**关于数据清洗项目的完成过程**

从接触该项目到结束完成的整个过程中，最让我印象深刻，同时让我吃很多亏的经验教训是：无论做什么项目，千万不要半途而废，至少每天都要游览一遍项目，否则在因为一段时间的缺席，会导致每次重新捡起项目时都需要重头到位再在自己脑海中过一遍思路，不仅浪费时间，而且容易出错。其他遇到的问题或是困难可以总结一下几点：

1、由于本人无法登陆国外网站(翻墙)，所以只能载入项目中给的各文本文件，(PS:其实如果可以的话，还是应该花点RMB买个软件然后去目标网站爬取所需的目标数据，这样更能贴近实战项目要求。)。在成功导入前两个文件后，在碰到数据类型为json的文本文件时，由于对该类型文件不熟悉，花了一些时间，首先用文本编辑器Atom打开该文件，了解json文件的类型以及其嵌套规律，然后决定采用read\_json方法提取数据；

2、在成功导入三个文件并用各种方法或函数查看三个数据集的时候，自己内心就开始有了无数的疑问，比如twitter中对于狗狗的评分千差万别，哪些是合理的，而哪些又是异常值；再比如对于狗狗名字的查询，发现一些例如a这些字符串是否是正确的等等；继续深入项目，发现自己对于这个项目越来越不清楚目标是什么，变得迷茫和无助。后来，通过求助于课程助教和网站论坛逐渐对于项目甚至自身的缺陷有了一些清晰的认识，例如对于狗狗评分应先做到正确性，再允许的情况下再去分析是否异常值等等；

3、第三个难点就是对于提取某些列中的关键数据时用到的正则表达式，比如twitter中的text列，需要提取出对应的狗狗地位、评分以及姓名。这对于我这个关于正则表达式一窍不通的学员来说，实在是犹如天书。幸亏得益于助教老师的公开课以及优达论坛各神秘大神前辈的经验，才能顺利的攻克此关；

其实，对于项目中各种难点也好，困境也罢，都是一种对于自身技能的磨练。在面对这些的时候，不要退缩，积极思考，勇于探讨，我认为优达的课程会是我们数据分析路上很好的一位导师。