# Analyse descriptive des profils de joueurs du Casino de Moncton

Travail remis dans le cadre du cours Analytique des affaires (MQT-6021)

*Nicolas Corneau-Tremblay*

*17 octobre 2017*

## Préambule

Ce travaille est effectué en R. Il utilise les librairies suivantes :

```
library(dplyr)
library(ggplot2)
library(readxl)
```

Les données sont importées ainsi :

```
profil <- read_xlsx("Donnees.xlsx", sheet = "Profils", col_names = TRUE, skip = 1)
```

### Question 1. Est-ce que certaines des variables retenues pour l'analyse doivent être recodées pour être traitées par des algorithmes analytiques? Si oui, faites le recodage dans votre fichier conception.

```
summary(profil)
```

```
##   Identifiant       NbVisites          Sexe             DepenseJeu
##  Min.   :  1.00   Min.   :  3.00   Length:100         Min.   : 138
##  1st Qu.: 25.75   1st Qu.: 11.00   Class :character   1st Qu.: 753
##  Median : 50.50   Median : 24.00   Mode  :character   Median :1366
##  Mean   : 50.50   Mean   : 25.71                      Mean   :1958
##  3rd Qu.: 75.25   3rd Qu.: 35.00                      3rd Qu.:2970
##  Max.   :100.00   Max.   :276.00                      Max.   :8352
##  DepensesAutres   JoueAuxTables  JoueAuxMachines
##  Min.   :  0.00   Min.   :0.00   Min.   :1
##  1st Qu.:  0.00   1st Qu.:0.00   1st Qu.:1
##  Median : 60.00   Median :0.00   Median :1
##  Mean   : 81.24   Mean   :0.18   Mean   :1
##  3rd Qu.:147.00   3rd Qu.:0.00   3rd Qu.:1
##  Max.   :264.00   Max.   :1.00   Max.   :1
```

Oui, la variable nominale *Sexe* doit être recodée pour devenir une variable dichotomique. Ceci est fait de la façon suivante :

```
profil <- profil %>%
  mutate(Femme = ifelse(Sexe == "F", 1, 0))
```

Nous avons ainsi créé une nouvelle variable qui prend la valeur zéro si l'individu est un homme et un si l'individu est une femme. Également, les variables *NbVisites*, *DepenseJeu* et *DepensesAutres* pourraient être normalisées afin de s'assurer que leur échelle respective n'affecte pas les analyses. Ceci est fait de la façon suivante :

```r
profil <- profil %>%
  mutate(NbVisites = (NbVisites - mean(NbVisites))/sd(NbVisites),
         DepenseJeu = (DepenseJeu - mean(DepenseJeu))/sd(DepenseJeu),
         DepensesAutres = (DepensesAutres - mean(DepensesAutres))/sd(DepensesAutres))
summary(profil)
```

```
##   Identifiant       NbVisites            Sexe             DepenseJeu
##  Min.   :  1.00   Min.   :-0.78992   Length:100         Min.   :-1.0872
##  1st Qu.: 25.75   1st Qu.:-0.51166   Class :character   1st Qu.:-0.7199
##  Median : 50.50   Median :-0.05948   Mode  :character   Median :-0.3537
##  Mean   : 50.50   Mean   : 0.00000                      Mean   : 0.0000
##  3rd Qu.: 75.25   3rd Qu.: 0.32313                      3rd Qu.: 0.6046
##  Max.   :100.00   Max.   : 8.70580                      Max.   : 3.8193
##
##  DepensesAutres    JoueAuxTables  JoueAuxMachines     Femme
##  Min.   :-0.9677   Min.   :0.00   Min.   :1       Min.   :0.0000
##  1st Qu.:-0.9677   1st Qu.:0.00   1st Qu.:1       1st Qu.:0.0000
##  Median :-0.2530   Median :0.00   Median :1       Median :0.0000
##  Mean   : 0.0000   Mean   :0.18   Mean   :1       Mean   :0.3636
##  3rd Qu.: 0.7833   3rd Qu.:0.00   3rd Qu.:1       3rd Qu.:1.0000
##  Max.   : 2.1769   Max.   :1.00   Max.   :1       Max.   :1.0000
##                                                   NA's   :1
```

Sinon, il ne semble pas y avoir à première vue de valeur aberrante dans les données.

## Question 2. On cherche à segmenter les clients (en fonction des données disponibles) afin d'obtenir entre 3 et 4 groupes de clients relativement homogènes entre eux.

### a. Est-ce que certaines variables sont non pertinentes pour la segmentation? Justifiez.

Oui, certaines variables sont non pertinentes pour la segmentation. Premièrement, la variable *Identifiant* est utile afin de structurer les données et de retrouver les individus de façon unique, mais elle n'est (généralement) pas utile lors de la segmentation.

Deuxièmement, la variable *JoueAuxMachines* ne contient aucune variation et prend systématiquement la valeur un. Elle ne peut donc pas être utilisée lors de la segmentation, puisqu'elle ne peut permettre l'identification de certains groupes.

