# Introduction

Dans le contexte actuel, la rétention des talents est un défi majeur pour de nombreuses entreprises, en particulier dans les domaines compétitifs comme la science des données. Avec un marché du travail en constante évolution, il devient essentiel pour les organisations de mieux comprendre les profils et comportements des candidats, afin d'adapter leurs stratégies de recrutement et de rétention.

Ce travail vise à explorer et à identifier les profils distincts des candidats en science des données à travers une classification non supervisée. Plusieurs questions guideront cette analyse :

* Quels sont les facteurs les plus déterminants pour différencier les profils des candidats ?
* Comment les variables sociodémographiques (âge, statut matrimonial, etc.) et professionnelles (type d'emploi, heures travaillées, etc.) influencent-elles les regroupements naturels des candidats ?
* Peut-on identifier des groupes homogènes en fonction des heures de formation suivies ou de l'expérience professionnelle ?

Pour répondre à ces problématiques, nous utiliserons des méthodes variées de classification non supervisée, notamment :

* Une analyse en composantes principales (ACP) pour réduire la dimensionnalité et visualiser les relations sous-jacentes dans les données.
* K-means une méthode de clustering pour segmenter la population en groupes ayant des caractéristiques similaires.
* Une classification hiérarchique ascendante (CAH).

# Objectifs de l’étude

L'objectif de cette étude est de comprendre les dynamiques qui régissent les différents profils de candidats en science des données. En utilisant des techniques de classification non supervisée, nous chercherons à :

* Identifier les groupes naturels au sein des candidats.
* Décrire les caractéristiques principales de chaque groupe.
* Explorer les relations entre les variables.

Cela permettra de mieux appréhender les comportements des candidats et comprendre les raisons de leur changement ou non d’emploi.

# **Contexte et présentation du jeu de données**

Ce travail s'appuie sur deux jeux de données : l'un provenant d'une analyse RH axée sur le changement d'emploi des scientifiques des données, et l'autre issu d'une enquête de recensement sur les caractéristiques sociodémographiques et professionnelles. Nous avons combiné ces données pour fournir une vue d'ensemble enrichie des candidats en science des données, nous expliquerons dans la suite comment nous avons procédé pour la fusion.

Les professionnels en science des données constituent une population particulièrement intéressante à analyser en raison de la forte demande dans ce secteur. Leur comportement face aux opportunités de formation, aux conditions de travail et aux perspectives de carrière est essentiel pour comprendre leur mobilité.

# Présentation du jeu de données final

## Processus de fusion des données

Pour construire le jeu de données final, nous avons effectué les étapes suivantes :

1. **Conservation des colonnes pertinentes du premier jeu de données** : Nous avons conservé toutes les colonnes du premier jeu de données à l'exception de **enrolled\_id** qui est un identifiant sans valeur analytique directe.
2. **Sélection des variables du deuxième jeu de données** : Nous avons ajouté les variables suivantes depuis le deuxième jeu de données :
   * **Age** : Permet d'explorer l'influence de l'âge sur les groupes naturels.
   * **Workclass** : Fournit des informations sur le type d'emploi occupé par le candidat (public, privé, etc.).
   * **Marital\_status** : Permet d'examiner les relations entre le statut matrimonial et d'autres caractéristiques des candidats.
   * **Hours\_per\_week** : Aide à évaluer les habitudes de travail des candidats.
   * **Race** : Donne un contexte sociodémographique supplémentaire.

Ces variables ont été choisies pour enrichir les données initiales avec des dimensions supplémentaires pertinentes pour l'analyse des profils.

1. **Échantillonnage de la population** : Pour simplifier l'analyse tout en restant représentatif des données, nous avons réduit notre population à un échantillon de 200 individus. Pour cela, nous avons utilisé une **approche d'échantillonnage stratifié**.

Cette méthode consiste à diviser les données en **groupes appelés strates**, basés sur une variable importante (par exemple, le revenu). Ensuite, nous sélectionnons des individus dans chaque groupe de façon proportionnelle à leur représentation dans l’ensemble des données. Cela permet d’obtenir un échantillon équilibré et fidèle à la population initiale.

Par exemple, si 60 % des individus de la population ont un revenu faible et 40 % un revenu élevé, notre échantillon de 200 conservera ces proportions (120 avec un revenu faible et 80 avec un revenu élevé). Cela évite que certains groupes soient sur- ou sous-représentés, ce qui pourrait fausser nos résultats.

L'échantillonnage stratifié est utile car il garantit un échantillon représentatif et facilite l’analyse avec une taille de données réduite mais équilibrée.

## Justification des choix

* **Suppression de la colonne enrolled\_id** : étant un identifiant unique, elle n'apporte aucune information exploitable dans notre analyse statistique.
* Nous ne supprimons pas la colonne **Target**, mais elle n’est pas utilisée dans notre analyse ; elle est utilisée uniquement pour la construction de notre échantillon représentatif du premier jeu de données. Elle n'est pas pertinente pour une classification non supervisée, car elle représente une variable cible utilisée dans les modèles supervisés.
* **Ajout des variables du deuxième jeu de données** : Ces variables complètent les dimensions sociodémographiques et professionnelles présentes dans le premier jeu de données, offrant une perspective plus riche pour identifier les groupes.
* **Limitation à 200 lignes** : Cela permet de simplifier les calculs et les visualisations, tout en conservant une diversité suffisante dans les données pour effectuer des analyses valides.

|  |  |  |
| --- | --- | --- |
| **Nom de la variable** | **Type** | **Description** |
| city | Catégorielle | Code de la ville du candidat. |
| city\_development\_index | Numérique | Indice de développement de la ville (échelle numérique). |
| gender | Catégorielle | Sexe du candidat. |
| relevent\_experience | Catégorielle | Expérience pertinente dans le domaine visé. |
| enrolled\_university | Catégorielle | Type de cours universitaire suivi, le cas échéant. |
| education\_level | Catégorielle | Niveau d'éducation du candidat. |
| major\_discipline | Catégorielle | Discipline majeure de l'éducation du candidat. |
| experience | Catégorielle | Expérience totale en années du candidat. |
| company\_size | Catégorielle | Taille de l'entreprise actuelle (nombre d'employés). |
| company\_type | Catégorielle | Type d'entreprise actuelle (privée, publique, etc.). |
| last\_new\_job | Catégorielle | Nombre d'années depuis le dernier changement d'emploi. |
| training\_hours | Numérique | Nombre d'heures de formation suivies par le candidat. |
| workclass | Catégorielle | Secteur d'emploi actuel (privé, public, etc.). |
| age | Numérique | Âge du candidat. |
| marital\_status | Catégorielle | État matrimonial du candidat. |
| hours\_per\_week | Numérique | Nombre moyen d'heures travaillées par semaine. |
| race | Catégorielle | Origine ethnique du candidat. |

Tableau 1: Description des variables du jeu de données.

Ce jeu de données présente une composition mixte avec des variables numériques et catégorielles, offrant ainsi une opportunité pour l’application des techniques d’analyse précédemment citées.

# 2. Pré-traitement des données

Le prétraitement est une étape cruciale dans l'analyse de données, car il permet de :

* **Nettoyer les données :** Corriger les erreurs, les valeurs manquantes, les incohérences.
* **Transformer les données :** Normaliser les données, créer de nouvelles variables, réduire la dimensionnalité.
* Présentation et choix des variables (actives et illustratives)

# Distribution des classes

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

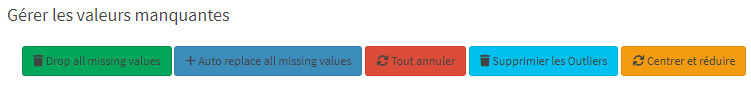
Figure 1: Distribution des classes.

Ce graphique représente la distribution des individus. Nous avons compté le nombre d’occurrences de chaque classe par rapport à la variable « target » ; on observe :

* **Distribution inégale :** La classe 0 est beaucoup plus fréquente que la classe 1. En effet, on compte 150 observations pour la classe 0 contre seulement 50 pour la classe 1. Cela signifie que la majorité des candidats dans notre dataset appartiennent à la classe 0(qui signifie que), ce qui est cohérant avec notre échantillonnage stratifié expliqué plus haut.

## Préparation des données

### Gérer les valeurs manquantes



Pour éviter que les valeurs manquantes influencent négativement l'analyse, nous les avons traitées de manière adaptée pour chaque type de variable.

Dans notre travail nous donnons la possibilité d’appliquer les traitements suivants :

* Supprimer toutes les valeurs manquantes : Elle consiste à supprimer toutes les lignes contenant des valeurs manquantes, elle est simple et rapide à mettre en œuvre, cependant elle peut entrainer une perte significative d’informations. Vu la taille de notre jeu de données, cette approche n’est pas pertinente pour notre analyse.

Avant : Nous avons initialement 200 entrées (individus dans le jeu de données).

Figure 2 : Jeu de données avec tous les individus.

Après : Nous avons 98 individus, 102 ont été supprimés.



Figure 3 : Jeu de données après suppression des valeurs manquantes.

* Remplacer automatiquement toutes les valeurs manquantes : Concernant les variables numériques, nous avons remplacé les valeurs manquantes pour la moyenne de la colonne correspondante. Pour les variables catégorielles, nous avons remplacé chaque valeur manquante par la modalité la plus fréquente de la colonne.

Avant : Nous avons les valeurs manquantes pour gender et enrolled\_university.

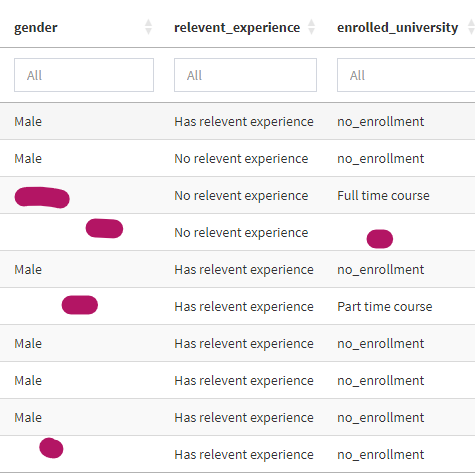


Figure 4 : Jeu de données avec valeurs manquantes sur le gender.

