Group\_descrip\_lin2.R

daitu

Tue Jun 21 23:16:48 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)  
})

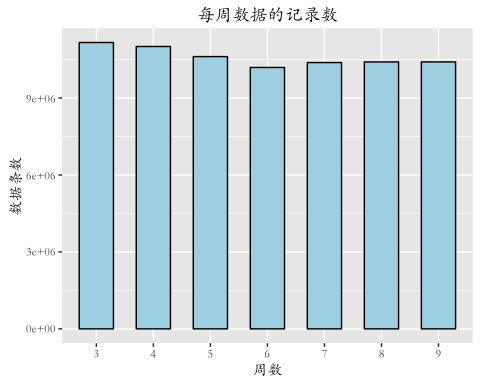
##   
Read 0.0% of 74180464 rows  
Read 4.4% of 74180464 rows  
Read 8.6% of 74180464 rows  
Read 12.9% of 74180464 rows  
Read 17.2% of 74180464 rows  
Read 21.6% of 74180464 rows  
Read 25.9% of 74180464 rows  
Read 30.2% of 74180464 rows  
Read 34.6% of 74180464 rows  
Read 38.9% of 74180464 rows  
Read 43.2% of 74180464 rows  
Read 47.5% of 74180464 rows  
Read 51.7% of 74180464 rows  
Read 56.0% of 74180464 rows  
Read 60.3% of 74180464 rows  
Read 64.6% of 74180464 rows  
Read 68.9% of 74180464 rows  
Read 73.2% of 74180464 rows  
Read 77.5% of 74180464 rows  
Read 81.9% of 74180464 rows  
Read 86.2% of 74180464 rows  
Read 90.5% of 74180464 rows  
Read 94.8% of 74180464 rows  
Read 99.1% of 74180464 rows  
Read 74180464 rows and 11 (of 11) columns from 2.980 GB file in 00:00:30

## user system elapsed   
## 26.827 2.294 29.126

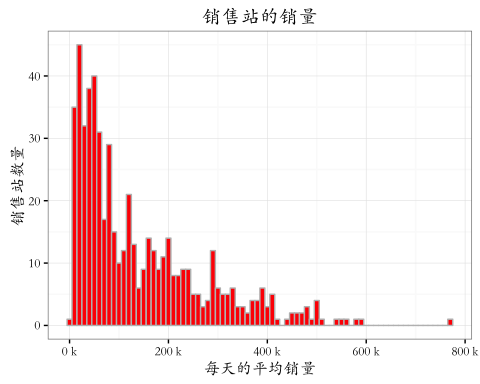
# Read 74180464 rows and 11 (of 11) columns from 2.980 GB file in 00:00:29  
# 用户 系统 流逝   
# 26.553 2.077 28.869  
  
## 2:读取客户名单数据  
cliente\_tabla <- fread("cliente\_tabla.csv",sep=",",header = TRUE)  
  
## 3:读取产品名单数据  
producto\_tabla <- fread("producto\_tabla.csv",sep=",",header = TRUE)  
  
## 4:读取城镇和国家（州）数据  
town\_state <- fread("town\_state.csv",sep=",",header = TRUE)

## 数据的描述统计####

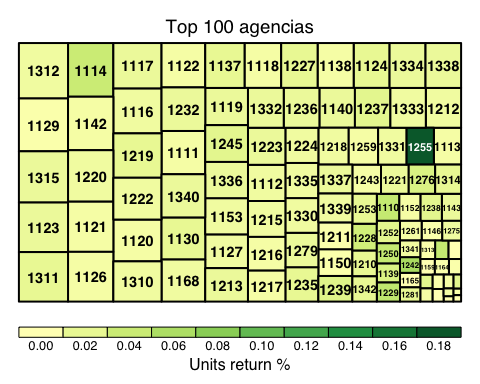
## 1:分析数据的周数：Semana  
Semana <- data.frame(table(traindata$Semana))  
colnames(Semana) <- c("Semana","Freq")  
# 3 4 5 6 7 8 9   
# 11165207 11009593 10615397 10191837 10382849 10406868 10408713   
## 条形图  
ggplot(data = Semana,aes(Semana,Freq)) +   
 geom\_bar(stat = "identity", width = 0.6,fill = "lightblue",colour = "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")



