Let us consider the dataframe mtcars, which comprises the fuel consumption and 10 aspects of design and performance for 32 automobiles (1970s models). The help file is given below

Describe how to perform a preliminary data analysis on this dataframe, using suitable R commands.

### Scopo e analisi delle variabili

Si vuole creare un modello statistico di regressione per analizzare i consumi medi di alcuni modelli di auto presenti nel dataframe mtcars.

La variabile risposta è mpg che rappresenta il numero di miglia (americane) percorse (in media) con un gallone di carburante.

Analizziamo le variabili

str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

Ci sono alcune variabili da convertire

mtcars$cyl <- as.integer(mtcars$cyl)  
mtcars$hp <- as.integer(mtcars$hp)  
mtcars$vs <- factor(mtcars$vs, labels = c("V-shaped", "straight"))  
mtcars$am <- factor(mtcars$am, labels = c("automatic", "manual"))  
mtcars$gear <- as.integer(mtcars$gear)  
mtcars$carb <- as.integer(mtcars$carb)  
str(mtcars)

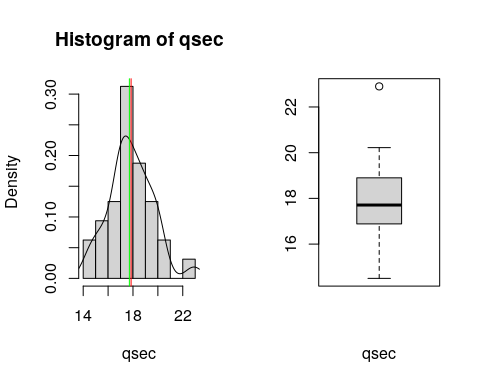
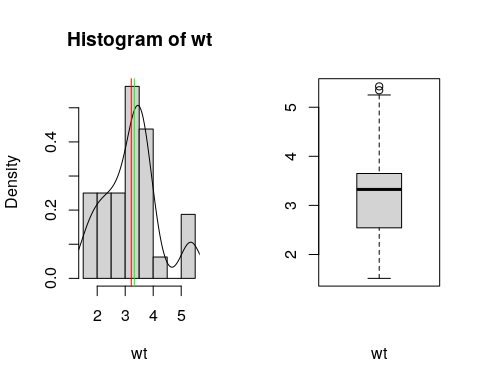
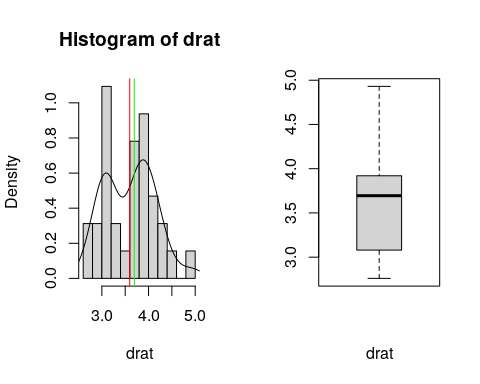
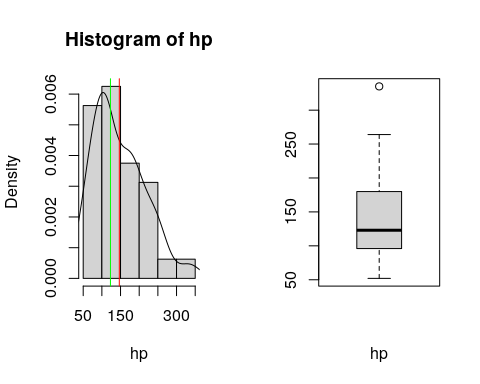
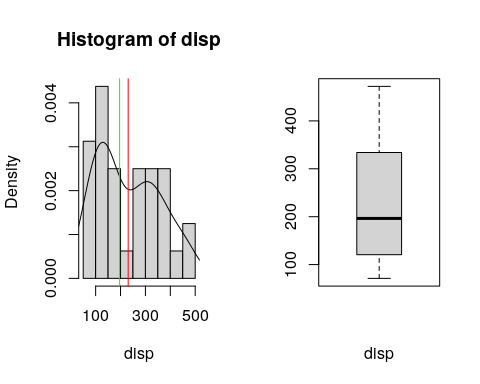
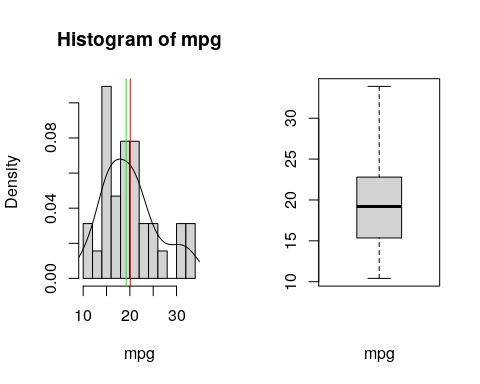
## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : int 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : int 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : Factor w/ 2 levels "V-shaped","straight": 1 1 2 2 1 2 1 2 2 2 ...  
## $ am : Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...  
## $ gear: int 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: int 4 4 1 1 2 1 4 2 2 4 ...

