Kaggle竞赛的“泰坦尼克号预测生还”是进行Pandas数据分析非常好的案例

1、不同性别、舱位和年龄的分布情况

2、不同性别、舱位和登录港口的获救比例比较，找到生还比例更高的特征有哪些

一、数据清洗

①筛选列

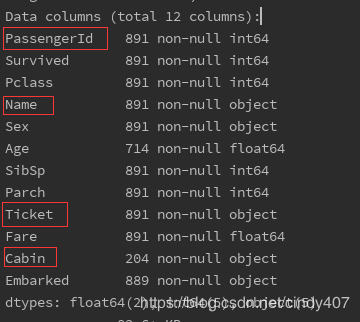
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

data = pd.read\_csv('train.csv')

print(data.info())



先看看各个特征代表的是什么意思：

PassengerId => 乘客ID，这个不会影响到存活，可删除

Pclass => 乘客等级(1/2/3等舱位)

Name => 乘客姓名，不会影响，可删除

Sex => 性别

Age => 年龄

SibSp => 堂兄弟/妹个数

Parch => 父母与小孩个数

Ticket => 船票信息 ，不会影响，可删除

Fare => 票价

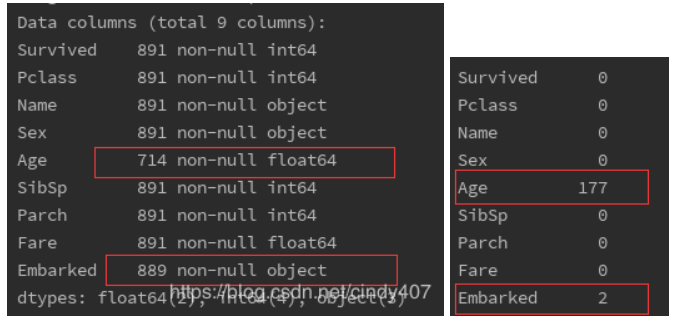
Cabin => 客舱 ，不会影响，可删除

Embarked => 登船港口

data.drop(['PassengerId', 'Ticket','Cabin'],axis=1,inplace=True)

print(data.info())

print(data.isnull().sum())

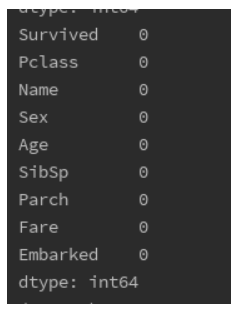


发现Age和Embarked有缺失值，Age缺失较多，达177个，年龄是分析相当重要的特征，不能删除，可以用填充法，这里选择填充均值；Embarked由于缺失值较少，直接删除或者填充都行， 这里选择删除。

data.Age = data.Age.fillna(data.Age.mean())

data.dropna(subset=['Embarked'],how='any',axis=0,inplace=True)

print(data.isnull().sum())



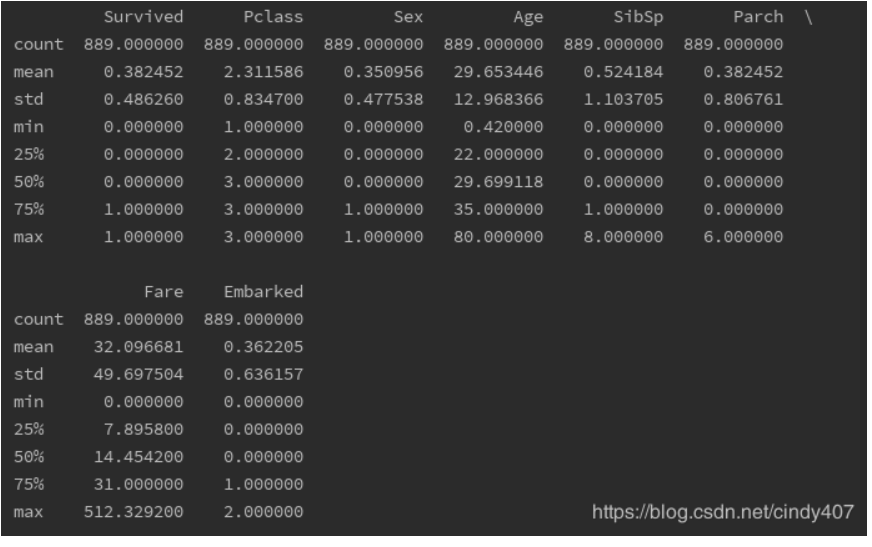
二、数据规整

①性别转换为数值，便于后续进行分析

print(data.Sex.unique())

data.loc[data['Sex']=='male','Sex']=0

data.loc[data['Sex']=='female','Sex']=1



②登录港口转换成数值，便于后续分析

print(data.Embarked.unique())

data.loc[data['Embarked']=='S','Embarked'] = 0

data.loc[data['Embarked']=='C','Embarked'] = 1

data.loc[data['Embarked']=='Q','Embarked'] = 2

pd.options.display.max\_rows=None

pd.options.display.max\_columns=None

print(data.describe())

