Group_descrip_lin2.R

daitu

Fri Jun 24 18:27:36 2016

```
## 2016年暑期课程设计
## 问题: Grupo Bimbo Inventory Demand
##   宾堡集团的库存需求
## 最大限度地提高销售和最大限度地减少烘焙食品的退回
## Daitu
## start:2016.06.21
## 参考借鉴kaggle上的公开程序


setwd("/Users/Daitu/数据分析/kaggle/Grupo Bimbo")
getwd()
```

```
## [1] "/Users/daitu/数据分析/kaggle/Grupo Bimbo"
```

```
## 加载包

library(data.table)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(treemap)
```

```
## Warning: package 'treemap' was built under R version 3.2.4
```

```
## 读取数据####
## 1:读取训练集
system.time({
  traindata <- fread("train.csv",sep=",",header = TRUE)
})
```

```
##
Read 0.0% of 74180464 rows
Read 4.4% of 74180464 rows
Read 8.8% of 74180464 rows
Read 13.3% of 74180464 rows
Read 17.7% of 74180464 rows
Read 22.1% of 74180464 rows
Read 26.6% of 74180464 rows
Read 31.0% of 74180464 rows
Read 35.4% of 74180464 rows
Read 39.8% of 74180464 rows
Read 44.3% of 74180464 rows
Read 48.6% of 74180464 rows
Read 52.9% of 74180464 rows
Read 57.3% of 74180464 rows
Read 61.6% of 74180464 rows
Read 65.9% of 74180464 rows
Read 70.3% of 74180464 rows
Read 74.6% of 74180464 rows
Read 78.8% of 74180464 rows
Read 83.1% of 74180464 rows
Read 87.4% of 74180464 rows
Read 91.5% of 74180464 rows
Read 95.8% of 74180464 rows
Read 74180464 rows and 11 (of 11) columns from 2.980 GB file in 00:00:30
```

```
##    user  system elapsed
## 26.756   2.868  32.726
```

```
head(traindata)
```

```
##    Semana Agencia_ID Canal_ID Ruta_SAK Cliente_ID Producto_ID
## 1:      3       1110        7     3301      15766        1212
## 2:      3       1110        7     3301      15766        1216
## 3:      3       1110        7     3301      15766        1238
## 4:      3       1110        7     3301      15766        1240
## 5:      3       1110        7     3301      15766        1242
## 6:      3       1110        7     3301      15766        1250
##    Venta_uni_hoy Venta_hoy Dev_uni_proxima Dev_proxima Demanda_uni_equil
## 1:             3     25.14               0           0                 3
## 2:             4     33.52               0           0                 4
## 3:             4     39.32               0           0                 4
## 4:             4     33.52               0           0                 4
## 5:             3     22.92               0           0                 3
## 6:             5     38.20               0           0                 5
```

2:读取客户名单数据
```
cliente_tabla <- fread("cliente_tabla.csv",sep=",",header = TRUE)
head(cliente_tabla)
```

```
##    Cliente_ID                     NombreCliente
## 1:         0                        SIN NOMBRE
## 2:         1                   OXXO XINANTECATL
## 3:         2                        SIN NOMBRE
## 4:         3                         EL MORENO
## 5:         4 SDN SER  DE ALIM  CUERPO SA CIA  DE INT
## 6:         4   SDN SER DE ALIM CUERPO SA CIA  DE INT
```

3:读取产品名单数据
```
producto_tabla <- fread("producto_tabla.csv",sep=",",header = TRUE)
head(producto_tabla)
```

```
##    Producto_ID                  NombreProducto
## 1:          0                 NO IDENTIFICADO 0
## 2:          9            Capuccino Moka 750g NES 9
## 3:         41 Bimbollos Ext sAjonjoli 6p 480g BIM 41
## 4:         53          Burritos Sincro 170g CU LON 53
## 5:         72      Div Tira Mini Doradita 4p 45g TR 72
## 6:         73       Pan Multigrano Linaza 540g BIM 73
```

4:读取城镇和国家（州）数据
```
town_state <- fread("town_state.csv",sep=",",header = TRUE)
head(town_state)
```

```
##    Agencia_ID              Town            State
## 1:       1110    2008 AG. LAGO FILT       MÉXICO, D.F.
## 2:       1111 2002 AG. AZCAPOTZALCO       MÉXICO, D.F.
## 3:       1112    2004 AG. CUAUTITLAN ESTADO DE MÉXICO
## 4:       1113    2008 AG. LAGO FILT       MÉXICO, D.F.
## 5:       1114  2029 AG.IZTAPALAPA 2       MÉXICO, D.F.
## 6:       1116  2011 AG. SAN ANTONIO       MÉXICO, D.F.
```

数据的描述统计####

1:分析数据的周数：Semana
```
Semana <- data.frame(table(traindata$Semana))
colnames(Semana) <- c("Semana","Freq")
# 3         4         5         6         7         8         9
# 11165207 11009593 10615397 10191837 10382849 10406868 10408713
## 条形图
ggplot(data = Semana,aes(Semana,Freq)) +
  geom_bar(stat = "identity", width = 0.6,fill = "lightblue",colour = "black") +
  theme_grey(base_family = "STKaiti") +
  scale_y_continuous() +
  labs(x="周数",y="数据条数",title = "每周数据的记录数")
```

每周数据的记录数
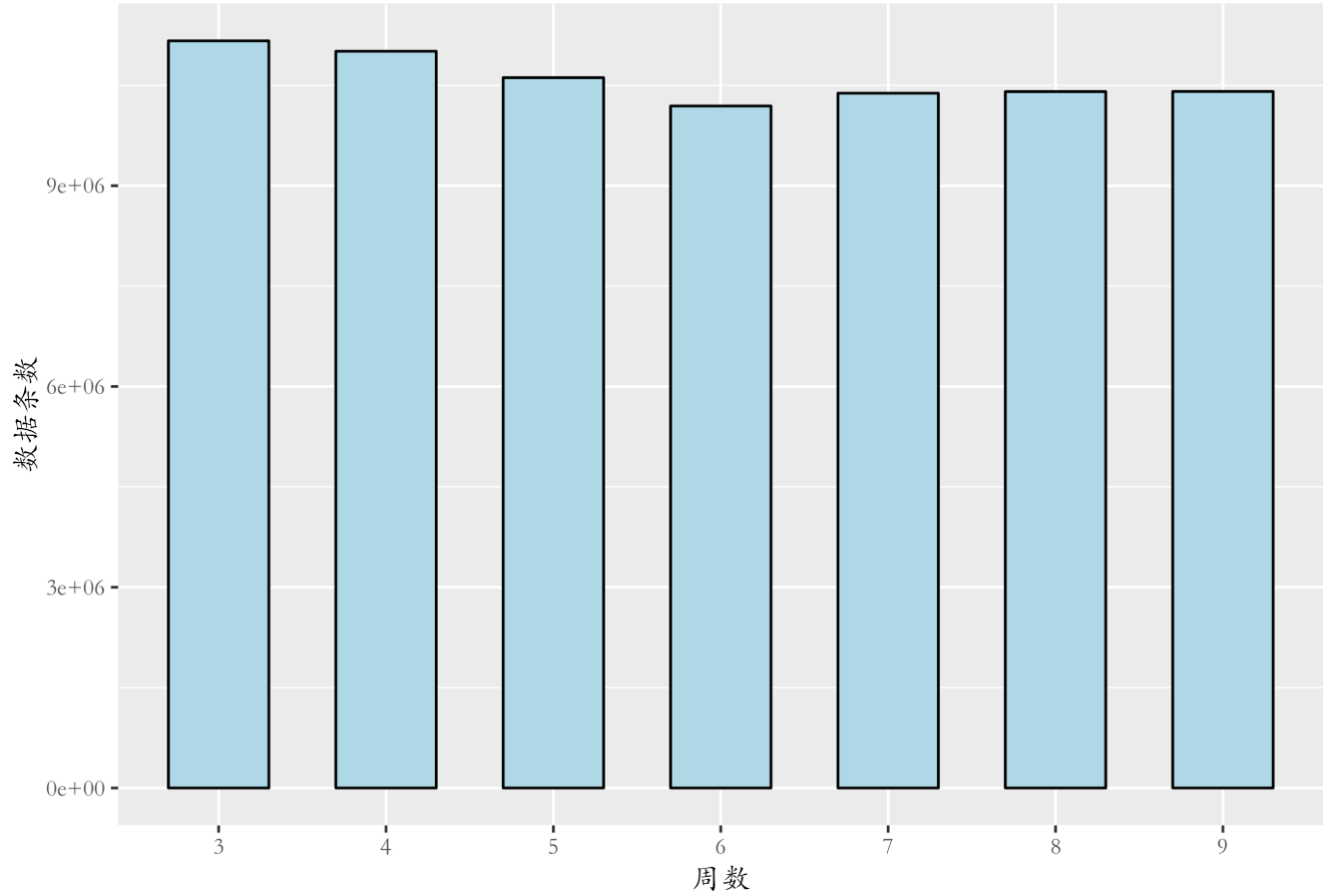

```
Semana2 <- traindata %>%
  group_by(Semana) %>%
  summarise(Units = sum(Venta_uni_hoy),
            Return_Units = sum(Dev_uni_proxima),
            NetU = sum(Demanda_uni_equil)) %>%
  mutate(Retern_Rate = Return_Units / (Return_Units + Units))
Semana2
```

```
## Source: local data table [7 x 5]
##
##   Semana     Units Return_Units     NetU Retern_Rate
##    (int)     (int)        (int)    (int)       (dbl)
## 1      3 78519996      1286055 77664309  0.01611476
## 2      4 80509571      1322852 79618866  0.01616538
## 3      5 78450353      1220967 77610637  0.01532505
## 4      6 74753832      1321510 73851129  0.01737107
## 5      7 77548038      1417450 76597014  0.01795025
## 6      8 76497828      1452369 75525105  0.01863201
## 7      9 75991700      1641371 75054450  0.02114268
```

```
dim(Semana2)   # 几个星期的数据
```

```
## [1] 7 5
```

```
## 可视化每周的情况
ggplot(data = Semana2,aes(Semana,NetU,fill = Retern_Rate)) +
  geom_bar(stat = "identity",color = "black",width = 0.7) +
  theme_bw(base_family = "STKaiti") +
  scale_y_continuous(labels=function(x)paste(x/100000, "m"))+
  scale_x_continuous(labels = function(x)paste("第",x,"星期")) +
  scale_fill_gradient(name="退回率", low="blue", high="red")+
  labs(x = "星期",y = "销量",title = "每星期销量")
```


每星期销量
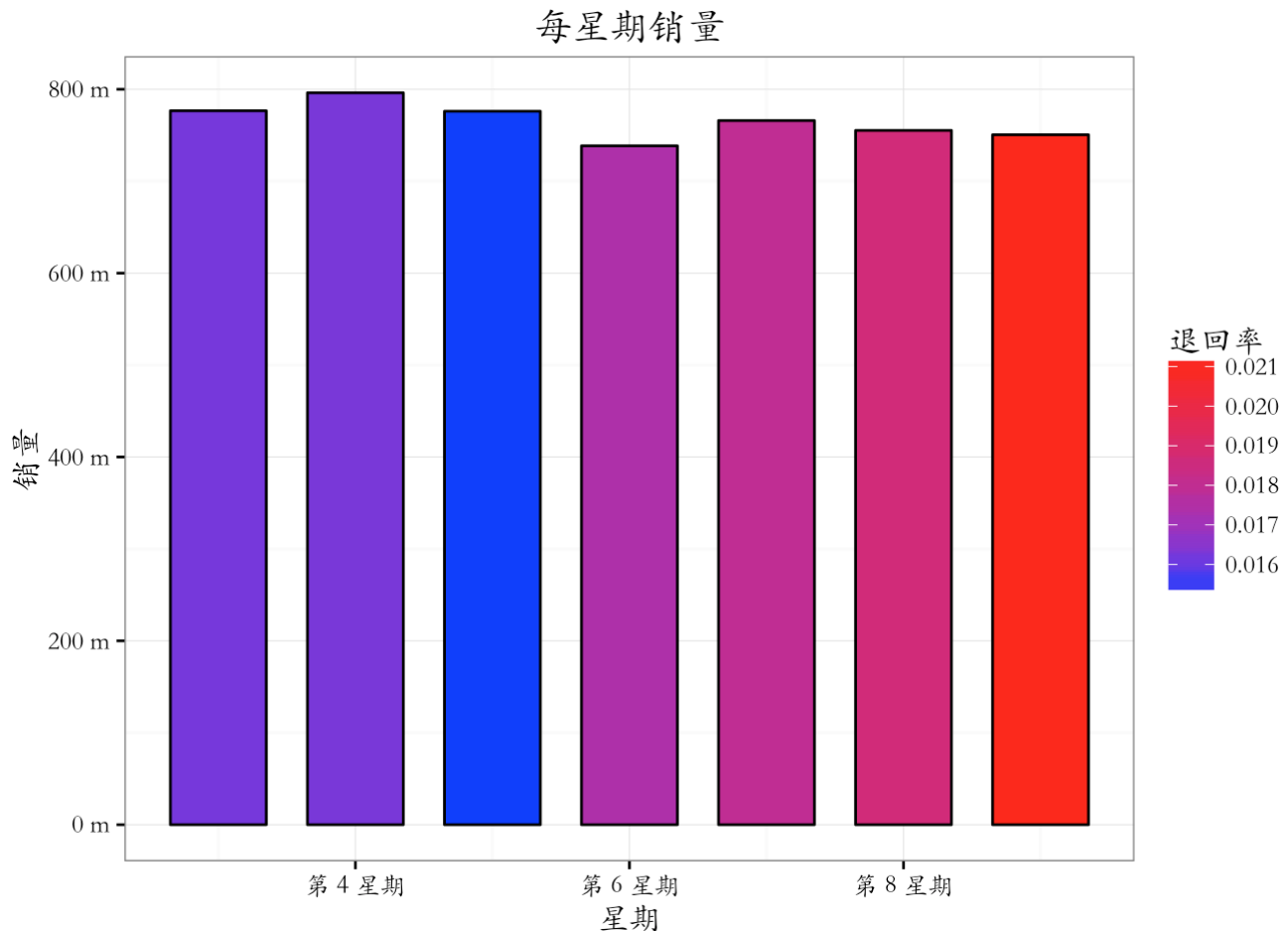
```
## 销售站的数据分析#####
## 2：销售站Agencia 和 州State
agencias <- traindata %>%
  group_by(Agencia_ID) %>%          # 数据按照销售战进行分组统计
  summarise(Units = sum(Venta_uni_hoy), # 总结多个值为一个值，units:本销售站的销量和
            Pesos = sum(Venta_hoy),   # 本周的销售量（比索）之和
            Return_Units = sum(Dev_uni_proxima),  # 下星期的返回量之和
            Return_Pesos = sum(Dev_proxima),  # 下星期的返回量（比索）之和
            Net = sum(Demanda_uni_equil)) %>%   # 调整后的需求和
  mutate(Net_Pesos = Pesos - Return_Pesos,  # mutate:添加新的变量
         Return_Rate = Return_Units / (Units+Return_Units)) %>%  # 添加变量退货比率
  arrange(desc(Units)) %>%       # 将数据按照变量Units的降序排列
  inner_join(town_state, by="Agencia_ID")   # 按照变量Agencia_ID，连接两个表，return all
 rows from x
head(agencias)
```

```
##    Agencia_ID    Units     Pesos Return_Units Return_Pesos       Net Net_Pesos
## 1       1110   877675   9274674        39900     214072.8    874523   9060601
## 2       1111  2720400  24070592        25231     264672.4   2701427  23805919
## 3       1112  1959534  16591688        23924     231897.4   1942114  16359791
## 4       1113  1442999  12094484        11865     117754.4   1434414  11976730
## 5       1114  3498170  62420320       150779    2480404.7   3363796  59939915
## 6       1116  3120201  27454358        37022     377100.6   3093985  27077257
##    Return_Rate                 Town            State
## 1 0.043484184    2008 AG. LAGO FILT        MÉXICO, D.F.
## 2 0.009189509 2002 AG. AZCAPOTZALCO        MÉXICO, D.F.
## 3 0.012061763    2004 AG. CUAUTITLAN ESTADO DE MÉXICO
## 4 0.008155401    2008 AG. LAGO FILT        MÉXICO, D.F.
## 5 0.041321213  2029 AG.IZTAPALAPA 2        MÉXICO, D.F.
## 6 0.011726128  2011 AG. SAN ANTONIO        MÉXICO, D.F.
```

```
dim(agencias)    # 多少个销售站
```

```
## [1] 552  10
```

```
## 可视化x：每天销量，y：销售站的数量
ggplot(agencias, aes(x=Units/7))+
  geom_histogram(fill="red", color="gray", binwidth=10000)+ #条形图的宽度为10000
  theme_bw(base_family = "STKaiti") +
  scale_x_continuous(labels=function(x)paste(x/1000, "k"))+
  labs(x = "每天的平均销量",y = "销售站数量",title = "销售站的销量")
```


