M1 IA – Parcours Data Science

Année 2022-2023

Rapport du Projet de semestre

Analyse de données avec R

Réalisé par :

GHEZALI Amina
VERA COSIO Ray Leonardo

INDEX

I . Introduction

R est un langage de programmation et un environnement de développement principalement utilisé pour l'analyse statistique et la représentation graphique des données. Il est largement utilisé dans la communauté scientifique des données et dans l'industrie pour effectuer des tâches de nettoyage, de transformation, d'analyse et de visualisation de données. Avec R, il est possible d'utiliser une grande variété de paquets et de bibliothèques spécialisés pour effectuer différents types d'analyses, allant des statistiques de base aux modèles d'apprentissage automatique. RStudio est un environnement de développement intégré (IDE) pour R qui fournit une interface facile à utiliser et des outils supplémentaires pour faciliter le travail avec le langage.

Objectifs

Nous cherchons à nous familiariser avec RStudio. Notre objectif principal est avant tout apprendre à manipuler et connaitre l'environnement. Précisément : Apprendre à utiliser le logiciel R pour analyser des données. Appliquer les méthodes de statistique descriptive, de prise de décision, d'analyse de la variance, de régression linéaire, d'analyse univariée et multivariée de données dans R.

Logiciels et outils de collaboration

RStudio instalé dans l'ordinateur

Pack Office pour le rapport

GitHub pour travailler en collaboration avec mon collègue

Meet pour prendre contact et faire des points.

1.1. Présentation de la Base de données

Cet ensemble de données viens de cette
[Kaggle] https://www.kaggle.com/competitions/kaggle-survey-2020/data

Il contient des informations sur près de 20 036 personnes 2020, 2021 et 2022 qui sont en relation aux metiers numeriques et data. Nous avons 355 colonnes dans notre jeu de données, dont nous avons extrait les colonnes qui concernent : l'âge, le sexe, le pays, le diplôme, le domaine d'action, l'expérience professionnelle, la taille de l'entreprise et le salaire. Tout cela en raison de la pertinence et de l'intérêt de notre étude.

1. 2. Énoncé du problème et problématique

Une étude statistique du salaire moyen basée sur l'âge, le sexe et les conditions de vie impliquerait la collecte de données sur le salaire, l'âge, le sexe et les conditions de vie d'un large échantillon d'individus, puis l'analyse des données pour déterminer les modèles ou les relations entre ces variables. Notre sommes utilisés des techniques statistiques telles que l'analyse de corrélation et de régression pour identifier toute relation significative entre les variables d'intérêt. De plus, l'étude peut également prendre en compte d'autres facteurs susceptibles d'influencer les résultats, tels que le niveau d'éducation, la profession, etc.

Donc notre problématique a été défini de cette façon

« La disparité salariale entre les femmes et les hommes qui exercent des activités dans le domaine de l'informatique en fonction de leurs âges, expériences professionnelles, taille d'entreprise et parcours professionnel »

1.3. Montage expérimental

Nous avons trouvées qu'il y a 46,44907 % de lignes avec au moins 1 valeur NA, donc c'est une chiffre important. Par conséquent, si nous éliminons directement les valeurs NA, cela peut diminuer la qualité de nos résultats.

	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4
1	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education th
2	1838	35-39	Man	Colombia	Doctoral degree
3	289287	30-34	Man	United States of America	Master's degree
4	860	35-39	Man	Argentina	Bachelor's degree
5	507	30-34	Man	United States of America	Master's degree
6	78	30-34	Man	Japan	Master's degree
7	401	30-34	Man	India	Bachelor's degree
8	748	22-24	Man	Brazil	Bachelor's degree
9	171196	25-29	Woman	China	Master's degree
10	762	35-39	Man	Germany	Doctoral degree
11	150	22-24	Man	China	No formal education past high school
12	7469	18-21	Man	India	Bachelor's degree
13	742	35-39	Man	United States of America	Doctoral degree
14	535	22-24	Man	Indonesia	Bachelor's degree
15	378	30-34	Man	Canada	Bachelor's degree
16	623	30-34	Man	Switzerland	Bachelor's degree

Showing 1 to 16 of 20,037 entries, 355 total columns

Figure 1 : Un aperçu de notre base de données originale

1.4 Prétraitement des données

Après avoir obtenu les colonnes mentionnées, changer les noms des colonnes, nous prévoyons de vérifier les données NA et leur importance pour justifier nos prochaines actions.

Ce que nous allons faire ensuite, c'est étudier chaque colonne, et vérifier la pertinence du NA et donc l'élimination des lignes correspondantes. Comme expliqué dans les commentaires du code, dans chaque cas sa justification est détaillée.

	Age	Sex	Pays	Diplome	Domaine	Experience_Prof	Taille_entreprise	Salaire
1	35-39	Man	Colombia	Doctoral degree	Student	5-10	0	0
2	30-34	Man	United States of America	Master's degree	Data Engineer	5-10	10000	100000-124999
3	35-39	Man	Argentina	Bachelor's degree	Software Engineer	10-20	1000-9999	15000-19999
4	30-34	Man	United States of America	Master's degree	Data Scientist	5-10	250-999	125000-149999
5	30-34	Man	Japan	Master's degree	Software Engineer	3-5	0	0
6	30-34	Man	India	Bachelor's degree	Data Analyst	1	0	0
7	22-24	Man	Brazil	Bachelor's degree	Student	3-5	0	0
8	25-29	Woman	China	Master's degree	Student	1	0	0
9	35-39	Man	Germany	Doctoral degree	Data Scientist	5-10	1000-9999	70000-79999
10	22-24	Man	China	No formal education past high school	Student	1	0	0
11	18-21	Man	India	Bachelor's degree	Student	1-2	0	0
12	35-39	Man	United States of America	Doctoral degree	Research Scientist	1-2	0-49	30000-39999
13	22-24	Man	Indonesia	Bachelor's degree	Student	1	0	0
14	30-34	Man	Canada	Bachelor's degree	Data Engineer	1	0-49	90000-99999
15	30-34	Man	Switzerland	Bachelor's degree	Other	1	50-249	70000-79999
16	18-21	Woman	India	Bachelor's degree	Student	1-2	0	0
17	25-29	Woman	Other	Bachelor's degree	Currently not employed	3-5	0	0
18	22-24	Man	Singapore	Bachelor's degree	Student	3-5	0	0
19	22-24	Man	India	Bachelor's degree	Student	1-2	0	0

Showing 1 to 19 of 19,313 entries, 12 total columns

Figure 2 : Données avec titres modifiés et contenu filtré

De plus, comme nos données proviennent d'une enquête, nous avons des intervalles dans certaines variables, ce qui nuit à notre étude. La solution que nous avons prise a été de créer d'autres tableaux avec la moyenne de ceux-ci. Ces détails sont expliqués dans les commentaires du code, où un filtrage et une création de colonnes ont été effectués pour chaque intervalle de variables.

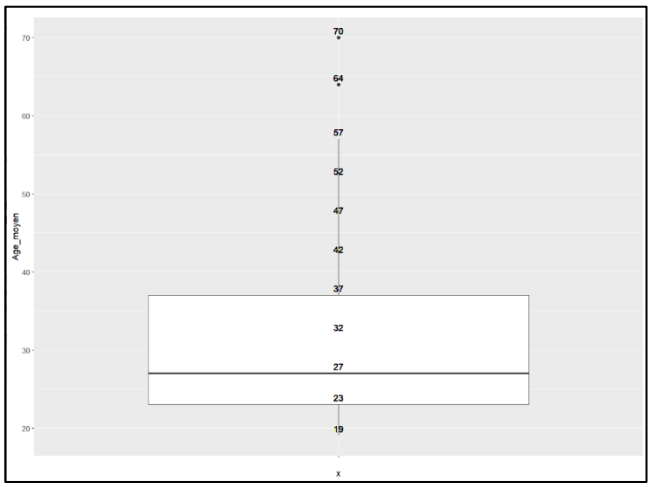
Age_moyen	Periode_experience	Nombre_employeurs_ENTREPRISE	Salaire_moyen
37	7	0	0
32	7	10000	112499
37	15	5499	17499
32	7	624	137499
32	4	0	0
32	1	0	0
23	4	0	0
27	1	0	0
37	7	5499	74999
23	1	0	0
19	1	0	0
37	1	24	34999
23	1	0	0
32	1	24	94999
32	1	149	74999
19	1	0	0
27	4	0	0
23	4	0	0
23	1	0	0

