projet carte de Kohonen

DIARRASSOUBA SAKARIA

13/03/2020

Contents

Présentation

Dans cette analyse, nous utiliserons les données sur les clients d'un centres commercial qui contiennent des données de base comme l'identité du client, l'âge, le sexe, le revenu annuel et le score des dépenses. L'objectif de cette analyse est d'identifier le segment de clientèle via la Carte de Kohonen, afin de comprendre quel est le segment de clientèle qui devient la cible de l'équipe marketing pour planifier les stratégies de marketing.

Importation des données

```
# importation des données
mall =read.csv("Mall_Customers.csv", header = TRUE)
mallCust=data.frame(mall[1],mall[2],mall[3],mall[4],mall[5])
names(mallCust)=c("ID","Sexe","Age","revenus_annuels","score_depense")
head(mallCust)
```

```
##   ID   Sexe Age revenus_annuels score_depense
## 1  1   Male  19              15            39
## 2  2   Male  21              15            81
## 3  3 Female  20              16             6
## 4  4 Female  23              16            77
## 5  5 Female  31              17            40
## 6  6 Female  22              17            76
```

nombre de variables manquantes

```
colSums(is.na(mallCust))
```

```
##              ID            Sexe             Age revenus_annuels
##               0               0               0               0
##   score_depense
##               0
```

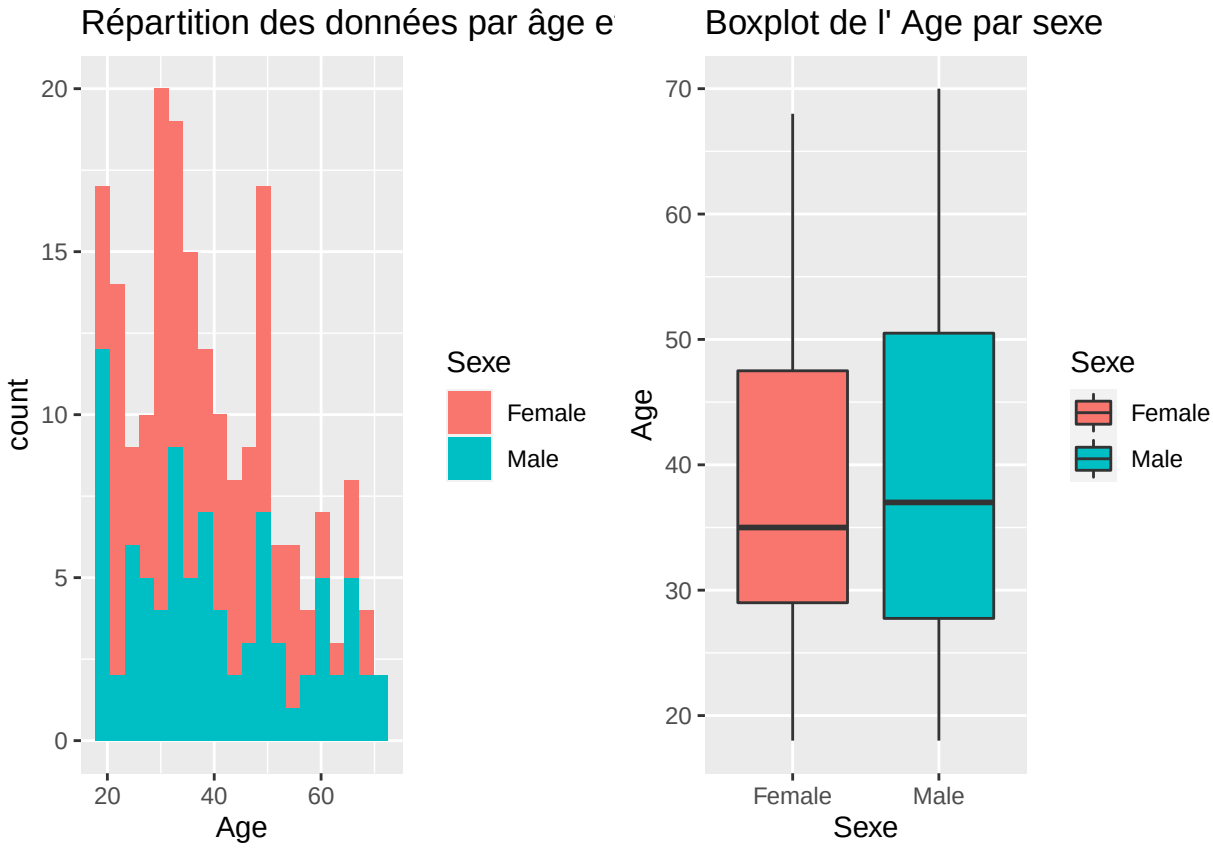
Age - Variable

Répartition des données par sexe.

```r
ageHist <- ggplot(mallCust, aes(Age, fill=Sexe)) + geom_histogram(bins = 20) +
          ggtitle("Répartition des données par âge et par sexe")

ageBox <- ggplot(mallCust, aes(x=Sexe, y=Age, fill = Sexe)) +
          geom_boxplot() +
          ggtitle("Boxplot de l' Age par sexe")

plot_grid(ageHist, ageBox)
```



