南京邮电大学

应用统计专业硕士

《统计软件应用》课程考核大作业

题　　目 小型超市订单数据分析

学　　号 1215083703

姓　　名 王飞翔

任课教师 黄宝凤

2015-2016 学年第 2 学期

评分项目	分值	得分
选题恰当		
设计科学		
内容完整		
论证能力		
文字流畅		
图表美观		
文献综述		
学术规范		
现场展示		
实用价值		
合计	100	

备注：任课老师可根据课程教学大纲、大作业要求合理分配各评分项目分值。

目录

超市销售数据分析

（本文使用通过 Rstudio rmakdown knit 自动生成 ）

查看并预处理数据

```
library(dplyr)

library(ggplot2)
library(chron)
library(reshape2)
makdata<-read.csv("E:/Rspace/data/某超市销售数据.csv", head=T)
head(makdata, 15)

(略)

summary(makdata)

（略）

str(makdata)
```

第一阶段是了解数据，通过 str()和 summary()函数查看数据特征。发现销售时间、商品名称、销售数量、销售金额中有不规范信息合计分舍去和积点印花，需要进一步清理。同时检查缺失值，未发现缺失值。尝试用 which()根据商品名称等检索时，无法检索发现商品名称前有空格。所以对数据进行以下操作：

```
sum(is.na(makdata))

## [1] 0

maktime<-as.character(makdata$销售时间)
makdata$销售时间<-strptime(maktime, "%Y%m%d%H%M%S")
makdata$商品名称<-sub("\\s", "", makdata$商品名称,  fixed = F)
makdata$商品货号<-sub("\\s", "", makdata$商品货号, fixed = F)
```

查看并删除**无效商品**：

```
invdata1<-subset(makdata, makdata$商品名称=="合计分舍去")
nrow(invdata1)

## [1] 618
```

```
invdata2<-subset(makdata，makdata$商品名称=="积点印花")
nrow(invdata2)

## [1] 265

makdata<-subset(makdata，makdata$商品名称!="合计分舍去"&makdata$商品名称!="积点印花")
```

同一订单中很多单品销售数量没有合并，所以需要**合并同一订单中商品**，同时发现存在退货产品，同一订单中的退货可以在合并同一订单商品时解决，而跨订单的退货很难消除，可以查看整体数量，发现数据量不大所以直接删除，对应以下操作:

```
valdata<-makdata %>%
  select(c(销售数量，销售金额)) %>%
  aggregate(by=list(单据号=makdata$单据号，商品货号=makdata$商品货号),
FUN=sum) %>%
  arrange(单据号) %>%
  merge(unique(makdata[，1:5])，by=c("单据号"，"商品货号")，all=FALSE)
 %>%
  mutate(销售单价=销售金额/销售数量)
nrow(subset(valdata，valdata$销售金额<0))

## [1] 74

valdata<-subset(valdata，valdata$销售金额>0)
summary(valdata)（略）
```

其中%>%为 Hadley Wickham 所作 gplyr 包中的管道操作符。可以传递变量减少中间变量。这里定义 valdata(validdata)，是对原始数据基本数据清理后的整洁数据。后面很多数据都将来自这里。这里根据 summary()信息，对很多 chr 变量因子化会方便以后分析和作图，所以：

```
valdata[，1]<-as.factor(valdata[，1])
valdata[，2]<-as.factor(valdata[，2])
valdata[，5]<-as.factor(valdata[，5])
valdata[，7]<-as.factor(valdata[，7])
```

此时得到**整洁的数据 valdata**，下面针对数据特征进行分析

数据特征与分析思路

根据源数据提供的每笔交易的时间，订单号，商品名称，销售量和销售额的信息可以通过简单的数据变型和加工得到以下易于进行二次分析的变量：

1. **时间类型数据**:时间序列，所在日、月、周、与小时
2. **交易类型数据**:销售金额，销售商品数，顾客数

3. **商品信息数据**:商品名称，商品销售量

本文针对这三种类型数据进行不同组合，从以下几个角度进行分析：

1.时间类型数据与交易类型数据结合

1.1(销售金额、商品数按种类、顾客数)-(14-10-18/15-02-28)关系
1.2(销售金额、商品数按种类、顾客数)-(1 日-31 日)关系
1.3 (销售金额、商品数按种类、顾客数)-(10 月-2 月)关系
1.4(销售金额、商品数按种类、顾客数)-(周一-周日)关系
1.5(销售金额、商品数按种类、顾客数)-(6 时-21 时)关系

2.时间类型数据与商品信息数据

2.1 热销产品-(2014-10-18 至 2015-02-28)关系
2.2 热销产品-(1 日-31 日)关系
2.3 热销产品-(10 月-2 月)关系
2.4 热销产品-(周一-周日)关系
2.5 热销产品-(6 时-21 时)关系

3.交易类型数据与商品信息数据

3.1 销售量-商品类型数目关系
3.2 关联规则挖掘

数据整理与分析

获取为日期对象为周几/月份/季度

这里使用了 chron 包的 days()，weekdays()等直接从 POSIX 格式时间获得对应时间元素的函数，并为方便后续作图将各因子设为有序因子。

```
valdata<-mutate(valdata，byday=days(销售时间)，byweek=weekdays(销售时间)，bymonths=months(销售时间)，byhours=hours(销售时间))
valdata[，9]<-as.factor(valdata[，9])
valdata[，10]<-factor(valdata[，10]，order=TRUE，levels = c("星期一"，"星期二"，"星期三"，"星期四"，"星期五"，"星期六"，"星期日"))#
valdata[，11]<-factor(valdata[，11]，order=TRUE，levels = c("十月"，"十一月"，"十二月"，"一月"，"二月"))
valdata[，12]<-as.factor(valdata[，12])
valdata$销售时间<-as.Date(valdata$销售时间)
```