Cela étant : on constate que les valeurs manquantes sont remplacées par la valeur la plus fréquente.

Une image contenant texte, capture d’écran, nombre, Police

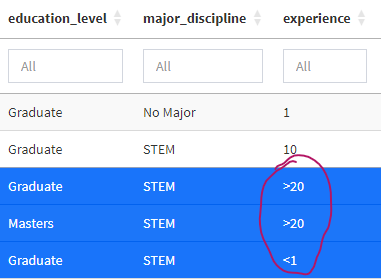
Description générée automatiquement

On voit bien que les données manquantes ont été remplacées par les modalités les plus fréquentes.

Figure 5 : Jeu de données après remplacement automatique.

* Tout annuler nous permet simplement d’annuler les traitements faits pour revenir sur le jeu de données à son état initial.
* Supprimer les Outliers : Nous avons exploité l’amplitude interquartile IQR=Q3−Q1, pour identifier et supprimer les outliers dans chaque colonne numérique. Les valeurs qui tombent en dehors de l'intervalle [𝑄1−1.5×IQR, 𝑄3+1.5×IQR] sont considérées comme des outliers et sont supprimées. Il est à noter que ces outliers sont présentés dans l’analyse univariée grâce aux boxplots.
* Centrer et réduire : cette action est utilisée dans le prétraitement pour standardiser les variables et soustrayant la moyenne et en divisant par l’écart-type.

##### Cas particulier Variable expérience



La variable expérience contenait des valeurs particulières, comme "<1", et ">20", qui ne sont pas directement utilisables pour des analyses numériques. Nous avons donc appliqué une transformation pour convertir "<1" en 0 et ">20" en 20, afin de les rendre cohérentes avec les autres valeurs numériques.

# 3. Analyse descriptive

## Analyse univariée

### Statistique des variables quantitatives

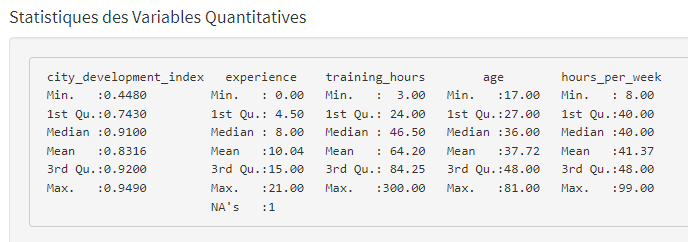


Figure 6 : Statistique descriptive des variables quantitatives.

Ce tableau présente la statistique descriptive de cinq variables quantitatives, ces statistiques permettent de résumer la distribution de chaque variable et d’avoir une première idée de leur comportement ;

Les statistiques présentées sont :

* **Minimum :** La plus petite valeur observée.
* **1er Quartile (Q1) :** Sépare les 25% de données les plus petites des 75% restantes.
* **Médiane :** Valeur centrale lorsque les données sont ordonnées. Sépare les 50% inférieurs des 50% supérieurs.
* **Moyenne :** Somme de toutes les valeurs divisées par le nombre de valeurs.
* **3ème Quartile (Q3) :** Sépare les 75% de données les plus petites des 25% restantes.
* **Maximum :** La plus grande valeur observée.
* **NA's:** Nombre de valeurs manquantes.

1. **City Development Index** :
   * Min : 0.448 ; Max : 0.949 ; Moyenne : 0.8316 ; Médiane : 0.91
   * La distribution semble légèrement biaisée vers des valeurs plus élevées, car la moyenne (0.8316) est proche du 3ème quartile (0.92) et de la médiane (0.91). Cela indique que la majorité des candidats provient de villes avec un index de développement urbain élevé, entre 0.794 et 0.92.
2. **Experience** :
   * Min : 0 années ; Max : 21 années ; Moyenne : 10.04 années ; Médiane : 8 années
   * L’expérience professionnelle varie de 0 à 21 années, la moyenne est légèrement supérieure à la médiane, ce qui suggère une asymétrie positive d’une part, et d’autre part la présence de quelques candidats avec une longue expérience (jusqu'à 21 ans) qui tirent la moyenne vers le haut.
3. **Training Hours** :
   * Min : 3 heures ; Max : 300 heures ; Moyenne : 64.20 heures ; Médiane : 46.50 heures.
   * Le nombre d’heures varie de 3 à 300 heures ; La distribution des heures de formation semble large, avec un écart considérable entre les candidats qui ont peu d'heures (1ère quartile à 24 heures) et ceux qui ont un grand nombre (jusqu'à 300 heures). La moyenne plus élevée que la médiane suggère également un biais, possiblement dû à des candidats qui accumulent de nombreuses heures de formation.
4. **Age**

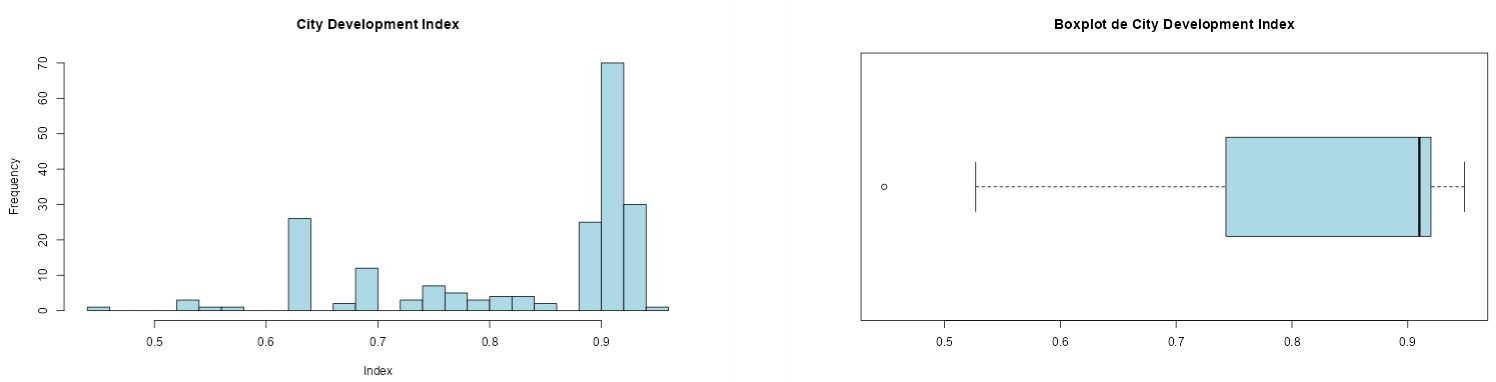
* Min : 17 ans, Max : 81ans, Médiane : 36 ans, Moyenne : 37.72 ans, Médiane : 36 ans
* L’age des individus varie de 17 à 81 ans.
* La distribution semble assez asymétrique, avec une médiane proche de la moyenne.

1. **Hours\_per\_week**

* Min : 8 heures, Max : 99 heures, Moyenne : 41.37 heures, Médiane : 40 heures.
* Le nombre d'heures travaillées par semaine varie de 8 à 99 heures.
* La distribution semble également assez symétrique, avec une médiane proche de la moyenne.

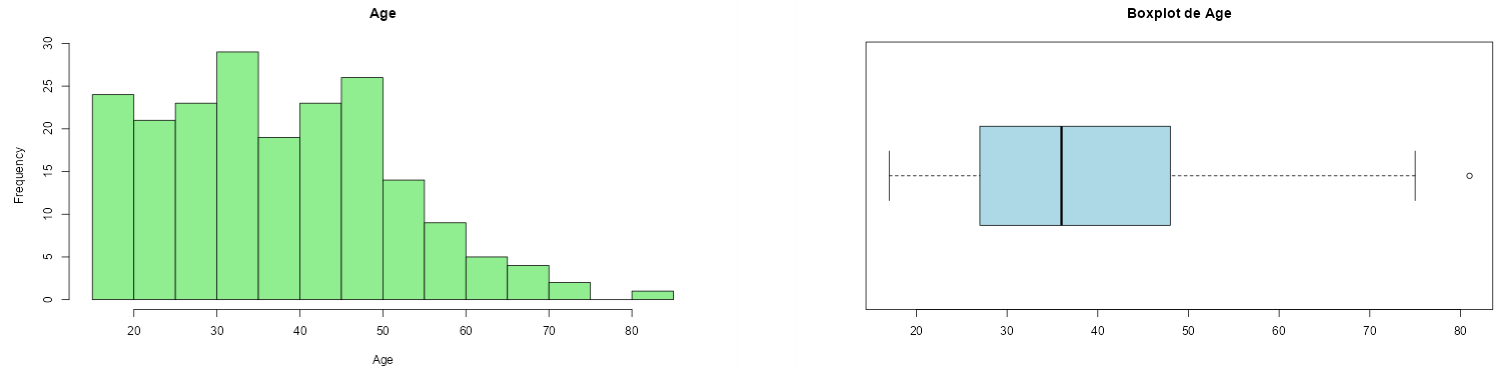
Ces statistiques montrent des distributions variées dans les données, et une tendance globale pour chaque variable qui pourrait influencer les décisions d'embauche ou de formation.