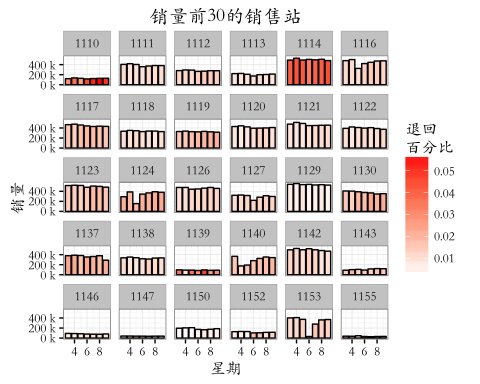
## 销售站的数据分析#####  
## 2 ：销售站Agencia 和 州State  
agencias <- traindata %>%  
 group\_by(Agencia\_ID) %>% # 数据按照销售战进行分组统计  
 summarise(Units = sum(Venta\_uni\_hoy), # 总结多个值为一个值，units：本销售站的销量和  
 Pesos = sum(Venta\_hoy), # 本周的销售量（比索）之和  
 Return\_Units = sum(Dev\_uni\_proxima), # 下星期的返回量之和  
 Return\_Pesos = 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   
  
## 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")



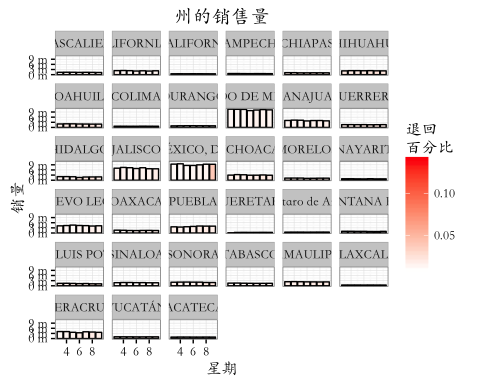
## 前100的销售站的销量树图  
treemap(agencias[1:100, ],   
 index=c("Agencia\_ID"), vSize="Units", vColor="Return\_Rate",   
 type="value", title.legend="Units return %", title="Top 100 agencias")



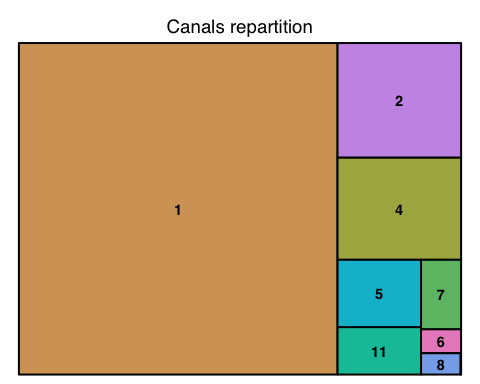
## 销售站的历史数据分析  
agencias\_history <- traindata %>%  
 group\_by(Agencia\_ID, Semana) %>% # 数据按照销售站、星期数，进行分组统计  
 summarise(Units = sum(Venta\_uni\_hoy), # 总结多个值为一个值，units：销量和  
 Pesos = sum(Venta\_hoy),  
 Return\_Units = sum(Dev\_uni\_proxima),  
 Return\_Pesos = sum(Dev\_proxima),  
 Net = sum(Demanda\_uni\_equil)) %>%  
 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")  
## 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\n百分比", low="white", high="red")+  
 ggtitle("销量前30的销售站") + ylab("销量") +xlab("星期")



