Analyse des Performances Académiques des Élèves

1. Description du dataset

Le dataset **StudentsPerformance.csv** contient des données sur les résultats académiques de plusieurs élèves, ainsi que des informations démographiques et socio-économiques. Il permet d'analyser l'influence de facteurs comme le sexe, le niveau d'éducation des parents, le type de repas et la préparation aux examens sur les performances scolaires des élèves.

1.1. Présentation des variables

Le dataset **StudentsPerformance.csv** contient des informations sur les performances académiques des élèves, ainsi que des caractéristiques démographiques et socio-économiques. Voici une description des principales variables :

Nom de la variable	Signification	Type	Plage de valeurs
gender	Sexe de l'élève	Qualitative	male, female
parental.level.of.education	Niveau d'éducation des parents	Qualitative	some high school, high school, some college, associate's degree, bachelor's degree, master's degree
lunch	Type de repas de l'élève	Qualitative	standard, free/reduced
test.preparation.course	Participation à un programme de préparation	Qualitative	completed, none
math.score	Score obtenu en mathématiques	Quantitative	0 à 100
reading.score	Score obtenu en lecture	Quantitative	0 à 100

writing.score	Score obtenu en écriture	Quantitative	0 à 100

1.2. Analyse Descriptive

a). Statistiques descriptives des variables numériques

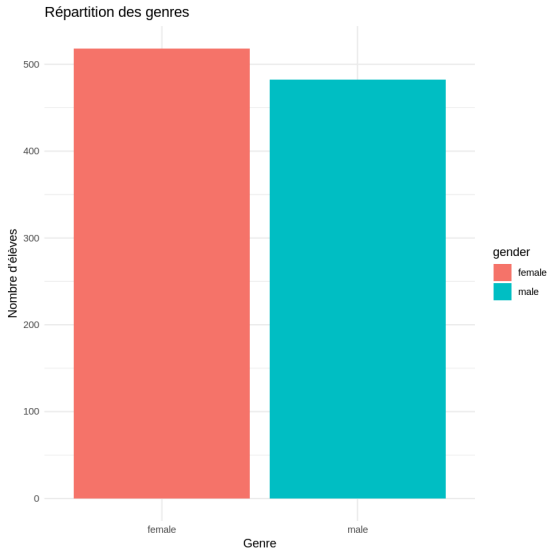
Les statistiques descriptives des variables numériques (math.score, reading.score, writing.score) sont résumées ci-dessous :

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
math.score	0.00	57.00	66.00	66.09	77.00	100.00
reading.score	17.00	59.00	70.00	69.17	79.00	100.00
writing.score	10.00	57.75	69.00	68.05	79.00	100.00

- **math.score :** Les scores varient de 0 à 100, indiquant une large dispersion des performances.De plus, la moyenne (66.09) et la médiane (66) sont proches, suggérant une distribution relativement équilibrée sans asymétrie majeure. Le premier quartile (57) et le troisième quartile (77) montrent que la majorité des élèves obtiennent des scores compris entre 57 et 77. Cependant, la présence d'un score minimum de 0 indique qu'il y a quelques élèves en grande difficulté.
- **reading.score :** Les scores s'étendent de 17 à 100, avec une moyenne de 69.17 et une médiane de 70. De plus, les quartiles montrent que la majorité des élèves ont des scores situés entre 59 et 79, un intervalle légèrement plus élevé que celui des mathématiques. Par rapport aux mathématiques, les résultats en lecture sont légèrement plus élevés, ce qui pourrait indiquer que la lecture est une compétence mieux maîtrisée par les élèves dans cet échantillon.
- **writing.score :** Les scores vont de 10 à 100, avec une moyenne de 68.05 et une médiane de 69. De plus, l'analyse des quartiles révèle que la majorité des élèves obtiennent entre 57.75 et 79, un schéma similaire à celui des résultats en lecture. Enfin, l'écart entre le minimum et le maximum est similaire à celui observé en lecture, ce qui indique que ces deux compétences sont étroitement liées et présentent des variations similaires entre les élèves.

