数据清洗：

质量

twitter_archive_enhanced 表格

1、 不包括转发数据(retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp，删除是转发的数据行)

处理：提取 retweeted_Status_Id 为空的数据

2、 timestamp 错误的数据类型

处理：用 to_datetime 转换成时间类型

3、 本实验中 source 不具有分析意义

处理：用 drop 删除

4、 本实验中不具有分析意义，清理列（retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp、in_reply_to_status_id 、in_reply_to_user_id 、expanded_urls,expanded_urls）

处理：用 drop 删除

5、 缺少狗的名称、评分和评级（无法清理）

6、 name 数据有误

处理：从 tweet_json.txt 提取名字信息替换

7、 rating_numerator、rating_denominator 数据有误

处理：从 tweet_json.txt 提取评分信息替换

8、 评级列数据有误

处理：从 tweet_json.txt 提取地位信息替换

9、 错误的数据类型，status(地位)应该是 catogary

处理：用.astype('category')修改数据类型

10、 需要删除无图片的推特

处理：将原数据集和包含图片信息的 image-predictions.tsv 数据集内链接，就剔除不包含图片信息的数据。

清洁度

1、 twitter_archive_enhanced 表格缺少喜爱和转发数两列

处理：从 tweet_json.txt 提取 retweet_count 和 favorite_count 数据添加到原数据集中。

2、 twitter_archive_enhanced 表格中表示地位的四列（doggo、floofer、pupper、puppo）表示同一个变量”地位 “

处理：用 melt 将四列合并成一列数据再去重。

3、 对于狗狗品种的猜测和喜爱数，转发数应该都属于 twitter_archive_enhanced 表格的一部分

处理：合并数据集