

## 1. 收集数据

(1) 导入所需的库

(2) 分别收集文件 1, 2, 3

- 文件 1 : 'twitter-archive-enhanced.csv', 保存为 dataframe
- 文件 2 : 通过 request 得到 url 的 response 的 content, 写入 file "image\_predictions.tsv", 保存数据为 dataframe 类型
- 文件 3 : 'tweet\_json.txt' 写入到 dic 类型, 再写入 dataframe

## 2. 数据清洁-目测评估

(1) df1, rating\_numerator 有非常大的数据, 比如 1776

(2) df1, name 有空值, 但是是字符串类型 ('None'), 还有 a, an, the, as 等错误名字。

(3) df1, retweeted\_status\_id, 有空值

(4) df1, retweeted\_status\_user\_id 有空值

(5) df3, contributors 有空值

(6) df3, display\_text\_range 应该是 2 个分开的列

## 3. 数据清洁-编程评估

方法 : 通过 df1.info(), .describe() 等, 发现问题包括 :

- (1) df1, rating\_denominator 最大数值为 170, 最小为 0, 数据质量有问题
- (2) df1 的 tweet\_id 应该是 object 类型, 而不是 float 类型
- (3) df1 的 time\_stamp 应该是时间类型
- (4) df3 的 created\_at 应该是时间类型

## 4. 数据清洗-质量问题

(1) df1 的 rating\_denominator

初步编程发现, rating\_denominator 只有 23 个数据  $\neq 10$ ,。  
结果 : 直接删除这些行。

(2) df1 的 name 有些为 'None', 还有的是 'a', 'an', 'such', 'the' 等等  
经检查,

①这些不是名字的，除了'None'，其他都是以小写字母开头的，数量有 834 个。

②有一个名字是'O'

结果：

①把这些不是名字的'name' 改为空值 None。

②返回 name 为'O'的行，查看 text 属性，发现 name 应该为'O'Malley"，将其名字更正为'O'Malley"

(3) df1 的 doggo, floofer, pupper, puppo 很多值应该为空值 None，而不是 object 类型的'None'

解决：把这些 object 类型的'None'，改为空值 None

(4) df1 的 timestamp 应该为 datetime 类型，而不是 object

解决：把 timestamp 改为了 datetime 类型

(5) df3 的 created\_at 应该为 datetime 类型，而不是 object

解决：把 created\_at 改为了 datetime 类型

(6) df3 的 lang（语言）最好是 category 类型，方便后续分析

解决：把 lang 改为 category 类型

(7) df3 的 possibly\_sensitive 应该为 bool 类型

解决：把 possibly\_sensitive 改为 bool 类型

(8) df3 的 possibly\_sensitive\_appealable 类型应该为 bool 类型

解决：把 possibly\_sensitive\_appealable 改为 bool 类型

## 5. 数据清洗-整洁问题

(1) df2 中 p1, p2, p3 中狗狗的品种写法不同，有的是大写首字母，有的全小写

解决：全部改为小写。

(2) df3 可以从 user 中提取出 user\_id

解决：从 user 中提取出了 user\_id

(3) df1, df2, df3 可以进行合并

- 把 df1, df2, df3 按照 tweet 的 id 进行合并

- 检查是否有转发的条目并删除。（即 df2.is\_quote\_status 为 true 的行需要删除）

## 6. 数据清洗结果检查

(1) df1.rating\_numerator.unique() 发现，有一些很大的值。

- 编程查找 > 100 的数值，发现只有 5 个，分别是 420, 666, 182, 1776, 420.

- 返回这 5 个数据所在行，进行查看，并未发现异常。

- 继续编程查找 > 15 的数值，发现只有 10 个值。

- note：在以后的分析中，可以将这些值单独拎出来。

**(2) 检查 df2 中，三个概率(p1, p2, p3) 是否都在 0-1 之间**

经验证，都在 0-1 之间。

**(3) 检查是否有无图片记录的数据包含在内**

- 经检查，df.jpg\_url 都不为空，即数据集中不包含没有图片的数据，符合要求。

**(4) 检查是否有转发的数据。**

- 经检查，df.is\_quote\_status 均为 false，即数据集中不包含转发的 tweet 数据，符合要求。

## 7. 存储清理后的主数据集

存储到 .csv 文件。

## 8. 分析和可视化如下：

-----

根据搜集和整理的 twitter 狗狗数据集，我作了如下分析：

**1. 整个数据集长度为 2055，其中 662 个 twitter 中不含狗狗 name，1393 个有 name。问题：是否**

**含有 name 的狗狗评分 ( rating\_numerator ) 会更高？**

**解答：**

由于已经对 rating\_denominator 作过了清洗，目前 rating\_denominator 均为 10，所以可以直接计算 rating\_numerator 来表示狗狗评分。

