R Notebook

Nguyen Thibaut

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Importation des librairies

```
library(plyr)
library(tidyr)
library(corrplot)

## corrplot 0.88 loaded

library(scales)
library(FactoMineR)
library(factoextra)

## Le chargement a nécessité le package : ggplot2

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

library(ggplot2)
library(reshape2)

##
## Attachement du package : 'reshape2'

## L'objet suivant est masqué depuis 'package:tidyr':
##
##     smiths

library(purrr)

##
## Attachement du package : 'purrr'

## L'objet suivant est masqué depuis 'package:scales':
##
##     discard

## L'objet suivant est masqué depuis 'package:plyr':
##
##     compact

library(caret)
```

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

##
## Attachement du package : 'caret'

## L'objet suivant est masqué depuis 'package:purrr':
##
##     lift

library(naniar)
library(logistf)
library(glmnet)

## Le chargement a nécessité le package : Matrix

##
## Attachement du package : 'Matrix'

## Les objets suivants sont masqués depuis 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-2

library(dplyr)

##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union

library(knitr)
library(rmarkdown)
library(tinytex)

# Chargement de la base de données notes.csv
billets <- read.table("data/notes.csv",header=TRUE, sep=",")

# Conversion de la colonne "is_genuine" en valeur booléenne
billets$is_genuine <- as.logical(billets$is_genuine)

# Identification des valeurs manquantes
billets[!complete.cases(billets),]
```

```
## [1] is_genuine   diagonal     height_left  height_right margin_low
## [6] margin_up    length
## <0 lignes> (ou 'row.names' de longueur nulle)
```

Partie 0 (Analyse)

```
# Taille du dataframe
dim(billets)
```

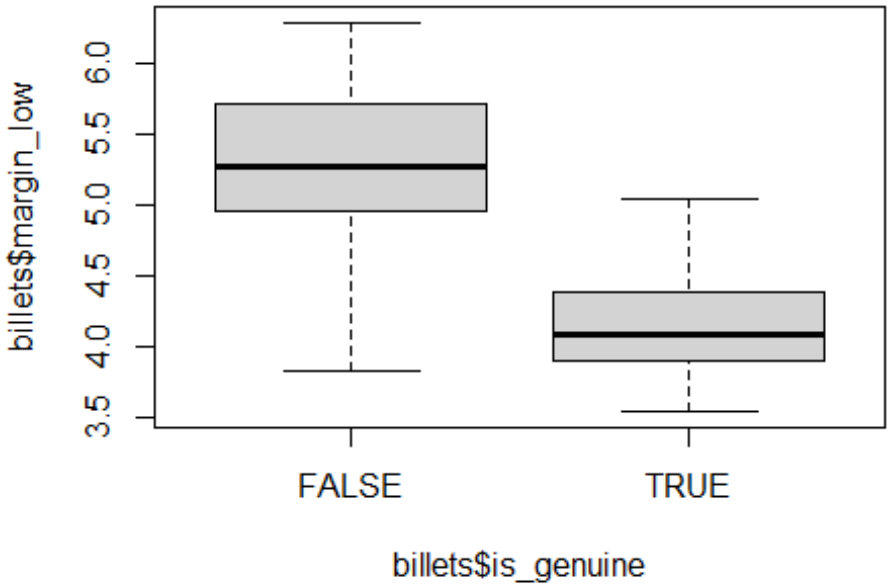
```
## [1] 170   7
```

```
# Analyse univariée
summary(billets)
```

```
##  is_genuine         diagonal      height_left     height_right
##  Mode :logical   Min.   :171.0   Min.   :103.2   Min.   :103.1
##  FALSE:70        1st Qu.:171.7   1st Qu.:103.8   1st Qu.:103.7
##  TRUE :100       Median :171.9   Median :104.1   Median :104.0
##                  Mean   :171.9   Mean   :104.1   Mean   :103.9
##                  3rd Qu.:172.1   3rd Qu.:104.3   3rd Qu.:104.2
##                  Max.   :173.0   Max.   :104.9   Max.   :105.0
##    margin_low      margin_up        length
##  Min.   :3.540   Min.   :2.270   Min.   :110.0
##  1st Qu.:4.050   1st Qu.:3.013   1st Qu.:111.9
##  Median :4.450   Median :3.170   Median :112.8
##  Mean   :4.612   Mean   :3.170   Mean   :112.6
##  3rd Qu.:5.128   3rd Qu.:3.330   3rd Qu.:113.3
##  Max.   :6.280   Max.   :3.680   Max.   :114.0
```

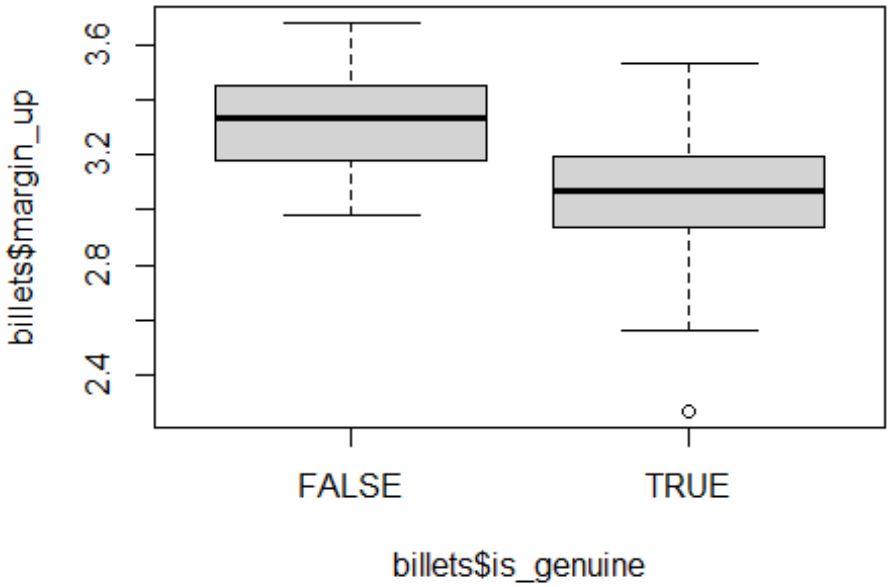
Parmi les 170 billets que comporte notre échantillon, 100 sont authentiques et 70 sont contrefaits.

```
# Analyse bivariée de l'authenticité des billets selon la taille du bord inférieur
boxplot(billets$margin_low ~ billets$is_genuine)
```



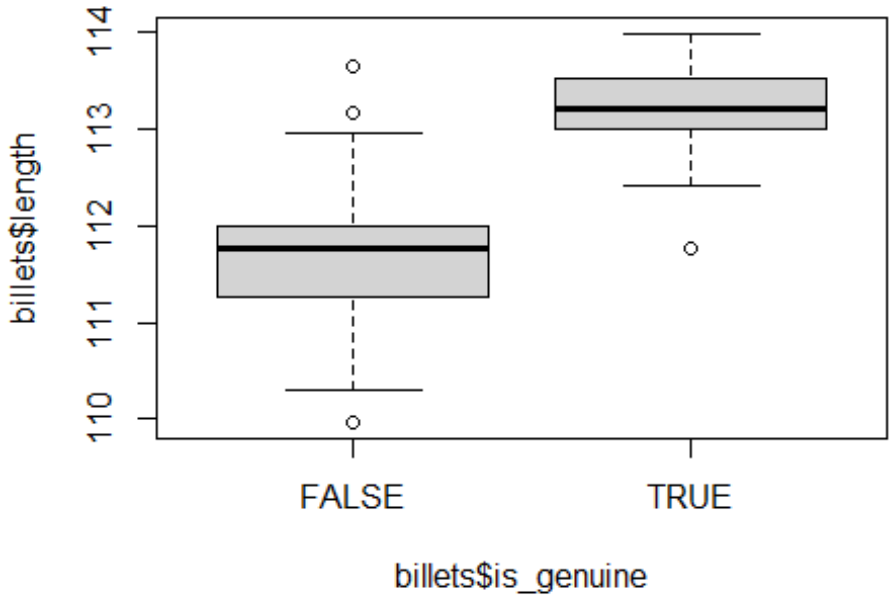
Les faux billets se distinguent nettement des vrais billets sur l'écart moyen du bord inférieur. Néanmoins, on remarque qu'une partie des résultats se chevauchent puisqu'un quart des faux billets ont un bord inférieur dont la taille s'étale sur des dimensions comparables aux bords des vrais billets.

```
# Analyse bivariée de l'authenticité selon l'écart à la marge supérieure
boxplot(billets$margin_up ~ billets$is_genuine)
```

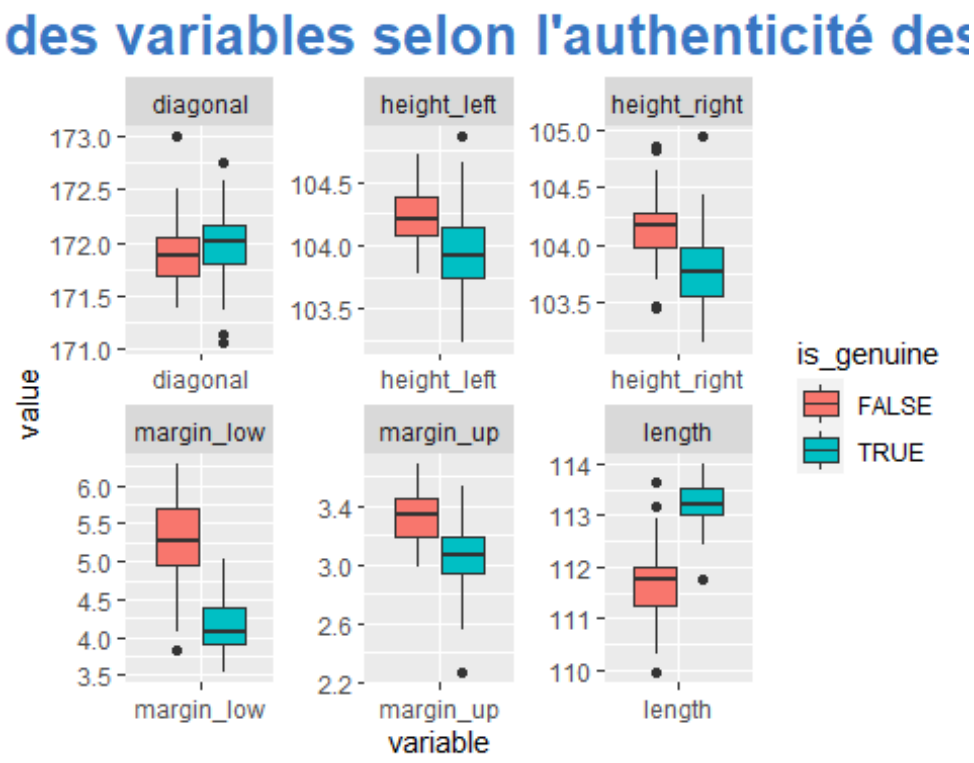


Sur les bords inférieurs également, on constate une taille moyenne bien supérieure du côté des faux billets. Toutefois, les dimensions relevées se chevauchent encore davantage sur cette partie des billets.

```
# Analyse bivariée de l'authenticité selon la longueur des billets
boxplot(billets$length ~ billets$is_genuine)
```




```
# Boxplots pour chaque variable et chaque condition d'authenticité
# Source: https://stackoverflow.com/questions/14604439/plot-multiple-boxplot-in-one-graph
billets_melt <- melt(billets, id.var="is_genuine")
p <- ggplot(data=billets_melt, aes(x=variable, y=value))+
  geom_boxplot(aes(fill=is_genuine))


# Division du graphique en plusieurs panneaux
p + facet_wrap( ~ variable, scales="free") + labs(title = "Amplitude des variables selon l'authenticité des billets") + theme(plot.title = element_text(color = '#3876C2', size=20, face='bold', hjust = 0.5)) +
ggsave("graphiques/graphique01_caracteristiques_billets.jpg", width = 16, height = 9)
```



Corrélations

```
# Conversion numérique des valeurs booléennes de la colonne "is_genuine" sur une copie du dataframe
# True=1 et False=0
billets_num <- billets
billets_num$is_genuine <- as.numeric(billets_num[,1])
head(billets_num)
```

```
##   is_genuine diagonal height_left height_right margin_low margin_up length
## 1          1   171.81      104.86       104.95       4.52      2.89 112.83
## 2          1   171.67      103.74       103.70       4.01      2.87 113.29
## 3          1   171.83      103.76       103.76       4.40      2.88 113.84
## 4          1   171.80      103.78       103.65       3.73      3.12 113.63
## 5          1   172.05      103.70       103.75       5.04      2.27 113.55
## 6          1   172.57      104.65       104.44       4.54      2.99 113.16
```

```
# Matrice de corrélations
billets_cor <- cor(billets_num, method ="pearson")

billets_cor
```

```
##              is_genuine    diagonal height_left height_right margin_low
## is_genuine    1.0000000  0.13922323  -0.4617300   -0.5513089 -0.8001108
## diagonal      0.1392232  1.00000000   0.3195838    0.2204180 -0.1810204
## height_left  -0.4617300  0.31958380   1.0000000    0.7343903  0.4245300
## height_right -0.5513089  0.22041801   0.7343903    1.0000000  0.5093752
```

```
## margin_low   -0.8001108 -0.18102040   0.4245300    0.5093752  1.0000000
## margin_up    -0.5828007 -0.02736555   0.3247876    0.3669179  0.1711128
## length        0.8257426  0.08029519  -0.4213873   -0.4170206 -0.6373517
##                margin_up      length
## is_genuine   -0.58280075  0.82574255
## diagonal     -0.02736555  0.08029519
## height_left   0.32478764 -0.42138735
## height_right  0.36691788 -0.41702056
## margin_low    0.17111283 -0.63735169
## margin_up     1.00000000 -0.52528385
## length       -0.52528385  1.00000000

# Tableau de corrélation en heatmap
png(height=600, width=600, file="graphiques/graphique03_heatmap.png", type =
"cairo")
corrplot(billets_cor, type="upper", tl.col="black", tl.srt=60, tl.pos = "lt")
```

Corrélation avec l'authenticité des billets:

- Corrélation très faible avec la diagonale
- Les hauteurs (gauche et droite) des billets ne sont que partiellement corrélées
- Les bords supérieurs, la différence entre les billets vrais et faux est assez significative
- La longueur des billets et leur marge inférieure semblent bien corrélées avec l'authenticité des billets

Analyse de la variance (ANOVA)

```
diagonal_anova <- aov(diagonal ~ is_genuine, billets)
summary(diagonal_anova)

