Untitled

January 23, 2022

[1]:
```
import pandas as pd
file = "./data/tested.csv"
df = pd.read_csv(file)
```

[2]:
```
df.head()
```

[2]:
```
   PassengerId  Survived  Pclass  \
0          892         0       3
1          893         1       3
2          894         0       2
3          895         0       3
4          896         1       3

                                          Name     Sex   Age  SibSp  Parch  \
0                              Kelly, Mr. James    male  34.5      0      0
1              Wilkes, Mrs. James (Ellen Needs)  female  47.0      1      0
2                     Myles, Mr. Thomas Francis    male  62.0      0      0
3                              Wirz, Mr. Albert    male  27.0      0      0
4  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female  22.0      1      1

    Ticket     Fare Cabin Embarked
0   330911   7.8292   NaN        Q
1   363272   7.0000   NaN        S
2   240276   9.6875   NaN        Q
3   315154   8.6625   NaN        S
4  3101298  12.2875   NaN        S
```

see the filename five lines

[4]:
```
change = {"male":1,"female":0}
df.replace(change)
```

[4]:
```
     PassengerId  Survived  Pclass  \
0            892         0       3
1            893         1       3
2            894         0       2
3            895         0       3
4            896         1       3
```

```
..           …       …     …
413         1305         0       3
414         1306         1       1
415         1307         0       3
416         1308         0       3
417         1309         0       3

                                             Name  Sex   Age  SibSp  Parch  \
0                                 Kelly, Mr. James    1  34.5      0      0
1                 Wilkes, Mrs. James (Ellen Needs)    0  47.0      1      0
2                        Myles, Mr. Thomas Francis    1  62.0      0      0
3                                 Wirz, Mr. Albert    1  27.0      0      0
4     Hirvonen, Mrs. Alexander (Helga E Lindqvist)    0  22.0      1      1
..                                             …  …    …    …    …
413                           Spector, Mr. Woolf    1   NaN      0      0
414                  Oliva y Ocana, Dona. Fermina    0  39.0      0      0
415                  Saether, Mr. Simon Sivertsen    1  38.5      0      0
416                           Ware, Mr. Frederick    1   NaN      0      0
417                      Peter, Master. Michael J    1   NaN      1      1

                Ticket      Fare Cabin Embarked
0               330911    7.8292   NaN        Q
1               363272    7.0000   NaN        S
2               240276    9.6875   NaN        Q
3               315154    8.6625   NaN        S
4              3101298   12.2875   NaN        S
..                 …       …   …      …
413          A.5. 3236    8.0500   NaN        S
414           PC 17758  108.9000  C105        C
415  SOTON/O.Q. 3101262    7.2500   NaN        S
416              359309    8.0500   NaN        S
417                2668   22.3583   NaN        C

[418 rows x 12 columns]
```

Change the sex into the number

[9]:
```
df_select = pd.DataFrame()
df_select['Name'] = ['Jack jose','Rose kk']
df_select['Age'] = [35,23]
df_select['Sex'] = [1,0]
```

[10]:
```
df_select
```

[10]:
```
         Name  Age  Sex
0  Jack jose   35    1
1    Rose kk   23    0
```

We can create the data like above

[12]:
```
newone = pd.Series(['Molly Moonly',40,1],index = ['Name','Age','Sex'])
df_select.append(newone,ignore_index=True)
```

[12]:
```
           Name  Age  Sex
0     Jack jose   35    1
1       Rose kk   23    0
2  Molly Moonly   40    1
```

Using the append method we can append the data to our existed data

[13]:
```
df.describe()
```

[13]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   418.000000  418.000000  418.000000  332.000000  418.000000
mean   1100.500000    0.363636    2.265550   30.272590    0.447368
std     120.810458    0.481622    0.841838   14.181209    0.896760
min     892.000000    0.000000    1.000000    0.170000    0.000000
25%     996.250000    0.000000    1.000000   21.000000    0.000000
50%    1100.500000    0.000000    3.000000   27.000000    0.000000
75%    1204.750000    1.000000    3.000000   39.000000    1.000000
max    1309.000000    1.000000    3.000000   76.000000    8.000000

