数据清洗项目：WeRateDogs

一、数据收集：

1. 已经提供的数据：twitter-archive-enhanced.csv。
    - ✧ 直接使用 pandas 读取，命名为 **archive**。
2. 使用编程方法下载数据：twitter 图像的预测数据。
    - ✧ 使用 requests.get 获取网页文件，命名为 **images**。
3. 使用编程方法爬取每条 twitter 的额外数据。
    - ✧ 目前 tweepy 无法使用，直接用 pandas 读取 tweet_json.txt。命名为 **extra**。

二、数据评估：

写在前面：在这两千多条数据中，大多数 twitter 文本都只提到了一只狗。但有少数有两条或多条狗，这就对应了多个 name，多个特征，多个分数。如果要进行细致的评估和清洗，多个参数也是要分开的，但是这个过程过于冗杂，有时还需要逐条查阅文本内容。考虑到以上情况，为了方便起见，除了狗的名字（name）考虑了多个名字的情况，狗的特征和狗的分数都不再考虑存在多个的情况，而是只取第一个。

1. 目测评估

直接观察这三个数据表格并找出其中的质量问题和整洁度问题。

1.1 archive 的质量问题：

- in_reply_to_status_id 和 inply_to_user_id 两列有大量的空值（NaN）
- retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp 三列有大量的空值（NaN）
- rating_numerator 和 rating_denominator 这两列有有些值是明显错误的（如分子为 1 分母为 2）
- name 列有些是错误的，例如，a,an
- name, doggo, floofer, pupper, puppo 这四列中的 None 应该是 NaN

1.2 archive 的整洁度问题：

- 应该把这三个表格合并成一个表格
- source 这一列中的值完全相同，没有存在的必要
- expanded_urls 这一列没有存在的必要
- doggo, floofer, pupper, puppo 这四列可以合并成一列

1.3 images 的质量和整洁度问题：

- 狗的品种名字格式不统一

1.4 extra 的质量问题和整洁度问题：没有问题

2. 编程评估

2.1 archive 的质量问题：

- timestamp 列的数据类型
- tweet_id，in_reply...,retweeted...共 5 列应该是 string 类型
- 狗的分数存在错误

2.2 images 的质量问题和整洁度问题：

- tweet_id 列数据类型应该是 string

2.3 extra 的质量问题和整洁度问题：没有明显问题

三、数据清洗

1. archive 的 in_reply 和 retweeted 中有大量的空值

需要查看一下这几列是因为什么原因导致的空值这么多。

经过查看，in_reply 不为空的行表示这条 twitter 是回复给其他人的， retweet 不为空的行表示这条 twitter 是转发的，而其他的都是原创的。**在回复和转发中也有很多条是对狗的评分，所以不能删除。**

但是这 5 列空值太多，而且对于数据分析有没有用处，使用 pandas 的 drop 方法删掉。

2. tweet_id 列应该是 string 类型：使用 astype 方法

这里出现了一个奇怪的现象：原来的 tweet_id 是 int，如果这里不转换成 string 而直接进行下面的清洗，到最后这一列变成了 float 格式，并且是以科学计数法表示的。我不清楚怎么再转换成 string 了（因为转成 string 还是科学计数法形式），所以这里先转换成 string。

3. name 列有一些是错误的

重新提取狗的名字：首字母大写，而且前面是“This is”, “name is”,“named”,“hello to”, 以及“Meet”。用这个规则，使用 extract 方法来提取狗的名字。另外有些是两个名字，也做了处理。

经过清洗，name 这一列中仍有 797 个空值，这些是因为 text 中没有提到狗的名字。

4. 狗的特征一个变量占用了四格

定义：使用 melt 方法将特征整理成一列。需要注意：

- ✧ 应该用 NaN 代替原来的 None

- ✧ 有几个狗是有两个特征的，为了方便处理，只保留第一个。

5. 对于狗的评分的提取

定义：狗的评分有一些错误的地方。重新从 text 中提取分数.

- ✧ 使用 extract 方法
- ✧ 提取的规则是: (?:\d\.)?\d{1,3})/[1-9]{1,2}0，即分子是 1-3 位的数字(可以是小数），分母是以 0 结尾的 1-3 的数字（但不能是 0）
- ✧ 然后删除原来的分数列
- ✧ 再把重新提取的分数列 merge 上去
- ✧ 把分数列的数据类型改成 int
- ✧ 最后再把分母不是 10 的行转换为分母为 10 的分数

6. 使用 drop 方法删除无用的 source 和 expanded_urls 这两列。

7.使用 to_datetime 方法修改 timestamp 为 timedate 类型

8. 使用 astype 方法需改 imges 表格的 tweet_id 为 string 类型

9. images 表格狗的名字格式不统一

统一成：中间没有“_”，且首字母大写(由于只用到 p1 列，所以只清洗 p1 列）。分别使用 replace 方法和 title 方法。

10.使用 merge 方法将 archive，images 和 extra 表格合并成一个表格，为了排除无图片的内容，merge 全部采用 inner 方式。最终的数据为 1971 行，20 列。