##              Df Sum Sq Mean Sq F value Pr(>F)
## is_genuine    1  0.306 0.30626   3.321 0.0702 .
## Residuals   168 15.494 0.09223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'anova révèle une p-valeur de 0.07. A un niveau de test de 5%, la variable "diagonal" varie donc peu selon que le billet est authentique ou non.

```
height_left_anova <- aov(height_left ~ is_genuine, billets)
summary(height_left_anova)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## is_genuine    1  3.204   3.204   45.52 2.33e-10 ***
## Residuals   168 11.823   0.070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

height_right_anova <- aov(height_right ~ is_genuine, billets)
summary(height_right_anova)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## is_genuine    1  5.627   5.627   73.36 6.67e-15 ***
## Residuals   168 12.887   0.077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
margin_low_anova <- aov(margin_low ~ is_genuine, billets)
summary(margin_low_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## is_genuine    1  53.33   53.33   298.9 <2e-16 ***
## Residuals   168  29.98    0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
margin_up_anova <- aov(margin_up ~ is_genuine, billets)
summary(margin_up_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## is_genuine    1  3.207   3.207   86.41 <2e-16 ***
## Residuals   168  6.235   0.037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
length_anova <- aov(length ~ is_genuine, billets)
summary(length_anova)
```

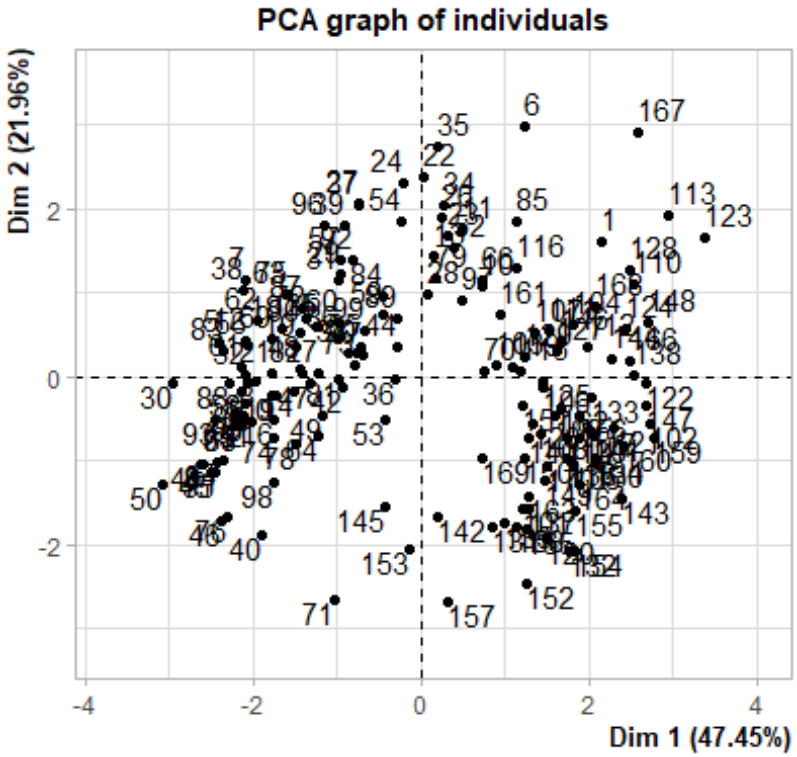
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## is_genuine    1  98.48   98.48   360.1 <2e-16 ***
## Residuals   168  45.95    0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

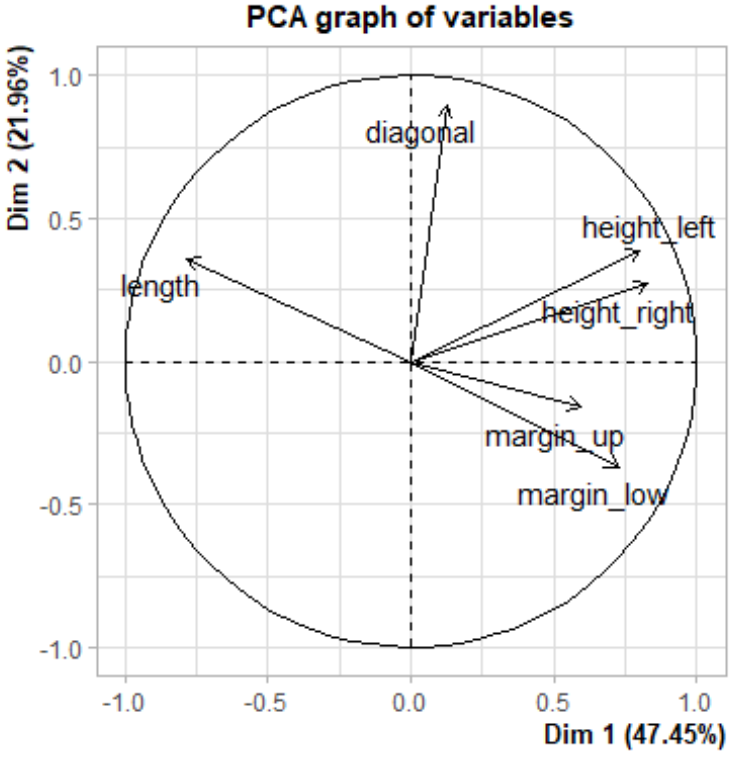
Sur le reste de variables, l'anova montre qu'à un niveau de test de 5% l'authenticité des billets engendre une modification substantielle des valeurs.

Partie 1

ACP

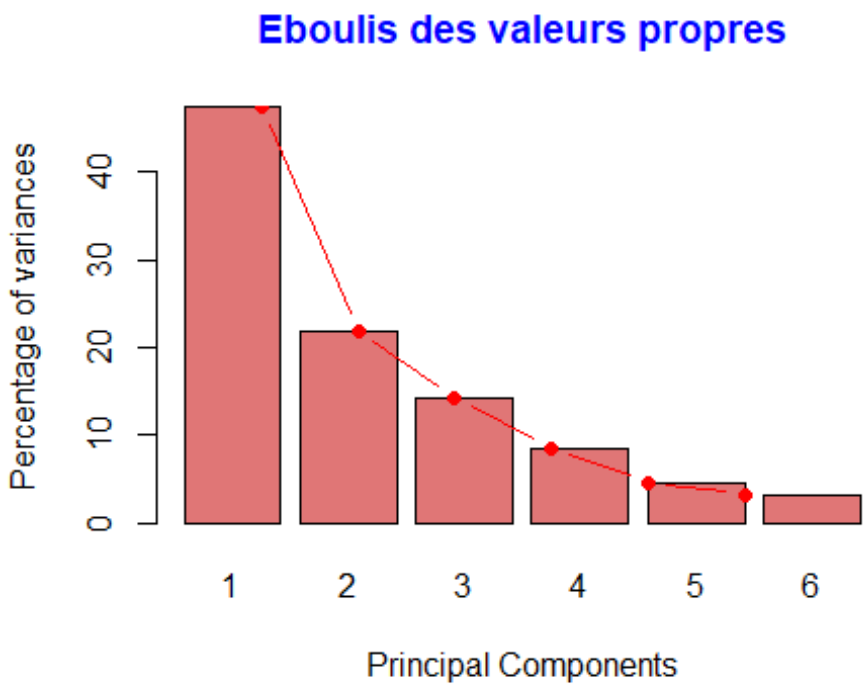
```
res.pca=PCA(billets[,2:7], scale.unit=TRUE, ncp=5, graph=TRUE, axes=c(1,2))
```





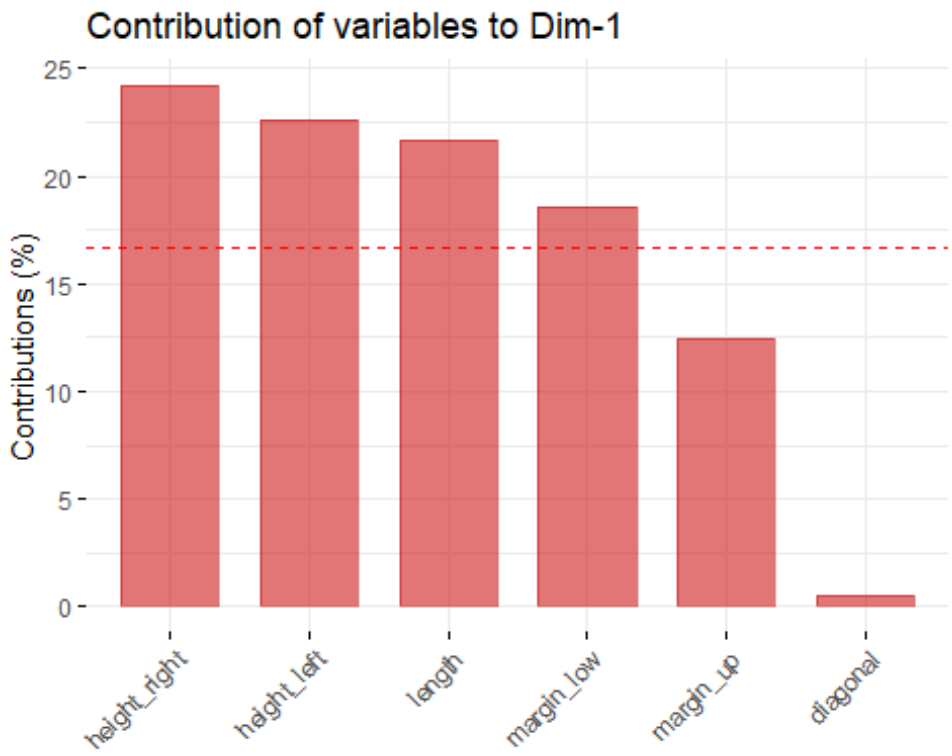

```
res.pca

## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 170 individuals, described by 6 variables
```

```
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"

# Éboulis des valeurs propres.
eigenvalues <- res.pca$eig
barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
        col = rgb(0.8,0.1,0.1,0.6),
        main = "Eboulis des valeurs propres",
        col.main = "blue",
        xlab = "Principal Components",
        ylab = "Percentage of variances")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
      type="b", pch=19, col = "red")
```



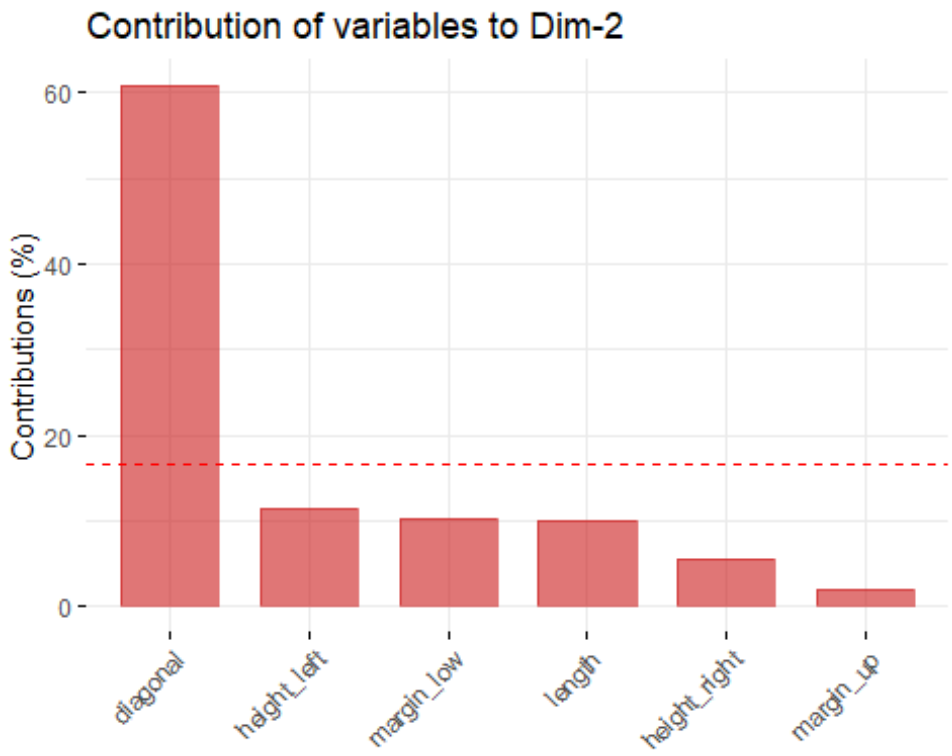
Le critère du coude dans l'éboulis nous invite à retenir les deux premiers axes, qui retiennent en tout 69,4% de l'inertie expliquée. En d'autres termes, près de 70% de la variabilité totale du nuage des individus est représenté par un plan en deux dimensions.

```
# Contribution des variables à la 1e Dimension
fviz_contrib(res.pca, fill = rgb(0.8,0.1,0.1,0.6),
             color = rgb(0.8,0.1,0.1,0.6),
             choice="var", axes = 1 )
```



Les variables de la hauteur, de la longueur et de la marge inférieure des billets contribuent de manière substantielle à la 1e Dimension de l'ACP.

```
# Description de la dimension
res.desc <- dimdesc(res.pca, axes =c(1,2), proba = 0.05)

# Description de la dimension 1
res.desc$Dim.1

## $quanti
##              correlation      p.value
## height_right   0.8298348 2.014843e-44
## height_left    0.8022997 1.725796e-39
## margin_low     0.7272578 2.923475e-29
## margin_up      0.5948294 1.200274e-17
## length        -0.7852090 8.381214e-37
##
## attr(,"class")
## [1] "condes" "list"

# Contribution des variables à la 2e Dimension
fviz_contrib(res.pca, fill = rgb(0.8,0.1,0.1,0.6),
             color = rgb(0.8,0.1,0.1,0.6),
             choice="var", axes = 2 )
```



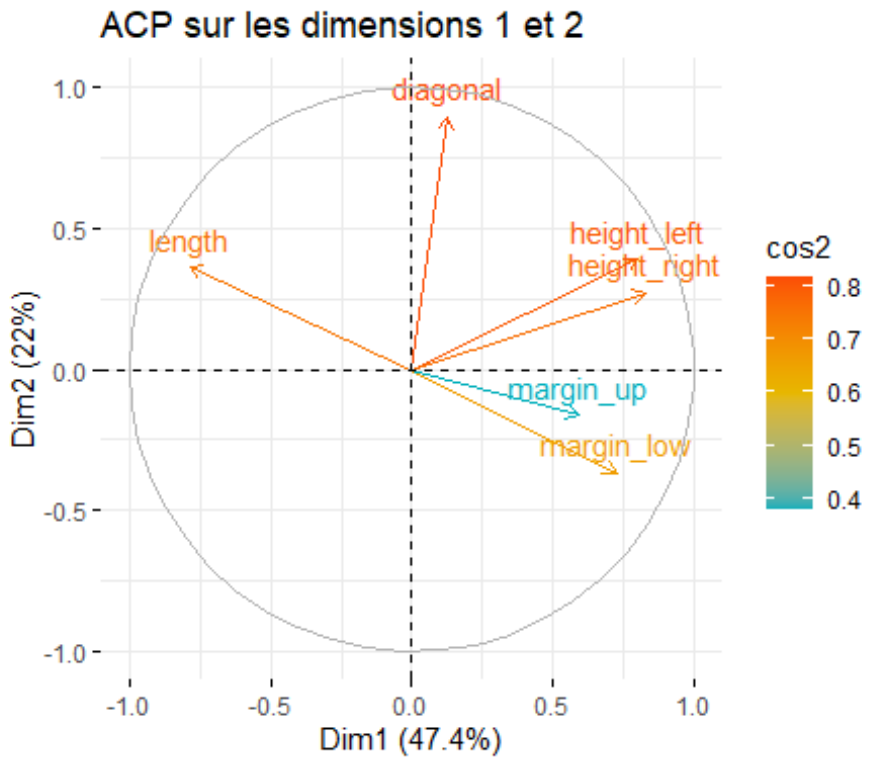
La 2e Dimension est nettement associée à la diagonale des billets, variable pourtant très peu corrélée avec l'authenticité des billets.

```
# Graphique des variables
var <- get_pca_var(res.pca)

# Visualisation du cos2 des variables et qualité de la représentation
png(height=600, width=600,
file="graphiques/graphique05_qualite_rep_variables.png", type = "cairo")
corrplot(var$cos2,
         is.corr=FALSE,
         main="Qualité de représentation des variables selon les dimensions",
         mar=c(3,0,3,0)
         )
```

Le tableau des couleurs montre qu'une majorité des variables est représentée dans la 1e dimension de notre ACP. Deux variables restent néanmoins sous-représentées: la diagonale et le bord supérieur qui sont mieux estimés respectivement dans les dimensions 2 et 3.

