
**Hamid Abdellaoui** **1ér Data Science**

# Rapport de Projet de la fin du 1er Semestre en Statistiques descriptives

## INTRODUCTION:

Faire des data Analysis périodiquement est un élément très important au sein des entreprises, car il permet d'avoir une vue globale et générale sur l'établissement entière, et par conséquent les responsables peuvent prendre des décisions ou bien prédire le future de l'entreprise.
L'analyse des données s'effectue grâce à des tableaux , des graphes , des secteurs, des histogrammes.... en interprétant leurs résultats.
On peut faire ceci avec des logiciel ou avec des langages des programmation , j'ai choisi le langage R pour obtenir un les travaux demandé dans ce projet , et j'ai utilisé **R studio** comme outil pour exécuter mes codes et manipuler les Datas.
Je vais commencer d'abord par une analyse univarié en étudiant les statistique de chaque variable puis on passe à une analyse bivariée ou je vais essayer de rétablie les relations entre les différentes variables en analysant les tableaux des fréquences et ensuite en retrouve les mêmes résultats mais en utilisant des graphiques, et à la fin je termine par une conclusion générale. Et voilà la structure de ce rapport:

Introduction
1. Statistiques uni-variées des variables
    1-1 Nbr d'années d'expérience(ancienneté)
    1-2 Salaire net en MAD
    1-3 Niveau d'étude
2. Statistiques bi-variées
    I. Tableau de fréquence de niveau d'étude en fonction de sexe
    II. Tableau de fréquence représentant les intervalles salaires en fonction du sexe
    III. Tableau d'effectif intervalles de salaires*Niveau d'étude
    IV. Tableaux de fréquence des salaires selon le sexe des employés pour chaque niveau d'étude
3. Comparaisons H vs F & Relations entres variables
Conclusion

**NB :** Les codes utilisé pour obtenir les tableaux/graphes est dans un fichier séparé , ils sont bien commentés et bien ordonnés dans l'ordre des tableaux et graphes de ce rapport

# 1. Statistiques uni-variées des variables

## 1-1 Variable: Nombre d'années d'expérience

- **Moyen, écart-type,médiane, étendu**

```
> library(psych)
> describe(Salaire_entreprise$'Nb_année_d'expérience')
   vars   n mean   sd median                 min  max range
X1    1 193 5.43 2.89   5.82                0.18 9.89  9.71
```

- Dans un nombre d'observations (**n**) de 193 employées dans cette entreprise on a le moyen (**mean**) des années d'expérience est 5.43 , l'écart type (**sd**) est de 2.89 qui représente une dispersion raisonnable autour de la moyen.
- 50% des employés ont une ancienneté inférieur à la médiane qui est est 5.82 et l'autre 50% est supérieur.

- **Les quartiles**

```
> quantile(Salaire_entreprise$'Nb_année_d'expérience')
       0%       25%       50%       75%      100%
0.1808362 2.8361867 5.8190726 7.8624722 9.8945161
```

*# La 1er et la dernière valeur(respec) de ce tableau représente le min et le max (respectivement), et*
*# 25% , 50%, 75%(respectivement) correspondent aux 1erquartiles , médiane et 3ém*
*# quartile(respectivement)*

75% des employées ont une ancienneté inférieur à **7.86** et l'autre partie de 25% est supérieur(selon 3em quartile).

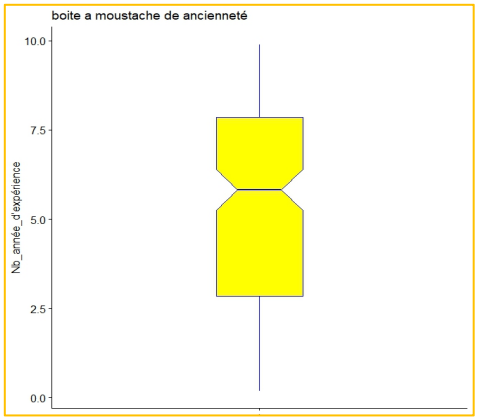
25% des employés ont une ancienneté inférieur à **2.83** et 75% restants supérieure(selon 1er quartile)

