一食堂3,4月份聚类分析

> setwd("C:/Users/Administrator/Desktop")

> c1201503<-read.csv("c1201503.csv")

> c1201504<-read.csv(“c1201504.csv”)

> school<-c1201503[,1] #产生数据集

> school<-c1201504[,1]

>table(school) #显示3月份school数据集列中各个值出现频次

工商学院 公管学院 管科学院 国际商学院 会计学院 金融学院 24589 30593 20232 3874 24627 21275

经济学院 数统学院 外语学院 文法学院 信息学院 艺术学院 27088 10653 5055 15129 26620 1108

>table(school) #显示4月份school数据集列中各个值出现频次

工商学院 公管学院 管科学院 国际商学院 会计学院 金融学院

26983 35471 20821 3849 29219 23477

经济学院 数统学院 外语学院 文法学院 信息学院 艺术学院 30450 11673 5392 16845 24825 1447

#画出饼图

> lbls<-c("工商学院","公管学院","管科学院","国际商学院","会计学院","金融学院","经济学院","数统学院","外语学院","文法学院","信息学院","艺术学院")

> slices<-c(24589,30593,20232,3874,24627,21275,27088,10653,5055,15129,26620,1108)

> slices<-c(26983,35471,20821,3849,29219,23477,30450,11673,5392,16845, 24825,1447)

#为饼图添加比例数值

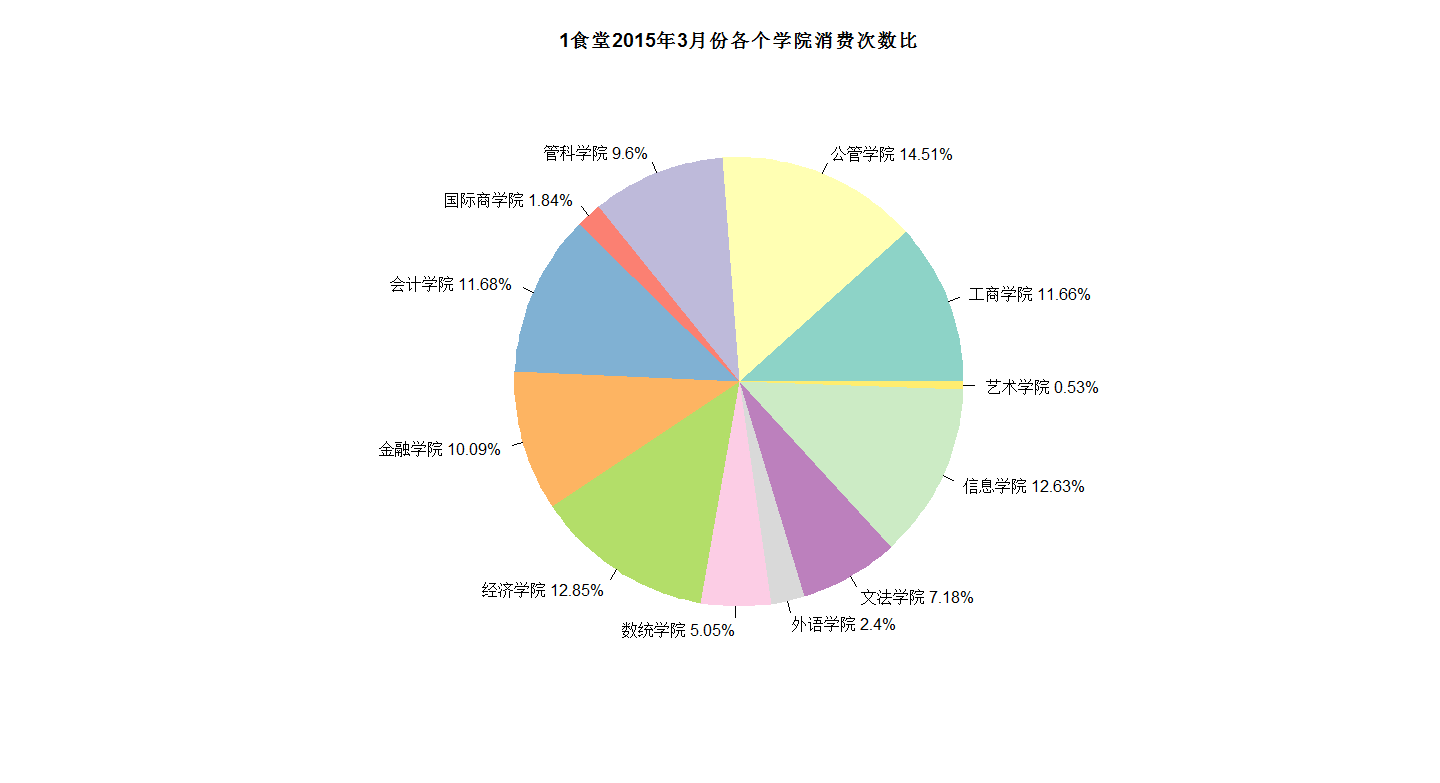
> pct<-round(slices/sum(slices)\*100,2)

> lbls2<-paste(lbls," ",pct,"%",sep="")

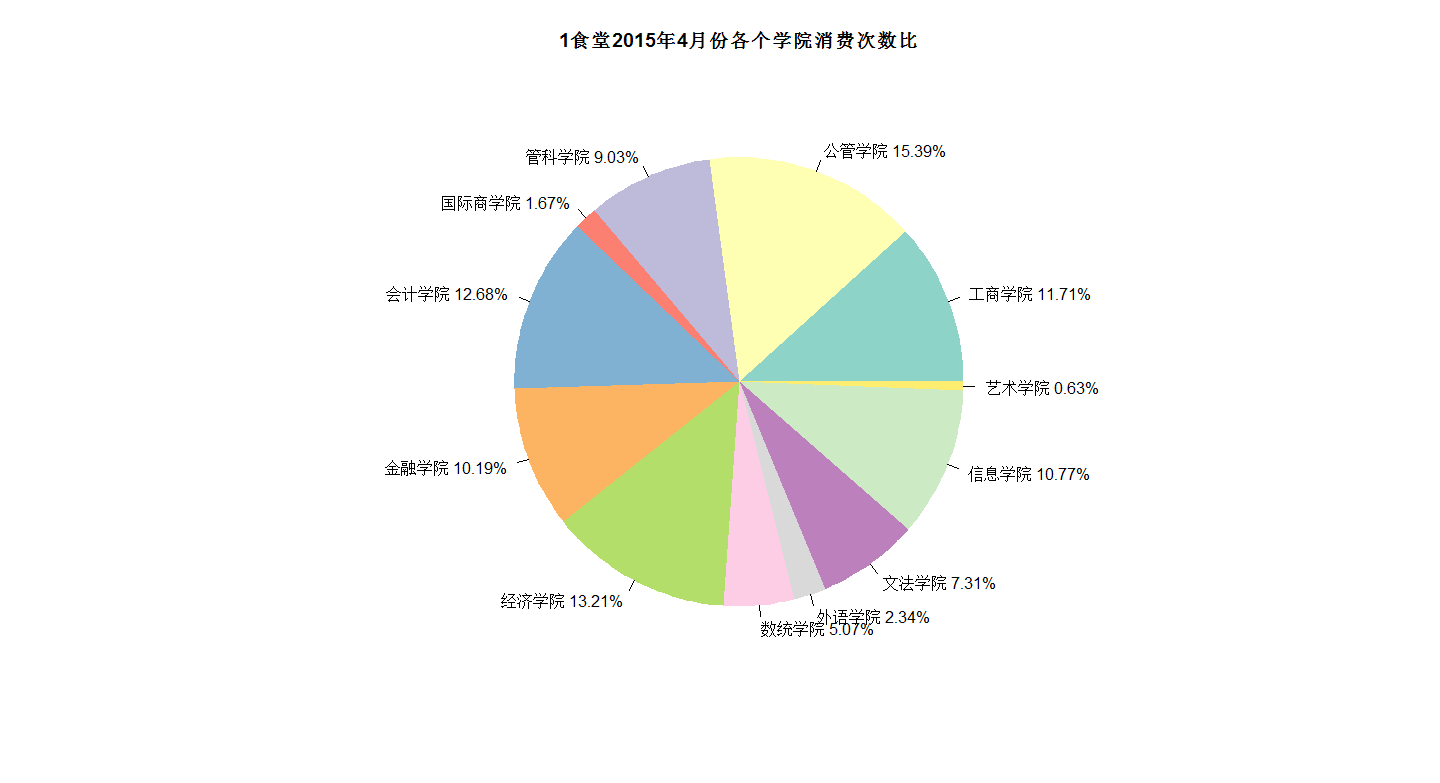
> b2p2 <- colorRamp(c("blue", "purple"))

> library(RColorBrewer)