把 rating\_numerator 按照 name 是否为空分为 2 类。分别计算 2 个类别的 mean, std, min, max, 结果如下：

没有 name 的狗狗：mean = 10.938, std = 16.137, min = 0, max = 420

有 name 的狗狗：mean = 12.070, std = 47.350, min = 2, max = 1776

接着，用统计工具对 2 类评分进行独立 T 检验，得到 T-test 的 p-value = 0.42, 即：两组数据不存在显著性差异。

鉴于清洗数据时，已发现，评分>50 的狗狗数据只有 10 个，而整个数据集有 2055 个。因此，考虑除去高评分的狗狗后，来检验剩余数据集是否由于显著性差异。清理后，计算得到：

没有 name 的狗狗：mean = 10.29, std = 2.63, min = 0, max = 15

有 name 的狗狗：mean = 10.74, std = 1.91, min = 2, max = 14

ttest, p-value = 9.24e-5 < 0.0001 即：两组数据存在显著性差异。有 name 的狗狗比没有 name

的狗狗评分高。

## 2. 狗狗的四种状态：pupper, puppo, doggo, floofer 哪种得到的评分更高？

pupper, puppo, doggo, floofer 分别有 222,24,80,8 个 items

解答：

采用和第 1. 相同的分析方法，得到，四种状态的 mean, std, min, meax 分别为：

```
pupper, [10.78, 2.068, 3, 27],
puppo,  [12.04, 1.241, 9, 14],
doggo,  [11.86, 1.579, 5, 14],
floofer,[11.88, 1.053, 10, 13]]
```

接着用 T-test，对 4 种 status，两两（共 6 种）分别分析得到：

```
【pupper,puppo】Ttest_indResult(statistic=-4.30, pvalue=0.00011562543427274383)
【pupper,floofer】Ttest_indResult(statistic=-2.59, pvalue=0.029297539394663664)
【pupper,doggo】Ttest_indResult(statistic=-4.80, pvalue=3.2077419722681211e-06)
【puppo,floofer】Ttest_indResult(statistic=0.35, pvalue=0.73101372773871087)
【puppo,doggo】Ttest_indResult(statistic=0.57, pvalue=0.57085397032660434)
【doggo,floofer】Ttest_indResult(statistic=-0.03, pvalue=0.97768750087474365)
```

结论：

- ① 从 Ttest 结果来看，pupper 和 puppo 的评分有显著性差异。Puppo 比 pupper 评分高。
- ② 但看均值,puppo 状态的分数最高。

## 3. favorite\_count 和 retweet\_count，哪种狗狗状态(pupper, puppo, floofer, doggo)的平均值最高？

解答：

【favorite\_count】采用和第 1. 相同的分析方法，得到，四种状态的 mean, std, min, meax 分别为：

```
Pupper:[7068, 10617, 0, 106481],
Puppo:[21700, 27028, 0, 132318],
Doggo:[17400, 20830, 0, 130533],
Floofer:[13652, 9803, 2255, 33209]
```

Favorite 点赞数最高前两名：puppo, doggo

【retweet\_count】采用和第 1. 相同的分析方法，得到，四种状态的 mean, std, min, meax 分别为：

```
Pupper:[2583, 3852, 82, 32705],
Puppo:[6923, 9909, 707, 47958],
Doggo: [7584, 12332, 718, 79116],
Floofer:[4745, 5315, 494, 18343]]
```

