

Rapport d’analyse de données – STA 101

Analyse des facteurs socio-démographiques influençant la mobilité professionnelle des profils du domaine de la Data Science.



FONKOUO SANOU TEMENA Loic

Liens des datasets :

[Adult Census Income](https://www.kaggle.com/datasets/uciml/adult-census-income)

[HR Analytics: Job Change of Data Scientists](https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists)

L’application est disponible sur le lien [suivant](https://loicsanou.shinyapps.io/Projet-STA101/)

<https://loicsanou.shinyapps.io/Projet-STA101/>

Table des matières

[Introduction 4](#_Toc188724583)

[Objectifs de l’étude 4](#_Toc188724584)

[Contexte et présentation du jeu de données 5](#_Toc188724585)

[1. Présentation du jeu de données final 5](#_Toc188724586)

[a) Processus de fusion des données 5](#_Toc188724587)

[b) Justification des choix 6](#_Toc188724588)

[2. Pré-traitement des données 7](#_Toc188724589)

[3. Distribution des classes 7](#_Toc188724590)

[Préparation des données 8](#_Toc188724591)

[Gérer les valeurs manquantes 8](#_Toc188724592)

[4. Analyse descriptive 10](#_Toc188724593)

[a. Analyse univariée 10](#_Toc188724594)

[i. Statistique des variables quantitatives 10](#_Toc188724595)

[ii. Variables qualitatives 14](#_Toc188724596)

[b. Analyse bivariée 17](#_Toc188724597)

[iii. Analyse des variables quantitatives Vs Qualitatives 17](#_Toc188724598)

[iv. Matrice de corrélation de Pearson et Spearman 18](#_Toc188724599)

[5. Analyse en composantes principales (ACP) 19](#_Toc188724600)

[a. Sélection de variables 19](#_Toc188724601)

[1. Variables actives 19](#_Toc188724602)

[2. Variables illustratives 19](#_Toc188724603)

[3. Justification des choix 19](#_Toc188724604)

[4. Illustration des choix 20](#_Toc188724605)

[b. Importance des composantes principales issues de l'ACP 20](#_Toc188724606)

[1. Interprétation des valeurs propres 20](#_Toc188724607)

[2. Proportion de variance expliquée 21](#_Toc188724608)

[3. Cumul de la variance expliquée 21](#_Toc188724609)

[4. Choix du nombre de composantes principales 21](#_Toc188724610)

[5. Pourcentage des variances expliquées par les composantes principales 21](#_Toc188724611)

[6. Plans factoriels des individus 22](#_Toc188724612)

[7. Cercle de corrélation 22](#_Toc188724613)

[8. Graphiques des Individus et Variables 23](#_Toc188724614)

[K-means 24](#_Toc188724615)

[1. Choix du nombre de clusters 24](#_Toc188724616)

[a. Tendance générale 24](#_Toc188724617)

[b. Identification du coude 24](#_Toc188724618)

[c. Choix du nombre de clusters 24](#_Toc188724619)

[d. Conclusion 25](#_Toc188724620)

[e. Clusters obtenus avec k = 3 25](#_Toc188724621)

[f. Projection sur deux dimensions principales (Dim1 et Dim2) : 25](#_Toc188724622)

[g. Analyse des clusters : 25](#_Toc188724623)

[h. Séparation des clusters : 25](#_Toc188724624)

[2. Analyse de la distribution suivant les variables age et training\_hours. 26](#_Toc188724625)

[a. Axes 26](#_Toc188724626)

[b. Clusters 26](#_Toc188724627)

[c. Interprétation des caractéristiques 26](#_Toc188724628)

[d. Conclusion partielle 26](#_Toc188724629)

[Classification Hiérarchique Ascendante 27](#_Toc188724630)

[1. Dendrogramme global de la CAH 27](#_Toc188724631)

[a. Structure hiérarchique 27](#_Toc188724632)

[b. Interprétation 27](#_Toc188724633)

[2. Dendrogramme découpé en 3 clusters 27](#_Toc188724634)

[a. Partition en trois clusters : 27](#_Toc188724635)

[b. Séparation des clusters 28](#_Toc188724636)

[c. Observation des distances 28](#_Toc188724637)

[d. Cohérence avec K-means : 28](#_Toc188724638)

[e. Conclusion partielle 28](#_Toc188724639)

[Synthèse et conclusion 28](#_Toc188724640)

[7. Annexes 29](#_Toc188724641)

[1. Analyse bivariée - Matrice de dispersion 29](#_Toc188724642)

[2. Tests Statistiques 30](#_Toc188724643)

[3. Corrélations entre Variables Quantitatives 31](#_Toc188724644)

[4. Corrélations entre Variables Qualitatives 33](#_Toc188724645)

[8. Guide d'installation et d'utilisation de l'application Shiny 37](#_Toc188724646)

[1. Préparation des fichiers 37](#_Toc188724647)

[2. Lancement de l’application 37](#_Toc188724648)

[3. Chargement des données 38](#_Toc188724649)

[4. Exploration de l’application 38](#_Toc188724650)

[5. Exploration des codes 38](#_Toc188724651)

[6. Option : Cloner le dépôt GitHub 39](#_Toc188724652)

[7. Support 39](#_Toc188724653)

# 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](https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists) provenant d'une analyse RH axée sur le changement d'emploi des scientifiques des données, et [l'autre](https://www.kaggle.com/datasets/uciml/adult-census-income) 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)

# 3. 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, ce qui est cohérant avec notre échantillonnage stratifié expliqué plus haut.

## Préparation des données

### Gérer les valeurs manquantes

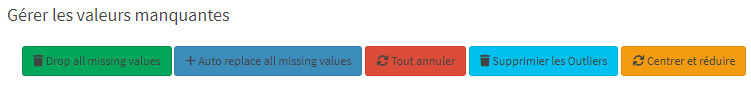


Figure 2 : Actions sur les valeurs manquantes.

