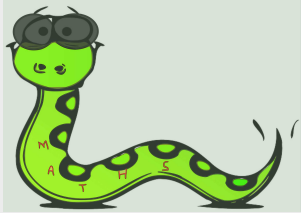


# Enseigner la statistique dans le secondaire avec Python

## Fouille de données - Visualisation - Modélisation

Bro Frédéric

(Lycée Henri-Moissan - Meaux - Académie de Créteil)

CFIES - 25 Septembre 2019

# Plan de la conférence

# La science des données

## Pourquoi ?

Le numérique est intégré dans nos vies et nous y laissons d'énormes traces à traiter !
La « **Science des données** » (Data Science) est la **nouvelle discipline** qui permet d'étudier ces nouvelles statistiques en utilisant :

- l'informatique
- les outils mathématiques
- la visualisation des données et la statistique

**Emergence de 2 nouveaux métiers :**

- le « Scientifique de données » (Data scientist)
- le « Journaliste de données » (Data journalist)

# Data Scientist

Dans n'importe quel classement des métiers les mieux payés se trouve en tête le **data scientist** !

## Son Rôle ?

- Fouiller
- Collecter les données
- Les Trier et ou les Croiser
- **Analyser des données** conséquentes pour :
  - **Modéliser** des phénomènes
  - **Prédire**
  - **Prendre des « bonnes » décisions**.

# Data Journalist

Dans n'importe quel journal d'investigation ou d'information, on **informe** « quasi quotidiennement » le « public » de résultats d'enquêtes ou études statistiques !

## Son Rôle ?

- Fouiller
- Collecter les données
- Filtrer les données
- Représenter pour mieux **expliquer** ces données « opaques » qui nous entourent
- **Lancer des alertes**

# Pourquoi la Data visualization ?

## Traduction

data visualization = **visualisation de données** !

## Data Visualization c'est avant tout

- utiliser des **graphiques pertinents** :
  - diagrammes en barres, circulaires, histogrammes
  - nuages de points
  - cartes choroplètes, etc.
    Exemples cartes : Nuage stations et Exemple cartes

  pour **résumer avec clarté** les statistiques étudiées et **soulever des problèmes** ou **enjeux majeurs** !
- permettre la comparaison d'individus
  *(boîtes à moustaches, courbes de Lorentz ...)*

# Nouveaux programmes

**Quelques capacités attendues :**

- Comparer deux séries statistiques, en s'appuyant sur des indicateurs ou sur des représentations graphiques données
- Pour des données réelles ou issues d'une simulation, calculer la proportion d'éléments compris dans

$$[m-2s\,;\,m+2s]$$

- Au moins un traitement statistique de fichiers de données individuelles anonymes
- Sélectionner des données selon un critère (filtre, ET, OU, NON)
- Dresser le tableau croisé de 2 variables et calculer les fréquences conditionnelles ou marginales

Voilà donc le ...

# Pandas

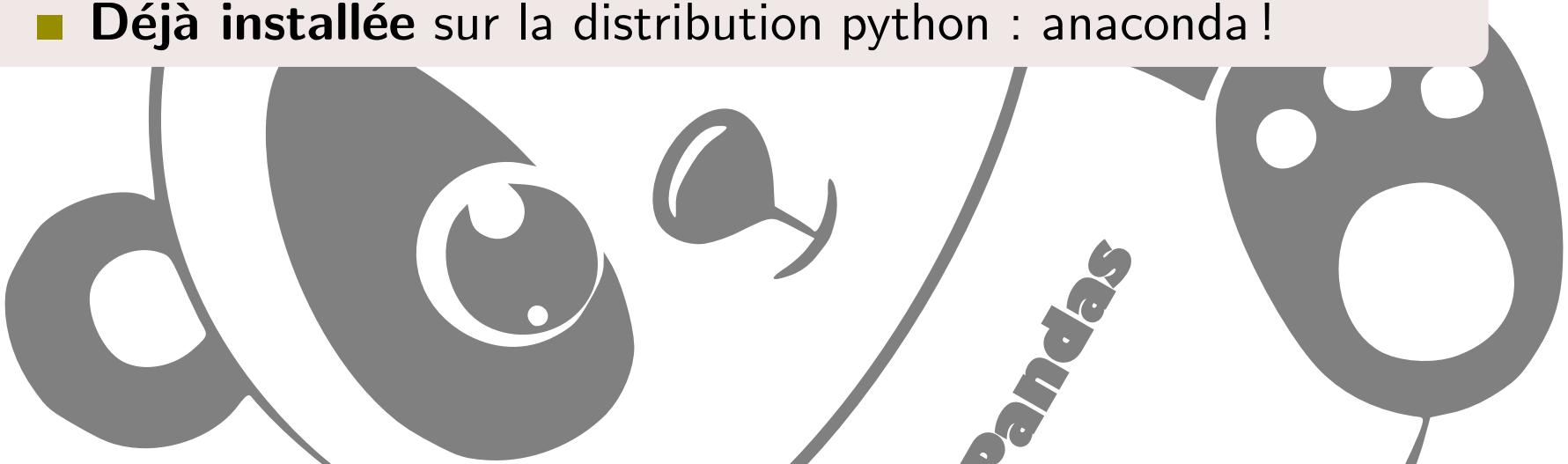
## C'est quoi ?