#### City Development Index



* **En analysant l’histogramme on observe :**
* **Asymétrie à droite :** La distribution est fortement penchée vers la droite. Cela signifie qu’un grand nombre de villes ont un indice de développement élevé, proche de 0,9, tandis que peu de villes ont un indice faible.
* **Mode autour de 0,9 :** Le pic de la distribution se situe autour de 0,9, indiquant que c'est la valeur de l'indice de développement la plus fréquente donc, la majorité des candidats sont issus d'une ville développée
* **Quelques valeurs aberrantes à gauche :** On observe quelques villes avec un indice de développement très faible, autour de 0,5. Ces villes pourraient être considérées comme des outliers.
* **L’analyse du Boxplot nous donne les observations clés suivantes :**
* **Asymétrie à droite :** La boîte est plus courte à gauche de la médiane qu'à droite. Cela indique une asymétrie positive, c'est-à-dire que la distribution est plus étirée vers les valeurs élevées (indice de développement élevé).
* **Médiane élevée :** La médiane est située vers la partie supérieure de la boîte, ce qui suggère que la plupart des villes ont un indice de développement élevé.
* **Peu d'outliers:** Il semble y avoir peu d'outliers, c'est-à-dire de valeurs très éloignées des autres comme on peut le voir aussi sur l'histogramme tout à gauche puis, sur la boîte à moustaches tout à gauche également.
* **Homogénéité relative :** La longueur de la boîte indique une certaine homogénéité dans les valeurs de l'indice de développement pour la majorité des villes.
* **Peu de disparités :** Les valeurs de l'indice de développement sont relativement concentrées autour de la médiane, ce qui suggère une faible dispersion des valeurs.
* **Quelques villes moins développées :** La présence de la moustache inférieure indique qu'il existe quelques villes avec un indice de développement plus faible, mais elles sont en minorité.

#### ****Age****



* **Cet histogramme nous donne une vue d'ensemble de la distribution des âges dans l'échantillon. On peut observer ceci :**
* **Mode autour de 30 ans :** La fréquence est maximale pour les individus âgés d'environ 30 ans. Cela suggère que ce groupe d'âge est le plus représenté dans l'échantillon.
* **Distribution asymétrique à droite :** La distribution est plus étirée vers les âges plus élevés. Cela signifie qu'il y a moins d'individus âgés que d'individus jeunes.
* **Répartition assez large :** Les âges sont répartis sur une large plage, allant d'environ 20 à 80 ans.
* **Groupe jeune :** L'échantillon est principalement composé d'individus jeunes et d'âge moyen.
* **Vieillissement de la population :** La diminution progressive de la fréquence avec l'âge suggère un vieillissement de la population.
* L’analyse de la boite à moustache nous permet de dire :
* **Médiane autour de 37 ans :** La ligne centrale de la boîte, représentant la médiane, est située autour de 37 ans. Cela signifie que la moitié des individus ont moins de 37 ans et l'autre moitié en a plus.
* **Intervalle interquartile relativement étroit :** La boîte est assez courte, ce qui indique que la moitié centrale des données (entre Q1 et Q3) est concentrée sur une plage d'âge relativement restreinte.
* **Asymétrie légère à droite :** La moustache supérieure est légèrement plus longue que la moustache inférieure, suggérant une légère asymétrie positive. Cela signifie qu'il y a quelques individus plus âgés que la moyenne, mais moins d'individus beaucoup plus jeunes.
* **Un outlier :** Le point isolé à droite de la moustache supérieure représente un individu dont l'âge est considéré comme une valeur aberrante (outlier), étant nettement supérieur à la majorité des autres individus.
* **Résumé :** L'âge moyen de la population étudiée semble être autour de 40 ans.
* **Homogénéité :** La majorité des individus sont concentrés autour de l'âge moyen, avec une dispersion relativement faible.  
  **Quelques individus plus âgés :** Il y a quelques personnes plus âgées, mais elles représentent une minorité.

#### Expérience

Une image contenant diagramme, Tracé, Plan, ligne

Description générée automatiquement

Une image contenant capture d’écran, Rectangle, diagramme, ligne

Description générée automatiquement

* Analyse de l’histogramme de l’expérience

La distribution des années d'expérience semble être bimodale, c'est-à-dire qu'il y a deux pics principaux.  
- **Premier pic** : Se situe autour de 4 à 6 ans d'expérience, avec des fréquences allant jusqu'à environ 20 individus. Cela pourrait indiquer un groupe de candidats relativement nouveaux (recrutement important) dans leur domaine, potentiellement des jeunes professionnels ou des personnes ayant récemment changé de carrière.

- **Second pic** : Très marqué à 20 ans d'expérience, avec une fréquence d'environ 35 individus. Ce pic pourrait être dû à une limite supérieure arbitraire dans la collecte des données ou à un effet de plafond où les individus ayant plus de 20 ans d'expérience sont tous regroupés dans cette catégorie. Cela pourrait représenter des candidats très expérimentés ou des seniors dans leur domaine.

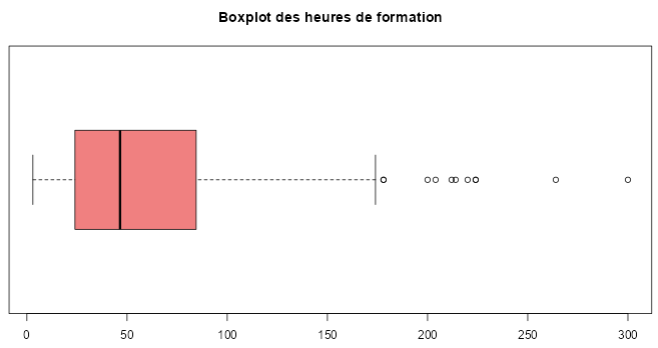
* Analyse du Boxplot de l’expérience :

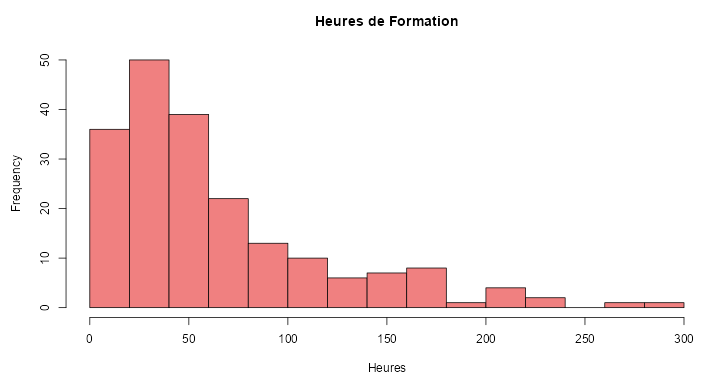
- **Distribution centrale** : La majorité des individus ont entre 5 et 15 ans d'expérience. Cela montre que la plupart des candidats ont une expérience modérée.

- **Symétrie** : Le boxplot semble relativement symétrique autour de la médiane, indiquant une distribution assez équilibrée des années d'expérience. Cela suggère qu'il n'y a pas de biais significatif vers des expériences très faibles ou très élevées.  
- **Étendue** : Les moustaches s'étendent de 0 à environ 20 ans, montrant que l'échantillon couvre une large gamme d'expériences professionnelles. Cela indique une diversité dans les niveaux d'expérience parmi les candidats.

* Interprétation
  + La médiane à environ 8 ans suggère que près de la moitié des individus ont 8 ans d'expérience ou moins, et l'autre moitié a plus de 8 ans d'expérience.
  + L'absence de points en dehors des moustaches suggère que les données sont relativement homogènes, sans expériences extrêmement faibles ou élevées par rapport au reste de l'échantillon. Cela renforce l'idée d'une distribution équilibrée des années d'expérience.

#### Heures de Formation





* **Histogramme** : l’analyse nous permet de voir :
  + **Forme de la distribution** : La distribution est asymétrique à droite (ou positivement asymétrique), ce qui signifie que la majorité des employés ont suivi un nombre relativement faible d'heures de formation, tandis que quelques employés ont suivi un nombre beaucoup plus élevé d'heures de formation.
  + **Pic de la distribution** : Le pic de la distribution se situe autour de 20 à 40 heures de formation, avec une fréquence maximale d'environ 50 employés.
  + **Concentration des faibles heures de formation** : Il y a une concentration notable d'employés ayant suivi entre 0 et 60 heures de formation. Cela pourrait indiquer que la plupart des employés ont suivi une formation de base ou minimale.
  + **Fréquence décroissante** : À mesure que le nombre d'heures de formation augmente, la fréquence des employés diminue progressivement. Cela montre que moins d'employés ont suivi des formations plus longues.
  + **Heures de formation élevées** : Un petit nombre d'employés ont suivi plus de 200 heures de formation, ce qui pourrait représenter des formations spécialisées ou avancées.
  + **Interprétation**: La distribution asymétrique suggère une diversité des besoins de formation parmi les employés, avec une majorité ayant des besoins de formation modérés et une minorité ayant des besoins de formation plus importants.
* **Boxplot**
* **Médiane** : La médiane des heures de formation se situe autour de 50 heures. Cela signifie que 50% des employés ont suivi 50 heures de formation ou moins, et les 50% restants ont suivi plus de 50 heures de formation.
* **Quartiles** : La boîte elle-même représente l'intervalle interquartile (IQR), c'est-à-dire la plage entre le premier quartile (25e percentile) et le troisième quartile (75e percentile). Ici, l'IQR va d'environ 25 à 75 heures. Cela indique que la moitié centrale des données se situe dans cette plage.
* **Moustaches** : Les lignes horizontales (moustaches) s'étendent des quartiles aux valeurs minimales et maximales, à l'exclusion des valeurs aberrantes. Elles montrent la dispersion des données. Dans ce cas, les moustaches s'étendent de 0 à environ 150 heures.
* **Valeurs aberrantes** : Les points individuels en dehors des moustaches représentent les valeurs aberrantes. Ici, il y a plusieurs valeurs aberrantes au-delà de 150 heures, allant jusqu'à environ 300 heures.
* Interprétation : Les valeurs aberrantes au-delà de 150 heures suggèrent que certains employés ont suivi des formations beaucoup plus longues et potentiellement spécialisées. Cela pourrait indiquer des besoins spécifiques ou des opportunités de développement professionnel pour certains rôles.

#### 4. ****Conclusion partielle****

Ces analyses descriptives offrent un premier aperçu de la population de notre jeu de données, en mettant en lumière des aspects tels que la répartition géographique (city development), l'expérience professionnelle et l'investissement en formation. Les distributions asymétriques et les valeurs extrêmes observées pour l’expérience et les heures de formation méritent une attention particulière pour de futures analyses, notamment pour les modèles de prédiction, car elles pourraient influencer les résultats. Pour une exploration plus approfondie et détaillée de ces données, nous vous invitons à consulter l'application Shiny, où des visualisations interactives et des analyses complémentaires sont disponibles.