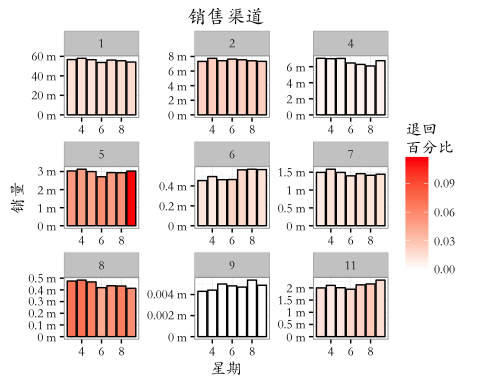
## 每个州的销售数据的分析  
states <- agencias\_history %>%  
 group\_by(State, Semana) %>% #数据按照州和星期分组  
 summarise(Units = sum(Units),  
 Pesos = sum(Pesos),  
 Return\_Units = sum(Return\_Units),  
 Return\_Pesos = sum(Return\_Pesos),  
 Net = sum(Net)) %>%  
 mutate(Avg\_Pesos = Pesos / Units,  
 Return\_Rate = Return\_Units / (Units+Return\_Units)) %>%  
 arrange(desc(Units)) # 数据按照销量排列  
  
## 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\n百分比", low="white", high="red")+  
 ggtitle("州的销售量")+ ylab("销量") +xlab("星期")



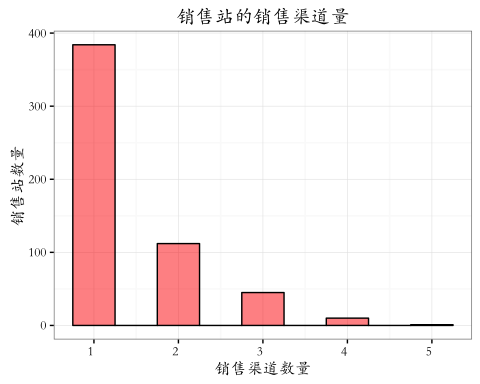
## 销售渠道的分析####  
  
canals <- traindata %>%  
 group\_by(Canal\_ID, Semana) %>% #根据销售渠道和星期进行分组  
 summarise(Units = sum(Venta\_uni\_hoy),  
 Pesos = sum(Venta\_hoy),  
 Return\_Units = sum(Dev\_uni\_proxima),  
 Return\_Pesos = sum(Dev\_proxima),  
 Net = sum(Demanda\_uni\_equil)) %>%  
 mutate(Net\_Pesos = Pesos - Return\_Pesos,  
 Avg\_Pesos = Pesos / Units,  
 Return\_Rate = Return\_Units / (Units+Return\_Units)) %>%  
 arrange(desc(Units))  
# 销售渠道1占据主要的销量  
treemap(canals, index=c("Canal\_ID"), vSize="Units", type="index",   
 title="Canals repartition")



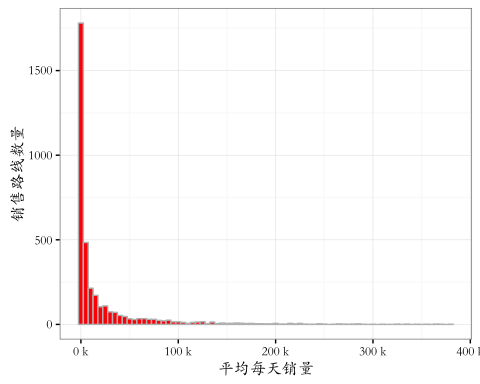
## 销售渠道的销量和星期和退货率  
ggplot(canals)+  
 geom\_bar(aes(x=Semana, y=Units, fill=Return\_Rate), stat="identity", color="black")+  
 theme\_bw(base\_family = "STKaiti") +  
 facet\_wrap(~Canal\_ID, scale="free")+  
 scale\_y\_continuous(labels=function(x)paste(x/1e6, "m"))+  
 scale\_fill\_gradient(name="退回\n百分比", low="white", high="red")+  
 ggtitle("销售渠道")+ ylab("销量") +xlab("星期")



