描述性的数据分析

周世祥

2020/5/24

小说数据分析

具体数据来自狗熊会微信公众号，输入关键词“网络小说”，查找。

变量有小说信息(人气排行，名称，作者，类型，总字数，性质，写作进程，授权状态，更新时间，内容简介)，会员评价(点击数，周点击数，评论数，评分)。

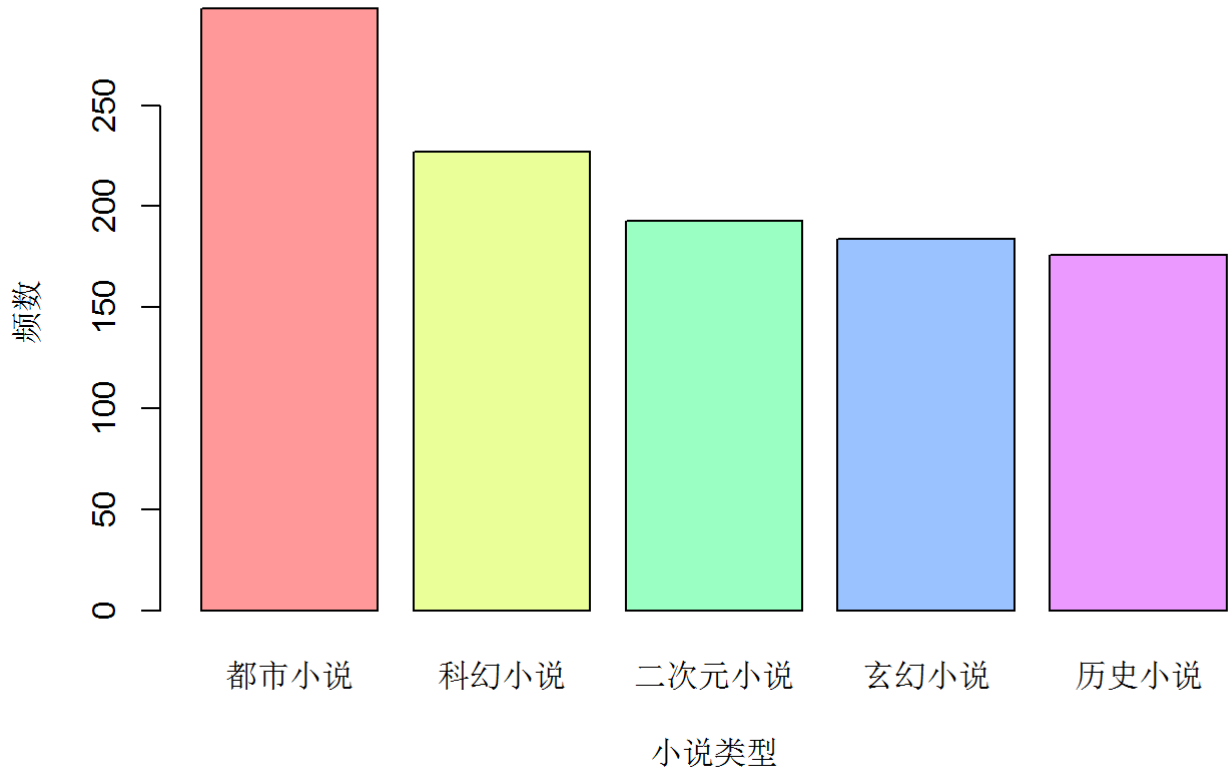
1. 数据准备

```
### 数据准备 ###
# 清空工作空间
rm(list = ls())
# 载入相关包及设定路径
# install.packages(plyr)
library(plyr)
# install.packages("reshape2")
library(reshape2)
# 读入数据
novel = read.csv("novel.csv", fileEncoding = "UTF-8")
# 数据查看与异常处理
head(novel)
```

```
##    人气排序 小说名称      作者 小说类型 总点击数 会员周点击数  总字数 评论数 评分
## 1        1 一念永恒      耳根 仙侠小说  4383898        10691 1155534 435429  9.8
## 2        2 斗战狂潮 骷髅精灵 仙侠小说  1678379        36587  422116  23159 10.0
## 3        3     天影      萧鼎 仙侠小说  1248708        32019  373763  25253  9.8
## 4        4 不朽凡人 鹅是老五 仙侠小说  2457382         9610  995669 146715  9.9
## 5        5 玄界之门      忘语 仙侠小说  3736897         6709 1784999 238113  9.8
## 6        6 龙王传说 唐家三少 玄幻小说  2968846         3080 1552654 293934  9.8
##   小说性质 写作进程 授权状态          更新时间
## 1 公众作品   连载中 A级签约 2016/10/23 11:50
## 2 公众作品   连载中 A级签约 2016/10/22 17:05
## 3 公众作品   连载中 A级签约 2016/10/23 10:40
## 4 公众作品   连载中 A级签约 2016/10/22 20:50
## 5 公众作品 新书上传 A级签约 2016/10/23 10:15
## 6 公众作品 新书上传 A级签约  2016/10/23 7:00
##
内容简介
## 1
    一念成沧海，一念化桑田。一念斩千魔，一念诛万仙。???    唯我念……永恒???    这是耳根继《仙逆》《求魔》《我欲封天》后，创作的第四部长篇小说《一念永恒》
## 2     双月当空，无限可能的英魂世界 \n???    孤寂黑暗，神秘古怪的嬉命小丑 \n???    百城联邦，三大帝国，异族横行，魂兽霸幽 \n???    这是一个英雄辈出的年代，人类卧薪尝胆重掌地球主权，孕育着进军高纬度的野望！ \n???    重点是……二年级的废柴学长王同学，如何使用嬉命轮盘，撬动整个世界，学妹们，请注意，学长来了！！！ \n???    斗战一群：21222419（两千人战力群）\n???    斗战二群：12962047 \n???    骷髅的微信公共号：kuloujingling00 \n???    新浪微博：骷髅精灵
## 3                                                                                       阴阳分天地，五行定乾坤。\n???    天穹之下岁月沧桑的中土神州，正是仙道昌盛的时代，亿万生灵欣欣向荣。\n???    纵横千万里间，总有人间一幕幕悲欢离合，在恢弘长生的仙道中上演着。\n???    有光便有暗，天穹之下光辉之中，仍有沉默的影子悄然前行着……\n???    新书上线！精彩万分！请各位书友多多投票支持！另外，大家可以添加微信公众号zhuxianxiaoding（诛仙萧鼎），QQ官方群 176378308 进行交流。
## 4
    在这里，拥有灵根才能修仙，所有凡根注定只是凡人。\n???    莫无忌，只有凡根，一介凡人！\n???    是就此老去，还是不甘？???
## 5
    天降神物！异血附体！???    群仙惊惧！万魔退避！???    一名从东洲大陆走出的少年。???    一具生死相依的红粉骷髅。???    一个立志成为至强者的故事。???    一段叱咤星河，大闹三界的传说。???    忘语新书，已完本《凡人修仙传》》《魔天记》。
## 6
    伴随着魂导科技的进步，斗罗大陆上的人类征服了海洋，又发现了两片大陆。魂兽也随着人类魂师的猎杀无度走向灭亡，沉睡无数年的魂兽之王在星斗大森林最后的净土苏醒，它要带领仅存的族人，向人类复仇！\n???    唐舞麟立志要成为一名强大的魂师，可当武魂觉醒时，苏醒的，却是……\n???    旷世之才，龙王之争，我们的龙王传说，将由此开始。\n???
```

