清洗过程

一、收集

1，WeRateDogs 推特档案

使用Requests模块，利用Github repo中对应的URL编程下载，命名为twitter-archive-enhanced.csv。

2，推特图片预测

使用Requests模块，利用Github repo中对应的URL编程下载，命名为image-predictions.tsv。

3，每条推特数据

使用Requests模块，利用Github repo中对应的URL编程下载，命名为tweet_json.txt。

使用json模块从数据中提取推文的retweet_count和favorite_count，并转换为DataFrame结构数据。

二、评估

使用目测评估及编程评估两种方式，发现以下的数据问题：

1，质量问题

- 数据冗余，比如没有附图、转发的以及与评级不相关的行，有效值很少与分析无关的列；
- 在name列及狗子评级列中空值为字符串 'None'，而不是 NaN；
- 狗子的评级数据提取错误，同一只狗子出现两种评级类型，并且有缺失值；
- 推特档案中部分狗子的名字提取错误，并且有缺失值；
- 推特档案中部分狗子的评分包含缺失值及无效值；
- source列数据冗余不清晰，应只包含推特来源；
- 类型错误，timestamp应该是时间类型而不是字符串；来源列，狗子评级列，推特图片预测数据中的p1,p2,p3列应为分类类型，而不是字符串。

2，整洁度

1）狗子的评级应为单独一列；

2）三个来源的数据应该在同一数据集中。

三、清理

1，清理冗余数据，即清理转发的以及与狗子评级不相关的推文；

2，清理会影响到数据合并、或合并后不易清理的质量问题，如冗余列、空值为字符串 'None'、名字提取错误、评级数据提取错误等；

3，整洁度问题清理，合并狗子评级、合并三个独立Dataframe(在合并过程中完成对不含图片推文的清理)；

4，清理剩余所有的质量问题。

四、保存数据

导出整理后的数据，命名为twitter_archive_master.csv。