收集数据

- WeRateDogs 的推特档案数据获取
  - twitter-archive-enhanced.csv
- 推特图像的预测数据
  - image-predictions.tsv
- 每条推特的额附加数据
  - tweet_json

对项目数据进行评估

质量

twitter_archive_enhanced 表格

1. tweet_id 应为 str 类型，而不是 int
2. 缺失值过多的列删除，in_reply_to_status_id，in_reply_to_user_id,retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp
3. 转发的信息的列应该删除 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp',
4. 数据部分缺失，expanded_urls 数据量为 2297
5. timestamp 的数据类型应该化成标准的时间格式，应进行数据转化
6. 评分分母不全为 10，可从 text 中重新提取
7. source 保留着 html 标签,应去除 html 标签
8. 狗狗名字存在缺失值,而且名字“a”,"an"应该不为狗狗名字，应该是信息提取错误

image_predictions 表格

1. image_predictions 表中的图片 url 存在重复，需要删除
2. tweet_id 列的数据类型不正确。

清洁度

- doggo、floofer、pupper、puppo 四列可融合成一列；
- 应该将 image_predictions 与 tweet_json 这两个表格应该与 twitter_archive_enhanced 合并，通过 tweet id 将 3 个表格合并为一个表格。

清理

1. tweet_id 应为 str 类型，而不是 int，进行数据转化

2. 缺失值过多的列删除，in_reply_to_status_id，in_reply_to_user_id,retweeted_status_id，
3. 转发的信息的列应该删除 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp',
4. 数据部分缺失，expanded_urls 数据量为 2297，进行去重处理
5. timestamp 的数据类型应该化成标准的时间格式，应进行数据转化
6. 评分分母不全为 10，可从 text 中重新提取，正则化提取
7. source 保留着 html 标签,应去除 html 标签，正则化提取
8. 狗狗名字存在缺失值,而且名字“a”,"an"应该不为狗狗名字，应该是信息提取错误，重新用正则化提取
9. image_predictions 表中的图片 url 存在重复，需要删除，去重处理
10. tweet_id 列的数据类型不正确，转化为 str 类型；
11. doggo、floofer、pupper、puppo 四列可融合成一列，融合后将这四个列删除
12. 应该将 image_predictions 与 tweet_json 这两个表格应该与 twitter_archive_enhanced 合并，通过 tweet id 将 3 个表格合并为一个表格。