# WeRateDogs数据整理报告

### 1. 收集数据

先导入WeRateDogs 的推特档案csv文件，然后从互联网下载推特图像的预测数据tsv文件，最后再导入每条推特的额外附加数据txt文件。

### 2. 评估数据

通过目测评估和.head()，.info()等编程方式评估3个表格的整体信息。然后分别对3个表格进行评估。

2.1 `archive_enhanced` 表格：

1. 通过.info()函数观察发现错误的数据类型(timestamp, twitter_id列)；
2. 大量的数据缺失（in*reply*to*status*id, in*reply*to*user*id, retweeted*status*user*id,* *retweeted*status_timestamp列）；
3. 查看rating*numerator*列和*rating*denominator列的数值分布；
4. 查看name列的数值分布，并与表格中的text对比，得出name列数据缺失，且与text列中的信息不符；

2.2 `image_predictions` 表格：

1. 查看jpg_url列的重复值

### 3. 清理数据

3.1**质量**

`archive_enhanced` 表格

- 错误的数据类型（timestamp列）：修改数据类型，删除多余的字符串；
- 大量数据缺失（in*reply*to*status*id, in*reply*to*user*id, retweeted*status*user*id,* *retweeted*status_timestamp列）：删除多余数据列；
- rating_denominator列数据不完整，分母有不是10的数据：列数值范围错误，重新在*text*中提取；
- rating_numerator列有些行数据提取错误，与text列中的信息不符：重新在*text*中提取；
- 狗狗名字(name列)数据缺失，且与text列中的信息不符：重新提取狗狗名字；
- 数据集中转发的Twitter信息无效：删除转发信息行；
- 无图片的Twitter信息无效：删除无图片的*Twitter*信息；
- tweet_id数据类型错误：修改数据类型；

`image_predictions` 表格

- jpg_url列有大量重复值：删除重复值；

3.2**整洁度**

- `archive_enhanced` 表格中的doggo，floofer，pupper，puppo 列可合并为一个变量
- `archive_enhanced`、`image_predictions`、`tweet` 三个表格可以合并