### Variables qualitatives

Une image contenant texte, Police, ligne, capture d’écran

Description générée automatiquement

Ici nous allons fournir une interprétation détaillée des statistiques descriptives pour les variables qualitatives.

#### 1. ****Gender****

Les données ne montrent pas la répartition exacte entre les genres, mais cette variable est importante pour évaluer la diversité.

**Interprétation** : Si un genre est surreprésenté, cela pourrait révéler un biais potentiel dans le domaine de la data science au sein de l'échantillon. Une analyse plus fine permettrait de voir si le genre influe sur d’autres variables comme l’expérience ou le niveau d’éducation.

#### 2. ****Relevant Experience****

Cette variable identifie les candidats ayant une expérience professionnelle pertinente dans leur domaine.

**Interprétation** : Une grande proportion de candidats avec une expérience pertinente suggérerait un niveau de préparation élevé. Inversement, une majorité sans expérience pertinente pourrait indiquer un bassin de candidats moins expérimentés mais potentiellement ouverts aux opportunités de formation et aux emplois débutants.

#### 3. ****Enrolled University****

Elle indique si les candidats sont actuellement inscrits dans une université, s’ils sont diplômés, ou s’ils n’ont jamais été inscrits.

**Interprétation** : Une forte proportion d’étudiants inscrits ou de diplômés récents pourrait suggérer que le domaine de la data science attire des talents encore en formation. Cela peut aussi refléter une tendance vers l'obtention de diplômes avancés pour entrer dans ce secteur.

#### 4. ****Education Level****

Cette variable représente le niveau d’éducation atteint (baccalauréat, master, etc.).

**Interprétation** : Si la majorité des candidats ont des diplômes de niveau master ou plus, cela pourrait indiquer un marché exigeant pour les postes de data science, privilégiant une haute qualification académique. Si un niveau plus faible est prédominant, cela pourrait suggérer un marché plus accessible aux profils moins qualifiés ou une demande élevée pour des rôles moins spécialisés.

#### 5. ****Major Discipline****

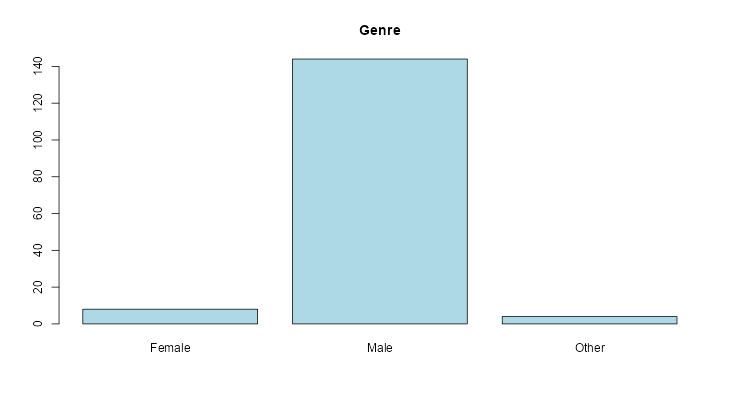
* Cette variable indique la discipline principale des candidats (ex : STEM, commerce, arts).
* **Interprétation** : Une majorité de candidats issus des domaines STEM (sciences, technologies, ingénierie, mathématiques) serait typique en data science, car ces domaines fournissent des compétences quantitatives. Une présence notable d'autres disciplines, comme le commerce, pourrait indiquer une diversité d’approches et de perspectives dans le domaine.

#### Conclusion partielle

Ces variables qualitatives offrent une vue d'ensemble des caractéristiques de l'échantillon de candidats. Elles permettront de comprendre l'impact de différents facteurs comme l'éducation, l'expérience, et le type d'entreprises précédentes sur l’intention des candidats de changer de poste. Les statistiques descriptives des variables qualitatives posent les bases pour explorer des relations plus complexes entre les caractéristiques des candidats et leur comportement vis-à-vis du marché du travail.

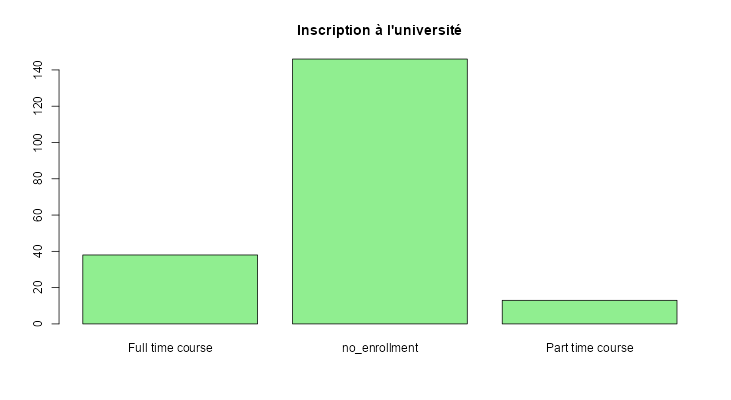
Voici une analyse des graphiques pour les variables qualitatives :

#### 1. ****Genre****



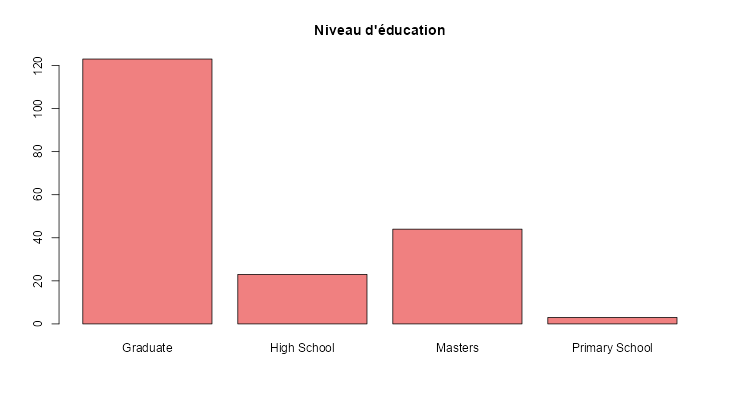
La majorité des individus de l'échantillon sont de genre masculin (plus de 70% des candidats de notre dataset.), ce qui montre un déséquilibre. Cela pourrait indiquer que les métiers de la data science sont encore dominés par les hommes.

#### 2. ****Inscription à l'université****



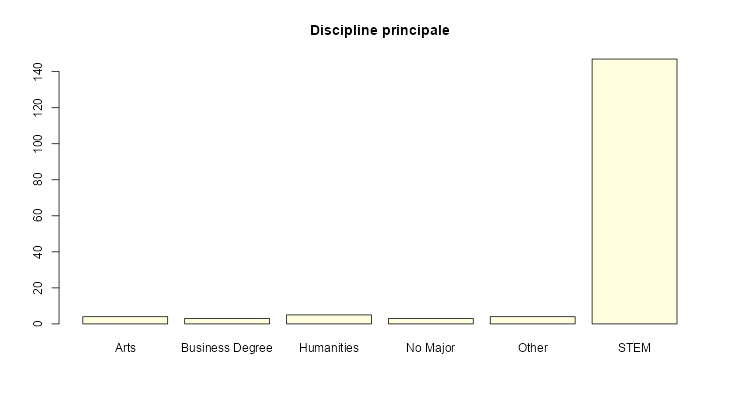
La plupart des candidats ne sont pas inscrits dans une université ('no\_enrollment'), tandis qu'une minorité suit des cours à temps plein ou à temps partiel. Cela pourrait indiquer que la majorité des candidats sont déjà sur le marché du travail plutôt qu'en formation. Il serait intéressant de voir si l’inscription (ou non) est liée à leur désir de changement de poste.

#### 3. ****Niveau d'éducation****



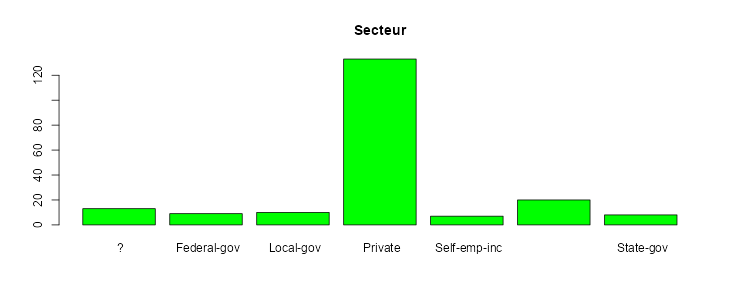
Le plus grand groupe de candidats a un diplôme de niveau 'Graduate', suivi des titulaires d'un 'Master', avec très peu de doctorats (PhD). La répartition montre que les postes en data science semblent majoritairement accessibles avec un diplôme de niveau licence ou master, tandis que le doctorat reste rare, ce qui pourrait être pertinent pour des postes plus spécialisés ou en recherche.

#### 4. ****Discipline principale****



La majorité des candidats viennent des domaines STEM (sciences, technologie, ingénierie et mathématiques), ce qui est attendu dans le domaine de la data science. Les autres disciplines, comme les arts, le commerce et les humanités, sont peu représentées. Cela peut limiter la diversité des perspectives dans le domaine, même si les profils non-STEM apportent souvent des compétences complémentaires utiles dans certains aspects de l'analyse de données.

1. Secteur



La majorité des candidats travaillent dans le secteur privé, avec une fréquence nettement plus élevée que les autres secteurs. Quelques candidats sont des travailleurs indépendants incorporés. On observe aussi qu'il y a quelques candidats dont le secteur d'emploi n'est pas spécifié (valeurs manquantes, ?) nous allons nous en occuper dans le prétraitement des données. Le secteur privé domine largement, indiquant que la plupart des candidats travaillent dans des entreprises privées. Les autres secteurs (travailleurs indépendants, gouvernements d'État, fédéral et local) sont nettement moins représentés.

