数据整理步骤

1.数据收集

按照项目要求收集三部分数据

- WeRateDogs 的推特档案数据直接加载
- 推特图像的预测数据使用Python的Requests库下载
- 每条推特的额外附加数据从提供的原文件加载，按项目要求提取必须数据

2.数据评估

对已经收集的三个表进行评估，其质量问题有：

- 有些列数据格式不正确
- 有些列存在缺失数据
- 有些列数据可能在初始提取时不正确，数据不准确

整洁度问题有：

- 三张表有相同的数据列
- 表示狗的成长阶段的数据列分成四列，可以用一列表示

3.数据清洗

针对数据评估出现的问题，参考上一次审阅结果，对数据清理进行以下操作

- 对三个表按照相同的列进行合并，方便后续清洗
- 将不正确的列数据格式调整为正确的格式
- 根据项目细节提取非转发数据，并将于存在大量缺失值的数据列删除
- 对数据不正确的来源列，使用正则表达式重新提取
- 对数据不正确的评分分子、分母，针对有问题的分母和分子数据，参考数据文本列信息进行重新修改
- 对数据不正确的名字列，观察文本列发现规律，使用正则表达式重新提取
- 对于表示不同狗成长阶段的数据，创建一个表示成长阶段的数据列，使用正则表达式从文本数据中提取四种阶段的字符，重新整理