wrangle_report

一： 收集数据

1:从现有的文件 twitter-archive-enhanced.txt 中读取数据并将其另存为 twitter-archive-enhanced.csv

2: 通过网络 URL 地址下载文件，从现有的文件 image_prediction.tsv，并把文件保存到本地；

3:从现有的文件 tweet_json.txt；读取 json 文件数据，并将其另存为 tweet_json.csv;

二： 评估数据

通过目测评估跟代码评估得出以下问题：

三个数据集先合并为一个数据集；合并为一个 twitter_archive_master 数据集；再进行清理

质量

twitter_archive_enhanced_master.csv 表格

- 包含的转发数据，需要清除
- jpg_url 有重复数据，删除；

- （ in_reply_to_status_id ， in_reply_to_user_id ， retweeted_status_id ， retweeted_status_user_id ， retweeted_status_timestamp）这几列数据缺失比较多，可删除掉这些列；
- timestamp 时间不是日期格式；
- source 来源列是个标签文本，去掉标签，只显示来源内容
- 评分分母列（rating_denominator）有个别数据不是 10；修正；
- 列 name 名字小写的是拼写错误，a,an the...none 等错误字段，改成 nan 值
- 列 p1，p2，p3 有大写有小写，改成小写

整洁度

- twitter_archive_enhanced_master.csv 表里面 doggo、floor、pupper、puppo 代表狗的类型，应该合并为一列
- twitter_archive_enhanced_master.csv 表里面 rating_numerator、rating_denominator ，都是评分，应该合并为一列 rateing 评分列
- twitter_archive_enhanced_master.tsv 表里面列 p1，p2，p3 这三列都是描述狗的种类应该为一列；p1_conf，p2_conf，p3_conf 这三列是也应该合并为一列；p1_dog，p2_dog，p3_dog 这三列也应该合并为一列；
-

三：清理数据

1:首先我对收集的数据进行备份，然后对各个数据用后缀 clean 重新 copy 一份数据进行清洗；

2:针对评估的问题，我逐一清洗完成

四：对清洗后的数据进行复查

重新对第一次清洗的数据进行重新目测评估跟代码评估；发现以下数据质量问题；

twitter_archive_master_clean 数据集

- 整合后数据的列 rating 存在个别数据比较大；有比例 177，42 不正常数据，修正，还有大部分都是取小数点一位的，把 rating 取小数点后一位；

对数据再次清理

五：对清洗完成的数据集进行保存

最后我把三个清洗完的数据集合并为一个数据集，并存储为 twitter_archive_master.csv 的文件中；

六：可视化分析

针对清洗的数据进行可视化分析