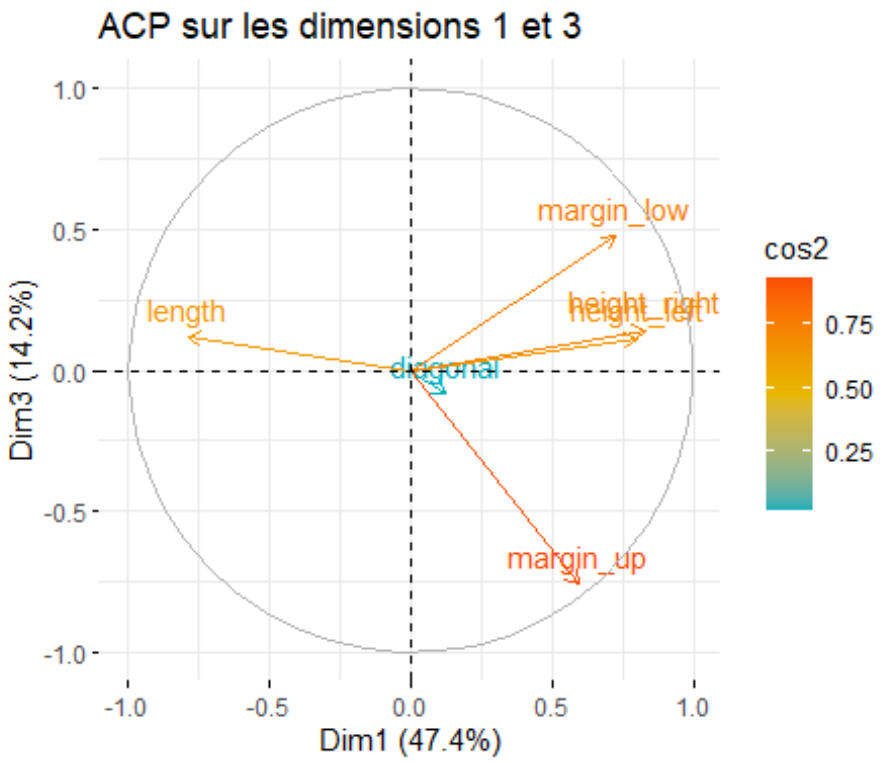
```
# Contribution aux composantes principales
png(height=600, width=600,
file="graphiques/graphique06_contrib_var_compos.png", type = "cairo")
corrplot(var$contrib,
         is.corr=FALSE,
         main="Contribution des variables aux composantes principales",
         mar=c(3,0,3,0)
         )
```

```r
# Carte factorielle - ACP sur les dimensions 1 & 2
fviz_pca_var(res.pca, col.var="cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             title="ACP sur les dimensions 1 et 2"
             )
```



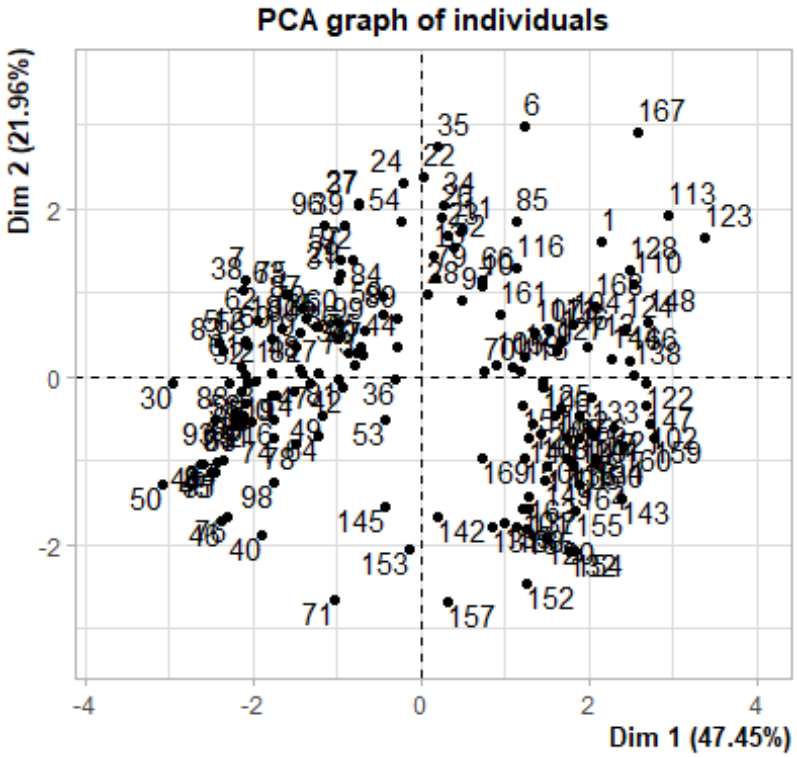
L'ACP sur les dimensions 1 et 2 montre la proximité entre les hauteurs gauches et droites d'une part et les marges inférieures et supérieures d'autre part, qui pourraient ainsi être regroupées pour former deux variables uniques.

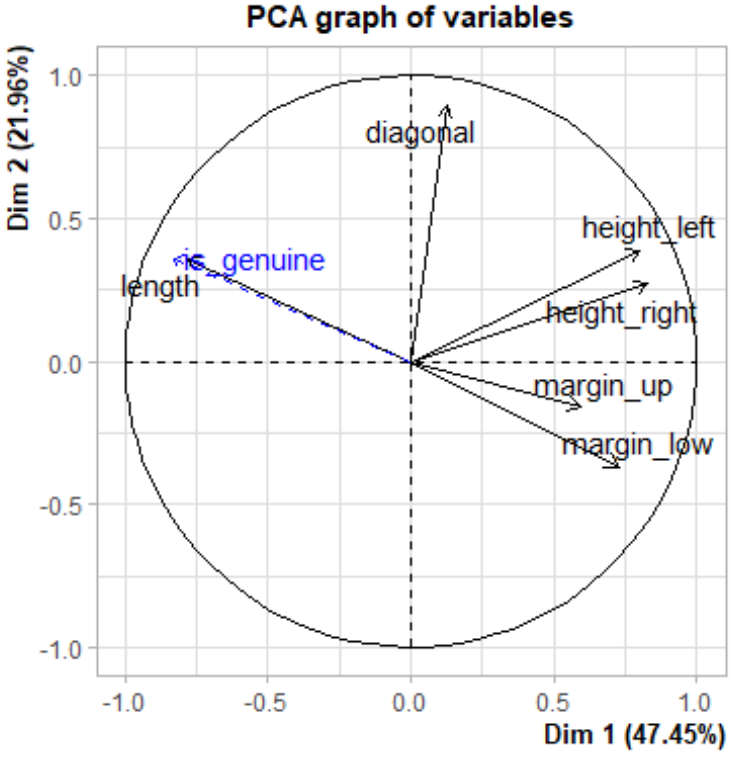
```r
# Carte factorielle - ACP sur les dimensions 1 & 3
fviz_pca_var(res.pca,
             col.var="cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             title="ACP sur les dimensions 1 et 3",
             axes=c(1,3)
             )
```



La variable “diagonale” est mal représentée sur ce plan. Nous remarquons également que les marges inférieures et supérieures n’ont plus la même proximité. Enfin, ce cercle confirme la forte corrélation entre la hauteur droite et la hauteur gauche.

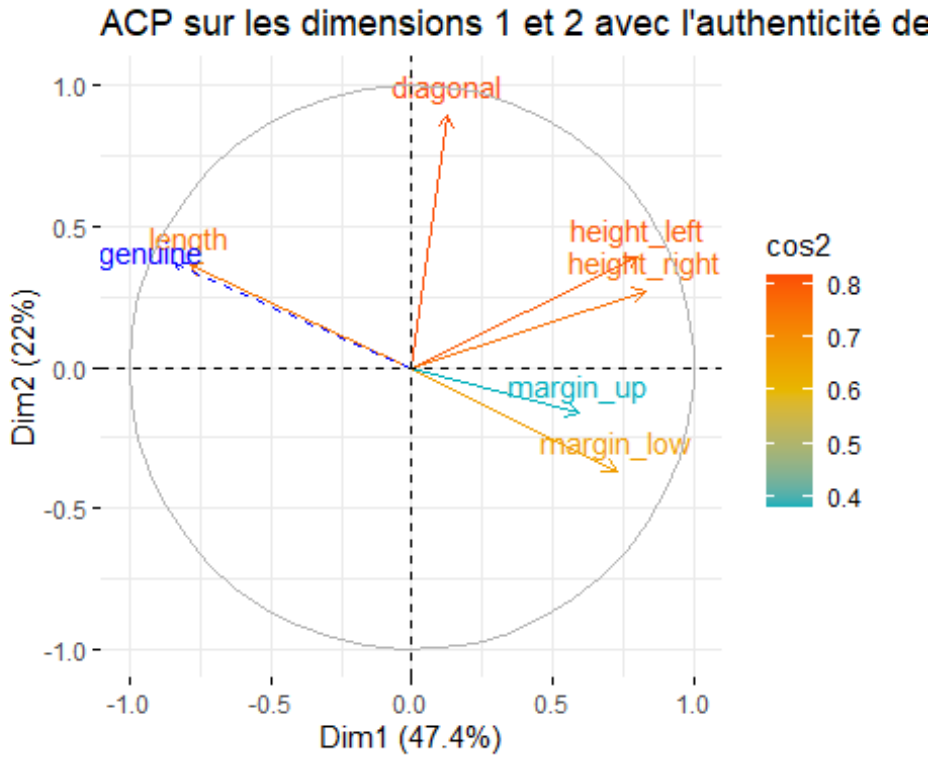
```
# Carte factorielle - ACP sur les dimensions 1 et 2 avec l'authenticité des billets comme variable illustrative
res.pca.quanti = PCA(billets_num, scale.unit = TRUE, ncp=5, quanti.sup = 1, graph=TRUE)
```



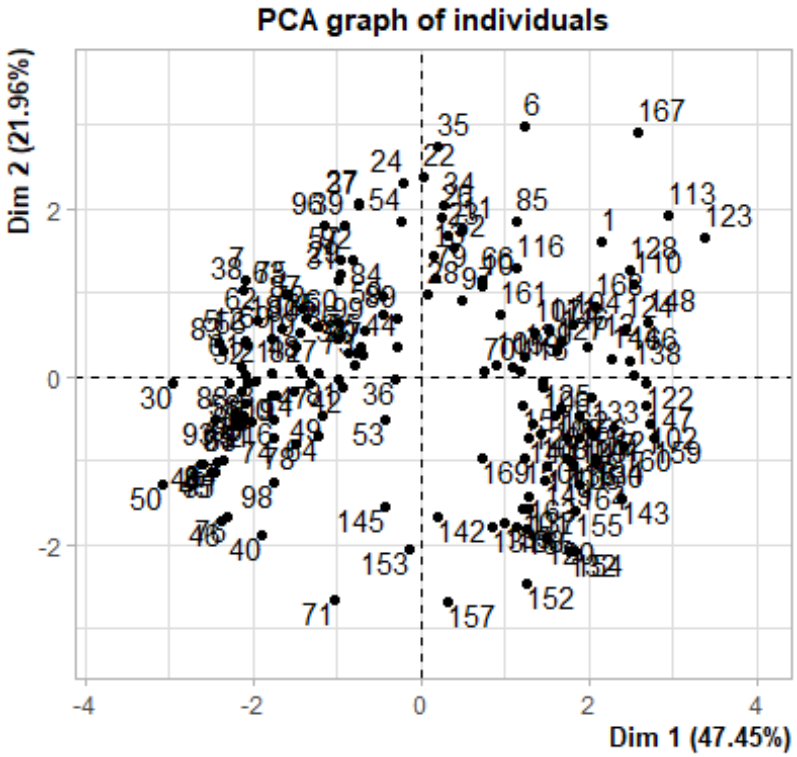


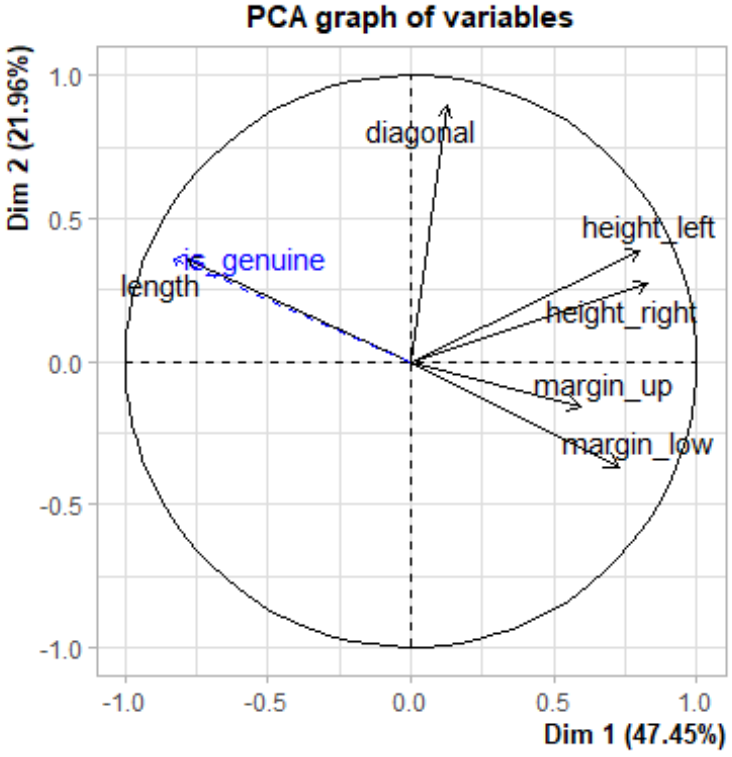

```
fviz_pca_var(res.pca.quanti,
             col.var="cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
```

```
            title="ACP sur les dimensions 1 et 2 avec l'authenticité des
billets comme illustration")
```



