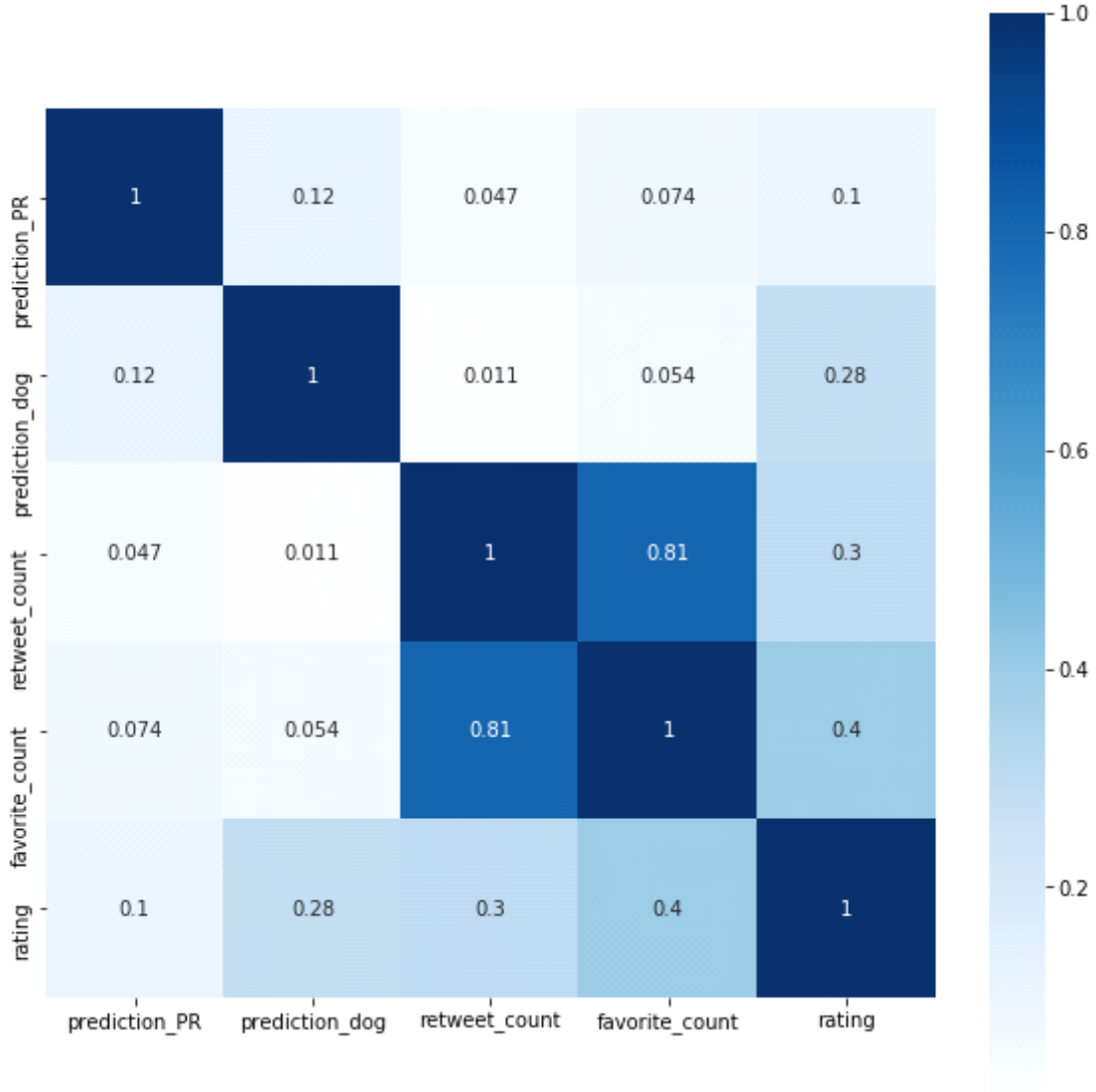
14.art_report

2018年6月6日星期三 10:31

一 探索性数据分析

1. **研究问题1：favorite_count点赞数与哪一个相关性最高？**

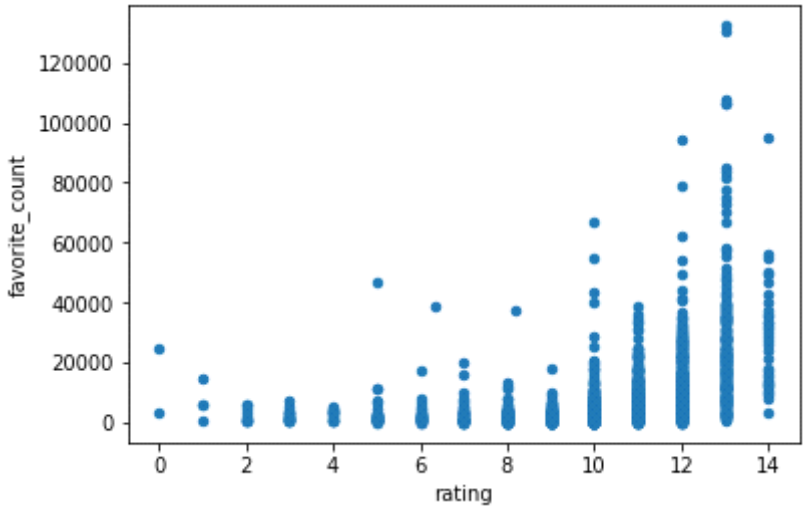
- 首先将清理干净的数据集twitter_archive_master复制一份
- 对数据集中的数值变量输出热力图探究变量之间的相关性，从图中得出以下结论：
  a. 从上图可以很容易的看出来favorite_count与retweet_count有极高的相关性，这与现实情况也相符，一般点赞多的会比计较有趣，大家也愿意转发给别人看。
  b. 意外的是，favorite_count与rating相关性貌似有一定的相关性，到底是怎样的相关性呢，在这个热力图中不明显，我们后续接着分析。
  c. 热力图见下：
  d.



