数据整理过程：

数据收集

1. 先获得 WeRateDogs 的推特档案，这是手头文件： twitter-archive-enhanced.csv，直接加载进工作区生成 df1 数据集。
2. 再获得推特图像的预测数据，从 url ：https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/image-predictions.tsv，运用 request()函数将网页文本下载到 image_predicted/images_predicted.tsv 文本文件中，在加载到 df2 数据集中。
3. 最后获得推特的额外附加数据，包括转发数、喜欢数，由于无法注册 tweet 账号，所以直接下载的 json 结构的 txt 文件，然后把 txt 格式运用 json.load()函数加载成 json 数据导入 df3 数据集中。

评估

df1：

1. 包含了没有图片或非原始评级的推特。
2. timestamp 列的数据类型错误。
3. tweet_id 数据类型错误，应该是字符串类型，而不是整型。
4. 狗狗的名字中有一半为 None 和 a,an,the 等无效字符串。
5. 狗狗的评分列没有计算出来，且分子分母异常：有的分子是小数但只取了小数点后面的数字、分子分母没简化、分子分母取值异常。
6. 太多无用的空值列。

df2:

1. df2 数据集中 jpg_url 列中有重复

df3：

1. df3 中 favourite_count 列有的行数据为零。

整洁问题：

1. df1 中 doggo,floofer,pupper,puppo 四列可以变成一列 stage。
2. 三个数据片段都是以 tweet_id 为观察单元，应该在同一表中。

清理

缺失数据：

1. 使用 .str.extract 和正则表达式，从 text 中重新提取评分，来完成评分的清理。

2. Name 中 none 不是有效值，为了标识缺失值，将所有的无效的名字用 np.nan 来替代。

清洁度:

1. 将四列中的 'None' 替换为空字符串 ''，然后将四列做向量加法，连接到一起组成 stage。
2. 用 merge()函数将三个数据集合并

质量:

1. timestamp 应该是时间数据类型，而不是对象，用 pd.to_datatime()函数转换。
2. 用 astype()函数把 tweet_id 列数据类型转换成字符串类型。
3. 选出 retweeted_status_id,retweeted_status_user_id 和 retweeted_status_timestamp 这三列为空值的行就可以删除所有转发条目；无图片的推特需要删除，对图片预测数据集进行 merge 时选择 inner 方式，这样可以删掉没有图片的推文条目；
4. 删除转发的推特后，这些问题也相应的解决了。
5. 无用的列可以直接删除