2. 单变量

```
## 定性变量--柱状图 ##
a = table(novel$小说类型)
a = a[order(a, decreasing = T)]
barplot(a[1:5], names.arg = names(a)[1:5], col = rainbow(5, alpha = 0.4), xlab = "小说类型", ylab="频数")
```

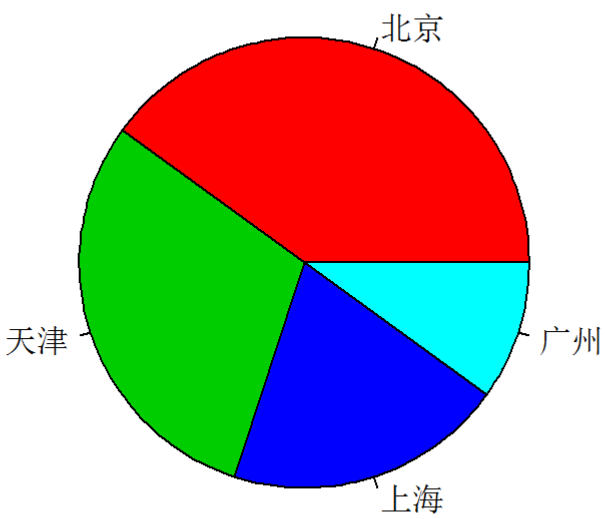



```
#names.arg参数定义每个柱子的名称
#col 定义颜色，rainbow彩虹图，alpha透明度
#main定义图标题
```

上面的图是不同类型小说分布的柱状图。可以了解目前的流行趋势。

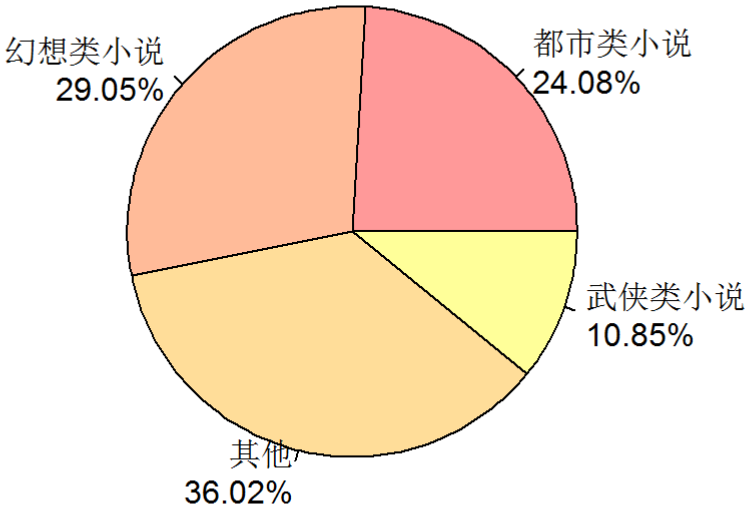
饼图

```
## 定性变量--饼图 ##
pie(c(4000, 3000, 2000, 1000), labels = c("北京", "天津", "上海", "广州"), main = "熊粉成员分布", col = 2:5)
```



复杂点的，结合小说数据，先合并类型，画饼图

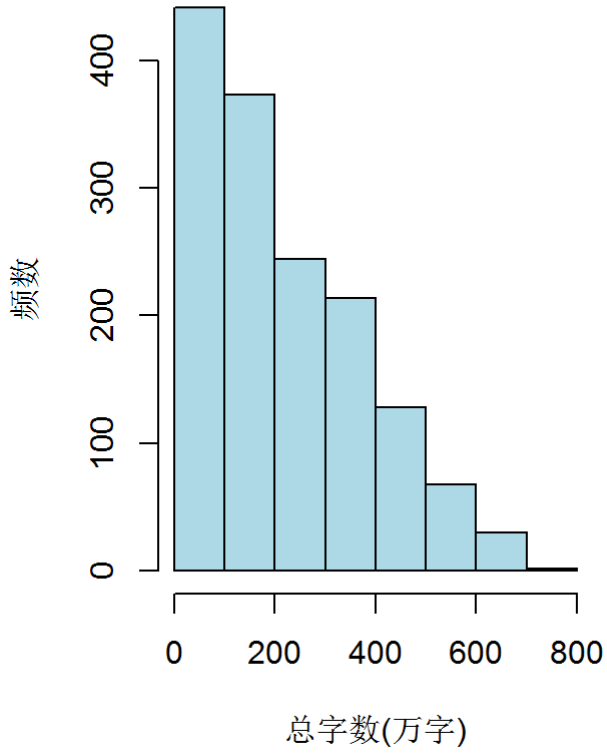
```
# 将小说类型进行简要合并
novel$'小说类别' = "其他"
novel$'小说类别'[novel$小说类型 == "都市小说" | novel$小说类型 == "职场小说"] = "都市类小说"
novel$'小说类别'[novel$小说类型 == "科幻小说" | novel$小说类型 == "玄幻小说" | novel$小说类型 =
= "奇幻小说"] = "幻想类小说"
novel$'小说类别'[novel$小说类型 == "武侠小说" | novel$小说类型 == "仙侠小说"] = "武侠类小说"
# 求出每一类所占百分比
ratio = table(novel$'小说类别') / sum(table(novel$'小说类别')) * 100
# 定义标签
label1 = names(ratio)
label2 = paste0(round(ratio, 2), "%")
# 画饼图
pie(ratio, col = heat.colors(5, alpha = 0.4), labels = paste(label1, label2, sep = "\n"), font
 = 1)
```

