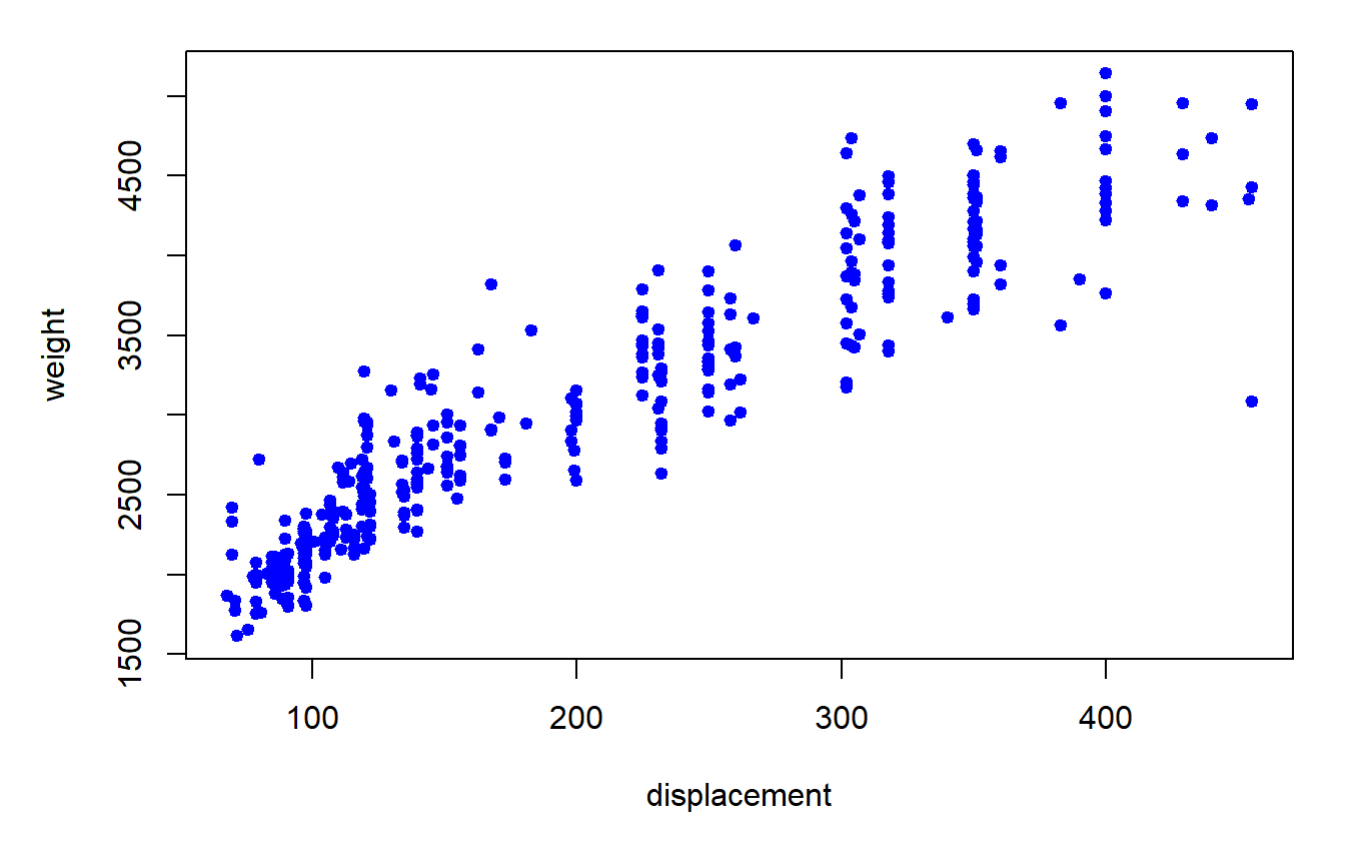
```
head(datos$`model year`)[1]
```

```
## [1] 70
```

As we can see, the oldest car in our dataset dates back to 1970.

2.The larger the displacement, the heavier the weight?

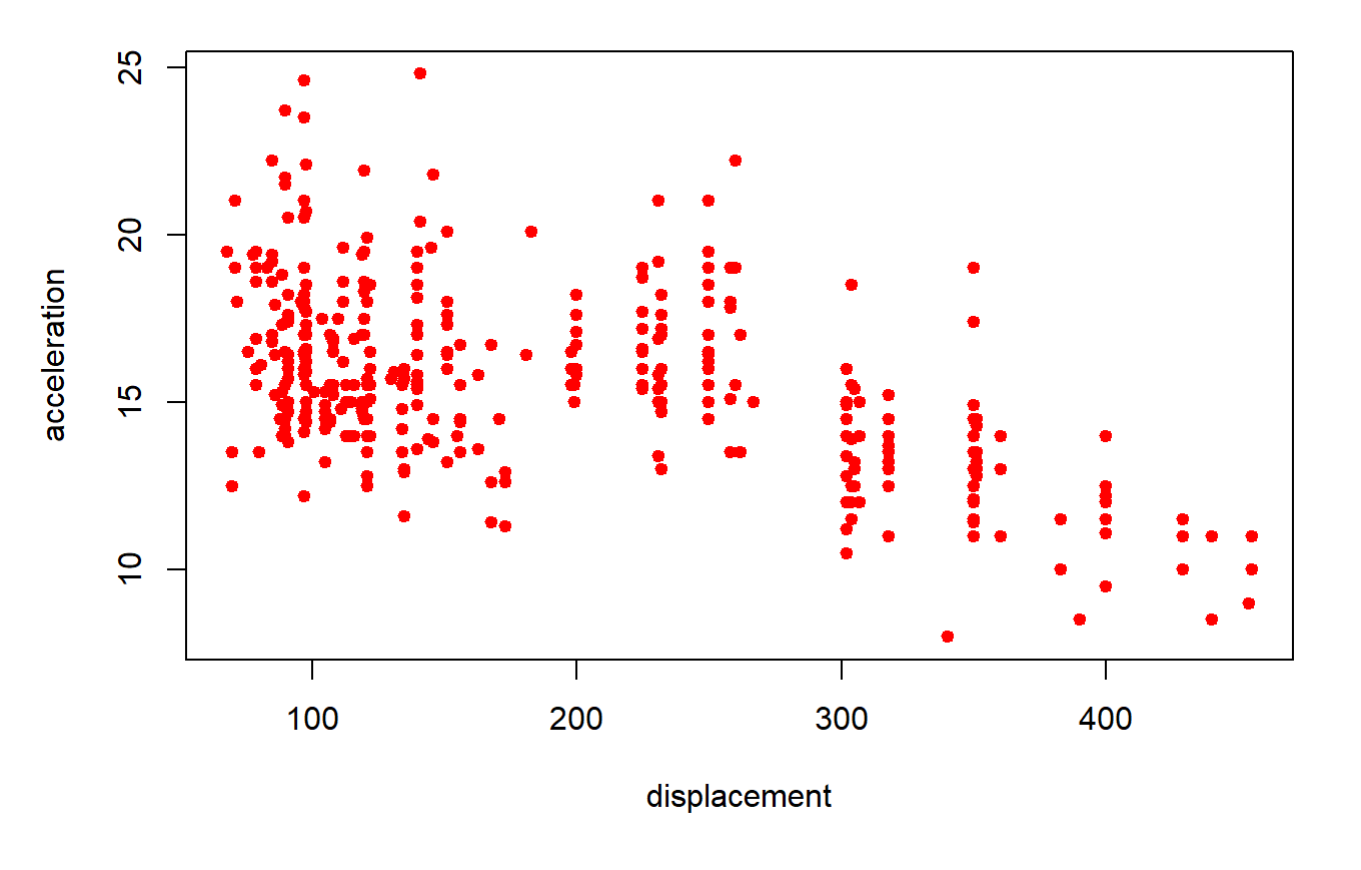
```
plot(x = datos$displacement, y = datos$weight , col = "blue", xlab = "displacement",
     ylab = "weight", type = 'p', lwd = 2, pch=20)
```



