# Projet Analyse de Données

Abdelhedi Youssef

Pour Commencer, après avoir téléchargé nos données sous format csv depuis Google Forms, nous devons les charger :

```
questionnaire <- read.csv("D:/A garder/maths/M1/Analyse de données/Formulaire.csv")
```

## Premier ACP

### Traitement des Données

Pour cela, nous devons simplement extraire le premier bloc (toutes les questions qui commencent par "Qu'est-ce qui influence votre envie de répondre à des questionnaires") :

```
colonnes <- grep(
  "^Qu\\.est\\.ce\\.qui\\.influence\\.votre\\.envie\\.de\\.répondre\\.à\\.des\\.questionnaires\\.",
  colnames(questionnaire),
  value = TRUE)
```

Et ensuite changer les valeurs pour qu'elles correspondent à l'échelle de Likert :

```
questionnaire <- questionnaire %>%
 mutate(across(all_of(colonnes), ~replace(., . == "Pas du tout d'accord", 1))) %>%
 mutate(across(all_of(colonnes), ~replace(., . == "Pas d'accord", 2))) %>%
 mutate(across(all_of(colonnes), ~replace(., . == "Neutre", 3))) %>%
 mutate(across(all_of(colonnes), ~replace(., . == "Plutôt D'accord", 4))) %>%
 mutate(across(all_of(colonnes), ~replace(., . == "Tout à fait D'accord", 5)))
```

Enfin changer le nom des colonnes pour qu'elles soient plus lisibles :

```
data_selectionnee <- questionnaire[colonnes]
nouveaux_noms <- c(
  "Clarté des Questions", "La Longueur", "Le créateur", "L'heure",
  "Le sujet","L'objectif","L'organisation","La langue","La Methode de Partage",
  "La récompense","Qualité des Qst","L'emetteur","Le temps","L'humeur","Reseau Social")
colnames(data_selectionnee) <- nouveaux_noms
```

Commençons par comprendre la data que nous avons maintenant :

```
summary(data_selectionnee)
```

```
##  Clarté des Questions La Longueur        Le créateur          L'heure
##  Length:40            Length:40          Length:40          Length:40
##  Class :character     Class :character   Class :character   Class :character
##  Mode  :character     Mode  :character   Mode  :character   Mode  :character
##    Le sujet          L'objectif        L'organisation      La langue
##  Length:40          Length:40          Length:40          Length:40
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  La Methode de Partage La récompense      Qualité des Qst     L'emetteur
##  Length:40              Length:40          Length:40          Length:40
##  Class :character       Class :character   Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character   Mode  :character
##    Le temps           L'humeur         Reseau Social
##  Length:40          Length:40          Length:40
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

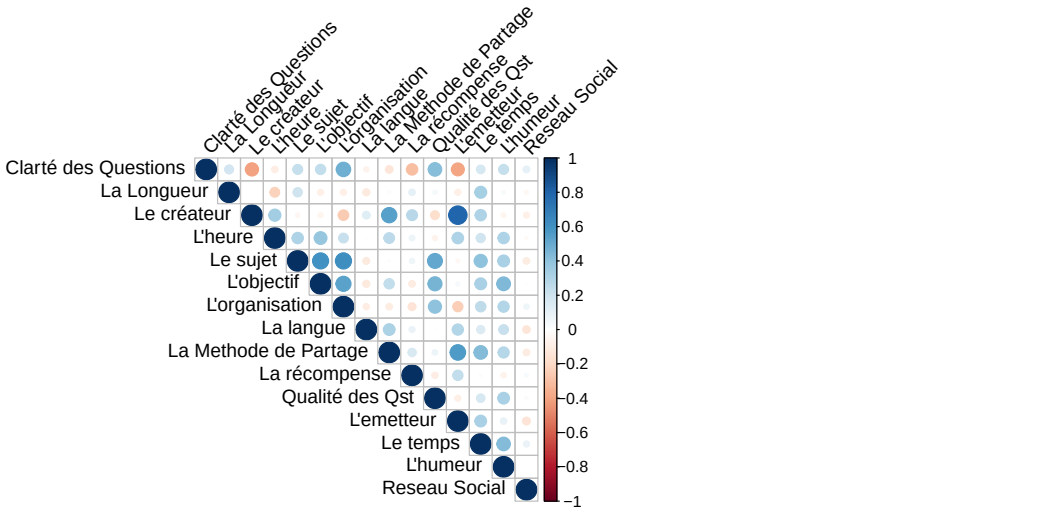
Nous remarquons donc qu'il faut que nous rendions toutes les valeurs numériques

