项目背景

清洗 WeRateDogs 推特数据，创建有趣且可靠的分析和可视化。这份推特档案很棒，但是只包含基本的推特信息。要达到 "Wow!" 的效果，在分析和可视化前，还需要收集额外的数据、然后进行评估和清洗。

我的数据整理

分为这四块：收集数据，评估数据，清洗数据，存储数据

收集数据

这里一共需要收集三个数据，如下所示

1. **twitter-archive-enhanced.csv**
   WeRateDogs 的推特档案，网页直接下载保存至本地。

2. **image_predictions.tsv**
   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。使用 requests 下载至本地。

3. **tweet_json.txt**
   每个推特的 JSON 数据。需要使用 twitter 的 API Tweepy 下载，但是这里我申请 Twitter 开发者账号，一直处于审核中，为此，我使用了 Udacity 线上提供的数据 tweet_json.txt。

评估数据

这里使用了目测评估和编程评估两种方法
目测评估：人眼查看的方式，查看某些行，某些列的数据，是否存在问题
编程评估：编程查看，比如 df.head(),df.info(),df.describe(),df.sample(5)

从以下两类（质量和整洁度）进行评估考量。

质量

a 完整性

是否记录了所有内容？是否缺少记录？是否丢失某个行、列或单元格？

b 有效性

已经做了记录，但却无效，即它们不符合定义的模式。模式是一组定义的数据规则。这些规则可以是真实世界约定成俗的事实（例如身高不可能是负数）和表格约定成俗的属性（例如表中的唯一键）。

c 准确性

不准确的数据是有效的，但仍然是错误的。这些数据符合定义的模式，但仍然不正确。例如：每个患者体重被多记录了 5 磅。虽然有失偏颇，这些数据仍然是有效的，但并不理想。

d 一致性

不一致的数据是有效和准确的，但是指代同一件事情的正确方式有多个。最好确保表内表示相同数据的列中的数据具有一致性，即采用标准格式。

整洁度

1. 每个变量构成一列
2. 每项观察构成一行
3. 每项类型的观察单元构成一个表格

我发现的一些问题如下

Tweet（指代源数据 tweet_json.txt）

质量：暂无

整洁度：

- tweet 的转发数和喜爱数可合并至 archive

Archive（指代源数据 twitter-archive-enhanced.csv）

质量

- 存在转发重复 tweet
- 错误的数据类型
- 狗狗地位列单元格数据不正确

- 部分列缺失
- source 简化
- 评分分子和分母存在异常值
- 狗狗名字需要修正

整洁度

- 狗狗地位列太多，可以合并为一列

Predications（指代源数据 image_predictions.tsv）

质量

- predications 中的 p1,p2,p3 对应的种类大小写不一致

整洁度

- predications 部分数据合并到 archive 中
- predications 可以更加精简

清洗数据

这里针对评估数据种发现的问题进行清理，每个问题清理的流程如下
定义：描述对此问题的看法，怎么处理
代码：将定义内容转为代码实现
测试 ：测试以上代码的结果，检查是否符合预期

关于以上的定义，代码以及测试方面的具体内容，需要查看 wrangle_act.ipynb

存储数据

已将清洗完毕的数据保存为 twitter_archive_master.csv 和一个附加文件 image_predications_condense.csv，此文件为图像预测文件的精简版本，与源文件 image_predictions.tsv 结构稍有不同，但内容相同