As we can see in the graph, there is a relationship between weight and cylinder capacity, the greater the weight, the greater the cylinder capacity of the vehicle. We can also observe that there are hardly any cars registered in our dataset that have displacement 3 and 5, and none that have displacement 7.The rest of the categories seem to have a similar number of samples.

3.More displacement, more acceleration?

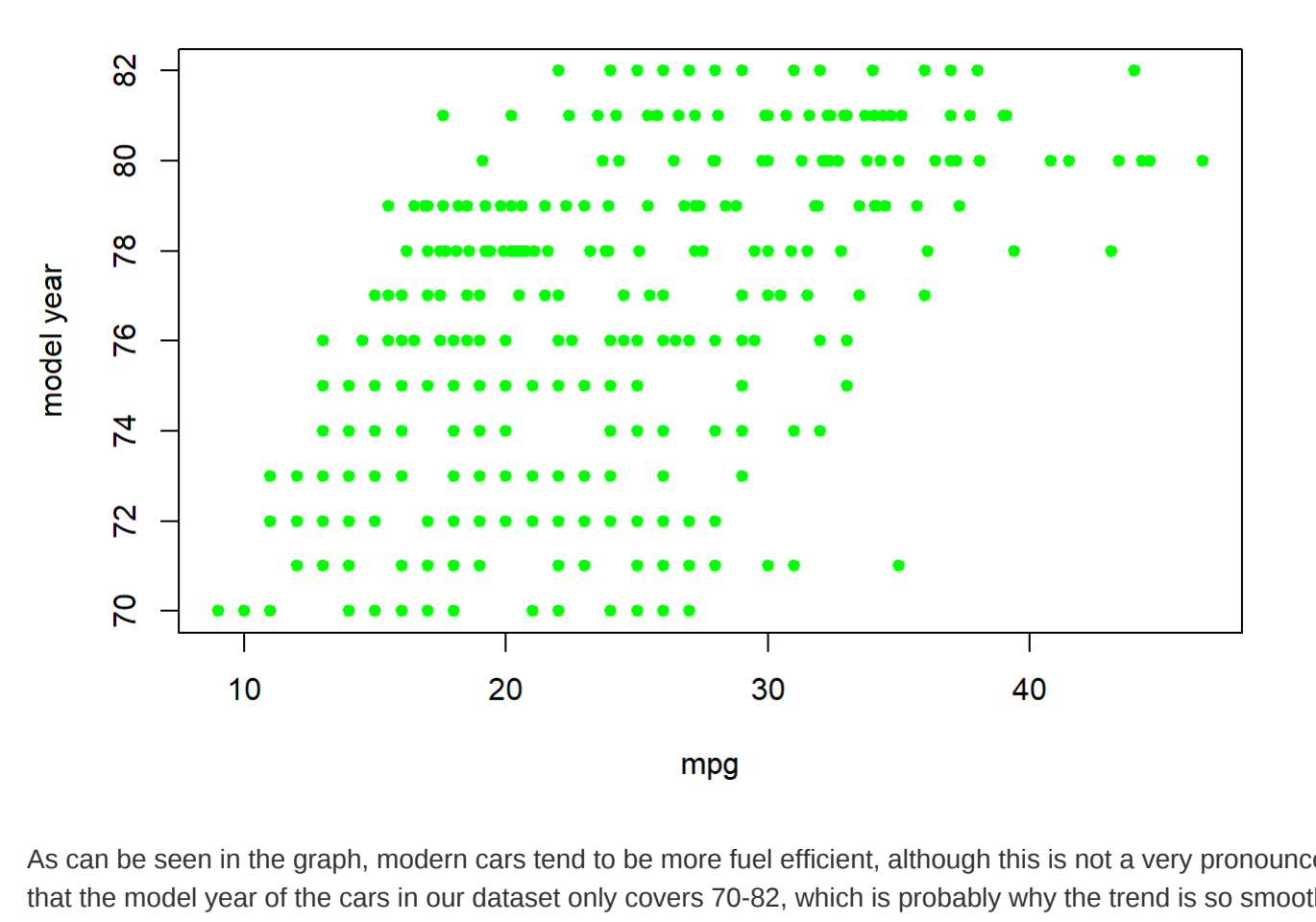
```
plot(x = datos$displacement, y = datos$acceleration , xlab = "displacement",
     ylab = "acceleration", col = "red",type = 'p', lwd = 2, pch=20)
```



Looking at the data we can see that the larger the displacement the less acceleration, this may seem strange since in principle the displacement increases the power of the engine, however if we consider the above graph we can conclude that because the weight of the vehicle also increases with the displacement, this power does not end up translating into more acceleration.

4.Do modern cars consume less fuel?

```
plot(x = datos$mpg, y = datos$`model year`, xlab = "mpg", ylab = "model year",
     col = "green",type = 'p', lwd = 2, pch=20)
```



As can be seen in the graph, modern cars tend to be more fuel efficient, although this is not a very pronounced trend either. It should also be noted that the model year of the cars in our dataset only covers 70-82, which is probably why the trend is so smooth, perhaps if we had more current data we would see a more pronounced trend.

