Twitter（WeRateDogs）数据清理报告

Wrangle_Report

评估

质量

`twitter_dog` 表格

- 数据集中有转发的，而我们需要的是原始评级，不包括转发
- 错误的数据类型（timestamp，大部分 id 列）
- 部分来自 text 的 rating_denominator 和 rating_numerator 解析错误
- rating_denominator 分母有些不为 10，主要是多个狗的图一起打分所致
- source 这一列应该简洁表示，如 iphone 等，数据类型为 category
- doggo、floofer、pupper、puppo 四个列的缺失数据格式设置不对， 应该为 category
- twitter 中有一个负分的评分（667550882905632768），这个不是狗，需要删除
- floof 类别太少，需要重新筛选

`image_prediction` 表格

- jpg_url 中有重复的数据
- 'id','id_str'重复，而且有的观察两列数据不一致

`extra_twitter` 表格

- `contributors`,`coordinates`,`geo` 等多列为空值
- 'id','id_str'重复，而且有的观察两列数据不一致
- favoriate_count 有很多值为 0，这是不正常的，在 retweeted_status 中需要重新解析

整洁度

- `twitte_dog` 表中的 doggo、floofer、pupper、puppo 四个列应该整合为一列
- `extra_twitter` 中的转发和喜欢添加到 `twitter_dog` 表格中
- `image_prediction` 中的 jpg_url 一列应该加入到 twitter_dog 中

清理

缺失数据

问题	定义（清理方法）
twitter_dogfloofer 列需要重新筛选	从 text 中提取 floof
twitter_dogdoggo、floofer、pupper、puppo 四个列的缺失数据格式设置不对	用 replace 方法把‘None’替换为 np.nan
extra_twitterfavoriate_count 有很多值为 0，这是不正常的，在 retweeted_status 中需要重新解析¶	从 retweeted_status 一列中解析数据

清洁度

问题	定义（清理方法）
witte_dog 表中的 doggo、floofer、pupper、puppo 四个列应该整合为一列	用 melt 方法将四个称谓整合成‘stage’一列，删除重复项
xtra_twitter 中的转发和喜欢两列添加到 twitter_dog 表格中	选取 extra_twitter 中的 'favorite_count', 'retweet_count' 两列添加到 twitter_dog 表格中
extra_twitter 中的转发和喜欢两列添加到 twitter_dog 表格中	选取 extra_twitter 中的 'favorite_count' , 'retweet_count' 两列添加到 twitter_dog 表格中

质量

问题	定义（清理方法）
twitter_dog 错误的数据类型（timestamp）	把 timestamp 相关的列转换为 datetime 类型，涉及到 id 的列转换为 int,这里将 np.nan 转换为 0。
twitter_dog 数据集中有转发的，而我们需要的是原始评级，不包括转发	删除 retweet 转发非零的行，删除转发相关的三个列
twitter_dog 中有一个负分的评分（667550882905632768），这个不是狗，需要删除	删除 id 是 667550882905632768 的行
image_predictionjpg_url 中有重复的数据	使用 drop_duplicated()方法去掉重复的行
twitter_dog 部分来自 text 的 rating_denominator 和 rating_numerator 解析错误	重新从'text'提取数值，对于群体的打分以及其他的选择丢弃数据(丢弃的数据中，大部分是群体打分)
twitter_dog source 这一列应该简洁表示，如 iphone 等，数据类型为 category	从 source 中选取来源，转为 category 列
twitter_dog 中有重复的 tweet_id，主要是 text 中有多个描述狗地位的词，需手动清理	因为重复的很少，而且很多有迷惑性，所以选择删除重复的（同一个 id 都删除）

twitter_dog 中 doggo、floofer、pupper、puppo 四个列的缺失数据格式设置不对，应该为 category	stage 这一列转换为 category 格式