Wrangle Report

背景资料： 这次项目的数据集是推特用户 @dog_rates 的档案, 推特昵称为 WeRateDogs。WeRateDogs 是一个推特主，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。清洗 WeRateDogs 推特数据，创建有趣且可靠的分析和可视化。

步骤如下：

1.对项目数据进行收集

WeRateDogs 项目中，数据来自 3 个不同的数据源：

1. twitter archive_enhanced 的 csv 文件（这个数据文件是直接提供给的）
2. 需要以编程方式下载的一份图像预测结果 tsv 文件（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。需要使用 Python 的 Requests 库和以下提供的 URL 来进行编程下载。）
3. 通过 tweepy API 构建的 tweet_json.txt 文件（每条推特的额外附加数据，包含转发数 retweet_count 和喜欢数 favorite_count。使用 WeRateDogs 推特档案中的推特 ID，使用 Python Tweepy 库查询 API 中每个推特的 JSON 数据，把所有 JSON 数据存储到一个名为 tweet_json.txt 的文件中。每个推特的 JSON 数据应当写入单独一行。然后将这个 .txt 文件逐行读入 tweet_data 中，用 tweet_data 创建 df_info 数据集。）

2.对项目数据进行评估

对 df_archive 数据集做基本的数据评估，找出质量及整洁度的问题。

- 目测的问题有狗的名字有一部分为小写单词 a，an，the，字母'O' 是错误名字。
- 通过编程的观察，发现一部分的列数据类型不正确，部分有数据缺失，有没有图片的 tweet 等。

以下为数据清理的详细内容

3.数据清理

列出了 8 个待清理的数据质量问题，及两个数据整洁度为题。

Part3.数据清理

质量

定义

df_archive

1. 从.info()看出数据集里包含了不正确的数据类型(in_reply_to_status_id 和 in_reply_to_status_user_id 应该为 int,tiemstamp 应该为 datetime，rating_numerator 和 rating_denominator 应该为 float)，需要用.astype()函数更改数据类型。
2. 数据集里包含了没有图片的 Tweets，数量少较少，所以这些 tweet 会被删除。
3. 有一部分狗的名字不正确(a,an,the ... ...)。通过查看 text 列发现有部分狗名字出现在 text 的内容中，名字前一般以'This is','Meet','name is','Say hello to','named'作为提示,需要用正则表达式从 text 列中筛选狗名，以作为 name 列内容的调整。另外，名字'O'应该为 O'Malley，需要更改。
4. 在 source 列中，4 个不同的 source 分别为 iPhone,Vine,Twitter Web Client 和 TweetDeck 的 URL,不容易读取。可以用.replace()函数简化为'iPhone','Vine'，'Web',和'TweetDeck'字段。
5. 在 rating_denominator 列和 rating_numerator 列中，通过.value_count()发现，denominator 大部分为 10。不为 10 的 denominator，有部分的评分分数出现在 text 列的内容里，需要用正则表达式从 text 列中筛分数。有部分需要转化为 2 为数，以作为 rating_numerator，rating_denominator，rating 列的调整。
6. tweet_id# 835152434251116546 及 746906459439529985 rating 为零，用.drop()函数删除这 2 行。
7. 数据集包含了转发 retweet 的行,要用.drop()删除。

df_image

- 没有明显的数据质量问题

df_info

1. df_info 包含 2352 行 tweet，df_archive 为 2356，有 4 条缺失，数量少，可以删除
2. tweet_id# 838085839343206401 的 'retweet_count'为零

整洁度

定义

1. 狗的“地位”分成了 4 列: doggo, floofer, pupper, puppo，有些狗会存在多种'status ‘，用逗号分隔开不同的 status 其次对于狗的地位中为 None 的值我们应该替换为 NaN。
2. 'tweet_info' 和 'image_predictions' 列要接拼到 'twitter_archive'。

Part4. 数据分析

做了 3 个可视化分析

3. 随着时间的变化流行度的趋势
4. 预测频率最高的狗品种
5. 狗地位的评分分布