- **Intervalle interquartille**

```
> IQR(Salaire_entreprise$'Nb_année_d'expérience')
[1] 5.026285
```

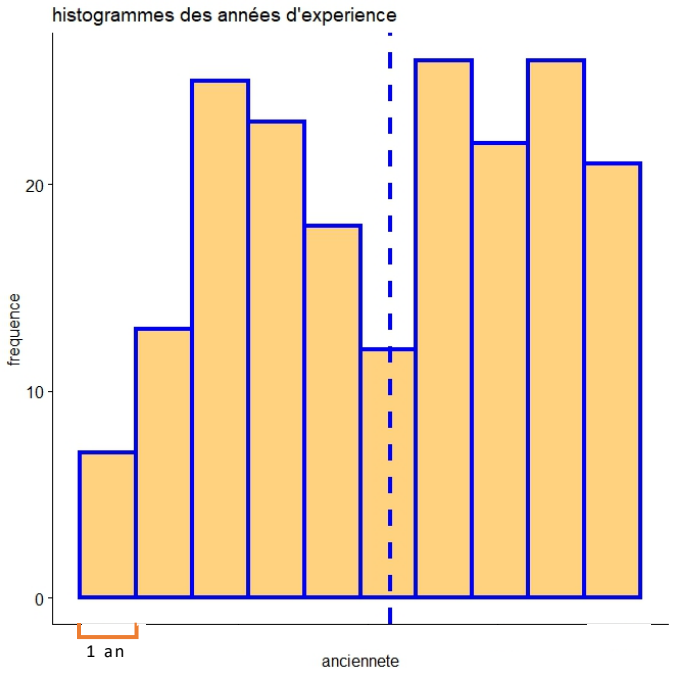
50 % des observations de l'ancienneté centrées autour de la médiane dans une intervalle de 5.02 ans

- **Boite à moustache**




La boite a moustache des années d'expérience montre plus clairement que 50% des employées ont une ancienneté centré / proche de la médiane cependant la personnes ayant une ancienneté inférieur à la médiane sont plus nombreux ( la partie bas du boite plus large que la partie haute), la boite donc à un peu près est asymétrique

- **Histogramme**




-La modalité plus fréquente est celui d'une ancienneté de 6 à 7 ans et de 8 à 9 ans, donc la série des observations est Bimodale en année d'expérience
-La ligne pointillée représente le moyen, on remarque qu'elle divise entre deux distribution la première est autour de 3 ans et la 2em est est entre 7 et 9ans

## 1-2 Variable:Salaire net en MAD

- **Moyen, écart-type,médiane, étendu**

```
> library(psych)
> describe(Salaire_entreprise$'Salaire_net_en_MAD')
   vars   n   mean      sd  median                 min      max   range
X1    1 193 5109.1 2024.79 4430.49                2931.15 11866.79 8935.64
```

→ Le moyen des salaires au sein de cette entreprise est 5109.1 MAD , les salaires sont fortement dispersés car l'écart type est égale à 2024.79, l' étendue est grande étant égale à 8930.64 MAD ce qui montre à nouveau la forte dispersion

- **Les quartiles/médiane**

```
> quantile(Salaire_entreprise$'Salaire_net_en_MAD')
       0%       25%       50%       75%      100%
 2931.153  3707.785  4430.486  5800.973 11866.789
```