- **Pandas est la bibliothèque de Python** créée en 2008 pour travailler les statistiques

$$\textbf{Pandas} = \textbf{Panel} + \textbf{Data}$$

**Panel** parce qu'on peut travailler avec des tables contenant **un ou plusieurs index** !

- **Déjà installée** sur la distribution python : anaconda !



# 3 principales commandes

## Puissance de pandas == Manipuler les données

| Commande | Rôle |
|---|---|
| `query` | Sélectionner des données |
| `groupby` | Regrouper des données |
| `crosstab` | Croiser 2 variables |

**Illustrations via document ressources de première**

Voir les activités :

**Titanic** - **les emplois en France** - **Sécurité routière** - **le Tennis**

# Un peu de thé ?

Une étude a été menée auprès de 300 consommateurs de thé.
Les résultats sont enregistrés dans le fichier `'the.csv'`.

In [1]:
```
import pandas as pa
import pylab as pl

T = pa.read_csv('the.csv')
T.head()
```

Out[1]:

| | tout_moment | variete | comment | sucre | forme | type | sexe | CSP | sportif | age | frequence | plus_pour_la_sante |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | noir | pur | 1 | sachet | inconnu | H | cadre moyen | 1 | 39 | 1/jour | 1 |
| **1** | 0 | noir | lait | 0 | sachet | variable | F | cadre moyen | 1 | 45 | 1/jour | 1 |
| **2** | 0 | parfumé | pur | 0 | sachet | variable | F | autre actif | 1 | 47 | + de 2/jour | 1 |
| **3** | 0 | parfumé | pur | 1 | sachet | variable | H | étudiant | 0 | 23 | 1/jour | 1 |
| **4** | 1 | parfumé | pur | 0 | sachet | variable | H | employé | 1 | 48 | + de 2/jour | 0 |

# Un peu de thé ?

## Sélectionner les femmes de plus de 30 ans :

```
In [2]:  T1 = T.query('sexe == "F" and age >= 30')
```

## Sélectionner les sportifs ou les moins de 25 ans :

```
In [3]:  T2 = T.query('age <= 25 or sportif == 1')
```

# Variable « sexe » croisée avec « comment »

In [4]:
```
E = pa.crosstab(T['sexe'],T['comment'])
E
```

Out[4]:

| sexe | comment autre | citron | lait | pur |
|---|---|---|---|---|
| **F** | 6 | 19 | 32 | 121 |
| **H** | 3 | 14 | 31 | 74 |

C'est le **tableau croisé des effectifs** des 2 variables : **sexe** et **comment**

**Pour la suite, on note :**

- $F$ : l'ensemble des **femmes**
- $C$ : l'ensemble des buveurs de thé avec du **citron**
- $L$ : l'ensemble des buveurs de thé avec du **lait**

# Transposer le tableau précédent

In [5]:
```
E.T
```

Out[5]:

| sexe | F | H |
|---|---|---|
| comment | | |
| autre | 6 | 3 |
| citron | 19 | 14 |
| lait | 32 | 31 |
| pur | 121 | 74 |

# Tableau marginal

In [6]:
```
M = pa.crosstab(T['sexe'],T['comment'],
                margins=True)
M
```

Out[6]:

| comment<br>sexe | autre | citron | lait | pur | All |
|---|---|---|---|---|---|
| **F** | 6 | 19 | 32 | 121 | 178 |
| **H** | 3 | 14 | 31 | 74 | 122 |
| **All** | 9 | 33 | 63 | 195 | 300 |

