数据可视化

1. 通过散点图可视化，来探索retweet_count / favorite_count与变量rating_numerator的关系：

经过观察，整体来说rating_numerator评分分子越大的情况下，转发和最受欢迎的数据整体越大。这在一定程度上表明了用户WeRateDogs的打分准则有一定的代表性，与普通网友的兴趣口味有一定的吻合程度。

Figure 1.1

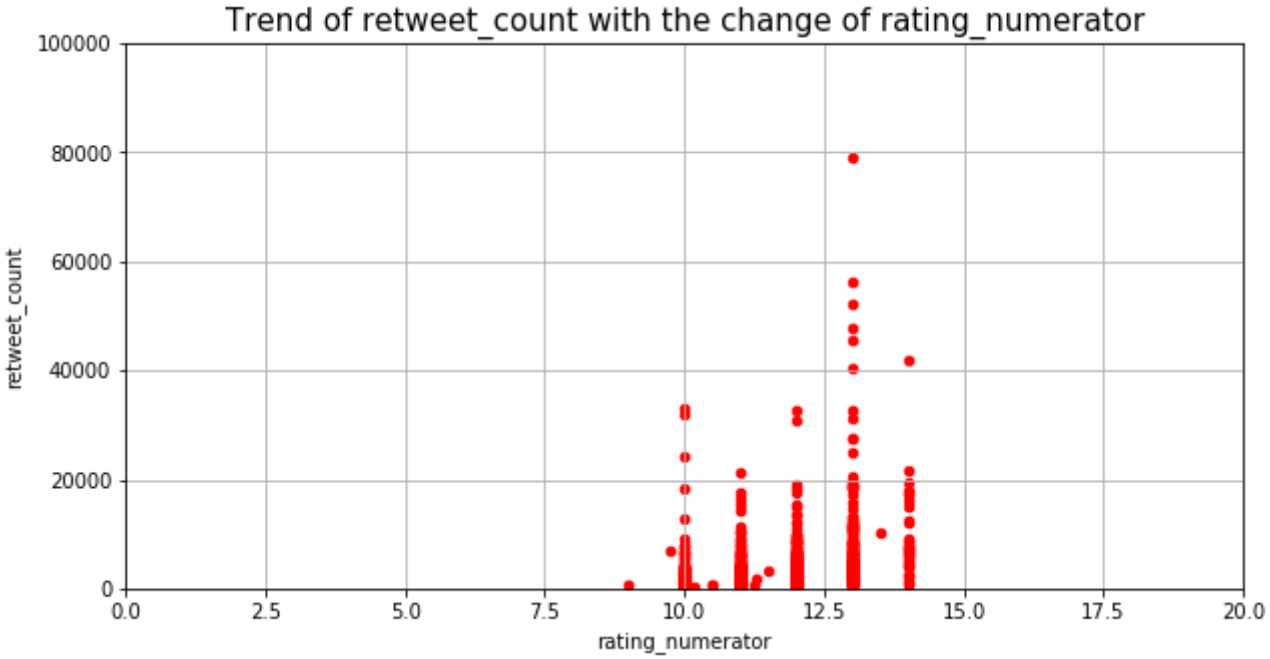
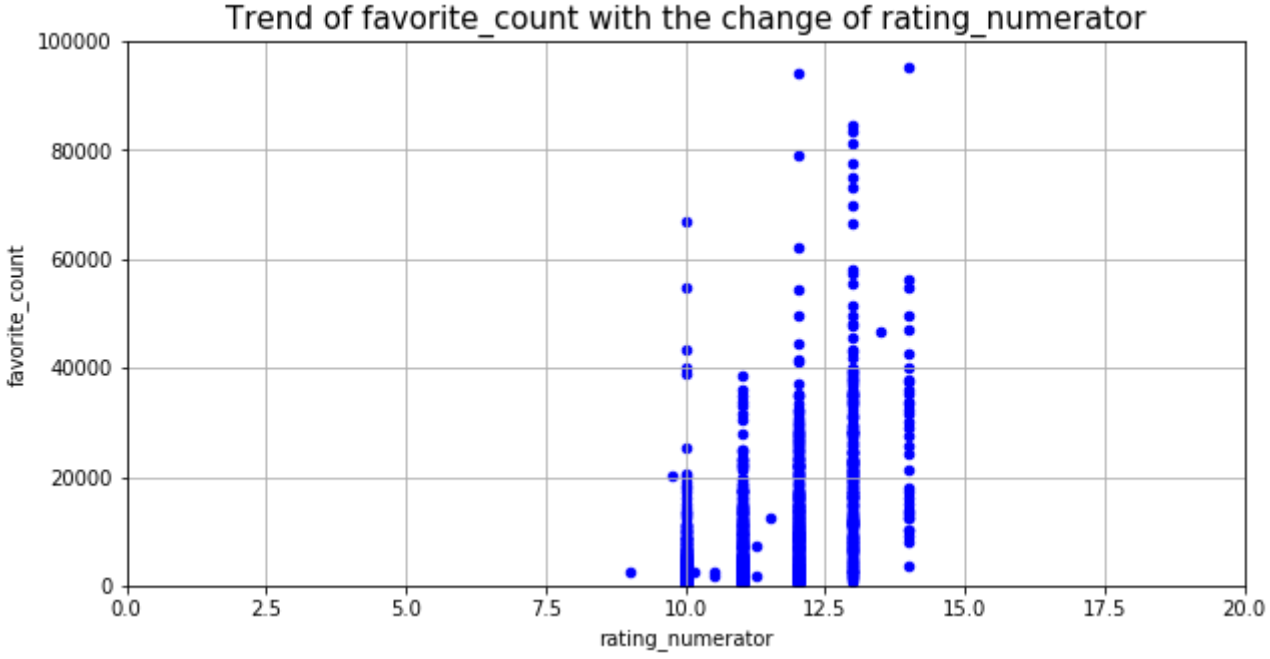