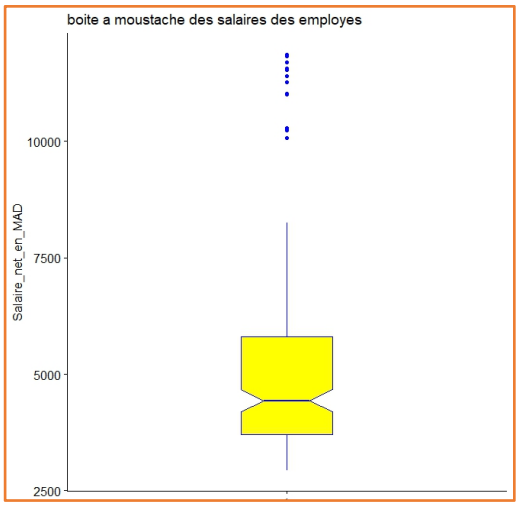
→ 50% des employés ont un salaire net inférieur à la médiane = **4430.48** et 50% restants ont des salaires supérieure.
→ 25% des employés ont un salaire net inférieur à la 1er quartile qui est **3707.7** MAD et l'autre75 % ont des salaires supérieur,
→ 75% des employés ont un salaire net inférieur à la 3em quartile qui est **5800.97 MAD** et l'autre25 % ont des salaires supérieur.

- **Intervalle interquartille**

```
> IQR(Salaire_entreprise$'Salaire_net_en_MAD')
[1] 2093.188
```

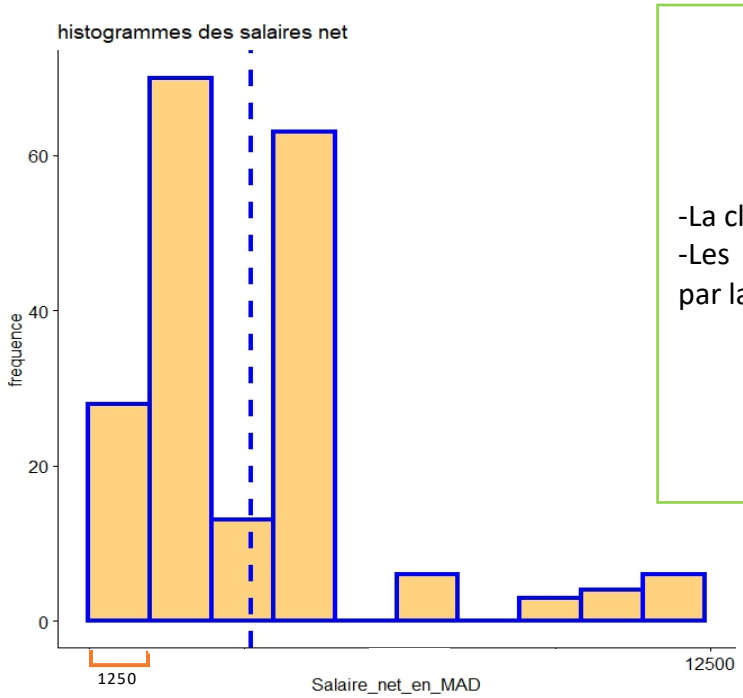
→ 50% des salariés sont centré dans un intervalle de longueur 2093.18

- **Boite à moustache**




-Boite a moustache des salaires net montre que 50% des salaires sont entre environ 3000 MAD et 6300 MAD
-Les points en bleu en haute représentent des points aberrantes des salaires de quelques employées ayant des salaires plus grande que leurs collègues, la boite est asymétrique

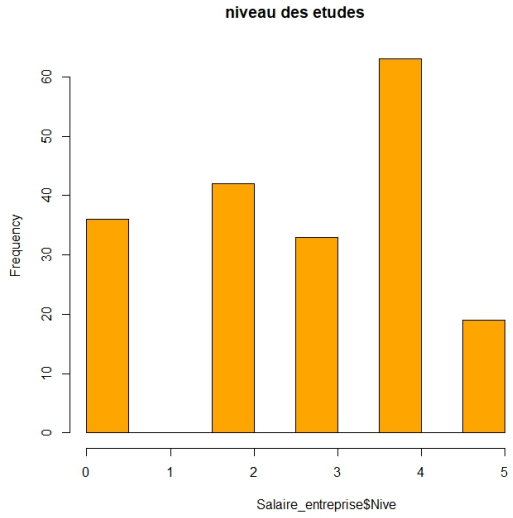
- **Histogramme**




-La classe modale est l'intervalle [1250,2500]
-Les salaires sont centré autour du moyen représenté par la ligne pointillée

