Atelier avec *R*

Note

Ce TP est à rendre en fin de séance.

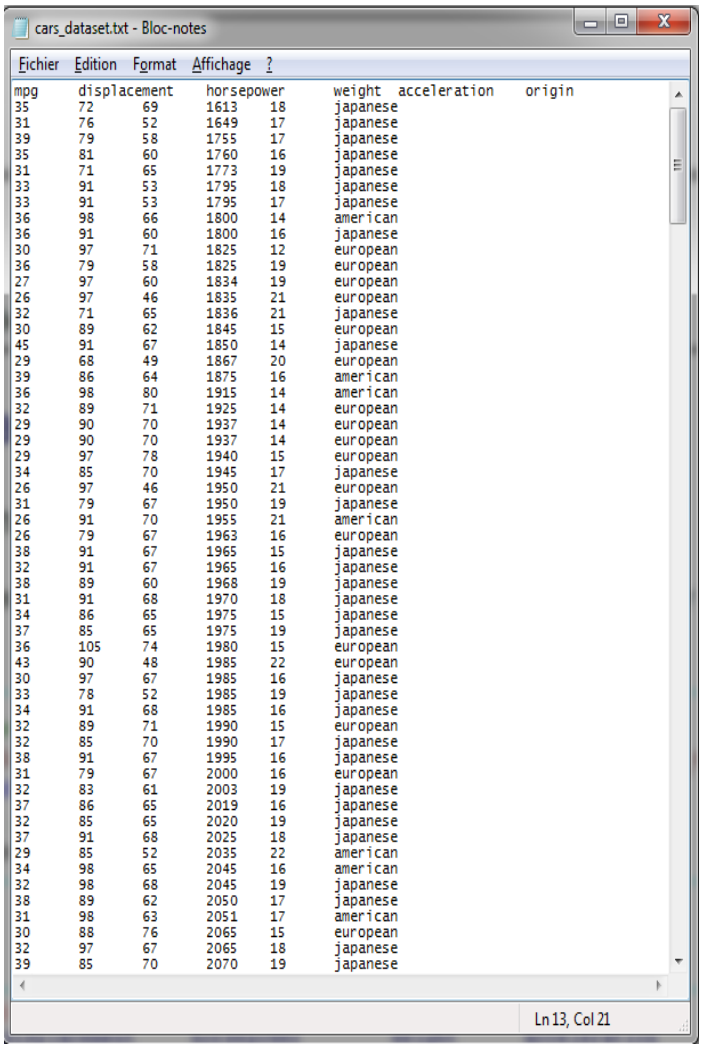
Objectifs généraux

Dans ce TP, nous allons appliquer la Méthodes des Centres Mobiles « K-Means » sur des échantillons de données et par la suite l'analyse en composantes principales (ACP).

Segmentation et Classification d'un ensemble de véhicules

On utilise le fichier « cars_dataset.txt », un fichier texte avec séparateur tabulation. Il décrit les caractéristiques de 392 véhicules. Les variables actives qui participeront au calcul sont :
La consommation (MPG, miles per galon, plus le chiffre est élevé, moins la voiture consomme) ;
La taille du moteur (DISPLACEMENT)
la puissance (HORSEPOWER)
le poids (WEIGHT)
l'accélération (ACCELERATION, le temps mis pour atteindre une certaine vitesse, plus le chiffre est faible plus la voiture est performante).
La variable illustrative « origine des véhicules » (ORIGIN : Japon, Europe, Etats Unis) servira à renforcer l'interprétation des groupes.

Sources :
http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cars_dataset.zip
http://lib.stat.cmu.edu/datasets/cars.desc



