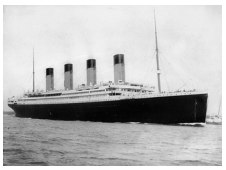
4. Traitement de données à l'aide d'un programme Python



Le RMS Titanic était un paquebot transatlantique britannique qui fit naufrage dans l'océan Atlantique Nord en 1912 à la suite d'une collision avec un iceberg. Du fait du manque de canots de sauvetage à bord, le nombre de victimes fût très important (1500 environ). Cette catastrophe a marqué les esprits du monde entier.
Nous allons, dans cette activité, exploiter avec Python (et son module « pandas ») le fichier de données structurées au format CSV contenant les informations sur les passagers du Titanic.

1. Exploitation du fichier contenant les données structurées des passagers

1.1. analyse de la table « Passager »

1. Copier les fichiers « **titanic.csv** »,« **01d_titanic.py** » et « **01d_titanic_passager.py** » du dossier commun SNT dans votre dossier SNT de votre dossier personnel.

2. Lancer Edupython et charger le fichier « **01d_titanic.py** » de votre dossier personnel .
Ce programme charge les données du fichier CSV et les stocke dans une sorte de variable appelée « data ».

3. A la suite du code de « **01d_titanic.py** », taper le code python permettant d'afficher le contenu de la variable « **data** ».

4. Après avoir exécuté le programme, se déplacer dans la console pour repérer les 6 descripteurs. Les écrire
5. A quoi correspond un objet de cette table ?

6. Pour le descripteur « sexe », à quoi correspond la valeur « 1 » ? Et la valeur « 2 » ?

7. Pour le descripteur «tarif», à quoi correspond la valeur indiquée pour chaque objet ?

8. Pour le descripteur « survie », à quoi correspondent la valeur « 0 » et la valeur « 1 » ?

Comment appelle-t-on le type de cette donnée ?

1.2. Rédaction d'un compte rendu portant sur un passager

9. `info=data.loc[X]` permet de mettre dans la variable « info » les données de l'objet (donc du passager) n°X.

Ecrire le code permettant de mettre dans une variable « info » les données du passager n°128, puis afficher cette variable. **Sauver le fichier sous le nom « 01d_titanic2.py »**

Ecrire ci-dessous un bref texte décrivant ce passager et ce qui lui est arrivé :

1.3. Automatisation de la rédaction du compte rendu portant sur un passager

10. Charger le programme Python « **01d_titanic_passager.py** »
Ce programme créé des variables correspondant aux données d'un passager (dont le n° est dans la variable « Passager »)

Compléter la zone "création des variables contenant les données du passager" pour récupérer toutes les données d'un passager.
Utiliser l'instruction « print » et les variables « nom », « passager », « sexe » etc… pour :

11. Afficher une phrase du type : « Le passager n°***0*** s'appelait ***Allen, Miss. Elisabeth Walton*** »

12. Idem avec : « ***Allen, Miss. Elisabeth Walton*** avait ***29.0 ans*** »

13. Idem avec : « ***Allen, Miss. Elisabeth Walton*** voyageait en classe ***1*** »

14. Idem avec : « ***Allen, Miss. Elisabeth Walton*** a payé son billet ***211.0*** £. »

15. Utiliser une condition pour afficher : «***Allen, Miss. Elisabeth Walton*** a survécu au naufrage» (ou «***n'a pas survécu***»)

Appeler le professeur pour lui montrer votre programme

16. Que doit-on modifier pour afficher le compte rendu portant sur le passager n°884 ? Le faire.

17. ***Pour les plus rapides:*** Modifier le code pour que l'**utilisateur** puisse choisir le n° du passager (utiliser un « input »)

2. Exploitation basique des données

Charger à nouveau le fichier « **01d_titanic.py** »

18. `stat = data.describe()` met dans une variable « stat » les statistiques calculées par python pour les données numériques uniquement.

Pour chaque descripteur (dont la donnée est de type numérique), on obtient :

Count	le nombre de données du descripteur
Mean	la moyenne des données
Std	l'écart-type des données (inutile en classe de seconde)
Min	la valeur mini des données
25%	25% des objets ont une valeur inférieure à celle-ci
50% et 75 %	50% et 75% des objets ont une valeur inférieure à celle-ci
Max	la valeur maxi des données

Ecrire le code permettant de créer la variable "stat" puis d'afficher les statistiques. **Sauver sous le nom « 01d_titanic_stat.py »**

19. En étudiant les statistiques calculées par Python, répondre aux questions suivantes :
a. Quel était l'âge moyen des passagers ?
b. Quelle est la moyenne de « survie » ? Exprimer ce chiffre en % (c'est-à-dire multiplier la valeur trouvée par 100) et justifier que ce naufrage eut un fort écho médiatique.

c. Quel était le tarif moyen payé ?
d. Quel était le tarif le plus élevé ?
e. 25% des passagers ont payé moins d'une certaine somme. Laquelle ?
f. La moitié des passagers ont payé moins d'un certaine somme. Laquelle ?

20. Pour tracer un histogramme des données d'un descripteur numérique (« tarif » par exemple), il faut taper :

```
data.hist(column='tarif', figsize= (9,6), bins = 20)
plt.show()
```

Pour retourner au « code », il faut fermer la fenêtre montrant l'histogramme

Taper ce code à la suite de « **01d_Titanic_Stat.py** »

Que pensez-vous de l'homogénéité des tarifs payés ? Quelle information cela nous apporte-t-il au niveau des inégalités sociales à l'époque ?

21. ***POUR LES PLUS RAPIDES*** **:** Adapter le code Python précédent pour afficher l'histogramme du descripteur âge.

Analyser cet histogramme. Quelle population composait principalement les passagers du Titanic ?

3. Exploitation plus fine des données

Charger à nouveau le fichier « **01d_titanic.py** » .
Sauver sous le nom **« 01d_titanic_groupes.py »**

22. Il n'y avait pas suffisamment de places dans les canots de sauvetage du Titanic pour tous les passagers et les membres de l'équipage (et certains canots sont partis à peine remplis). On souhaite examiner l'influence de la classe sociale des passagers sur l'obtention d'une place sur un canot de sauvetage. On va donc extraire les données **groupées** par « classe » et afficher la moyenne des données numériques correspondant aux différents descripteurs

Taper le code suivant :

```
groupe=data.groupby(['classe']).mean()
print(groupe)
```

En étudiant le % de survie de chaque classe, conclure quant à l'influence de la classe sociale sur les places dans les canots.

23. ***POUR LES PLUS RAPIDES*** Dans le film de James Cameron, on voit que les femmes embarquent davantage sur les canots que les hommes. On peut donc supposer que la fréquence de survie pour les femmes a été supérieure à celle des hommes... Est-ce la réalité ? Répondre en adoptant une démarche similaire à celle utilisée à la question précédente.

24. ***POUR LES PLUS RAPIDES²*** La fréquence de survie chez les femmes a-t-elle été indépendante de la classe dans laquelle voyageaient les passagères ? Pour répondre à cette question, extraire les données comme précédemment en indiquant deux descripteurs « classe », « sexe » pour pouvoir faire un deuxième tri (les mettre à la suite, séparés par une virgule : `'classe','sexe'`)