一：本项目提供了三个文本数据集：twitter-archive-enhanced（主数据集，包括推特的各项信息），image-predictions（狗狗的品种预测信息），image-predictions（推特的转发数和评论数），读入后生成 3 个 dataframe 文件：df1,df2,df3。

三个表格信息如下：

df1:

<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

df2:

```
<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
```

df3:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 3 columns):
tweet_id          2352 non-null object
retweet_count     2352 non-null int64
favorite_count    2352 non-null int64
dtypes: int64(2), object(1)
memory usage: 55.2+ KB
```

二：导入后，发现如下问题需要清理：

- **质量问题：**

- **表 1：**

1) 1，表 1 这些列的数据类型需要修改：tweet_id，timestamp，retweeted_status_timestamp，doggo，floofer，pupper，puppo。 -
2) 2，查看到有 181 条 tweet 是转发的，需要删除。
3) 3，评分的分子分母存在多处错误，跟提取文本不一致，需要更改。
4) 4，发现 name 列很多错误，例如 the,a，考虑通过正则表达式提取名字信息。
5) 5，expand url 列部分值为空，后续跟表 2 采用 Inner 方式合并，可以直接过滤掉没有图片的行，所以暂不处理。

- **表 2：**

1) 1，表 2tweet_id 数据类型有误。
2) 2，表 2 的图片链接有重复项，后续跟表 1 合并采用 Inner 方式直接清理（表 1 提前删除了转发的行）。
3) 3，p1,p2,p3 狗狗的品种有_号，部分没有大写。

- **表 3：**

  - 无

- **整洁度：**

- 表 1：

1) 1，通过初步了解，表 1 的结构存在问题，后四列狗狗种类可以归为一列来描述。
2) 2,部分列不需要,或者处理后不需要,可以删除：in_reply_to_status_id，in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp
3) 3，表 1，表 3，表 3 可以通过 tweet_id 结合为一张表。方便分析。

- 表 2:

1) 1，根据 tweet_id 外的其他列可以提取出两列，表示可能性最高的狗狗品种预测。

- 表 3：无

三：通过以上清洗，最后合并三个 dataframe 输出文本数据 twitter_archive_master.csv，作为分析的素材。信息如下：

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1993 entries, 0 to 1992
Data columns (total 25 columns):
推特 ID　　1993 non-null object
时间　　1993 non-null datetime64[ns]
来源　　1993 non-null object
文本　　1993 non-null object
链接　　1993 non-null object
分子　　1993 non-null float64
分母　　1993 non-null float64
狗狗名字　　1377 non-null object
评分　　1993 non-null float64
狗狗种类　　1993 non-null object
图片链接　　1993 non-null object
预测图片编号　　1993 non-null int64
P1 品种　　1993 non-null object
P1 可信度　　1993 non-null float64
P1 是否为狗　　1993 non-null bool
P2 品种　　1993 non-null object
P2 可信度　　1993 non-null float64
P2 是否为狗　　1993 non-null bool
P3 品种　　1993 non-null object
P3 可信度　　1993 non-null float64
P3 是否为狗　　1993 non-null bool
推测品种　　1993 non-null object

可信度　　1993 non-null float64

转发数　　1993 non-null int64

点赞数　　1993 non-null int64

dtypes: bool(3), datetime64[ns](1), float64(7), int64(3), object(11)

memory usage: 364.0+ KB