Il ressort de répartition de que les femmes sont les plus fréquentes dans ce centre commercial ce qui est en adéquation avec la réalité

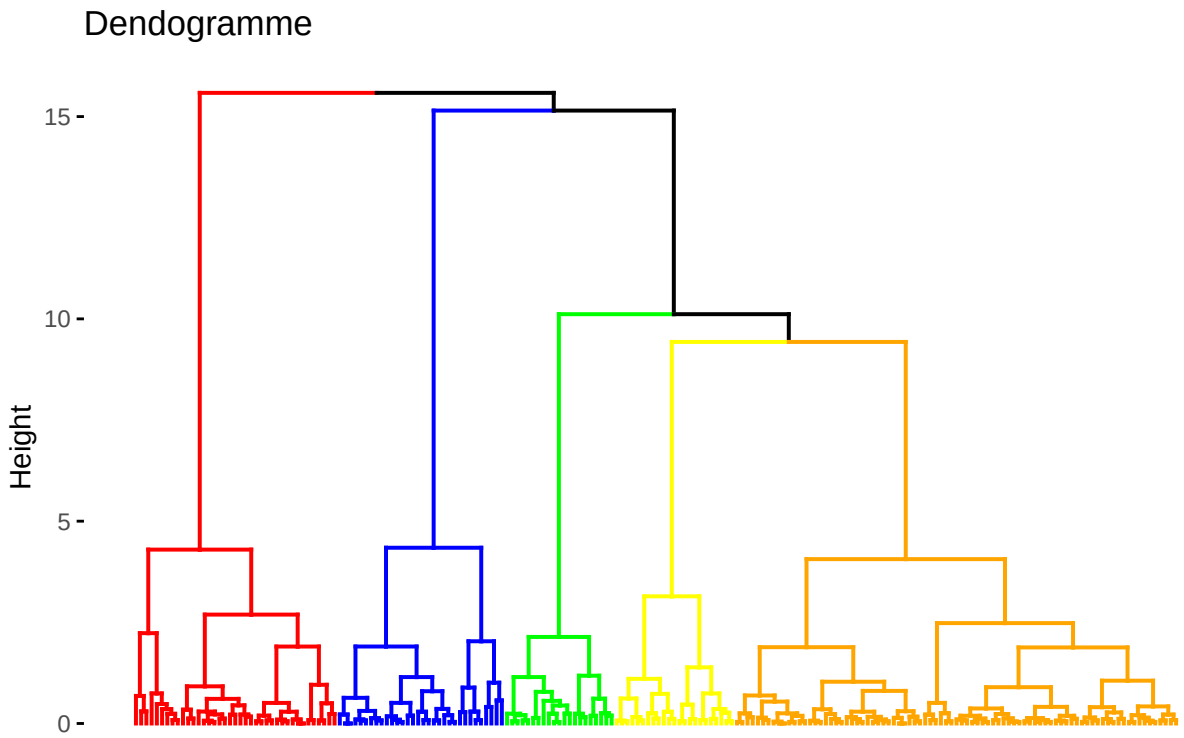
Méthode utilisé pour le clustering

Choix des revenus et scores dépenses annuels pour le regroupement des sujets et l'échelle des données ## Méthode du Dendrogramme

Utilisation du Dendrogramme pour connaitre le nombre de classes des clients

```r
# sacle :pour rendre les données à la meme échelle
custom=scale(mallCust[,c("revenus_annuels", "score_depense")], center = T, scale = T)
den<-hclust(dist(custom,method="euclidean"),method='ward.D2')
fviz_dend(den,show_labels = FALSE,main="Dendogramme",
          k = 5,
          k_colors = c("red","blue","green","yellow","orange")

          )
```

