Pandas_analyse_du_titanic

August 10, 2020

Document rédigé par BOUNGOTO BIBAYI Yoanne

0.1 Utilisation de Pandas (Ressemble à excel mais beaucoup plus puissant et pratique)

0.1.1 Objectif: Analyse des données du titanic

Analyse du dataset des passagers du titanic. Dataset contenant les informations concernant les passagers à bord du navire (Le sexe, l'age, la classe, et l'information s'ils ont survécu ou non etc).

0.1.2 Importation des librairies numpy, pandas pour la manipluation des données et matplotlib pour la visualisation des graphes

[89]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

0.1.3 Importer le dataset (Format Excel)

[71]:
```
titanic= pd.read_excel('titanic3.xls')
```

0.1.4 Verification de l'importation du dataset en ayant un aperçu des 3 premières lignes du tableau

[74]:
```
titanic.head(3)
```

[74]:
```
   pclass  survived                             name     sex      age  sibsp  \
0       1         1    Allen, Miss. Elisabeth Walton  female  29.0000      0
1       1         1   Allison, Master. Hudson Trevor    male   0.9167      1
2       1         0     Allison, Miss. Helen Loraine  female   2.0000      1

   parch  ticket      fare    cabin embarked boat  body  \
0      0   24160  211.3375       B5        S    2   NaN
1      2  113781  151.5500  C22 C26        S   11   NaN
2      2  113781  151.5500  C22 C26        S  NaN   NaN

                        home.dest
```

```
0                     St Louis, MO
1  Montreal, PQ / Chesterville, ON
2  Montreal, PQ / Chesterville, ON
```

0.1.5 Dimensions de notre dataset

[75]:
```
titanic.shape
print(titanic.shape)
print("Notre fichier excel importé contient 1309 lignes pour 1309 passagers et␣
 ↪14 colonnes")
```

```
(1309, 14)
Notre fichier excel importé contient 1309 lignes pour 1309 passagers et 14
colonnes
```

0.1.6 Suppression des colonnes dont les données ne nous seront pas utiles pour notre analyse

[76]:
```
titanic=titanic.drop(['name', 'sibsp', 'parch', 'ticket', 'fare', 'cabin',␣
 ↪'embarked', 'boat', 'body', 'home.dest'], axis=1)
```

[77]:
```
#Au finale on a garder la classe, le sexe et l'âge des passagers

titanic.head(4)
```

[77]:

	pclass	survived	sex	age
0	1	1	female	29.0000
1	1	1	male	0.9167
2	1	0	female	2.0000
3	1	0	male	30.0000

0.1.7 Statistiques de base pour chacunes de nos colonnes

[23]:
```
titanic.describe()
```

[23]:

	pclass	survived	age
count	1309.000000	1309.000000	1046.000000
mean	2.294882	0.381971	29.881135
std	0.837836	0.486055	14.413500
min	1.000000	0.000000	0.166700
25%	2.000000	0.000000	21.000000
50%	3.000000	0.000000	28.000000
75%	3.000000	1.000000	39.000000
max	3.000000	1.000000	80.000000

Observations: Mean,survived: Seulement 38 pourcent des passagers ont survécu. Mean,age: La moyenne d'age à bord du titanic était de 29 ans. Min,age et Max,age: la personne la plus jeune avait moins 1 ans age<0.16 et la plus agée 80 ans. Count: la colonne pclass 1309 lignes, survived

1309 et age 1046, ceci nous indique qu'il nous manque des données concernant l'age de nos passager, peut-être que ces données n'avaient pas été enregistrées, peut-être qu'ils ont été perdu, peu importe.Deux options s'offrent à nous; - soit on remplace toutes les valeurs manquantes par une valeur par defaut, exemple: l'ages moyen 29 ans (faut garder à l'esprit que sa va corrompre notre dataset, peut-être que ces passagers là n'avaient pas 30 ans, peut-être qu'ils étaient en dessous de 10 ans d'age, on ne le sait pas. -soit supprimer toutes les lignes des valeurs manquantes. C'est dommage on perd des données, mais parfois veut mieux un peut des données plutôt que de corrompre la réalité des chose.

