

# WeRateDogs 推特数据清洗报告

## 一、收集

本项目有 3 个数据来源：

**tweets** 表格：WeRateDogs 的推特档案 `twitter_archive_enhanced.csv`

**json** 表格：推特图像的预测数据 `image-predictions.tsv`

**predictions** 表格：每条推特的额外附加数据（Json 格式）`json.txt`

## 二、评估

经评估，上述三组数据有以下问题：

### 1. 质量

**tweets 表格**

- `tweet_id` 是整型，而不是字符串
- 有 281 条推特记录中的 `tweet_id` 在 **predictions** 表格中不存在。
- 有 4 条推特记录中的 `tweet_id` 在 **json** 表格中不存在。
- 含有 181 条转发数据
- 分子有很多超过 20 的值
- 分母有很多不是 10
- 狗狗姓名提取不完整、不准确
- `in_reply_to_status_id`、`in_reply_to_user_id`、`source`、`expanded_urls` 列与分析内容不太相关

**json 表格**

- `id_str` 是整型，而不是字符串
- 含有 177 条转发数据
- `contributors`、`coordinates` 和 `geo` 列为空
- `favorited`、`retweeted` 和 `truncated` 三列的值均为 false

**predictions 表格**

- `tweet_id` 是整型，而不是字符串

### 2. 清洁度

- `tweets` 表格中的 *doggo*、*floofer*、*pupper* 和 *puppo* 四列应合并为一个变量
- `tweets` 表格、`json` 表格和 `predictions` 表格 应合并为一个表格

### 三、清洗

本次数据清洗时，处理清洁度与质量问题交替进行，先解决了 `tweets` 表格的清洁度问题，然后删除了各表格中的无用、重复数据以免合并表格时出问题，在处理完 3 个表格的质量问题后，对 3 个表格进行合并，得到 1994 条数据。

具体过程如下：

1. `tweets` 表格中的 *doggo*、*floofer*、*pupper* 和 *puppo* 四列应合并为一个变量：狗狗地位 `stage`。
2. 删除 `tweets` 表格的 181 条转发数据，即只保留 *retweeted\_status\_id* 为空值的行
3. 根据推文用正则表达式重新提取狗狗评分（分子和分母）及狗狗姓名。
4. 删除 `json` 表格的 177 条转发数据，即只保留 *retweeted\_status* 为空值的行，并只保留 *id*、*favorite\_count*、*retweeted\_status* 列的数据。
5. 用 `inner join` 的方式合并 3 个表格。
6. 将 *tweet\_id* 转为字符串，将 *timestamp* 转为日期格式。
7. 将合并表格保存为 `twitter_archive_master.csv` 文件。