```
data_selectionnee <- apply(data_selectionnee, 2, function(x) as.numeric(as.character(x)))
```

### Analyse des Données

Voyons ce que nous propose la matrice de corrélation :

```
matrice_correlation <- cor(data_selectionnee)
corrplot(matrice_correlation, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
```

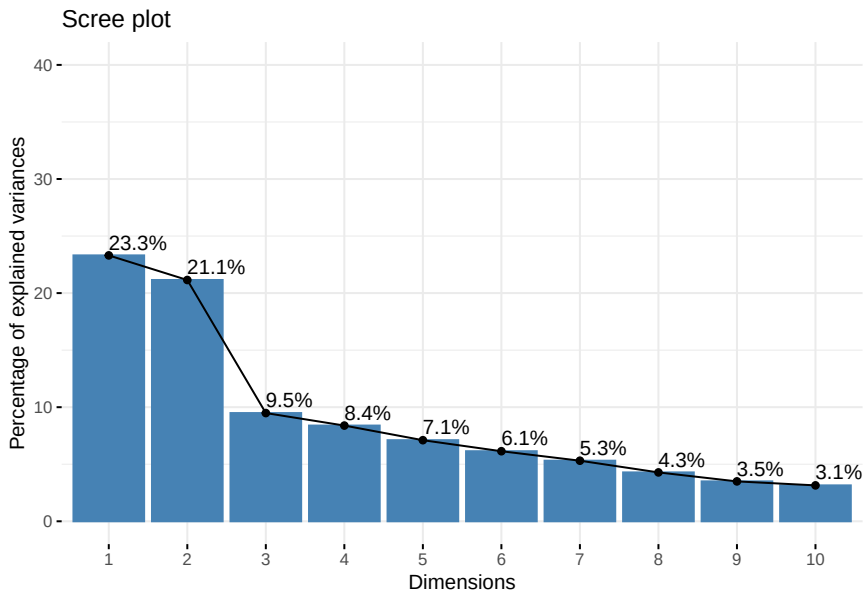



Nous pouvons voir notamment que la variable "créateur du formulaire" est très liée à la variable "émetteur" à environ 80% ainsi qu'à la "méthode de partage" à 55%

Nous pouvons aussi remarquer que les variables "Sujet","Objectif" et "Organisation" sont corrélées à hauteur d'environ 50%

Maintenant effectuons l'ACP :

```
acp_result <- PCA(data_selectionnee, graph = FALSE)
fviz_eig(acp_result, addlabels = TRUE, ylim = c(0, 40))
```




Ici nous pouvons choisir de garder 2 ou 3 axes. En effet, ils ont tous des Valeurs Propres supérieures à 1 (critère de Kaiser). Le coude se trouve au niveau du 3ème axe et l'inertie cumulée est autour des 50% malgré le fait qu'il y ait 15 variables. Choisissons-en 2.

**Interpretation**

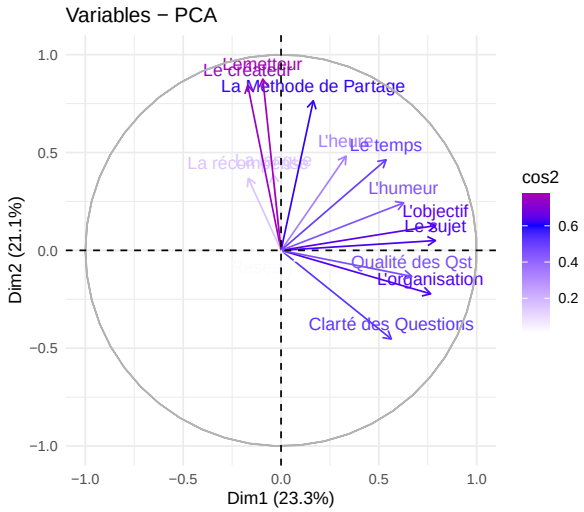
```
acp_result$var$coord[, 1:2]
```

