Research on Celebrities with Huge Followings

Reporter：DongXingchen　ZhengYihang　Time：December 13th,2020

CONTENTS

Background

「 The reason why we choose this topic 」

Present of Celebrities

- In rencent years, celebrities with huge following have become a very popular concept.

- When an actor becomes a star, the audience's attention shifts from acting skills and works to the star himself. In order to fully tap the commercial value of stars and extract profits to the maximum extent, capital vigorously cultivates "fan economy" and “fans culture" in the entertainment market. Under the strict control of corporatization and organization, the fanatical "fan support" effect makes every word, every move and every smile of the stars get a surprising market premium.

- In the hot “fan economy”, “follow” has become the source of profit for entertainment capital. The deep binding of entertainment circle and capital market makes the liquidity of flow even more powerful. From “the film and television industry” to the capital market, the scale of profit reaping by entertainment capital has expanded exponentially.

Data Collection

「Method of getting data　Details of dataset_1」

Method of getting dataset_1

- Dataset_1 comes from a website: https://www.chinaindex.net/idol/MyIdol （中国娱乐指数）
- In order to prevent crawlers and other technologies, the website cancels the login on the web, can only query from the mobile phone.
- So we can only collect data manually.
- We have selected 145 stars in China, including actors, singers; idol, elitists; top-star、popular celebrity 、 outmoded stars… The samples are abundant and widely distributed.
- We selected 13 covariances for each stars (some may have the missing values).
- The final dataset dataset_1

Variable of dataset_1

- Name
  name of the celebrity
- Sex
  the gender of the celebrity
  “0”means female, ”1” means male
- Field
  the field of the celebrity
  “1”means actor, ”2” means singer
- Type
  the type of the celebrity
  “1”means idol, “2” means elitists
- Busi_index
  the comprehensive index of the commercial value of the celebrity
  calculated by the weight of prof_index, heat_index, endor_index, pub_index

(depend on October,2020)

Variable of dataset_1

- Prof_index

  The scores of artists in the calculation period are calculated by weighting their contribution to the box office of films, TV series ratings, variety show ratings, TV plays, variety shows, awards and history.

- Heat_index

  The heat index is weighted by the indexes such as the number of active dehydrated fans, media exposure, dehydration hot discussion, and search volume.

- Endor_index

  The score of individual word-of-mouth of artists in a certain calculation period is calculated according to the six dimensions of public praise index, marriage and love index, words and deeds index, shape index, personality index and professional skill index.

- Pub_index

  The endorsement score of artists in the calculation cycle is calculated by weighted calculation of the number of brands, brand level, endorsement method, endorsement area and endorsement effect.

(depend on October,2020)

Variable of dataset_1

- Fans
  The number of active dehydrated fans daily on the first week in November.
- Female
  Proportion of female fans daily on the first week in November.
- Red
  The number of red active dehydrated fans daily on the first week in November.
- Black
  The number of red active dehydrated fans daily on the first week in November.

(depend on the first week in November,2020)

Data Management of dataset_1

- For sex, field, type, we use the command "fre" to check missing value and frequency of discrete variables.

```
. /*Data Management*/
. fre sex

sex — sex
```

		Freq.	Percent	Valid	Cum.
Valid	0	58	40.00	40.00	40.00
	1	87	60.00	60.00	100.00
	Total	145	100.00	100.00	

```
. fre field

field — field
```

		Freq.	Percent	Valid	Cum.
Valid	1	79	54.48	54.48	54.48
	2	66	45.52	45.52	100.00
	Total	145	100.00	100.00	

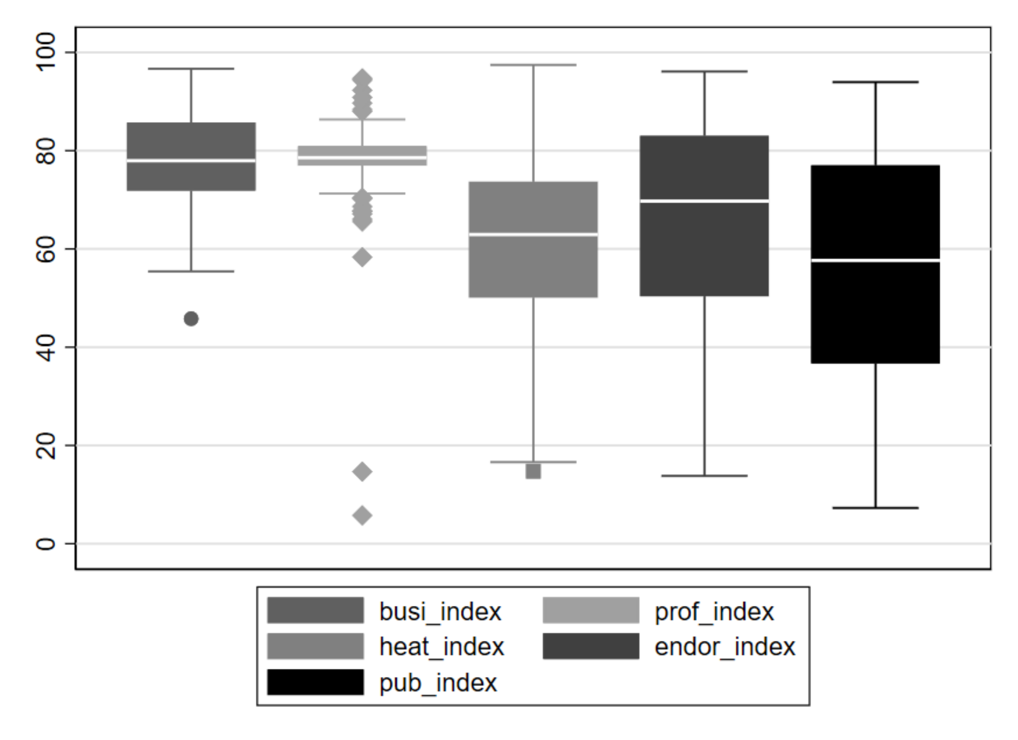
```
. fre type

type — type
```

		Freq.	Percent	Valid	Cum.
Valid	1	68	46.90	46.90	46.90
	2	77	53.10	53.10	100.00
	Total	145	100.00	100.00	

- Do not find mistakes and omissions

Data Management of dataset_1

- For busi_index, prof_index, heat_index, endor_index, pub_index, we use command "graph box" to check.



- We find that there are two extremely low point in prof_index, they are "成果" and "姚琛".But we check the data again from the website, they are truly low as we have seen.

Data Management of dataset_1

- For fans, female, red, black, we use the command "fre" to check missing value and frequency of values.

- We find some missing value by the command "list" :

```
. list name if female==.

     +--------+
     |   name |
     |--------|
 85. | 王志文 |
     +--------+
```

```
. list name if red==.

      +--------+
      |   name |
      |--------|
  54. | 马德华 |
  85. | 王志文 |
  94. | 许亚军 |
 126. |   朱丹 |
      +--------+
```

```
. list name if black==.

      +--------+
      |   name |
      |--------|
  31. |   金星 |
  54. | 马德华 |
  85. | 王志文 |
  94. | 许亚军 |
 126. |   朱丹 |
      +--------+
```

- Then we replace black=0 of "金星" who red!=0 but black is the missing value.

- The final dataset dataset_1

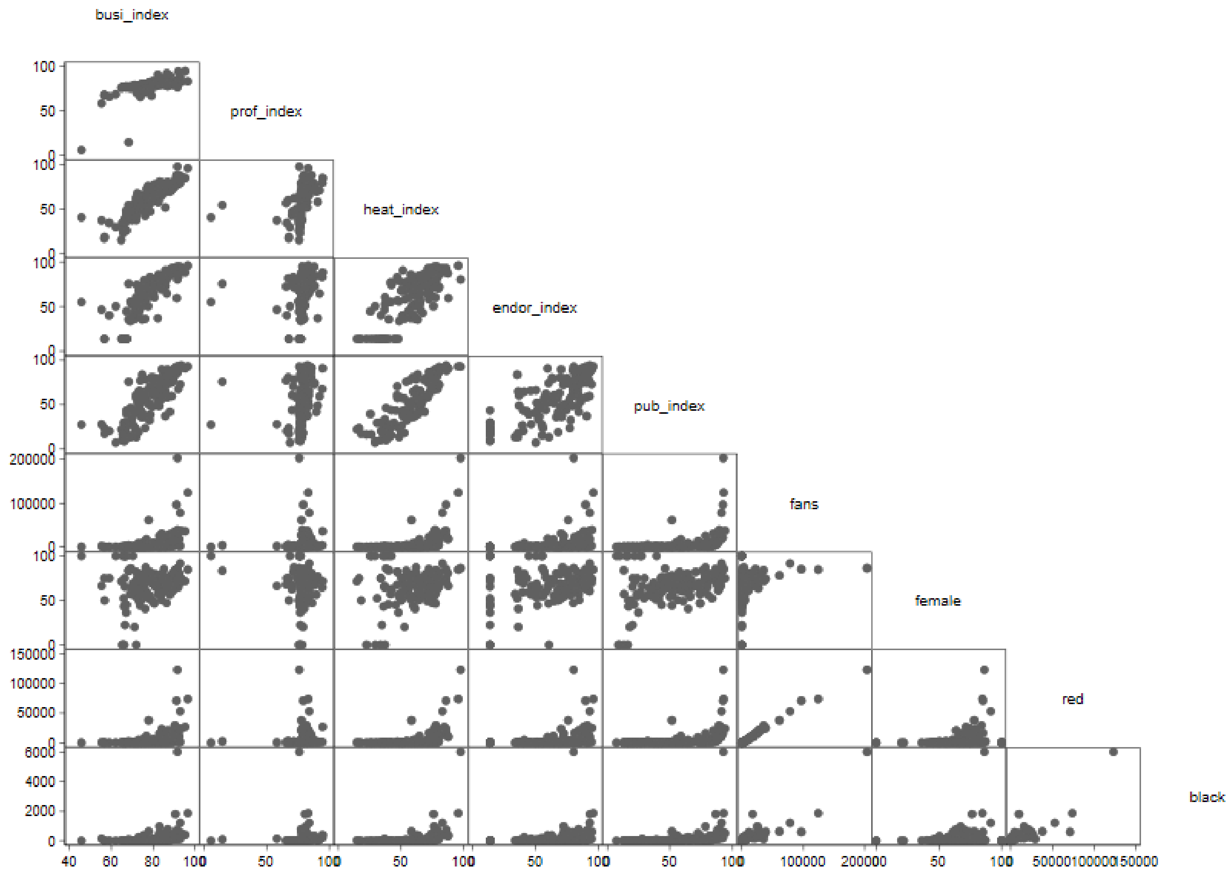
Analysis of dataset_1

「 Prof_index   The threshold of endor_index   Black fans 」

Before specific questions

Analysis of dataset_1

- We draw the scatter matrix using the command "graph matrix" , finding that fans, red and black should do log-transformation to fix the model. And female is NOT suitable to be analyzed as a single variable.

