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

- id应该是字符串类型、favorite_count、retweet_count应该是int类型，retweeted应该是bool类型
- 只需要不包括转发的数据，即保留retweeted为False的数据
- 只需要包含图片的原始评级，即保留expanded_url为空的数据
- created_at应该是date类型
- 从full_text中提取分数时，分子为浮点数，分母为10
- 从full_text中提取分数时，如果分数出现多次，那么则取均值作为分数
- 从full_text中提取狗的名字

清洁度

- 拆分分数，分别得出分子rating_numerator和分母rating_denominator
- source字段内容是html文本，只需要获取该html的text部分即可
- 清理了分子小于分母的数据，清理了分子过大的数据

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，暂时没有想到好的方法处理这样的情况。

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