             Parch        Fare
count  418.000000  417.000000
mean     0.392344   35.627188
std      0.981429   55.907576
min      0.000000    0.000000
25%      0.000000    7.895800
50%      0.000000   14.454200
75%      0.000000   31.500000
max      9.000000  512.329200
```

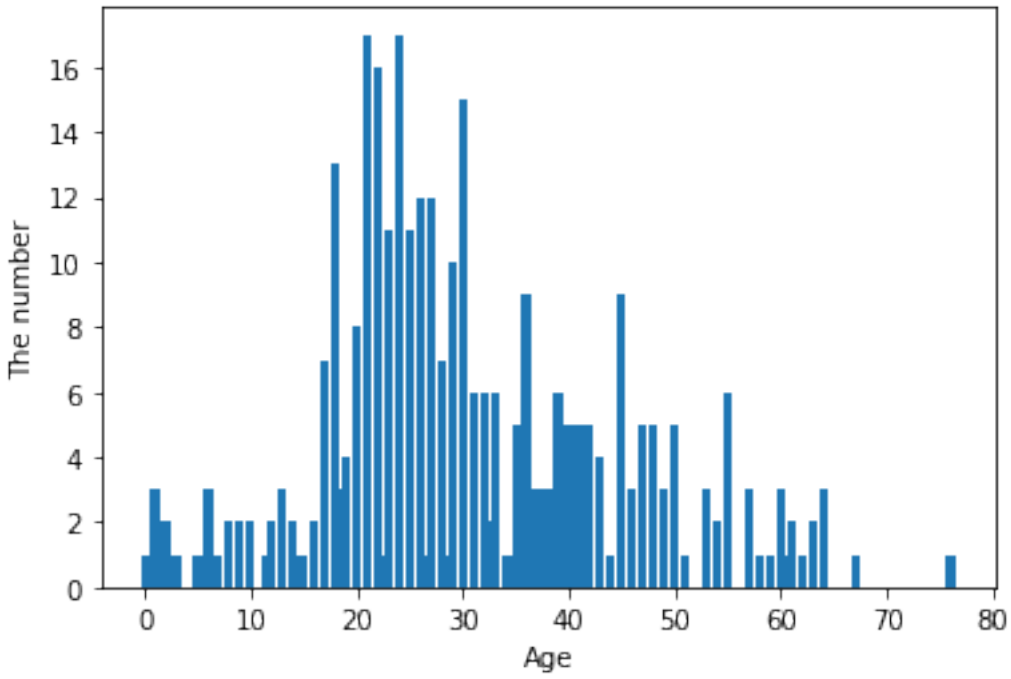
You can use the describe function to see the basic situation for our data

[41]:
```
data = df['Age'].value_counts()
data.to_numpy().shape
data.index
```

[41]:
```
Float64Index([21.0, 24.0, 22.0, 30.0, 18.0, 27.0, 26.0, 25.0, 23.0, 29.0, 45.0,
              36.0, 20.0, 17.0, 28.0, 31.0, 39.0, 33.0, 32.0, 55.0, 50.0, 48.0,
              41.0, 47.0, 42.0, 40.0, 35.0, 43.0, 19.0, 60.0, 37.0, 46.0, 53.0,
              49.0,  6.0, 64.0, 38.0, 13.0, 57.0, 18.5,  1.0,  8.0, 12.0, 54.0,
              61.0, 10.0, 14.0, 16.0,  2.0,  9.0, 63.0, 32.5, 59.0, 0.17, 58.0,
               3.0, 44.0,  5.0, 0.83, 0.75, 14.5, 36.5, 51.0, 0.92, 34.5, 67.0,
              40.5, 0.33, 11.5, 34.0, 15.0,  7.0, 60.5, 26.5, 76.0, 28.5, 22.5,
              62.0, 38.5],
             dtype='float64')
```

[44]:
```
from matplotlib import pyplot as plt
plt.xlabel("Age")
plt.ylabel("The number")
plt.bar(data.index,data.to_numpy())
```

[44]:
```
<BarContainer object of 79 artists>
```



[15]:
```
df[df['Age'].isnull()].head(2)
```

[15]:
```
    PassengerId  Survived  Pclass                                     Name  \
10          902         0       3                      Ilieff, Mr. Ylio
22          914         1       1  Flegenheim, Mrs. Alfred (Antoinette)

       Sex  Age  SibSp  Parch    Ticket     Fare Cabin Embarked
10    male  NaN      0      0    349220   7.8958   NaN        S
22  female  NaN      0      0  PC 17598  31.6833   NaN        S
```

[21]:
```
del_num = df[df['Age'].isnull()]
del_num.index
```

[21]:
```
Int64Index([ 10,  22,  29,  33,  36,  39,  41,  47,  54,  58,  65,  76,  83,
             84,  85,  88,  91,  93, 102, 107, 108, 111, 116, 121, 124, 127,
            132, 133, 146, 148, 151, 160, 163, 168, 170, 173, 183, 188, 191,
            199, 200, 205, 211, 216, 219, 225, 227, 233, 243, 244, 249, 255,
            256, 265, 266, 267, 268, 271, 273, 274, 282, 286, 288, 289, 290,
```

