Twitter数据爬取，清洗与分析

简介

我们将要整理 (以及分析和可视化) 的数据集是推特用户 @dog_rates 的档案, 推特昵称为 WeRateDogs。他以诙谐幽默的方式对人们的宠物狗评分。这些评分通常以 10 作为分母。但是分子则一般大于 10：11/10、12/10、13/10 等等。
WeRateDogs下载了他们的推特档案，并通过电子邮件发送给优达学城，专门为本项目使用。这个档案是基本的推特数据（推特 ID、时间戳、推特文本等），包含了截止到 2017 年 4 月 1 日的 5000 多条推特。

目标

清洗 WeRateDogs 推特数据，创建有趣且可靠的分析和可视化。（但该档案只包含基本的推特信息，还需要收集额外的数据。）

我们在这个项目中的任务如下：
数据整理，其中包括：
收集数据，评估数据，清洗数据。
对清洗过的数据进行储存、分析和可视化。
书面报告 (1) 数据整理工作 (2) 数据分析和可视化

步骤与细节

1.数据收集

收集下面所述的三份数据：

1. WeRateDogs 的推特档案。这个数据文件是直接提供的，详见twitter_archive_enhanced.csv。
2. 推特图像的预测数据，这个文件你需要使用 Python 的 Requests 库和以下提供的 URL 来进行编程下载。
3. 每条推特的额外附加数据，要包含转发数（retweet_count）和喜欢数（favorite_count）这两列。推荐使用Twitter API。

2.数据评估

收集上述三个数据集之后，使用目测评估和编程评估的方式，对数据进行质量和清洁度的评估。
完整地评估和清理整个数据集将需要大量时间，出于学习和实践的考虑，本项目只是评估和清理此数据集中的8个质量问题和2个整洁度问题。

3.数据清洗

对你在评估时列出的每个问题进行清洗。展示清洗的过程,结果应该为一个优质干净整洁的主数据集（pandas DataFrame 类型）
根据整洁数据的规则要求，本项目的数据清理应该包括将三个数据片段进行合并。

4.对项目数据进行存储、分析和可视化

将清理后的数据集存储到 CSV 文件中，命名为 twitter_archive_master.csv。如果有其他观察对象的数据集存在，需要多个表格，那么要给这些文件合理命名。
对清洗后的数据进行分析和可视化。生成至少 3 个见解和 1 个可视化。

细节：

我们只需要含有图片的原始评级 (不包括转发)。尽管数据集中有 5000 多条数据，但是并不是所有都是狗狗评分，并且其中有一些是转发。
如果分子评级超过分母评级，不需要进行清洗。这个 特殊评分系统 是 WeRateDogs 人气度较高的主要原因。同样，也不需要删除分子小于分母的数据。
不必收集 2017 年 8 月 1 日之后的数据，你可以收集到这些推特的基本信息，但是你不能收集到这些推特对应的图像预测数据，因为你没有图像预测算法的使用权限。
不要在项目提交中包含你的推特 API 密钥和访问令牌（可以用 * 号代替）。

以下为数据清理和分析的具体步骤：

对三种数据类型的数据集分别进行导入，并数据集进行评估。通过查看发现了至少以下 8 种数据质量问题和 2 种数据清洁度问题。

数据质量问题

- 1.我们只需要含有图片的原始评级，所以需要删除一些转发的数据， 有一些列可以删除
- 2.多列数据缺失严重
- 3.source table: 数据需要提取，整理成Ipone客户端,Web客户端,VINE, tweet deck四个选项
- 4.评分的提取不太准确
- 5.text table: 既有文字，又有网址和评分，应当拆分
- 6.name table: 有的name提取不正确
- 7.timestamp table : 数据类型转换为时间类型
- 8.有时同一条数据中， 含有多个狗的地位数据， 正常情况下只能有doggo，floofer，pupper，puppo 中的一个

数据整洁度问题

- 1.推特数据中，doggo，floofer，pupper，puppo栏应当删掉，建立新的地位栏，并填入每条所对应的’doggo floofer pupper puppo'其中一个.
- 2.补充数据中的两个计数栏可以合并到推特数据中.

针对以上出现的问题，都分别运用编程的方式，一一进行了处理。在数据处理完成后，还进行了检查验证。

最后通过三个表格中共同的 tweet_id 列，将三个表格合并在一起，同时删除了重复列和一 些不相关的列，存储在 twitter_archive_master 文件里。