### b. On vous demande de choisir au maximum (2) variables pour segmenter les clients. Indiquez les deux variables les plus pertinentes et justifiez. (1 points)

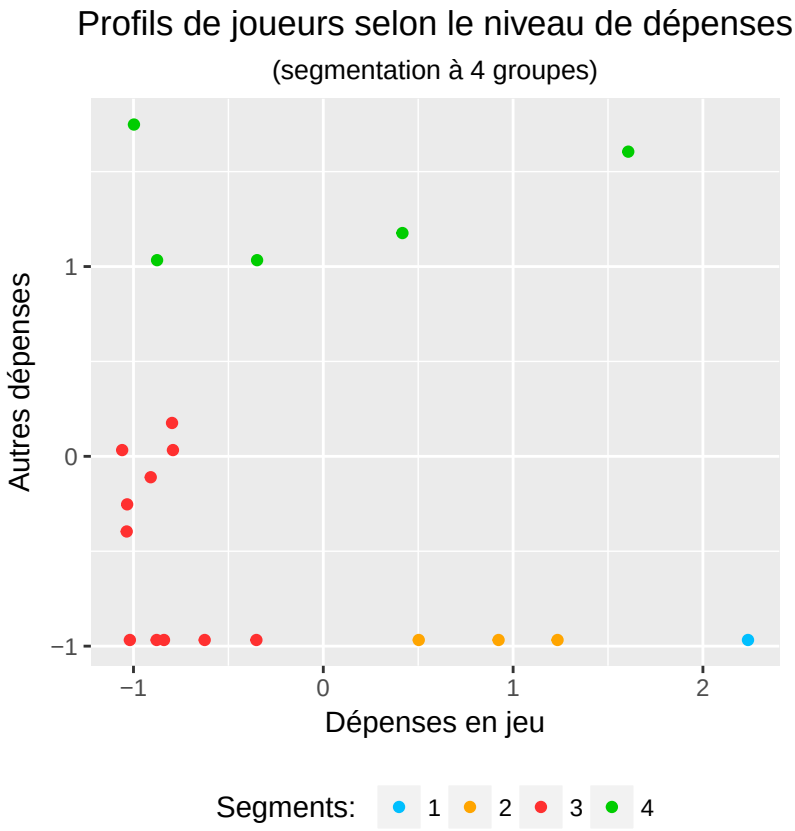
Selon nous, les deux variables les plus importantes pour segmenter les joueurs sont *DepenseJeu* et *DepensesAutres*. En effet, si le casino est intéressé par ses profits, ces deux variables sont déterminantes dans sa réussite financière, puisqu'elles réprésentent des sources de revenus. Elles permettent donc de répondre à la question "qui dépense combien et comment?"

## Question 3. Réalisez la segmentation sur vos deux variables à l'aide de la méthode des k-moyennes, en utilisant seulement les 20 premières observations :

### a. à k=4 groupes.

Le code suivant permet d'effectuer la segmentation à l'aide des 20 premières observations. La segmentation est faite pour quatre groupes sur la base des dépenses de jeu et des autres dépenses. Les quatre groupes formés par la segmentation sont présentés ensuite dans une figure.

```
# sélection des variables et des observations
top.20 <- profil %>%
  select(c("DepenseJeu", "DepensesAutres")) %>%
  head(., 20)
# définition d'un seed pour que les résultats soient toujours les mêmes
set.seed(222)
# segmentation à l'aide de la fonction kmeans
estim <- kmeans(top.20, centers = 4)
top.20 <- top.20 %>%
  mutate(cluster = as.factor(estim$cluster))
# visualisation des résultats
ggplot(top.20) +
  geom_point(aes(DepenseJeu, DepensesAutres, color = cluster)) +
  scale_color_manual(breaks = c("1", "2", "3", "4"),
                     values = c("1" = "deepskyblue1", "2" = "orange",
                                "3" = "firebrick1", "4" = "green3"),
                     name = "Segments: ") +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5)) +
  labs(title = "Profils de joueurs selon le niveau de dépenses",
       subtitle = "(segmentation à 4 groupes)") +
  ylab(label = "Autres dépenses") +
  xlab(label = "Dépenses en jeu") +
  coord_fixed()
```




**Question 4. Identifiez brièvement (quelques lignes au maximum) les caractéristiques propres à chaque segment qui permettent de le différencier des autres.**

Le premier segment (en bleu) est possiblement formé de joueurs professionnels ou de gens très fortunés dépensant des montants importants dans le jeu. Le deuxième segment (en orange) parait être des joueurs sérieux ne dépensant qu'en jeu. Le troisième segment semble être des clients qui viennent au casino à des fins récréatives. Ils dépensent relativement peu et forment le groupe le plus nombreux. Le quatrième et dernier segment (en vert) semble lui-aussi être présent dans une perspective récréative puisqu'il dépense de façon marquée, notamment dans les autres dépenses. Il semble toutefois plus fortuné que le troisième segment.

Évidemment, cette analyse est limitée par le nombre restreint d'observations utilisées pour effectuer la segmentation.

**Question 5. Supposons que le décideur vous demande s'il est possible d'obtenir d'autres segmentations à partir des mêmes données, que lui répondriez-vous ? Justifiez votre réponse.**

Oui il est possible d'effectuer d'autres segmentations à partir des mêmes données. D'abord, seulement deux variables ont été utilisées pour effectuer les segments précédents. D'autres variables disponibles pourraient être mises à profit afin d'obtenir des profils plus fins et détaillés. Par ailleurs, il est possible de changer le nombre de segments afin de voir si quatre groupes est le nombre le plus approprié pour représenter les données que nous possédons. Ce processus pourrait être effectué à l'aide de la méthode de validation croisée.