Analyse des facteurs influençant le salaire : Étude du dataset CPS 1985

Table des matières

1. Description du dataset

Le dataset utilisé dans cette analyse provient de l'enquête Current Population Survey (CPS) de 1985. Il contient des informations sur les salaires et les caractéristiques démographiques et professionnelles de 534 individus aux États-Unis.
Le dataset contient un mélange de variables quantitatives et qualitatives, permettant d'étudier les facteurs influençant le salaire tels que l'éducation, l'expérience, le secteur d'activité, le genre et l'adhésion syndicale

1.1. Présentation des variables

Le dataset comporte 12 variables, dont certaines sont numériques (quantitatives) et d'autres catégorielles (qualitatives). Voici leur description :

Nom	Type	Description	valeurs
rownames	Numérique	Identifiant unique de l'individu	1, 2, 3, ... 534
wage	Numérique	Salaire horaire de l'individu (en dollars)	4.00, 6.67, 7.50
education	Numérique	Nombre d'années d'éducation complétées	8, 12, 16
experience	Numérique	Nombre d'années d'expérience professionnelle	1, 10, 21
age	Numérique	Âge de l'individu	19, 35, 57
ethnicity	Catégorielle	Origine ethnique de l'individu	*cauc*, *hispanic*
region	Catégorielle	Région géographique de résidence	other
gender	Catégorielle	Sexe de l'individu	*male*, *female*
occupation	Catégorielle	Catégorie d'emploi occupée	worker
sector	Catégorielle	Secteur d'activité	*manufacturing*, *other*
union	Catégorielle	Adhésion à un	*yes*, *no*

		syndicat	
married	Catégorielle	Statut matrimonial	*yes*, *no*

1.2. Analyse Descriptive

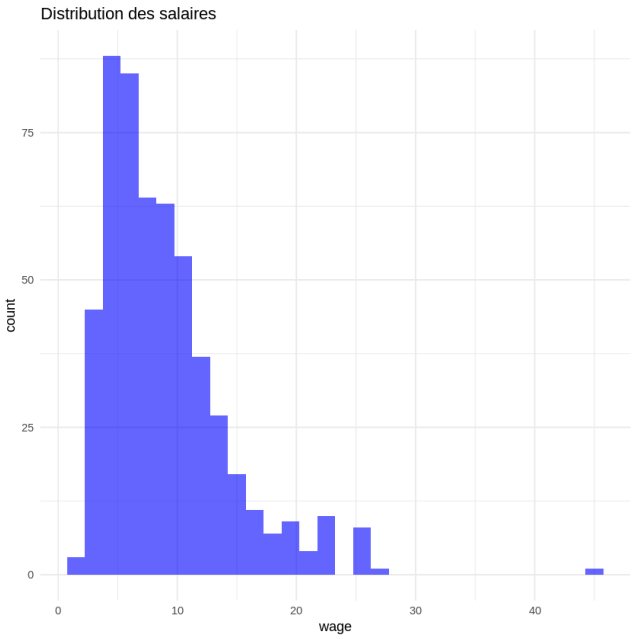
a). Statistiques descriptives des variables numériques

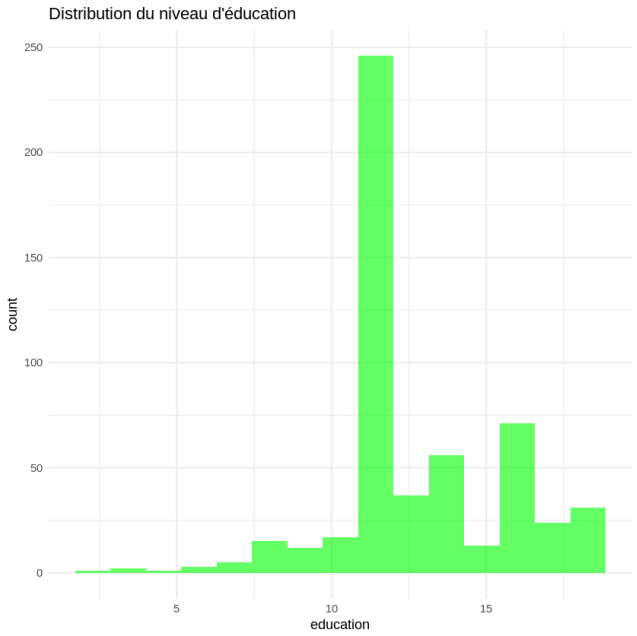
Les statistiques descriptives des variables numériques (wage, education, experience, age) sont résumées ci-dessous :

