Wrangle Report

一、项目背景

推特用户 WeRateDog 一个以诙谐幽默的方式对人们的宠物狗评级，拥有四百多万关注者。该项目主要通过 Python 对 WeRateDogs 的推特数据集进行收集，清洗，并分析和可视化数据。

二. 数据整理过程

根据项目要求，只需要含有图片的原始评级 (不包括转发)，对数据进行了处理。

数据整理过程包括三个步骤：数据收集，数据评估和数据清理。

1. 数据收集

(1) 手动下载包含该账号推特档案文件的 twitter-archive-enhanced.csv 文件，读取并获得 tweet_id.

(2) 通过 tweepy 获取相对应 tweet_id 的推特数据，将 json 格式的数据写入 txt 文件并读取。

(3) 通过 python4 以编程的方式变成下载包含对图片狗狗品种预测的 image-predictions.tsv 并读取其中数据。

2. 数据评估

以目测和编程两个方式对数据进行评估，发现数据在质量和整洁度方面存在的问题。

(1) 质量方面问题体现在：数据不完整，数据错误等，如 tweet_archive 中狗狗名字错误)，tweet_archive 含有一些转发的 tweet 数据

(2) 整洁度方面存在的问题体现在：数据结构问题，如狗狗的地位：doggo、floofer、pupper，puppo 应该作为一列。

3. 数据清理

数据清理包含三个步骤：

(1) 定义：将问题转化为具体的清理所要实现的任务

(2) 代码：以代码来实现定义中的任务

(3) 测试：通过测试，用目测及代码来检验是否达到了清理的目标。

通过数据清理，实现了将三个不同表格的数据整合到一个表格，去除了转发的 tweet 数据，清除了错误的数据内容等任务，得到一个质量和整洁度都有所提升的表格。