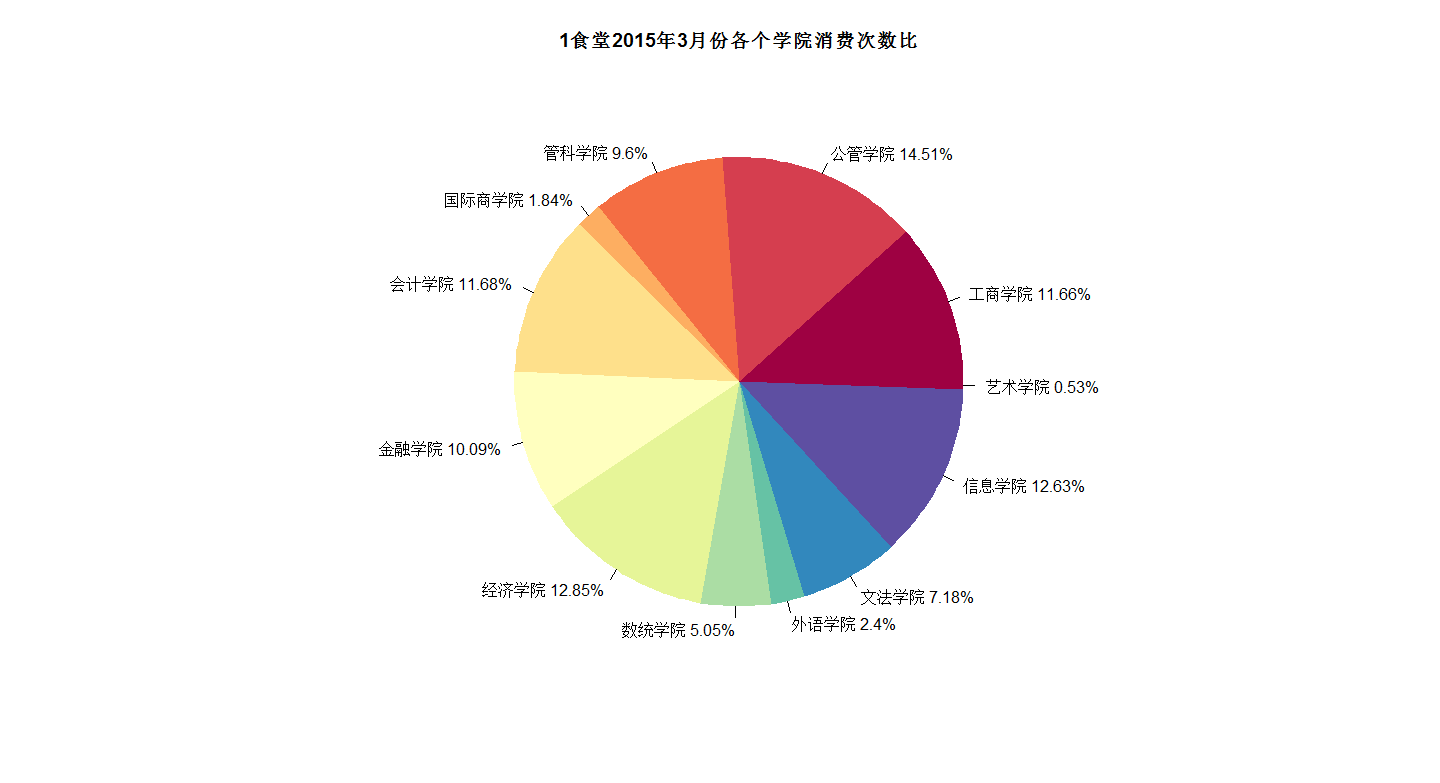
> pie(slices,labels=lbls2,col=brewer.pal(12,"Set3"),border=FALSE,main="1食堂2015年3月份各个学院消费次数比")



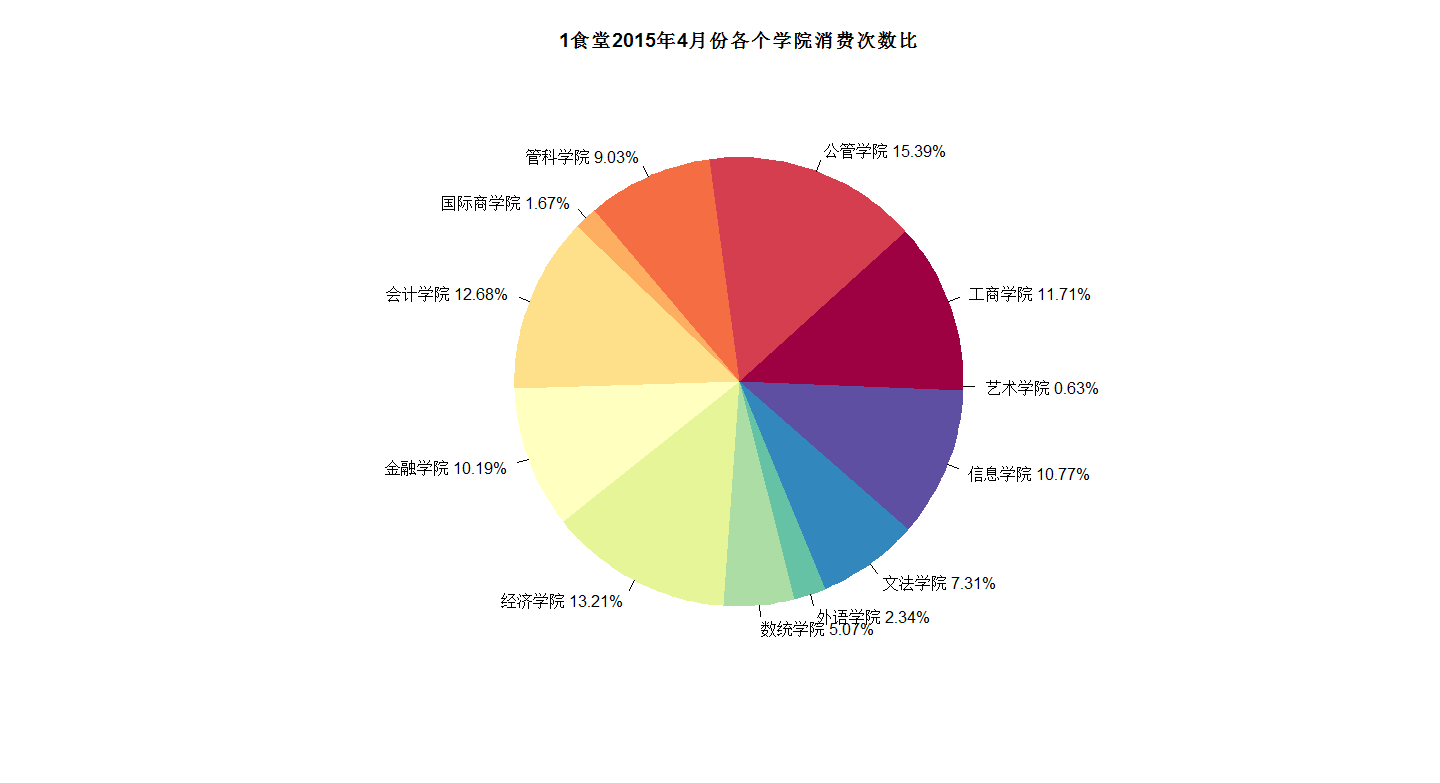
> pie(slices,labels=lbls2,col=brewer.pal(12,"Set3"),border=FALSE,main="1食堂2015年4月份各个学院消费次数比")



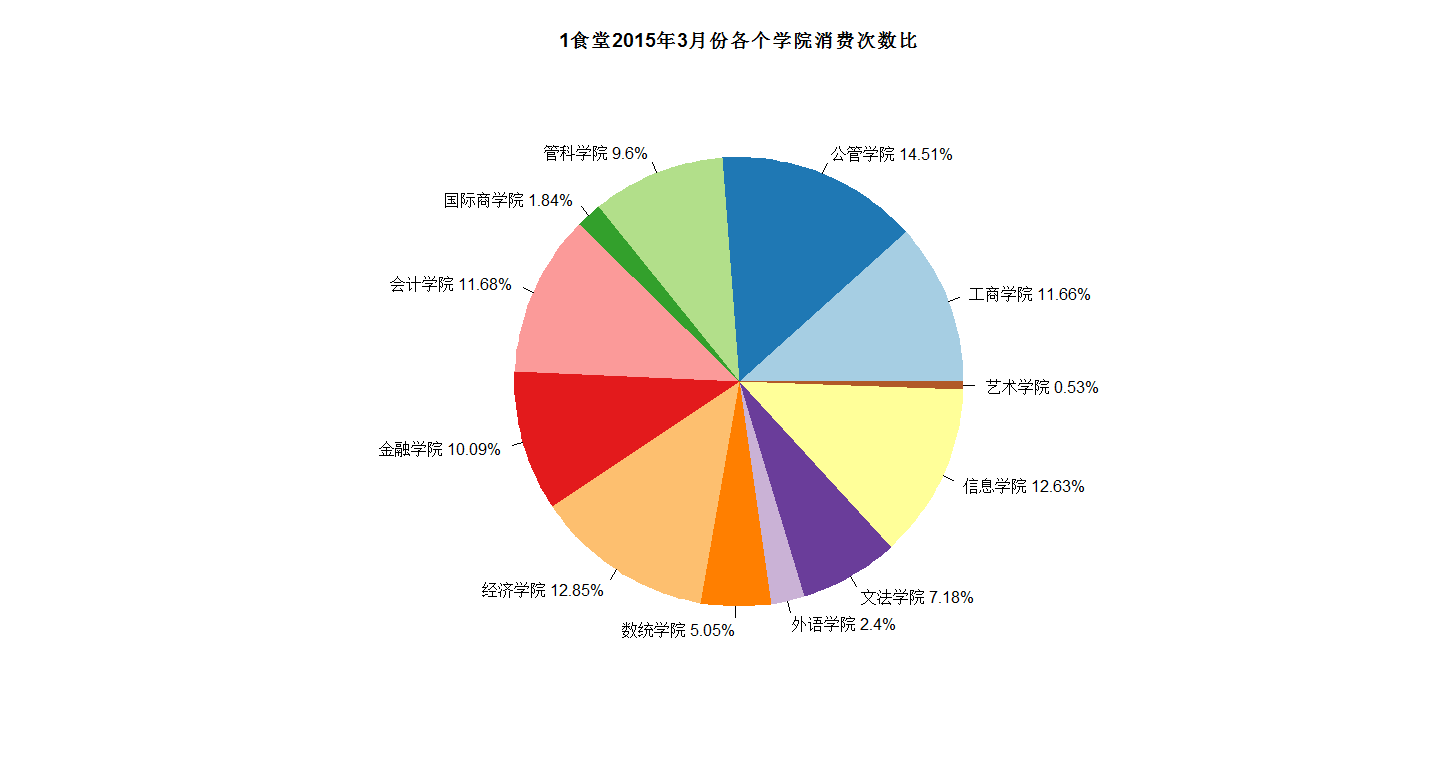
> pie(slices,labels=lbls2,col=brewer.pal(11,"Spectral"),border=FALSE,main="1食堂2015年3月份各个学院消费次数比")



