# Résumé des méthodes de l’analyse univariée :

Table des matières

[Résumé des méthodes de l’analyse univariée : 1](#_Toc145512835)

[1. Introduction 2](#_Toc145512836)

[2. Import des données 2](#_Toc145512837)

[3. Exploration des données. 3](#_Toc145512838)

[4. Variables quantitatives 3](#_Toc145512839)

[4.1. Graphiques 3](#_Toc145512840)

[4.2. Indicateur de position 4](#_Toc145512841)

[4.3. Indicateur de dispersion 5](#_Toc145512842)

[4.4. Centrer et réduire 5](#_Toc145512843)

[5. Variables qualitatives 5](#_Toc145512844)

[6. Conclusion 5](#_Toc145512845)

## Introduction

Lors du module de Data Science Révision nous somme amener à lier les contenues théoriques avec les exercices pratiques à réaliser. Nous tenterons dans ce document d’appuyer les connaissances utilisées dans les révisions en voyant dans un cas concret comment elles sont utilisées.

Mais rappelons tout d’abord que la méthodologie d’un data analyste consiste à :

* Explorer les données
* Résumer les données
* Prendre une décision
* Communiquer les résultats.

Nous verrons donc tout d’abord la réalisation de l’import des données puis la manière dont un data analyste peu Résumer de manière pertinente leurs contenus

Dans le cadre de ce TP, les packages « Pandas » et « matplotlib » ont été utilisés. Le jeu de donnée FrenchCities.csv, composé des données météo de 32 villes, composera la base de notre étude.

## Import des données

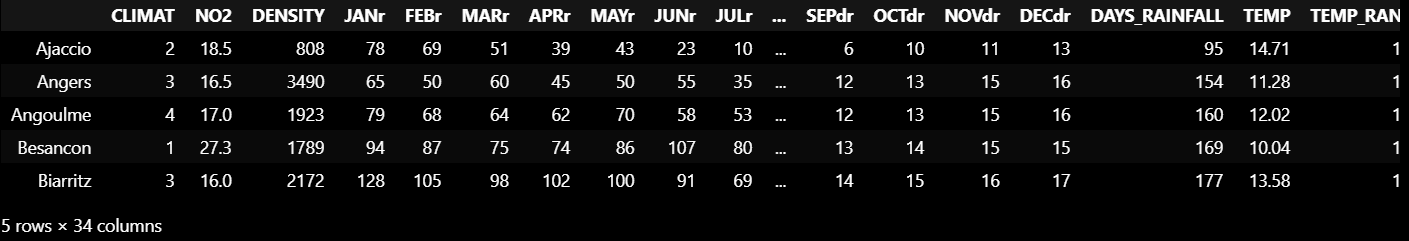
Afin d’apporter les données à étudiant dans python, on utilise le package Pandas qui simplifie la création et la manipulation de dataframe.

Suite à l’import des packages,



Il suffit d’utiliser read\_csv, une fonction de pandas, pour lire un csv et le récupérer sous forme de dataframe exemple :



Il en sort notre dataframe sur Python. 

## Exploration des données.

Il est commun de réaliser une exploration des données afin de vérifier la qualité et la pertinence de celles-ci. Pour ce faire on peut afficher l’index et les colonnes pour se familiariser avec les variables. On peut aussi, grâce à a fonction .info() afficher les types des variables composants le tableau. 

Cela permet de se rendre compte de la qualité des données (s’il y a beaucoup de valeurs manquantes) et des types de données (à première vue) données (quantitative, qualitative). L’étude univarié des données est différente suivant leurs types.

## Variables quantitatives

Il existe plusieurs moyens plus ou moins pertinent suivant les cas, d’étudier les variables quantitatives d’un data frame.

### Graphiques

On peut ainsi réaliser des graphiques pour rendre l’analyse de données plus visuel. Pour les variables discrètes, les diagrammes bâton sont pertinent lorsque l’histogramme se prête plus à l’analyse de variables quantitatives continues.



### Indicateur de position

Il est également nécessaire, lors de l’exploration des données de vérifier les indicateurs de position (**Moyenne**, **Mode**, **Médiane** et **Quartiles**)

La fonction .describe() permet de récupérer facilement ces informations : 

Le graph boxplot ou boite à moustaches est une manière très visuelle de réaliser cette étude.



Cette étude de position nous a permis éliminer une variable aberrante, rendant notre jeu de données plus propre.

### Indicateur de dispersion

Une fois la compréhension de ce que représente chaque variable ainsi que ce que sont leurs valeurs « normales », il est possible de trouver les indicateurs de dispersions. (**Étendu**, **écart-type**, **variance**, **écart-median**)

### Centrer et réduire

Enfin afin de rendre les variables plus pertinentes à étudier et à comprendre, il est fortement conseiller de centrer et réduire les variables. (Seulement les variables quantitatives !!)



## Variables qualitatives

Le traitement des variables qualitative est différents. La plupart du temps, dans ces cas, les informations numériques parlent moins que les représentations graphiques. C’est pour cela que pour étudier la variable ‘CLIMAT’ une étude graphique est choisie. Un diagramme de secteur nous permet de se rendre compte des proportions de villes parmi quatre types de climats différents : 

## Conclusion

En somme le data scientiste doit pour chaque jeu de données, mener une étude rigoureuse avant de pouvoir décider de la pertinence d’une étude dessus. Il doit aussi prévoir lesquelles sont appropriées ou non en fonction de chaque variables (quantitative ou qualitative).