```
            292, 297, 301, 304, 312, 332, 339, 342, 344, 357, 358, 365, 366,
            380, 382, 384, 408, 410, 413, 416, 417],
           dtype='int64')
```

df.drop(del_num.index,axis=0)

1 /

, axis = 0, , axis 1

[46]:
```
df.describe()
```

[46]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   418.000000  418.000000  418.000000  332.000000  418.000000
mean   1100.500000    0.363636    2.265550   30.272590    0.447368
std     120.810458    0.481622    0.841838   14.181209    0.896760
min     892.000000    0.000000    1.000000    0.170000    0.000000
25%     996.250000    0.000000    1.000000   21.000000    0.000000
50%    1100.500000    0.000000    3.000000   27.000000    0.000000
75%    1204.750000    1.000000    3.000000   39.000000    1.000000
max    1309.000000    1.000000    3.000000   76.000000    8.000000

             Parch        Fare
count  418.000000  417.000000
mean     0.392344   35.627188
std      0.981429   55.907576
min      0.000000    0.000000
25%      0.000000    7.895800
50%      0.000000   14.454200
75%      0.000000   31.500000
max      9.000000  512.329200
```

[49]:
```
df = df.drop(['Age'],axis=1)
df.describe()
```

[49]:
```
       PassengerId    Survived      Pclass       SibSp       Parch        Fare
count   418.000000  418.000000  418.000000  418.000000  418.000000  417.000000
mean   1100.500000    0.363636    2.265550    0.447368    0.392344   35.627188
std     120.810458    0.481622    0.841838    0.896760    0.981429   55.907576
min     892.000000    0.000000    1.000000    0.000000    0.000000    0.000000
25%     996.250000    0.000000    1.000000    0.000000    0.000000    7.895800
50%    1100.500000    0.000000    3.000000    0.000000    0.000000   14.454200
75%    1204.750000    1.000000    3.000000    1.000000    0.000000   31.500000
max    1309.000000    1.000000    3.000000    8.000000    9.000000  512.329200
```

[55]:
```
df.head(2)
```

[55]:
```
   PassengerId  Survived  Pclass                          Name   Sex  SibSp  \
0          892         0       3             Kelly, Mr. James  male      0
2          894         0       2  Myles, Mr. Thomas Francis  male      0

   Parch  Ticket    Fare Cabin Embarked
0      0  330911  7.8292   NaN        Q
2      0  240276  9.6875   NaN        Q
```

[59]:
```
df.drop([2],axis=0)
```

[59]:
```
     PassengerId  Survived  Pclass  \
0            892         0       3
3            895         0       3
4            896         1       3
5            897         0       3
6            898         1       3
..           …       …     …
413         1305         0       3
414         1306         1       1
415         1307         0       3
416         1308         0       3
417         1309         0       3

                                             Name     Sex  SibSp  Parch  \
0                                Kelly, Mr. James    male      0      0
3                                Wirz, Mr. Albert    male      0      0
4    Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female      1      1
5                      Svensson, Mr. Johan Cervin    male      0      0
6                            Connolly, Miss. Kate  female      0      0
..                                            …     …    …     …
413                          Spector, Mr. Woolf    male      0      0
414                  Oliva y Ocana, Dona. Fermina  female      0      0
415                  Saether, Mr. Simon Sivertsen    male      0      0
416                           Ware, Mr. Frederick    male      0      0
417                     Peter, Master. Michael J    male      1      1

                 Ticket      Fare Cabin Embarked
0                330911    7.8292   NaN        Q
3                315154    8.6625   NaN        S
4               3101298   12.2875   NaN        S
5                  7538    9.2250   NaN        S
6                330972    7.6292   NaN        Q
..                  …       …   …      …
413           A.5. 3236    8.0500   NaN        S
414            PC 17758  108.9000  C105        C
415  SOTON/O.Q. 3101262    7.2500   NaN        S
416              359309    8.0500   NaN        S
```

```
417                2668   22.3583   NaN        C

[416 rows x 11 columns]
```

Be remind that the index will not change after the delete

[60]:
```
Group = df.groupby('Sex').mean()
```

We can use the group function to divide the data

[62]:
```
Group
```

[62]:

Sex	PassengerId	Survived	Pclass	SibSp	Parch	Fare
female	1098.139073	1.0	2.139073	0.562914	0.602649	50.030796
male	1102.620301	0.0	2.334586	0.379699	0.274436	27.527877

[ ]:
```

```