Tweet 文件没有用 API 抓取，直接下载的
这样的话，共有三个数据集，分别是：
twitter-archive-enhanced.csv
image-predictions.tsv
tweet_json.txt

先清洗单独的数据集，再合并到一起

首先是 twitter-archive-enhanced.csv

首先用.info()功能，查看数据集每列的数据类型，发现 tweet_id 应该为 str 类型，timestamp 应该为 datetime 类型，属于质量问题
in_reply_to_status_id　in_reply_to_user_id 数量很少，对最后的结论也没有什么太大关系，所以后面会删掉这两列
狗狗评级因为分母不全为 10，不便于比较，可以将分子分母两列合为一列，便于比较评级
同样用.vlaue_counts()功能查看狗狗的名字，发现 a 的名字有 55 个，这个不像是狗狗名字，怀疑数据提取过程中出错导致

.duplicated()功能查看得出图片链接有重复，后面需要删掉
根据项目的要求，有 2 个问题是需要明确记录出来（且在后面清洗中进行清洗）的。
1，数据中包含了转发的数据
2，数据中包含了没有照片的数据

查看 source 列，发现里面包含不必要的 html 内容
Text 列，包含 url 链接，后面删掉这些链接和不必要的 html 内容

狗狗的地位可以合并为一列

只有一个观察对象 tweet_id,所以可以将三个数据集可以合并到一起，以 tweet_id 作为主键