@WeRateDogs 数据分析

宠物狗评分的分析和可视化

Kipmin | 数据分析（进阶） | 2019-02-18

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

```
In [10]: image_prediction.head(10)
```

Out[10]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	1	Bernese_mountain_dog	0.651137	True	English_springer	0.263788	
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	1	box_turtle	0.933012	False	mud_turtle	0.045885	
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	1	chow	0.692517	True	Tibetan_mastiff	0.058279	
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	1	shopping_cart	0.962465	False	shopping_basket	0.014594	
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	1	miniature_poodle	0.201493	True	komondor	0.192305	

image_prediction

```
In [11]: twitter_archive_enhanced.head(10)
```

Out[11]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retwee
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	Twitter for iPhone	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	Twitter for iPhone	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10 https://t.co/0Xxu71qeIV	
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co/wUnZnhtVJB	
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51	Twitter for iPhone	This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us	

tweet_json

```
In [12]: tweet_json.head(10)
```

Out[12]:

	id	retweet_count	favorite_count
0	892420643555336193	8287	37950
0	892177421306343426	6119	32593
0	891815181378084864	4054	24539
0	891689557279858688	8416	41293
0	891327558926688256	9129	39484
0	891087950875897856	3035	19839
0	890971913173991426	2010	11599
0	890729181411237888	18383	64028
0	890609185150312448	4175	27249
0	890240255349198849	7190	31269

将数据整合到一张表 twitter_archive_master 中

twitter_archive_master

```
In [291]: twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2064 entries, 0 to 2063
Data columns (total 23 columns):
tweet_id              2064 non-null int64
timestamp             2064 non-null datetime64[ns]
source                2064 non-null object
text                  2064 non-null object
expanded_urls         2064 non-null object
rating_numerator      2063 non-null float64
rating_denominator    2064 non-null int64
name                  1421 non-null object
stage                 357 non-null object
jpg_url               2064 non-null object
img_num               2064 non-null int64
p1                    2064 non-null object
p1_conf               2064 non-null float64
p1_dog                2064 non-null bool
p2                    2064 non-null object
p2_conf               2064 non-null float64
p2_dog                2064 non-null bool
p3                    2064 non-null object
p3_conf               2064 non-null float64
p3_dog                2064 non-null bool
id                    2064 non-null int64
retweet_count         2064 non-null int64
favorite_count        2064 non-null int64
dtypes: bool(3), datetime64[ns](1), float64(4), int64(6), object(9)
memory usage: 344.7+ KB
```