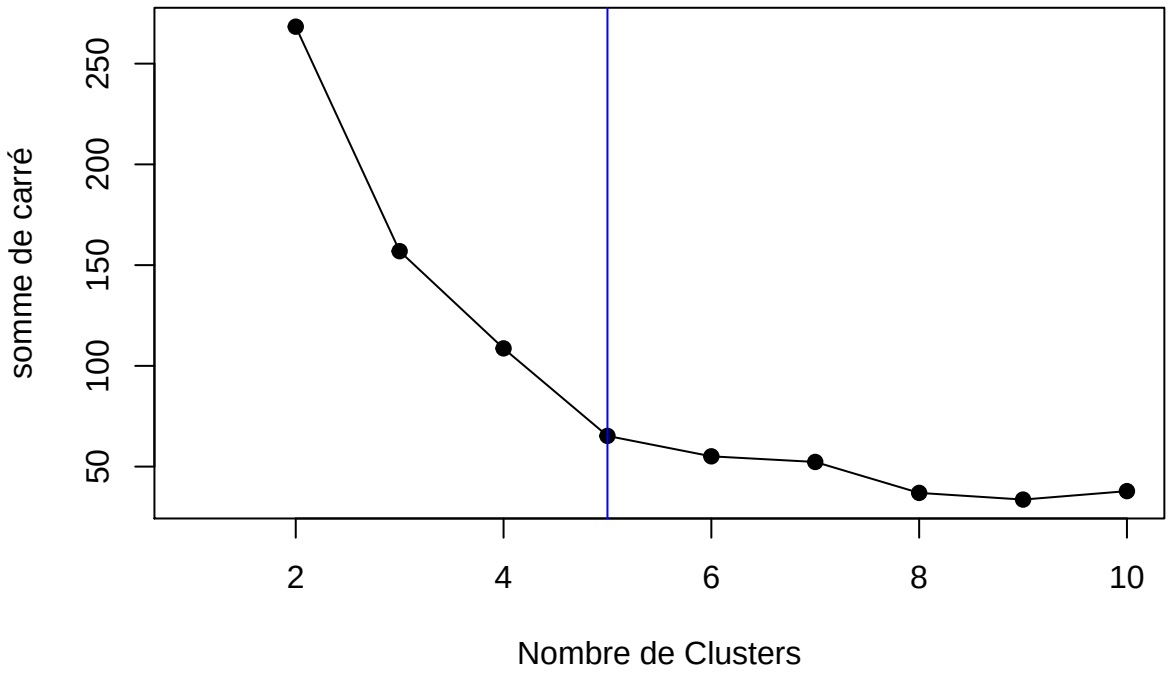


Au vu de l'arbre hiérarchique, on peut tailler l'arbre pour construire des classes.
** Pour une hauteur de coupe 10, on définit 3 classes Pour une hauteur de coupe 5, on définit 5 classes **

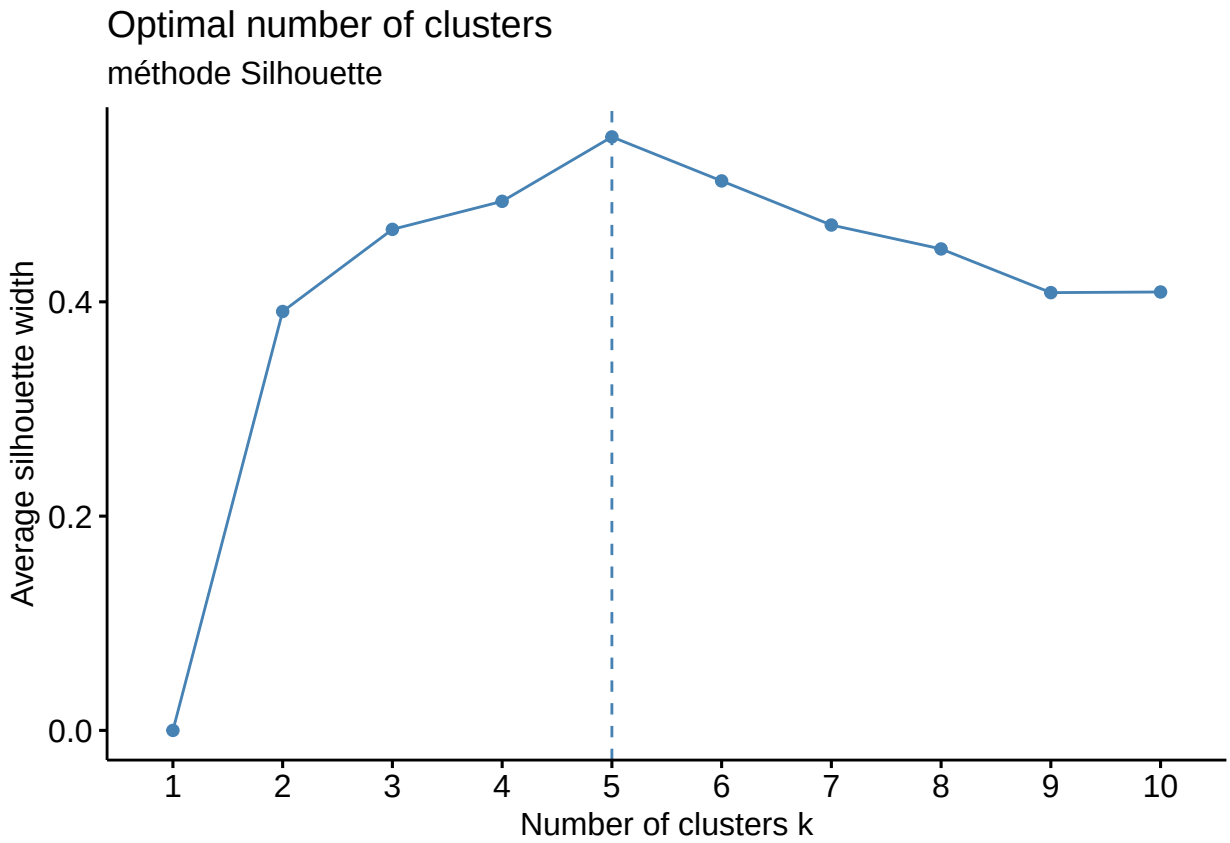
Méthode du Kmeans (Elbow,Sihouette)

