Université de Thiès **Année Académique 2019-2020**
U.F.R S.E.S **Master 1 SDA**

Travail Pratique de SDD4142-Statistiques

Groupe 3 **: Souhoude OUEDRAOGO**

Ismael YODA

Lassana BA

1. Enregistrons les données dans un format adapté pour une lecture par la suite avec Python et Vérifions si nous avons une structure de 799 observations et 26 variables.

   Nous avons enregistrés nos données dans un format adapte et importer sur python grâce a "pandas.read_csv", avec "shape" nous constatons effectivement une structure de 799 observations et 26 variables.

2. Changeons les types des variables

   Nous avons changé le types des variables avec la fonction "astype()" et après vérification avec "dtype" nous obtenons les résultats attendues.

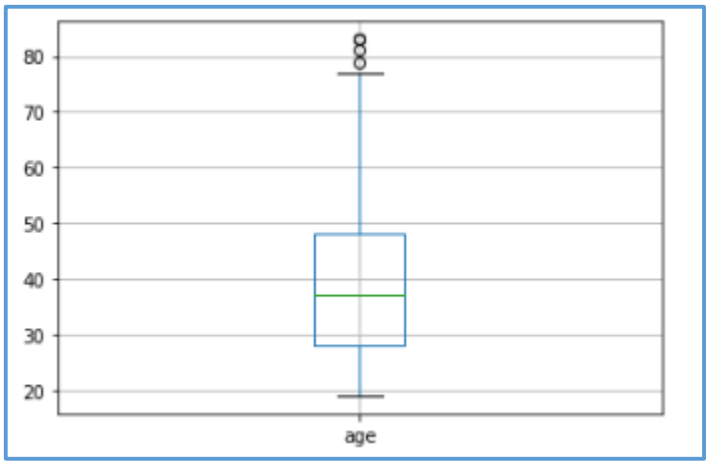
age	float64	ago.cons	category
prof	category	ptsd.cons	category
duree	category	alc.cons	category
discip	category	subst.cons	category
n.enfant	float64	scz.cons	category
n.fratrie	int64	char	category
ecole	category	rs	category
separation	category	ed	category
juge.enfant	category	dr	category
place	category	suicide.s	float64
abus	category	suicide.hr	category
grav.cons	category	suicide.past	category
dep.cons	category	dur.interv	float64

3. . Calculons la moyenne, la variance, et l'écart type pour chacune des variables age, n.enfant, n.fratrie, dur.interv. et donnons les 3 premiers quantiles pour la variable age.

```
Pour la variable age

age moyen de ces prisonniers est :  38.89962358845671
la variance de l'age est :  176.1630738859178
l'eccart type de l'age est :  13.27264381673515
le 1er  quantile de la variable age est : 28.0
le 2eme  quantile de la variable age est :  37.0
le 3eme  quantile de la variable age est :  48.0

Pour la variable n.fratrie

la taille moyenne de la fratrie de ces prisonniers est ;  4.286608260325407
la variance du taille de la fratrie est :  11.828994628767783
l'eccart type du taille de la fratrie est  :  3.439330549506369

Pour la variable dur.inetrv

la duree moyenne de l'interview est :  61.89185580774366
la variance de la duree de l'interview est :  386.3768228577131
l'eccart du taille typede la duree de l'interview est :  19.656470254288106
```

Les calculs ont étés faites en utilisant la librairie numpy et pour le cas spécifique des quartiles, pour la variable age nous avons utilisés "nanquantile" au dépend de "quantile" car la variable contenait des valeurs manquantes.

4. Traçons le boxplot pour la variable age et en tirez des conclusions ?



Nous avons tracé le boxplot avec "boxplot", après avoir visualisé ce nombre avec "value_count" nous obtenons 4 valeurs aberrantes.

5. Affichons les données pour les agriculteurs qui ont plus de 2 enfants.

```
      age         prof duree discip  n.enfant  n.fratrie ecole separation  \
14   64.0  agriculteur   NaN    0.0       3.0          2   1.0        0.0
311  42.0  agriculteur   4.0    0.0       3.0          6   1.0        0.0
390  36.0  agriculteur   4.0    1.0       3.0          4   3.0        1.0
441  79.0  agriculteur   5.0    0.0       5.0          6   2.0        0.0

    juge.enfant place  ... subst.cons scz.cons char   rs   ed   dr suicide.s  \
14          0.0   0.0  ...          0        0  1.0  1.0  1.0  3.0       0.0
311         0.0   0.0  ...          0        0  2.0  1.0  3.0  2.0       3.0
390         1.0   1.0  ...          1        0  1.0  NaN  3.0  1.0       0.0
441         0.0   0.0  ...          0        0  1.0  2.0  1.0  1.0       0.0

    suicide.hr suicide.past dur.interv
14         0.0          0.0       80.0
311        1.0          0.0        NaN
390        0.0          0.0        NaN
441        0.0          0.0       85.0
```