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
wage	1.00	5.25	7.78	9.02	11.25	44.50
education	2.00	12.00	12.00	13.02	15.00	18.00
education	0.00	8.00	15.00	13.02	26.00	55.00
age	18.00	28.00	35.00	36.83	44.00	64.00

- **Salaire** (wage) : Le salaire minimum est de 1.00, tandis que le salaire maximum est de 44.50. La médiane (7.78) est légèrement inférieure à la moyenne (9.02), ce qui suggère une distribution légèrement asymétrique à droite, avec quelques salaires élevés tirant la moyenne vers le haut.
- **Éducation** (education) : Le niveau d'éducation varie de 2 à 18 ans. La médiane et le premier quartile sont à 12 ans, ce qui indique que la majorité des individus ont au moins un niveau d'éducation secondaire. La moyenne est de 13.02 ans, ce qui suggère que certains individus ont un niveau d'éducation supérieur.
- **Expérience** (experience) : L'expérience professionnelle varie de 0 à 55 ans. La médiane est de 15 ans, ce qui indique que la moitié des individus ont plus de 15 ans d'expérience. La moyenne est légèrement plus élevée (17.82 ans), ce qui suggère une distribution légèrement asymétrique à droite.
- **Âge** (age) : L'âge des individus varie de 18 à 64 ans. La médiane est de 35 ans, ce qui indique que la moitié des individus ont plus de 35 ans. La moyenne est de 36.83 ans, ce qui suggère une distribution relativement symétrique.

b. Distribution des salaires (wage) et du niveau d'éducation (education)

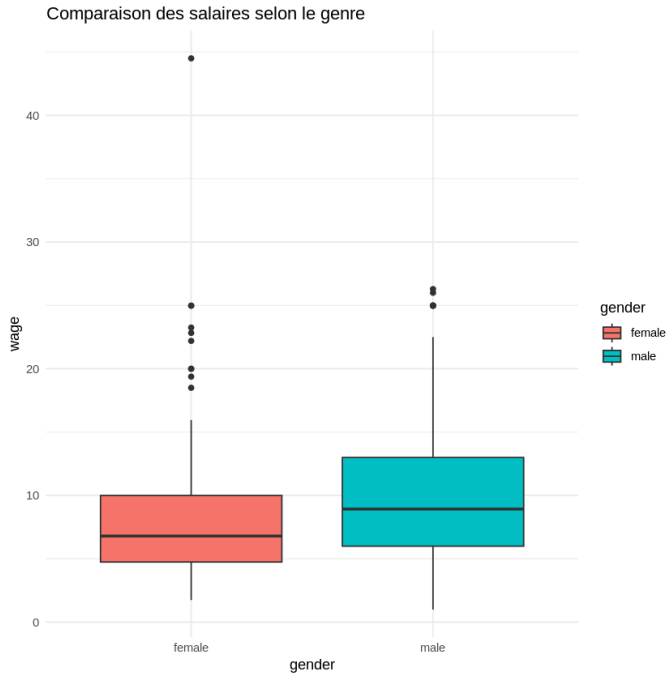




- **Distribution des salaires (wage):** Le graphique de la distribution des salaires montre une asymétrie à droite, avec une concentration des salaires dans les valeurs basses (entre 0 et 15). Quelques valeurs extrêmes (salaire > 30) tirent la queue de la distribution vers la droite. Cela confirme l'observation faite dans les statistiques descriptives, où la moyenne est supérieure à la médiane.
  - La majorité des individus ont un salaire relativement bas, mais il existe quelques individus avec des salaires beaucoup plus élevés, ce qui peut indiquer des inégalités salariales dans le dataset.