Trouver le meilleur k pour le Kmeans

```
wss <- function(data, maxCluster = 10) {
    # Initialisation avec les somme de  carré
    SSw <- (nrow(data) - 1) * sum(apply(data, 2, var))
    SSw <- vector()
    for (i in 2:maxCluster) {
        SSw[i] <- sum(kmeans(data, centers = i)$withinss)# application de kmean
    }
    plot(1:maxCluster, SSw, type = "o", xlab = "Nombre de Clusters", ylab = "somme de carré", pch=19)
    abline(v=5, col="blue")
}
set.seed(100)
wss(custom)
```




```
# méthode Silhouette
fviz_nbclust(mallCust[, 4:5], kmeans, method = "silhouette")+
  labs(subtitle = "méthode Silhouette ")
```

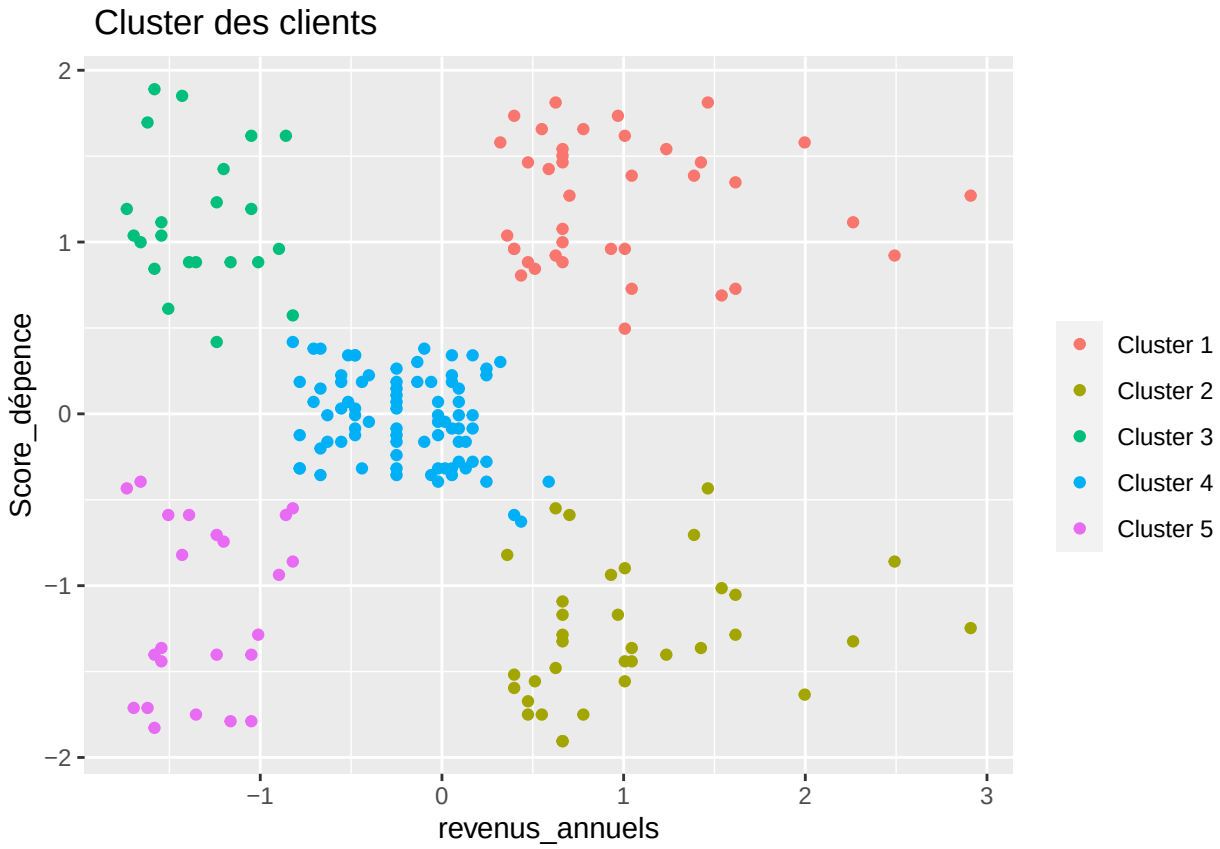


D'après les résultats de la méthode du Elbow (Coude) et Silhouette, on peut voir que le coude est un cercle de flexion k=5, donc k=5 est le nombre de groupes de clients que nous utilisons dans ce cas pour la suite .

Représentation du score de dépense en fonction du revnus selon les 5 clusters des clients

```
set.seed(111)
da=data.frame(custom)

cust.KM<-kmeans(custom,5)
ggplot(da, aes(x = revenus_annuels, y = score_depense)) +
    geom_point(stat = "identity", aes(color = as.factor(cust.KM$cluster))) +
    scale_color_discrete(name=" ",
                         breaks=c("1", "2", "3", "4", "5"),
                         labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5")) +
    ggtitle(" Cluster des clients")+
