本项目是 针对 推特账号 WeRateDogs 的数据展开数据分析，采用多种数据收集手段获取原始数据，客观评估后进行清洗，最终进行合理分析并提供可视化结果。

1. 收集

WeRateDogs 推特档案中包括需要分析的5000多条基本信息。根据Twitter API可以获取更为丰富和重要的信息（例如转发量和点赞量）。另外本项目已经提供了一个图像预测文件，该预测以神经网络为理论基础，得到了每个推文图片中对象的种类和可靠度。
这三个信息来源分别采用不同的数据收集手段来一一解决。

本项目采用的收集手段

1.1 平面文件的数据收集

平面数据文件 twitter-archive-enhanced.csv 中提供了本项目的基本信息。python的pandas包提供了丰富的读写平面文件的函数，如.read_csv, .to_csv等等，并有相应参数灵活获取数据。

1.2 利用Twitter API进行数据收集

使用 Tweepy 查询推特 API 获取 WeRateDogs 推特档案以外的重要数据，其中tweet_id、转发量和点赞量是关注重点。。利用API获取的每个推特 ID 数据类型为 JSON data，采用json库，来读取数据。最终将数据存tweet_json.txt 文件中。

1.3 利用Requests库进行数据收集

该项目中图像预测文件是需要从网络链接中下载获取的，这就要用到Requests库。

小结

数据获取手段一般是读写文件，通过HTML和API，使用数据库。本项目采用了其中的两种，结果可靠。

2. 评估

概述

评估数据是数据整理的第二步。评估重点主要有两点：质量问题（即内容问题）和整洁度（即结构性问题）。评估的手段主要是目测评估和编程评估。
一定要先处理整洁度，再处理质量问题。
通过pandas库中.info(), describe(), .duplicated(), sort_values(), .value_counts(), .sample()可以参看从整体和细节去查看数据问题。
另外需要注意的是，tweet_id 是将各个数据来源融合到一起的重要column。

整洁度

使分析难以进行的结构性问题都是整洁维度需要关注。
tweet_json和image_predictions中的部分column应当和WeRateDogs 推特档案中的基本信息合并到一个DataFrame中，这样方便数据分析，简化文件间查询的繁琐。
狗的地位共有四种，原dataframe中分为了四列。应当合并为一列。

质量

质量问题主要体现在数据类型，部分数据明显错误应当剔除，部分转发推文数据行应当剔除。

小结

整洁度问题一定要先于质量问题处理。数据类型问题也一定要重视。这两点会极大影响之后的数据清洗。另外有关数据整体评估的函数（如.describe()）和value有关的函数结合起来使用对发现异常值情况有所帮助。

3. 清洗

针对两大数据问题采用不同的清洗函数。在明确了数据问题后，就是对症下药了。

针对整洁度的清洗

数据合并运用到的函数主要有.merge()，并结合.rename()。

针对质量的清洗

其中运用到的函数主要有.replace(), .apply(), to_datetime(), .astype(), .loc()。

小结

函数运用可以考虑更多的手段。例如数据合并还可以考虑的函数有.concat(), .combine_first；数据清洗时用到的apply函数还可以考虑用map函数替换。

4. 总结

经过数据收集、评估和清洗三大步骤了，就可以得到“整洁和干净”的数据，可供之后分析所用。这三大步骤是反复和迭代的过程，但是每一步操作一定要有目的性。