Figure 3 : Colonnes ajoutées avec des données filtrées

II. Réalisation du modèle de régression linéaire

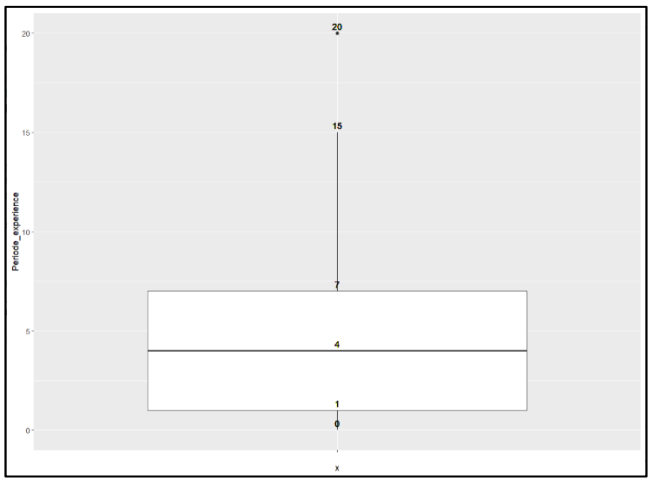
Etudes statistiques préalables

Age moyen & Boite a moustaches Age_moyen



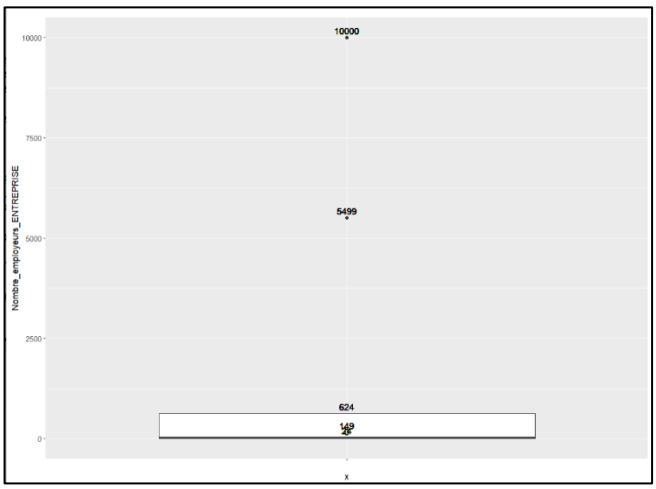
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.00	23.00	27.00	30.94	37.00	70.00

Periode_experience & Boite a moustache



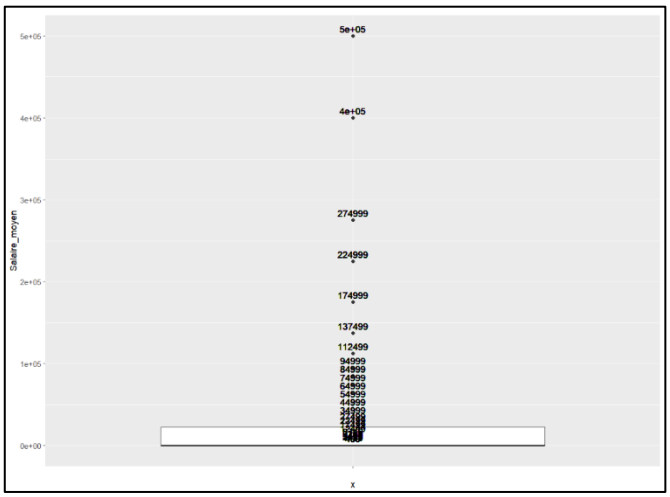
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.000	4.000	4.829	7.000	20.000

Nombre_employeurs_ENTREPRISE & Boite a moustache



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	24	1718	624	10000

Salaire_moyen & Boite a moustache




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0       0     499   23987   22499  500000
```

A. Régression simple

Filtrage des données par rapport aux femmes et hommes :

A.1. Décrire les relations statistiques

a) Les données statistiques de la base de données de chaque champs par rapport aux femmes .

```
> summary(DBWOMAN)
     Age               Sex                Pays              Diplome            Domaine
 Length:3805        Length:3805        Length:3805        Length:3805        Length:3805
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
 Experience_Prof    Taille_entreprise    Salaire            Age_moyen     Periode_experience
 Length:3805        Length:3805        Length:3805        Min.   :19.0   Min.   : 0.000
 Class :character   Class :character   Class :character   1st Qu.:23.0   1st Qu.: 1.000
 Mode  :character   Mode  :character   Mode  :character   Median :27.0   Median : 1.000
                                                          Mean   :28.7   Mean   : 3.465
                                                          3rd Qu.:32.0   3rd Qu.: 4.000
                                                          Max.   :70.0   Max.   :20.000
```

```
 Nombre_employeurs_ENTREPRISE Salaire_moyen
 Min.   :    0               Min.   :     0
 1st Qu.:    0               1st Qu.:     0
 Median :    0               Median :     0
 Mean   : 1301               Mean   : 14549
 3rd Qu.:  149               3rd Qu.:  4499
 Max.   :10000               Max.   :500000
```

b) Les données statistiques de la base de données de chaque champs par rapport aux hommes.

```
> summary(DBMAN)
     Age               Sex                Pays              Diplome            Domaine
 Length:15508       Length:15508       Length:15508       Length:15508       Length:15508
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
Experience_Prof    Taille_entreprise    Salaire           Age_moyen     Periode_experience
Length:15508       Length:15508       Length:15508       Min.   :19.00   Min.   : 0.000
Class :character   Class :character   Class :character   1st Qu.:23.00   1st Qu.: 1.000
Mode  :character   Mode  :character   Mode  :character   Median :27.00   Median : 4.000
                                                         Mean   :31.49   Mean   : 5.164
                                                         3rd Qu.:37.00   3rd Qu.: 7.000
                                                         Max.   :70.00   Max.   :20.000
```

```
Nombre_employeurs_ENTREPRISE Salaire_moyen
Min.   :    0                Min.   :     0
1st Qu.:    0                1st Qu.:     0
Median :   24                Median :   499
Mean   : 1821                Mean   : 26303
3rd Qu.:  624                3rd Qu.: 27499
Max.   :10000                Max.   :500000
```

D'après les résultats affichés on peut regarder que en fonction des moyennes, les hommes gagnent plus que les femmes dans le monde.

c) Le coefficient de corrélation de Pearson ne peut être calculé qu'entre deux variables quantitatives et le sexe est une variable qualitative. Au lieu de cela, nous pouvons utiliser une technique statistique appelée analyse croisée (test de student) pour étudier la relation entre le salaire et le sexe. Cette technique vous permet de voir s'il existe une différence statistiquement significative dans le salaire moyen entre les hommes et les femmes, et nous fournit une mesure de l'association entre ces deux variables.