销售站的销量
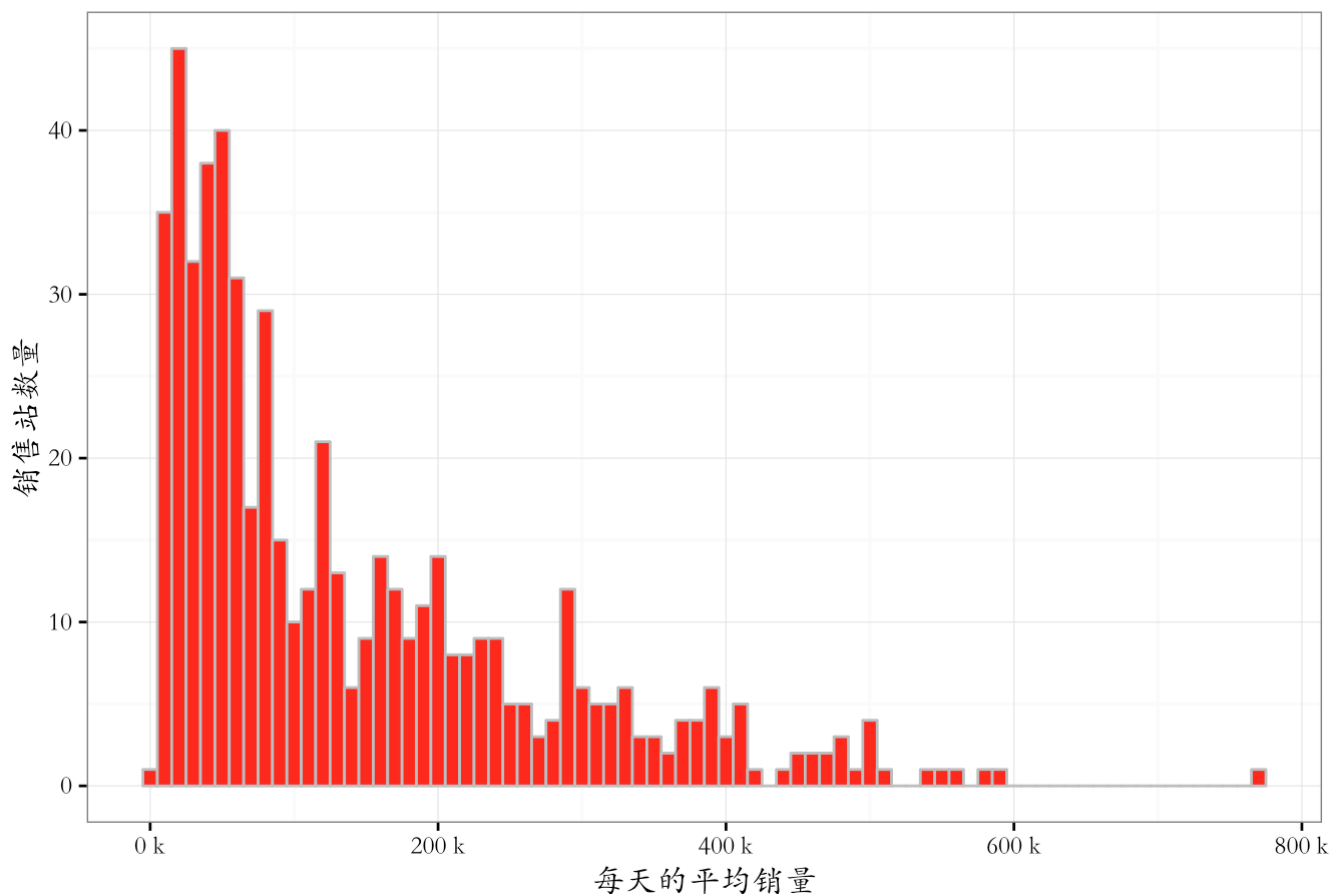
```
## 前100的销售站的销量树图
treemap(agencias[1:100, ],
        index=c("Agencia_ID"), vSize="Units", vColor="Return_Rate",
        type="value", title.legend="Units return rate", title="Top 100 agencias")
```

Top 100 agencias
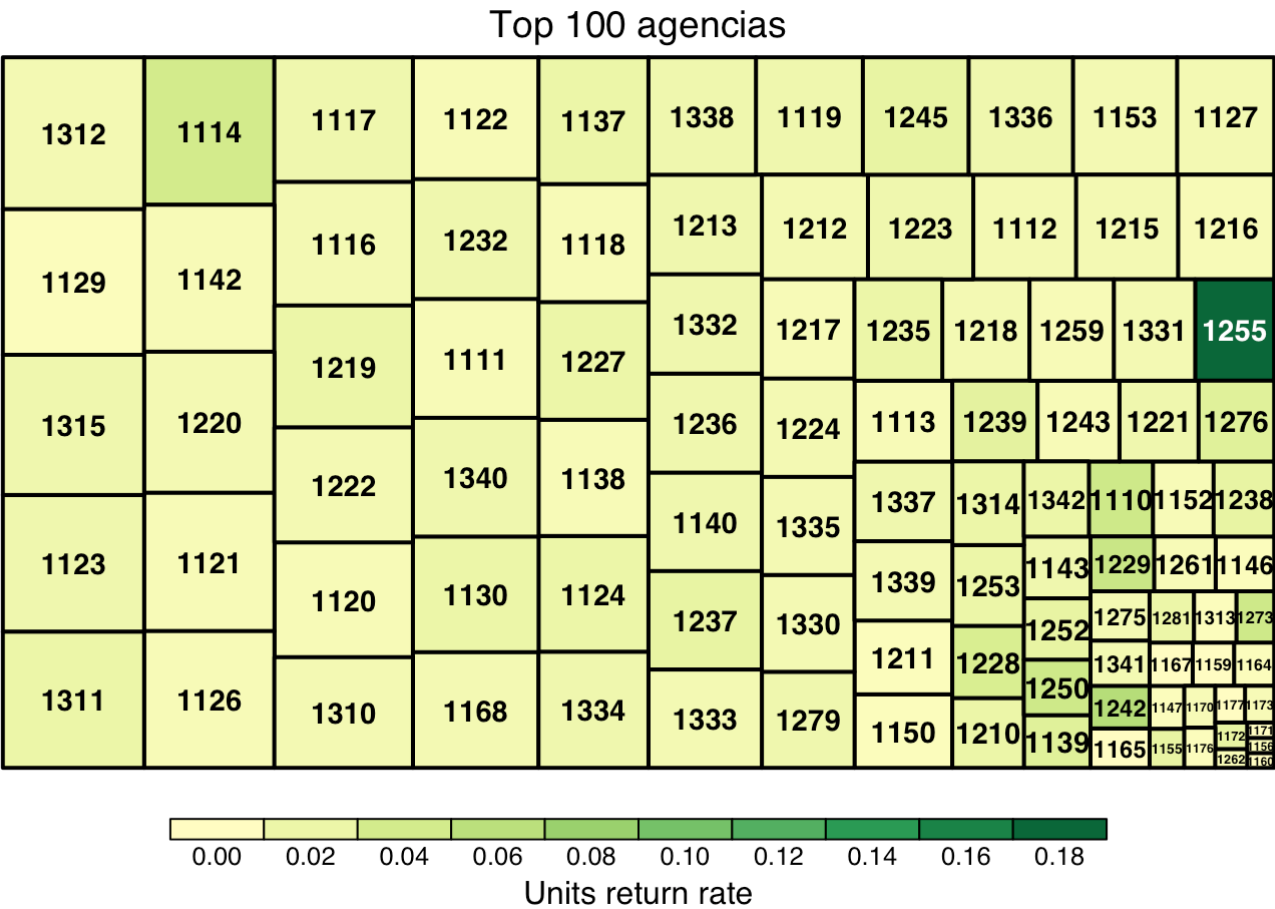
Units return rate

```
## 所有销售站的销量树图
treemap(agencias,
        index=c("Agencia_ID"), vSize="Units", vColor="Return_Rate",
        type="value", title.legend="Units return rate", title="All agencias")
```

All agencias

Units return rate

0.00   0.05   0.10   0.15   0.20
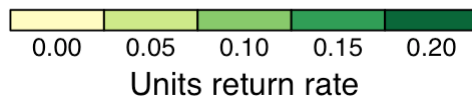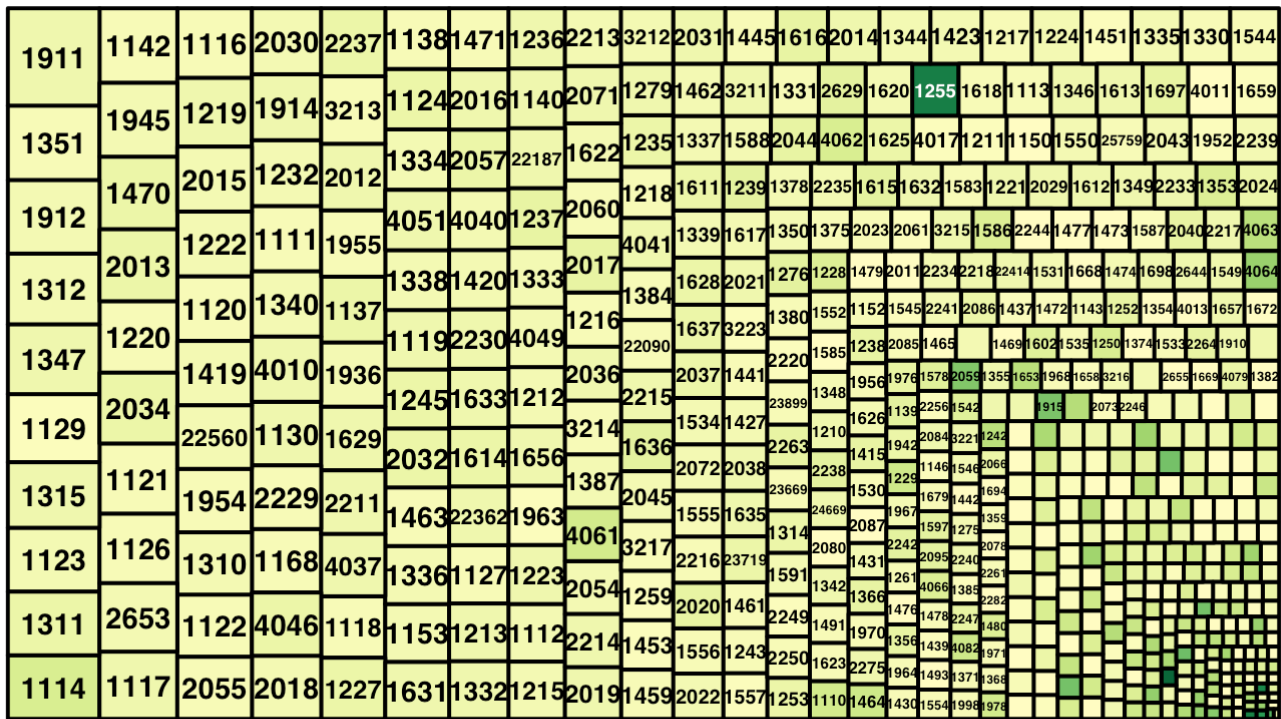
```
## 销售站的历史数据分析
agencias_history <- traindata %>%
  group_by(Agencia_ID, Semana) %>%   # 数据按照销售站、星期数，进行分组统计
  summarise(Units = sum(Venta_uni_hoy),   # 总结多个值为一个值，units: 销量和
          Pesos = sum(Venta_hoy),
          Return_Units = sum(Dev_uni_proxima),
          Return_Pesos = sum(Dev_proxima),
          Net = sum(Demanda_uni_equil)) %>%
  mutate(Net_Pesos = Pesos - Return_Pesos,
        Avg_Pesos = Pesos / Units,
        Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(Agencia_ID, Semana) %>% # 将数据按照变量 销售站、星期数的降序排列
  inner_join(town_state, by="Agencia_ID")
head(agencias_history)
```

```
##    Agencia_ID Semana  Units   Pesos Return_Units Return_Pesos     Net
## 1       1110      3 120285 1296739         3577     29712.03 119951
## 2       1110      4 135788 1385112         5260     29932.82 135327
## 3       1110      5 127420 1345680         3342     26249.25 127077
## 4       1110      6 115255 1239051         5721     26595.87 114865
## 5       1110      7 122955 1297072         6950     34845.26 122513
## 6       1110      8 127277 1345696         8030     35157.98 126735
##   Net_Pesos Avg_Pesos Return_Rate                  Town        State
## 1   1267027  10.78055  0.02887891 2008 AG. LAGO FILT MÉXICO, D.F.
## 2   1355179  10.20055  0.03729227 2008 AG. LAGO FILT MÉXICO, D.F.
## 3   1319431  10.56098  0.02555788 2008 AG. LAGO FILT MÉXICO, D.F.
## 4   1212456  10.75052  0.04729037 2008 AG. LAGO FILT MÉXICO, D.F.
## 5   1262227  10.54916  0.05350064 2008 AG. LAGO FILT MÉXICO, D.F.
## 6   1310538  10.57297  0.05934652 2008 AG. LAGO FILT MÉXICO, D.F.
```

```
dim(agencias_history)
```

```
## [1] 3863   12
```

```
## 取出销售量前30的销售站ID
top30agencias <- agencias$Agencia_ID[1:30]

## 销量前30的销售站每周的销量和退货率图
ggplot(agencias_history %>% filter(Agencia_ID %in% top30agencias))+
  geom_bar(aes(x=Semana, y=Units, fill=Return_Rate), stat="identity", color="black")+
  theme_bw(base_family = "STKaiti") +
  facet_wrap(~Agencia_ID)+    # 按照销售站划分成子图
  scale_y_continuous(labels=function(x)paste(x/1000, "k"))+
  scale_fill_gradient(name="退回率", low="white", high="red")+
  ggtitle("销量前30的销售站") + ylab("销量") +xlab("星期")
```

销量前30的销售站
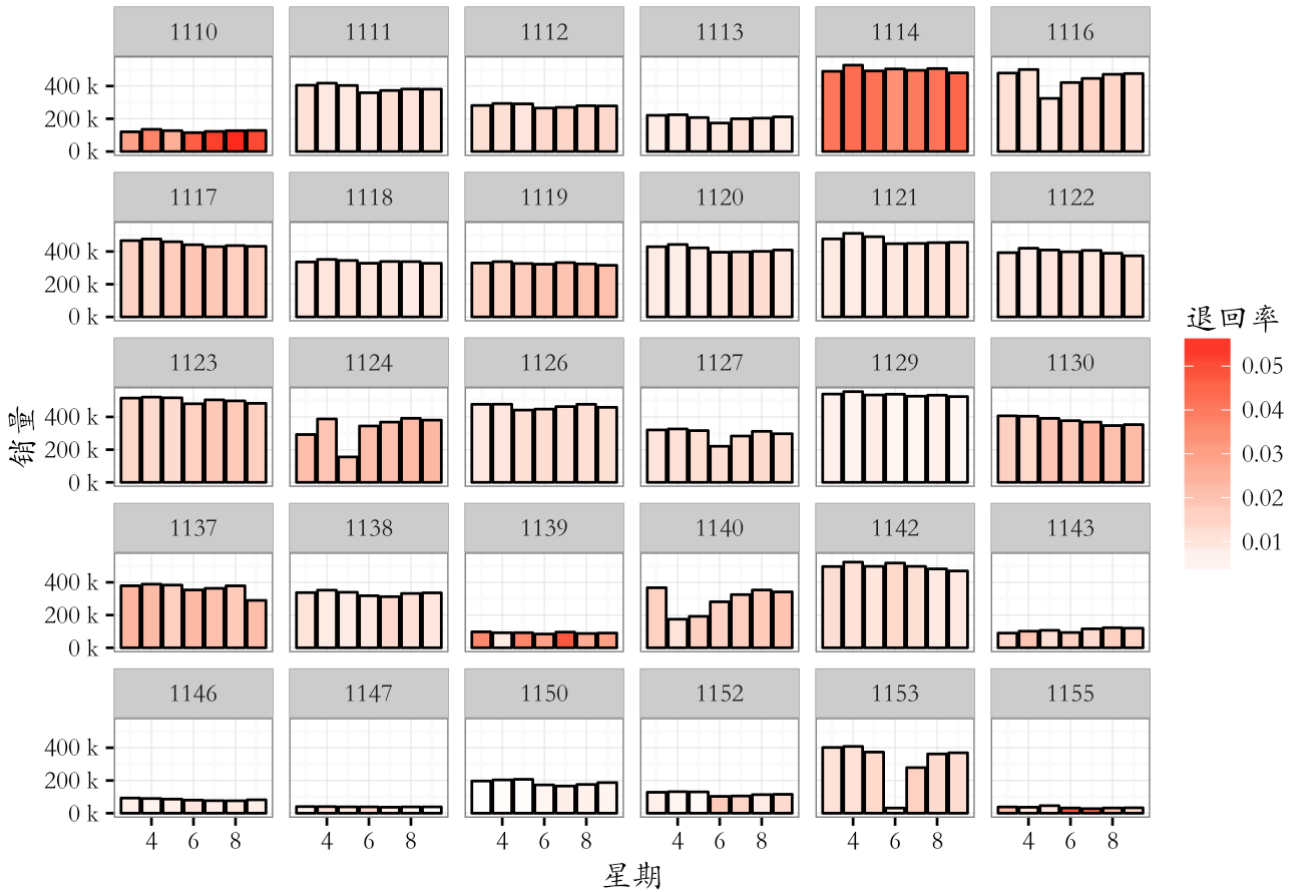

```
## 每个州的销售数据的分析
states <- agencias_history %>%
  group_by(State, Semana) %>%        #数据按照州和星期分组
  summarise(Units = sum(Units),
            Pesos = sum(Pesos),
            Return_Units = sum(Return_Units),
            Return_Pesos = sum(Return_Pesos),
            Net = sum(Net)) %>%
  mutate(Avg_Pesos = Pesos / Units,
         Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(desc(Units))   # 数据按照销量排列
head(states)
```

```
##               State Semana     Units      Pesos Return_Units Return_Pesos
## 1 ESTADO DE MÉXICO      3 10653345 103437148       154026      1510612
## 2 ESTADO DE MÉXICO      4 10584696 100758414       148409      1484139
## 3 ESTADO DE MÉXICO      5 10523973  99693511       144172      1449122
## 4 ESTADO DE MÉXICO      8 10422406  98640885       168741      1600855
## 5 ESTADO DE MÉXICO      9 10252045  98114659       160559      1523816
## 6 ESTADO DE MÉXICO      7 10156108  96457302       175260      1698870
##        Net Avg_Pesos Return_Rate
## 1 10552261  9.709359  0.01425194
## 2 10489671  9.519254  0.01382722
## 3 10424028  9.472992  0.01351425
## 4 10310433  9.464310  0.01593227
## 5 10144764  9.570252  0.01541968
## 6 10040746  9.497467  0.01696387
```

```
dim(states)[1] /7   # 多少个州
```

```
## [1] 33
```

```
## 地点－－星期 －－退回百分比 图像可视化
ggplot(states)+
  geom_bar(aes(x=Semana, y=Units, fill=Return_Rate), stat="identity", color="black")+
  theme_bw(base_family = "STKaiti") +
  facet_wrap(~State)+
  scale_y_continuous(labels=function(x)paste(x/1e6, "m"))+
  scale_fill_gradient(name="退回率", low="white", high="red")+
  ggtitle("州的销售量")+ ylab("销量") +xlab("星期")
```