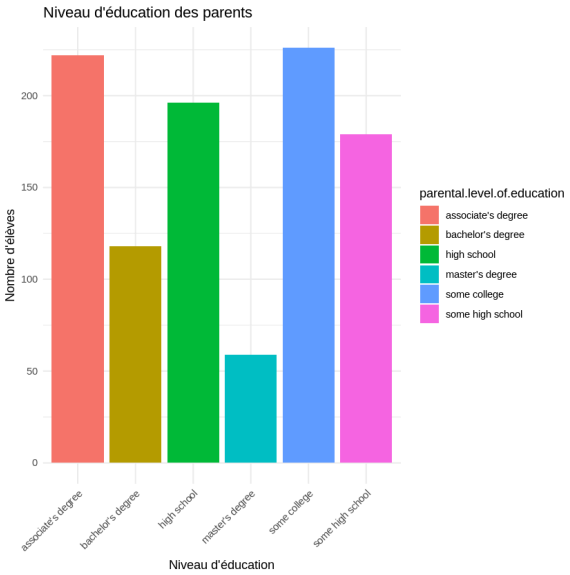
b. Répartition des variables qualitatives :

- **Répartition des genres (gender) :**



Le premier graphique représente la distribution des élèves selon leur genre. Il montre que le nombre d'élèves de sexe féminin est légèrement supérieur à celui des élèves de sexe masculin. En effet, environ 520 élèves sont de genre féminin, tandis qu'environ 480 élèves sont de genre masculin. Cette répartition relativement équilibrée suggère une présence équivalente des deux genres dans l'échantillon étudié.

- **Répartition du niveau d'éducation des parents (parental.level.of.education) :**

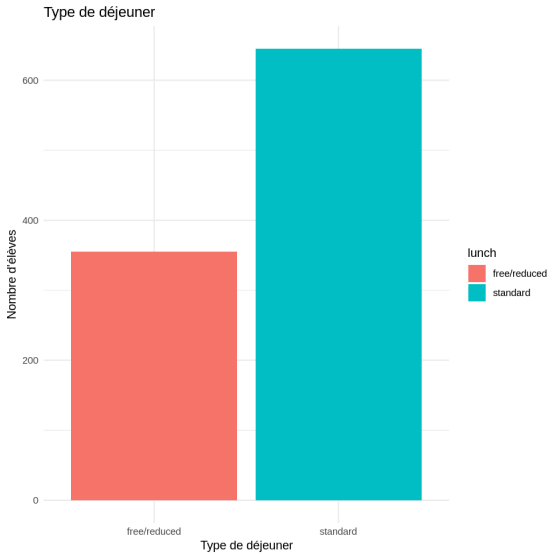


Ce premier graphique représente la répartition des niveaux d'éducation des parents en fonction du nombre d'élèves. On observe les tendances suivantes :

1. **Niveau d'éducation prédominant :** Les catégories "associate's degree" et "some college" comptent le plus grand nombre d'élèves, dépassant les 200 individus chacune. Cela suggère que ces niveaux d'éducation sont les plus répandus parmi les parents des élèves étudiés.
2. **Forte présence du niveau "high school" :** Le nombre d'élèves dont les parents ont atteint le niveau "high school" est également élevé, se situant juste en dessous de 200.
3. **Moins de parents titulaires d'un master :** La catégorie "master's degree" représente le plus faible effectif, avec un peu plus de 50 élèves, ce qui indique que peu de parents possèdent un diplôme de niveau master.
4. **Répartition équilibrée des autres niveaux :** Les niveaux "bachelor's degree" et "some high school" affichent des valeurs intermédiaires, avec respectivement environ 120 et 175 élèves.