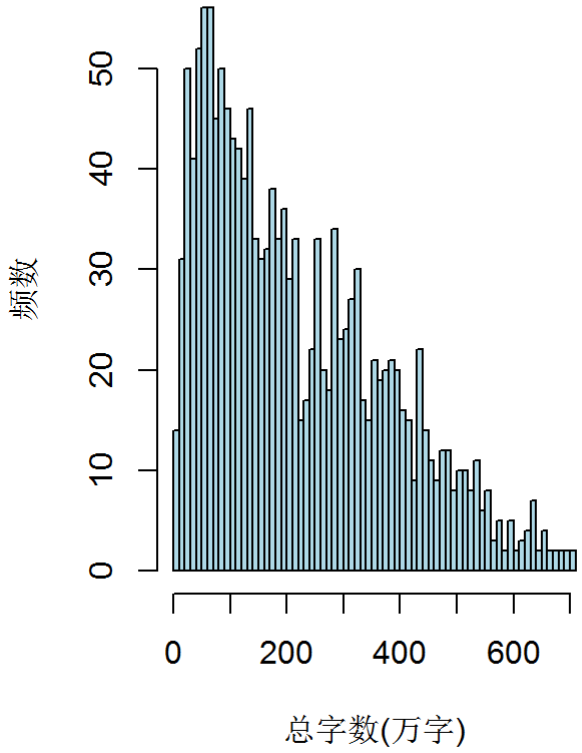


当然不是任何一个定性变量都适合画饼图。

单个定量变量

```
## 定量变量--直方图 ##
novel$总字数 = novel$总字数 / 10000
par(mfrow = c(1, 2))
chara = sort(novel$总字数)[1:1500]  # 去掉异常值
hist(chara, breaks = 10, xlab = "总字数(万字)", ylab = "频数", main = "", col = "lightblue")
hist(chara, breaks = 100, xlab = "总字数(万字)", ylab = "频数", main = "", col = "lightblue")
```





对横截面数据来说，最重要的就是它的分布，直方图能直观地展示数据的分布形态即异常。

R语言画直方图命令是hist().

折线图

```
## 定量变量--折线图 ##
par(mfrow = c(1, 1))
# 画时间序列图
data(AirPassengers)
head(AirPassengers)
```

```
## [1] 112 118 132 129 121 135
```

```
## [1] 112 118 132 129 121 135
class(AirPassengers)
```