州的销售量
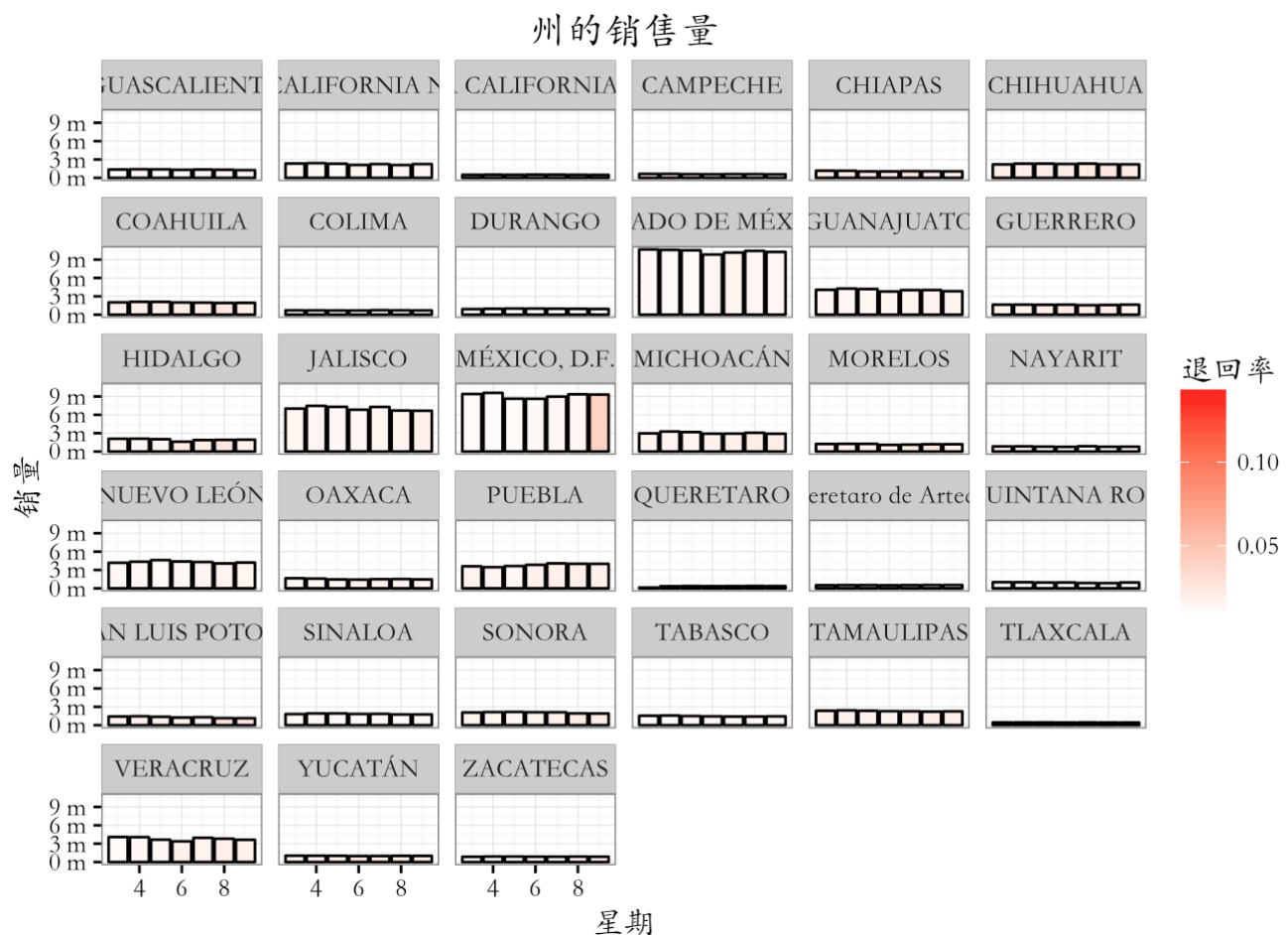
```
## 州的销量树图
states2 <- states %>%
  group_by(State) %>%
  summarise(Units = sum(Units),
            Pesos = sum(Pesos),
            Return_Units = sum(Return_Units),
            Return_Pesos = sum(Return_Pesos),
            Net = sum(Net)) %>%
  mutate(Avg_Pesos = Pesos / Units,
         Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(desc(Units))    # 数据按照销量排列

head(states2)
```

```
##                 State     Units      Pesos Return_Units Return_Pesos        Net
## 1 ESTADO DE MÉXICO  72421491  692290237      1108790     10840260  71685681
## 2    MÉXICO, D.F.   63859463  641420811      1155641      9808764  63216897
## 3         JALISCO   49266415  429703973       837289      7840476  48701496
## 4      NUEVO LEÓN   30105293  286344224       466455      5252757  29756845
## 5      GUANAJUATO   28215533  241668605       481674      4447663  27884613
## 6          PUEBLA   26631917  226746531       502023      4189637  26293428
##   Avg_Pesos Return_Rate
## 1  9.559182  0.01507937
## 2 10.044256  0.01777496
## 3  8.722047  0.01671112
## 4  9.511425  0.01525771
## 5  8.565091  0.01678470
## 6  8.514090  0.01850166
```

```
dim(states2) # 多少个州
```

```
## [1] 33  8
```

```
treemap(states2,index = c("State"),vSize = "Units",vColor = "Return_Rate",
        type = "value",palette=c("#FFFFFF","#FFFFFF","#FF0000"),
        title.legend="Units return rate", title="All State")
```


All State
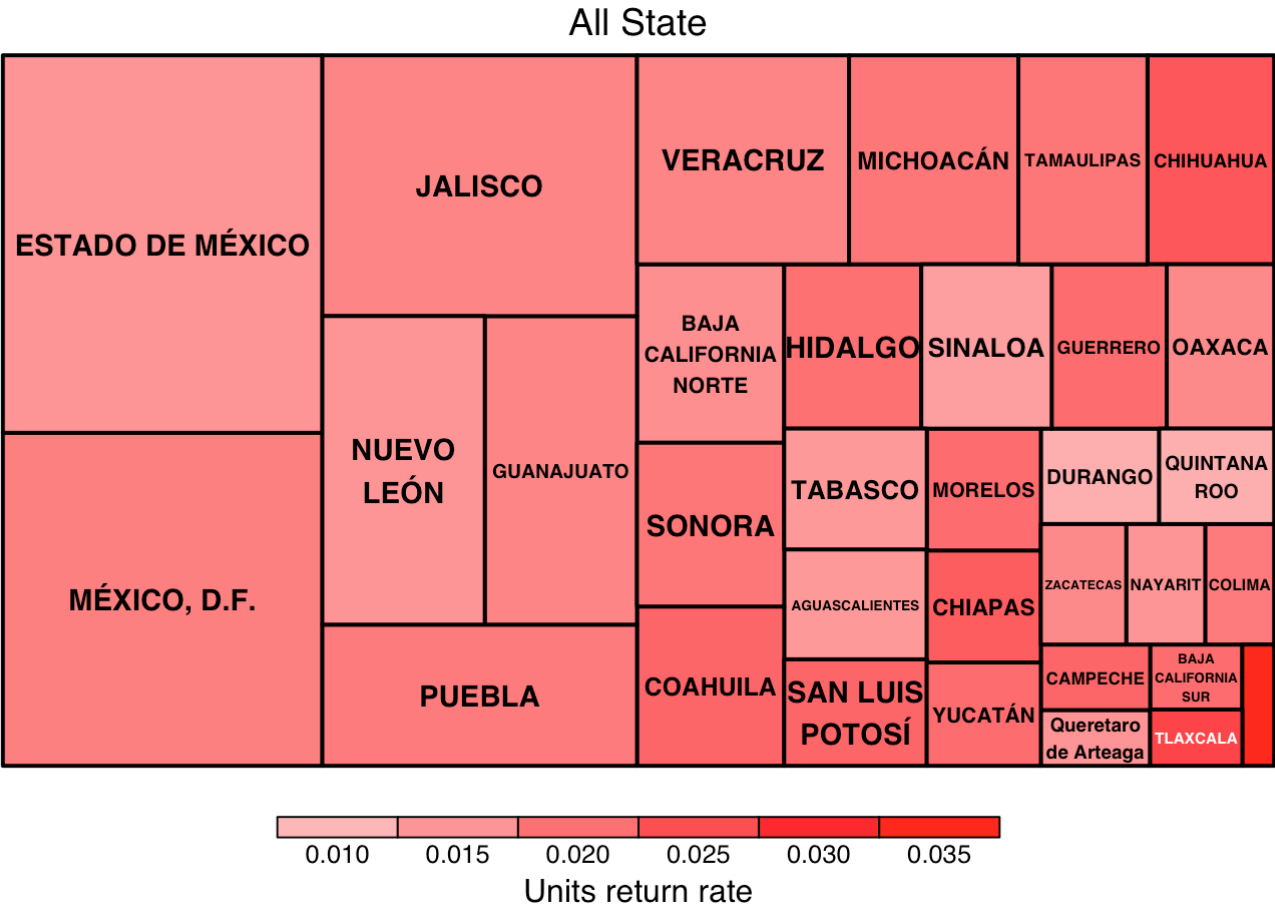
```
## 销售渠道的分析####

canals2 <- traindata %>%
  group_by(Canal_ID) %>%   #根据销售渠道和星期进行分组
  summarise(Units = sum(Venta_uni_hoy),
            Pesos = sum(Venta_hoy),
            Return_Units = sum(Dev_uni_proxima),
            Return_Pesos = sum(Dev_proxima),
            Net = sum(Demanda_uni_equil)) %>%
  mutate(Net_Pesos = Pesos - Return_Pesos,
         Avg_Pesos = Pesos / Units,
         Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(desc(Units))
canals2
```

```
## Source: local data table [9 x 9]
##
##    Canal_ID       Units        Pesos Return_Units Return_Pesos         Net
##       (int)       (int)        (dbl)        (int)        (dbl)       (int)
## 1         1   390757290   3143719571      6177725   54588903.9   386646625
## 2         2    52402129    940084794      1226250   19762509.8    51384359
## 3         4    46663250    471003920       272376    2554061.4    46468113
## 4         5    20774423    172359549      1344109    9170591.8    19891458
## 5        11    14661539    165875147       252079    2379442.9    14589584
## 6         7    10282022    119925189       132817    1588764.4    10249097
## 7         6     3569312     28573289        54311     456051.2     3549277
## 8         8     3127856     41589248       202907    1724391.6     3109500
## 9         9       33497      1533784            0          0.0       33497
## Variables not shown: Net_Pesos (dbl), Avg_Pesos (dbl), Return_Rate (dbl)
```
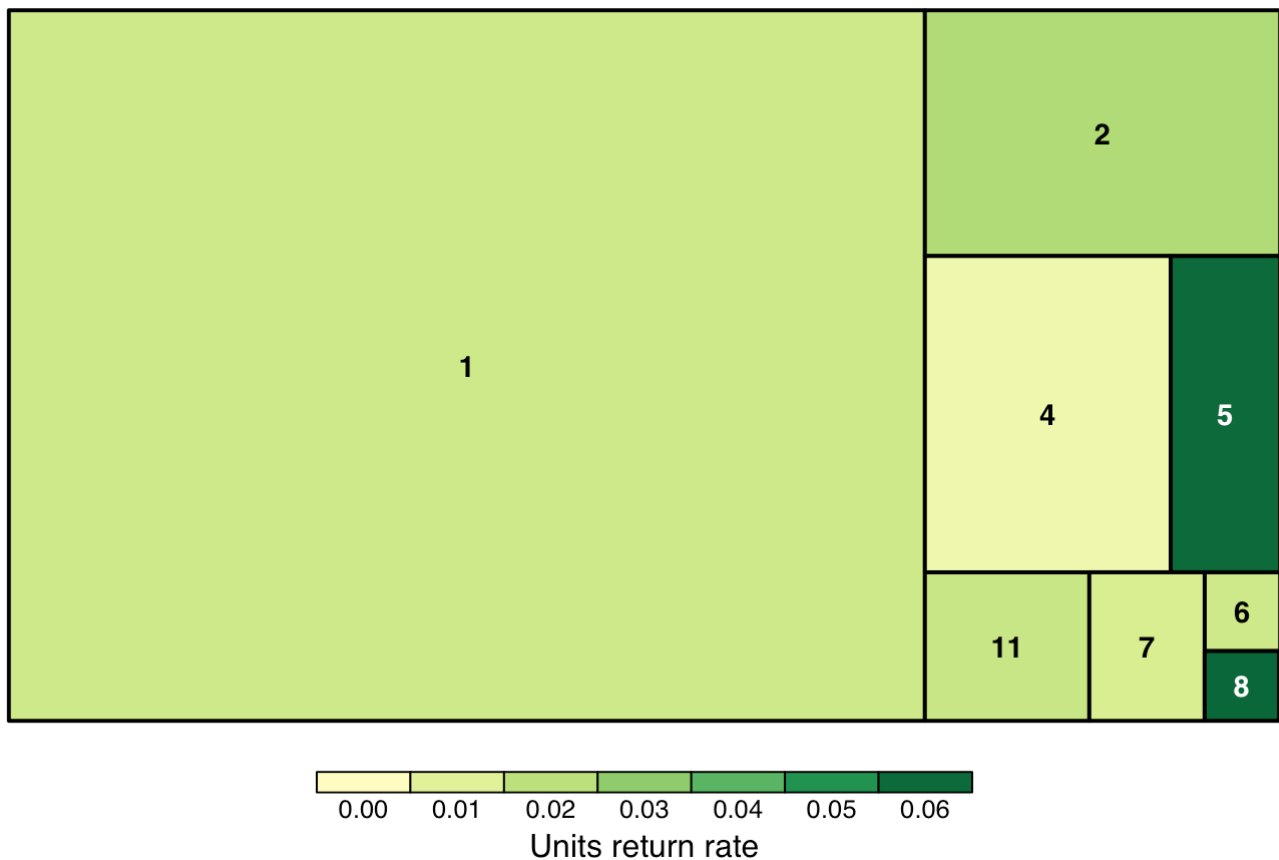
```
dim(canals2)
```

```
## [1] 9 9
```

```
# 销售渠道1占据主要的销量
treemap(canals2, index=c("Canal_ID"), vSize="Units", vColor = "Return_Rate", type="va
lue",
        #palette=c("#FFFFFF","#FFFFFF","#FF0000"),
        title.legend="Units return rate",
        title="Canals repartition")
```