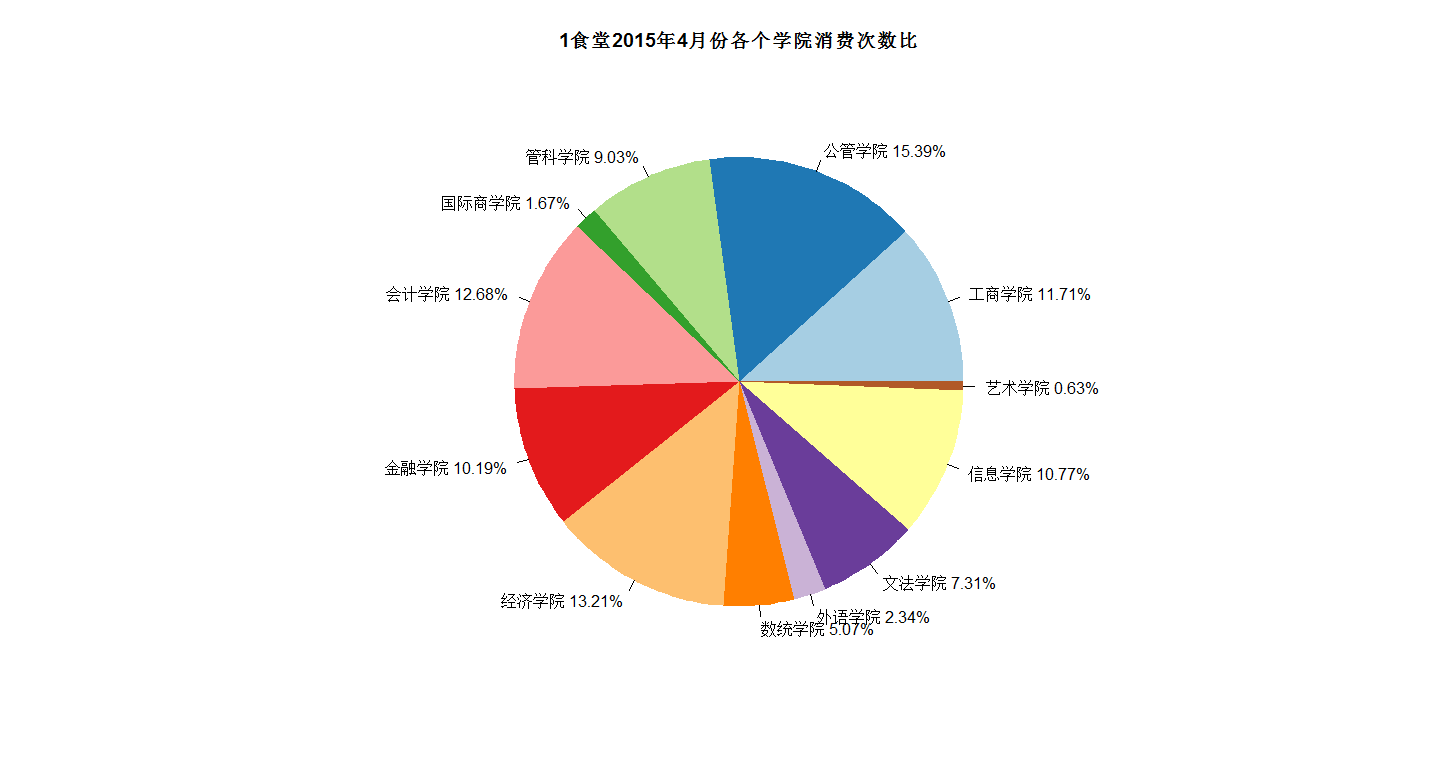
> pie(slices,labels=lbls2,col=brewer.pal(11,"Spectral"), border=FALSE,main="1食堂2015年4月份各个学院消费次数比")



> pie(slices,labels=lbls2,col=brewer.pal(12,"Paired"),border=FALSE,main="1食堂2015年3月份各个学院消费次数比")



> pie(slices,labels=lbls2,col=brewer.pal(12,"Paired"),border=FALSE,main="1食堂2015年4月份各个学院消费次数比")

