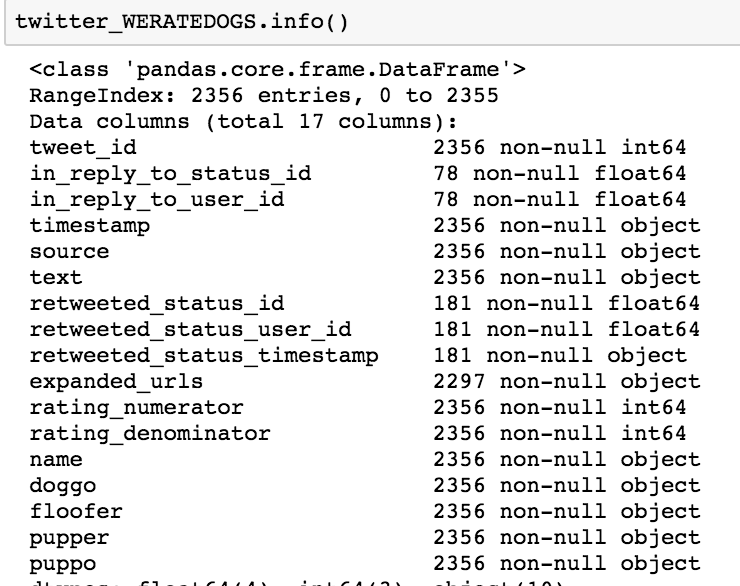
数据清洗报告

# 一、数据收集

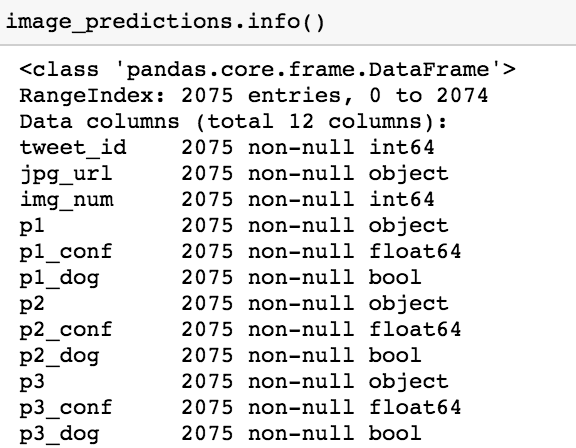
收据收集包含三个表格，1) twitter\_WERATEGODS, 2)

1. “**twitter\_WERATEGODS**”: WeRateDogs的推特档案，包含了此博主的微博数据，从官网直接下载
2. “**image\_predictions”**: 神经网络数据，通过机器学习判断图片中狗的类型，为也是从官网提供的链接下载，但采用的是Python的Requests方法
3. “**tweet\_data**”: 每条推特的数据，包括tweet\_id，被赞次数和被转次数，数据集从官网直接下载，使用pandas中的read\_json读取

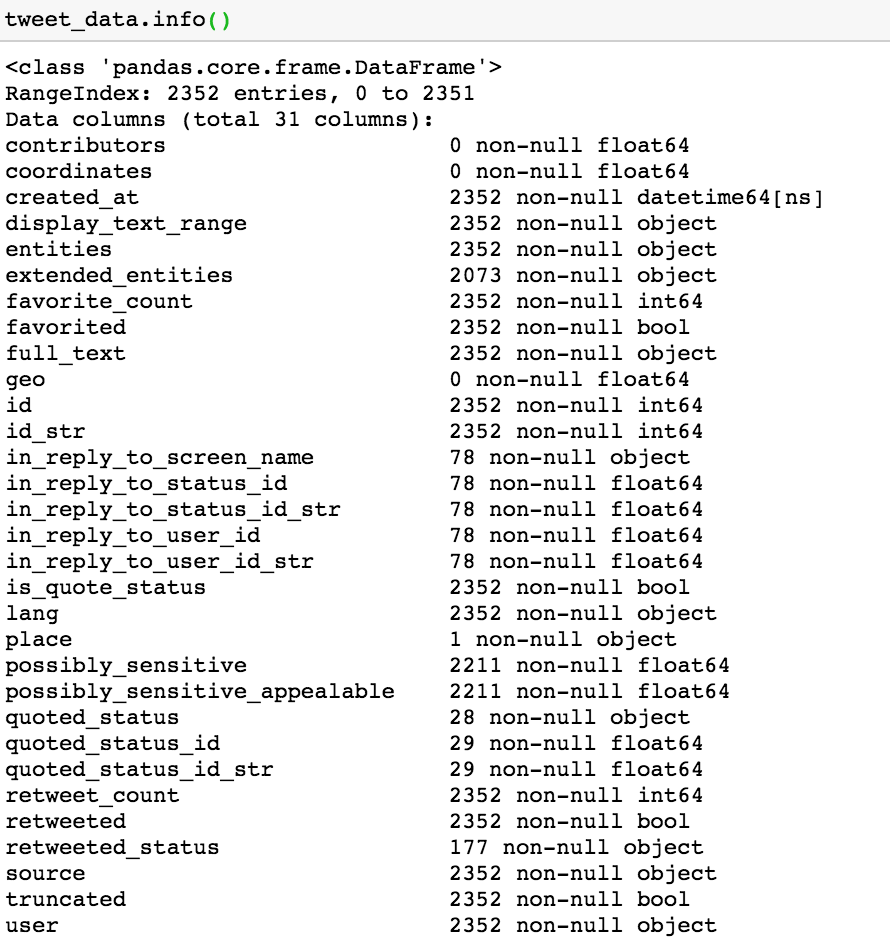
twitter\_WERATEDOGS基本信息:



