Analyse du dataset StudentPerformance

Ahmed Ben Yaflah & Islem Ridene

2025-05-07

Introduction

Cette étude vise à analyser le dataset “StudentPerformance” qui contient des informations sur les performances académiques d’étudiants et leurs caractéristiques socio-démographiques. Notre objectif est d’identifier les facteurs influençant les résultats scolaires et d’explorer les relations entre les différentes variables.

Pour mener cette analyse, nous utilisons plusieurs méthodologies statistiques vues en cours, notamment les analyses univariées (statistiques descriptives, histogrammes, diagrammes) et bivariées (corrélations, régressions simples, tests du khi-deux). Nous complétons cette approche par des techniques d’analyse multivariée comme l’Analyse en Composantes Principales (ACP), l’Analyse Factorielle des Correspondances (AFC) et la régression multiple.

Cette démarche méthodologique nous permettra d’extraire des informations pertinentes sur les déterminants de la réussite scolaire et potentiellement d’orienter des interventions éducatives ciblées.

NB!!: Nous avons utilisé la majorité des fonctions et des bibliothéques vues en classe avec l’ajout d’autres

1. Chargement et exploration des données

- Chargement des packages nécessaires

```
library(plotly)
library(vcd)
library(tidyverse)
library(psych)
library(corrplot)
library(ggplot2)
library(MASS)
library(FactoMineR)
library(factoextra)
library(ca)
```

- Chargement des données

```
data <- read.csv("./StudentsPerformance.csv", header = TRUE, sep = ",")
```

- Aperçu des données

```
head(data)
```

```
##   gender race.ethnicity parental.level.of.education       lunch
## 1 female        group B           bachelor's degree    standard
## 2 female        group C                some college    standard
## 3 female        group B             master's degree    standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college    standard
## 6 female        group B          associate's degree    standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

- Structure des données

```
str(data)
```

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "female" "female" "female" "male" ...
##  $ race.ethnicity             : chr  "group B" "group C" "group B" "group A" ...
##  $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree"
"associate's degree" ...
##  $ lunch                      : chr  "standard" "standard" "standard" "free/reduced" ...
##  $ test.preparation.course    : chr  "none" "completed" "none" "none" ...
##  $ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
##  $ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
##  $ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
```

- Dimensions du dataframe

```
dim(data)
```

```
## [1] 1000    8
```

- Noms des colonnes

```
names(data)
```

```
## [1] "gender"                     "race.ethnicity"
## [3] "parental.level.of.education" "lunch"
## [5] "test.preparation.course"     "math.score"
## [7] "reading.score"               "writing.score"
```

- Analyse descriptive globale

```
describe(data)
```

```
##                              vars    n  mean    sd median trimmed   mad min max
## gender*                         1 1000  1.48  0.50      1    1.48  0.00   1   2
## race.ethnicity*                 2 1000  3.17  1.16      3    3.20  1.48   1   5
## parental.level.of.education*    3 1000  3.49  1.83      3    3.48  2.97   1   6
## lunch*                          4 1000  1.65  0.48      2    1.68  0.00   1   2
## test.preparation.course*        5 1000  1.64  0.48      2    1.68  0.00   1   2
## math.score                      6 1000 66.09 15.16     66   66.38 14.83   0 100
## reading.score                   7 1000 69.17 14.60     70   69.50 14.83  17 100
## writing.score                   8 1000 68.05 15.20     69   68.41 16.31  10 100
##                              range  skew kurtosis   se
## gender*                          1  0.07    -2.00 0.02
## race.ethnicity*                  4 -0.14    -0.75 0.04
## parental.level.of.education*     5 -0.03    -1.45 0.06
## lunch*                           1 -0.61    -1.64 0.02
## test.preparation.course*         1 -0.59    -1.65 0.02
## math.score                     100 -0.28     0.26 0.48
## reading.score                   83 -0.26    -0.08 0.46
## writing.score                   90 -0.29    -0.05 0.48
```

2. Préparation des données

- Conversion des scores en variables numériques

```
data$math_score <- as.numeric(data$'math.score')
data$reading_score <- as.numeric(data$'reading.score')
data$writing_score <- as.numeric(data$'writing.score')
```

- Création d'une variable pour le score total

```
data$total_score <- data$math_score + data$reading_score + data$writing_score
```

- Vérification des valeurs manquantes

```
sum(is.na(data))
```

```
## [1] 0
```

**Ne contient aucune valeur manquante (NA) dans aucune colonne.

3. Analyses univariées

3.1. Variables quantitatives (numériques)

- Statistiques descriptives complètes pour les scores

```
summary(data["math_score"])
```

```
##    math_score
##  Min.   :  0.00
##  1st Qu.: 57.00
##  Median : 66.00
##  Mean   : 66.09
##  3rd Qu.: 77.00
##  Max.   :100.00
```

```
summary(data["reading_score"])
```

```
##  reading_score
##  Min.   : 17.00
##  1st Qu.: 59.00
##  Median : 70.00
##  Mean   : 69.17
##  3rd Qu.: 79.00
##  Max.   :100.00
```

```
summary(data["writing_score"])
```

```
##  writing_score
##  Min.   : 10.00
##  1st Qu.: 57.75
##  Median : 69.00
##  Mean   : 68.05
##  3rd Qu.: 79.00
##  Max.   :100.00
```

```
summary(data["total_score"])
```

```
##   total_score
##  Min.   : 27.0
##  1st Qu.:175.0
##  Median :205.0
##  Mean   :203.3
##  3rd Qu.:233.0
##  Max.   :300.0
```

- Calcul des écarts-types

```
sd(data$math_score)
```

```
## [1] 15.16308
```

```
sd(data$reading_score)
```

```
## [1] 14.60019
```

```
sd(data$writing_score)
```

```
## [1] 15.19566
```

```
sd(data$total_score)
```

```
## [1] 42.77198
```

- Calcul des quartiles

```
sapply(data[, c("math_score", "reading_score", "writing_score", "total_score")],
       function(x) quantile(x, probs = c(0.25, 0.5, 0.75)))
```

```
##     math_score reading_score writing_score total_score
## 25%         57            59         57.75         175
## 50%         66            70         69.00         205
## 75%         77            79         79.00         233
```

- Intervalles de confiance pour les moyennes: Quantifier la précision de l'estimation

```
t.test(data$math_score)$conf.int#On estime avec 95 % de confiance que la vraie moyenne des sc
ores en mathématiques dans la population se situe entre 65.15 et 67.03.
```

```
## [1] 65.14806 67.02994
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(data$reading_score)$conf.int#La vraie moyenne des scores en lecture a 95 % de chances
d'être comprise entre 68.26 et 70.08.
```

```
## [1] 68.26299 70.07501
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(data$writing_score)$conf.int#La vraie moyenne des scores en écriture a 95 % de chances
d'être comprise entre 67.11 et 69.00.
```

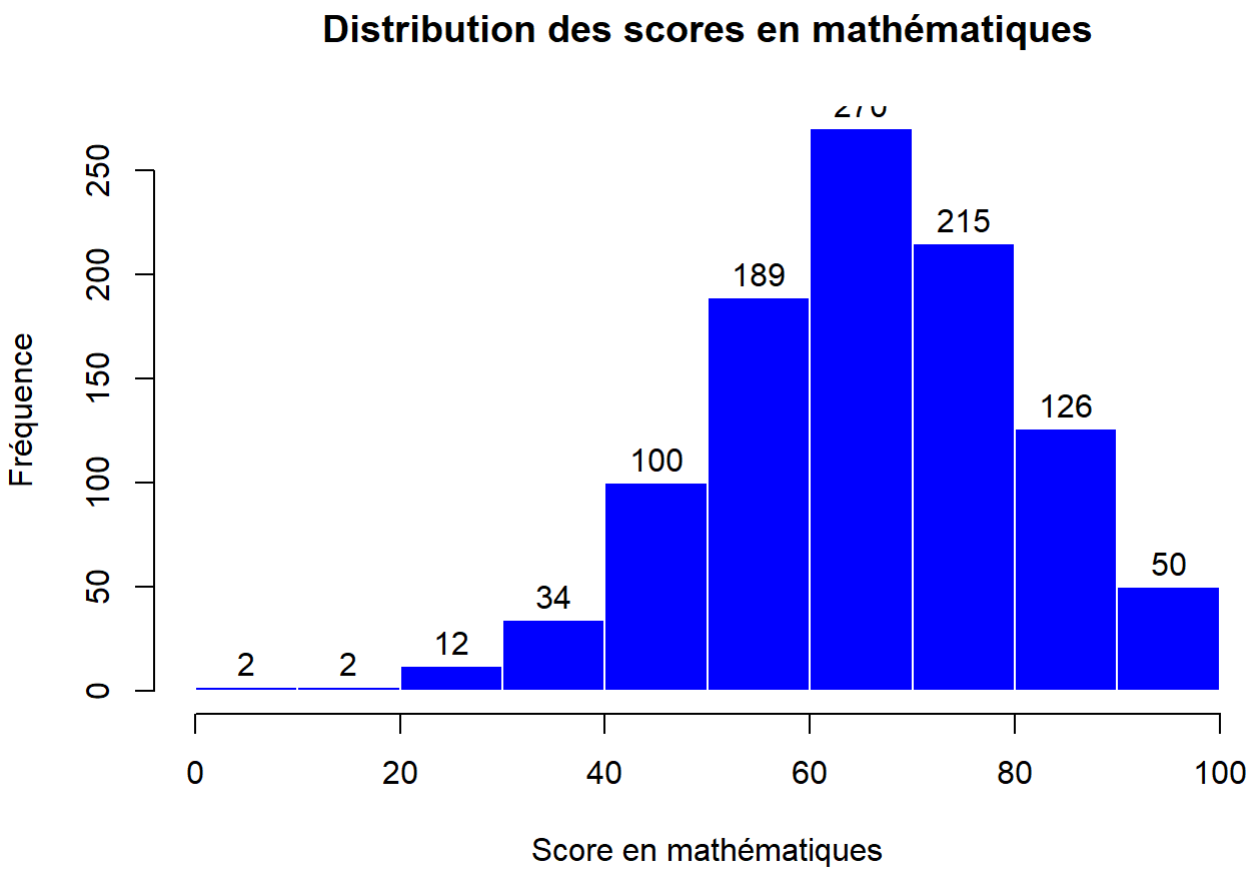
```
## [1] 67.11104 68.99696
## attr(,"conf.level")
## [1] 0.95
```

**Les scores de lecture ont la moyenne estimée la plus élevée (autour de 69.17), suivis de l'écriture (~68.05) et des mathématiques (~66.09).

**Les largeurs des intervalles sont similaires, ce qui suggère une précision comparable pour les trois estimations. Représentations graphiques

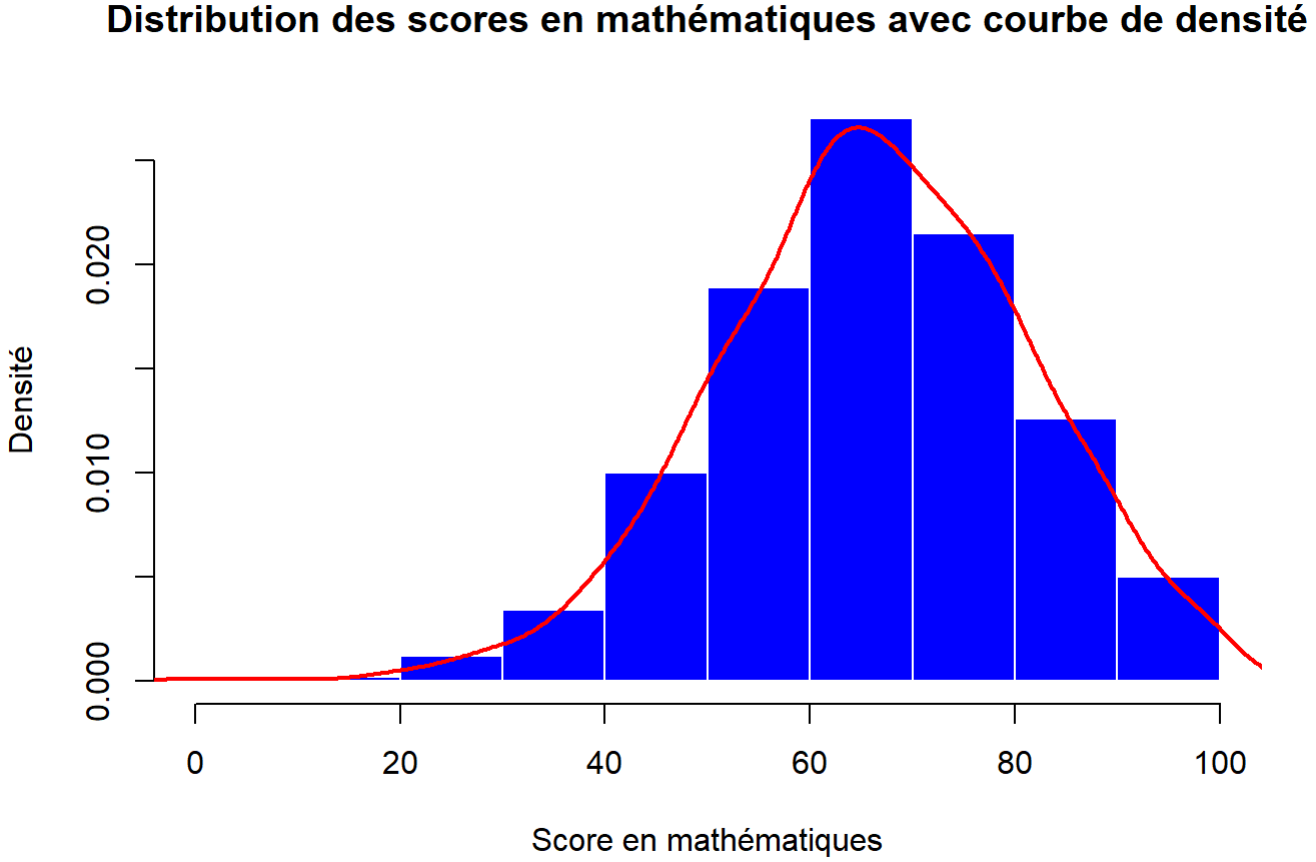
- Histogramme des scores en mathématiques

```
hist(data$math_score,
     col = "blue",
     border = "white",
     main = "Distribution des scores en mathématiques",
     xlab = "Score en mathématiques",
     ylab = "Fréquence",
     labels = TRUE)
```



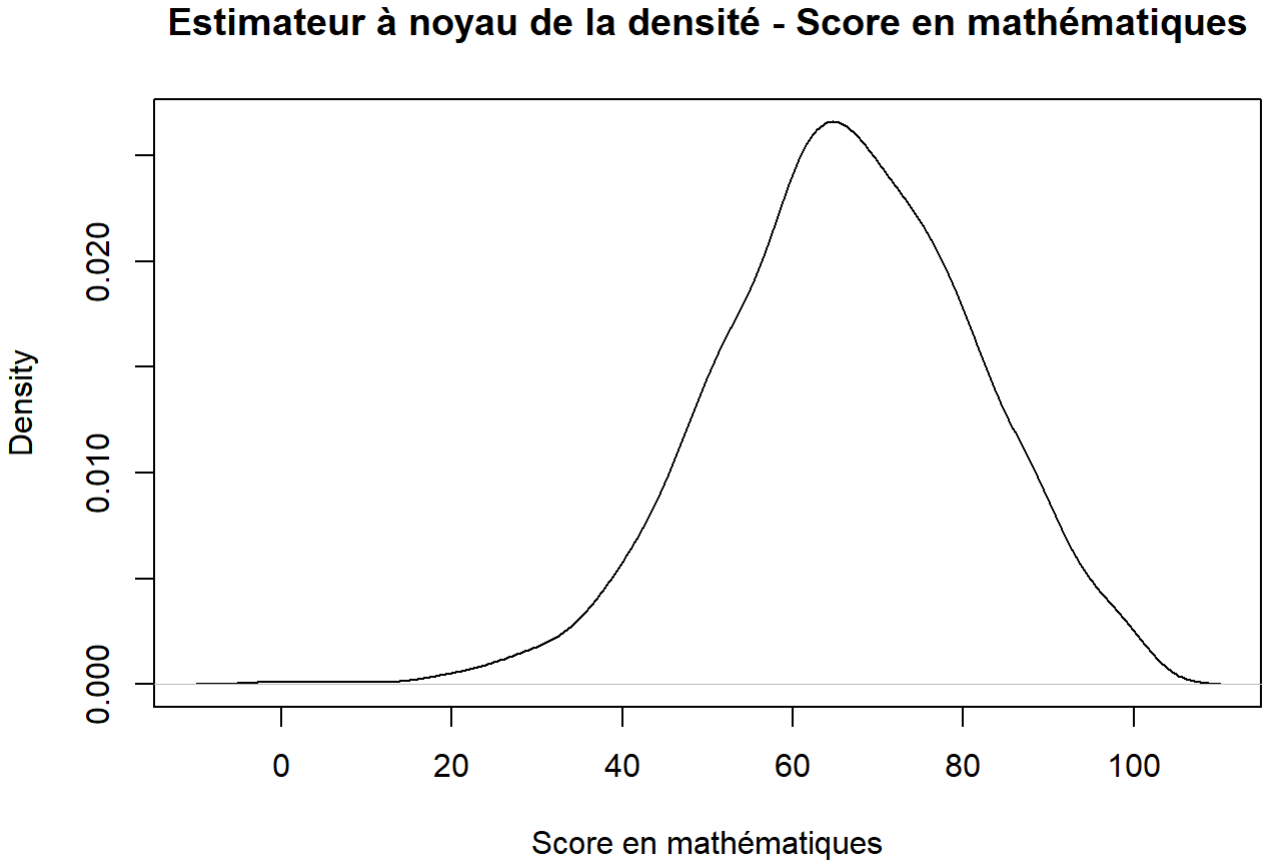
- Histogramme avec courbe de densité

```
hist(data$math_score,
     col = "blue",
     border = "white",
     prob = TRUE,
     main = "Distribution des scores en mathématiques avec courbe de densité",
     xlab = "Score en mathématiques",
     ylab = "Densité")
lines(density(data$math_score), lwd = 2, col = "red")
```



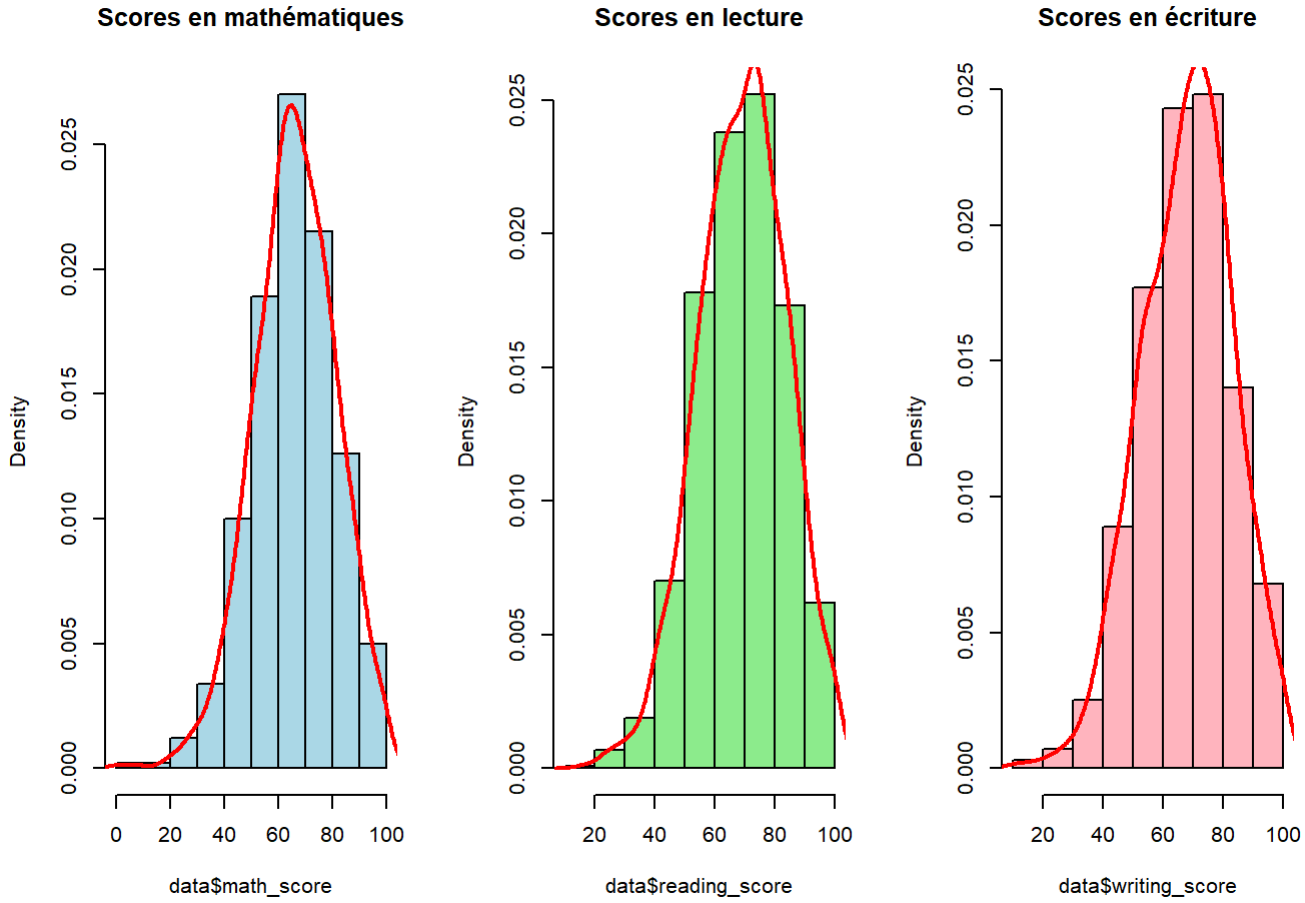
- Estimateur à noyau de la densité

```
plot(density(data$math_score),
     main = "Estimateur à noyau de la densité - Score en mathématiques",
     xlab = "Score en mathématiques")
```



