WeRateDogs 的推特档案包括了5000 多条推特的基本信息。 档案中有一列包含每个推特的文本，用这一列数据提取了评分、狗的名字和“地位”。

推特json部分，我们根据tweet_id提取了转发数和喜爱数。

图像预测部分根据了图片进行了狗狗种类的预测，并给出了预测的可信度。

我们分别从本地读取推特档案csv、API收集转发数和喜爱数、网上下载图像预测tsv文件。然后将上述三个表格根据tweet_id进行合并，成一个完整的数据框。

数据评估流程

1. 查看数据格式是否合适，比如tweet_id应为int64
2. 查看数据是否重复
3. 挨个查看每一列的数据有否异常，并记录问题供后续清洗
    1. 查看twitter发送时间是否连续
    2. 查看text是否重复
    3. 查看评分分子是否过大或者过小，是否单位一致
    4. 查看评分分母是否不为10
    5. 查看姓名列是否提取正确
    6. 查看转发数和喜爱数有没有异常值
    7. 查看狗狗的doggo、floofer、pupper、puppo四列是否完整，和满足独热编码的要求
    8. 查看一下预测文件的性质，比如可信度的特征、有多少不是狗

整理出以下问题：

> 1. **质量问题1**：
>   1. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id 的数据格式应该是int64
>   2. 存在转发的条目需要删除
>   3. 存在无图片的推特需要删除
>   4. 分子出现极端值 1776
>   5. 分子单位不一致，即不为 1/10
>   6. 分数提取不全正确
>   7. 部分名字提取不正确，有 765个"None" 和 67个"a"，,还有一个叫'space'，不是大写，可能有问题
>   8. 根据项目介绍，狗狗的stage不全，还有"blep", "snoot" ,"BlepiPen"
>   9. stage提取出错：
>   有些时候n_stage为0；
>
>   还有些情况，存在狗狗有两个stage，查看后，发现这种情况对应的情况包括：图片中有2只狗狗、作者进行比较的时候提及了另一stage等。
>   10. 漏将floof归类为floofer
>   11. df_twt_json表格的tweet_id 存在重复
>
> **整洁问题1**：
>
>   1. 不应该将狗的stage分成doggo、floofer、pupper、puppo四列，以独热编码方式记录，应该记为一列‘stage’，而为真的原列名为内容。
>   2. 需要将三个表合为一张表。