```
> t.test(x=DB$Salaire_moyen[DB$Sex == 'Man'],y=DB$Salaire_moyen[DB$Sex == 'Woman'])

        Welch Two Sample t-test

data:  DB$Salaire_moyen[DB$Sex == "Man"] and DB$Salaire_moyen[DB$Sex == "Woman"]
t = 14.894, df = 7655.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10206.73 13300.55
sample estimates:
mean of x mean of y
 26303.08  14549.43
```

*Les résultats du test incluent **la valeur de t calculée (8.6824)**, **le degré de liberté (2672.9)**, la **valeur p (inférieure à 2.2e-16), l'hypothèse alternative (la différence des moyennes n'est pas égale à zéro)** et **l'intervalle de confiance à 95%** (10261.56 à 16248.69).*

*La valeur p est inférieure au niveau de signification établi (généralement 0,05), ce qui **signifie qu'il y a des preuves statistiquement significatives pour rejeter l'hypothèse nulle** selon laquelle les deux moyennes sont égales. Cela suggère qu'il y a une différence significative des salaires moyens entre les hommes et les femmes dans les données que vous analysez.*

*L'intervalle de confiance à 95% indique qu'avec **95% de confiance**, la différence réelle entre les moyennes des salaires des hommes et des femmes se **situe entre 10261.56 et 16248.69**.*

***En résumé**, les résultats suggèrent qu'**il y a une différence significative des salaires moyens entre les hommes et les femmes** dans les données. La différence spécifique se situe dans l'intervalle de confiance fourni.*

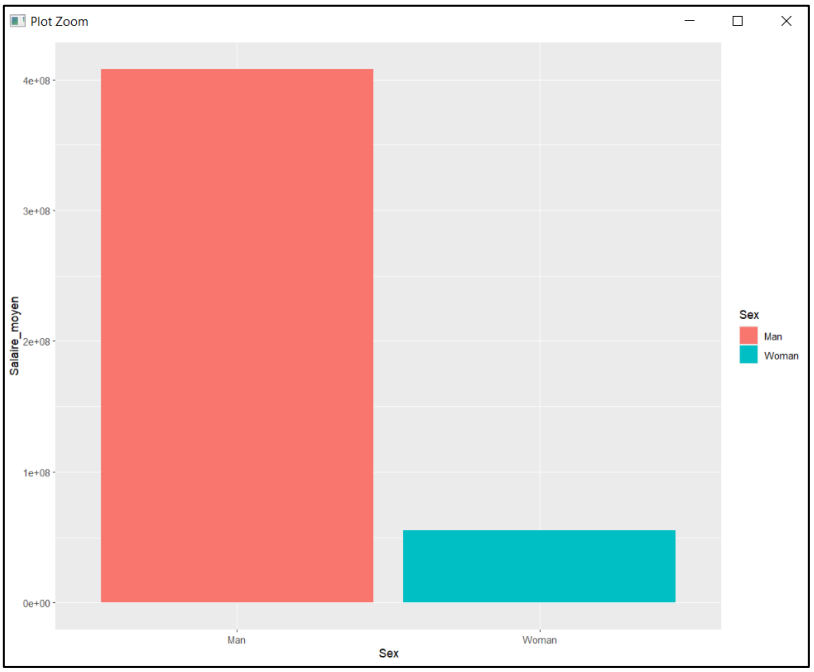


Figure 4 : Représentation graphique d'un tableau de consistance.

Nous avons trouvé un coefficient de détermination de 0,007906188, qui indique que la relation entre les deux variables du tableau de contingence est très faible. Cela signifie que la variable indépendante ne peut pas prédire avec précision la variable dépendante.

d) Etude des salaires des femmes en fonction de la période d'expérience :

L'équation de la droite de la régression linéaire nous permet d'estimer le salaire moyen d'une femme qui exerçant de l'activité dans le domaine informatique dans le monde d'entier est :

y = 3703.345 * x + 1718.635

```
> # Determiner la covariance entre les variables salaires et experiencess: 52963.68
> cov(DBWOMAN$Salaire_moyen,DBWOMAN$Periode_experience)
[1] 55823.03
```

```
> # Determiner le coefficient de correlation: 0.3100646
> cor(DBWOMAN$Salaire_moyen,DBWOMAN$Periode_experience)
[1] 0.3100646
```

```
> # Calculer le coefficient de determination: 0.09614005
> summary(model1)$r.squared
[1] 0.09614005
```

```
> # Afficher les mesures statistiques pour le modele lineaire obtenu:
> summary(model1)

Call:
lm(formula = DBWOMAN$Salaire_moyen ~ DBWOMAN$Periode_experience)

Residuals:
   Min     1Q Median     3Q    Max
-60954 -15553  -7633  -4826 492367

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4826.2      788.0   6.125    1e-09 ***
DBWOMAN$Periode_experience  2806.4      139.5  20.112   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38380 on 3803 degrees of freedom
Multiple R-squared:  0.09614,   Adjusted R-squared:  0.0959
F-statistic: 404.5 on 1 and 3803 DF,  p-value: < 2.2e-16
```

A.2. Vérifier les hypothèses de validité du modèle de régression linéaire

Le modelé linéaire entre Salaire et Age est de qualité médiocre, il faut ajouter d'autres variables explicatives.

Pour améliorer sa performance, la valeur p- value de la variable expérience professionnelle : Le p-value est inferieur a 2 "p-value: < 2.2e-16" donc on peut rejeter l'hypothèse nulle, néanmoins les hypothèses de validités sont encore très peu fiables.

Pour le deuxième modèle entre « le salaire et l’expérience », on peut en déduire qu'il existe une relation positive entre le temps d'expérience et le salaire moyen des femmes, et que le modèle est statistiquement significatif pour expliquer cette relation.

1. Le modèle s'ajuste aux données, car la valeur de R-squared (coefficient de détermination) est de 0,09614, ce qui indique un ajustement modéré. La valeur ajustée de R-squared est de 0,0959, ce qui indique que le modèle n'est pas sur ajusté.
2. L'erreur standard résiduelle est de 38380, ce qui indique que la différence entre les valeurs observées et les valeurs prédites du modèle est modérée.
3. La valeur de la statistique F est de 404,5, avec une p-valeur de < 2.2e-16, ce qui indique que le modèle est statistiquement significatif.

A.3. Valider le modèle de régression

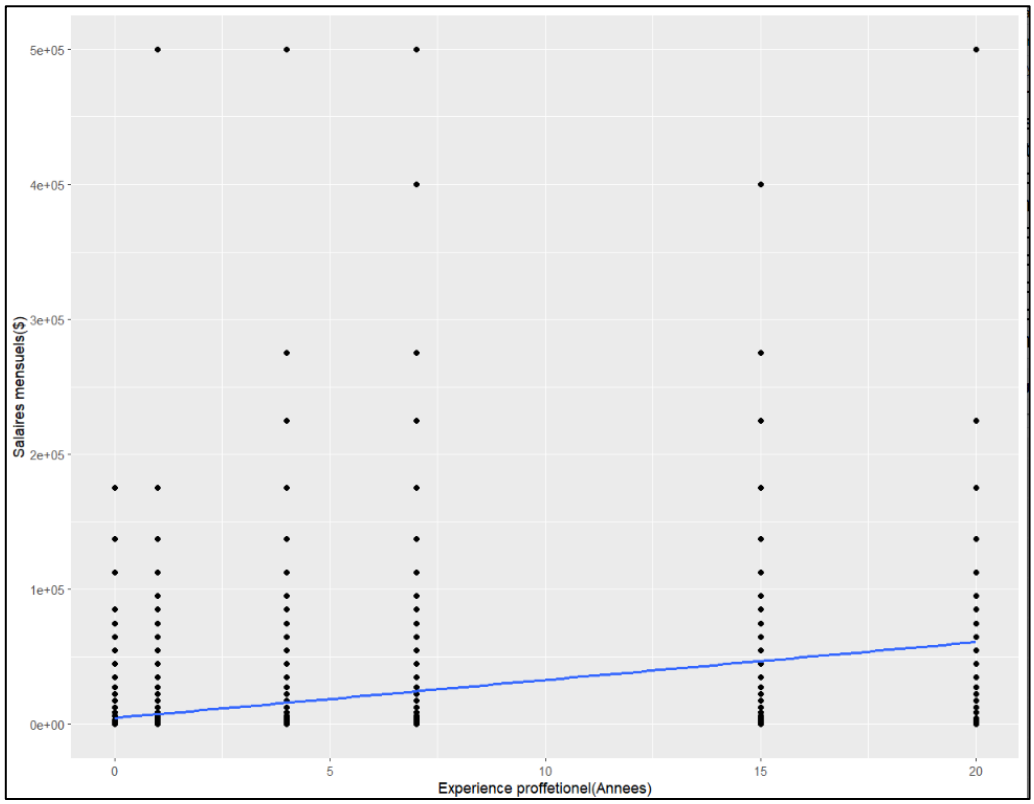


Figure 5 : Etude des salaires des femmes en fonction de la période d'expérience

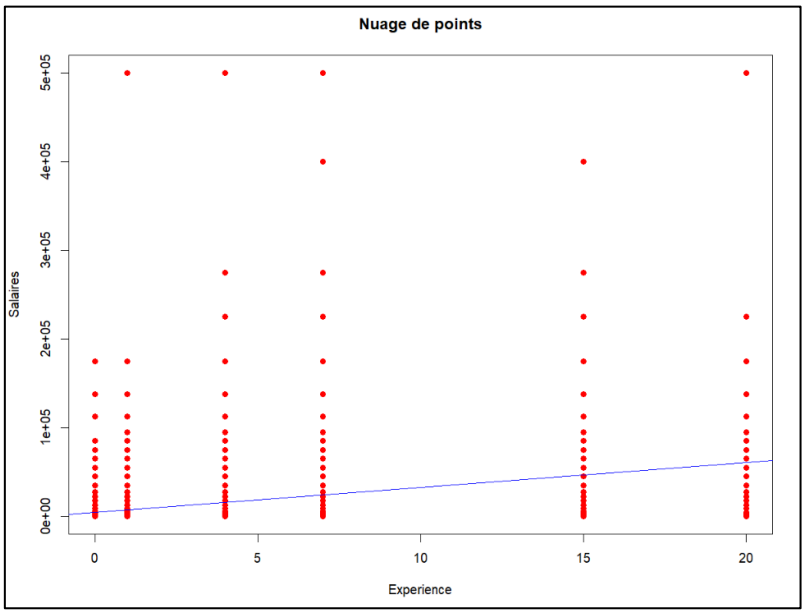


Figure 6 : Nuage des points Etude des salaires des femmes en fonction de la période d'expérience

A.4. Evaluer les points qui exercent une grande influence sur la régression afin de les écarter s'il s'agit de points potentiellement aberrants

Nous allons utiliser la fonction cooks.distance() : cette fonction calcule les distances de cuisson pour chaque point de données. Les points avec des valeurs plus élevées indiquent une plus

grande influence sur la régression. Les points avec des valeurs extrêmement élevées peuvent être supprimés.