5.How much does a car consume on average?

```
mean(datos$mpg)
```

```
## [1] 23.44592
```

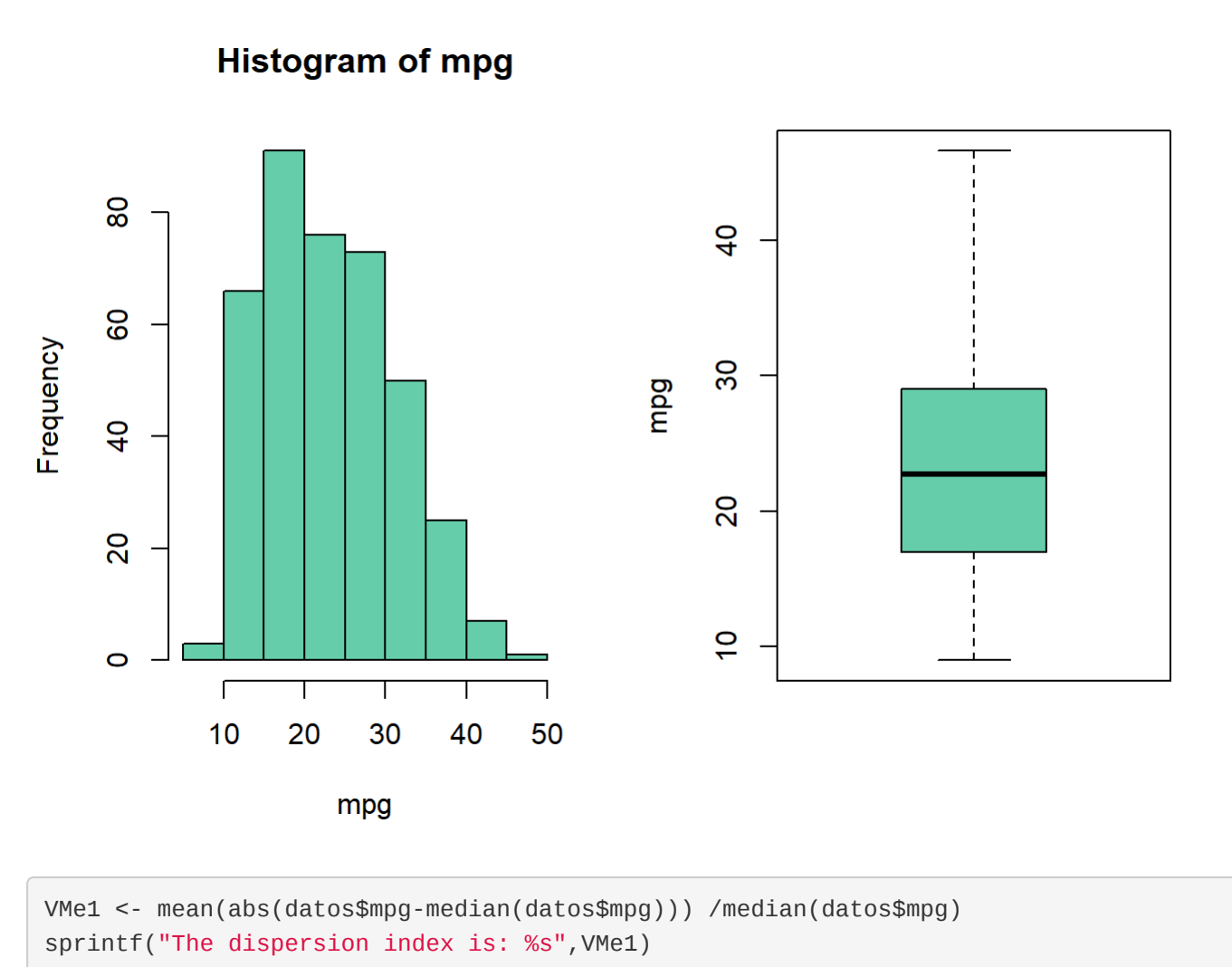
To find this data we simply calculated the arithmetic mean using the mean function. The result obtained was 23.44592, so cars registered from 1970 to 1982 consumed on average 23.44592 mpg.

Part three.

The variables

MPG

```
par(mfrow = c(1, 2))
hist(datos$mpg, xlab = "mpg", main = "Histogram of mpg", col = ("aquamarine3") )
boxplot(datos$mpg, ylab = "mpg", col = ("aquamarine3"))
```




```
VMe1 <- mean(abs(datos$mpg-median(datos$mpg))) /median(datos$mpg)
sprintf("The dispersion index is: %s",VMe1)
```

```
## [1] "The dispersion index is: 0.286768333707109"
```