三、数据分析

① 存活人数分布

# 存活情况

plt.rcParams['font.sans-serif'] = ['SimHei']

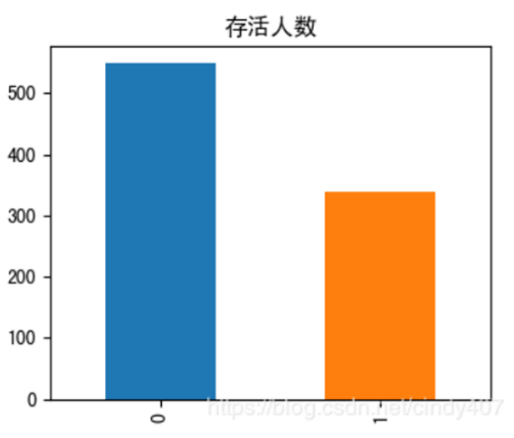
plt.rcParams['font.serif'] = ['SimHei']

fig,axes = plt.subplots(2,3,figsize=(20,30))

plt.subplot(2,3,1)

data['Survived'].value\_counts().plot(kind='Bar',alpha=0.8)

plt.title('存活人数')

1

大概有330人获救，比例1/3左右

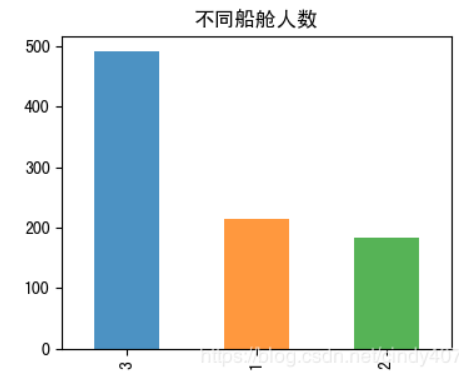
②不同船舱的人数分布

# 不同船舱的人数分布

plt.subplot(2,3,2)

data['Pclass'].value\_counts().plot(kind='Bar',alpha=0.8)

plt.title('不同船舱人数')



3等舱人数最多，这也符合人们的收入情况

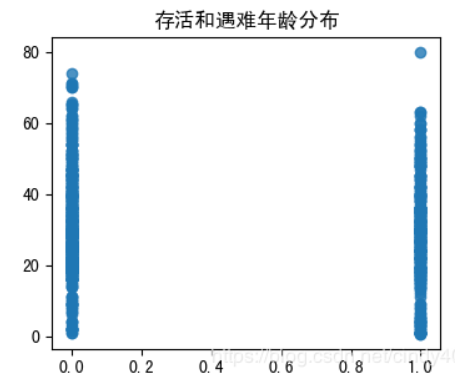
③存活和遇难的年龄分布

# 统计存活和遇难的年龄分布

plt.subplot(2,3,3)

plt.scatter(data['Survived'],data['Age'],alpha=0.8)

plt.title('存活和遇难年龄分布')



获救与遇难的年龄分布都比较广，没有太大区别

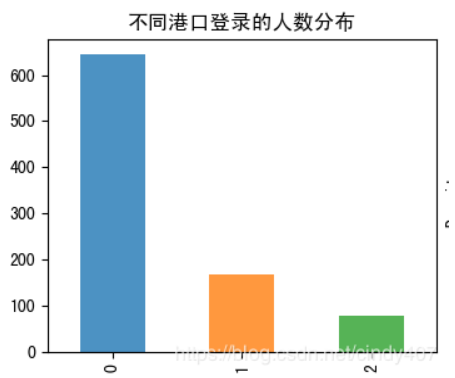
④不同港口登录的人数分布

#不同港口登录的人数分布

plt.subplot(2,3,4)

data['Embarked'].value\_counts().plot(kind='Bar',alpha=0.8)

plt.title('不同港口登录的人数分布')



