如果审阅老师你看了我的 wrangle_act.ipynb,以下内容就不用再看了.

因为我所有的文字解释,解决问题的思路都再 wrangle_act.ipynb 里有详细说明.

我觉得文字只有结合代码运行结果才更清晰明白.

数据存在的问题及解决思路简述

一. 质量

1. 有效性

- 存在非 tweet 网站的原始数据，从其他网站转发到 tweet 的数据不能用。删掉

2. 完整性

- archive 表格中的 in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls 这几列，存在大量缺失值。删掉

- archive 中的 name 含有大量值为“None”，说明有缺失信息。从 `text` 中匹配,筛选并补充

- 小狗的地位数据含大量缺失信息。从 `text` 中匹配,筛选并补充

3. 一致性

- 三张表的 id 列数据类型和列名称不统一。统一为 `tweet_id`

- archive 表格中的 timestamp 列内容是表时间，但数据类型不是时间型。`pd.to_datetime(timestamp)`

4. 准确性

- name 一列里有 745 个 None，55 个 a，7 个 an，这些通常不是狗的名字。从 `text` 中匹配,筛选并补充

- 根据 rating_numerator 和 rating_dominator，再结合 text 原文内容发现，评分的选取存在错误，有的数据是日期如 9/11、11/15/15；有的数据如 960/00、3 1/2 这些也不是评分。从 `text` 中匹配,筛选并补充

二. 整洁度

定义：“不符合“每个变量构成一列，每个观察值构成一行，每个观察单元构成一个表格”

- doggo，floofer，pupper，puppo 都是狗的地位，应该单独为一列。从 `text` 中匹配,筛选,赋值给新建的 `state` 列

- 很多 text 列中含有两个及以上的 rate，name，state 数据。从 `text` 中匹配,筛选,赋值给相关列

附录：评分/姓名/地位的多值问题

根据 text 列（附加 expanded_urls 列）找到原文后发现，同一条推特里存在多个评分、姓名、地位的数据。这样的情况通常是因为：

1. 同一条推特里提到了两个及以上的描述对象，即多只宠物狗，因此这条推特里就含有两个及以上的评分/姓名/地位；
2. 还有一种情况是，可能真正有效的评分/姓名/地位其实只有一个值。

评分例子："Meet Eve. She's a raging alcoholic 8/10 (would be 11/10 but pupper alcoholism is a tragic issue that I can't condone)"
姓名例子："This is Cermet, Paesh, and Morple. They are absolute h*ckin superstars. 14/10 for all"
地位例子："He's a sophisticated doggo. Also pointier than your average pupper"

解决思路：

为了数据的整洁度，我决定每条推特只取一个评分/姓名/地位的值。大部分情况下，每条推特的**第一个**评分/姓名/地位值相对更有效。例如：

- 评分例子中，8/10 是有效的评分，11/10 是虚拟语气的表达，相对无效。
- 姓名例子中，14/10 是对 Cermet, Paesh, and Morple 这三只狗的整体评分。任取一个姓名都有效。
- 地位例子中，doggo 是对主要描述对象的地位评价，而 pupper 对应的是其它对象。。

又因为 findall()提取的是列表，而大多数评分/姓名/地位只有一个值，我需要从大量列表里取出那单个值，这样步序繁琐而且比较耗时。所以最后用 str.extract()获取每条推特 text 中的第一个分数/姓名/地位。然后再筛选出异常值及其索引，根据索引重新到 text 里找出第二个替代值。

关于姓名/地位数据的二次筛选：

1. 关于`姓名`的名词，姓名是不能被量化的自然语言，最后筛选出来可能存在少量异常值。

我没有对第一次匹配后的数据进行第二次筛选，因为对我而言，知识超纲了。这个部分知识涉及自然语言处理的知识，才能更好地判断：这个名词是不是合理的名字？是指代人还是指代狗？

2. 关于`地位`的名词，它们基本上是确定的/较为精确的字母组合。

基本上可以说：能匹配到的都是，没有匹配到的，就不是。所以也没有进行二次筛选。