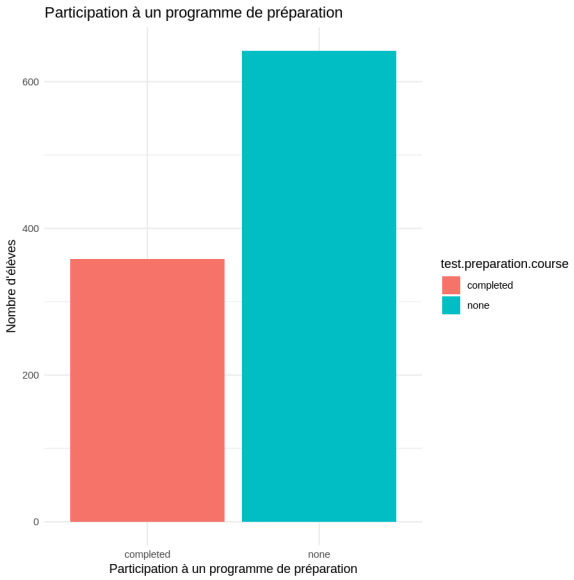
En conclusion, la majorité des parents ont un niveau d'éducation allant de "some high school" à "some college", tandis que les diplômes de niveau master restent relativement rares.

- **Répartition du type de repas de l'élève (lunch) :**



Le troisième graphique présente la répartition des élèves en fonction du type de déjeuner qu'ils reçoivent. Il apparaît qu'un nombre significatif d'élèves bénéficient d'un déjeuner standard (environ 650 élèves), tandis qu'une proportion moindre (environ 350 élèves) reçoit un déjeuner gratuit ou à tarif réduit. Cette répartition peut être un indicateur du niveau socio-économique des élèves, suggérant qu'une part importante d'entre eux n'ont pas besoin d'une aide alimentaire, tandis qu'une fraction non négligeable bénéficie d'un soutien financier pour leur repas.

- **Répartition de la participation à un programme de préparation (test.preparation.course) :**



Le dernier graphique illustre la répartition des élèves selon leur participation à un programme de préparation. On observe que la majorité des élèves (plus de 600) n'ont pas suivi de programme de préparation, tandis qu'environ 350 élèves ont complété un tel programme. Cette tendance suggère que la participation à des cours de préparation reste relativement faible par rapport au nombre total d'élèves,

2. Prétraitement des données

Avant d'effectuer des analyses plus approfondies, il est essentiel de préparer les données afin d'assurer leur qualité et leur compatibilité avec les méthodes d'analyse de données. Cette étape comprend la vérification des valeurs manquantes, la normalisation des variables quantitatives et la transformation des variables qualitatives en facteurs.

2.1. Vérification des données manquantes

La première étape du prétraitement consiste à identifier d'éventuelles valeurs manquantes dans le jeu de données. À cette fin, nous avons appliqué la fonction is.na() sur l'ensemble des observations, suivie d'une sommation pour déterminer le nombre total de valeurs manquantes. On a obtenu un résultat égal à 0, ce qui indique qu'aucune donnée manquante n'est présente dans notre dataset, ce qui simplifie le processus de préparation des données en évitant des techniques d'imputation ou d'exclusion.

2.2. Normalisation des Variables Quantitatives

Afin d'assurer une comparabilité entre les différentes variables numériques et d'éviter que certaines d'entre elles n'aient une influence disproportionnée sur les analyses futures, nous avons appliqué une normalisation (ou standardisation) aux trois variables quantitatives. La normalisation a été effectuée à l'aide de la fonction scale(). Après cette transformation, chaque variable possède une moyenne proche de 0 et un écart-type de 1.

2.3. Transformation des Variables Qualitatives en Facteurs

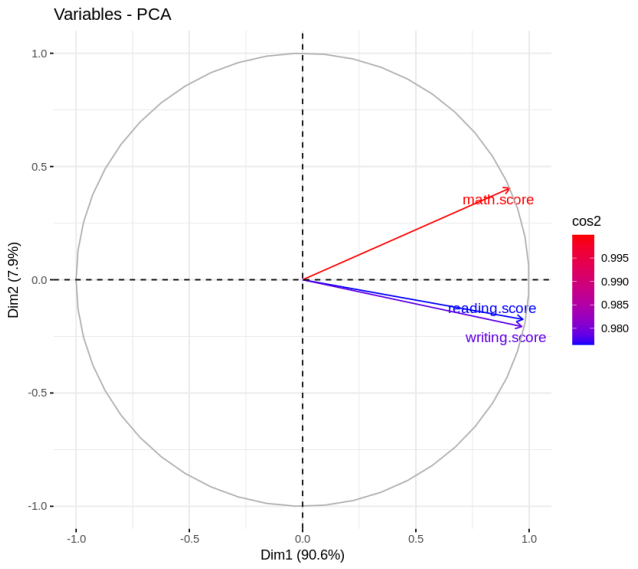
Les variables qualitatives sont converties en facteurs. Cette transformation garantit que ces variables seront correctement interprétées comme des catégories distinctes dans les analyses ultérieures.

