

Charger des données avec Python

Atelier théorique #3





Les données

Il y a 3 principaux types de données:

1- Données tabulaires : Tout ce qu'on peut retrouver dans un tableau Excel. Peut être constituer de chiffres (numérique) ou de mots (texte).

2- Texte : Données comprenant des séquences de mots, phrases, paragraphes

3- Image : Comprends des photos, vidéos, images satellites, etc.

Dans le cadre de ce cours, on se concentrera sur les données tabulaires numériques.



Format des fichiers de données

Les données peuvent être enregistrées dans plusieurs formats différents selon les besoins.

Le format le plus couramment utilisé est .CSV

- .CSV signifie “*comma-separated values*”
- C’est globalement un fichier texte dont les **valeurs sont séparées par une virgule**, que l’on nomme séparateur en programmation

1	M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15
2	M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7
3	F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9
4	M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10
5	I,0.33,0.255,0.08,0.205,0.0895,0.0395,0.055,7
6	I,0.425,0.3,0.095,0.3515,0.141,0.0775,0.12,8



Jeux de données

Jeu de données principal utilisé en exemple

Données sur le diabète chez les indiens Pima

Téléchargement : <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

Jeu de données pratique

Données sur l'authenticité de billets de banques

Téléchargement :

- <https://archive.ics.uci.edu/ml/datasets/banknote+authentication#>
- <https://www.kaggle.com/shantanuss/banknote-authentication-uci>



Vérification du jeu de données

Avant d'importer le jeu de données dans un Notebook, il est important d'y jeter un coup d'oeil dans le bloc note avant de l'importer pour connaître ses caractéristiques.

```
1 Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome
2 6,148,72,35,0,33.6,0.627,50,1
3 1,85,66,29,0,26.6,0.351,31,0
4 8,183,64,0,0,23.3,0.672,32,1
5 1,89,66,23,94,28.1,0.167,21,0
6 0,137,40,35,168,43.1,2.288,33,1
7 5,116,74,0,0,25.6,0.201,30,0
8 3,78,50,32,88,31,0.248,26,1
9 10,115,0,0,0,35.3,0.134,29,0
10 2,197,70,45,543,30.5,0.158,53,1
```

2 constats à faire ici :

1. Les titres de colonnes sont déjà présents.
2. Les données sont séparées par des virgules.

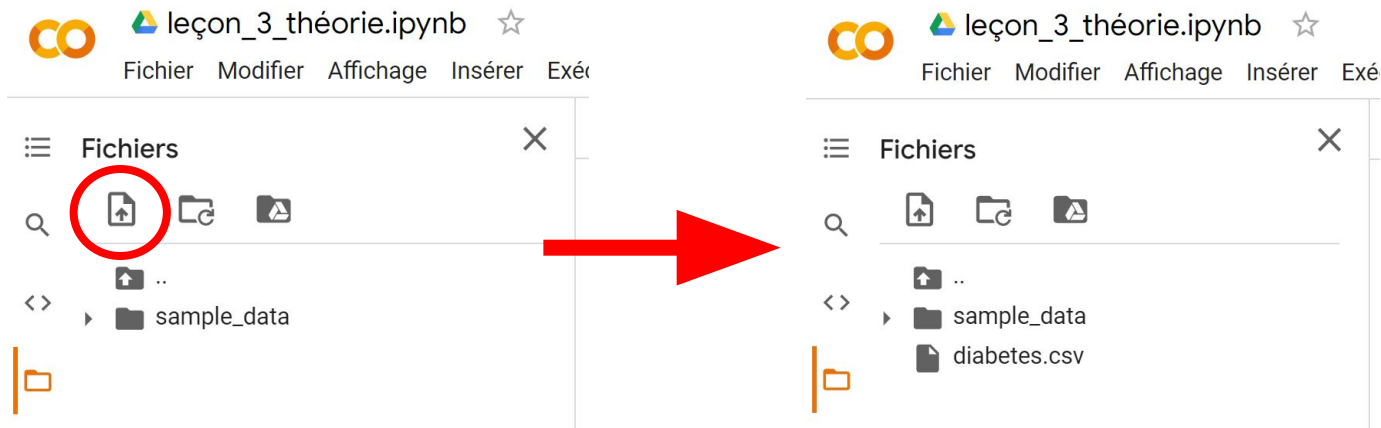


Téléchargement des données

Méthode 1 - Version manuelle

Les données sont disponibles localement sur votre ordinateur.

Utiliser le menu de gauche pour sélectionner le jeu de données manuellement.





Téléchargement des données

Méthode 2 - Version code

Les données sont disponibles localement sur votre ordinateur.

Étapes à suivre :

1. Télécharger le jeu de données sur votre ordinateur local.
2. Importer le jeu de données en mémoire dans le Notebook.
3. Télécharger le jeu de données dans l'environnement de travail du Notebook.

```
# Importer le jeu de données à partir de votre ordinateur local
from google.colab import files
uploaded = files.upload()
```

```
# Télécharger le jeu de données
import io
data = pd.read_csv(io.BytesIO(uploaded['diabetes.csv']))
```



Importation du jeu de données

Maintenant qu'on a vu le jeu de données brutes, il est temps de l'importer dans un Notebook.



EXERCICE - Importer un jeu de données réel dans un Notebook



Jeux de données pour aller plus loin

Répertoire de 500+ jeux de données par l'Université de Californie Irvine (UCI)

<http://archive.ics.uci.edu/ml/>



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems


[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)

[View ALL Data Sets](#)

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 585 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:  In Collaboration With:  **Rexa.info**
Research • People • Connections

Latest News:	Newest Data Sets:	Most Popular Data Sets (hits since 2007):
<p>09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!</p> <p>04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!</p> <p>03-01-2010: Note from donor regarding Netflix data</p> <p>10-16-2009: Two new data sets have been added.</p> <p>09-14-2009: Several data sets have been added.</p> <p>03-24-2008: New data sets have been added!</p> <p>06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope</p>	<p>02-17-2021:  Hungarian Chickenpox Cases</p> <p>12-09-2020:  Myocardial infarction complications</p> <p>10-14-2020:  Gait Classification</p>	<p>3977916:  Iris</p> <p>2146440:  Adult</p> <p>1660143:  Wine</p>

Jeux de données pour aller plus loin

Répertoire de centaines de jeux de données générés par le public sur Kaggle

<https://www.kaggle.com/datasets>

- *Attention! Puisque les jeux de données sur Kaggle sont créés par le public, il est possible qu'ils ne soient pas optimaux et prêts pour utilisation immédiate.*

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset

Your Work

Search datasets

Filters

Datasets

Tasks

Computer Science

Education

Classification

Computer Vision

NLP

Data Visualization

Trending Datasets

See All



Protests against mass tourism(2014-2017)

IshaDS · Updated an hour ago
Usability 8.2 · 30 KB
1 Task · 1 File (CSV)

1



Covid-19 period air-traffic dataset

IshaDS · Updated 3 hours ago
Usability 9.4 · 177 MB
1 Task · 1 File (CSV)

2



Dogecoin historical data (08.03.2019 - 05.05.2021)

Zackery M · Updated 4 hours ago
Usability 8.2 · 10 KB
1 File (CSV)

7



GDP annual growth for each country (1960 - 2020)

Zackery M · Updated 4 hours ago
Usability 9.7 · 111 KB
1 Task · 1 File (CSV)

4

