R Notebook

2.1

2.1.1 Basic Concepts

Simple practical problem:

`Identify a relationship that allows us to predict the consumption of fuel, or equivalently, the distanc`

Some of the data are numerical:

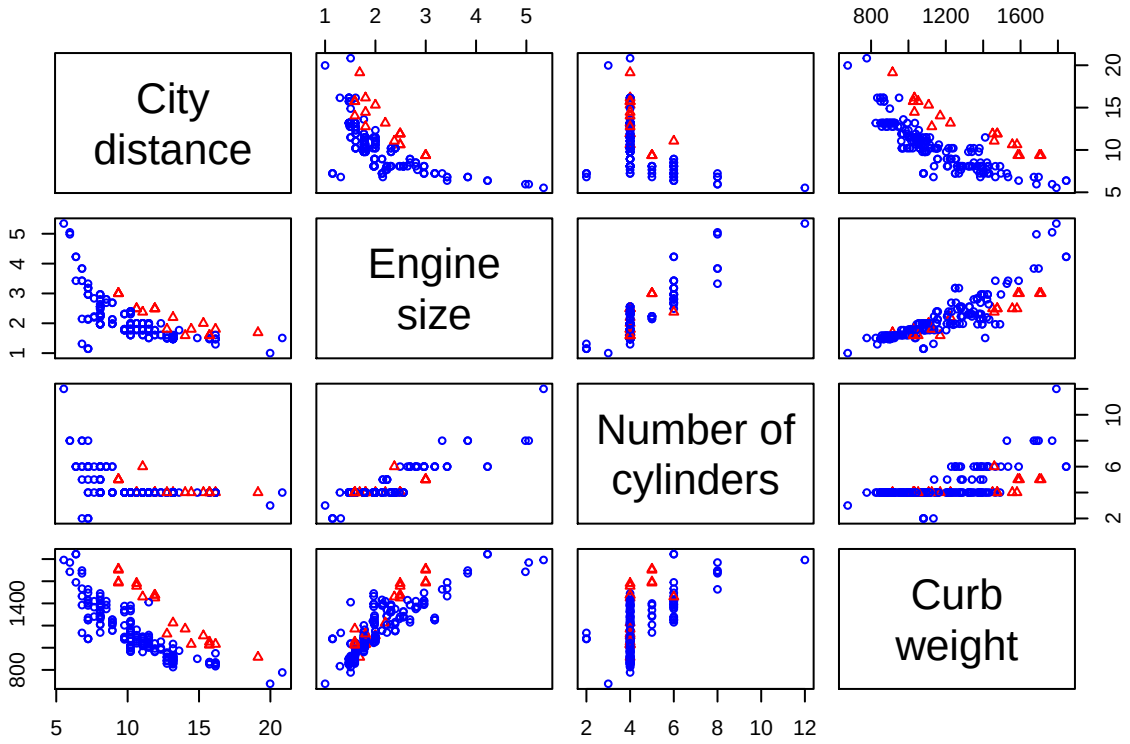
- Quantitative and continuous: City distance, engine size, and curb weight (kg)
- Quantitative and discrete: number of cylinders

Matrix of scaterplots of some variables of car data, stratified by fuel type.

```
auto <- read.table("http://azzalini.stat.unipd.it/Book-DM/auto.dat", header = TRUE)
attach(auto)
#summary(auto)
# Sample size
n = nrow(auto)

# Create dummy variable for fuel: disel = False, gasoline = TRUE
d = fuel == "gas"


# Scater-plot matrix
pairs(auto[ , c("city.distance", "engine.size","n.cylinders","curb.weight")],
      labels = c("City\ndistance", "Engine\nsize","Number of\ncylinders", "Curb\nweight"),
      col = ifelse(d, 'blue', 'red'), pch = ifelse(d, 1, 2), # 1 is a circle, 2 is a triangle
      cex = 10/sqrt(n))
```



Some are qualitative:

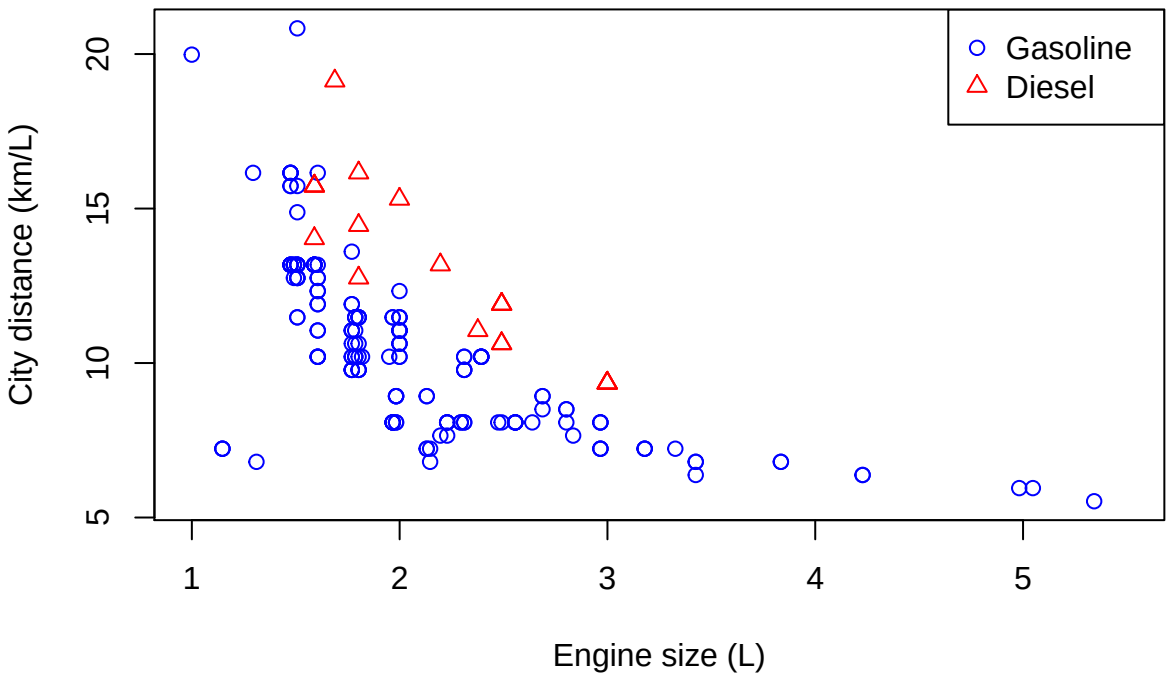
- fuel type: diesel and gasoline

We will in first phase consider only two explanatory variables:

- Engine size, fuel type

To study the relationship between quantitative variables we usually make a grafic representation.

```
### figure 2.2 ###
plot(engine.size, city.distance, type = "n", ylab = "City distance (km/L)",
     xlab = "Engine size (L)", xlim = c(1, 5.5))
points(engine.size[d], city.distance[d], col = 4, pch = 1)
points(engine.size[!d], city.distance[!d], col = 2, pch = 2)
legend('topright', pch = c(1, 2), col = c(4, 2),
       legend = c("Gasoline  ","Diesel"))
```



We first suggest a simple linear regression line: $y = \beta_0 + \beta_1 x + \epsilon$, where y represents city distance, x fuel type, and ϵ is a nonobservable random 'error', with expected value 0 and variance σ^2. For simplicity we consider no correlation between error terms and y. We need to find a setimator for the unknown variales β_0 and β_1. To do so wee need to user the method of least squares. Wich means the finding the min of the function:

$$B(\beta) = \Sigma_{i=1}^{n}\{y_i - f(x_i;\beta)\}^2 = ||y - f(x;\beta)||^2$$

The last expressison is showing the matrix notation for representing
$y = (y1, ..., y_n)^T; (f(x_1;\beta), ..., f(x_m;\beta))^T;.$

But from what we can see in the plot of 'city distance against engine size', a linear model might not be the best.