TP2 – Naufrage du Titanic (Python) Thème 4 : Données structurées

CORRECTION

Le but de ce TP est d'utiliser Python et deux de ses librairies (pandas et matplotlib) pour traiter les données des passagers du Titanic.

I. Préparatifs

1. Copier sur votre espace de travail les fichiers « analyse.py » et « titanic.csv » qui sont situés sur le réseau.

2. Ouvrir EduPython, en haut, sélectionner « Outils », puis encore « Outils » puis « Installation d'un nouveau module ». Saisissez « 1 » pour « Votre choix » puis tapez « pandas » pour nom du module.

II. Analyse des données avec Python

1. Fermer EduPython puis le rouvrir. Cliquer sur « Fichier », « Ouvrir » puis sélectionner le fichier « analyse.py ». Exécuter ce programme. A quoi sert la fonction « `mean` » ?

La fonction `mean` sert à calculer la valeur moyenne du descripteur (« survie » dans ce cas). Ici, cela revient à donner la proportion de personnes ayant survécu.

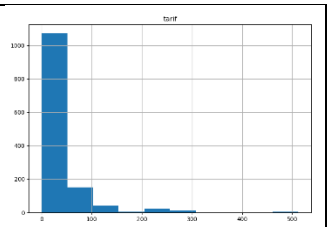
Il n'y avait pas suffisamment de places dans les canots de sauvetage du Titanic pour tous les passagers et les membres de l'équipage (et certains canots sont partis à peine remplis). On souhaite examiner l'influence de la classe sociale des passagers sur l'obtention d'une place sur un canot de sauvetage.

2. Ajouter les lignes de code suivantes au script « analyse.py » puis exécuter le programme :

```
infos_titanic.hist(column = 'tarif', figsize = (9,6), bins = 10)
plt.show()
```

Comment interpréter ce résultat ?

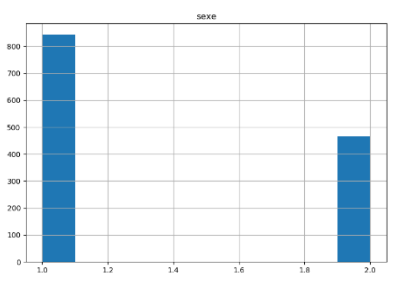
Il y a un grand déséquilibre des tarifs, signe qu'il y avait une forte inégalité sociale entre les passagers.



3. Modifier le script afin d'obtenir un histogramme de la répartition par sexe. Quel sexe était majoritaire à bord du Titanic ?

```
infos_titanic.hist(column = 'sexe', figsize = (9,6), bins =
10)
plt.show()
```

On remarque qu'il y avait presque deux fois plus d'hommes que de femmes à bord du Titanic.



4. Supprimer les deux dernières lignes ajoutées dans les questions précédentes et ajouter les lignes suivantes au script :

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

Quelle est la fréquence de survie pour chaque classe ? Que peut-on en conclure ?

classe	survie	sexe	age	tarif
1	0.619195	1.445820	39.165493	87.594427
2	0.429603	1.382671	29.517241	21.314079
3	0.255289	1.304654	24.842315	13.336158

Il y avait 62% de survivants en 1ère classe, 43% en 2nde classe et seulement 26% en 3ème classe. Les passagers les plus aisés avaient donc plus de chance de survivre.

5. Dans le film de James Cameron, lors de l'évacuation du Titanic, 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.

On saisit le code suivant :

```
groupe = infos_titanic.groupby(['sexe']).mean()
print(groupe)
```

sexe	classe	survie	age	tarif
1	2.372479	0.190985	30.607903	26.232779
2	2.154506	0.727468	28.693299	46.246781

Seulement 19 % des hommes ont survécu au naufrage alors que 73 % des femmes ont survécu (toutes classes confondues). L'illustration de James Cameron est donc fidèle à la réalité, les canots étaient essentiellement composés de femmes.

6. Supprimer les deux lignes ajoutées à la question précédentes et ajouter les lignes suivantes à la place :

```
groupe = infos_titanic.groupby(['classe', 'sexe']).mean()
print(groupe)
```

La fréquence de survie chez les femmes a-t-elle été indépendante de la classe dans laquelle voyageaient les passagères ?

Là encore, les inégalités sont très fortes selon la classe : alors que 49 % des femmes de 3ème classe vont survivre, 89 % des femmes de 2nde classe et97 % des femmes de 1ère classe vont survivre. Les canots de sauvetage étaient donc essentiellement composés de passagères de 1ère et 2ème classe.

classe	sexe	survie	age	tarif
1	1	0.340782	41.039735	70.011173
	2	0.965278	37.037594	109.451389
2	1	0.146199	30.829114	20.064327
	2	0.886792	27.504854	23.330189
3	1	0.152130	25.994269	12.449187
	2	0.490741	22.197368	15.356481

7. Pour ceux qui ont terminé toute l'activité… Rendez-vous sur le site data.education.gouv.fr et télécharger les données de réussite au baccalauréat au format CSV. Utiliser Python afin d'analyser ces données. En particulier, on étudiera la réussite en fonction de l'âge, en fonction du sexe et en fonction de l'académie.