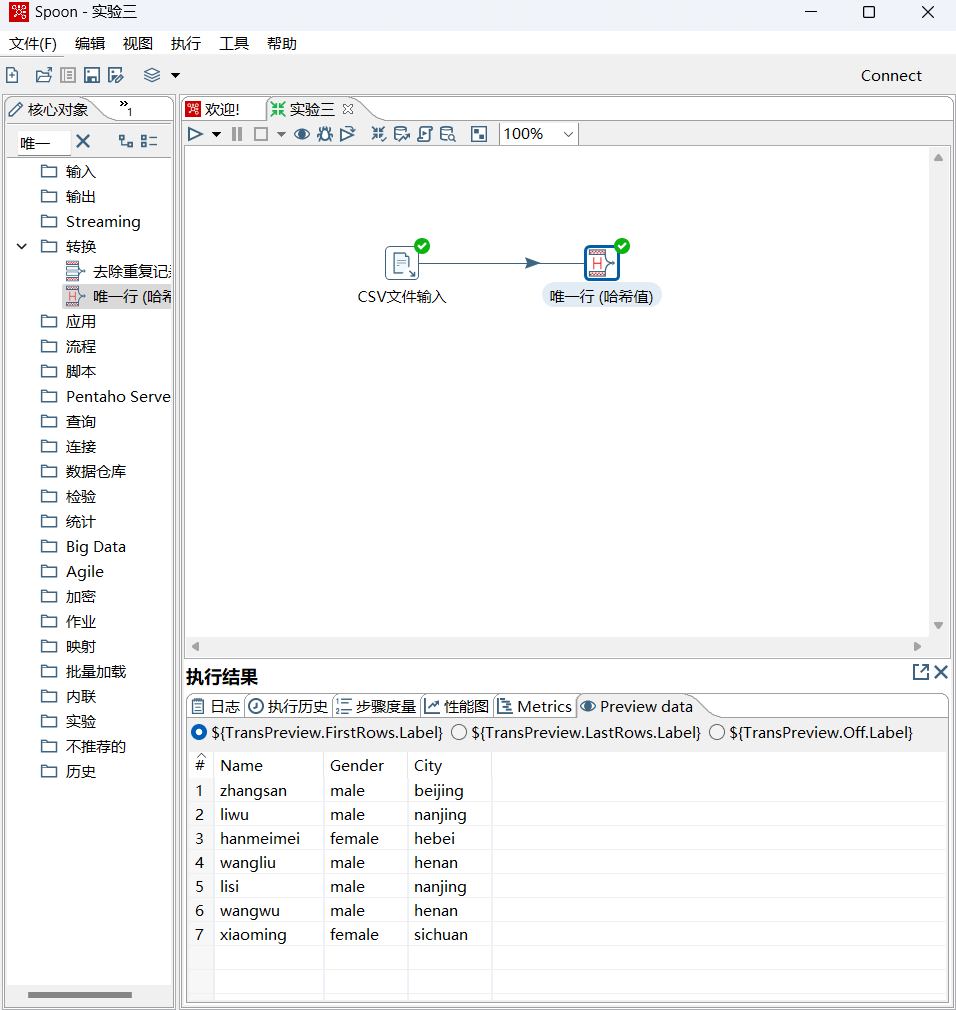
**项目三 清洗与检验（4学时）**

1. **实验目的**
2. 掌握数据去重的方法
3. 掌握数据缺失值处理的方法
4. 掌握数据异常值处理的方法
