BDA-LD

Benjamin Bernaud

2023-11-27

```
setwd("C:/Users/Berna/Documents/BDA-LD/userscripts")
```

BDA-LD - Partie Analyse de données

Chargement des librairies

```
install.packages("fpc", repos = "http://cran.us.r-project.org")
```

```
## Installation du package dans 'C:/Users/Berna/AppData/Local/R/win-library/4.3'
## (car 'lib' n'est pas spécifié)
```

```
## le package 'fpc' a été décompressé et les sommes MD5 ont été vérifiées avec succés
##
## Les packages binaires téléchargés sont dans
##  C:\Users\Berna\AppData\Local\Temp\RtmpCGO16i\downloaded_packages
```

```
install.packages("regclass", repos = "http://cran.us.r-project.org")
```

```
## Installation du package dans 'C:/Users/Berna/AppData/Local/R/win-library/4.3'
## (car 'lib' n'est pas spécifié)
```

```
## le package 'regclass' a été décompressé et les sommes MD5 ont été vérifiées avec succés
##
## Les packages binaires téléchargés sont dans
##  C:\Users\Berna\AppData\Local\Temp\RtmpCGO16i\downloaded_packages
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(regclass)
```

```
## Warning: le package 'regclass' a été compilé avec la version R 4.3.2
```

```
## Le chargement a nécessité le package : bestglm
```

```
## Warning: le package 'bestglm' a été compilé avec la version R 4.3.2
## Le chargement a nécessité le package : leaps
## Warning: le package 'leaps' a été compilé avec la version R 4.3.2
## Le chargement a nécessité le package : VGAM
## Warning: le package 'VGAM' a été compilé avec la version R 4.3.2
## Le chargement a nécessité le package : stats4
## Le chargement a nécessité le package : splines
## Le chargement a nécessité le package : rpart
## Le chargement a nécessité le package : randomForest
## Warning: le package 'randomForest' a été compilé avec la version R 4.3.2
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attachement du package : 'randomForest'
##
## L'objet suivant est masqué depuis 'package:dplyr':
##
##     combine
##
## L'objet suivant est masqué depuis 'package:ggplot2':
##
##     margin
##
## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

```
library(cluster)
```

```
## Warning: le package 'cluster' a été compilé avec la version R 4.3.2
```

```
library(ggplot2)
library(fpc)
```

```
## Warning: le package 'fpc' a été compilé avec la version R 4.3.2
```

1) Analyse exploratoire des données :

L'analyse exploratoire des données vous permettra d'identifier d'éventuels problèmes dans les données (valeurs incohérentes, codage des valeurs manquantes, etc.) et découvrir d'éventuelles propriétés de l'espace des données (valeurs doublons, variables liées, variables d'importance particulière ou bien inutiles, etc.). Appliquez pour cela les différentes méthodes d'analyse exploratoire des données vues en cours (statistiques descriptives, histogrammes, nuages de points, boîtes à moustaches, etc.).

Chargement des données

```
catalogue <- read.csv("../files/Catalogue.csv", header = TRUE, sep = ",", dec = ".", stringsAsFactors =
head(catalogue)
```

```
##   marque    nom puissance       longueur nbPlaces nbPortes couleur occasion
## 1  Volvo S80 T6       272 tr\xe8s longue        5        5   blanc    false
## 2  Volvo S80 T6       272 tr\xe8s longue        5        5    noir    false
```

```
## 3  Volvo S80 T6       272 tr\xe8s longue        5        5   rouge    false
## 4  Volvo S80 T6       272 tr\xe8s longue        5        5    gris     true
## 5  Volvo S80 T6       272 tr\xe8s longue        5        5    bleu     true
## 6  Volvo S80 T6       272 tr\xe8s longue        5        5    gris    false
##    prix
## 1 50500
## 2 50500
## 3 50500
## 4 35350
## 5 35350
## 6 50500
```