- 探究过程中发现很在意c问题，于是继续探究，引发了问题2的探究

2. **研究问题2：favorite_count点赞数与rating有无相关性？**

- 首先将清理干净的数据集twitter_archive_master复制一份

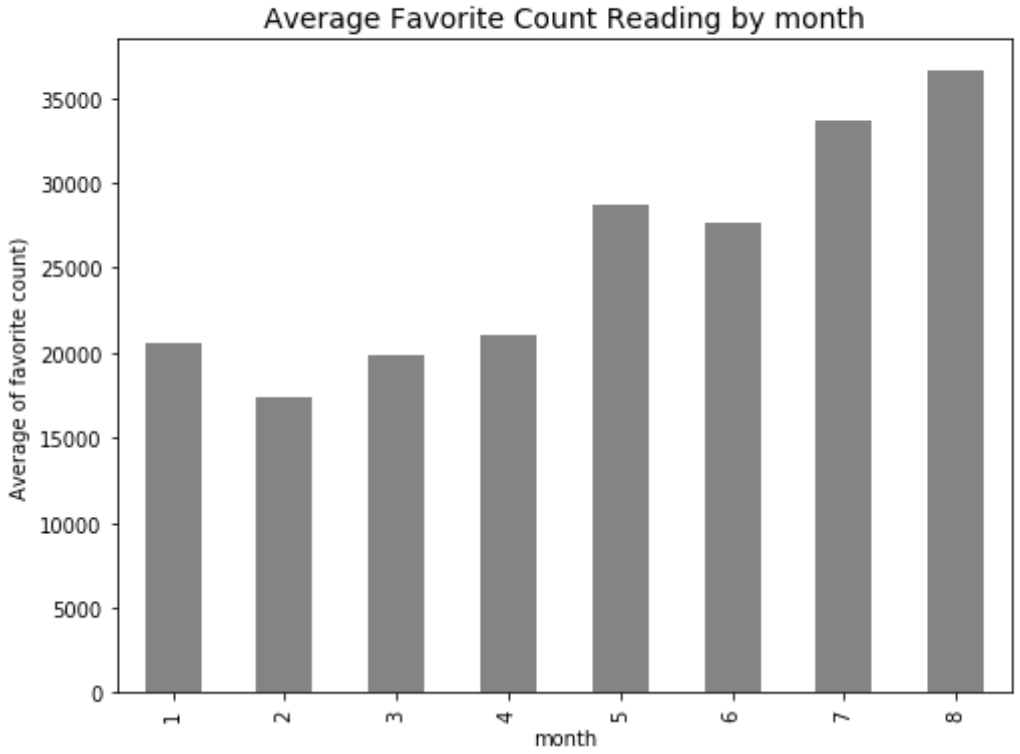
- 研究这个题目的原因是我想看下被自己主人非常钟爱的狗狗（rating得分高），到底会不会被其他人也喜欢（点赞数高）？
- 输出散点图，相关性比较明显，可视化图形见下：
-


- 对图形的解读见下：
  a. 从上图可以很容易的看出来favorite_count与rating有一定相关性的，图形大致符合一个负偏态；
  b. 也就是说你非常喜欢自己的狗狗，给到高分，就有可能得到更多的点赞；想反的要是你自己都不喜欢自己的狗狗，肯定别人也不会喜欢
  c. 从图形中可以看出得分再13分左右的狗狗比较受大家的青睐，毕竟自己的主人喜欢，也不过分吹捧，这样会比较招大家喜欢。