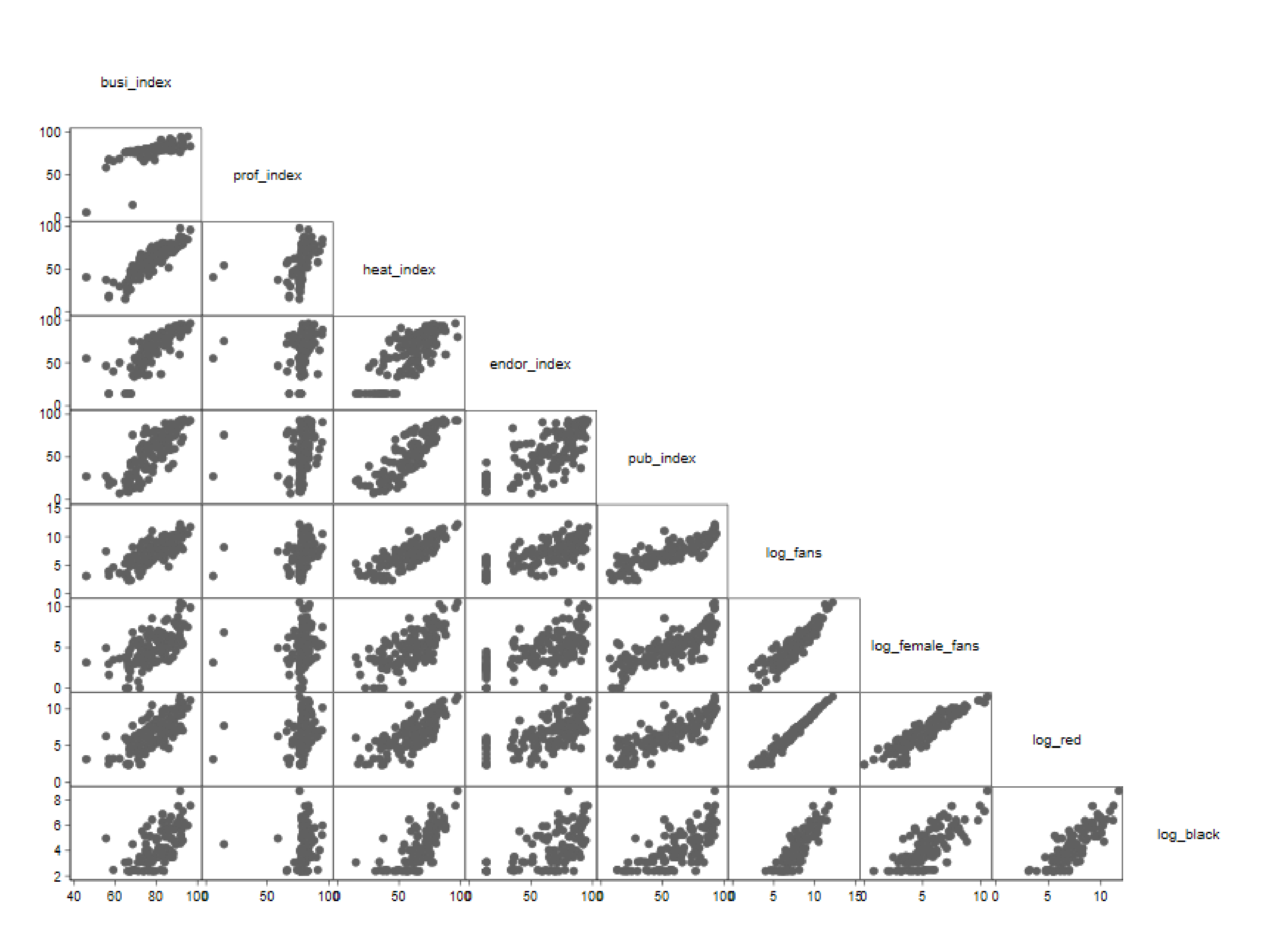


(before transformation)

Before specific questions

Analysis of dataset_1

- We draw the scatter matrix using the command "graph matrix" , finding that fans, red and black should do log-transformation to fix the model. And female is NOT suitable to be analyzed as a single variable.

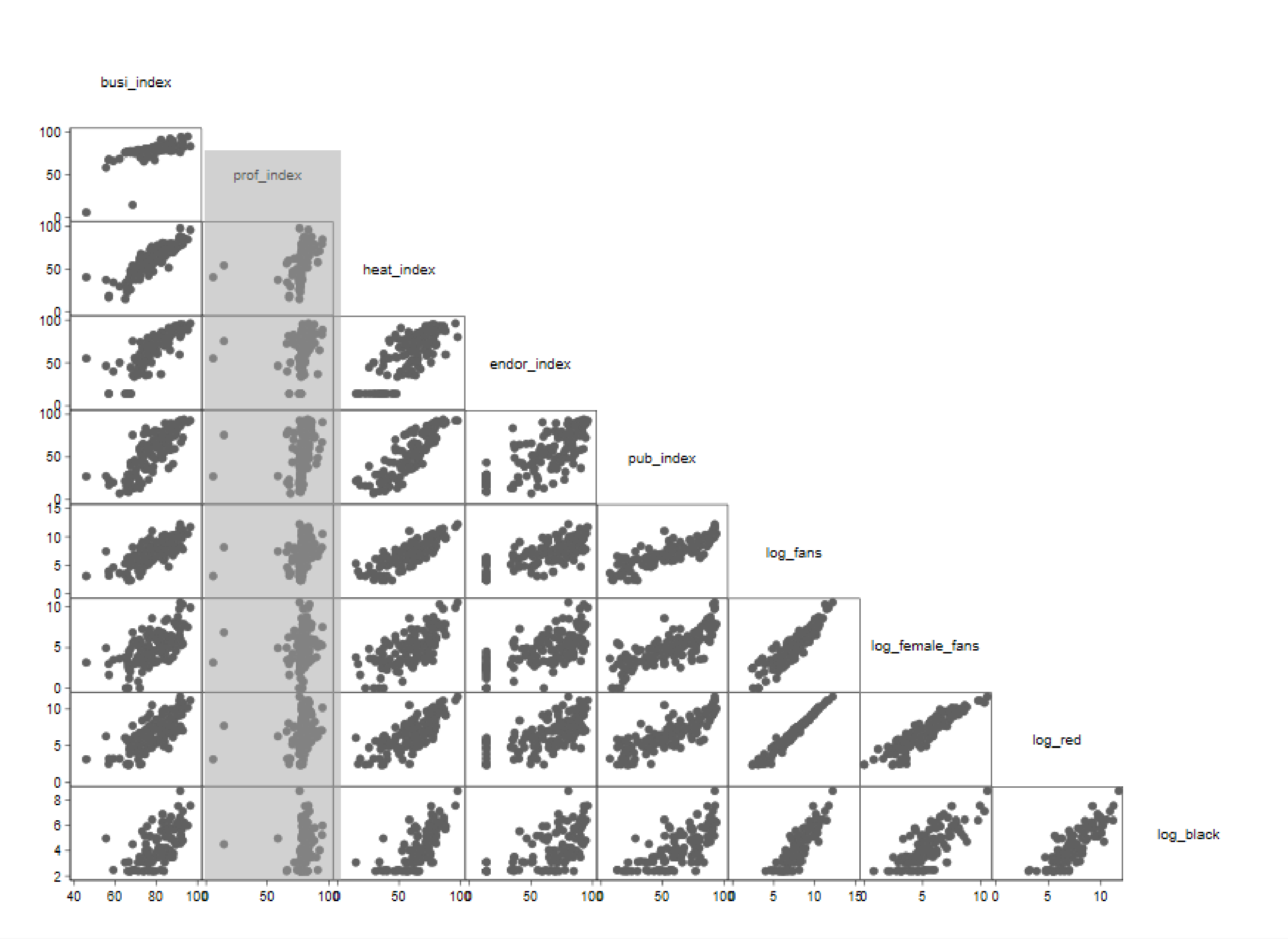


(after transformation)

Prof_index

Analysis of dataset_1

- The first interesting question we find is about the prof_index, it seems that prof_index has NOT any relationship with other variables.

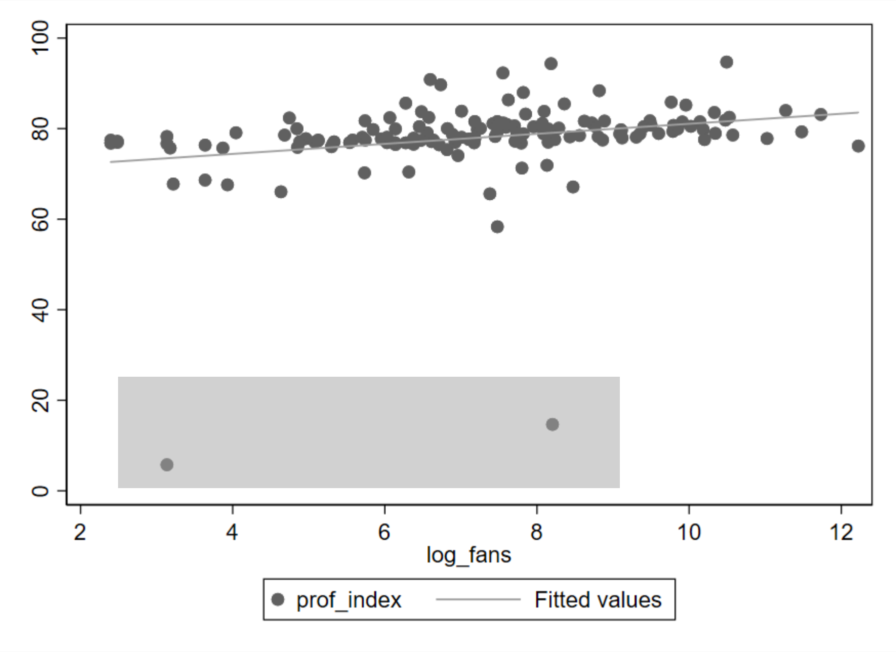


(after transformation)

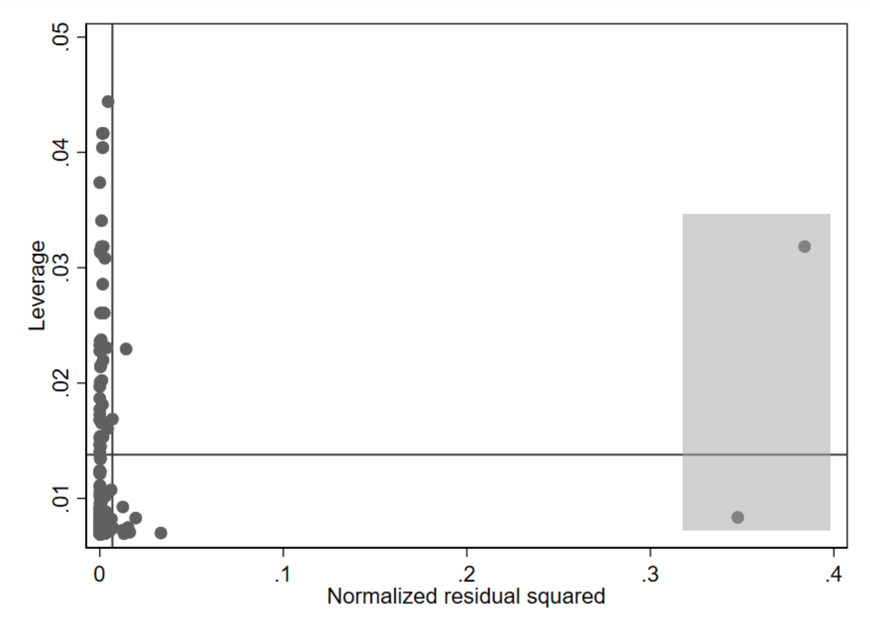
Prof_index

Analysis of dataset_1

- In the scatter graph, we find two points who may have influence for our outcome, so we test the influence using command "lvr2plot" ,and decide to delete these two points.