- Histogrammes pour les autres scores

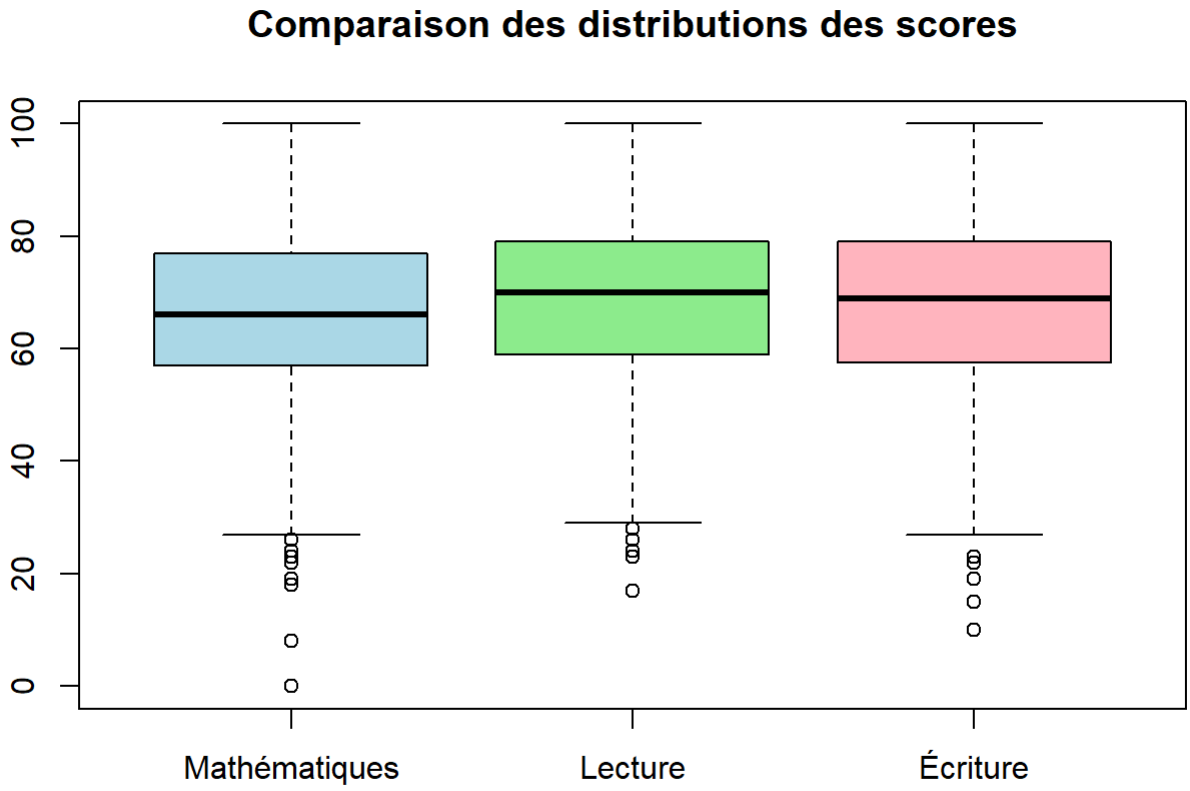
```
par(mfrow = c(1, 3))
hist(data$math_score, main = "Scores en mathématiques", col = "lightblue", prob = TRUE)
lines(density(data$math_score), lwd = 2, col = "red")

hist(data$reading_score, main = "Scores en lecture", col = "lightgreen", prob = TRUE)
lines(density(data$reading_score), lwd = 2, col = "red")

hist(data$writing_score, main = "Scores en écriture", col = "lightpink", prob = TRUE)
lines(density(data$writing_score), lwd = 2, col = "red")
```




```
par(mfrow = c(1, 1))#nettoyer l'espace graphique avant de créer un nouveau plot et éviter que les graphiques suivants s'affichent dans une grille précédente
```

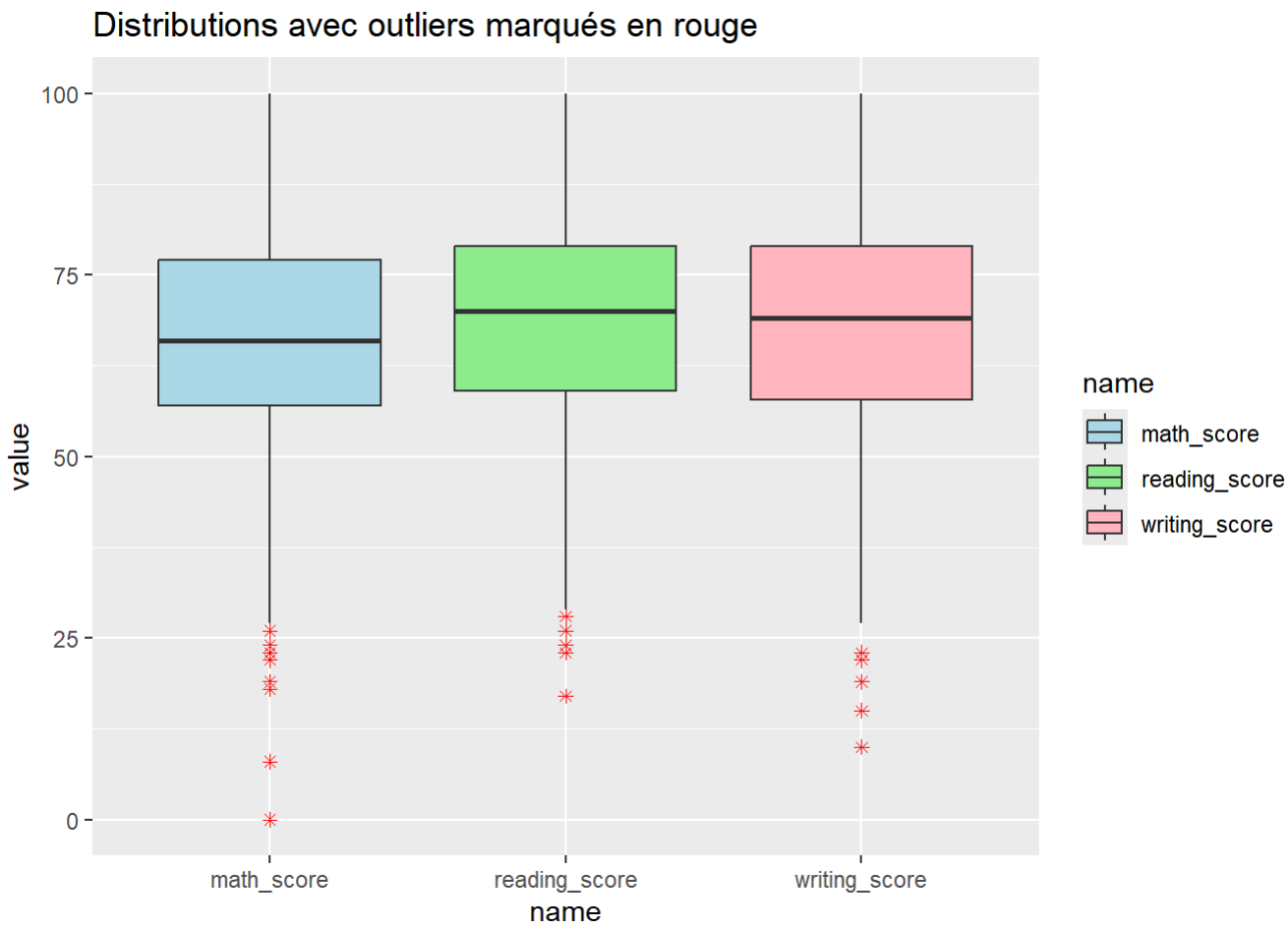
- Boîtes à moustaches pour comparer les distributions

```
boxplot(data$math_score, data$reading_score, data$writing_score,
        names = c("Mathématiques", "Lecture", "Écriture"),
        col = c("lightblue", "lightgreen", "lightpink"),
        main = "Comparaison des distributions des scores")
```



- Les Outliers : observations qui s'écartent significativement du reste des données

```
data_long <- tidyr::pivot_longer(data, cols = c(math_score, reading_score, writing_score))

ggplot(data_long, aes(x = name, y = value, fill = name)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 8) +
  scale_fill_manual(values = c("lightblue", "lightgreen", "lightpink")) +
  labs(title = "Distributions avec outliers marqués en rouge")
```



3.2 Variables qualitatives(catégorielles)

- Tableau de fréquences pour le genre

```
table_gender <- table(data$gender)
table_gender
```

```
##
## female   male
##    518    482
```

```
prop.table(table_gender) * 100  # Pourcentages
```

```
##
## female   male
##   51.8   48.2
```

- Tableau de fréquences pour le groupe ethnique

```
table_race <- table(data$`race.ethnicity`)
table_race
```

```
##
## group A group B group C group D group E
##      89     190     319     262     140
```

```
prop.table(table_race) * 100  # Pourcentages
```

```
##
## group A group B group C group D group E
##     8.9    19.0    31.9    26.2    14.0
```

- Tableau de fréquences pour le niveau d'éducation des parents

```
table_education <- table(data$`parental.level.of.education`)
table_education
```

```
##
## associate's degree  bachelor's degree        high school    master's degree
##                222                118                196                 59
##       some college   some high school
##                226                179
```

```
prop.table(table_education) * 100  # Pourcentages
```

```
##
## associate's degree  bachelor's degree        high school    master's degree
##               22.2               11.8               19.6                5.9
##       some college   some high school
##               22.6               17.9
```

- Tableau de fréquences pour le type de déjeuner

```
table_lunch <- table(data$lunch)
table_lunch
```

```
##
## free/reduced     standard
##          355          645
```

```
prop.table(table_lunch) * 100  # Pourcentages
```

```
##
## free/reduced     standard
##         35.5         64.5
```

-Tableau de fréquences pour la préparation aux tests

```
table_prep <- table(data$`test.preparation.course`)
table_prep
```

```
##
## completed      none
##       358       642
```

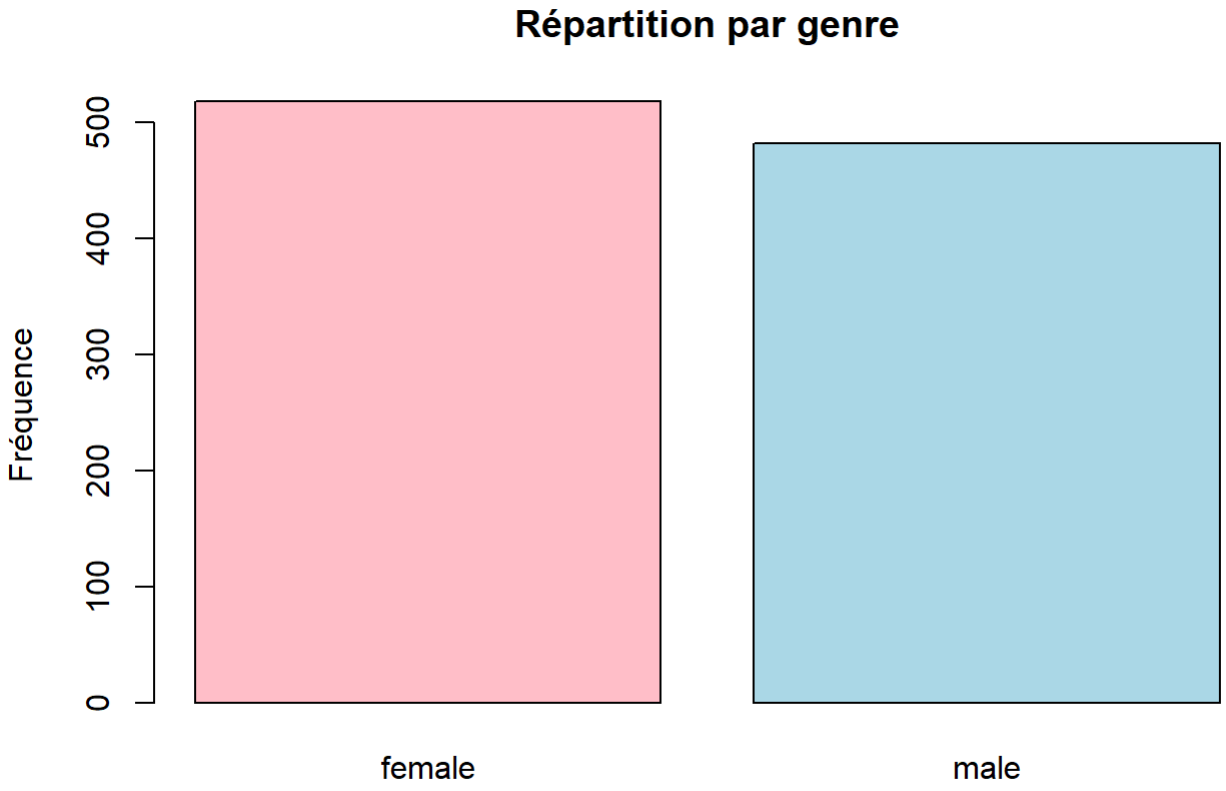
```
prop.table(table_prep) * 100  # Pourcentages
```

```
##
## completed      none
##      35.8      64.2
```

Représentations graphiques

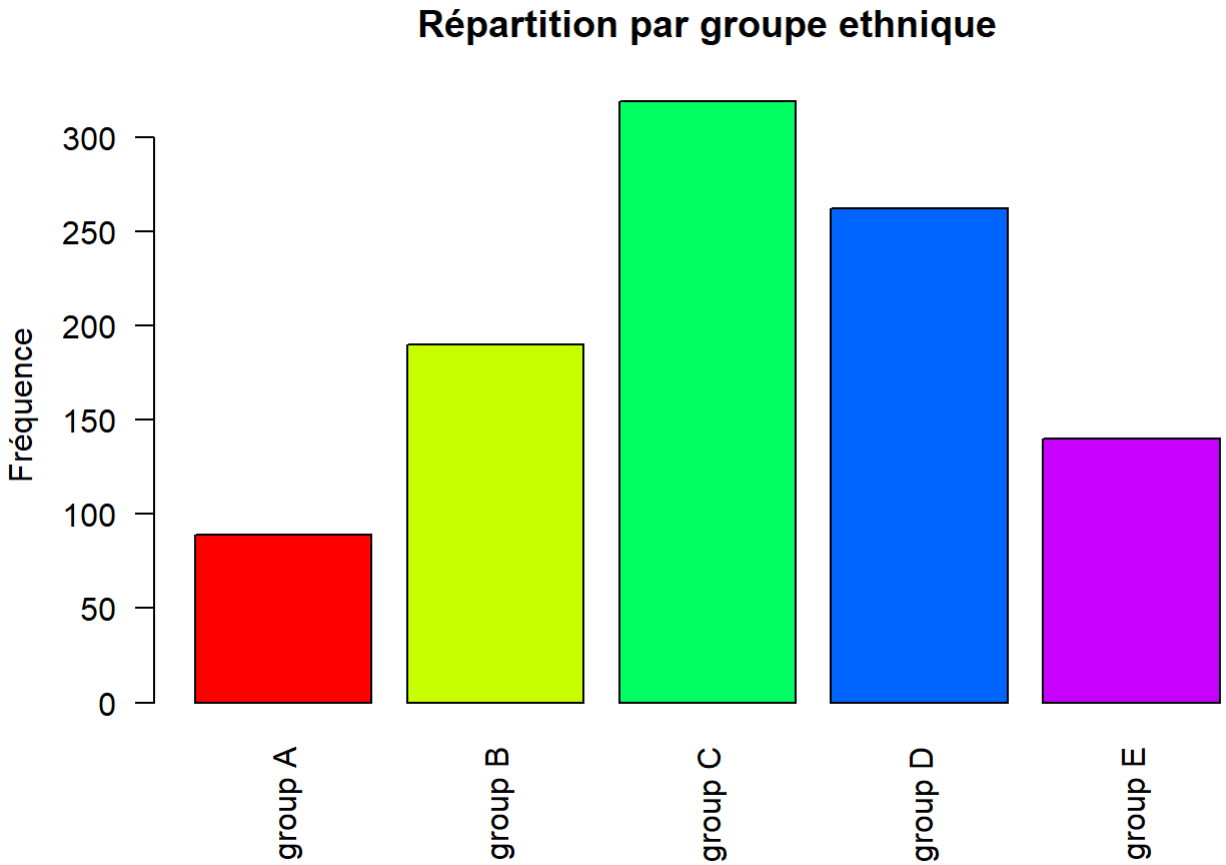
- Diagramme en barres pour le genre

```
barplot(table(data$gender),
        col = c("pink", "lightblue"),
        main = "Répartition par genre",
        ylab = "Fréquence")
```



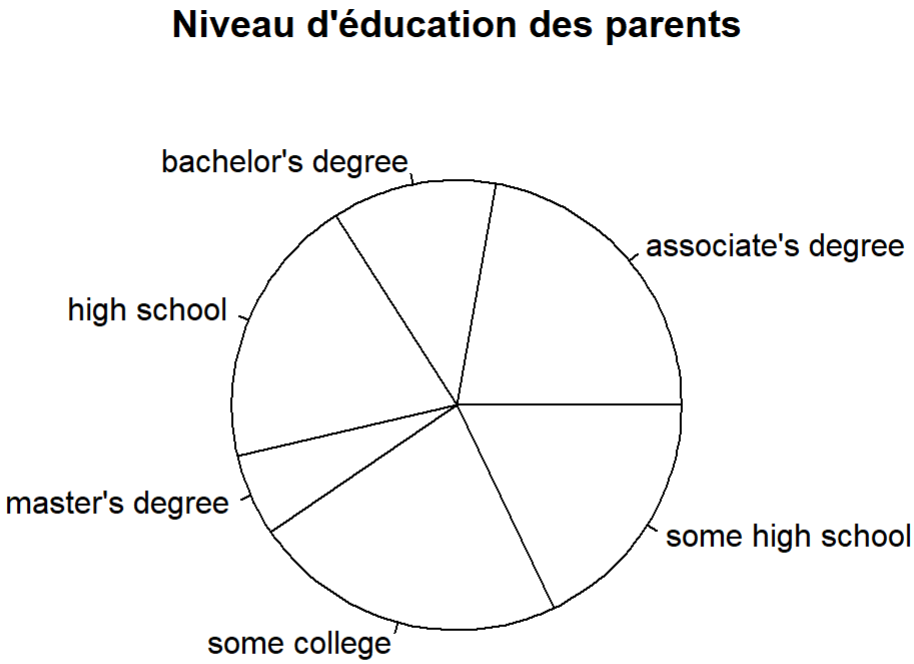
- Diagramme en barres pour le groupe ethnique

```
barplot(table(data$`race.ethnicity`),
        col = rainbow(length(levels(factor(data$`race.ethnicity`)))),
        main = "Répartition par groupe ethnique",
        las = 2,
        ylab = "Fréquence")
```