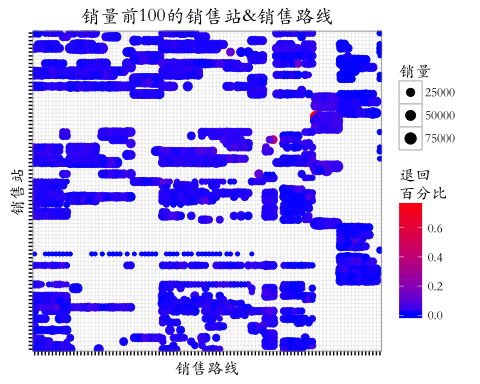
## 销售渠道和销售站分析####  
agencias\_canals <- traindata %>%  
 group\_by(Agencia\_ID) %>%  
 summarise(n\_canals = n\_distinct(Canal\_ID)) #添加该销售渠道有多少销售站  
  
## 销售渠道有多少销售站可视化  
ggplot(agencias\_canals)+  
 geom\_histogram(aes(x=n\_canals), fill="red", color="black", alpha="0.5", binwidth=0.5)+  
 theme\_bw(base\_family = "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")



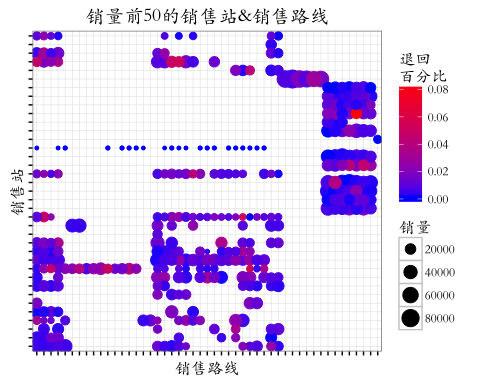
# 大部分销售站有1条销售渠道，只有很少的销售站有超过三条的销售渠道  
  
## 销售路线的分析####  
routes <- traindata %>% group\_by(Ruta\_SAK) %>%  
 summarise(n\_Agencias = n\_distinct(Agencia\_ID), #销售路线有多少销售站  
 n\_Clients = n\_distinct(Cliente\_ID), #销售路线有多少客户  
 Units=sum(Venta\_uni\_hoy), #销售路线的销售量  
 Return\_Units = sum(Dev\_uni\_proxima)) %>% #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  
  
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")



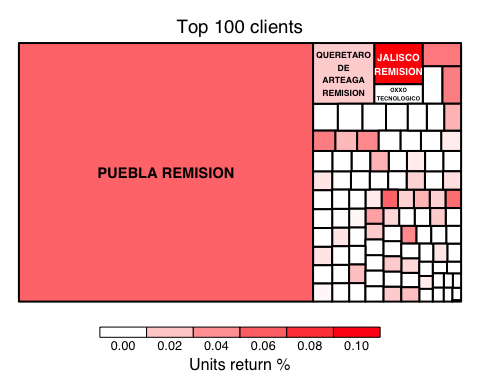
## 大部分的销售路线的销售量并不多，超过2/3的销售路线每天的销售量不超过10千  
  
  
## 销售路线和销售站#####  
routes\_agencias <- traindata %>% group\_by(Ruta\_SAK, Agencia\_ID) %>%  
 summarise(count=n(), #当前分组的观测数  
 n\_Clients = n\_distinct(Cliente\_ID), # 客户数量  
 Units=sum(Venta\_uni\_hoy), #销售量求和  
 Return\_Units = sum(Dev\_uni\_proxima)) %>%  
 mutate(Return\_Rate = Return\_Units / (Units+Return\_Units)) %>%  
 arrange(desc(Units))  
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\n百分比", 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")