  xlab("revenus_annuels")+ylab("Score_dépence")
```



Ajoutons la colonne cluster

```
# Adding 'Cluster' column
mallCust$Cluster <- cust.KM$cluster
head(mallCust)
```

```
##   ID   Sexe Age revenus_annuels score_depense Cluster
## 1  1   Male  19              15            39       5
## 2  2   Male  21              15            81       3
## 3  3 Female  20              16             6       5
## 4  4 Female  23              16            77       3
## 5  5 Female  31              17            40       5
## 6  6 Female  22              17            76       3
```

La phase d'apprentissage consiste à trouver les paramètres du modèle optimal :

Grid : la grille, sa taille, sa forme, le type fonctionde voisinage, etc. . .

Rlen : le nombre de fois que l'ensemble des données sera présenté au réseau

Alpha : le pas de l'apprentissage pour contrôler la vitesse d'apprentissage

Radius : le rayon du voisinage

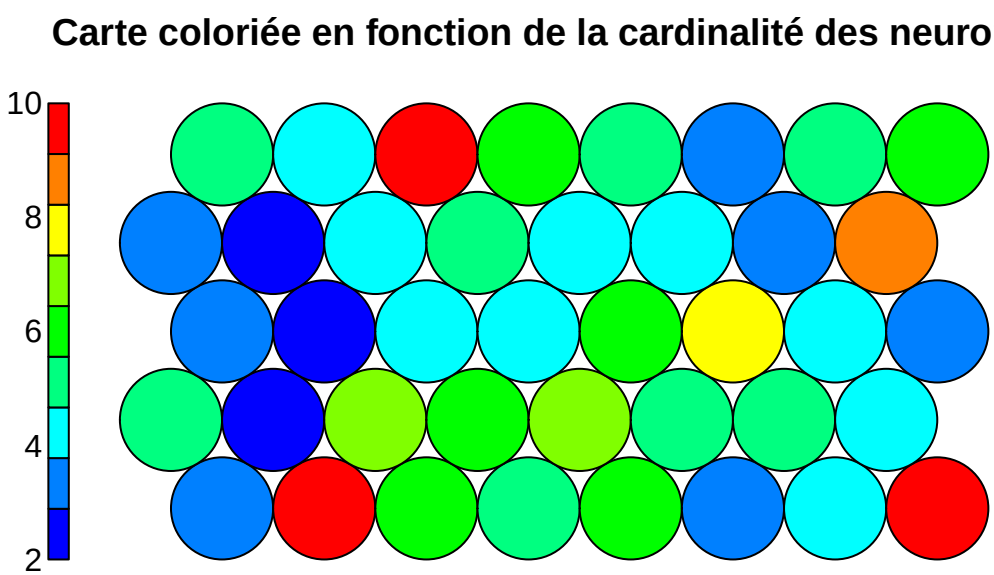
Init : les valeurs initiales des vecteurs référents

```
set.seed(111)
# Choix du type de carte et de sa taille
som.grid <- somgrid(8, 5, topo="hexagonal", neighbourhood.fct="bubble")
# Apprentissage
som.model <- som(custom, grid =som.grid, rlen=1000, alpha=c(0.05,0.01), keep.data = TRUE)
# Palette de couleur pour l'affichage des cartes
coolBlueHotRed <- function(n, alpha = 1) {rainbow(n, end=4/6, alpha=alpha)[n:1]}
```