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 3 : Jeu de données avec tous les individus.

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



Figure 4 : 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.

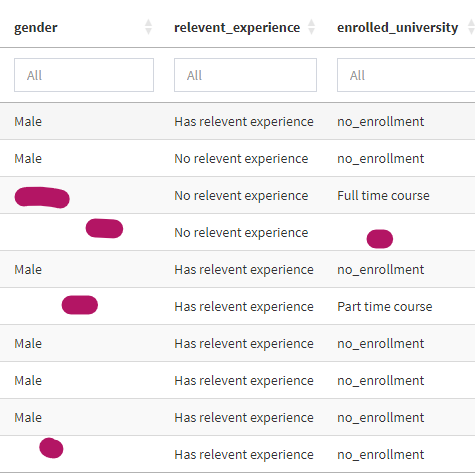


Figure 5 : Jeu de données avec valeurs manquantes sur le gender.

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

Figure 6 : Jeu de données après remplacement automatique.

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

* 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.

# Analyse descriptive

## Analyse univariée

Dans cette section, le but est de décrire les variables de façon univariée. Les graphiques présentés ici ne constituent qu’un échantillon représentatif de la multitude de visualisations disponibles dans l’application Shiny que nous vous invitons à jeter un coup [d’œil](https://loicsanou.shinyapps.io/Projet-STA101/).

### Statistique des variables quantitatives

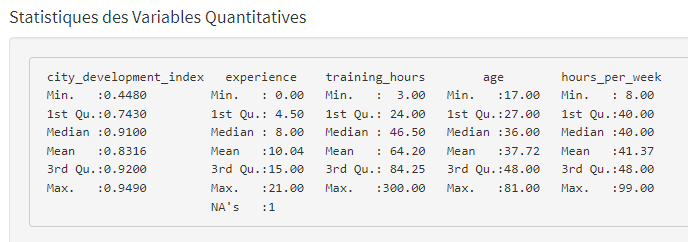


Figure 7 : Statistique descriptive des variables quantitatives.

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.

#### City Development Index

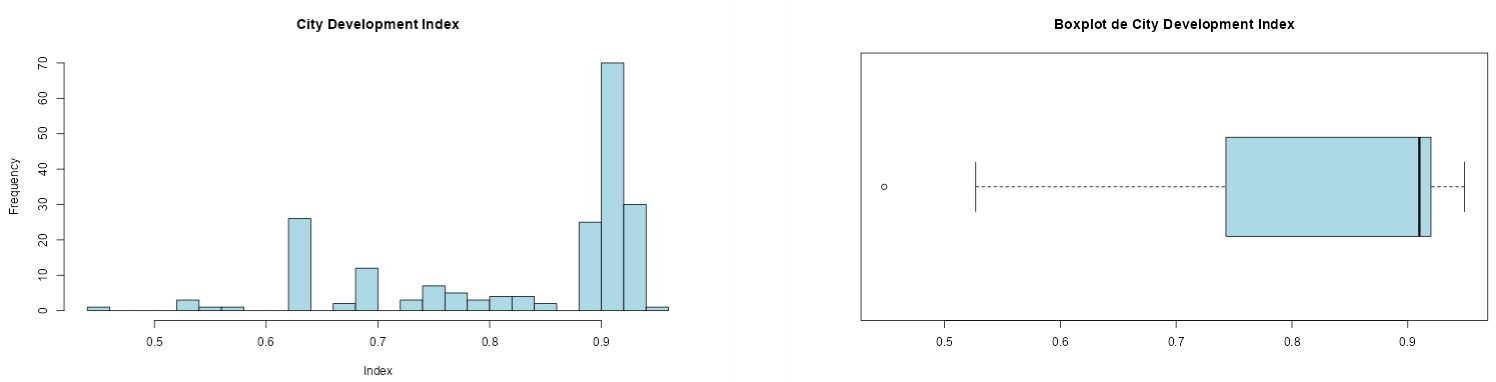


Figure 8 : Histogramme et boxplot de city\_development\_index.

* **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 : On observe 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 y’a 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****

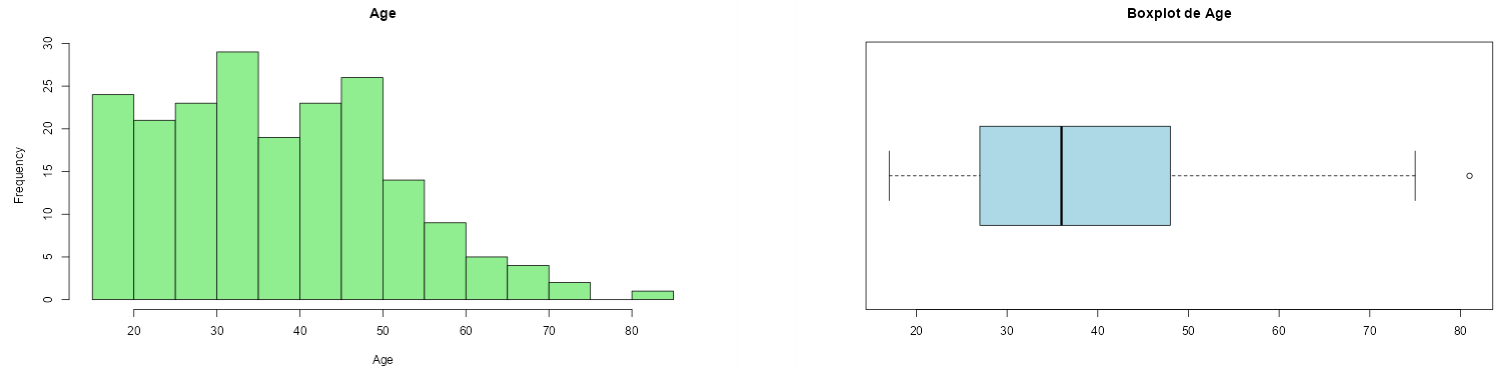


Figure 9 : Histogramme et boxplot de age.