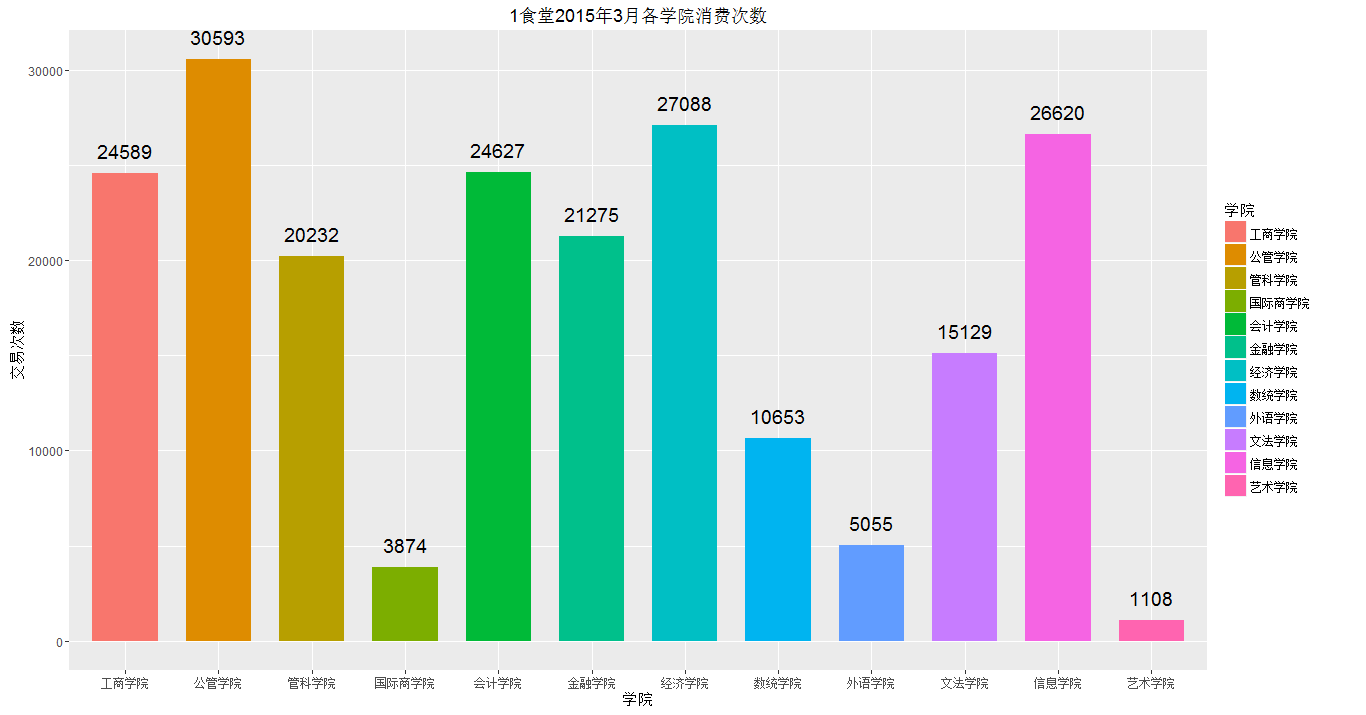


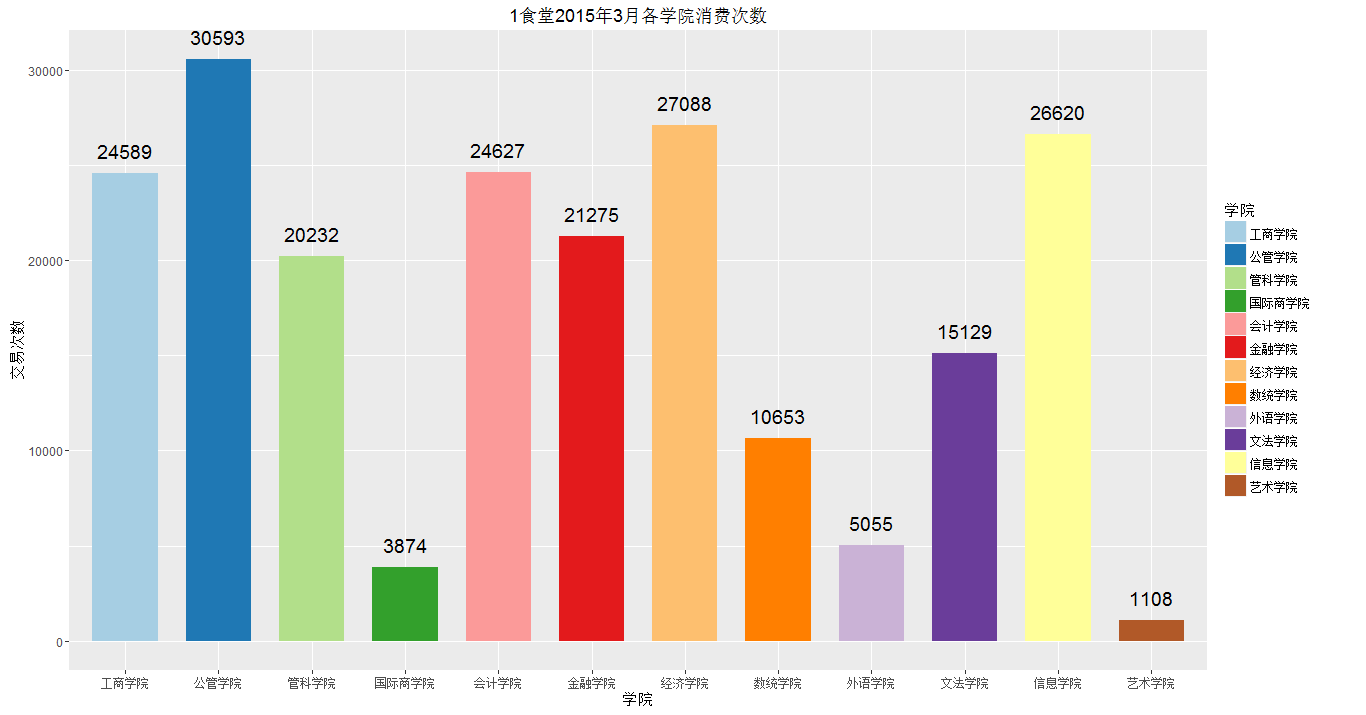
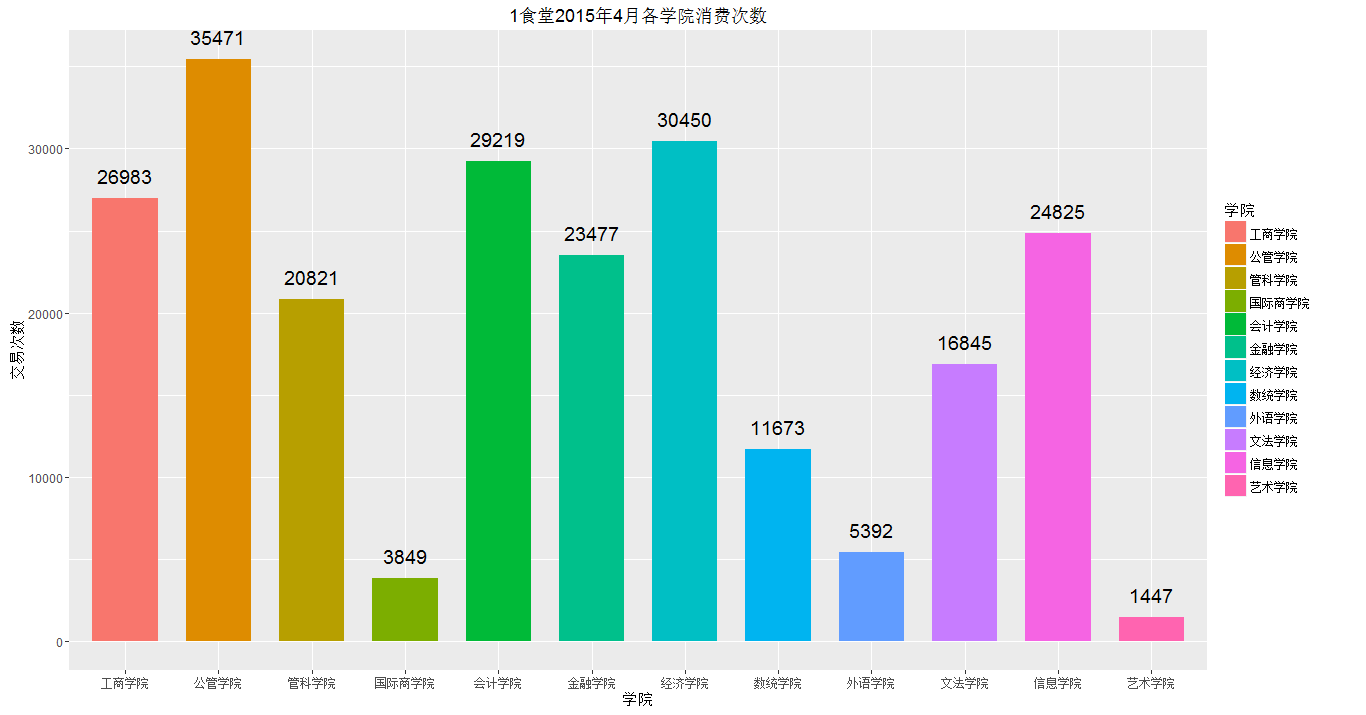
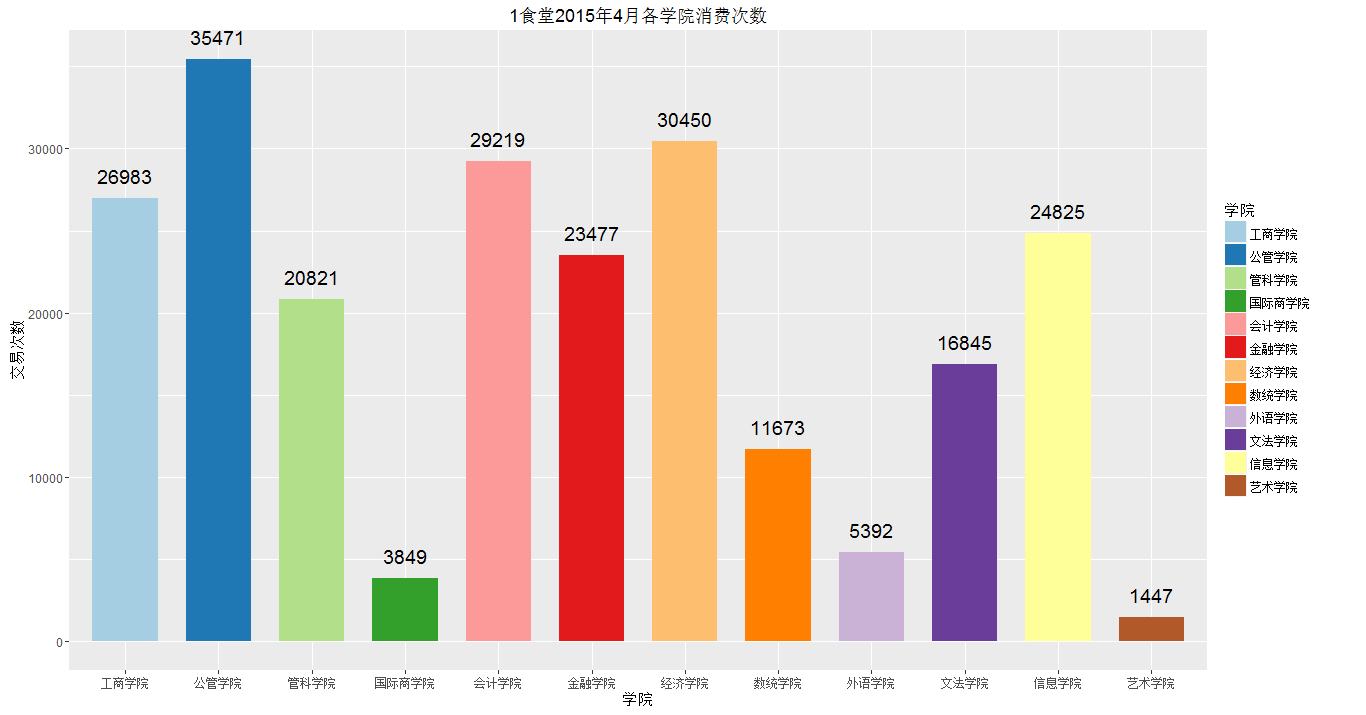
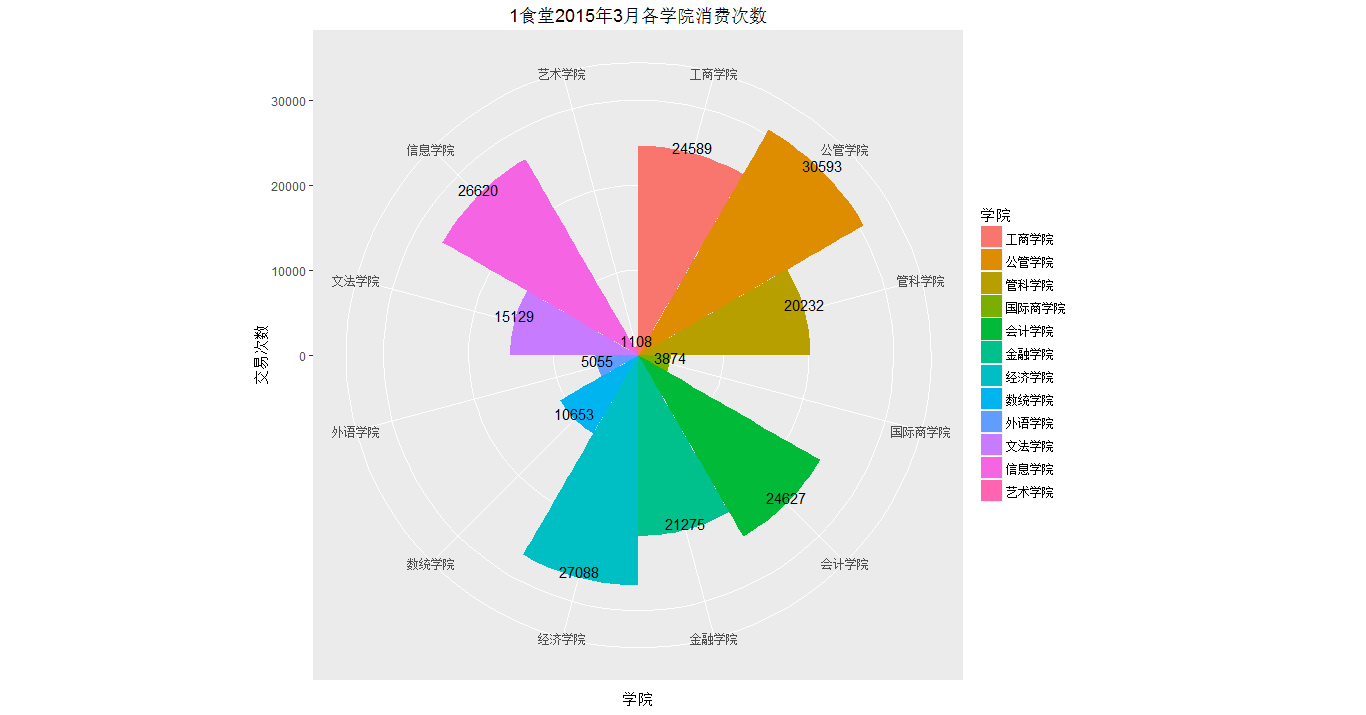
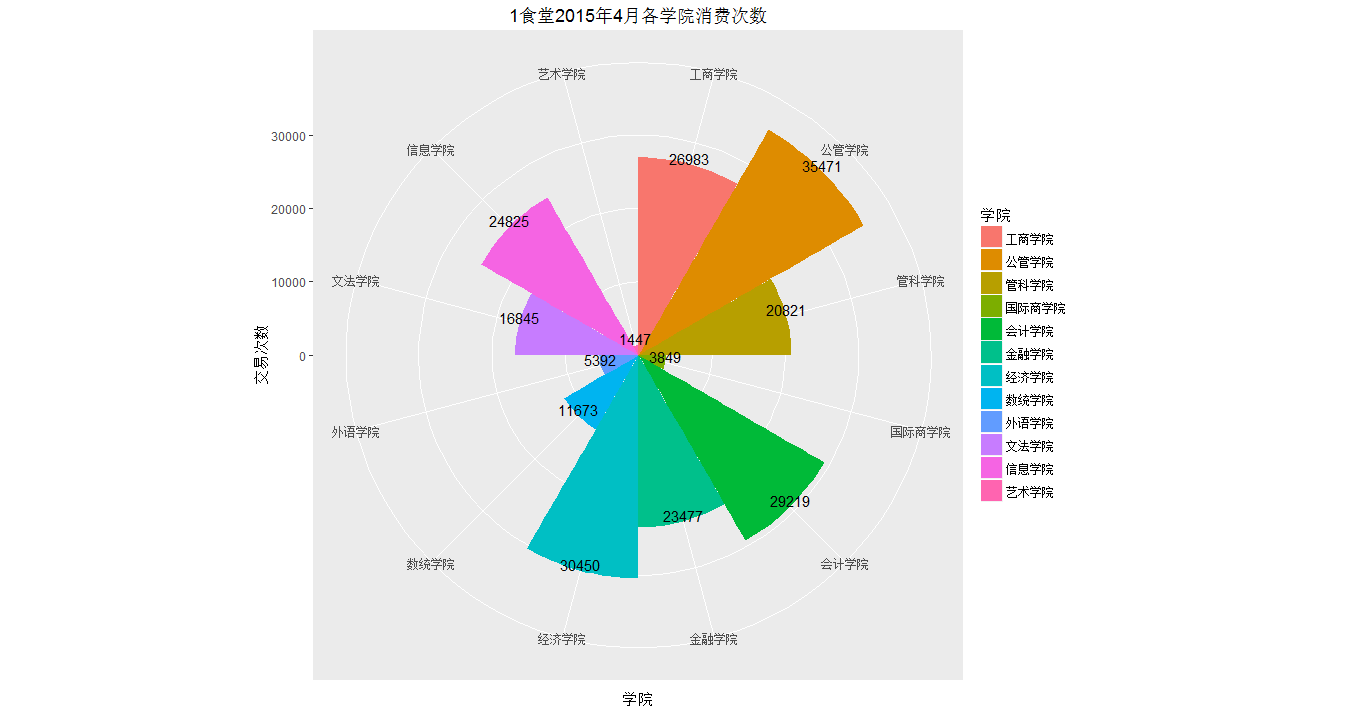