```
> cor(DBWOMAN_clean$Salaire_moyen,DBWOMAN_clean$Periode_experience)
[1] 0.3194753
>
```

Figure 7 : Légère amélioration

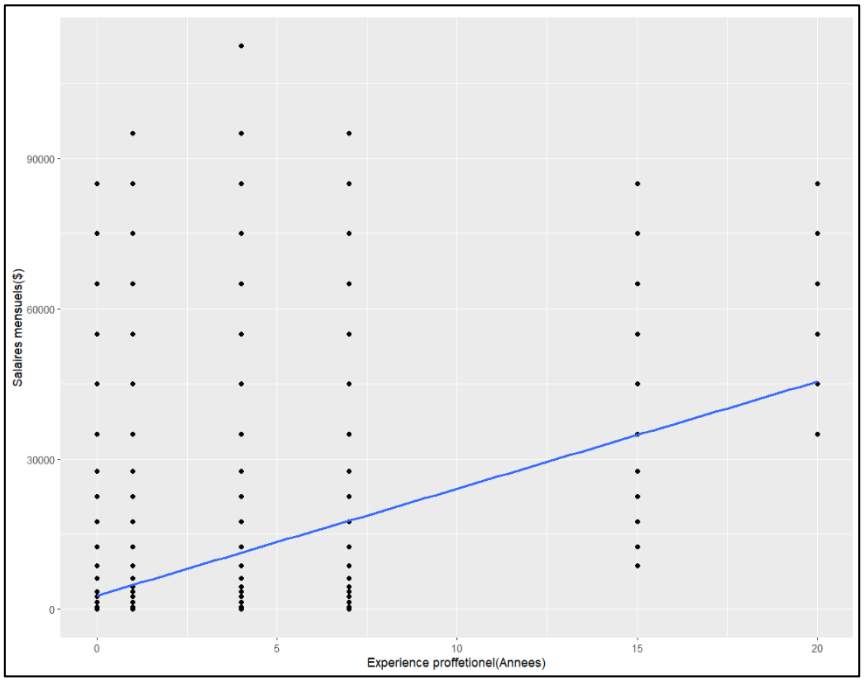


Figure 8 : Graphique après suppression de certaines valeurs aberrantes

B. Régression multiple

Pour cette étude nous allons faire un étude dans les variables :

Salaire_moyen vs Sex, Age_moyen,Pays, Diplome, Domaine, Periode_experience et Nombre_employeurs_ENTREPRISE de façon générale.

B.1. Décrire les relations statistiques

On commence par l'analyse de la variance puis les résultats avec summary du model « aov » :

```
> summary(modelo_aov)
                            Df    Sum Sq   Mean Sq F value Pr(>F)
Sex                          1 4.221e+11 4.221e+11  242.59 <2e-16 ***
Age_moyen                    1 6.387e+12 6.387e+12 3670.79 <2e-16 ***
Periode_experience           1 2.995e+12 2.995e+12 1721.24 <2e-16 ***
Nombre_employeurs_ENTREPRISE 1 2.703e+12 2.703e+12 1553.42 <2e-16 ***
Diplome                      6 1.915e+11 3.192e+10   18.35 <2e-16 ***
Domaine                     13 1.431e+12 1.101e+11   63.25 <2e-16 ***
Pays                        54 5.791e+12 1.072e+11   61.63 <2e-16 ***
Residuals                19235 3.347e+13 1.740e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Les résultats de l'ANOVA indiquent qu'il existe des différences significatives de salaire entre les groupes dans toutes les variables indépendantes à l'exception du Diplôme.
En particulier, la valeur "Pr(>F)" dans le tableau indique la probabilité d'obtenir une valeur F aussi élevée ou supérieure due au hasard si l'hypothèse nulle d'égalité des moyennes est vraie.
Une très petite valeur Pr(>F) (inférieure à 0,05) indique qu'il est très peu probable que la différence de moyenne soit due au hasard, et nous pouvons donc rejeter l'hypothèse nulle et conclure qu'il existe de réelles différences entre les groupes.

a) Ensuite on peut faire des tableaux de contingence :

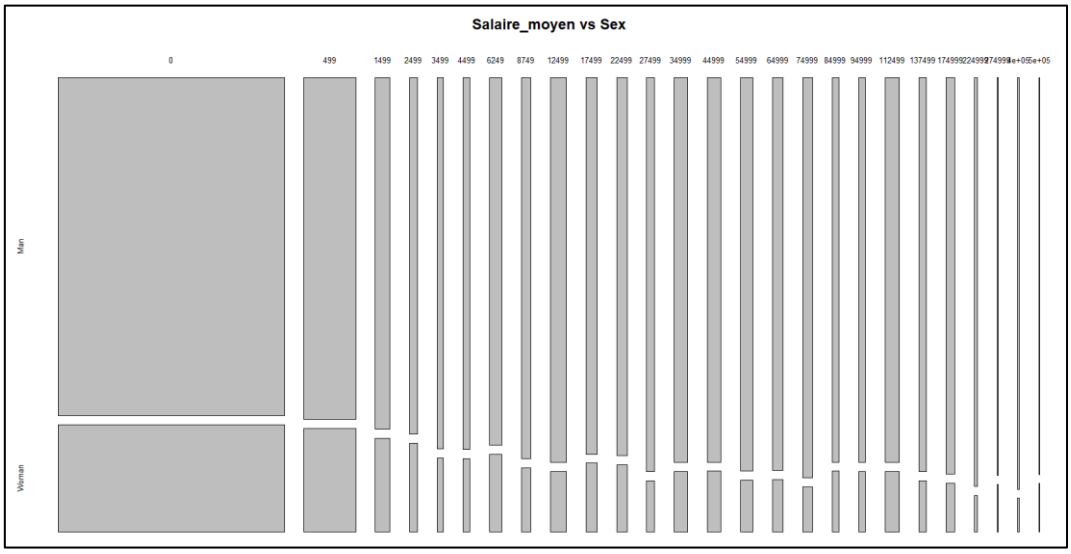


Figure 9 : « Salaire moyen » vs, « Sex »