tweet_id：推特 ID

timestamp：发推时间

source：使用何种设备发送

text：推文内容

expanded_urls：推文链接

rating_numerator：评分分子

rating_denominator：评分分母

name：宠物名

stage：狗的地位分类

jpg_url： 是预测的图像资源链接

img_num： 最可信的预测结果对应的图像编号 → 1 推特中的第一张图片

p1： 是算法对推特中图片的一号预测 → 金毛犬

p1_conf： 是算法的一号预测的可信度 → 95%

p1_dog： 是一号预测该图片是否属于“狗”（有可能是其他物种，比如熊、马等） → True 真

p2： 是算法对推特中图片预测的第二种可能性 → 拉布拉多犬

p2_conf： 是算法的二号预测的可信度 → 1%

p2_dog： 是二号预测该图片是否属于“狗” → True 真

依次类推

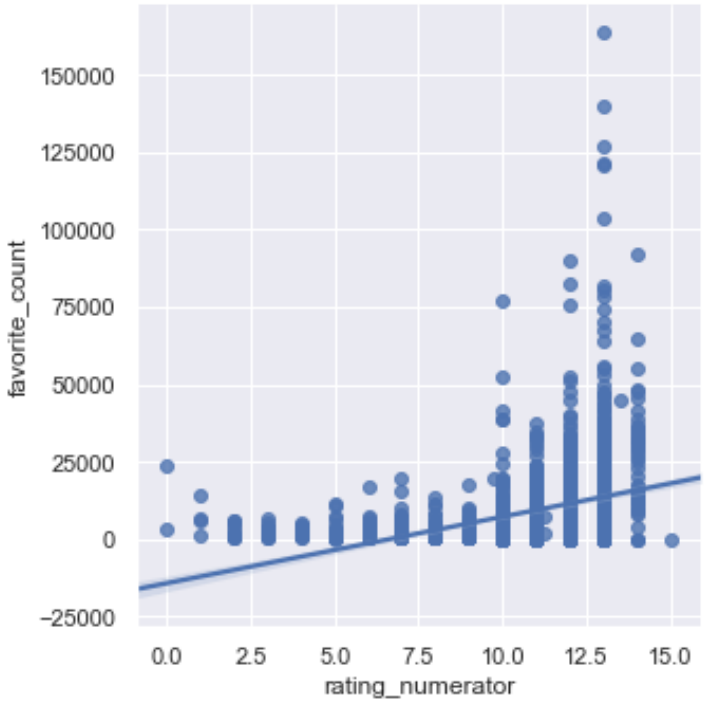
id：等于 tweet_id

retweet_count：转发该推的数量

favorite_count：点赞该推的数量

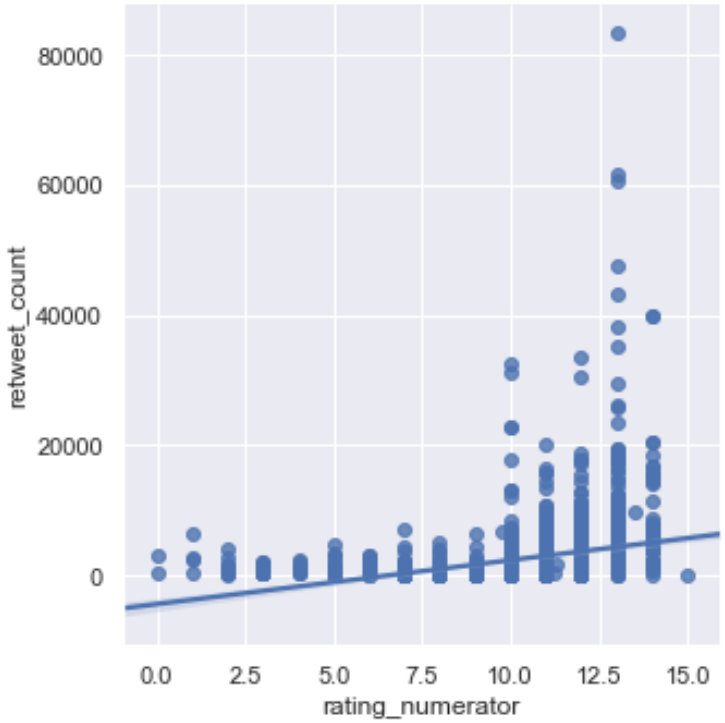
分析与可视化

- **评分高低和点赞数的关系**



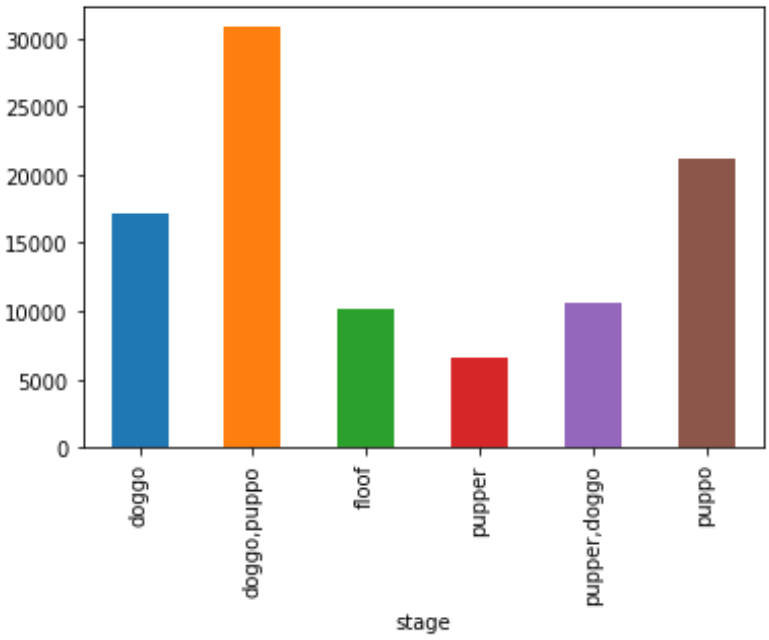
点赞和评分呈现正相关性

