Titanic

May 8, 2018

```
In [52]: import pandas as pd
         import matplotlib.pyplot as plt
         from decimal import *
```

```
In [3]: df = pd.read_csv('titanic-data.csv')
```

```
In [13]: # check how many passengers are there

         passenger_count = len(df.index)
         print(passenger_count)
```

```
891
```

```
In [5]: # check if some columns has empty data

        df.count()
```

```
Out[5]: PassengerId    891
        Survived       891
        Pclass         891
        Name           891
        Sex            891
        Age            714
        SibSp          891
        Parch          891
        Ticket         891
        Fare           891
        Cabin          204
        Embarked       889
        dtype: int64
```

it seems that Cabin is incomplete data, may not able to analysis on that column
found 2 records that has no emarkation value

```
In [6]: df.loc[~df['Embarked'].isin(['S','C','Q'])]
```

```
Out[6]:      PassengerId  Survived  Pclass                                      Name  \
        61            62         1       1                       Icard, Miss. Amelie
```

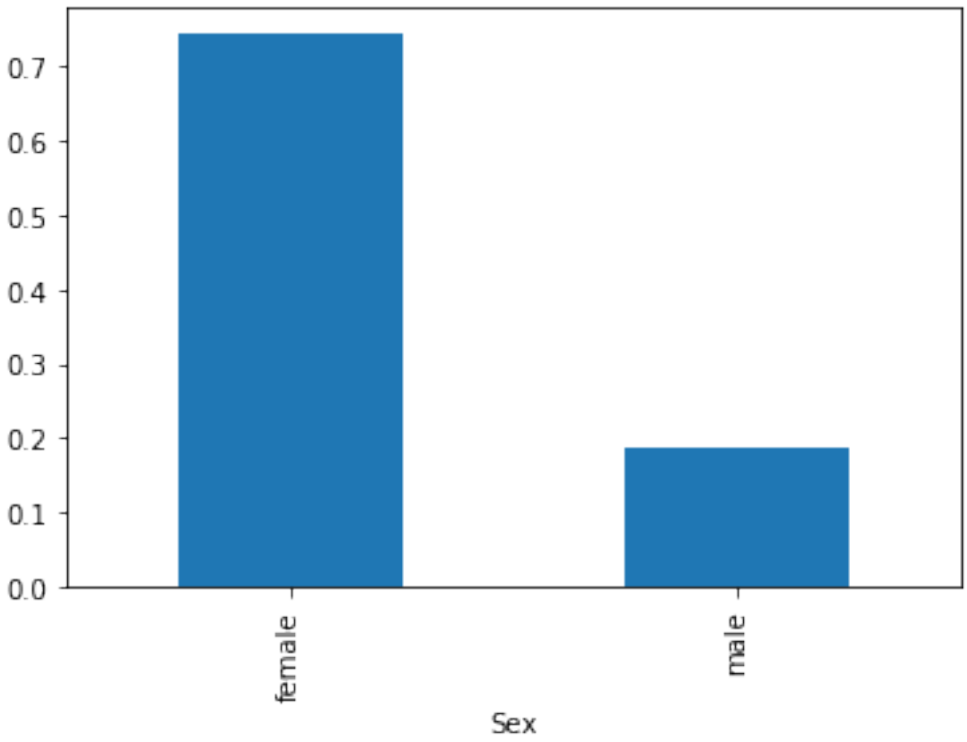
```
829          830         1       1  Stone, Mrs. George Nelson (Martha Evelyn)

        Sex   Age  SibSp  Parch  Ticket  Fare Cabin Embarked
61   female  38.0      0      0  113572  80.0   B28      NaN
829  female  62.0      0      0  113572  80.0   B28      NaN
```

is female or male more likely to survive?

```
In [40]: # https://stackoverflow.com/questions/10373660/converting-a-pandas-groupby-object-to-
         # https://stackoverflow.com/questions/38174155/group-dataframe-and-get-sum-and-count
         # https://stackoverflow.com/questions/18504967/pandas-dataframe-create-new-columns-an

         def agg_sex_by_pclass(data, pclass=None):
             '''
             aggregate passengers count and survival rate by passenger cabin class
             if class is not provided, take all passengers into calculation
             '''
             df = None
             if not pclass == None:
                 df = data.loc[data['Pclass'] == pclass]
             else:
                 df = data
             agg_sex = df.groupby('Sex').agg({'PassengerId': 'count', 'Survived': 'sum'})
             agg_sex['survival_rate'] = agg_sex['Survived'] / agg_sex['PassengerId']
             return agg_sex

In [39]: agg_sex = agg_sex_by_pclass(df)
         agg_sex

Out[39]:         PassengerId  Survived  survival_rate
         Sex
         female          314       233       0.742038
         male            577       109       0.188908
```

female had a much larger chance to survive(compared to total female/male count)? how come?

```
In [35]: agg_sex.survival_rate.plot(kind='bar')
         plt.show()
```



how is age related with survival rate?