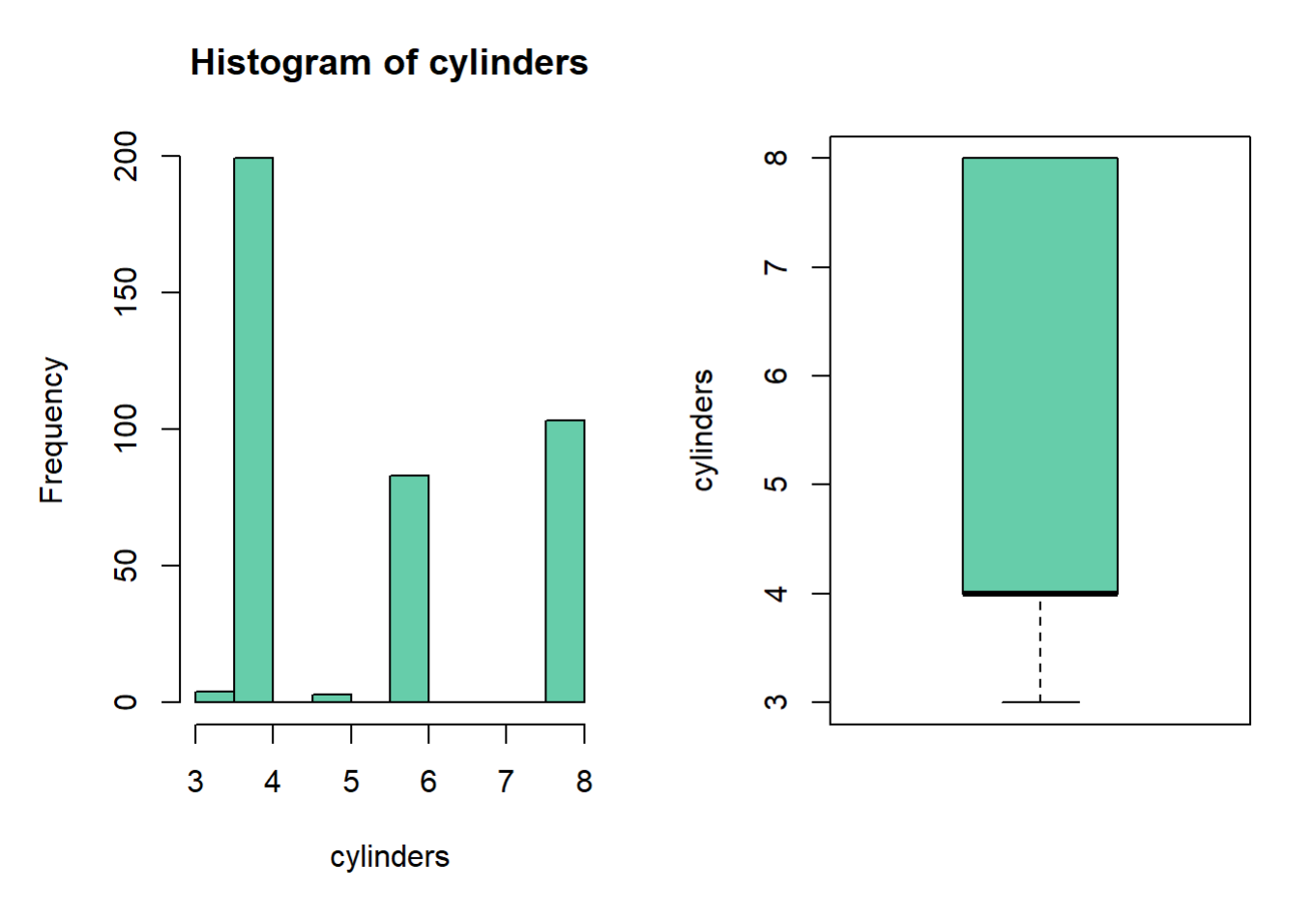
To check the dispersion of the variables we use the dispersion index with respect to the median. This index measures the number of times that the median is contained in the deviation with respect to the median of the distribution, so that larger values indicate a greater dispersion and therefore a lower representativeness of the median. The formula for calculating this index is as follows:

$$VMe = \frac{\sum |Xi - Me| * ni}{N * Me}.$$

Applying this formula we have obtained a dispersion index of 0.286768333707109 with respect to the median. We can therefore say that the data have little dispersion. As for the centre, if we look at the boxplot, we can see that the mean of the data is around 23 mpg, as we had obtained previously. As for the symmetry of the data, looking at the histogram we can see that the representation of the data is quite symmetrical, some asymmetries can be seen, but not too considerable. It seems that the data follow a normal distribution.

Cylinders

```
par(mfrow = c(1, 2))
hist(datos$cylinders, xlab = "cylinders", main = "Histogram of cylinders", col = ("aquamarine3"))
boxplot(datos$cylinders, ylab = "cylinders", col = ("aquamarine3"))
```



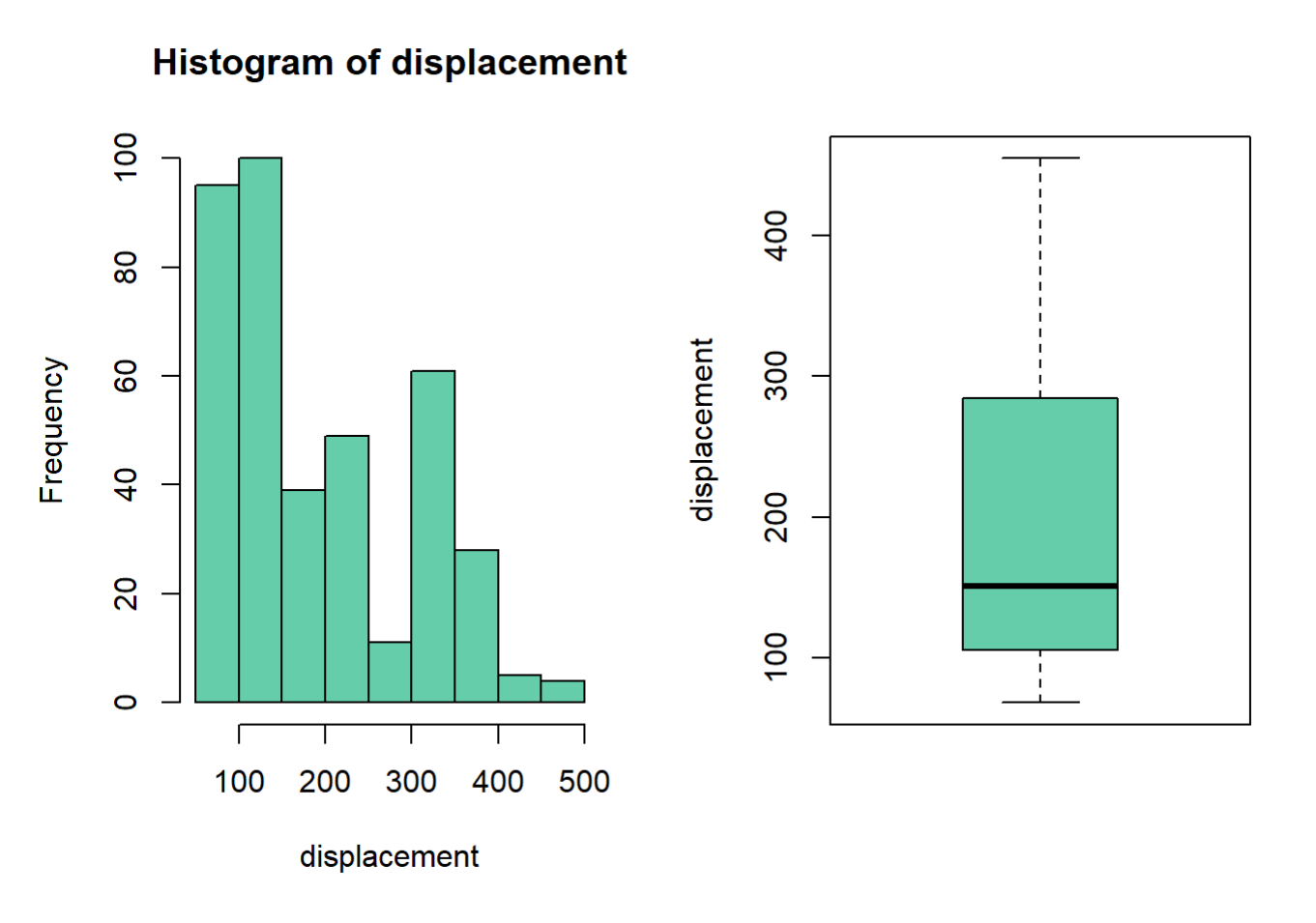

```
VMe2 <- mean(abs(datos$cylinders-median(datos$cylinders))) /median(datos$cylinders)
sprintf("The dispersion index is: %s",VMe2)
```

```
## [1] "The dispersion index is: 0.373086734693878"
```