- **Distribution du niveau d'éducation:** Le graphique de la distribution du niveau d'éducation montre une concentration autour de 12 ans, ce qui correspond probablement à un diplôme d'études secondaires. Il y a également un pic autour de 16 ans, ce qui pourrait correspondre à un diplôme universitaire.
  - La plupart des individus ont un niveau d'éducation secondaire, mais un nombre significatif a poursuivi des études supérieures. Cela peut influencer les salaires et l'expérience professionnelle.

c. Comparaison des salaires selon le genre (gender)



Le boxplot comparant les salaires selon le genre montre que :

- Les hommes ont une médiane de salaire légèrement plus élevée que les femmes.
- La distribution des salaires des hommes est plus étendue, avec des valeurs extrêmes plus élevées que celles des femmes.

Il semble y avoir une disparité salariale entre les genres, avec les hommes ayant tendance à gagner plus que les femmes en moyenne.

d. La matrice de corrélation

- **Corrélation entre wage (salaire) et education (niveau d'éducation) (Valeur : 0.38):** Il existe une corrélation positive modérée entre le salaire et le niveau d'éducation. Cela

signifie que, en général, les individus avec un niveau d'éducation plus élevé ont tendance à gagner plus. Cette corrélation est attendue, car l'éducation est souvent un facteur clé dans la détermination du salaire.

- **Corrélation entre wage (salaire) et experience (expérience professionnelle) (Valeur : 0.09):** La corrélation entre le salaire et l'expérience professionnelle est très faible. Cela suggère que l'expérience professionnelle, dans ce dataset, n'a pas un impact significatif sur le salaire. Cela peut être surprenant, car on s'attend généralement à ce que l'expérience influence positivement le salaire.
- **Corrélation entre wage (salaire) et age (âge) (Valeur : 0.18):** Il existe une corrélation positive faible entre le salaire et l'âge. Cela signifie que, en moyenne, les individus plus âgés ont tendance à gagner un peu plus que les plus jeunes. Cependant, cette corrélation est faible, ce qui indique que l'âge n'est pas un facteur déterminant du salaire dans ce dataset.

2. Prétraitement des données

Le prétraitement des données est une étape cruciale pour garantir la qualité et la pertinence des analyses. Dans cette section, nous décrivons les différentes étapes de prétraitement appliquées au dataset.

2.1 Suppression de variables non pertinentes

Trois variables ont été supprimées du dataset, car elles n'apportent pas d'informations pertinentes pour l'analyse :

- **rownames** : Cette variable représente uniquement un identifiant pour chaque individu. Elle ne contient aucune information exploitable pour l'analyse et a donc été supprimée.
- **region** : Cette variable ne comporte qu'une seule modalité (other) pour l'ensemble des individus. Une variable constante n'apporte aucune variabilité et ne permet donc pas de différencier les observations. Par conséquent, elle a été exclue de l'analyse.
- **occupation** : Tous les individus de l'échantillon sont classés dans la même catégorie (worker). Comme cette variable est uniforme, elle ne contribue pas à l'explication des variations de salaire ou d'autres caractéristiques étudiées. Elle a donc été supprimée.

2.2. Vérification des données manquantes

Avant de procéder à l'analyse, nous avons vérifié la présence de données manquantes dans le dataset. Aucune valeur manquante n'a été détectée dans les colonnes restantes. Cela signifie que toutes les observations sont complètes et peuvent être utilisées pour les analyses ultérieures.