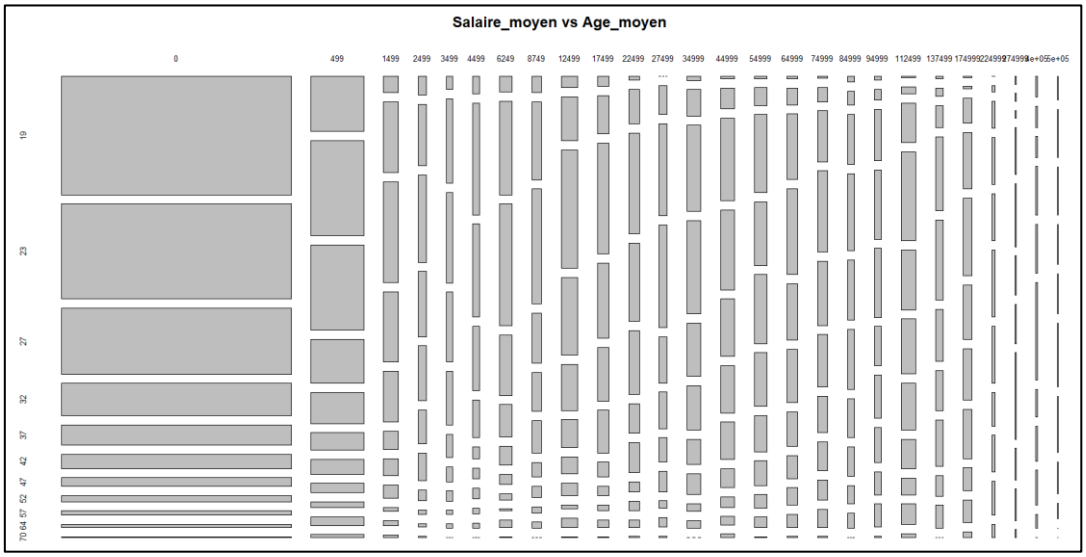


Figure 10 : « Salaire moyen » vs, « Age moyen »

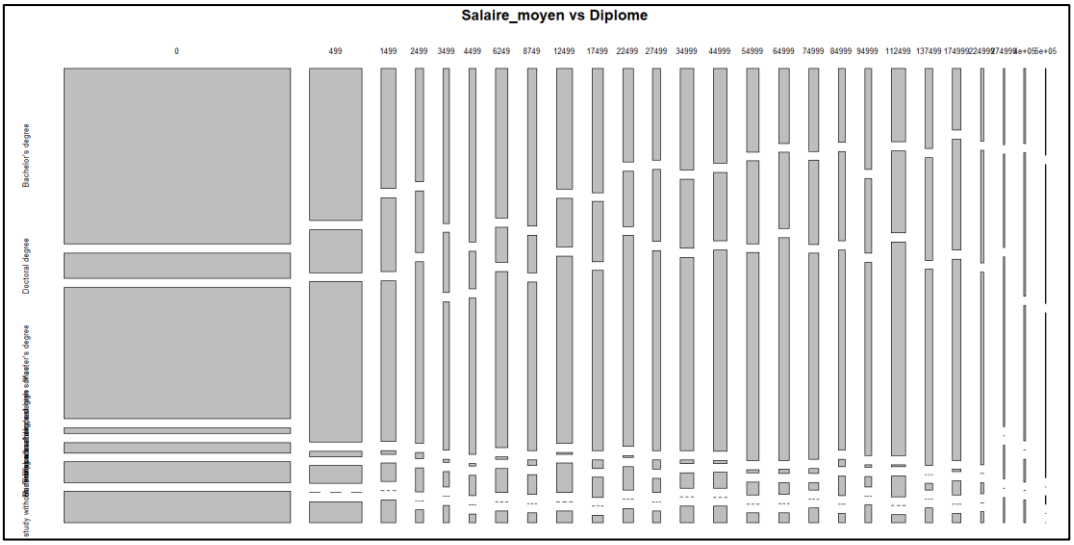


Figure 11 : « Salaire moyen » vs, « Diplôme»

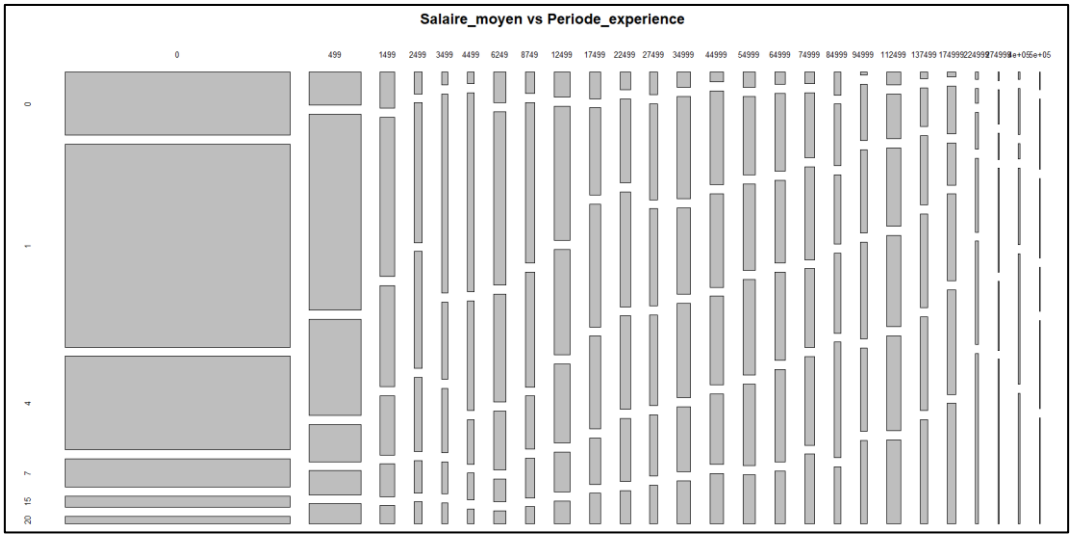


Figure 12 : « Salaire moyen » vs, « Période expérience»

b) Etudier les corrélations linéaires entre les variables explicatives avec ggpairs( )

La colonne 'Pays' a plus de niveaux (55) que le seuil (15) autorisé, donc nous allons supprimer la colonne.

```
> ggpairs(DB, columns = c("Salaire_moyen","Sex","Age_moyen", "Diplome", "Periode_experience", "Nombre_
employeurs_ENTREPRISE"))
```

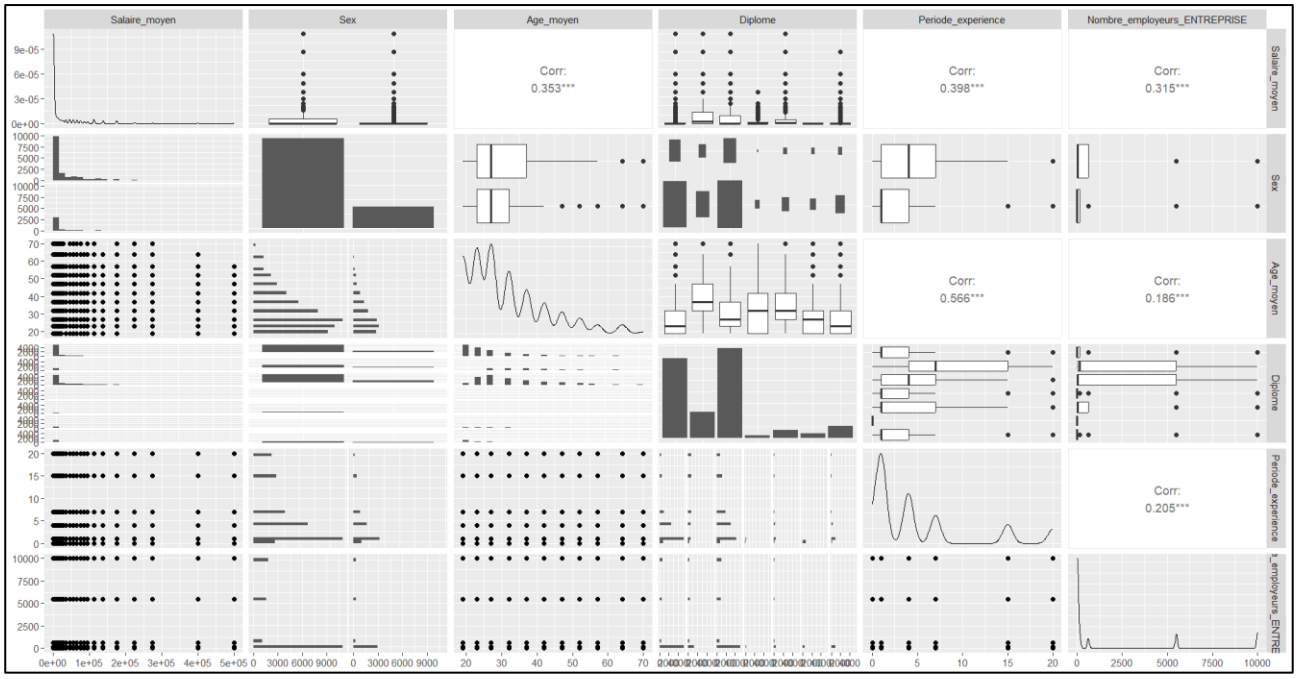


Figure 13 : Etude des corrélations linéaires avec ggpairs( )