* **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 37 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

Figure 10 : Histogramme et boxplot de Expérience.

* 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.

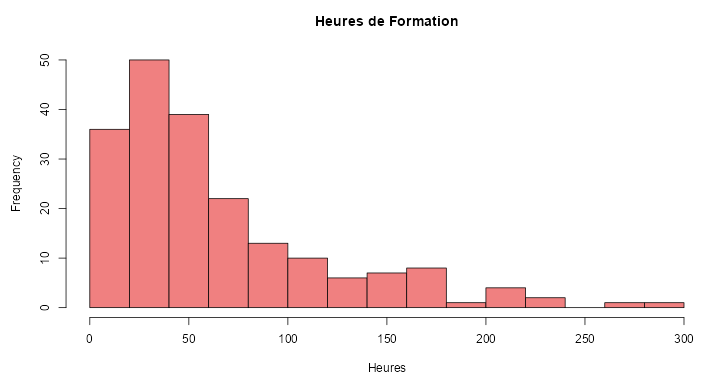
* Analyse du Boxplot de l’expérience :

- **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



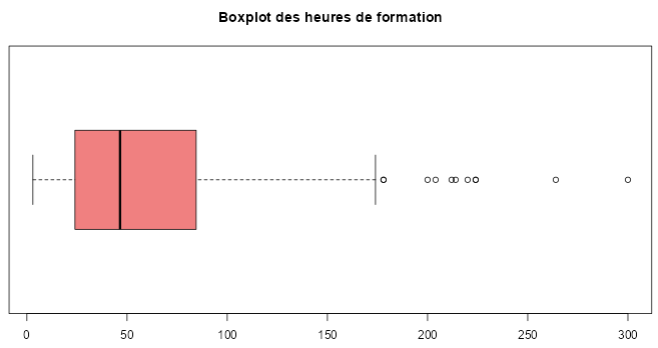


Figure 11 : Histogramme et boxplot de Formation.

* **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.

#### ****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

Figure 12 : Statistique des variables qualitatives.

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

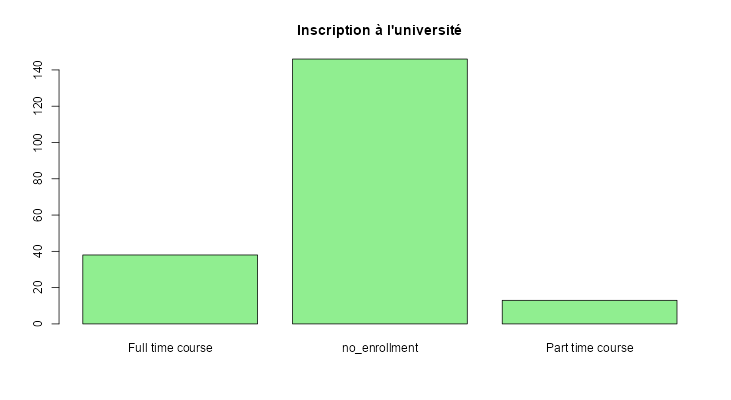
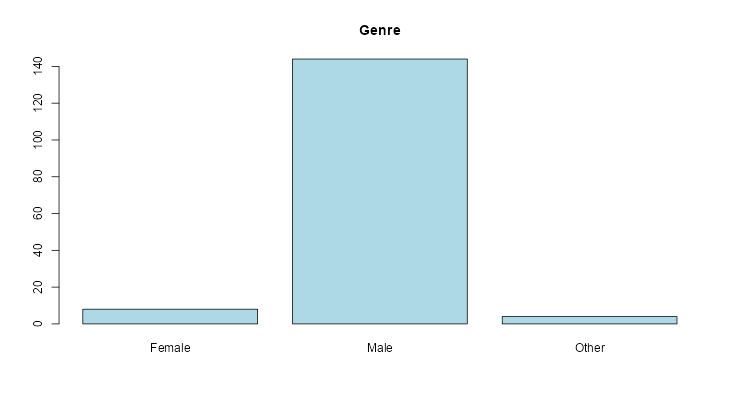


Figure 13 : Histogramme des variables Genre et Inscription à l'université.

#### 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****

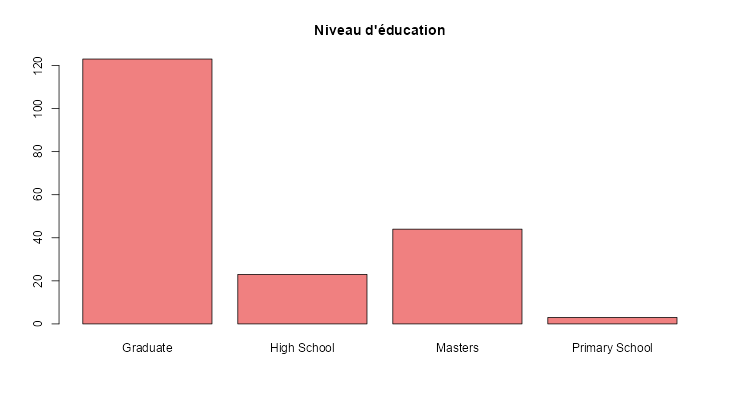


Figure 14: Histogramme du niveau d'éducation.

