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. 数据评估和清理

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，并对数据集进行清理。

(1) 质量方面：

① tweet_archive 中有记录的 rating_denominator 为 0,为无效数据，于是将该条数据删除

② tweet_archive 有 2356 条数据，收集到的 tweet_gathered 只有 2334 条数据，都包含一些没有图片的推特数据，通过 merge 将三个表格以 inner 的形式合并，确保合并后的表格拥有一致的推特记录，并都含有图片，符合项目要求

③ tweet_archive 中的名字错误和 None 设置为 NaN

④ tweet_archive 中 source 的内容难以分辨,把难以辨认的 source 内容替换为容易识别的 source

⑤ 部分列格式错误，将狗狗地位的格式由 str 转为 category，timestamp 格式由 str 转为 datetime

⑥ tweet_archive 含有一些转发的 tweet 数据，将这些转发的推特数据删除，以达到项目要求。

⑦ 删除了 tweet_archive 一些无关的列

(2) 整洁度方面：

① 狗狗的地位：doggo、floofer、pupper，puppo 通过 melt 合并为 stage 列

② tweet_archive 包含狗狗姓名等信息，而 tweet_gathered 包含转发数和点赞数，image 包含图片预测信息，通过合并将三个表格整合为一个完整的 dataframe。