2.3. Standardisation des variables numériques

Pour faciliter la comparaison des variables numériques et améliorer la performance des modèles statistiques (notamment la régression linéaire et l'ACP), nous avons standardisé les variables numériques. La standardisation consiste à transformer les données de sorte que chaque variable ait une moyenne de 0 et un écart-type de 1. Cette transformation est particulièrement utile lorsque les variables ont des échelles différentes (par exemple, l'âge en années et le salaire en dollars).

Les variables numériques standardisées sont les suivantes :wage (salaire), education (niveau d'éducation), experience (expérience professionnelle),age (âge)

La standardisation a été réalisée en utilisant la formule suivante pour chaque variable :

$$z = \frac{x - \mu}{\sigma}$$

ou x est la valeur originale, μ est la moyenne de la variable, σ est l'écart-type de la variable.
Cette étape permet de s'assurer que toutes les variables numériques sont sur la même échelle, ce qui est essentiel pour les analyses multivariées comme la régression linéaire et l'ACP.

3. Régression linéaire

3.1. Etapes suivies

On a effectué une régression linéaire pour prédire le salaire (wage) en fonction de plusieurs variables explicatives. Voici les étapes que nous avons suivies :

- **Prétraitement des données :** On a converti les variables qualitatives (gender, sector, union) en facteurs à l'aide de as.factor(). Cela permet à R de traiter ces variables comme des variables catégorielles dans le modèle de régression. Les variables numériques (education, experience, age) ont été laissées telles quelles, car elles sont déjà adaptées à la régression linéaire.
- **Création du modèle :** On a utilisé la fonction lm() pour créer un modèle de régression linéaire. La formule wage ~ education + experience + age + gender + sector + union indique que : wage est la variable dépendante (cible) et education, experience, age, gender, sector, et union sont les variables indépendantes (explicatives).
- **Résumé du modèle :** On a utilisé summary(model) pour afficher les résultats détaillés de la régression linéaire, y compris : Les coefficients estimés pour chaque variable, les erreurs standards, les valeurs de t, et les p-values. Et le R2 (coefficient de détermination) qui mesure la qualité de l'ajustement du modèle.

```r
# ------------------ REGRESSION LINEAIRE ------------------
dfR <- df %>%
  mutate(
    gender = as.factor(gender),
    sector = as.factor(sector),
    union = as.factor(union)
  )
# Modèle de régression linéaire : salaire en fonction de plusieurs variables
model <- lm(wage ~ education + experience + age + gender + sector + union, data = dfR)
summary(model)  # Résumé du modèle
```

3.2. Interprétation des résultats:

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.298  -2.701  -0.667   1.960  37.923