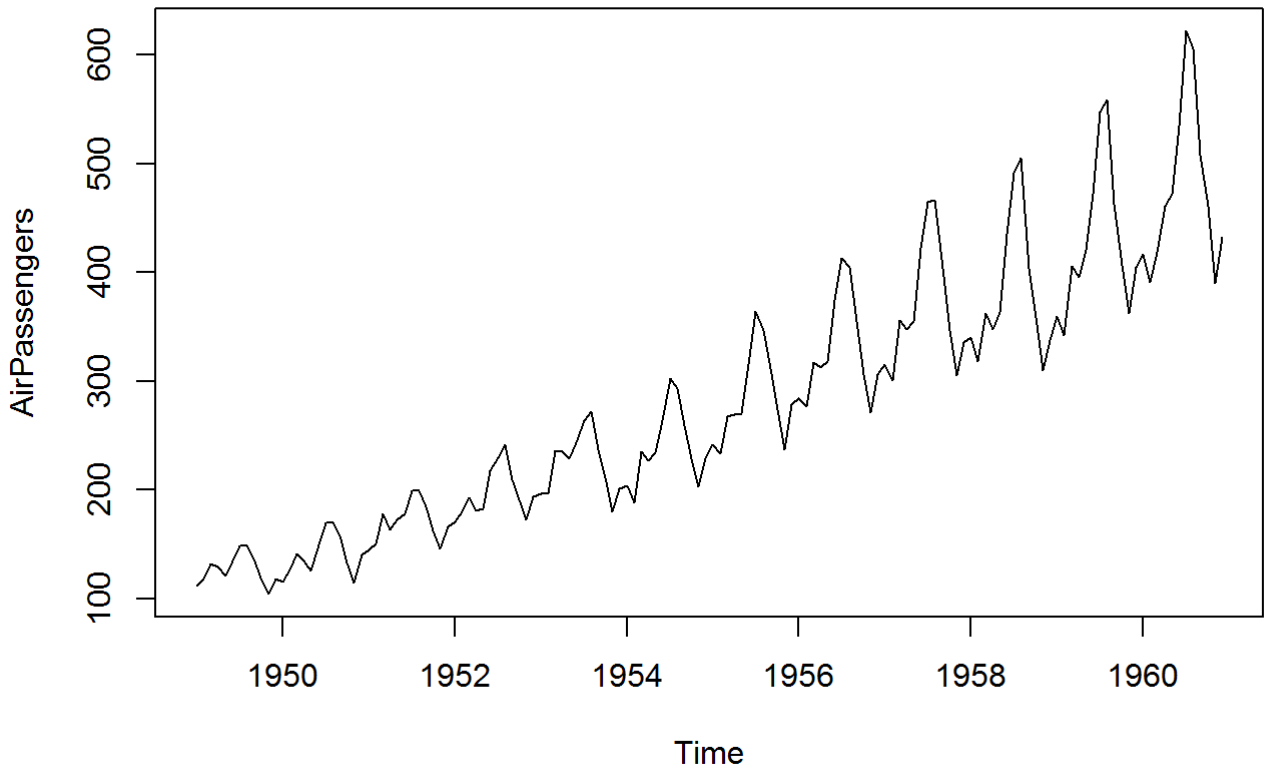
```
## [1] "ts"
```

```
## [1] "ts"
plot(AirPassengers)
# 人民的名义百度搜索指数图
# install.packages(zoo)
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.0.0
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```



针对时间序列，观察指标随时间变化的趋势，看趋势，走向。

R中的tz格式可以直接采用plot(x)。

参考：https://www.cnblogs.com/luhuajun/p/8504187.html (https://www.cnblogs.com/luhuajun/p/8504187.html)

如果数据仅仅是一个普通的向量，又该如何将其变成可用于画图及后续时间序列分析的数据格式？如果是年月或季度数据，可以采用tz()函数直接转换，如果数据是天数据或者不等间隔的时序数据，可以选择另一个包zoo来生成。

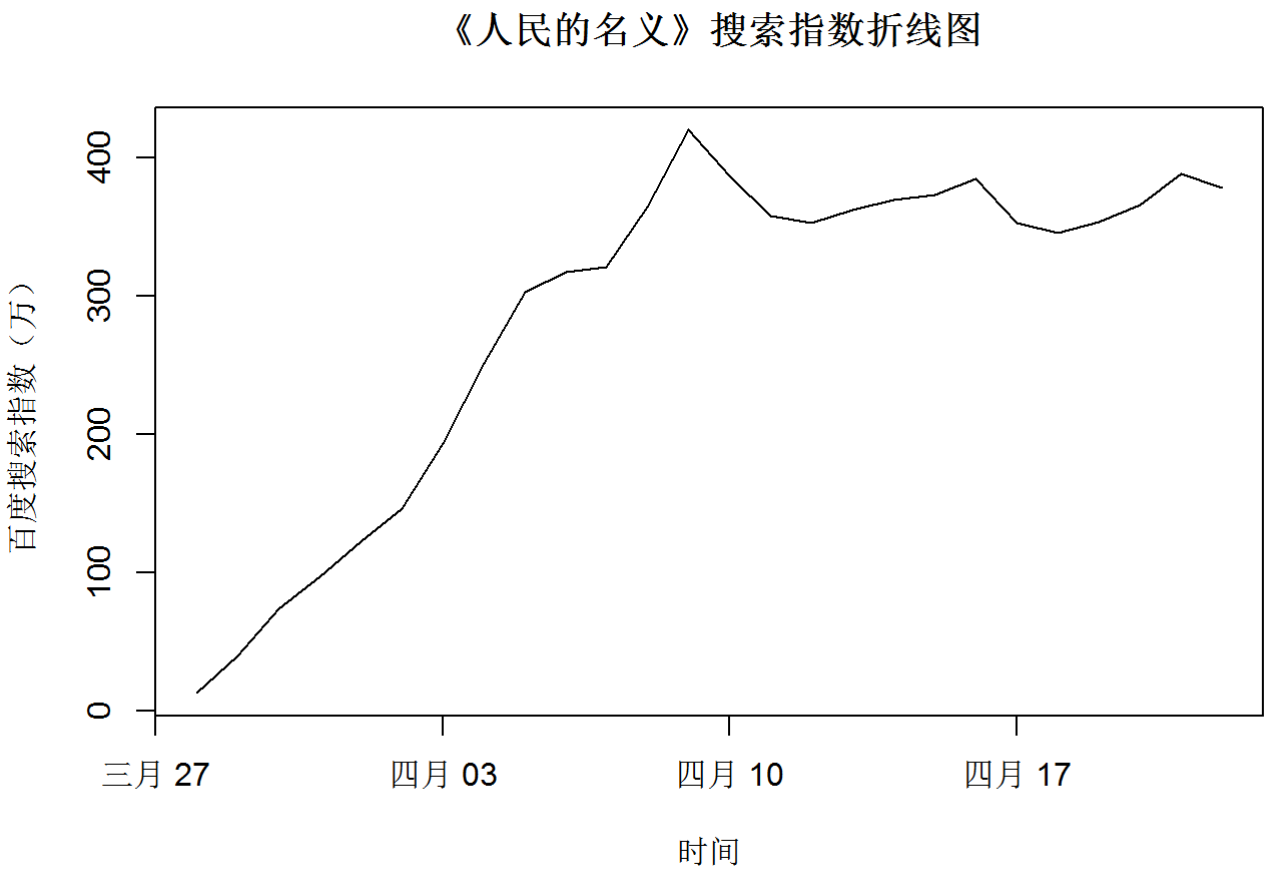
以《人民的名义》的百度搜索指数为例示范后一种情况。

首先，设定好时间标签data，使用zoo函数将时间标签及对应的数据“组合”在一起，将数据改为时间序列格式后，采用plot函数。

```
# 将搜索指数index变成时间序列格式
index = c(127910, 395976, 740802, 966845, 1223419, 1465722, 1931489, 2514324, 3024847, 3174056,
3208696, 3644736, 4198117, 3868350, 3576440, 3524784, 3621275, 3695967, 3728965, 3845193, 35255
79, 3452680, 3535350, 3655541, 3884779, 3780629) / 10000
date = seq(as.Date("2017-3-28"), length = 26, by = "day")
people_index = zoo(index, date)
class(people_index)
```

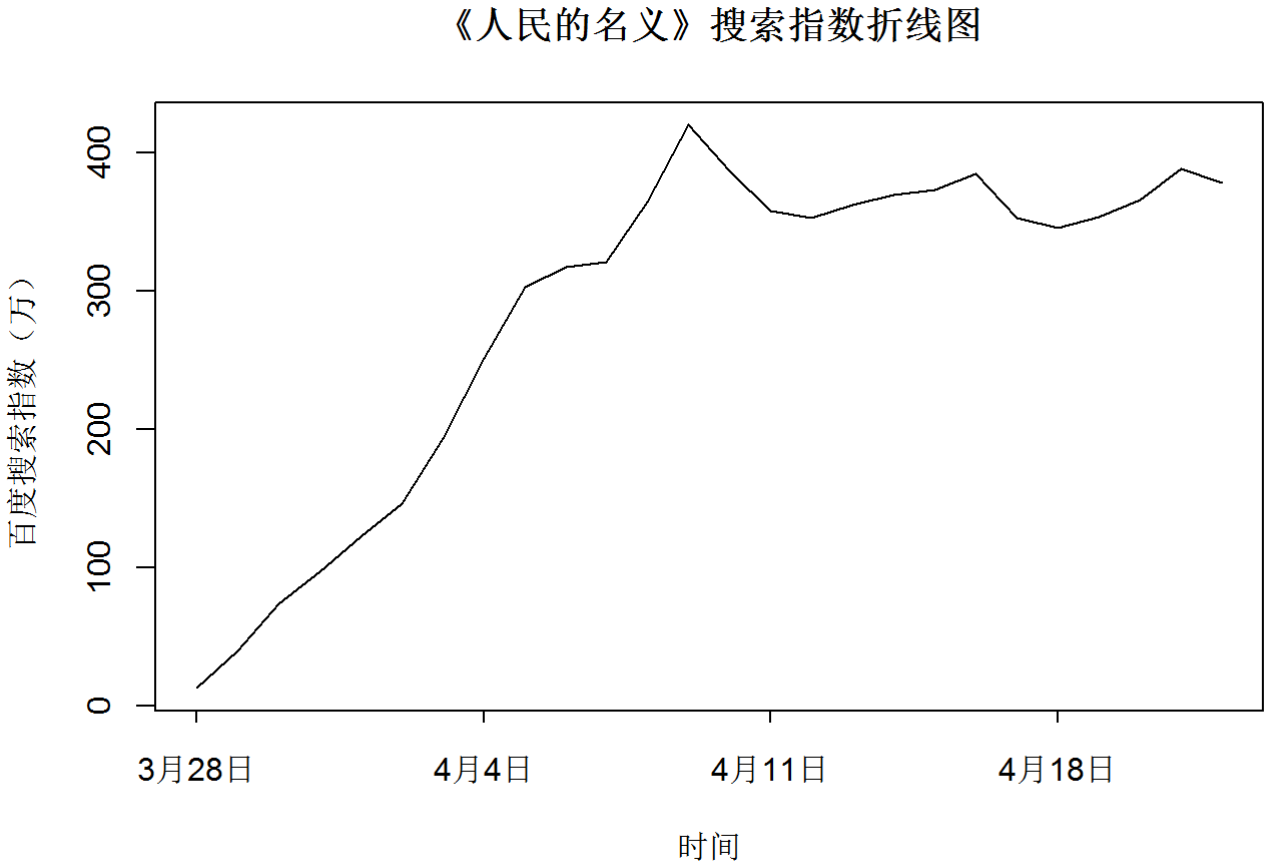
```
## [1] "zoo"
```