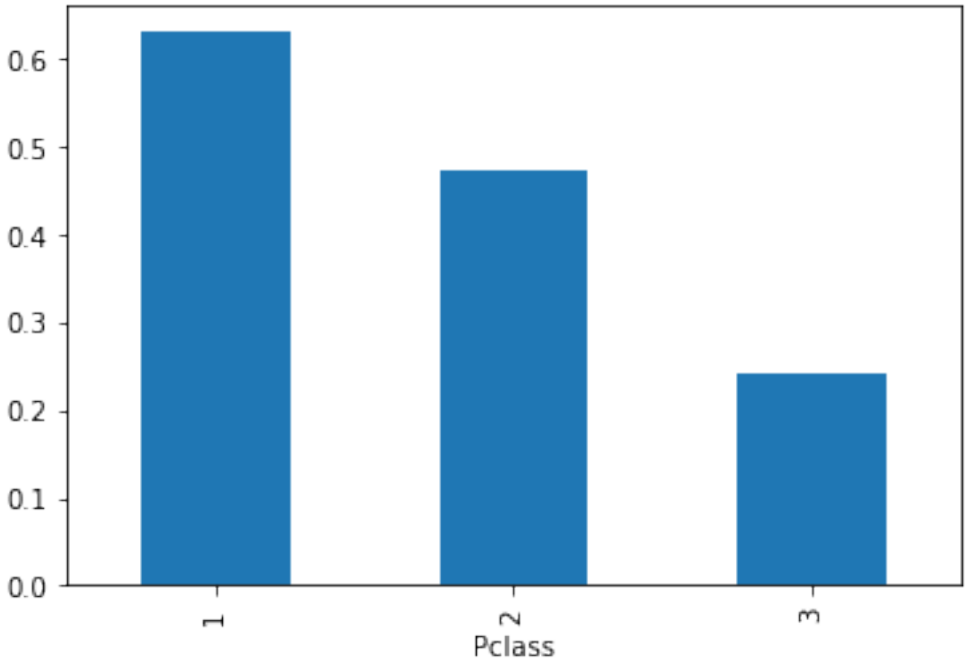
```
In [57]: df['Age'].describe()

Out[57]: count    714.000000
         mean      29.699118
         std       14.526497
         min        0.420000
         25%       20.125000
         50%       28.000000
         75%       38.000000
         max       80.000000
         Name: Age, dtype: float64

In [61]: agg_age_younger = df.loc[df['Age']<8].agg({'PassengerId': 'count', 'Survived': 'sum'})
         Decimal(agg_age_younger['Survived']) / agg_age_younger['PassengerId']

         # agg_age['Survived'].hist(by=agg_age['Age'])
         # plt.show()

Out[61]: Decimal('0.68')

In [57]: agg_age_elder = df.loc[df['Age']>50].agg({'PassengerId': 'count', 'Survived': 'sum'})
         Decimal(agg_age_elder['Survived']) / agg_age_elder['PassengerId']
```

```
Out[57]: Decimal('0.34375')
```

```
In [36]: agg_pclass = df.groupby('Pclass').agg({'PassengerId': 'count', 'Survived': 'sum'})
         agg_pclass['survival_rate'] = agg_pclass['Survived'] / agg_pclass['PassengerId']

         agg_pclass
```

```
Out[36]:         PassengerId  Survived  survival_rate
         Pclass
         1              216       136       0.629630
         2              184        87       0.472826
         3              491       119       0.242363
```

```
In [25]: agg_pclass.survival_rate.plot(kind='bar')
         plt.show()
```



the above bar chart shows that the 3rd class passengers had a less chance to survive compared to those in 1st class and 2nd class

```
In [41]: agg_1st_class_sex = agg_sex_by_pclass(df, pclass=1)
         agg_1st_class_sex
```

```
Out[41]:         PassengerId  Survived  survival_rate
         Sex
         female           94        91       0.968085
         male            122        45       0.368852
```

```
In [42]: agg_2nd_class_sex = agg_sex_by_pclass(df, pclass=2)
         agg_2nd_class_sex
```

Out[42]:

Sex	PassengerId	Survived	survival_rate
female	76	70	0.921053
male	108	17	0.157407

```
In [43]: agg_3rd_class_sex = agg_sex_by_pclass(df, pclass=3)
         agg_3rd_class_sex
```

Out[43]:

Sex	PassengerId	Survived	survival_rate
female	144	72	0.500000
male	347	47	0.135447

1. 72 2. 62 3. 4. (8)(50)