5. **实验内容**

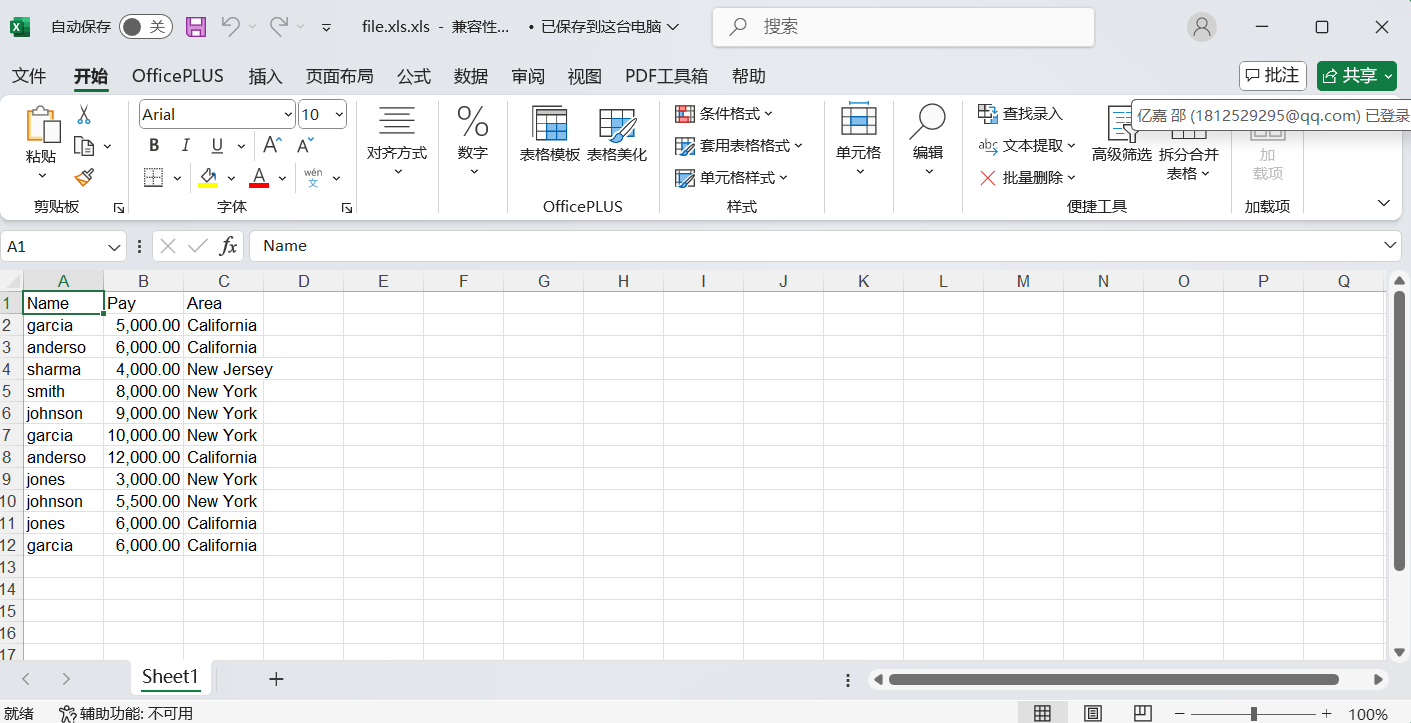
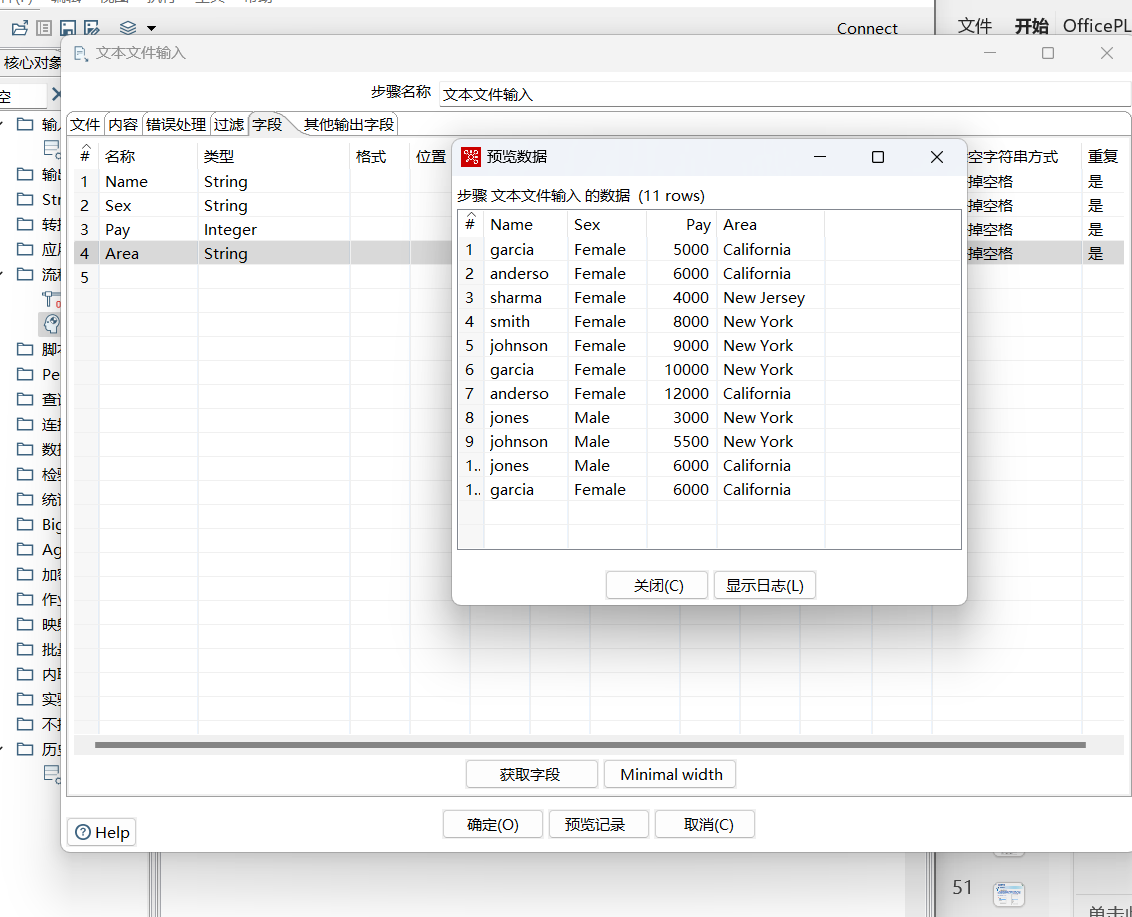
**（一）动手操作Kettle工具，实现数据的完全去重。**