mpg	displacement	horsepower	weight	acceleration	origin
35	72	69	1613	18	japanese
31	76	52	1649	17	japanese
39	79	58	1755	17	japanese
35	81	60	1760	16	japanese
31	71	65	1773	19	japanese
33	91	53	1795	18	japanese
33	91	53	1795	17	japanese
36	98	66	1800	14	american
36	91	60	1800	16	japanese
30	97	71	1825	12	european
36	79	58	1825	19	european
27	97	60	1834	19	european
26	97	46	1835	21	european
32	71	65	1836	21	japanese
30	89	62	1845	15	european
45	91	67	1850	14	japanese
29	68	49	1867	20	european
39	86	64	1875	16	american
36	98	80	1915	14	american
32	89	71	1925	14	european
29	90	70	1937	14	european
29	90	70	1937	14	european
29	97	78	1940	15	european
34	85	70	1945	17	japanese
26	97	46	1950	21	european
31	79	67	1950	19	japanese
26	91	70	1955	21	american
26	79	67	1963	16	european
38	91	67	1965	15	japanese
32	91	67	1965	16	japanese
38	89	60	1968	19	japanese
31	91	68	1970	18	japanese
34	86	65	1975	15	japanese
37	85	65	1975	19	japanese
36	105	74	1980	15	european
43	90	48	1985	22	european
30	97	67	1985	16	japanese
33	78	52	1985	19	japanese
34	91	68	1985	16	japanese
32	89	71	1990	15	european
32	85	70	1990	17	japanese
38	91	67	1995	16	japanese
31	79	67	2000	16	european
32	83	61	2003	19	japanese
37	86	65	2019	16	japanese
32	85	65	2020	19	japanese
37	91	68	2025	18	japanese
29	85	52	2035	22	american
34	98	65	2045	16	american
32	98	68	2045	19	japanese
38	89	62	2050	17	japanese
31	98	63	2051	17	american
30	88	76	2065	15	european
32	97	67	2065	18	japanese
39	85	70	2070	19	japanese

Objectifs spécifiques

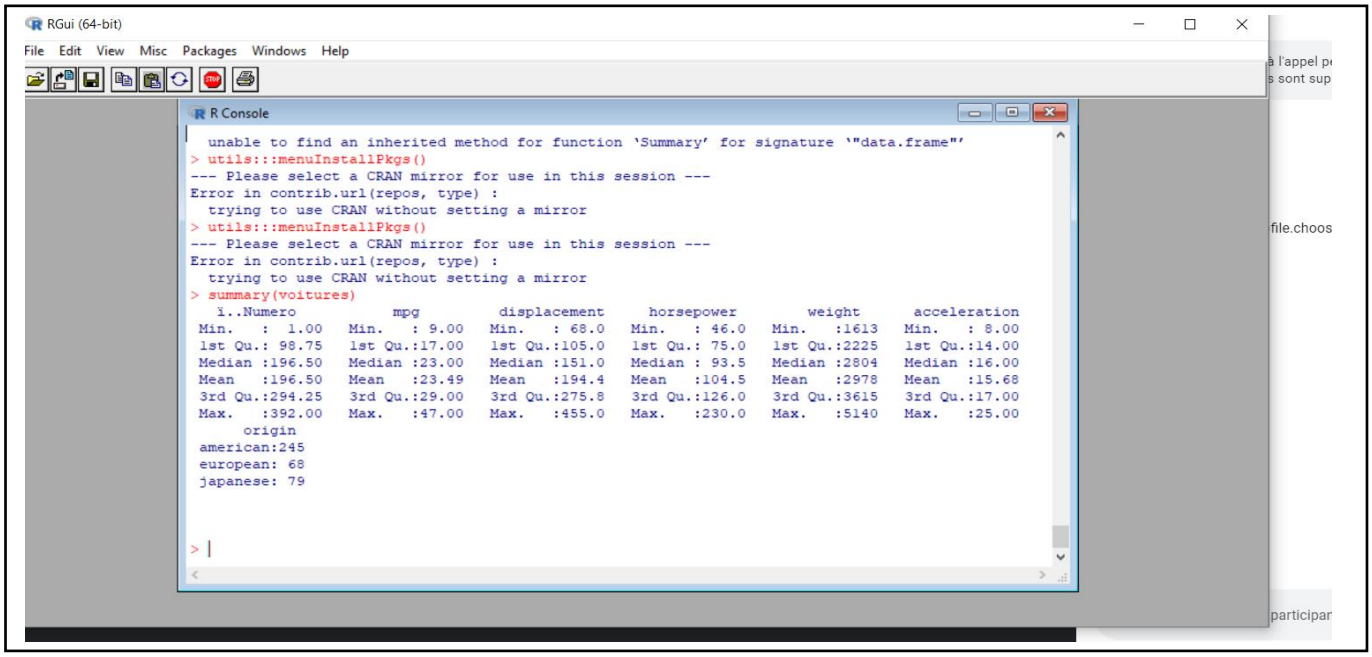
- ✔ Importer les données
- ✔ Réaliser quelques statistiques descriptives sur les variables actives ;
- ✔ Centrer et réduire les variables
- ✔ Réaliser la classification automatique via les K-Means sur les variables transformées, en décidant nous même du nombre de classes
- ✔ Visualiser les données avec la nouvelle colonne représentant la classe d'appartenance des individus
- ✔ Illustrer les classes à l'aide des variables actives, via des statistiques descriptives comparatives et des graphiques judicieusement choisis
- ✔ Croiser la partition obtenue avec une variable catégorielle illustrative
- ✔ Exporter les données, avec la colonne additionnelle, dans un nouveau fichier