```
str(catalogue)
```

Catalogue

```
## 'data.frame':    270 obs. of  9 variables:
##  $ marque   : Factor w/ 21 levels "Hyunda\xef","Audi",..: 21 21 21 21 21 21 21 21 21 21 ...
##  $ nom      : Factor w/ 32 levels "1007 1.4","120i",..: 26 26 26 26 26 26 26 26 26 26 ...
##  $ puissance: int  272 272 272 272 272 272 272 272 272 272 ...
##  $ longueur : Factor w/ 4 levels "courte","longue",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ nbPlaces : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ nbPortes : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ couleur  : Factor w/ 5 levels "blanc","bleu",..: 1 4 5 3 2 3 2 5 1 4 ...
##  $ occasion : Factor w/ 2 levels "false","true": 1 1 1 2 2 1 1 2 2 2 ...
##  $ prix     : int  50500 50500 50500 35350 35350 50500 50500 35350 35350 35350 ...
```

```
summary(catalogue)
```

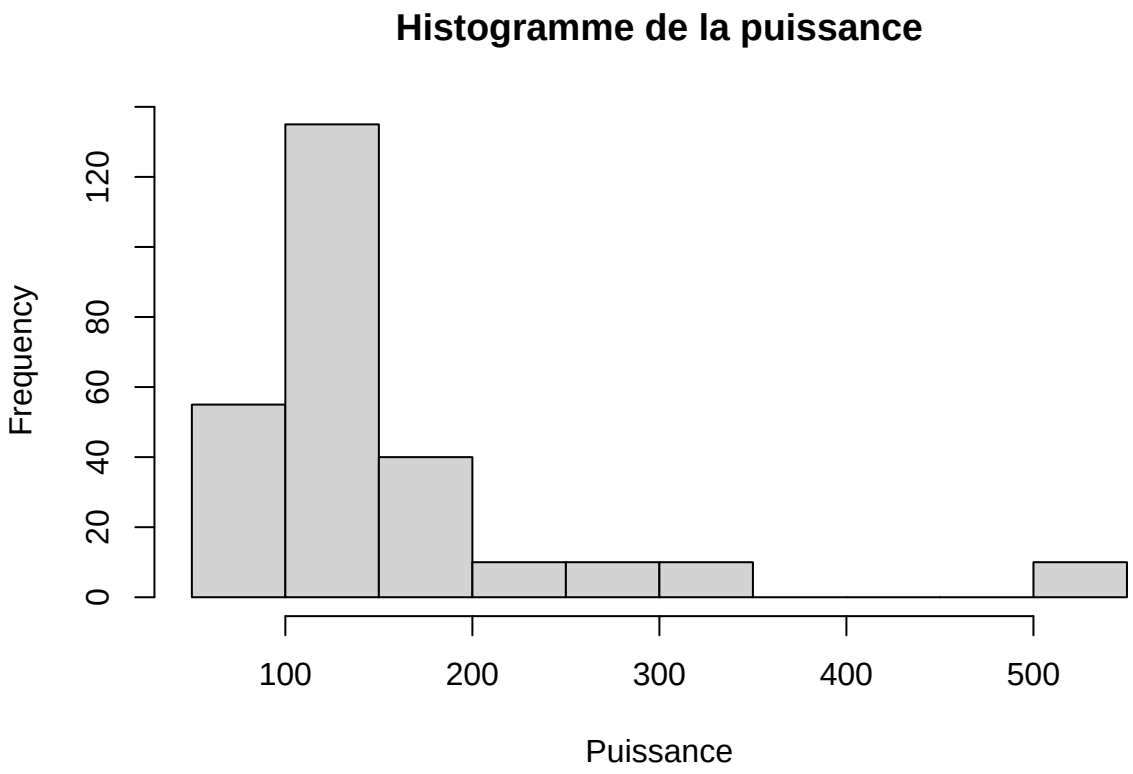
```
##         marque            nom        puissance               longueur
##  Renault   : 40   1007 1.4  : 10   Min.   : 55.0   courte        :60
##  Volkswagen: 40   120i      : 10   1st Qu.:109.0   longue        :90
##  Audi      : 20   9.3 1.8T  : 10   Median :147.0   moyenne       :70
##  BMW       : 20   A2 1.4    : 10   Mean   :157.6   tr\xe8s longue:50
##  Mercedes  : 20   A200      : 10   3rd Qu.:170.0
##  Nissan    : 15   A3 2.0 FSI: 10   Max.   :507.0
##  (Other)   :115   (Other)   :210
##     nbPlaces        nbPortes      couleur    occasion        prix
##  Min.   :5.000   Min.   :3.000   blanc:54   false:160   Min.   :  7500
##  1st Qu.:5.000   1st Qu.:5.000   bleu :54   true :110   1st Qu.: 16029
##  Median :5.000   Median :5.000   gris :54               Median : 20598
##  Mean   :5.222   Mean   :4.815   noir :54               Mean   : 26668
##  3rd Qu.:5.000   3rd Qu.:5.000   rouge:54               3rd Qu.: 30000
##  Max.   :7.000   Max.   :5.000                          Max.   :101300
##
```

Statistiques descriptives pour une colonne spécifique, par exemple, 'puissance'

