Analyse de données
Le Titanic

Comme vous le savez certainement, le RMS Titanic a coulé lors de son voyage inaugural dans la nuit du 14 au 15 avril 1912 (plus d'informations voir la page wikipedia de la catastrophe)

Ce qui est intéressant, c'est que l'on a récupéré le manifeste du navire qui contient de nombreuses informations. Certains s'amusent à prédire les personnes qui vont survivre à partir des informations de ce manifeste et y arrivent relativement bien. La prédiction de la survie des passagers est un exercice de Data Science intéressant mais il dépasse le niveau de ce cours. Je vous propose ici à partir des éléments que nous avons vus de tirer et visualiser des informations.

Chargement des données et exploration

Les données sont stockées au format csv dans un fichier 'data-titanic.csv' qui vous sera fourni par votre enseignant. Utilisez pandas vous lire ces données et affichez par exemples les 10 premières lignes.

Le fichier comporte 1309 lignes. La page wikipedia nous indique 953 passagers et 899 membres d'équipage ce qui semble indiqué que notre base est incomplète, ce qui est généralement souvent le cas.

Voici quelques informations sur les colonnes de notre manifeste du RMS Titanic :

- PassengerId : numéro unique de passage
- Survived : 1,0 si le passager a survécu 0 sinon
- Pclass : classe du passager
- Name : nom du passager
- Sex : sexe du passager
- Age : âge du passager
- SibSp : nombre de parents/épouses présents sur le HMS Titanic
- Parch : nombre de parents / enfants
- Ticket : numéro du ticket

- Fare : montant du billet
- Cabin : catégorie de la cabine
- Embarked : port d'embarquement (C= Cherbourg, Q=Queentown, S = Southampton)

Quels sont les données entier, réel, objets ?

Donnez une description des données concernant le Titanic

Calculez le nombre de personnes qui ont survécu, calculez le pourcentage

J'ai trouvé sur 891 données renseignées, 549 décès et 342 survivants.
décès 61,61 %
survivants ; 38,38 %

La méthode groupby

Il est temps d'introduire une nouvelle méthode qui nous permet de travailler sur les DataFrame de manière agréable. Il s'agit de la méthode groupy() qui permet de grouper les données suivant certains critères.

Généralement, on part du principe que groupby() est basé sur trois étapes : séparation, application et combinaison.

La **séparation** est la partie la plus simple, on sépare notre DataFrame en fonction d'un critère (généralement une ou plusieurs colonnes). Ensuite, on **applique** des fonctions sir les groupes obtenus à l'étape précédente. Finalement, on **combine** les résultats obtenus pour chaque groupe.

La feuille de (Python pour le Data Scientist – E. Jakobowicz), vous donne quelques-une des fonctions qu'il est possible d'appeler après avoir réalisé la séparation.

Méthode/Propriété	Utilisation
`.agg()`	Application de plusieurs fonctions et obtention d'un DataFrame
`.apply()`	Application de n'importe quelle transformation sur les données
`.count()`	Comptage du nombre de lignes par groupe
`.cummax() / .cummin() / .cumprod()/. cumsum()`	Réalisation des calculs cumulés
`.describe()`	Equivalent de la méthode .describe() des DataFrame
`.first()`	Affichage des premiers éléments de chaque groupe
`.groups`	Affichage des groupes sous forme de dictionnaire
`.max() / .min() / .median() / .mean() / .quantile() / .std() / .var()`	Méthodes statistiques de base pour faire vos calculs. On peut aussi les intégrer sous forme de liste de chaînes de caractères dans la méthode .agg().
`.ngroups`	Affichage du nombre de groupes
`.nth()`	Affichage du n^{ième} élément de chaque groupe (si un groupe a moins de n éléments alors le groupe n'est pas affiché)
`.sum()`	Calcul de la somme des valeurs du groupe
`.tail()`	Affichage des dernières lignes de chaque groupe

par exemple pour grouper les éléments de la DataFrame suivant la classe des cabines, on pourra écrire :

```
Entrée [25]: test = données.groupby('Pclass')
             test.sample(3)
```

Out[25]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1218	1219	NaN	1	Rosenshine, Mr. George (Mr George Thorne")"	male	46.0	0	0
669	670	1.0	1	Taylor, Mrs. Elmer Zebley (Juliet Cummins Wright)	female	NaN	1	0
1003	1004	NaN	1	Evans, Miss. Edith Corse	female	36.0	0	0
1227	1228	NaN	2	de Brito, Mr. Jose Joaquim	male	32.0	0	0
772	773	0.0	2	Mack, Mrs. (Mary)	female	57.0	0	0
791	792	0.0	2	Gaskell, Mr. Alfred	male	16.0	0	0
888	889	0.0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2
601	602	0.0	3	Slabenoff, Mr. Petco	male	NaN	0	0
300	301	1.0	3	Kelly, Miss. Anna Katherine "Annie Kate"	female	NaN	0	0

Calculer pour chaque classe de cabines, le pourcentage de passager survivants

Voici les valeurs que j'ai trouvées :
Nombre de passagers survivants :

- 1ère classe : 136 ~ 62,96%
- 2ème classe : 87 ~ 47%
- 3ème classe : 119 ~ 24%

Calculez le pourcentage d'hommes et de femmes survivantes

J'ai trouvé qu'il y avait 314 femmes et que le le pourcentage de survie était des74,2 %

Calculez le nombre est le pourcentage de survivants suivant le lieu d'embarquement

Rappelez-vous que le C signifie Cherbourg ; Q Queentown et S Southhampton

J'ai trouvé que 55 % des personnes embarquées à Cherbourg avaient survécu, 38 % à Queenstown et 33 % à Southampton.

