WeRateDogs 项目数据清理过程

项目概述：WeRateDogs 项目是对推特用户 @dog_rates 的宠物狗的数据进行处理分析，从而得出相关结论

项目的原始数据集分别是：推特档案 twitter_archive_enhanced.csv、Twitter API 的附加数据 tweet_json.txt、图像预测文件image_predictions.tsv

清理评估思路 ：先对三个数据集评估，用电子表格和预览观察数据，也用编程的方式评估数据，清洗数据后分析

其中需要处理的问题点有

数据质量问题

1. df_archive_clean 含有属于转发的推文条目
2. df_archive_clean 存在没有图片的推文条目
3. ID 应为字符串而不是整数，in_reply_to_status_id, in_reply_to_user_id 的数据量太少去掉
4. 时间戳的数据类型应为 datatime 非对象
5. 修改分子为浮点数
6. 分母不为10的评级为异常值，可以个别修改，但是其总数并不大所以可以去掉
7. 分子大于等于24的也多为异常值，结合分母的异常值一起去掉后，高效且不会影响整体分析因为异常值少
8. name 中的 None, a, an, the 不是 non-null

数据清洁度问题

1. 这三个数据集应该基于相同的 ID 进行合并
2. doggo, floofer, pupper, puppo 应该合并为一列

最后使用 pandas 和 matplotlib 库对数据进行分析