数据整理过程

> 创建一个 300-600 字的书面报告，命名为 wrangle_report.pdf，在该报告中简要描述你的数据整理过程。这份报告可以看作是一份内部文档，供你的团队成员查看交流。

01 收集数据

- 以编程方式下载推特图像的预测数据，命名为 `image` 表
- 收集 WeRateDogs 的推特档案，该部分数据由课程组提供，命名为 `archive` 表
- 通过 API 构建文件：每条推特的额外附加数据，命名为 `tweet` 表
    - 不会使用 API 下载，借助论坛帖子，直接获取了数据。（占个坑，晚点探索 API 构建文件）。论坛帖子；数据下载来源

02 评估数据

使用目测评估和编程评估的方式，对数据的质量和整洁度进行评估，运用以下函数

- .info()
- .value_counts()
- .head()
- .tail()
- .describe()
- .count()

得到以下待清理的问题

质量

1. `archive` 表中，in*reply*to*status*id、in*reply*to*user*id、retweeted*status*id、retweeted*status*user*id*、*retweeted*status_timestamp 缺失数据
2. `archive` 表中，timestamp 列有 +0000，多余
3. `archive` 表中，source 列数据意义不大
4. `archive` 表中，745 条记录的名字为 None，55 条记录的名字为 a，数据缺失
5. `image` 表中，tweet_id 应为字符串，而不是 int 整数
6. `image` 表中，部分 jpg_url 相同
7. `tweet` 表中，favorites、retweets 数据类型应为整数
8. `archive` 表中，rating*denominator*、*rating*numerator 列部分数值异常，甚至出现三位数、四位数

清洁度

1. 以 tweet_id 为准，合并三个表格（根据tidy data的第3项要求：观察单位按表格组织（即：一个种类的观察形成一个单独的表格），这个项目里用到的数据集的主要观察是“对狗狗照片进行评分”，根据这个观察主题，而3个表格的变量都是围绕“对狗狗照片进行评分”这个主题的，所以将3个表格里的相关变量合并到一个表格就更符合tidy data的第3项说明。）
2. `archive` 表中，doggo、floofer、pupper、puppo 四列需合并，因为 doggo、floofer、pupper、puppo 四列代表狗狗等级，属于同一变量，应集中在一列中

03 清洗数据

运用编程方式，对上述质量和整洁度问题进行清洗，用到以下函数

- .copy()
- .astype()
- .drop()
- .drop_duplicates()
- .extract()
- .replace()
- .melt()
- .merge()

04 分析和可视化

对整理后的数据进行分析和可视化，应用的函数为

- .value_counts()
- .sort_values()
- .groupby()
- .mean()

CHANGELOG

- 190325 创建