```
## [1] "zoo"
plot(people_index, xlab = "时间", ylab = "百度搜索指数（万）", main = "《人民的名义》搜索指数折线图")
```



如果对横轴显示时间格式不满意，可以通过axis函数中tick和label_name参数来定义标签。

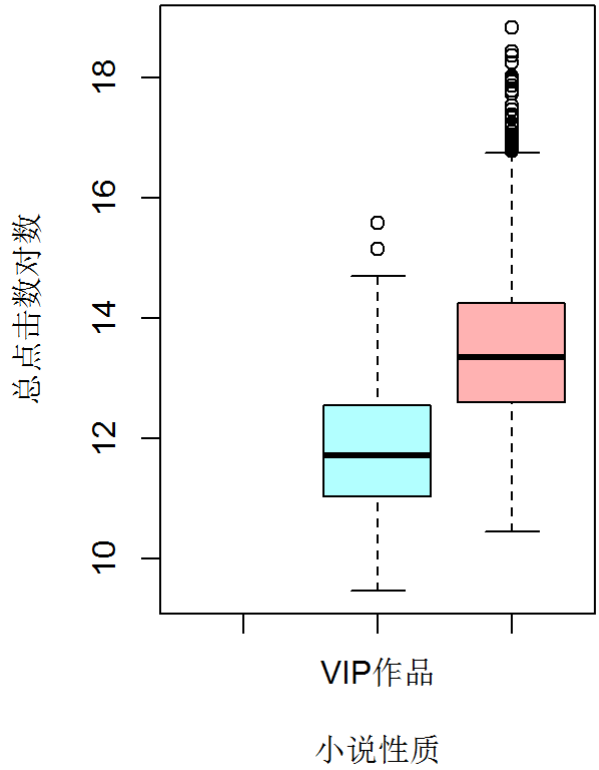
```
# 更改坐标轴显示内容
plot(people_index, xaxt = "n", xlab = "时间", ylab = "百度搜索指数（万）", main = "《人民的名义》搜索指数折线图")
times = date #or directly times = x.Date
ticks = seq(times[1], times[length(times)], by = "weeks")  # month, weeks, year etc.
label_name = c("3月28日", "4月4日", "4月11日", "4月18日")
axis(1, at = ticks, labels = label_name, tcl = -0.3)
```

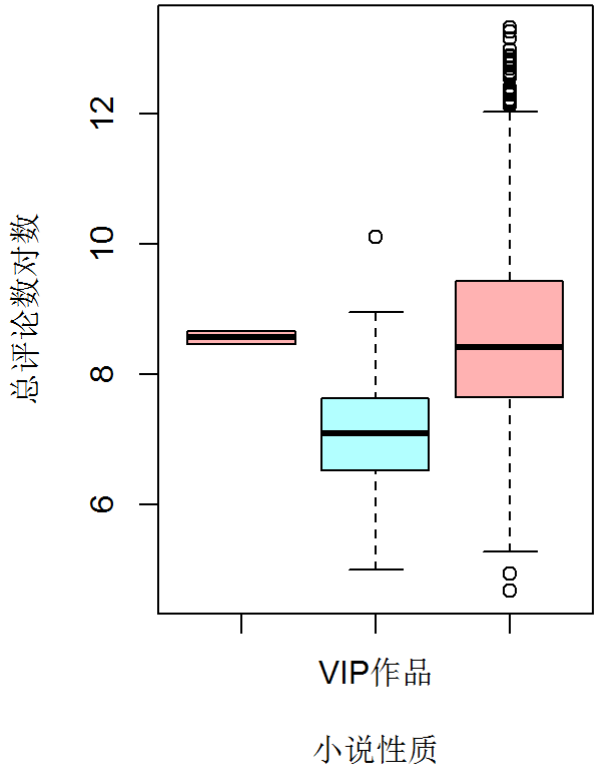


两个变量