B.2. Vérifier les hypothèses de validité du modèle de régression linéaire

Dans notre cas, si notre objectif est de trouver une relation entre l'écart salarial entre les hommes et les femmes, les variables indépendantes les plus pertinentes seraient la variable "Sexe" et la variable "Age_moyen", puisque ce sont celles qui ont la valeur F la plus élevée .et ont un coefficient plus élevé dans le modèle. Cependant, il est important de considérer également toute relation potentielle entre les variables indépendantes et les variables qualitatives

B.3. Valider le modèle de régression

a) La regression multiples pour femmes

Regression multiple des salaires des femmes dans le monde dans le domaine informatique.

(Salaire_moyen vs Periode_experience + Age_moyen + Nombre_employeurs_ENTREPRISE)

a) Etude pour les femmes

```
> summary(model2)

Call:
lm(formula = Salaire_moyen ~ Periode_experience + Age_moyen +
    Nombre_employeurs_ENTREPRISE, data = DBWOMAN)

Residuals:
   Min     1Q Median     3Q    Max
-84850 -11178  -2904   2117 489108

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.944e+04  1.940e+03  -10.02   <2e-16 ***
Periode_experience          1.599e+03  1.480e+02   10.80   <2e-16 ***
Age_moyen                   8.275e+02  7.056e+01   11.73   <2e-16 ***
Nombre_employeurs_ENTREPRISE  3.608e+00  1.999e-01   18.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36110 on 3801 degrees of freedom
Multiple R-squared:  0.2006,    Adjusted R-squared:  0.1999
F-statistic: 317.9 on 3 and 3801 DF,  p-value: < 2.2e-16
```

Figure 14 : Coeff de correlation 0.2006

Le modèle global est plus pertinent que le modèle simple vu l'augmentation du p-value

b) Etude pour les hommes

```
> summary(model2)

Call:
lm(formula = Salaire_moyen ~ Periode_experience + Age_moyen +
    Nombre_employeurs_ENTREPRISE, data = DBMAN)

Residuals:
    Min      1Q  Median      3Q     Max
-120131  -17239   -6091     422  496872

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.695e+04  1.196e+03  -14.17   <2e-16 ***
Periode_experience          2.445e+03  7.986e+01   30.61   <2e-16 ***
Age_moyen                   7.627e+02  4.198e+01   18.17   <2e-16 ***
Nombre_employeurs_ENTREPRISE  3.629e+00  1.156e-01   31.40   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48110 on 15504 degrees of freedom
Multiple R-squared:  0.2327,    Adjusted R-squared:  0.2326
F-statistic:  1567 on 3 and 15504 DF,  p-value: < 2.2e-16
```

Le modèle global est plus pertinent que le modèle simple vu l'augmentation du p-value

B.4. Evaluer les points qui ont une grande influence sur la régression afin de les écarter s’il s’agit de points potentiellement aberrants

Nous pouvons utiliser la matrice de corrélation pour évaluer les valeurs aberrantes dans une régression multiple.

La matrice de corrélation montre les relations linéaires entre les variables de l'ensemble de données. Les valeurs varient entre -1 et 1, où une valeur proche de 1 indique une forte corrélation positive entre deux variables, une valeur proche de -1 indique une forte corrélation négative entre deux

variables, et une valeur proche de 0 indique une faible corrélation ou pas de corrélation entre les deux variables.

a) Etude pour les femmes

```
> mat_cor
                           Age_moyen Periode_experience Nombre_employeurs_ENTREPRISE
Age_moyen                  1.0000000          0.4471780                    0.1441786
Periode_experience         0.4471780          1.0000000                    0.1802514
Nombre_employeurs_ENTREPRISE 0.1441786          0.1802514                    1.0000000
Salaire_moyen              0.3081417          0.3100646                    0.3261409
                           Salaire_moyen
Age_moyen                      0.3081417
Periode_experience             0.3100646
Nombre_employeurs_ENTREPRISE   0.3261409
Salaire_moyen                  1.0000000
```

Figure 15 : Matrice de corrélation

```
> #représentation graphique de la matrice de corrélation qui utilise des symboles
> # pour indiquer le degré de relation entre les variables.
> symnum(mat_cor, abbr.colnames=FALSE)
                           Age_moyen Periode_experience Nombre_employeurs_ENTREPRISE
Age_moyen                  1
Periode_experience         .         1
Nombre_employeurs_ENTREPRISE                           1
Salaire_moyen              .         .                 .
                           Salaire_moyen
Age_moyen
Periode_experience
Nombre_employeurs_ENTREPRISE
Salaire_moyen              1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

Figure 16 : Représentation graphique Matrice de corrélation

Dans ce cas, on observe qu'il existe une relation positive modérée entre la variable «Période_expérience » et « Salaire_moyen » (0,31) et une relation positive modérée entre « Nom_employeurs_ENTREPRISE » et « Salaire_moyen » (0,3261409).

Ensuite une corrélation avec corrplot :

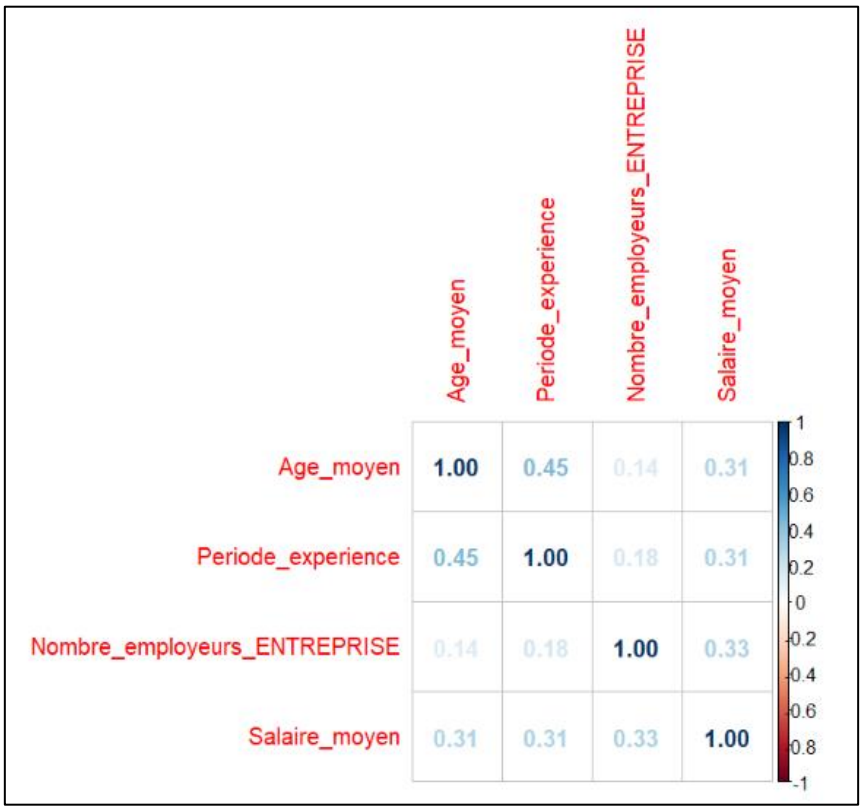


Figure 17 : Graphique la corrélation avec corrplot( )

Dans ce cas, on constate que la corrélation entre l'âge moyen et l'expérience **est modérée (0,45)** et **la corrélation entre l'expérience et le salaire moyen est modérée (0,31**). La corrélation entre **l'âge moyen et le nombre d'employeurs est faible (0,14)** et la corrélation entre **le nombre d'employeurs et le salaire moyen est modérée (0,33).**

On trouve le VIF :