In this case we are faced with totally asymmetric data, with a lot of data in some classes and very little or no data in others. In the boxplot we can see that the mean revolves around 4 cylinders, although this is also obvious in the histogram as practically all the samples belong to this class. In this case, the dispersion index with respect to the median is 0.373086734693878.

Displacement

```
par(mfrow = c(1, 2))
hist(datos$displacement, xlab = "displacement", main = "Histogram of displacement"
     , col = ("aquamarine3"))
boxplot(datos$displacement, ylab = "displacement", col = ("aquamarine3"))
```



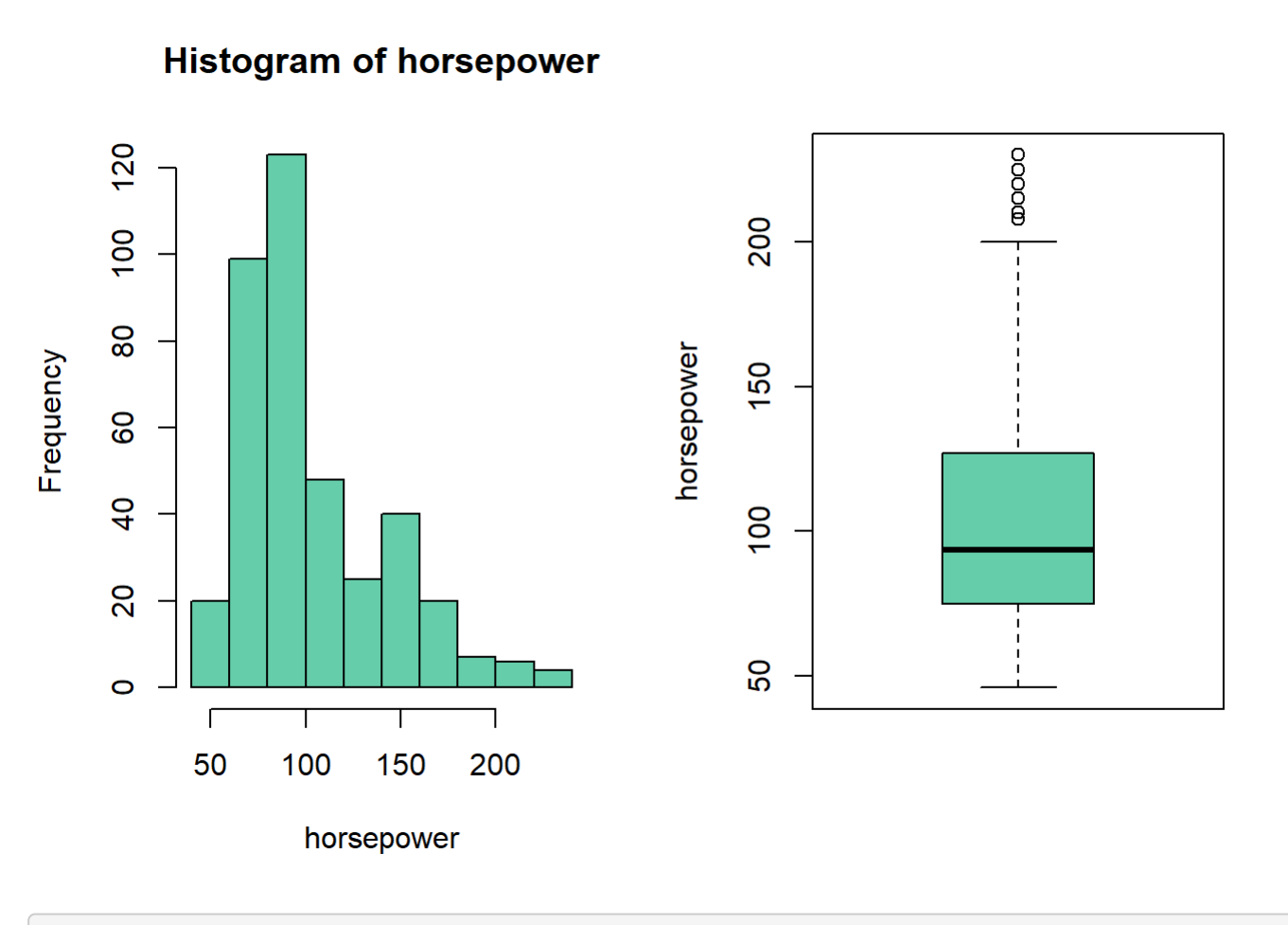

```
VMe3 <- mean(abs(datos$displacement-median(datos$displacement))) /median(datos$displacement)
sprintf("The dispersion index is: %s",VMe3)
```

```
## [1] "The dispersion index is: 0.578904243816732"
```

Looking at the histogram we see that the data are quite asymmetrical, the number of samples of low displacement is much higher than those of high displacement, with some intermediate peaks. The mean of the data is around 160, and in this case we see that the index of dispersion of the data with respect to the median has is high, in this case we have 0.578904243816732, this may be due to the asymmetry of the data, median representation is low.