Figure 1.2



```
Conclusion:
- twitter转发数量与评分高低成正比
- twittewr被选为最喜爱的数量与评分高低成正比
```

2. twitter用户WeRateDogs的在线活跃时间段分析

该用户在的活跃时间有明显的差异性，打分主要集中在凌晨时间段。经过细致观察可以看出，该用户在凌晨5点之后几乎很少有活跃，早上7点到中午12点之间数据缺失，说明该用户在该时间段内从来没有进行过评论（推测：很大可能在睡觉），猜测用户WeRateDogs应该是一个典型的“夜猫”。

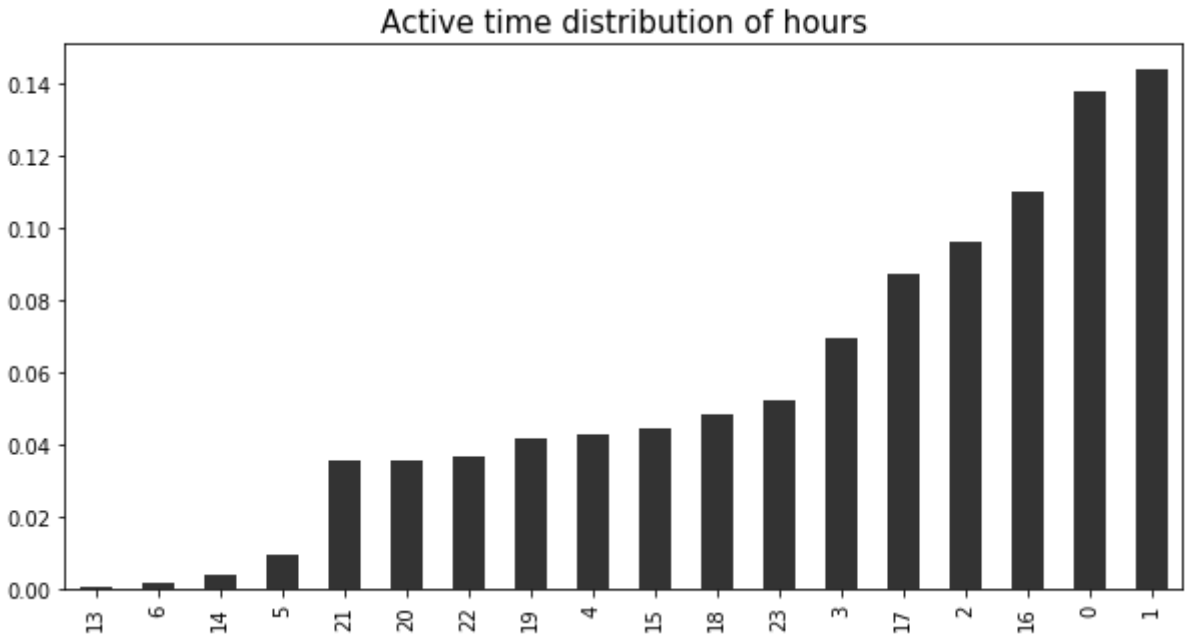