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

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

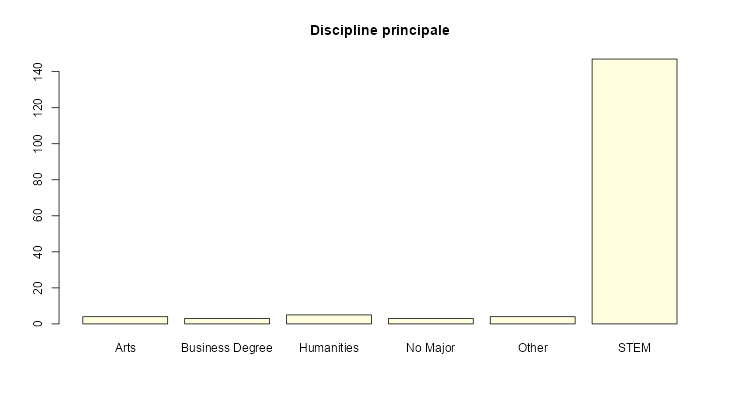


Figure 15 : Histogramme de la Discipline principale.

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.

#### 5. Secteur

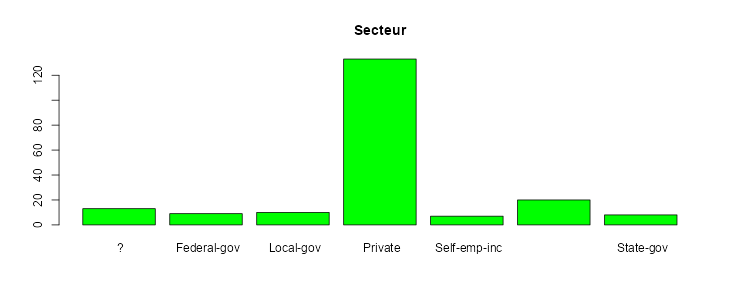


Figure 16 : Histogramme de la variable Secteur.

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 nous sommes occupés 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.

#### 6. 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 nous serons 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 graphiques ainsi que les analyses correspondantes dans l’application shiny.

## Analyse bivariée

### Analyse des variables quantitatives Vs Qualitatives

#### Expérience par Genre

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

Description générée automatiquement

Figure 17 : 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 est légèrement plus élevée pour les femmes que pour les hommes et les autres genres. Les boîtes ont à peu près la même largeur pour les trois groupes, ce qui indique une dispersion des données relativement similaire. Il n'y a pas de valeurs aberrantes très éloignées des boîtes, ce qui suggère que les données sont relativement homogènes.

#### 2. city\_development\_index par niveau d’expérience

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

Description générée automatiquement

Figure 18 : city\_development\_index par niveau d’expérience

Les niveaux d'éducation 'High School', 'Primary School' et 'Masters' présentent des médianes presque similaires, indiquant que les candidats avec ces niveaux d'éducation tendent à vivre dans des villes avec un indice de développement un peu moins élevé par rapport au niveau Graduate. Le niveau d'éducation 'Graduate' présente une médiane légèrement supérieure, suggérant que les candidats avec ce niveau d'éducation vivent dans des villes avec un indice de développement plus élévé.

#### 3. Heure de formation par niveau d’éducation

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

Description générée automatiquement

Figure 19 : Heure de formation par niveau d’éducation

* Les candidats avec un niveau d'éducation 'Masters', 'Graduate' et 'High School' tendent à suivre un nombre médian d'heures de formation similaire (environ 50 heures), mais avec une variabilité plus grande pour les diplômés de masters.
* Les candidats avec un niveau d'éducation 'Primary School' tendent à suivre un nombre médian d'heures de formation très faible (environ 5 heures), avec une variabilité élevée.
* Il y a une variabilité notable dans les heures de formation pour tous les niveaux d'éducation, avec des valeurs aberrantes indiquant que certains candidats suivent beaucoup plus d'heures de formation que la médiane.

#### 4. 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.

### Matrice de corrélation de Pearson et Spearman

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

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquement

Figure 20: Matrice de corrélation – Méthode de Pearson et Spearman.

Figure 21: Matrice de corrélation – Méthode de Pearson et Spearman.

#### 1. Méthode de Pearson

experience vs age (0.21) : Une faible corrélation positive, ce qui est logique, car plus une personne est âgée, plus elle a accumulé d'expérience professionnelle. Autres relations (comme experience vs training\_hours, hours\_per\_week, etc.) montrent des coefficients très faibles (proches de 0), indiquant une absence de relation significative. age vs training\_hours (0.08) : Une corrélation positive très faible, cohérente avec le scatterplot précédemment analysé. age vs hours\_per\_week (0.16) : Une faible corrélation positive, suggérant que les heures travaillées par semaine augmentent légèrement avec l'âge. age vs experience (0.21) : Déjà commenté, mais reste la relation la plus notable avec age. Les relations entre le city\_development\_index et d'autres variables sont faibles ou proches de 0, montrant que cet indice n'a pas d'impact direct sur des variables comme l'expérience ou les heures de travail. Absence de relations significatives : Les variables comme training\_hours, hours\_per\_week et city\_development\_index ne montrent pas de fortes corrélations entre elles ou avec d'autres variables. Aucune des variables ne présente une corrélation forte avec une autre. Cela suggère que ces variables sont relativement indépendantes les unes des autres. Les relations observées (par exemple, entre age et experience) sont attendues et intuitives, mais restent faibles.

#### Méthode de Spearman

Les corrélations sont globalement faibles (valeurs proches de 0), ce qui suggère une faible relation linéaire ou monotone entre les variables. Une corrélation légèrement positive est observée entre age et experience (0.21), ce qui est attendu car l'expérience tend à augmenter avec l'âge. Les autres corrélations sont négligeables (< 0.2), notamment entre city\_development\_index et les autres variables.

#### Conclusion partielle

Les matrices de Pearson et Spearman présentées plus haut montrent qu’il y’a pas de corrélation forte entre les variables, dans la partie suivante nous allons explorer les relations entre les variables après projection dans le plan.

# 5. Analyse en composantes principales (ACP)

## a. Sélection de variables

### 1. 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.

### 3. 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.

## b. Importance des composantes principales issues de l'ACP

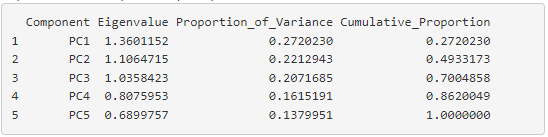


Figure 22 : Importance des composantes principales.