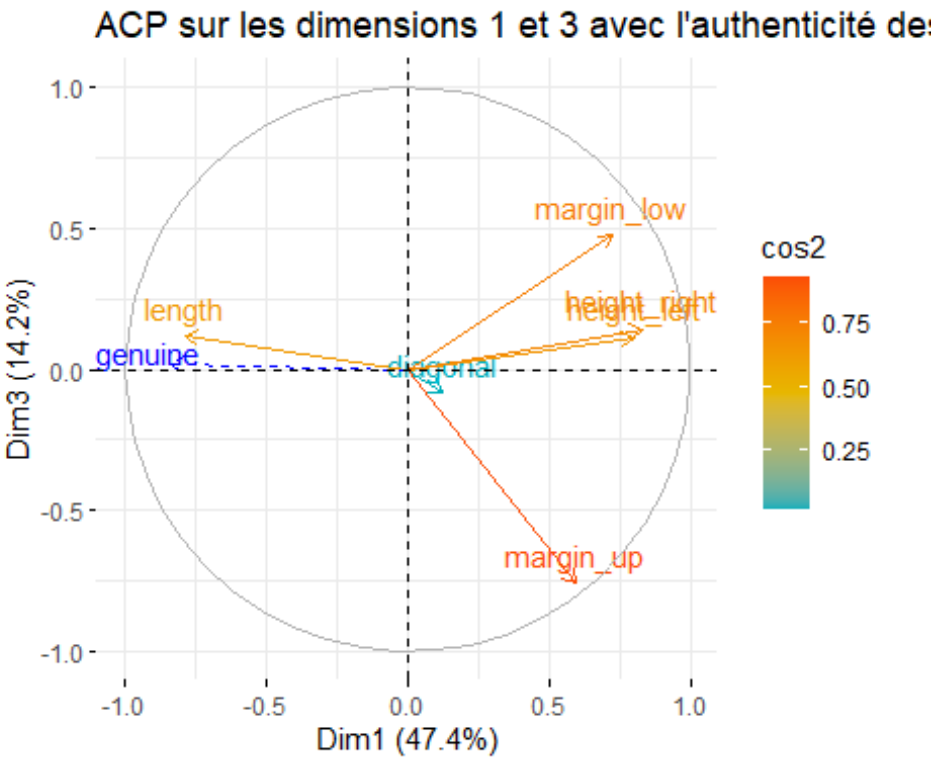

```
# Carte factorielle - ACP sur les dimensions 1 et 3 avec l'authenticité des
billets comme variable illustrative
res.pca.quanti = PCA(billets_num, scale.unit = TRUE, ncp=5, quanti.sup = 1,
graph=TRUE)
```






```
fviz_pca_var(res.pca.quanti,
             col.var="cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             title="ACP sur les dimensions 1 et 3 avec l'authenticité des
```

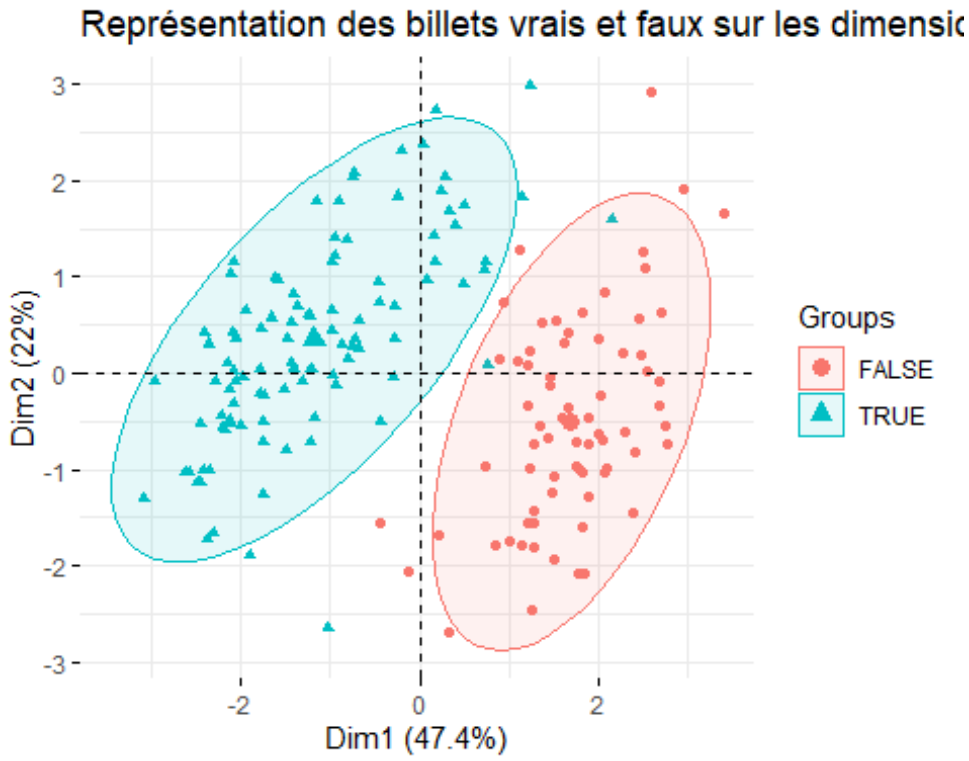
```
billets comme illustration",
             axes=c(1,3)
            )
```



Les deux plans factoriels présentent une corrélation entre la variable d'authenticité et la longueur des billets. Le cercle de corrélation représentant les dimensions 1 et 2 confirme la corrélation négative entre cette authenticité et les variables des bords.

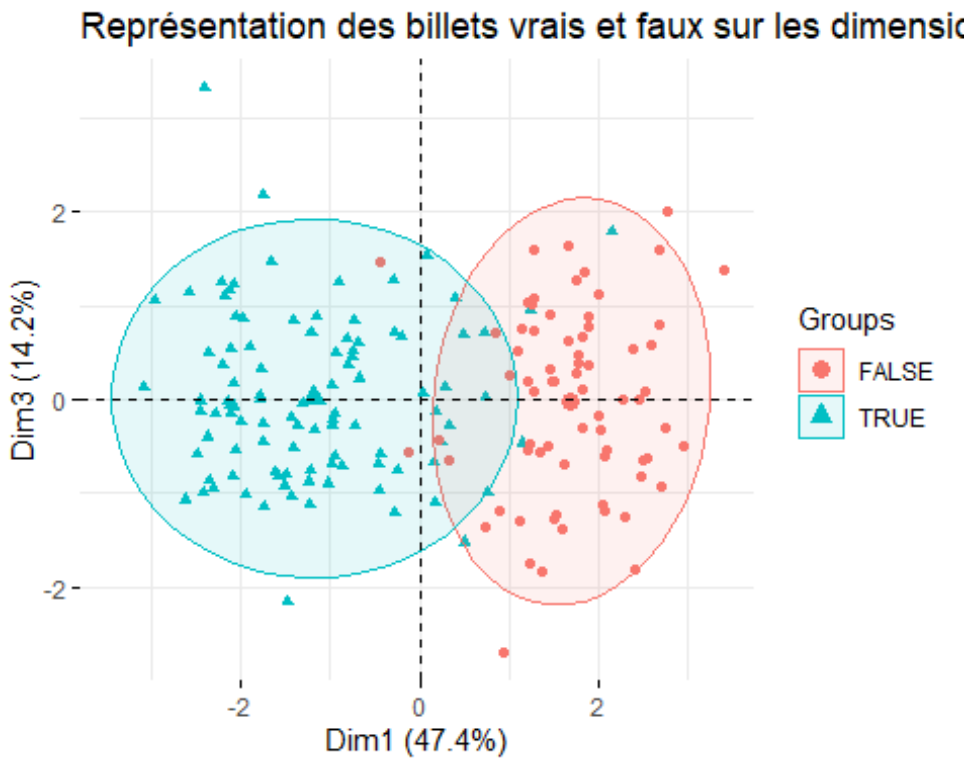
Cette représentation factorielle confirme que les variables "longueur", "bord inférieur" et "bord supérieur" permettent une détection plus précise des faux billets.

```
# Représentation par ellipses des billets sur plan factoriel selon leur
authenticité
fviz_pca_ind(res.pca,
             label="Genuineness", habillage=billets$is_genuine,
             addEllipses=TRUE, ellipse.level=0.90, axes=c(1,2),
             title="Représentation des billets vrais et faux sur les
dimensions 1 et 2")
```



Sur ce plan en deux dimensions, la représentation graphique en ellipses distingue assez nettement deux groupes de billets. Hormis deux billets authentiques positionnés dans l'ellipse qui regroupe les billets dont les caractéristiques suggèrent une contrefaçon, ce plan des dimensions 1 et 2 semble relativement bien permettre la distinction entre les vrais billets et les faux billets.

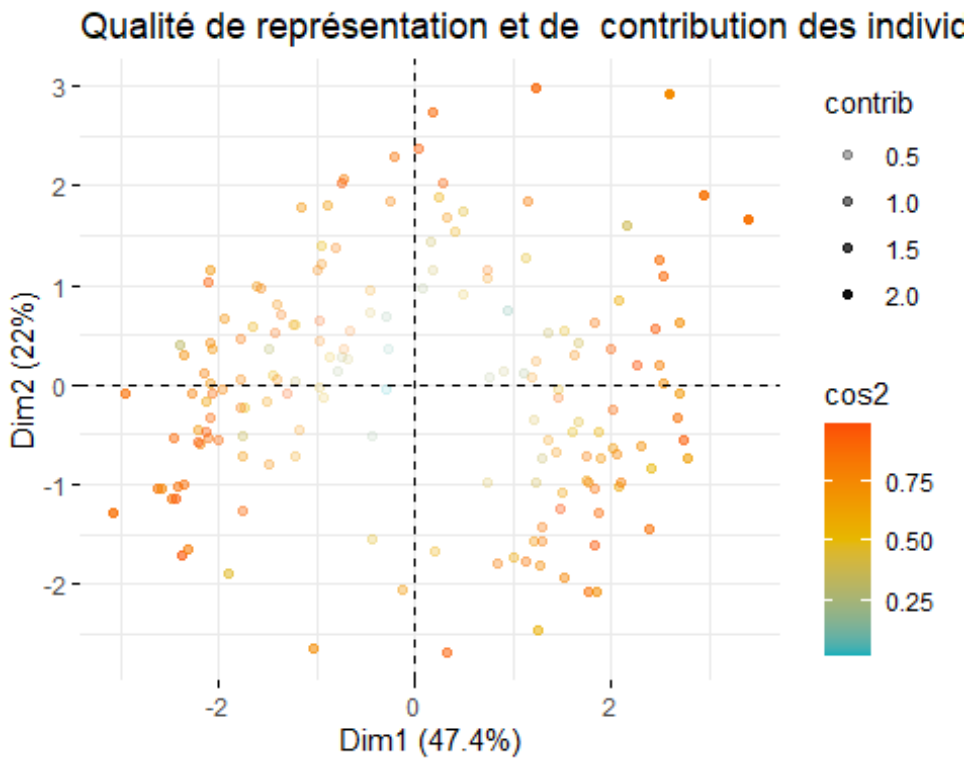
```
# Représentation par ellipses des billets selon leur authenticité sur les
dimensions 1 et 3
fviz_pca_ind(res.pca,
             label="Genuineness", habillage=billets$is_genuine,
             addEllipses=TRUE, ellipse.level=0.90, axes=c(1,3),
             title="Représentation des billets vrais et faux sur les
dimensions 1 et 3")
```



Prenant en considération les dimensions 1 et 3, ce plan factoriel montre un relatif chevauchement des vrais et faux billets. La dimension 3 semble donc en mesure de déterminer l'authenticité des billets avec une précision moindre que la dimension 2.

Nous savons que la variable la mieux représentée dans la dimension 2 est la diagonale des billets. L'ACP nous a montré que cette variable présentait des résultats décorrélés de la longueur et des bords des billets. Nous pourrions donc en déduire que la diagonale des billets constitue un facteur permettant de distinguer les vrais des faux billets. L'analyse bivariée portant sur la diagonale a pourtant montré que le critère d'authenticité n'entraînait pas une forte modification des valeurs de la diagonale des billets. Nous savons toutefois que la diagonale peut varier selon la hauteur et la longueur des billets, et comme l'ACP a montré que la longueur constituait un des principaux critères de détection des billets contrefaits. Nous pouvons conclure que les faux billets compensent les variations de longueur par une modification de la hauteur, ce qui impacte par la même occasion la diagonale.

```
# Qualité de représentation et de contribution des individus
fviz_pca_ind (res.pca,
              axes=c(1,2),
              col.ind = "cos2",
              alpha.ind = "contrib",
              label="Genuineness",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              title="Qualité de représentation et de  contribution des
individus"
              )
```



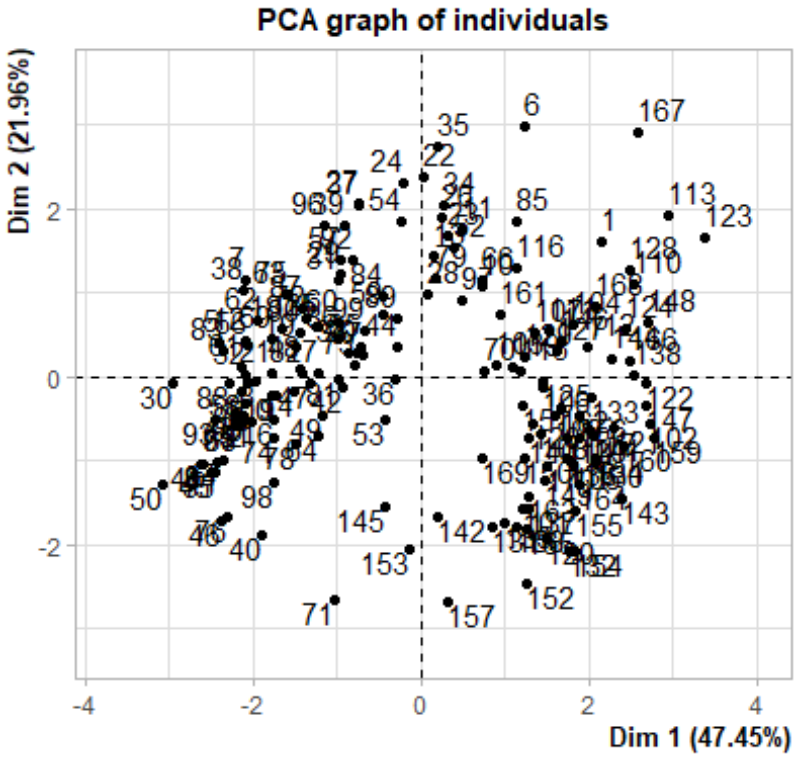
Le graphique montre que les billets situés sur la partie centrale du plan factoriel ont une plus faible qualité de représentation et de contribution que les billets situés plus en marge sur les dimensions 1 et 2.

