优达数据清洗项目经验分享

这个星期一直在完成优达学城的数据清洗的项目，项目一共有三个数据集，一:从 tweeter_archive_enhanced.csv 中导入的 WeRateDogs 的推特档案的数据，二：通过 Python 的 Request 导入的根据神经网络对推特图像的预测的每个狗狗品种的预测结果，三：我们使用 python 的通过 Tweet 的 API(tweepy)导入的每条推特的额外数据(JSON 格式)。

在经过对数据的依次评估和清洗后，我最终将三个数据集整合成一个新的数据集。

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1976 entries, 0 to 1975
Data columns (total 28 columns):
tweet_id                1976 non-null int64
in_reply_to_status_id    22 non-null float64
in_reply_to_user_id      22 non-null float64
timestamp               1976 non-null datetime64[ns]
source                  1976 non-null object
text                    1976 non-null object
expanded_urls           1976 non-null object
rating_numerator        1976 non-null int64
rating_denominator      1976 non-null int64
name                    1976 non-null object
doggo                   1976 non-null object
floofer                 1976 non-null object
pupper                  1976 non-null object
puppo                   1976 non-null object
create_time             1976 non-null datetime64[ns]
retweet_count           1976 non-null int64
favorite_count          1976 non-null int64
jpg_url                 1976 non-null object
img_num                 1976 non-null int64
p1                      1976 non-null object
p1_conf                 1976 non-null float64
p1_dog                  1976 non-null bool
p2                      1976 non-null object
p2_conf                 1976 non-null float64
p2_dog                  1976 non-null bool
p3                      1976 non-null object
p3_conf                 1976 non-null float64
p3_dog                  1976 non-null bool
dtypes: bool(3), datetime64[ns](2), float64(5), int64(6), object(12)
memory usage: 407.2+ KB
```

	expanded_urls	rating_numerator	rating_denominator	name	...	img_num	p1	p1_conf	p1_dog
	https://twitter.com/dog_rates/status/668633411...	10	10	Churlie	...	1	Pekinese	0.589011	True
	https://twitter.com/dog_rates/status/685325112...	10	10	None	...	1	golden_retriever	0.586937	True
	https://twitter.com/dog_rates/status/696754882...	10	10	Reptar	...	1	weasel	0.137832	False
	https://twitter.com/dog_rates/status/879050749...	11	10	Steven	...	1	tabby	0.311861	False

In [48]:
```
final_df.sample(20)
```

Out[48]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text
1411	678389028614488064	NaN	NaN	2015-12-20 01:38:42	http://twitter.com/download/iphone" rel="nofol...	This is Bella. She just learned that her final...
1415	678278586130948096	NaN	NaN	2015-12-19 18:19:51	http://twitter.com/download/iphone" rel="nofol...	Another spooky pupper here. Most definitely fl...
1218	689143371370250240	NaN	NaN	2016-01-18 17:52:38	http://twitter.com/download/iphone" rel="nofol...	Meet Trip. He likes wearing costumes that aren...
228	836380477523124226	NaN	NaN	2017-02-28 01:00:19	http://twitter.com/download/iphone" rel="nofol...	This is Ava. She just blasted off. Streamline ...
1075	700518061187723268	NaN	NaN	2016-02-19 03:11:35	http://twitter.com/download/iphone" rel="nofol...	This is Vincent. He's the man your girl is wit...
747	747594051852075008	NaN	NaN	2016-06-28 00:54:46	http://twitter.com/download/iphone" rel="nofol...	Again w the sharks guys. This week is about do...
538	780192070812196864	NaN	NaN	2016-09-25 23:47:39	http://twitter.com/download/iphone" rel="nofol...	We only rate dogs. Pls stop sending in non-can...
827	735274964362878976	NaN	NaN	2016-05-25 01:03:06	http://twitter.com/download/iphone" rel="nofol...	We only rate dogs. Please stop sending in your...
1887	667534815156183040	NaN	NaN	2015-11-20 02:47:56	http://twitter.com" rel="nofollow">Twitter Web...	This is Frank (pronounced "Frong"). Too many b...
1207	689661964914655233	NaN	NaN	2016-01-20 04:13:20	http://twitter.com/download/iphone" rel="nofol...	Meet Luca. He's a Butternut Scooperfloof. Glor...

	expanded_urls	rating_numerator	rating_denominator	name	...	img_num	p1	p1_conf	p1_dog
	https://twitter.com/dog_rates/status/668633411...	10	10	Churlie	...	1	Pekinese	0.589011	True
	https://twitter.com/dog_rates/status/685325112...	10	10	None	...	1	golden_retriever	0.586937	True
	https://twitter.com/dog_rates/status/696754882...	10	10	Reptar	...	1	weasel	0.137832	False
	https://twitter.com/dog_rates/status/879050749...	11	10	Steven	...	1	tabby	0.311861	False

	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
	Shih-Tzu	0.390987	True	Japanese_spaniel	0.003310	True
	Labrador_retriever	0.398260	True	kuvasz	0.005410	True
	toy_poodle	0.098378	True	Scottish_deerhound	0.097397	True
	window_screen	0.169123	False	Egyptian_cat	0.132932	False

图 数据集最终展示

随后我按自己的想法对该数据集进行了不同角度的分析和总结：

一、 tweet 用户与 tweet 转发数和喜爱数的相关性分析

我将数据集的评分分子，tweet 转发数，tweet 喜爱数单独提取出来并且通过 pandas 的 corr 函数进行相关性分析