La librairie kohonen de R présente plusieurs types de visualisation permettant de mesurer de la pertinence de la carte obtenue.

La carte coloriée en fonction de la cardinalité (le nombre d'individus capturés par un neurone) des neurones permet de mésurer la qualité de carte. La distribution de la cardinalité doit être uniforme. Des neurones présentant des cardinalités assez importantes montrent que la taille de carte est petite. De même, la présence de beaucoup de neurones avec des cardinalités nulles suggère que la carte est trop grande.

```
# Affichage des cartes
plot(som.model, type="count", main="Carte coloriée en fonction de la cardinalité des neurones", palette
```



On peut aussi représenter les neurones dans l'espace des données à l'aide des vecteurs référents. Cela permet de voir la qualité de la quantification vectorielle de l'espace d'entrée.

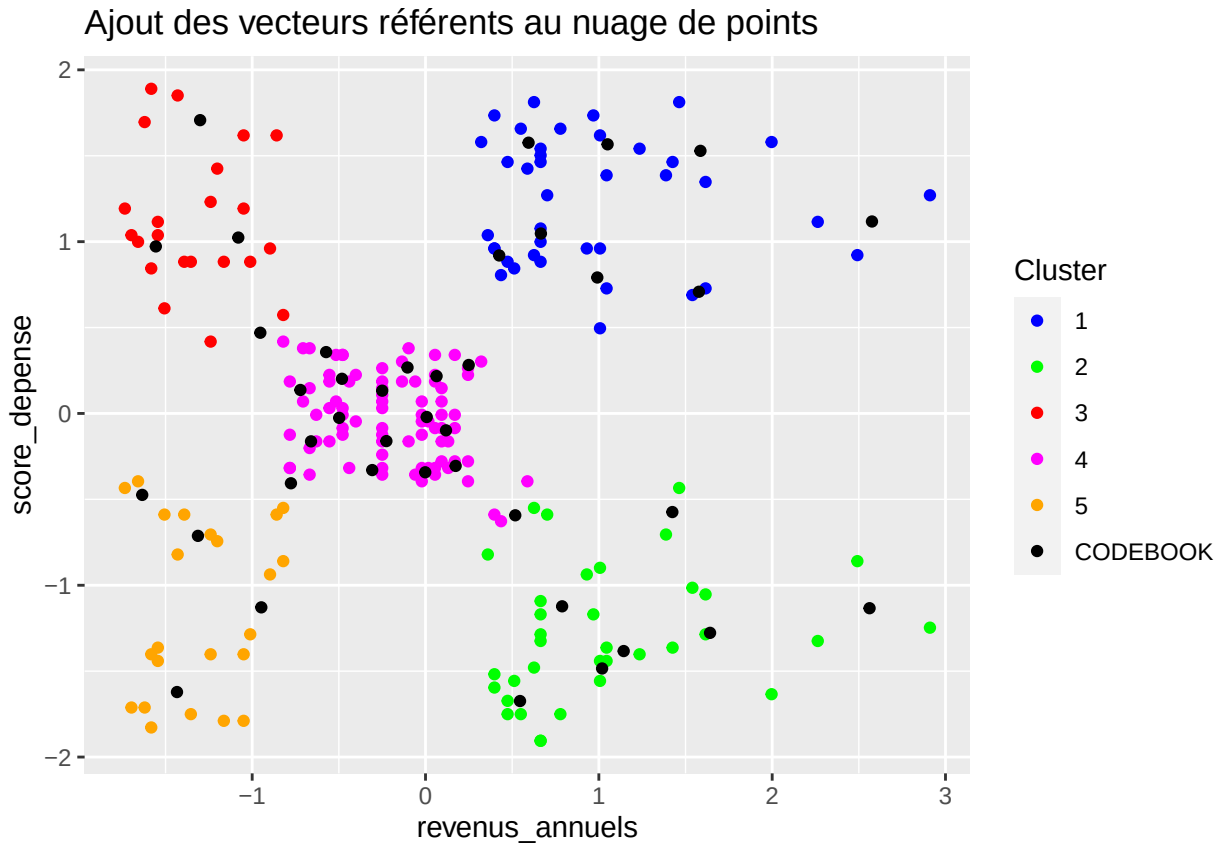
```
df = data.frame(som.model$codes)

df1=data.frame(df[,1],df[,2])
colnames(df1) = c("revenus_annuels", "score_depense")
df1$Cluster = "CODEBOOK"
```

```r
dfA = data.frame(custom) %>% setNames(nm=c("revenus_annuels", "score_depense"))
dfA$Cluster = mallCust[,c(4,5,6)]$Cluster

df1 = rbind(dfA,df1)

colours = c('blue', 'green', 'red','magenta','orange', 'black')
qplot(x = revenus_annuels, y = score_depense, color = Cluster, data = df1, geom = "point") + scale_colo
```