scatter of prof_index



lvr2plot test

Prof_index

Analysis of dataset_1

- Then we do the regression of these models between prof_index and other variables.
- We choose the examples: prof_index & log_fans   prof_index & endor_index

```
. **选择不分析姚琛与成果
. regress prof_index log_fans if name!="成果"&name!="姚琛"

      Source |       SS           df       MS      Number of obs   =       143
-------------+----------------------------------   F(1, 141)       =     18.43
       Model |  393.181425         1  393.181425   Prob > F        =    0.0000
    Residual |  3008.17024       141  21.3345407   R-squared       =    0.1156
-------------+----------------------------------   Adj R-squared   =    0.1093
       Total |  3401.35166       142  23.9531807   Root MSE        =    4.6189

------------------------------------------------------------------------------
  prof_index |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    log_fans |    .777413    .181091     4.29   0.000     .4194084    1.135418
       _cons |   73.30028   1.366699    53.63   0.000     70.59841    76.00215
------------------------------------------------------------------------------
```

prof_index & log_fans

```
. **选择不分析姚琛与成果
. regress prof_index endor_index if name!="成果"&name!="姚琛"

      Source |       SS           df       MS      Number of obs   =       143
-------------+----------------------------------   F(1, 141)       =     23.49
       Model |  485.745649         1  485.745649   Prob > F        =    0.0000
    Residual |  2915.60601       141  20.6780568   R-squared       =    0.1428
-------------+----------------------------------   Adj R-squared   =    0.1367
       Total |  3401.35166       142  23.9531807   Root MSE        =    4.5473

------------------------------------------------------------------------------
  prof_index |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 endor_index |   .0738849   .0152442     4.85   0.000      .043748    .1040217
       _cons |   74.27309    1.03301    71.90   0.000      72.2309    76.31528
------------------------------------------------------------------------------
```

prof_index & endor_index

Prof_index

- Unfortunately, we have come to the conclusion that the professionalism of a star has little to do with the number of fans, endorsements and other indicators.
- This is an decade of followings, not professionalism.

```
. **选择不分析姚琛与成果
. regress prof_index log_fans if name!="成果"&name!="姚琛"

      Source |       SS           df       MS      Number of obs   =       143
-------------+----------------------------------   F(1, 141)       =     18.43
       Model |  393.181425         1  393.181425   Prob > F        =    0.0000
    Residual |  3008.17024       141  21.3345407   R-squared       =    0.1156
-------------+----------------------------------   Adj R-squared   =    0.1093
       Total |  3401.35166       142  23.9531807   Root MSE        =    4.6189

------------------------------------------------------------------------------
  prof_index |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    log_fans |    .777413    .181091     4.29   0.000     .4194084    1.135418
       _cons |   73.30028   1.366699    53.63   0.000     70.59841    76.00215
------------------------------------------------------------------------------
```

prof_index & log_fans

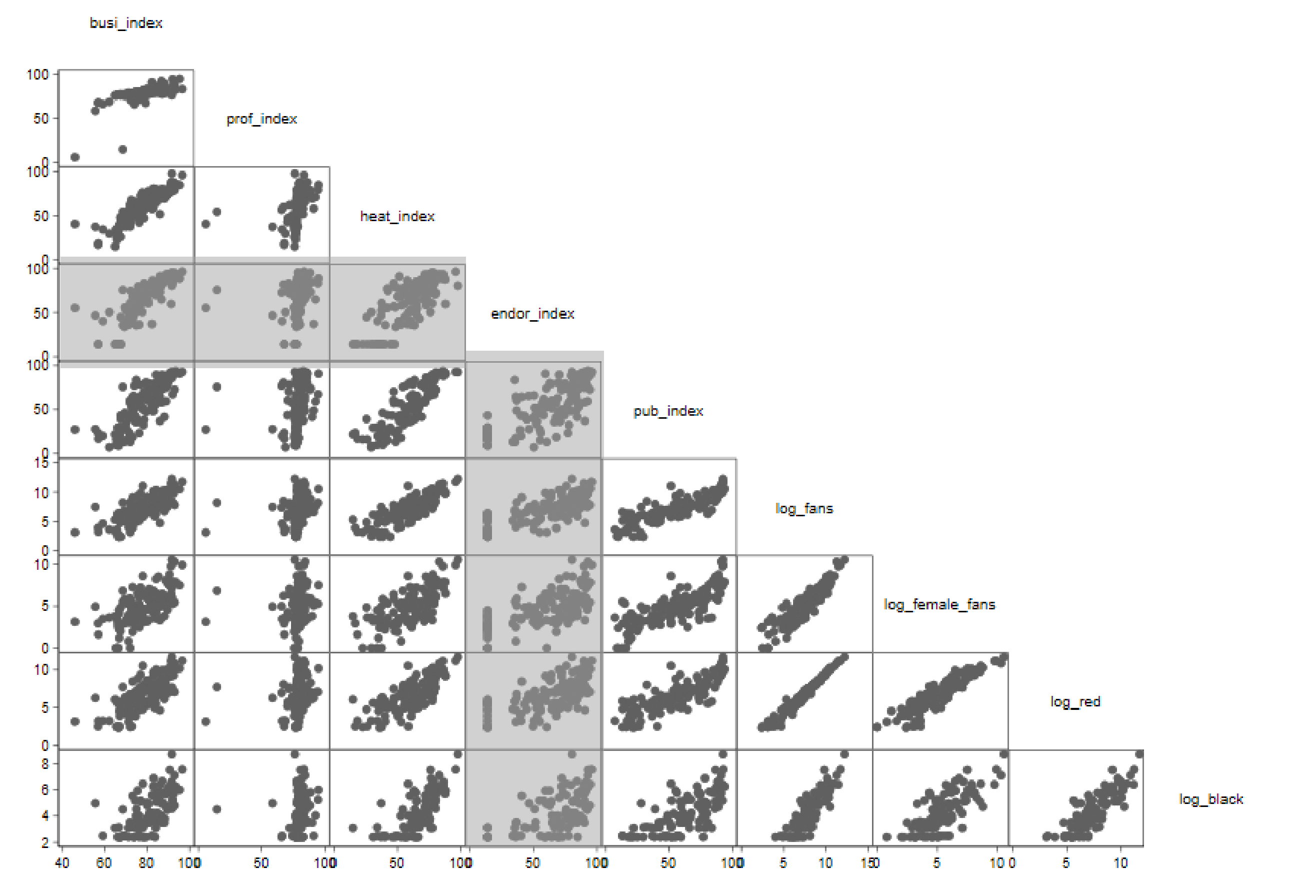
```
. **选择不分析姚琛与成果
. regress prof_index endor_index if name!="成果"&name!="姚琛"

      Source |       SS           df       MS      Number of obs   =       143
-------------+----------------------------------   F(1, 141)       =     23.49
       Model |  485.745649         1  485.745649   Prob > F        =    0.0000
    Residual |  2915.60601       141  20.6780568   R-squared       =    0.1428
-------------+----------------------------------   Adj R-squared   =    0.1367
       Total |  3401.35166       142  23.9531807   Root MSE        =    4.5473

------------------------------------------------------------------------------
  prof_index |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 endor_index |   .0738849   .0152442     4.85   0.000      .043748    .1040217
       _cons |   74.27309    1.03301    71.90   0.000      72.2309    76.31528
------------------------------------------------------------------------------
```

prof_index & endor_index

The threshold of endor_index

Analysis of dataset_1

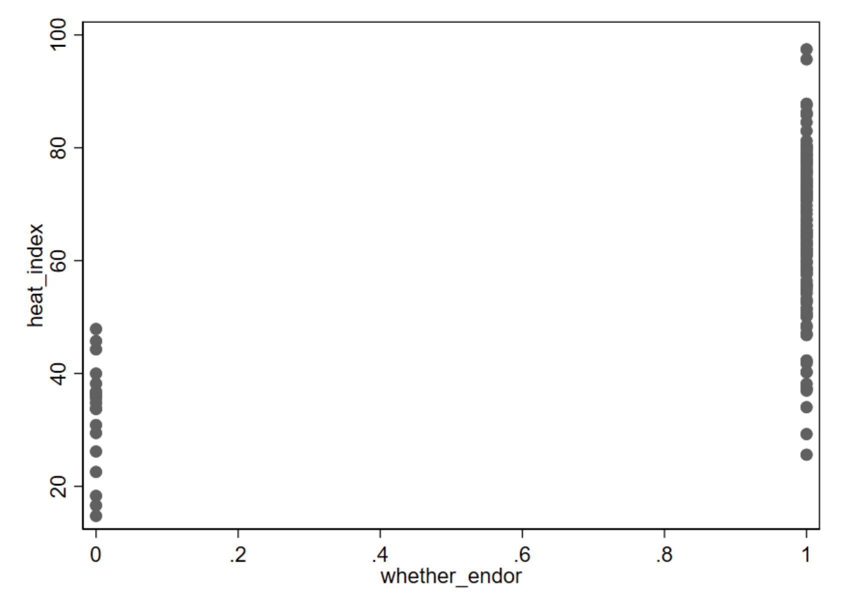
- We notice that there is a series of outlier continuous points. Through consulting, we know that this series of points represents the stars who can not receive endorsement. So, does this kind of alienation mean a kind of "threshold" that can receive endorsement?