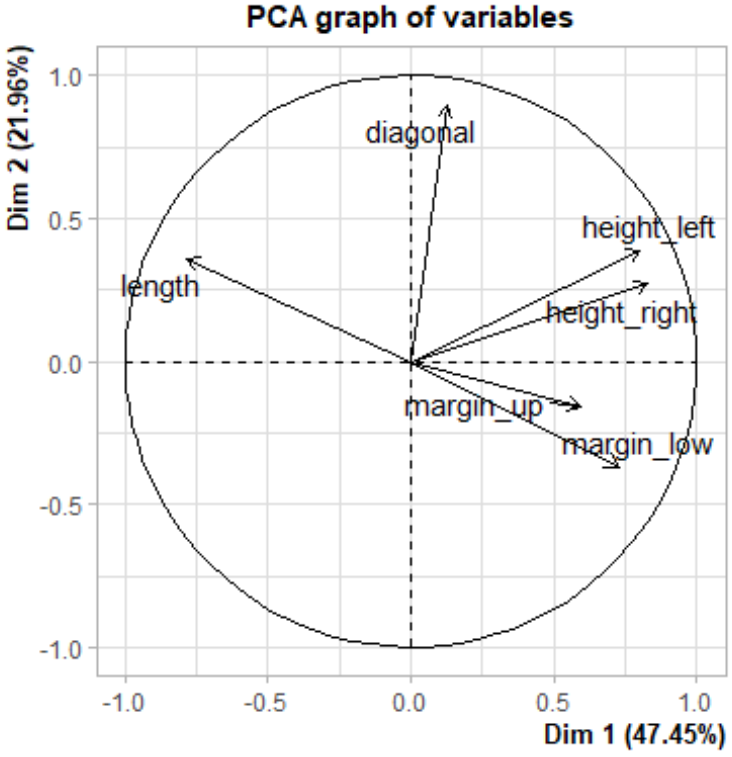
Partie 2

```
# Mise à l'échelle des données (centrage et réduction)
scaled_billets <- as.data.frame(scale(billets[,2:7], center=T, scale=T))
head(scaled_billets)
```

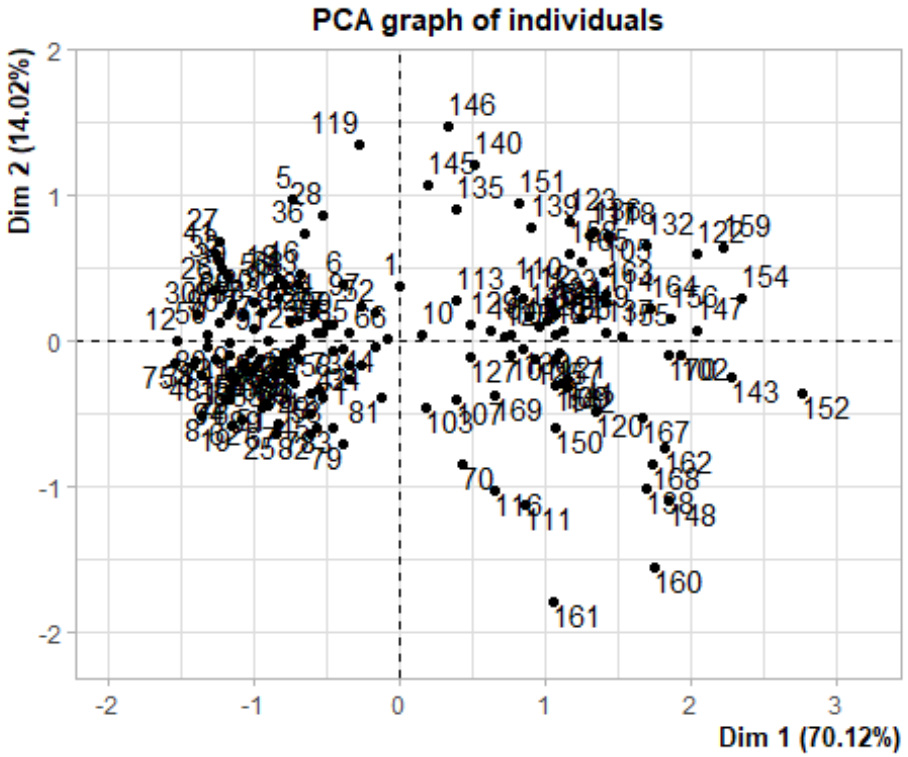
```
##     diagonal height_left height_right margin_low  margin_up    length
## 1 -0.4270822    2.661591    3.0874422 -0.1312025 -1.1863689 0.2808035
## 2 -0.8849450   -1.094464   -0.6892183 -0.8575920 -1.2709851 0.7783978
## 3 -0.3616732   -1.027391   -0.5079386 -0.3021177 -1.2286770 1.3733475
## 4 -0.4597866   -0.960319   -0.8402848 -1.2563941 -0.2132826 1.1461849
## 5  0.3578256   -1.228609   -0.5381519  0.6094299 -3.8094710 1.0596467
## 6  2.0584590    1.957331    1.5465647 -0.1027167 -0.7632879 0.6377733
```

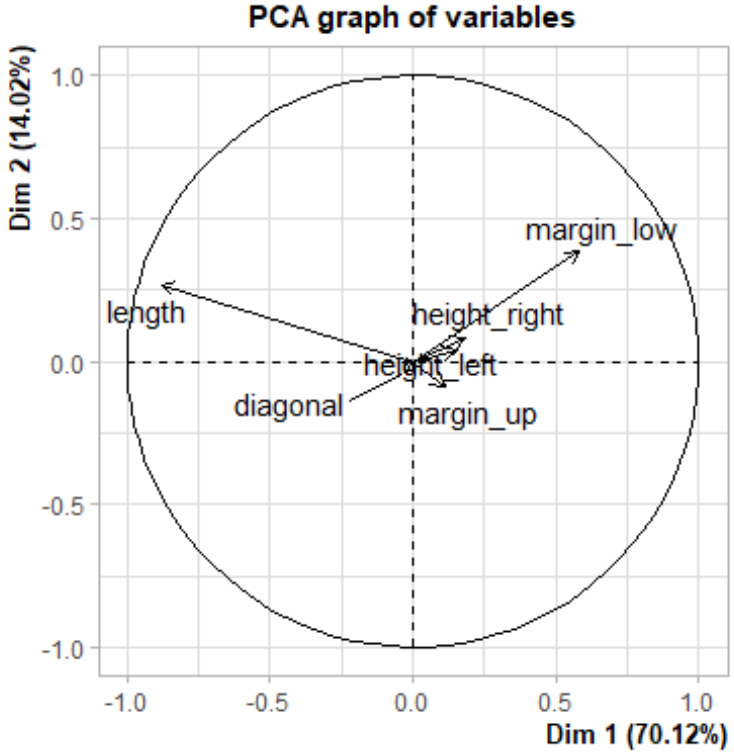
```
#ACP sur les données brutes et les données centrées réduites
res.pca <- PCA(billets[,2:7], scale.unit=TRUE, ncp=5, graph=TRUE)
```






```
nonscaled.res.pca <- PCA(billets[,2:7], scale.unit=FALSE, ncp=5, graph=TRUE)
```

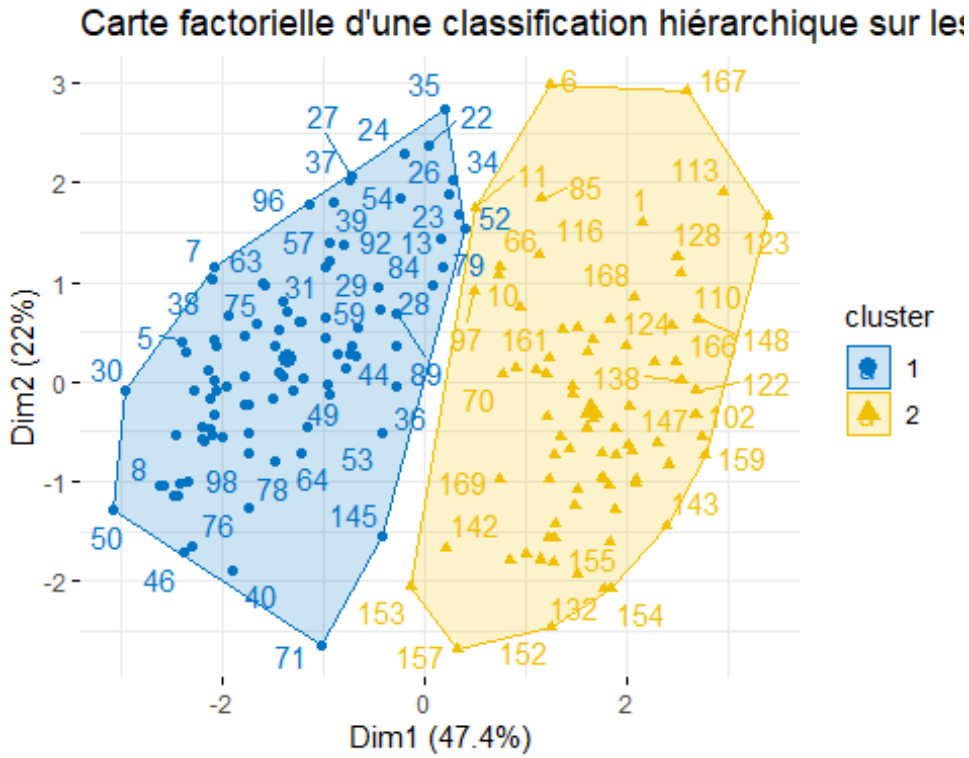





```r
# HCPC sur les données brutes et les données centrées réduites
res.hcpc <- HCPC(res.pca, nb.clust = 2, graph = FALSE)
nonscaled.res.hcpc <- HCPC(nonscaled.res.pca, nb.clust = 2, graph = FALSE)
```

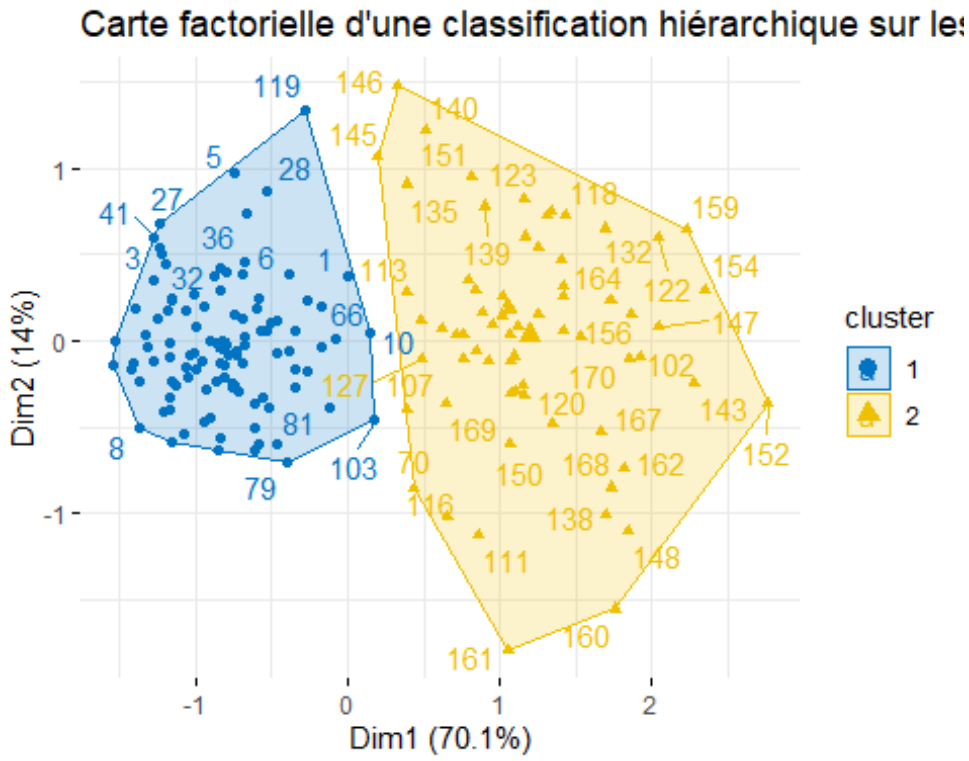
```
# Visualisation des individus par groupes colorés sur les données centrées et réduites
fviz_cluster(res.hcpc,
             repel = TRUE,
             show.clust.cent = TRUE,
             palette = "jco",
             ggtheme = theme_minimal(),
             main = "Carte factorielle d'une classification hiérarchique sur les données centrées et réduites"
             )
```

```
## Warning: ggrepel: 95 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
# Visualisation des individus par groupes colorés sur les données centrées et réduites
fviz_cluster(nonscaled.res.hcpc,
             repel = TRUE,
             show.clust.cent = TRUE,
             palette = "jco",
             ggtheme = theme_minimal(),
             main = "Carte factorielle d'une classification hiérarchique sur les données brutes"
             )
```

```
## Warning: ggrepel: 119 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
# Restrictions et préparation des données pour la jointure
billets_clust <- res.hcpc$data.clust
billets_clust <- billets_clust %>%
  select(clust)

nonscaled.billets_clust <- nonscaled.res.hcpc$data.clust
nonscaled.billets_clust <- nonscaled.billets_clust %>%
  select(clust)
```

K-Means

```
#K-Means sur les valeurs centrées réduites
set.seed(42)
kmeans_billets <- kmeans(scaled_billets, centers=2, nstart = 1)
kmeans_billets

## K-means clustering with 2 clusters of sizes 93, 77
##
## Cluster means:
##      diagonal height_left height_right margin_low  margin_up     length
## 1  0.05293535  -0.5337241   -0.5950049 -0.6735060 -0.5303659  0.7246604
## 2 -0.06393490   0.6446278    0.7186422  0.8134553  0.6405718 -0.8752392
##
## Clustering vector:
##   [1] 2 1 1 1 1 2 1 1 1 2 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
##  [38] 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 2 1 1 1 2 1 1 1
```

```
##  [75] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 2 2 2 2 2 2 2
2 2 2
## [112] 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 1
2 2 2
## [149] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 311.3286 312.2003
##  (between_SS / total_SS =  38.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

La méthode des K-means distingue 2 clusters dont les effectifs respectifs sont de 93 et 77 billets

```
#K-Means sur les valeurs brutes
set.seed(42)
kmeans_nonscaled_billets <- kmeans(billets_num[,2:7], centers=2, nstart = 1)
kmeans_nonscaled_billets
```

```
## K-means clustering with 2 clusters of sizes 69, 101
##
## Cluster means:
##   diagonal height_left height_right margin_low margin_up   length
## 1 171.8907    104.2267     104.1461   5.275362  3.332899 111.6252
## 2 171.9747    103.9568     103.7792   4.159010  3.059406 113.2161
##
## 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
##  [38] 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 1 2
2 2 2
##  [75] 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 1 1 2 1 1 1 1 1
1 1 1
## [112] 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
## [149] 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] 63.8563 55.9585
##  (between_SS / total_SS =  58.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