```
##                            Dim.1       Dim.2
## Clarté des Questions   0.56383704 -0.45299770
## La Longueur            0.10965769 -0.04924791
## Le créateur           -0.17012742  0.84340094
## L'heure                0.33224758  0.48130477
## Le sujet               0.78898775  0.05010600
## L'objectif             0.79027269  0.12698524
## L'organisation         0.76468307 -0.22213604
## La langue             -0.03800761  0.38723987
## La Methode de Partage  0.16475089  0.76464913
## La récompense         -0.16745352  0.36824138
## Qualité des Qst        0.66854851 -0.13288838
## L'emetteur            -0.09317590  0.87535397
## Le temps               0.53690219  0.46306713
## L'humeur               0.62670619  0.24286429
## Reseau Social          0.03506089 -0.16112179
```

Nous pouvons ici voir que la première dimension est en rapport avec "Le sujet", "L'objectif", "L'organisation", "La récompense", "Qualité des Qst" et "L'humeur". Nous pouvons donc l'appeler Fond et Timing. La seconde est en rapport avec "Le créateur", "La Methode de Partage" et "L'emetteur". Nous pouvons donc l'appeler La conception.

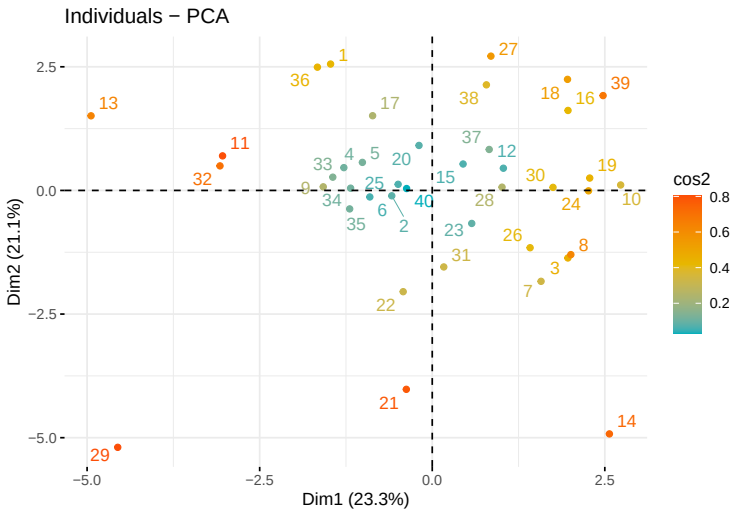
Nous Pouvons remarquer ces mêmes informations dans le graphique ci-dessous :

```
fviz_pca_var(acp_result, col.var="cos2") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.6) +
  theme_minimal()
```




Maintenant, Essayons de comprendre les individus :

```
fviz_pca_ind(acp_result, col.ind = "cos2",
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
 repel = TRUE
 )
```




Nous Pouvons remarquer qu'il y a un groupe de personnes qui donne peu d'importance à la Conception et encore moins au fond et au Timing, Ils sont donc désintéressés de l'action même de remplir un formulaire. Et plus on s'éloigne, moins il y a d'individus, leur repartition suit certainement une loi normale..

## Second ACP

### Traitement des Données

Pour cela, nous devons simplement extraire le second bloc (de la colonne 21 à 35)

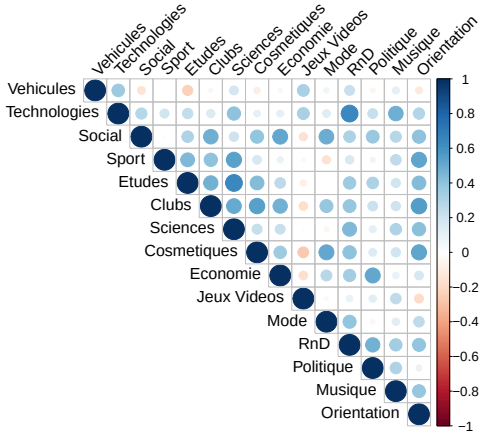
```
data_selectionnee <- questionnaire[, 21:35]
nouveaux_noms <- c(
    "Vehicules", "Technologies", "Social", "Sport","Etudes","Clubs","Sciences",
    "Cosmetiques","Economie","Jeux Videos","Mode","RnD","Politique","Musique","Orientation")
colnames(data_selectionnee) <- nouveaux_noms
summary(data_selectionnee)
```

