**实验11——Pandas数据分析库1**

|  |  |  |  |
| --- | --- | --- | --- |
| 学号 |  | 姓名 |  |
| **一、实验目的**  掌握Pandas数据分析库中的Series和DataFrame数据结构，掌握缺失值的处理、统计函数、读写文件、数据整理以及分组统计。 | | | |
| **二、实验准备**  连接网络的计算机一台 | | | |
| **三、实验步骤**  启动Spyder，在IPython交互环境下完成下列练习。  1．Series基本练习  In: import pandas as pd  In: from pandas import DataFrame, Series  In: pd.\_\_version\_\_ # 显示Pandas版本  In: s = Series(np.arange(6), index=list('abcdef'))  In: s.index # 显示索引  In: s.values # 显示数据  In: s['a'], s[0], s['a':'c'], s[0:2] # 索引访问  In: s['c':'e']=3 # 赋值修改  In: s.sum(), s.mean(), s.median() # 统计函数  In: s.value\_counts() # 统计数据的频次  In: s.unique() # 返回不重复的数据列  In: s.head(3) # 显示头3个数据  In: s.tail(2) # 显示末尾2个数据  In: s2 = Series(np.arange(12, 30, 3), index=list('abcabc'))  In: s + s2 # 按索引对齐运算  In: s \* s2  2．创建和访问数据框DataFrame  In: data = {'apple':[1100, 1050, 1200], 'huawei':[1250, 1300, 1328], 'oppo':[800, 850, 750]}  In: df = DataFrame(data, index=['一月', '二月', '三月']) # index标签  In: df.apple, df.loc['一月']  In: df.iloc[0], df.iloc[:, 1]  In: df.at['二月', 'oppo'], df.iat[1, 2]  In: df.query('huawei>1300')  In: df.index, df.columns  In: df.to\_csv('m5.csv', encoding='cp936') # 保存文件  In: df2 = pd.read\_csv('m5.csv', index\_col=0, encoding='cp936') # 读取文件  3．nan缺失值处理  构造一个含有nan值的Series，练习isnull()、notnull()、dropna()、fillna()等函数。  b = Series(np.array([3, 4, np.nan, np.nan, None]))  b.isnull()  b.notnull()  b.dropna()  b.fillna(0)  4．统计函数  构造数据框，练习sum()、count()、mean()、median()、max()、min()、var()等函数。  data = {'id':[1,2,3], 'name':list('ABC'), 'score1':[89,85,99] , 'score2':[79,95,88]}  df = pd.DataFrame(data)  df.set\_index('id', inplace=True)  df.sum(); df.count()  df.mean(); df.median()  df.max(); df.min()  df.var()  5．读写文件  （1）构造一个文本数据文件，练习read\_csv()、to\_csv()函数。  有文本文件fee.csv，内容如下:  *姓名,月份,当月费用,余额*  *张三, 10 ,35.6 ,200.5*  *李四, 10 ,80 , 24.2*  *王五, 10 ,56 ,125.6*  df = pd.read\_csv('fee.csv', encoding='GBK')  df.to\_csv('fee2.csv', encoding='GBK', index=False)  （2）构造一个Excel文件，练习read\_excel()、to\_excel()函数。  将上题数据保存为Excel文件 fee.xlsx。  df = pd.read\_excel('fee.xlsx')  df.to\_excel('fee2.xlsx', index=False)  6．数据整理  （1）练习数据框的drop()、pop()、reindex()、reset\_index()、set\_index()、append()方法。  