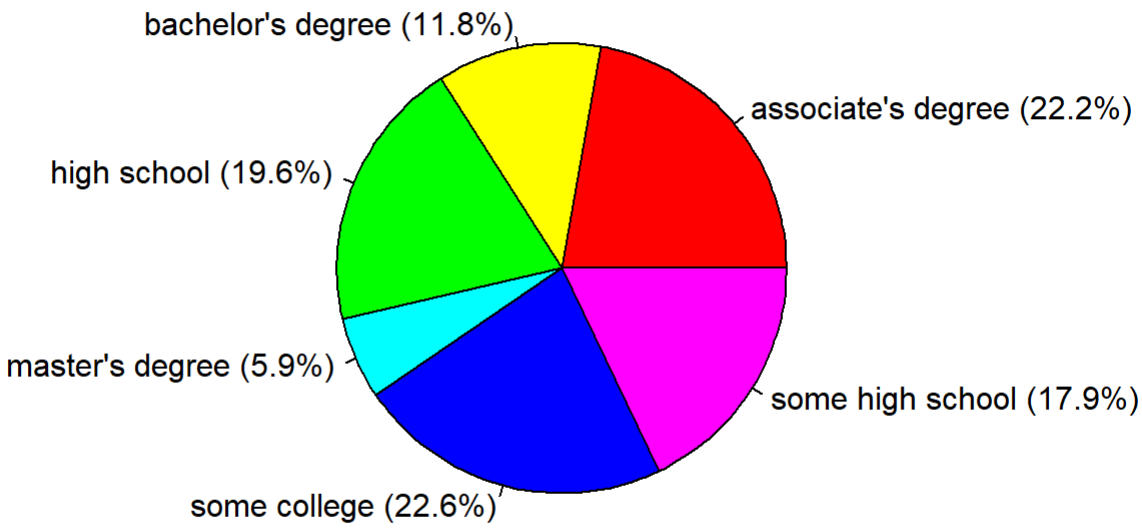
- Diagramme en secteurs pour le niveau d'éducation des parents

```
pie(table(data$`parental.level.of.education`),
    col = rainbow(length(levels(factor(data$`parental.leve.of.education`)))),
    main = "Niveau d'éducation des parents")
```



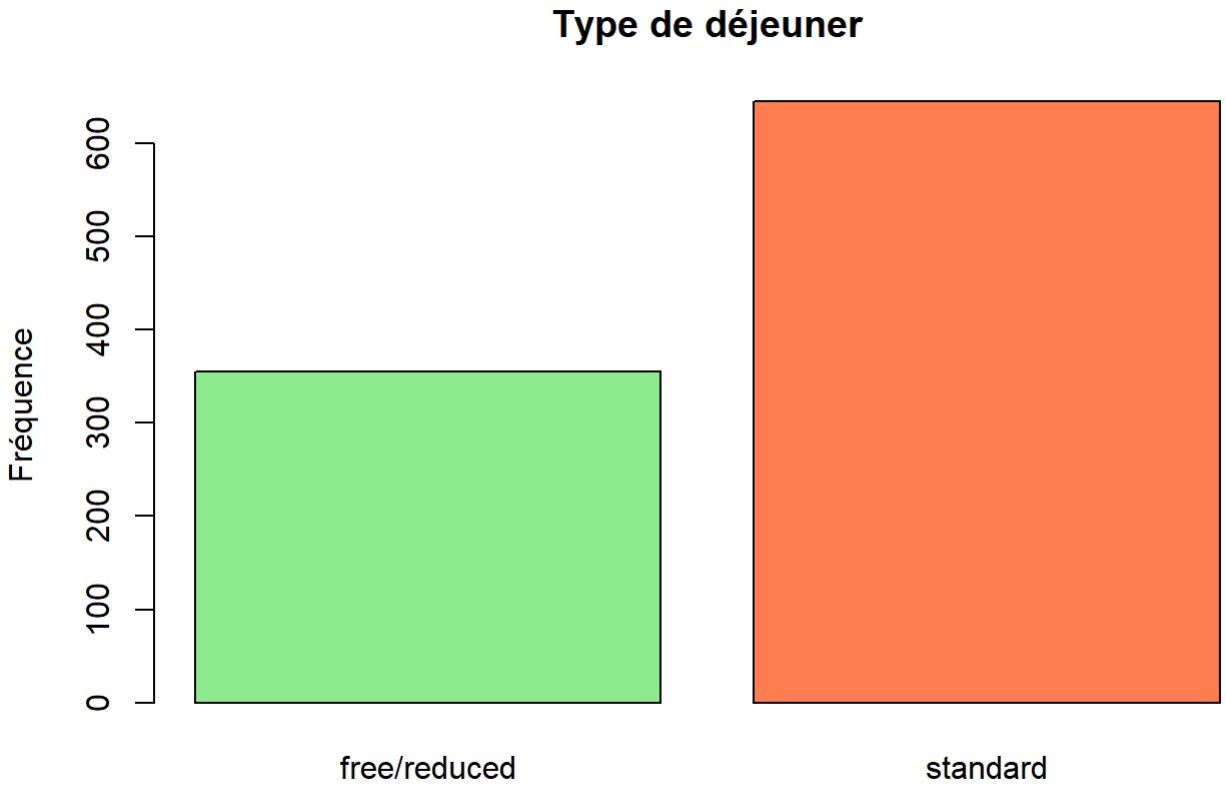
- Ajout des pourcentages au diagramme en secteurs

```
education_counts <- table(data$`parental.level.of.education`)
education_labels <- names(education_counts)
education_pct <- round(prop.table(education_counts) * 100, 1)
education_labels <- paste(education_labels, " (", education_pct, "%)", sep = "")

pie(education_counts,
    labels = education_labels,
    col = rainbow(length(education_counts)),
    main = "Niveau d'éducation des parents (avec pourcentages)")
```



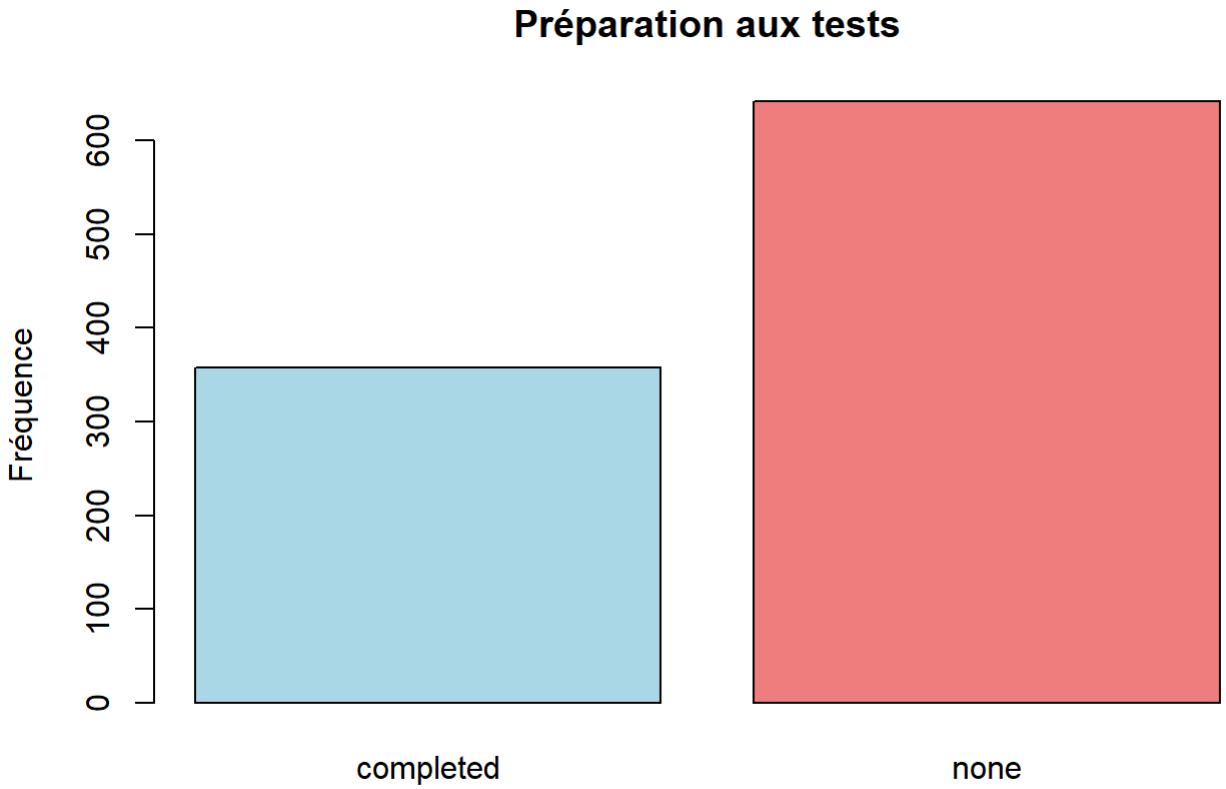
- Diagramme en barres pour le type de déjeuner

```
barplot(table(data$lunch),
        col = c("lightgreen", "coral"),
        main = "Type de déjeuner",
        ylab = "Fréquence")
```



- Diagramme en barres pour la préparation aux tests

```
barplot(table(data$`test.preparation.course`),
        col = c("lightblue", "lightcoral"),
        main = "Préparation aux tests",
        ylab = "Fréquence")
```



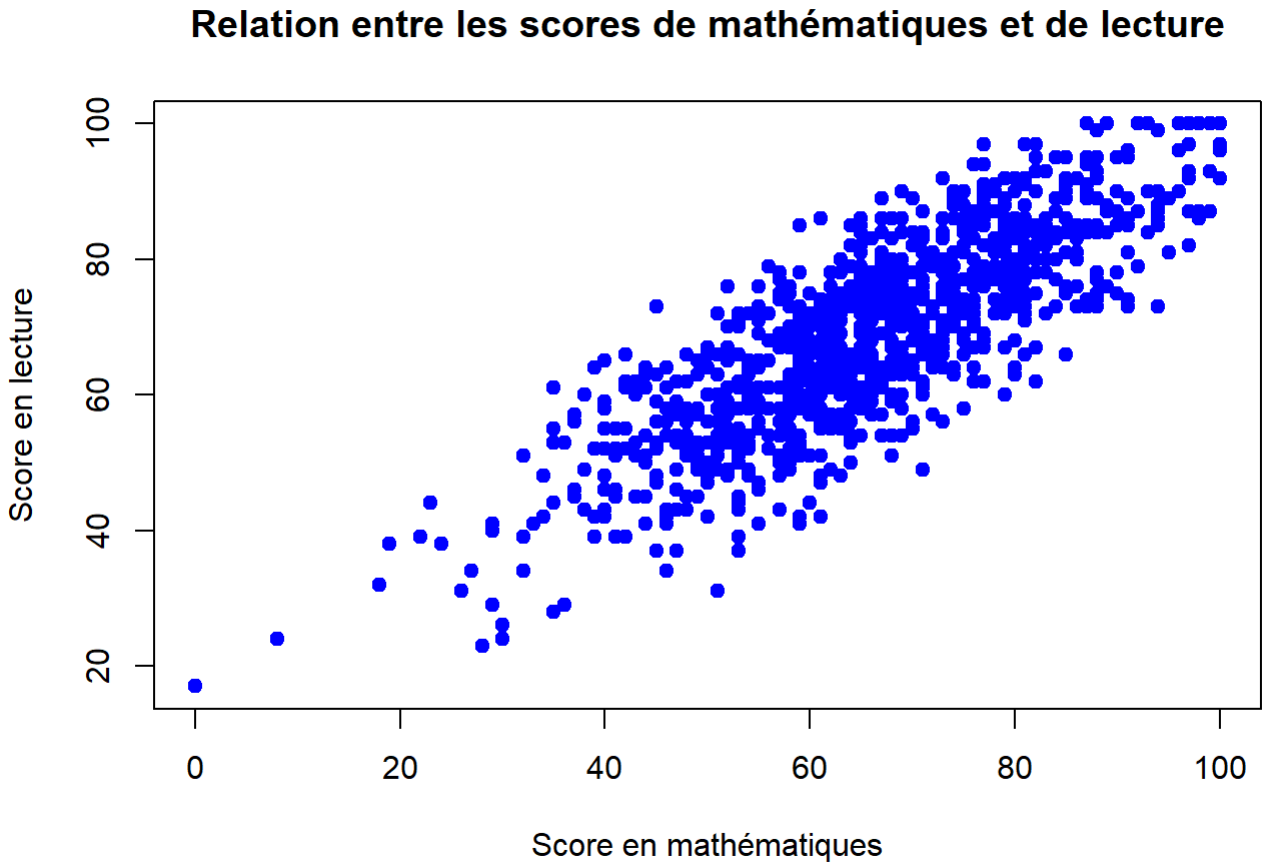
4. Analyses bivariées

4.1 Relations entre variables quantitatives

Représentations graphiques

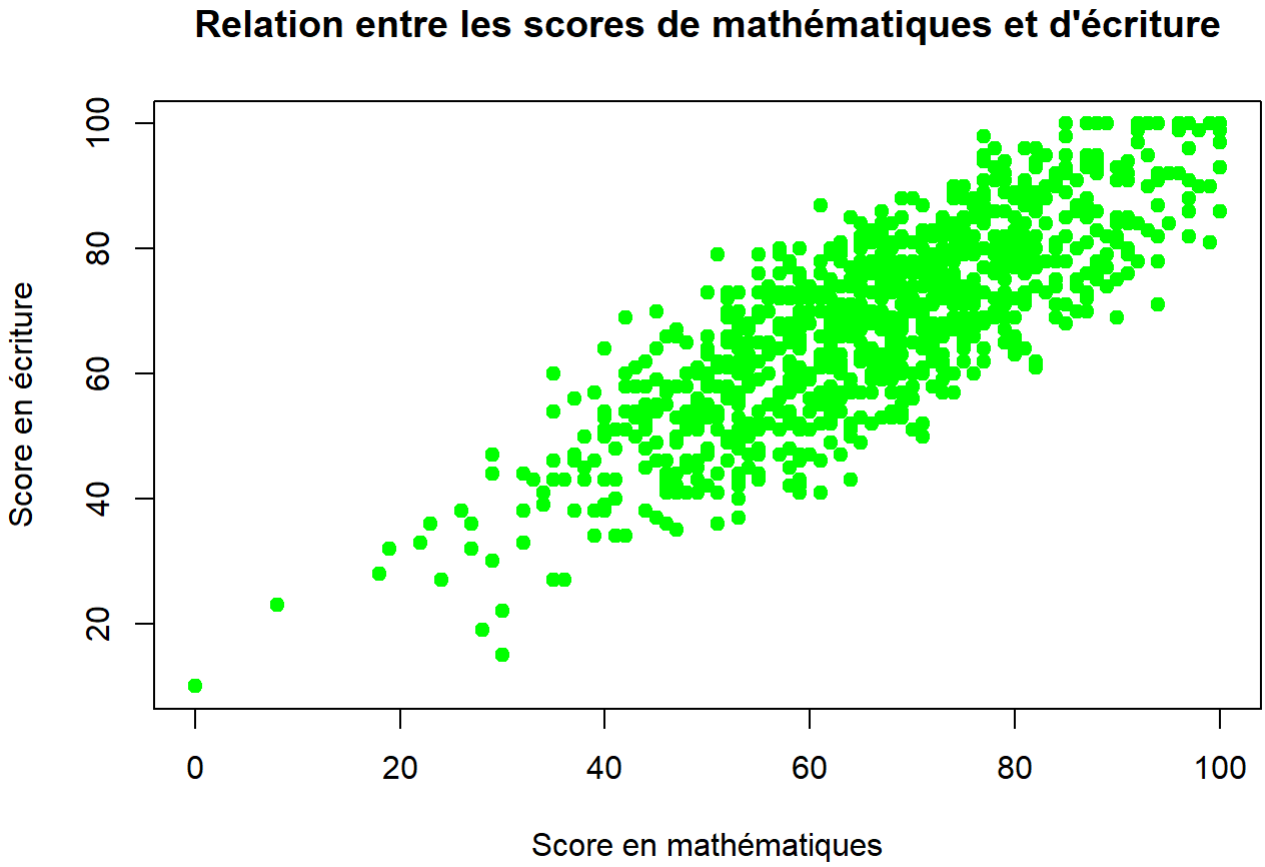
- Nuage de points entre les scores de mathématiques et de lecture

```
plot(data$math_score, data$reading_score,
     pch = 19, col = "blue",
     main = "Relation entre les scores de mathématiques et de lecture",
     xlab = "Score en mathématiques",
     ylab = "Score en lecture")
```



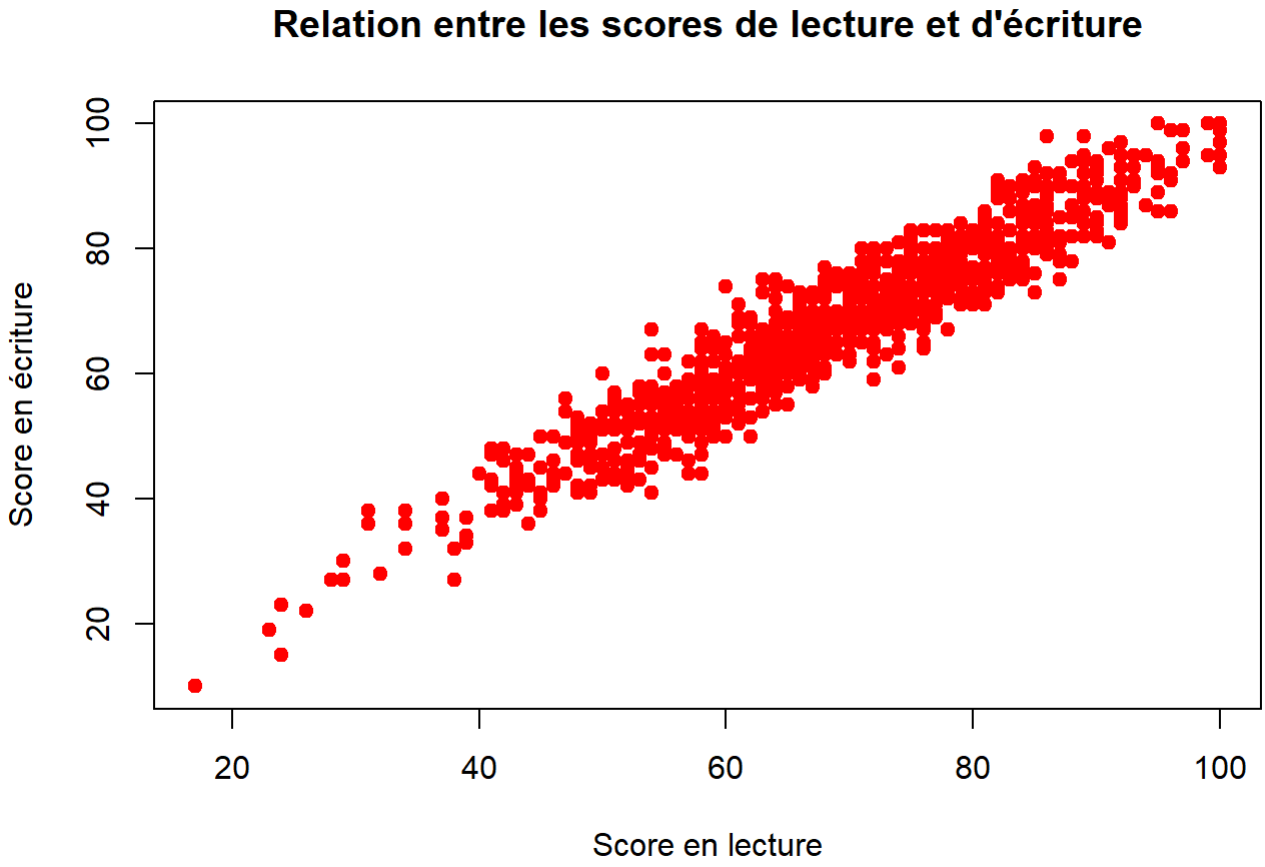
- Nuage de points entre les scores de mathématiques et d'écriture

```
plot(data$math_score, data$writing_score,
     pch = 19, col = "green",
     main = "Relation entre les scores de mathématiques et d'écriture",
     xlab = "Score en mathématiques",
     ylab = "Score en écriture")
```