0.1.8 Elimination des lignes des données manquantes (valeurs manquantes)

[78]:
```
titanic=titanic.dropna(axis=0)
```

[79]:
```
#Vérification des dimensions de notre dataset avec les données supprimées

print("On peut voir que les dimensions sont maintenant de 1046 lignes et 4␣
 ↪colonnes")
print(titanic.shape)
```

```
On peut voir que les dimensions sont maintenant de 1046 lignes et 4 colonnes
(1046, 4)
```

Statistiques de notre nouveau dataset (colonne pclass, survived et age ont tous 1046 valeurs)

[80]:
```
titanic.describe()
```

[80]:

	pclass	survived	age
count	1046.000000	1046.000000	1046.000000
mean	2.207457	0.408222	29.881135
std	0.841497	0.491740	14.413500
min	1.000000	0.000000	0.166700
25%	1.000000	0.000000	21.000000
50%	2.000000	0.000000	28.000000
75%	3.000000	1.000000	39.000000
max	3.000000	1.000000	80.000000

Observation: Les colonnes pclass, survived, age ont tous 1046 lignes, on n'est passé de 1309 à 1046. Ceci à pour effet de changer les statistiques de notre dataset. Vous pouvez constater par exemple que la moyen des survivants est passée à 40 pourcent. C'est pas bien grave tant que nous gardons à l'esprit que c'est une moyenne sur les 1046 et pas 1309 passagers.

0.1.9 Affichage du nombre de survivant par classe

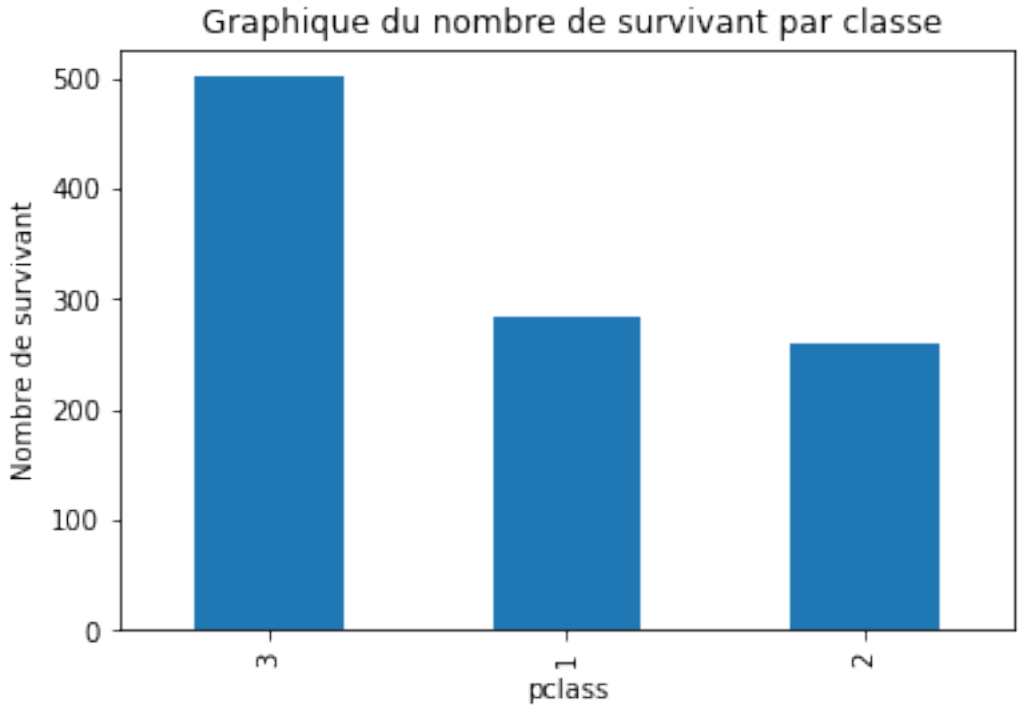
[81]:
```
print("3ème classe: 501 survivants, 1ère classe: 284 et 2ème classe: 261")
print(titanic['pclass'].value_counts())
```

```
3ème classe: 501 survivants, 1ère classe: 284 et 2ème classe: 261
3    501
```

```
1    284
2    261
Name: pclass, dtype: int64
```