1.时间类型数据与交易类型数据结合

1.1(销售金额、商品数按种类、顾客数)-(2014-10-18 至 2015-02-28)关系

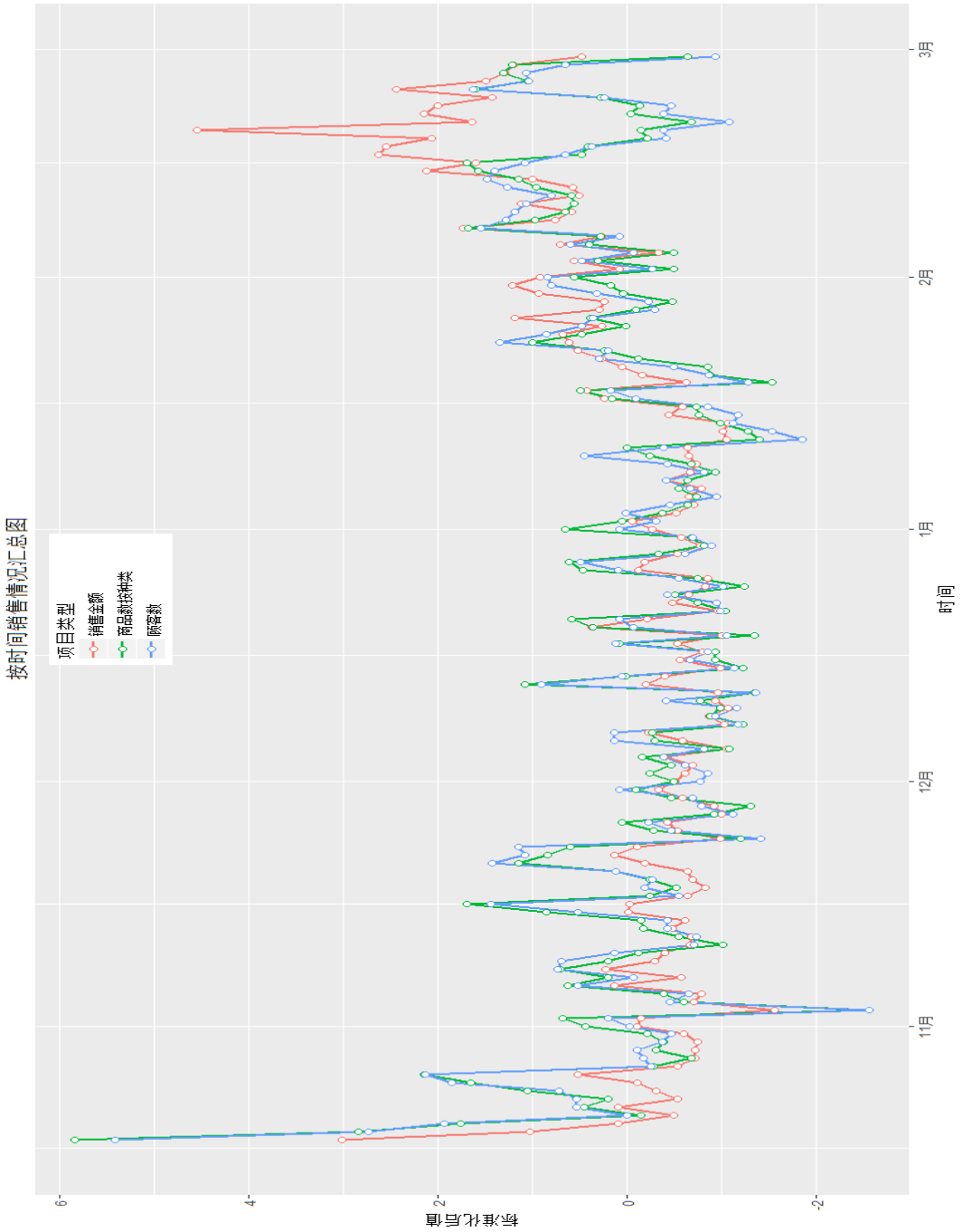
首先根据销售时间对 valdata 中数据进行行金额和数量汇总。 这里使用了 dplyr 包中的 group_by()，summarise()，和管道符%>%，分别进行分组，汇总和结果传递。其中每一订单号一般为一个独立用户的购买，所以这里用不同的订单号计算顾客数。

```
smbt<-tapply(valdata$销售金额，valdata$销售时间，sum)
snbt<-tapply(valdata$单据号，valdata$销售时间，length)
sbt<-rbind(smbt，snbt)%>%t()%>%as.data.frame()
names(sbt)<-c("销售金额"，"商品数按种类")
sbt$时间<-rownames(sbt)
rownames(sbt)<-NULL
sbt$时间<-strptime(sbt$时间，"%Y-%m-%d")
spbt<-valdata%>%
  group_by(销售时间)%>%
  summarise(订单数=n_distinct(单据号))
names(spbt)<-c("时间"，"顾客数")
sbt<-cbind(sbt，spbt[，2])
str(sbt)

（略）.
```

使用 **ggplot2** 作图

```
head(sbt，5)

##   销售金额 商品数按种类       时间 顾客数
## 1  9030.78         985 2014-10-18    505
## 2  5669.19         699 2014-10-19    372
## 3  4103.67         597 2014-10-20    332
## 4  3105.26         415 2014-10-21    236
## 5  4095.77         473 2014-10-22    263

sbt[，c(1，2，4)]<-scale(sbt[，c(1，2，4)])
sbt$时间<-as.Date(sbt$时间)
md<-melt(sbt，id.vars=c("时间")，
         measure.vars=c("销售金额"，  "商品数按种类"，  "顾客数")，
         variable.name="项目类型"，  value.name="值")
x<-md$时间
y<-md$值
z<-md$项目类型
ggplot(md，aes(x=x，y=y，color=z，group=z))+
```

```
geom_line(size=1)+
geom_point(size=2, shape=21, fill="white")+
labs(y="标准化后值", x="时间", color="项目类型")+
theme(legend.position=c(0.5, 1), legend.justification=c(0.5, 1))+
ggtitle("按时间销售情况汇总图")
```



从该图可以发现：

1.2 月 18 日明显的销售额和顾客数背离，猜测该日为节日，采购年货导致客单价提高，后续可以继续验证

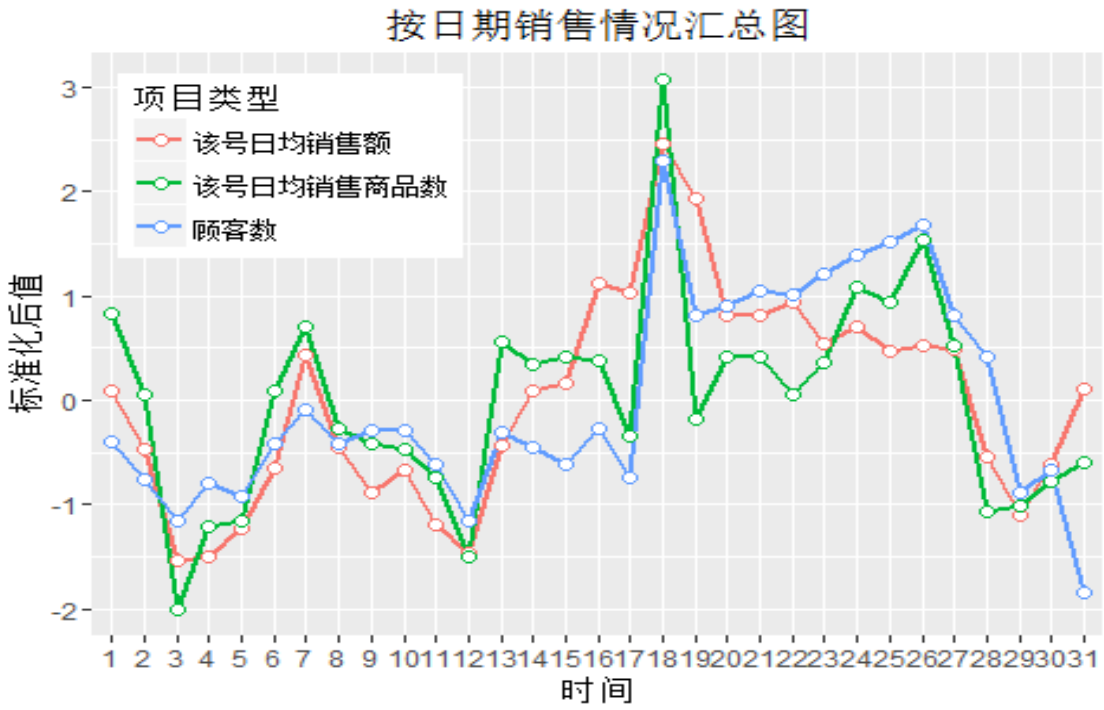
2.每日销售具有波浪式周期性，周期为一天

3.2014-11-01 销售明显下滑，猜测该日为双十一活动日，购买显著下降
4.10 月-2 月销售整体上呈现先下降后上升的特征，猜测和气温有关

1.2(销售金额、商品数按种类、顾客数)-(1 日-31 日)关系

```
smbd<-tapply(valdata$销售金额，valdata$byday，sum)
snbd<-tapply(valdata$单据号，valdata$byday，length)
sbd<-valdata%>%
  select(销售时间，byday)%>%
  unique()%>%
  .$byday%>%
  table()%>%
  rbind(smbd，snbd，.)%>%
  t()%>%
  as.data.frame()
names(sbd)<-c("销售金额"，"商品数按种类"，"频数")
sbd<-sbd%>%
  mutate(该号日均销售额=销售金额/频数，该号日均销售商品数=商品数按种类/频
数)%>%
  cbind(时间=row.names(.)，.)
row.names(sbd)=NULL
spbd<-valdata%>%
  group_by(byday)%>%
  summarise(订单数=n_distinct(单据号))
names(spbd)<-c("时间"，"顾客数")
sbd<-cbind(sbd，spbd[，2])
sbd$时间<-factor(sbd$时间，order=TRUE，levels = c(1:31))
str(sbd)

## 'data.frame':    31 obs. of  7 variables:
##  $ 时间             : Ord.factor w/ 31 levels "1"<"2"<"3"<"4"<..: 1
 2 3 4 5 6 7 8 9 10 ...
##  $ 销售金额         : num  15833 14534 12057 12152 12760 ...
##  $ 商品数按种类     : num  1828 1717 1424 1537 1545 ...
##  $ 频数             : num  4 4 4 4 4 4 4 4 4 4 ...
##  $ 该号日均销售额   : num  3958 3633 3014 3038 3190 ...
##  $ 该号日均销售商品数: num  457 429 356 384 386 ...
##  $ 顾客数           : int  951 884 812 878 855 948 1005 947 971 970
 ...

head(sbd，5)

sbd[，5:7]<-scale(sbd[，5:7])
md<-melt(sbd，  id.vars=c("时间"),
         measure.vars=c("该号日均销售额"，  "该号日均销售商品数"，  "顾客数
"),
```

```
        variable.name="项目类型",  value.name="值")
x<-md$时间
y<-md$值
z<-md$项目类型
ggplot(md, aes(x=x, y=y, color=z, group=z))+
  geom_line(size=1)+
  geom_point(size=2, shape=21, fill="white")+
  labs(y="标准化后值", x="时间", color="项目类型")+
  theme(legend.position=c(0, 1), legend.justification=c(0, 1))+
  ggtitle("按日期销售情况汇总图")
```