A ce stade, nous ne savons pas encore si la méthode des K-means repère mieux les billets authentiques et contrefaits en se basant sur les données brutes ou les données centrées-réduites, mais les données brutes donne une répartition (69/101) plus proche de la répartition de notre échantillon d'origine (70/100) que les données centrées-réduites (77/93).

```
# Assignation des valeurs du clustering au dataframe principal
billets_clust$klust <- factor(kmeans_billets$cluster)
billets_clust
```

```
##     clust klust
## 1       2     2
## 2       1     1
## 3       1     1
## 4       1     1
## 5       1     1
## 6       2     2
## 7       1     1
## 8       1     1
## 9       1     1
## 10      2     2
## 11      2     2
## 12      1     1
## 13      1     1
## 14      1     1
## 15      1     1
## 16      1     1
## 17      1     1
## 18      1     1
## 19      1     1
## 20      1     1
## 21      1     1
## 22      1     1
## 23      1     1
## 24      1     1
## 25      1     1
## 26      1     1
## 27      1     1
## 28      1     1
## 29      1     1
## 30      1     1
## 31      1     1
## 32      1     1
## 33      1     1
## 34      1     1
## 35      1     1
## 36      1     1
## 37      1     1
## 38      1     1
## 39      1     1
```

```
## 40      1     1
## 41      1     1
## 42      1     1
## 43      1     1
## 44      1     1
## 45      1     1
## 46      1     1
## 47      1     1
## 48      1     1
## 49      1     1
## 50      1     1
## 51      1     1
## 52      1     1
## 53      1     1
## 54      1     1
## 55      1     1
## 56      1     1
## 57      1     1
## 58      1     1
## 59      1     1
## 60      1     1
## 61      1     1
## 62      1     1
## 63      1     1
## 64      1     1
## 65      1     1
## 66      2     2
## 67      1     1
## 68      1     1
## 69      1     1
## 70      2     2
## 71      1     1
## 72      1     1
## 73      1     1
## 74      1     1
## 75      1     1
## 76      1     1
## 77      1     1
## 78      1     1
## 79      1     1
## 80      1     1
## 81      1     1
## 82      1     1
## 83      1     1
## 84      1     1
## 85      2     2
## 86      1     1
## 87      1     1
## 88      1     1
## 89      1     1
```

```
## 90      1     1
## 91      1     1
## 92      1     1
## 93      1     1
## 94      1     1
## 95      1     1
## 96      1     1
## 97      2     2
## 98      1     1
## 99      1     1
## 100     1     1
## 101     2     2
## 102     2     2
## 103     2     2
## 104     2     2
## 105     2     2
## 106     2     2
## 107     2     2
## 108     2     2
## 109     2     2
## 110     2     2
## 111     2     2
## 112     2     2
## 113     2     2
## 114     2     2
## 115     2     2
## 116     2     2
## 117     2     2
## 118     2     2
## 119     2     2
## 120     2     2
## 121     2     2
## 122     2     2
## 123     2     2
## 124     2     2
## 125     2     2
## 126     2     2
## 127     2     2
## 128     2     2
## 129     2     2
## 130     2     2
## 131     2     2
## 132     2     2
## 133     2     2
## 134     2     2
## 135     2     2
## 136     2     2
## 137     2     2
## 138     2     2
## 139     2     2
```

```
## 140     2     2
## 141     2     2
## 142     2     2
## 143     2     2
## 144     2     2
## 145     1     1
## 146     2     2
## 147     2     2
## 148     2     2
## 149     2     2
## 150     2     2
## 151     2     2
## 152     2     2
## 153     2     2
## 154     2     2
## 155     2     2
## 156     2     2
## 157     2     2
## 158     2     2
## 159     2     2
## 160     2     2
## 161     2     2
## 162     2     2
## 163     2     2
## 164     2     2
## 165     2     2
## 166     2     2
## 167     2     2
## 168     2     2
## 169     2     2
## 170     2     2
```

```
# Assignation des valeurs du clustering des données brutes au dataframe principal
nonscaled.billets_clust$klust <- factor(kmeans_nonscaled_billets$cluster)
nonscaled.billets_clust
```

```
##     clust klust
## 1       1     2
## 2       1     2
## 3       1     2
## 4       1     2
## 5       1     2
## 6       1     2
## 7       1     2
## 8       1     2
## 9       1     2
## 10      1     2
## 11      1     2
## 12      1     2
```

```
## 13      1     2
## 14      1     2
## 15      1     2
## 16      1     2
## 17      1     2
## 18      1     2
## 19      1     2
## 20      1     2
## 21      1     2
## 22      1     2
## 23      1     2
## 24      1     2
## 25      1     2
## 26      1     2
## 27      1     2
## 28      1     2
## 29      1     2
## 30      1     2
## 31      1     2
## 32      1     2
## 33      1     2
## 34      1     2
## 35      1     2
## 36      1     2
## 37      1     2
## 38      1     2
## 39      1     2
## 40      1     2
## 41      1     2
## 42      1     2
## 43      1     2
## 44      1     2
## 45      1     2
## 46      1     2
## 47      1     2
## 48      1     2
## 49      1     2
## 50      1     2
## 51      1     2
## 52      1     2
## 53      1     2
## 54      1     2
## 55      1     2
## 56      1     2
## 57      1     2
## 58      1     2
## 59      1     2
## 60      1     2
## 61      1     2
## 62      1     2
```

```
## 63      1     2
## 64      1     2
## 65      1     2
## 66      1     2
## 67      1     2
## 68      1     2
## 69      1     2
## 70      2     1
## 71      1     2
## 72      1     2
## 73      1     2
## 74      1     2
## 75      1     2
## 76      1     2
## 77      1     2
## 78      1     2
## 79      1     2
## 80      1     2
## 81      1     2
## 82      1     2
## 83      1     2
## 84      1     2
## 85      1     2
## 86      1     2
## 87      1     2
## 88      1     2
## 89      1     2
## 90      1     2
## 91      1     2
## 92      1     2
## 93      1     2
## 94      1     2
## 95      1     2
## 96      1     2
## 97      1     2
## 98      1     2
## 99      1     2
## 100     1     2
## 101     2     1
## 102     2     1
## 103     1     2
## 104     2     1
## 105     2     1
## 106     2     1
## 107     2     1
## 108     2     1
## 109     2     1
## 110     2     1
## 111     2     1
## 112     2     1
```

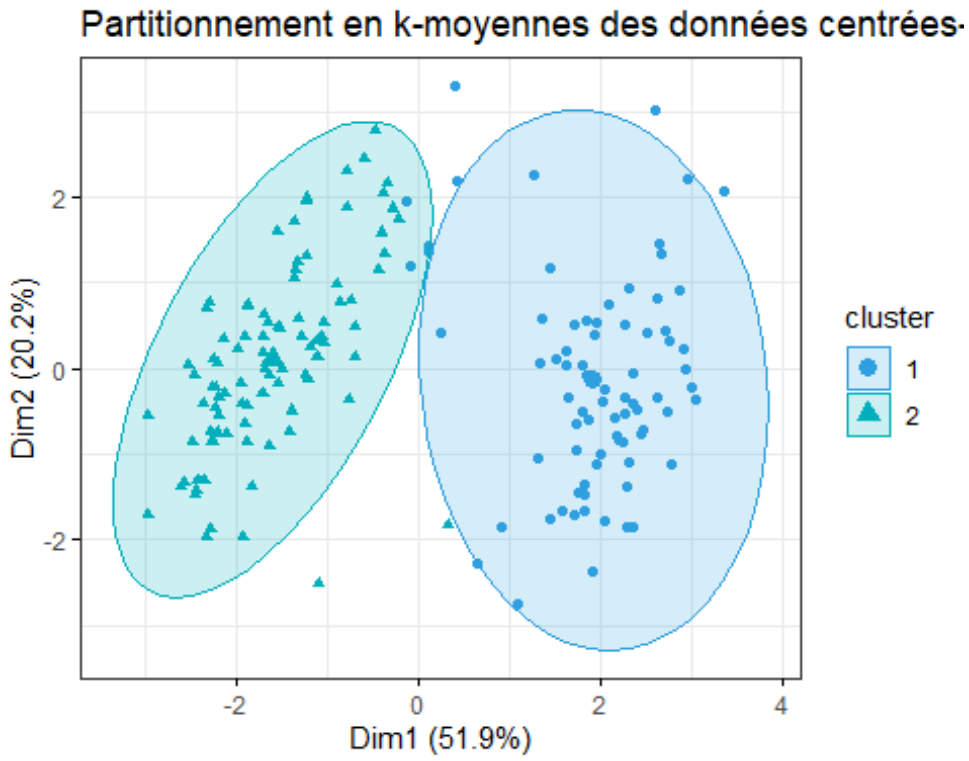
```
## 113     2     1
## 114     2     1
## 115     2     1
## 116     2     1
## 117     2     1
## 118     2     1
## 119     1     2
## 120     2     1
## 121     2     1
## 122     2     1
## 123     2     1
## 124     2     1
## 125     2     1
## 126     2     1
## 127     2     1
## 128     2     1
## 129     2     1
## 130     2     1
## 131     2     1
## 132     2     1
## 133     2     1
## 134     2     1
## 135     2     1
## 136     2     1
## 137     2     1
## 138     2     1
## 139     2     1
## 140     2     1
## 141     2     1
## 142     2     1
## 143     2     1
## 144     2     1
## 145     2     1
## 146     2     1
## 147     2     1
## 148     2     1
## 149     2     1
## 150     2     1
## 151     2     1
## 152     2     1
## 153     2     1
## 154     2     1
## 155     2     1
## 156     2     1
## 157     2     1
## 158     2     1
## 159     2     1
## 160     2     1
## 161     2     1
## 162     2     1
```

```
## 163     2     1
## 164     2     1
## 165     2     1
## 166     2     1
## 167     2     1
## 168     2     1
## 169     2     1
## 170     2     1
```

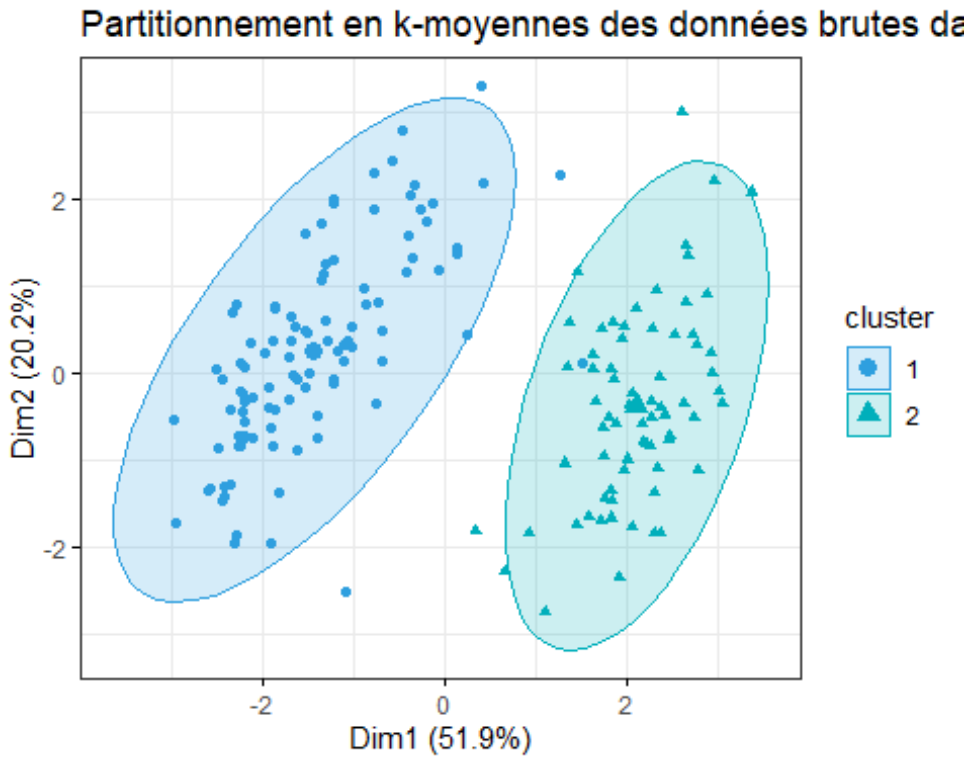
```r
# Jointures avec le dataframe principal
billets_clust <- merge(billets_clust, billets_num, by.x = 0, by.y = 0, all.x
= TRUE, all.y = TRUE)

nonscaled.billets_clust <- merge(nonscaled.billets_clust, billets_num, by.x =
0, by.y = 0, all.x = TRUE, all.y = TRUE)

# Restriction aux colonnes de prédiction (selon les méthodes HCPC et K-means)
et à la variable d'authenticité numérique
billets_clust <- billets_clust %>%
  select(is_genuine, clust, klust)

nonscaled.billets_clust <- nonscaled.billets_clust %>%
  select(is_genuine, clust, klust)
```

Clustering et visualisation selon la méthode des K-means

```r
# Clustering K-means
res.km <- kmeans(scaled_billets, 2, nstart=25)

# Clustering K-means des valeurs brutes
nonscaled.res.km <- kmeans(billets_num, 2, nstart=25)

# Visualisation des clusters dans un plan à 2 dimensions
fviz_cluster(res.km, data = billets_num,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
             geom = "point",
             ellipse.type = "norm",
             habillage=billets$is_genuine,
             ggtheme = theme_bw(),
             main = "Partitionnement en k-moyennes des données centrées-
réduites dans un plan à 2 dimensions"
             )
```




```r
# Visualisation des clusters des données brutes dans un plan à 2 dimensions
fviz_cluster(nonscaled.res.km, data = billets_num,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
             geom = "point",
             ellipse.type = "norm",
             habillage=billets$is_genuine,
             ggtheme = theme_bw(),
             main = "Partitionnement en k-moyennes des données brutes dans un plan à 2 dimensions"
             )
```



Matrices de confusion

Matrices de confusion sur les données centrées et réduites

```
# Matrice de confusion avec les clusters issus d'une classification hiérarchique
MC_hcpc = table(billets_clust$clust, billets_clust$is_genuine)
MC_hcpc
```

```
##
##      0  1
##   1  1 92
##   2 69  8
```

En colonne, le 0 désigne les billets dont on sait qu'ils sont faux, tandis que le 1 liste les billets authentiques. En ligne, le 1 indique que les billets détectés comme vrais, et le 2 désignent ceux qui ont été désignés faux par notre modèle.

Sur 100 billets authentiques, 92 d'entre eux ont été correctement désignés. Le taux de vrais positifs ("rappel" ou "sensibilité") est donc de 92%, avec une précision d'environ 99% (92 sont positifs parmi les 93 que notre modèle a jugé positifs). La F-mesure est de 0.48.

