数据清理报告

# 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。

# 数据评估

## 质量问题

### **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

### **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。

### **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。

## 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

# 数据清理

## 清理数据质量问题

### 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。

### 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。

### 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

### 清理一致性问题[¶](http://localhost:8888/files/DAND1/P2/wrangle_act_Ver1.0.html#%E6%B8%85%E7%90%86%E4%B8%80%E8%87%B4%E6%80%A7%E9%97%AE%E9%A2%98)

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

### 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。

## 清理数据整洁度问题

### 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。

### 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'