Canals repartition

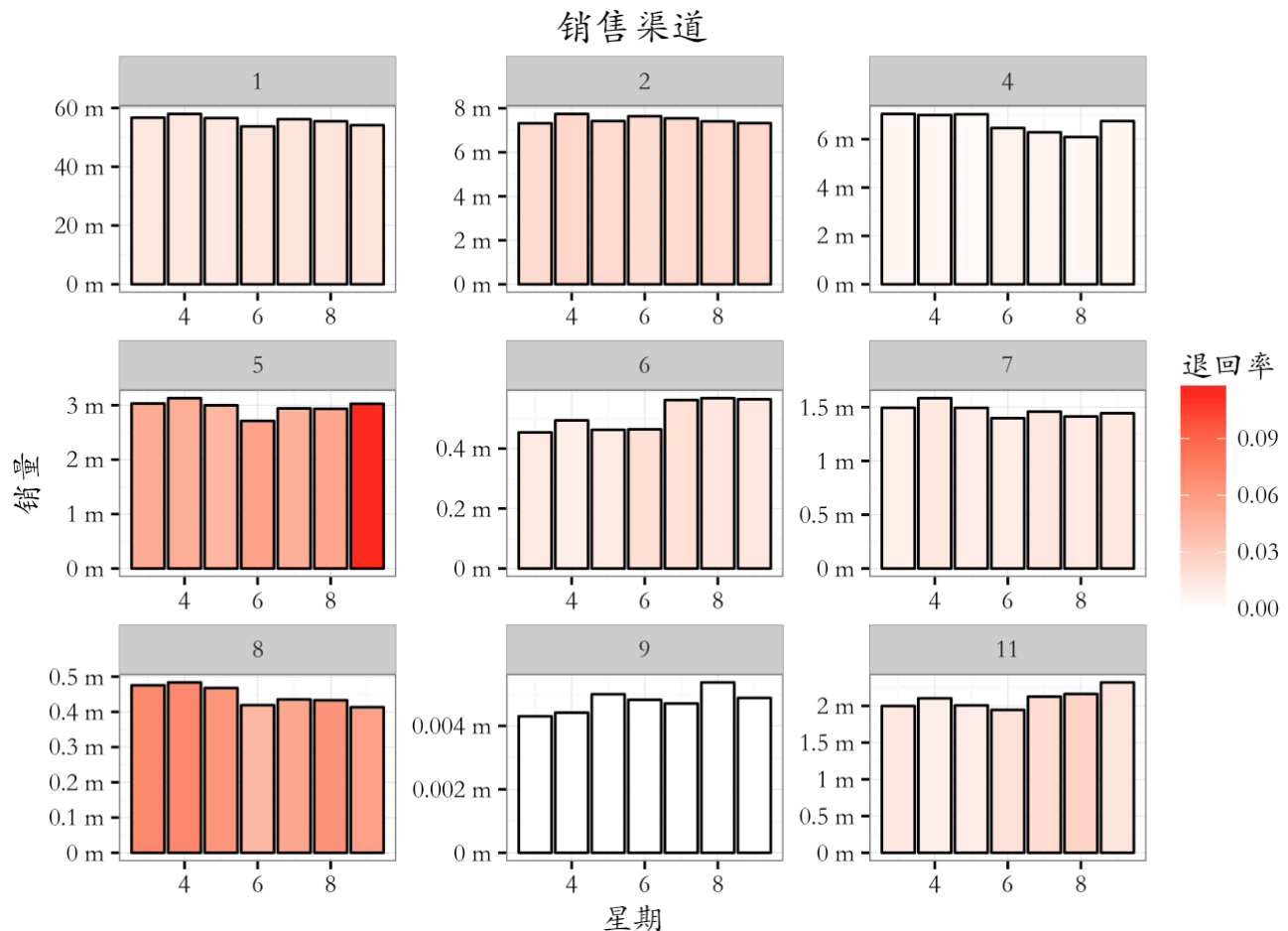Units return rate

```
##
canals <- traindata %>%
  group_by(Canal_ID, Semana) %>%   #根据销售渠道和星期进行分组
  summarise(Units = sum(Venta_uni_hoy),
            Pesos = sum(Venta_hoy),
            Return_Units = sum(Dev_uni_proxima),
            Return_Pesos = sum(Dev_proxima),
            Net = sum(Demanda_uni_equil)) %>%
  mutate(Net_Pesos = Pesos - Return_Pesos,
         Avg_Pesos = Pesos / Units,
         Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(desc(Units))
head(canals)
```

```
##   Canal_ID Semana     Units      Pesos Return_Units Return_Pesos       Net
## 1        1      4 57970962  469960421       839777      7536925  57415516
## 2        1      3 56697977  472277894       835051      7612406  56154862
## 3        1      5 56565253  450590664       798399      7198662  56026267
## 4        1      7 56188416  444780203       942701      8242810  55557236
## 5        1      8 55482753  441614833       957245      8294457  54845666
## 6        1      9 54142315  437679187       932279      7972306  53522411
##   Net_Pesos Avg_Pesos Return_Rate
## 1 462423496  8.106825  0.01427931
## 2 464665489  8.329713  0.01451429
## 3 443392001  7.965856  0.01391820
## 4 436537393  7.915870  0.01650066
## 5 433320376  7.959497  0.01696040
## 6 429706881  8.083865  0.01692757
```

```
## 销售渠道的销量和星期和退货率
ggplot(canals)+
  geom_bar(aes(x=Semana, y=Units, fill=Return_Rate), stat="identity", color="black")+
  theme_bw(base_family = "STKaiti") +
  facet_wrap(~Canal_ID, scale="free")+
  scale_y_continuous(labels=function(x)paste(x/1e6, "m"))+
  scale_fill_gradient(name="退回率", low="white", high="red")+
  ggtitle("销售渠道")+ ylab("销量") +xlab("星期")
```



```
## 销售渠道和销售站分析####
agencias_canals <- traindata %>%
  group_by(Agencia_ID) %>%
  summarise(n_canals = n_distinct(Canal_ID)) #添加该销售渠道有多少销售站
head(agencias)
```

```
##    Agencia_ID    Units     Pesos Return_Units Return_Pesos      Net Net_Pesos
## 1       1110   877675   9274674        39900     214072.8   874523   9060601
## 2       1111  2720400  24070592        25231     264672.4  2701427  23805919
## 3       1112  1959534  16591688        23924     231897.4  1942114  16359791
## 4       1113  1442999  12094484        11865     117754.4  1434414  11976730
## 5       1114  3498170  62420320       150779    2480404.7  3363796  59939915
## 6       1116  3120201  27454358        37022     377100.6  3093985  27077257
##    Return_Rate                 Town              State
## 1 0.043484184    2008 AG. LAGO FILT      MÉXICO, D.F.
## 2 0.009189509 2002 AG. AZCAPOTZALCO      MÉXICO, D.F.
## 3 0.012061763   2004 AG. CUAUTITLAN ESTADO DE MÉXICO
## 4 0.008155401    2008 AG. LAGO FILT      MÉXICO, D.F.
## 5 0.041321213  2029 AG.IZTAPALAPA 2      MÉXICO, D.F.
## 6 0.011726128  2011 AG. SAN ANTONIO      MÉXICO, D.F.
```

```
##  销售渠道有多少销售站可视化
ggplot(agencias_canals)+
  geom_histogram(aes(x=n_canals), fill="red", color="black", alpha="0.5", binwidth=0.
5)+
  theme_bw(base_family = "STKaiti") +
  scale_x_continuous(breaks=1:5)+
  scale_y_continuous()+
  theme(axis.text.x=element_text(hjust=1)) +
  labs(x = "销售渠道数量",y = "销售站数量",title = "销售站的销售渠道量")
```


销售站的销售渠道量
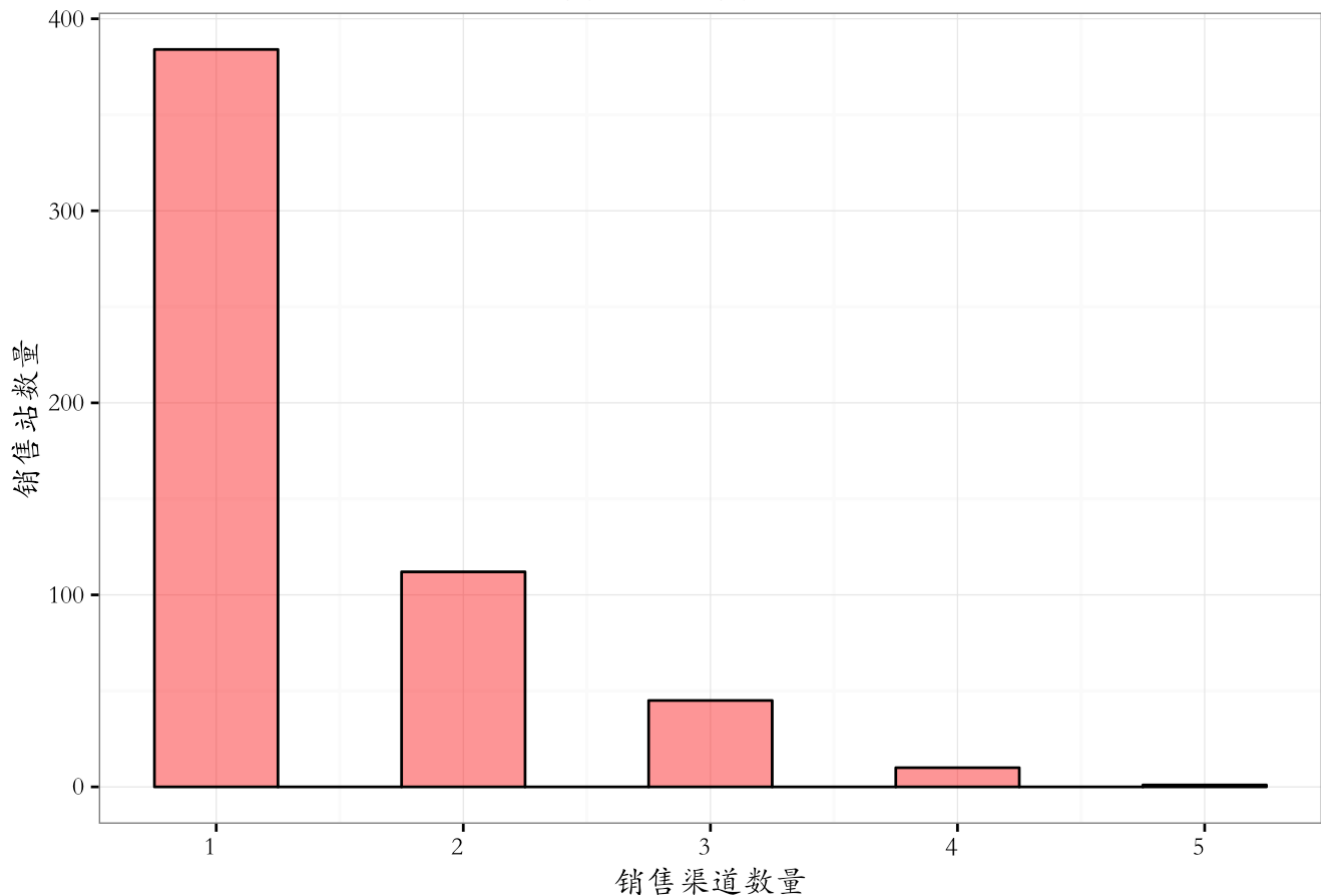
```
#  大部分销售站有1条销售渠道，只有很少的销售站有超过三条的销售渠道

## 销售路线的分析####
routes <- traindata %>% group_by(Ruta_SAK) %>%
  summarise(n_Agencias = n_distinct(Agencia_ID), #销售路线有多少销售站
            n_Clients = n_distinct(Cliente_ID), #销售路线有多少客户
            Units=sum(Venta_uni_hoy),  #销售路线的销售量
            Return_Units = sum(Dev_uni_proxima)) %>%  #销售路线的销售量退货量
  mutate(Return_Rate = Return_Units / (Units+Return_Units)) %>% # 添加退货率变量
  arrange(desc(Units))   # 按照销量排序
head(routes)
```

```
##   Ruta_SAK n_Agencias n_Clients   Units Return_Units Return_Rate
## 1     1101         82      5404 2645921        30981 0.011573453
## 2     6601         80      1104 2577239        10601 0.004096467
## 3     1102         82      5670 2571009        36771 0.014100499
## 4     1103         79      5434 2568287        37902 0.014543074
## 5     3001         49        20 2481974       402419 0.139516009
## 6     3002         49        25 2405887       125932 0.049739733
```

```
ggplot(routes, aes(x=Units/7))+
  geom_histogram(fill="red", color="gray", binwidth=5000)+
  theme_bw(base_family = "STKaiti") +
  scale_x_continuous(labels=function(x)paste(x/1000, "k"))+
  scale_y_continuous()+
  labs(x = "平均每天销量",y = "销售路线数量")
```
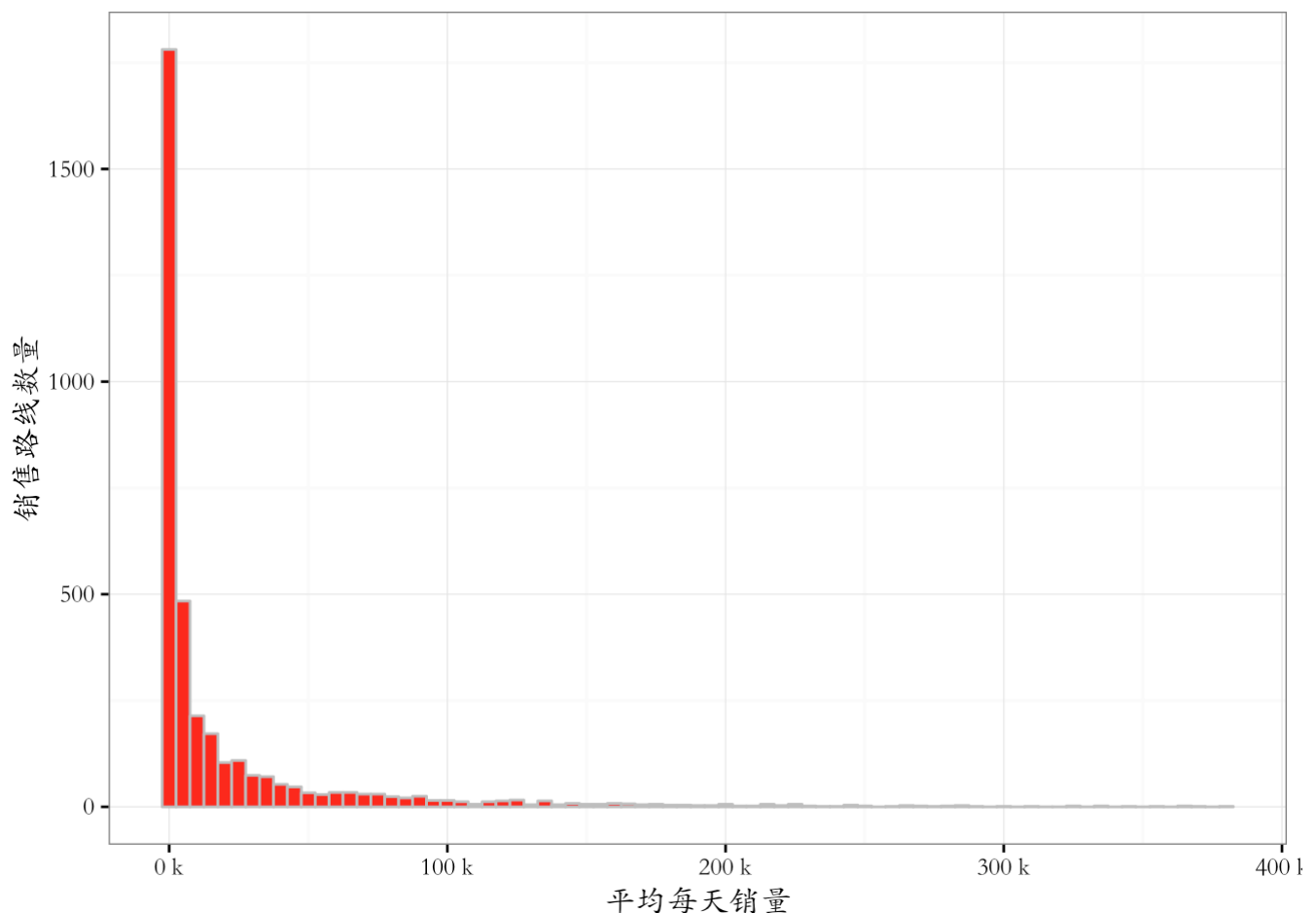
```
## 大部分的销售路线的销售量并不多，超过2/3的销售路线每天的销售量不超过10千


## 销售路线和销售站#####
routes_agencias <- traindata %>% group_by(Ruta_SAK, Agencia_ID) %>%
  summarise(count=n(),          #当前分组的观测数
            n_Clients = n_distinct(Cliente_ID), # 客户数量
            Units=sum(Venta_uni_hoy),   #销售量求和
            Return_Units = sum(Dev_uni_proxima)) %>%
  mutate(Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(desc(Units))
head(routes_agencias)
```

```
##   Ruta_SAK Agencia_ID count n_Clients  Units Return_Units  Return_Rate
## 1      900      22362  6953         9 201679            0 0.000000e+00
## 2        3       1142   773         1 163195            0 0.000000e+00
## 3      900      22560  2744         5 148973            0 0.000000e+00
## 4        1       1168   748         1 145380            8 5.502517e-05
## 5        8       1114   736         1 125192          174 1.387936e-03
## 6      900      22090  5043         6 111042            0 0.000000e+00
```

```
top100routes <- routes$Ruta_SAK[1:100]   # 销量前100的路线
top100agencias <- agencias$Agencia_ID[1:100] # 销量前100的销售站
## 可视化
ggplot(routes_agencias %>%
        filter(Ruta_SAK %in% top100routes, Agencia_ID %in% top100agencias))+
  geom_point(aes(x=as.character(Ruta_SAK),
                 y=as.character(Agencia_ID),
                 size=Units, color=Return_Rate))+
  theme_bw(base_family = "STKaiti") +
  scale_color_gradient(name="退回率", low="blue", high="red")+
  scale_size_continuous(name = "销量",range = c(1,4)) +
  theme(axis.line=element_blank(),
        axis.text.x=element_blank(),
        axis.text.y=element_blank()) +
  labs(x = "销售路线",y = "销售站",title = "销量前100的销售站&销售路线")
```