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -4.0433     6.8025  -0.594  0.55251
education             1.3600     1.1099   1.225  0.22099
experience            0.5105     1.1102   0.460  0.64585
age                  -0.4045     1.1092  -0.365  0.71550
gendermale            2.0758     0.3971   5.228 2.48e-07 ***
sectormanufacturing   0.6354     1.0163   0.625  0.53211
sectorother          -0.4361     0.9554  -0.457  0.64822
unionyes              1.4324     0.5095   2.811  0.00512 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.416 on 526 degrees of freedom
Multiple R-squared:  0.2713,    Adjusted R-squared:  0.2616
F-statistic: 27.97 on 7 and 526 DF,  p-value: < 2.2e-16
```

- **Qualité du modèle :**
  - Le coefficient de détermination R2 est de 27.13 %, ce qui signifie que 27.13 % de la variabilité du salaire est expliquée par les variables incluses dans le modèle. Cela indique que le modèle explique une partie modérée de la variation du salaire, mais une grande part de variabilité reste inexpliquée. Cela suggère que d'autres facteurs, non pris en compte dans ce modèle, pourraient également influencer le salaire.
  - De plus, la F-statistique est de 27.97 avec une p-value < 2.2e-16, ce qui indique que le modèle dans son ensemble est statistiquement significatif (p<0.001). Cela confirme qu'au moins une des variables explicatives a un effet significatif sur le salaire.

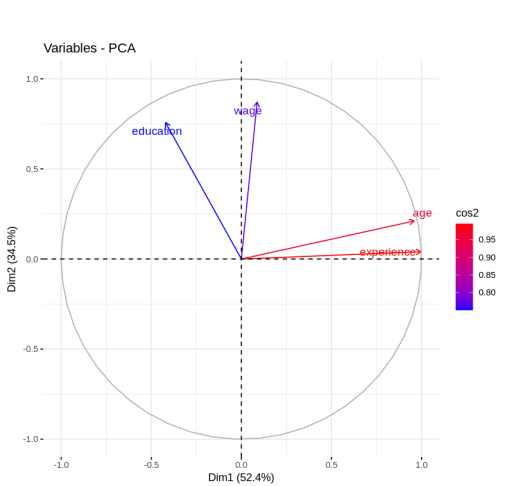
- **Variables significatives :** Seules les variables gendermale et unionyes ont un impact statistiquement significatif sur le salaire, avec des p-values inférieures à 0.05. Par conséquent, l'hypothèse nulle est rejetée pour ces variables, et nous pouvons conclure que :
    - Les hommes gagnent significativement plus que les femmes. Le coefficient de gendermale est positif, ce qui indique que les hommes ont un salaire horaire plus élevé que les femmes.
    - Les individus syndiqués gagnent significativement plus que les non-syndiqués. Le coefficient de unionyes est également positif, ce qui suggère que l'appartenance à un syndicat est associée à un salaire horaire plus élevé.