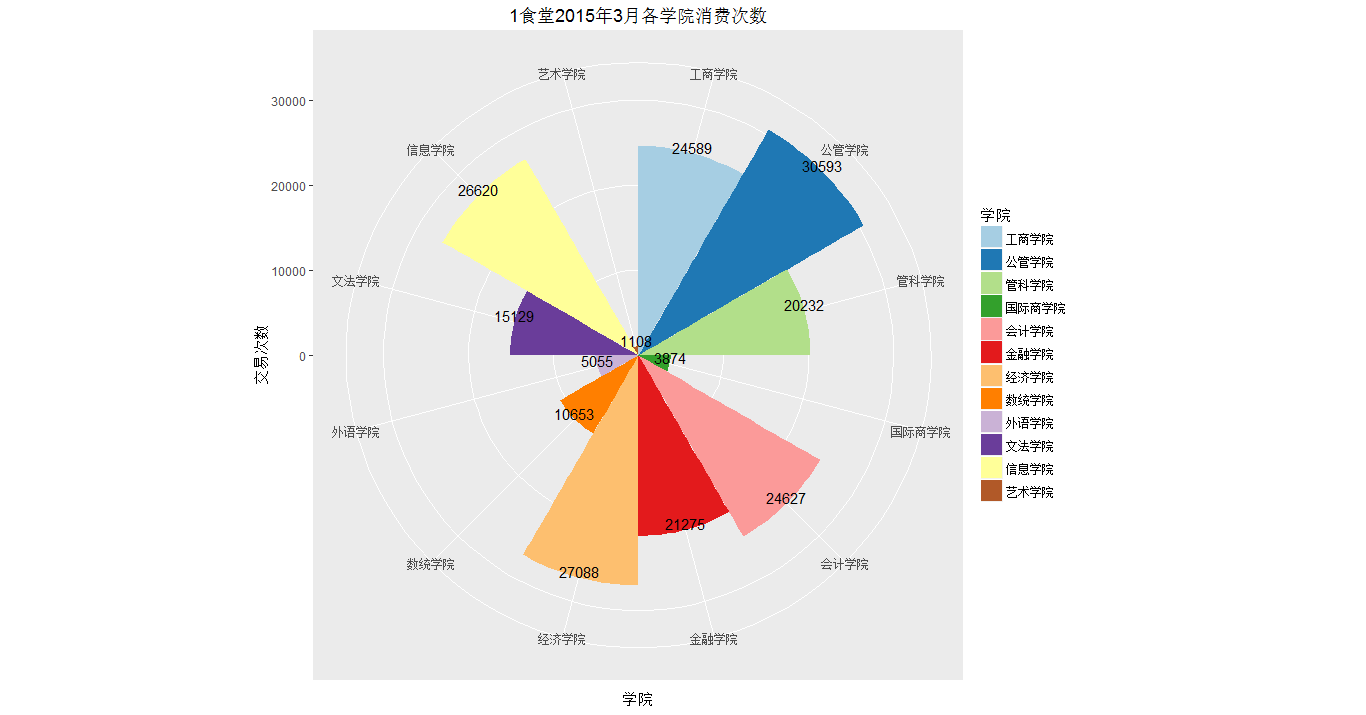
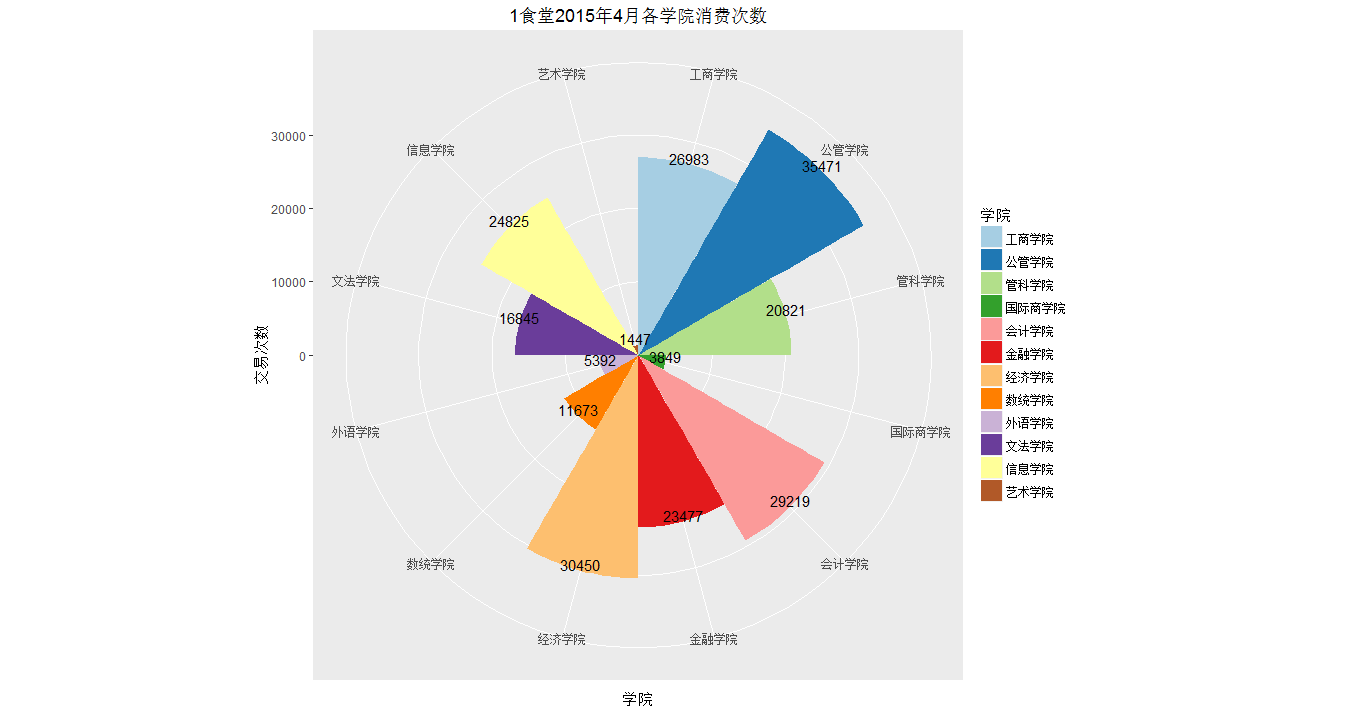
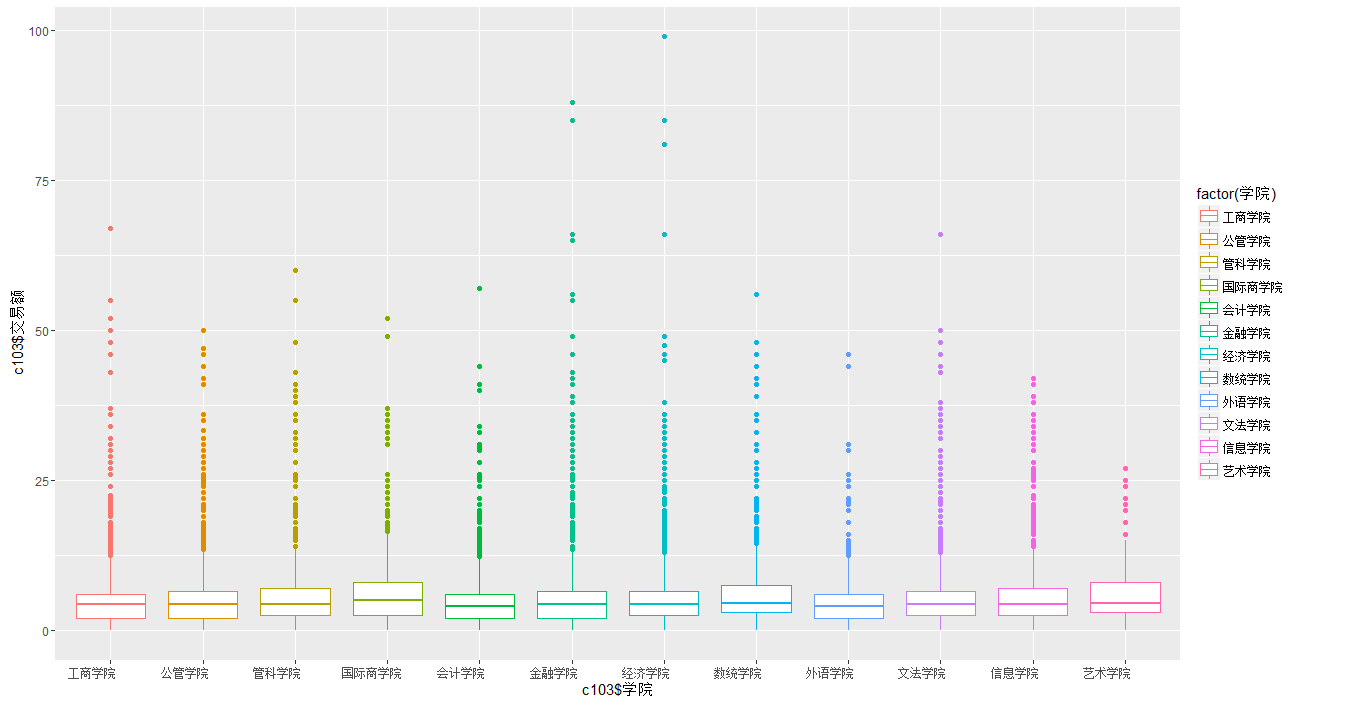
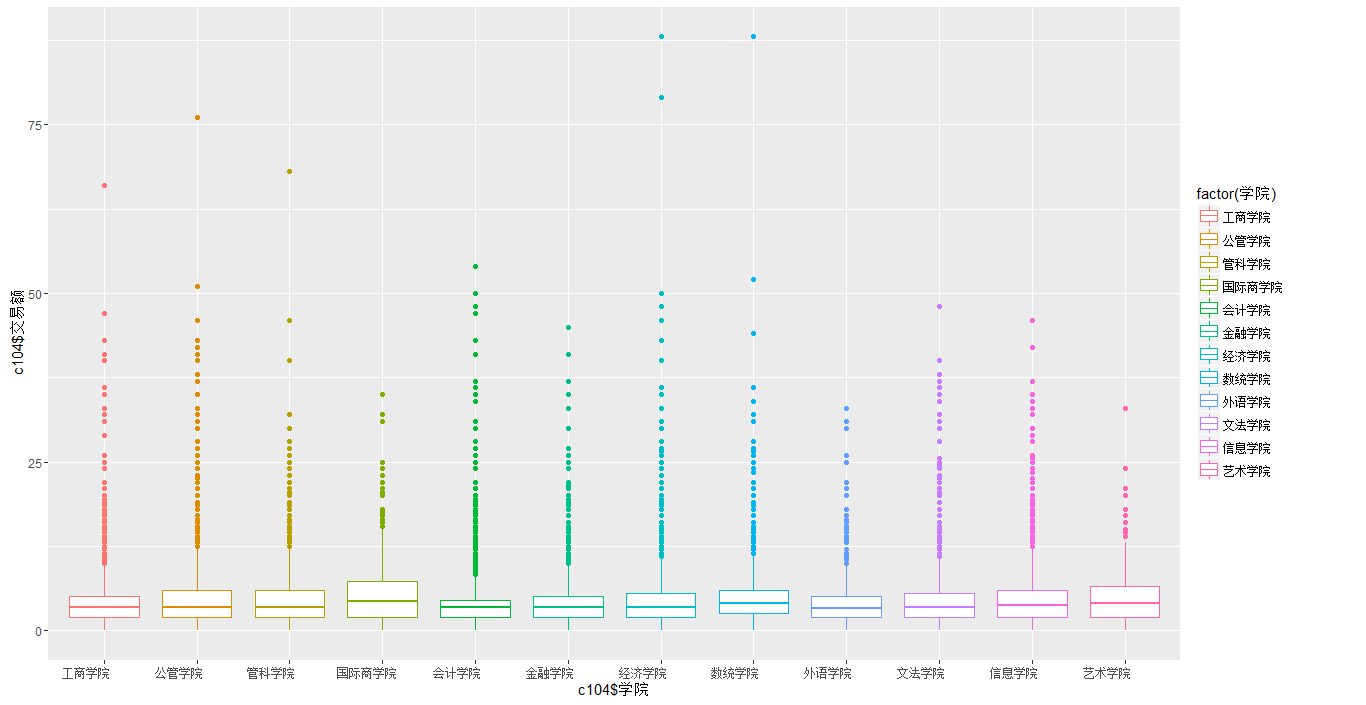
如上图可以简单的看出各个学院2015年3月份和4月份在一食堂消费的人数比，公管学院，会计学院，信息学院，经济学院占相当大的比例，法得出各个学院消费人数占各自学院总人数的比，像经济学院和会计学院本身计数就会比艺术学院和外语学院多。