### 1. 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.

### 2. 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.

### 3. 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.

### 4. 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.

### 5. Pourcentage des variances expliquées par les composantes principales

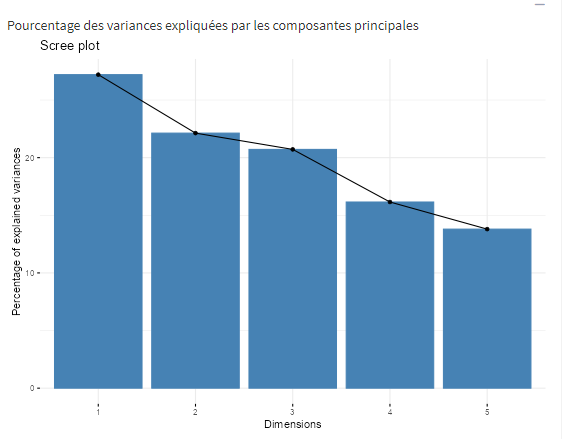


Figure 23 : Pourcentage des variances expliquées.

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.

### 6. Plans factoriels des individus

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

Description générée automatiquement

Une image contenant diagramme, texte, cercle, ligne

Description générée automatiquement

Figure 24 : Répartition des individus sur le plan factoriel.

#### a. 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.

#### b. 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.
* Le graphique à droite mon la répartition des individus suivant la variable gender.

### 7. Cercle de corrélation

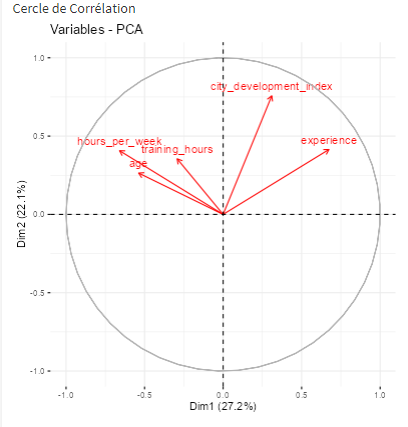


Figure 25 : Cercle des corrélations.

#### a. 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**.

#### b. 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.

### 8. Graphiques des Individus et Variables

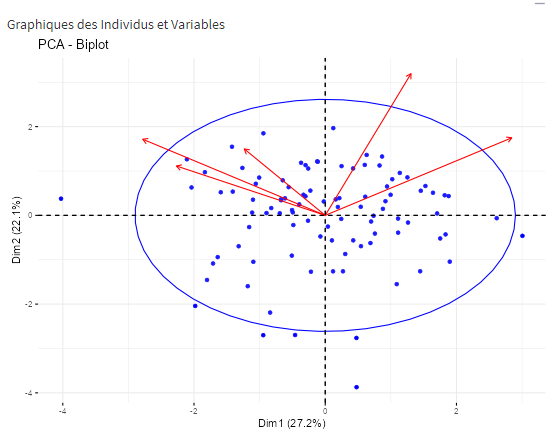


Figure 26: Graphique des individus et variables.

#### a. 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.

#### b. 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.

# K-means

## 1. Choix du nombre de clusters

Une image contenant texte, ligne, Tracé, diagramme

Description générée automatiquement

Figure 27 : Slider pour choix des clusters.

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

### a. 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.

### b. 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.

### c. 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.

### d. 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.

### e. Clusters obtenus avec k = 3

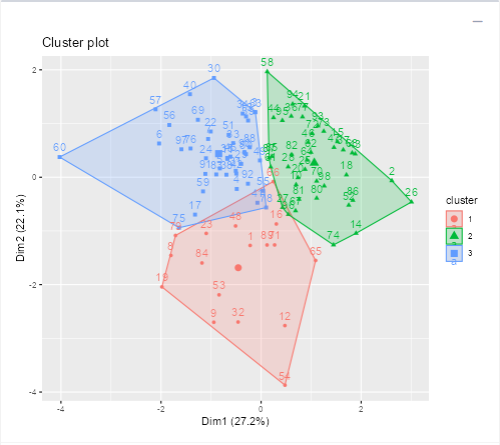


Figure 28: Clusters obtenus avec k = 3.

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

### f. 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.

### g. 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.

### h. 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).

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

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

Description générée automatiquement

Figure 29 : Distribution suivante Age et training\_hours

### a. 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.

### b. 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**.

### c. 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.

### d. Conclusion partielle

Dans cette partie il était question de réaliser la méthode K-means, nous retenons qu’avec k = 3 pour le nombre de clusters nous avons une séparation des groupes optimales. Dans la suite il sera question de faire une classification historique ascendante pour séparer les différents groupes.

# Classification Hiérarchique Ascendante

## 1. Dendrogramme global de la CAH

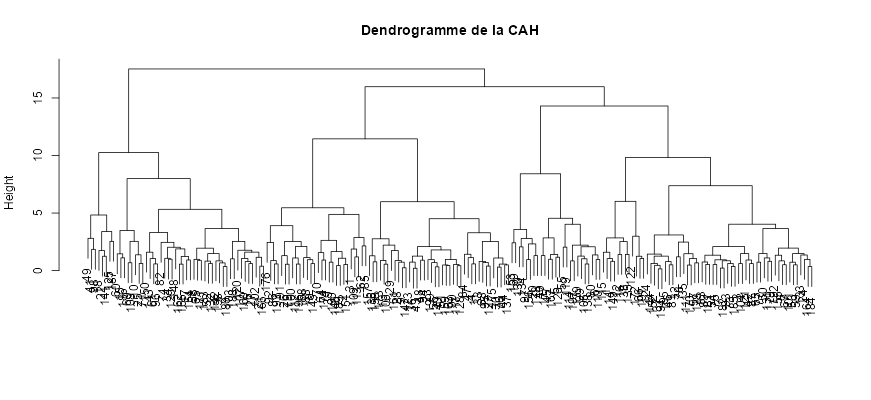


Figure 30 : Dendrogramme de la CAH.