```
summary(catalogue$puissance)
```

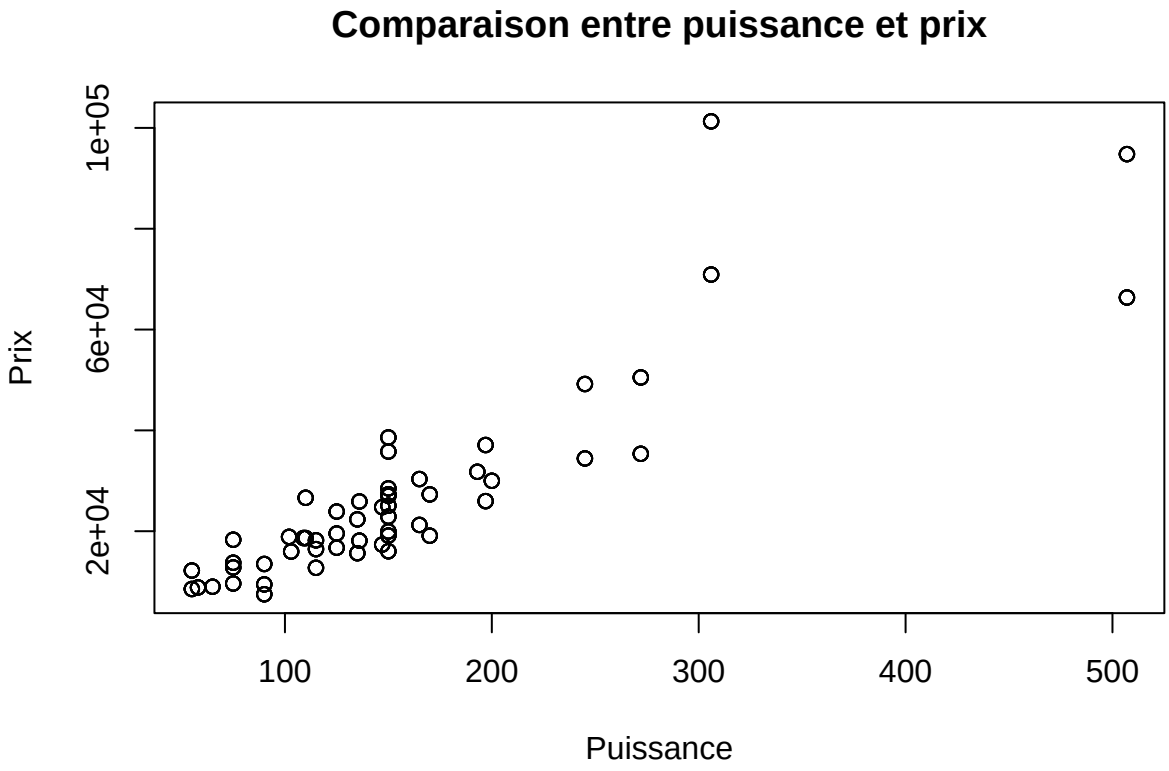
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   109.0   147.0   157.6   170.0   507.0
```

```
hist(catalogue$puissance, main="Histogramme de la puissance", xlab="Puissance")
```