* c103<-sqlQuery(db2,'select \* from c103')
* c104<-sqlQuery(db2,'select \* from c104')
* d103<-c103 %>% group\_by(学院) %>% summarise(交易次数=n())
* d104<-c104 %>% group\_by(学院) %>% summarise(交易次数=n())

1. **柱状图**

* p1<- ggplot(d103,aes(学院,交易次数,fill=学院))+geom\_bar(stat="identity",width = 0.7)+labs(x = "学院", y = "交易次数", title = "1食堂2015年3月各学院消费次数")+geom\_text(aes(label=round(交易次数)),vjust=-1,colour="black",size=5)



* p1 +scale\_fill\_brewer(palette = "Paired")
* 
* p2<-ggplot(d104,aes(学院,交易次数,fill=学院))+geom\_bar(stat="identity",width = 0.7)+labs(x = "学院", y = "交易次数", title = "1食堂2015年4月各学院消费次数")+geom\_text(aes(label=round(交易次数)),vjust=-1,colour="black",size=5)
* p2
* 
* p2+scale\_fill\_brewer(palette = "Paired")
* 
* p1<-ggplot(d103,aes(学院,交易次数,fill=学院))+geom\_bar(stat="identity",width = 1)+coord\_polar()+labs(x = "学院", y = "交易次数", title = "3食堂2015年3月各学院消费次数")+scale\_fill\_discrete()+geom\_text(aes(label=round(交易次数)),vjust=0,colour="black",size=4)
* p1
* 
* p2<-ggplot(d104,aes(学院,交易次数,fill=学院))+geom\_bar(stat="identity",width = 1)+coord\_polar()+labs(x = "学院", y = "交易次数", title = "3食堂2015年4月各学院消费次数")+scale\_fill\_discrete()+geom\_text(aes(label=round(交易次数)),vjust=0,colour="black",size=4)
* p2
* 
* p1+scale\_fill\_brewer(palette = "Paired")
* 
* p2+scale\_fill\_brewer(palette = "Paired")
* 
* 一食堂：
* > c103<-read.csv("C103.csv")
* > p<-ggplot(c103,aes(c103$学院,c103$交易额))
* > p+geom\_boxplot(aes(c103$学院,c103$交易额,colour=factor(学院)))
* 
* > c104<-read.csv("C104.csv")
* > p<-ggplot(c104,aes(c104$学院,c104$交易额))
* > p+geom\_boxplot(aes(c104$学院,c104$交易额,colour=factor(学院)))
* 

#对数据集进行初步统计

#检查数据的维度

> dim(canteen)

[1] 210843 3

> #显示数据集中每个变量的分布情况

> summary(canteen)

schoolID 交易额 机号

Min. : 1.000 Min. : 0.000 Min. : 1.00

1st Qu.: 3.000 1st Qu.: 2.300 1st Qu.:10.00

Median : 5.000 Median : 4.300 Median :19.00

Mean : 5.263 Mean : 4.787 Mean :22.97

3rd Qu.: 8.000 3rd Qu.: 6.500 3rd Qu.:33.00

Max. :12.000 Max. :99.000 Max. :59.00

> canteen1<-data.frame(x=c("工商学院","公管学院","管科学院","国际商学院","会计学院","金融学院","经济学院","数统学院","外语学院","文法学院","信息学院","艺术学院"),y=c(24589,30593,20232,3874,24627,21275))

> p=ggplot(canteen1,aes(x,y,fill=y))