```
> lm.beta(model2)

Call:
lm(formula = Salaire_moyen ~ Periode_experience + Age_moyen +
    Nombre_employeurs_ENTREPRISE, data = DBWOMAN)

Standardized Coefficients::
                 (Intercept)           Periode_experience                   Age_moyen
                          NA                    0.1767191                   0.1906501
Nombre_employeurs_ENTREPRISE
                   0.2667994
```

Figure 18 : VIF du modele

```
> # Utiliser  des intervalles de confiance a laide de la fonction confint  pour voir  pour   chaque
 î²
> confint(model2)
                                    2.5 %        97.5 %
(Intercept)                -23242.933961 -15634.161335
Periode_experience          1309.272851   1889.708096
Age_moyen                    689.147027    965.820298
Nombre_employeurs_ENTREPRISE   3.215821      3.999649
```

Figure 19 : Intervalles de confiance avec confint( )

b) Etude pour les hommes

```
> mat_cor
                            Age_moyen Periode_experience Nombre_employeurs_ENTREPRISE
Age_moyen                   1.0000000          0.5780319                    0.1875565
Periode_experience          0.5780319          1.0000000                    0.2028602
Nombre_employeurs_ENTREPRISE 0.1875565          0.2028602                    1.0000000
Salaire_moyen               0.3533267          0.4025732                    0.3098728
                            Salaire_moyen
Age_moyen                       0.3533267
Periode_experience              0.4025732
Nombre_employeurs_ENTREPRISE    0.3098728
Salaire_moyen                   1.0000000
```

Figure 20 : Matrice de corrélation

```
> symnum(mat_cor, abbr.colnames=FALSE)
                            Age_moyen Periode_experience Nombre_employeurs_ENTREPRISE
Age_moyen                   1
Periode_experience          .         1
Nombre_employeurs_ENTREPRISE                           1
Salaire_moyen               .         .                .
                            Salaire_moyen
Age_moyen
Periode_experience
Nombre_employeurs_ENTREPRISE
Salaire_moyen               1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

Figure 21 : Représentation graphique Matrice de corrélation

Ensuite une corrélation avec corrplot :

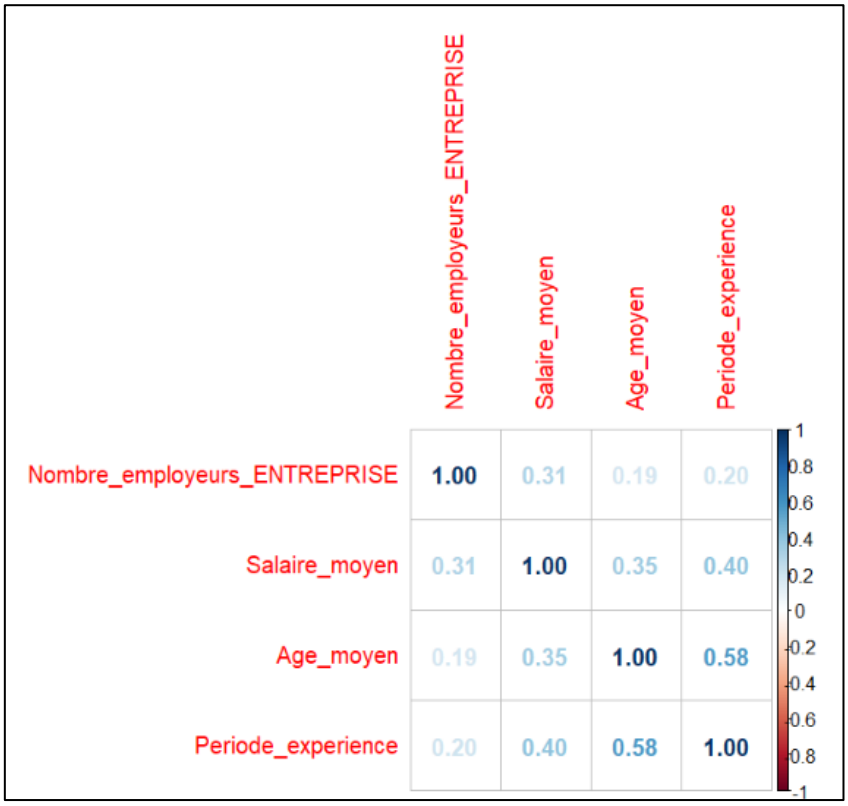


Figure 22 : Graphique la corrélation avec corrplot( )

Le symbole "." représente une faible corrélation (inférieure à 0,3), "," représente une corrélation modérée (entre 0,3 et 0,6), "+" représente une forte corrélation (entre 0,6 et 0,8) et "*" représente

une très forte corrélation (supérieure à 0,8). Dans ce cas, on peut voir que toutes les variables ont une faible corrélation entre elles, ce qui indique qu'il n'y a pas de relation forte entre elles.

On trouve le VIF :

```
> lm.beta(model2)

Call:
lm(formula = Salaire_moyen ~ Periode_experience + Age_moyen +
    Nombre_employeurs_ENTREPRISE, data = DBMAN)

Standardized Coefficients::
                 (Intercept)           Periode_experience                   Age_moyen
                          NA                    0.2657383                   0.1572450
Nombre_employeurs_ENTREPRISE
                   0.2264728
```

Figure 23 : VIF du modèle

```
> # Utiliser des intervalles de confiance a  laide de la fonction confint pour voir pour chaque Î²
> confint(model2)
                                  2.5 %        97.5 %
(Intercept)                -19290.696072 -14603.386772
Periode_experience           2288.034927   2601.115339
Age_moyen                     680.451932    845.024064
Nombre_employeurs_ENTREPRISE    3.402841      3.855946
```

Figure 24 : Intervalles de confiance avec confint( )

III. Estimation de la pertinence du modèle

3.1. Utilisez un indice pour étudier sa pertinence

Evaluation des multi colinéaire par les VIF

```
> vif(model2)
          Periode_experience                   Age_moyen Nombre_employeurs_ENTREPRISE
                    1.271710                    1.256511                    1.039007
```

```
> check_collinearity(model)
# Check for Multicollinearity

Low Correlation

                      Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
                 Age_moyen 2.08 [2.03, 2.12]         1.44      0.48     [0.47, 0.49]
       Periode_experience 1.74 [1.71, 1.77]         1.32      0.58     [0.56, 0.59]
 Nombre_employeurs_ENTREPRISE 1.26 [1.24, 1.28]         1.12      0.80     [0.78, 0.81]
                       Sex 1.05 [1.04, 1.07]         1.02      0.95     [0.94, 0.97]
                      Pays 1.71 [1.68, 1.75]         1.31      0.58     [0.57, 0.60]
                  Domaine 5.97 [5.82, 6.13]         2.44      0.17     [0.16, 0.17]

Moderate Correlation

    Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
 Diplome 4.38 [4.27, 4.49]         2.09      0.23     [0.22, 0.23]
```

Figure 25 : Etude de multi colinéarité

Il est recommandé d'examiner d'autres aspects du modèle et des données pour déterminer si ces points sont réellement aberrants.