#### Conclusion partielle

Ces graphiques permettent de constater des tendances importantes dans l'échantillon, notamment un déséquilibre en termes de genre, un fort taux de non-inscription à l'université, une majorité de candidats ayant un niveau d’éducation de licence ou master, et une forte représentation des disciplines STEM. Ces observations pourront être utiles pour interpréter les analyses ultérieures et comprendre comment ces variables qualitatives influencent la probabilité de changement de poste des candidats. Nous avons ajouté plusieurs autres graphique ainsi que les analyses correspondantes dans l’application shiny.

## Analyse bivariée

### Corrélations entre variables quantitatives

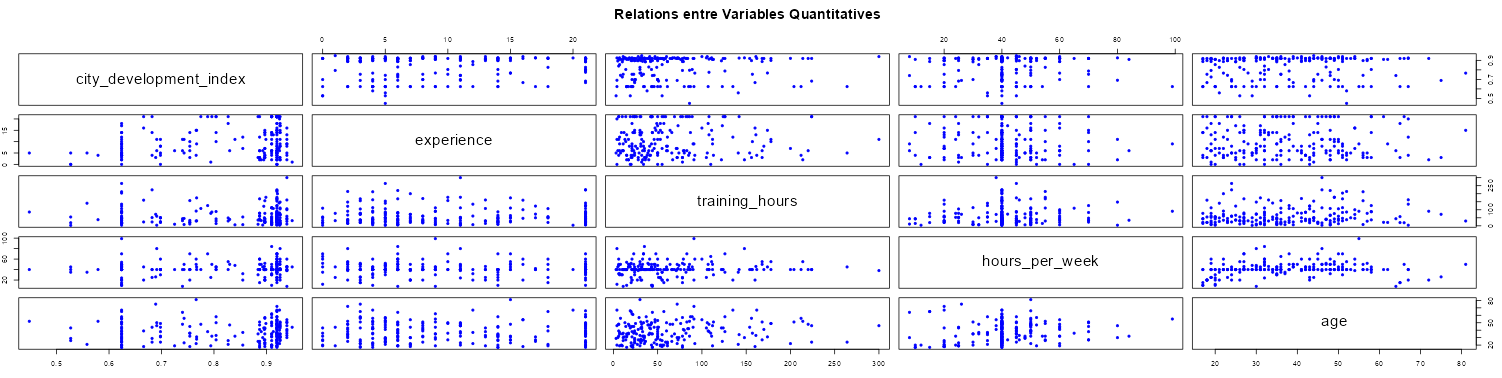
Une image contenant texte, Police, nombre, capture d’écran

Description générée automatiquement

**Analyse des Corrélations :**

* **City\_development\_index (Indice de développement de la ville) :**
  + **Faible corrélation avec expérience** (*r = 0.21*) : Les candidats des villes à indice élevé peuvent avoir légèrement plus d'expérience.
  + **Très faible corrélation avec les autres variables** (*r ≈ 0.03 − 0.06*) : Cet indice n'est pas un facteur déterminant pour ces variables.
* **expérience (Expérience totale) :**
  + **Très faible corrélation avec training\_hours** (*r = 0.02*) et **age** (*r = 0.02*) : Cela peut suggérer que l'expérience professionnelle n'est pas directement liée au nombre d'heures de formation ou à l'âge des candidats.
  + **Légère corrélation négative avec hours\_per\_week** (*r = −0.05*) : Cela peut indiquer que les candidats avec plus d'expérience travaillent légèrement moins d'heures, bien que l'effet soit faible.
* **training\_hours (Heures de formation) :**
  + **Faible corrélation avec hours\_per\_week** (*r = 0.08*) : Cela suggère que les candidats ayant plus d'heures de formation travaillent aussi légèrement plus d'heures par semaine.
  + **Corrélation modérée avec age** (*r = 0.13*) : Cela peut refléter que les candidats plus âgés ont tendance à suivre plus de formations.
* **hours\_per\_week (Heures travaillées par semaine) :**
  + **Corrélation modérée avec age** (*r = 0.15*) : Les candidats plus âgés ont tendance à travailler légèrement plus d'heures par semaine.
* **age (Âge du candidat) :**
  + **Corrélation faible ou très faible avec toutes les autres variables** (*r < 0.16*) : Cela montre que l'âge n'est pas fortement associé aux autres variables quantitatives dans cet ensemble.

**Conclusion globale :** Les corrélations entre les variables quantitatives sont globalement faibles à modérées. Cela indique qu'il n'y a pas de relations fortes ou linéaires entre ces variables dans notre échantillon.



**Analyse Visuelle de la matrice de dispersion :**

* **city\_development\_index :**
  + Aucune relation claire ou tendance visible avec les autres variables.
  + La dispersion des points est relativement homogène, ce qui confirme les faibles corrélations avec d'autres variables comme experience, training\_hours, ou hours\_per\_week.
* **experience :**
  + Pas de tendance apparente avec training\_hours, ce qui reflète une absence de relation linéaire notable.
  + Avec age, les points montrent une tendance légèrement croissante, ce qui est **logique car plus une personne est âgée, plus elle a de chances d'avoir accumulé de l'expérience**.
* **training\_hours :**
  + Aucune tendance forte avec les autres variables, y compris hours\_per\_week. Les points sont dispersés.
  + Avec age, une faible augmentation des heures de formation est visible pour certains groupes d'âge, mais les écarts restent faibles.
* **hours\_per\_week :**
  + Légère augmentation visible en fonction de age : les personnes plus âgées semblent travailler un peu plus d'heures par semaine en moyenne.
* **age :**
  + Les relations sont faibles avec les autres variables, bien que les tendances logiques (comme l'accumulation d'expérience avec l'âge) soient faiblement visibles.

**Observations Générales :**

* La matrice confirme les corrélations numériques faibles que nous avons calculées précédemment.
* Il n'y a pas de relation linéaire forte entre les variables quantitatives. La variabilité semble être due à des facteurs multiples ou non linéaires.
* Les variables semblent relativement indépendantes les unes des autres, à l'exception de quelques tendances légères (par exemple, âge vs expérience).

#### 1. city\_development\_index vs. experience

* La distribution semble assez dispersée, avec une très légère tendance positive, confirmée par la corrélation modérée de **0,33**. Cela pourrait indiquer qu'il y a une association positive entre le niveau de développement de la ville et l'expérience des individus, mais elle reste faible.

#### 2. city\_development\_index vs. training\_hours

* Les points semblent distribués de manière assez aléatoire sans tendance apparente, ce qui est cohérent avec la très faible corrélation de **-0,005**. Cela indique qu’il n'y a pas de relation significative entre le niveau de développement de la ville et les heures de formation.

#### 3. experience vs. training\_hours

* Bien que les points montrent une certaine densité dans les faibles valeurs d'expérience et d'heures de formation, il n'y a pas de tendance claire, ce qui est cohérent avec la corrélation quasi nulle de **-0,002**. Il semble que l'expérience et le nombre d'heures de formation ne soient pas significativement liés dans ce contexte.

#### Conclusion partielle

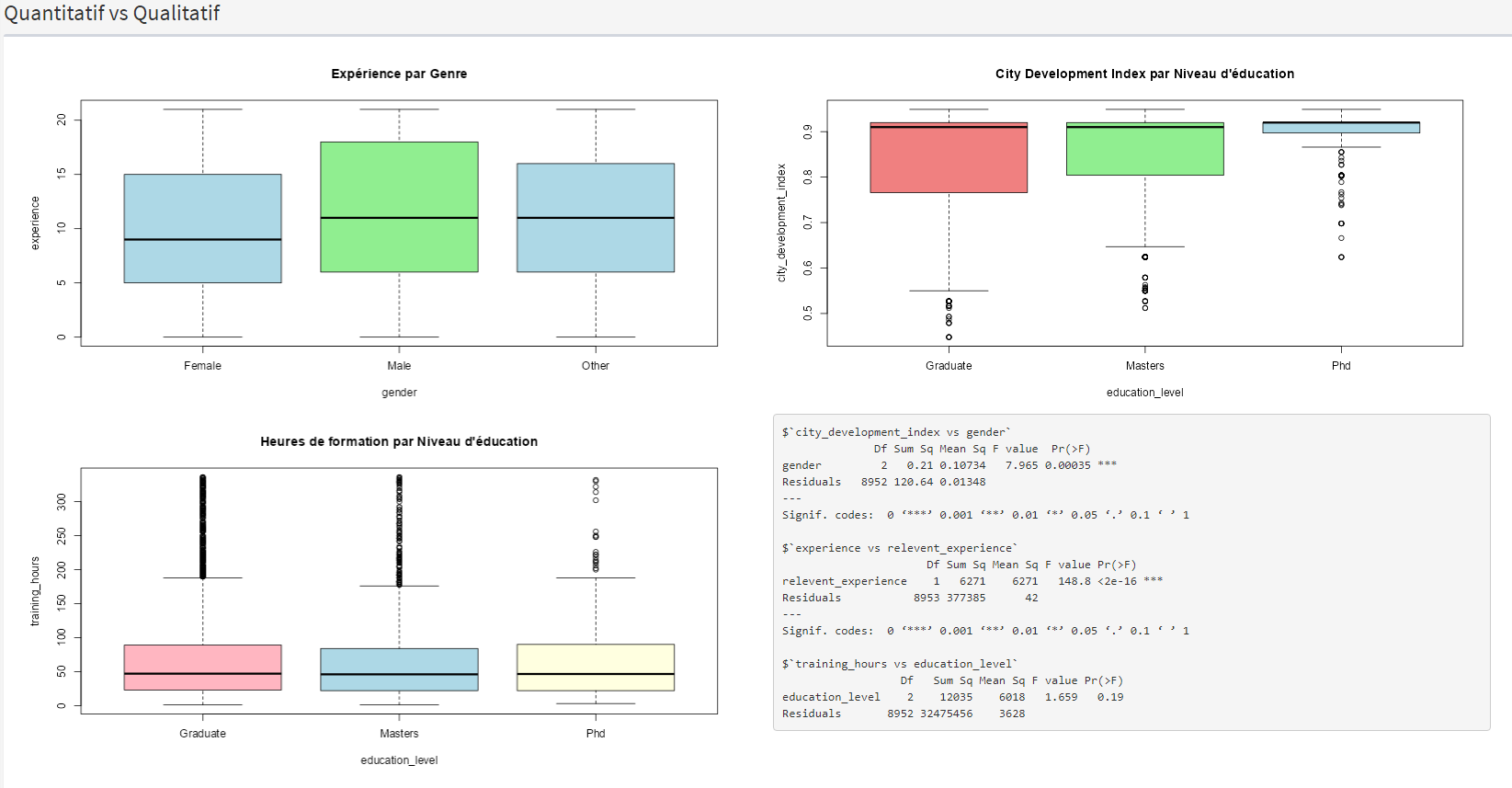
* Les relations entre les variables quantitatives sont très faibles, à l'exception de la relation modérée entre city\_development\_index et experience. Cela peut suggérer que ces variables ne sont pas directement liées entre elles, et leur influence sur la variable cible pourrait être mieux explorée en combinaison avec d'autres variables qualitatives ou en utilisant des méthodes de modélisation avancée.