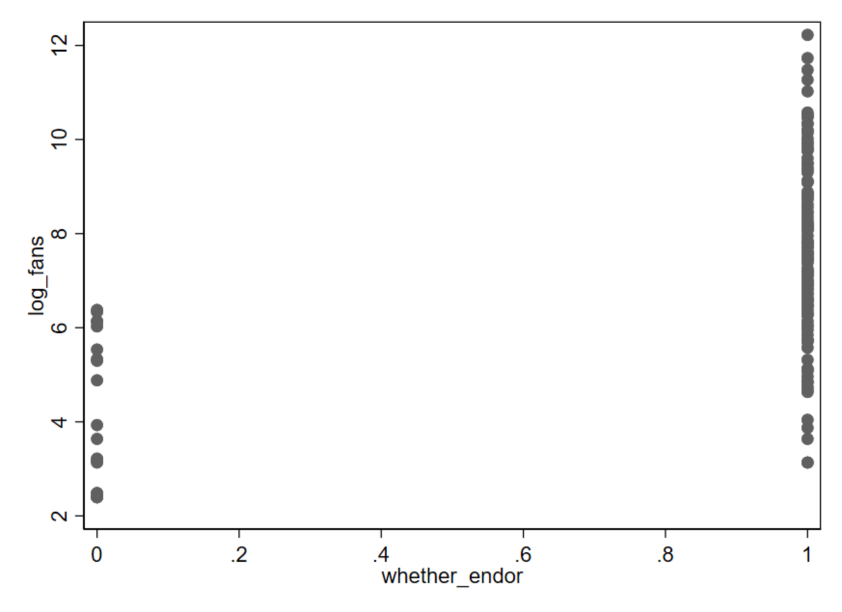
(after transformation)

The threshold of endor_index

- We discretize the endor_index, and mark the endorser index as 1 and those without endorsement as 0.
- Shown in the scatter graph, we believe that there is a clear threshold in endor_index.



heat_index



log_fans

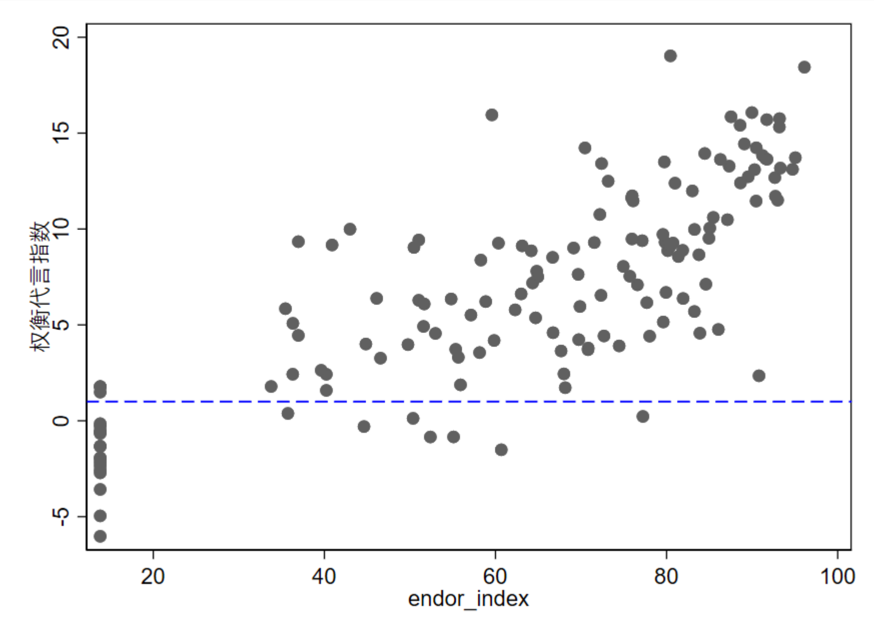
The threshold of endor_index

- We used all the covariances to do logistics regression, and through regression diagnosis, we deleted covariances with large p-value and obvious multicollinearity.
- Finally, we get the result: whether_endor = f(whether_endor, heat_index, type, log_fans)

```
Logistic regression                             Number of obs     =        145
                                                LR chi2(3)        =      77.86
                                                Prob > chi2       =     0.0000
Log likelihood = -17.382398                     Pseudo R2         =     0.6913
```

whether_endor	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
heat_index	.2332206	.0627503	3.72	0.000	.1102322	.356209
type	-3.346821	1.300203	-2.57	0.010	-5.895172	-.7984688
log_fans	.3490019	.3243503	1.08	0.282	-.286713	.9847169
_cons	-4.617545	2.316404	-1.99	0.046	-9.157612	-.0774767

regression result



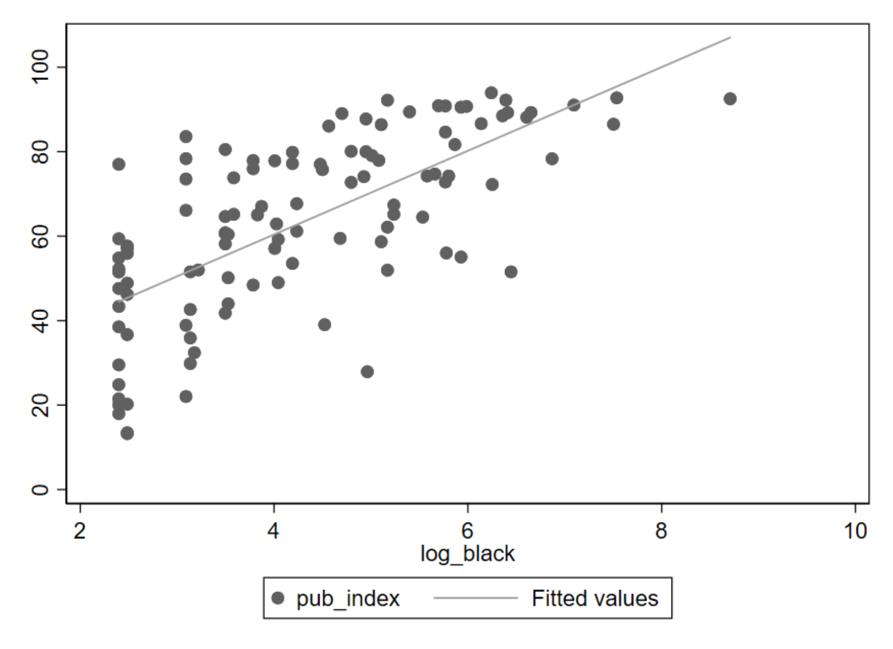
predictor

Black fans

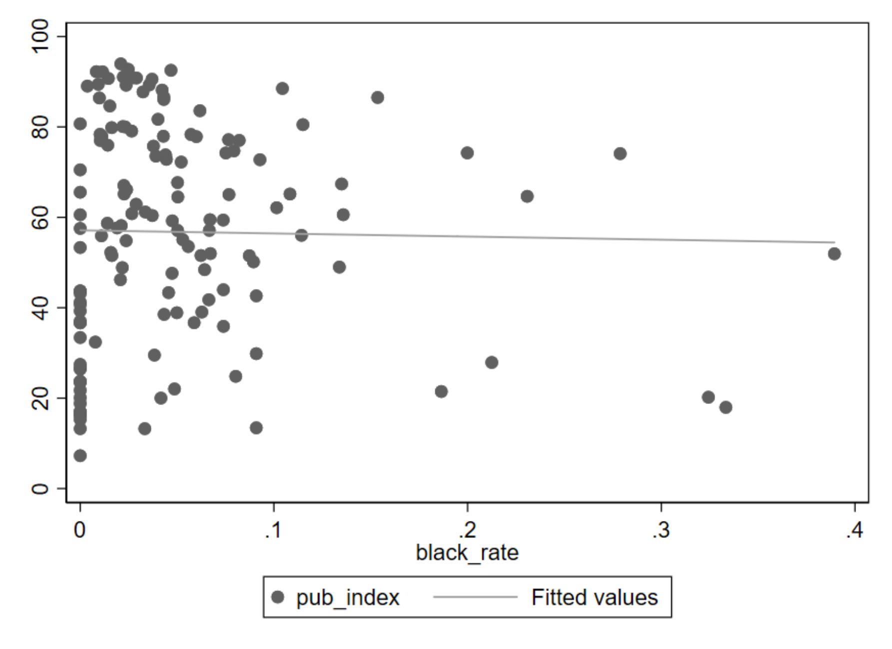
Analysis of dataset_1

- We first care about the the pub_index & black.

- We believe that the black represents the folk's evaluation of stars' word-of-mouth, while the pub_index represents the industry's evaluation of whether the stars' word-of-mouth is good or not.



Black & pub_index

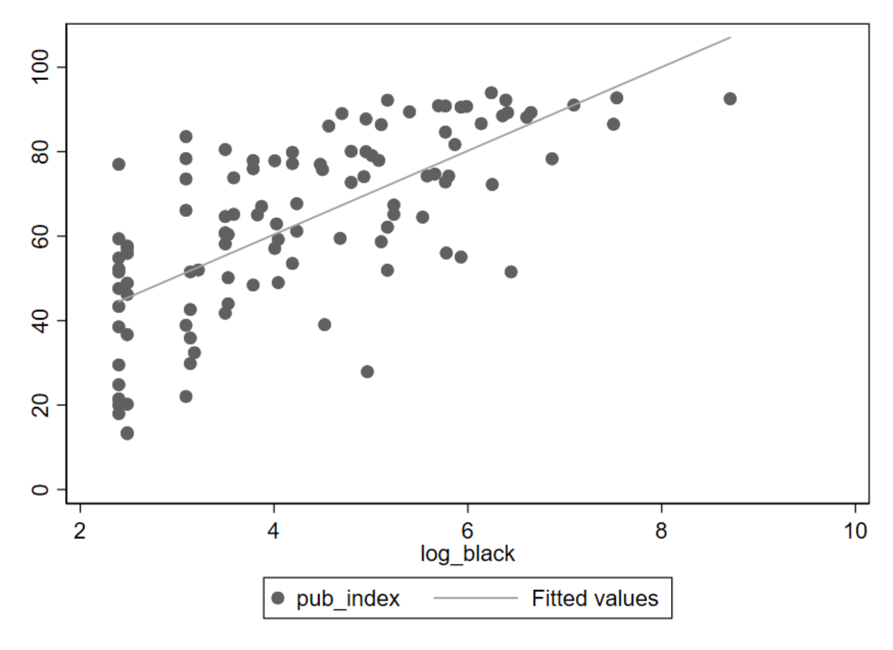


Black rate & pub_index

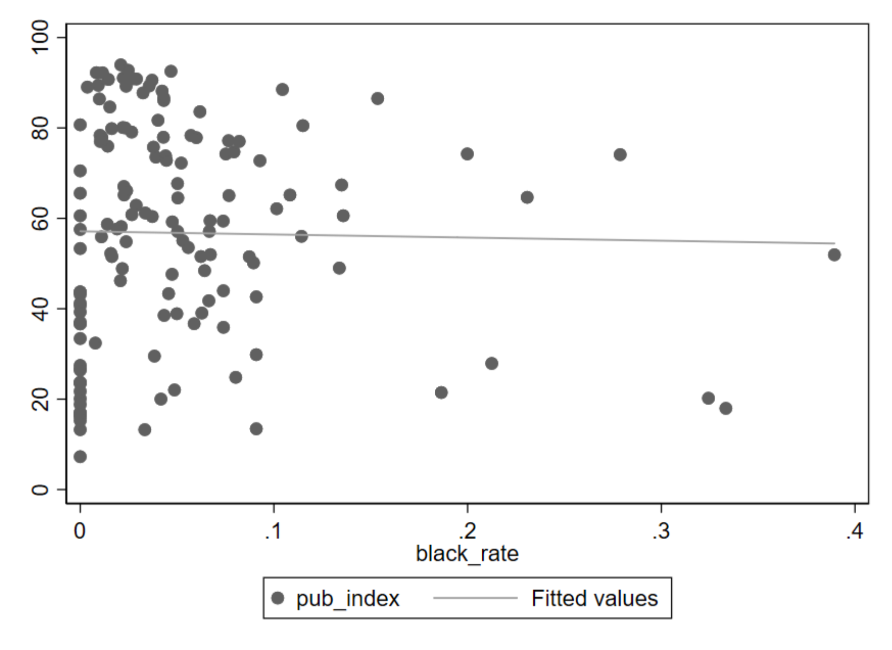
Black fans

Analysis of dataset_1

- there is no relationship between the industry's judgment criteria (pub_index) and the black/red ratio.
- .Even black has become a measure of the popularity of stars