Horsepower

```
par(mfrow = c(1, 2))
hist(datos$horsepower, xlab = "horsepower", main = "Histogram of horsepower"
     , col = ("aquamarine3"))
boxplot(datos$horsepower, ylab = "horsepower", col = ("aquamarine3"))
```



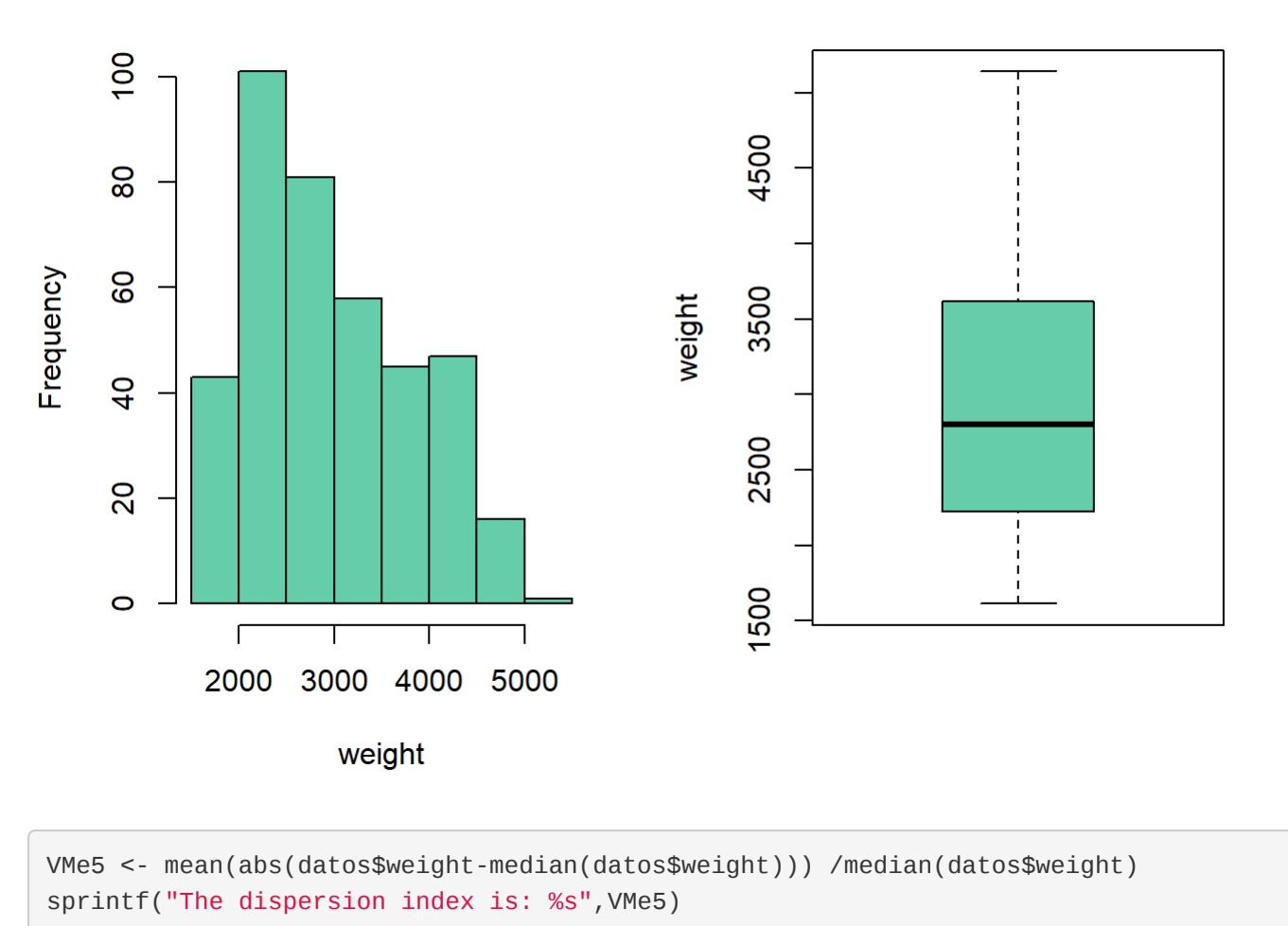

```
VMe4 <- mean(abs(datos$horsepower-median(datos$horsepower))) /median(datos$horsepower)
sprintf("The dispersion index is: %s",VMe4)
```

```
## [1] "The dispersion index is: 0.306777256357088"
```

In this case, if we look at the histogram, we see a distribution more similar to that of the horsepower, although in this case it is somewhat more asymmetrical. In the boxplot graph we can see that the average is around 90 horsepower, in this case we can see how we find in the dataset samples with outliers in very high displacements. As for the asymmetry of the data, as we have already said, we have a value closer to that of the horsepower, the index is 0.30677725256357088.

Weight

```
par(mfrow = c(1, 2))
hist(datos$weight, xlab = "weight", main = "Histogram of weight"
     , col = ("aquamarine3"))
boxplot(datos$weight, ylab = "weight", col = ("aquamarine3"))
```



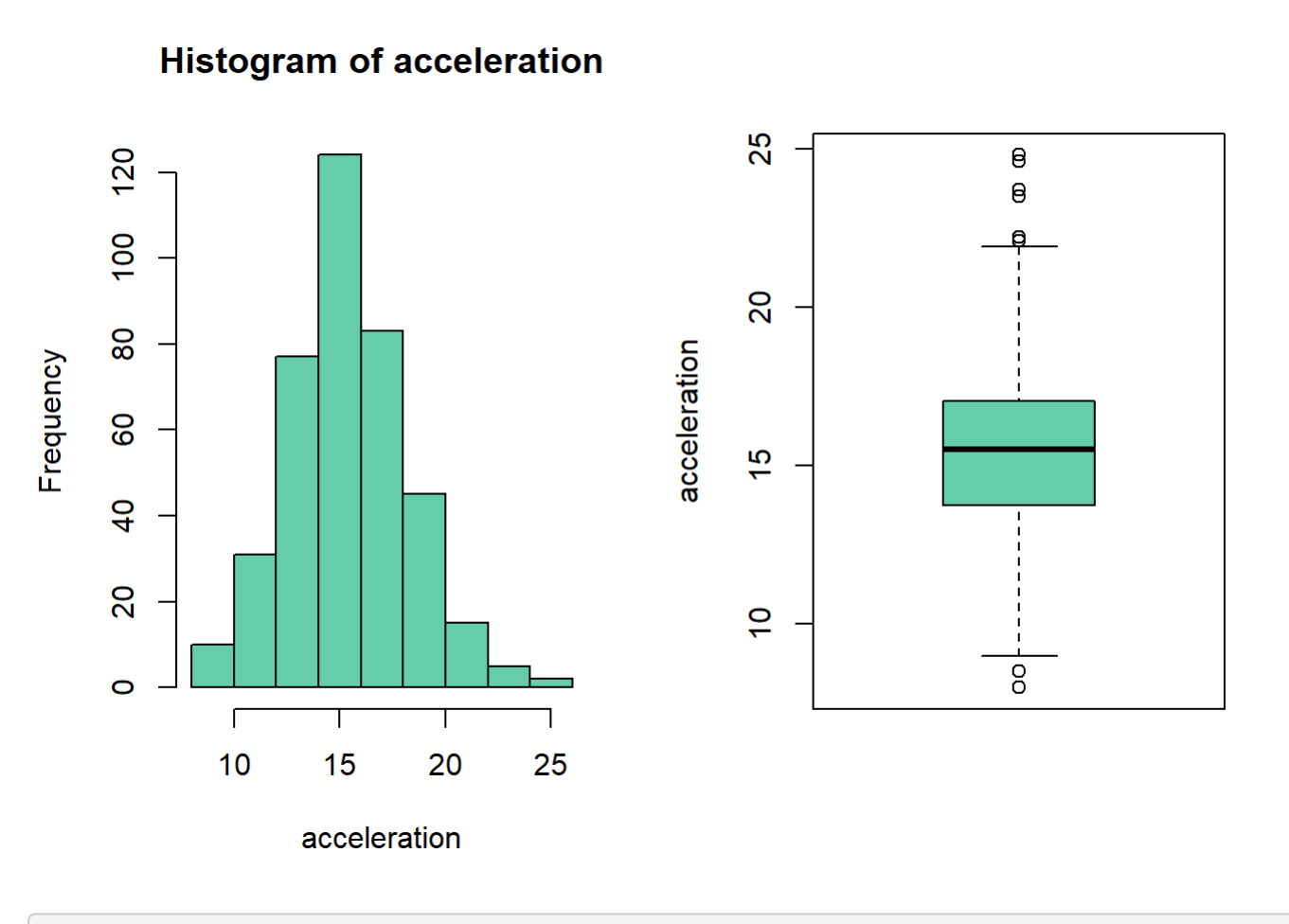

```
VMe5 <- mean(abs(datos$weight-median(datos$weight))) /median(datos$weight)
sprintf("The dispersion index is: %s",VMe5)
```

```
## [1] "The dispersion index is: 0.254012841091493"
```