## 1-3 Variable:Niveau d'étude

- **Histogramme du variable : niveau d'étude**




-Le mode de ce histogramme est bac+4 , ce qui montre que les employées titulaire d'un bac+4 sont est la plus fréquentes dans cette entreprise
-les employées ayant bac+5 sont les moins fréquentes

**Conclusion 1:** Les employés de cette entreprise ont une ancienneté distribué autour de la moyen 5.43 ans avec une petite dispersion, par contre leurs salaires sont mal distribué avec une forte dispersion, cette dispersion est dû à l'existence d'un petit nombre de personnes ayant **BAC+5** et qui touchent des salaires loin et plus grand que les autres

# 4. Statistique bi-variées

## i. Tableau de fréquence de niveau d'étude en fonction de sexe :

Le premier caractère Niveau d'étude(Nive) ayant comme modalités: **0:bac** , **2:bac+2** , **3:bac+3**, **4:bac+4** , **5:bac+5**

Le 2em est le sexe d'employé (gender) ayant comme modalités: **1** qui revient à **homme** et **0** pour signifier **femme**

| Salaire_entreprise.gender | Salaire_entreprise.Nive 0 | 2 | 3 | 4 | 5 | Sum |
|---|---|---|---|---|---|---|
| 0 | 17 | 23 | 14 | 26 | 7 | 87 |
| 1 | 19 | 19 | 19 | 37 | 12 | 106 |
| Sum | 36 | 42 | 33 | 63 | 19 | 193 |

-Sur un série de 193 observations contenant 106 hommes et 87 femmes on remarque que les femmes plus nombreux que les hommes dans la catégorie bac+2 , et l'inverse dans les autre niveaux d'étude

## *ii.* Tableau de fréquence représentant les intervalles salaires en fonction du sexe:*(0:femme et 1:homme)*

```
> addmargins(tab2)
                      Salaire_entreprise.Sal_int
Salaire_entreprise.gender (0,1.5e+03] (1.5e+03,3e+03] (3e+03,4.5e+03] (4.5e+03,6e+03] (6e+03,7.5e+03] (7.5e+03,9e+03] (9e+03,1.05e+04] (1.05e+04,1.2e+04] Sum
                        0           0               1              49              26               4               2               3                2  87
                        1           0               3              45              37               9               4               0                8 106
                        Sum         0               4              94              63              13               6               3               10 193
> |
```

→ D'après le tableau des effectifs qui donne le nombre des femmes/hommes dans chaque intervalle, Il semble que les femmes sont plus nombreuse dans la catégorie correspondante a l'intervalle des salaires **[3000,5000[** et **[9000,10500[** , et dans les autres intervalles les hommes qui prédominent surtout les catégories des salaires les plus grands.

### iii. Tableau d'effectif intervalles de salaires*Niveau d'étude

```
> addmargins(tab4)
                          Salaire_entreprise.Nive
Salaire_entreprise.Sal_int   0   2   3   4   5 Sum
        (0,1.5e+03]          0   0   0   0   0   0
        (1.5e+03,3e+03]      4   0   0   0   0   4
        (3e+03,4.5e+03]     32  42  20   0   0  94
        (4.5e+03,6e+03]      0   0  13  50   0  63
        (6e+03,7.5e+03]      0   0   0  13   0  13
        (7.5e+03,9e+03]      0   0   0   0   6   6
        (9e+03,1.05e+04]     0   0   0   0   3   3
        (1.05e+04,1.2e+04]   0   0   0   0  10  10
        Sum                 36  42  33  63  19 193
```