### Corrélations entre variables quantitatives et qualitatives

Ce graphique de corrélation montre les relations entre les différentes variables de manière binaire et continue, codées par couleur pour indiquer la force et la direction de la corrélation (de -1 à 1). Voici quelques observations basées sur cette matrice :

1. **Corrélations positives fortes :**
   * Certaines catégories de variables catégorielles semblent être corrélées avec des sous-catégories. Par exemple, il pourrait y avoir des corrélations positives entre certains niveaux de education\_level et des types de company\_type, ce qui peut indiquer une certaine association entre ces catégories.
2. **Corrélations négatives :**
   * Il y a aussi quelques corrélations négatives significatives entre des sous-catégories, qui peuvent indiquer une opposition entre certaines caractéristiques. Par exemple, les catégories comme company\_type\_Public Sector et company\_type\_Pvt Ltd montrent une corrélation négative (car elles sont mutuellement exclusives).
3. **Variable cible (target) :**
   * La variable target ne semble pas montrer de corrélations très fortes avec les autres variables dans cette matrice, bien qu’il pourrait y avoir quelques relations plus faibles avec certaines catégories comme relevent\_experience\_Has relevent experience.
4. **Corrélation avec city\_development\_index :**
   * city\_development\_index ne semble pas avoir de corrélations marquées avec les autres variables binaires ou catégorielles, ce qui pourrait indiquer que cet indice est relativement indépendant par rapport aux autres attributs dans le contexte de ce dataset.
5. **Autres corrélations notables :**
   * Certaines corrélations faibles apparaissent, mais elles sont peu marquées. Par exemple, il peut y avoir une relation entre training\_hours et les niveaux d'éducation ou les types d'enrôlement universitaire.

En somme, la matrice indique des relations modérées et faibles, sans corrélations très marquées à exploiter immédiatement pour la variable cible target. Cela suggère que des méthodes avancées comme la régression logistique, les arbres de décision ou les modèles basés sur le machine learning pourraient être nécessaires pour capturer des relations non linéaires ou plus subtiles.



#### 1. Expérience par Genre

* **Graphique (Boxplot)** : Ce graphique montre la répartition des années d'expérience pour chaque genre. On peut noter que la médiane de l'expérience semble légèrement plus élevée pour les hommes que pour les femmes et les autres genres. Les intervalles interquartiles et les valeurs extrêmes sont également visibles.
* **Interprétation** : Il pourrait y avoir une différence dans l'expérience entre les genres, ce qui peut être lié à divers facteurs, tels que les secteurs ou les rôles où les individus travaillent.

#### 2. City Development Index par Niveau d'Éducation

* **Graphique (Boxplot)** : Ce graphique compare l'indice de développement de la ville en fonction des niveaux d'éducation (Graduate, Masters, PhD). On remarque que la médiane de l'indice est similaire pour les niveaux Graduate et Masters, avec une légère augmentation pour PhD. Cependant, la distribution présente des valeurs aberrantes pour chaque niveau.
* **Interprétation** : Il est possible que les individus ayant un niveau de doctorat (PhD) viennent de villes avec un indice de développement légèrement plus élevé. Cela pourrait être lié à l'accessibilité à l'éducation de haut niveau dans les grandes villes.
* **Test ANOVA (Résultats)** : Le tableau montre une valeur de p très faible (0.00035), suggérant une différence significative dans l'indice de développement de la ville en fonction du genre.

#### 3. Heures de Formation par Niveau d'Éducation

* **Graphique (Boxplot)** : Ce graphique présente la répartition des heures de formation en fonction des niveaux d'éducation. La médiane des heures de formation est similaire entre les trois groupes, mais il y a des valeurs aberrantes importantes, notamment pour les niveaux Masters et PhD.
* **Interprétation** : Bien que les heures de formation ne varient pas de manière significative en fonction du niveau d'éducation, il semble que certaines personnes, en particulier celles avec un niveau de doctorat, effectuent des heures de formation bien au-delà de la moyenne.
* **Test ANOVA (Résultats)** : Le p-value de 0.19 indique que les différences dans les heures de formation entre les niveaux d'éducation ne sont pas significatives.

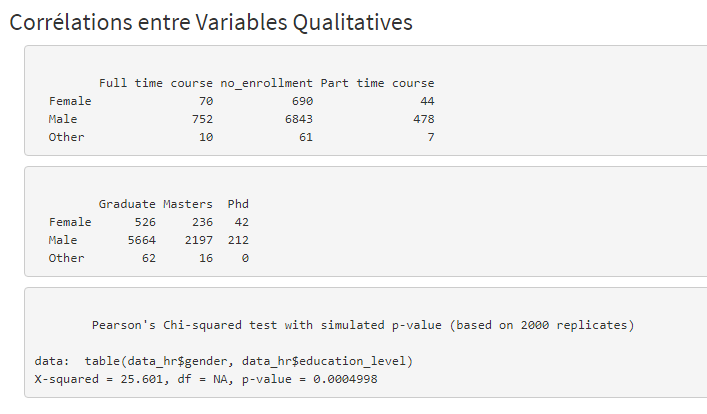
#### 4. Résultats ANOVA supplémentaires

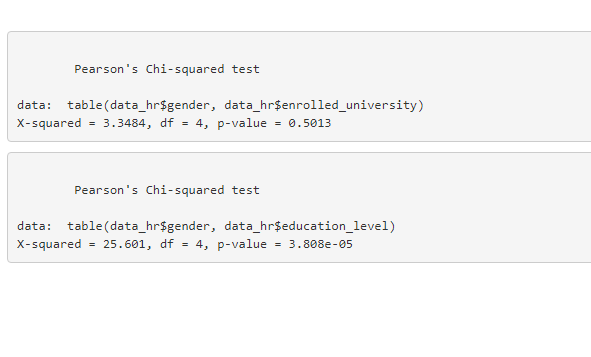
* **City Development Index vs Gender** : Ce test ANOVA a un p-value de 0.00035, indiquant que le genre a un effet significatif sur l'indice de développement de la ville.
* **Experience vs Relevant Experience** : Avec un p-value de < 2e-16, ce test montre une forte association entre l'expérience et l'expérience pertinente, ce qui est intuitivement logique.
* **Training Hours vs Education Level** : Avec un p-value de 0.19, il n'y a pas de différence significative dans les heures de formation selon le niveau d'éducation.

#### Conclusion

* **Variables significatives** : Le gender et education\_level montrent des différences significatives dans certaines variables (par exemple, city\_development\_index).
* **Variables non significatives** : Le niveau d'éducation n'a pas d'effet significatif sur les heures de formation, ce qui suggère que les heures de formation peuvent être influencées par d'autres facteurs indépendants du niveau d'éducation.

### Corrélations entre variables qualitatives





Les tests de corrélation entre les variables qualitatives fournissent des informations sur les relations entre le genre, l'inscription universitaire et le niveau d'éducation dans le jeu de données.

1. **Genre et Inscription à l'Université** :
   * La table de contingence montre les effectifs par genre et par statut d'inscription à l'université (cours à temps plein, pas d'inscription, cours à temps partiel).
   * Le test du khi-deux de Pearson (Chi-squared) pour cette association donne un X2=3.3484X^2 = 3.3484X2=3.3484, avec 4 degrés de liberté et une valeur p de 0.5013.
   * La p-valeur (0.5013) est bien supérieure au seuil habituel de 0.05, indiquant qu'il n'y a pas de relation statistiquement significative entre le genre et le statut d'inscription universitaire. Autrement dit, le statut d'inscription à l'université ne dépend pas du genre dans ce jeu de données.
2. **Genre et Niveau d'Éducation** :
   * La table de contingence indique les effectifs par genre et niveau d'éducation (Graduate, Masters, PhD).
   * Le test du khi-deux pour cette association montre un X2=25.601X^2 = 25.601X2=25.601, avec 4 degrés de liberté, et une p-valeur de 3.808e-05 (ou 0.00003808).
   * La p-valeur est très faible, bien en dessous du seuil de 0.05, indiquant une relation statistiquement significative entre le genre et le niveau d'éducation. Cela suggère que la distribution des niveaux d'éducation diffère en fonction du genre.
   * Le test de khi-deux avec valeur simulée (basée sur 2000 répétitions) confirme également cette signification avec une p-valeur de 0.0004998, soutenant le fait que le genre et le niveau d'éducation sont associés.

**En résumé** :

* Genre et Inscription à l'Université : Pas de relation significative.
* Genre et Niveau d'Éducation : Relation significative.

# 4. Analyse en composantes principales (ACP)

## Sélection de variables

### Variables actives

Les variables actives doivent être des variables quantitatives ou ordinales ayant une pertinence directe pour l’analyse des comportements ou des regroupements.

