Wrange_report

收集数据：
从三个来源收集数据：
1. 从 Github repo 中通过 Python request 库获得
twitter-archive-enhanced.csv 和 image-predictions.tsv 两份文件存到本地
2. 通过 Tweepy 获得扩展字段：目的是获得转推数和喜爱数：
    a. 通过 api.user_timeline（限制数量为 20）循环抓取 json 数据，该方法有上限 3K2 条，最终只获得最新的 3K2 条推特的完整 json 数据保存为'all_tweets_3k2_json.txt'，事实上该文件在后期数据评估、清洗中并未用到
    b. 通过第一步获得的 “'twitter-archive-enhanced.csv'” 文件，拿到 tweet_id，借助 api.statuses_lookup（限制数为 100）获得对应的推特的完整 json 数据，保存为'tweet_json.txt'
3. 通过 Python I/O 将之前的三份文档：
    twitter-archive-enhanced.csv
    image-predictions.tsv
    tweet_json.txt
    导入 Python，以备评估使用

评估数据
整洁度
1. tweet_json.txt 中需要的数据有 tweet_id,转推数 retweet_count，喜爱数 favorite_count
2. archive 表中，doggo/floofer/pupper/puppo 应为一列'stage '
3. archive 表中有用的列 :tweet_id, timestamp, text, rating_numerator, rating_denominator, name，stage（需处理）
4. image 表中， p1_dog/p2_dog/p3_dog 中第一个为 true 的所对应的 p 应作为 p_species（狗狗的品种）
5. image 表中有用的列：tweet_id，p_species（需处理）
6. 三张整理后的表格应该为一张表格

##清理数据
整洁度
1. Tweet_json.txt 中有每个 tweet 的全部基本信息，通过切分、正则表达式，分别截取 tweet_id,转推数 retweet_count，喜爱数 favorite_count，并将这三项输出到文件 'df_json.csv'
2. Archive 表中，将 doggo/floofer/pupper/puppo 为 None 的列替换为空后拼接，获得 stage 属性
3. 删除 archive 表中无用的列
4. Image 表中，循环查看 p1_dog/p2_dog/p3_dog 中第一个为 true 对应的 p 列名称作为狗狗的种类，如果三者都为 False 则标记品种为 Unknown
5. 删除 image 表中无用的列
6. 从 archive 表开始，根据 tweet_id，依次合并 image 表和 json 表，获得文件 “good_dogs.csv”

评估数据

Quality

- index=385,评分 24/7 是 7 天 24 小时，该狗狗没有评分，但路人给出评分 11 可作为参考
- index=800,评分 9/11 是幸存数，实际评分为 14
- index=891,正文中两组分数，实际评分为 13;index=1973,实际评分为 9
- index=1328,11/7 是便利店 7-11，实际评分为 10
- 其他评分基线>=40 数据，是狗狗的打包评分。评分需要改成狗狗们的平均分
- 修正评分后，评分基线统一改为 10
- tweet_id 数据类型应为 string
- timestamp 数据类型应为 datatime
- p_species 字段中写法不统一，有大写、有小写、有首字母大写，还有下划线、减号
- name 字段中，狗狗名字除了 None(546),还有 a(55),an(6),the(7),统一改为 Unknown

清理数据

Quality

- index=385,rating_numerator 为 11
- index=800，rating_numerator 为 14
- index=891，rating_numerator 为 13
- index=1973,rating_numerator 为 9
- index=1328,rating_numerator 为 10
- 打包评分改为平均分：rating_denominator>=40 的项，rating_numerator==round(rating_numerator*10/rating_denominator) #四舍五入
- 评分基线全部改为 10，令所有 rating_denominator==10
- tweet_id 数据类型改为 string
- timestamp 数据类型改为 datatime
- p_species 字段中改为全部小写，减号、空格替换为下划线
- name 字段中，狗狗名字除了 None(546),还有 a(55),an(6),the(7),统一改为 Unknown

最终获得清理后的数据 “twitter_archive_master.csv”