summary(mtcars)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs am   
## Min. :2.760 Min. :1.513 Min. :14.50 V-shaped:18 automatic:19   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 straight:14 manual :13   
## Median :3.695 Median :3.325 Median :17.71   
## Mean :3.597 Mean :3.217 Mean :17.85   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90   
## Max. :4.930 Max. :5.424 Max. :22.90   
## gear carb   
## Min. :3.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:2.000   
## Median :4.000 Median :2.000   
## Mean :3.688 Mean :2.812   
## 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :5.000 Max. :8.000

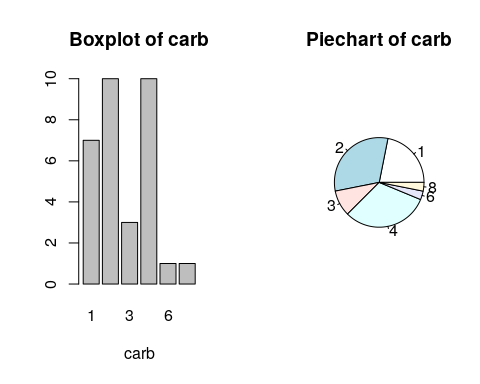
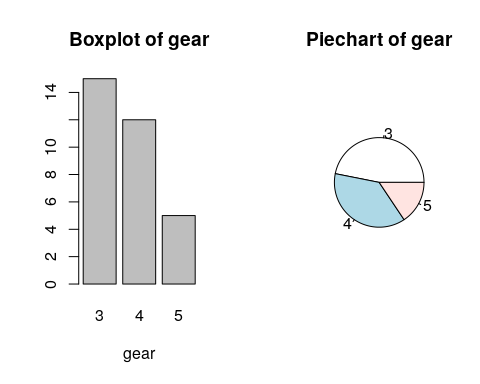
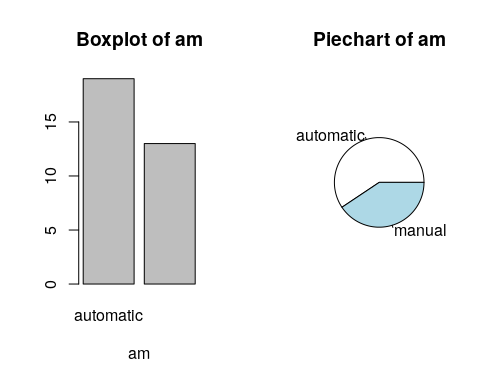
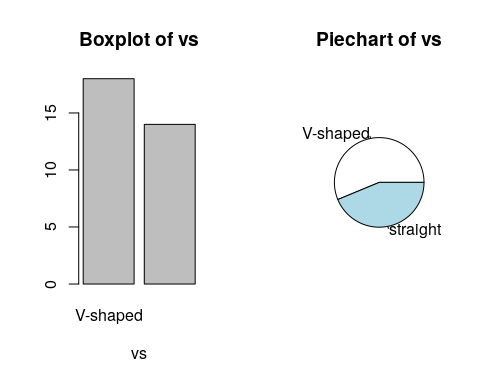
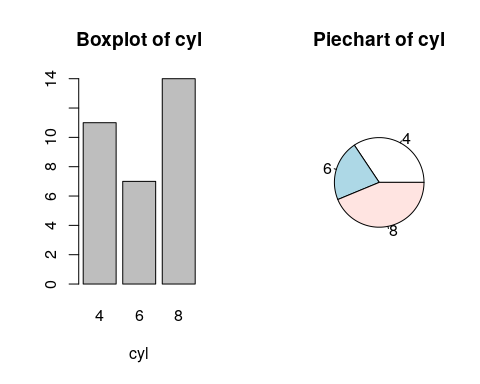
### Distribuzione delle variabili

source("./../functions.R", local = knitr::knit\_global())  
  
numeric.graph(mtcars, c(1, 3, 4, 5, 6, 7))