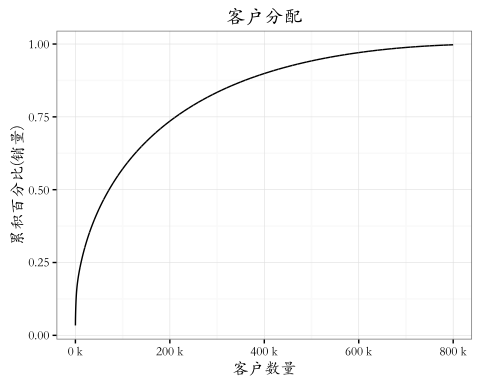
top50routes <- routes$Ruta\_SAK[1:50] # 销量前100的路线  
top50agencias <- agencias$Agencia\_ID[1:50] # 销量前100的销售站  
## 可视化  
ggplot(routes\_agencias %>%   
 filter(Ruta\_SAK %in% top50routes, Agencia\_ID %in% top50agencias))+  
 geom\_point(aes(x=as.character(Ruta\_SAK),   
 y=as.character(Agencia\_ID),   
 size=Units, color=Return\_Rate))+  
 theme\_bw(base\_family = "STKaiti") +  
 scale\_color\_gradient(name="退回\n百分比", 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")



## 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  
# 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 %", title="Top 100 clients")



## 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("累积百分比(销量)")



## 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")   
  
table(agencias\_by\_client$n\_agencias)

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

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

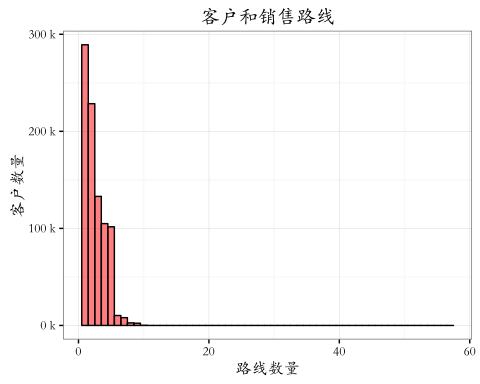
# 1 2 3 4   
# 874022 6516 65 1   
  
# 很少有销售站有同一个客户通过多个渠道。  
clients\_agencies\_canals <- traindata %>%  
 group\_by(Cliente\_ID, Agencia\_ID) %>%  
 summarise(n\_canals = n\_distinct(Canal\_ID))  
  
table(clients\_agencies\_canals$n\_canals)

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

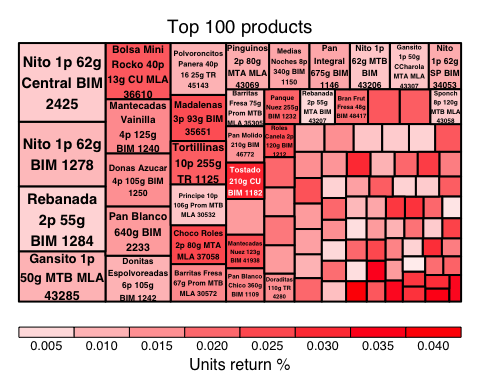
## 客户和路线分析#####  
clients\_routes <- traindata %>%  
 group\_by(Cliente\_ID) %>%  
 summarise(n\_routes = n\_distinct(Ruta\_SAK))  
  
## 大多数客户只有不到5个仓库的交货，但超过240个客户的工作与10个仓库或更多。  
sum(clients\_routes$n\_routes >= 10)

## [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")



## 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))  
  
products$NombreProducto <- factor(as.character(products$NombreProducto), levels=products$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 %", title="Top 100 products")

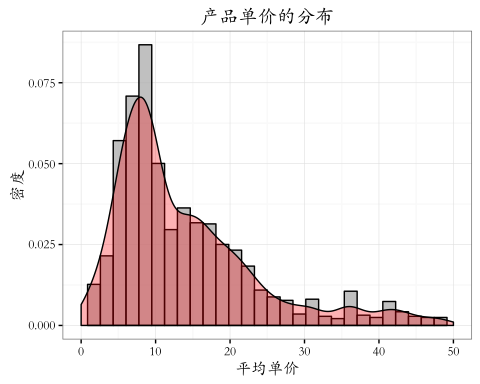


## 产品的家的密度分布  
ggplot(products, aes(x=Avg\_Pesos))+  
 geom\_histogram(aes(y=..density..), fill="gray", color="black", alpha="0.8")+  
 geom\_density(fill="red", alpha="0.3")+  
 theme\_bw(base\_family = "STKaiti") +  
 scale\_x\_continuous(lim=c(0, 50))+  
 ggtitle("产品单价的分布")+ xlab("平均单价")+ylab("密度")

