一．结论

1.1 可能存在 Selection bias

从 37,178 个关注三体话题用户的描述性统计中发现：

1)行业上：互联网,计算机，金融和科技行业用户居多；
2)教育上：大学及以上学历人数较多，知名高校占一定比重；
3)地区上：一二线城市用户较多

但知乎 app 本身的用户群体同样具有类似特征，即：

1)互联网从业者居多
2)具有良好教育背景的白领及大学生较多
3)都市用户较多

(*来源：《知乎产品分析报告》https://zhuanlan.zhihu.com/p/25844273*)

所以得出：
1)抽取的用户虽然关注了'三体'这一具体话题，但能够反映知乎用户的普遍特征;
2)反之可以推断，整个知乎用户群体对'三体'应该具有较高的兴趣度。

即，有可能使用知乎，就有可能会关注‘三体’这一类话题，因为这个群体本身就具有较大的好奇心。

二．用户指标分析

2.1 描述性统计结论

```
userinfo.describe()
```

	answer	articles	follower	following	voteup	thanked	favorited
count	37178.000000	37178.000000	37178.000000	37178.000000	37178.000000	37178.000000	3.717800e+04
mean	30.155818	1.260692	386.867637	109.666846	1035.904756	193.361961	5.912669e+02
std	108.300571	12.242286	6951.232235	369.255743	10280.623145	1835.399066	8.577348e+03
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00
25%	0.000000	0.000000	6.000000	9.000000	0.000000	0.000000	0.000000e+00
50%	6.000000	0.000000	23.000000	32.000000	18.000000	5.000000	7.000000e+00
75%	24.000000	0.000000	102.000000	103.000000	250.000000	59.000000	1.050000e+02
max	5847.000000	1067.000000	568275.000000	43909.000000	612928.000000	114131.000000	1.028586e+06

结论：
从中位数和 75%位数发现，大部分用户活跃度不是特别高,有少量网络大 V 和活跃用户

answer:用户回答问题数　　articles:用户文章数量　　follower:用户的关注者数量
following:用户关注其他用户的数量　　voteup:用户获得的点赞个数
thanked:用户被感谢的次数　　favorited:用户被收藏的次数

2.2 指标相关系数矩阵

```
. correlate answer articles follower following voteup thanked favorited
(obs=37178)
```

	answer	articles	follower	follow~g	voteup	thanked	favori~d
answer	1.0000						
articles	0.0923	1.0000					
follower	0.1525	0.1389	1.0000				
following	0.1008	-0.0026	0.0433	1.0000			
voteup	0.2441	0.1452	0.6931	0.0335	1.0000		
thanked	0.2466	0.1296	0.7691	0.0360	0.9522	1.0000	
favorited	0.1504	0.1673	0.7337	0.0358	0.7034	0.7898	1.0000

结论：

1）answer，articles, following 与其他变量都基本不相关

2）follower, voteup, thanked, favorited 四个变量之间非常相关

2.3 K-means 聚类分析

结论:

大多数用户为 少量 followers（<100），少量 answers(<10)的普通用户；

极少数用户(3.5%) followers >1k, 0.7%用户 followers >5k，0.3%用户 followers>10k

2.3.1 算法原理

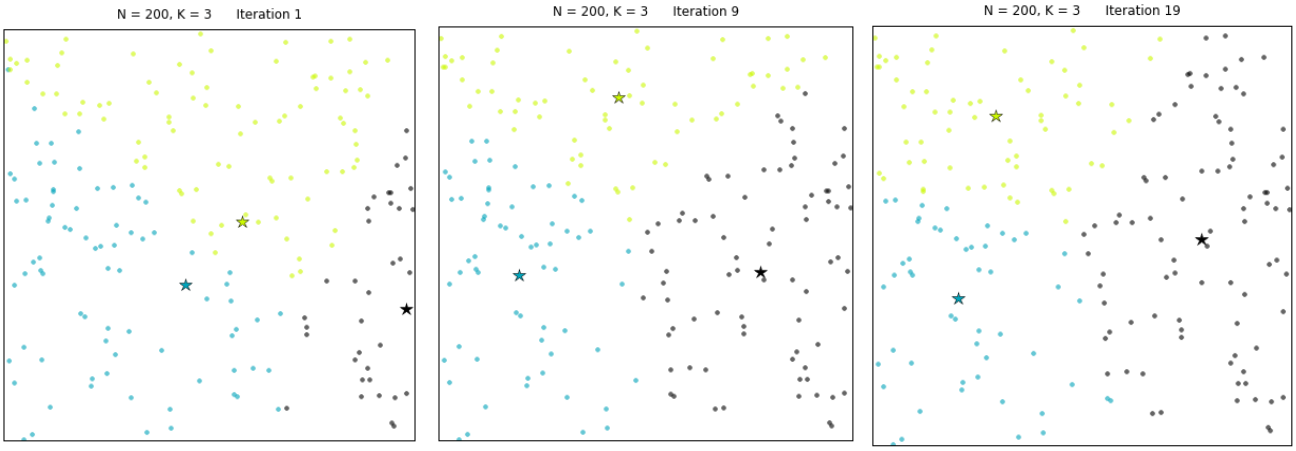
将数据进行分组聚类。组内的数据相似性越大，组间的差别越大，则聚类效果越好。

算法步骤：

1.随机选择 K 个聚类中心
2.将每个数据点分类至离其最近的聚类中心
3.根据第 2 步的分类，重新计算每一个分类的中心
4.用第 3 步中的分类中心重新为所有数据点分类
5.重复 3，4 两步直至收敛

示例：

人为预先设定 k 值，***代表同一类数据点的中心**，**同种颜色代表同一类点**；通过多次迭代最终中心不再变化，达到收敛状态；



2.3.2 回答问题数与关注者数

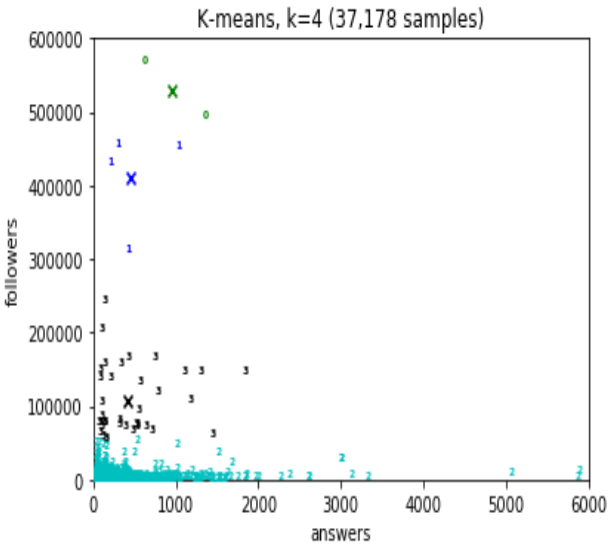
结论： 大多数用户为 少量 followers，少量 answers；

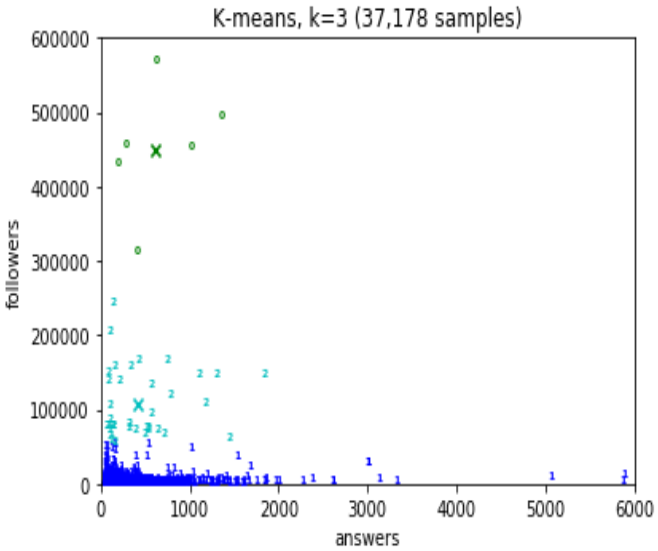
极少数用户 followers >5k, answers>1k;