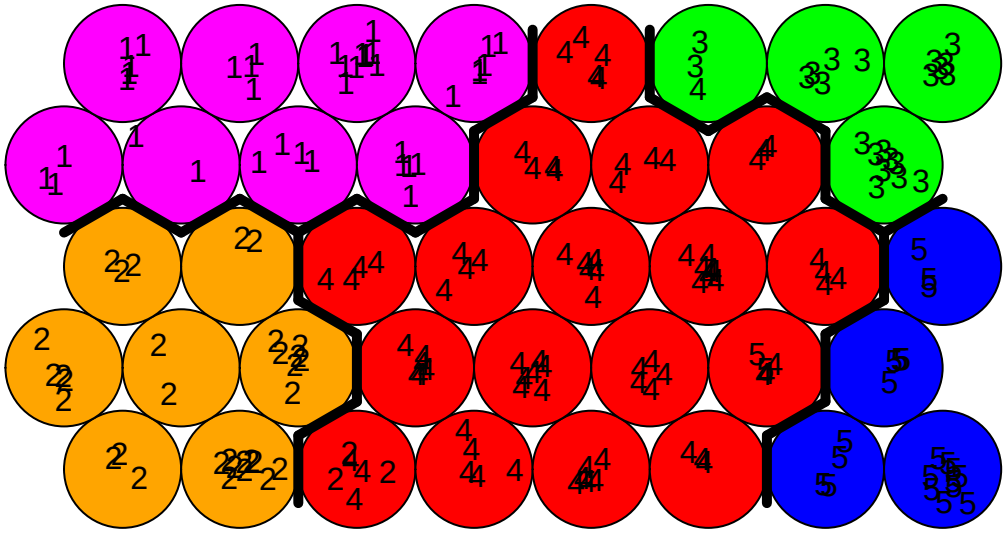


La dernière partie consiste à appliquer la classification automatique des données à partir des cartes de Kohonen. L'idée est de retrouver les cinq groupes qu'on a créés au début du l'exemple.

La décroissance de l'inertie intra-classe suggère un nombre de classe égale à 5

```r
set.seed(111)
# On fixe le nombre de classes à 3
code.books = data.frame(som.model$codes) %>% setNames(nm=c("revenus", "depense"))
model.kmeans <- kmeans(x = code.books, centers = 5, nstart=100)
plot(som.model, type="mapping",
bgcol = c('blue', 'green', 'red','magenta','orange')[model.kmeans$cluster],
labels = mallCust$Cluster,
main = "Les clusters sur la carte de kohonen")
add.cluster.boundaries(som.model, clustering = model.kmeans$cluster)
```

Les clusters sur la carte de kohonen



Les cartes topologiques de Kohonen constituent un outil puissant de réduction de dimension et de classification automatique. Elles peuvent être vues comme :

Une extension non linéaire de l'ACP (Analyse en Composantes Principales) dans le cadre d'une réduction de dimension Une extension non linéaire de K-means dans le cadre d'une classification automatique

On retrouve bien les cinq groupes définis au début l'exemple. Les neurones fournissent une classification plus fine que les k-means.