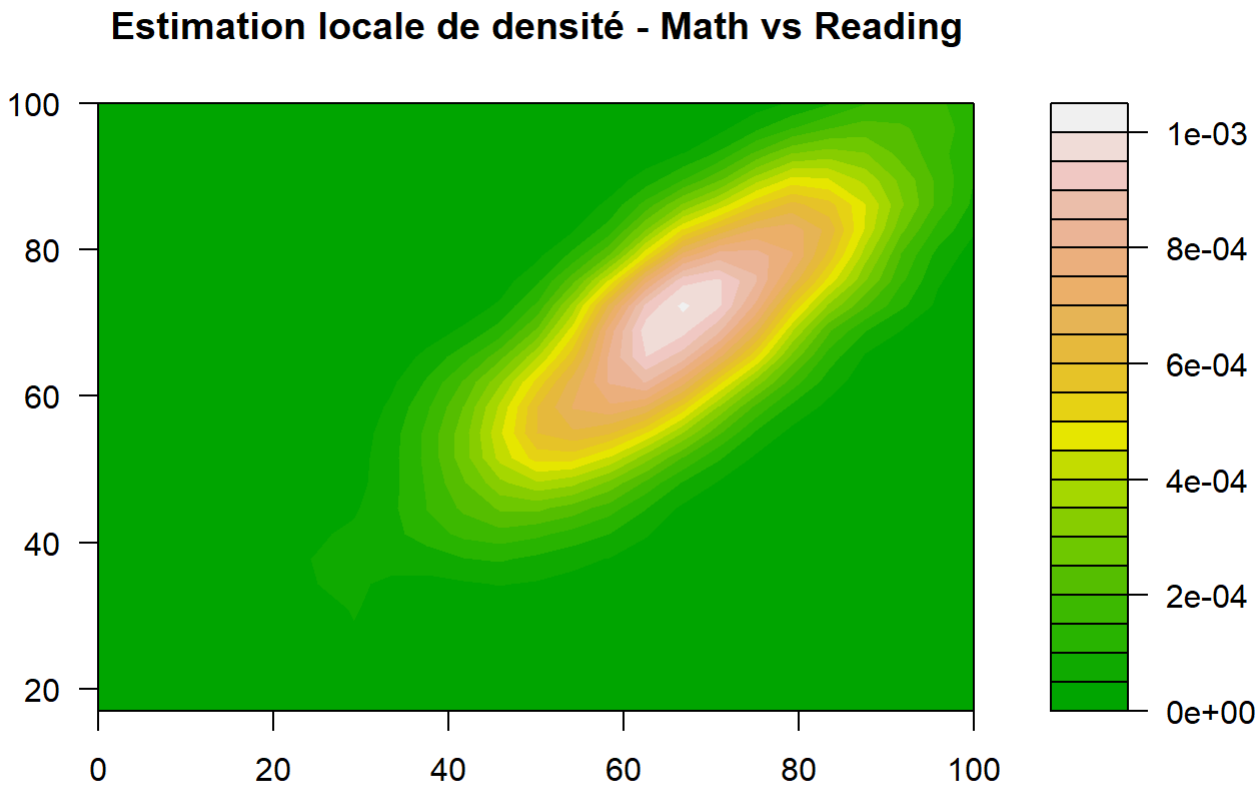
- Nuage de points entre les scores de lecture et d’écriture

```
plot(data$reading_score, data$writing_score,
     pch = 19, col = "red",
     main = "Relation entre les scores de lecture et d'écriture",
     xlab = "Score en lecture",
     ylab = "Score en écriture")
```



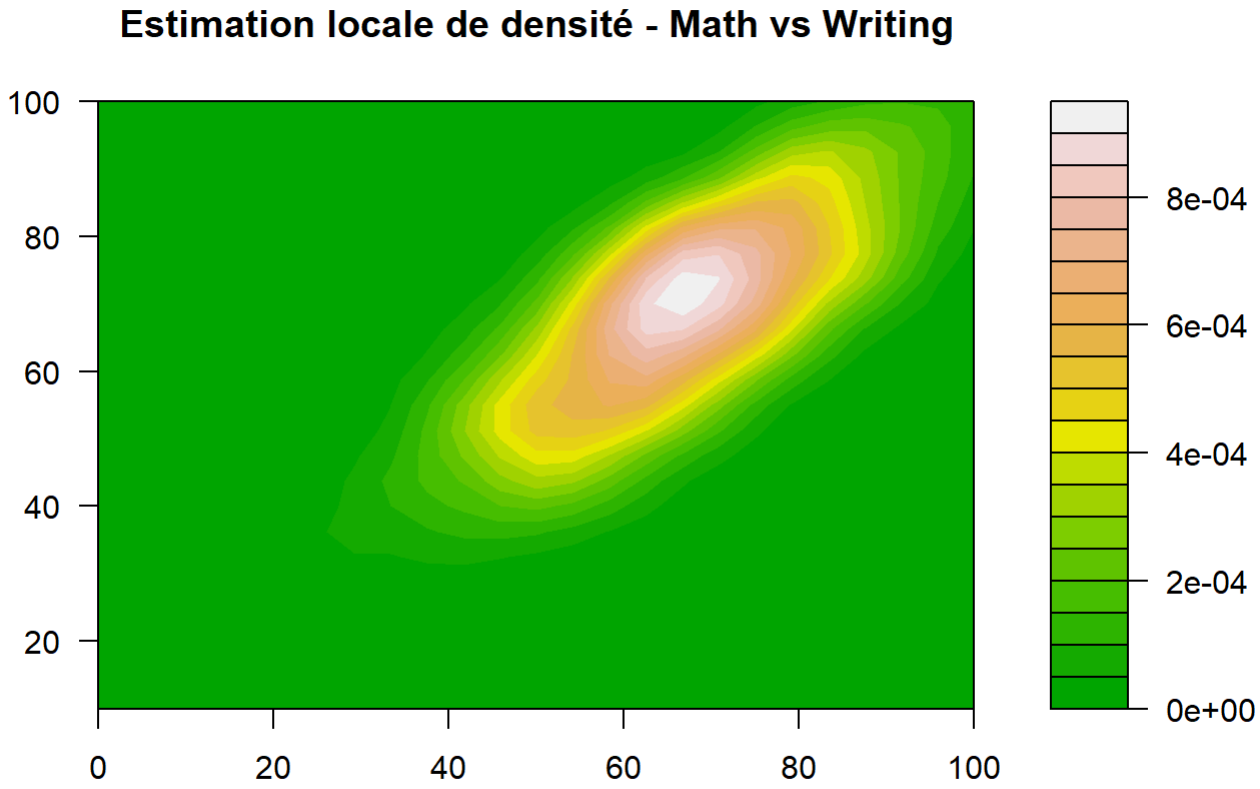
- Estimation locale de densité pour math et reading

```
tmp <- data[, c("math_score", "reading_score")]
tmp <- tmp[complete.cases(tmp), ]
filled.contour(kde2d(tmp$math_score, tmp$reading_score),
               color = terrain.colors,
               main = "Estimation locale de densité - Math vs Reading")
```

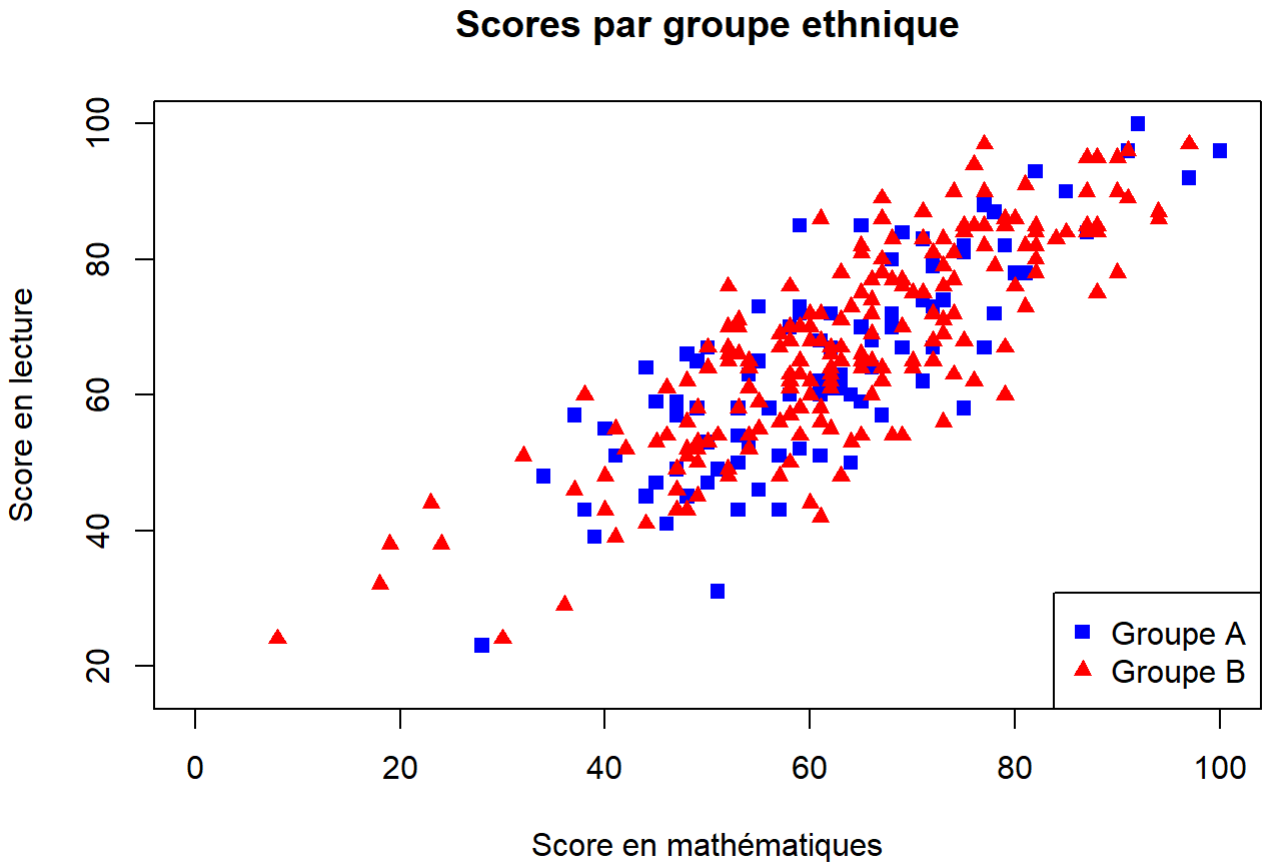


- Estimation locale de densité pour math et writing

```
tmp <- data[, c("math_score", "writing_score")]
tmp <- tmp[complete.cases(tmp), ]
filled.contour(kde2d(tmp$math_score, tmp$writing_score),
               color = terrain.colors,
               main = "Estimation locale de densité - Math vs Writing")
```




```
# Nuage de points avec coloration par groupe ethnique
# Sélection de deux groupes pour l'exemple
group_a <- data$`race.ethnicity` == "group A"
group_b <- data$`race.ethnicity` == "group B"

plot(data$math_score[group_a], data$reading_score[group_a],
     pch = 15, col = "blue",
     main = "Scores par groupe ethnique",
     xlab = "Score en mathématiques",
     ylab = "Score en lecture",
     xlim = range(data$math_score),
     ylim = range(data$reading_score))
points(data$math_score[group_b], data$reading_score[group_b],
       pch = 17, col = "red")
legend("bottomright",
       legend = c("Groupe A", "Groupe B"),
       col = c("blue", "red"),
       pch = c(15, 17))
```



Coefficients de corrélation

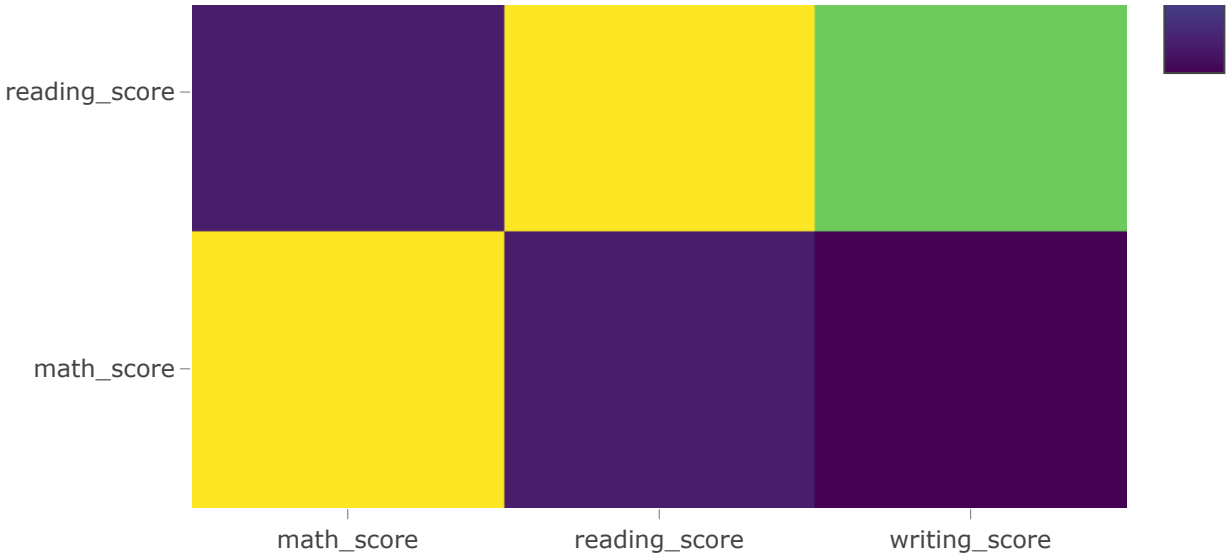
- Matrice de corrélation entre les scores

```
cor_matrix <- cor(data[, c("math_score", "reading_score", "writing_score")],
                  use = "complete.obs")
cor_matrix
```

```
##               math_score reading_score writing_score
## math_score     1.0000000     0.8175797     0.8026420
## reading_score  0.8175797     1.0000000     0.9545981
## writing_score  0.8026420     0.9545981     1.0000000
```

- Visualisation de la matrice de corrélation

```
plot_ly(z = cor_matrix, type = "heatmap",
        x = colnames(cor_matrix),
        y = rownames(cor_matrix))
```



- Coefficient de corrélation entre mathématiques et lecture

```
cor_math_reading <- cor(data$math_score, data$reading_score, use = "complete.obs")
cat("Coefficient de corrélation entre math et reading:", cor_math_reading, "\n")
```

```
## Coefficient de corrélation entre math et reading: 0.8175797
```

- Coefficient de corrélation entre mathématiques et écriture

```
cor_math_writing <- cor(data$math_score, data$writing_score, use = "complete.obs")
cat("Coefficient de corrélation entre math et writing:", cor_math_writing, "\n")
```

```
## Coefficient de corrélation entre math et writing: 0.802642
```

- Coefficient de corrélation entre lecture et écriture

```
cor_reading_writing <- cor(data$reading_score, data$writing_score, use = "complete.obs")
cat("Coefficient de corrélation entre reading et writing:", cor_reading_writing, "\n")
```

```
## Coefficient de corrélation entre reading et writing: 0.9545981
```

- Interprétation des coefficients de corrélation

```
cat("\nInterprétation des coefficients de corrélation:\n")
```

```
##
## Interprétation des coefficients de corrélation:
```

```
if(cor_math_reading > 0.7) {
  cat(" Forte corrélation positive entre math et reading\n")
} else if(cor_math_reading > 0.3) {
  cat(" Corrélation modérée positive entre math et reading\n")
} else {
  cat(" Faible corrélation entre math et reading\n")
}
```

```
##  Forte corrélation positive entre math et reading
```

```
if(cor_math_writing > 0.7) {
  cat(" Forte corrélation positive entre math et writing\n")
} else if(cor_math_writing > 0.3) {
  cat(" Corrélation modérée positive entre math et writing\n")
} else {
  cat(" Faible corrélation entre math et writing\n")
}
```

```
##  Forte corrélation positive entre math et writing
```

```
if(cor_reading_writing > 0.7) {
  cat(" Forte corrélation positive entre reading et writing\n")
} else if(cor_reading_writing > 0.3) {
  cat(" Corrélation modérée positive entre reading et writing\n")
} else {
  cat(" Faible corrélation entre reading et writing\n")
}
```

```
##  Forte corrélation positive entre reading et writing
```

Régression linéaire simple

- Régression linéaire: score en lecture en fonction du score en mathématiques

```
reg <- lm(reading_score ~ math_score, data = data)
summary(reg)
```

```
##
## Call:
## lm(formula = reading_score ~ math_score, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2905  -5.8011   0.1139   6.0341  21.4117
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.14181    1.19000   14.40   <2e-16 ***
## math_score   0.78723    0.01755   44.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.411 on 998 degrees of freedom
## Multiple R-squared:  0.6684, Adjusted R-squared:  0.6681
## F-statistic:  2012 on 1 and 998 DF,  p-value: < 2.2e-16
```

- Interprétation des coefficients

```
cat("Équation de la droite de régression: reading_score =",
    round(reg$coefficients[1], 2), "+",
    round(reg$coefficients[2], 2), "* math_score\n")
```

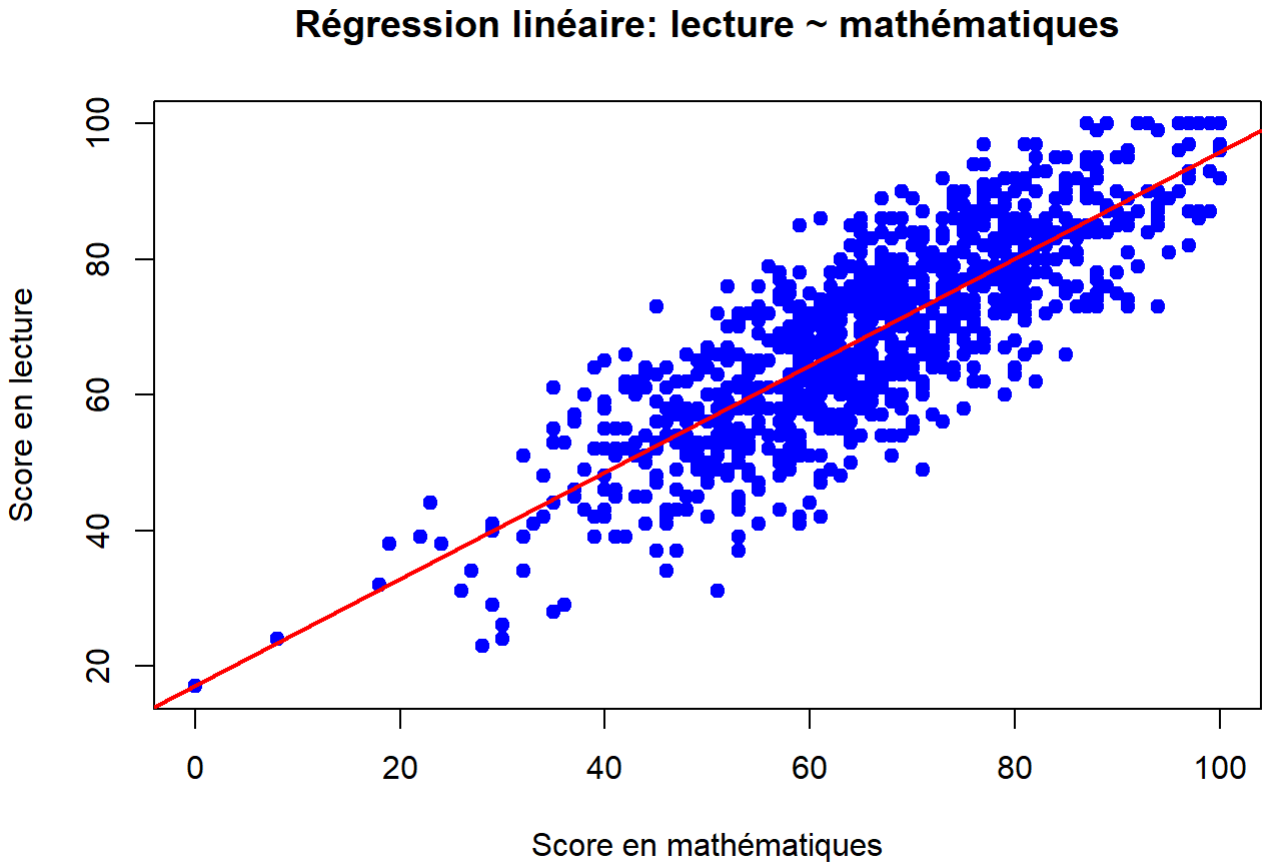
```
## Équation de la droite de régression: reading_score = 17.14 + 0.79 * math_score
```

```
cat("R² =", round(summary(reg)$r.squared, 4),
    "- Cela signifie que", round(summary(reg)$r.squared * 100, 2),
    "% de la variance du score en lecture est expliquée par le score en mathématiques.\n")
```

```
## R² = 0.6684 - Cela signifie que 66.84 % de la variance du score en lecture est expliquée p
ar le score en mathématiques.
```