3.2. Valider pertinence avec les p-values et R² ajusté

Après avoir fait un résumé de notre étude multi colinéaire :

```
Residual standard error: 41710 on 19235 degrees of freedom
Multiple R-squared:  0.3731,    Adjusted R-squared:  0.3706
F-statistic: 148.7 on 77 and 19235 DF,  p-value: < 2.2e-16
```

Figure 26 : P-value <2.2^{e} -16 , Multiple R-squared = 0.3731

Ce résultat indique que le modèle de régression multiple a un bon ajustement puisque le R au carré est élevé, ce qui signifie que le modèle explique une grande quantité de variabilité dans les données. La statistique F est élevée et sa valeur p est extrêmement petite, ce qui signifie qu'il est fort probable que les coefficients du modèle ne soient pas nuls et que le modèle soit significatif. L'erreur résiduelle standard est de 41 710, ce qui indique qu'en moyenne, le modèle fait une erreur de 41 710 $ dans la prévision des salaires. Cependant, il convient de noter que le R au carré ne mesure que la proportion de variabilité dans les données qui est expliquée par le modèle, et n'indique pas nécessairement que le modèle est utile pour faire des prédictions précises.

IV. Prédictions

4.1. Utilisez la commande « predict » à partir d'un modèle de régression

On va prédire les salaires des femmes et des hommes qui exercent une activité dans le domaine informatique dans le monde entier avec de différentes tranche d'Age.

```
#Dans ce graphique, vous pouvez comparer les prédictions du modèle (axe des x)
# avec les valeurs réelles (axe des y) des données.
library(ggplot2)
ggplot(data = DBMAN, aes(x = predictions, y = Salaire_moyen)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Predicciones vs Salaire_moyen")
```

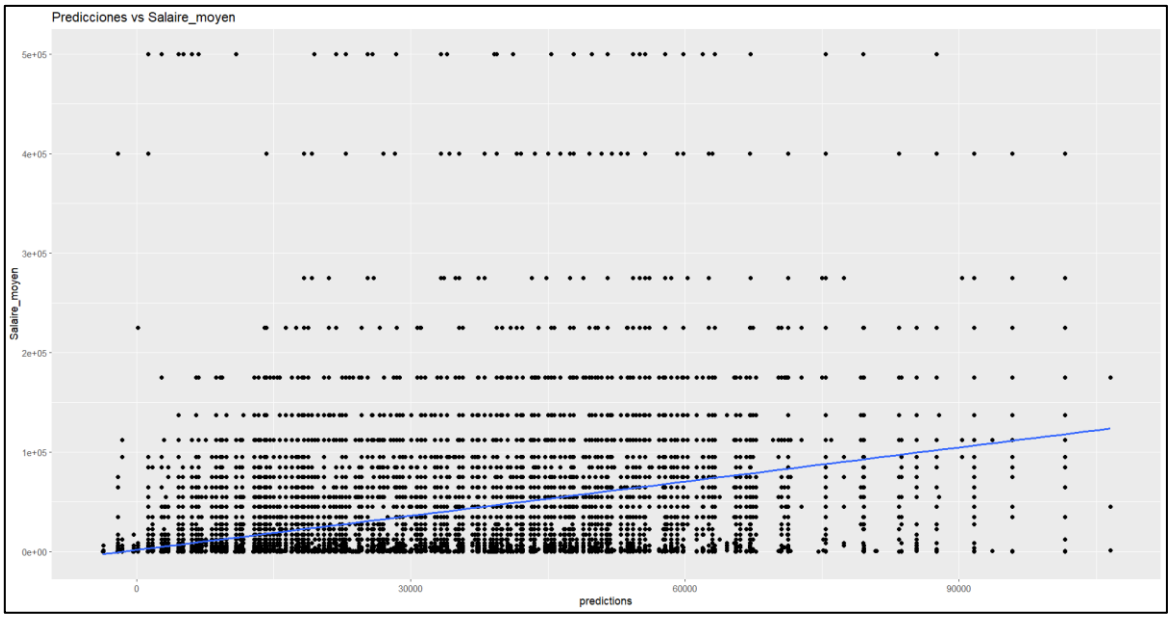


Figure 27 : Représentation graphique des résultats des prédictions

4.2. Valider pertinence avec les p-values et R² ajusté

```
> summary(model)

Call:
lm(formula = Salaire_moyen ~ Age_moyen + Periode_experience +
    Nombre_employeurs_ENTREPRISE, data = DBMAN)

Residuals:
    Min      1Q  Median      3Q     Max
-120131  -17239   -6091     422  496872

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.695e+04  1.196e+03  -14.17   <2e-16 ***
Age_moyen                   7.627e+02  4.198e+01   18.17   <2e-16 ***
Periode_experience          2.445e+03  7.986e+01   30.61   <2e-16 ***
Nombre_employeurs_ENTREPRISE  3.629e+00  1.156e-01   31.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48110 on 15504 degrees of freedom
Multiple R-squared:  0.2327,    Adjusted R-squared:  0.2326
F-statistic:  1567 on 3 and 15504 DF,  p-value: < 2.2e-16
```

Figure 28 : les p-values < 2.2^{e}-16 et R² ajusté = 0.2326

On constate que les variables indépendantes « Age_moyen », « Periode_experience » et « Nombre_employeurs_ENTREPRISE » sont significatives dans le modèle, puisque leurs p-values sont très petites. Le R au carré du modèle est de 0,2327, ce qui indique que 23,27 % de la variabilité du salaire moyen est expliquée par les variables indépendantes du modèle. L'erreur type résiduelle est de 48110 et la valeur F est de 1567 avec une très petite valeur de p, ce qui indique que le modèle est statistiquement significatif.

V. Conclusions et perspectives

5.1. Conclusions sur l'étude

On a créé un modèle de régression linéaire pour prédire les salaires des femmes qui travaillent dans le domaine de l'informatique dans le monde en fonction de l'expérience professionnelle, l'âge moyen et la taille de l'entreprise . Nous utilisons la fonction predict( ) pour effectuer une prédiction sur une base de donnes qui touche au minimum 499$ et maximum 500000$ (je précise que elles ne sont pas nombreuses il y a 5 femmes dans le monde d'après les informations qu'on a eu sur la jeu de données).
La prédiction obtenue est 700$ et 122150.9$ pour les profils supérieurs.

Cela signifie que notre modèle prévoit une augmentation de salaires de les femmes en se basant # sur les données de la base "DBWOMAN" que nous avons utilisés pour créer le modèle.
Cependant, il est important de noter que cette prédiction n'est qu'une estimation et peut ne pas refléter les salaires réels. Il est important de prendre en compte d'autres facteurs tels que la situation du pays (les pays développées, les pays en guerre, situations culturels, politiques, chômage, etc.), la réputation et l'activité de l'entreprise, le nombre des femmes dans le domaine informatique au sein de l'entreprise ... pour obtenir une estimation plus précise.

Nous pouvons également ajouter que travailler avec un vrai sondage a été un défi, car nous avons dû faire un filtrage rigoureux de nos données, en raison de la grande quantité de données et des réponses multiples. Nous considérons que travailler avec des données réelles nous permet de enrichir nos connaissances et notre expérience et de nous pousser à en savoir plus à ce sujet.

D'autre part, en ce qui concerne le filtrage que nous avons effectué sur nos données, nous avons conclu qu'il existe une grande différence statistique entre travailler directement avec les données et travailler avec les données filtrées de manière générale, et même une plus grande différence lorsque nous avons filtré colonne par colonne, comme nous l'avons fait dans notre travail. Par conséquent, nous recommandons pour un futur travail de toujours tenir compte de la pertinence de nos données manquantes et du type de réponses que nous avons par colonne.

5.1. Travail futur et perspectives

Après avoir effectué ce travail, nous avons comme suggestion : premièrement, créer un index avec les objectifs à atteindre pour avoir un ordre et également tracer des objectifs sur ce que nous cherchons à faire dans notre étude. Deuxièmement, commentez autant que possible en fonction de la pertinence de l'explication du code. Cela permet de revoir les étapes, de corriger les erreurs et de suivre toute notre progression. Troisièmement, les graphiques peuvent être très dynamiques et explicatifs, mais il faut toujours tenir compte de l'importance de ce que l'on veut communiquer avec ces résultats.

Une façon d'améliorer ce travail serait d'ajouter des données avec plus d'années, ce qui nous permettrait d'améliorer une prédiction d'un point de vue temporel, car nos données sont basées sur 3 ans d'étude.

VI. ANNEXE :

Lien GITHUB du code :

https://github.com/Leonardo-VERA/TravailR/blob/774dac8c4724da231c9fad8680443774b0b8cfca/analisys_r_ghezali_vera_m1_ia