1. Importation des données et statistiques descriptives : (sans package additionnel spécifique)

```
# Importation des données
Voitures<-read.table(file="cars_origin" , header=T,sep="/T",dec=".")

# Description et statistiques descriptives
summary(voitures)
```

Donner les statistiques descriptives :



2. Centrage et réduction :

Pour centrer et réduire les données, on doit commencer tout d'abord par la construction d'une fonction *«centrage_reduction»* qui centre et réduit une colonne, qu'on applique à l'ensemble des variables actives avec **apply(……..)**

Pour ce faire, on propose d'exécuter la fonction de standardisation de colonne suivante :

```
# Exécuter la fonction de standardisation d'une colonne
centrage_reduction<- function(x)
                    {return ((x-mean(x))/sqrt(var(x)))
                    }
```

Obtention du tableau des données centrées et réduites

```
# Appliquer pour produire le tableau des données centrées et réduites
```

voitures.cr <-apply(voitures[,2 :6],2,centrage_reduction)

Vérification des moyennes
Moyenne<-apply(voitures.cr,2,mean)

Vérification de la variance
variance<-apply(voitures.cr,2,var)

```
+ > variance<-apply(voitures.cr,2,var)
> moyenne
          mpg  displacement    horsepower        weight  acceleration
-2.400263e-16  7.642464e-17 -1.767797e-16 -7.163479e-18  1.160824e-16
> variance
          mpg displacement   horsepower       weight acceleration
            1            1            1            1            1
>
```

Interprétation de la moyenne et de la variance des colonnes :

Les moyennes ont réduits

3. Application de la méthode des Centres Mobiles (K-Means)

Maintenant, on lance l'algorithme K-Means sur les variables centrées et réduites.

On propose de concevoir une partition de deux groupes (deux clusters), en se limitant à 40 itérations.

i. Expliquer le principe d'utilisation de la fonction R « **kmeans**(…) »

