数据整理过程

下面简述一下我做《数据清洗》这个项目时的步骤和过程。

第一步，理解题意，理解清楚项目的要求。我在理清题意这部分花去了过多的时间，导致后面接下来的开始编程的时间延迟，也让自己越来越紧张，这样不太好。这个项目中刚开始主要是对三个文件之间的关系比较混乱，不知道这三个文件应该分开来整理分析还是合并后整理分析，也不知道各个字段的意思，因此而更加的头疼。后续反复研读和在助教的帮助下理解清楚了题意，才开始推进项目。

第二步，收集数据。这次的这个步骤没什么问题，一共三个文件，都通过编程的方式下载到本地上，然后完成了这个步骤。

第三步，评估。这一步花的时间也是很多的，主要是感觉自己目前对数据还很不敏感，对各个字段的意思也很不清楚。刚读取打开的表格的时候脑子里面一篇空白，看了很久才渐渐的一个一个收集评估发现的问题，然后再把他们分类成质量问题和整洁性的问题。主要是分析得少，咋一看很多字段的表，很多数据量的表心理就害怕，嗯，慢慢看了后可以找到一个一个出现的问题了，然后最后汇总起来，相当于是为下一步清洗过程做了一份任务清单。

第四步，清洗过程。清洗过程包括定义，编码和测试，并且可以迭代进行这个过程。因为第三部已经收集好了评估时发现的问题，第四部的定义就可以直接根据第三部汇总的问题一个一个来解决。这一步主要的问题是对 Pandas 各功能的熟悉，我也卡了很长时间，特别

是关于 iloc 这个，反复不会就找助教，助教帮助下，我渐渐的把收集到的问题一个一个解决下去，心理也越来越自信。

第五步，分析。这一步就可以把整理好的数据拿来分析了，所以一开始最好要知道自己想分析什么，整理的时候就是为了达成能够分析的目的而整理相关的地方的。这样下来顺势就可以开始分析自己整理好的数据了。

以上就是我的项目解题过程。