数据清理报告

1 数据收集：

收集了三个数据集，分别是 twitter-archive-enhanced，image-prediction 和 tweet_json，并将这三个数据集分别加载为 df_raw，df_img，df-supl 三个 DataFrame 文件中。

2 数据评估

2.1 质量问题

2.1.1 twitter-archive-enhanced 表格

1. 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_timestamp', 'expanded_urls', 'retweeted_status_timestamp', 'source'这些列不需要。
2. 只保留'retweeted_status_user_id', 'retweeted_status_id'为 null 的记录，因为这两个列如果不是空值的记录，表明记录中相关的 tweet 内容都是转发，会与原内容重复。
3. 有些犬只没有名字，但是 name 列填充的是’an’, ‘the’或者’a’，而不是空值或者 None。
4. 有些犬只没有等级分类，数值是 None。
5. rating_numerator，rating_denominator 这两个列存在异常值，分别大于均值 13 和 10
6. timestamp 列的类型是 object。
7. tweet_id 列的类型是 int64

2.1.2 image-prediction 表格

1. 由于 p1 列和相关预测分数表明犬只最可能的种类，因此不需要’p2’, ’p2_conf’,’ p3’, ’p3_conf’, ’p2_dog’, ’p3_dog’这些列。
2. ‘img_num’ 表示对应的图片的编号，不需要这个列。

2.1.3 tweet_json 表格

1. ‘id’, 应该为’tweet_id’,应与其他两个数据集相应的列保持一致。

2.2 整洁度问题

1. twitter-archive-enhanced，image-prediction，tweet_json 三个表格的观察对象都是 tweet_id。
2. 表格 twitter-archive-enhanced 中，犬只的等级分布在四个列中。
3. 表格 image-prediction 中使用 p1，p1_dog 两个列均用来表述犬只的种类

3 数据清理

3.1 清理数据质量问题

3.1.1 清理不需要的特征

- 从 df_raw 中过滤掉'retweeted_status_id','retweeted_status_user_id'非空的记录
- 使用 drop 方法从 df_raw 删除下列不需要的列： ‘in_reply_to_status_id’. ‘in_reply_to_user_id’, ‘retweeted_status_id’,’retweeted_status_user_id’, ‘retweeted_status_timestamp’, ‘expanded_urls’, ‘timestamp’, ‘retweeted_status_timestamp’, ‘source’
- 从 df_img 中删除 p2，p2_conf，p3，p3_conf，p2_dog，p3_dog，img_num 这些列
- 从 df_supl 删除除了下列三个列 favorite_count，retweet_count，id 之外的其他列。

3.1.2 清理空值

- replace 方法对 df_raw 中的 name 中的 a,an,the 替换为"None"
- 使用 dropna 方法清理空值以及相应的记录。

3.1.3 清理数据类型

- 使用 astype 或者 to_datatime 方法将 df_raw 中的'timestamp'的数据类型改为 Datetime 类型
- 使用 astype 方法将 df_raw，df_img 中的'tweet_id'，以及 df_supl 中的'id'的数据类型改为 str 类型

3.1.4 清理一致性问题¶

- 使用 str.title()将 df_img 中'p1'中犬的类型的格式进行统一化处理。

3.1.5 清理异常值问题

- 过滤掉 df_raw 中'rating_numerator'以及'rating_denominator'中分子大于 20，分母大于 10 的记录。

3.2 清理数据整洁度问题

3.2.1 清理多个列表示一个特征的问题

- 过滤掉 df_img 文件中'p1_dog'为 False 的记录
- 创建 df_temp，是 df_raw 的切片，包含'doggo','floofer','pupper','puppo'四个列，并添加'stage'列
- 使用 apply 方法将 df_raw 中'doggo','floofer','pupper','puppo'四个列合并为'stage',
- 使用 drop 方法，删除 df_raw 中的'doggo','floofer','pupper','puppo'四个列，并将 df_temp 中的'stage'合并到 df_raw 中。

3.2.2 清理多个数据表描述统一个 observation 的问题

- 使用 drop 方法清理 df_img 中不必要的列:'p1_dog','p1_conf'
- 使用 rename 方法，将 df_supl 中的'id'改为'tweet_id'
- 使用 merge 函数，将 df_raw，df_supl,df_img,三个数据集，按照'tweed_id'进行左连接再进行内连接
- 使用 rename 方法，将 df_img 中的'p1'改为'dog_type'