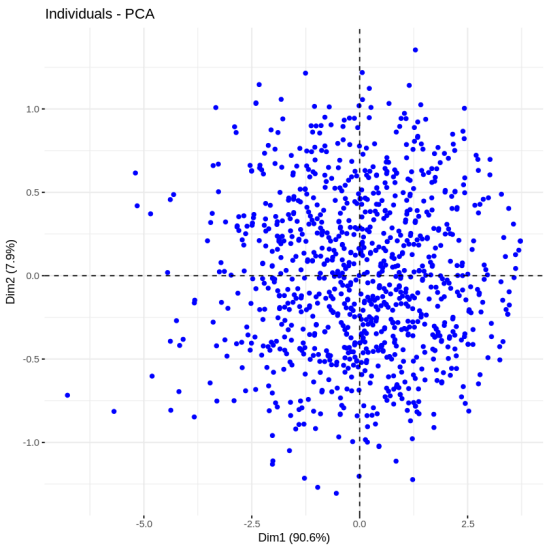
3. Analyse en Composantes Principales

3.1. Étapes suivies

Nous avons réalisé une Analyse en Composantes Principales (ACP) pour explorer les relations entre les variables numériques de notre dataset (math.score, reading.score, writing.score). L'ACP est une méthode de réduction de dimension qui permet de visualiser les variables dans un espace de dimension réduite tout en conservant l'essentiel de l'information. Voici les étapes que nous avons suivies :

- **Prétraitement des données :** Les variables numériques ont été standardisées (moyenne = 0, écart-type = 1) pour s'assurer qu'elles soient sur la même échelle.
- **Application de l'ACP :** Nous avons utilisé la fonction PCA() du package FactoMineR pour effectuer l'ACP. Les résultats de l'ACP incluent les valeurs propres, les contributions des variables aux axes, et les coordonnées des individus et des variables dans le plan factoriel.
- **Visualisation des résultats :** Nous avons utilisé le package factoextra pour visualiser les contributions des variables aux axes et les individus dans le plan factoriel.





3.2. Interprétation des résultats

a. Valeurs propres et variance expliquée

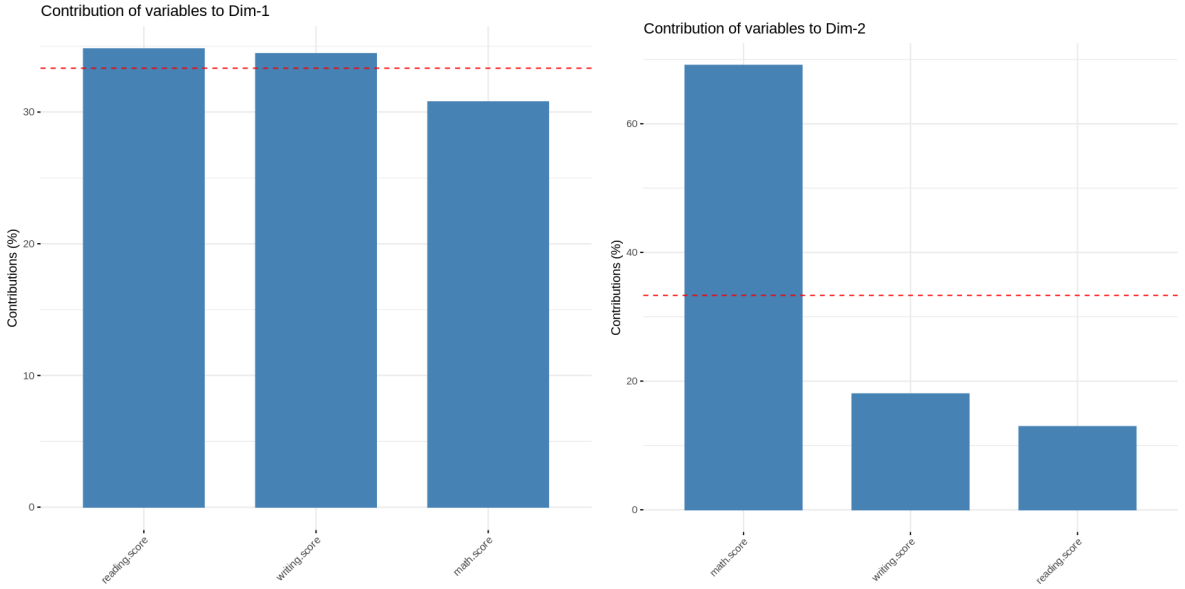
L'ACP a généré trois composantes principales (axes), mais nous nous concentrons sur les deux premières, qui expliquent la majeure partie de la variance :