（1）通过Kettle工具，消除CSV文件merge.csv中完全重复的数据。



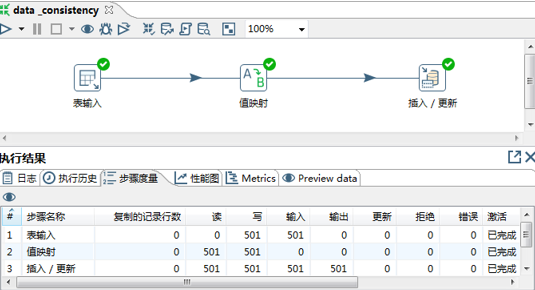
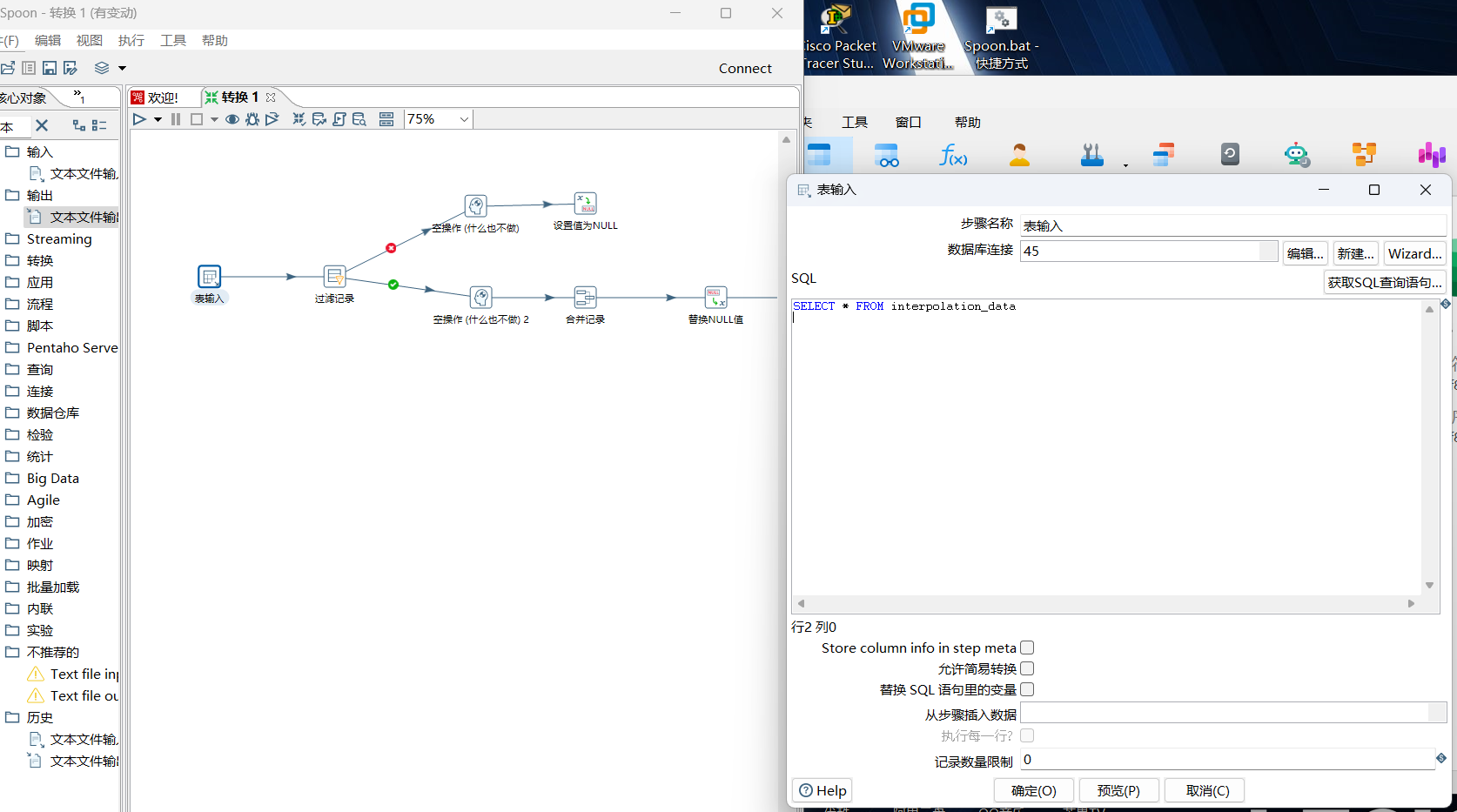
**（二）动手操作Kettle工具，实现去除缺失值、填充缺失值。**