```
##    Vehicules      Technologies      Social          Sport          Etudes
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.00   Min.   :1.00
##  1st Qu.:1.750   1st Qu.:3.00   1st Qu.:3.750   1st Qu.:3.00   1st Qu.:3.00
##  Median :3.000   Median :4.00   Median :4.500   Median :4.00   Median :4.00
##  Mean   :2.875   Mean   :4.05   Mean   :4.125   Mean   :3.55   Mean   :3.85
##  3rd Qu.:4.000   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:5.00
##  Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.00   Max.   :5.00
##      Clubs         Sciences      Cosmetiques       Economie      Jeux Videos
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0
##  1st Qu.:2.00   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.0
##  Median :3.00   Median :4.000   Median :2.000   Median :3.000   Median :3.0
##  Mean   :3.25   Mean   :4.025   Mean   :2.375   Mean   :3.025   Mean   :2.9
##  3rd Qu.:4.25   3rd Qu.:5.000   3rd Qu.:3.250   3rd Qu.:4.000   3rd Qu.:4.0
##  Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.0
##       Mode            RnD          Politique        Musique       Orientation
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0
##  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.0
##  Median :3.000   Median :4.000   Median :3.000   Median :4.000   Median :3.0
##  Mean   :3.075   Mean   :3.725   Mean   :2.875   Mean   :3.975   Mean   :3.2
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.0
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.0
```

### Analyse des Données

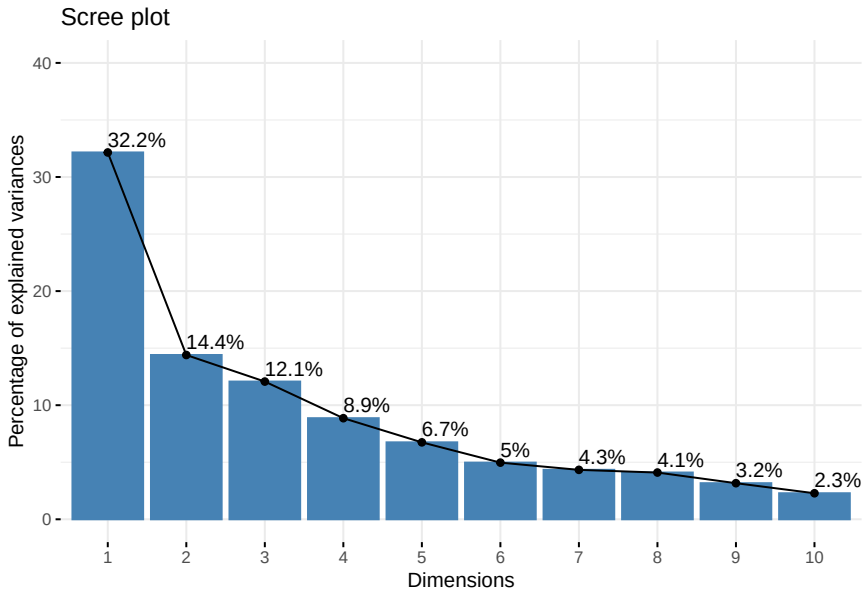
Voyons ce que nous propose la matrice de corrélation :

```
matrice_correlation <- cor(data_selectionnee)
corrplot(matrice_correlation, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
```



Maintenant effectuons l'ACP :