> p+geom\_bar(stat="identity")+

+ geom\_abline(intercept=0,slope=0,size=1,colour='gray')+

+ geom\_text(aes(label=y),hjust=0.5,vjust=-0.5)+

+ scale\_y\_continuous(limits=c(1000,38000))+

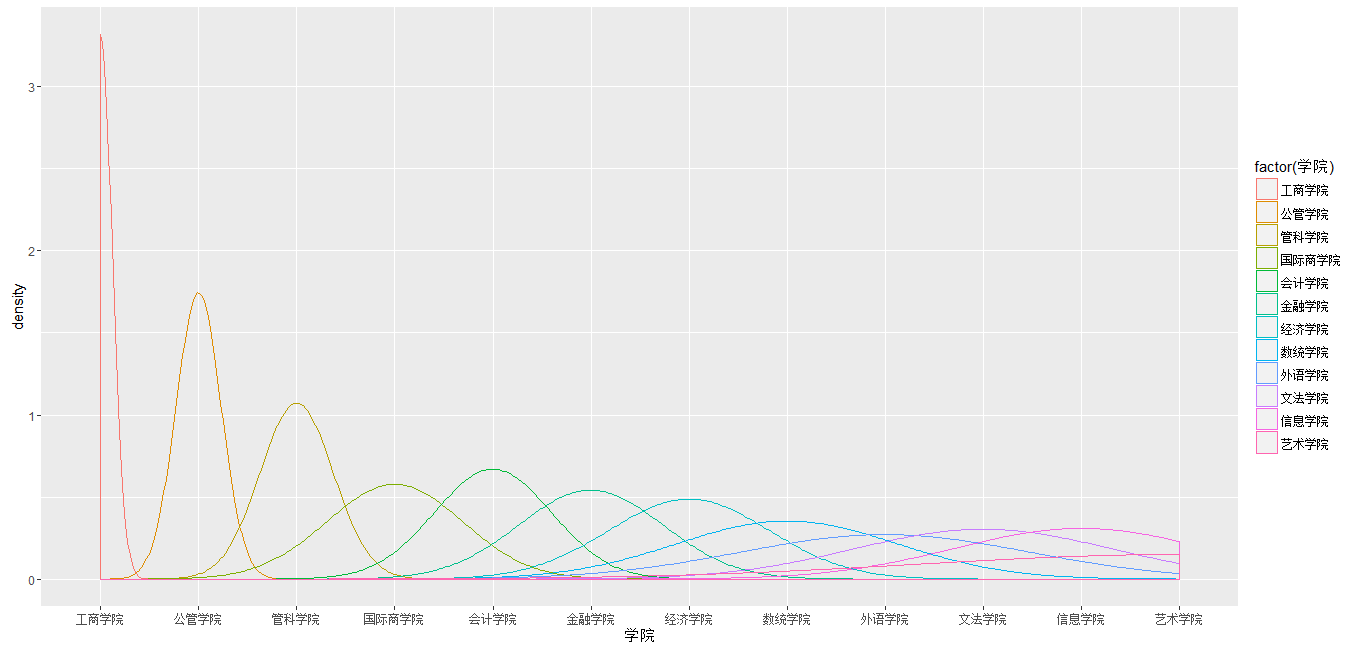
+ labs(x='年份',y='消费次数')+

+ opts(title="1食堂2015年3月各学院学生消费次数")

#折线密度图

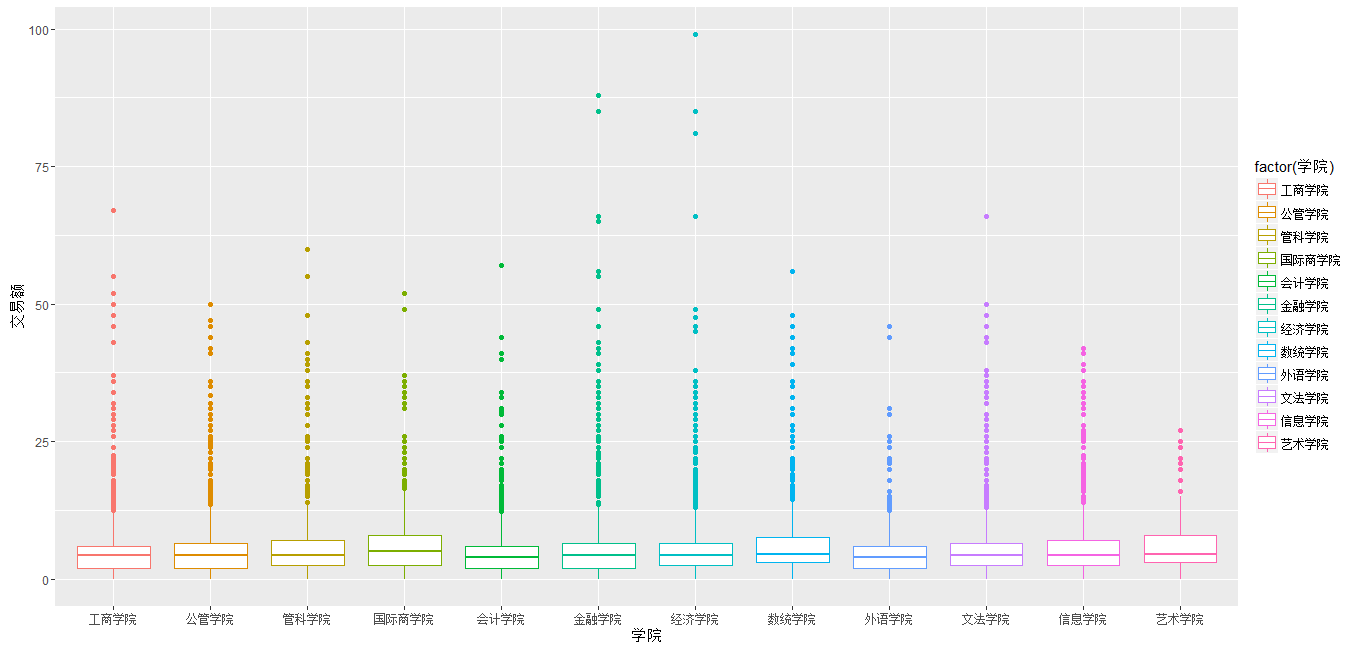
> library(ggplot2)

> qplot(学院,data=canteen,geom="density",colour=factor(学院))



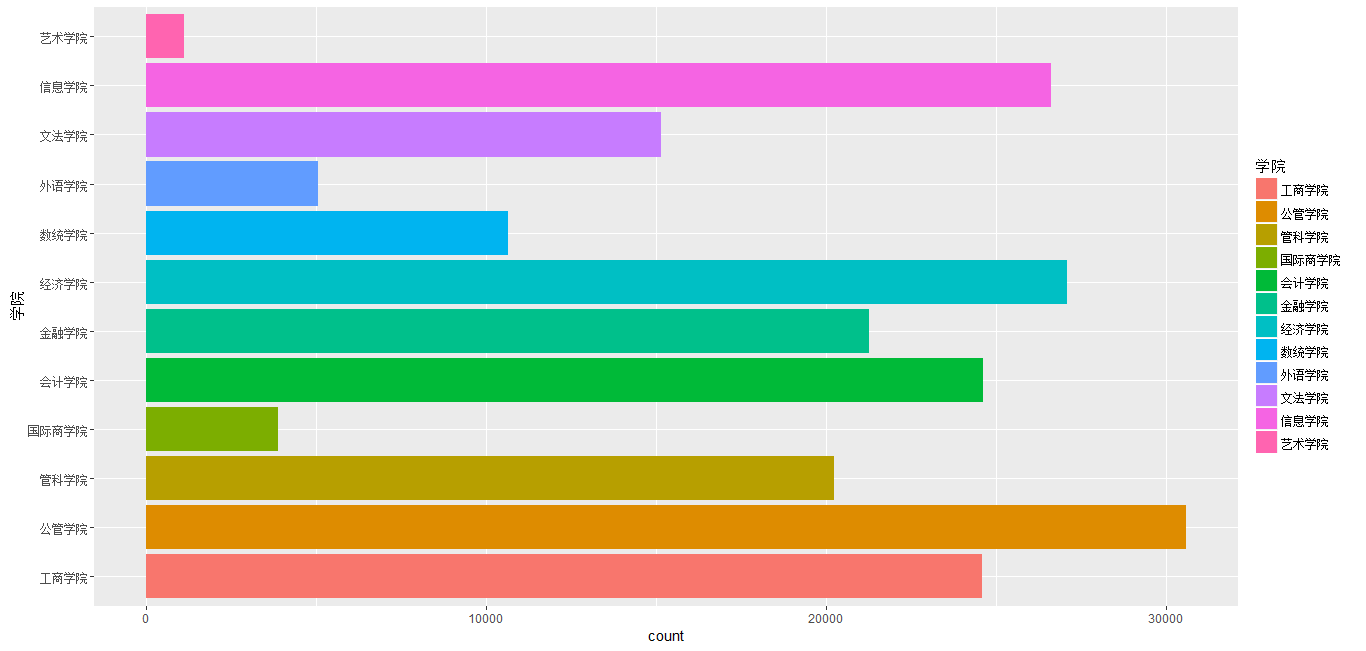
如上图可以看出从左到右，各个学院在一食堂消费的比率呈现从高到低的变化，工商学院的消费比率最高，艺术学院消费比率最低。

> p+geom\_boxplot(aes(x=学院,y=交易额,colour=factor(学院)))