If we look at the histogram we can see that the distribution of the weights of the cars in our samples is close to a normal distribution, with a mean according to the boxplot of about 2700 kg. The dispersion index in this case is even lower than the previous one as we find more samples contained in the median.

Acceleration

```
par(mfrow = c(1, 2))
hist(datos$acceleration, xlab = "acceleration", main = "Histogram of acceleration"
     , col = ("aquamarine3"))
boxplot(datos$acceleration, ylab = "acceleration", col = ("aquamarine3"))
```



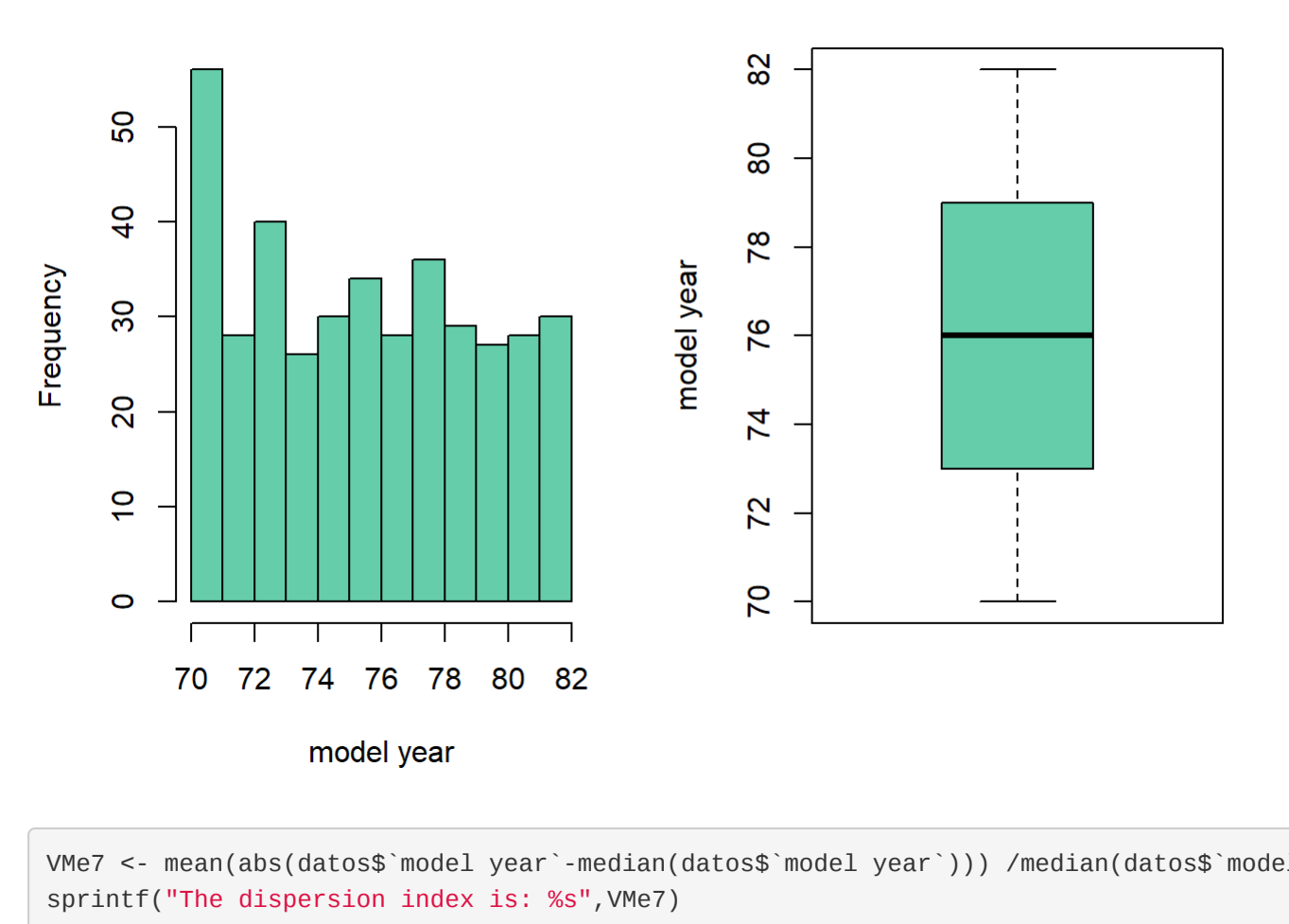

```
VMe6 <- mean(abs(datos$acceleration-median(datos$acceleration))) /median(datos$acceleration)
sprintf("The dispersion index is: %s",VMe6)
```

```
## [1] "The dispersion index is: 0.137820934825543"
```

The acceleration data present a normal distribution, fairly symmetrical and with an average of around 16. The dispersion index is 0.137820934825543.