```
acp_result <- PCA(data_selectionnee, graph = FALSE)
fviz_eig(acp_result, addlabels = TRUE, ylim = c(0, 40))
```




Ici nous pouvons choisir de garder 2. En effet, ils ont tous des Valeurs Propres supérieures à 1 (critère de Kaiser). Le coude se trouve au niveau du 2ème axe et l'inertie cumulée est autour des 47% malgré le fait qu'il y ait 15 variables.

**Interpretation**

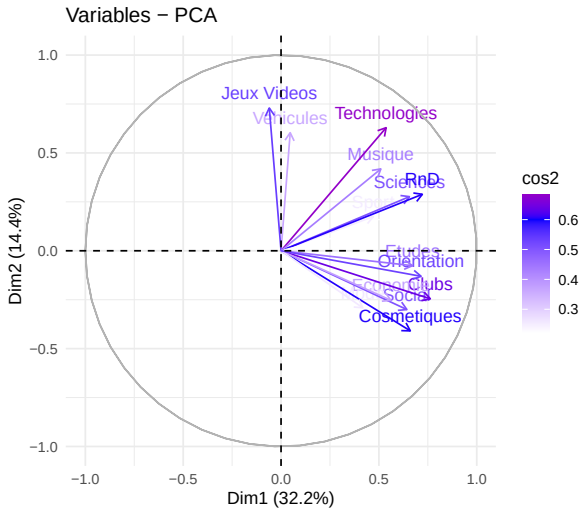
```
acp_result$var$coord[, 1:2]
```

```
##                    Dim.1      Dim.2
## Vehicules     0.04670966  0.6034721
## Technologies  0.53697435  0.6290448
## Social        0.64294174 -0.3004132
## Sport         0.47159414  0.1746898
## Etudes        0.67138709 -0.0761582
## Clubs         0.76274493 -0.2454059
## Sciences      0.65615110  0.2766407
## Cosmetiques   0.66060207 -0.4093864
## Economie      0.56114097 -0.2530836
## Jeux Videos  -0.06014481  0.7296112
## Mode          0.41773924 -0.2913497
## RnD           0.72199663  0.2886653
## Politique     0.45871405  0.1118671
## Musique       0.50997039  0.4182240
## Orientation   0.71555008 -0.1304903
```

Nous pouvons ici voir que la première dimension est en rapport avec "Etudes", "Clubs", "Sciences", "Cosmetiques", "Economie", "RnD" et "Orientation". Nous pouvons donc l'appeler Culture. La seconde est en rapport avec "Vehicules","Technologies" et "Jeux Videos". Nous pouvons donc l'appeler "Divertissements Technologiques".

Nous Pouvons remarquer ces mêmes informations dans le graphique ci-dessous :

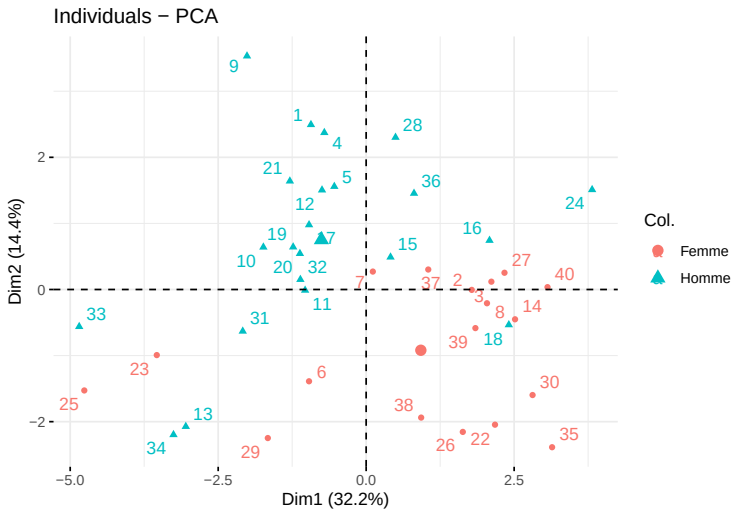
```
fviz_pca_var(acp_result, col.var="cos2") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.6) +
  theme_minimal()
```




Maintenant, Essayons de comprendre les individus :

```
colors <- ifelse(questionnaire$Sexe == "M", "Homme", "Femme")

fviz_pca_ind(acp_result, col.ind = colors, repel = TRUE)
```




Nous pouvons remarquer qu'il y a un groupe de Personnes qui est plutot intéressé par la culture et plutot neutre vis-à-vis du divertissement technologique, tandis qu'un second a tendance à ne pas aimer la culture mais plutot etre intéressé par le divertissement technologique. Enfin, il y a ceux qui ne sont interessés par aucun des deux. Nous remarquons que le premier groupe est essentiellement constitué de femmes, tandis que le second d'hommes.

## ACM

### Traitement des Données

Commençons par extraire uniquement les colonnes dont nous avons besoin :