→ On voit l'augmentation des salaires proportionnellement à l'augmentation de niveau d'étude, et on peut dire qu'il y' a une relation linéaire entre salaire et niveau d'étude

### iv. Tableaux de fréquence des salaires selon le sexe des employés pour chaque niveau d'étude:

```
> addmargins(tab3)
, , Salaire_entreprise.Nive = 0

                      Salaire_entreprise.Sal_int
Salaire_entreprise.gender (0,1.5e+03] (1.5e+03,3e+03] (3e+03,4.5e+03] (4.5e+03,6e+03] (6e+03,7.5e+03] (7.5e+03,9e+03] (9e+03,1.05e+04] (1.05e+04,1.2e+04] Sum
                        0           0               1              16               0               0               0                0  17
                        1           0               3              16               0               0               0                0  19
                        Sum         0               4              32               0               0               0                0  36

, , Salaire_entreprise.Nive = 2

                      Salaire_entreprise.Sal_int
Salaire_entreprise.gender (0,1.5e+03] (1.5e+03,3e+03] (3e+03,4.5e+03] (4.5e+03,6e+03] (6e+03,7.5e+03] (7.5e+03,9e+03] (9e+03,1.05e+04] (1.05e+04,1.2e+04] Sum
                        0           0               0              23               0               0               0                0  23
                        1           0               0              19               0               0               0                0  19
                        Sum         0               0              42               0               0               0                0  42

, , Salaire_entreprise.Nive = 3

                      Salaire_entreprise.Sal_int
Salaire_entreprise.gender (0,1.5e+03] (1.5e+03,3e+03] (3e+03,4.5e+03] (4.5e+03,6e+03] (6e+03,7.5e+03] (7.5e+03,9e+03] (9e+03,1.05e+04] (1.05e+04,1.2e+04] Sum
                        0           0               0              10               4               0               0                0  14
                        1           0               0              10               9               0               0                0  19
                        Sum         0               0              20              13               0               0                0  33

, , Salaire_entreprise.Nive = 4

                      Salaire_entreprise.Sal_int
Salaire_entreprise.gender (0,1.5e+03] (1.5e+03,3e+03] (3e+03,4.5e+03] (4.5e+03,6e+03] (6e+03,7.5e+03] (7.5e+03,9e+03] (9e+03,1.05e+04] (1.05e+04,1.2e+04] Sum
                        0           0               0               0              22               4               0                0  26
                        1           0               0               0              28               9               0                0  37
                        Sum         0               0               0              50              13               0                0  63

, , Salaire_entreprise.Nive = 5

                      Salaire_entreprise.Sal_int
Salaire_entreprise.gender (0,1.5e+03] (1.5e+03,3e+03] (3e+03,4.5e+03] (4.5e+03,6e+03] (6e+03,7.5e+03] (7.5e+03,9e+03] (9e+03,1.05e+04] (1.05e+04,1.2e+04] Sum
                        0           0               0               0               0               0               2                3   2   7
                        1           0               0               0               0               0               4                0   8  12
                        Sum         0               0               0               0               0               6                3  10  19
```

→ Dans la catégorie de **BAC+2** les femmes sont dominants, tous les salaires dans l'intervalle **[3000,500[**

→ Dans le niveau **BAC+3** on remarque que **47.36%** des hommes **(**9 hommes parmi 19 **)** ont pu entrer dans la catégorie des salaires **[4500,6000[,** alors que seulement **28.57%** des femmes (4 femmes parmi 14 ) qui ont pu faire la même chose.

→ La même chose se répète dans **BAC+4** avec des pourcentages **24.3%** des hommes (9 parmi 37) et **15.3%** des femmes (4 parmi 26 femmes) pour l'intervalle des salaires **[6000,7500[**

→ Dans la table correspondant au **BAC+5** on a **66.7%** des hommes(8 hommes parmi 12) contre seulement **28.85%** des femmes (2 femmes parmi 7) **touchent un salaire dans [10500,12000]**

Alors **100%** des salaire de la catégorie **[9000,10500[**sont des femmes.

