清洗与分析数据

Kipmin | 数据分析（进阶） | 2018-02-11

收集数据

下载已有的数据文件 **twitter_archive_enhanced.csv**

从网络中收集其他数据 **image-predictions.tsv**

从 twitter-archive-enhanced.csv 文件中获取的 tweet_id 来获取每条推特详细信息保存到 **tweet_json.txt**

评估数据

将数据文件用 pandas 读取到 jupyter 中，生成 3 张表，分别是：

twitter_archive_enhanced.csv→twitter_archive_enhanced

image-predictions.tsv→image_prediction

tweet_json.txt→tweet_json

通过 pd.options 将表修改成能展示全部内容

检查数据格式类型发现如下问题并提出了解决方案：

质量问题

`twitter_archive_enhanced`

- 把所有的 None 替换为空值 NaN
- 评分的分母有时是 10，有时不是，分子实际上有小数的情况，需要重新抓取替换，并删除了异常分数 1776 和 420
- 狗的名字不应该是 a,an,the，把这些都替换为空值 NaN
- 狗的地位信息有大量缺失，留在整洁度解决
- source 列有过多无用 html 代码，删除 html 代码
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 总数太小，可以删除这些行
- timestamp 列不是 date 数据类型，转换为 date 类型
- 有 59 张照片链接缺失，删除缺失行

`image_prediction`

- 有 66 张照片地址重复，可以删除这些行

`tweet_json`

- id 列改为 int 类型

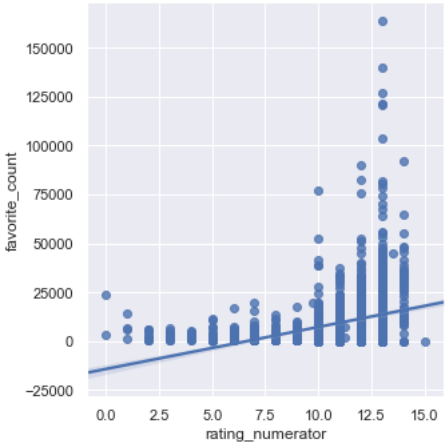
整洁度问题

- 狗狗的地位不需要用四列(doggo,floofer,pupper,puppo)来表示,可以整合为一列
- `image_prediction` 和 `tweet_json` 两表可以合并到 `twitter_archive_enhanced`

分析数据

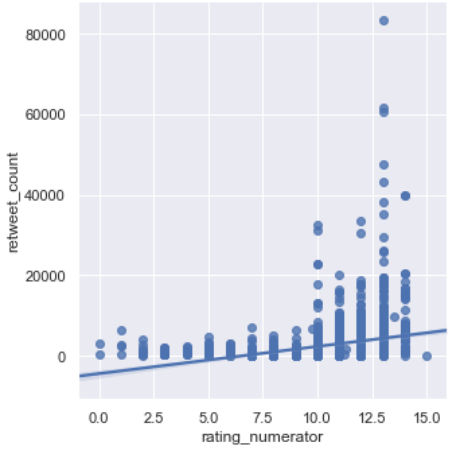
提出 4 个问题：

- 评分高低和点赞数的关系



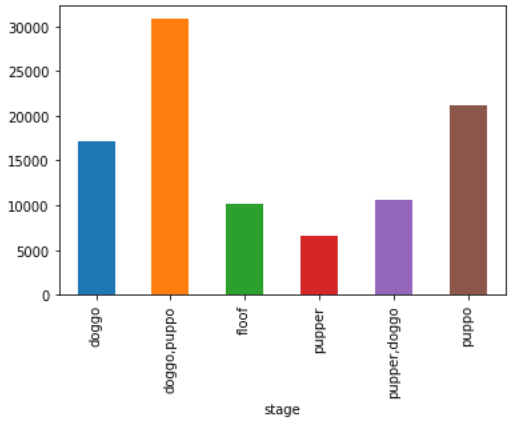
点赞和评分呈现正相关性

- 评分高低和转发数的关系



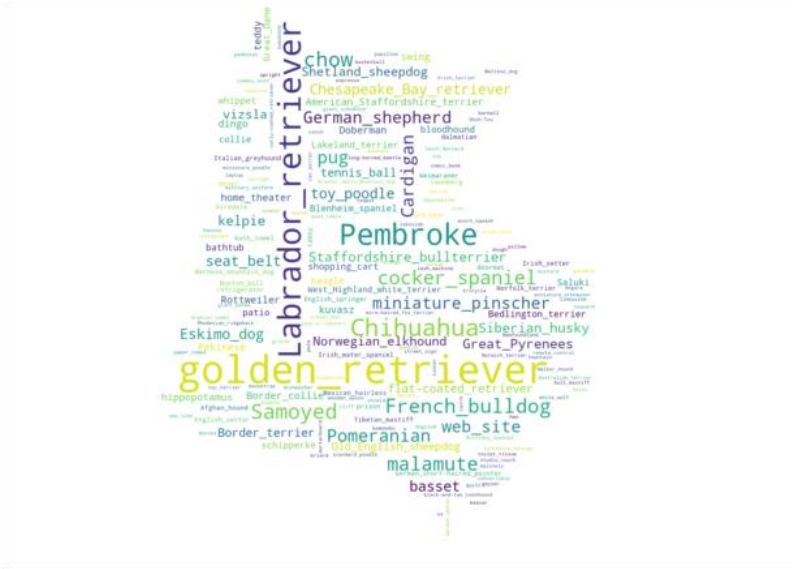
转发和评分呈正相关性

- 地位不同跟人们喜欢程度的关系



人们最喜欢 doggo 与 puppo 地位的狗，最不受欢迎的是 pupper 地位的狗

- 点赞超过平均数的狗中，什么品类的狗狗最多



人们最喜欢且最常看到的狗狗种类是 golden_retriever