### a. 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.

### b. 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.

## 2. Dendrogramme découpé en 3 clusters

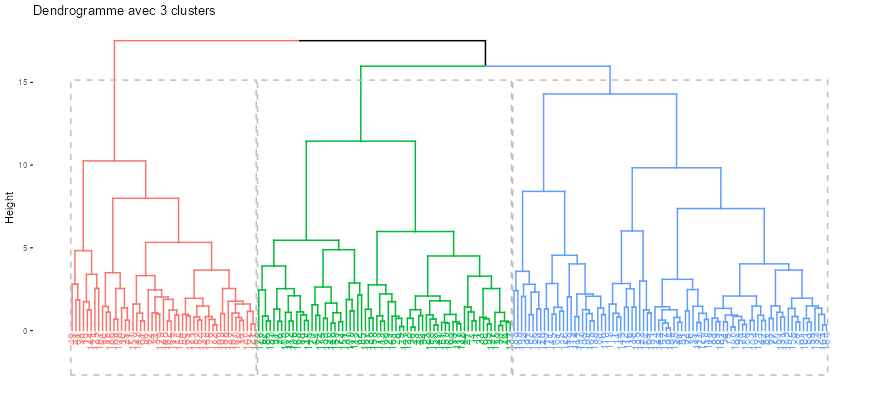


Figure 31: Dendrogramme après coupure.

### a. 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.

### b. Séparation des clusters

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

Justification du choix de 3 clusters :

### c. Observation des distances

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

### d. 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.

### e. Conclusion partielle

Après réalisation de la Classification hiérarchique ascendante nous retenons 3 groupes après coupure du dendogramme présenté plus haute ce qui est en phase avec l’analyse faite dans K-means.

# Synthèse et conclusion

Ce rapport avait pour objectif d'explorer les caractéristiques distinctives des profils de candidats en data science. Grâce à une combinaison d'analyses statistiques exploratoires, nous avons mis en évidence des patterns significatifs et des regroupements naturels au sein de cette population.

L'analyse en composantes principales a révélé que les variables quantitatives telles que l'âge, les heures de formation et l'indice de développement urbain jouent un rôle prépondérant dans la différenciation des profils. Ces facteurs soulignent l'importance de l'expérience professionnelle et de l'engagement dans des formations continues pour se démarquer dans ce domaine.

Les méthodes de clustering, quant à elles, ont permis d'identifier des groupes homogènes de candidats, caractérisés par des combinaisons spécifiques de variables sociodémographiques et professionnelles. Nous avons ainsi pu mettre en évidence trois profils types :

* **Les profils « jeunes et en formation »** : Ces candidats, souvent célibataires et débutants, investissent massivement dans leur développement professionnel en consacrant un temps important à la formation.
* **Les profils « expérimentés et actifs »** : Ce groupe se compose d'individus plus âgés, généralement en poste, qui bénéficient d'une expérience solide mais dont l'engagement en formation est moins marqué.
* **Les profils « hétérogènes »** : Ce dernier groupe rassemble des individus aux profils plus variés, ne se rattachant pas clairement à l'un des deux groupes précédents.

# 7. Annexes

## 1. Analyse bivariée - Matrice de dispersion

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

Description générée automatiquement

Figure 32 : Analyse bivariée - Matrice de dispersion.

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 l’âge : 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.
* Les variables sont relativement indépendantes les unes des autres, à l'exception de quelques tendances légères (par exemple, âge vs expérience).

## 2. Tests Statistiques

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

Description générée automatiquement

Figure 33 : Analyse des Tests Statistiques

* **city\_development\_index vs gender:**
  + **p-value** : 0.131
  + La p-value est supérieure à 0.05, ce qui suggère qu'il n'y a pas de relation statistiquement significative entre l'indice de développement urbain et le sexe.
* **experience vs relevant\_experience :**
  + **p-value** : 1.03e-10
  + La p-value est très faible, ce qui indique une relation statistiquement significative entre l'expérience et l'expérience pertinente.
* **training\_hours vs education\_level :**
  + **p-value** : 0.88
  + La p-value est supérieure à 0.05, ce qui suggère qu'il n'y a pas de relation statistiquement significative entre les heures de formation et le niveau d'éducation.
* **hours\_per\_week vs marital\_status :**
  + **p-value** : 1.23e-05
  + La p-value est très faible, ce qui indique une relation statistiquement significative entre le nombre d'heures travaillées par semaine et le statut marital.
* **age vs workclass :**
  + **p-value** : 0.000117
  + La p-value est très faible, ce qui indique une relation statistiquement significative entre l'âge et la catégorie socioprofessionnelle.

Résumé

* En résumé, seuls les tests impliquant **experience vs relevant\_experience**, **hours\_per\_week vs marital\_status** et **age vs workclass** montrent des relations statistiquement significatives, tandis que les autres comparaisons ne le sont pas.

## 3. Corrélations entre Variables Quantitatives

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

Description générée automatiquement

Figure 34 : Corrélation entre les variables quantitatives.

#### a. city\_development\_index (Indice de développement de la ville) :

* + **Faible corrélation avec experience** (*r = 0.21*) : Les candidats des villes à indice élevé pourraient 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.

#### b. experience (Expérience totale) :

* + **Très faible corrélation avec training\_hours** (*r = 0.02*) et **age** (*r = 0.02*) : Cela suggère 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 indique que les candidats avec plus d'expérience travaillent légèrement moins d'heures, bien que l'effet soit faible.

#### c. 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 reflète que les candidats plus âgés ont tendance à suivre plus de formations.

#### d. 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.

#### e. 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.

#### f. Conclusion

* 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.

