**一、实验目的及要求**

1. 掌握对数据集成、清洗概念的理解

2. 掌握数据清洗，样本数据构建基础方法

3. 掌握基本的统计方法

4. 按照既定格式书写实验报告

**二、实验设备与平台**

1. 实验设备：计算机；

2. 平台：Windows 10操作系统

**三、实验内容**

**题目**：广州大学某班有同学100人，现要从两个数据源汇总学生数据。第一个数据源在数据库中，第二个数据源在txt文件中，两个数据源课程存在缺失、冗余和不一致性，请用C/C++/Java程序实现对两个数据源的一致性合并以及每个学生样本的数值量化。

数据库表：ID (int), 姓名(string), 家乡(string:限定为Beijing / Guangzhou / Shenzhen / Shanghai), 性别（string:boy/girl）、身高（float:单位是cm)）、课程1成绩（float）、课程2成绩（float）、...、课程10成绩(float)、体能测试成绩（string：bad/general/good/excellent）；其中课程1-课程5为百分制，课程6-课程10为十分制。

txt文件：ID(string：6位学号)，性别（string:male/female）、身高（string:单位是m)）、课程1成绩（string）、课程2成绩（string）、...、课程10成绩(string)、体能测试成绩（string：差/一般/良好/优秀）；其中课程1-课程5为百分制，课程6-课程10为十分制。

两个数据源合并后读入内存，并统计：

1. 学生中家乡在Beijing的所有课程的平均成绩。
2. 学生中家乡在广州，课程1在80分以上，且课程9在9分以上的男同学的数量。(备注：该处做了修正，课程10数据为空，更改为课程9)
3. 比较广州和上海两地女生的平均体能测试成绩，哪个地区的更强些？
4. 学习成绩和体能测试成绩，两者的相关性是多少？（九门课的成绩分别与体能成绩计算相关性）

**实验实施**：

（在此详述平台，技术栈，思路，处理逻辑等等）

平台：Windows10

技术栈：python

思路：先对不同源的数据进行读取、合并和存储，然后按题目要求对存储的数据进行计算。

处理逻辑：先读取，后处理，再计算。

**实验一二问代码：**

import pandas as pd

df=pd.read\_csv('合并后的数据.csv',encoding='gbk')

# print(df)

df1=df[(df['City'])=='Beijing']

df2=df[(df['City']=='Guangzhou')&(df['C1']>=80)&(df['C10']>=9)&(df['Gender']=='boy')].shape[0]

# print(df1)

C1\_mean=df1['C1'].mean()

C2\_mean=df1['C2'].mean()

C3\_mean=df1['C3'].mean()

C4\_mean=df1['C4'].mean()

C5\_mean=df1['C5'].mean()

C6\_mean=df1['C6'].mean()

C7\_mean=df1['C7'].mean()

C8\_mean=df1['C8'].mean()

C9\_mean=df1['C9'].mean()

C10\_mean=df1['C10'].mean()

print('北京学生C1课程平均成绩 % .2f'% C1\_mean)

print('北京学生C2课程平均成绩 %.2f'% C2\_mean)

print('北京学生C3课程平均成绩 % .2f'% C3\_mean)

print('北京学生C4课程平均成绩 % .2f'% C4\_mean)

print('北京学生C5课程平均成绩 % .2f'% C5\_mean)

print('北京学生C6课程平均成绩 % .2f'% C6\_mean)

print('北京学生C7课程平均成绩 % .2f'% C7\_mean)

print('北京学生C8课程平均成绩 % .2f'% C8\_mean)

print('北京学生C9课程平均成绩 % .2f'% C9\_mean)

print('北京学生C10课程平均成绩 % .2f'% C10\_mean)

print('广州男学生C1成绩80分以上并且C10成绩9分以上数量为%.2f'%df2)

**实验三四问代码：**

import pandas as pd

df=pd.read\_csv('合并后的数据.csv',encoding='gbk')

df1=df[(df['City']=='Guangzhou')&(df['Gender']=='girl')]

df2=df[(df['City']=='Shanghai')&(df['Gender']=='girl')]

