Wrangle report

数据收集：

twitter-archive-enhanced.csv：直接下载用 pd.read_csv 导入
image_predictions.tsv：用 requests 包和文件的 url 来写入文件然后利用 pd.read_csv 读取导入
tweet_json.txt：利用 Twitter API 获取，得到文件并读取，这里需要将自己需要的列提取出来放入 dataframe 里。

数据评估：

一共有 2356 行数据。利用.describe()函数和.info()函数查看该 dataframe 的基本情况。发现一些不合理的地方：

1. Tweet ID#835246439529840640 评分中的分母为 0，需要修正
2. Tweet ID# 835152434251116546, 和 Tweet ID# 746906459439529985 评分中没有分子
3. twitimestamp 变量类型错误
4. dog name 有问题，很多狗的名字为 None，如果没有名字应该修正为 NaN
5. 还有些狗的名字是 a，猜测是从 Twitter 中抓取狗的名字时位置出错
6. 数据包括了转发的 Twitter，我们需要移除

image_predictions：
没有什么问题，可以考虑将 tweet_id 类型从 int 修改为 str
tweet_json：
变量 lang 命名不规范，并且类型不合适
整洁度：
Dog “stage”有四列，很不必要，可以考虑将四列合为一列
Rating 也有两列分别是分子和分母，可以考虑将其合并

数据清洗：

1. 将三组数据分别复制并命名为_clean,然后利用 merge 函数将三组数据合并到 twitter_archive_clean 中，合并依靠数据之间 tweet_id 链接。
2. 检查新的 twitter_archive_clean，利用.drop_duplicates()去除重复的行。
3. 利用 isnull()函数讲非转发的 Twitter 保留，转发的 Twitter 删除。
4. 将不含图片的行删除
5. 与转发相关的'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'删除以使数据更加整洁。
6. 利用.str.extract()函数从 text 中提取出 dog stage 信息创建新的列 dog_stage，并将'doggo', 'floofer', 'pupper', 'puppo'删除
7. 创造新的列 rating，利用分子分母计算 rating，并将 'rating_numerator','rating_denominator'删除

8. 修改变量“lang”为“language
9. 修改列宽度属性，以使得每列都能完整地呈现
10. 从 text 重新抓取 name 并将其赋给 name 变量
11. 将 source 中的 url 去除使得 source 更易理解
12. 将"source", "timestamp", "dog_stage", "tweet_id", "in_reply_to_status_id", "in_reply_to_user_id"这些变量类型修改为更合适的
13. 将清洗后的数据一用.to_csv()存到 twitter_archive_master.csv