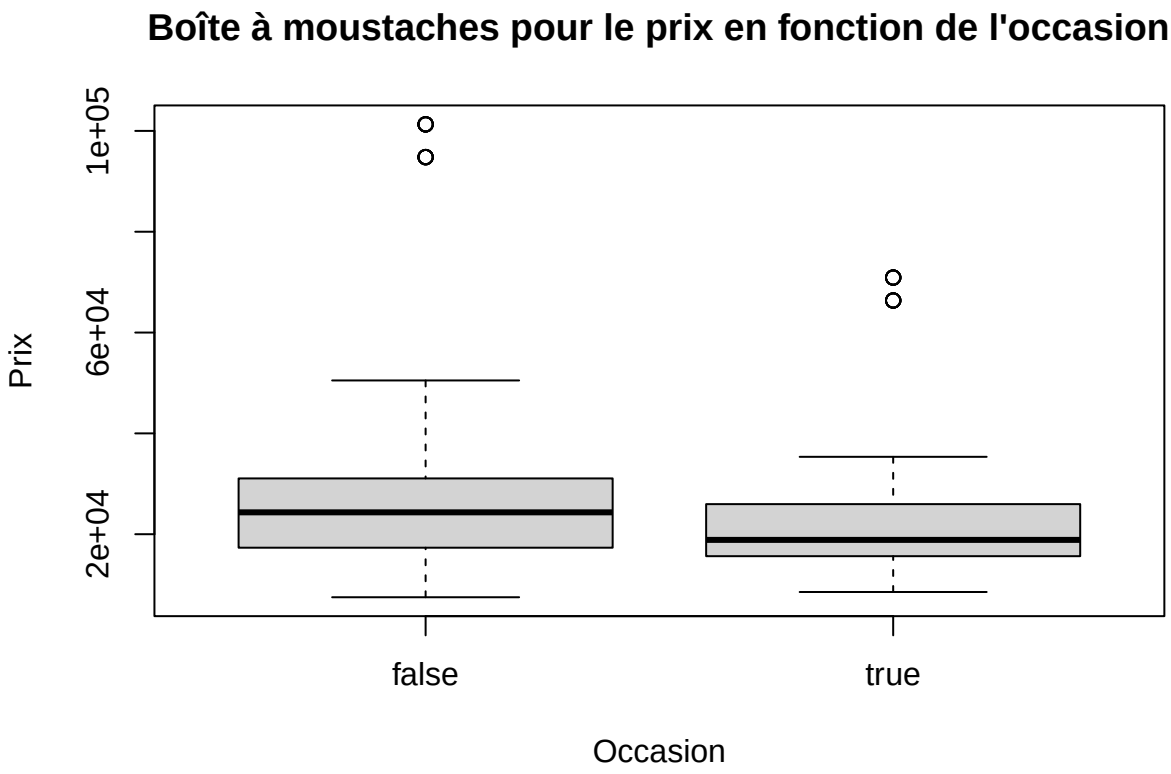
Comparaison entre puissance et prix

```
plot(catalogue$puissance, catalogue$prix, main="Comparaison entre puissance et prix", xlab="Puissance",
```



Prix en fonction du statut occasion

```
boxplot(catalogue$prix ~ catalogue$occasion, main="Boîte à moustaches pour le prix en fonction de l'occa
```



On vérifie si il y a des données manquantes

```
sum(is.na(catalogue))
```

```
## [1] 0
```

On vérifie si il y a des lignes en double

```
doublons <- duplicated(catalogue)

catalogue[doublons, ]
```

```
## [1] marque    nom       puissance longueur  nbPlaces  nbPortes  couleur
## [8] occasion  prix
## <0 lignes> (ou 'row.names' de longueur nulle)
```

Recherche de données liées :

```
matrice_cor <- cor(catalogue[, sapply(catalogue, is.numeric)])

print(matrice_cor)
```

```
##             puissance    nbPlaces  nbPortes        prix
## puissance  1.00000000 -0.05708192 0.3109884  0.87545111
## nbPlaces  -0.05708192  1.00000000 0.1129385 -0.08189026
## nbPortes   0.31098839  0.11293849 1.0000000  0.27147998
## prix       0.87545111 -0.08189026 0.2714800  1.00000000
```

La puissance et le prix ont l'air assez liés !

2) Identification des catégories de véhicules :

Vous devez à partir des informations du Catalogue identifier des catégories de véhicules (citadine, routière, sportive, etc.) en fonction de leur taille, puissance, prix, etc. Ces catégories doivent correspondre à divers besoins de la part des clients (une grande voiture pour les familles nombreuses, une petite voiture pour circuler en ville, etc.). Ces catégories de véhicules constitueront les classes à prédire durant les étapes suivantes du processus.

