本项目中通过三种方式收集数据，
1. 编程下载：使用编程方式从Github下载image-predictions.tsv，并导入DataFrame中。
2. API接口方式下载：通过twitter API成功下载数据，在后面的清理分析中采用项目提供的数据集
3. 手工下载twitter-archive-enhanced.csv文件导入DataFrame

评估数据集后发现6个质量问题和3个整洁度问题：
质量问题：
- 狗狗的名字提取不准确，出现a、all、the等名字
- source来源不清晰，应为“Twitter for iPhone”，“Vine - Make a Scene”等
- 记录中包含转发数据、回复数据和没有图片的数据
- 错误的数据类型：如id应为字符、时间应为时间格式

`image-predictions` 表格
- 可信度保留的小数位数应一致，统一保留4位小数
- p1、p2、p3 首字母大小写不一致

整洁度问题
- twitter_data中doggo，floofer，pupper、puppo列应融合到单独的type列
- 转发数、喜爱数与twitter_data应为同一类数据
- image-predictions中的数据与twitter_data应为同一类数据

狗狗名字部分提取不准确，出现了“a”, “all”, “the”等名字，目测这些记录中不存在名字信息，则将这些记录的名字列设置为None。

之后依次处理剩余质量问题，通过Googe查询相关函数与方法，处理起来比较容易。

最后处理整洁度问题，使用melt函数将twitter_data中doggo，floofer，pupper、puppo列应融合到单独的type列，但有些记录中存在多条地位信息的情况处理遇到了问题。三个数据集为同一类数据，将三个数据按tweet_id合并。在合并过程中因tweet_id为字符数据类型而报错，将tweet_id转换为数值类型后，合并成功。