

Titanic数据集分析

□ 前期准备

- 数据下载
- 软件准备：python3.6+ anaconda 或 使用集成开发环境 pycharm

□ 导入数据&查看基本信息

```
wocanmei.py ×  
1 import numpy as np  
2 import pandas as pd  
3 import matplotlib.pyplot as plt  
4 data_src='D://titanic-data.csv'  
5 df = pd.read_csv(data_src,header=0) # 导入数据  
6 print(df.info()) # 查看数据集的基本信息,  
7 print(df.describe()) # 查看数据的摘要信息  
8 print(df.head()) # 查看前几行数据, 方便了解数据具体情况  
9
```



Titanic数据集分析

- 运行结果：从数据集的基本信息可以看出，Age \ Cabin \ Embarked 是存在缺失值的，其中Cabin字段缺失值过多。常用的方法是去除和补齐，数值型的数据是可以根据统计学的方法或者机器学习的方法将其进行补齐的

titanic						
	PassengerId	Survived	Pclass	Age	SibSp	\
count	891.000000	891.000000	891.000000	714.000000	891.000000	
mean	446.000000	0.383838	2.308642	29.699118	0.523008	
std	257.353842	0.486592	0.836071	14.526497	1.102743	
min	1.000000	0.000000	1.000000	0.420000	0.000000	
25%	223.500000	0.000000	2.000000	20.125000	0.000000	
50%	446.000000	0.000000	3.000000	28.000000	0.000000	
75%	668.500000	1.000000	3.000000	38.000000	1.000000	
max	891.000000	1.000000	3.000000	80.000000	8.000000	
	Parch	Fare				
count	891.000000	891.000000				
mean	0.381594	32.204208				
std	0.806057	49.693429				
min	0.000000	0.000000				
25%	0.000000	7.910400				
50%	0.000000	14.454200				
75%	0.000000	31.000000				
max	6.000000	512.329200				



Titanic数据集分析

□ 分析乘客存活率与各单变量之间的关系

- 查看总存活率: `survived_rate = float(df['Survived'].sum()) / df['Survived'].count()`

`Print('survived_rate: ',survived_rate)`

- 输出结果: `survived_rate: 0.383838383838`

```
7 x=[df[(df.Pclass==1)]['Pclass'].size,df[(df.Pclass==2)]['Pclass'].size,df[(df.Pclass==3)]['Pclass'].size]
8 y=[df[(df.Pclass==1) & (df.Survived == 1)]['Pclass'].size,df[(df.Pclass==2) & (df.Survived == 1)]['Pclass'].size,df[(df.Pclass == 3) & (df.Survived == 1)]['Pclass'].size]
9 print('1 Pclass number:' + str(x[0]) + '      ' + '1 Pclass survive:' + str(y[0]) + '      ' + '1 Pclass survive rat:', float(y[0]) / x[0])
10 print('2 Pclass number:' + str(x[1]) + '      ' + '2 Pclass survive:' + str(y[1]) + '      ' + '2 Pclass survive rat:', float(y[1]) / x[1])
11 print('3 Pclass number:' + str(x[2]) + '      ' + '3 Pclass survive:' + str(y[2]) + '      ' + '3 Pclass survive rat:', float(y[2]) / x[2])
12
13 Pclass_survived_rate = (df.groupby(['Pclass']).sum() / df.groupby(['Pclass']).count())['Survived']
14 Pclass_survived_rate.plot(kind='bar')
15 plt.title('Pclass_survived_rate')
16 plt.show()
```



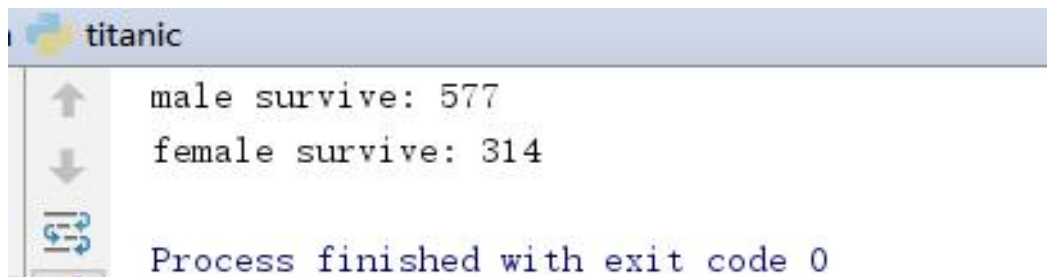
Titanic数据集分析

□ 分析乘客存活率与各单变量之间的关系

• 性别与存活率关系

```
7 male_survived=df[(df.Sex=='male')]['Sex'].size
8 female_survived=df[(df.Sex=='female')]['Sex'].size
9 print('male survive:',male_survived)
10 print('female survive:',female_survived)
11 Sex_survived_rate = (df.groupby(['Sex']).sum() / df.groupby(['Sex']).count())['Survived']
12 Sex_survived_rate.plot(kind='bar')
13 plt.title('Sex_survived_rate')
14 plt.show()
```

• 运行结果



```
titanic
↑ male survive: 577
↓ female survive: 314
↻ Process finished with exit code 0
```

