

Tidying

1) Chaque colonne \Leftrightarrow une variable

ex): M15 : Nasulin - 15 ans
heure | jour | mois \Rightarrow date

2) Chaque ligne \Leftrightarrow une observation

Ville 12-25 25-52

Paris 1,6 2,3

Ville	Classe d'Âge	Population
Paris	12-25	1,6

3) Chaque type d'observation (\Rightarrow une table)

Nbre d'éoutes de morleauc par jour

par plateforme

. Revenus générés par jour par plateforme

4) One Hot Encoding s/ variables

Catégories

Checking the data

1) Missing Data:

- . + 10% de val manquantes
- . \exists autre colonne/variable corrélée avec celle ci ?

Si pas le cas : ~~10%~~ 30%

2) Outliers:

$[Q1 - 1,5 \cdot IQR, Q3 + 1,5 \cdot IQR]$

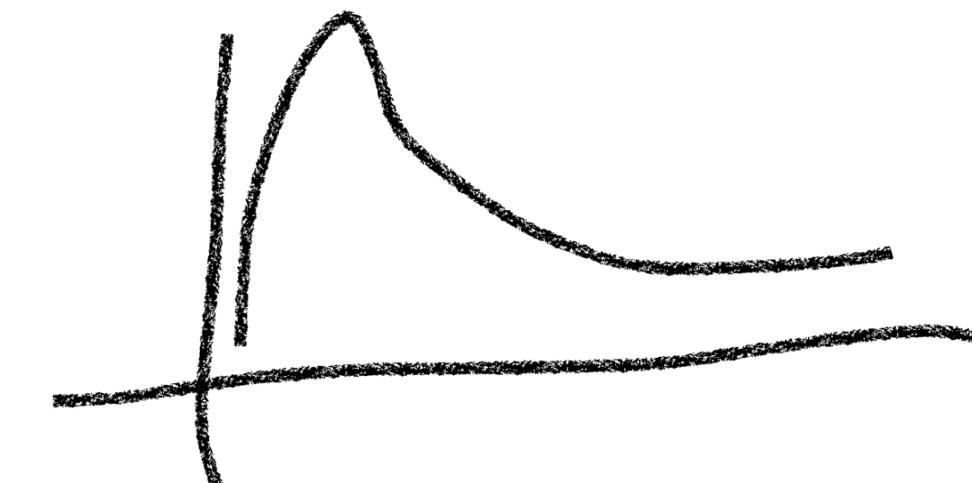
Moyennes, Écart-Type, Corrélations

3) Darblons / Erreurs de la beling / etc.

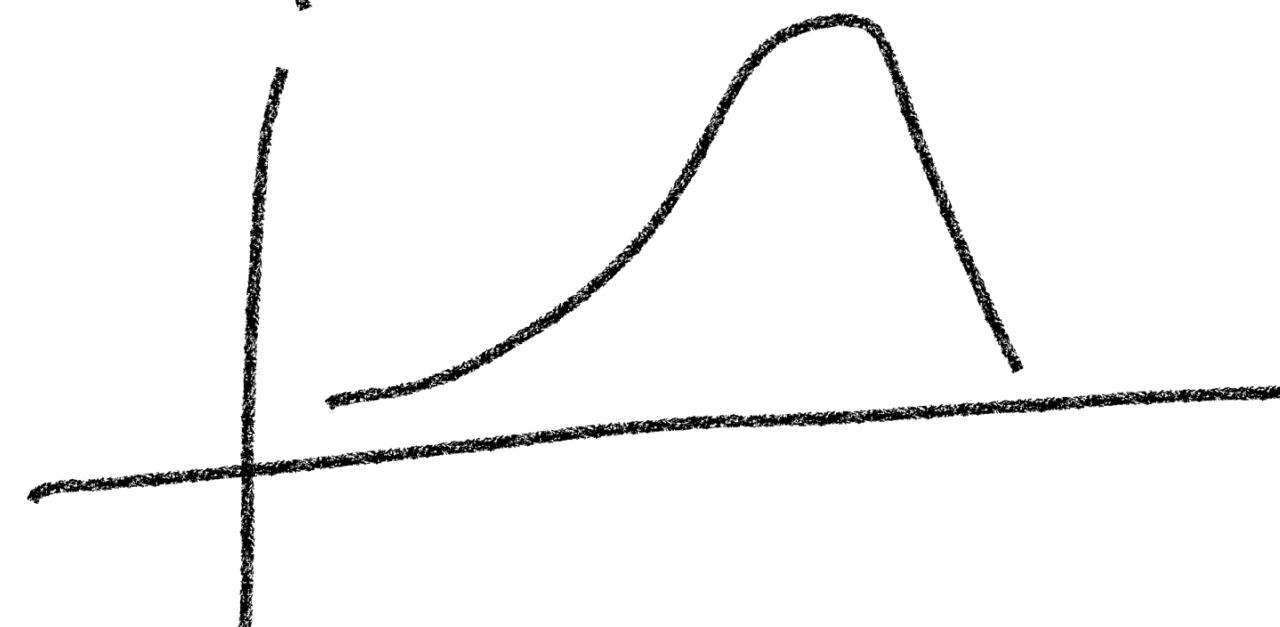
Value-count()

4) Variables numériques :

Si bris, Alors :



Concave (\log ,
 \sqrt{x})



Convex (x^2 ,
 e^x)

5) Nécessité de créer de nouvelles variables:

t	x	y	v_x	v_y	a_x	a_y
---	---	---	-------	-------	-------	-------

⚠ Toutes les opérations et les choix que vous effectuez doivent être notés.