```
### 两个变量 ###

## 定性与定量变量--分组箱线图 ##
# 将画板分成1行2列
par(mfrow = c(1, 2))
# 不同性质的小说总点击数和评论数有差别吗
boxplot(log(总点击数) ~ 小说性质, data = novel, col = rainbow(2, alpha = 0.3), ylab = "总点击数对数")
boxplot(log(评论数) ~ 小说性质, data = novel, col = rainbow(2, alpha = 0.3), ylab = "总评论数对数")
```





切分面板，两变量其实就是两个单变量结合在一起，理论上可以将两个变量各放一张图，然后摆在一起对着看，怎么让两幅图甚至多幅图在一起看，用到切分面板的功能，par(mfrom=c(a,b))函数将画图的屏幕切分成a行b列个小格子，然后每画一幅图就放在一个小格子里。

定性变量和定量变量之间的关系：如比较不同教育水平的收入差异，不同地段的房价差异。分组箱线图。

下面比较属于VIP作品还是公众作品点击数更多？评论数多少？

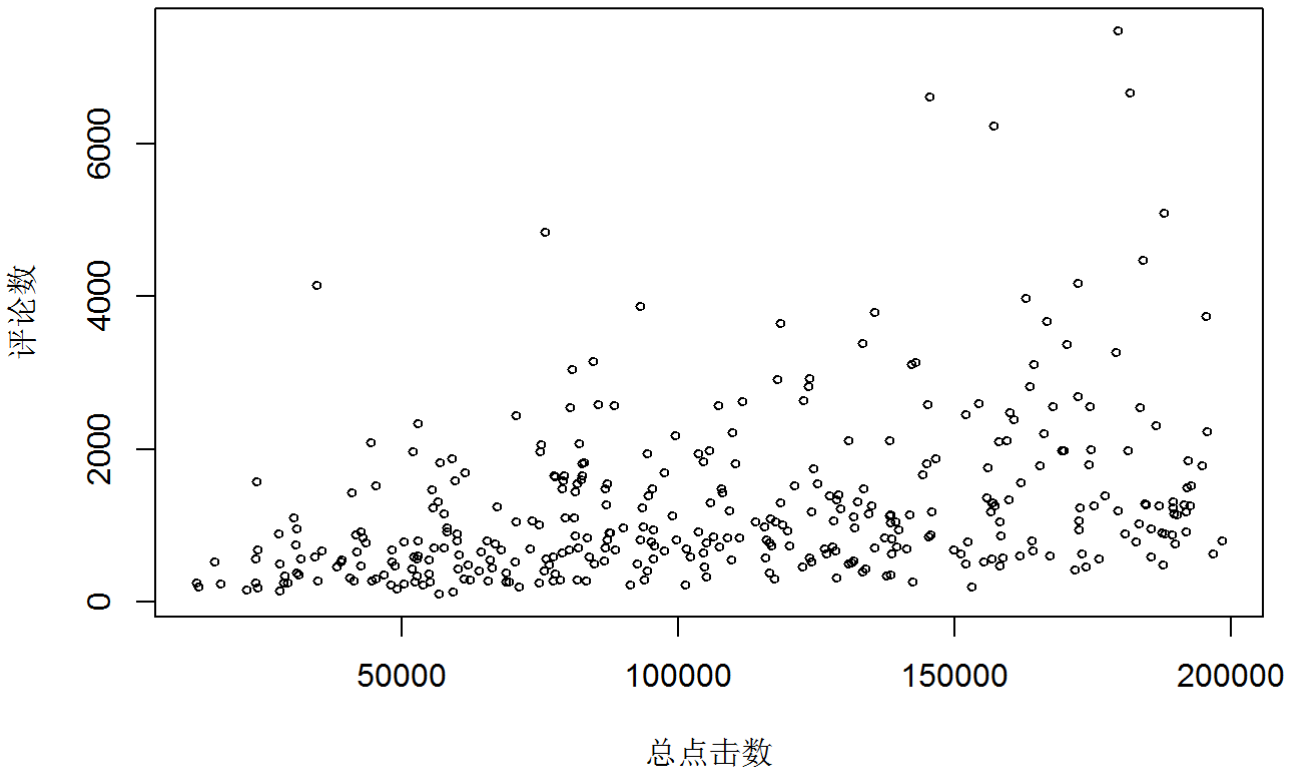
平均来看，每篇公众作品无论是点击数还是评论数都显著比VIP作品多。

画分组箱线图的命令是boxplot()可以用公式形式表示： boxplot(y~x),其中y是要比对的定量变量，x是分组变量。

两个定量变量

```
# 将画板恢复
par(mfrow = c(1, 1))