从该图可以发现：

1.18 号销售额远好于其它日期，猜测为特定节日，查看日历验证为 2015 年除夕，为采购年货高峰

2.日均销售额、商品数、顾客数有波浪式周期性，周期类似

3.每月月中销售普遍好于月初和月末，猜测可能人月初发工资，因而有月中高于月初和月末的现象

4.月中客单价明显高于月中和月末，猜测同上，刚发工资购买力较高

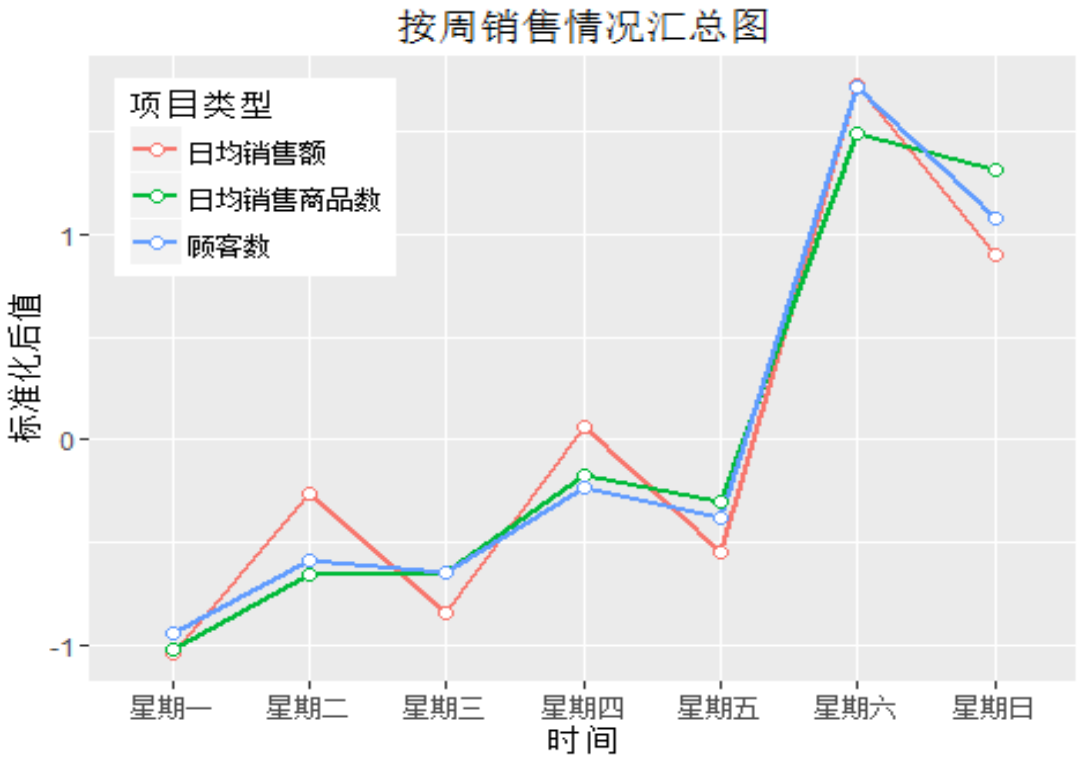
1.3(销售金额、商品数按种类、顾客数)-(周一-周日)关系

```
smbw<-tapply(valdata$销售金额, valdata$byweek, sum)
snbw<-tapply(valdata$单据号, valdata$byweek, length)
sbw<-valdata%>%
  select(销售时间, byweek)%>%
  unique()%>%
  .$byweek%>%
```

```
  table()%>%
  rbind(smbw, snbw, .)%>%
  t()%>%
  as.data.frame()
names(sbw)<-c("销售金额", "商品数按种类", "频数")
sbw$时间<-row.names(sbw)
row.names(sbw)=NULL
sbw<-mutate(sbw, 日均销售额=销售金额/频数, 日均销售商品数=商品数按种类/频数)
spbw<-valdata%>%
  group_by(byweek)%>%
  summarise(订单数=n_distinct(单据号))
names(spbw)<-c("时间", "顾客数")
sbw<-cbind(sbw, spbw[, 2])
sbw$时间<-factor(sbw$时间, order=TRUE, levels = c("星期一", "星期二", "星期三", "星期四", "星期五", "星期六", "星期日"))

head(sbw, 5)
```

```
##   销售金额 商品数按种类 频数   时间 日均销售额 日均销售商品数 顾客数
## 1 65838.90         7233   19 星期一   3465.205       380.6842   4033
## 2 72619.70         7559   19 星期二   3822.089       397.8421   4218
## 3 67517.55         7568   19 星期三   3553.555       398.3158   4186
## 4 75441.84         7991   19 星期四   3970.623       420.5789   4403
## 5 70081.75         7878   19 星期五   3688.513       414.6316   4325
```

```
sbw[, 5:7]<-scale(sbw[, 5:7])
md<-melt(sbw,  id.vars=c("时间"),
         measure.vars=c("日均销售额",  "日均销售商品数",  "顾客数"),
         variable.name="项目类型",  value.name="值")
x<-md$时间
y<-md$值
z<-md$项目类型
ggplot(md, aes(x=x, y=y, color=z, group=z))+
  geom_line(size=1)+
  geom_point(size=2, shape=21, fill="white")+
  labs(y="标准化后值", x="时间", color="项目类型")+
  theme(legend.position=c(0, 1), legend.justification=c(0, 1))+
  ggtitle("按周销售情况汇总图")
```



