**WeRateDog项目数据清洗**

1. **数据收集**
   1. WeRateDogs推特档案数据直接导入twitter-archive-enhanced.csv文件
   2. 通过URL编程下载推特图像的预测数据，即根据神经网络，对出现在每个推特中狗的品种（或其他物体、动物等）进行预测的结果。
   3. 通过Tweepy库下载额外的推特数据，至少要包含转发数（retweet\_count）和喜欢数（favorite\_count）
2. **数据评估**

通过目测评估和编程评估的方式对数据进行质量及整洁度的评估

* 1. 质量：
     1. twitter\_archive 表格
* tweet\_id是整数格式
* in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp 列有太多缺失值
* 只需要含有图片的原始评级 (不包括转发)
* timestamp是字符串格式不是时间格式，并且都有+0000
* source列太多多余的信息，如http连接地址
* expanded\_urls列中有缺失值或者一行中内容重复
* rating\_numerator和rating\_denominator的最大值最小值异常
* name有很多None或者a/an
* doggo, floofer, pupper, puppo中有很多None
* 不需要2017年8月1日后的数据
  + 1. *image\_predictions* 表格
* tweet\_id和 img\_num是整数格式
* jpg\_url列有重复项
  + 1. tweet\_extra\_data表格
* tweet\_id是整数格式
  1. 整洁度：
* 狗狗阶段分为了四列 (doggo, floofer, pupper, puppo)
* 三张表格可以整合

1. **数据清理**
   1. 数值缺失问题

* 只需要含有图片的原始评级 (不包括转发)：删除包含in\_reply\_to\_status\_id, in\_reply\_to\_user\_id的行
* in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp 列有太多缺失值：删除列
  1. 整洁度
* 狗狗阶段分为了四列 (doggo, floofer, pupper, puppo)：重新从文中提取stage的信息，放入stage列中，删除原有的四列
* 三张表格可以整合：放在最后合并
  1. 质量
     1. Twitter\_archive表格
* tweet\_id是整数格式：修改tweet\_id的格式为字符串
* timestamp是字符串格式不是时间格式，并且都有+0000：用pd.to\_datetime转换成时间格式，并只选取+0000之前的字符
* source列太多多余的信息，如http连接地址：从source中截取正确的信息，比如Twitter for iphone, Twitter Web Clien等
* expanded\_urls列中有缺失值：保留expanded\_urls为notnull的行
* rating\_numerator和rating\_denominator的最大值最小值异常：从文中重新提取分数，检查异常值并修改
* name有很多None或者a/an：重新从文中匹配狗狗名字
* doggo, floofer, pupper, puppo中有很多None：四列已删除
* 不需要2017年8月1日后的数据：在合并三张表后删除2017/8/1之后的数据
  + 1. image\_predictions表格
* tweet\_id是整数格式：修改tweet\_id的格式为字符串
* jpg\_url列有重复项：用drop\_duplicates删除重复项
  + 1. tweet\_extra\_data表格
* tweet\_id是整数格式：修改tweet\_id的格式为字符串

1. **数据合并及储存**

* 合并三张表格并删除2017/8/1之后的数据
* 储存数据到twitter\_archive\_master.csv中