## 两个定量变量--散点图 ##
# 去除较大的异常值后画图
test = novel[novel$评论数 < 8000 & novel$总点击数 < 200000, ]
x = test$总点击数
y = test$评论数
plot(x, y, pch = 1, cex = 0.6, xlab = "总点击数", ylab = "评论数")
```



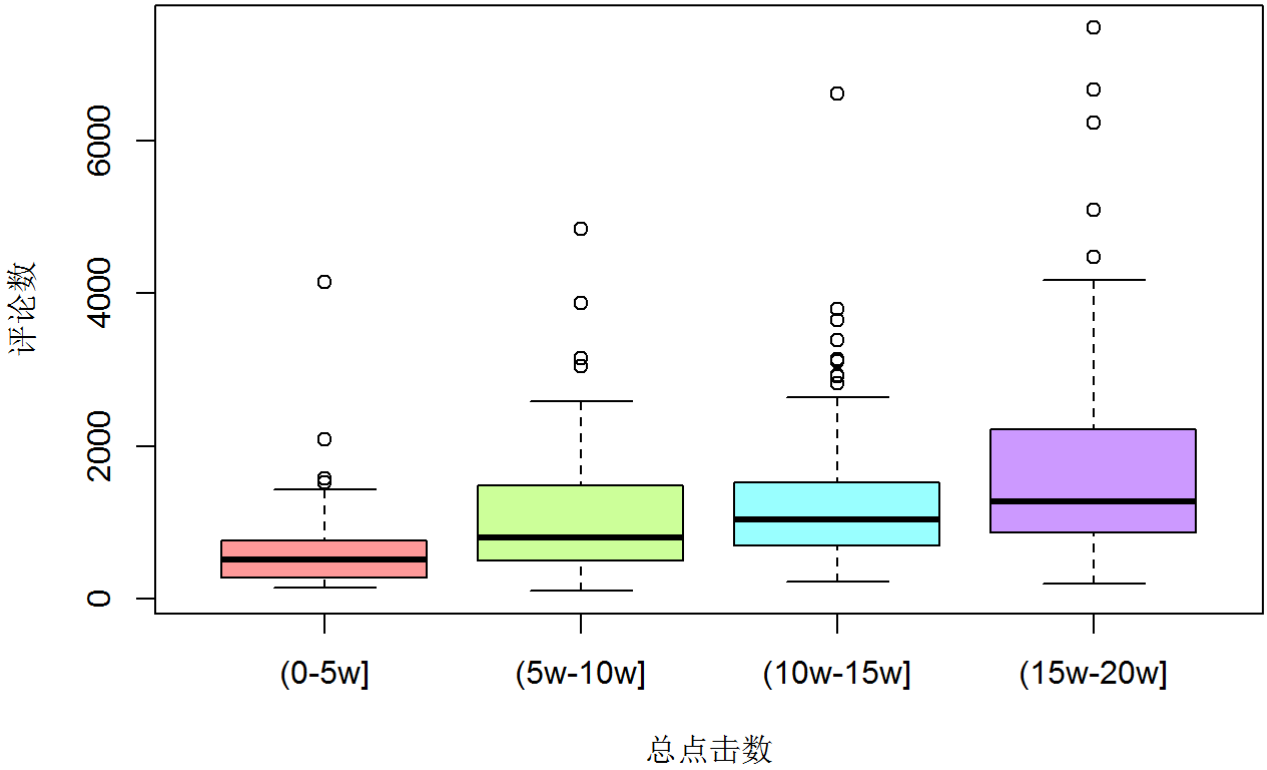
col颜色设置，pch设定点的形状，cex符号的大小。

看小说的总点击数和评论数有何关联。选取总评论数在8000以下，且总点击数在20万以下的小数数据示范。

有点正相关的迹象，相关程度不高。

可以考虑把连续变量离散化，分组，变成定性变量，将总点击数离散化，再与评论数做箱线图，相关关系更明显。

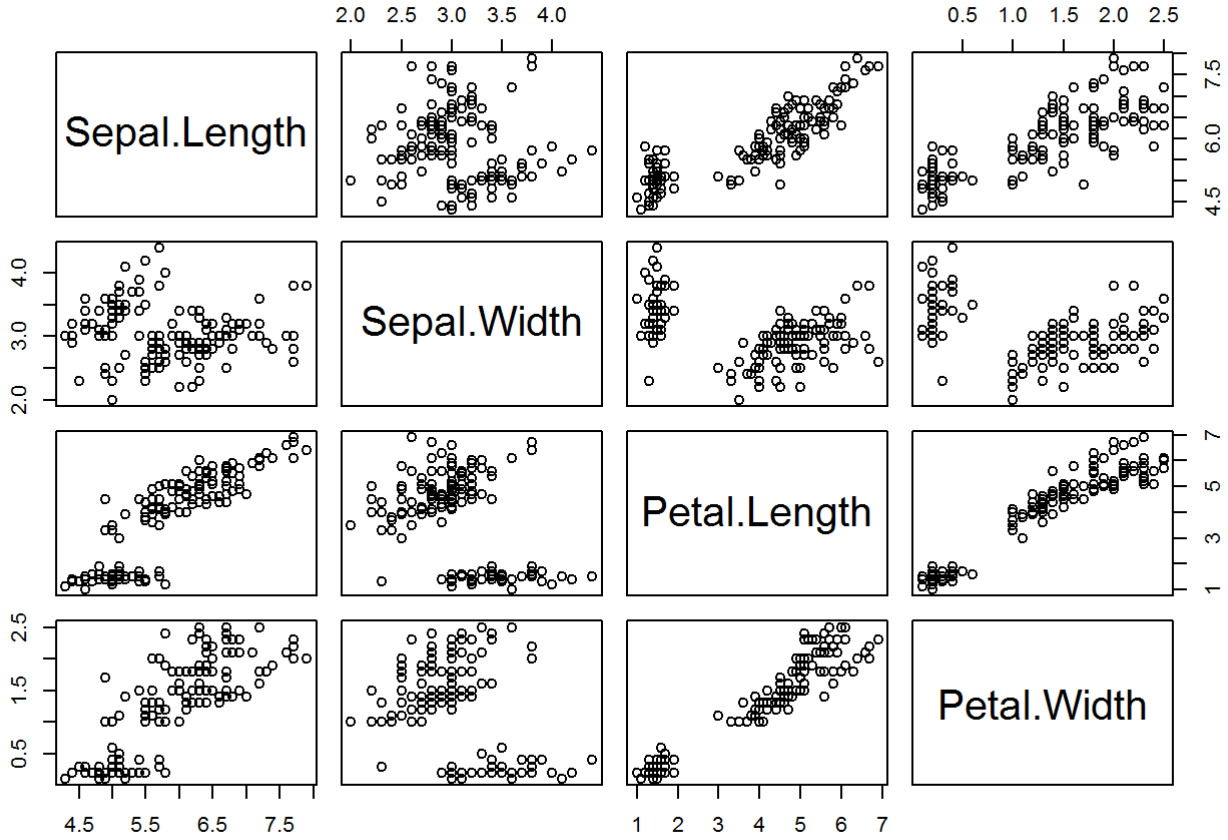
```
# 分组做分组箱线图
aa = cut(x, breaks = c(0, 50000, 100000, 150000, 200000), labels = c("(0-5w]", "(5w-10w]", "(10w-15w]", "(15w-20w]"))
boxplot(y ~ aa, col = rainbow(4, alpha = 0.4), xlab = "总点击数", ylab = "评论数")
```



实践中如果有多个变量，两两看相关图太麻烦，plot(data.frame)就可以画散点图矩阵。

一次性观察所有变量的相关关系。