从该图可以发现：

1.日均销售额、商品数和顾客数具有波浪式周期性，周期为 1 天

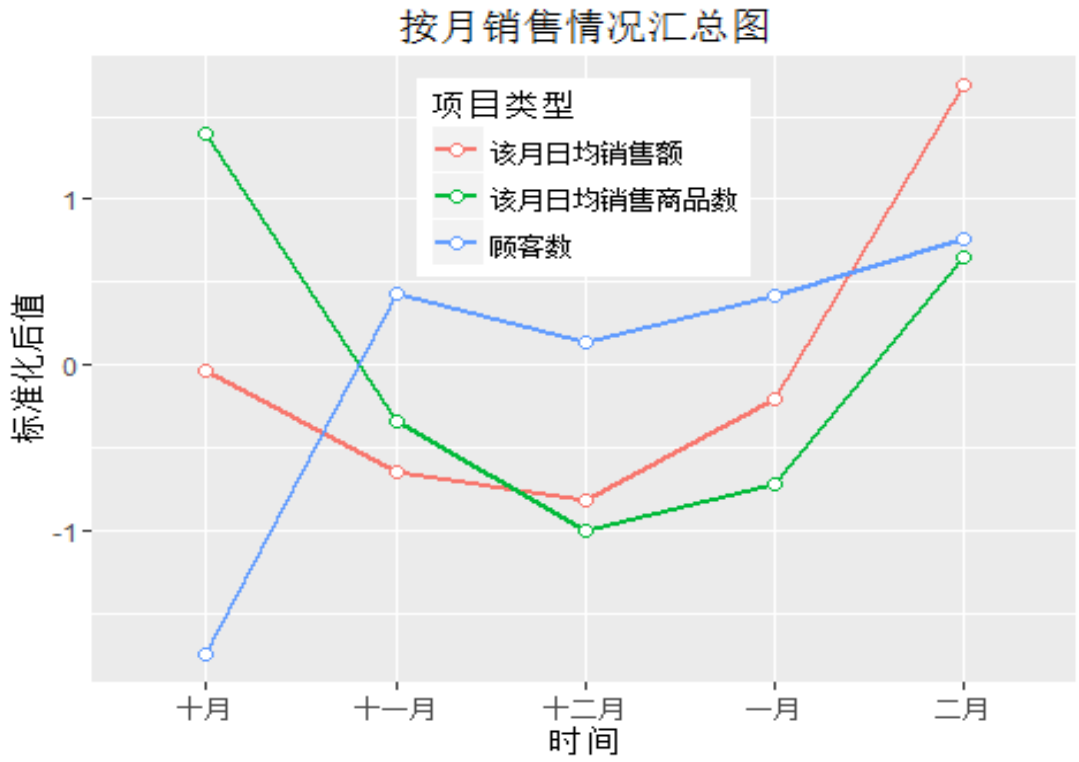
2.周末的销售情况远好于工作日

3.销售随着周一到周五递升，周六达到高峰

1.4(销售金额、商品数按种类、顾客数)-(10 月-2 月)关系

```
smbm<-tapply(valdata$销售金额，valdata$bymonths，sum)
snbm<-tapply(valdata$单据号，valdata$bymonths，length)
sbm<-valdata%>%
  select(销售时间，bymonths)%>%
  unique()%>%
  .$bymonths%>%
  table()%>%
  rbind(smbm，snbm，.)%>%
  t()%>%
  as.data.frame()
names(sbm)<-c("销售金额"，"商品数按种类"，"频数")
sbm$时间<-row.names(sbm)
row.names(sbm)=NULL
sbm<-mutate(sbm，该月日均销售额=销售金额/频数，该月日均销售商品数=商品数
按种类/频数)
spbm<-valdata%>%
  group_by(bymonths)%>%
  summarise(订单数=n_distinct(单据号))
names(spbm)<-c("时间"，"顾客数")
```

```
sbm<-cbind(sbm, spbm[, 2])
sbm$时间<-factor(sbm$时间, order=TRUE, levels = c("十月", "十一月", "十二月", "一月", "二月"))
head(sbm, 5)

sbm[, 5:7]<-scale(sbm[, 5:7])
md<-melt(sbm,  id.vars=c("时间"),
         measure.vars=c("该月日均销售额",  "该月日均销售商品数",  "顾客数"),
         variable.name="项目类型",  value.name="值")
x<-md$时间
y<-md$值
z<-md$项目类型
ggplot(md, aes(x=x, y=y, color=z, group=z))+
  geom_line(size=1)+
  geom_point(size=2, shape=21, fill="white")+
  labs(y="标准化后值", x="时间", color="项目类型")+
  theme(legend.position=c(0.5, 1), legend.justification=c(0.5, 1))+
  ggtitle("按月销售情况汇总图")
```



从该图可以发现：

1.该月日均销售额、该月日均销售数、顾客数不同月销售差异较大

2.日均销售额和日均商品数在十二月达到低点回升，猜测可能随温度降低，人们能量消耗降低，去超市频次不变而单次购买变少。

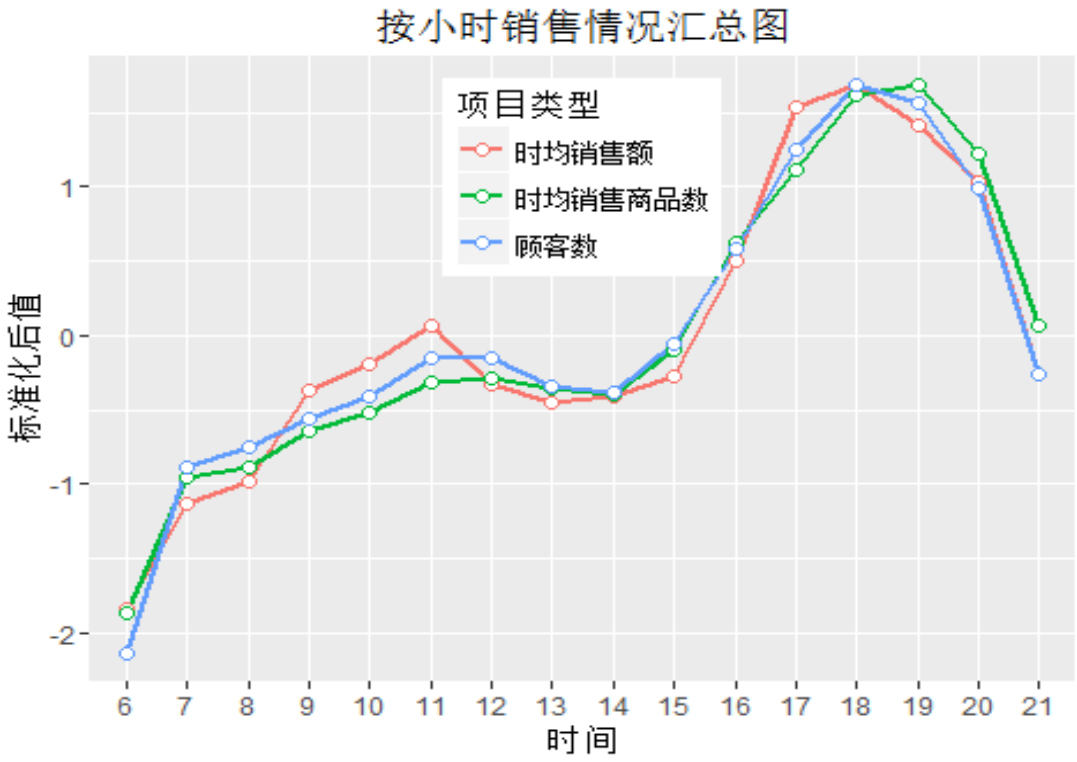
3.一月二月虽然气温依然很低却销售回升可能和节日有关

4.二月销售额显著高于其他月份，猜测节日采购年货有关

5.二月顾客数没有随销售额显著增加，猜测该超市为社区型中小超市，消费群体稳定

1.5(销售金额、商品数按种类、顾客数)-(6 时-21 时)关系

```
smbh<-tapply(valdata$销售金额，valdata$byhours，sum)
snbh<-tapply(valdata$单据号，valdata$byhours，length)
sbh<-valdata%>%
  select(销售时间，byhours)%>%
  unique()%>%
  .$byhours%>%
  table()%>%
  rbind(smbh，snbh，.)%>%
  t()%>%
  as.data.frame()
names(sbh)<-c("销售金额"，"商品数按种类"，"频数")
sbh$时间<-row.names(sbh)
row.names(sbh)=NULL
sbh<-mutate(sbh，时均销售额=销售金额/频数，时均销售商品数=商品数按种类/频数)
spbh<-valdata%>%
  group_by(byhours)%>%
  summarise(订单数=n_distinct(单据号))
names(spbh)<-c("时间"，"顾客数")
sbh<-cbind(sbh，spbh[，2])
sbh$时间<-factor(sbh$时间，order=TRUE，levels = c(6:21))

head(sbh，5)

sbh[，5:7]<-scale(sbh[，5:7])
md<-melt(sbh， id.vars=c("时间")，
         measure.vars=c("时均销售额"， "时均销售商品数"， "顾客数")，
         variable.name="项目类型"， value.name="值")
x<-md$时间
y<-md$值
z<-md$项目类型
ggplot(md，aes(x=x，y=y，color=z，group=z))+
  geom_line(size=1)+
  geom_point(size=2，shape=21，fill="white")+
  labs(y="标准化后值"，x="时间"，color="项目类型")+
  theme(legend.position=c(0.5，1)，legend.justification=c(0.5，1))+
  ggtitle("按小时销售情况汇总图")
```