销量前100的销售站&销售路线


```
top50routes <- routes$Ruta_SAK[1:50]   # 销量前50的路线
top50agencias <- agencias$Agencia_ID[1:50] # 销量前50的销售站
## 可视化
ggplot(routes_agencias %>%
        filter(Ruta_SAK %in% top50routes, Agencia_ID %in% top50agencias))+
  geom_point(aes(x=as.character(Ruta_SAK),
                 y=as.character(Agencia_ID),
                 size=Units, color=Return_Rate))+
  theme_bw(base_family = "STKaiti") +
  scale_color_gradient(name="退回率", low="blue", high="red")+
  scale_size_continuous(name = "销量",range = c(1,6)) +
  theme(axis.line=element_blank(),
        axis.text.x=element_blank(),
        axis.text.y=element_blank()) +
  labs(x = "销售路线",y = "销售站",title = "销量前50的销售站&销售路线")
```
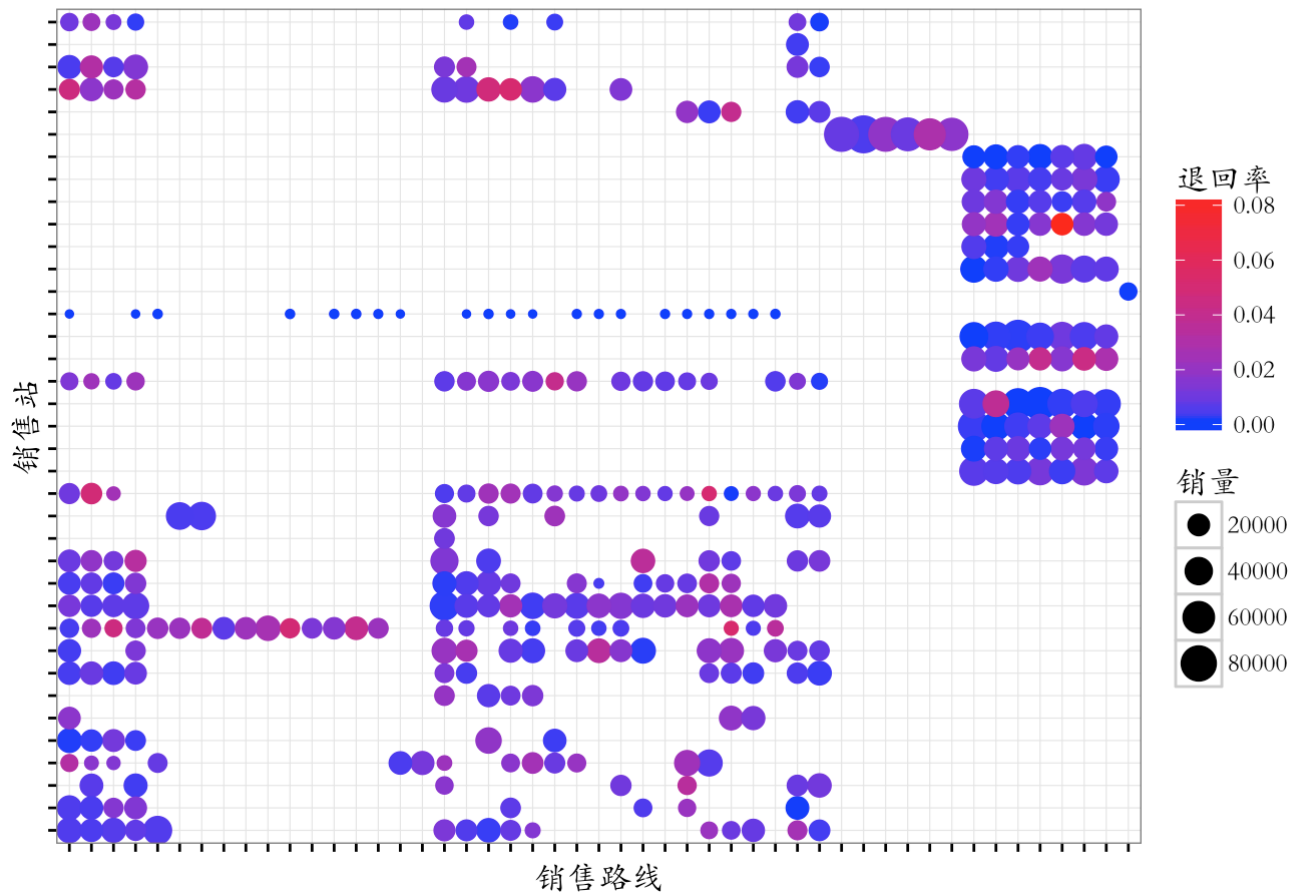
销量前50的销售站&销售路线

销售站 (y轴)
销售路线 (x轴)

退回率
0.08
0.06
0.04
0.02
0.00

销量
● 20000
● 40000
● 60000
● 80000

```
## 对客户数据进行分析#####
sales <- traindata %>%        #客户数据
  group_by(Cliente_ID) %>%    # 按照客户id分组
  summarise(Units = sum(Venta_uni_hoy),
            Pesos = sum(Venta_hoy),  # 本周销售金额
            Return_Units = sum(Dev_uni_proxima),
            Return_Pesos = sum(Dev_proxima),  # 下星期的退回金额
            Net = sum(Demanda_uni_equil)) %>%
  mutate(Return_Rate = Return_Units / (Units+Return_Units),
         Avg_Pesos = Pesos / Units) %>%    # 单价
  mutate(Net_Pesos = Pesos - Return_Pesos) %>%  # 实际销售金额
  inner_join(cliente_tabla, by="Cliente_ID") %>%
  arrange(desc(Pesos))  # 本周销售金额排序
head(sales)
```

```
##    Cliente_ID     Units       Pesos Return_Units Return_Pesos       Net
## 1     653378 18650001 154662268      1131794    7367474.15 17866224
## 2     653039   909671   7697623        16642     143066.78   893756
## 3     827594    69264   4814696            0          0.00    69264
## 4     652850   490617   4018867        59664     495570.18   440039
## 5    1216931   232517   3325557           49        395.64   232472
## 6    5903732   162633   2931618            0          0.00   162633
##    Return_Rate Avg_Pesos Net_Pesos
## 1 0.0572139182  8.292883 147294794
## 2 0.0179658496  8.461986   7554557
## 3 0.0000000000 69.512248   4814696
## 4 0.1084246049  8.191454   3523296
## 5 0.0002106929 14.302425   3325161
## 6 0.0000000000 18.025974   2931618
##                                       NombreCliente
## 1                                    PUEBLA REMISION
## 2                       QUERETARO DE ARTEAGA REMISION
## 3                       MC DONALDS ANTONIO CUAUTITLAN
## 4                                    JALISCO REMISION
## 5                                    OXXO TECNOLOGICO
## 6 WAL MART SUPER CENTER DOMINGO DIEZ CUERNAVACA
```
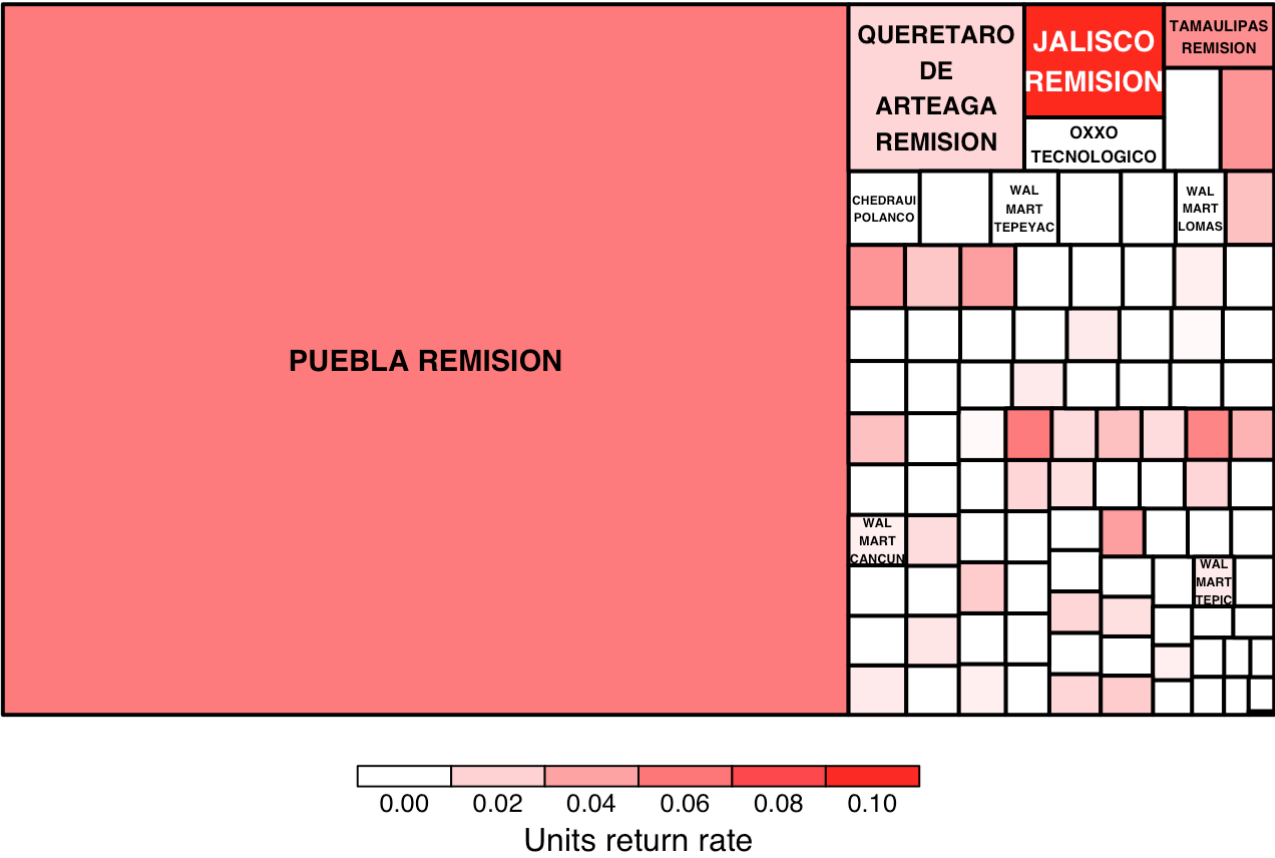
```
dim(sales)
```

```
## [1] 885416       10
```

```
# 花费量前100个客户的树形图
# 可见有一个大客户: Puebla Remision
treemap(sales[1:100, ],
       index=c("NombreCliente"), vSize="Units", vColor="Return_Rate",
       palette=c("#FFFFFF","#FFFFFF","#FF0000"),
       type="value", title.legend="Units return rate", title="Top 100 clients")
```

Top 100 clients

Units return rate

```
## 客户的累积消耗量
sales$Cum_Units <- cumsum(sales$Units) / sum(sales$Units)  # 累积百分比
s <- seq(1, 800000, 100)    # 约有80万个客户
ggplot()+geom_line(aes(x=s, y=sales$Cum_Units[s]))+
  theme_bw(base_family = "STKaiti") +
  scale_x_continuous(labels=function(x) paste(x/1000, "k"))+
  ggtitle("客户分配")+ xlab("客户数量")+ylab("累积比率")
```

客户分配
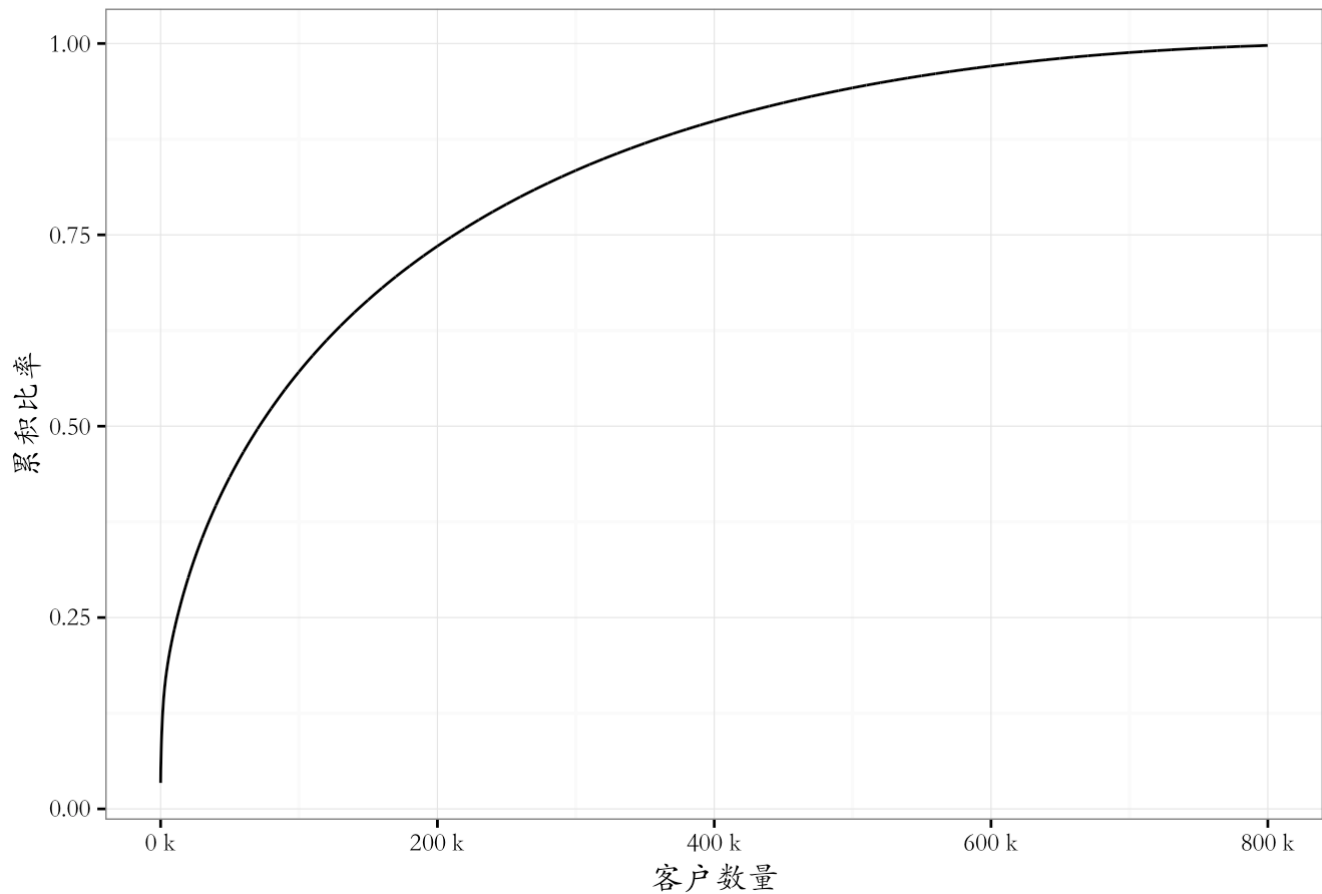

```
## 前20万客户约贡献了75%的销售量
```

```
## 客户和销售站分析####
agencias_by_client <- traindata %>%
  group_by(Cliente_ID) %>%   #按照客户id分组
  summarise(n_agencias = n_distinct(Agencia_ID)) %>% #多少个销售站
  inner_join(cliente_tabla, by="Cliente_ID")
head(agencias_by_client)
```

```
##   Cliente_ID n_agencias                      NombreCliente
## 1         26          2 BODEGA COMERCIAL MEXICANA TOLUCA
## 2         60          2                  SAMS CLUB TOLUCA
## 3         65          2                 WAL MART METEPEC
## 4        101          1                   WAL MART TOLUCA
## 5        105          1  SUPER KOMPRAS SAN BUENAVENTURA
## 6        106          1                         ISSSTE 21
```

```
dim(agencias_by_client)
```

```
## [1] 885416       3
```

```
# 单个客户使用销售站的数量
# 大部分的客户只从一个销售站购买，只有几个客户购买狗的销售站 >= 5
table(agencias_by_client$n_agencias)
```

```
##
##      1       2       3       4       5       9      62
## 844113   37510    3771      19       1       1       1
```

```
## 使用销售站多的客户信息
agencias_by_client %>% filter(n_agencias %in% c(5, 9, 62))  #返回符合条件的行
```

```
## Source: local data table [3 x 3]
##
##    Cliente_ID n_agencias                 NombreCliente
##         (int)      (int)                         (chr)
## 1     188391          9           DESAYUNOS ESCOLARES
## 2     653378         62               PUEBLA REMISION
## 3    1274327          5 COMERCIALIZADORA LA PUERTA DEL SOL
```

```
# Cliente_ID n_agencias                 NombreCliente
# (int)          (int)                         (chr)
# 1     188391          9           DESAYUNOS ESCOLARES
# 2     653378         62               PUEBLA REMISION
# 3    1274327          5 COMERCIALIZADORA LA PUERTA DEL SOL

##  客户和购买渠道分析#####
clients_canals <- traindata %>%
group_by(Cliente_ID) %>%
  summarise(n_canals = n_distinct(Canal_ID))

## 大多数客户只有一个购买渠道。不同的销售渠道可以为一个客户提供服务。
table(clients_canals$n_canals)
```

```
##
##      1       2       3       4
## 874022    6516      65       1
```

```
# 1       2       3       4
# 874022    6516      65       1

# 很少有销售站有同一个客户通过多个渠道。
clients_agencies_canals <- traindata %>%
  group_by(Cliente_ID, Agencia_ID) %>%
  summarise(n_canals = n_distinct(Canal_ID))

table(clients_agencies_canals$n_canals)
```

```
##
##      1       2       3
## 922108    3271       3
```

```
## 客户和路线分析#####
clients_routes <- traindata %>%
  group_by(Cliente_ID) %>%
  summarise(n_routes = n_distinct(Ruta_SAK))
head(clients_routes)
```

```
##   Cliente_ID n_routes
## 1      15766        1
## 2      22926        2
## 3      24080        1
## 4      24695        1
## 5      50379        1
## 6      50395        1
```
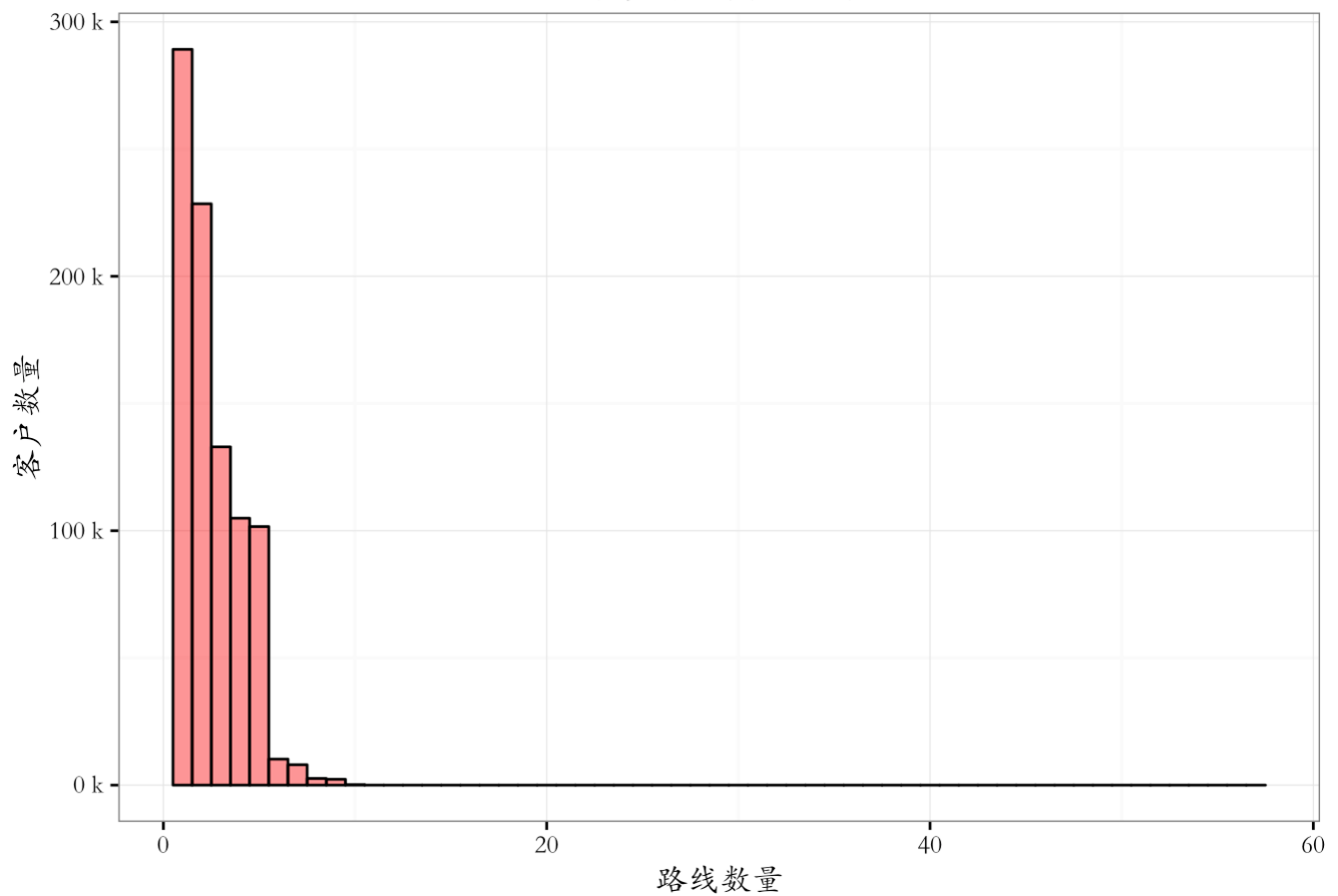
```
dim(clients_routes)
```

```
## [1] 880604        2
```

```
## 大多数客户只有不到5个仓库的交货，但超过240个客户的工作与10个仓库或更多。
sum(clients_routes$n_routes >= 10)
```

```
## [1] 242
```

```
ggplot(clients_routes)+
  geom_histogram(aes(x=n_routes), fill="red", color="black", alpha="0.5",
binwidth=1)+
  theme_bw(base_family = "STKaiti") +
  scale_y_continuous(labels=function(x) paste(x/1000, "k"))+
  ggtitle("客户和销售路线")+ xlab("路线数量")+ylab("客户数量")
```

