wrangle report

回头看本次作业，两个部分花了特别长时间

1. 理解题目要干什么？刚开始看了很多遍都不知道题目究竟要干什么， 比如一下要求：
2. 我们只需要含有图片的原始评级 (不包括转发)。
3. 充分评估和清洗整个数据集需要巨大努力，所以只有一些问题 (至少 8 个质量问题和 2 个清洁度问题) 的子集需要进行评估和清洗。
4. 根据清洗数据的规则，清洗包括合并数据的独立内容。
5. 如果分子评级超过分母评级，不需要进行清洗。这个 特殊评级系统 是 WeRateDogs 人气度较高的主要原因。

本次经验：一定一定从观察数据开始， 即使没有理解题目，先看看各个数据有些什么，什么地方是重合的，什么地方可以相互关联起来，哪些数据奇怪但是对整体影响不大可直接丢掉

1. 花了太多精力和时间去办法清理text的内容，但是题目本身木的可能并不在此。 本次经验：多练习，多练习，找更多数据，更多类型的数据练习

具体步骤：

1. 下载和读取数据
2. 通过反复观察实验，或者对数据进行多次采样，了解基本的数据结构，内容，格式等
3. 找到最终会用到的数据后，寻找相关性，并保证相关性格式等一致（tweet_id）
4. 清洗无关数据，再合并保存
5. 观察最终数据，哪些数据能相关等，用可视化等检

注意：

一定注意修改Dataframe时取新名字，避免有错误要回头查看，数据不完整