从该图可以发现：

1.销售时间从早上 6 点到 21 点

2.时均销售额、时均商品数、顾客数在不同时段间没有明显差异

3.11-15 点时间段，为销售低点；17-20 点为销售高点。

2.时间类型数据与商品信息数据结合

统计商品销售信息

整体产品销售量情况，包括销量前 10 的产品名称，及不同销售量的产品分布情况

```
snbp<-valdata%>%
  group_by(商品名称)%>%
  summarise(销售量=n())%>%
  arrange(desc(销售量))
head(snbp，10)
```

```
## Source: local data frame [10 x 2]
##
##               商品名称 销售量
##                 (fctr)  (int)
## 1                 烤肠   2352
## 2  华润苏果中号购物袋   1806
## 3           怡宝纯净水    694
## 4            五香鸡蛋    594
```

```
## 5          可口可乐汽水     530
## 6                甜不辣     472
## 7     苏烟(五星红杉树)     423
## 8              百事可乐     395
## 9     五香茶叶蛋(小吃)     373
## 10 华润苏果大号购物袋     321
```

```
str(snbp)（略）
```

可以看到在所有时间内，销量前十的产品为:烤肠，中号购物袋，怡宝纯净水等。通过 head(snbp，n)可以继续查看销量前 n 的产品，发现热销产品主要以小额的食品饮料为主，符合社区小型超市的销售特征，商品数较多可以通过词云展示。

词云展示热销产品

通过 devtools 安装 github 上的包 wordcloud2，方式为：

```
devtools::install_github("lchiffon/wordcloud2")
library(wordcloud2)
wordcloud2(data = snbp)
```

得到下图：

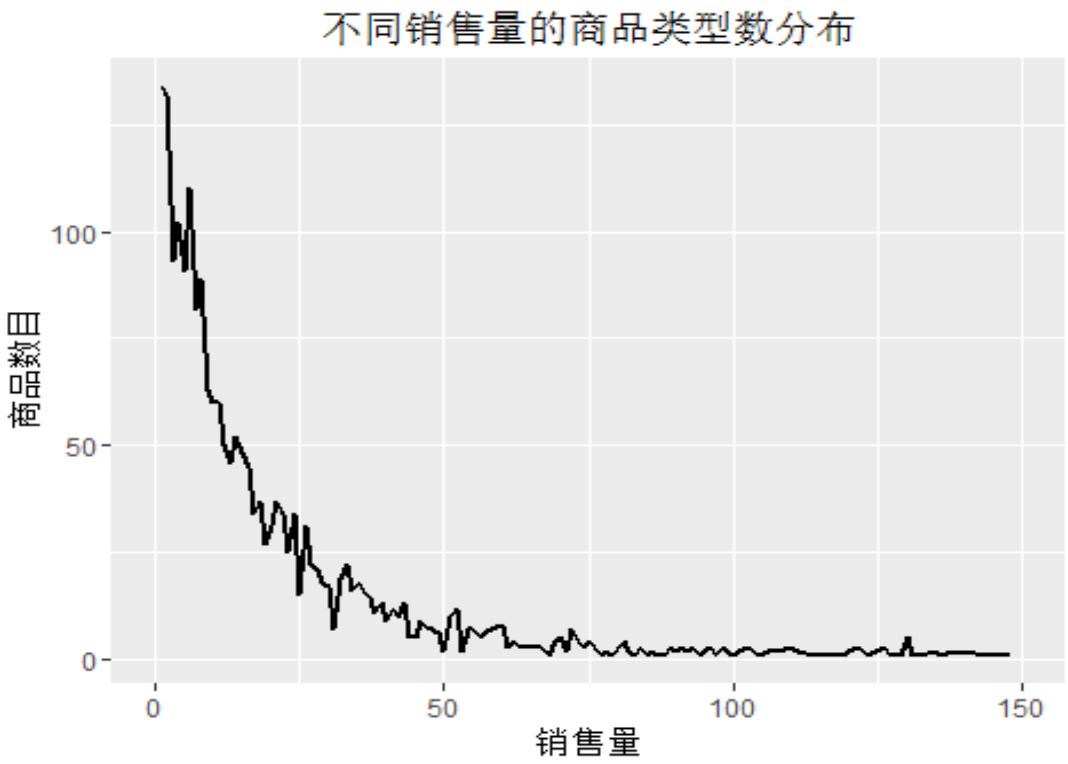
从该图可以发现，产品之间销量差异较大，绝大大部分产品销量较小，可能具有传统的长尾效应或幂率分布的特征。

由于产品总数有 2 千多种，不方便分析所有产品的销售与时间的关系。所以方便起见，首先选取热销的 50 件商品为分析样本，按时间类型分别分析热销 50 件商品与销售时间/日期/月/周/时的销售。 **选取总销售量前 50 的产品**从总销量排名数据 snbp 中选择销量前 50 的产品名称为数据 hsp50(hot sell product 50)，方便后续通过 filter 匹配产品名称选取特定分析样本。

```
hsp50<-snbp[1:50, 1]
hsp50<-hsp50$商品名称
```

不同销售量的商品类型数分布

```
snbp<-as.data.frame(snbp)
snd<-snbp%>%
  group_by(snbp[, 2])%>%
  summarise(商品数=n_distinct(商品名称))
names(snd)<-c("销售量", "商品数")

ggplot(snd, aes(snd$销售量, snd$商品数))+
  geom_line(size=1)+
  xlim(0, 150)+
  labs(y="商品数目", x="销售量")+
  ggtitle("不同销售量的商品类型数分布")

## Warning: Removed 40 rows containing missing values (geom_path).
```



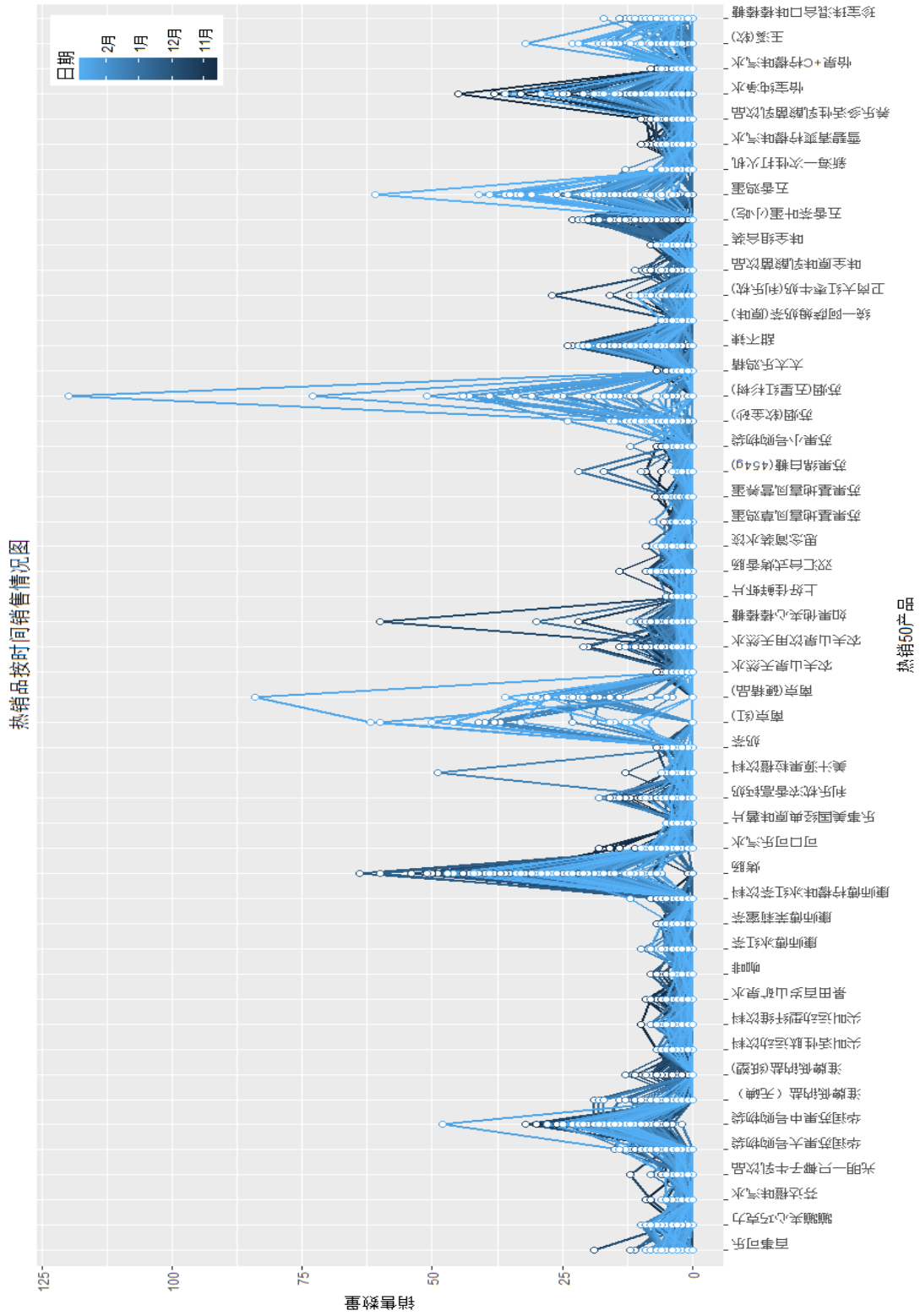
从该图可以发现，不同销售量的商品类型数符合典型的幂率分布或者说长尾效应，大部分产品的销售量都小于 50。这里除了受现实中的长尾效应影响，也和产品能分类过于详细和数据量有关，本次数据有 2115(按商品名称)种商品，31682 条交易数据，会加剧长尾效应。