Premièrement on définit des critères d'identification : - Puissance : Utiliser la puissance du véhicule pour distinguer entre des catégories telles que citadine, routière, sportive, etc. - Taille : Utiliser la longueur du véhicule pour également contribuer à la classification. Les citadines ont tendances a être plus courtes que les routières par exemple. - Prix : Utiliser la fourchette de prix pour distinguer entre des catégories de véhicules plus abordables ou haut de gamme.

Pour définir les seuils de classification, on part du principe que l'ensemble proposé est homogène et représentatif. Ainsi on peut se baser sur ces données pour définir différentes classes.

Différentes classes de puissances

```
summary(catalogue$puissance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   109.0   147.0   157.6   170.0   507.0
```

Différentes classes de tailles

```
summary(catalogue$longueur)
```

```
##         courte         longue        moyenne tr\xe8s longue
##             60             90             70             50
```

Différentes classes de prix

```
summary(catalogue$prix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7500   16029   20598   26668   30000  101300
```

```
# On définit les seuils en fonction des stats affichées juste au dessus
seuil_puissance <- c(0, 100, 170, Inf)
seuil_prix <- c(0, 16000, 30000, Inf)

# Assigner les catégories en fonction des seuils
catalogue$puissance_class <- cut(catalogue$puissance, breaks = seuil_puissance, labels = c("Puissance fa
catalogue$prix_class <- cut(catalogue$prix, breaks = seuil_prix, labels = c("Prix faible", "Prix moyen"

head(catalogue, 20)
```