### ****2. Variables illustratives****

Les variables illustratives doivent être des variables qualitatives servant à enrichir l’interprétation des groupes ou axes factoriels a posteriori. Ces variables n’interviennent pas dans la construction des composantes principales mais peuvent aider à donner du sens aux résultats.

### Justification des choix

|  |  |
| --- | --- |
| Type | Variables choisies |
| Variables actives | city\_development\_index, training\_hours, age, hours\_per\_week, experience |
| Variables illustratrives | gender, relevent\_experience, education\_level, major\_discipline, company\_type |

Les variables sélectionnées pour être actives sont celles qui contribuent directement à expliquer la variabilité des individus dans les dimensions principales, de plus ces variables sont compatibles avec l’ACP. Ces variables quantifiables permettent de maximiser l’inertie des composantes et de capturer l’essence des comportements dans l’analyse.

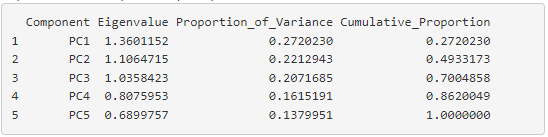
Les variables illustratives choisies permettent d’analyser les relations entre les axes factoriels et des caractéristiques spécifiques des individus. Elles ne biaisent pas l’analyse mais permettent une interprétation plus riche des clusters ou des composantes principales.

### Illustration des choix

Dans le cercle des corrélations (voir partie suivante) :

* Les flèches représentant les variables actives montrent leur contribution à chaque composante principale (Dim1 et Dim2). Par exemple :
  + **training\_hours** et **hours\_per\_week** sont corrélées positivement avec Dim1.
  + **age** et **experience** peuvent influencer Dim2.
* Les variables qualitatives illustratives peuvent être projetées pour observer si elles se distinguent en fonction des axes factoriels.

## Importance des composantes principales issues de l'ACP



### Interprétation des valeurs propres :

* Les valeurs propres représentent l'inertie expliquée par chaque composante principale (**PC**).
* Une valeur propre supérieure à 1 indique qu'une composante explique plus de variance qu'une variable initiale unique.
* Ici, les trois premières composantes (**PC1**, **PC2**, et **PC3**) ont des valeurs propres > 1, ce qui suggère qu'elles sont importantes pour résumer l'information du jeu de données.

### Proportion de variance expliquée :

* La colonne "**Proportion\_of\_Variance**" montre la part de variance totale expliquée par chaque composante :
  + **PC1** : 27,2 %
  + **PC2** : 22,1 %
  + **PC3** : 20,7 %
* Cela signifie que **PC1**, **PC2**, et **PC3** expliquent ensemble environ 70 % de la variance totale.

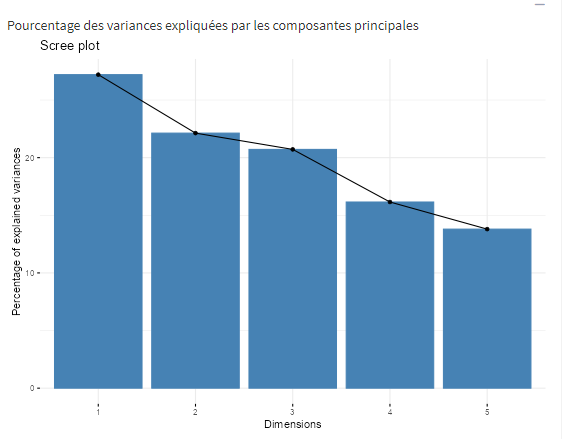
### Cumul de la variance expliquée :

* La colonne "**Cumulative\_Proportion**" montre le pourcentage cumulé de variance expliquée :
  + Les trois premières composantes (**PC1** à **PC3**) expliquent 70 % de la variance cumulée.
  + Avec **PC4** et **PC5**, toute la variance est expliquée (100 %), mais les deux dernières composantes ajoutent moins d'information significative.

### Choix du nombre de composantes principales :

* Généralement, on conserve les composantes qui expliquent environ 70 à 80 % de la variance cumulée ou dont les valeurs propres sont supérieures à 1.
* Ici, on peut justifier de conserver **PC1**, **PC2**, et **PC3** (avec une variance cumulative de 70 %) :
  + Elles contiennent l'essentiel de l'information.
  + Réduire les dimensions à ces trois composantes est à la fois efficace et informatif.
* On observe un "coude" clair après la 3ᵉ dimension. Cela suggère qu’après les trois premières composantes principales, l’ajout d’autres dimensions apporte des gains marginaux en termes de variance expliquée.

### Pourcentage des variances expliquées par les composantes principales



Le graphique suit une décroissance (ou coude), indiquant que les premières composantes principales expliquent une part importante de la variance totale.

La pente diminue progressivement après les premières dimensions, ce qui montre une contribution réduite des composantes supplémentaires.

Cet histogramme nous confirme que les **trois premières composantes principales** suffisent à résumer efficacement l'information du jeu de données, avec un compromis optimal entre la réduction dimensionnelle et la préservation de la variance expliquée.

### Plans factoriels des individus

Une image contenant diagramme, texte, cercle, ligne

Description générée automatiquement

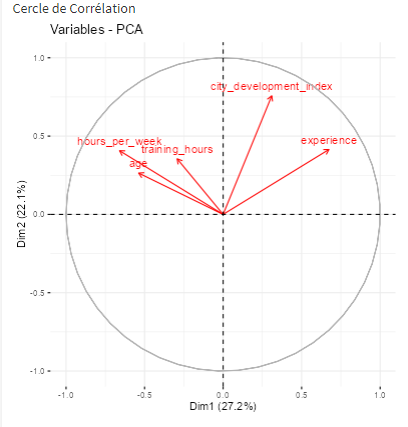
#### Interprétation :

* Ce graphique représente les individus projetés sur les deux premières dimensions principales (**Dim1** et **Dim2**).
* L’ellipse bleue montre la zone où se concentrent la majorité des individus.
* Les points correspondent aux individus, et leur position indique leur relation vis-à-vis des axes principaux.

#### Analyse :

* Les individus proches du centre (origine) ont des valeurs moyennes sur toutes les variables.
* Les individus éloignés du centre dans une direction donnée sont fortement influencés par les variables corrélées à cette direction.

### Cercle de corrélation



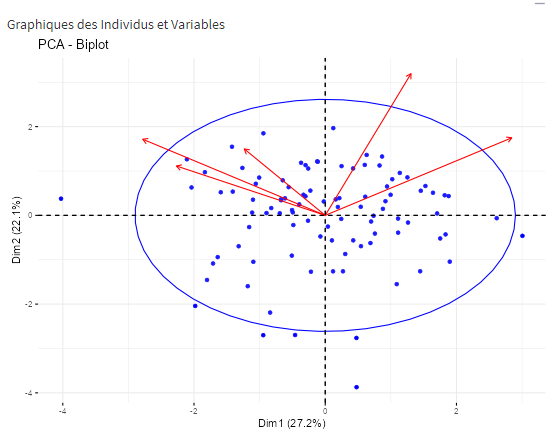
#### Interprétation :

* Ce graphique montre la contribution et la corrélation des variables d’origine avec les dimensions principales (**Dim1** et **Dim2**).
* Chaque flèche représente une variable, et sa longueur indique sa contribution.
* Les variables proches de la circonférence du cercle sont mieux représentées par les axes **Dim1** et **Dim2**.

#### Analyse :

* **Dim1 (27,2 %)** :
  + Les variables fortement corrélées avec **Dim1** incluent probablement *city\_development\_index* et *experience*.
  + Cela signifie que cette dimension peut refléter des aspects liés au développement urbain et à l'expérience.
* **Dim2 (22,1 %)** :
  + *training\_hours* semble contribuer davantage à cette dimension.
  + Elle pourrait refléter un aspect lié à l'engagement dans la formation.
* Les flèches opposées (angles de 180°) indiquent des corrélations négatives. Par exemple, si *age* est opposé à *training\_hours*, cela pourrait indiquer une relation négative entre ces deux variables.

### Graphiques des Individus et Variables



#### Interprétation :

* Ce graphique combine les deux précédents, superposant les individus et les variables.
* Il permet d'interpréter comment les variables influencent les regroupements des individus.

#### Analyse :

* Les individus situés près d'une flèche (variable) sont fortement influencés par cette variable.
* Par exemple, des individus proches de la flèche de *training\_hours* sont ceux ayant consacré plus de temps à leur formation.

### Conclusion partielle

A faire

# K-means

## Choix du nombre de clusters

Une image contenant texte, ligne, Tracé, diagramme

Description générée automatiquement

La courbe montre la somme des carrés intra-cluster (variance interne des clusters) en fonction du nombre de clusters **k**.

Interprétation du graphique

### Tendance générale :

* Lorsque le nombre de clusters augmente, la variance interne diminue (c’est attendu, car plus il y a de clusters, plus les points sont proches de leurs centres).
* Cependant, cette diminution n’est pas linéaire : elle ralentit progressivement.

### Identification du coude :

* Le "coude" est le point où la diminution de la variance interne commence à ralentir significativement.
* Sur ce graphique, le coude semble se situer autour de **k = 3** ou **k = 4**.
* Avant ce point, l'ajout de clusters réduit considérablement la variance interne.
* Après ce point, l'ajout de clusters apporte peu d'amélioration en termes de compacité.

### Choix du nombre de clusters

Le coude est une bonne indication du nombre optimal de clusters, car il correspond à un équilibre entre la simplicité du modèle (moins de clusters) et sa capacité à expliquer la structure des données (compacité des clusters).

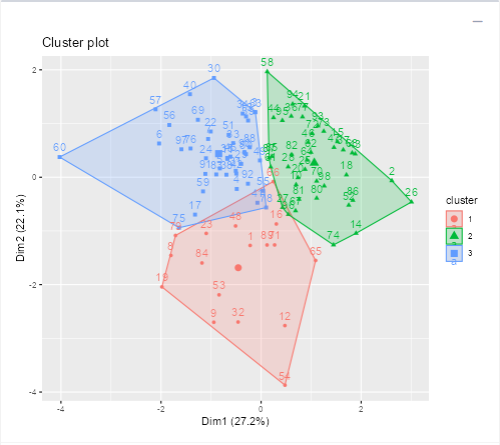
Justification :

* **k = 3** : Une réduction importante de la variance interne est encore observable, tout en maintenant une simplicité du modèle.
* **k = 4** : Peut aussi être un choix acceptable, mais au-delà, les gains deviennent négligeables.