## [1] "Skewness indexes"  
## mpg disp hp drat wt qsec   
## 0.6404399 0.4002724 0.7614356 0.2788734 0.4437855 0.3870456   
## [1] "3 - Kurtosis indexes"  
## mpg disp hp drat wt qsec   
## 0.20053321 1.08968266 -0.05223273 0.56488390 -0.17247054 -0.55375334

categorical.graph(mtcars, c(2, 8, 9, 10, 11))

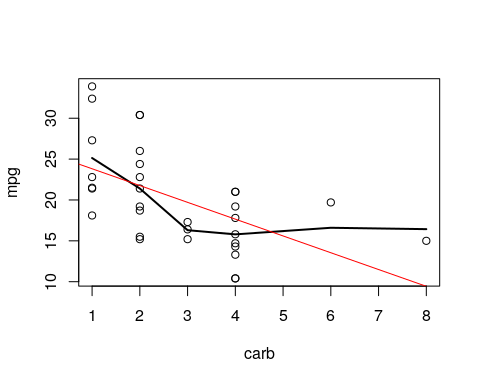
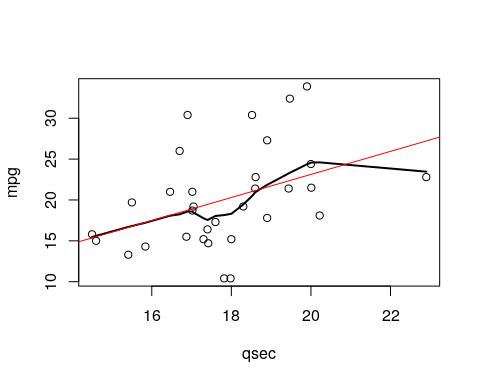
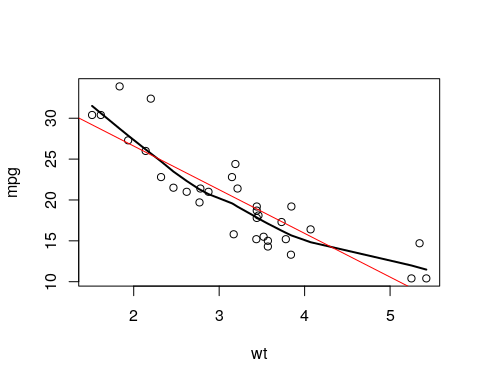
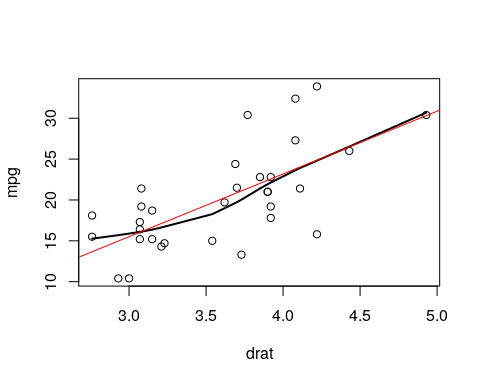
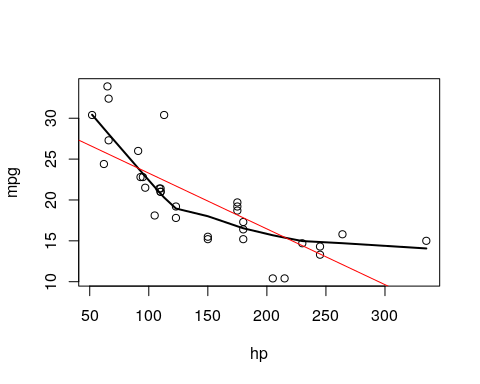
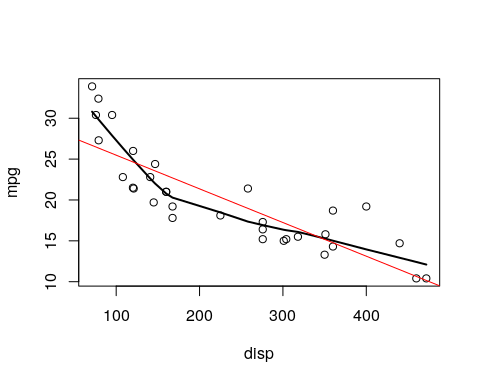
 #### Osservazioni

