

ANALYSE DE DONNÉES AVEC PYTHON

Enseignant : Louis RAYNAL

Contact : l-raynal@ices.fr

Public : L1 Maths














Ressources : <https://github.com/LouisRaynal/pythonData>

Établissement : ICES

Année : 2024-2025

TP 3 : Analyse du jeu de données Titanic

Mémos pour Mac :

{ =  +  } =  +  | =  +  + 
[=  +  + ] =  +  + 

Selon la version du clavier  peut être à remplacer par 

Au cas où vous avez un problème avec les fenêtres de Spyder :

View > Window Layouts > Spyder Default Layout

Pratique

Ouvrir le navigateur Anaconda (*Anaconda-navigator*), puis Spyder.

Ouvrir un nouveau fichier *.py* dans Spyder, que vous sauvegarderez sur votre bureau avec le nom *TP3_nom_prénom.py*.

Ajouter en début du fichier, votre nom et prénom en commentaire, puis les lignes d'import des librairies que nous pourrions utiliser lors de ce TP, soit :

```
# Votre nom et prénom ici

import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as ss
```

Ajouter également cette ligne qui vous permettra d'afficher toutes les colonnes des DataFrames pandas.

```
pd.options.display.max_columns = None
```

Exécuter ces codes.

Pour ce TP, avant de répondre à chaque question, veuillez à écrire le numéro de la question en commentaire. Par exemple :

```
# Question X :
```

Pour les questions nécessitant de commenter des résultats, le faire dans votre fichier Python, dans des lignes de commentaires.

Présentation du jeu de données Titanic

Lors de ce TP vous allez analyser des données concernant des passagers du Titanic. Il s'agit d'un échantillon de 891 passagers différents, parmi les 2240 qui occupaient le Titanic avant qu'il ne heurte un iceberg et ne coule.

Les variables suivantes sont recueillies sur chacun des passagers :

- **PassengerId** : un identifiant unique propre à chaque passager
- **Survived** : une variable indiquant si le passager a survécu ou non (1 = Survécu, 0 = Décédé)
- **Pclass** : la classe du passager encodée de la manière suivante :
 - 1 = première classe
 - 2 = deuxième classe
 - 3 = troisième classe
- **Name** : le nom du passager
- **Sex** : le sexe du passager (**male** ou **female**)
- **Age** : l'âge du passager en année (peut être un nombre décimal)
- **SibSp** : le nombre de frères/soeurs/époux/épouses à bord du Titanic pour ce passager
- **Parch** : le nombre de parents/enfants à bord du Titanic pour ce passager
- **Ticket** : le numéro du ticket du passager
- **Fare** : le prix du ticket du passager en livres (£)
- **Cabin** : le numéro de cabine du passager
- **Embarked** : le port où a embarqué le passager, encodé de la manière suivante :
 - C = Cherbourg
 - Q = Queenstown
 - S = Southampton

Chargement et nettoyage du jeu de données

1. Les données se trouvent dans le fichier *titanic.csv*, à l'adresse <https://github.com/LouisRaynal/pythonData/tree/main/TP3>.
Rendez-vous à cette adresse avec le navigateur web **Google chrome** et sauvegardez ce fichier de la manière suivante :
 - clic gauche sur le nom du fichier *titanic.csv*
 - faire ensuite un clic gauche sur le bouton
 - puis faire un **clic droit** puis **Enregistrer sous...**, et sauvegarder le fichier sur votre bureau avec le nom *titanic.csv* (ajouter si besoin l'extension .csv si elle n'est pas présente)

Le fichier se trouve également sur le Classroom dédié.

2. Chargez dans Spyder le fichier *titanic.csv* afin d'obtenir un DataFrame **pandas** que vous stockerez dans un variable nommée **df**. Prendre en compte dans votre import les trois points suivants :
 - la première ligne du fichier contient le nom de vos variables (l'entête)
 - la dernière ligne du fichier est un commentaire à ne pas importer dans Python
 - le séparateur de données utilisé est le ;
3. Vérifiez avec une ligne de code Python que votre DataFrame **df** contient bien 893 lignes (hors entête) et 12 colonnes.