2.1 热销产品-(2014-10-18 至 2015-02-28)关系

这里使用了 dplyr 包中的 select()和 reshape2 包中的 melt()和 cast()函数。Melt()可以将宽型 wide 数据整合成长型 long 数据，而 cast()则可以将长型数据整合成宽型数据。

数据：hsbt(hot sell by time)

```
hsbt<-valdata%>%
  select(商品名称，销售数量，销售时间)%>%
  dcast(销售时间~商品名称，value.var="销售数量"，sum)%>%
  as.data.frame()
md<-hsbt%>%
  melt(id.vars=c("销售时间")，variable.name="商品名称"， value.name="
销售数量")%>%
  filter(商品名称 %in% hsp50)
x<-md$商品名称
y<-md$销售数量
z<-md$销售时间
ggplot(md，aes(x=x，y=y，color=z，group=z))+
  geom_line(size=1)+
  geom_point(size=2，shape=21，fill="white")+
  labs(y="销售数量"，x="热销 50 产品"，color="日期")+
  theme(legend.position=c(1，1)，legend.justification=c(1，1))+
  theme(axis.text.x=element_text(angle=90，hjust=1))+
  ggtitle("热销品按时间销售情况图")

hsbt[1:5，1:3]

##      销售时间 100 支家庭装吸管 2 粒南孚 5#碱电池 15A2B
## 1 2014-10-18              0                  1
## 2 2014-10-19              0                  1
## 3 2014-10-20              0                  0
## 4 2014-10-21              0                  2
## 5 2014-10-22              0                  1
```



从该图有以下发现：

1.大部分热销产品不同时间的销售情况没有显著差异

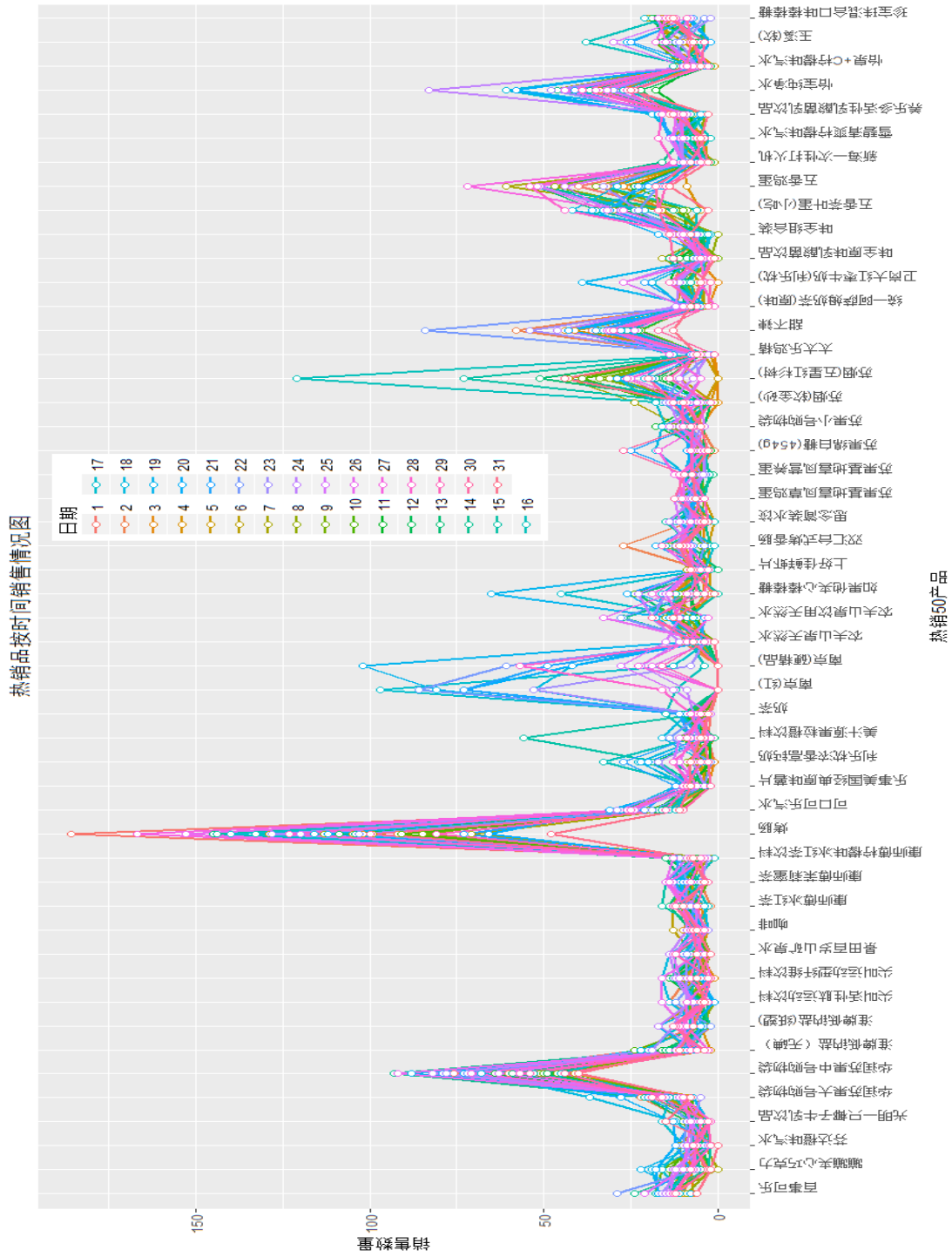
2.运动型饮料、可口可乐、纯净水等饮品在 10-12 月份销售好于其它月份，猜测和气温降低，人们不愿消费常温饮料有关

3.烟草类产品在 1-2 月销售明显好于其它月份，猜测一是和节日习俗有关，也可能是企业单位年终考核总结、个体户企业年终收账在这几个月，普遍工作量和压力较大有关。

2.2 热销产品-(1 日-31 日)关系

数据:hsbd(hot sell by day)。数据结构同上。

```
hsbd<-valdata%>%
  select(商品名称，销售数量，byday)%>%
  dcast(byday~商品名称，value.var="销售数量"，sum)%>%
  as.data.frame()
md<-hsbd%>%
  melt(id.vars=c("byday")，variable.name="商品名称"， value.name="销售数量")%>%
  filter(商品名称 %in% hsp50)
x<-md$商品名称
y<-md$销售数量
z<-md$byday
ggplot(md，aes(x=x，y=y，color=z，group=z))+
  geom_line(size=1)+
  geom_point(size=2，shape=21，fill="white")+
  labs(y="销售数量"，x="热销 50 产品"，color="日期")+
  theme(legend.position=c(0.6，1)，legend.justification=c(0.6，1))+
  theme(axis.text.x=element_text(angle=90，hjust=1))+
  ggtitle("热销品按时间销售情况图")
```