Figure 2.1

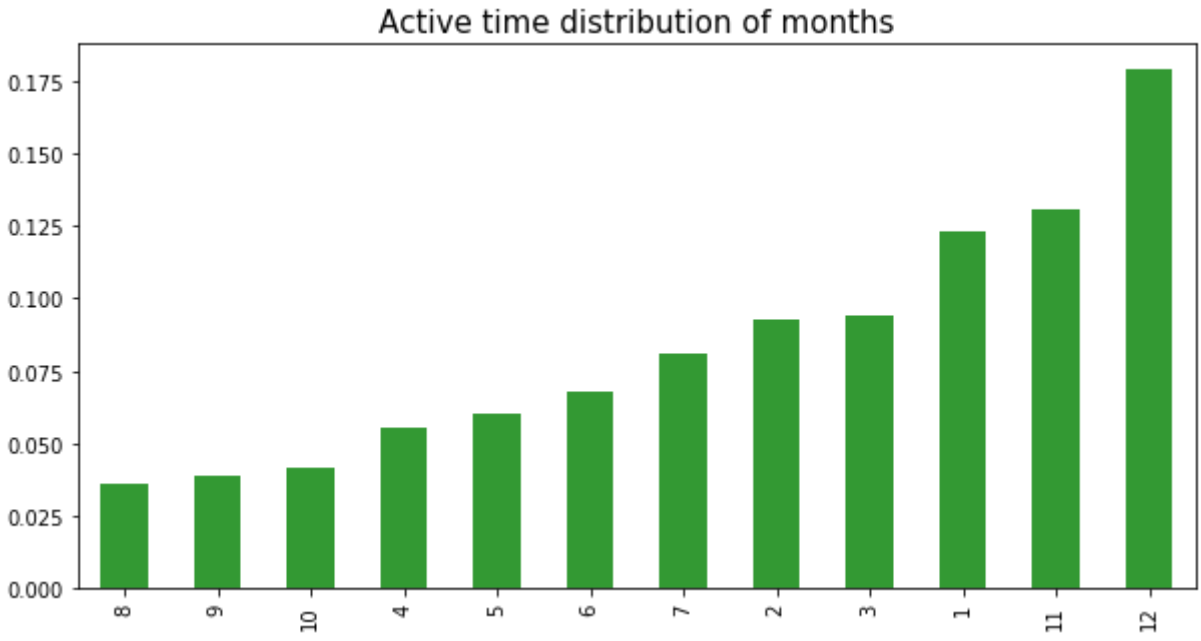


Figure 2.2

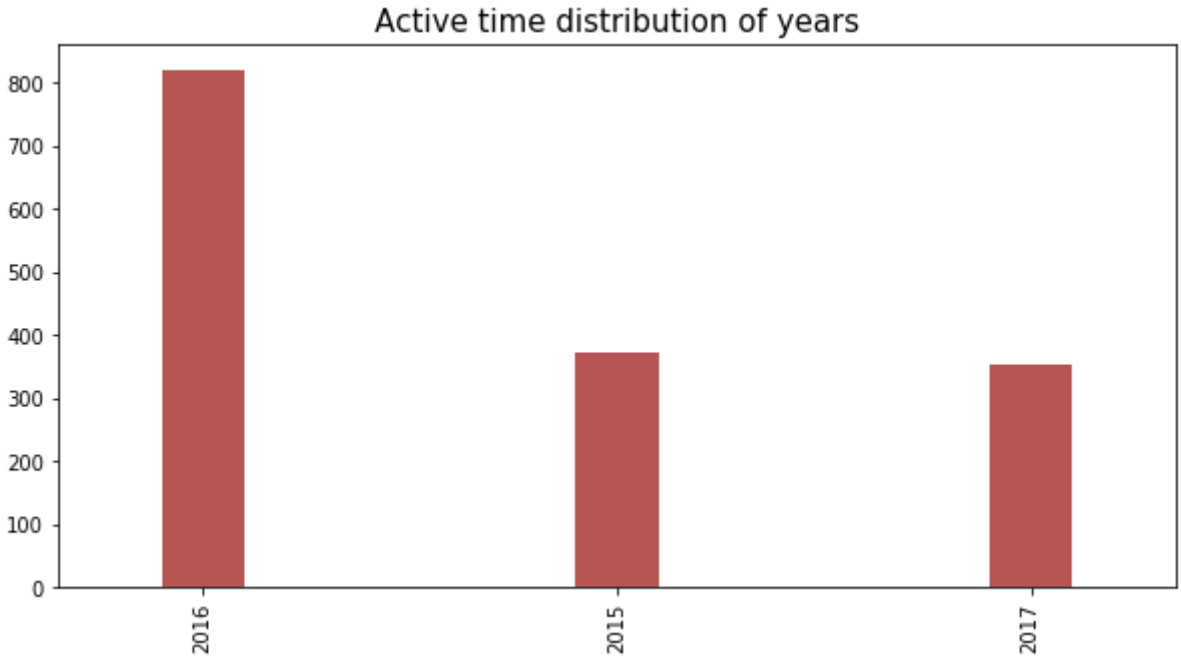


Figure 2.3

```
Conclusion:
- 推特用户WeRateDogs通常最活跃时间在一天内的凌晨时间段（0点 - 1点）
- 推特用户WeRateDogs在一年内打分最多的时间段为年底时间段（11 - 12月份）
- 从2015年到2017年，该用户WeRateDogs的评分活动逐渐减少，热度下降
```

