**组员信息**

李华辉（组长） 冼海俊

**分工信息**

李华辉：代码的实现部分+文档编写（80%）

冼海俊：代码的规范化+文档编写+作业提交（20%）

**作业题目和内容**

* + 题目：多源数据集成、清洗和统计
  + 内容：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。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10在9分以上的男同学的数量。

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？

**作业环境**

* + 文件说明：一个python 文件 main.py
  + 函数说明：自己编写了一个计算相关性的函数 cor(a,b) 调用了计算开根号的函数math.sqrt()，计算列表所有元素和的sum()，计算列表长度的len()，导入了pandas包，使用了它的导入数据源不同的函数 .read\_excel（）和.read\_csv（），数据源去重函数 drop\_duplicates（），多数据源合并函数merge（），空值替换函数 fillna（），本次实验主要涉及的主要技术为pandas 里面的dataframe 的相关数据处理操作。

**难题与解决**

在刚开始的时候不知道如何对数据进行操作，然后去了解了pandas库之后，学习了许多对数操作的方法。

刚开始不知道以什么标准去合并两个数据源，后来观察两个数据源的ID发现，他们的差值都是202000，所以就可以统一ID，以ID为唯一标准去合并两个数据源，运用数据库的外连接思想。

**总结**

通过本次实验，对多源数据源的合并、处理、清理有了初步的认识，对pandas库也有了初步的了解，掌握了数据清洗的具体步骤和套路。