Black & pub_index



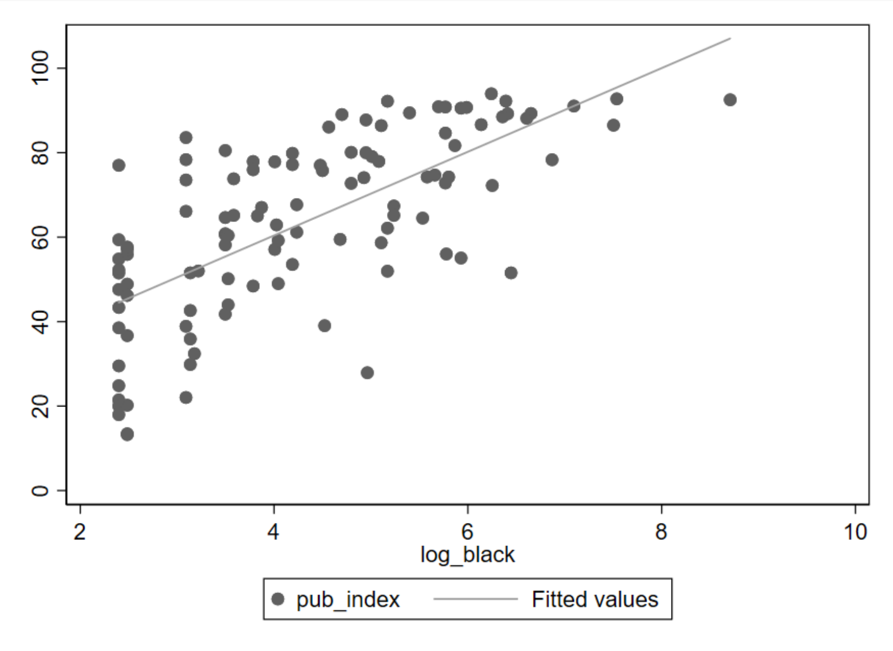
Black rate & pub_index

Black fans

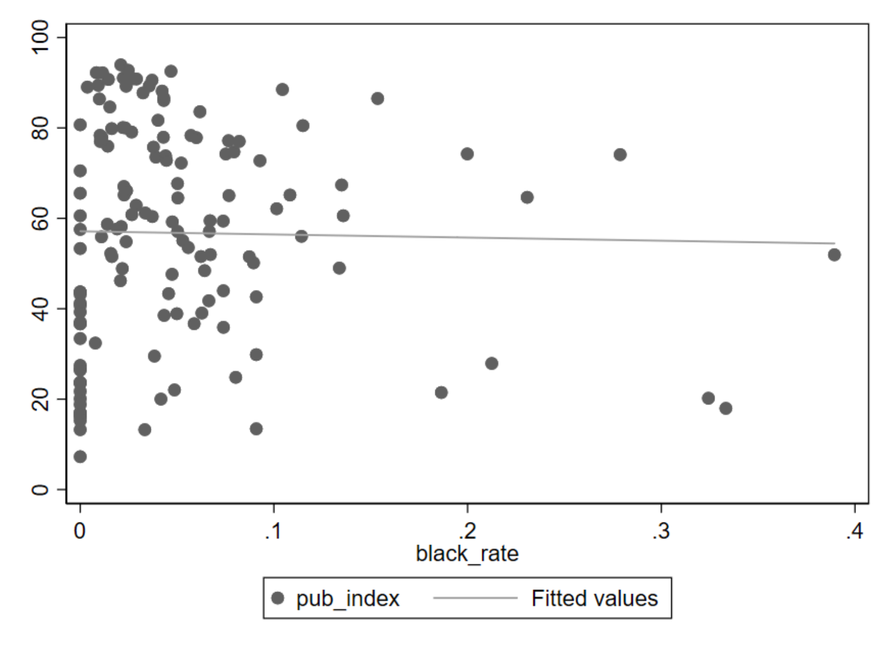
Analysis of dataset_1

- So in the entertainment industry, "a bad reputation is better than no reputation" is really a truth.

- But it's NOT means that the black has not influence for business-value of stars. For example, "肖战".



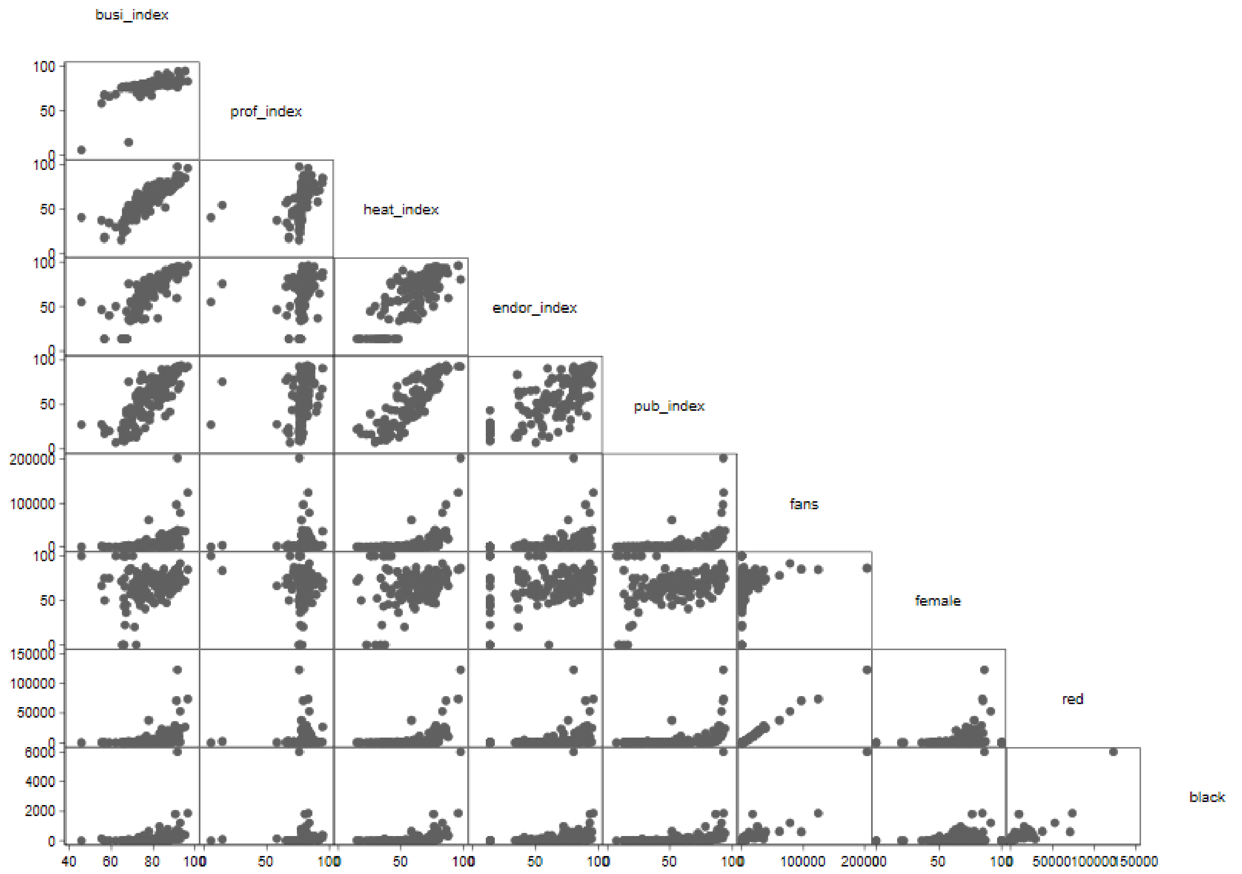
Black & pub_index



Black rate & pub_index

Analysis of dataset_1

- Expect the question we analysis right now, there are also many questions we can find. If you are interested in it, We welcome further research from you.



(before transformation)

Data Collection

「 Method of getting data　Details of dataset_2 」

Method of getting dataset_2

The original list collected from web weibo :

We have to collect data manually for subsequent crawler.
the original list: id_dataset.xlsx including information as follows:

name	user_id	container_id	gender	follwers_count
韩东	7317173686	1076037317173686	F	565709

URL format:
https://m.weibo.cn/api/container/getIndex?type=uid&value=2710029517&containerid=1076031263498570

To improve the efficiency, we use a loop in the code.
Each artist's weibo information is independently save in a .csv file.
Finally we collect 50 csv files and name them after artists's name.

```
import requests
import json
import csv
from w3lib.html import remove_tags
import time
```

Variable of dataset_2

Original Dataset Format:

原文	转发数	评论数	点赞数	发文时间
往后余生......	100万+	343911	609323	2019/5/23

In fact this data is a time series data, so we have to set a base date to change the format of time variable.
What's more, the encoding scheme of our CSV is GB2312, which is not available for Stata to recognise, therefore we have to change the encoding scheme to UTF-8 with the help of Notepad ++. This transform need to be done when the whole dataset need no more change, otherwise the encoding format will return to GB2312.
To find more useful information, we manually classify the type of the posts into 5 types, and this variable is named after 'ptype'.
One more you need to be careful:
Some stars set up their weibo "visible for the last half a year", for example "肖战", "林心如", we fail to make indepth analysis about them.

原文	转发数	评论数	点赞数	发文时间	reference	date	ptype

Data Management of dataset_2

原文	转发数	评论数	点赞数	发文时间	reference	date	ptype

- 原文
  contents of their posts
- 转发数
  number of reposts, number greater than 1e6 are treated as "100万+"
- 评论数
  number of comments, number greater than 1e6 are treated as "100万+"
- 点赞数
  number of likes do not have missing value problem
- 发文时间、reference、date
  date created to make time-series analysis
- ptype
  the comprehensive index of the commercial value of the celebrity
  calculated by the weight of prof_index, heat_index, endor_index, pub_index
  ps: since some artists posted too much blogs, we only classify the posts starting from 2019/1/1 to the collecting time(2020/11/29).

```
import delimit "王力宏.csv",encoding(utf-8) clear
```

- **example of the 5 types:**(use 刘昊然's weibo content)

contents	ptype
分享图片	1
#唐人街探案3# “名侦探·秦风”终于上线啦 今晚有什么离奇的案子在等着我们唐探家族呢？20:20#快乐大本营#唐人街探案秦风就位！	2
#国家公祭日#勿忘历史 守护和平	3
探索#TOMFORD惹火派对#满溢钻光的梦幻空间，多重色泽，演绎我的百变风格，开启#天猫小黑盒#感受@TOMFORDBEAUTY 致奢银熠系列新品的闪耀。	4
思诚哥新作 明年暑假见！ #外太空的莫扎特官宣#	5

在我的队友的建议下，将微博内容分成了四类：1代表分享生活，2代表发行作品，3代表时事相关（比如转发一些正能量微博、微博自动发送生日快乐等），4代表广告赞助，5圈内联动（商业互吹）

Verification of dataset_2

- codebook ptype

```
ptype                                                      (unlabeled)

                  type:  numeric (byte)

                 range:  [1,5]                     units:  1
         unique values:  5                     missing .:  1,831/1,962

            tabulation:  Freq.  Value
                            43  1
                            61  2
                            15  3
                             3  4
                             9  5
                         1,831  .
```