Information pour le cluster 1

```
#summary(cluster1)
```

```
   Gender       Age         Annual_Income  Spending_Score
Female:14   Min.   :19.00   Min.   :15.0   Min.   : 3.00
Male  : 9   1st Qu.:35.50   1st Qu.:19.5   1st Qu.: 9.50
            Median :46.00   Median :25.0   Median :17.00
            Mean   :45.22   Mean   :26.3   Mean   :20.91
            3rd Qu.:53.50   3rd Qu.:33.0   3rd Qu.:33.50
            Max.   :67.00   Max.   :39.0   Max.   :40.00
```

Information pour le cluster 2

```
#summary(cluster2)
```

Gender	Age	Annual_Income	Spending_Score
Female:48	Min. :18.00	Min. :39.0	Min. :34.00
Male :33	1st Qu.:27.00	1st Qu.:48.0	1st Qu.:44.00
	Median :46.00	Median :54.0	Median :50.00
	Mean :42.72	Mean :55.3	Mean :49.52
	3rd Qu.:54.00	3rd Qu.:62.0	3rd Qu.:55.00
	Max. :70.00	Max. :76.0	Max. :61.00

Figure 1: résumé du cluster 3.

Gender	Age	Annual_Income	Spending_Score
Female:16	Min. :19.00	Min. : 70.0	Min. : 1.00
Male :19	1st Qu.:34.00	1st Qu.: 77.5	1st Qu.:10.00
	Median :42.00	Median : 85.0	Median :16.00
	Mean :41.11	Mean : 88.2	Mean :17.11
	3rd Qu.:47.50	3rd Qu.: 97.5	3rd Qu.:23.50
	Max. :59.00	Max. :137.0	Max. :39.00

Information pour le cluster 3

```
#summary(cluster3)
```

Information pour le cluster 4

```
#summary(cluster4)
```

Gender	Age	Annual_Income	Spending_Score
Female:13	Min. :18.00	Min. :15.00	Min. :61.00
Male : 9	1st Qu.:21.25	1st Qu.:19.25	1st Qu.:73.00
	Median :23.50	Median :24.50	Median :77.00
	Mean :25.27	Mean :25.73	Mean :79.36
	3rd Qu.:29.75	3rd Qu.:32.25	3rd Qu.:85.75
	Max. :35.00	Max. :39.00	Max. :99.00

Information pour le cluster 5

```
#summary(cluster5)
```

Gender	Age	Annual_Income	Spending_Score
Female:21	Min. :27.00	Min. : 69.00	Min. :63.00
Male :18	1st Qu.:30.00	1st Qu.: 75.50	1st Qu.:74.50
	Median :32.00	Median : 79.00	Median :83.00
	Mean :32.69	Mean : 86.54	Mean :82.13
	3rd Qu.:35.50	3rd Qu.: 95.00	3rd Qu.:90.00
	Max. :40.00	Max. :137.00	Max. :97.00

Figure 2: résumé du cluster 5.

Interprétation pour le groupe/segment de clients :

Cluster 1. Les clients ayant un revenu annuel élevé mais un score de dépenses faible.

Cluster 2. Clients ayant un revenu annuel moyen et un score moyen en matière de dépenses.

Cluster 3. Clients ayant un faible revenu annuel et un faible niveau de dépenses.

Cluster 4. Clients ayant un revenu annuel faible mais un niveau de dépenses élevé.

Cluster 5. Clients ayant un revenu annuel élevé et un score élevé en matière de dépenses.

les données fournies ci-dessus, nous pourrions prendre un résumé d'analyse qui peut être utilisé pour le plan de marketing comme suit :

J'ai pu constater que le pourcentage de clients féminins (56%) est légèrement supérieur à celui des clients masculins (44%), ce qui nous a permis de cibler davantage les clients masculins pour les campagnes de marketing ou les promotions que les clients féminins, même si le pourcentage différent n'est pas trop important. Dans ce cas, nous pouvons également choisir judicieusement la cible des clients masculins en combinant des facteurs liés à leur âge et à leur groupe.

Le centre commercial peut mener des campagnes de marketing ou des programmes de fidélisation auprès des clients des groupes 5 et 4, qui sont des clients ayant un niveau de dépenses élevé, en particulier les clients âgés de 20 et 30 ans, afin de conserver cette clientèle et d'augmenter les possibilités de vente.

Les groupes 1 et 3 ont en général un faible niveau de dépenses, malgré leur niveau de revenu, et les clients sont généralement âgés de plus de 40 ans. Avec ces données, nous pourrions envisager de rechercher et d'ajouter certaines marques qui sont populaires parmi les clients de ces âges, et de mener des campagnes pour les cibler avec les bons produits.

Comme nous pouvons le voir à partir des données ci-dessus, le groupe 2 a un score de dépenses moyen, afin d'augmenter les ventes dans ce groupe de clients, nous pourrions leur donner quelques promotions pour les encourager à acheter plus de produits.