- Axe 1 (Dim1) : Explique 90.6 % de la variance totale.
- Axe 2 (Dim2) : Explique 7.9 % de la variance totale.

Ensemble, ces deux axes expliquent 98.5 % de la variabilité totale des données, ce qui est très satisfaisant pour une analyse en deux dimensions.

b. Contributions des variables aux axes

Les contributions des variables aux axes nous permettent de comprendre quelles variables sont les plus importantes pour expliquer chaque axe.



Variable	Contributions à l'axe 1 (Dim1)
math.score	Contribution de 30.77 %
reading.score	Contribution de 34.80 %
writing.score	Contribution de 34.43 %

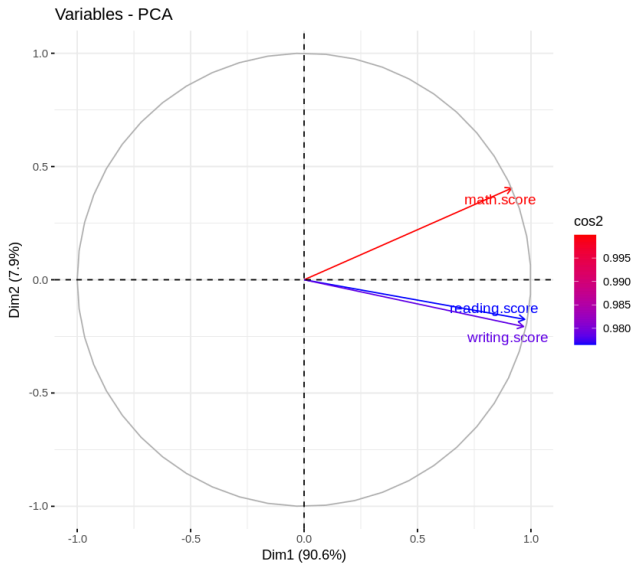
- **Interprétation :** L'axe 1 est principalement expliqué par les variables reading.score et writing.score. Cet axe représente une dimension liée aux compétences en écriture et lecture.
  - Les individus avec des coordonnées positives sur cet axe ont tendance à avoir plus de compétence en lecture et écriture, tandis que ceux avec des coordonnées négatives sont moins bons.

Variable	Contributions à l'axe 2 (Dim2)
math.score	Contribution de 69.07 %
reading.score	Contribution de 12.92 %
writing.score	Contribution de 18.01 %

- **Interprétation :** L'axe 2 est principalement expliqué par la variable math.score. Cet axe représente une dimension liée aux compétences en mathématiques.
  - Les individus avec des coordonnées positives sur cet axe ont tendance à avoir plus de compétences en mathématiques, tandis que ceux avec des coordonnées négatives sont moins bons.