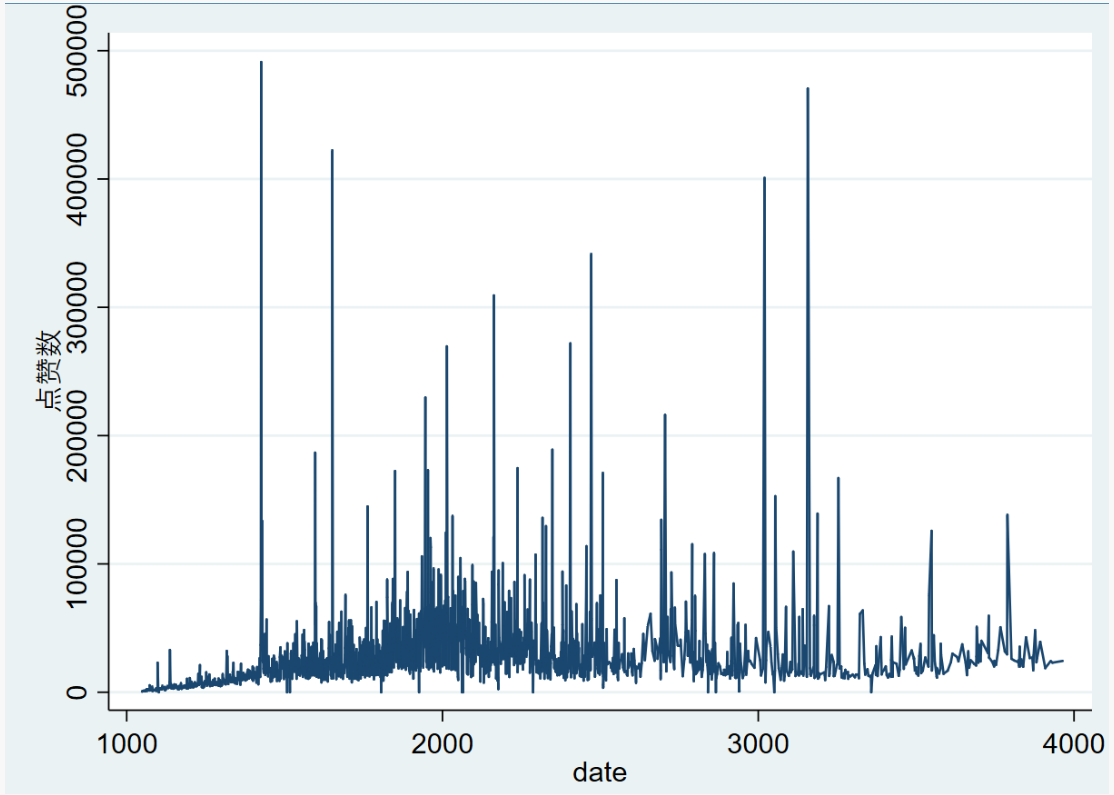
tabulation helps us to see the different types of blog frequency, as well as the proportion of career and advertising in the artist's career.

Verification of dataset_2

- FOR artists who belong to “idol”, this process may be very useful
- *转换一下数据格式为numeric
- destring(评论数),ignore("100万+") gen(comments)
- //replace comments = 1000000 if 评论数=="100万+"

- destring(转发数),ignore("100万+") gen(reposts)
- //replace reposts = 1000000 if 转发数=="100万+"
- twoway line 点赞数 date

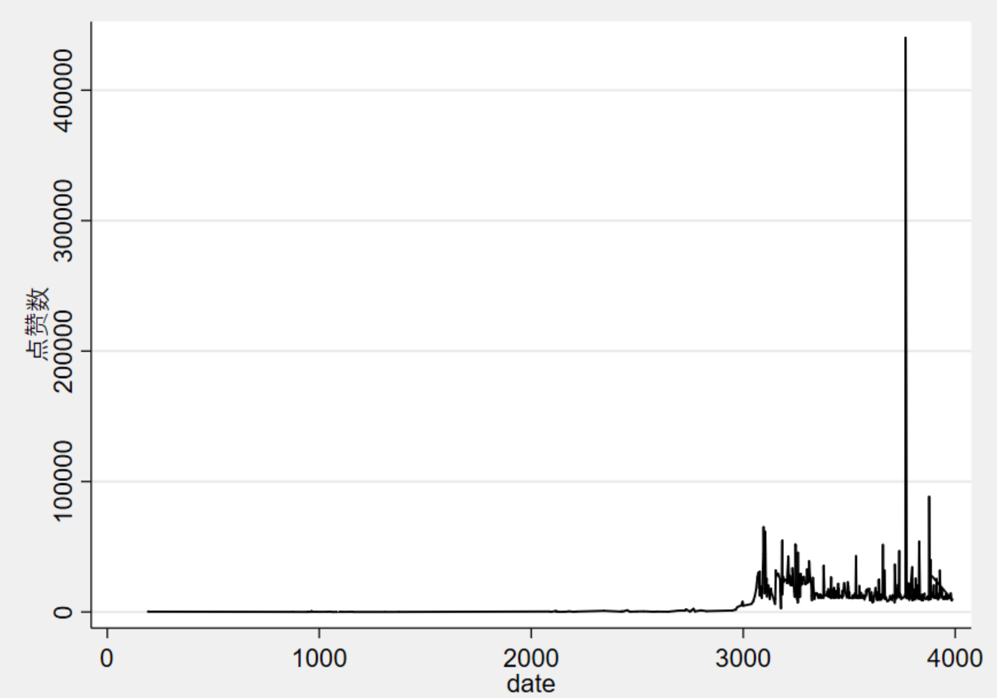
Result visualization



Verification of dataset_2

- sometimes artists have one post whose # of likes is far more than other posts, like the situation below, we tabulate this post and observe, then drop it.
- e.g. data of 王菊

61522	1	0.21	99.38
65133	1	0.21	99.59
88664	1	0.21	99.79
440258	1	0.21	100.00
Total	483	100.00	



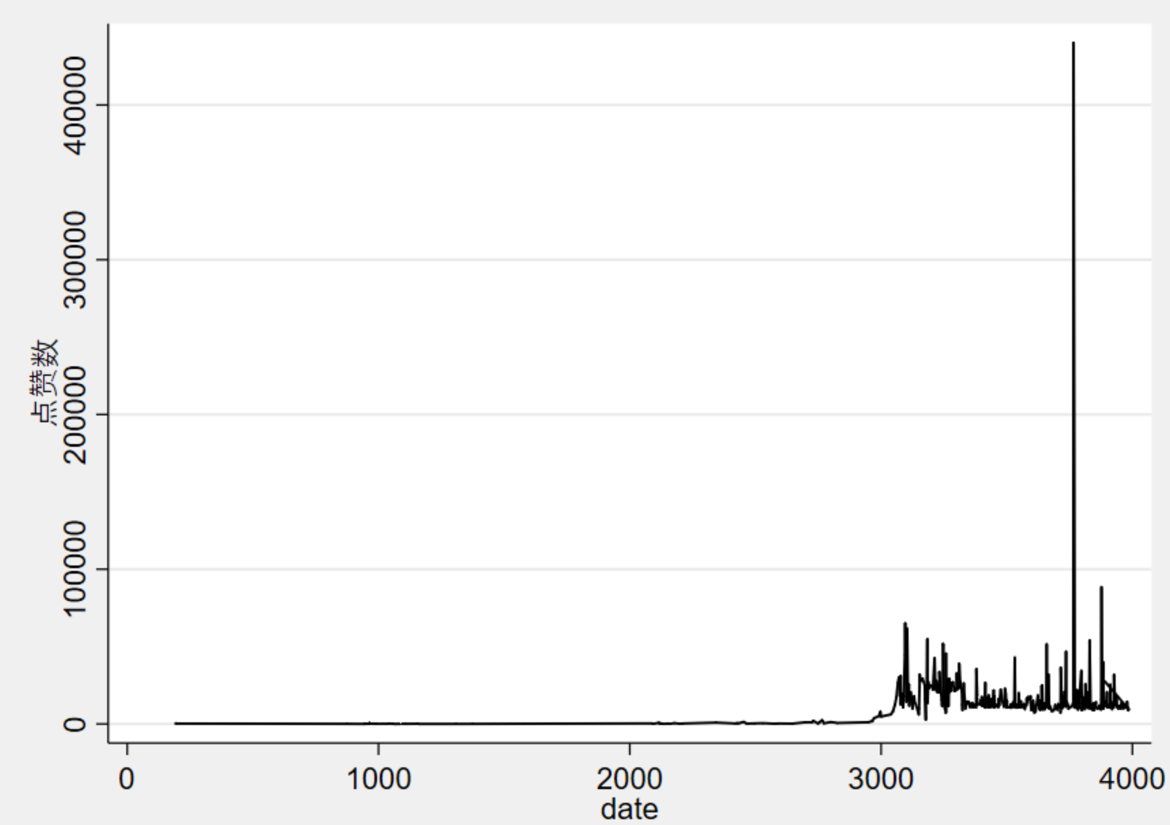

```
. list if 点赞数==440258
```

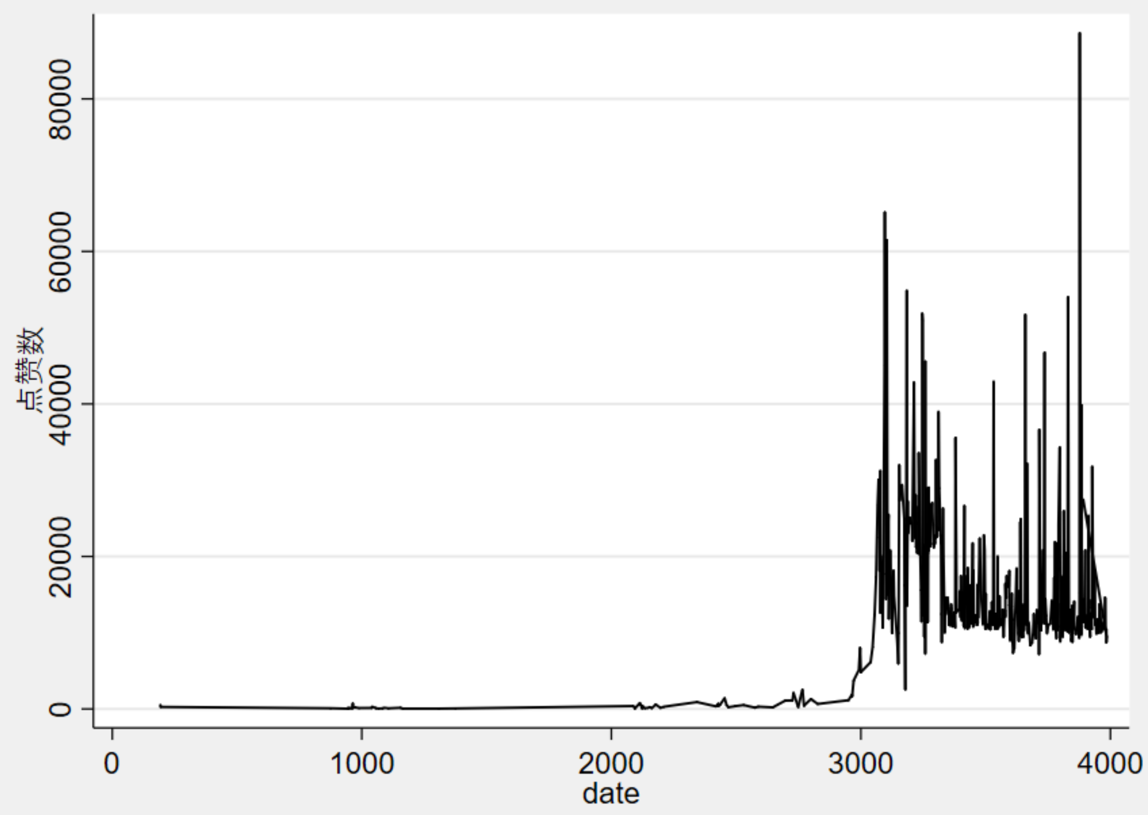
	原文	转发数	评论数	点赞数	发文时间	refere~e	date
102.	鲍毓明案现在什么进展？还有后续吗？看最新的媒体报道停在了上周...	79639	9202	440258	2020/4/24	2010/1/1	3766

ptype
3

Verification of dataset_2

- before and after we drop the outlier, We can see the trend more clearly






```
drop if 点赞数==440258
```