```
In [47]: df_1.head()
```

Out[47]:

	rating_numerator	retweet_count	favorite_count
0	13	8842	39492
1	13	6480	33786
2	12	4301	25445
3	13	8925	42863
4	12	9721	41016

	rating_numerator	retweet_count	favorite_count
rating_numerator	1.000000	0.024073	0.023062
retweet_count	0.024073	1.000000	0.914563
favorite_count	0.023062	0.914563	1.000000

可以看出：

第一：从上面的对 df_1 的相关性分析可以看出 tweet 用户的评分与他们的 tweet 的转发数和喜爱数有着正相关的关系，但是他们的相关性都不太明显

第二：tweet 用户对 tweet 的喜爱数和转发数有着很明显的正相关关系，说明 tweet 用户在喜爱一条 tweet 的时候他们有很大可能会选择转发该 tweet

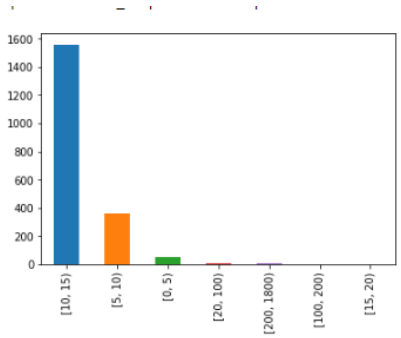
二、 tweet 用户对狗狗们的评分分布

我先大概观察了 tweet 用户评分的统计

```
In [49]: final_df.rating_numerator.value_counts()
Out[49]: 12      450
         10      419
         11      396
         13      261
         9       150
         8        95
         7        51
         14       35
         5        33
         6        32
         3        19
         4        15
         2         9
         1         4
         0         2
         420       1
         26        1
         27        1
         75        1
         1776      1
         Name: rating_numerator, dtype: int64
```

然后我通过 pandas 的分段统计函数 cut，按我预先分好的区间进行统计

```
[10, 15)       1561
[5, 10)         361
[0, 5)           49
[20, 100)         3
[200, 1800)       2
[100, 200)        0
[15, 20)          0
Name: rating_numerator, dtype: int64
```



从上面的统计可以看出：

tweet 用户对狗狗们的评分主要分布在[10, 15)这个区间内，其次分布在[5，10)和[0，5)这个区间内，有少数用户评分打的比较高，分布在[20，1800)区间中。

可以看得出用户们普遍喜欢这个活动的特殊的评分，他们对狗狗的评分基本都超过 10，有些用户的评分尤其高，甚至达到 1776.

三、tweet 用户评分高的图片中，狗占的比重高吗

我先是按 tweet 用户的评分排序，然后将他们按每 100 个人分成一个区间，观察他们的 tweet 的预测结果是否为狗

```
top_1 = sorted_df.iloc[0:100].p1_dog.value_counts()
top_2 = sorted_df.iloc[101:200].p1_dog.value_counts()
top_3 = sorted_df.iloc[201:300].p1_dog.value_counts()
top_4 = sorted_df.iloc[301:400].p1_dog.value_counts()
top_5 = sorted_df.iloc[401:500].p1_dog.value_counts()
top_6 = sorted_df.iloc[501:600].p1_dog.value_counts()
top_7 = sorted_df.iloc[601:700].p1_dog.value_counts()
top_8 = sorted_df.iloc[701:800].p1_dog.value_counts()
top_9 = sorted_df.iloc[801:900].p1_dog.value_counts()
top_10 = sorted_df.iloc[901:1000].p1_dog.value_counts()
top_11 = sorted_df.iloc[1001:1100].p1_dog.value_counts()
top_12 = sorted_df.iloc[1101:1200].p1_dog.value_counts()
top_13 = sorted_df.iloc[1201:1300].p1_dog.value_counts()
top_14 = sorted_df.iloc[1301:1400].p1_dog.value_counts()
top_15 = sorted_df.iloc[1400:1500].p1_dog.value_counts()
top_16 = sorted_df.iloc[1501:1600].p1_dog.value_counts()
top_17 = sorted_df.iloc[1601:1700].p1_dog.value_counts()
top_18 = sorted_df.iloc[1701:1800].p1_dog.value_counts()
top_19 = sorted_df.iloc[1801:1900].p1_dog.value_counts()
top_20 = sorted_df.iloc[1901:2000].p1_dog.value_counts()
```

然后将所有的 Series 重新合并成一个新的 DataFrame

	top_1	top_2	top_3	top_4	top_5	top_6	top_7	top_8	top_9	top_10	top_11	top_12	top_13	top_14	top_15	top_16	top_17	top_18	top_19	top_20
False	25	20	17	14	17	13	18	18	25	23	19	27	26	25	20	25	23	39	54	60
True	75	79	82	85	82	86	81	81	74	76	80	72	73	74	80	74	76	60	45	15

由上面的统计分析可以看到，从 top_1 到 top_17 中狗狗占的比重都比较高，基本分布在（20%，25%）之间，区别不大，但是在后面 top_18 到 top_20 中狗狗占的比重较小。

结论:说明 tweet 用户对图片的评价的高低与图片是不是狗的关系

不是很大。