K 值选取理由： 可能有以下四类用户,区分明显时取 k=4,不明显时 k=3

1)网络大 V：answer 多，follower 多 2）名人：answer 少,follower 多

3)活跃用户:answer 多，follower 少 4）普通不活跃用户: answer,follower 均少

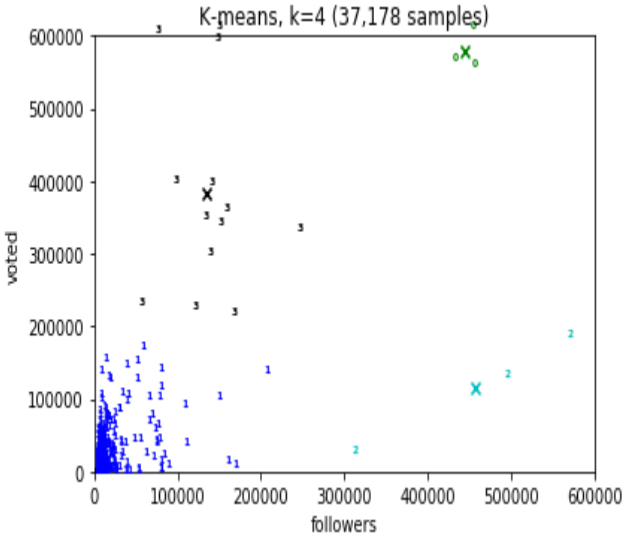


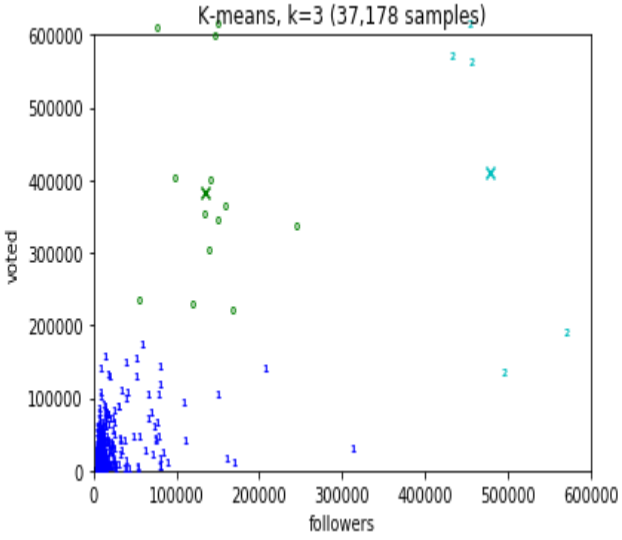


2.3.3 被点赞个数与关注者数

结论： followers 与 voted 高度正相关;

绝大多数用户 followers 与 voted 较低

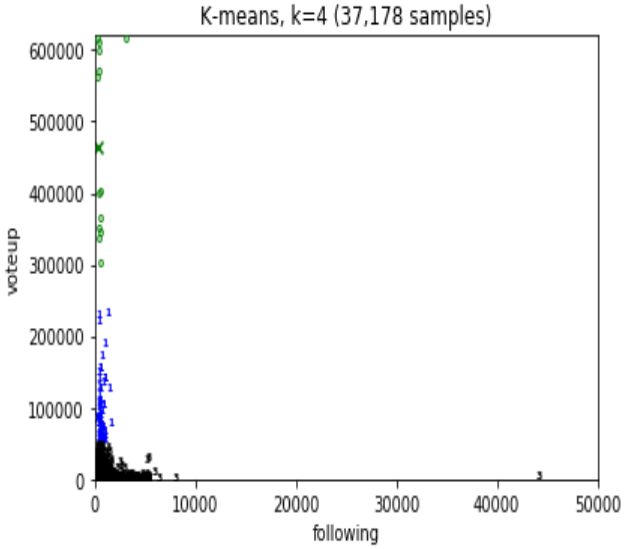


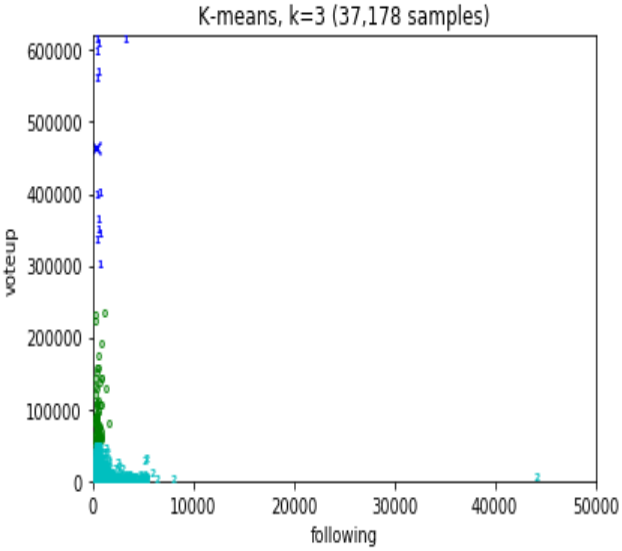


2.3.4 被点赞个数与关注其他用户个数

结论： 用户平均获赞数 voteup 显著高于 following;

多数用户获赞数仍较少;





