Wrangle Report

整个项目清洗过程分为三个部分：收集数据、评估数据、清洗数据。

一、 收集数据

1.读取项目提供的 WeRateDogs 推特档案。

2.使用 requests 库从指定的 url 下载推特图像的预测数据。

3. 因为无法连接 twitter, 所以直接使用了项目提供的 tweet_json.txt 文件。因为对 json 文件的处理了解的不够，所以在网上搜寻相关信息后，使用了先将 json 文件中每行信息添加进表格，再将表格转化为 DataFrame 的方法。

二、 评估数据

通过目测评估和编程评估来评估数据，发现的问题请见 wrangle_act.ipynb。大部分内容错误的问题，主要是因为推文中出现了多次关键字，导致数据提取时有误。如推文中包含分数形式的日期，日期会被提取为评分。

三、 清理数据

在处理整洁度问题“WeRateDogs_archive 最后 4 列狗的地位属

于同一变量，可合并为一列”时，遇到了问题，最终采用了先将数据集一分为二，分开处理后再合并的方法。依照第一次审阅的建议，将有多个地位的狗标记为 multiple。因为对正则表达式的掌握不足，对审阅建议中提取狗的名字和狗的多次评分等问题没有解决。