Concernant les faux billets, la spécificité s'élève à 98.5%. 69 billets ont été détectés comme étant faux sur les 70 faux billets que contient notre échantillon. La précision est cependant plus faible (89.6%) car 8 billets jugés faux sont en fait authentiques. La F-mesure est alors de 0.47.

```
# Matrice de confusion dont le clustering est issu de la méthode des K-means
MC_kmeans = table(billets_clust$klust, billets_clust$is_genuine)
MC_kmeans

##
##      0  1
##   1  1 92
##   2 69  8
```

D'après la matrice de confusion, la méthode des K-means apporte exactement les mêmes résultats que le HCPC sur les données centrées et réduites.

Matrices de confusion sur les données brutes

```
# Matrice de confusion avec les clusters issus d'une classification
hiérarchique
MC_hcpc = table(nonscaled.billets_clust$clust,
nonscaled.billets_clust$is_genuine)
MC_hcpc

##
##      0  1
##   1  2 99
##   2 68  1
```

La matrice de confusion montre que la détection des billets est globalement plus efficace en se basant sur les données brutes. En effet, la sensibilité des billets authentiques s'élève à 99% avec une précision d'environ 98%. La F-mesure est de 0.49.

Concernant les faux billets, la spécificité est de 97% (68 des 70 faux billets sont identifiés comme contrefaits) avec une précision de 98.5% (sur les 69 billets jugés faux, 68 sont contrefaits). La F-mesure est ici de 0.49.

Elle récolte la combinaison de spécificités et de sensibilités la plus élevée.

```
# Matrice de confusion dont le clustering est issu de la méthode des K-means
MC_kmeans = table(nonscaled.billets_clust$klust,
nonscaled.billets_clust$is_genuine)
MC_kmeans

##
##      0  1
##   1 68  1
##   2  2 99
```

La méthode des K-means présente les mêmes résultats que la précédente méthode avec les données brutes. La clusterisation a en effet provoqué une inversion des valeurs par rapport aux précédents résultats. L'intitulé n°1 en colonne correspond aux billets jugés faux par notre méthode de détection tandis que le libellé 2 en colonne correspond aux livres désignés comme étant authentiques.

Est-il pour autant plus intéressant de réaliser ces méthodes en se basant sur les données brutes? Si l'objectif de l'exercice est de détecter les billets faux, on remarque que la spécificité sur les billets contrefaits est finalement plus élevée sur les données centrées-réduites (98,5%) que sur les données brutes (97%).

Régression logistique à variables multiples

```
# Régression de la variable d'authenticité en fonction des autres variables
reg_multi <- glm(is_genuine~., data=billets_num, family = binomial(link =
"logit"))

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(reg_multi)

##
## Call:
## glm(formula = is_genuine ~ ., family = binomial(link = "logit"),
##     data = billets_num)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -7.465e-05  -2.100e-08   2.100e-08   2.100e-08   7.553e-05
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4746.47 8556000.82  -0.001    1.000
## diagonal          15.04   69484.85   0.000    1.000
## height_left      -59.09  133557.62   0.000    1.000
## height_right      43.04   72860.71   0.001    1.000
## margin_low      -131.68   69995.63  -0.002    0.998
## margin_up       -217.08   77520.99  -0.003    0.998
## length            45.75   22283.83   0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.3035e+02  on 169  degrees of freedom
## Residual deviance: 2.0247e-08  on 163  degrees of freedom
## AIC: 14
##
## Number of Fisher Scoring iterations: 25
```

L'algorithme n'a pas réalisé la convergence. Nous allons néanmoins utiliser ce modèle pour réaliser notre prédiction.

```
# Régression de la variable d'authenticité en fonction des autres variables
reg_multi <- glm(is_genuine~margin_low+margin_up+length, data=billets_num,
family = binomial(link = "logit"))

## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(reg_multi)
```

```
##
## Call:
## glm(formula = is_genuine ~ margin_low + margin_up + length, family =
binomial(link = "logit"),
##     data = billets_num)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.260e-04  -2.100e-08   2.100e-08   2.100e-08   1.291e-04
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5537.30 2968595.16  -0.002    0.999
## margin_low     -176.96   51301.98  -0.003    0.997
## margin_up      -288.77  108266.26  -0.003    0.998
## length           64.83   26529.00   0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.3035e+02  on 169  degrees of freedom
## Residual deviance: 3.4724e-08  on 166  degrees of freedom
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

Partie 3

Subdivision apprentissage / test

```
# Indexation des billets en apprentissage
set.seed(42)
trainIndex <- createDataPartition(billets$is_genuine, p=0.7,list=F)

print(length(trainIndex))
```

```
## [1] 119
```

```
# Partition du data frame des billets en apprentissage
billetsTrain <- billets[trainIndex,]
dim(billetsTrain)
```

```
## [1] 119   7
```

```
# Partition du data frame des billets en test
billetsTest <- billets[-trainIndex,]
dim(billetsTest)
```

```
## [1] 51  7

# Distribution de l'authenticité dans l'échantillon d'apprentissage
prop.table(table(billetsTrain$is_genuine))

##
##     FALSE      TRUE
## 0.4117647 0.5882353

# Distribution de l'authenticité dans l'échantillon test
prop.table(table(billetsTest$is_genuine))

##
##     FALSE      TRUE
## 0.4117647 0.5882353
```

Nous remarquons que la distribution des billets vrais et faux est la même dans l'échantillon test et l'échantillon d'apprentissage.

Modélisation par validation croisée sur toutes les variables

```
# Paramètre du processus d'apprentissage
fitControl <- trainControl(method="cv",number=10)

# Conversion de la variable d'authenticité en valeur factorielle.

billetsTrain$is_genuine <- as.factor(billetsTrain$is_genuine)
billetsTest$is_genuine <- as.factor(billetsTest$is_genuine)

# Apprentissage par régression logistique sur toutes les variables
fitControl <- trainControl(method="cv",number=10)

m_lr <- train(is_genuine ~ .,
              data = billetsTrain,
              method="glm",
              control = list(maxit = 50),
              trControl=fitControl)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
print(m_lr)
```

```
## Generalized Linear Model
##
## 119 samples
##   6 predictor
##   2 classes: 'FALSE', 'TRUE'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 107, 107, 107, 107, 108, 107, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9825758  0.9637089
```

```
# Performance sur l'échantillon test
confusionMatrix(data = predict(m_lr,newdata = billetsTest),reference =
billetsTest$is_genuine, positive="FALSE")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE    21    1
##      TRUE      0   29
##
##                Accuracy : 0.9804
##                  95% CI : (0.8955, 0.9995)
##     No Information Rate : 0.5882
##     P-Value [Acc > NIR] : 6.483e-11
##
##                   Kappa : 0.9598
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 1.0000
##             Specificity : 0.9667
##          Pos Pred Value : 0.9545
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4118
##          Detection Rate : 0.4118
##    Detection Prevalence : 0.4314
```

```
##       Balanced Accuracy : 0.9833
##
##        'Positive' Class : FALSE
##
```

D'après la matrice de confusion, le taux de succès sur l'ensemble des variables est de 98%.

```
summary(m_lr)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -8.165e-06  -2.110e-08   2.110e-08   2.110e-08   8.501e-06
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.071e+01  5.790e+07       0        1
## diagonal      2.339e+00  1.925e+05       0        1
## height_left   5.206e+01  3.514e+05       0        1
## height_right -8.056e+01  4.491e+05       0        1
## margin_low   -6.456e+01  2.003e+05       0        1
## margin_up    -1.443e+02  5.161e+05       0        1
## length        2.932e+01  2.398e+05       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.6124e+02  on 118  degrees of freedom
## Residual deviance: 2.6852e-10  on 112  degrees of freedom
## AIC: 14
##
## Number of Fisher Scoring iterations: 28
```

```
# Importance des variables
varImp(m_lr)
```

```
## glm variable importance
##
##              Overall
## margin_low    100.00
## margin_up      86.24
## height_right   53.92
## height_left    43.84
## length         35.51
## diagonal        0.00
```

La variable "diagonal" semble avoir très peu d'influence sur la variation des échantillons dans la prédiction.

```r
# Affichage du modèle obtenu
print(m_lr$finalModel)
```

```
##
## Call:  NULL
##
## Coefficients:
##  (Intercept)      diagonal   height_left  height_right    margin_low
##       20.707         2.339        52.064       -80.564       -64.560
##    margin_up        length
##     -144.338        29.322
##
## Degrees of Freedom: 118 Total (i.e. Null);  112 Residual
## Null Deviance:       161.2
## Residual Deviance: 2.685e-10     AIC: 14
```

L'AIC, privilégié pour comparer la pertinence de différents modèles, s'élève ici à 14.

```r
#Méthode intégrée de sélection
m_lrs <- train(is_genuine ~ ., data = billetsTrain, method="glm", control=
list(maxit = 50), trControl = fitControl)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Performance sur l'échantillon test
confusionMatrix(data = predict(m_lrs,newdata = billetsTest),reference =
billetsTest$is_genuine, positive="FALSE")
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction FALSE TRUE
##      FALSE    21    1
##      TRUE      0   29
##
##                Accuracy : 0.9804
##                  95% CI : (0.8955, 0.9995)
##     No Information Rate : 0.5882
##     P-Value [Acc > NIR] : 6.483e-11
##
##                   Kappa : 0.9598
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 1.0000
##             Specificity : 0.9667
##          Pos Pred Value : 0.9545
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4118
##          Detection Rate : 0.4118
##    Detection Prevalence : 0.4314
##       Balanced Accuracy : 0.9833
##
##        'Positive' Class : FALSE
##
```

L'accuracy s'élève ici à 98%.

Modélisation par validation croisée sur les 3 variables les plus significatives

Le modèle de régression logistique utilisé dans la partie précédente nous a permis de constater que 3 des 6 variables contribuaient plus largement à la prédiction des faux billets. Nous allons donc ici tester un modèle se basant uniquement sur ces 3 variables.

```
# évaluation par rééchantillonnage
fitControl <- trainControl(method="cv",number=10)
m_lr <- train(is_genuine ~ margin_low+margin_up+length, data =
billetsTrain,method="glm",control= list(maxit = 50),trControl=fitControl)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
print(m_lr)
```

```
## Generalized Linear Model
##
## 119 samples
##   3 predictor
##   2 classes: 'FALSE', 'TRUE'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 107, 107, 107, 108, 107, 107, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9916667  0.9823529
```

D'après la matrice de confusion, le taux de succès sur les variables "margin_low', 'margin_up' et 'length' est de 99%.

```
# Importance des variables
varImp(m_lr)
```

```
## glm variable importance
##
##            Overall
## margin_low 100.000
## length       4.557
## margin_up    0.000
```

La variable "margin low" semble avoir une influence particulièrement significative sur la variation des échantillons dans la prédiction.

```
# Affichage du modèle obtenu
print(m_lr$finalModel)
```

```
##
## Call:  NULL
##
## Coefficients:
## (Intercept)   margin_low    margin_up       length
```

```
##    -5512.43       -81.17      -142.86        56.42
##
## Degrees of Freedom: 118 Total (i.e. Null);  115 Residual
## Null Deviance:       161.2
## Residual Deviance: 3.823e-10     AIC: 8
```

L'AIC s'élève ici à 8, ce qui confirmerait que le modèle ici choisi est plus adapté à la prédiction avec nos données.

```
# Performance sur l'échantillon test
print(confusionMatrix(data = predict(m_lr,newdata = billetsTest),reference =
billetsTest$is_genuine, positive="FALSE"))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE    21    1
##      TRUE      0   29
##
##                Accuracy : 0.9804
##                  95% CI : (0.8955, 0.9995)
##     No Information Rate : 0.5882
##     P-Value [Acc > NIR] : 6.483e-11
##
##                   Kappa : 0.9598
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 1.0000
##             Specificity : 0.9667
##          Pos Pred Value : 0.9545
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4118
##          Detection Rate : 0.4118
##    Detection Prevalence : 0.4314
##       Balanced Accuracy : 0.9833
##
##        'Positive' Class : FALSE
##
```

Malrgé la diminution du nombre de variables, le taux de succès est de 98%, tout comme la régression réalisée sur l'ensemble des variables. La simplification du modèle n'a donc pas entraîné une chute de l'accuracy.

Prédiction des 5 billets à tester

```
# Lecture des tableaux
billets <- read.table("data/notes.csv",header = TRUE, sep = ",")
billetsTest5 <- read.table("data/5_notes.csv",header=TRUE, sep=",")
```

```r
row.names(billetsTest5) <- billetsTest5$id
billetsTest5$id <- NULL
head(billetsTest5)
```

```
##     diagonal height_left height_right margin_low margin_up length
## A_1   171.76      104.01       103.54       5.21      3.30 111.42
## A_2   171.87      104.17       104.13       6.00      3.31 112.09
## A_3   172.00      104.58       104.29       4.99      3.39 111.57
## A_4   172.49      104.55       104.34       4.44      3.03 113.20
## A_5   171.65      103.63       103.56       3.77      3.16 113.33
```

```r
# Prédiction par régression logistique
pred <- predict(m_lr,newdata=billetsTest5)
table(pred)
```

```
## pred
## FALSE  TRUE
##     3     2
```

```r
# Jointure des lignes
billets_new <- bind_rows(billets, billetsTest5)

billets_new
```

```
##        is_genuine diagonal height_left height_right margin_low margin_up
length
## ...1         True   171.81      104.86       104.95       4.52      2.89
112.83
## ...2         True   171.67      103.74       103.70       4.01      2.87
113.29
## ...3         True   171.83      103.76       103.76       4.40      2.88
113.84
## ...4         True   171.80      103.78       103.65       3.73      3.12
113.63
## ...5         True   172.05      103.70       103.75       5.04      2.27
113.55
## ...6         True   172.57      104.65       104.44       4.54      2.99
113.16
## ...7         True   172.38      103.55       103.80       3.97      2.90
113.30
## ...8         True   171.58      103.65       103.37       3.54      3.19
113.38
## ...9         True   171.96      103.51       103.75       4.06      3.33
113.53
## ...10        True   172.14      104.34       104.20       4.63      3.02
112.47
## ...11        True   172.27      104.29       104.22       3.89      3.53
113.50
## ...12        True   172.07      103.64       103.67       3.86      3.20
113.83
## ...13        True   172.19      104.61       103.69       4.00      3.26
```