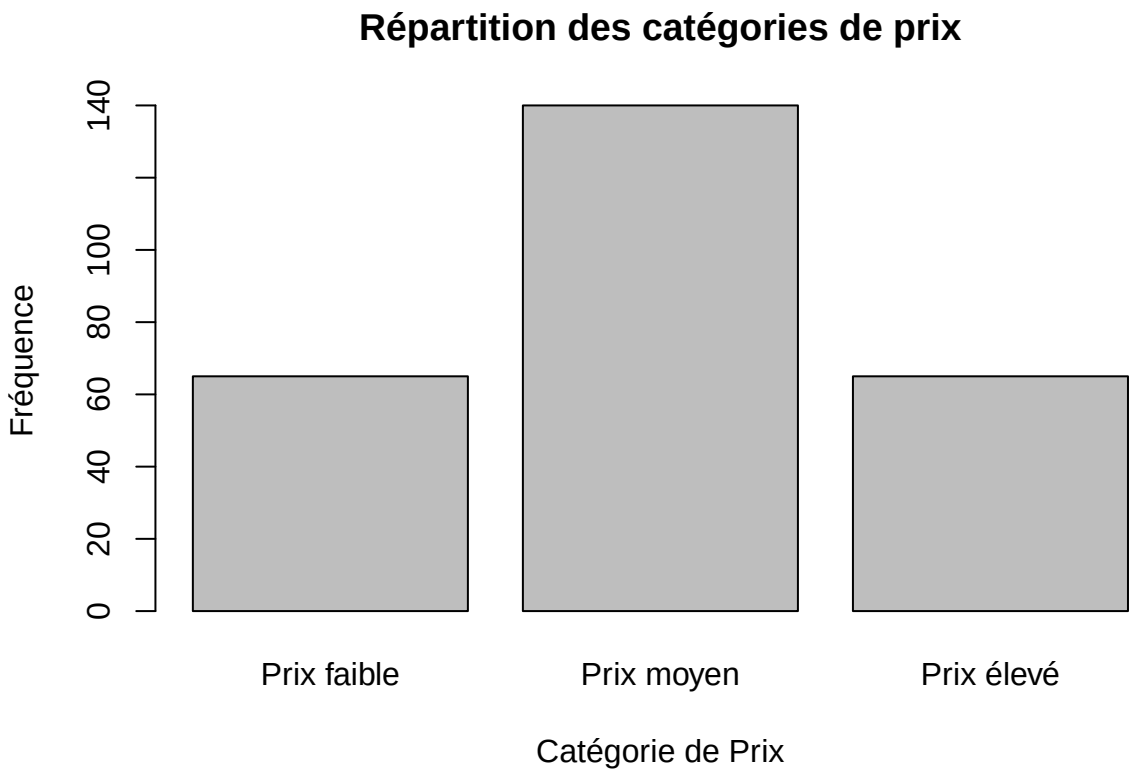
```
##        marque            nom puissance       longueur nbPlaces nbPortes couleur
## 1       Volvo         S80 T6       272 tr\xe8s longue        5        5   blanc
## 2       Volvo         S80 T6       272 tr\xe8s longue        5        5    noir
## 3       Volvo         S80 T6       272 tr\xe8s longue        5        5   rouge
## 4       Volvo         S80 T6       272 tr\xe8s longue        5        5    gris
## 5       Volvo         S80 T6       272 tr\xe8s longue        5        5    bleu
## 6       Volvo         S80 T6       272 tr\xe8s longue        5        5    gris
## 7       Volvo         S80 T6       272 tr\xe8s longue        5        5    bleu
## 8       Volvo         S80 T6       272 tr\xe8s longue        5        5   rouge
## 9       Volvo         S80 T6       272 tr\xe8s longue        5        5   blanc
## 10      Volvo         S80 T6       272 tr\xe8s longue        5        5    noir
## 11 Volkswagen Touran 2.0 FSI       150         longue        7        5   rouge
```

```
## 12 Volkswagen Touran 2.0 FSI       150         longue        7        5    gris
## 13 Volkswagen Touran 2.0 FSI       150         longue        7        5    bleu
## 14 Volkswagen Touran 2.0 FSI       150         longue        7        5    gris
## 15 Volkswagen Touran 2.0 FSI       150         longue        7        5    bleu
## 16 Volkswagen Touran 2.0 FSI       150         longue        7        5   blanc
## 17 Volkswagen Touran 2.0 FSI       150         longue        7        5    noir
## 18 Volkswagen Touran 2.0 FSI       150         longue        7        5   rouge
## 19 Volkswagen Touran 2.0 FSI       150         longue        7        5   blanc
## 20 Volkswagen Touran 2.0 FSI       150         longue        7        5    noir
##    occasion  prix   puissance_class prix_class
## 1     false 50500  Puissance élevée Prix élevé
## 2     false 50500  Puissance élevée Prix élevé
## 3     false 50500  Puissance élevée Prix élevé
## 4      true 35350  Puissance élevée Prix élevé
## 5      true 35350  Puissance élevée Prix élevé
## 6     false 50500  Puissance élevée Prix élevé
## 7     false 50500  Puissance élevée Prix élevé
## 8      true 35350  Puissance élevée Prix élevé
## 9      true 35350  Puissance élevée Prix élevé
## 10     true 35350  Puissance élevée Prix élevé
## 11    false 27340 Puissance moyenne Prix moyen
## 12     true 19138 Puissance moyenne Prix moyen
## 13     true 19138 Puissance moyenne Prix moyen
## 14    false 27340 Puissance moyenne Prix moyen
## 15    false 27340 Puissance moyenne Prix moyen
## 16     true 19138 Puissance moyenne Prix moyen
## 17     true 19138 Puissance moyenne Prix moyen
## 18     true 19138 Puissance moyenne Prix moyen
## 19    false 27340 Puissance moyenne Prix moyen
## 20    false 27340 Puissance moyenne Prix moyen
```

```r
# Tableau croisé entre la catégorie de puissance et la catégorie de taille
table_croisee <- table(catalogue$puissance_class, catalogue$longueur)
print(table_croisee)
```

```
##
##                     courte longue moyenne tr\xe8s longue
##   Puissance faible      50      0       5              0
##   Puissance moyenne     10     80      65              0
##   Puissance élevée       0     10       0             50
```

```r
# Diagramme en barres pour visualiser la répartition des catégories de prix
barplot(table(catalogue$prix_class), main="Répartition des catégories de prix", xlab="Catégorie de Prix
```