# Fréquences conditionnelles

Out[6]:

| comment<br>sexe | autre | citron | lait | pur | All |
|---|---|---|---|---|---|
| **F** | 6 | 19 | 32 | 121 | 178 |
| **H** | 3 | 14 | 31 | 74 | 122 |
| **All** | 9 | 33 | 63 | 195 | 300 |

$F$ : ensemble des femmes  $L$ : buveurs de thé avec du lait

## Calculs à la « main » :

- $f_F(L) = \dfrac{32}{178} \approx 0,18$
- $f_H(L) = \dfrac{31}{122} \approx 0,25$

## Calculs avec python :

In [7]:
```
Freq = M['lait']/M['All']
Freq
```

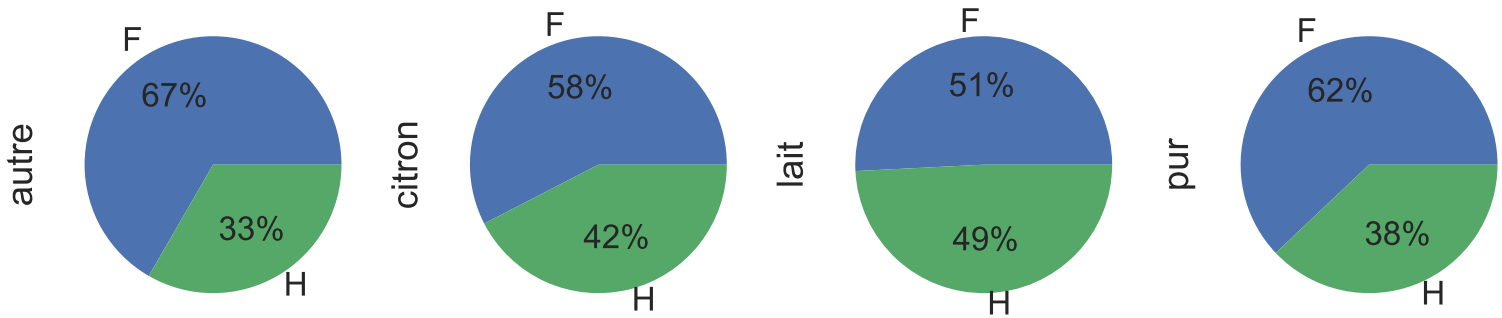
Out[7]:

| sexe | |
|---|---|
| F | 0.179775 |
| H | 0.254098 |
| All | 0.210000 |

# Visualisation des fréquences

**In [8]:**
```
E.plot.pie(subplots=True,legend=None,
                    autopct='%.0f%%')
```

**Out[8]:**




**Interprétation :**

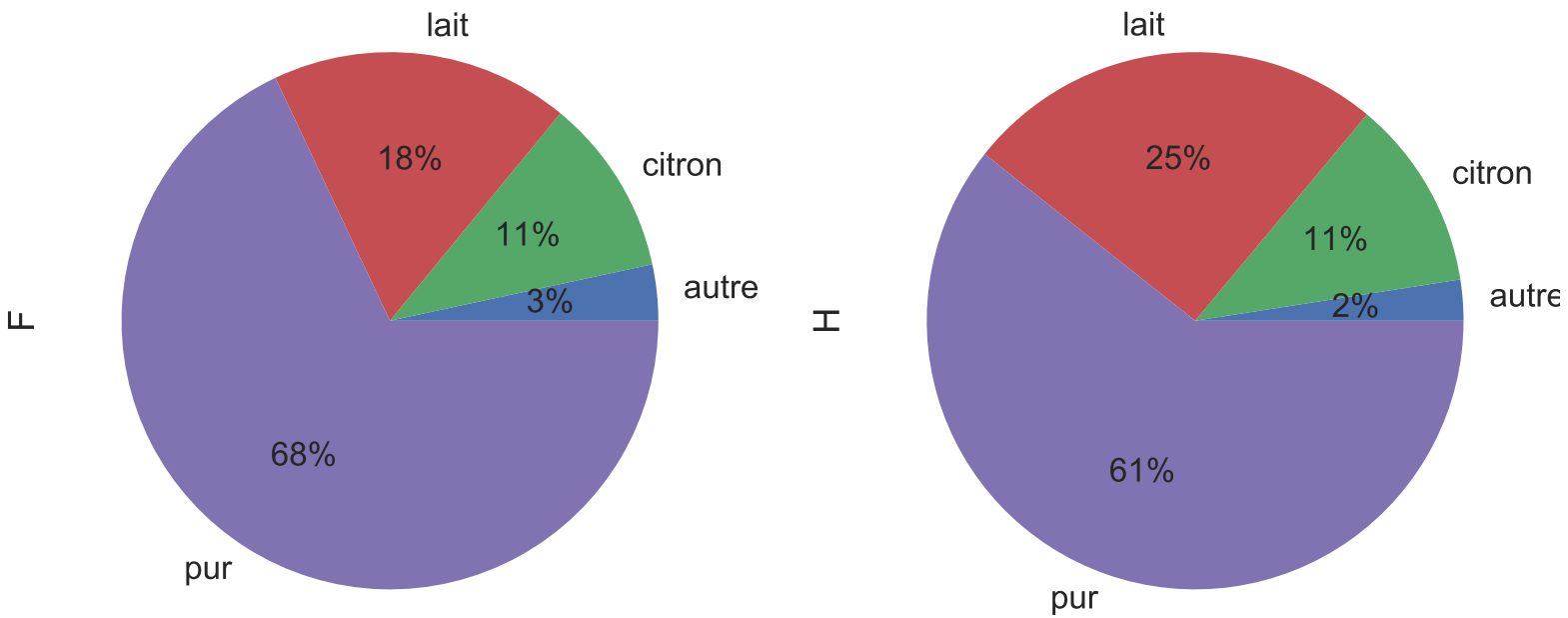
**Diagramme 2 :** $f_C(F) = 58\%$, $f_C(H) = 42\%$