2.3.4 被点赞个数与被收藏数

结论： 二者高度 正相关；

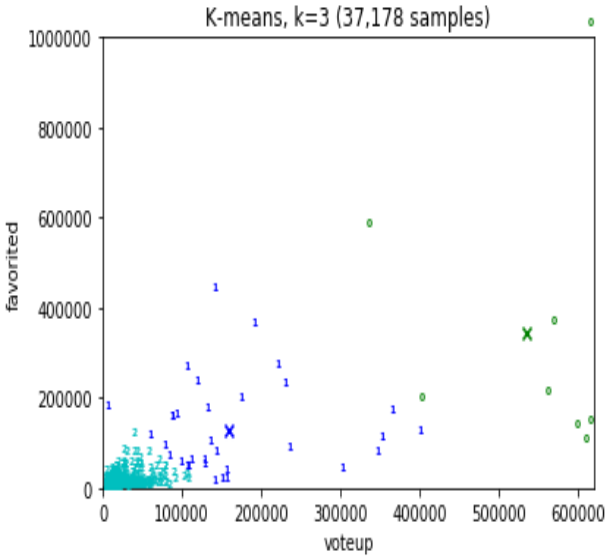
大部分用户两者值都较小；

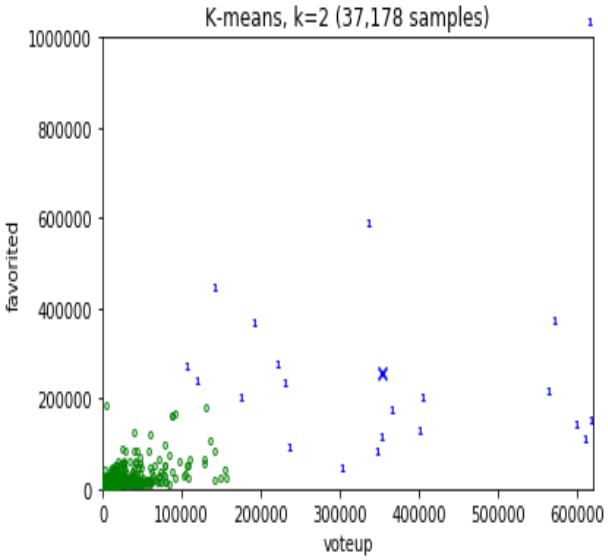
K 值选取理由： 可能主要有以下两大类用户

1）回答多，质量高：voteup 多，favorited 多

2）回答少，质量低：voteup 少，favorited 少

另外有可能点赞次数很多，但没有都被收藏

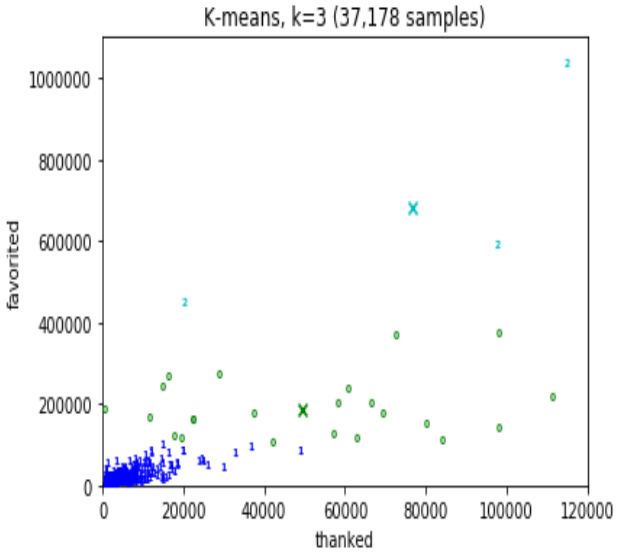


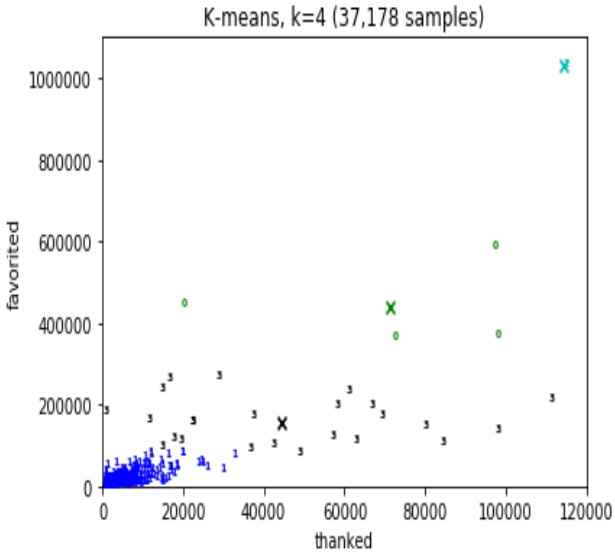


5.被感谢次数与被收藏次数

结论： 二者高度 正相关；大部分用户两者值都较小；

k 值选取理由类似；





三．用户个性签名 headlines 文本分析

结论：用户签名与用户行业，教育描述性统计特征相一致，即：

1）互联网，计算机（以狗自称）,设计，金融等行业用户较多　2)白领与大学生用户较多