数据整理报告

一、数据收集

收集下面描述的三份数据：

1.WeRateDogs 推特档案。

项目提供了 csv 文件 twitter-archive-enhanced，直接读取到 Jupyter Notebook，命名为 twitter。

2.推特图片预测，即根据神经网络，出现在每个推特中狗的品种 (或其他物体、动物等)。

使用 Python 的 Requests 库和提供的 URL 来进行编程下载，然后读取到 Jupyter Notebook，命名为 image_predictions。

3.每条推特的数据，至少要包含转发数（retweet_count）和喜欢数（favorite_count），以及任何你觉得有趣的额外数据。

从 txt 文件 tweet_json 中获取数据 tweet_id，retweet_count，favorite_count，读取到 Jupyter Notebook，命名为 tweet_extra。

二、数据评估

通过目测评估和编程评估发现：

质量问题：

twitter 表格：

- 存在转发的推特需要删除；
- in_reply_to_status_id 、 in_reply_to_user_id 、 retweeted_status_id 、 retweeted_status_user_id、retweeted_status_timestamp 类型不对；
- in_reply_to_status_id、in_reply_to_user_id、expanded_urls 信息缺省；
- timestamp 类型错误；
- name 有的条目提取错误，且有缺省值；
- doggo、floofer、pupper、puppo 有缺省值,有的条目并列存在；
- rating_denominator 有的不是 10；
- rating_numerator 存在异常值；
- 无图片的 twitter 需要删除。

image_predictions 表格：

- p1,p2,p3 中狗名字格式不一致，首字母要大写。

tweet_extra 表格：

- 缺少记录，暂时无法修改

清洁度问题：

- twitter 表格中的 doggo、floofer、pupper、puppo 应该用狗的地位 stage 来表示；
- tweet_extra 表格中的 favorite_count 和 retweet_count 加入 twitter 表格中；
- image_predictions 表也应该合并到 twitter 表格中，并且 twitter 表格去掉没有 image_predictions 的部分。

三、数据清洗

数据清洗首先处理缺失数据，然后清理整洁度问题，最后清理质量问题。
首先将三份原始的数据 copy，保存原始数据的副本。

1.处理缺失数据

- twitter：name 有的条目提取错误，且有缺省值。

处理方法：用 str.extract()和正则表达式从 text 中重新提取 name。

2.清洁度问题

- twitter：doggo、floofer、pupper、puppo 应该用狗的地位 stage 来表示，doggo、floofer、pupper、puppo 有缺省值，有的条目并列存在。

处理方法：doggo、floofer、pupper、puppo 应该用狗的地位 stage 来表示，删除这四列，并用 str.findall 函数和正则表达式重新提取“stage”信息，再用 apply 函数针对 stage 这一列做出修改，将未提取出来的修改为 NaN，有多重地位的将这些地位信息用逗号连接。

- twitter 表格中应该去掉 retweeted_status_id 不为空的行。

处理方法：我们分析的是原始数据，即不包含转发的数据，因此在 twitter 的 retweeted_status_id 不为空（转发条目）的都要删除。

- tweet_extra 表格中的 favorite_count 和 retweet_count 加入 twitte 表格中。

处理方法：用 merge 函数将 tweet_extra_clean 表中的 favorite_count 和 retweet_count 加入 twitter_clean 表格。

- image_predictions 表也应该合并到 twitter 表格中，去掉 twitter 表格中没有 image_predictions 的部分。

处理方法：用 merge 函数将 image_predictions_clean 合并到 twitter_clean 表格，how 为 inner。

3. 质量问题

- in_reply_to_status_id 、in_reply_to_user_id 、retweeted_status_id 、retweeted_status_user_id、retweeted_status_timestamp 类型不对。
- in_reply_to_status_id、in_reply_to_user_id、expanded_urls 信息缺省

处理方法：考虑到在分析数据时不会用到这几项，此处用 drop 函数直接删除。

- timestamp 类型错误

处理方法：使用 pd.to_datetime 将 timestamp 转为 datetime 数据类型。

- rating_denominator 有的不是 10
- -rating_numerator 存在异常值

处理方法：使用正则表达式和 extract 函数重新提取两项评分，再次检查异常值，进一步分析和处理。无法确定原始值的，虑用众数替换。

- p1,p2,p3 中狗的名字格式不一致，首字母要大写

处理方法：使用 title 函数，将 p1,p2,p3 中名字的每个字的首字母大写。

三、保存数据

将清洗完成后的数据保存在 twitter_archive_master.csv。