Model year

```
par(mfrow = c(1, 2))
hist(datos$`model year`, xlab = "model year", main = "Histogram of model year"
     , col = ("aquamarine3"))
boxplot(datos$`model year`, ylab = "model year"
        , col = ("aquamarine3"))
```



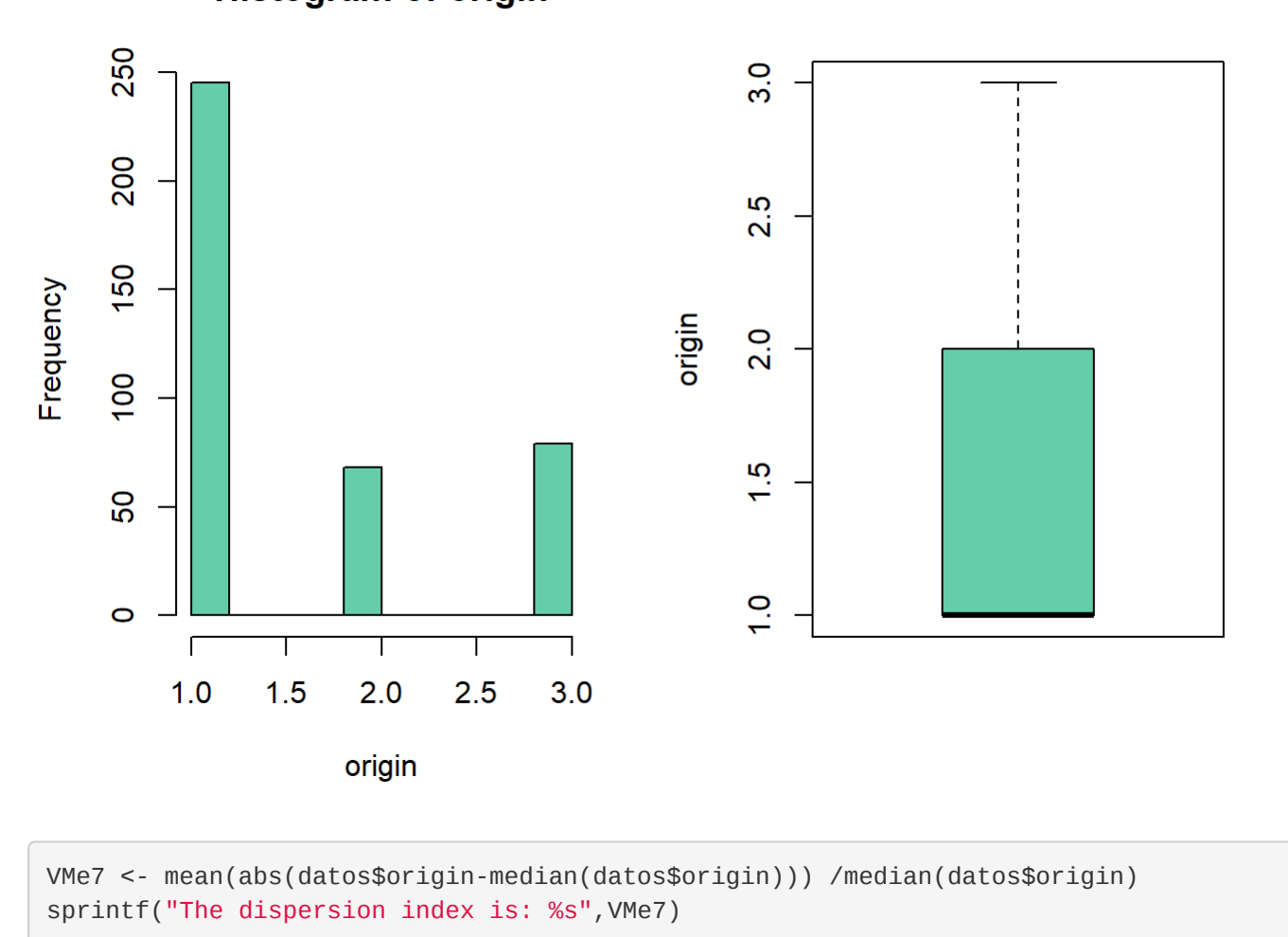

```
VMe7 <- mean(abs(datos$`model year`-median(datos$`model year`))) /median(datos$`model year`)
sprintf("The dispersion index is: %s",VMe7)
```

```
## [1] "The dispersion index is: 0.0415547798066595"
```

As for the model years of our samples, except for a peak in the 1970s, we can see that the distribution is uniform, with the mean according to the boxplot around the year 1976. And the dispersion index obtained is quite low 0.0415547798066595, so the representation of the median is considerable.

Origin

```
par(mfrow = c(1, 2))
hist(datos$origin, xlab = "origin", main = "Histogram of origin"
     , col = ("aquamarine3"))
boxplot(datos$origin, ylab = "origin", col = ("aquamarine3"))
```




```
VMe7 <- mean(abs(datos$origin-median(datos$origin))) /median(datos$origin)
sprintf("The dispersion index is: %s",VMe7)
```

```
## [1] "The dispersion index is: 0.576530612244898"
```

Finally, with respect to the origin, we find very asymmetric data and a rather high dispersion index 0.576530612244898, therefore with a low representation of the median. According to the boxplot graph, the mean is around 1.0, and this is evident simply by looking at the histogram, as the asymmetry is notably towards this value.

Correlation matrix

```
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(datos[1:8]), method = "number")
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
mpg	1.00	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
cylinders	-0.78	1.00	0.95	0.84	0.90	-0.50	-0.35	-0.57
displacement	-0.81	0.95	1.00	0.90	0.93	-0.54	-0.37	-0.61
horsepower	-0.78	0.84	0.90	1.00	0.86	-0.69	-0.42	-0.46
weight	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31	-0.59
acceleration	0.42	-0.50	-0.54	-0.69	-0.42	1.00	0.29	0.21
model year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1.00	0.18
origin	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1.00

The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. The correlation coefficient is a unitless value between -1 and 1. The closer r is to zero, the weaker the linear relationship. Positive values indicate a positive correlation, where the values of both variables tend to increase together. Negative values indicate a negative correlation, where the values of one variable tend to increase while the values of the other variable decrease. Values of 1 and -1 represent a positive and negative "perfect" correlation, respectively. Here we can see for example, as we have seen before, that the ratio between displacement and weight is high. On the other hand we can see that the model year and the weight are hardly related. This matrix is very useful because at a glance we can see which variables are strongly related and which are hardly related.