数据清理过程简介

1. 数据收集

在本次项目中首先下载了 twitter-archive-enhanced.csv 文件。 并且使用requests库下载了神经网络识别狗狗图片信息的表格 image-predictions.tsv。 由于无法申请推特API故直接下载 tweet_json.txt 文件，并且将json文件逐行导入DataFrame，并保存为 twitter_json.csv.

2. 数据评估

首先使用目测评估和编程评估，发现

质量问题

`twitter-archive-enhanced.csv` 表格中的问题：

1. 多列数据没有值，而且没有实际意义。
2. 删除转发的推特文章，其中 `retweet_status_id`， `retweet_status_user_id` 和 `retweeted_statue_timestamp` 中不为空的就是转发的文章。
3. `timestamp` 列中的+0000需要删除。
4. `timestamp` 列需要修改类型。
5. `rating_denominator` 和 `rating_numerator` 列有部分值识别不对，需要重新进行用正则表达式获取。并将识别后分别赋值给 rating_denominator 和 rating_numerator 列，然后转换这两列类型为int类型，然后将 rating_numerator/rating_denominator 得到 rating_num 列具体的分数,并删除 rating_num 列异常值。

`twitter_json` 表格中的问题：

1. `created_at` 列日期月份和星期使用了英文缩写。
2. `display_text_range` 列需要将文本换成单个数字，而不是一个范围。

`image_predictions_clean` 表格中的问题：

1. 删除p2及p3关联的所有列。

清洁度问题

`twitter_archive_enhanced` 表中的问题

1. 将 `twitter_json` 表中的 `favorite_count` 列和 `retweet_count` 列合并到 `twitter_archive_enhanced` 表中。
2. 将 `image_predictions` 表合并到 `twitter_archive_enhanced` 表中。

3. 数据清理

质量问题

对 `twitter-archive-enhanced.csv` 表格进行清理

1. 将没有太多实际意义的列删除，这些列为：'in_reply_to_status_id', 'in_reply_to_user_id'。
2. 删除转发的推特文章，判断retweet_statue_id的数据是否为空值，并得到一个布尔值列，根据这个布尔值列，选取整个表格中的行。然后删除 retweeted_status_id，retweeted_status_user_id 和 retweeted_statue_timestamp 三列。
3. twitter_archive_enhanced: timestamp列中的+0000需要删除。
4. 将 timestamp 列通过 to_datetime 函数转换为 datetime64 类型。
5. rating_denominator 和rating_numerator 列有部分值识别不对，需要重新进行用正则表达式获取。然后用split('/')将获取的列分开，并且用第一列除以第二列。得到 rating_num 列。然后删除该列里的异常值所在的行。

对 `twitter_json` 表格进行清理

1. created_at 列日期月份和星期使用了英文缩写。将 twitter_archive_enhanced 表格中的 timestamp 列根据 tweet_id 列合并到 twitter_json表格中，并且将twitter_json表格中的 created_at 列删除。将合并过来的 timestamp 列改名为，created_at。
2. display_text_range 列的所有值均为一个范围如[0, 132] ，这其实并不方便分析，应该将具体的推特文长度用一个数值进行表示，故将[0,132] 中的第二位数提取出来，由于该列的类型其实为 list 类型，故无法进行 split() 分割，所以直接使用索引方式，选择第二个值，即 132 并且赋值给当前位置。

对 `image_predictions_clean` 表格进行清理:

1. 删除p2及p3关联的所有列。由于p1的分析已经基本确定该图片是否为狗狗，p2 p3分析基本属于辅助，故为了排除干扰，故将p2 p3相关列删除。

清洁度问题

将 twitter_json 表中的 favorite_count 列和 retweet_count 列合并到 twitter_archive_enhanced 表中。

将image_predictions表合并到twitter_archive_enhanced表中，通过merge函数。并且删除多余的项。