客户和销售路线


```
## 对集团销售的产品进行分析#####

products <- traindata %>% group_by(Producto_ID) %>%   #根据生产的产品进行分组
  summarise(Units = sum(Venta_uni_hoy),   # 销量
            Pesos = sum(Venta_hoy),     # 卖出的总钱数
            Return_Units = sum(Dev_uni_proxima),  # 被退回的总量
            Return_Pesos = sum(Dev_proxima),  #备退回的总钱数
            Net = sum(Demanda_uni_equil)) %>%   #调整后的我总需求
  mutate(Avg_Pesos = Pesos / Units,  # 每种产品的单价
         Return_Rate = Return_Units / (Units+Return_Units)) %>%   # 退货率
  filter(!is.nan(Avg_Pesos)) %>% #剔除没有单价的商品
  inner_join(producto_tabla, by="Producto_ID") %>%
  arrange(desc(Units))
head(products)
```

```
##   Producto_ID     Units      Pesos Return_Units Return_Pesos      Net
## 1        2425  23860309  107365673       170005     765022.5  23728674
## 2        1278  19660667   88448180       178937     805123.9  19536596
## 3        1284  19625219   59297775       122273     387284.7  19539579
## 4       43285  15259454   80239869       158415     834036.1  15157951
## 5       36610  12836368    9884190       297745     229263.6  12680243
## 6        1240  12313013  104695281       195520    1800498.9  12167245
##   Avg_Pesos Return_Rate                      NombreProducto
## 1 4.4997604 0.007074606      Nito 1p 62g Central BIM 2425
## 2 4.4987375 0.009019182               Nito 1p 62g BIM 1278
## 3 3.0215089 0.006191824           Rebanada 2p 55g BIM 1284
## 4 5.2583709 0.010274766       Gansito 1p 50g MTB MLA 43285
## 5 0.7700145 0.022669593 Bolsa Mini Rocko 40p 13g CU MLA 36610
## 6 8.5028157 0.015630930  Mantecadas Vainilla 4p 125g BIM 1240
```
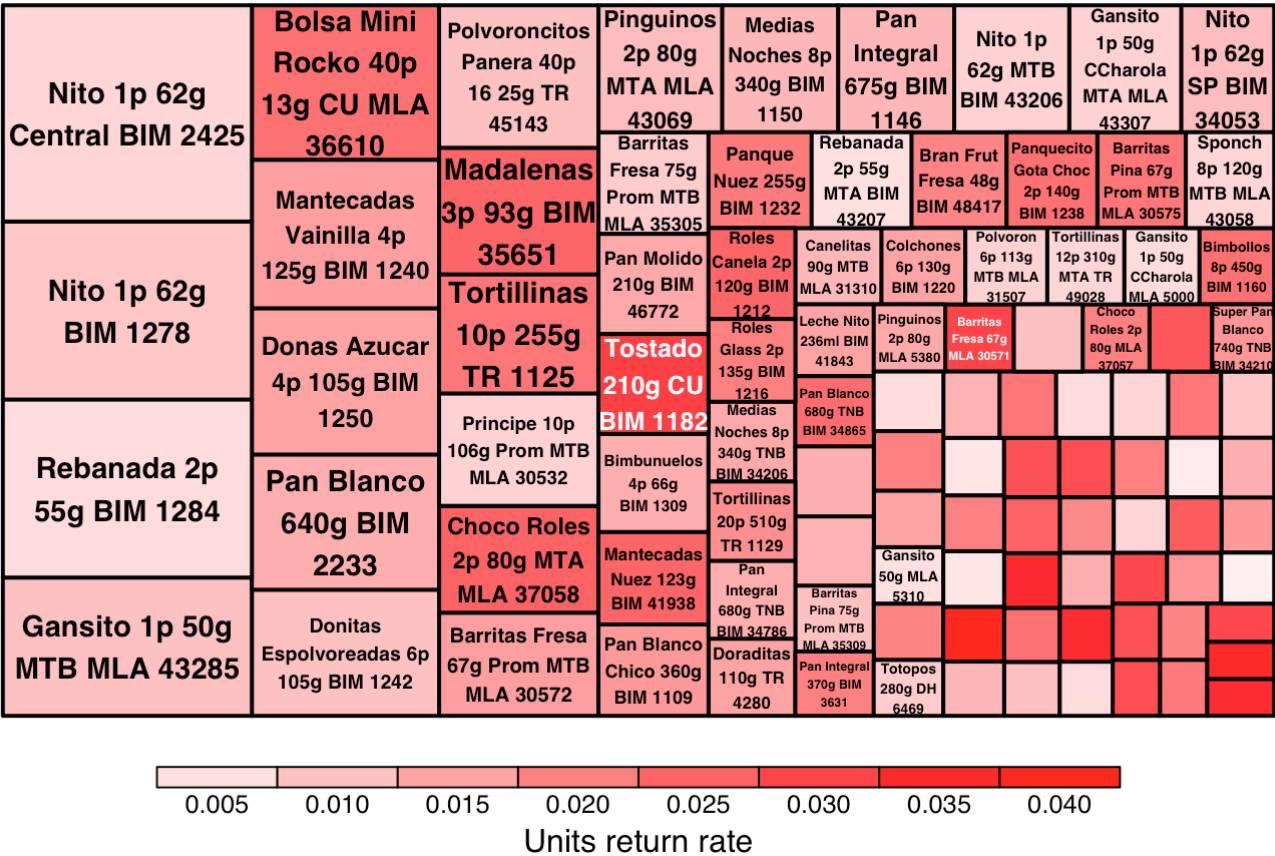
```
dim(products)   # 产品数量
```

```
## [1] 1719    9
```

```
products$NombreProducto <- factor(as.character(products$NombreProducto), levels=produ
cts$NombreProducto)

# 销量前100的产品树图
treemap(products[1:100, ],
        index=c("NombreProducto"), vSize="Units", vColor="Return_Rate",
        palette=c("#FFFFFF","#FFFFFF","#FF0000"),
        type="value", title.legend="Units return rate", title="Top 100 products")
```


Top 100 products

```
## 产品的家的密度分布
ggplot(products, aes(x=Avg_Pesos))+
  geom_histogram(aes(y=..density..), fill="gray", color="black", alpha="0.8")+
  geom_density(fill="red", alpha="0.3")+
  theme_bw(base_family = "STKaiti") +
  scale_x_continuous(lim=c(0, 50))+
  ggtitle("产品单价的分布")+ xlab("平均单价")+ylab("密度")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 74 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 74 rows containing non-finite values (stat_density).
```


产品单价的分布
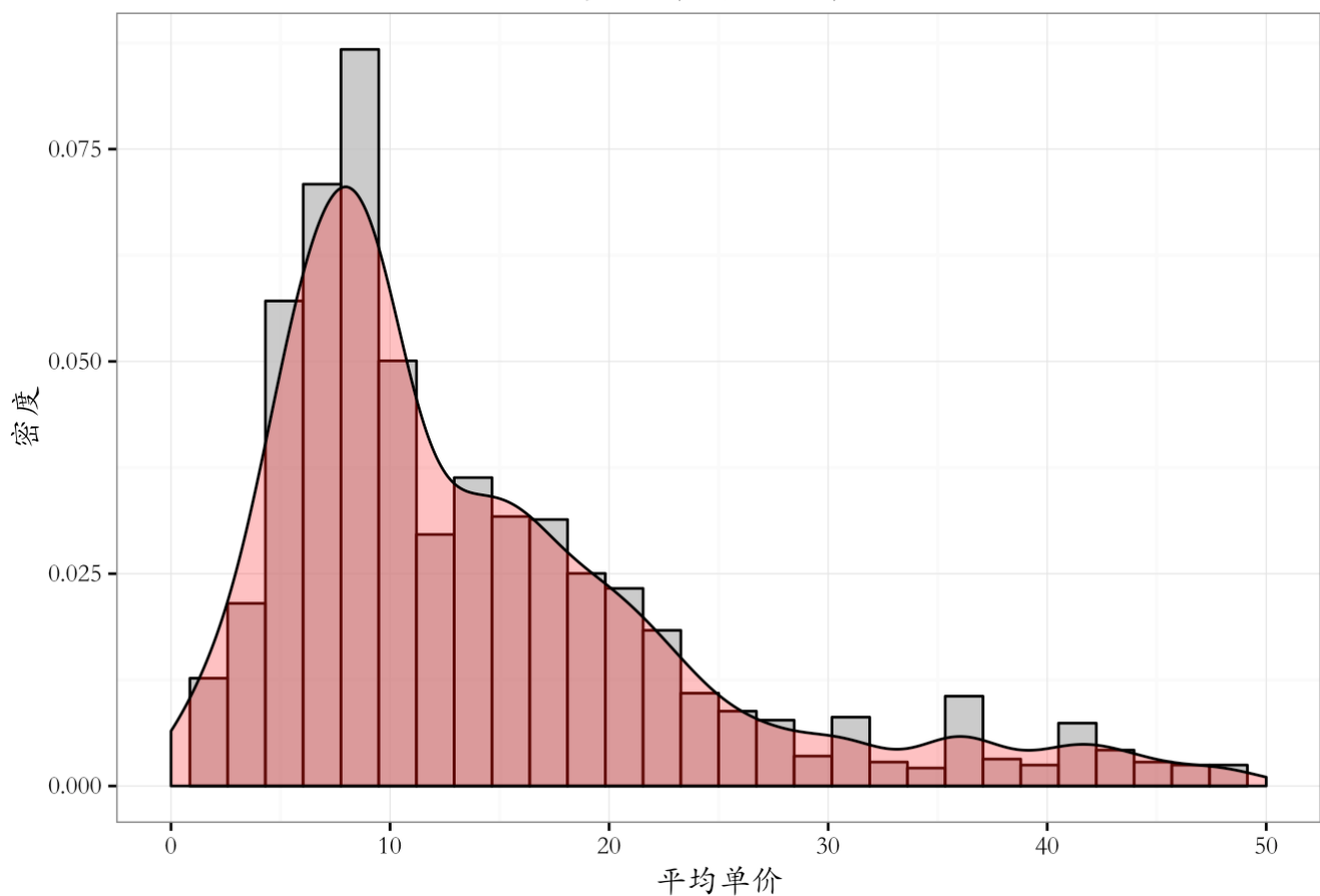
```
## 产品和销售站
products_agencies <- traindata %>% group_by(Agencia_ID) %>%
  summarise(n_products = n_distinct(Producto_ID))
head(products_agencies)
```

```
##    Agencia_ID n_products
## 1       1110        214
## 2       1111        208
## 3       1112        199
## 4       1113        202
## 5       1114        222
## 6       1116        200
```
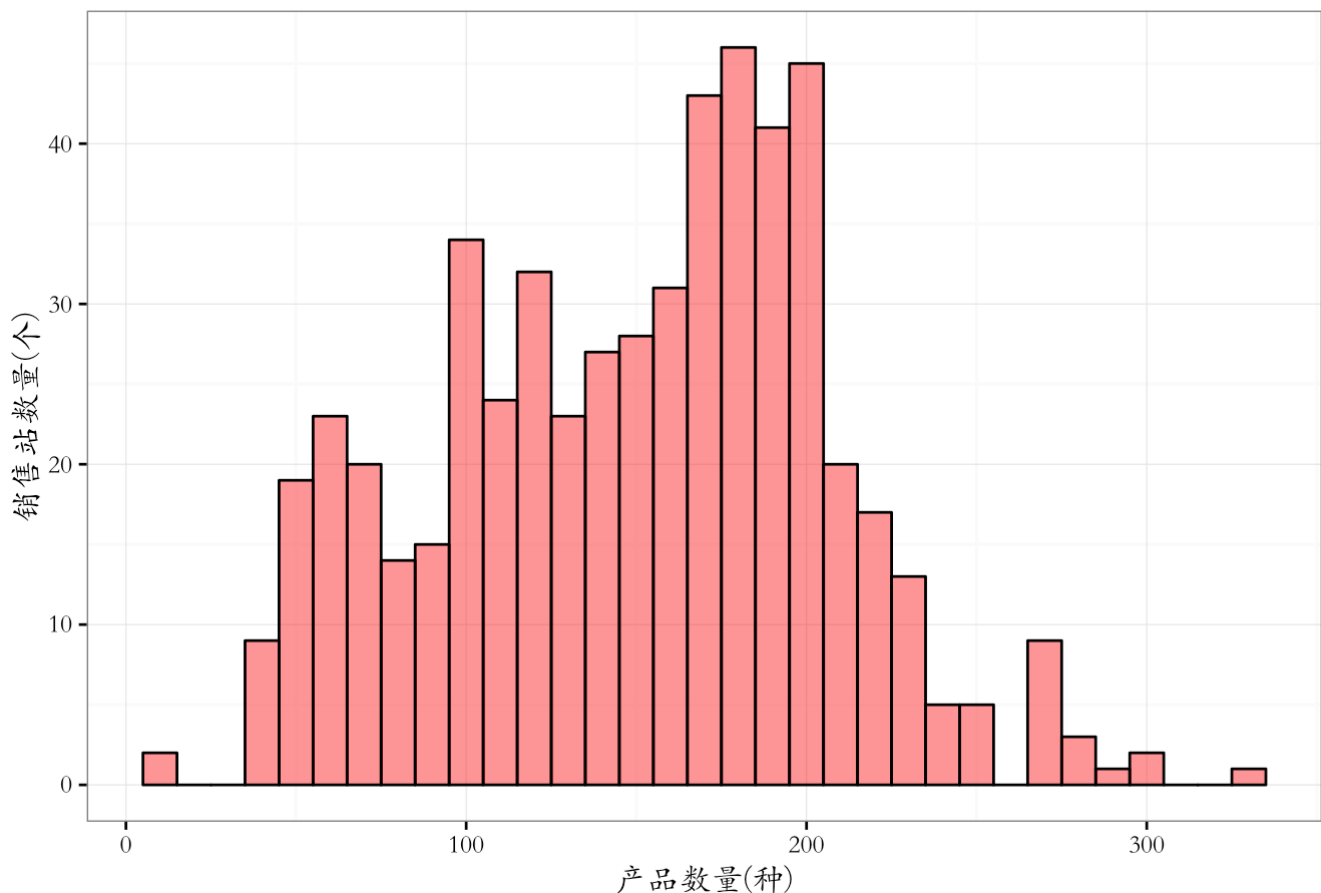
```
dim(products_agencies)
```

```
## [1] 552    2
```

```
## 大多数销售站会卖100~200种产品
ggplot(products_agencies)+
  geom_histogram(aes(x = n_products), fill="red", color="black", alpha="0.5", binwidt
h=10)+
  theme_bw(base_family = "STKaiti") +
  ggtitle("销售站出售的产品数量")+ xlab("产品数量(种)")+ylab("销售站数量(个)")
```

