Wrangle report

简介：

推特用户 WeRateDogs 以诙谐幽默的方式对人们的宠物狗评级。这些评级通常以 10 作为分母。WeRateDogs 拥有四百多万关注者，曾受到国际媒体的报道。

本文是基于 WeRateDogs 的数据集，使用 Python 及 Pyhton 上的库，在 Jupyternote book 平台的基础上进行的数据分析过程，其主要包括：

- 数据收集
- 数据检查
- 数据清洗
- 数据存储

数据收集：

本文主要收集的是三个数据集，分别是
Twitter 基本信息：twitter-archive-enhanced.csv
图片预测信息：image_predictions.tsv（url 下载）
Twitter 附加信息：tweet_json.txt

数据检查：

三个数据集分别生成三个对应 python 中的 DataFrame，分别是：
twitter-archive-enhanced.csv：df_twitter_enhanced
image_predictions.tsv（url 下载）：image_predictions
tweet_json.txt：tweets_jsons

通过数据的检查分别发现了以下的数据质量和整洁度问题：

df_twitter_enhanced

- tidiness

1. 新建一个 state 列,整合 doggo、floofer、pupper、puppo

- quality

1. rating_denominator 异常值
2. doggo、floofer、pupper、puppo 列缺失
3. 去除 retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp，in_reply_to_status_id，in_reply_to_user_id 几列无用列
4. tweet_id 应为 char，in_reply_to_status_id，in_reply_to_user_id 数据类型应为 char

5. timestamp 应为 datetime
6. source 的词文提取
7. 将 name 中的无意义词提取转为 None
8. 转 name 中的 None 为 NaN
9. 删除 expanded_urls 中的 NaN 值

image_predictions

- tidiness

1. 把 image_predictions， df_twitter_enhanced 和 tweet_json 连接成一个表，twitter_archive_master

- quality

1. tweet_id 为 char

数据清洗：

- **Quality：**

定义 1：

为了减少数据集中的无用数据，去除 retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp，in_reply_to_status_id，in_reply_to_user_id 几列无用列

定义 2：

除了分子中用户对自己的小狗狗的过度热爱的高分值外，分母中的 rating_denominator 通过检查发现，存在大量的异常值，在此定义分母不为 10 的值全部为异常值，并删去。

定义 3：

通过数据检查发现，有部分的数据的 expanded_urls 中存在 NaN 值，将其认为找不到数据源的错误数据，删除该条数据

定义 4：

通过数据检查发现，name 列中，有许多的数据的提取出的并非是姓名，而是例如：a、the、an、O 等的错误数据，在这里定义若是全小写和全大写的 name 的数据为错误无意义数据，将 name 中的无意义词提取并转为 None

定义 5：

在 DataFrame 中应将 None 数据全部转为统一规格的 NaN, 所以将 name 中的 None 转为 NaN

定义 6：

检查数据，发现部分数据所村粗的数据类型，并不是正确的，所以需要转换数据类型，

如：tweet_id 应为 char，in_reply_to_status_id，in_reply_to_user_id 数据类型应为 char，timestamp 应为 datetime

定义 7：

通过检查发现 Source 中的 URL 重要分为四大类，而且有一定的规律，可以通过正则表达式，将 source 的词文提取。分别是：Twitter for iPhone；Vine - Make a Scene；Twitter Web Client；TweetDeck。

定义 8：

发现 doggo、floofer、pupper、puppo 列有大量的数据缺失，通过后文中的清洁度整理建立 state 列后，删除这几列。

- **Tidiness**

定义 1：

建立一个 state 列整合 doggo、floofer、pupper、puppo 这四列的数据，若是不是这四列的话，则定义为 None，方法是通过正则表达式提取 Text 中的内容。

定义 2：

为了方便数据的整体分析和整理把 image_predictions，df_twitter_enhanced 和 tweet_jsons 连接成一个表并命名为，twitter_archive_master

数据存储：

将整理好的数据存储到一个 twitter_archive_master.csv 的文件中。