# print(df1)

# print(df2)

def cons2num(x):

if x=='excellent':

return 90

elif x=='good':

return 85

elif x=='general':

return 70

elif x=='bad':

return 60

GZ\_PE\_mean=df1['Constitution'].apply(cons2num).mean()

SH\_PE\_mean=df2['Constitution'].apply(cons2num).mean()

if GZ\_PE\_mean>SH\_PE\_mean:

print('广州女生平均体能测试成绩比上海女生好\r\n平均体能测试成绩分别为%.2f、%.2f\r\n'% (GZ\_PE\_mean,SH\_PE\_mean))

elif GZ\_PE\_mean==SH\_PE\_mean:

print('广州女生平均体能成绩与上海女生实力相当\r\n平均体能测试成绩分别为%.2f、%.2f\r\n'% (GZ\_PE\_mean,SH\_PE\_mean))

else:

print('广州女生平均体能成绩比上海女生差\r\n平均体能测试成绩分别为%.2f、%.2f\r\n'% (GZ\_PE\_mean,SH\_PE\_mean))

# print(GZ\_PE\_mean)

# print(SH\_PE\_mean)

# print(df1)

col\_name1=['C1','C2','C3','C4','C5']

col\_name2=['C6','C7','C8','C9','C10']

mul\_10\_grade=df[col\_name2].apply(lambda x:x\*10)

academic\_mean=(df[col\_name1].sum().sum()+mul\_10\_grade.sum().sum())/1000

pe\_mean=df['Constitution'].apply(cons2num).mean()

each\_aca\_mean=df[col\_name1].apply(lambda x:x.sum(),axis=1)+mul\_10\_grade.apply(lambda x:x.sum(),axis=1)

each\_aca\_mean=each\_aca\_mean.apply(lambda x:x-academic\_mean)

each\_pe\_mean=df['Constitution'].apply(cons2num)

each\_pe\_mean=each\_pe\_mean.apply(lambda x:x-pe\_mean)

cov=(each\_aca\_mean\*each\_pe\_mean/104).sum()

std\_aca=pow(pow(each\_aca\_mean,2).sum()/104,0.5)

std\_pr=pow(pow(each\_pe\_mean,2).sum()/104,0.5)

std\_aca\_pe=pow(std\_aca\*std\_pr,0.5)

correlation=cov/std\_aca\_pe

# print(each\_aca\_mean)

# print(each\_pe\_mean)

# print(pe\_mean)

print("学习成绩和体能测试成绩的相关系数：%s" % correlation)

if correlation<0.3:

print("得出结论：学习成绩与体能成绩不相关")

elif correlation>=0.3 and correlation<=0.8:

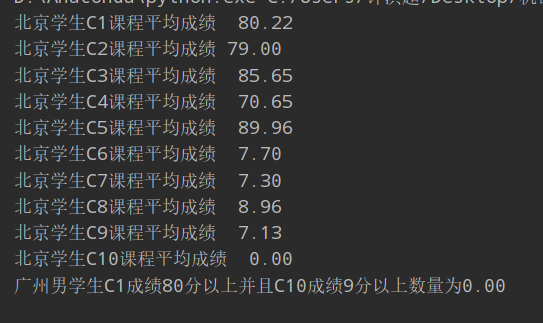
print("得出结论：学习成绩与体能成绩弱相关")

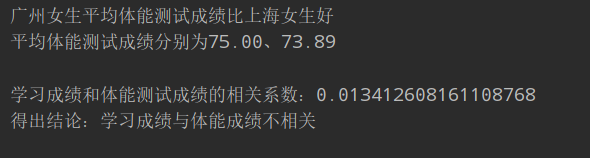
else:

print("得出结论：学习成绩与体能成绩相关")