Retweet 转发数最高前两名：doggo, puppo

同样，分别采用 Ttest 两两进行比较分析，得到：

【favorite】以下 6 组分别是：pupper-puppo, pupper-floofer, pupper-doggo, puppo-floofer, puppo-doggo, doggo-floofer:

```
Ttest_indResult(statistic=-2.5756885009475177, pvalue=0.016667593328511851)
Ttest_indResult(statistic=-1.7447710119304134, pvalue=0.12152662688794623)
Ttest_indResult(statistic=-4.2174636408553452, pvalue=5.6812270244030561e-05)
Ttest_indResult(statistic=1.1932072533572344, pvalue=0.24238244702467776)
Ttest_indResult(statistic=0.70439732409777667, pvalue=0.48638232394644521)
```

```
Ttest_indResult(statistic=0.85498340207604362, pvalue=0.4074472456297511)
```

【rewteet】以下6组分别是：pupper-puppo, pupper-floofer, pupper-doggo, puppo-floofer, puppo-doggo, doggo-floofer:

```
Ttest_indResult(statistic=-2.0836943758867221, pvalue=0.048134865407796282)
```

```
Ttest_indResult(statistic=-1.0671675089577393, pvalue=0.32020703351806906)
```

```
Ttest_indResult(statistic=-3.5430377344657566, pvalue=0.00064677912143014079)
```

```
Ttest_indResult(statistic=0.75556948982200156, pvalue=0.45788108895012303)
```

```
Ttest_indResult(statistic=-0.265972274084469, pvalue=0.79145647754029391)
```

```
Ttest_indResult(statistic=1.1629628956451805, pvalue=0.26304893254071665)
```

结论：在 95%的置信区间内，puppo, doggo 的两者的点赞数和转发数 均明显高于 pupper 的点赞数和转发数。说明大家更喜爱 puppo 和 doggo 的狗狗。

#### 4. 分析转发与点赞数与 tweet 时间的相关性

解答：

我提取了 tweet 的时间的星期几 1：7，月份 1：12 和发布的 hour 0：23。

分别统计在每个时间里的 favorite 和 rewteet 数。

其中蓝色：平均点赞数；黄色：平均转发数

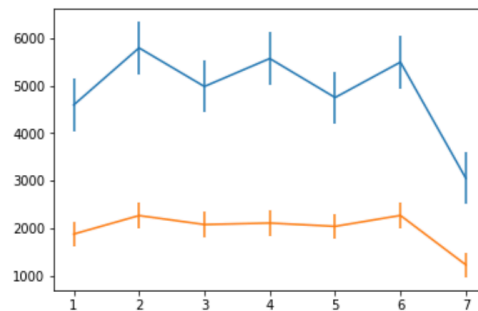


图 1：周一到周日的平均点赞数和转发数

结论 1：周日的点赞数和转发数最低。

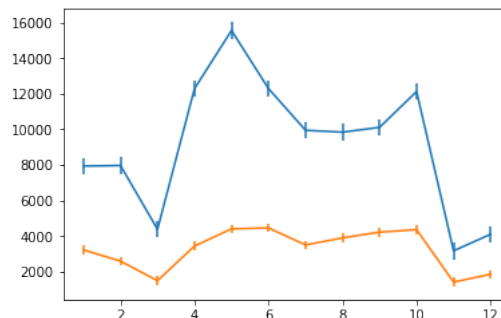


图 2：1 月到 12 月的平均点赞数和转发数

结论 2：3 月份和 11 月的点赞数和转发数最低。5 月份和 10 月份出现 2 个峰值。

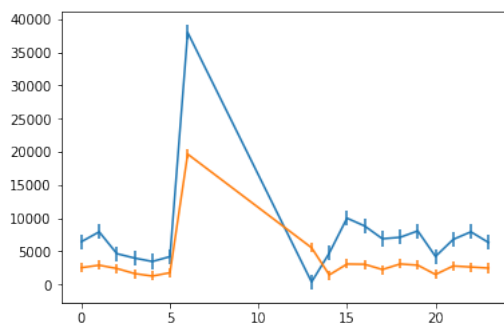


图 3：0 点到 23 点的平均点赞数和转发数

结论 3：舍弃 5 到 13 点的结果（目测是累加结果），发现在晚上 20 点有一个低值点，可能因为是晚餐时间，大家顾不上刷 twitter.

5. 该数据集还可以做很多其他分析，比如：

- (1) 哪些 user\_id 发布的 twitter 评分数，点赞数，转发数最高？
- (2) 哪些狗的品种最受欢迎（评分数，点赞数，转发数）？
- (3) possibly\_sensitive 和 possibly\_sensitive\_appealable 的狗狗特征是什么？