wrangle_report

缺少数据

tweet_data 表格中缺少记录，在程序中将有异常的 tweet_id 都存放在 error_tweet_id.txt 文件,将此文件内容导入 DataFrame 中去,并合并到 tweet_data 中去。

清洁度

1.twitter 表格中 doggo,floofer,pupper,puppo 应该用狗的 stage 来表示

由于这几列的值存在很多不正确的情况，所以在此直接删除这几列，并新增 stage 一列，在后续的质量问题中再进行解决，写入正确的狗的“地位”。

2.tweet_data 表格中的数据应该整合到 twitter 表格中，属于 twitter 表格的一部分

利用 merge 方法按照 tweet_id 将 tweet_data 表格中的数据合并到 twitter 表格中。

3.twitter 表格中应该去掉 retweeted_status_id 不为空的行

利用 drop 函数去除 retweeted_status_id 不为 NaN 的行。

质量

1.错误的数据类型（timestamp, retweeted_status_timestamp, tweet_id)

使用 pd.to_datetime 方法将 timestamp 和 retweeted_status_timestamp 转为 datetime 数据类型，tweet_id 转为 str。

2.评级存在不正确的情况，如 1663 行；

3.rating_denominator 数据有的不是 10；

4.rating_denominator 有数据等于 0。

由于 2，3，4 都属于狗的评级问题，因此放在一起处理。使用正则表达式和 pandas 的 str.extract 方法从 text 一列中提取评级分数，在这里统一采用 10 分制，即分母为 10，其他分数都转为分母为 10 的情况。

5.存在不正确的 name 数据，例如 index 等于 22 的一行里 name 的值为 such

使用正则表达式和 pandas 的 str.extract 方法从 text 一列中提取 name,对于图中存在两个名字的情况，使用’,’符号来连接，如：name1,name2。

6.doggo,floofer,pupper,puppo 存在数据不正确的情况

使用正则表达式和 pandas 的 str.extract 方法 text 一列中提取 doggo, floofer,

pupper, puppo，保存到 stage。

7.text 中存在网址链接 url

使用正则表达式和 pandas 的 repalce 方法匹配 url，然后删除它。

8.对于空值应采用相同的表示（存在 NaN 与 None 两种方式）

此问题在上面的清理过程中已经得到解决了。

9.p1,p2,p3 中狗的名字用空格来连接而不是下划线_

使用 repalce 方法，将 p1,p2,p3 中名字的下划线用空格来替代。

10.p1,p2,p3 中狗名字首字母要大写，保持一致的格式

使用 title 方法，将 p1,p2,p3 中名字的每个字的首字母大写。

再次清洁度

1.twitter 表格中 rating_numerator 和 rating_denominator 应该合并到一起，用 rating 表示

使用 rating_numerator/rating_denominator 的形式来表示 rating,然后再删除这两列。

2.image_predictions 表格也应该合并到 twitter 表格中，并且 twitter 表格去掉没有 image_predictions 的部分

使用 merge 方法将 image_predictions 表格合并到 twitter 表格中去。