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

Description générée automatiquement

Figure 35: Analyse entre Age et City\_development\_index

Dans ce graphique on observe que certaines zones ont plus de points et ces zones influencent davantage la pente de la droite, de plus la corrélation est de 0.02 qui est presque nulle. Cela indique qu’il n’existe pas de relation linéaire significative entre les deux variables.

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

Description générée automatiquement

Figure 36: Analyse entre Age et Training\_hours

#### g. Analyse et interprétation

* Corrélation : Le coefficient de corrélation de 0.14 indique une relation très faible et positive entre l'âge et les heures de formation. En d'autres termes, il existe une légère tendance montrant que les heures de formation augmentent légèrement avec l'âge, mais cette relation est loin d'être significative.
* Répartition des données : La majorité des points se concentrent dans les jeunes tranches d'âge (moins de 40 ans). Les heures de formation varient considérablement pour tous les âges, avec certains individus ayant jusqu'à 300 heures de formation, indépendamment de leur âge.
* Ligne de tendance : La pente de la ligne de régression est faible, ce qui confirme que l'âge a un impact minime sur les heures de formation. L'intervalle de confiance (zone grisée) montre une incertitude importante dans la prédiction des heures de formation en fonction de l'âge.

## 4. Corrélations entre Variables Qualitatives

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

Description générée automatiquement

Figure 37 : Répartition des valeurs manquantes par colonne.

Certaines colonnes, comme gender, major\_discipline et company\_size, présentent un nombre significatif de valeurs manquantes. Cela signifie qu'une grande partie des lignes de données ne contient pas d'information pour ces colonnes spécifiques. Colonnes avec peu de valeurs manquantes : À l'inverse, d'autres colonnes comme experience, enrolled\_university ont très peu de valeurs manquantes. Ces colonnes sont donc plus complètes et pourraient être plus fiables pour l'analyse.

Une image contenant diagramme, ligne, Tracé, Plan

Description générée automatiquement

Figure 38 : Analyse univariée – Histogramme de Hours\_per\_week.

L'histogramme nous donne une vision d'ensemble de la distribution du nombre d'heures travaillées par semaine dans l'échantillon. On peut observer que :

* **Distribution asymétrique à droite :** La majorité des individus travaillent moins d'heures, et la queue de distribution s'étend vers les nombres d'heures plus élevés. Cela signifie qu'il y a plus de personnes travaillant un nombre d'heures faible ou moyen dans l'échantillon.
* **Pic de la distribution autour de 40 heures :** Le nombre d'heures le plus fréquent dans l'échantillon est d'environ 40 heures.
* **Étendue des heures travaillées :** Les nombres d'heures travaillées se situent entre environ 20 et 100 heures, ce qui donne une idée de la diversité des horaires dans l'échantillon.
* **Résumé :** La distribution est asymétrique à droite, ce qui indique une plus grande proportion d'individus travaillant à temps plein ou à temps partiel, et moins d'individus effectuant de nombreuses heures supplémentaires. Cela reflète une diversité dans les contrats de travail et les modes de vie des individus de l'échantillon.

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

Description générée automatiquement

Figure 39 : Analyse univariée - Boxplot de Hours\_per\_week.

Le boxplot nous fournit des informations plus précises sur les quantiles de la distribution, notamment :

* **Médiane autour de 40 heures :** La ligne centrale de la boîte, représentant la médiane, est située autour de 40 heures. Cela signifie que la moitié des individus travaillent moins de 40 heures par semaine et l'autre moitié en travaille 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 horaire 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 travaillant plus d'heures que la moyenne, mais moins d'individus travaillant beaucoup moins d'heures.
* **Outliers :** Les points isolés à gauche et à droite de la boîte représentent des individus dont le nombre d'heures travaillées est considéré comme une valeur aberrante (outlier), étant nettement inférieur ou supérieur à la majorité des autres individus.

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

Description générée automatiquement

Figure 40: Table de contingence par genre et par statut d’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), on voit qu'il y a bien plus de personnes qui ne sont pas inscrites.

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

Description générée automatiquement

Figure 41 : Test du Chi Carré de Pearson.

Test du Chi Carré de Pearson, Degrés de liberté (df) : 4, p-value : 0.228.

Interprétation

* La p-value est supérieure à 0.05, ce qui indique qu'il n'y a pas de relation statistiquement significative entre le sexe (**gender**) et le statut d'inscription à une université (**enrolled\_university**).
* En d'autres termes, on ne peut pas rejeter l'hypothèse nulle selon laquelle les deux variables sont indépendantes.

Une image contenant texte, capture d’écran, ligne

Description générée automatiquement

Figure 42 : Table de contingence par genre et par niveau d'éducation.

La table de contingence indique les effectifs par genre et niveau d'éducation (Graduate, Masters, Primary School). On voit qu'il y a bien plus d'effectifs ayant un niveau d'éducation 'graduate'

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

Description générée automatiquement

Figure 43 : Test du Chi carré de Pearson pour gender et education\_level

Test du Chi Carré de Pearson : Degrés de liberté (df) : 6, p-value : 0.7308

Interprétation

* La p-value est supérieure à 0.05, ce qui indique qu'il n'y a pas de relation statistiquement significative entre le sexe (**gender**) et le niveau d'éducation (**education\_level**).
* En d'autres termes, l'hypothèse nulle (selon laquelle les deux variables sont indépendantes) ne peut pas être rejetée, ce qui suggère que ces deux variables sont indépendantes l'une de l'autre dans notre échantillon de données.

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

Description générée automatiquement

Figure 44 : Test du Chi carré de Pearson pour last\_new\_job et marital\_status

Test du Chi Carré de Pearson avec p-value simulée : p-value : 0.5432.