```
colonnes_a_retirer <- c(
     "Combien.de.temps.au.maximum.vous.passez.à.répondre.à.un.questionnaire..",
     "Comment.choisissez.vous.les.questionnaires.auxquels.vous.décidez.de.répondre.",
     "Quelles.informations.seriez.vous.réticents.à.fournir.dans.un.questionnaire..même.de.manière.anony
)
questionnaire <- questionnaire[, !names(questionnaire) %in% colonnes_a_retirer]
colonnes <- cbind(questionnaire[, 36:ncol(questionnaire)], questionnaire[, c(2, 3, 4)])
```

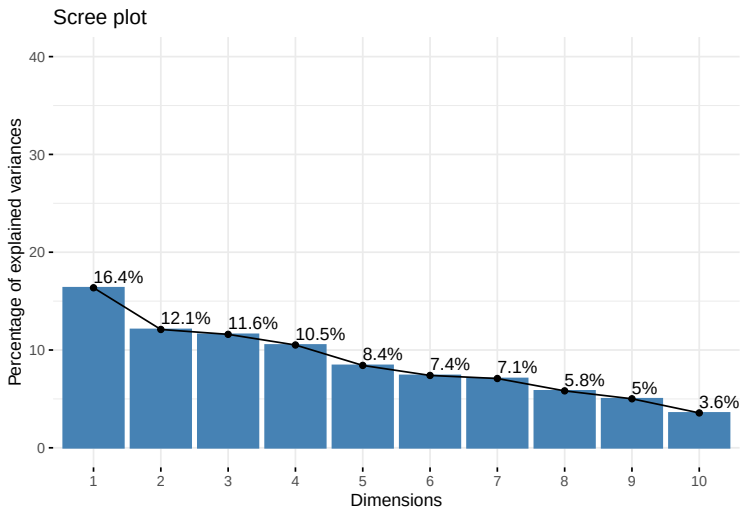
Ensuite, pour que ce soit plus lisible, nous allons changer le nom des colonnes :

```
nouveaux_noms<-c(
  "abandon_longueur","type_questionnaire","fausses_informations",
  "anonymat","creation","importance_opinion","sondage","reflexion","humeur",
  "createur","type_question","longueur","sexe","age","SSP"
                  )
colnames(colonnes) <- nouveaux_noms
```

### Analyse des Données

Une fois cela a été éffectué, réalisons une ACM sur les preferences des individus :

