WeRateDogs 推特数据清洗与分析

本项目对 WeRateDogs 的推特档案的数据进行清洗并进行探索分析。项目主要步骤如下：

一. 文件收集

需要收集的文件有三个：

- 手头文件 `twitter_archive_enhanced.csv`，其中包含了一些主要的推特信息，是本次清洗的主要数据，其中的评分、地位和名字等数据是从 `text` 原文中提取的，但是提取的并不好，需清洗并清洗
- 编程下载收集互联网文件：`image-predictions.tsv`，其中包含了推特图像预测信息，根据推特中的图片预测出狗狗种类
- 推特转发数（`retweet_count`）和喜欢数（`favorite_count`）等额外推特信息，可以通过查询 API 收集，但由于无法访问 Twitter ，本项目直接使用的项目可供下载的 `tweet_json.txt` 文件，从中提取出转发数和喜欢数等所需数据。

二. 评估

通过目测评估和编程评估，发现需要清洗的数据质量及整洁度问题如下：

质量

- `df1` 表格

– 从‘in_reply_to_status_id’等列可以看出，有 78 行数据是评论，不是原始评级

– 从’retweeted_status_id‘等列可以看出，有 181 行数据是转发，不是原始评级

– expanded_url 列有缺失值

– 评分分子中有异常值（0、1、2、3、4、5、6、7、8、9）

– 评分分母中有异常值（11、50、80、20 等）

– 狗狗名字列有缺失值

– 狗狗名字列有错误值（a,an,the,such 等）

- `df2` 表格 `df3` 表格

– 表格 df1 有 2356 行数据，表格 df3 有 2352 行数据，表格 df2 只有 2075 行数据

整洁度

– 表格 df1：狗狗地位，一个变量表示成了 4 列（doggo、floofer、pupper、puppo）

– 表格 df3：twitter id 列名称为 id,其他两个表格此列名称为 tweet_id

– 只需一个表格，而当前有三个表格，需将三个表格合并

三. 数据清洗

使用代码逐一对评估发现的问题进行清洗，并最终生成名为’twitter_archive_enhanced_master’的数据集，供后续探索分析使用

四. 探索分析

探索的三个问题为：

- 最人气的狗狗名字前十名是什么
- favorite_count 与 retweet_count 是否相关
- 在有狗狗地位的数据中，评分前三的狗狗都是什么地位?

探索结果：

- 最人气的名字前十分别是 Charlie,Lucy,Cooper,Oliver,Tucker,Sadie,Winston,Penny,Daisy,Bo
- favorite_count 与 retweet_count 正相关
- 在有狗狗地位的数据中，评分最高的前三名评分分子分别为 27，14，14；对应的狗狗地位分别为 pupper，doggo,pupper