一、收集数据

输入数据

- twitter_archive_enhanced.csv
  项目提供的数据，其中包括评分、地位、名字和转发状态的信息，但是存在许多错误和缺失的数据。
- tweet_json.txt
  项目提供的数据，缺失的数据较多，仅保留使用点赞和转发的数据。
- image_predictions.tsv
  通过提供的地址，通过网络接口进行下载，其中包含了对于狗的识别情况（分类，可行度，识别的结果），图片地址，图片数量等信息。

输出数据

- twitter_archive_master.csv
  合并清洗之后的数据

二、数据评估

质量问题：

- witter_archive_enhanced
  - name 列存在在很多重复名称，并且存在为None的记录
  - in_reply_to_status_id/in_reply_to_user_id/retweeted_status_id/retweeted_status_timestamp/ retweeted_status_user_id 数据缺失严重
  - retweeted_status_id：存在非空字段，存在转发的twitter记录
  - soure：为html标记内容<a
  - rating_denominator:存在非10的数据记录
  - doggo，floofer，pupper，puppo：存在四个stage均为None的记录，以及同一条记录属于多个stage的状况
  - name:存在a,o等无意义的单词
  - timestamp:时间戳的类型为string类型
- image_predictions
  - jpg_url：图片的url存在重复
  - p1,p2,p3:字符存在首字母大小写不一致的问题
  - p1_dog，p2_dog，p3_dog：存在预测失败的情况
- tweet_json
  - contributors,coordinates,geo,in_reply_to_screen_name,in_reply_to_status_id,in_reply_to_status_id_str,in_reply_to_user_id,in_reply_to_user_id_str,place,quoted_status,quoted_status_id,quoted_status_id_str,retweeted_status 数据缺失严重

整洁度问题：

- witter_archive_enhanced、image_predictions、tweet_json 中关于tweet id的标识不一致
- timestamp 转化为datetime类型
- doggo，floofer，pupper，puppo进行归一化合并为一列
- iwitter_archive_enhanced、image_predictions、tweet_json 合并到一个dataframe处理

三、数据清理

质量问题

- 删除数据：
  - 对于质量问题中提到的缺失严重的列进行删除
- 存在转发的数据记录，对于转发的数据进行删除，避免重复统计和处理
- 图片预测中存在重复的数据记录
- 缺失处理：
  - 对于缺失的数据（name，stage等），通过正则表达从text中进行提取
- 数据修正：
  - 对于数据一致性存在问题的数据进行编码的统一处理（大小写，下划线等）
- 手动处理：
  - 对部分存在两个stage类型的数据进行手动处理
  - 对部分name名称存在单个字母的数据进行手动处理

整洁度问题

- 对于部分变量（tweet id）统一命名
- 合并stage的四种类型
- 合并三个数据源的数据