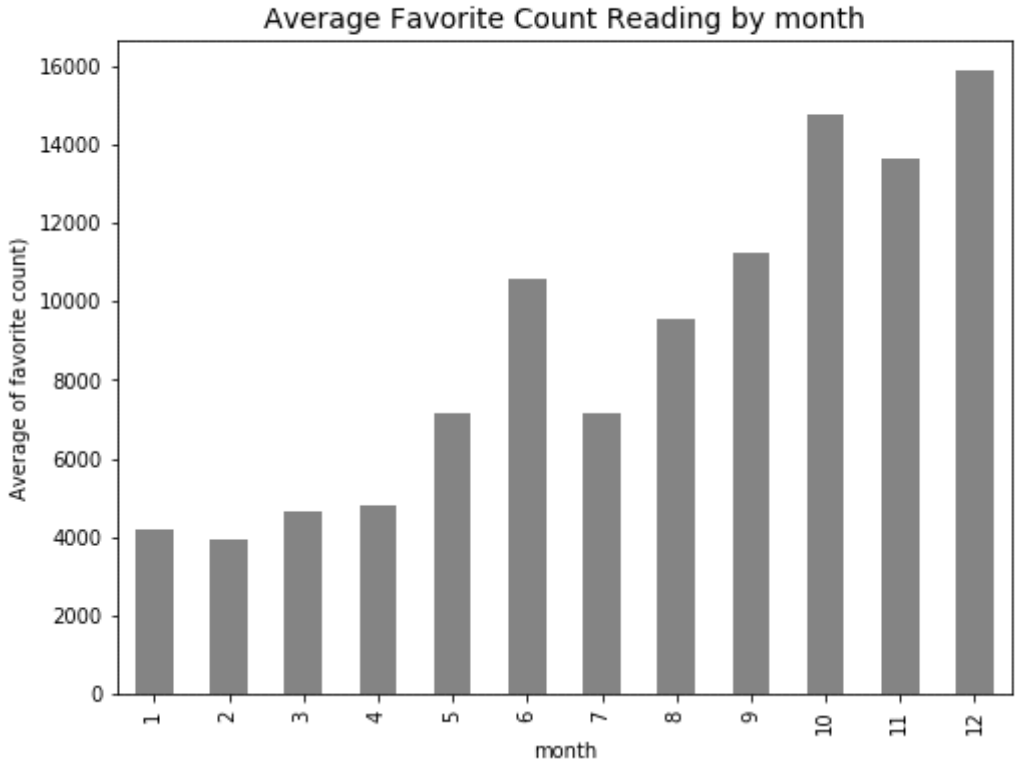
3. 研究问题3：什么时间段（哪年哪月）这个活动达到的热度的顶峰。

- 这毕竟是一个全民点赞的活动，于是我对这个活动的热度非常感兴趣。这个问题的目标是找出哪一个月的favorite_count点赞数的均值最高；
- 首先将清理干净的数据集twitter_archive_master复制一份；
- 将timestamp列分解成只显示年份和月份的两列，这样便于后面作为key取出相应的数据，同时删除对这个问题研究无关的列；
- 定义删选数据的reading_stats函数，这个函数可以输入需要筛选的df和筛选条件，然后给你输出符合该筛选条件的df、该df的箱线图以及一些统计参数；
- 利用reading_stats函数分别取出2015-2017年的数据，供后面可视化使用；
- 定义一个可视化函数：univariate_plot函数，这个函数可以输入需要筛选的df和key，然后给你输出符合该key的直方图，于是我进行2015-2017年每年的月份直方图输出，得到如下三个图：

```
univariate_plot(df_2017 , 'month', 'grey')
```




```
univariate_plot(df_2016 , 'month', 'grey')
```



- 结合相线图和柱状图的解读见下：
  a. 从三张箱线图可以看出，这个活动虽然始发于2015年，但是直到2018年才是真正大火起来，因为2016年一共有1310条推特，点赞数的均值却比2017年466条推特的均值少太多。
  b. 结合后面的柱状图可以得出，从2016年7月开始，这个活动就持续上升，一直延续到2017年9月，由于后面数据看不到了，但是估计上升态势应该还在持续。
  c. 由此可得出，从目前数据集来看最高的点赞月是2017年9月。

二 得出结论

- 首先需要声明这次数据分析的数据清洗部分，可能不能完全清洗干净，所以可能会影响到后续的可视化分析阶段，但是我已经尽量避免使用不易清洗干净的数据了；
- 与favorite_count相关性最高的是retweet_count，有趣的是favorite_count也许与rating也有一定的相关性，但更细致的情况在热力图中看不出来，所以再第二年问中再细致分析。
- 狗狗得分（rating）在13分左右的狗狗比较受大家的青睐，毕竟自己的主人喜欢，也不过分吹捧，这样会比较招大家喜欢，而得分过低的狗狗大家不会太去关注；
- 这个活动虽然始发于2015年，但是直到2018年才是真正大火起来，从2016年7月开始，这个活动的热度就持续上升，一直延续到2017年9月，由于后面数据看不到了，但是估计上升态势应该还在持续；
- 这是一个活动策划者值得深挖的好例子。