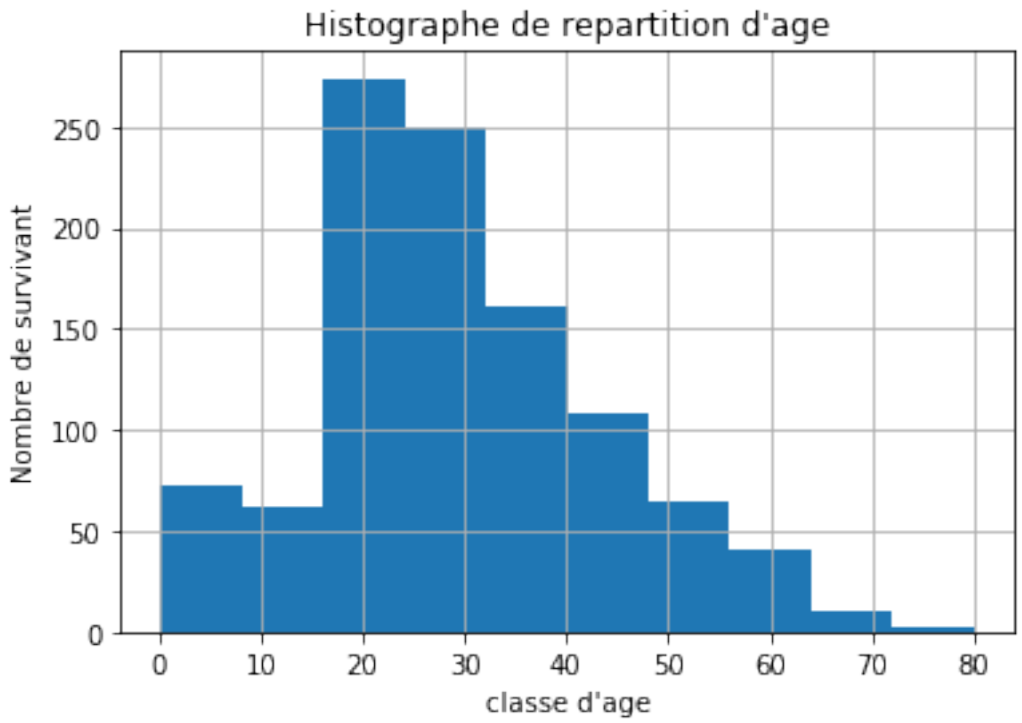
Affichage du nombre de survivant par classe avec un graphique

[82]:
```
titanic['pclass'].value_counts().plot.bar()
plt.title(' Graphique du nombre de survivant par classe')
plt.xlabel('pclass')
plt.ylabel('Nombre de survivant')
plt.show()
```



0.1.10 Répartition de l'age suivant un histographe

[53]:
```
titanic['age'].hist()
plt.xlabel("classe d'age")
plt.ylabel("Nombre de survivant")
plt.title("Histographe de repartition d'age")
plt.show()
```



Observations: On peut voir que le nombre de passagers entre 20-30 ans étaient plus nombreux 250. Entre 0-15 ans un peu plus de 50 Et très peu à plus de 65 ans

0.1.11 On peut faire une analyse en regroupant les gens suivant leur sexe

[83]:
```
titanic.groupby(['sex']).mean()
```

[83]:
```
          pclass  survived        age
sex
female  2.048969  0.752577  28.687071
male    2.300912  0.205167  30.585233
```

Observations: 75 pourcent des femmes d'une moyenne d'age de 28 ans ont survécu et seul 20 pourcent des hommes d'une moyenne de 30 d'age de ans ont survécu

[86]:
```
# On peut aller plus loin en regroupant par sexe et par classe
titanic.groupby(['sex','pclass']).mean()
```

[86]:
```
                survived        age
sex    pclass
female 1        0.962406  37.037594
       2        0.893204  27.499191
       3        0.473684  22.185307
male   1        0.350993  41.029250
```

```
2       0.145570  30.815401
3       0.169054  25.962273
```

Observations: 96 pourcent des femmes de 1ère classe ont survécu et un peu plus de 15 pourcent des hommes de 2ème et 3ème classe ont echappé à la mort. Le mot d'odre qui était "les femmes et les enfants d'abord" est verifié, mais bon on peut aussi voir que c'était les femmes et hommes d'une classe plus aisée. "Je faisais de l'humour"

[87]:
```
#Regroupement de tout les passagers mineurs selon le sex et leur classe
titanic[titanic['age'] < 18].groupby(['sex', 'pclass']).mean()
```

[87]:
```
                survived        age
sex    pclass
female 1       0.875000  14.125000
       2       1.000000   8.273150
       3       0.543478   8.416667
male   1       0.857143   9.845243
       2       0.733333   6.222220
       3       0.233333   9.838888
```

A ce niveau vous pouviez maintenant de vous même interpreter les resultats. J'ai voulu juste montrer qu'avec pandas on peut aligner des petits bouts de code mais mis ensemble on peut obtenir quelque chose de beaucoup plus complexe.

[88]:
```
#Affichage des mineurs par classe
titanic[titanic['age'] < 18]['pclass'].value_counts()
```

[88]:
```
3    106
2     33
1     15
Name: pclass, dtype: int64
```

1 Conclusions:

Voila, quelques que fonctions de la librairie pandas, il est possible d'aller encore plus loin. Ce travail est effectué dans Jupyter nootebook et le rendu des importations des tableaux sont beaucoup plus agréable à voir que la version convertie en pdf? qui ne laisse apparaître que les valeurs. de connaitre un peu plus sur les passagers du titanic, on pourrait aller encore plus loin.