数据清洗与其他技术细节报告

数据集的问题：

经过观察，我们发现数据集存在这些问题：

质量问题

1. `dogrates_lite` 数据集中，`name` 列有大量空值，和错误的情况；
2. `dogrates_lite` 数据集中，部分数据错误的提取了其他包含“/”的文本作为分数，这些数据被保存在了 `tweets_mulpitple_number` 数据集中；
3. `dogrates_lite` 数据集中，有一行数据虽然包含数字，但是一条筹款的推文，并不包含评分；这条数据被保存在了 `tweets_fund_raise` 数据集中；
4. `dogrates_lite` 数据集中，部分数据存在多只狗狗统一打（总）分的情况，导致这些条目的分子和分母显著较高；这些数据储存在 `tweets_multiple_dogs` 中；
5. `dogrates_lite` 数据集中，部分数据存在推文主题真的是狗狗的情况下，有分子提取错误的情况；主要体现在这些分数为了某些纪念日等，使用了特殊的小数分数；这些数据储存在 `tweets_wrong_numerator_dog` 中；
6. `dogrates_lite` 数据集中，部分数据在主题可能不是狗的情况下，有分子提取错误的情况；这些数据储存在了 `tweets_wrong_numerator_NOT_dog` 数据集中；
7. `dogrates_lite` 数据集中，`timestamp` 列数据类型错误；
8. `dogrates_lite` 数据集中，有部分数据属于转发的推特，与原始数据重复；
9. `dogrates_lite` 数据集中，部分数据里只有一只狗，却因为text列中包含两个狗狗分类的信息，而拥有两个分类；这些数据被储存在了 `_1dog_2stage` 数据集中；
10. `dogrates_lite` 数据集中，部分数据，一条推特对两只处于不同生长阶段的打了同样的分数，因此拥有两个分类；这些数据被储存在了 `_2dogs_1tweet` 数据集中；
11. `retweets_lite` 数据集中，`id` 列应更名为 `tweets_id`，与 `dogrates_lite` 和 `breeds_lite` 保持一致；
12. `dogrates_lite` 数据集中，有部分数据包含两组正确的分数，但只提取了一组；这些数据被保存在了 `tweets_multiple_number_case2` 数据集中（`tweets_multiple_number` 中的数据不在此列）。

整洁度

1. `dogrates_lite`数据集中，`doggo``floofer``pupper``puppo`四列是一个变量的观察结果，应该被储存在一列中；
2. 包含转发和点赞信息的`retweets_lite`数据集和`dogrates_lite`数据集应当合并，因其观察的而对象是相同的。

数据清洗

针对质量问题的操作：

- 删除影响分析的数据
    - 删除了一条tweet_id为810984652412424192的推特，因其是一条筹款推文，并不包含这份报告的分析范围之内；
    - 删除了两条评分显著异常的推文，其中一条庆祝了美国独立日，另一条似乎是为名人；
    - 在验证转发数据不包含任何未知的信息后，我们抓住转发推特都带有“RT @”的特征删除了他们
- 补全空值
    - 通过优化的正则表达式重新提取了狗狗的名字：我们重做了name列，存在多只狗狗的推文，其name列值使用了&将两名字连接（可能影响了名字频率统计）；
- 修复错误
    - 修复语言表达造成的的一只狗狗对应多个分类问题（人工辨别，手动清理）
    - 修复推文中就是有两只不同分类的狗狗的问题：我们为这类推文单独建立了一个分类（使用&连接不同的分类名称），顺便解决了狗狗生长状态四列需要合并为一列的问题；
    - 修复推文中存在多个“/”导致分数提取错误的问题：我们提取了存在该项错误的行，使用专为这项任务涉及的正则表达式重新提取分数信息，并将这些分数信息更新回原数据集；
    - 修复小数点导致的分数提取错误的问题：显然在推特账号的发展过程当中，推主的打分标准发生过一些变化；
    - 修复推文中包含两组正确的分数，但只提取了一组的问题：我们使用一个新的正则表达式为这些行提取了这些分数，并将它们暂存在新的分数列中；
    - 修复了一条推文对多只狗狗打总分的情况：我们抛弃了原来的分数系统，直接计算了每一条推文所有分数的平均分作为推文的唯一分数；在修复这一问题的过程中，也顺便解决了上一条错误修复导致的一条推文有两个分数的问题；
    - 修复了dogrates数据集下`timestamp`列数据类型错误的问题：将其修正为`datatime`数据类型；
    - 统一了不同数据集中代表同一变量的不同列名称；
    - 修改了breeds_clean数据集中不适宜的名称。

针对整洁度问题的操作：

- 将标明狗狗生长状态的四列合并为一列（已在针对质量问题修复的过程中修复）
- 将项目涵盖的三个数据集合并，因其观察的对象本质上是一样的。

对数据集的进一步探索

在以上工作完成之后，我们确实得到了一个清洗干净的数据集，但它还不足以解答我们的问题：有没有因素能帮助我们确定，这条推文会更受欢迎？为此，我们在上述工作的基础之上，进一步提取了如下信息，以帮助我们进一步探索数据集：

- **通过人称和物主代词，利用正则表达式，从text列提取推文主体的性别信息**；这一步骤的执行效果似乎比网上其他同学的效果要好，有更少的遗漏，并通过人工辨别修复了一些错误；
- **启用全新变量"转赞比"**：通过计算转发/点赞的比例，得出某一条推文更深层次受欢迎程度；
    - 通过这项指标，一般的分析中我们也无需再看单独的转发数量；
    - 通过这一指标，我们能更好的衡量推特账号核心粉丝的变化情况，并为深入分析打下基础；
- **计算了推文点赞量和转发量的各项里程碑**，以帮助我们更好的理解推特账号的发展历程；
- **将评分分组**，以衡量不同分数之间的点赞量和转赞比情况；1分，7分和13分一定代表着三种不同的推文风格，其受欢迎程度的不同显然值得我们进一步探索；这项分析与回归分析的结论结合，可能能让我们有更进一步的思考；
- **将推文是否是狗进行分类**：我们注意到，图像预测机器学习的数据集里，对同样图像做了三次预测。我们发现，综合考虑其三次预测的结果，能够更好的说明推文的实质内容。我们根据图像预测的结果将推文分成了三大类：不是狗（三次预测结果全部为False），有可能是狗（三次预测结果有一次为True），和是狗（三次预测结果全部为True）。根据我们列举的例子，这三类实质上代表了不同类型的内容，也决定了推主在表达时的推文风格可能也有所不同，而这很可能能够帮助我们进一步不同推文的探索受欢迎程度的区别。

保存数据集

在对数据的全部操作完成之后，我们将主数据集按照要求保存为`twitter_archive_master.csv`

其他可能值得一提的技术细节

- 能够快速通过主键跨数据集筛选和调取数据的`ISIN`函数。其逻辑和可拓展性都大大强于常见于CSDN的利用`join`去数据集差集的方法；
- 绘制饼状图，并自动忽略指定比例下注解的`value_count`函数；
- 使用plotly绘制的可交互可视化；
- 使用seaborn绘制复杂的包含多个子图的可视化图像，和对可视化颜色的全方位主动控制。