# *Traiter les données : les essentiels*

## Lire les données

```
import pandas as pd
df = pd.read_csv("fichier.csv",encoding='latin1')
df.shape
df.columns
```

## Corriger des valeurs

```
df=df["ville"].str.title()
```

## Retenir certaines variables

```
df=df[["nom","genre","ville"]]
```

## Retenir certaines observations

```
df=df.query("ville == 'Laval'")
```

## Éliminer les observations manquantes

```
df=df.dropna(subset=["connais_python"])
```

## Regrouper des valeurs

```
dico={"Magog":"Reste du Québec",
      "Sherbrooke":'Reste du Québec',
      "Montréal":"Métropole",
       "Laval":"Métropole"}
df["region"]=df["ville"].map(dico).fillna(pd.NA)
```

## Sauvegarder les résultats

```
df.to_csv("donnees.csv", index=False, encoding="latin1")
t6.to_csv("age_moyen_tableau.csv", index=False, encoding="latin1")
```

# Obtenir des résultats

## Une variable qualitative

```
t1=df["genre"].value_counts()
t2=df["genre"].value_counts(normalize=True)
t3=round(df["genre"].value_counts(normalize=True)*100)
```

## Une variable quantitative

```
moyenne=df["age"].mean()
mediane=df["age"].median()
```

## Deux variables qualitatives

```
t4=pd.crosstab(df["ville"],df["joue_musique"],margins=True,margins_name="Total")
t5=round(pd.crosstab(df["ville"],df["joue_musique"],normalize="columns")*100)
t6=round(pd.crosstab(df["ville"],df["joue_musique"],normalize="index")*100)
```

## Deux variables: une qualitative et l'autre quantitative

```
t7=df.groupby("genre")["age"].mean().reset_index()
```

## Deux variables quantitatives

```
correlation = df["age"].corr(df["nb_ordi"])
```

## Un graphique univarié

```
import matplotlib.pyplot as plt
couleurs=["red","blue"]
t3.plot(kind="bar",color=couleurs)
plt.title("Genre", fontsize=16)
plt.xlabel("Genre")
plt.xticks(rotation=0)
plt.ylabel("%")
plt.show()
plt.savefig("mon_graphique.png")
```

## Un graphique bivarié

```
import matplotlib.pyplot as plt
couleurs=["red","blue"]
t7.plot(kind="bar",color=couleurs)
plt.title("Âge moyen par genre")
plt.xlabel("Genres")
plt.ylabel("Âges")
plt.grid(axis='y', linestyle="--")
plt.xticks(rotation=0)
plt.show()
plt.savefig("mon_graphique.png")
```

La source des données de l'exemple

Comprendre les notions de variable et de liste

Comprendre la notion de dataframe

La démarche d'ensemble

Page Google Colab pour un accès à la page qui illustre cet aide-mémoire :ED_Fondamentaux_pandas.ipynb - Colab

Jean-Herman Guay