- **Variables non significatives :**Les variables education, experience, age, sectormanufacturing, et sectorother n'ont pas d'impact significatif sur le salaire dans ce modèle, car leurs p-values sont supérieures à 0.05. Par conséquent, l'hypothèse nulle est acceptée pour ces variables, ce qui signifie qu'elles ne contribuent pas de manière significative à l'explication du salaire dans ce contexte.

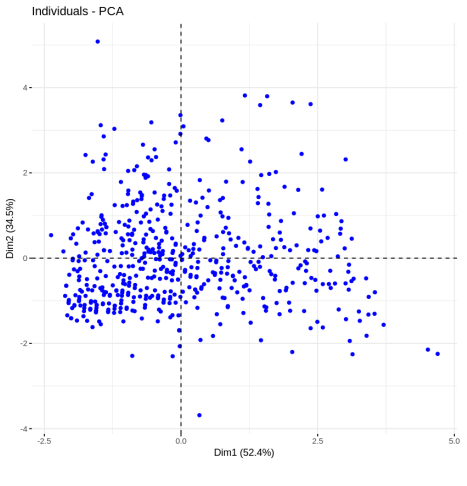
4. Analyse en Composantes Principales

4.1. Étapes suivies

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

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





4.2. Interprétation des résultats

a. Valeurs propres et variance expliquée

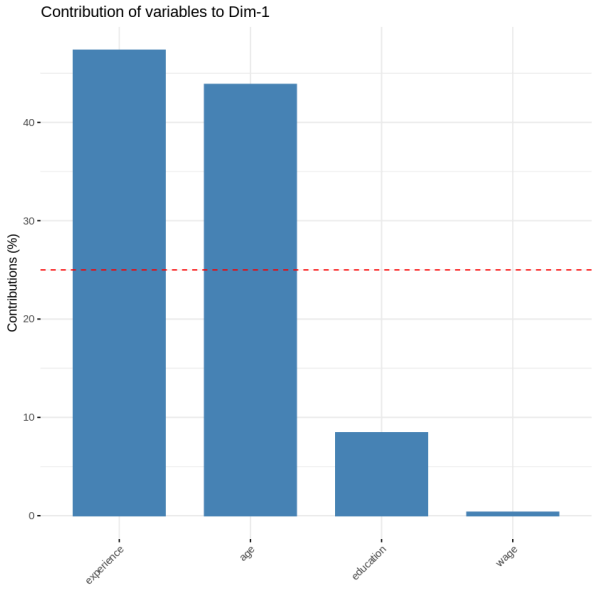
L'ACP a généré quatre composantes principales (axes), mais nous nous concentrons sur les deux premières, qui expliquent la majeure partie de la variance :

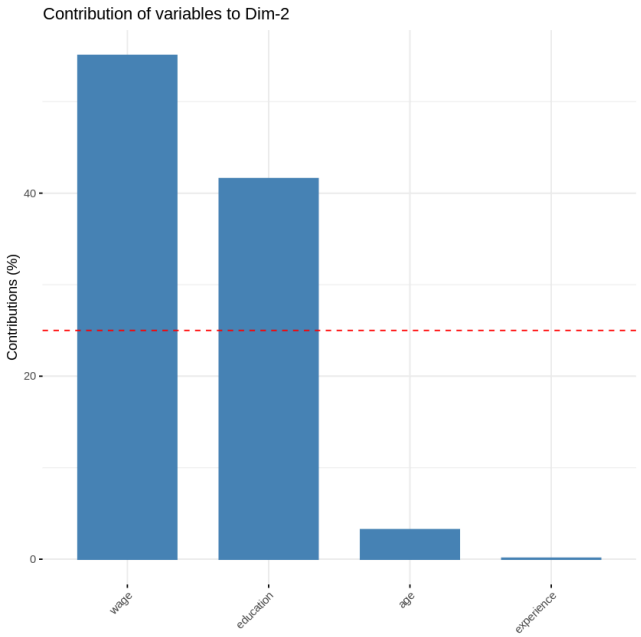
- Axe 1 (Dim1) : Explique 52.4 % de la variance totale.
- Axe 2 (Dim2) : Explique 34.5 % de la variance totale.

Ensemble, ces deux axes expliquent 86.9 % de la variabilité totale des données, ce qui est très satisfaisant pour une analyse en deux dimensions.

b. Contributions des variables aux axes

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





Variable	Contributions à l'axe 1 (Dim1)
experience	Contribution de 47.34 %
age	Contribution de 43.86 %
education	Contribution de 8.44 %
wage	Contribution de 0.36 %

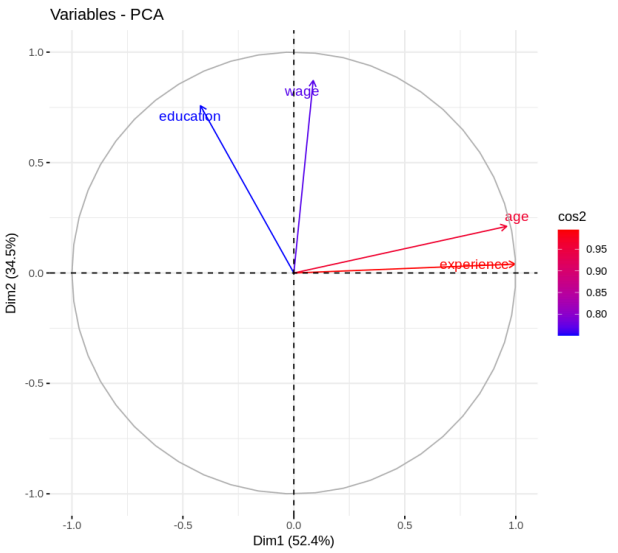
- **Interprétation :** L'axe 1 est principalement expliqué par les variables experience et age. Cet axe représente une dimension liée à l'expérience professionnelle et à l'âge.
  - Les individus avec des coordonnées positives sur cet axe ont tendance à avoir plus d'expérience et à être plus âgés, tandis que ceux avec des coordonnées négatives ont moins d'expérience et sont plus jeunes.

Variable	Contributions à l'axe 2 (Dim2)
wage	Contribution de 55.06 %
education	Contribution de 41.59 %
age	Contribution de 3.23 %
experience	Contribution de 0.12 %