（1）通过Kettle工具，去除原始数据集revenue.txt中的缺失值（去掉性别所在的列，并删除其他三个属性中存在缺失值的行记录）。

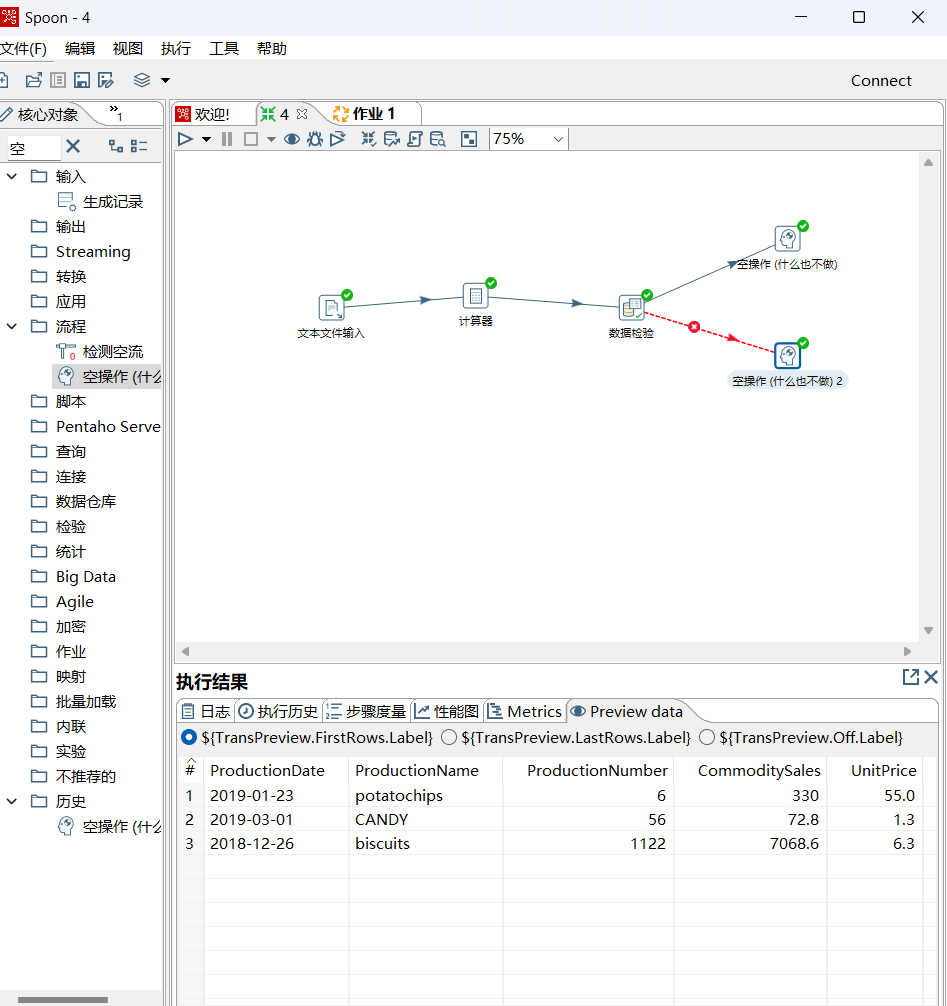
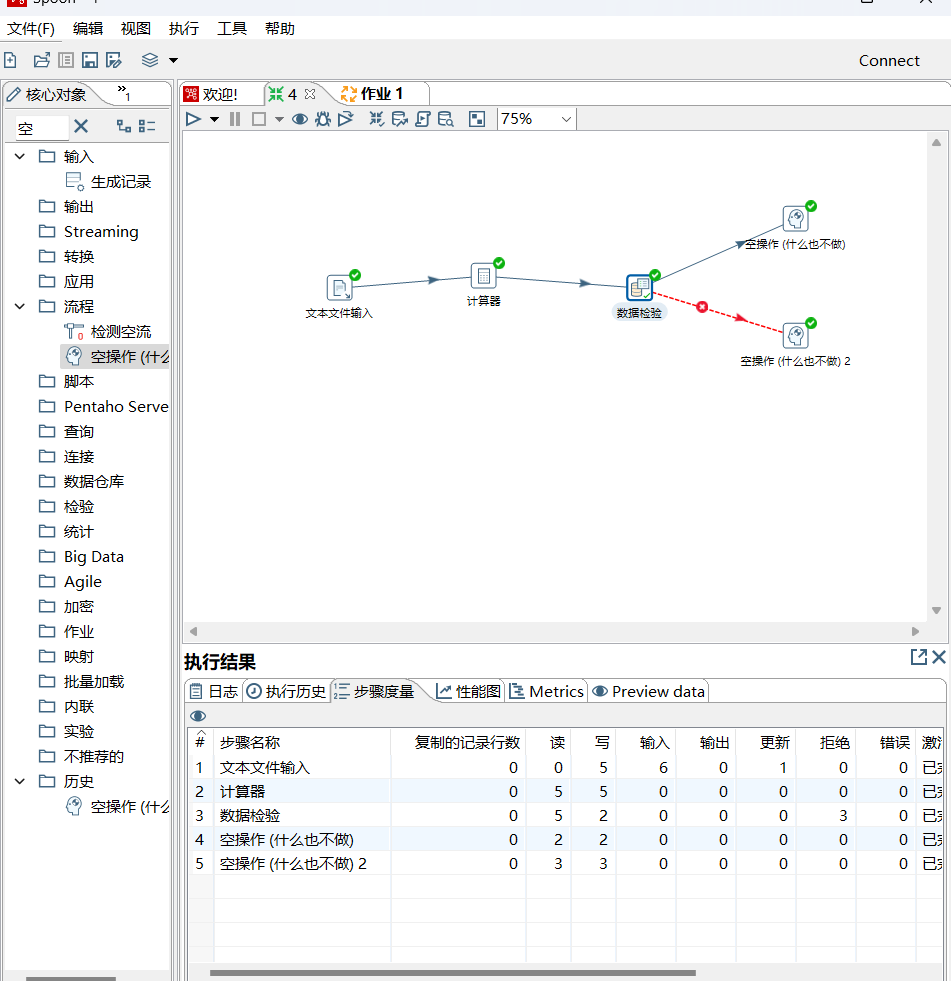
（2）通过Kettle工具，使用平均值填充法对文件people\_survey.txt中的缺失

**（三）动手操作Kettle工具，实现删除包含异常值的记录和修补异常值。**

（1）通过Kettle工具，替换和修改数据表interpolation\_data中的身高异常值，并将数据以文本文件输出。（需先执行interpolation\_data.sql语句，创建对应的表数据，其中身高的正常取值范围[114,226]，将异常值替换为平均值170）。



**（四）动手操作Kettle工具，实现删除包含异常值的记录和修补异常值。**

通过Kettle工具，对“salesData.txt”文件中的数据进行规范化检验，具体要求为：（1）由产品销售额CommoditySales和销量ProductionNumber计算产品单价（2）检验出单价大于10的产品；（3）检验出日期在2019-01-01之前的产品。

2、实验过程中出现的问题以及解决方法；

数据库错误，进行检查重做。

1. 实验心得。

掌握了去重方法后，我明白了如何精准地识别重复的数据行或数据记录，通过合适的工具和算法（如利用编程语言中的特定函数或者数据库的去重语句等），将冗余的数据剔除，使得数据更加精炼、准确，为后续深入分析奠定坚实基础。