3. 最受欢迎的狗狗名字排行前10

经过观察，在该数据集中有大约有1/3的狗狗是没有成功提取到名字的，因此仅仅对有名字的狗狗来进行分析，发现狗主人取名字的问题上，也存在一些“比较流行”的名字，比如Charlie, Lucy, Cooper, Oliver, Tucker... 都多次重名，说明这些名字在狗狗名字中比较普遍。

Figure 3.1

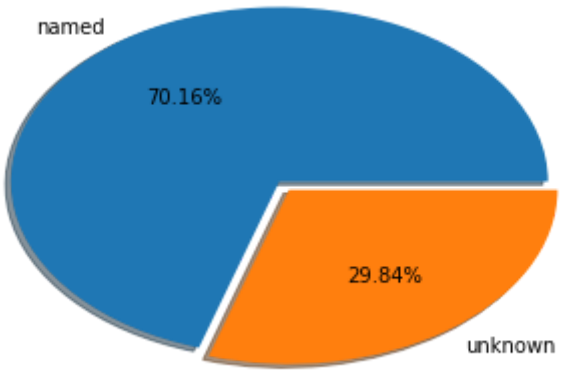


Fig_6

Figure 3.2

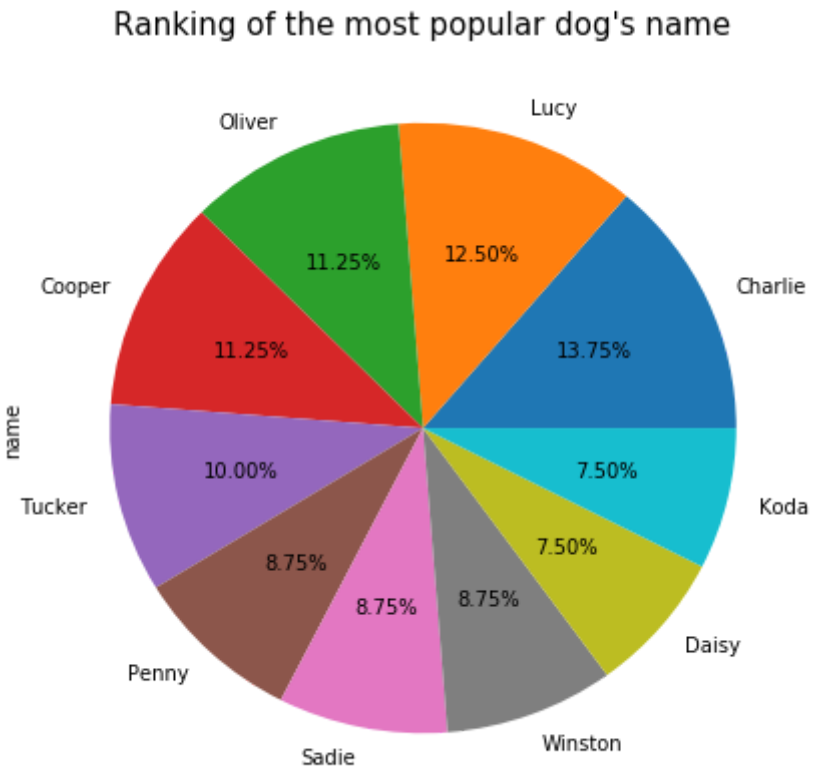


Fig_7

```
Conclusion
- 在该数据集中，有名字的狗狗大约占70%
- 在有名字的狗狗中，Charlie，Lucy, Oliver, Cooper, Tucker, Penny, Winston,
Daisy, Koda都是比较受欢迎的名字
```

探索性可视化心得：

刚开始准备进行可视化部分的工作时，不知道从哪里入手，对哪些变量进行分析，经过一定的思考后，决定先尝试对int或float类型的变量进行联系，试图找出其中的规律。

经过一定的尝试后联想到转发数和最喜爱数在一定程度上反应了大众对该图片的欢迎程度，就拿这两个指标和twitter用户WeRateDogs给出的评分放在一张散点图上，得到了一定的结论。

以此类推，将时间戳timestamp中的年月日分开，经过可视化后即可明显地反应出该用户的日常作息规律。