从该图有以下发现：

1.大部分热销品在不同日期的销售最大最小值差异显著，与按周分布对比可以明显发现。

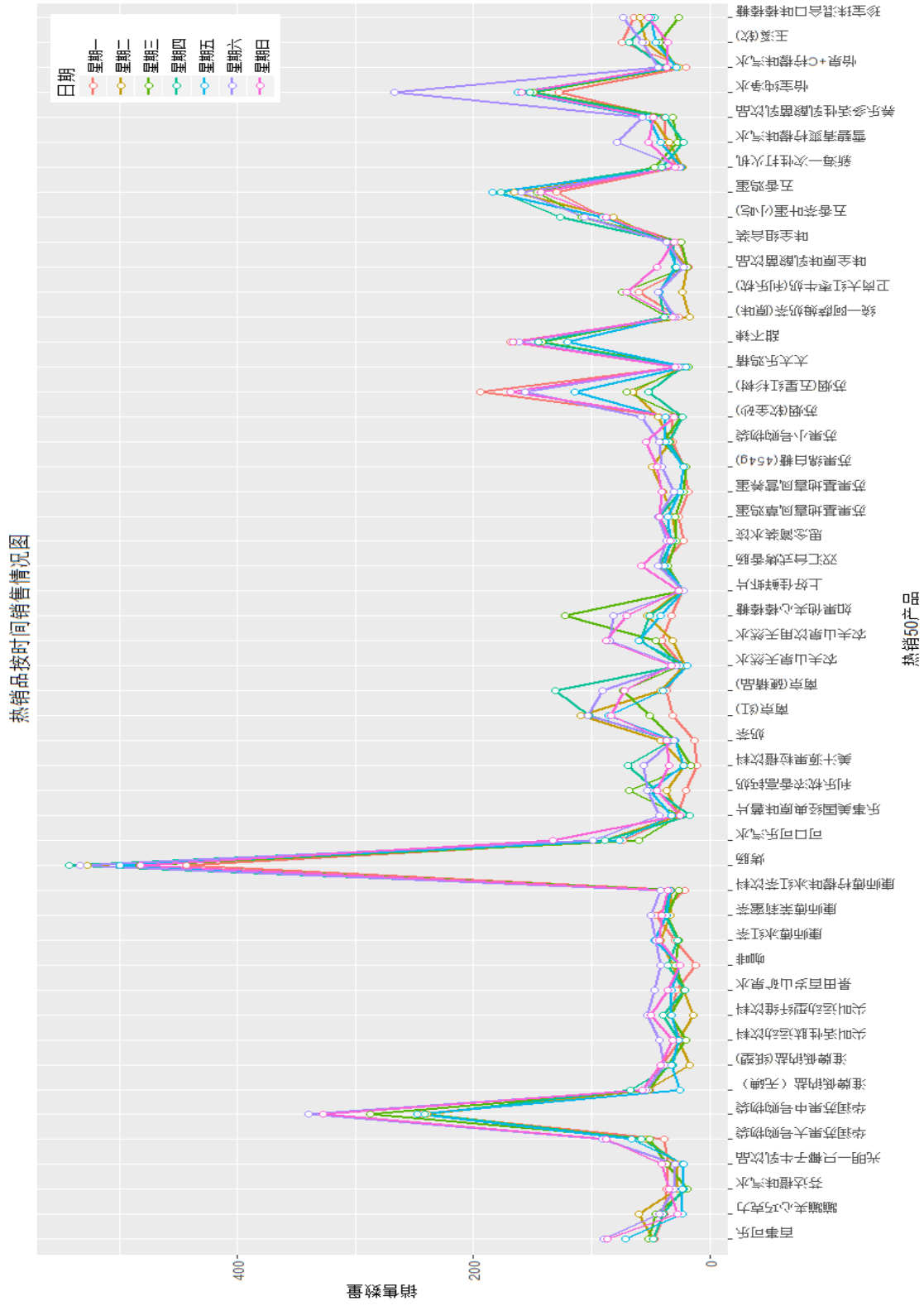
2.很多热销品在 17-21 号的销售较多，结合月份销售情况可以猜测 2015 年的春节在这附近。由于总体数据量较少，节日相关产品购买影响了本日期的销售总数，查询日历得到 2015 春节为 2 月 19，也与数据相符。

3.在 17-21 号左右的节日采购时间段内，可以发现，美汁源果粒橙、南京系列烟、五星红杉树、卫岗大红枣利乐枕、玉溪等销售明显较多，是大家常常备的年货。

2.3 热销产品-(周一-周日)关系

数据：hsbw(hot sell by week)

```
hsbw<-valdata%>%
  select(商品名称，销售数量，byweek)%>%
  dcast(byweek~商品名称，value.var="销售数量"，sum)%>%
  as.data.frame()
md<-hsbw%>%
  melt(id.vars=c("byweek")，variable.name="商品名称"， value.name="销售数量")%>%
  filter(商品名称 %in% hsp50)
x<-md$商品名称
y<-md$销售数量
z<-md$byweek

ggplot(md，aes(x=x，y=y，color=z，group=z))+
  geom_line(size=1)+
  geom_point(size=2，shape=21，fill="white")+
  labs(y="销售数量"，x="热销 50 产品"，color="日期")+
  theme(legend.position=c(1，1)，legend.justification=c(1，1))+
  theme(axis.text.x=element_text(angle=90，hjust=1))+
  ggtitle("热销品按时间销售情况图")
```



