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

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
• image_df表格只需保留最佳预测结果
• 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replytweets，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. 保留p1预测结果，如果p1_dog为false，则继续查看p2,p3预测结果。如果3个预测结果都不为狗，则保留NaN值。将预测结果分别储存于breed，conf列中，分别表示预测的品种和预测的可信度。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。