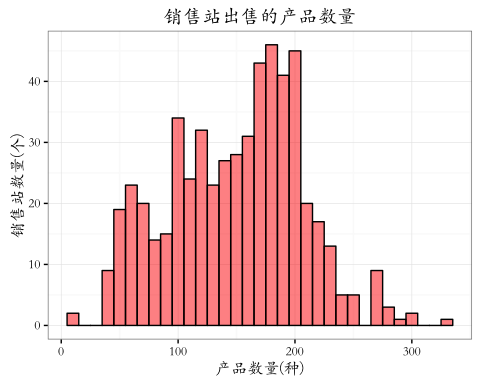
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 74 rows containing non-finite values (stat\_bin).

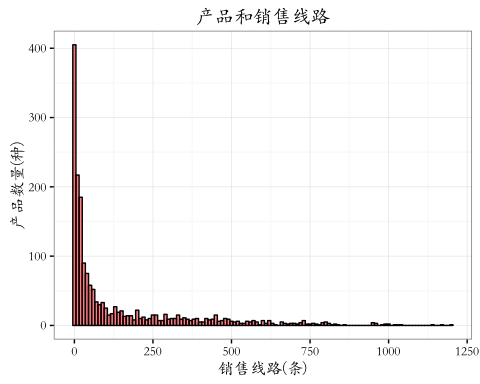
## Warning: Removed 74 rows containing non-finite values (stat\_density).



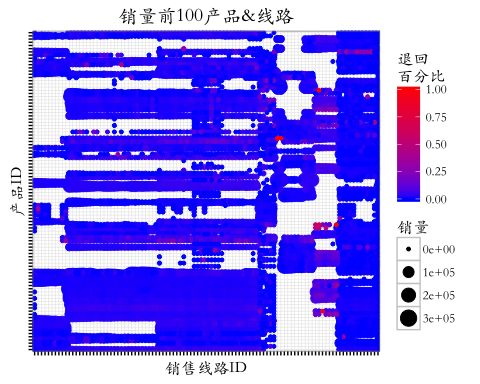
## 产品和销售站  
products\_agencies <- traindata %>% group\_by(Agencia\_ID) %>%  
 summarise(n\_products = n\_distinct(Producto\_ID))  
  
## 大多数销售站会卖100～200种产品  
ggplot(products\_agencies)+  
 geom\_histogram(aes(x = n\_products), fill="red", color="black", alpha="0.5", binwidth=10)+  
 theme\_bw(base\_family = "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)")



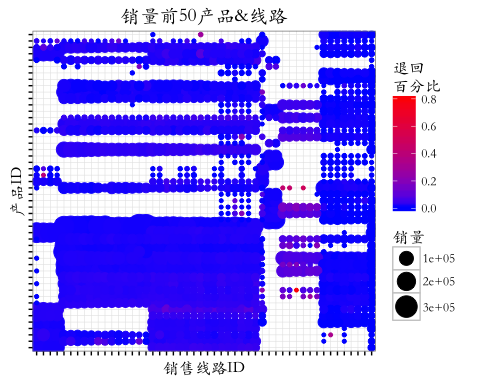
## 产品和销售路线  
routes\_products <- traindata %>% group\_by(Producto\_ID) %>%  
 summarise(n\_routes = n\_distinct(Ruta\_SAK))  
## 大部分的产品只有几条销售路线，只有几种产品的销售路线很多  
ggplot(routes\_products)+  
 geom\_histogram(aes(x=n\_routes), fill="red", color="black", alpha="0.5", binwidth=10)+  
 theme\_bw(base\_family = "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))  
  
  
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\n百分比", 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")



## 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\n百分比", 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")



## 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")  
  
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)")

