Wrangle_report

收集

数据集共包含三方面的数据

- WeRateDogs 的推特档案所包括的 5000 多条推特的基本信息，通过 pandas 的 read_csv 方法储存于 df 中
- JSON 格式的 tweeter 附加数据，使用 json 库读取 id，fav_count, retweet_count 三列，储存于 tweet_df 中
- 使用 request 库从所提供的 url 中编程下载的 tsv 文件，通过 pandas 的 read_csv 方法储存于 image_df 中

评估

质量

df 表格

- 可视化评估发现 rating_denominator 和 rating_nominator 并非均小于等于 10，也就是说评级并不是 10 分制。从原文可以发现部分评级是小数，而 dataframe 中的评级类型为 int，因此评级不完成正确
- 可视化评估发现 doggo/floofer/pupper/puppo 分类中存在大量 None 值
- 可视化评估发现 text 中 floofer 类的狗狗有时也被称为 floof，但 floofer 列中未将其纳入
- 编程评估发现 source 仅分为 iphone/web client/tweetdeck 三类，但原始数据杂乱
- 编程评估发现 expanded url 有 137 个重复值
- 编程评估发现共有 181 行 retweet 和 78 行 reply，因数据集只需要原始评级，故可以删除
- 编程评估发现 name 列有姓名错误，如‘a’，‘the’，‘an’
- 编程评估发现 timestamp 不是时间日期格式

image_df 表格

- 编程评估发现 img_url 有 66 个重复值
- 编程评估发现共 2075 行数据，少于 df 的 2356 条，存在缺失值
- 可视化评估发现每条 tweet 配有 4 张图，即 img_num 有 4 个，但仅有 P1-P3 的预测值***(无法清理)***

整洁度

- 可视化评估发现 doggo/floofer/pupper/puppo 分类不应分为 4 列，而应归为一列 stage
- 可视化评估发现只需要与 img_num 相关的预测值，其他列可以清除
- 3 个表格可以根据 tweeter id 进行合并

清洗

1. 根据 id 合并 3 个 dataframe，新表命名为 df_clean。
2. 我们只需要含有图片的原始评级 (不包括转发)，因此删除 jpg_url 为空的行、删除 retweeted_status_id 不为空的行。
3. 此时 reply 仅有 23 行有数据，再次通过可视化评估检查这些 reply tweets，发现大部分是对原推中评级的修改，为避免一只狗有多行数据，删除 reply 相关的行。
4. 删除无用的列，包括'id', 'in_reply_to_status_id','in_reply_to_user_id', 'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'。
5. 根据 img_num 来选择相应的预测值，若 pre_dog 为 4，则赋空值。将预测结果分别储存于 breed，conf，pre_dog 三列中，分别表示预测的品种、预测的可信度、以及预测该图片是否属于狗。删除 'img_num','p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog'列。
6. 从 text 中用正则表达式来匹配 puppo,pupper,floofer/floof,doggo，并将匹配结果储存于 stage 列，删除原'doggo','floofer','pupper','puppo'四列。
7. 从 text 中根据正则表达式重新提取评级的分子和分母， 由于存在分子大于分母的情况，单独比较分子或分母不利于判断评级，我们用分子除以分母，结果大于等于 1 即视为满分，并将该结果储存于 rating 中。
8. 把 source 列的三个取值替换为可读性更高的 iPhone，Web Client，TweetDeck。
9. 修改错误的数据类型：fav_count，retweet_count 修改为 int、timestamp 修改为 datetime、source, stage 修改为 category。
10. 使用正则表达式，在 text 列中根据常见的语法重新为宠物名赋值：

    This is (name)
    Meet (name)
    Named (name)
    Say hello to (name)
    为没有注明宠物名字赋空值
    为避免错取关键字 a、the，只选择以大写字母开头的名字。

储存

将清洗完成的表格储存于 csv 文件中，命名为 twitter_archive_master.csv。