```
res.MCA <- MCA(colonnes, quanti.sup = c(12), graph = FALSE)
fviz_eig(res.MCA, addlabels = TRUE, ylim = c(0, 40))
```

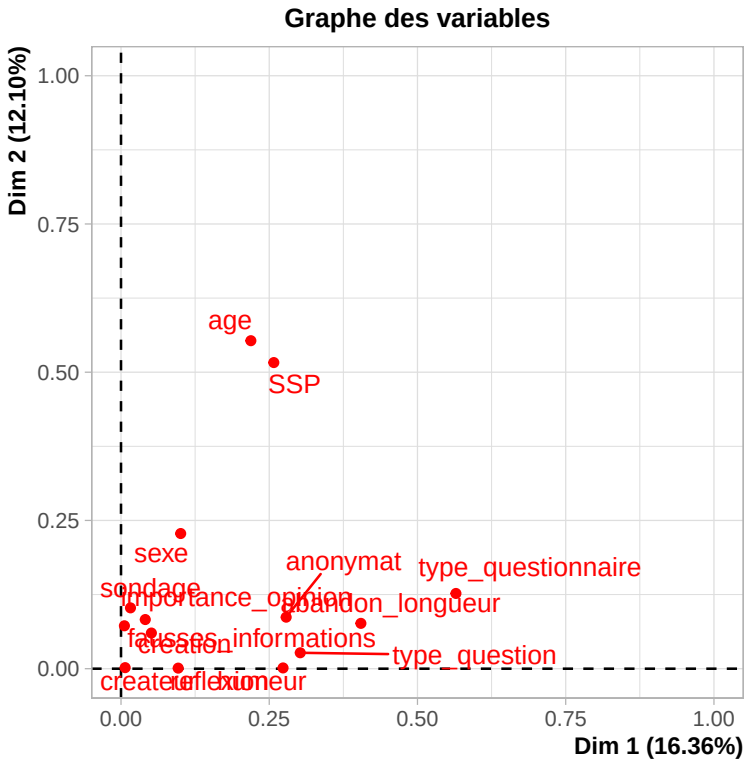



Nous Remarquons donc que le coude est au bout de la 2ème dimension et que l'inertie cumulée est d'environ 30% malgré la présence de 12 variables (ce qui reste quand même faible)

### Interpretation

Voyons ce que chacune des 2 premières dimensions représente :

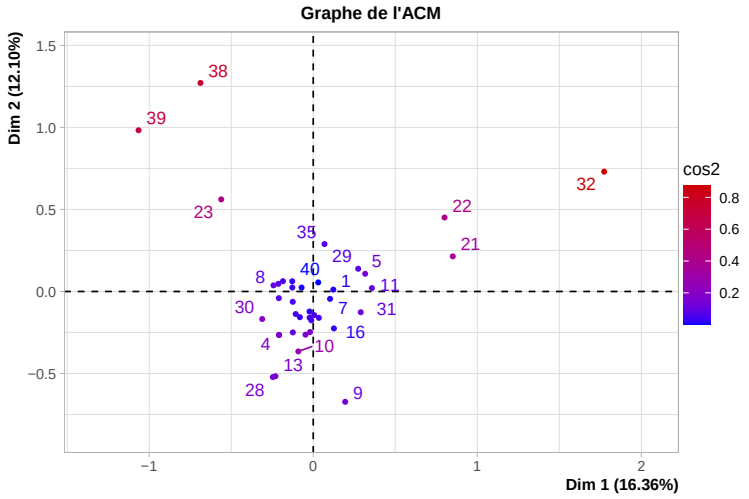
```
plot.MCA(res.MCA, choix='var',invisible='quanti.sup',title="Graphe des variables")
```




Nous pouvons voir que la première dimension est influencée par les préférences du type de questionnaires (en ligne ou papier) mais aussi par la préférence de longueur du questionnaire. Quant à la deuxième dimension elle est influencée par la catégorie d'age et la situation socio-professionnelle de l'individu

```
plot.MCA(res.MCA,invisible= 'var',habillage='cos2',title="Graphe de l'ACM",label =c('ind'))
```

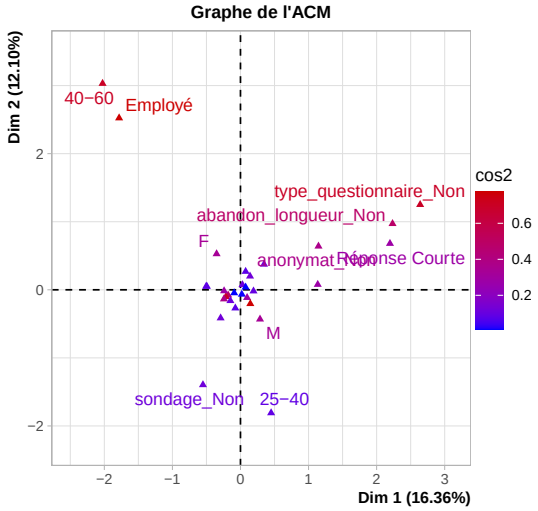
```
## Warning: ggrepel: 18 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




Nous Pouvons remarquer que la plupart des individus se concentrent au centre du graphique avec un faible cos2 (donc la représentation n'est pas très effective), et que les individus qui sont plus loin du centre sont mieux représentés

```
plot.MCA(res.MCA,invisible= 'ind',habillage='cos2',title="Graphe de l'ACM",label =c('var'))
```

```
## Warning: ggrepel: 20 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




Nous Pouvons donc voir que les individus 38 et 39 ont en commun le fait d'être employés et qu'ils sont dans la tranche d'age 40-60 et que l'individu 32 préfère les questionnaires papiers et qu'il a en commun avec l'individu 22 et 21 les réponses courtes et l'abandon des questionnaires longs