# 《机器学习与数据挖掘实验》

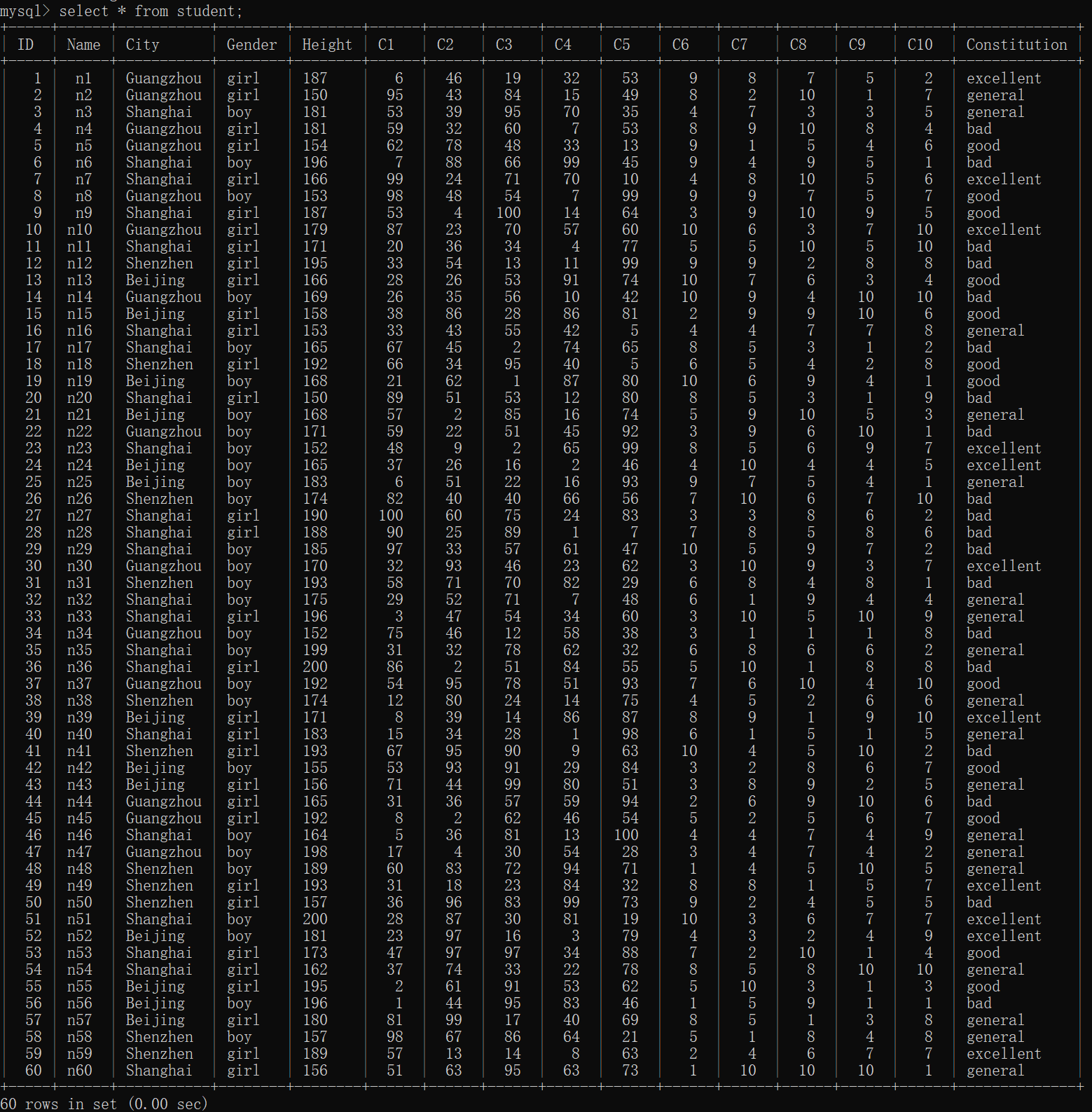
## 实验一 《多源数据集成、清洗和统计》

**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。

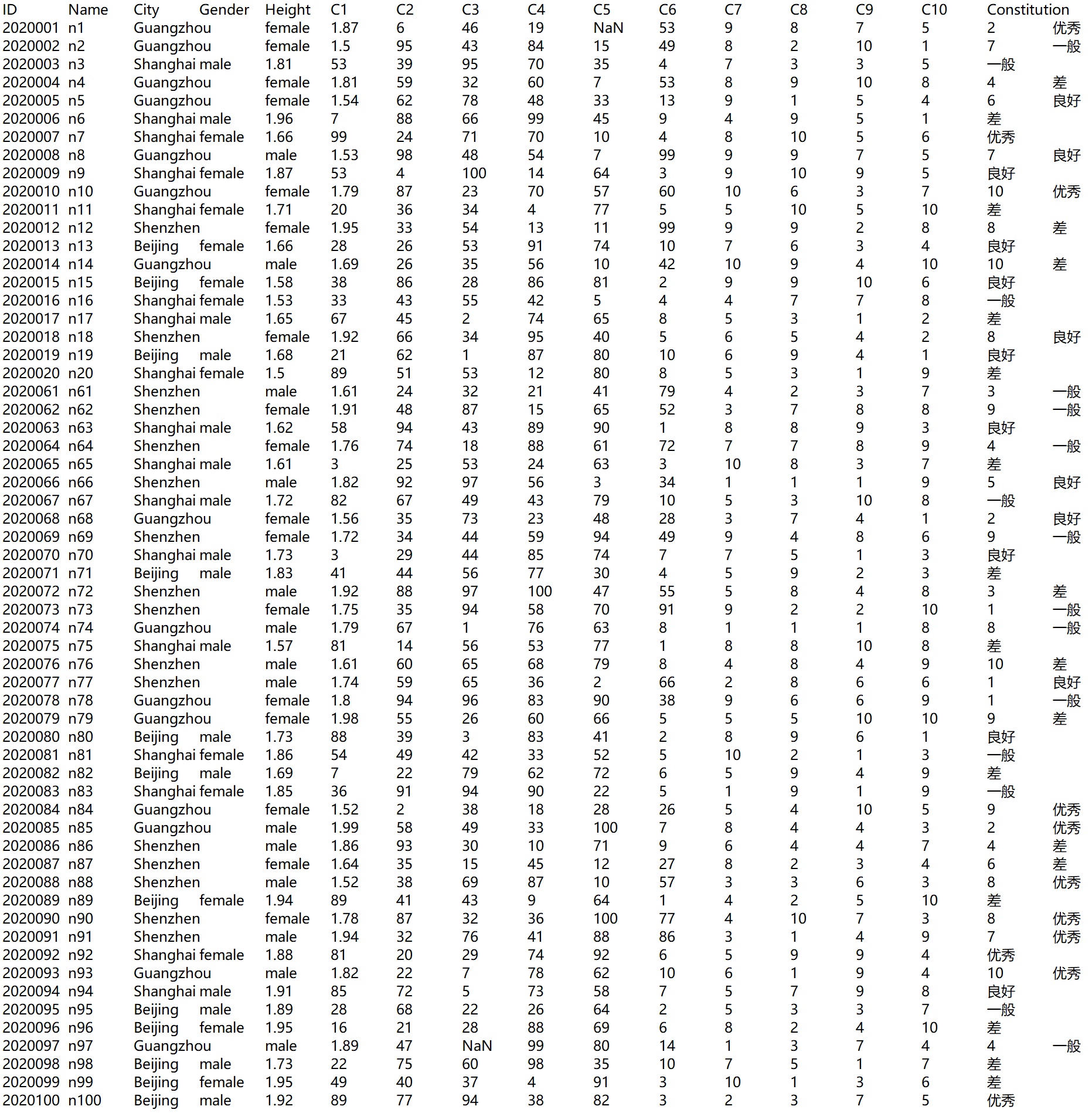
**数据库表：**ID (int), 姓名(string), 家乡(string:限定为Beijing / Guangzhou / Shenzhen / Shanghai), 性别（string:boy/girl）、身高（float:单位是cm)）、课程1成绩（float）、课程2成绩（float）、...、课程10成绩(float)、体能测试成绩（string：bad/general/good/excellent）；其中课程1-课程5为百分制，课程6-课程10为十分制。

**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。

数据库中Student表数据



student.txt中

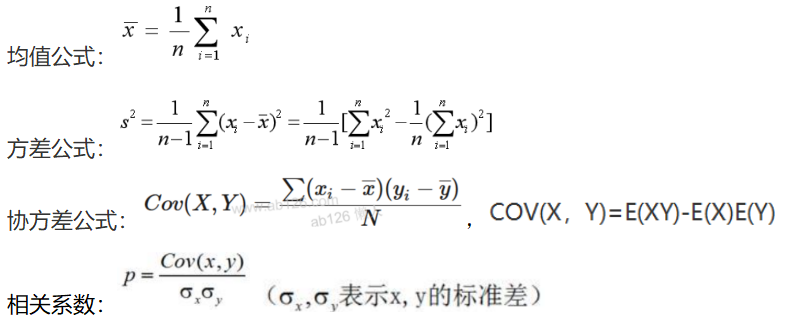


两个数据源合并后读入内存，并统计：

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Shenzhen，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？



### 代码实现：