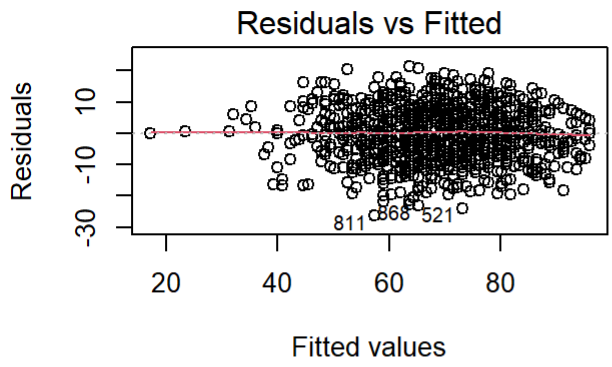
- Visualisation de la régression

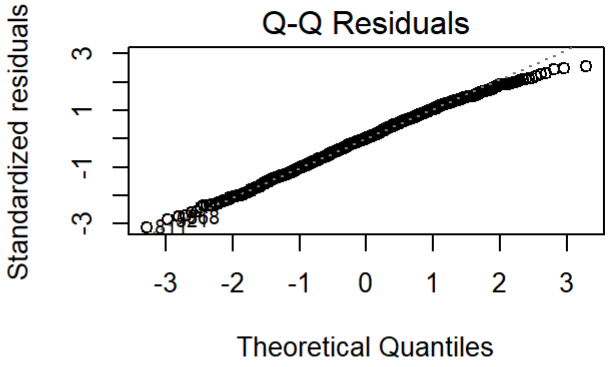
```
plot(data$math_score, data$reading_score,
     pch = 19, col = "blue",
     main = "Régression linéaire: lecture ~ mathématiques",
     xlab = "Score en mathématiques",
     ylab = "Score en lecture")
abline(reg, col = "red", lwd = 2)
```

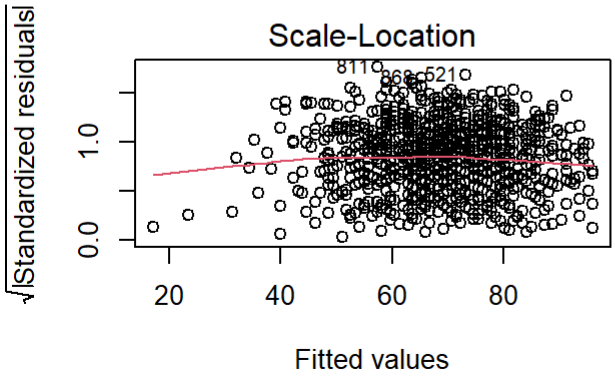


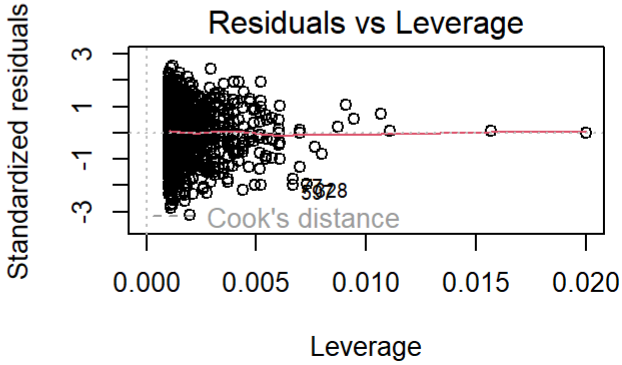
- Analyse des résidus

```
par(mfrow = c(2, 2))
plot(reg)
```










```
par(mfrow = c(1, 1))
```

- Interprétation des résidus

```
cat("\nAnalyse des résidus:\n")
```

```
##
## Analyse des résidus:
```

```
cat("- Résidus vs Valeurs ajustées: ",
    ifelse(shapiro.test(reg$residuals)$p.value > 0.05,
           "Les résidus semblent homoscédastiques (variance constante).",
           "Les résidus présentent de l'hétéroscédasticité."), "\n")
```

```
## - Résidus vs Valeurs ajustées:  Les résidus présentent de l'hétéroscédasticité.
```

```
cat("- Normal Q-Q: ",
    ifelse(shapiro.test(reg$residuals)$p.value > 0.05,
           "Les résidus suivent approximativement une distribution normale.",
           "Les résidus ne suivent pas une distribution normale."), "\n")
```

```
## - Normal Q-Q:  Les résidus ne suivent pas une distribution normale.
```

- Prédictions

```
new_data <- data.frame(math_score = c(50, 70, 90))
predictions <- predict(reg, newdata = new_data, interval = "confidence")
predictions_df <- cbind(new_data, predictions)
predictions_df
```

```
##   math_score      fit      lwr      upr
## 1         50 56.50327 55.74204 57.26450
## 2         70 72.24785 71.70880 72.78691
## 3         90 87.99244 87.01746 88.96742
```

4.2 Relations entre variables qualitatives

- Tableau de contingence entre genre et préparation aux tests

```
table_gender_prep <- table(data$gender, data$`test.preparation.course`)
table_gender_prep
```

```
##
##          completed none
##   female       184  334
##   male         174  308
```

- Proportions

```
prop.table(table_gender_prep)
```

```
##
##          completed  none
##   female     0.184 0.334
##   male       0.174 0.308
```

```
prop.table(table_gender_prep, margin = 1)  # Proportions par ligne
```

```
##
##          completed      none
##   female 0.3552124 0.6447876
##   male   0.3609959 0.6390041
```

```
prop.table(table_gender_prep, margin = 2)  # Proportions par colonne
```

```
##
##          completed      none
##   female 0.5139665 0.5202492
##   male   0.4860335 0.4797508
```

- Test du khi-deux

```
chi_test <- chisq.test(table_gender_prep)
chi_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_gender_prep
## X-squared = 0.015529, df = 1, p-value = 0.9008
```

- Interprétation du test du khi-deux

```
cat("Test du khi-deux:\n")
```

```
## Test du khi-deux:
```

```
cat("- Valeur du khi-deux:", chi_test$statistic, "\n")
```

```
## - Valeur du khi-deux: 0.0155292
```

```
cat("- Degrés de liberté:", chi_test$parameter, "\n")
```

```
## - Degrés de liberté: 1
```

```
cat("- p-value:", chi_test$p.value, "\n")
```

```
## - p-value: 0.9008274
```

```
cat("- Interprétation: ",
    ifelse(chi_test$p.value < 0.05,
           "Il existe une relation significative entre le genre et la préparation aux test
s.",
           "Il n'y a pas de relation significative entre le genre et la préparation aux test
s."), "\n")
```

```
## - Interprétation:  Il n'y a pas de relation significative entre le genre et la préparation
aux tests.
```

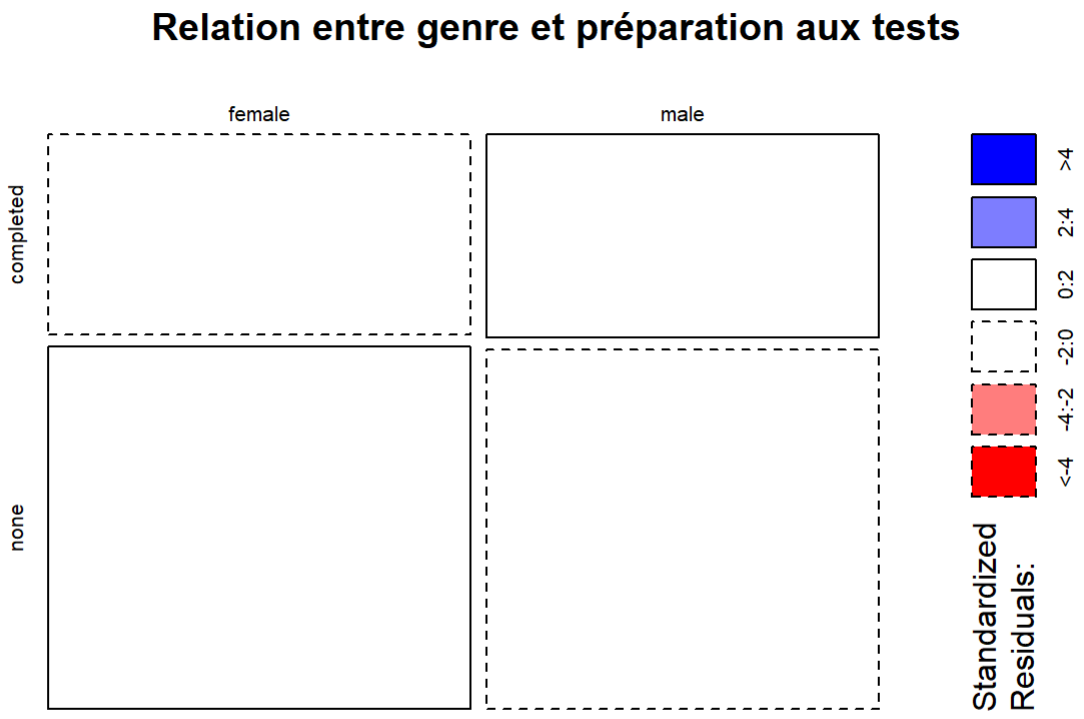
- Test exact de Fisher

```
fisher_test <- fisher.test(table_gender_prep)
fisher_test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_gender_prep
## p-value = 0.895
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7465866 1.2739441
## sample estimates:
## odds ratio
##  0.9751694
```

- Graphique en mosaïque

```
mosaicplot (table_gender_prep, data = data, shade = TRUE, main="Relation entre genre et prépa
ration aux tests")
```



- Tableau de contingence entre groupe ethnique et niveau d'éducation des parents

```
table_race_edu <- table(data$`race.ethnicity`, data$`parental.level.of.education`)
table_race_edu
```

```
##
##            associate's degree bachelor's degree high school master's degree
##   group A                  14                 12          18                3
##   group B                  41                 20          48                6
##   group C                  78                 40          64               19
##   group D                  50                 28          44               23
##   group E                  39                 18          22                8
##
##            some college some high school
##   group A            18                24
##   group B            37                38
##   group C            69                49
##   group D            67                50
##   group E            35                18
```

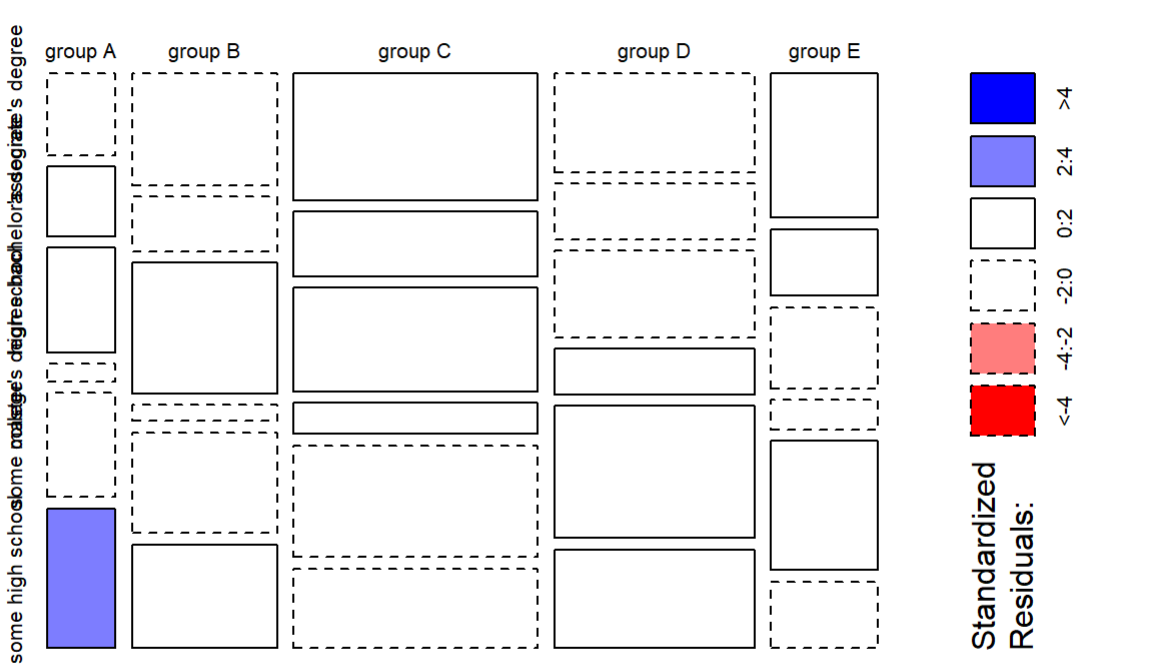
- Test du khi-deux

```
chi_test_race_edu <- chisq.test(table_race_edu)
chi_test_race_edu
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_race_edu
## X-squared = 29.459, df = 20, p-value = 0.07911
```

- Graphique en mosaïque

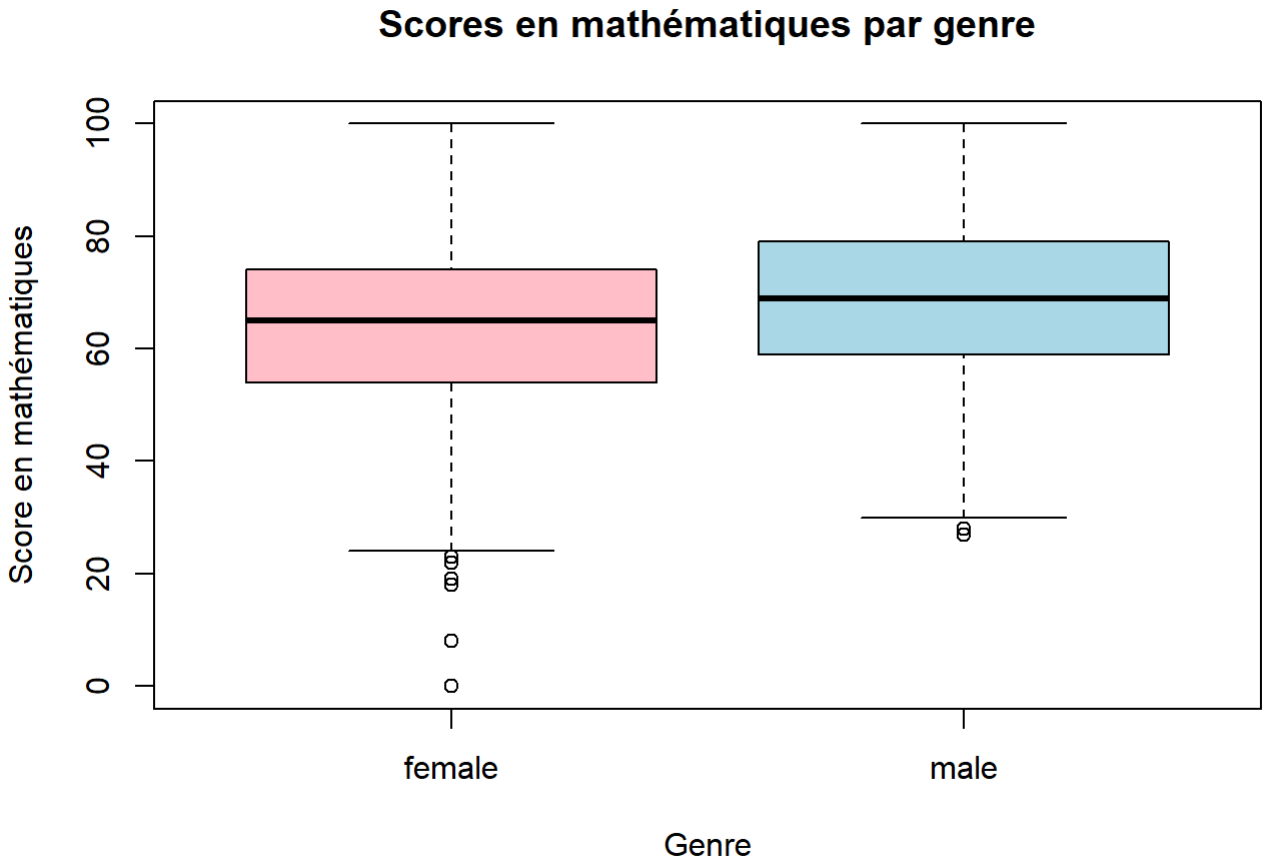
```
mosaicplot(table_race_edu,
           shade = TRUE,
           main = "Relation entre groupe ethnique et niveau d'éducation des parents")
```



4.3 Relations entre variables qualitatives et quantitatives

- Boîtes à moustaches des scores en mathématiques par genre

```
boxplot(math_score ~ gender, data = data,
        col = c("pink", "lightblue"),
        main = "Scores en mathématiques par genre",
        xlab = "Genre",
        ylab = "Score en mathématiques")
```



- Statistiques descriptives par genre

```
by(data$math_score, data$gender, summary)
```