Verification of dataset_2

- How to obtain data when we want to analysise correlation between ptype and “likes”

```
*为了直观区分不同种类微博的点赞量，将其数据分割成五个子集
gen tp1 = 点赞数 if ptype==1
gen tp2 = 点赞数 if ptype==2
gen tp3 = 点赞数 if ptype==3
gen tp4 = 点赞数 if ptype==4
gen tp5 = 点赞数 if ptype==5
summarize tp1 tp2 tp3 tp4 tp5
```

```
*标签手动标记至2019/1/1，即2019全年至2020/11/29日所有微博，其余数据舍去
sort 点赞数,stable
drop if(ptype==.)
summarize 点赞数
*分析点赞数最多和最少的105条微博（试一下）
tabulate ptype in -10/L
tabulate ptype in 1/10
```

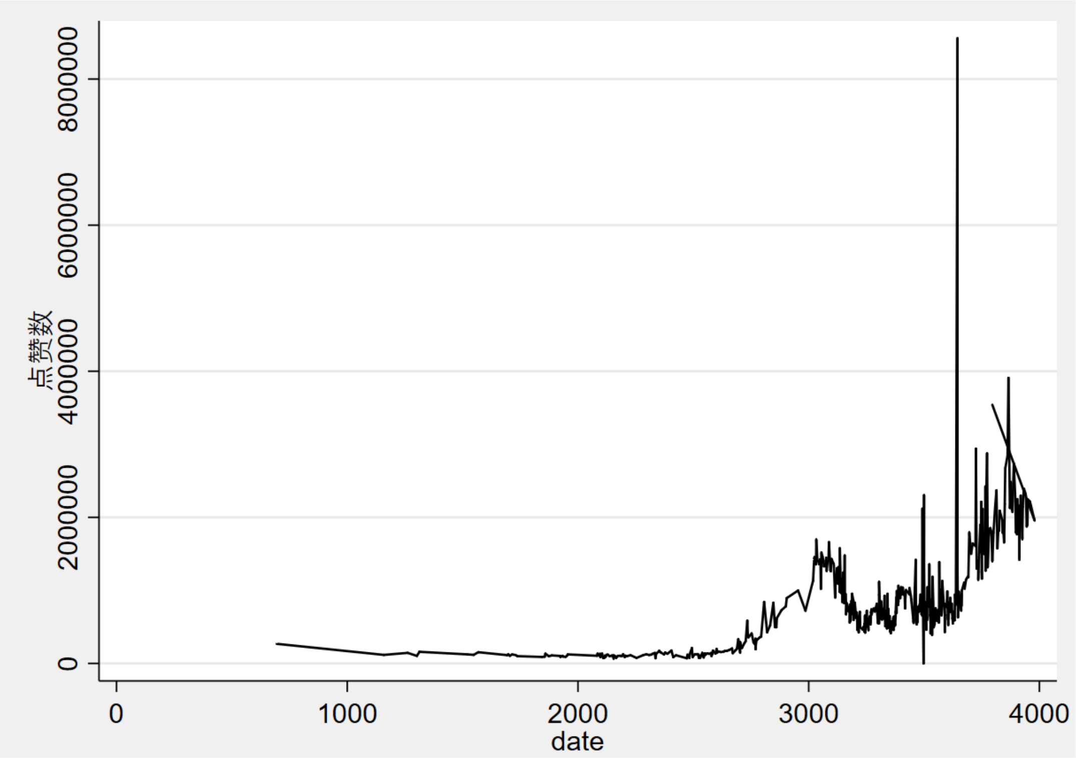
Analysis of dataset_2

「 」

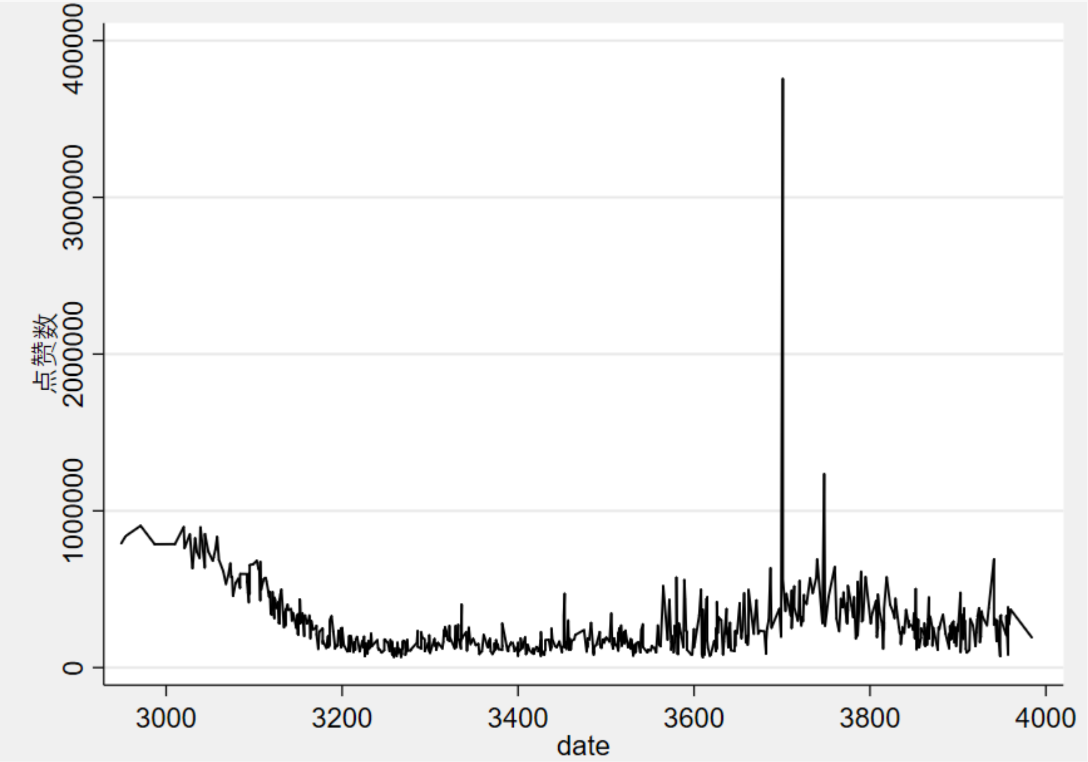
Other specific questions

Analysis of dataset_2

- From the twoway line of artists' like-date data, we can see a general change in sentiment.



time series plot of 蔡徐坤's likes



time series plot of 黄明昊's likes

Other specific questions

- Fans' preference of contents posted by artists e.g. 王菊

. summarize 点赞数

Variable	Obs	Mean	Std. Dev.	Min	Max
点赞数	305	14705.52	7938.31	7188	88664

Variable	Obs	Mean	Std. Dev.	Min	Max
tp1	120	13836.15	5662.36	8003	46727
tp2	74	15536.89	10351.19	8802	88664
tp3	45	13627.38	8359.603	7188	54019
tp4	44	16798.48	8363.416	8327	51731
tp5	22	14670.55	7072.631	7362	39843

Female idol: daily share post make up the most contents of their posts, therefore most(10/120) and least(12/120) likes. Fans are interested in her commercial posts(6/44) and her works(10/74).

. tabulate ptype in -30/L

ptype	Freq.	Percent	Cum.
1	10	33.33	33.33
2	10	33.33	66.67
3	3	10.00	76.67
4	6	20.00	96.67
5	1	3.33	100.00
Total	30	100.00	

. tabulate ptype in 1/30

ptype	Freq.	Percent	Cum.
1	12	40.00	40.00
2	5	16.67	56.67
3	7	23.33	80.00
4	4	13.33	93.33
5	2	6.67	100.00
Total	30	100.00	

- Fans' preference of contents posted by artists e.g. 刘昊然

```
. summarize 点赞数
```

Variable	Obs	Mean	Std. Dev.	Min	Max
点赞数	211	115829.9	110310.5	5458	696449

Variable	Obs	Mean	Std. Dev.	Min	Max
tp1	27	252088.6	157845.6	26557	584551
tp2	59	82868.24	59384.24	21766	310963
tp3	26	116709.7	153844.8	24420	696449
tp4	73	101194.2	50076	5458	239591
tp5	26	89340.81	115149.4	29466	617882

```
. tabulate ptype in -20/L
```

ptype	Freq.	Percent	Cum.
1	12	60.00	60.00
2	2	10.00	70.00
3	3	15.00	85.00
4	2	10.00	95.00
5	1	5.00	100.00
Total	20	100.00	

```
. tabulate ptype in 1/20
```

ptype	Freq.	Percent	Cum.
1	1	5.00	5.00
2	11	55.00	60.00
3	3	15.00	75.00
4	4	20.00	95.00
5	1	5.00	100.00
Total	20	100.00	

Actor: Fans are more interested in his daily share posts(12/27) while his propoganda for his works receive fewer likes(11/59).

