数据清理简报:

数据的评估

评估方式：目测评估、编程评估

发现的问题及对应的处理方法:

质量

完整性：
表：tweet_archive:
Name 一列存在空缺值（None）
- 利用 sample 方法运行几遍查找有无需要更正的部分，有则更正，其余则当作空值。

expanded_urls 存在空缺
- 根据网址规则利用 tweet_id 补全更替

表：image_recognition:
记录数目比 tweet_archive 少了近 300 条
- 由于无法获取图像预测数据，无需补全

有效性：
表：tweet_archive:
Timestamp 和 retweeted_status_timestamp 末尾有+0000
- 去除+0000

Source 中仍存在 html 标签
- 去除 html 标签

expanded_urls 存在带有多余字符的网址，以及网址重复问题
- 前面已经处理，无需清理

Rating_denominator 存在着为 0 以及不为 10 的数
- 筛选出这些数据，按照推文修改

诸多列存在数据格式不正确的问题
- 更正格式

表：iamge_recongnition:
多个图片链接重复出现
- 根据该列丢弃重复项

tweet_id 数据类型不正确
- 更正格式

准确性:
表：tweet_archive:
Name 一列存在叫做 a、an、my、the、this、his、actually、incredibly、O、such、light、life 的。
- 筛选出来更正

表：addtional_info:
Favorite_count 中存在大量为 0 的数据
- 更换为空值

一致性:
表：tweet_archive:
rating_numerator、rating_denominator 两列存在着对一次多只狗狗评分的情况，导致评分分母大于 10 的情况。同时存在评分分母不为 10 的情况。标准不一致
-用这两列计算比例作为评分

整洁度

表：tweet_archive:
in_reply_to_status_id 和 retweeted_status_id、in_reply_to_user_id 和 retweeted_status_user_id 应当分别合并
- 将这些列合并

Doggo、floofer、pupper、puppo 都是狗狗的种类，应该为一个变量
- 将这些列合并

表：additonal_info:
此表的内容应属于 tweet_archive 的一部分，因此不应单独构成一张表
- 将其连接到 tweet_archive 表中