销售站出售的产品数量


```
## 产品和销售路线
routes_products <- traindata %>% group_by(Producto_ID) %>%
  summarise(n_routes = n_distinct(Ruta_SAK))
table(routes_products$n_routes)
```

```
## 
##     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
##   155    87    67    40    56    36    26    19    16    15    26    14    16    34    15
##    16    17    18    19    20    21    22    23    24    25    26    27    28    29    30
##    18    21    46    14    23    13    13    12    14    11     6    12     5    13     8
##    31    32    33    34    35    36    37    38    39    40    41    42    43    44    45
##    10     9     6     9    12     4     8     5     7     5     7    11     7    11    10
##    46    47    48    49    50    51    52    53    54    55    56    57    58    59    60
##     9    14     4     5     3     5     5     4     6     3     6     8     7     3     5
##    61    63    64    65    66    67    68    69    70    71    72    73    74    75    76
##     7     7     4     5     1     5     2     2     8     4     3     2     4     3     4
##    77    78    79    80    82    83    84    85    86    87    88    89    90    91    92
##     1     2     3     2     2     4     5     7     2     5     2     3     4     5     5
##    93    94    95    96    97    98    99   100   101   102   103   104   105   106   108
##     1     2     4     5     3     2     2     2     1     2     2     3     3     2     1
##   109   111   113   114   115   116   117   118   119   122   123   124   125   126   127
##     2     3     2     3     2     4     1     2     2     1     1     1     5     3     2
##   128   129   130   131   132   133   134   135   136   137   138   139   140   141   142
##     2     1     4     3     3     5     1     3     1     2     1     2     1     1     4
##   143   144   145   146   147   148   149   150   151   152   153   154   155   156   157
##     1     5     1     2     2     2     1     2     3     1     3     2     3     2     1
##   158   160   161   162   164   167   168   169   170   171   173   174   177   178   179
##     4     2     1     1     2     3     1     2     3     2     2     1     1     2     4
##   181   182   183   184   185   186   190   192   193   194   195   196   197   198   200
##     2     2     1     1     1     1     1     2     2     1     1     2     4     4     1
##   201   202   203   204   205   206   208   209   213   214   215   216   217   218   219
##     2     2     1     4     2     1     2     3     2     1     1     1     2     1     3
##   221   223   225   226   227   230   231   235   237   238   239   240   243   245   247
##     1     2     2     1     3     1     1     2     2     3     1     1     2     1     1
##   248   249   251   252   253   254   255   256   257   258   260   261   262   263   264
##     1     1     2     1     2     3     4     1     2     2     2     1     4     2     1
##   266   269   270   273   274   275   276   277   279   283   284   286   287   288   289
##     1     2     1     1     1     1     1     3     1     1     1     1     1     4     2
##   291   292   293   294   297   298   299   300   303   305   306   307   308   312   313
##     2     2     1     3     2     2     1     1     2     1     1     1     1     1     5
##   315   318   320   321   322   325   326   327   328   330   331   332   334   336   337
##     1     2     3     1     1     3     2     3     4     1     1     3     1     1     2
##   340   341   343   345   348   349   351   352   353   354   356   357   358   360   361
##     2     1     2     1     1     1     5     2     1     1     2     1     1     1     1
##   364   365   369   370   372   373   374   375   376   380   381   382   383   385   387
##     1     2     1     1     1     2     1     1     1     2     1     1     2     2     5
##   388   391   392   397   398   400   402   405   408   410   413   415   418   420   421
##     2     2     1     1     1     1     1     1     2     1     1     1     2     2     1
##   423   424   425   428   431   432   433   434   435   437   438   439   442   443   444
##     1     2     2     1     1     2     2     1     1     1     2     1     1     1     1
##   445   446   449   450   451   452   453   454   455   456   459   460   462   463   467
##     2     1     1     1     1     2     2     4     3     1     1     1     1     2     1
##   468   470   471   474   478   479   482   483   485   486   488   490   492   493   494
##     1     1     3     1     2     2     3     2     1     2     1     1     1     2     1
##   495   497   498   499   500   503   504   509   511   515   516   520   521   522   525
##     1     1     1     1     1     1     1     2     1     2     1     2     1     1     1
##   529   535   539   540   541   546   547   549   553   564   565   567   572   574   576
##     1     2     1     1     1     2     2     1     1     2     3     4     2     1     1
##   579   580   582   593   596   598   600   601   602   604   608   615   616   619   621
##     2     1     1     1     1     1     1     2     1     1     2     1     2     1     2
##   622   624   632   635   637   656   658   662   665   669   671   674   680   682   689
##     1     1     1     2     1     1     1     1     2     1     1     1     1     1     1
```

```
##  690  691  696  697  702  707  708  716  719  722  724  729  731  733  735
##    1    1    1    1    1    1    1    1    1    1    1    2    2    1    2
##  736  744  748  756  765  771  773  782  789  791  793  798  799  803  804
##    1    1    2    2    1    1    1    1    2    1    1    2    1    1    1
##  807  811  818  830  832  843  861  947  955  956  957  964  979  991  994
##    1    2    1    1    1    1    1    3    1    1    1    1    1    1    1
##  997 1001 1020 1029 1036 1144 1166 1196
##    1    1    1    1    1    1    1    1
```

```
## 大部分的产品只有几条销售路线，只有几种产品的销售路线很多
ggplot(routes_products)+
  geom_histogram(aes(x=n_routes), fill="red", color="black", alpha="0.5",
binwidth=10)+
  theme_bw(base_family = "STKaiti") +
  ggtitle("产品和销售线路")+ xlab("销售线路(条)")+ylab("产品数量(种)")
```

产品和销售线路
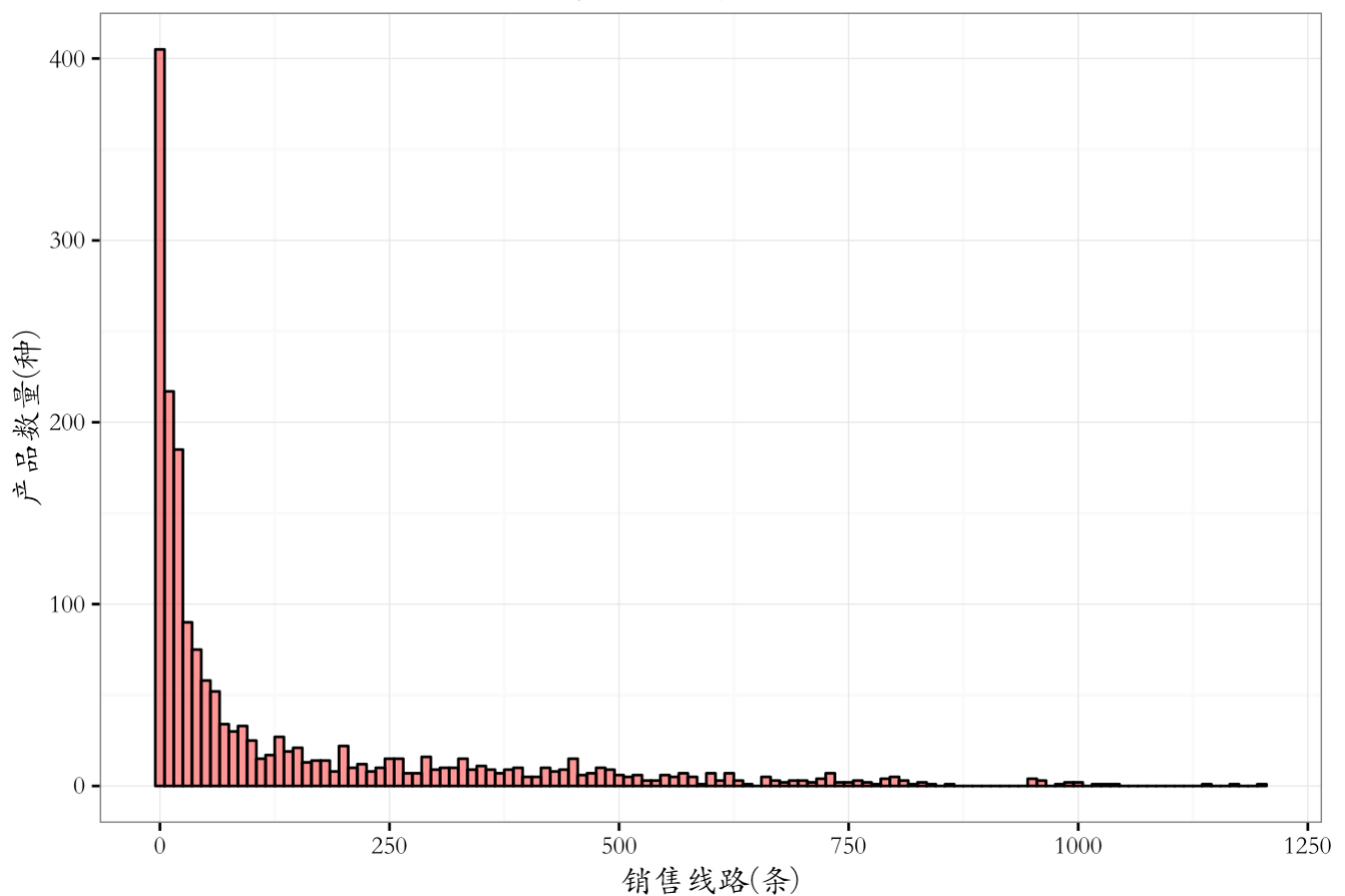

```
## 产品和销售路线 top100
routes.products <- traindata %>% group_by(Ruta_SAK, Producto_ID) %>%
  summarise(count=n(),
            n_Agencias = n_distinct(Agencia_ID),
            n_Clients = n_distinct(Cliente_ID),
            Units=sum(Venta_uni_hoy),
            Return_Units = sum(Dev_uni_proxima)) %>%
  mutate(Return_Rate = Return_Units / (Units+Return_Units)) %>%
  arrange(desc(Units))
head(routes.products)
```

```
##    Ruta_SAK Producto_ID count n_Agencias n_Clients  Units Return_Units
## 1     1154        2425 12062         27       1967 388887         2604
## 2     1151        2425 12759         27       2070 377794         2172
## 3     1155        2425 10914         24       1792 372039         2737
## 4     1152        2425 12196         26       2012 370659         2282
## 5     1156        2425 10844         24       1780 361474         2787
## 6     1153        2425 11219         24       1830 352817         2316
##    Return_Rate
## 1 0.006651494
## 2 0.005716301
## 3 0.007303029
## 4 0.006118930
## 5 0.007651107
## 6 0.006521500
```

```
dim(routes.products)
```

```
## [1] 250265       8
```
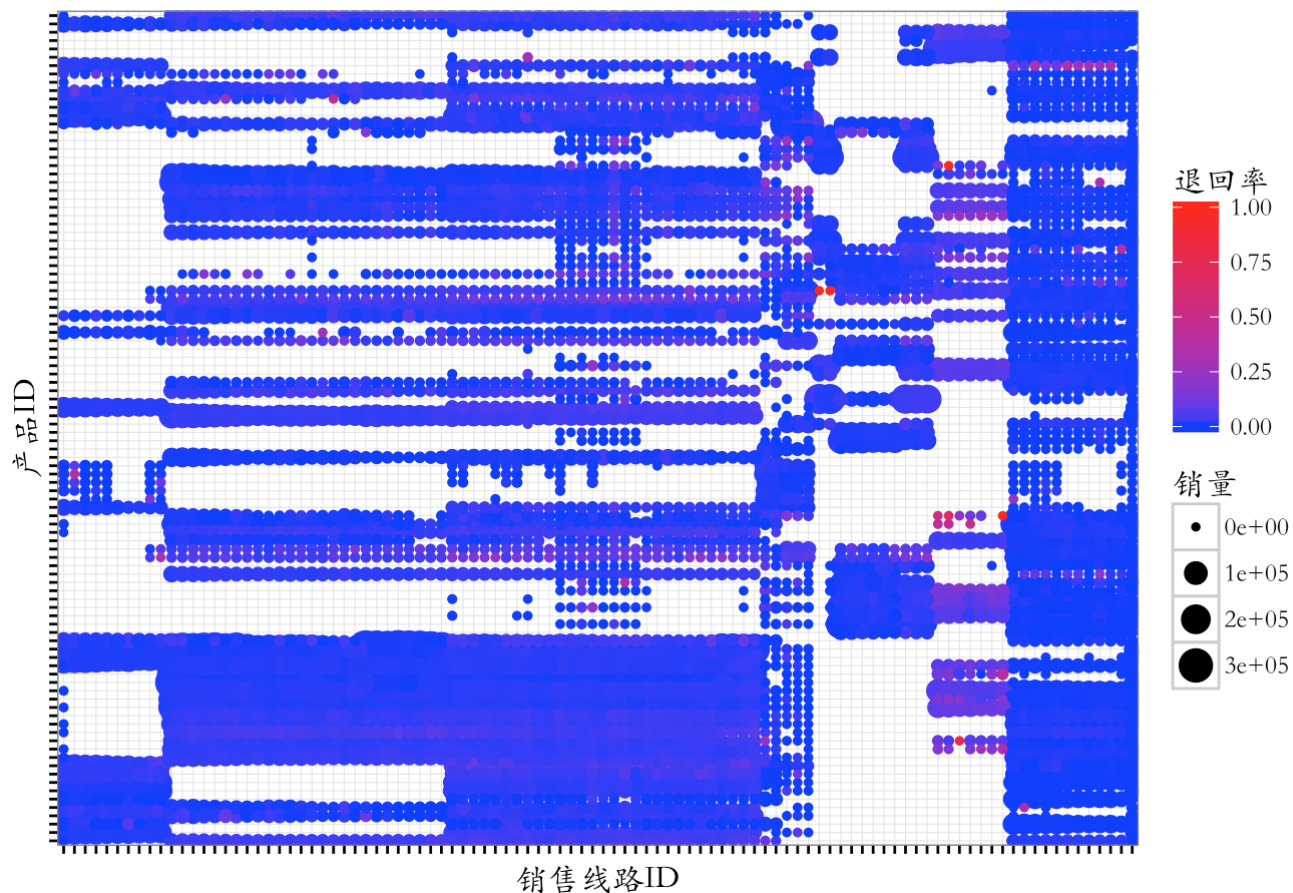
```
top100routes <- routes$Ruta_SAK[1:100]
top100products <- products$Producto_ID[1:100]

ggplot(routes.products %>%
        filter(Ruta_SAK %in% top100routes, Producto_ID %in% top100products))+
  geom_point(aes(x=as.character(Ruta_SAK),
                 y=as.character(Producto_ID),
                 size=Units, color=Return_Rate))+
  theme_bw(base_family = "STKaiti")+
  scale_color_gradient(name="退回率", low="blue", high="red")+
  scale_size_continuous(name = "销量",range = c(1,6)) +
  theme(axis.line=element_blank(),
        axis.text.x=element_blank(),
        axis.text.y=element_blank()) +
  ggtitle("销量前100产品&线路")+ xlab("销售线路ID")+ylab("产品ID")
```

销量前100产品&线路


```
## 产品和销售路线 top50
top50routes <- routes$Ruta_SAK[1:50]
top50products <- products$Producto_ID[1:50]

ggplot(routes.products %>%
        filter(Ruta_SAK %in% top50routes, Producto_ID %in% top50products))+
  geom_point(aes(x=as.character(Ruta_SAK),
                y=as.character(Producto_ID),
                size=Units, color=Return_Rate))+
  theme_bw(base_family = "STKaiti")+
  scale_color_gradient(name="退回率", low="blue", high="red")+
  scale_size_continuous(name = "销量",range = c(1,8)) +
  theme(axis.line=element_blank(),
        axis.text.x=element_blank(),
        axis.text.y=element_blank()) +
  ggtitle("销量前50产品&线路")+ xlab("销售线路ID")+ylab("产品ID")
```