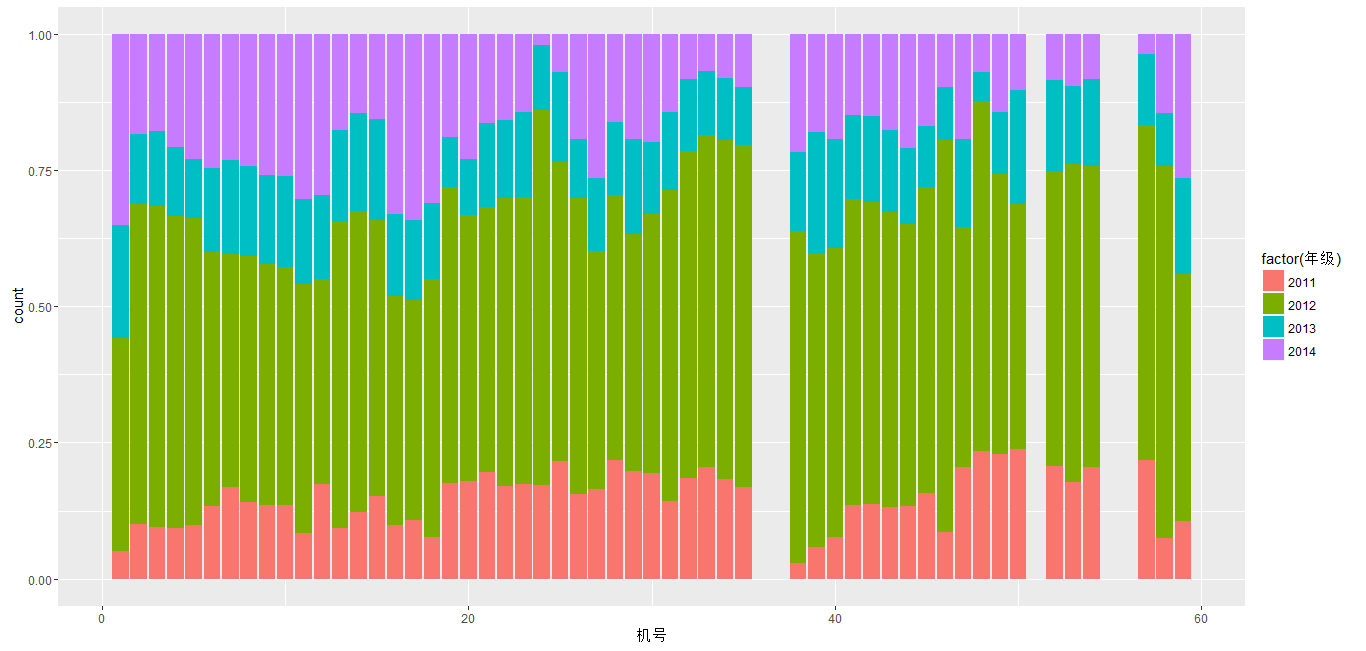
> p<-p+geom\_bar(aes(x=学院,fill=学院))+coord\_flip()

> p



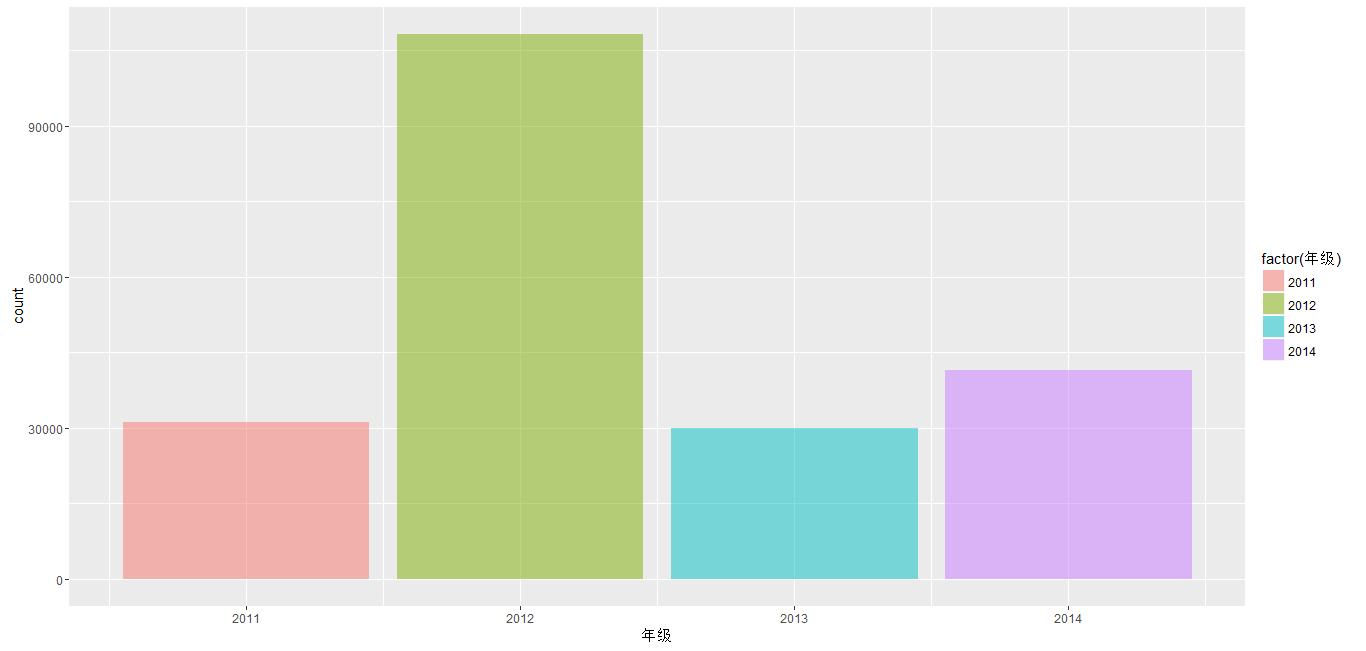
> p<-ggplot(canteen)

> p+geom\_bar(position='fill')



> p<-ggplot(canteen,aes(年级,fill=factor(年级)))

> p+geom\_bar(position='identity',alpha=0.5)



K均值聚类过程可以划分为以下5个步骤：

1. 指定聚类个数
2. 随机产生k个划分
3. 计算每个划分的中心
4. 将观测点分配到距离簇中心最近的一个簇中
5. 重复步骤2),3),4)，直到WCSS(组内平方和)基本不发生变化（或最小化）

#对交易额进行K均值聚类

金融学院1 工商学院2 公管学院3 管科学院4 会计学院5 信息学院 6

外语学院7 数统学院8 文法学院9 经济学院10 国际商学院11 艺术学院12

#对数据集canteen进行K均值聚类

K均值方法处理数据

> canteen<-read.csv('test.csv')

> fit=kmeans(canteen,5) #把数据集canteen随机分成5个簇

> fit #显示每个簇中的数据集数量

K-means clustering with 5 clusters of sizes 63160, 17056, 30200, 39068, 61359

#k-means算法产生了5个聚类

Cluster means: #每个聚类中各个列值生成的最终平均值

schoolID 交易额 机号

1 5.252961 3.610180 6.604338

2 5.331848 8.522107 54.497186

3 5.218113 4.750378 42.700828

4 5.328735 7.155673 28.179636

5 5.232941 3.469856 18.011196

Clustering vector:#每行记录所属的聚类（1代表属于第一个聚类，以此类推）

[1] 4 4 5 5 4 4 5 3 3 3 3 3 5 5 2 1 5 3 3 3 3 4 4 4 4 5 4 4 4 4 5 1 1 5 5 4 3 3 3 3 3 3 3 4 3 2 2 2 2 4 1 1 5 5 5 5 4 3 1 2 3 3 3 3 3 3 3 3 3 3 4 3 3 4 5 5 1 3 1 3

[81] 3 5 2 5 4 5 5 4 4 5 5 4 5 4 5 5 1 1 1 5 1 5 1 4 5 4 3 3 4 4 5 2 1 5 2 1 1 3 3 3 2 2 4 2 2 2 2 2 2 4 4 1 1 4 3 3 1 5 5 4 4 4 4 3 5 2 1 3 5 3 4 3 2 4 4 1 4 1 3 4

[161] 1 4 4 4 1 1 5 5 1 3 4 4 1 3 3 3 3 3 2 2 3 3 3 4 4 2 1 4 4 5 3 3 4 2 3 3 3 1 5 5 4 4 4 4 5 4 4 1 1

1 1 1 1 1 3 4 1 1 3 1 3 3 2 2 1 3 3 1 1 4 1 4 1 4 4 5 5 5 5 5

[ reached getOption("max.print") -- omitted 110844 entries ]

Within cluster sum of squares by cluster:#每个聚类内部的距离平方和

[1] 1442051.8 638281.3 470109.9 1238500.8 1378908.1

(between\_SS / total\_SS = 90.4 %)

#组间的距离平方和占了整体平方和的90.4%，也就是说各个聚类间的距离做到了最大

Available components:#运行kmeans函数返回的对象所包含的各个组成部分

[1] "cluster" "centers" "totss" "withinss"

[5] "tot.withinss" "betweenss" "size" "iter"

[9] "ifault"

#”cluster”是一个整数向量，用于表示记录所属的聚类

#“centers”是一个矩阵，表示每聚类中各个变量的中心点

#“totss”表示所生成聚类的总体距离平方和

#“withinss”表示各个聚类组内的距离平方和

#“tot.withinss”表示聚类组内的距离平方和总量

#“betweenss”表示聚类组间的聚类平方和总量

#“size”表示每个聚类组中成员的数量

#衔接R与C++

> library(Rcpp)

#加载颜色程序包

> library(RColorBrewer)

#加载wordcloud包

> library(wordcloud)

> data<-read.csv('1食堂机号.csv')

#使用Set1主题模板

> rc=brewer.pal(9,"Set1")

> wordcloud(data$NO,data$frequency,scale=c(5,0.5),min.freq=-Inf,max.words=Inf,colors=rc)

#函数介绍

wordcloud(words,freq,scale=c(4,0.5),min.freq=3,max.words=Inf,random.order=TRUE,random.color=FALSE,  
rot.per=.1,colors="black",ordered.colors=FALSE,use.r.layout=FALSE,...)

参数：  
words：词  
freq：词频  
scale：c(最大字号，最小字号)  
min.freq:最小限制频数，低于频数的词不被显示  
max.words：显示词的数量  
random.order：T：乱序，F:按频数降序排列  
random.color：T:任意选择颜色，F:基于频数选择颜色  
rot.per：T：水平，F:旋转90度

