本次项目首先是收集数据的部分，使用到了 BeautifulSoup 和 Requests 库获得了第一个数据集，其余数据集以读入 tsv,json 格式获得。

接着是对数据的目测评估和编程评估，目测主要使用 sample 函数随机抽取条目，编程评估主要使用.info()，.describe()，.count_values()，.duplicated()函数进行评估。发现数据集在数据类型，缺失值，数值准确性，整洁度上均有问题。比如说 source 列出现了多余的文本,例如 HTML 标签和网址文本；expanded_urls 列存在部分空值,需要删除；ID 的数据类型应该为 string 而不为 integar;in_reply_to_status_id 以及 in_reply_to_user_id 列缺失值过多,需要删除；name 列出现了'a',明显不为名字,需要重新从 text 中提取宠物名；第二个数据表 jpg_url 列有 66 个重复行,需要删除；doggo,floofer,pupper,puppo 为类型数据,可以合并为新的一列，并且缺失值较多，需要重新从文本中提取；所有表格中观察对象相同，可以将三个数据片段进行合并。

本次项目中印象很深的一部分是根据 HTML 的规则从网页内容中抓取所需的数据，还有运用类似正则表达式从文本中取出需要的部分，比如狗狗的名字、类型等等。包括通过对表格逻辑性的揣摩，对表格进行合并(pd.merge)，以及最后独立结论的生成，对相关性的探索，包括狗狗评分与点赞数、转发数的关系。总体来说，数据抓取和清洗是数据科学家耗时最长的部分，也是灵活性技巧性非常高的一部分，在未来的学习中，我也将继续在此方向努力。