c. Visualisation des variables dans le plan factoriel

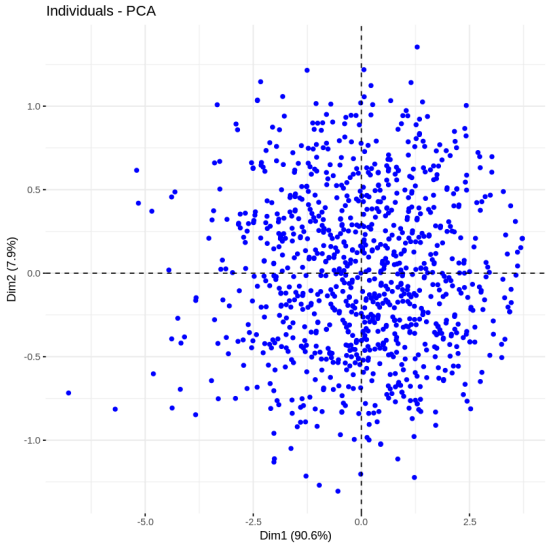
Dans le plan factoriel (Dim1 vs Dim2), les variables sont représentées par des flèches. Les variables proches les unes des autres sont fortement corrélées.



- reading.score et writing.score sont proches l'une de l'autre, ce qui confirme leur forte corrélation avec l'axe 1.

d. Visualisation des individus dans le plan factoriel

Les individus sont représentés par des points dans le plan factoriel. Les individus proches les uns des autres ont des profils similaires selon les variables.



- **Analyse des axes :**
  ➔ Axe 1 (Dim1 - 52.4%) : Représente une dimension liée aux compétences en lecture et en écriture.
  - Individus à droite (valeurs positives) : Ils ont de bonnes compétences en lecture et écriture.
  - Individus à gauche (valeurs négatives) : Ils ont de moins bonnes compétences dans ces matières.
  ➔ Axe 2 (Dim2 - 34.5%) : Représente une dimension liée aux compétences en mathématiques.
  - Individus en haut (valeurs positives) : Ils ont de bonnes compétences en mathématiques.
  - Individus en bas (valeurs négatives) : Ils ont de moins bonnes compétences en mathématiques.
- **Interprétation des groupes d'individus :**
  ➔ Individus dans le quadrant supérieur droit (positifs sur Dim1 et Dim2) : Excellents en mathématiques, lecture et écriture. Ce sont des élèves globalement performants.
  ➔ Individus dans le quadrant supérieur gauche (négatifs sur Dim1, positifs surDim2): Bons en mathématiques, mais faibles en lecture et écriture.
  ➔ Individus dans le quadrant inférieur droit (positifs sur Dim1, négatifs sur Dim2) : Bons en lecture et écriture, mais faibles en mathématiques.
  ➔ Individus dans le quadrant inférieur gauche (négatifs sur Dim1 et Dim2) :Faibles en lecture, écriture et mathématiques.
- **Distribution générale :**La majorité des individus semblent concentrés autour de l'origine, suggérant une population relativement homogène avec quelques variations en fonction des compétences spécifiques. Il y a aussi quelques individus éloignés du

centre, qui représentent probablement des cas extrêmes (très bons ou très faibles dans certaines compétences).

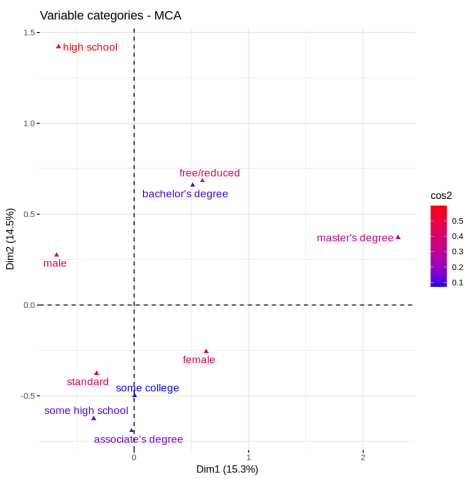
4. Analyse Factorielle des Correspondances Multiples (AFCM)

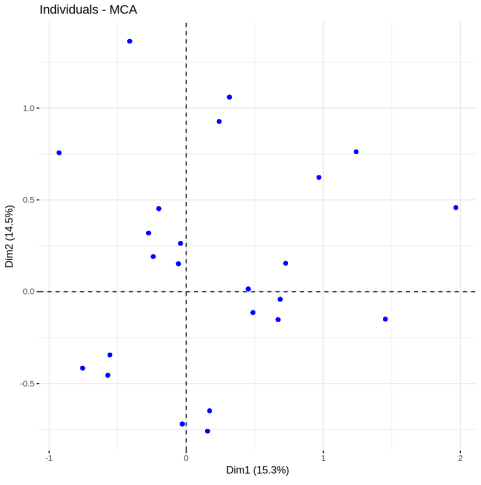
4.1. Étapes suivies

L'Analyse Factorielle des Correspondances Multiples (AFCM) a été réalisée pour explorer les relations entre les variables qualitatives de notre dataset : le genre, le niveau d'éducation des parents et le type de déjeuner.

