pandas_groupby

June 29, 2022

1 Pandas et groupby

Petit tour de passe passe autour d'un groupby et des valeurs manquantes qui ne sont plus prises en compte depuis les dernières versions.

[1]:
```
from jyquickhelper import add_notebook_menu
add_notebook_menu()
```

[1]:
```
<IPython.core.display.HTML object>
```

1.1 groupby et valeur manquantes

[2]:
```
import pandas
data = [{"a":1, "b":2}, {"a":10, "b":20}, {"b":3}, {"b":4}]
df = pandas.DataFrame(data)
df
```

[2]:

	a	b
0	1.0	2
1	10.0	20
2	NaN	3
3	NaN	4

[3]:
```
df.groupby("a").sum()
```

[3]:

a	b
1.0	2
10.0	20

Les valeurs manquantes ont disparu et c'est le comportement attendu d'après groupby and missing values. Il est possible de ocrriger le tir avec la fonction implémenté dans ce module.

[4]:
```
from pandas_streaming.df import pandas_groupby_nan
pandas_groupby_nan(df, "a").sum()
```

[4]:

	a	b
0	1.0	2
1	10.0	20
2	NaN	7

L'astuce consiste à remplacer les valeurs manquantes par d'autres non utilisées dans le dataframe, à grouper, puis à leur redonner leur valeurs initiales. Le code de la fonction n'est pas très propre car il modifie des variables que l'utilisateur n'est pas censé modifier. Il est possible que la fonction "casse" pour des versions ultérieures. Le code utilise quelques variables non documentation du module pandas.

[5]: