# 数据仓库与数据挖掘 第一次作业

姓名：王宸昊

学号：2019214541

时间：2019.10.19

## 数据预处理

首先，先通过观察，分析几张表当中数据存在的噪音类型，然后通过脚本针对数据中的噪音进行分类，再通过脚本进行处理。

在这里通过Python的脚本对数据进行分析和处理，其中利用pandas库对csv文件进行数据分析。

从脚本中的数据可以分析出，首先解决的问题是数据的冗余问题，在每张表中都存在一些行的数据冗余，在这里处理的方法是pandas中的drop\_duplicates()方法，在导入csv文件后先进行一步去重的工作。