S港口获救的人数最多，可能有距离上的优势，其余两个人数较少

⑤不同舱位的年龄分布

# 不同舱位的年龄分布

plt.subplot(2,3,5)

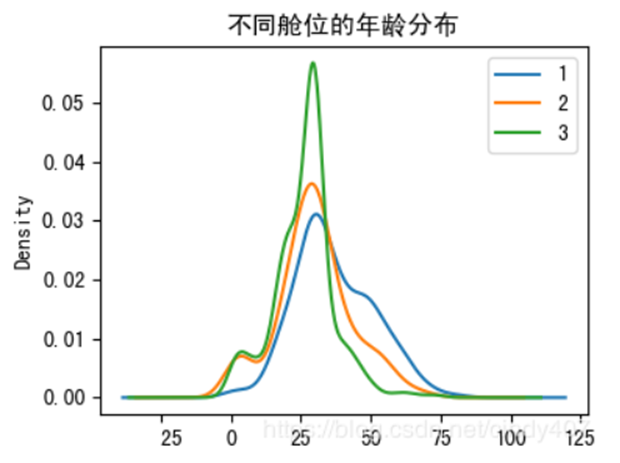
data['Age'][data['Pclass']==1].plot(kind='kde',label='1')

data['Age'][data['Pclass']==2].plot(kind='kde',label='2')

data['Age'][data['Pclass']==3].plot(kind='kde',label='3')

plt.title('不同舱位的年龄分布')

plt.legend(loc='best')

1

一等舱和二等舱年龄要稍长一些，三等舱要年轻化一点，这与不同年龄段累计财富比较一致

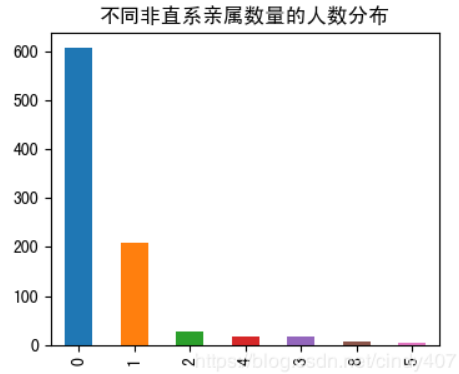
⑥不同非直系亲属数量的人数分布

# 不同非直系亲属数量的人数分布

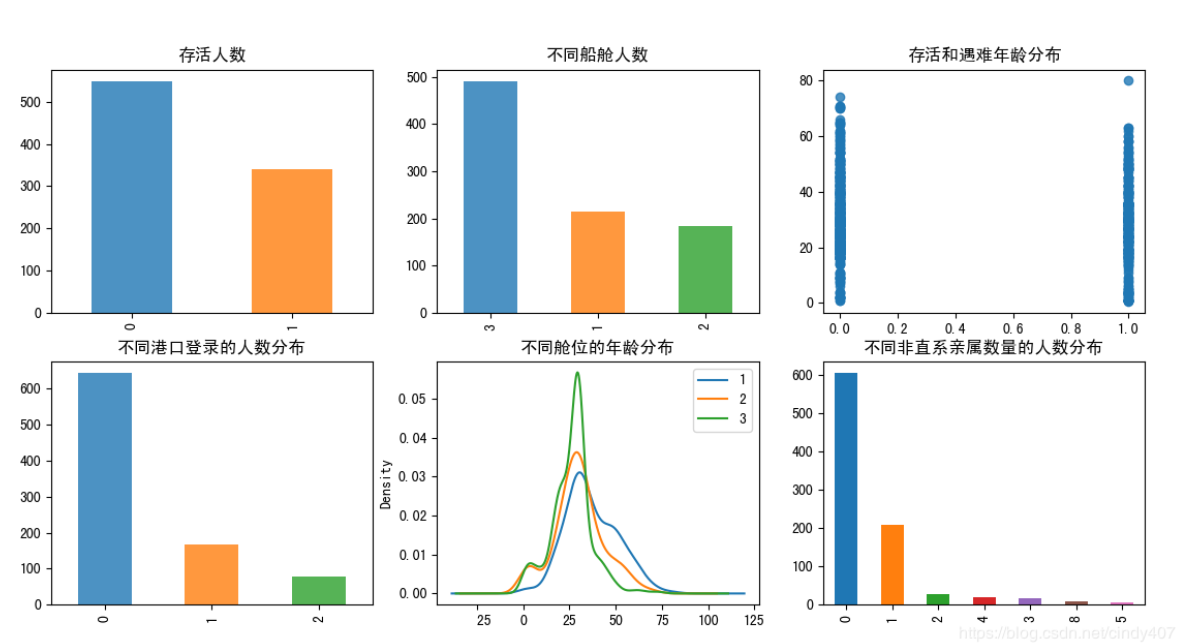
plt.subplot(2,3,6)

data['SibSp'].value\_counts().plot(kind='bar')

plt.title('不同非直系亲属数量的人数分布')