L'AFCM permet de réduire la dimensionnalité des données et de visualiser les associations entre les différentes modalités. Les étapes suivies sont :

- **Prétraitement des données :** Factorisation des variables quantitatives et leur sélection.
- **Application de l'AFCM :** Utilisation de la fonction MCA() du package FactoMineR.
- **Visualisation des résultats :** À travers les valeurs propres, les contributions des variables aux axes, et la représentation des individus et modalités dans le plan factoriel.





4.2. Interprétation des résultats

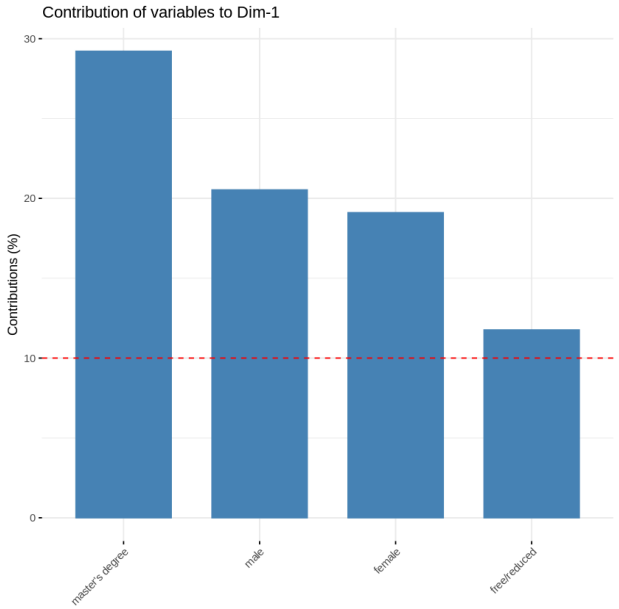
a. Valeurs propres et variance expliquée

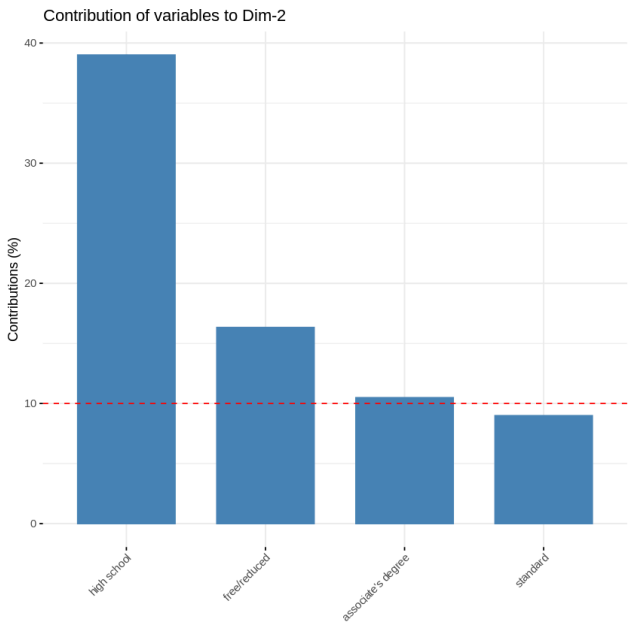
L'AFCM extrait plusieurs dimensions, mais nous nous concentrons sur les deux premières, qui expliquent la majeure partie de l'inertie (variance).

- Axe 1 (Dim1) : Explique 15.3 % de la variance totale.
- Axe 2 (Dim2) : Explique 14.5 % de la variance totale.

Ensemble, ces deux axes expliquent 29.8 % de la variabilité totale des données.