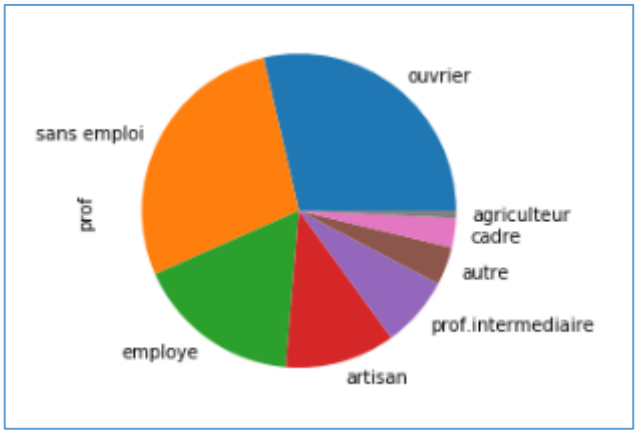
Nous dénombrons quatre (04) personnes respectant ces conditions, nous avons puis filtrer les données grâce à ".loc[,]".

6. Calculons les fréquences des modalités de la variable prof et donnons la catégorie modale

```
ouvrier              227
sans emploi          222
employe              135
artisan               90
prof.intermediaire    58
autre                 31
cadre                 24
agriculteur            6
```

La catégorie modale est ouvrier.

7. Traçons le diagramme circulaire de la variable profession



8. Donnons les moyennes des âges par profession

prof	age
agriculteur	48.833333
artisan	45.111111
autre	34.935484
cadre	50.083333
employe	38.711111
ouvrier	37.396476
prof.intermediaire	43.258621
sans emploi	35.896396

9. Donnons la table des effectifs pour les variables prof incluant les ”NaN”.

```
ouvrier              227
sans emploi          222
employe              135
artisan               90
prof.intermediaire    58
autre                 31
cadre                 24
NaN                    6
agriculteur            6
Name: prof, dtype: int64
```

10. Donnons le nombre de ”Nan” pour chaque variable.

```
age 2
prof 6
duree 223
discip 6
n.enfant 26
ecole 5
separation 11
juge.enfant 5
place 7
abus 7
```

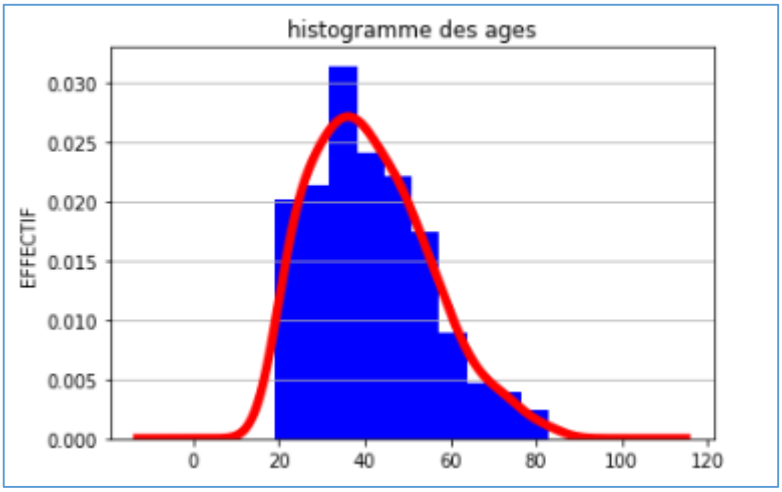
```
grav.cons 4
char 96
rs 103
ed 107
dr 111
suicide.s 41
suicide.hr 39
suicide.past 14
dur.interv 50
```

Les variables qui ont le plus de valeurs manquantes sont la durée, ed, dr.

11. Supprimer toutes les lignes contenant des ”Nan”.

Apres suppression on se retrouve avec 403 observations
(403, 26)

12. Traçons l'histogramme et la densité de la variable age sur la même figure.



13. Discrétisons la variable age. Pour ce faire nous ajouterons une variable dans le DataFrame des données une nouvelle variable nommée age_classe. Cette variable aura 4 classes :[min(age), Q1],]Q1, Q2],]Q2, Q3],]Q3, max(age)].

```
7      D
8      C
12     A
13     B
16     D
      ..
793    A
795    C
796    A
797    B
798    D
Name: age_classe, Length: 403, dtype: category
Categories (4, object): [A, B, C, D]
```

14. Donnons les fréquences des modalités de la nouvelle variable age_classe.

```
A    107
B    104
D     96
C     96
```