**Diagramme 3 :** $f_L(F) = 51\%$, $f_L(H) = 49\%$

# Visualisation des fréquences

```
In [9]: E.T.plot.pie(subplots=True,legend=None,
                     autopct='%.0f%%')
```

Out[9]:




**Interprétation :**

**Diagramme 1 :** $f_F(L) = 18\%$, $f_F(C) = 11\%$

**Diagramme 2 :** $f_H(L) = 25\%$, $f_H(C) = 11\%$

# Regroupement selon « sexe » : calcul de moyennes

$P$ : ensemble des personnes qui pensent que « boire du thé est un plus pour la santé »

```
In [10]: T.groupby('sexe').mean()
```

Out[10]:

| sexe | tout_moment | sucre | sportif | age | plus_pour_la_sante |
|---|---|---|---|---|---|
| F | 0.353933 | 0.398876 | 0.544944 | 37.353933 | 0.713483 |
| H | 0.327869 | 0.606557 | 0.672131 | 36.614754 | 0.680328 |

- Âge moyen femmes : 37 ans contre 36 ans pour les hommes.
- $f_F(P) \approx 71\%$ et $f_H(P) \approx 68\%$

# Regroupement selon « sexe » et « sucre »

```
In [11]: T.groupby(['sexe','sucre']).mean()
```

Out[11]:

| | | tout_moment | sportif | age | plus_pour_la_sante |
|---|---|---|---|---|---|
| sexe | sucre | | | | |
| F | 0 | 0.336449 | 0.551402 | 40.757009 | 0.757009 |
| | 1 | 0.380282 | 0.535211 | 32.225352 | 0.647887 |
| H | 0 | 0.312500 | 0.645833 | 41.145833 | 0.687500 |
| | 1 | 0.337838 | 0.689189 | 33.675676 | 0.675676 |

$S$ : ensemble des personnes qui sucrent leur thé

- En moyenne, les personnes qui sucrent leur thé sont plus jeunes.
- $f_{F\cap\overline{S}}(P)\approx 76\%$ et $f_{F\cap S}(P)\approx 65\%$
- $f_{H\cap\overline{S}}(P)\approx 69\%$ et $f_{H\cap S}(P)\approx 68\%$

# Apports pour nos élèves ?

## Travailler avec pandas c'est

- Faire vivre les maths dans de nombreux domaines :

  Physique - SVT - Géographie - Economie - Informatique ...

- Manipuler toutes les compétences :
  - Chercher
  - Modéliser
  - Représenter
  - Raisonner
  - Calculer
  - Communiquer
- Sensibiliser et faire travailler autrement les élèves :

  Élève = Futur citoyen !

- Donner des clefs à un futur oral ...

# Plan de la conférence


# Plan de la conférence

# Merci pour votre attention.

- Toutes les ressources sont dans le **feuillet** avec un **petit memento pandas**.
- Des suppléments :
  - BRO F. et REMY C. (2016), Python et les 40 Problèmes Mathématiques, Ellipse.
  - Conférence SFDS : UPEM Mercredi 14 Mars 2018

    «Statistique en mouvement : leçons du passé et perspectives d'avenir»

**Je fais des stats...**