没有非直系亲属一起前行的占了绝大部分，其次是与一位同行的



⑦ 不同性别的获救情况分布

# 不同性别的获救情况分布

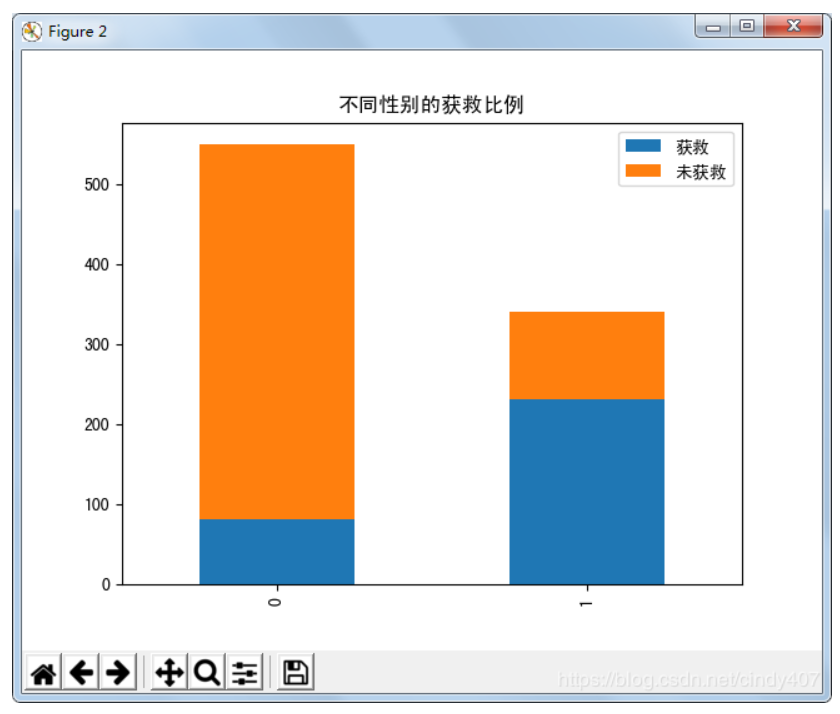
survive\_0 = data['Survived'][data['Sex']==0].value\_counts()

survive\_1 = data['Survived'][data['Sex']==1].value\_counts()

data1 =pd.DataFrame({'获救': survive\_1,'未获救': survive\_0})

data1.plot(kind='Bar',stacked=True)

plt.title('不同性别的获救比例')



