twitter_archive：WeRateDog 的推特档案，给定的数据集

image_predictions：图片预测信息文档

tweet_json：每条推特的额外附加数据

df：合并后数据

数据评估以及数据

1. 推特档案

twitter_archive.head()
twitter_archive.info()
twitter_archive.source.value_counts()
twitter_archive.name.value_counts()
twitter_archive.rating_numerator.value_counts()
twitter_archive.rating_denominator.value_counts()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                    2356 non-null int64
in_reply_to_status_id       78 non-null float64
in_reply_to_user_id         78 non-null float64
timestamp                   2356 non-null object
source                      2356 non-null object
text                        2356 non-null object
retweeted_status_id         181 non-null float64
retweeted_status_user_id    181 non-null float64
retweeted_status_timestamp  181 non-null object
expanded_urls               2297 non-null object
rating_numerator            2356 non-null int64
rating_denominator          2356 non-null int64
name                        2356 non-null object
doggo                       2356 non-null object
floofer                     2356 non-null object
pupper                      2356 non-null object
puppo                       2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

缺失值较多的主要集中在转发的数据，source 不应为 HTML 格式，特殊评分分子分母值出现错误，name 列数据不整洁，doggo、floofer、pupper、poppo 有清洁度问题。

2.预测文件

image_predictions.head()
image_predictions.sample(5)
image_predictions.p1.value_counts()
image_predictions.describe()
image_predictions.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

未出现数据缺失情况，p1、p2、p3 以及其预测数据过于复杂可以选取可信度最高的数据。

3.附加推特文件

tweet_json.head()
tweet_json.sample(5)
tweet_json.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 3 columns):
tweet_id          2352 non-null int64
retweet_count     2352 non-null int64
favorite_count    2352 non-null int64
dtypes: int64(3)
memory usage: 55.2 KB
```

数据格式正确，未出现数据缺失。

清洁度

1.三个数据表可以根据 tweet_id 合并为一个数据集 df。

2.doggo、floofer、pupper、puppo 不应作为四列，从 text 中提取数据建立狗的 status 的新列，并删除以上四列。

最终需要清理的数据集 df 的简要信息如下：

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 27 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
rating_denominator            2356 non-null int64
name                          2356 non-null object
jpg_url                       2075 non-null object
img_num                       2075 non-null float64
p1                            2075 non-null object
p1_conf                       2075 non-null float64
p1_dog                        2075 non-null object
p2                            2075 non-null object
p2_conf                       2075 non-null float64
p2_dog                        2075 non-null object
p3                            2075 non-null object
p3_conf                       2075 non-null float64
p3_dog                        2075 non-null object
retweet_count                 2352 non-null float64
favorite_count                2352 non-null float64
status                        423 non-null object
dtypes: float64(10), int64(3), object(14)
memory usage: 515.4+ KB
```

整洁度

1. 首先删除转发数据，retweeted_status_id 列若不为空值，则代表是转发数据。检查 tweet_id 的重复，删除空值较多的列，以下五列并不需要
retweeted_status_id
retweeted_status_user_id
retweeted_status_timestamp
in_reply_to_status_id
in_reply_to_user_id

2.数据的行数应该与包含预测数据的 jpg_url 相同，因此删除 jpg_url 为空的行，

完成 1,2 步骤后数据集信息如下所示：

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 2355
Data columns (total 22 columns):
tweet_id              1994 non-null int64
timestamp             1994 non-null object
source                1994 non-null object
text                  1994 non-null object
expanded_urls         1994 non-null object
rating_numerator      1994 non-null int64
rating_denominator    1994 non-null int64
name                  1994 non-null object
jpg_url               1994 non-null object
img_num               1994 non-null float64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null object
p2                    1994 non-null object
p2_conf               1994 non-null float64
p2_dog                1994 non-null object
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null object
retweet_count         1994 non-null float64
favorite_count        1994 non-null float64
status                342 non-null object
dtypes: float64(6), int64(3), object(13)
memory usage: 358.3+ KB
```

3. source 列是 html 格式，不便观察。因此通过正则表达式，用 str.extract()提出标签内的内容重新放入 source 列中。

4. 特殊计分系统，整理步骤如下：
1) 调整分子分母，检查分母不等于 10 的行
2) 分母为 int，分子为 float 形式，通过正则表达式找到 text 中的浮点数值
3) 检查观察到的异常值，即分子为 1776 和 420 的数据
4) 创建新列 score 为分子分母的比值

5. 筛选预测的数据，变为两列 breed 以及 p_conf。选取检测结果为狗的可信度最高的狗的品种，以及其相应的数值，若前三可信度都不是狗狗，则数据缺失。最后再删除 p1、p2、p3 以及其相关数据。

6. 统一 breed 格式，均设置为首字母大写，连接符号设置为空格。

7. 整理 name 列，寻找 text 列中命名规律，正则匹配。

8. 根据需要，删去不研究的列，根据所属列的数据，重新定义数据集的数据类型。

twitter_archive_master：清理完的数据集

简要信息如下：

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1991 entries, 0 to 2355
Data columns (total 13 columns):
tweet_id              1991 non-null int64
timestamp             1991 non-null datetime64[ns]
source                1991 non-null category
rating_numerator      1991 non-null float64
rating_denominator    1991 non-null int64
name                  1354 non-null object
img_num               1991 non-null int64
retweet_count         1991 non-null int64
favorite_count        1991 non-null int64
status                342 non-null object
score                 1991 non-null float64
breed                 1685 non-null object
p_conf                1991 non-null float64
dtypes: category(1), datetime64[ns](1), float64(3), int64(5), object(3)
memory usage: 204.3+ KB
```