data = {'apple':[1100, 1050, 1200], 'huawei':[1250, 1300, 1328], 'oppo':[800, 850, 750]}  df = DataFrame(data, index=['一月', '二月', '三月'])  df.drop('一月') # 删除行'一月', 不直接改变df  df.pop('oppo') # 删除 oppo列, 直接改变df  df.reindex(index=['一月','三月','四月']) # 剔除‘二月’，新增‘四月’  df.reset\_index(inplace=True) # 将原索引列'月份'变为数据列，索引改为整数序号索引  df.set\_index('apple') # 将apple列设为索引列  df.append({'index':'四月','apple':1400,'huawei':1500}, ignore\_index=True)  （2）构造含有重复值的数据框，练习针对重复值的duplicated()、drop\_duplicates()方法。  df = DataFrame({'product':['A', 'B', 'C','A'], 'color':['r', 'b','r','r']})  df.duplicated()  df.drop\_duplicates()  （3）构造数据框，练习sort\_values()和sort\_index()排序方法。  df = DataFrame({'id':list('axc'), 'quantity':[300, 240,400],'ptype':list('BCB')})  df.set\_index('id',inplace=True)  df.sort\_index()  df.sort\_values('quantity')  df.sort\_values(by='ptype')  （4）构造两个数据框，含有共同的列，练习merge的各种连接方法。  df1 = DataFrame({'color': ['r', 'b', 'w', 'w'], 'c1': range(4)})  df2 = DataFrame({'color': ['b', 'w', 'b'], 'c2': range(2, 5)})  df3 = pd.merge(df1, df2)  （5）构造区间[0, 100]内的80个年龄随机整数，按区间[0, 10, 18, 35, 50, 70, 100]做pd.cut()分段统计。  ar=np.random.randint(0,101,80)  bins=[0, 10, 18, 35, 50, 70, 100]  ar\_cat=pd.cut(ar,bins=bins)  print(ar\_cat)  print(ar\_cat.value\_counts())  7．分组统计  构造一个至少含有3个班级名册的Excel成绩表（姓名，学号，班级，性别，课程1，课程2）。用Pandas读入该表，完成如下计算：  （1）计算每个同学的平均分［平均分 =（课程1 + 课程2）/2］。  （2）使用rank()方法按平均分给出一个全年级的名次排名。  （3）按班级统计每门功课的平均分。  （4）统计男生、女生人数，并给出男生、女生的平均成绩。  注：stu.xlsx 文件在随书的配套数据文件中。  import pandas as pd  df = pd.read\_excel('stu.xlsx')  df['平均分'] = (df['课程1'] + df['课程2']) / 2  df['排名'] = df['平均分'].rank(ascending=False) # 按平均分降序得到排名列  df2 = df.sort\_values(by='排名') # 按排名列得到名次表df2  print('名次为:\n', df2)  print(df.groupby('班级')['课程1', '课程2'].mean()) # 按班级统计每门课的平均分  print(df.groupby('性别').size()) # 按性别统计人数  print(df.groupby('性别')['课程1', '课程2'].mean()) # 按性别求平均分  8. 世界杯球队数据分析  在文件“2018世界杯球队数据.csv”中存储了2018年世界杯32支参赛球队的相关数据，请利用pandas库和matplotlib库编程实现如下功能：  （1）利用pandas库的read\_csv()函数读取“2018世界杯球队数据.csv”中的数据，并存入一个DataFrame对象中。  （2）输出净胜球（进球减去失球）大于0的球队。  （3）输出被罚红牌的球队。  （4）输出进球成功率（进球数/射门数）超过10%的球队及其进球数和射门数。  （5）输出进球数超过平均数且被罚黄牌少于5张的球队及其进球数和黄牌数。  （6）按照进球数降序输出所有球队及进球信息。  （7）按照所属区进行分组，按升序统计输出每个区的进球数。 | | | |
| **四、实验过程及截图****[学生填写]**  【只需要给出第8题“世界杯球队数据分析”的截图】 | | | |
| **五、实验结论与分析讨论[学生填写]**  **（1）列出本次实验你认为重要的概念、术语。**  **（2）谈谈你认为本次实验需要注意的地方。**  **（3）实验过程中是否存在问题，如果有，你解决的办法是什么？** | | | |