?Kmeans

Clustering sur un ensemble de données

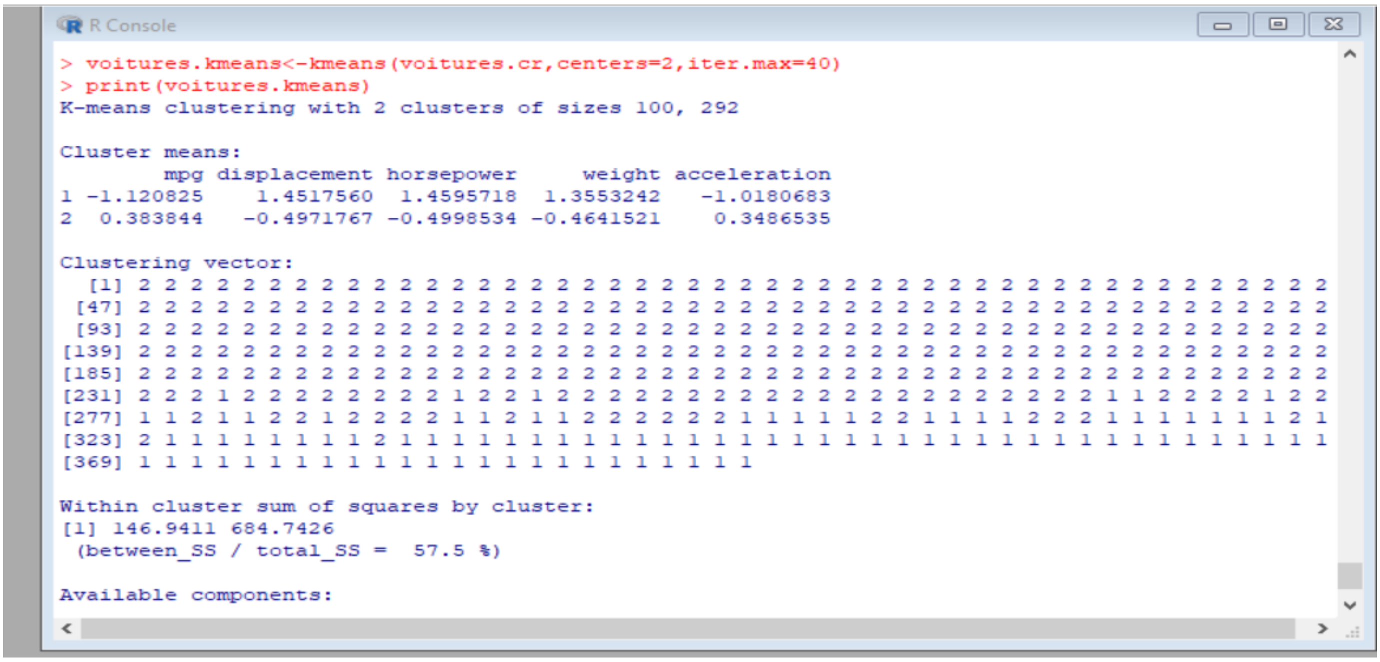
ii. En déduire le code suivant :

```
# K-means en deux groupes
nb.classes <- 2
voitures.kmeans <- kmeans(voitures.cr,centres=2,iter.max=40)
```

R affiche, entre autres : le nombre d'observations dans chaque groupe, 100 et 292 ; les moyennes conditionnellement aux groupes pour chaque variable active, celles qui ont été utilisées pour le calcul c.-à-d. les variables centrées et réduites, ce n'est pas très exploitable pour nous ; la classe associée à chaque observation, nous exploiterons cette colonne par la suite.

En effet, dans la mesure où l'algorithme repose sur une heuristique, le choix des centres de départ notamment pouvant influer sur le résultat final, les classes peuvent être légèrement différentes à la sortie. Il peut en être de même si nous choisissons un autre algorithme d'optimisation.

iii. Donner l'imprime écran : (K-means clustering, cluster means, clustering vector)



```
R Console
> voitures.kmeans<-kmeans(voitures.cr,centers=2,iter.max=40)
> print(voitures.kmeans)
K-means clustering with 2 clusters of sizes 100, 292

Cluster means:
        mpg displacement horsepower     weight acceleration
1 -1.120825    1.4517560  1.4595718  1.3553242   -1.0180683
2  0.383844   -0.4971767 -0.4998534 -0.4641521    0.3486535

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [47] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [93] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[139] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[185] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[231] 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 1 2 2
[277] 1 1 2 1 1 2 2 1 2 2 2 2 1 1 2 1 1 2 2 2 2 2 2 1 1 1 1 1 2 2 1 1 1 1 2 2 2 1 1 1 1 1 1 1 2 1
[323] 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[369] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 146.9411 684.7426
 (between_SS / total_SS =  57.5 %)

Available components:
```

4. Interprétation des groupes d'appartenance :

✔ Pour l'interprétation des groupes, on calcule les moyennes conditionnelles des variables actives originelles. On les collecte dans une seule matrice à l'aide des commandes suivantes :

```
#récupération des groupes d'apparetenance

groupe <- as.factor(voitures.kmeans$cluster)
```

✔ Pour croiser les clusters avec la variable catégorielle illustrative **ORIGIN**, on introduit la commande « table » :

```
#croisement des clusters avec la variable illustrative catégorielle
table(voitures$origin,voitures.kmeans$cluster)
```