Finalement, grâce a ces classifications, on pourrait affecter une classe à chacun des véhicules :

```
library(dplyr)
```

```
catalogue$classe <- NA

catalogue <- catalogue %>%
  mutate(classe = case_when(
    puissance_class == "Puissance élevée" & longueur != "Courte" ~ "Routière",
    puissance_class == "Puissance élevée" & longueur == "Courte" & prix_class == "Prix élevé" ~ "Sporti
    puissance_class == "Puissance moyenne" & longueur == "Courte" & prix_class == "Prix élevé" ~ "Sport
    puissance_class == "Puissance moyenne" & longueur == "Courte" & prix_class != "Prix élevé" ~ "Citad
    puissance_class == "Puissance moyenne" & longueur != "Courte" ~ "Routière",
    puissance_class == "Puissance faible" ~ "Citadine",
    TRUE ~ "?"
  ))

head(catalogue, 25)
```

```
##        marque           nom puissance       longueur nbPlaces nbPortes couleur
## 1       Volvo         S80 T6       272 tr\xe8s longue        5        5   blanc
## 2       Volvo         S80 T6       272 tr\xe8s longue        5        5    noir
## 3       Volvo         S80 T6       272 tr\xe8s longue        5        5   rouge
## 4       Volvo         S80 T6       272 tr\xe8s longue        5        5    gris
## 5       Volvo         S80 T6       272 tr\xe8s longue        5        5    bleu
## 6       Volvo         S80 T6       272 tr\xe8s longue        5        5    gris
## 7       Volvo         S80 T6       272 tr\xe8s longue        5        5    bleu
```

```
## 8       Volvo         S80 T6       272 tr\xe8s longue        5        5   rouge
## 9       Volvo         S80 T6       272 tr\xe8s longue        5        5   blanc
## 10      Volvo         S80 T6       272 tr\xe8s longue        5        5    noir
## 11 Volkswagen Touran 2.0 FSI       150         longue        7        5   rouge
## 12 Volkswagen Touran 2.0 FSI       150         longue        7        5    gris
## 13 Volkswagen Touran 2.0 FSI       150         longue        7        5    bleu
## 14 Volkswagen Touran 2.0 FSI       150         longue        7        5    gris
## 15 Volkswagen Touran 2.0 FSI       150         longue        7        5    bleu
## 16 Volkswagen Touran 2.0 FSI       150         longue        7        5   blanc
## 17 Volkswagen Touran 2.0 FSI       150         longue        7        5    noir
## 18 Volkswagen Touran 2.0 FSI       150         longue        7        5   rouge
## 19 Volkswagen Touran 2.0 FSI       150         longue        7        5   blanc
## 20 Volkswagen Touran 2.0 FSI       150         longue        7        5    noir
## 21 Volkswagen    Polo 1.2 6V        55         courte        5        3   blanc
## 22 Volkswagen    Polo 1.2 6V        55         courte        5        3   blanc
## 23 Volkswagen    Polo 1.2 6V        55         courte        5        3    noir
## 24 Volkswagen    Polo 1.2 6V        55         courte        5        3    noir
## 25 Volkswagen    Polo 1.2 6V        55         courte        5        3    bleu
##    occasion  prix   puissance_class  prix_class   classe
## 1     false 50500  Puissance élevée  Prix élevé Routière
## 2     false 50500  Puissance élevée  Prix élevé Routière
## 3     false 50500  Puissance élevée  Prix élevé Routière
## 4      true 35350  Puissance élevée  Prix élevé Routière
## 5      true 35350  Puissance élevée  Prix élevé Routière
## 6     false 50500  Puissance élevée  Prix élevé Routière
## 7     false 50500  Puissance élevée  Prix élevé Routière
## 8      true 35350  Puissance élevée  Prix élevé Routière
## 9      true 35350  Puissance élevée  Prix élevé Routière
## 10     true 35350  Puissance élevée  Prix élevé Routière
## 11    false 27340 Puissance moyenne  Prix moyen Routière
## 12     true 19138 Puissance moyenne  Prix moyen Routière
## 13     true 19138 Puissance moyenne  Prix moyen Routière
## 14    false 27340 Puissance moyenne  Prix moyen Routière
## 15    false 27340 Puissance moyenne  Prix moyen Routière
## 16     true 19138 Puissance moyenne  Prix moyen Routière
## 17     true 19138 Puissance moyenne  Prix moyen Routière
## 18     true 19138 Puissance moyenne  Prix moyen Routière
## 19    false 27340 Puissance moyenne  Prix moyen Routière
## 20    false 27340 Puissance moyenne  Prix moyen Routière
## 21     true  8540  Puissance faible Prix faible Citadine
## 22    false 12200  Puissance faible Prix faible Citadine
## 23    false 12200  Puissance faible Prix faible Citadine
## 24     true  8540  Puissance faible Prix faible Citadine
## 25     true  8540  Puissance faible Prix faible Citadine
```