数据整理报告

一、整体评估

这份数据复杂度比较高，包括dog_rating推特存档数据、补充数据以及图像预测数据三个文件，需要分别应对数据质量和整洁度问题。其中课程组提供的`dog_predict`文件几乎没有问题。重点是`twitter_archive.csv`这个文件，存在数据质量和整洁度的问题。数据质量包括数据缺失、格式错误、内容冗杂等问题。

二、整理过程

`twitter_archive.csv`

- 其中`in_reply_to_status_id`等五列id和事件戳几乎缺失90%以上，而且转推id和时间分析意义不大，以上诸列数据删除。
- 然后是`source`一列还带有`<a href=`等杂乱元素，通过使用正则提取方法，取两组尖括号中的值
- 然后通过`info()`看出timestamp数据类型错误，转为datetime
- 随后又整理出狗狗类型大量缺失的问题，但都被「None」掩盖，鉴于这部分变量对于之后评判狗狗的可爱度直接相关，所以不删除，先将None替换为「NaN」，为之后整洁度整理做准备。
- 如上所言，描述狗狗地位的四个参数放在了`doggo\floofer\pupper\pupo`四列里面，影响随后分类描述和归因，所以使用`melt`将狗狗的position 集中到一列。
- 本以为这个文件应该没有问题，但最后在可视化的过程中发现了狗狗`rating_numerator`的异常值27，回溯调查发现是11.27，于是迭代更正。这里的经验就是，最好在一开始就通过`describe()`或者可视化了解异常值。

使用逐行读取`tweet_json.txt`获得的`twitter_add`文件一开始所有变量都混在一列中。

- 通过`map(lambda x:x.split(','))`方法一一提取出来
- 将`tweet_id\retweet_count\favorite_count`几个关键变量中的杂项提出，数据类型转回整数。

最后，使用`pd.merge`方法将三个文件基于`tweet_id`相连，由于想进行地位分析，只保留了有地位记录的394条数据。