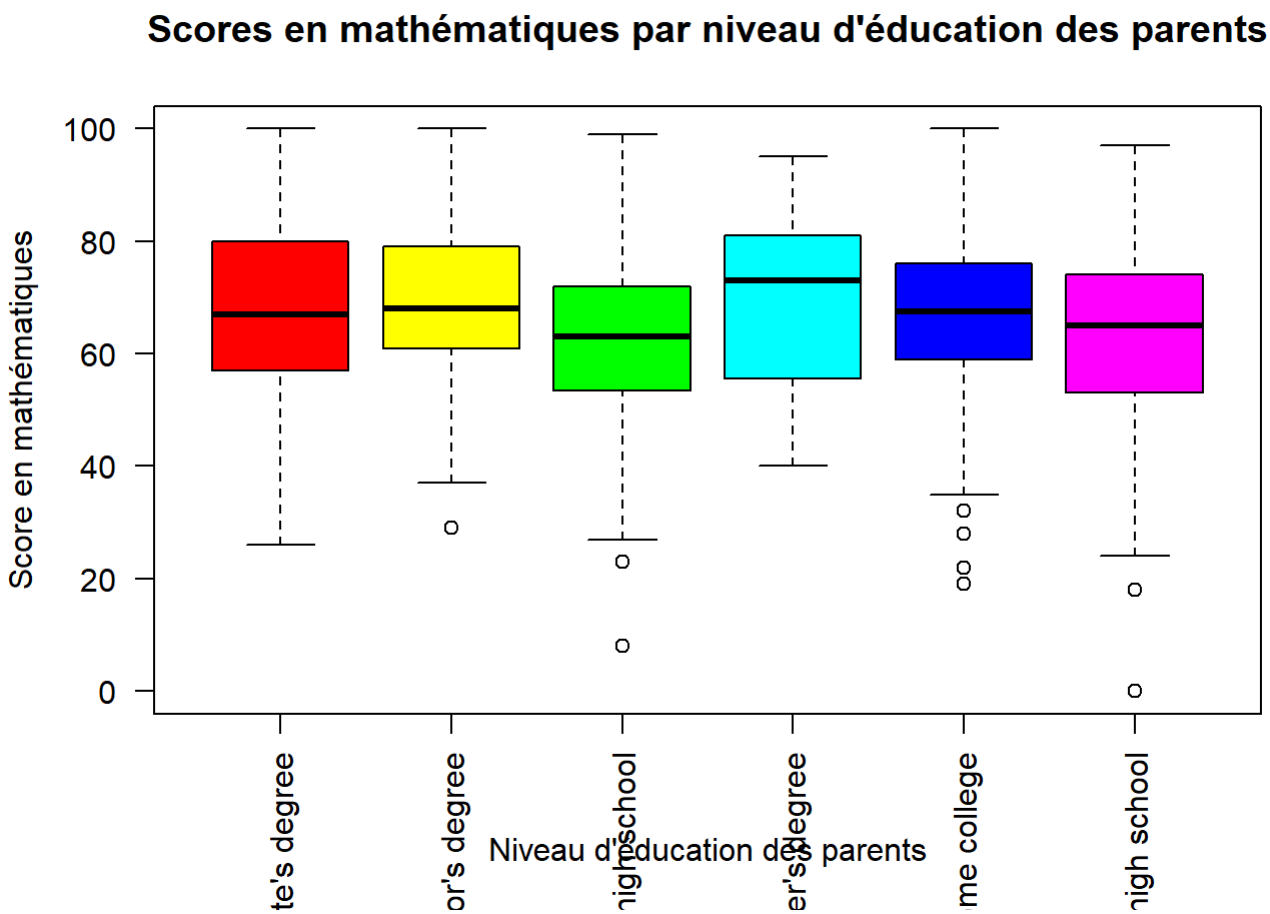
```
## data$gender: female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   54.00   65.00   63.63   74.00  100.00
## ------------------------------------------------------------
## data$gender: male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   59.00   69.00   68.73   79.00  100.00
```

```
by(data$math_score, data$gender, sd)
```

```
## data$gender: female
## [1] 15.49145
## ------------------------------------------------------------
## data$gender: male
## [1] 14.35628
```

- Boîtes à moustaches des scores en mathématiques par niveau d'éducation des parents

```
boxplot(math_score ~ `parental.level.of.education`, data = data,
        col = rainbow(length(levels(factor(data$`parental.level.of.education`)))),
        main = "Scores en mathématiques par niveau d'éducation des parents",
        xlab = "Niveau d'éducation des parents",
        ylab = "Score en mathématiques",
        las = 2)
```



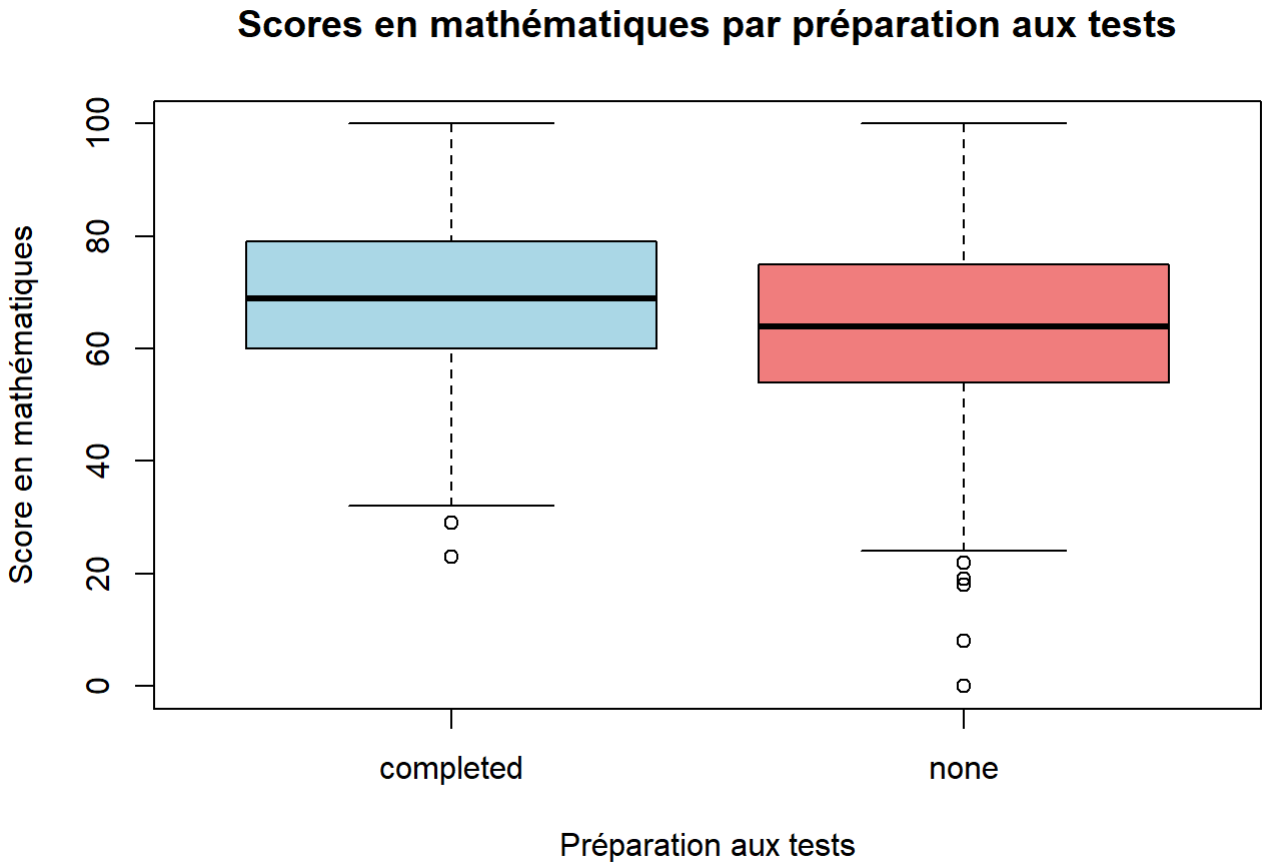
- Statistiques descriptives par niveau d’éducation

```
by(data$math_score, data$`parental.level.of.education`, summary)
```

```
## data$parental.level.of.education: associate's degree
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   57.00   67.00   67.88   80.00  100.00
## ------------------------------------------------------------
## data$parental.level.of.education: bachelor's degree
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   61.00   68.00   69.39   79.00  100.00
## ------------------------------------------------------------
## data$parental.level.of.education: high school
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   53.75   63.00   62.14   72.00   99.00
## ------------------------------------------------------------
## data$parental.level.of.education: master's degree
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.00   55.50   73.00   69.75   81.00   95.00
## ------------------------------------------------------------
## data$parental.level.of.education: some college
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   59.00   67.50   67.13   76.00  100.00
## ------------------------------------------------------------
## data$parental.level.of.education: some high school
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    53.0    65.0    63.5    74.0    97.0
```

- Boîtes à moustaches des scores en mathématiques par préparation aux tests

```
boxplot(math_score ~ `test.preparation.course`, data = data,
        col = c("lightblue", "lightcoral"),
        main = "Scores en mathématiques par préparation aux tests",
        xlab = "Préparation aux tests",
        ylab = "Score en mathématiques")
```



- Statistiques descriptives par préparation aux tests

```
by(data$math_score, data$`test.preparation.course`, summary)
```

```
## data$test.preparation.course: completed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.0    60.0    69.0    69.7    79.0   100.0
## ------------------------------------------------------------
## data$test.preparation.course: none
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   54.00   64.00   64.08   74.75  100.00
```

```
by(data$math_score, data$`test.preparation.course`, sd)
```

```
## data$test.preparation.course: completed
## [1] 14.4447
## ------------------------------------------------------------
## data$test.preparation.course: none
## [1] 15.19238
```

- Comparaison des moyennes par groupe (t-test)

```
t.test(math_score ~ gender, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  math_score by gender
## t = -5.398, df = 997.98, p-value = 8.421e-08
## alternative hypothesis: true difference in means between group female and group male is no
t equal to 0
## 95 percent confidence interval:
##  -6.947209 -3.242813
## sample estimates:
## mean in group female   mean in group male
##             63.63320             68.72822
```

```
t.test(math_score ~ `test.preparation.course`, data = data)
```

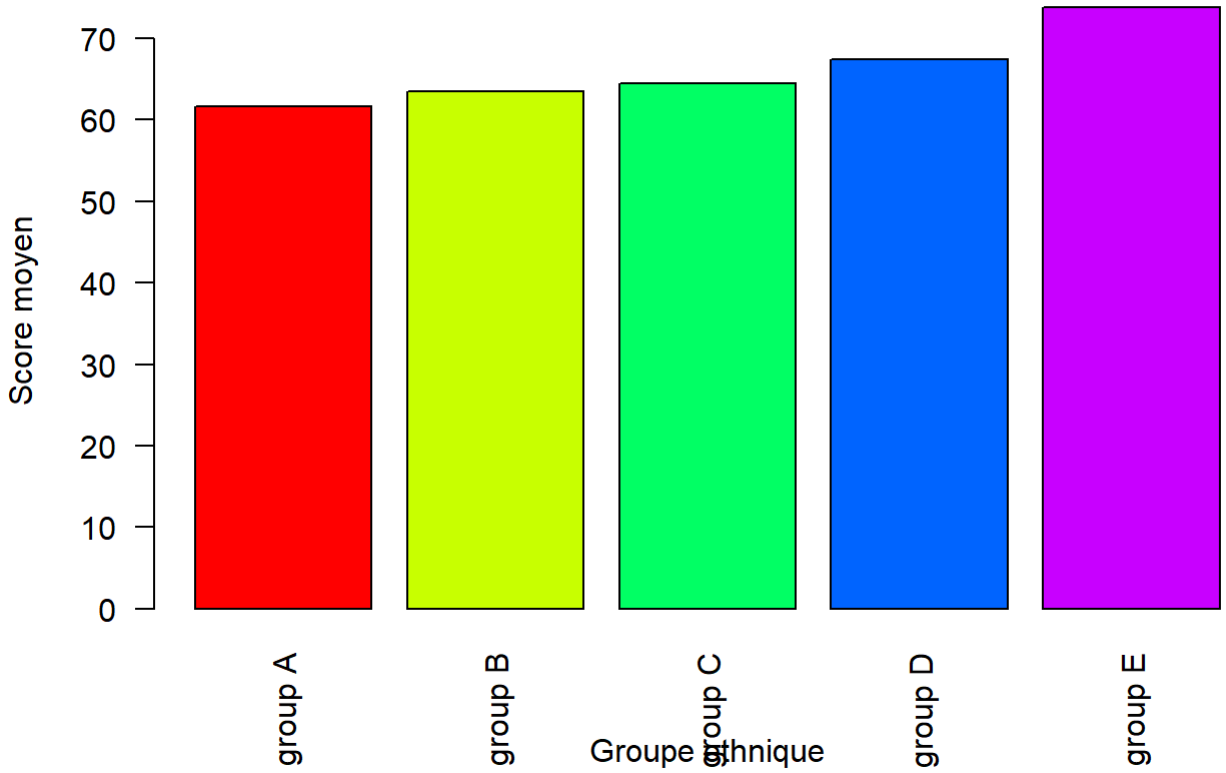
```
##
##  Welch Two Sample t-test
##
## data:  math_score by test.preparation.course
## t = 5.787, df = 770.08, p-value = 1.043e-08
## alternative hypothesis: true difference in means between group completed and group none is
not equal to 0
## 95 percent confidence interval:
##  3.712041 7.523257
## sample estimates:
## mean in group completed      mean in group none
##                69.69553                64.07788
```

- Visualisation des scores moyens par groupe ethnique

```
tapply(data$math_score, data$`race.ethnicity`, mean)
```

```
##  group A  group B  group C  group D  group E
## 61.62921 63.45263 64.46395 67.36260 73.82143
```

```
barplot(tapply(data$math_score, data$`race.ethnicity`, mean),
        col = rainbow(length(levels(factor(data$`race.ethnicity`)))),
        main = "Score moyen en mathématiques par groupe ethnique",
        xlab = "Groupe ethnique",
        ylab = "Score moyen",
        las = 2)
```



5. Analyse en Composantes Principales (ACP)

- Sélection des variables quantitatives

```
data_pca <- data[, c("math_score", "reading_score", "writing_score")]
```

- Réalisation de l'ACP

```
resultat <- PCA(data_pca, scale.unit = TRUE, graph = FALSE)
```

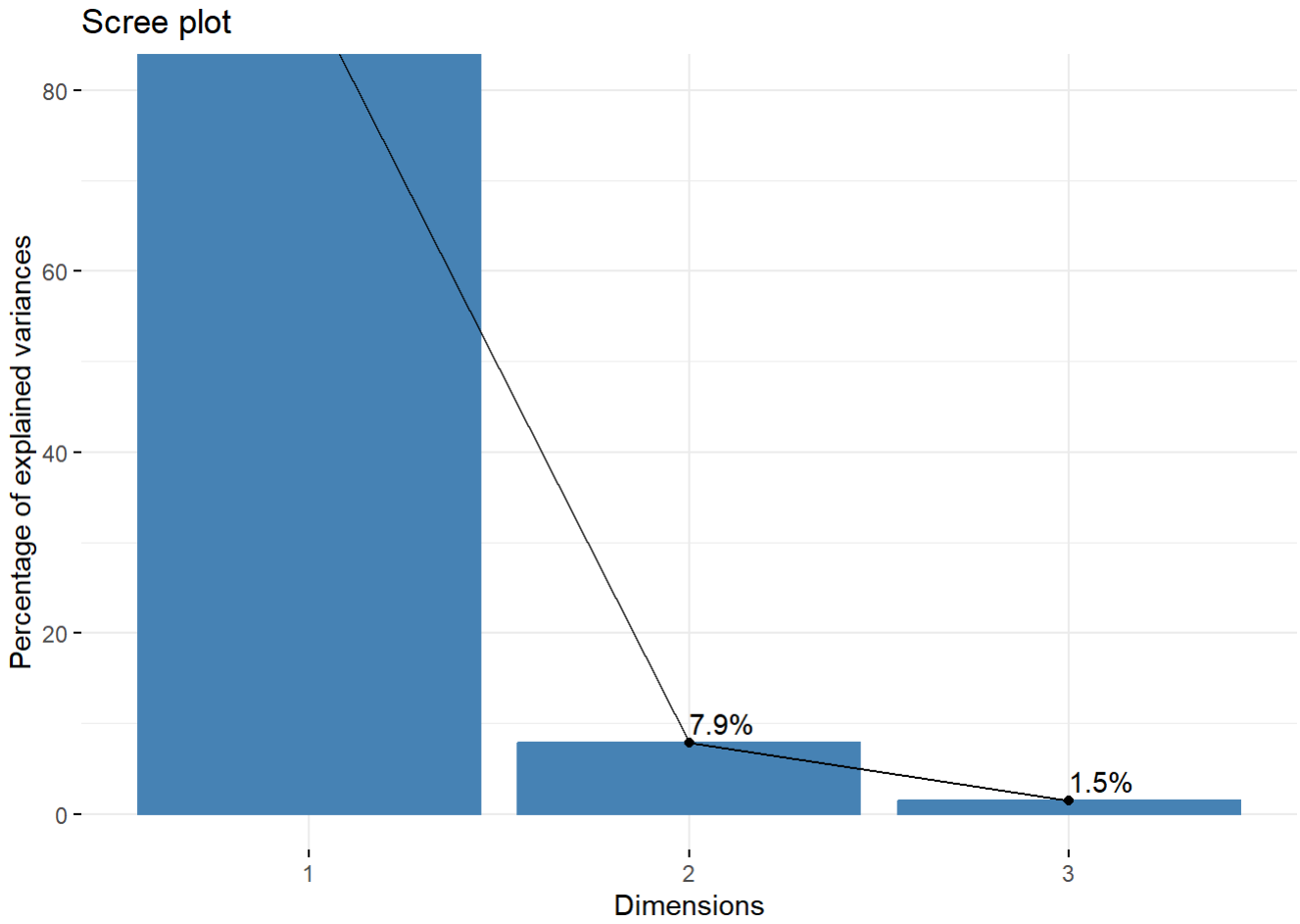
- Valeurs propres

```
eigenvalues <- get_eigenvalue(resultat)
eigenvalues
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 2.71842975        90.614325                    90.61433
## Dim.2 0.23658513         7.886171                    98.50050
## Dim.3 0.04498512         1.499504                   100.00000
```

- Visualisation du scree plot

```
fviz_eig(resultat, addlabels = TRUE, ylim = c(0, 80))
```



- Pourcentage de variance expliquée

```
cat("Pourcentage de variance expliquée par les composantes principales:\n")
```

```
## Pourcentage de variance expliquée par les composantes principales:
```

```
cat("- PC1:", round(eigenvalues[1, 2], 2), "%\n")
```

```
## - PC1: 90.61 %
```

```
cat("- PC2:", round(eigenvalues[2, 2], 2), "%\n")
```

```
## - PC2: 7.89 %
```

```
cat("- PC3:", round(eigenvalues[3, 2], 2), "%\n")
```

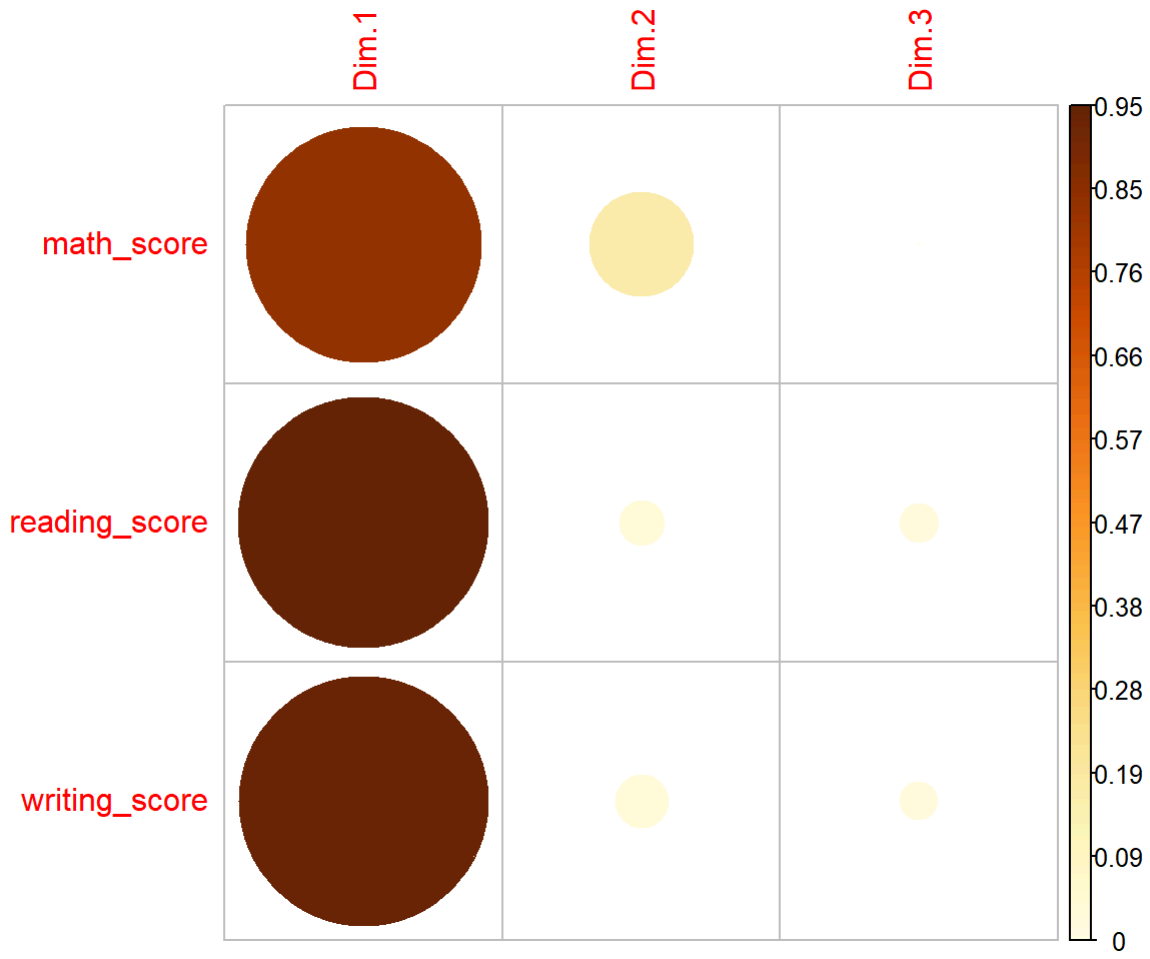
```
## - PC3: 1.5 %
```

```
cat("- Variance cumulée (PC1+PC2):", round(eigenvalues[2, 3], 2), "%\n")
```