4. Affichez les 10 premières lignes de votre DataFrame `df`. En déduire quelle est la classe du 10ième passager et s'il a survécu.
5. Renommez les 12 colonnes avec le nom de votre choix, puis vérifiez que le changement a bien été effectif en affichant les noms de colonnes de `df`.
6. Certaines lignes de ce jeu de données sont en double. Comptez combien de lignes sont dédoublées dans `df`, puis supprimez les doublons dans `df`. Vérifiez que vous n'avez plus de doublons.
7. Certaines données sont manquantes.
 - (a) Comptez combien il y a de valeurs manquantes dans chaque colonne de `df`. Vous devriez avoir des valeurs manquantes dans les colonnes correspondant à l'âge, le numéro de cabine et le port d'embarquement.
 - (b) La colonne avec le numéro de cabine devrait avoir 687 valeurs manquantes. Étant donné ce nombre important, nous souhaitons supprimer cette colonne. Supprimez donc cette colonne, et vérifiez que la suppression a bien été effective sur `df`.
 - (c) La colonne avec le port d'embarquement devrait contenir 2 valeurs manquantes. Nous souhaitons remplacer ces valeurs manquantes par le mode de cette colonne. La valeur de son mode s'obtient de la manière suivante (dans le cas où votre colonne s'appelle `portEmbarquement`) :

```
df.portEmbarquement.mode()[0] # le [0] permet d'extraire la valeur
```
 - Remplacez les deux valeurs manquantes de cette colonne par son mode.
 - Les deux valeurs manquantes concernaient les passagers avec pour identifiant passager 62 et 830. Extraire les lignes de données correspondant à ces deux passagers pour vérifier que leurs port d'embarquement est maintenant bien 'S'.
 - (d) Il devrait nous rester 177 valeurs manquantes dans la colonne avec l'âge. Supprimez les lignes concernées.
 - (e) Vérifiez que vous n'avez maintenant plus de valeurs manquantes, puis que `df` contient désormais 714 lignes et 11 colonnes.
8. Nous voulons maintenant décoder les valeurs de la colonne contenant le port d'embarquement, pour avoir le nom complet du port (soit le transcodage C = Cherbourg, Q = Queenstown, S = Southampton). Effectuez ce transcodage pour cette colonne, puis vérifiez que ce changement a bien été effectué en appliquant la méthode `unique()` sur cette colonne.

Étude de l'âge des passagers

9. Quel est l'âge moyen, ainsi que l'âge médian des passagers ? Calculer ensuite l'écart-type des âges.
10. Quel est l'âge du plus jeune passager ? Quel est l'âge du plus vieux passager ? Quelle est l'étendue des âges ?
11. Tracez l'histogramme de la variable âge, en densité de fréquence et en utilisant 20 classes. Commentez ce graphique.
12. Tracez la boîte à moustache des âges. Grâce à ce graphique, donnez un intervalle d'âges contenant environ 50% des âges des passagers ?
13. Extraire de `df` toutes les informations sur le passager le plus âgé. A-t-il survécu ?

Étude de la survie des passagers

14. Quel est le taux de survie des passagers du Titanic ?
Astuce : cela correspond à la moyenne de la colonne indiquant la survie (cela équivaut au nombre de 1 divisé par le nombre total de passagers).
15. Calculez le taux de survie selon le sexe des passagers. Commentez les résultats.
16. Calculez le taux de survie selon la classe des passagers. Commentez les résultats.
17. Afin de vérifier si les jeunes passagers ont eu plus de chance de survie que les passagers plus vieux, créez deux classes d'âges : $[0, 20]$ pour les *jeunes*, et $]20, 80]$ pour les *vieux*. Sur ces classes d'âges, calculez le taux de survie des passagers correspondants. Commentez les résultats.
18. Vous aurez remarqué que les chances de survie étaient nettement différentes selon le sexe du passager. Afin de confirmer cette dépendance, effectuez un test d'indépendance du χ^2 entre les variables survie et sexe, puis commentez les résultats.
19. De la même manière faire un test d'indépendance entre la variable survie et les deux groupes d'âges précédemment créés ($[0, 20]$, $]20, 80]$). Commentez les résultats.
20. Pour finir, calculez le taux de survie selon les modalités jointes des variables sexe et classe. En déduire le taux de survie des femmes de première classe, ainsi que le taux de survie des hommes de troisième classe.

Étude du prix du ticket

21. Tracez l'histogramme des prix de ticket en densité de fréquence avec 10 classes. Commentez ce graphique.
22. Calculez le prix moyen du ticket, puis le prix médian. D'où provient cette forte différence entre la moyenne et la médiane ?
23. Tracez le nuage de points avec l'âge en abscisse et le prix du ticket en ordonnée.
24. Calculez la covariance entre l'âge et le prix du ticket. Interprétez cette valeur de covariance.
25. Nous voulons maintenant ajuster une droite de régression linéaire suivant la relation :

$$\text{prix} = a \times \text{âge} + b + \epsilon$$

Ajustez un tel modèle linéaire.

26. Prédire le prix du ticket pour deux passagers, un âgé de 0 an et un autre de 100 ans.
27. Tracez sur un même graphique, à la fois le nuage de points entre l'âge et le prix du ticket, ainsi que la droite linéaire que vous venez d'ajuster.
28. Pensez-vous que le modèle linéaire décrit bien la relation entre l'âge et le prix du ticket ? Calculez la corrélation entre l'âge et le prix du ticket pour appuyer vos propos.
29. Pour finir, nous voulons étudier si la classe du passager a une influence sur le prix du ticket.
 - (a) Grâce à des boîtes à moustaches, représentez les distributions du prix selon chaque classe de passagers. Comparez ces boîtes à moustaches avec celle du prix toutes classes confondues. Le prix semble-t-il impacté par la classe des passagers ?
 - (b) Pour finir calculez le coefficient de détermination R^2 afin de quantifier l'impact qu'à la classe sur le prix du ticket. Commentez cette valeur de R^2 que vous venez de trouver.