image\_predictions基本信息



**tweet\_data基本信息**



# 二、 数据评估

通过对三个表格的观察，可以发现有如下问题：

### 数据质量：

表格”twitter\_WERATEGODS”中有如下质量问题：

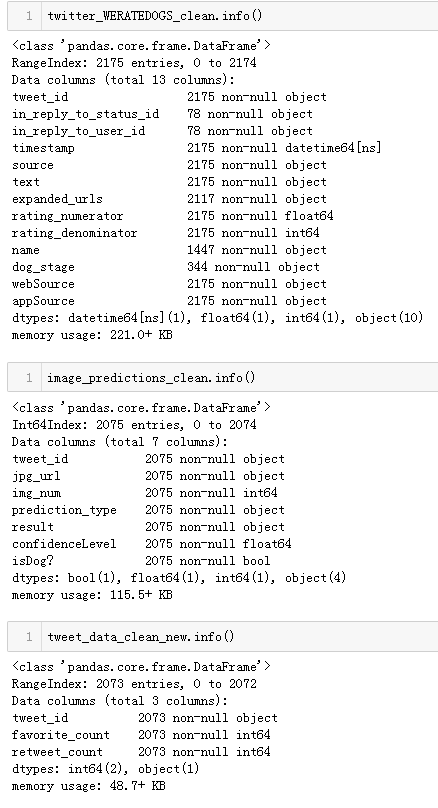
1. 转发的推特要从表格中剔除
2. Tweet id的数据类型应该是string
3. In\_reply数据为float，应该改成int并且转化为string
4. Time-stamp在里面是string类型，应该转化为datetime类型
5. 表格中content source一栏中的<ahref =xxx, >来自于提取中的问题，为多余信息，并且content source包含了网址和实用平台两个信息
6. dog’s name 中有提取错误，比如 ‘a’, ‘an’, ‘such’明显不是狗的名字
7. Rating Denominator中有2 和0这两个错误，来源于提取了错误的信息，这两行的rating数据都要修改
8. Rating numerator没有正确处理小数，比如9.75被错误的取成了75
9. Text中“/r/n”这样的冗杂信息会影响分析

### 数据整洁度：

1. “twitter\_WERATEDOGS”: 最后4列的信息应该被合并成一列”dog\_stage”
2. “image\_predictions”: 每个变量应该为1列，所有(p1,p2,p3), (p1\_conf, p2\_conf, p3\_conf), (p1\_dog, p2\_dog, p3\_dog) 应该合并成3列
3. “tweet\_data”: 中包含太多不需要的信息，提取其中的tweet\_id， favourite\_count 和retweet\_count来制作一张新的表格

# 三、 数据清洗

经过清理后的三张表格的简要形式如下：



# 四、构建新表twitter\_archive\_master

## 重新检查

1. 质量问题
2. 整理过程中发现Tweet\_data\_clean\_new 有重复的tweet\_id
3. 表中有79个facorite\_count为0。0与其他数值相差太远
4. 整洁度
   1. 将三表以tweet\_id为key合并成一张twitter\_archieve\_master表格

## 清理

1. 发现质量问题的a) 和b）的解决办法是一样的，只要将favorite\_count的行去除就行，因为所有重复tweet\_id的行都是只重复了1次，并且其中一行的favorite\_count数值为0
2. 用pd.merge中inner的方法，以twitter\_WERATEDOG\_cl0065an表格为基础，建立新表格