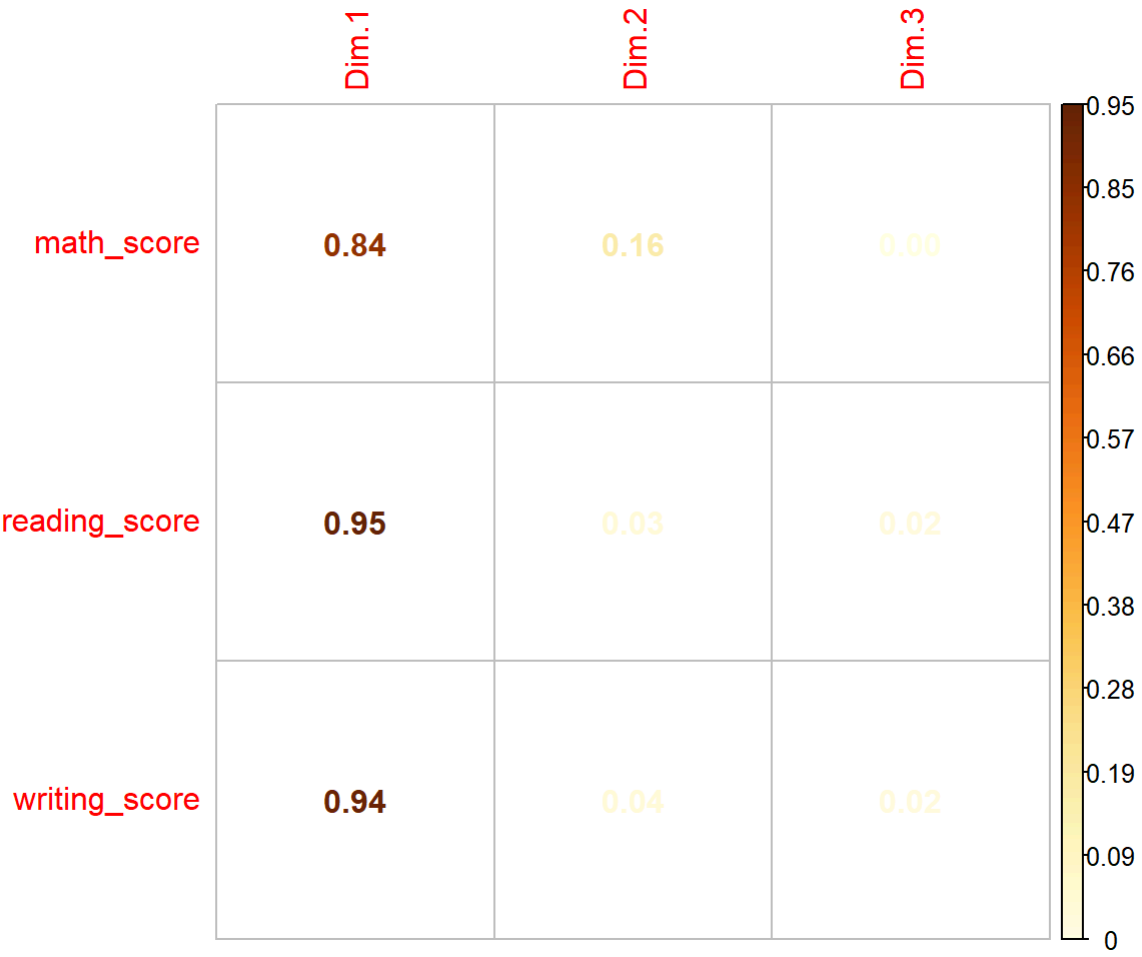
```
## - Variance cumulée (PC1+PC2): 98.5 %
```

- Graphique des variables

```
var <- get_pca_var(resultat)
corrplot(var$cos2, is.corr = FALSE)
```

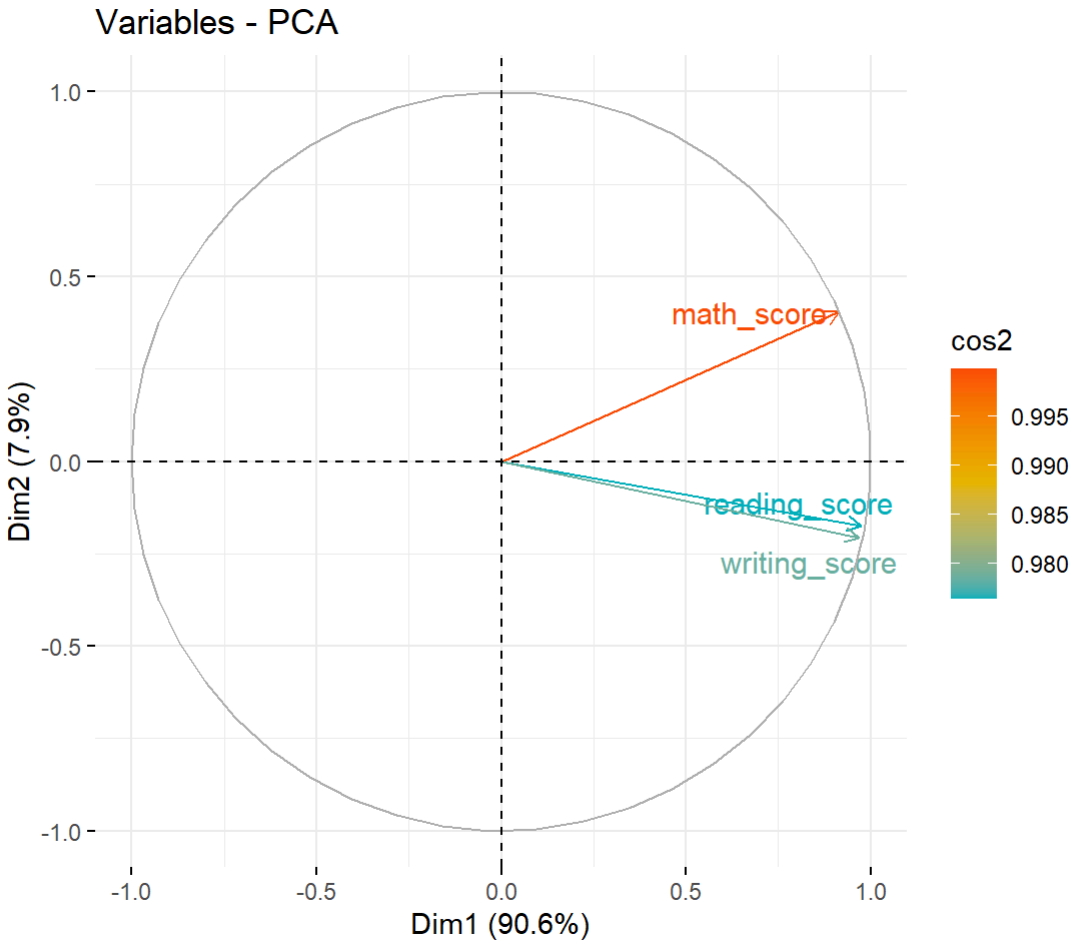



```
corrplot(var$cos2, is.corr = FALSE, method = "number")
```



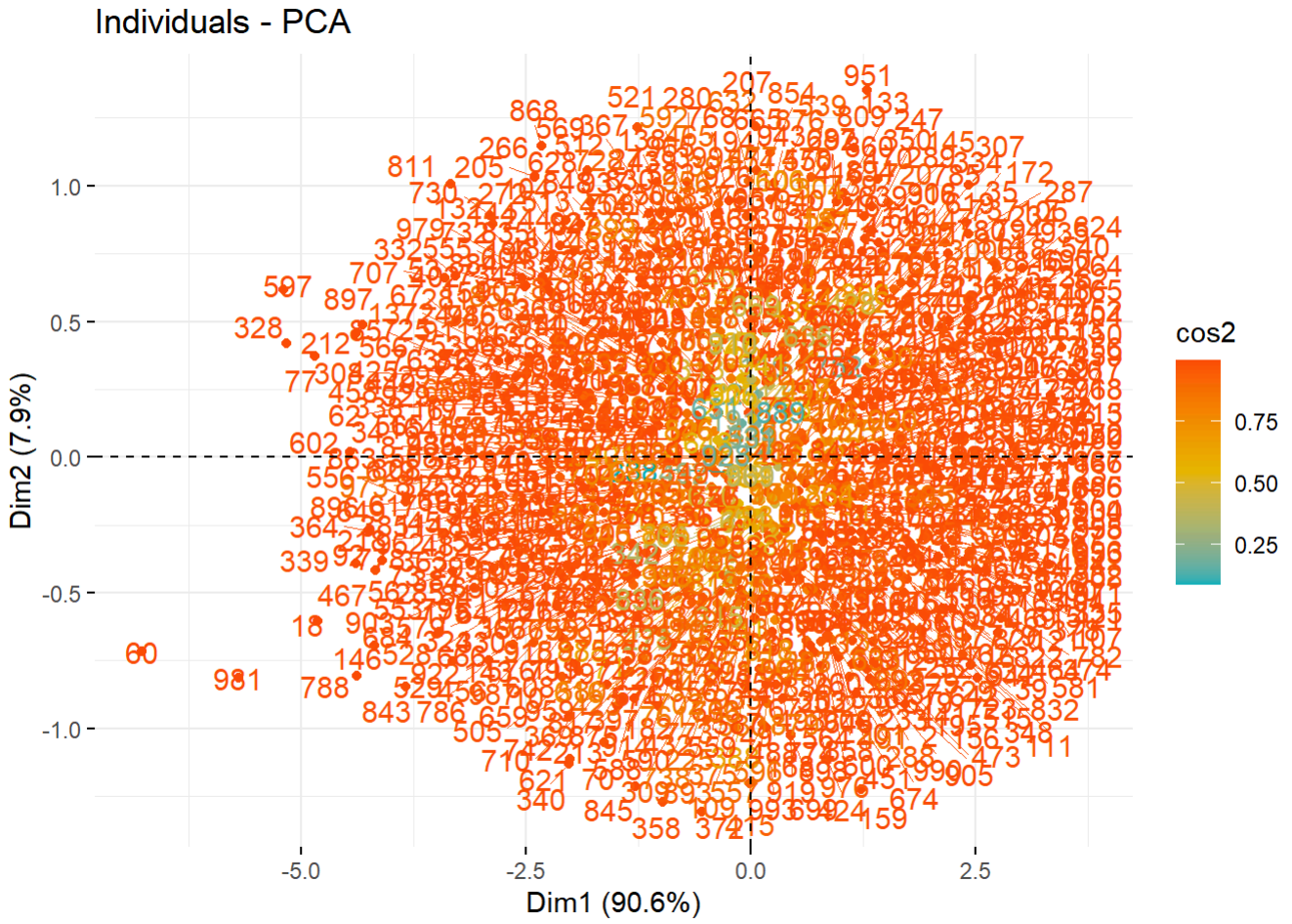
- Qualité de représentation des variables

```
fviz_pca_var(resultat, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



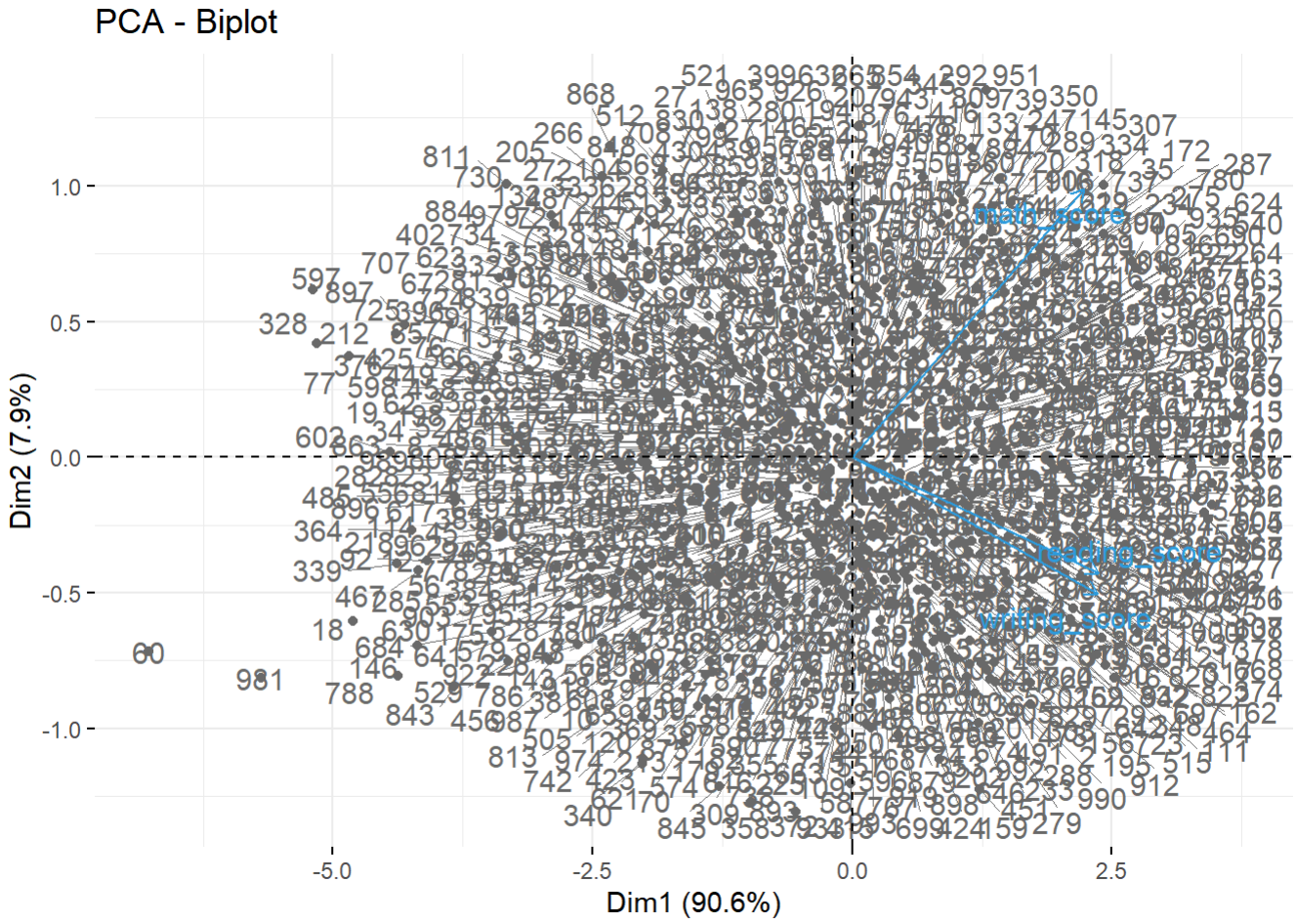
- Graphique des individus

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



- Biplot

```
fviz_pca_biplot(resultat,
                col.var = "#2E9FDF",
                col.ind = "#696969",
                repel = TRUE)
```



- Interprétation des axes

```
cat("\nInterprétation des axes principaux:\n")
```

```
##
## Interprétation des axes principaux:
```

```
cat("- PC1 (", round(eigenvalues[1, 2], 2), "% de variance): représente le niveau général de performance académique\n")
```

```
## - PC1 ( 90.61 % de variance): représente le niveau général de performance académique
```

```
cat("- PC2 (", round(eigenvalues[2, 2], 2), "% de variance): représente la différence entre les compétences mathématiques et linguistiques\n")
```

```
## - PC2 ( 7.89 % de variance): représente la différence entre les compétences mathématiques et linguistiques
```

6. Analyse Factorielle des Correspondances (AFC)

- Tableau de contingence pour l'AFC

```
table_race_edu <- table(data$`race.ethnicity`, data$`parental.level.of.education`)
```

- Réalisation de l'AFC

```
afc <- CA(table_race_edu, graph = FALSE)
```

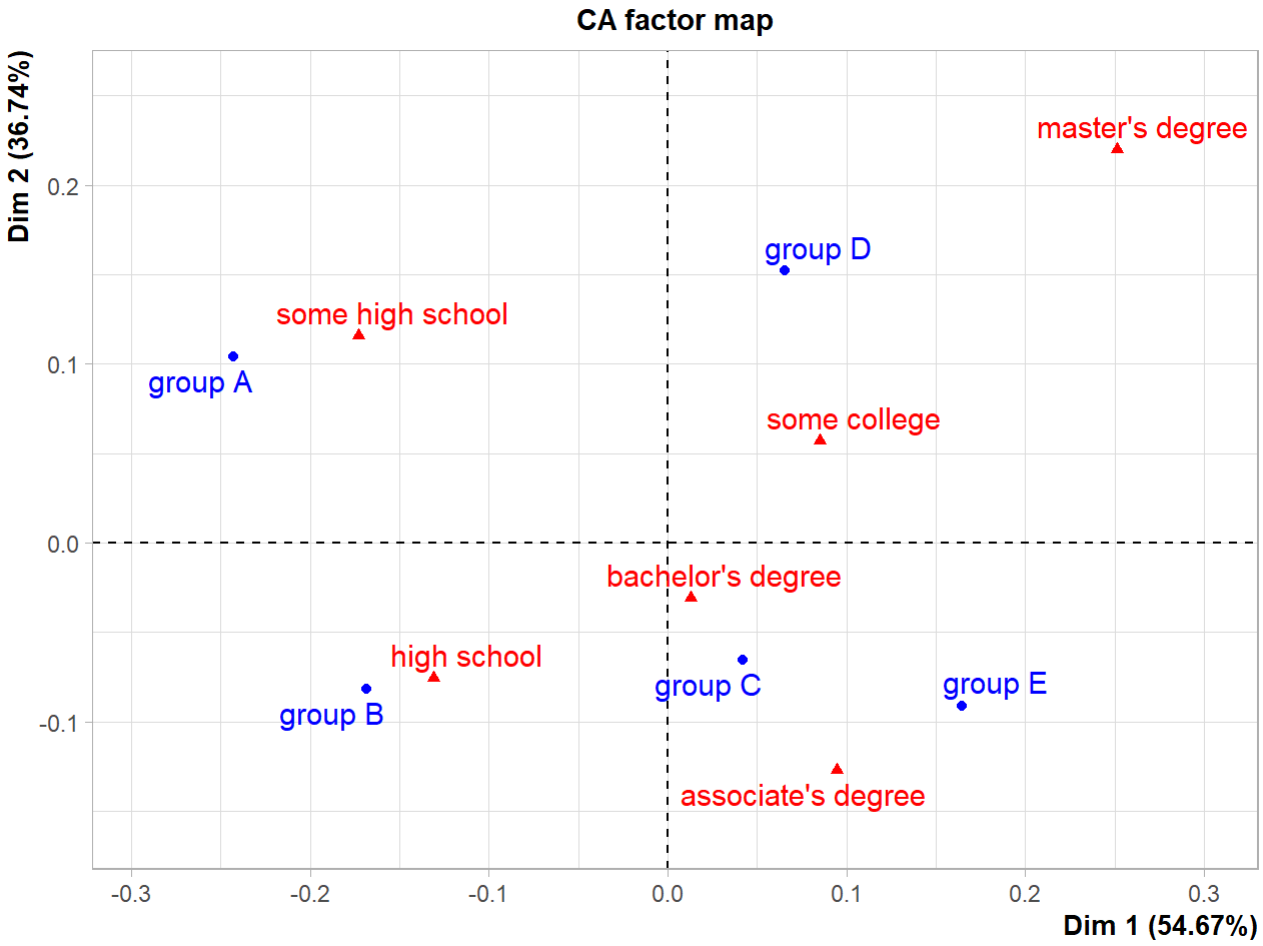
- Valeurs propres et pourcentage d'inertie