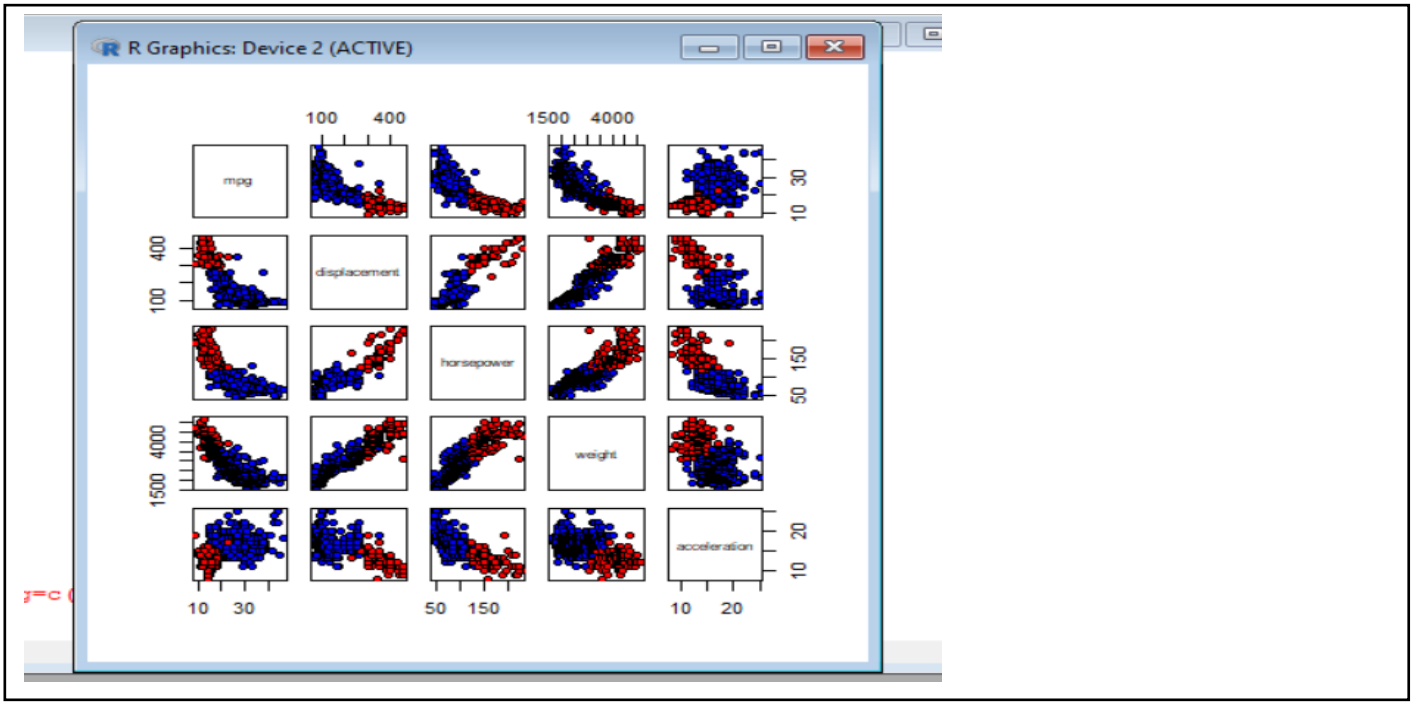
✔ Compléter le tableau de contingence suivant :

Un tableau de contingence est une méthode de représentation de données découlant d'un comptage. Les données sont rassemblées dans un tableau avec le caractère auquel elles sont reliées.

On pratique des études sur plusieurs caractères, en essayant alors de déterminer s'il existe une quelconque liaison entre eux. Pour cela on étudie les individus recensant plusieurs caractères à la fois.

Pour projeter les points, illustrés selon leur groupe d'appartenance, dans les plans formés par les couples de variables, R démontre toute sa puissance. La commande utilisée est **« pairs »** le résultat est riche d'enseignements : les variables sont pour la plupart fortement corrélées, presque tous les couples de variables permettent de distinguer les groupes :

```
#graphique des variables 2 à 2 avec groupe d'appartenance
>pairs(voitures[,2:6],pch=21,bg=c("red","blue")[groupe])
```



5. Combinaison ACP / K-Means :

Dans le but de trouver un outil permettant de bien situer les groupes, on propose de projeter les points dans le premier plan factoriel de l'Analyse en Composantes Principales.