- **Interprétation :** L'axe 2 est principalement expliqué par les variables wage et education. Cet axe représente une dimension socio-économique liée au salaire et au niveau d'éducation.
  - Les individus avec des coordonnées positives sur cet axe ont tendance à avoir des salaires plus élevés et un niveau d'éducation plus élevé, tandis que ceux avec des coordonnées négatives ont des salaires plus faibles et un niveau d'éducation plus faible.

c. Visualisation des variables dans le plan factoriel

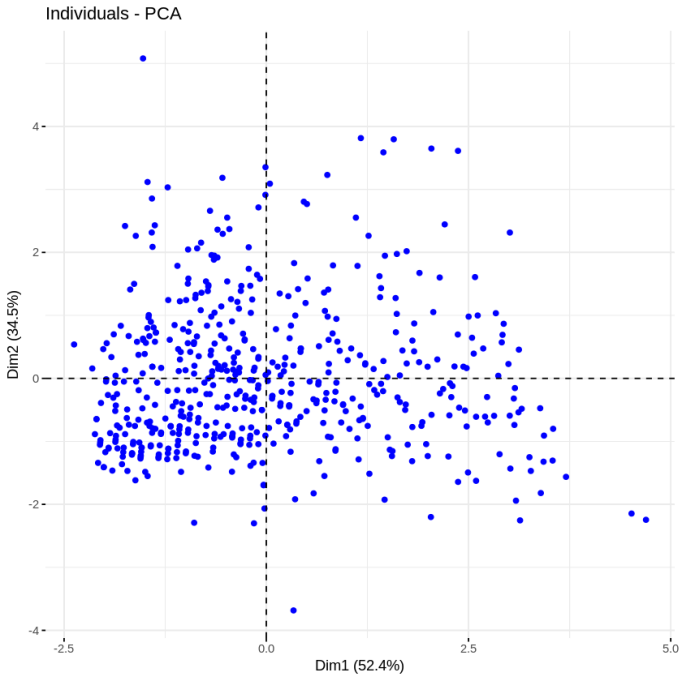
Dans le plan factoriel (Dim1 vs Dim2), les variables sont représentées par des flèches. Les variables proches les unes des autres sont fortement corrélées.



- experience et age sont proches l'une de l'autre, ce qui confirme leur forte corrélation avec l'axe 1.
- wage et education sont proches l'une de l'autre, ce qui confirme leur forte corrélation avec l'axe 2.

d. Visualisation des individus dans le plan factoriel

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



- Les individus avec des coordonnées positives sur l'axe 1 ont plus d'expérience et sont plus âgés.
- Les individus avec des coordonnées positives sur l'axe 2 ont des salaires plus élevés et un niveau d'éducation plus élevé.

5. ANOVA :

Afin de tester si les différences de salaire entre les groupes sont statistiquement significatives, nous avons réalisé une Analyse de la Variance (ANOVA) à deux facteurs (genre et appartenance syndicale).

5.1.Résultats du test ANOVA:

Facteur	Df	Somme des carrés	Moyenne des carrés	F-Value	p-Value

Genre	1	22.5	22.480	23.810	1.41e-06
Union	1	9.2	9.167	9.709	0.00193
Résidus	531	501.4	0.944		

5.2. Interprétation des résultats

- La p-value du facteur "genre" (1.41e-06) est très faible, ce qui signifie que les salaires des hommes et des femmes sont significativement différents.
- La p-value du facteur "union" (0.00193) est également inférieure à 0.05, ce qui indique que l'appartenance syndicale influence aussi significativement le salaire.
- L'analyse confirme que les différences observées dans la section descriptive et la régression sont bien statistiquement significatives.