* Non si rilevano particolari outlier se non per hp e qsec che ne hanno uno, mentre wt ne presenta due molto vicini al baffo superiore.
* Per le variabili numeriche si nota una generale assimmetria verso destra (più o meno accentuata a seconda della variabile in analisi).
* Gli indici di curtosi mostrano una buona normalità delle curve per quanto riguarda hp, wt e mpg.
* Nonostante a prima vista qsec sembra la variabile migliore dal punto di vista della simmetria e della curtosi, i valori di questi indici sono probabilmente influenzati dall’outlier
* Tra le variabili categoriali si può notare, in generale, una buona distribuzione di queste ultime nel dataset.
* Discorso a parte vale per carb, dove ci sono due valori (estremi) con un campione ciascuno.

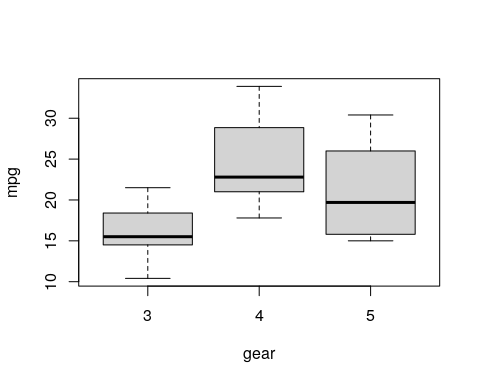
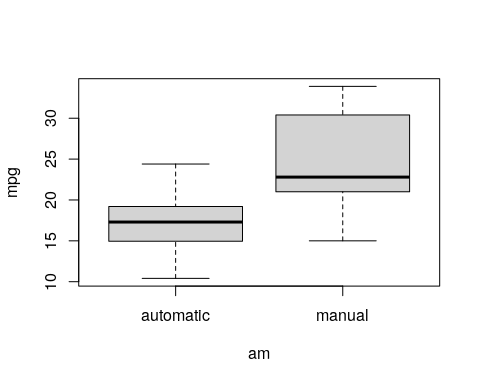
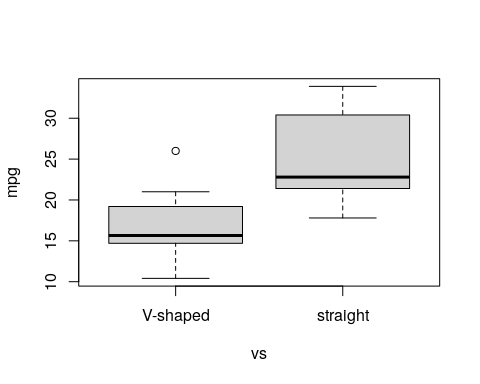
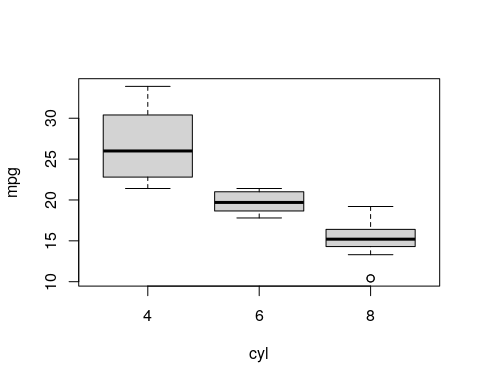
### Correlazioni tra variabili

### Correlazioni tra la variabile risposta (mpg) e le altre variabili

scatterplots.graph(1, c(3, 4, 5, 6, 7, 11), mtcars)



boxplots.graph(1, c(2, 8, 9, 10), mtcars)

 #### Osservazioni Dai grafici emergono le seguenti correlazioni con mpg:

* con disp e hp, sembra esserci correlazione, ma non di tipo lineare;
* con wt, sembra esserci una correlazione di tipo lineare;
* con drat la curva lowess suggerisce una correlazione lineare, tuttavia la distribuzione dei punti suggerisce che probabilmente è molto debole;
* con cyl, vs e am sembrano esserci correlazioni;
* non sembrano esserci correlazioni con qsec, carb e gear;

### Collinearità tra le variabili

Visto l’elevato numero di variabili esplicative sarebbe in caso di rivolgersi a esperti del settore per avere le seguenti informazioni:

* quali sono le variabili di maggiore importanza,
* quali potrebbero essere le collinearità presenti tra le variabili.