数据整理报告

数据收集

该项目一共有3个数据文件：

1. WeRateDogs 的推特档案，该文件可以通过直接下载的方式获取到，保存为 `twitter-archive-enhanced.csv`；
2. 推特图像的预测数据，该文件可以使用 Python 的 Requests 库和 URL 来进行编程下载，保存为 `image-predictions.tsv`；
3. 每条推特的`retweet_count`，`favorite_count`，该数据可以通过 Tweet API 来获取得到，保存为 `tweet_json.txt`；

将以上3个文件分别保存到项目根目录。

数据评估

将收集来的3个文件分别读入到3个 Pandas DataFrame 里，并分别观察 DataFrame 里存在的问题，包括质量和清洁度两个方面的问题，整理如下：

质量

1. `twitter-archive-enhanced.csv`: `rating_denominator` 有些是 10，有些不是 10
2. `twitter-archive-enhanced.csv`: `rating_numerator` 缺少小数点精度
3. `twitter-archive-enhanced.csv`: `name` 大小写不统一
4. `twitter-archive-enhanced.csv`: `name=None` 未识别成 NaN
5. `twitter-archive-enhanced.csv`: `tweet_id` 类型是 int，应该是 str
6. `twitter-archive-enhanced.csv`: 有些狗狗存在两个stage（如 tweet_id=854010172552949760）
7. `twitter-archive-enhanced.csv`: 狗狗分类 `None` 未识别成 NaN
8. `tweet_json.txt`: `display_text_range` 类型是 list，应该是 int

清洁度

1. 狗狗的 `Stage` 用了四列来表示（doggo, floofer, pupper, puppo）
2. 三个数据集都是以 `tweet_id` 为观察单位，需要合并成一个表

数据整理

清理过程比较简单，按照以下给出的模板来清理：

1-定义

2-代码

3-测试

最终清理干净的数据，保存到项目根目录：`cleaned_tweets.csv`。到此位置，数据收集，数据评估和数据整理的工作就已经完成了👏🏻。