从图上可看出，船上的男性大概是女性的1.5倍，男性获救的比例不到20%，女性获救的比例达到70%以上，说明船上的男士真的很绅士，把存活的机会留给了女性

⑧ 不同等级舱位的获救情况

# 不同等级舱位的获救情况

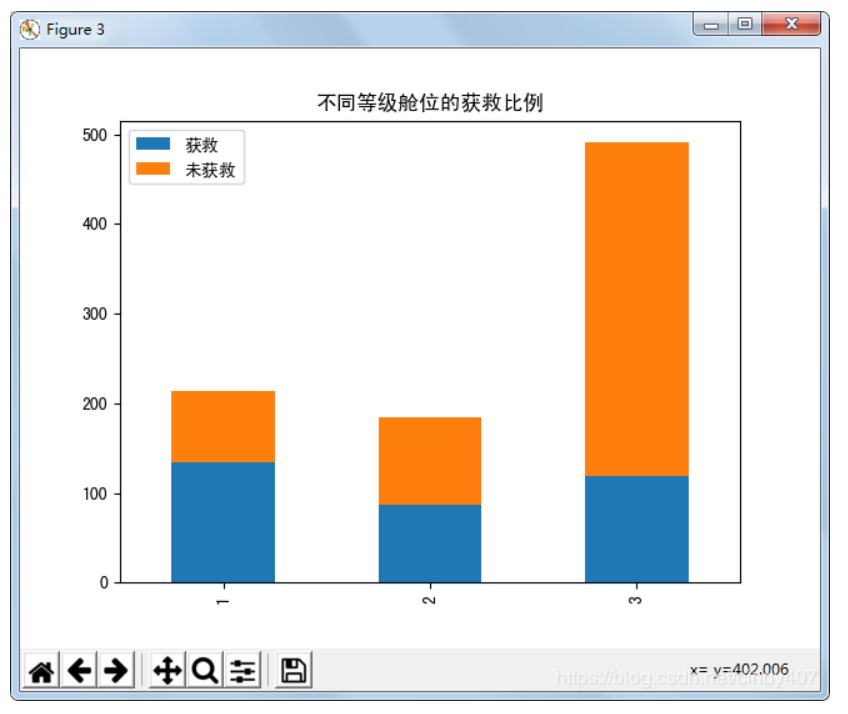
survive\_0 = data['Pclass'][data['Survived']==0].value\_counts()

survive\_1 = data['Pclass'][data['Survived']==1].value\_counts()

data2 =pd.DataFrame({'获救': survive\_1,'未获救': survive\_0})

data2.plot(kind='Bar',stacked=True)

plt.title('不同等级舱位的获救比例')



船舱等级越高，获救比例越高，说明越有钱，越容易得救

⑨不同登录港口的获救情况

# 不同登录港口的获救情况

survive\_0 = data['Embarked'][data['Survived']==0].value\_counts()

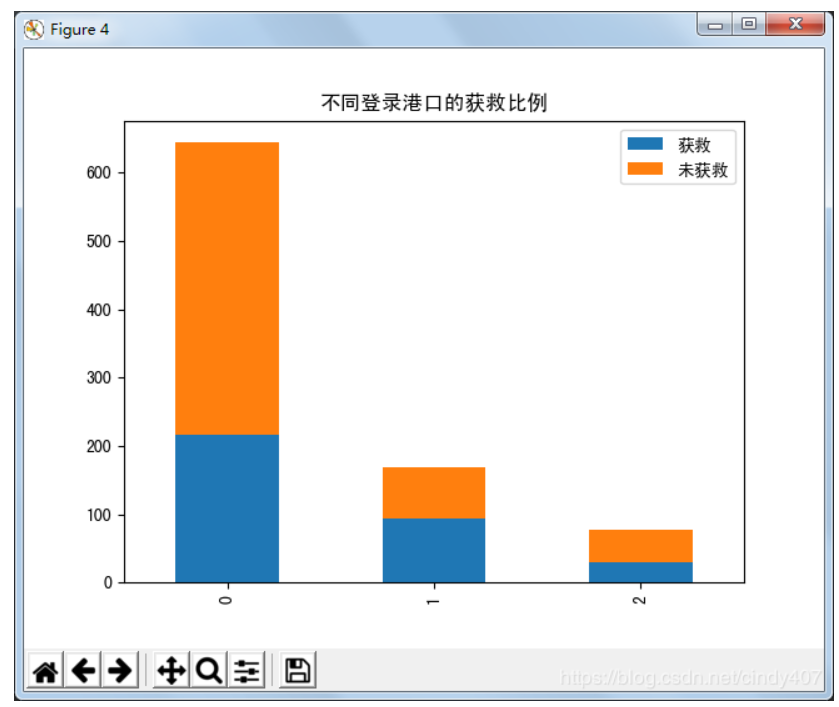
survive\_1 = data['Embarked'][data['Survived']==1].value\_counts()

data3 =pd.DataFrame({'获救': survive\_1,'未获救': survive\_0})

data3.plot(kind='bar',stacked=True)

plt.title('不同登录港口的获救比例')

plt.show()



S港口登录的人数最多，但是获救的比例却最低，C港口获救的比例是最高的，大概60%，Q港口为35%左右

上述就是基于Survived，Sex，Age，Embarked等字段的描述性分析，可以发现女生、头等舱、从C港口登录的乘客获救可能性更高，后续还会用机器学习的方法建模，找到最佳特征。

————————————————

版权声明：本文为CSDN博主「数据小斑马」的原创文章，遵循 CC 4.0 BY-SA 版权协议，转载请附上原文出处链接及本声明。

原文链接：https://blog.csdn.net/cindy407/article/details/91387707