Other specific questions

Analysis of dataset_2

- Fans' preference of contents posted by artists e.g. 范丞丞

```
. summarize 点赞数
```

Variable	Obs	Mean	Std. Dev.	Min	Max
点赞数	421	218444.3	176119.5	51295	2728952

```
. summarize 点赞数
```

Variable	Obs	Mean	Std. Dev.	Min	Max
点赞数	420	212466.9	126547.1	51295	1328127

```
. summarize tp1 tp2 tp3 tp4 tp5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
tp1	105	274404.9	110936	116459	791864
tp2	140	191457.2	113540.7	83226	957361
tp3	63	161825.2	126232.5	73211	826101
tp4	73	220407.2	85461.54	99415	704006
tp5	38	189788.2	199048.3	51295	1328127

```
. tabulate ptype in -20/L
```

ptype	Freq.	Percent	Cum.
1	10	50.00	50.00
2	5	25.00	75.00
3	2	10.00	85.00
4	2	10.00	95.00
5	1	5.00	100.00
Total	20	100.00	

```
. tabulate ptype in 1/20
```

ptype	Freq.	Percent	Cum.
2	5	25.00	25.00
3	11	55.00	80.00
5	4	20.00	100.00
Total	20	100.00	

Male idol: Fans are more interested in his daily share posts(10/105) while his propoganda for his works receive fewer likes(11/63).

Other specific questions

- Fans' preference of contents posted by artists e.g. 周冬雨

. summarize 点赞数

Variable	Obs	Mean	Std. Dev.	Min	Max
点赞数	339	48779.66	56698.61	6552	784277

. summarize tp1 tp2 tp3 tp4 tp5

Variable	Obs	Mean	Std. Dev.	Min	Max
tp1	84	79817.67	96938.81	14431	784277
tp2	74	40803.82	34799.67	6552	255869
tp3	48	32331.65	24975.35	12186	135214
tp4	101	39298.86	20369	12136	145744
tp5	32	40344.81	34750.82	11027	198769

. tabulate ptype in -30/L

ptype	Freq.	Percent	Cum.
1	22	73.33	73.33
2	3	10.00	83.33
3	2	6.67	90.00
4	2	6.67	96.67
5	1	3.33	100.00
Total	30	100.00	

. tabulate ptype in 1/30

ptype	Freq.	Percent	Cum.
1	3	10.00	10.00
2	8	26.67	36.67
3	9	30.00	66.67
4	5	16.67	83.33
5	5	16.67	100.00
Total	30	100.00	

Actress: Fans are more interested in his daily share posts(22/84) while not so concerned about her comments on current events(9/48).

118.

原文	转发数	reposts	评论数	点赞数
拔完智齿五天后脸依然肿成这样↓ 有人遇到过这种情况吗这样出现还有人认识吗？	27470	2747	70488	784277

发文时间	date	ptype
2020/1/5	3656	1

Other specific questions

Analysis of dataset_2

- Fans' preference of contents posted by artists e.g. 范丞丞 and 蔡徐坤

	原文	转发数
420.	今天是选c位出道，都去投@杨迪 ！！！不然三天见不到我！！！#青春环游记# #青春环游记2# 范丞丞Adam0616的微博视频	577208

reposts	评论数	comments	点赞数	发文时间	date	ptype	tp1	tp2	tp3	tp4	tp5
57728	127454	27454	1328127	2020/7/17	3850	5	.	.	.	.	1328127

1236294	1	0.15	99.85
3755868	1	0.15	100.00
Total	689	100.00	

```
. list if 点赞数==3755868
```

	原文	转发数	reposts	评论数	comments	点赞数	发文时间	date
182.	今天是我的生日，来祝福我吧！	775078	77578	376619	37669	3755868	2020/2/19	3701

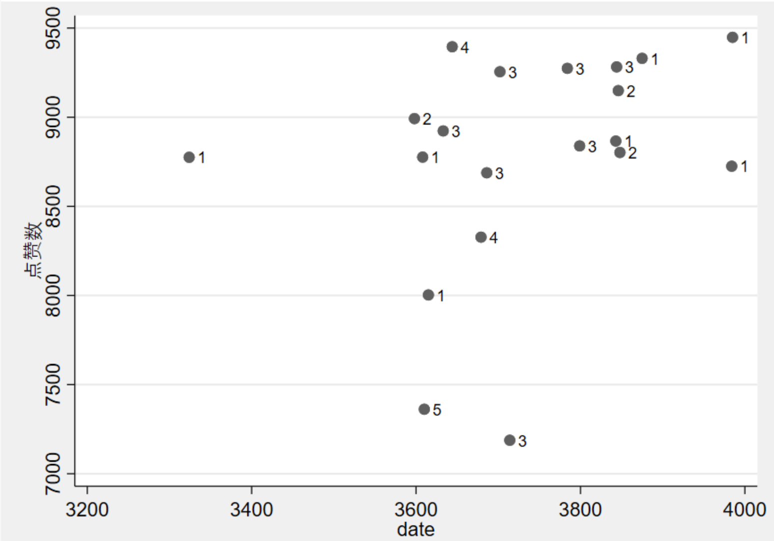
```
. list if 点赞数==8556942
```

	原文	转发数	reposts	评论数	comments	点赞数	发文时间	refernce	date
93.	我回来了 请多关照!	100万+	.	100万+	.	8556942	2019/12/25	2010/1/1	3645

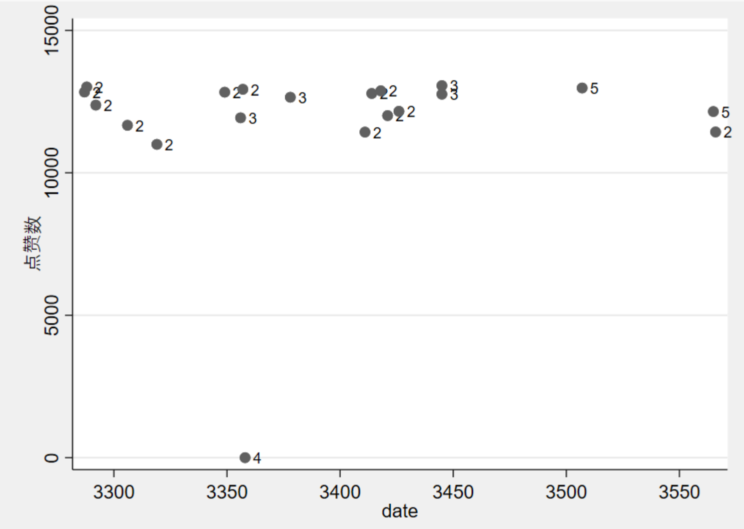
Other specific questions

Analysis of dataset_2

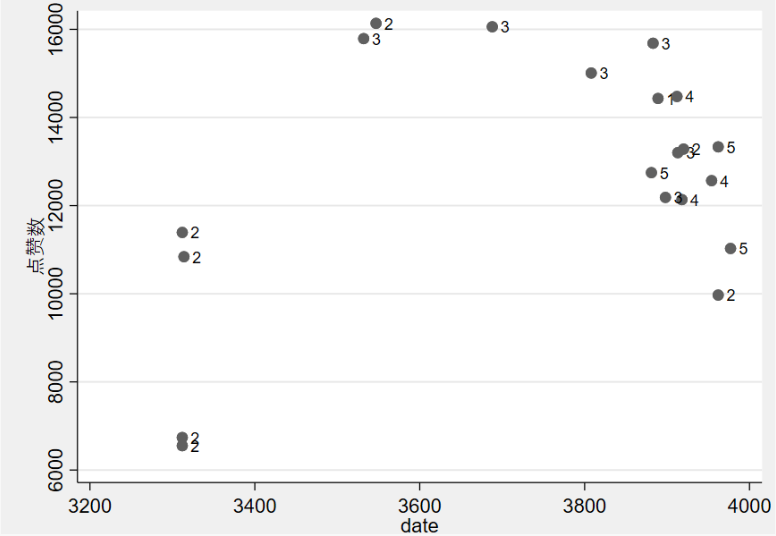
```
twoway scatter 点赞数 date in 1/20
```



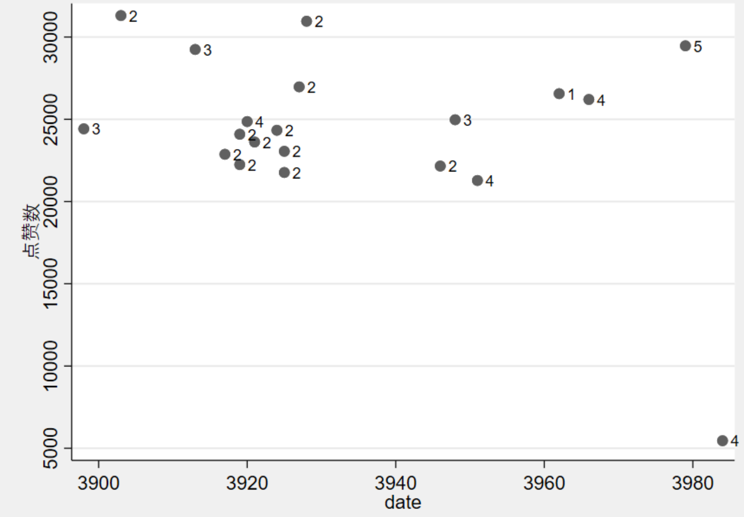
Wang Ju



Wang Lihong



Zhou Dongyu



Liu Haoran

Conclusion

「 what we find from the project 」

Main conclusion in dataset_1

- Daily active fans be treated as log-transformation.
- The professionalism of a star has little to do with the number of fans, endorsements and other indicators.
- There is a threshold of endor_index for stars. whether_endor = f(whether_endor, heat_index, type, log_fans)
- There is no relationship between the industry's judgment criteria (pub_index) and the black/red ratio. Even black has become a measure of the popularity of stars.

Main conclusion in dataset_2

Other interesting result

- There has a weak relationship between female-ratio and busi_index. The higher busi_index is, the higher female-ratio is.
- For the personal data of Xiaozhan, we find that the black increase his heat_index, but decrease his busi_index finally.
- Breaking news takes a short but huge following in 1-2 days, significant improvement in 3-5 days. But the influence is hardly more than 1 week. e.g 罗志祥、易烊千玺、丁真

Other interesting result

- **Idol's Rubbing Heat Phenomenon**： WangJu commented on Bao yu Ming's events and became her thumb up's most popular weibo post, and the data was much higher than her other micro blog data.
- **Analysise Fan's Preference to Set Up A Character**： Fans' preferences can be seen according to fans' interactive weibo data on artists, so as to decide whether to conduct interactive marketing. e.g. 博君一肖 after the broadcast of 陈情令
- **Different Types of Artists Have Quite Different Fan Orientation**:actor,actress, male idol,female idol. e.g. 刘昊然、、范丞丞、王菊
- **Sometimes data goes wrong because of fans' behaviors**: comments, reposts and others may lead you to a wrong conclusion. e.g. 周冬雨's lowest likes posts distribution
- **About dataset annotation**： classification about each post really influence our judgement

Something about our subject:

- The data obtained by using the crawler has a very strong timeliness，therefore replication may be hard if you want to use crawler and create a new dataset.
- There are still some bugs in our crawlers, 3 of the 50 artists' dataset went wrong with incorrect data or missing values.
- Deviating from the direction: due to the limited ability of us, we fail to collect more valid information, e.g. the detail of the comments and proportion of the fans of one specific artist or a specific post. If you are interested in this subject too, you may browse below links in Zhihu and find how other researchers obtain and analysize data.

- https://zhuanlan.zhihu.com/p/150514700 当数据爬虫遇上娱乐圈：用微博大数据带你看《乘风破浪的姐姐》
- https://zhuanlan.zhihu.com/p/265555609 亲眼见证明星微博“发大水”：用R爬虫记录艺人微博数据注水全过程
- https://zhuanlan.zhihu.com/p/37914996 为了知道胡歌粉丝的男女比率，爬了三百万微博数据

- I am a curiosity-driven person. These interesting data in life make me interested to deepen the exploration of these things, and the study of large amounts of data in life can also enable us to gain a lot of new discoveries.Maybe that's the beauty of statistics

Reference

「 website　code 」

THANK YOU FOR WATCHING

Reporter: **Time: December 13th,2020**

Team Members

DongXingchen ZhengYihang