```
112.91
## ...14        True   171.82      103.78       103.76       3.81      3.25
113.36
## ...15        True   172.04      103.94       103.76       3.81      3.24
113.41
## ...16        True   171.60      103.85       103.91       4.56      2.56
113.27
## ...17        True   171.69      103.90       104.13       4.07      2.92
113.52
## ...18        True   172.05      103.90       103.76       4.52      2.71
113.42
## ...19        True   172.15      103.65       103.66       3.60      3.50
113.24
## ...20        True   171.75      104.16       104.00       4.19      3.03
113.55
## ...21        True   172.03      103.87       103.40       4.29      3.01
113.09
## ...22        True   172.49      104.44       103.98       4.08      3.07
113.16
## ...23        True   172.24      104.51       103.97       4.18      3.22
113.21
## ...24        True   172.59      104.22       104.01       4.47      2.95
113.19
## ...25        True   172.13      103.76       103.85       3.65      3.24
112.92
## ...26        True   172.21      104.28       104.37       4.06      3.30
113.92
## ...27        True   172.41      104.14       104.06       4.45      2.94
113.98
## ...28        True   172.02      104.23       104.26       4.92      2.89
113.49
## ...29        True   172.14      104.01       104.00       3.64      3.16
113.37
## ...30        True   171.84      103.75       103.38       4.08      2.70
113.72
## ...31        True   172.19      104.05       103.81       3.90      3.22
113.52
## ...32        True   171.82      103.77       103.79       4.36      2.77
113.79
## ...33        True   172.01      104.03       103.67       3.90      3.18
112.61
## ...34        True   172.49      104.33       104.03       4.28      3.07
112.71
## ...35        True   172.75      104.33       103.97       4.34      3.14
113.12
## ...36        True   171.66      104.17       104.16       4.75      2.94
113.52
## ...37        True   172.40      104.19       103.98       4.08      2.93
113.44
## ...38        True   172.20      103.93       103.49       3.80      2.99
```

```
113.63
## ...39        True   172.21      104.27       104.01       4.23      2.79
113.78
## ...40        True   171.13      104.28       103.14       4.16      2.92
113.00
## ...41        True   171.51      103.85       103.36       4.49      2.80
113.87
## ...42        True   171.81      104.10       103.69       4.29      2.95
112.72
## ...43        True   171.88      103.66       103.52       4.66      2.75
113.25
## ...44        True   171.91      104.34       103.77       4.45      2.95
112.66
## ...45        True   171.79      103.51       103.25       4.05      3.08
112.71
## ...46        True   171.44      103.52       103.49       4.09      3.12
113.23
## ...47        True   171.85      103.90       103.74       4.13      3.07
113.15
## ...48        True   171.81      103.91       103.78       3.66      3.28
113.59
## ...49        True   171.73      103.82       103.85       3.97      3.12
112.85
## ...50        True   171.59      103.23       103.64       4.01      2.94
113.59
## ...51        True   171.71      103.83       103.51       3.80      3.02
113.01
## ...52        True   172.22      104.48       104.06       4.59      2.91
112.82
## ...53        True   171.59      104.06       103.99       3.93      3.24
112.80
## ...54        True   172.32      104.16       104.14       3.78      3.25
113.38
## ...55        True   171.62      103.49       103.58       3.95      3.00
113.10
## ...56        True   172.14      103.74       103.52       4.56      2.83
113.43
## ...57        True   172.53      103.99       103.55       4.50      3.10
113.03
## ...58        True   171.97      103.69       103.54       4.39      2.70
112.79
## ...59        True   172.09      104.06       103.90       3.97      3.32
113.09
## ...60        True   172.07      103.97       103.84       4.05      3.12
113.18
## ...61        True   172.11      103.67       103.43       4.19      2.98
113.09
## ...62        True   172.22      103.75       103.49       3.69      3.17
113.14
## ...63        True   172.33      103.83       103.54       3.98      3.18
```

```
113.31
## ...64        True   171.65      103.95       103.61       4.03      3.25
113.06
## ...65        True   171.99      103.97       103.89       4.22      3.17
113.12
## ...66        True   172.16      104.43       104.06       4.51      3.19
112.69
## ...67        True   171.73      103.60       103.34       3.82      3.15
112.89
## ...68        True   171.79      103.74       103.48       4.60      2.80
113.35
## ...69        True   172.05      103.72       103.81       4.21      2.97
113.61
## ...70        True   171.94      104.11       104.16       4.08      3.35
111.76
## ...71        True   171.04      103.84       103.64       4.22      3.36
112.70
## ...72        True   172.17      103.93       103.62       4.06      3.08
113.10
## ...73        True   171.97      103.69       104.17       4.32      3.00
112.82
## ...74        True   171.52      103.92       103.66       3.81      3.15
113.54
## ...75        True   172.10      103.94       103.75       3.66      3.20
113.78
## ...76        True   171.35      103.70       103.43       3.71      3.22
113.28
## ...77        True   171.92      103.93       104.06       4.38      2.97
113.01
## ...78        True   171.69      103.85       103.53       3.86      3.19
112.68
## ...79        True   172.16      104.39       103.85       3.77      3.32
112.55
## ...80        True   171.70      103.88       103.56       3.89      3.03
113.60
## ...81        True   172.07      103.74       103.76       4.30      3.09
112.41
## ...82        True   171.95      103.84       103.68       3.79      3.09
112.68
## ...83        True   172.17      103.75       103.29       4.43      2.88
113.38
## ...84        True   172.14      104.06       103.96       3.97      3.24
113.07
## ...85        True   172.30      104.58       104.17       4.36      3.33
112.98
## ...86        True   172.10      103.95       103.72       4.49      3.07
113.15
## ...87        True   172.16      103.92       103.76       4.35      2.84
113.21
## ...88        True   172.02      103.73       103.31       4.35      3.07
```

```
113.62
## ...89        True   171.91      104.28       103.92       3.64      3.36
113.15
## ...90        True   171.99      103.91       103.79       4.05      3.11
113.67
## ...91        True   171.77      103.73       103.48       4.21      2.92
113.24
## ...92        True   172.30      104.04       103.93       4.33      2.92
113.19
## ...93        True   171.86      103.47       103.59       4.04      2.97
113.22
## ...94        True   171.64      103.58       103.46       3.72      3.20
113.30
## ...95        True   171.79      103.65       103.61       4.19      3.06
113.60
## ...96        True   172.49      103.92       103.91       4.42      2.84
113.38
## ...97        True   172.00      104.32       104.26       4.53      3.04
112.93
## ...98        True   171.49      103.77       103.60       4.01      3.09
112.95
## ...99        True   172.10      103.98       103.86       4.47      3.06
113.00
## ...100       True   171.81      103.96       103.47       4.00      3.00
113.10
## ...101      False   171.45      104.03       104.26       4.88      3.44
111.92
## ...102      False   171.97      104.38       104.18       5.59      3.47
110.98
## ...103      False   171.94      104.21       104.10       4.28      3.47
112.23
## ...104      False   172.04      104.34       104.48       4.88      3.28
112.15
## ...105      False   171.75      104.16       104.23       5.75      3.25
111.68
## ...106      False   171.99      104.18       104.20       5.26      3.23
111.83
## ...107      False   172.22      104.17       104.07       4.52      3.67
112.13
## ...108      False   171.79      104.05       104.30       5.02      3.44
112.01
## ...109      False   172.04      104.17       103.90       5.05      3.62
111.56
## ...110      False   172.22      104.41       104.64       5.20      3.37
112.20
## ...111      False   172.10      104.30       104.21       4.07      3.41
111.27
## ...112      False   172.09      104.40       104.21       5.28      3.41
112.11
## ...113      False   172.32      104.60       104.83       4.84      3.51
```

```
112.55
## ...114      False   171.89      104.32      103.94      5.64      3.30
111.56
## ...115      False   172.10      104.22      103.99      5.26      3.24
111.94
## ...116      False   172.43      104.32      103.95      4.13      3.39
111.50
## ...117      False   171.78      104.51      104.06      5.90      3.18
111.91
## ...118      False   171.75      104.36      104.02      6.00      3.13
111.79
## ...119      False   171.83      104.39      104.17      5.51      3.33
113.64
## ...120      False   171.51      104.13      103.90      4.99      3.60
111.23
## ...121      False   171.84      104.23      104.31      5.10      3.68
111.72
## ...122      False   172.07      104.50      104.23      6.19      3.07
111.21
## ...123      False   172.29      104.72      104.86      5.71      3.16
112.15
## ...124      False   172.05      104.60      104.32      5.12      3.35
111.78
## ...125      False   171.74      104.40      104.39      4.87      3.06
112.00
## ...126      False   171.91      103.99      104.23      5.01      3.42
111.77
## ...127      False   171.99      104.28      104.32      4.71      3.45
112.18
## ...128      False   172.40      104.55      104.22      5.18      3.51
111.94
## ...129      False   172.08      104.15      104.17      4.96      3.40
112.29
## ...130      False   171.62      104.14      104.45      4.94      3.66
111.93
## ...131      False   171.43      104.14      103.95      5.34      3.14
111.76
## ...132      False   171.56      104.17      103.87      6.16      3.38
111.55
## ...133      False   171.94      104.37      104.14      5.37      3.46
111.94
## ...134      False   171.69      104.17      104.37      5.31      3.54
111.89
## ...135      False   171.38      104.04      104.20      5.54      3.38
112.80
## ...136      False   171.86      104.12      104.10      6.01      3.34
111.91
## ...137      False   171.69      103.87      104.16      5.46      3.31
111.42
## ...138      False   171.94      104.56      104.25      4.60      3.37
```

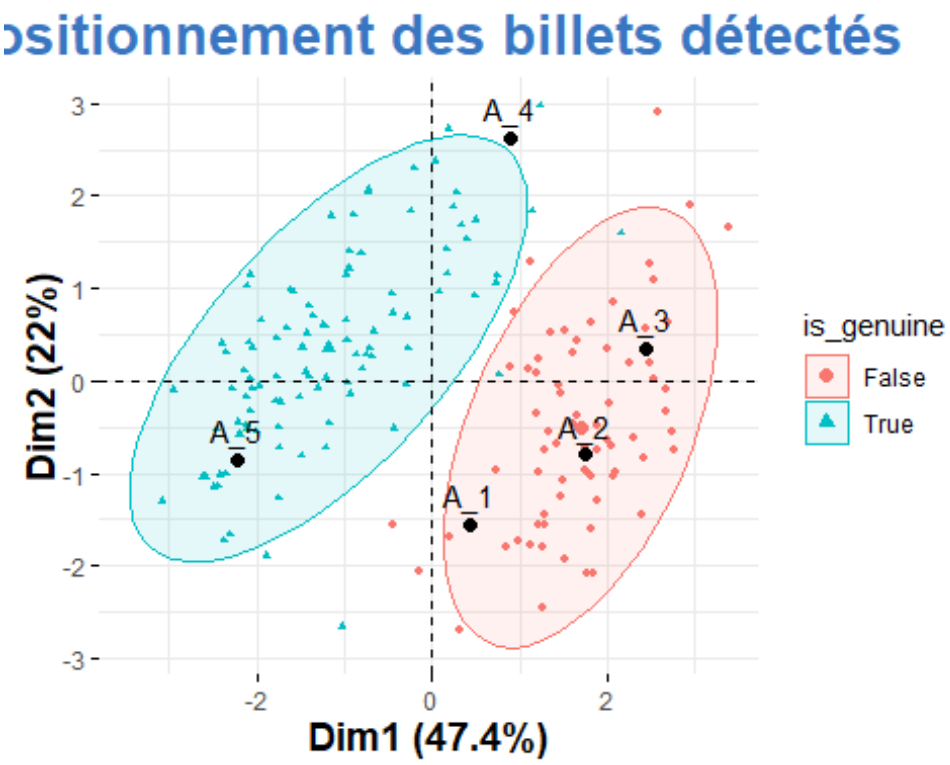
```
110.64
## ...139      False   171.65      104.32      104.38      5.65      3.24
112.30
## ...140      False   171.60      104.37      104.20      5.82      3.08
112.84
## ...141      False   171.83      104.18      104.26      5.00      3.60
111.55
## ...142      False   171.74      103.96      103.47      5.14      3.30
111.40
## ...143      False   171.69      104.18      104.28      5.62      3.23
110.53
## ...144      False   172.00      104.46      104.30      5.27      3.37
111.85
## ...145      False   171.56      103.80      103.87      5.66      2.98
112.95
## ...146      False   171.95      104.47      104.34      5.92      3.10
113.17
## ...147      False   171.98      104.44      104.26      5.75      3.20
110.93
## ...148      False   172.25      104.52      104.22      4.65      3.43
110.48
## ...149      False   171.67      104.16      104.08      5.42      3.30
111.63
## ...150      False   171.91      103.91      103.98      4.78      3.65
111.41
## ...151      False   171.95      104.26      103.97      5.88      3.16
112.44
## ...152      False   171.68      103.89      103.70      5.97      3.03
109.97
## ...153      False   171.67      103.79      103.44      5.13      3.32
111.47
## ...154      False   171.61      104.04      104.06      6.19      3.08
110.73
## ...155      False   171.62      104.21      103.99      5.50      3.45
111.35
## ...156      False   172.10      103.98      104.28      5.78      3.16
111.09
## ...157      False   171.38      103.78      103.70      5.22      3.43
111.60
## ...158      False   171.53      104.03      104.05      5.77      3.22
111.93
## ...159      False   171.84      104.32      104.50      6.28      3.00
111.06
## ...160      False   171.72      104.46      104.12      4.21      3.61
110.31
## ...161      False   172.50      104.07      103.71      3.82      3.63
110.74
## ...162      False   171.92      104.37      104.05      4.95      3.04
110.61
## ...163      False   171.67      104.12      103.98      5.68      3.18
```

```
111.55
## ...164      False   171.78      104.07      104.16      5.77      3.30
111.27
## ...165      False   171.43      104.26      103.97      5.73      3.14
111.82
## ...166      False   172.11      104.23      104.45      5.24      3.58
111.78
## ...167      False   173.01      104.59      104.31      5.04      3.05
110.91
## ...168      False   172.47      104.27      104.10      4.88      3.33
110.68
## ...169      False   171.82      103.97      103.88      4.73      3.55
111.87
## ...170      False   171.96      104.00      103.95      5.63      3.26
110.96
## A_1          <NA>   171.76      104.01      103.54      5.21      3.30
111.42
## A_2          <NA>   171.87      104.17      104.13      6.00      3.31
112.09
## A_3          <NA>   172.00      104.58      104.29      4.99      3.39
111.57
## A_4          <NA>   172.49      104.55      104.34      4.44      3.03
113.20
## A_5          <NA>   171.65      103.63      103.56      3.77      3.16
113.33
```

```
# ACP
res.pca.test=PCA(billets_new[,1:7], quali.sup = 1, ind.sup = 171:175,
scale.unit=TRUE, graph=FALSE, axes=c(1,2))
```

```
## Warning in PCA(billets_new[, 1:7], quali.sup = 1, ind.sup = 171:175,
scale.unit
## = TRUE, : Missing values are imputed by the mean of the variable: you
should use
## the imputePCA function of the missMDA package
```

```
# Représentation par ellipses des billets sur plan factoriel selon leur
authenticité
p <- fviz_pca_ind(res.pca.test, geom.ind = "point", pointsize = 1, habillage
= 1, addEllipses=TRUE, ellipse.level=0.90)

p <- fviz_add(p, res.pca.test$ind.sup$coord, color = "black") + labs(title =
"Positionnement des billets détectés") + theme(plot.title =
element_text(color = '#3876C2', size=20, face='bold', hjust = 0.5),
axis.title.x = element_text(color="black", size = 14, face = "bold", hjust =
0.5), axis.title.y = element_text(color = "black", size = 14, face = "bold",
vjust = 0.5)) + ggsave("graphiques/graphique00_billets_detectes.jpg", width =
16, height = 9)

p
```



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.