```
summary(afc)
```

```
##
## Call:
## CA(X = table_race_edu, graph = FALSE)
##
## The chi square of independence between the two variables is equal to 29.45866 (p-value =
0.07911305 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4
## Variance               0.016   0.011   0.002   0.000
## % of var.             54.668  36.742   6.954   1.637
## Cumulative % of var.  54.668  91.409  98.363 100.000
##
## Rows
##                     Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## group A            |     7.183 | -0.243 32.519  0.729 |  0.104  8.896  0.134 |
## group B            |     7.114 | -0.168 33.463  0.757 | -0.082 11.667  0.177 |
## group C            |     2.195 |  0.042  3.489  0.256 | -0.066 12.656  0.624 |
## group D            |     7.383 |  0.065  6.976  0.152 |  0.152 56.071  0.822 |
## group E            |     5.583 |  0.165 23.554  0.679 | -0.091 10.710  0.208 |
##                     Dim.3    ctr   cos2
## group A            -0.105 47.497  0.135 |
## group B             0.045 18.496  0.053 |
## group C             0.007  0.800  0.007 |
## group D             0.026  8.851  0.025 |
## group E            -0.060 24.356  0.089 |
##
## Columns
##                      Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## associate's degree |     5.604 |  0.095 12.344  0.355 | -0.127 33.157  0.640 |
## bachelor's degree  |     1.030 |  0.013  0.125  0.020 | -0.031  1.044  0.110 |
## high school        |     5.293 | -0.131 20.931  0.637 | -0.076 10.329  0.211 |
## master's degree    |     7.076 |  0.251 23.148  0.527 |  0.220 26.383  0.404 |
## some college       |     2.558 |  0.085 10.210  0.643 |  0.057  6.850  0.290 |
## some high school   |     7.898 | -0.173 33.240  0.678 |  0.116 22.236  0.305 |
##                     Dim.3    ctr   cos2
## associate's degree -0.009  0.808  0.003 |
## bachelor's degree  -0.077 34.421  0.684 |
## high school         0.064 38.714  0.150 |
## master's degree     0.082 19.178  0.056 |
## some college       -0.006  0.407  0.003 |
## some high school   -0.027  6.473  0.017 |
```

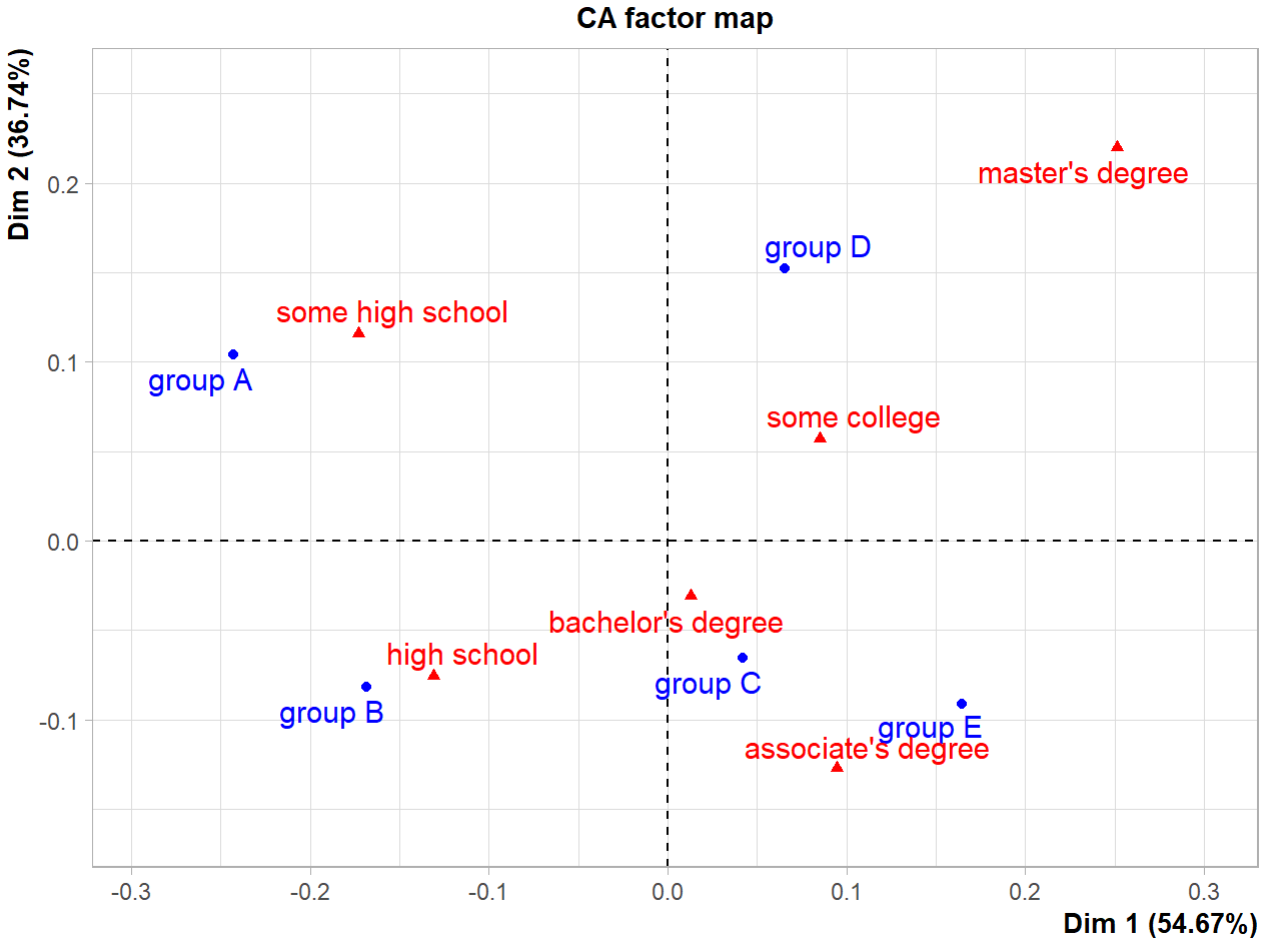
- Visualisation des résultats

```
plot(afc, main = "AFC: Groupe ethnique et niveau d'éducation des parents")
```



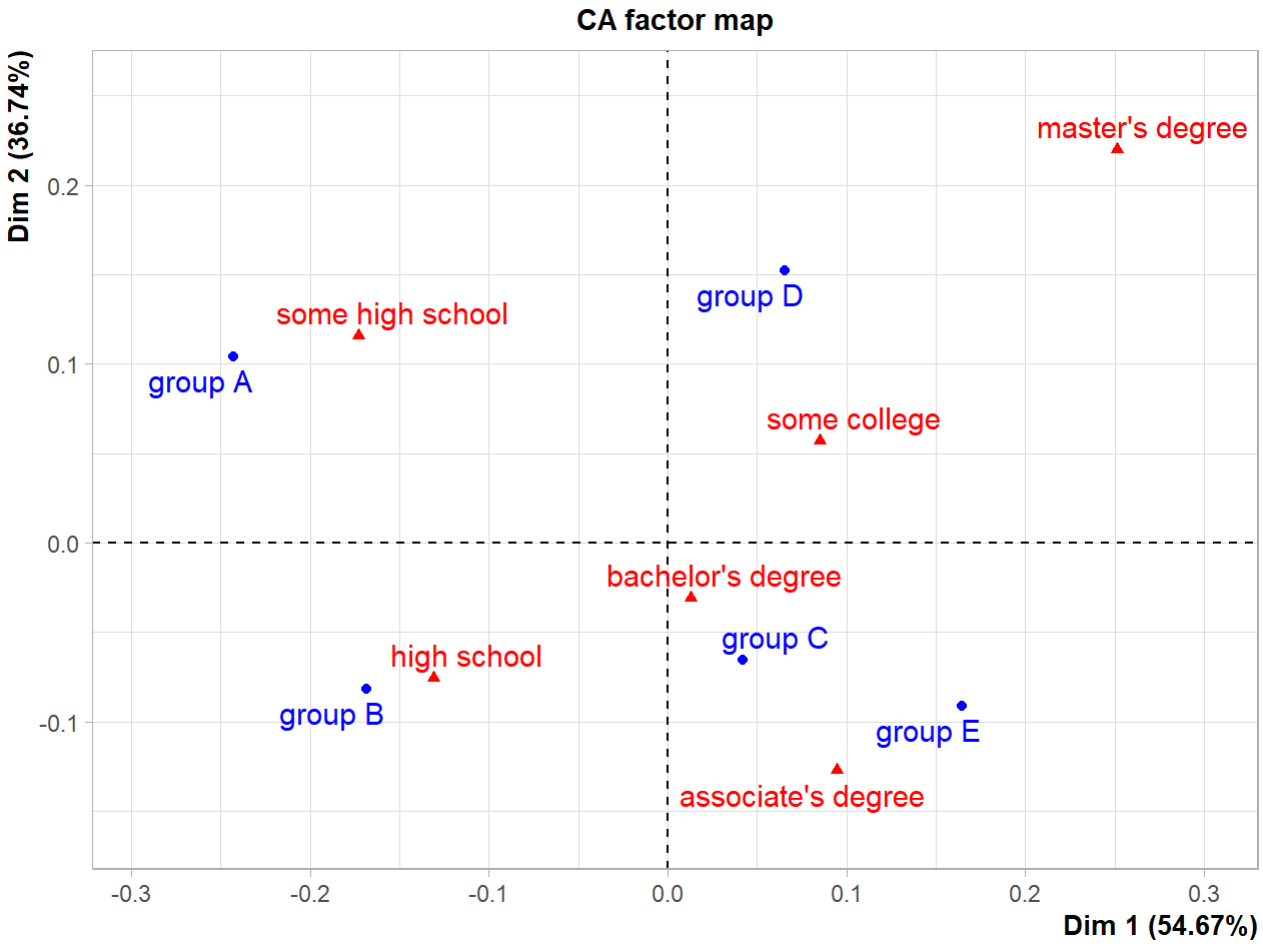
- Graphique des lignes (groupes ethniques)

```
plot(afc, axes = c(1, 2), col.row = "blue", col.col = "red",
     map = "rowprincipal", main = "AFC: Projection des groupes ethniques")
```



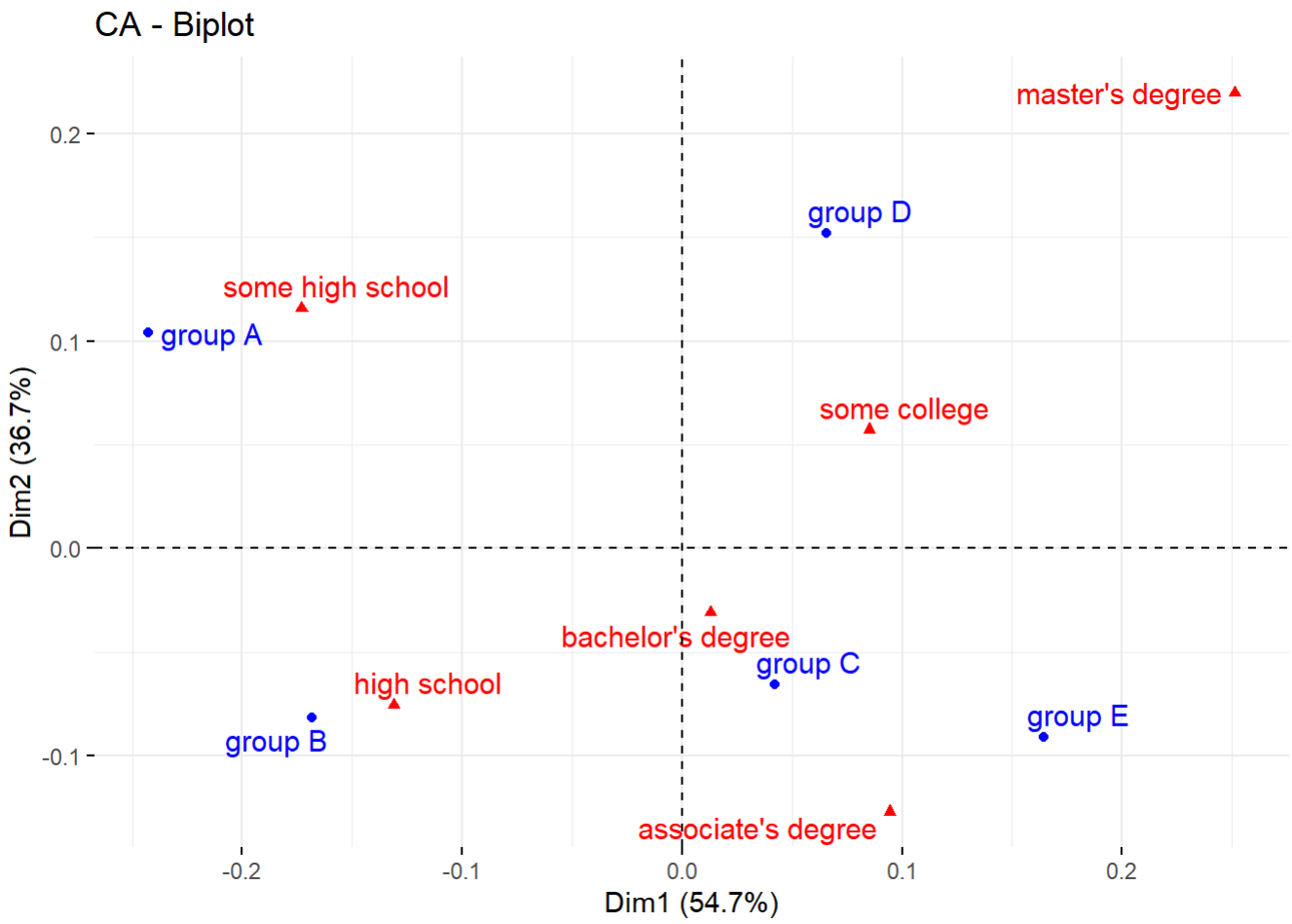
- Graphique des colonnes (niveaux d'éducation)

```
plot(afc, axes = c(1, 2), col.row = "blue", col.col = "red",
     map = "colprincipal", main = "AFC: Projection des niveaux d'éducation")
```



- Biplot

```
fviz_ca_biplot(afc, repel = TRUE)
```



7. Intervalles de confiance

- Intervalle de confiance pour la moyenne des scores en mathématiques

```
t_test_math <- t.test(data$math_score)
t_test_math
```

```
##
##  One Sample t-test
##
## data:  data$math_score
## t = 137.83, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  65.14806 67.02994
## sample estimates:
## mean of x
##    66.089
```

```
cat("Intervalle de confiance à 95% pour la moyenne des scores en mathématiques:\n")
```

```
## Intervalle de confiance à 95% pour la moyenne des scores en mathématiques:
```

```
cat("[", round(t_test_math$conf.int[1], 2), ", ", round(t_test_math$conf.int[2], 2), "]\n")
```

```
## [ 65.15 ,  67.03 ]
```

- Intervalle de confiance pour la moyenne des scores en lecture

```
t_test_reading <- t.test(data$reading_score)
t_test_reading
```

```
##
##  One Sample t-test
##
## data:  data$reading_score
## t = 149.81, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  68.26299 70.07501
## sample estimates:
## mean of x
##    69.169
```

```
cat("Intervalle de confiance à 95% pour la moyenne des scores en lecture:\n")
```

```
## Intervalle de confiance à 95% pour la moyenne des scores en lecture:
```

```
cat("[", round(t_test_reading$conf.int[1], 2), ", ", round(t_test_reading$conf.int[2], 2), "]
\n")
```

```
## [ 68.26 ,  70.08 ]
```

- Intervalle de confiance pour la moyenne des scores en écriture

```
t_test_writing <- t.test(data$writing_score)
t_test_writing
```

```
##
##  One Sample t-test
##
## data:  data$writing_score
## t = 141.62, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  67.11104 68.99696
## sample estimates:
## mean of x
##    68.054
```

```
cat("Intervalle de confiance à 95% pour la moyenne des scores en écriture:\n")
```

```
## Intervalle de confiance à 95% pour la moyenne des scores en écriture:
```

```
cat("[", round(t_test_writing$conf.int[1], 2), ", ", round(t_test_writing$conf.int[2], 2), "]
\n")
```

```
## [ 67.11 ,  69 ]
```

- Intervalle de confiance pour la différence des moyennes (math vs reading)

```
t_test_diff <- t.test(data$math_score, data$reading_score, paired = TRUE)
t_test_diff
```

```
##
##  Paired t-test
##
## data:  data$math_score and data$reading_score
## t = -10.816, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.638791 -2.521209
## sample estimates:
## mean difference
##           -3.08
```

```
cat("Intervalle de confiance à 95% pour la différence des moyennes (math - reading):\n")
```

```
## Intervalle de confiance à 95% pour la différence des moyennes (math - reading):
```

```
cat("[", round(t_test_diff$conf.int[1], 2), ", ", round(t_test_diff$conf.int[2], 2), "]\n")
```

```
## [ -3.64 ,  -2.52 ]
```

8. Conclusion

- Résumé de l'exploration des données

L'analyse du dataset StudentPerformance a permis d'explorer un ensemble de 1000 observations comportant 8 variables, dont 5 qualitatives (genre, groupe ethnique, niveau d'éducation des parents, type de repas, préparation aux tests) et 3 quantitatives (scores en mathématiques, lecture et écriture). Les données ne présentaient aucune valeur manquante, facilitant ainsi l'analyse.

- Analyses univariées

Les analyses univariées ont révélé des distributions relativement normales pour les trois scores académiques, avec des moyennes de 66,09 pour les mathématiques, 69,17 pour la lecture et 68,05 pour l'écriture. La répartition des variables qualitatives a montré une légère prédominance de femmes (51,8%), une plus forte représentation du groupe ethnique C (31,9%), et une majorité d'étudiants n'ayant pas suivi de préparation aux tests (64,2%). Les repas standard concernaient 64,5% des étudiants, suggérant une population majoritairement issue de milieux non défavorisés.

- Analyses bivariées

Les analyses bivariées ont mis en évidence des corrélations fortes entre les trois scores académiques, particulièrement entre la lecture et l'écriture ($r = 0,95$). La régression linéaire a montré que le score en mathématiques explique 66,8% de la variance du score en lecture, confirmant l'interdépendance des compétences académiques.

Les tests statistiques ont révélé des différences significatives de performance selon le genre ($p < 0,001$), avec des scores en mathématiques plus élevés chez les garçons, et selon la préparation aux tests ($p < 0,001$), avec un avantage marqué pour les étudiants ayant suivi une préparation. En revanche, aucune relation significative n'a été détectée entre le genre et la participation à la préparation aux tests ($p = 0,90$).

- Analyses multivariées

L'Analyse en Composantes Principales (ACP) a permis d'identifier une structure forte dans les données, avec une première composante expliquant 90,6% de la variance totale et représentant le niveau général de performance académique. La seconde composante (7,9% de variance) distingue les compétences mathématiques des compétences linguistiques.

L'Analyse Factorielle des Correspondances (AFC) a exploré les relations entre groupes ethniques et niveaux d'éducation parentale, révélant certaines associations, bien que le test du khi-deux n'ait pas atteint le seuil de significativité conventionnel ($p = 0,079$).

- Intervalles de confiance

Les intervalles de confiance à 95% ont permis de quantifier la précision des estimations des scores moyens : [65,15 - 67,03] pour les mathématiques, [68,26 - 70,08] pour la lecture et [67,11 - 69,00] pour l'écriture. La différence entre les scores de mathématiques et de lecture est significative, avec un intervalle de confiance de [-3,64 - -2,52].

- Implications générales

Cette analyse complète a démontré l'influence significative de facteurs socio-démographiques et éducatifs sur la performance académique des étudiants. Les résultats suggèrent que des interventions ciblées, notamment la préparation aux tests et le soutien aux étudiants issus de milieux défavorisés, pourraient contribuer à réduire les écarts de performance. L'interdépendance des compétences académiques souligne également l'importance d'une approche pédagogique intégrée, où le renforcement d'une compétence peut avoir des effets positifs sur les autres.

Les méthodes statistiques employées, allant des analyses descriptives simples aux techniques multivariées plus complexes, ont permis d'extraire des informations pertinentes et complémentaires, offrant une vision globale des facteurs influençant la réussite scolaire dans ce dataset.