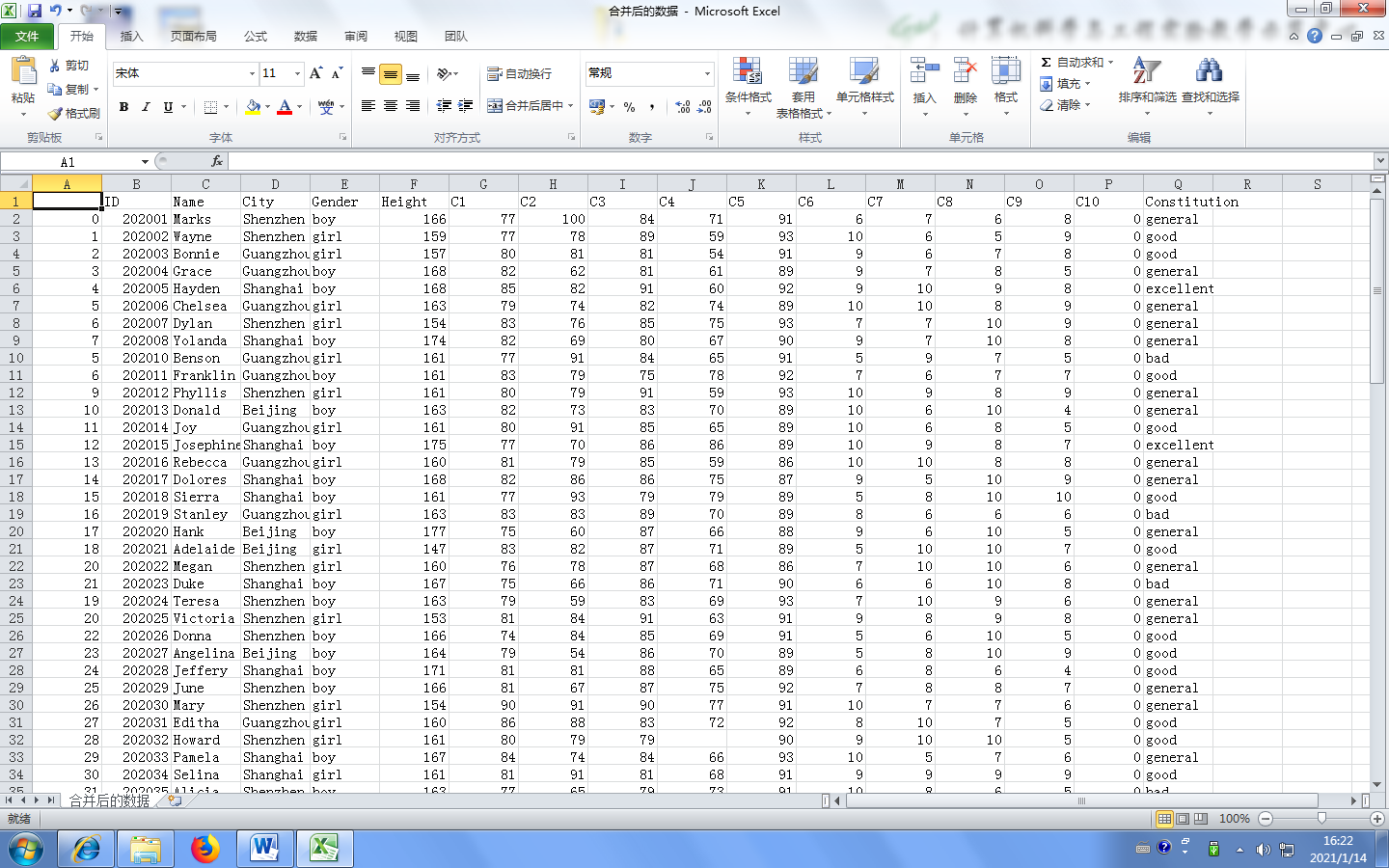
**运行结果**：

（运行结果、现象的说明与截图）





合并后的数据：



**实验总结**：

（遇到的问题，有什么感想和收获等等）

对数据的合并与处理有了实际的认知，对数据挖掘与机器学习有了更加深入的理解。

|  |
| --- |
| 本次实验用Python完成，先进行了数据库的连接，在数据库中导入csv表后，进行数据库表和txt表的合并，之后对合并的表进行清洗，对冗余和不一致性的进行并运算然后再删除重复，填补缺失的数据(数值按列填补列浮点数平均值，体能成绩空值填补bad），然后进行数据的规范化，对单位不同的数值进行一致规范，之后再进行相关的统计计算，求得对应问题的平均数，对体能测试的文字表述进行数值赋分再进行计算，求出学习成绩与数值体能成绩的协方差与标准差，最后依照公式得出相关系数。 |