数据清理报告

在这次的对 tweet 上 WeRateDogs 的数据的清理项目中，我们手头上一共有三份数据来源：一:从 tweeter_archive_enhanced.csv 中导入的 WeRateDogs 的推特档案的数据，二：通过 Python 的 Request 导入的根据神经网络对推特图像的预测的每个狗狗品种的预测结果，三：我们使用 python 的通过 Tweet 的 API(tweepy)导入的每条推特的额外数据(JSON 格式)。

我决定采取的思路是先对每个数据集依次评估他们的质量和整洁度问题，最后再将三个数据集拼接起来进行最后的数据分析和可视化。

首先，我先是观察了 data_1 的表头和表尾的几条数据的全部内容大概了解了一下数据集的形式，数据集内容包括：推特的基本信息

tweet_ID,text,source,tweet,timestamp,in_reply_to_staus_id,in_reply_to_user_id(表示是否为回复其他的 tweet),retweeted_staus_id,retweeted_user_id(表示是否为转发其他的 tweet)，以及狗的名字和品种 name。然后观察了每列是否有空的信息以及他们的基本格式，我发现以下几个问题：①timestamp 中的时间格式不对，应该为 datetime；②in_reply_to_status_id 和 in_reply_to_user_id 列的类型应该为 int；③ retweeted_status_id 和 retweeted_status_user_id 列的类型应该为 int。随后，我继续根据我对每列的内容的了解继续对每列具体

观察，我发现：在 name 列中有些狗的名字叫做 a，such，that，于是我通过布尔索引看他们那一行的内容，我发现这可能是作者编程的错误，他将 is 后面的单词默认为狗的名字，我决定通过 pandas 的字符串替换将他们全部替换为 None。通过观察项目动机，我发现只需要含有图片的原始评级（不包括转发)。尽管数据集中有多条数据，但是并不是所有都是狗评分，并且其中有一些是转发，于是我只选择”in_reply_to_status_id"和"retweeted_status_id"列为空的行。最后，通过观察我发现在评分的分母列中有些行的不为 10，这不合常理，我决定要将他们所在行删除。随后在整洁度的评估中我发现 Source 列中的内容表达不明确，他的格式应修改为 Https 网址的形式，于是我决定通过字符串截取的形式选取其中的网址。至此，我完成了对 data_1 数据集的清洗。

随后我对数据集二进行了观察和评估，但经过对他们内容和格式的仔细观察，我没有发现任何的质量和整洁度上的问题，只能说这个数据集制作的十分好。

然后，我以同样的方法开始了对数据集三的观察和评估，由于数据集三是我自己从 JSON 中提取的部分内容，因此没有什么太大的问题，主要是 create_time 中时间的格式不正确，我也是打算通过 date 库中的函数将其转化为 datetime 格式。

最后，我依次对三个数据集进行了数据的清洗，并且将数据集二和数据集三通过左连接的方法作为额外信息添加到数据集一中。至此我完成了该项目对数据集的清洗。