- **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**

使用 wordcloud 生成词云

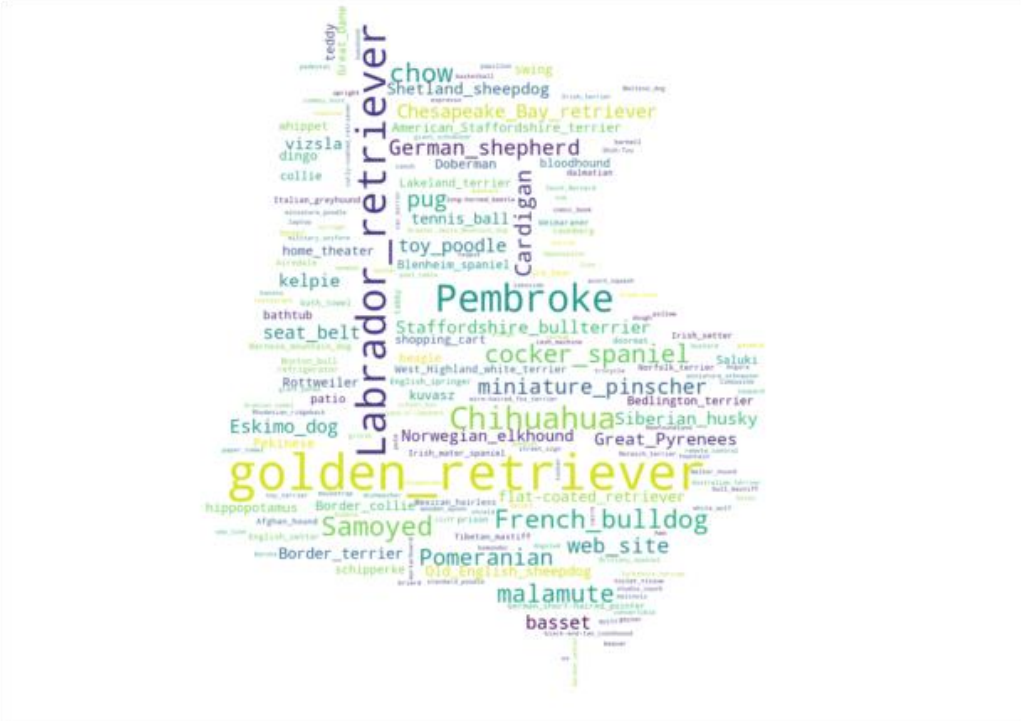
```
fav_mean = twitter_archive_master.favorite_count.mean()
variety = twitter_archive_master[twitter_archive_master.favorite_count >= fav_mean]
data = variety.groupby(['p1']).count()

data = data.tweet_id.sort_values(ascending=False)
variety_dict = data.to_dict()

dog_image = np.array(Image.open('dog_white.jpg'))

wc = WordCloud(background_color="white", max_words=200, mask=dog_image)
wc.generate_from_frequencies(variety_dict);

plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.rcParams['figure.figsize'] = (30.0, 8.0)
plt.show()
```



人们最喜欢且最常看到的狗狗种类是 golden_retriever