Pour ce faire, appliquer les lignes de commandes suivantes :

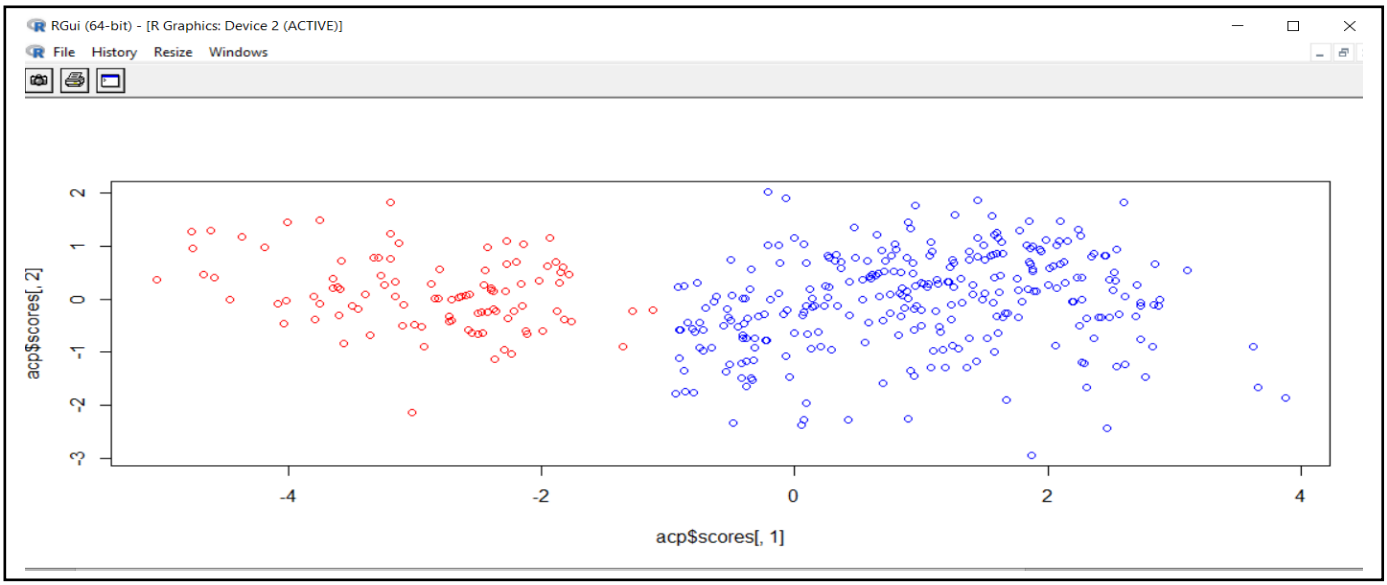
```
#ACP sur les données centrées réduites
Acp<- princomp(voitures.cr,cor=T,scores=T)
Print(acp)

#pour obtenir les valeurs propres
Print(acp$sdev ^2)
```

```
#pour obtenir les corrélations sur le premier axe
Print(acp$loading[,1]*acp$sdev[1])

#graphique dans le premier plan factoriel, avec mise en évidence des groupes
plot(acp$scores[,1],acp$scores[,2],type="p",pch=21,col=c("red","blue")[groupe])
```

✔ Le graphique obtenu donne:



6. Exportation des données :

Durant cette dernière étape du processus K-means, on exporte l'ensemble des données dans un seul fichier en fusionnant la base initiale avec la colonne additionnelle produite par la typologie.

```
#exportation des données avec le cluster d'appartenance
voitures.export<- cbind(voitures,groupe)
write.table(voitures.export,file="export_r.txt",sep="\t",dec=".",row.names=F)
```

✔ Vérifier la création du nouveau fichier « export_r.txt » sur votre répertoire courant, et le faire joindre avec votre compte-rendu.