Allez plus loin dans l'analyse de ces valeurs et donnez pour cahque port d'embarquement le pourcentage d'hommes et de femmes qui ont survécu.

Quel est l'âge du capitaine ?

Quel est l'âge moyen des passagers ?

Et l'âge moyen des personnes ayant survécu, décédé ?

29,8 ans

28,34 ans pour les survivants

30,62 ans pour ceux n'ayant pas survécu

Quel est l'âge du plus jeune et du passager le plus âgé ? Quel est son nom ?

0,17 ans Dean, Miss. Elizabeth Gladys Millvina

et 80 ans

Données incohérentes et manquantes

La méthode isnull appliqué sur un DataFrame permet de retourner un tableau de booléens indiquant l'absence de données (True) ou la présence de données (False).

En écrivant un petit script python, retournez le nombre de valeurs manquantes de chaque colonne. Voici les valeurs que j'ai trouvées :

```
[0, 418, 0, 0, 0, 263, 0, 0, 0, 1, 1014, 2]
```

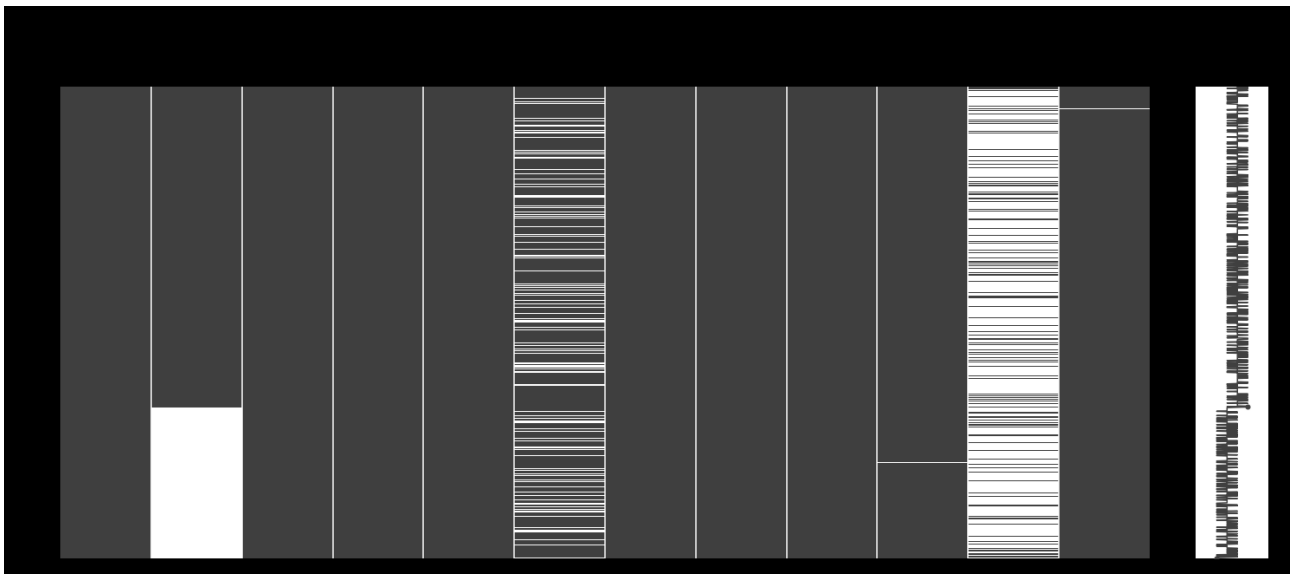
Il y a beaucoup de valeurs manquantes pour l'avant-dernière colonne (le numéro de la cabine, mais a priori ce n'est pas très important)

Il existe un package missingno qui permet de visualiser les valeurs manquantes dans une Series et une DataFrame. Essayez de l'importez avec votre programme d'installation préféré :

pip install missingno

import missingno as msno

Avec la commande matrix de ce module, visualisez les sonnées manquantes. Est-ce cohérent avec le calcul effectué précédemment ? Qu'en déduisez-vous ?



De même, utilisez la commande heatmap de ce module sur nos données. Qu'est-ce que cela représente à votre avis ?

Correction de la base de données

Il y a certains éléments de notre base que nous ne savons pas corriger et qu'il est inutile de corriger (typiquement la colonne Survived) mais nous pouvons par exemple modifier certains données pour avoir un DataFrame plus « cohérent ».

- Déjà nous allons commencer par enlever les colonnes qui nous gênent et n'apportent pas d'informations mais bien sûr, on peut se tromper. Avec la commande drop, enlevez les colonnes Ticket et Cabin
- Je vous propose de remplacer les âges manquants par l'âge moyen ou médian de chaque classe