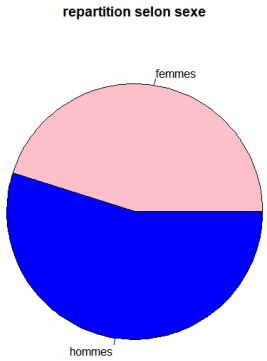
**Conclusion 2:** Dans chaque niveau d'étude on voit une supériorité remarquables des hommes sur le compte des femmes dans les salaires net, le niveau scolaire étant le même et tous les autres conditions sont identiques on peut alors s'interroger sur cette comparaison.

Pourquoi cette différence est ce qu c'est la sexisme? ou bien c'est juste parce que les hommes sont plus compétents et plus performants que les femmes?

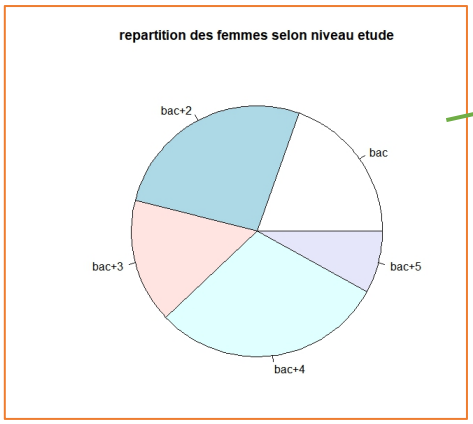
## 3. Comparaisons H vs F & Relations entres variables

- **Répartition des femmes et des hommes**

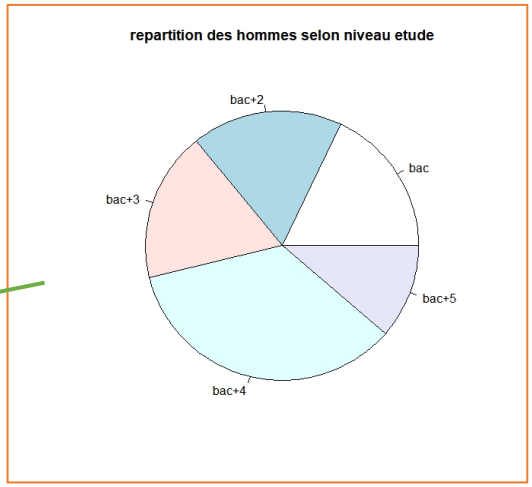
Avec 106 hommes et 87 femmes, les hommes prédomine par 54.9% contre 45% des femmes




- **Répartition des niveaux d'étude selon sexe par des secteurs**

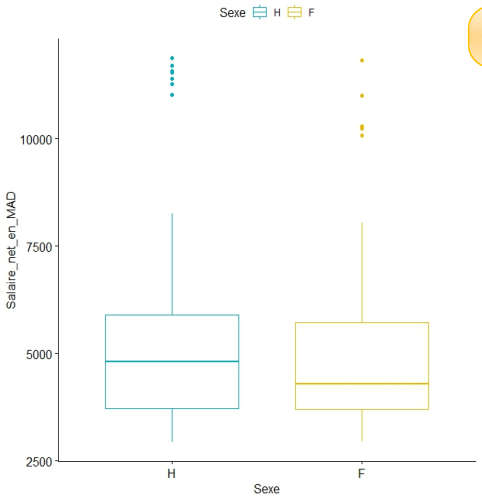



les femmes ayant bac+2 et bac+4 sont plus fréquentes, et BAC+5 ne présent qu'une très petite