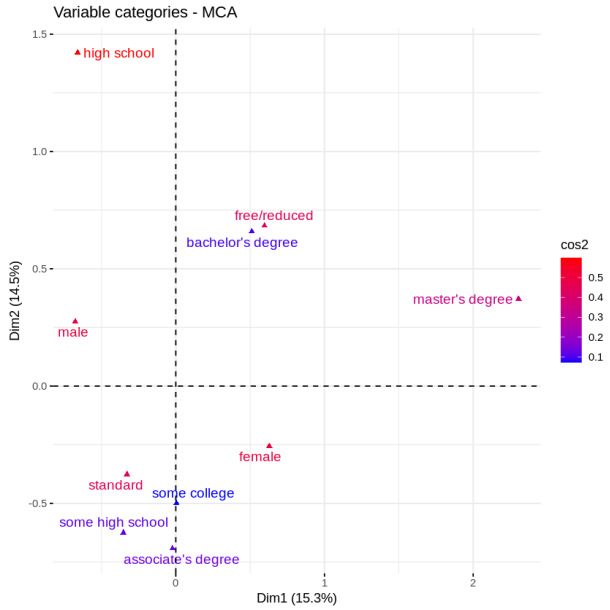
b. Contributions des variables aux axes





- **L'axe 1** est fortement influencé par le niveau d'éducation et le genre.
  - La modalité "Master's degree" contribue le plus à Dim1.
  - Le genre (male/female) joue aussi un rôle important dans cette première dimension.
  ➔ Et donc cet axe représente les étudiants ayant des parents qui ont un master et il oppose les étudiants hommes aux étudiant femmes.
- **L'axe 2** est fortement influencé par le type de déjeuner et le niveau d'éducation.
  -La modalité "High school" contribue fortement à Dim2 (39%), indiquant une distinction importante entre les individus dont les parents ont un niveau d'éducation plus faible.
  - Le type de déjeuner ("Free/reduced") contribue également à Dim2, suggérant une relation entre les conditions socio-économiques et cette dimension.
  - La modalité associate's degree contribue également à Dim2.
  ➔ Et donc cet axe oppose les étudiants qui ont des parents avec un niveau bac aux étudiants qui ont un diplôme d'associé. Il représente aussi les étudiants qui prennent le déjeuner gratuitement ou à un prix réduit.

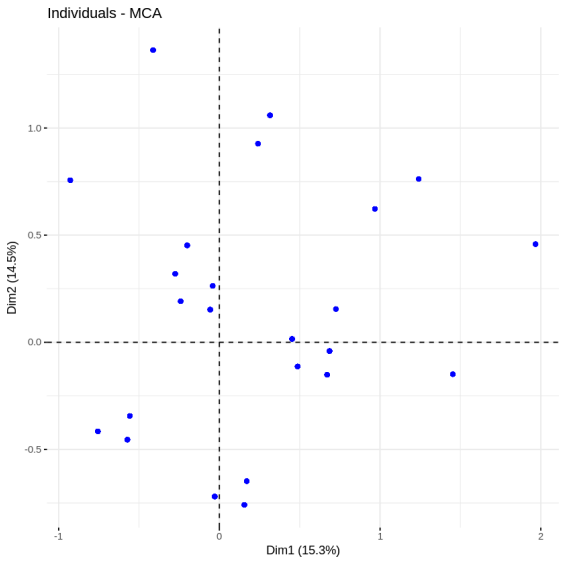
c. Visualisation des variables dans le plan factoriel



Dans le plan les modalités sont positionnées selon leur proximité :

- free/reduced et bachelor's degree sont très proche l'un de l'autre et donc les étudiants qui prennent un déjeuner gratuit ou à un prix réduit ont plus de chances d'avoir un parents avec un bachelor's degree
- some high school, some college et associate degree sont aussi proches l'une de l'autre ce qui veut dire que les étudiants qui ont des parents avec un associate's degree ou qui sont diplômés d'une université ou du lycée ont un comportement similaire. Par exemple, ces variables sont proches de la variable standard et donc ces étudiants ont plus de chance de prendre un petit déjeuner standard

d. Visualisation des individus dans le plan factoriel



-