wrangle_report

第一步：收集数据
由于没有办法使用tweet的api，所以tweet_json.txt从github获取。使用request模块下载。

第二步：评估数据
使用pandas和json模块，读取tweet_json.txt文件，按行转换成字典，再将这个字典转换为pandas.Series类型并append到dataframe中。使用describe()、info()、sample()查看数据。

第三步：清理数据
查看数据集，发现有以下的数据问题：

质量

- in_reply_to_status_id、in_reply_to_user_id、retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp的缺失的数据较多
- expanded_urls的数据，有的记录存在多个相同的url，应该去重，而且expanded_urls存在数据缺失
- rating_numerator的数据，翻看xls格式数据发现有的数值小于rating_denominator
- rating_denominator的数据，发现值为0或者不为10的分母，0不能当分母
- text的数据，存在多个分数时，只获取了一个分数，这里应该要取多个分数的平均值
- timestamp的数据是object类型，要转换成date类型的数据
- name的数据，有的是a或者an，需要把这样的数据改成null。例如This is Adele.提取name为Adele；但是This is a rare Arctic Wubberfloof.提取a是不正确的，不能通过This is .的方式来提取狗的名字。
- doggo/floofer/pupper/puppo的数据，有的是'None'字符串，其实应该替换为null，表示数据的缺失。
- tweet_json的id，要转换成int64类型，以便跟twitter_archive_enhanced进行merge

清洁度

- source字段内容是html文本，只需要获取该html的text部分即可
- twitter_archive_enhanced缺少`favorite_count`和`retweet_count`，需要从tweet_json.txt中获取后合并到twitter_archive_enhanced中

注：提取狗的名字时，使用This is <dog_name>.的表达式去匹配的，这样会导致有些不是这样写的tweet，狗名字获取失败，现在多加了几种匹配模式（This is a|an、named、Meet等等），尽可能的获取到狗的名字。

最后：把清理好的数据，使用to_csv()保存到twitter_archive_master.csv文件中。