Interprétation

* La p-value est supérieure à 0.05, ce qui suggère qu'il n'y a pas de relation statistiquement significative entre le dernier emploi occupé (**last\_new\_job**) et le statut marital (**marital\_status**).
* Ainsi, l'hypothèse nulle (selon laquelle les deux variables sont indépendantes) ne peut pas être rejetée. Cela signifie que, dans notre jeu de données, le statut marital et le dernier emploi occupé sont indépendants l'un de l'autre.

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

Description générée automatiquement

Figure 45 : Test du Chi carré de Pearson avec P-value simulée pour Worklass et marital\_status

Test du Chi Carré de Pearson avec p-value simulée : Degrés de liberté (df) : Non spécifiés (NA), calculés de manière simulée, p-value : 0.04598

Interprétation

* La p-value est inférieure à 0.05, ce qui indique une relation statistiquement significative entre la catégorie socioprofessionnelle (**workclass**) et le statut marital (**marital\_status**).
* En d'autres termes, l'hypothèse nulle (selon laquelle les deux variables sont indépendantes) peut être rejetée. Cela suggère que ces deux variables sont liées dans notre jeu de données.

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

Description générée automatiquement

Figure 46 : Test du Chi carré de Pearson avec P-value simulée pour Worklass et relevent\_experience

Test du Chi Carré de Pearson avec p-value simulée :

* Degrés de liberté (df) : Non spécifiés (NA), calculés de manière simulée, p-value : 0.3033

Interprétation

* La p-value est supérieure à 0.05, ce qui indique qu'il n'y a pas de relation statistiquement significative entre l'expérience pertinente (**relevant\_experience**) et la catégorie socioprofessionnelle (**workclass**).
* L'hypothèse nulle (selon laquelle les deux variables sont indépendantes) ne peut pas être rejetée, ce qui suggère que ces deux variables sont indépendantes.

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

Description générée automatiquement

Figure 47 : Test du Chi carré de Pearson avec p-valeu simulée pour race et company\_type

Test du Chi Carré de Pearson avec p-value simulée : Degrés de liberté (df) : Non spécifiés (NA), calculés de manière simulée, p-value : 0.6442

Interprétation

* La p-value est supérieure à 0.05, ce qui indique qu'il n'y a pas de relation statistiquement significative entre la race (**race**) et le type d'entreprise (**company\_type**).
* L'hypothèse nulle (selon laquelle les deux variables sont indépendantes) ne peut pas être rejetée, ce qui suggère que ces deux variables sont indépendantes.

# 8. Guide d'installation et d'utilisation de l'application Shiny

Ce guide vous permettra de configurer et d'utiliser correctement l'application Shiny associée à cette étude.

## 1. Préparation des fichiers

1. Téléchargez l’archive contenant l’application.
2. Décompressez le dossier ZIP à l'emplacement de votre choix. Assurez-vous que tous les fichiers nécessaires sont extraits et bien organisés.

## 2. Lancement de l’application

1. **Ouvrez R ou RStudio** :
   * Si RStudio n'est pas installé, vous pouvez le télécharger depuis [le site officiel](https://www.rstudio.com).
2. **Ouvrez le fichier principal** :
   * Dans RStudio, ouvrez le fichier MainApp.R situé dans le dossier décompressé.
3. **Lancez l'application** :
   * Cliquez sur le bouton "Run App" en haut à droite de l’interface RStudio ou exécutez la commande suivante dans la console : source("MainApp.R")

## 3. Chargement des données

Une fois que l’interface de l’application s’affiche dans votre navigateur ou dans une fenêtre dédiée :

1. Naviguez dans l’interface jusqu’à la section de chargement des données.

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

Description générée automatiquement

Figure 48: Bouton pour charger le jeu de données.

1. **Sélectionnez le fichier de données** :
   * Rendez-vous dans le dossier data (inclus dans le dossier décompressé).
   * Chargez le fichier **data\_final.csv** (utilisé dans le cadre de cette étude).

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

Figure 49: Jeu de données à charger.

1. Une fois les données chargées correctement, vous pourrez explorer toutes les fonctionnalités de l'application.

## 4. Exploration de l’application

Une fois les données chargées, vous pouvez :

* **Visualiser et analyser les données** grâce aux graphiques et aux outils proposés par l'application.
* Parcourir les différentes sections de l'application pour explorer les résultats de l’analyse statistique réalisée.

## 5. Exploration des codes

Si vous souhaitez consulter les codes de l'application, vous pouvez explorer les fichiers dans le dossier. Voici un quelques exemples de leur contenu :

* **ACP.r** : Contient le code lié à l’analyse en composantes principales (ACP).
* **Kmeans.r** : Traite l'algorithme de segmentation K-means.
* **CAH.r** : Contient le code lié à la classification ascendante hiérarchique (CAH).
* **Autres fichiers** :
  + **analyseBiVariee.r** : Analyse bivariée des données.
  + **analyseUniVariee.r** : Analyse univariée des données.

Ces fichiers sont accessibles directement depuis le dossier décompressé.

## 6. Option : Cloner le dépôt GitHub

Si vous préférez récupérer le code directement depuis GitHub :

1. Copiez le [lien du dépôt GitHub](https://github.com/loicosquare/Projet-STA101/tree/Branche_stable).
2. Clonez le dépôt dans votre machine en exécutant la commande suivante dans un terminal ou dans RStudio : git clone <URL\_DU\_DEPOT\_GITHUB>
3. Accédez au dossier cloné, ouvrez le fichier MainApp.R dans R ou RStudio, puis suivez les étapes mentionnées précédemment pour lancer l’application.

## 7. Support

En cas de problème, assurez-vous :

* Que tous les fichiers nécessaires (scripts, données, etc.) sont présents.
* Que les packages requis pour exécuter l’application sont installés dans R. Vous pouvez installer les dépendances en exécutant : install.packages(c("librairie1", "librairie2", "librairie3", etc..))

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement

Figure 50 : Code utilisé pour construire l'échantillon.