### Conclusion

En utilisant la méthode du coude, nous pouvons justifier le choix de **k = 3** comme le nombre optimal de clusters pour notre analyse de K-means. Cela permet un bon équilibre entre la qualité des clusters et la simplicité du modèle. Si nécessaire, **k = 4** peut être exploré comme alternative pour affiner les clusters.

### Clusters obtenus avec k = 3



Cette visualisation montre les clusters obtenus avec **k = 3** lors de l'analyse K-means. Expliquaons :

### Projection sur deux dimensions principales (Dim1 et Dim2) :

* Ce graphique montre les données projetées dans un espace réduit à deux dimensions principales, permettant de visualiser les clusters sur un plan.
* Les clusters **1**, **2** et **3** sont représentés par des couleurs différentes (rouge, vert, bleu).
* Chaque cluster est délimité par une enveloppe convexe qui contient les points appartenant à ce cluster.

### Analyse des clusters :

* **Cluster 1 (rouge)** : Concentré dans une région spécifique du graphe, indiquant des individus relativement homogènes dans leurs caractéristiques.
* **Cluster 2 (vert)** : Montre un groupe dense avec une certaine variabilité dans la dimension **Dim1**.
* **Cluster 3 (bleu)** : Bien séparé, suggérant que les individus partagent des similarités spécifiques qui les différencient des autres clusters.

### Séparation des clusters :

* Les clusters sont bien séparés, indiquant une bonne segmentation des données par l'algorithme K-means.
* Cependant, quelques chevauchements peuvent indiquer des observations limites (frontières entre clusters).

# Analyse de la distribution suivant les variables age et training\_hours.

Une image contenant capture d’écran, texte, diagramme, Tracé

Description générée automatiquement

## Axes

* Les variables sélectionnées sont **age** (axe **x**) et **training\_hours** (axe **y**).
* Cela permet de visualiser comment les individus se regroupent en fonction de ces deux variables spécifiques.

## Clusters

* **Cluster 1 (rouge)** : Correspond aux individus ayant généralement un âge moyen et un nombre de **training\_hours** élevé.
* **Cluster 2 (vert)** : Correspond à des individus plus jeunes avec des **training\_hours** modérés.
* **Cluster 3 (bleu)** : Correspond à des individus plus âgés et avec un nombre plus faible de **training\_hours**.

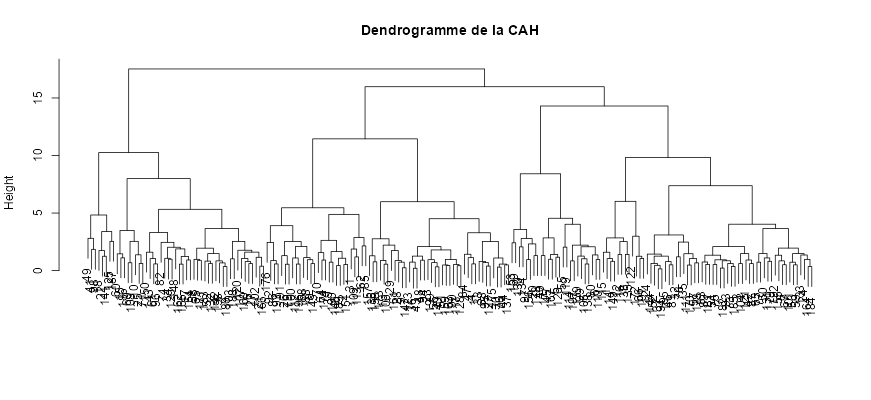
## Interprétation des caractéristiques :

* Les clusters révèlent des groupes distincts d’individus en fonction de l’âge et du temps de formation.
* Par exemple, les individus plus jeunes (vert) ont tendance à avoir un profil distinct des individus plus âgés (bleu), ce qui peut correspondre à différentes catégories dans les données, comme des niveaux d'expérience ou des priorités différentes.

## Conclusion partielle

# Classification Hiérarchique ascendante

## Dendrogramme global de la CAH



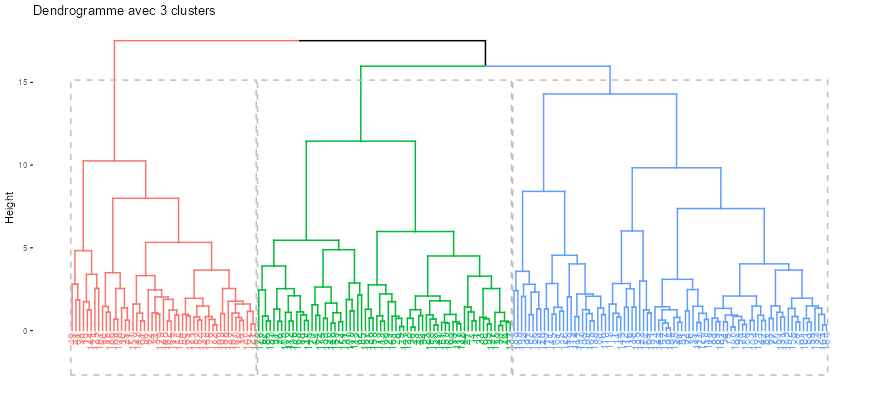
### Structure hiérarchique :

* Le dendrogramme illustre la hiérarchie entre les observations.
* Les observations (individus) sont représentées en bas, tandis que les regroupements successifs sont indiqués par des fusions verticales.
* L'axe des ordonnées (**Height**) représente la distance ou l'inertie nécessaire pour fusionner des clusters.

### Interprétation :

* En regardant le dendrogramme, on observe des niveaux où des regroupements importants se forment (grandes branches).
* Cela suggère que le jeu de données peut être divisé en groupes bien distincts à ces niveaux.

## Dendrogramme découpé en 3 clusters



### Partition en trois clusters :

* En coupant le dendrogramme à un certain niveau (lignes pointillées), les données sont divisées en **3 clusters** (rouge, vert, bleu).
* Chaque couleur représente un cluster distinct.

### Séparation des clusters

* Les clusters semblent bien séparés, avec des branches qui ne se chevauchent pas significativement.
* Le choix de **3 clusters** semble cohérent, basé sur l'observation visuelle et les regroupements naturels des données.

Justification du choix de 3 clusters :

### Observation des distances

* Le choix du nombre de clusters se base sur le "saut" des distances entre les fusions (coupure au niveau des grandes branches).
* Ici, couper le dendrogramme en **3** semble optimal pour capturer les regroupements majeurs.

### Cohérence avec K-means :

* Ce découpage en **3 clusters** est cohérent avec les résultats obtenus précédemment via K-means, confirmant la robustesse de cette segmentation.

### Conclusion partielle

# 6. Synthèse et conclusion

* Résumé des principales observations
* Pistes de réflexion et améliorations futures

# 7. Annexes

* Tableaux et graphiques supplémentaires
* Codes et méthodes de calcul utilisées

Voici la fusion que j’ai faite sur ChatGPT.

### ****PLAN FINAL DÉTAILLÉ****

1. **Introduction**
   * Contexte de l’étude
   * Objectifs de l’étude
   * Description générale des données et de la problématique
2. **Présentation des Données**
   * Description du jeu de données
   * Types de variables et leurs caractéristiques
   * Description des individus
   * Prétraitement des variables
     + Normalisation / Centrage-Réduction
     + Imputation des données manquantes (si applicable)
     + Sélection des variables actives et illustratives
3. **Analyse Descriptive des Données**
   * **Analyse Univariée**
     + Étude des variables quantitatives
     + Étude des variables qualitatives
     + Résumé et bilan de l’analyse univariée
   * **Analyse Bivariée**
     + Lien entre les variables quantitatives
     + Lien entre variables quantitatives et qualitatives
     + Lien entre variables qualitatives
     + Matrice des coefficients de corrélation
     + Bilan de l’analyse bivariée
4. **Analyse Factorielle et Réduction de Dimensionnalité**
   * **Analyse en Composantes Principales (ACP)**
     + Objectifs de l’ACP
     + Choix des variables actives et illustratives
     + Standardisation des variables
     + Étude des valeurs propres et choix du nombre d’axes à analyser
     + Représentation des individus et des variables
       - Interprétation des axes et des plans factoriels
       - Graphes des individus et des variables sur les axes principaux
       - Analyse des corrélations entre axes et variables
     + Analyse des contributions des variables et des individus aux axes
     + Bilan et conclusion de l’ACP
   * **Analyse des Correspondances Multiples (ACM)** (si applicable)
     + Choix du nombre de dimensions
     + Étude des dimensions et représentation graphique
     + Interprétation des plans factoriels de l’ACM
5. **Classification Non-Supervisée**
   * Objectifs de la classification
   * Choix de la méthode de classification (k-means, CAH, etc.)
   * Choix des distances et du nombre de classes
   * Description des classes par les variables
     + Variables influençant le plus chaque classe
     + Analyse des individus spécifiques dans chaque classe
     + Comparaison et interprétation des classes
   * Utilisation conjointe de la classification et de l’ACP
     + Analyse du profil des classes dans les plans factoriels
     + Visualisation des classes sur les axes de l’ACP
   * Bilan et conclusion de la classification
6. **Discussion et Conclusion Générale**
   * Résumé des principaux résultats
   * Limites de l’analyse
   * Perspectives d’amélioration et futures analyses
7. **Annexes**
   * Données supplémentaires et tables descriptives
   * Représentation graphique des analyses (nuages de points, dendrogrammes, etc.)
   * Résultats supplémentaires et codes utilisés (si applicable)
   * Tableau des contributions des variables et des individus