销量前50产品&线路
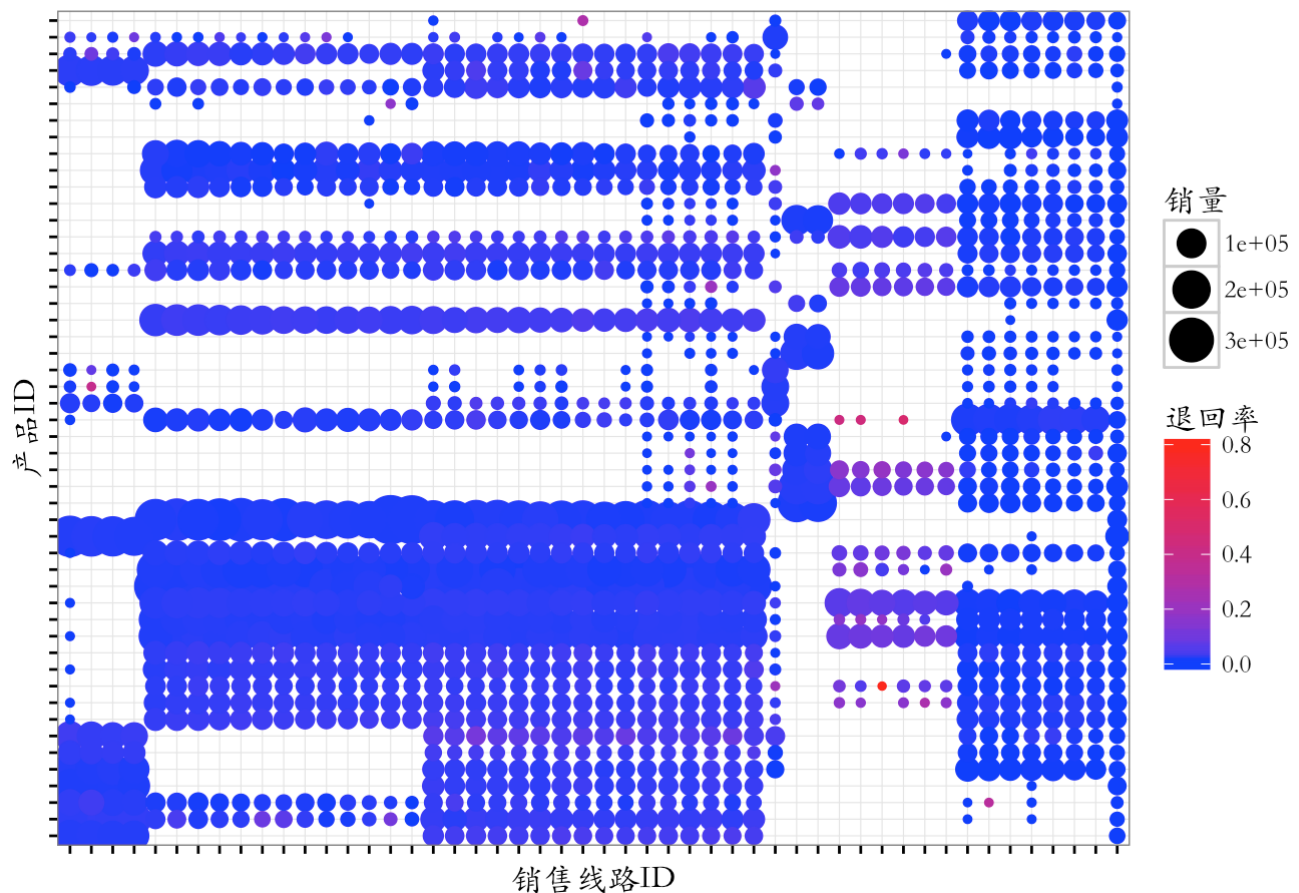
产品和客户

```
products_by_client <- traindata %>%
group_by(Cliente_ID) %>%
  summarise(n_products = n_distinct(Producto_ID)) %>%
  inner_join(cliente_tabla, by="Cliente_ID")
head(products_by_client)
```

```
##   Cliente_ID n_products                   NombreCliente
## 1         26         51 BODEGA COMERCIAL MEXICANA TOLUCA
## 2         60         34                 SAMS CLUB TOLUCA
## 3         65        112                 WAL MART METEPEC
## 4        101          4                  WAL MART TOLUCA
## 5        105         72     SUPER KOMPRAS SAN BUENAVENTURA
## 6        106         21                        ISSSTE 21
```
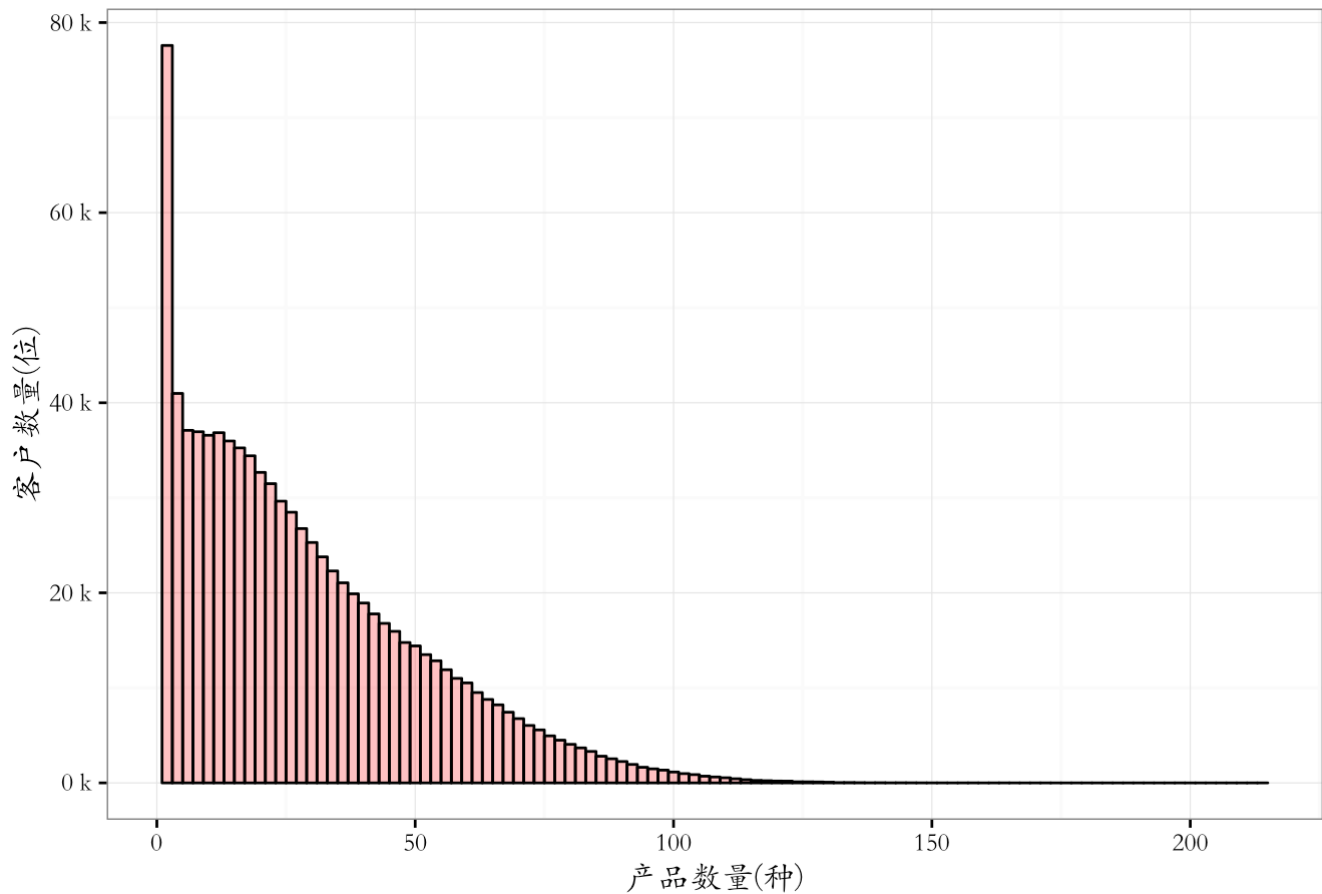
```
dim(products_by_client)
```

```
## [1] 885416       3
```

```
ggplot(products_by_client)+
  geom_histogram(aes(x=n_products), fill="red", color="black", alpha="0.3",
binwidth=2)+
  theme_bw(base_family = "STKaiti")+
  scale_y_continuous(labels=function(x)paste(x/1000, "k"))+
  ggtitle("产品量所对应的客户量")+ xlab("产品数量(种)")+ylab("客户数量(位)")
```

产品量所对应的客户量


```
## 需求量的数据分布####
Demanda_uni_equil <- traindata %>%
  group_by(Producto_ID,Semana,Agencia_ID,Canal_ID,Ruta_SAK) %>%
  summarise(Demanda = sum(Demanda_uni_equil),
            logDemanda = log(sum(Demanda_uni_equil+1)))   #销量取对数，均值
head(Demanda_uni_equil)
```

```
##    Producto_ID Semana Agencia_ID Canal_ID Ruta_SAK Demanda logDemanda
## 1      1212       3      1110        7      3301       55   4.290459
## 2      1216       3      1110        7      3301       27   3.610918
## 3      1238       3      1110        7      3301       49   4.127134
## 4      1240       3      1110        7      3301      109   4.867534
## 5      1242       3      1110        7      3301       37   3.912023
## 6      1250       3      1110        7      3301      165   5.225747
```

```
dim(Demanda_uni_equil)
```

```
## [1] 5363216         7
```

```
summary(Demanda_uni_equil$Demanda)
```

```
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.00    14.00    40.00    99.93   104.00  15120.00
```

```
summary(Demanda_uni_equil$logDemanda)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.890   3.892   3.824   4.828   9.624
```

```
ggplot(Demanda_uni_equil,aes(x =Demanda)) +
  geom_histogram(aes(y = ..density..),fill="gray", color="black", alpha="0.8",binwidt
h = 1) +
  geom_density(fill="red", alpha="0.3")+
  theme_bw(base_family = "STKaiti") +
  #scale_x_continuous(lim=c(0, 100))+
  ggtitle("产品需求量的分布")+ xlab("实际需求量")+ylab("密度")
```
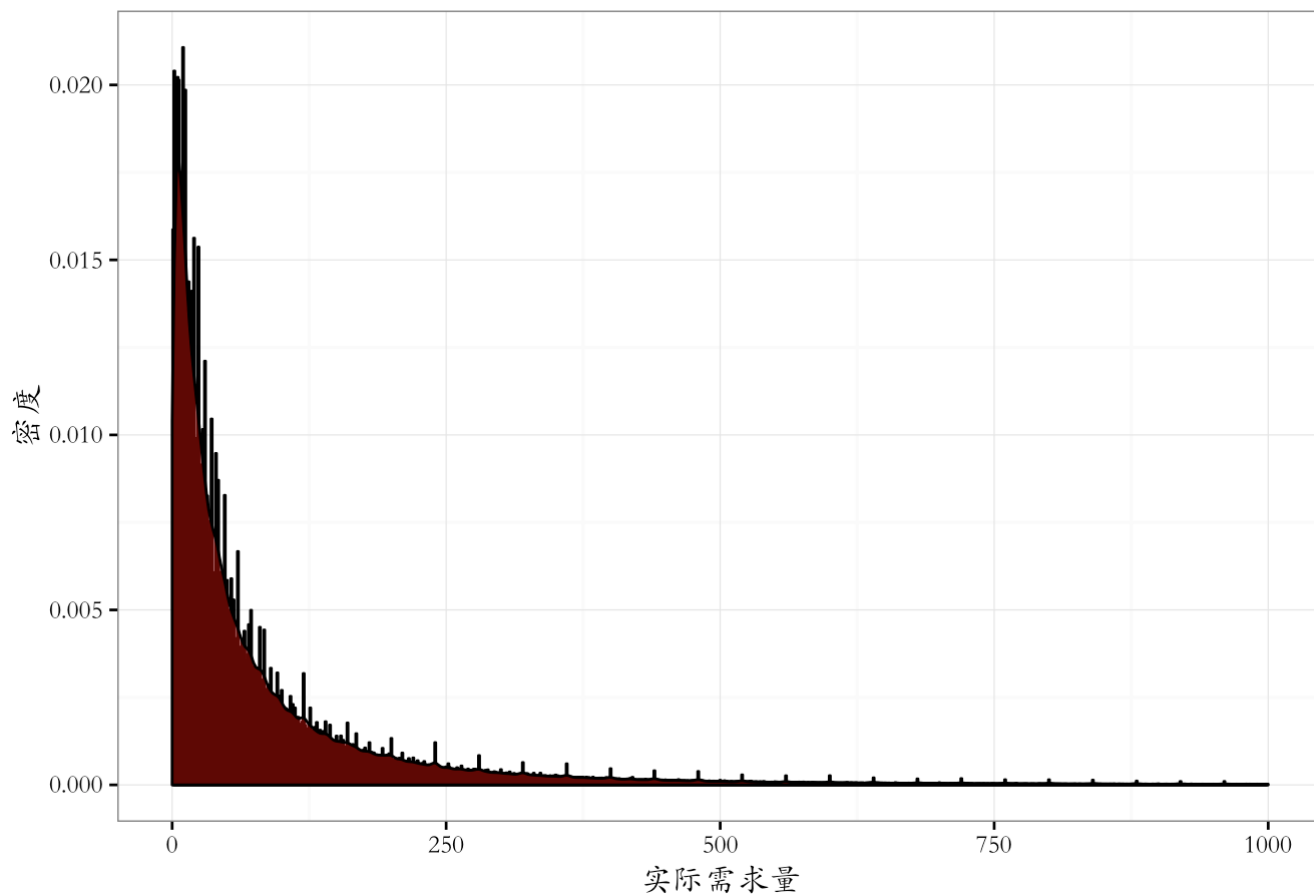
产品需求量的分布
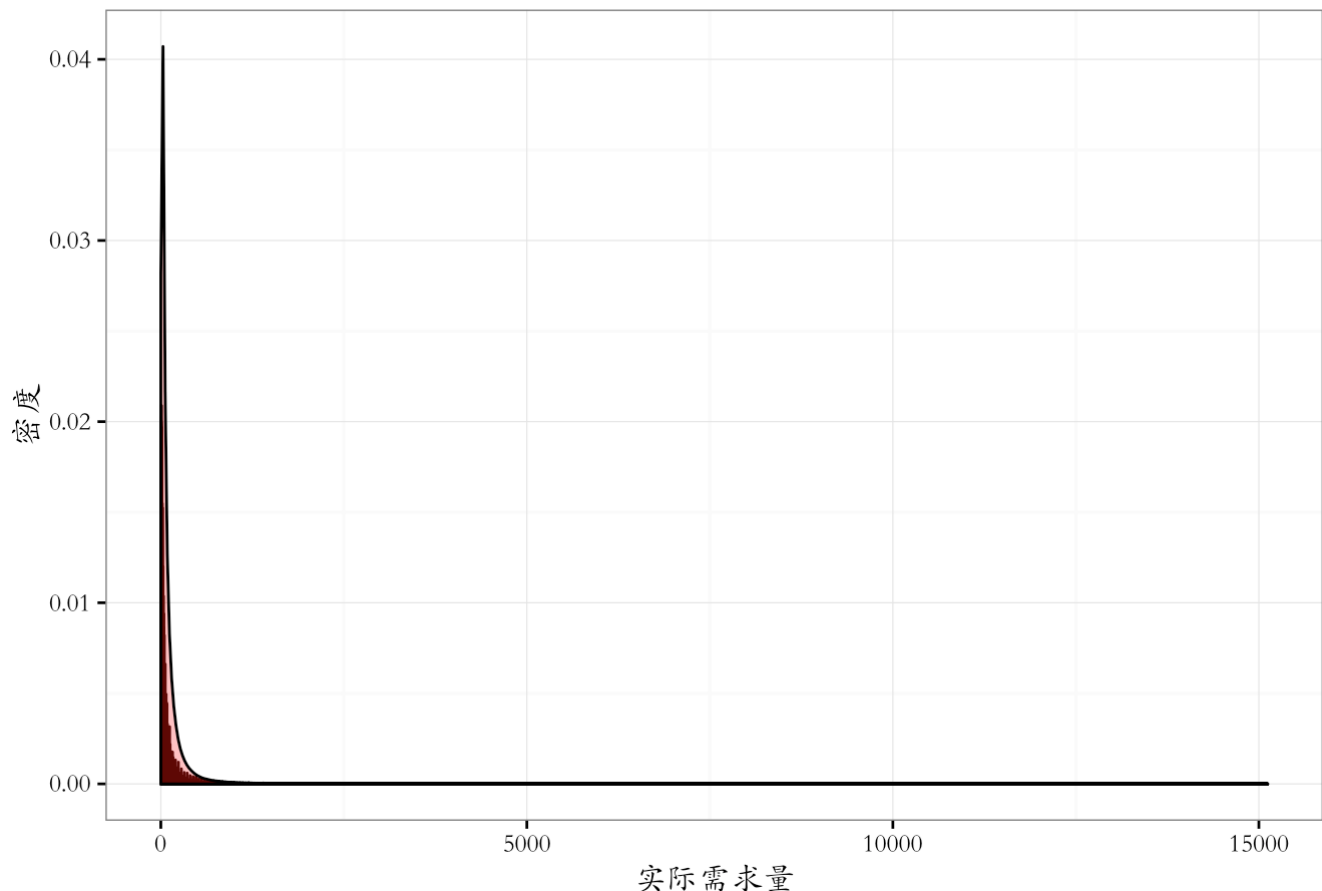

```
ggplot(Demanda_uni_equil,aes(x =Demanda)) +
  geom_histogram(aes(y = ..density..),fill="gray", color="black", alpha="0.8",binwidt
h = 1) +
  geom_density(fill="red", alpha="0.3")+
  theme_bw(base_family = "STKaiti") +
  scale_x_continuous(lim=c(0, 1000))+
  ggtitle("产品需求量的分布")+ xlab("实际需求量")+ylab("密度")
```

```
## Warning: Removed 47788 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 47788 rows containing non-finite values (stat_density).
```

产品需求量的分布


```
ggplot(Demanda_uni_equil,aes(x =logDemanda)) +
  geom_histogram(aes(y = ..density..),fill="gray", color="black", alpha="0.8",binwidt
h = 0.4) +
  geom_density(fill="red", alpha="0.2")+
  theme_bw(base_family = "STKaiti") +
  #scale_x_continuous(lim=c(0, 50))+
  ggtitle("产品需求量的分布")+ xlab("取对数后实际需求量")+ylab("密度")
```

产品需求量的分布