les hommes ayant bac+4 suivi par bac+3 sont plus fréquents ,

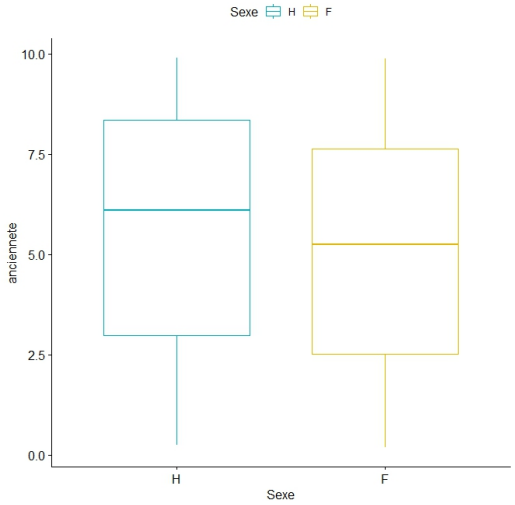
- **Comparaison par les boites à moustaches**




**Pour les salaires net**

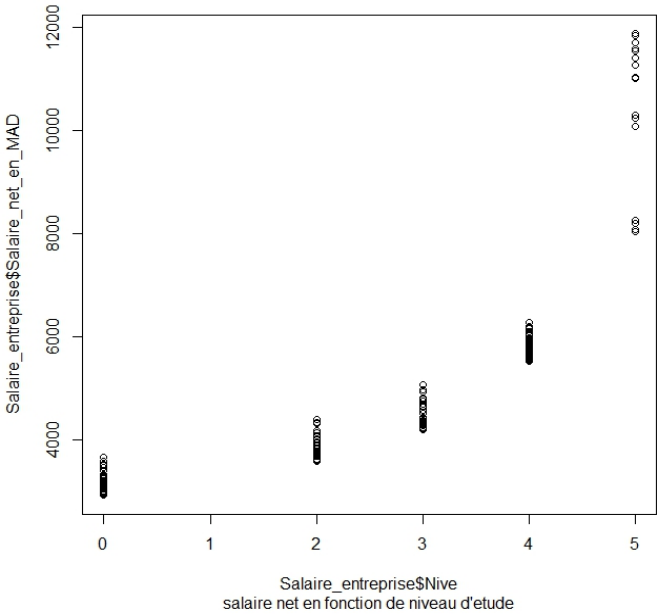
-Selon le dessin des deux boxplot on remarque que la médiane des salaires des hommes est légèrement plus grand que celle des femmes
- Il semble que la boite à moustache des hommes est symétrique par contre celle des femmes est asymétrique

**Pour l'ancienneté**




-Selon le dessin des deux boxplot de l'ancienneté on trouve que la médiane des années d'expérience des hommes est plus grand que celle des femmes
- en générale 50% des homme sont centré sur un nombre d'années d'expérience plus grand que les femmes

## ◆ Relation entre salaire et niveau d'étude




- Ce graphique représente employés selon leurs niveaux d'étude (0 signifie BAC, 1 signifie BAC+2 .... et ainsi de suite) et leurs salaires net en MAD
- On remarque clairement d'après ce graphique que les employées ayant un niveau d'étude plus haute ont une forte chance d'avoir un salaire plus grand ce qui coïncide avec la réalité, le salaire donc est croissante en fonction de niveau d'étude

**Conclusion 3:** Une supériorité pour les hommes dans les niveaux d'étude de BAC+4 et BAC+5 et BAC+4 ce qui explique peut la supériorité dans les moyens des salaires et de l'ancienneté

**Conclusion générale :** l'ancienneté dans cette entreprise est de moyen 5.4 ans , avec une supériorité des hommes. cette supériorité s'étale aussi sur le niveau d'étude qui entraîne encore une supériorité en générale dans les salaires grâce au relation linéaire croissante entre salaire et niveau d'étude. Cependant dans les mêmes niveau d'études il y' a des différence de salaire entre femmes et homme ce qui peut être explicable par une sorte de sexisme dans cette entreprise ou bien par la supériorité dans le rendement des hommes. Au niveau des salaire il' y a une forte dispersion à cause d'un petit groupe des personnes ayant BAC+5 et qui touchent des salaires plus grands et plus loin de la moyen et de la médiane.