```
# 散点图矩阵
plot(iris[, 1:4])
```



两个定性变量

可以用柱状图的变形—堆积柱状图和并列柱状图来表现。

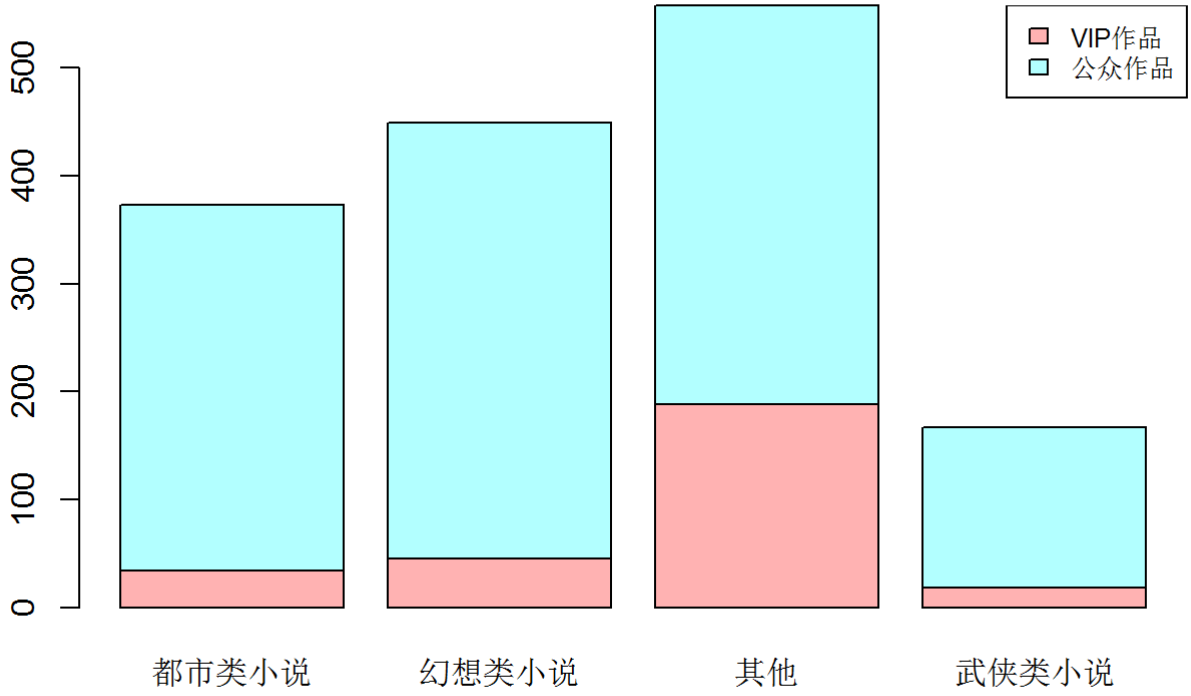
```
## 两个定性变量--柱状图 ##
a = ddply(novel, .(小说类别,小说性质), nrow)#
#ddply()函数位于plyr包，用于对data.frame进行分组统计，与tapply有些类似
d = dcast(a, 小说性质 ~ 小说类别)[-1, -1]
```

```
## Using V1 as value column: use value.var to override.
```

```
##  Using V1 as value column: use value.var to override.
rownames(d) = c("VIP作品", "大众作品")
(d = as.matrix(d))
```

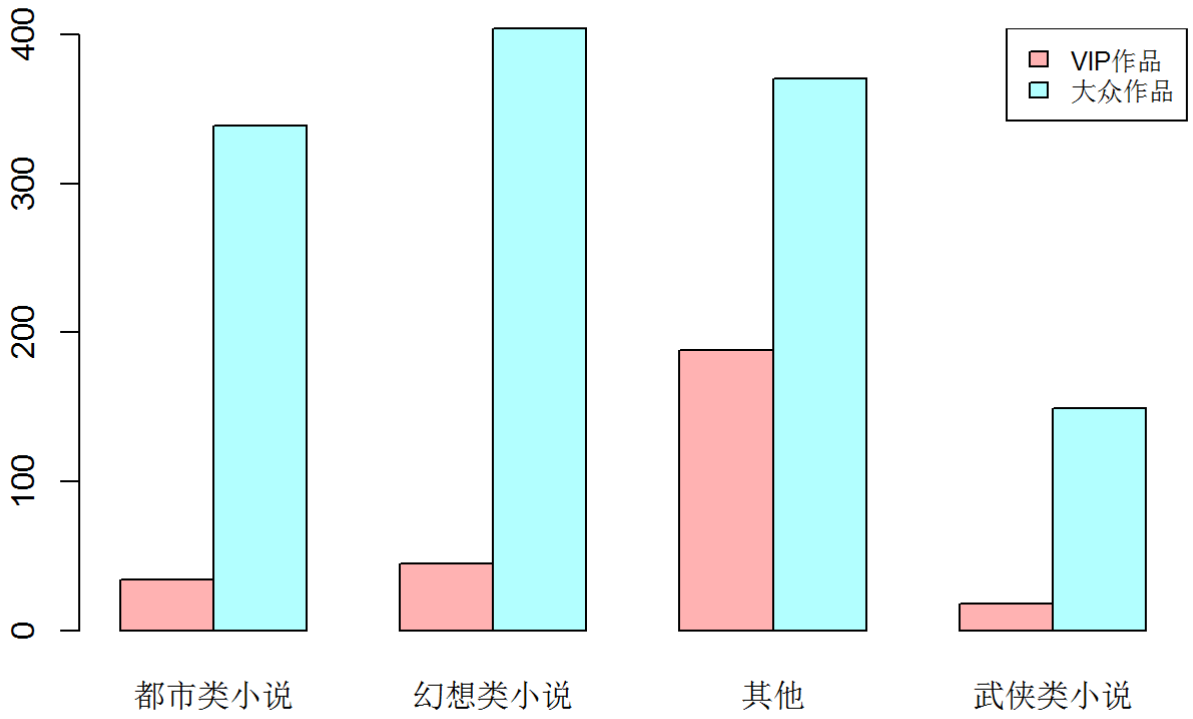
```
##           都市类小说 幻想类小说 其他 武侠类小说
## VIP作品          34         45  188          18
## 大众作品        339        404  370         149
```

```
##            都市类小说 幻想类小说 其他 武侠类小说
##  VIP作品           34         45  188          18
##  大众作品         339        404  370         149
# beside = T，按列累计
barplot(d, beside = F, col = rainbow(2, alpha = 0.3))
legend("topright", legend = c("VIP作品", "公众作品"),
       fill = rainbow(2, alpha = 0.3), cex = 0.8)
```



包含两个说变量“小说性质”及四个水平变量“小说类别”作图。

```
# beside = F，按列并列 ，加这个参数就可以
barplot(d, beside = T, col = rainbow(2, alpha = 0.3))
legend("topright",legend= c("VIP作品", "大众作品"),
       fill = rainbow(2, alpha = 0.3), cex = 0.8)
```



参考文献：

朱雪宁，《R语言从数据思维到数据实战》，中国人民大学出版社，2019