**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，即有281条数据没有图片信息。
* 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，只保留包含包含图片的原始评级。

#### 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。