从该图有以下发现：

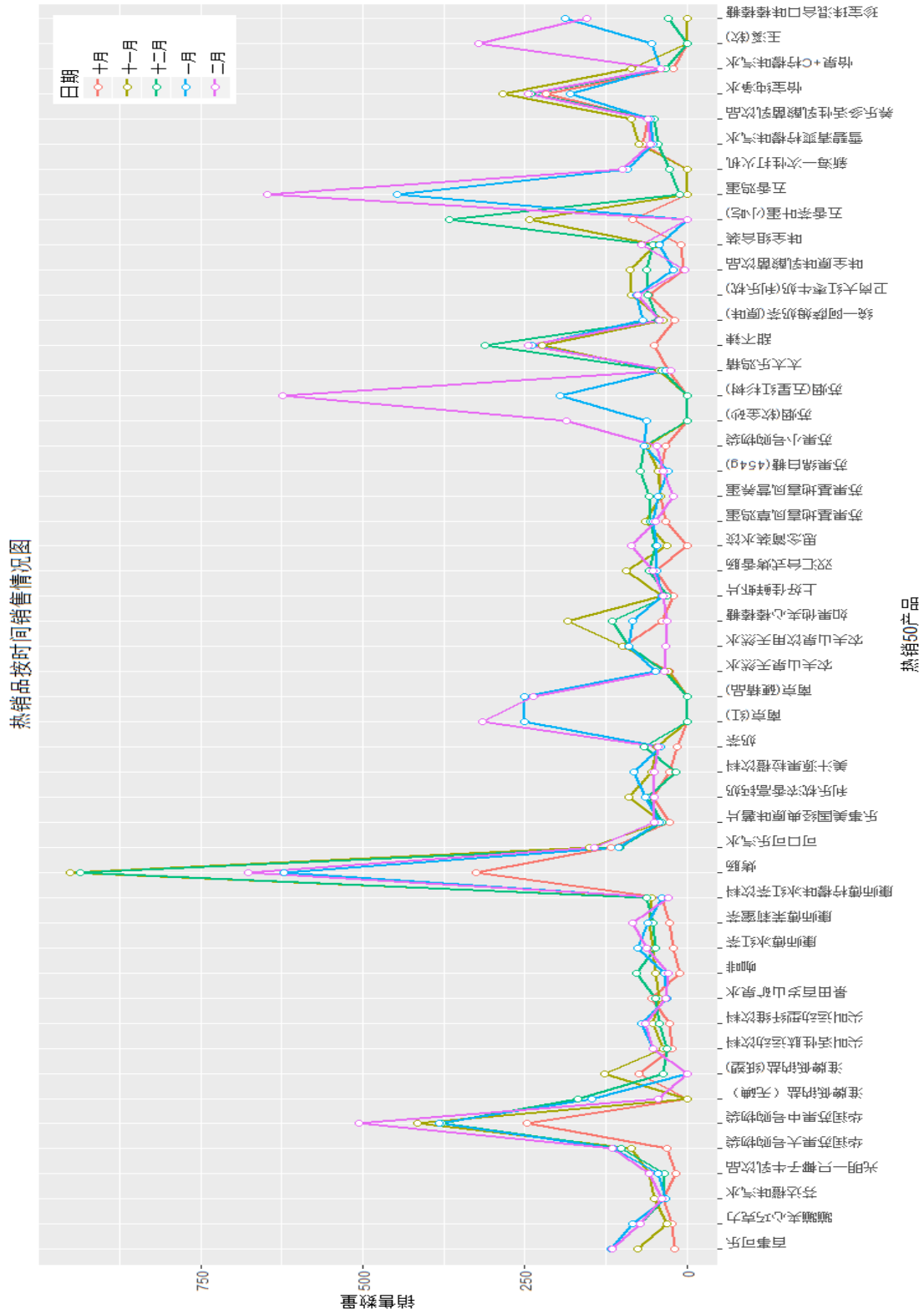
1.大部分热销产品在不同周几的销售情况没有显著差别

2.如果他夹心棒棒糖在周三的销售教多，南京(硬精品)在周五销售较多，怡宝纯净水在周六销售显著增加，原因不明。

3.除南京(红)外，其它烟草制品普遍周末的销售好于工作日

2.4 热销产品-(10 月-2 月)关系

```
hsbm<-valdata%>%
  select(商品名称，销售数量，bymonths)%>%
  dcast(bymonths~商品名称，value.var="销售数量"，sum)%>%
  as.data.frame()

md<-hsbm%>%
  melt(id.vars=c("bymonths")，variable.name="商品名称"，  value.name="
销售数量")%>%
  filter(商品名称 %in% hsp50)

x<-md$商品名称
y<-md$销售数量
z<-md$bymonths

ggplot(md，aes(x=x，y=y，color=z，group=z))+
  geom_line(size=1)+
  geom_point(size=2，shape=21，fill="white")+
  labs(y="销售数量"，x="热销 50 产品"，color="日期")+
  theme(legend.position=c(1，1)，legend.justification=c(1，1))+
  theme(axis.text.x=element_text(angle=90，hjust=1))+
  ggtitle("热销品按时间销售情况图")
```



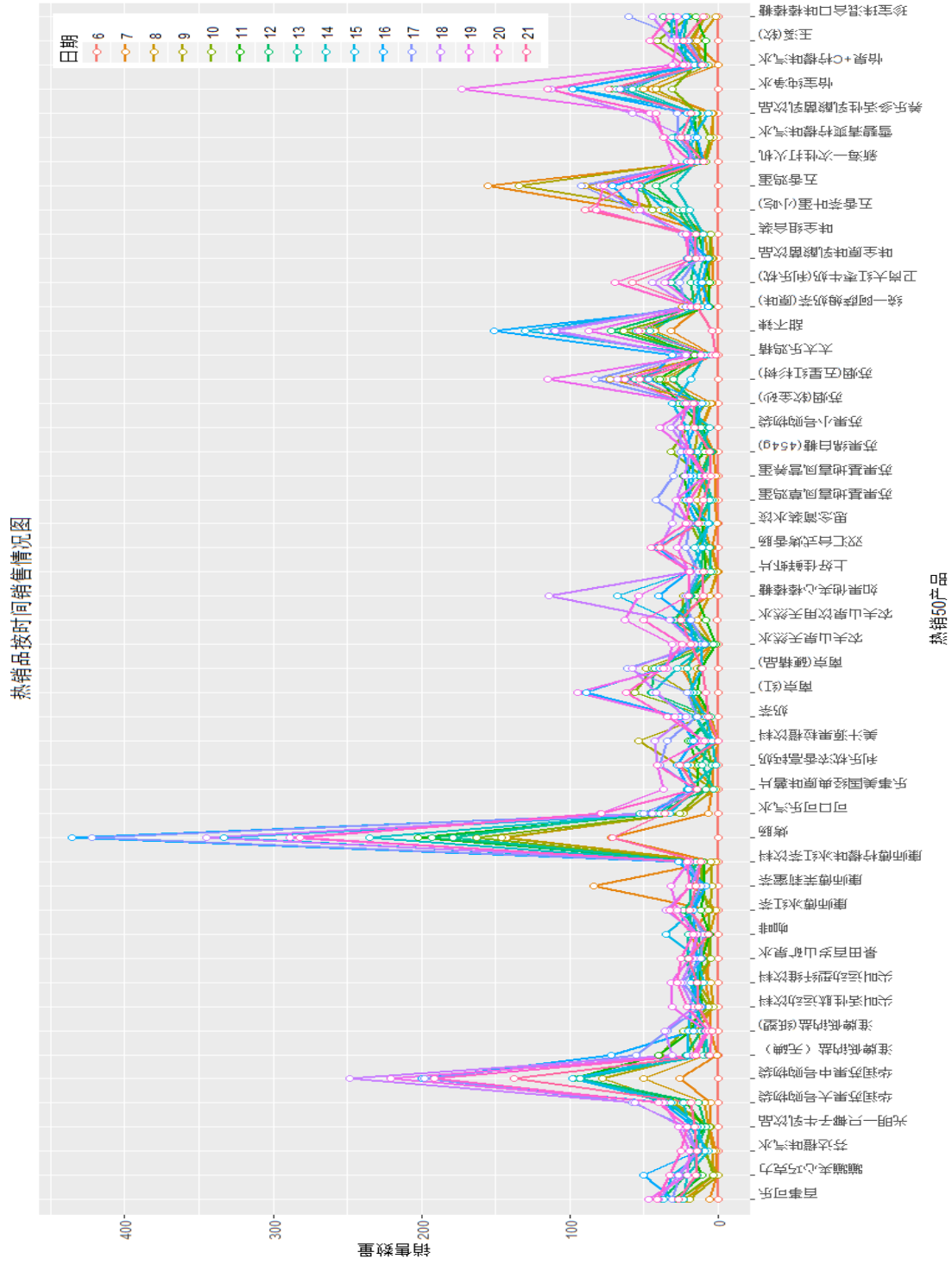
从该图中有以下发现：

1.南京(红)，苏烟，红杉树，玉溪等烟类产品在一二月份中销售最多，猜测这两月有传统节日春节，所以烟草销售好。而在其它月份销售很差，表明该类产品销售有明显周期性，受节日影响较大。

2.烤肠在十一月到二月的销售最多，猜测天气变冷，很多人更喜欢吃热的烤肠
3.运动型饮料，茶类饮料，果汁类饮料，每个月的销售情况接近，而且销售总量接近，表明不同饮料类型偏好的消费者的比例可能类似，而且市场比较成熟和稳定了
4.苏果小号购物袋销售比较稳定，中号和大号购物袋在一二月份明显增多，猜测随着天气变冷，人们为了不冻手及携带方便，更愿意购买购物袋
5.如果他夹心棒棒糖在十一月的销售远好于其它月，猜测该月有光棍节，很多单身女同学会购买棒棒糖作为朋友或自己的节日礼物

2.5 热销产品-(6 时-21 时)关系

```
hsbh<-valdata%>%
  select(商品名称，销售数量，byhours)%>%
  dcast(byhours~商品名称，value.var="销售数量"，sum)%>%
  as.data.frame()
md<-hsbh%>%
  melt(id.vars=c("byhours")，variable.name="商品名称"， value.name="销售数量")%>%
  filter(商品名称 %in% hsp50)
x<-md$商品名称
y<-md$销售数量
z<-md$byhours
ggplot(md，aes(x=x，y=y，color=z，group=z))+
  geom_line(size=1)+
  geom_point(size=2，shape=21，fill="white")+
  labs(y="销售数量"，x="热销 50 产品"，color="日期")+
  theme(legend.position=c(1，1)，legend.justification=c(1，1))+
  theme(axis.text.x=element_text(angle=90，hjust=1))+
  ggtitle("热销品按时间销售情况图")
```



从该图中有以下发现：

1.大部分热销品在不同时段的销售情况没有显著差别

2.购物袋和怡宝纯净水分布类似，猜测为纯净水较重，买了纯净水同时一般会买塑料袋，所以相关性较强。

3.早上八九点左右康师傅茉莉蜜茶，烤肠和五香鸡蛋销售最多，猜测这个时间段为上班族的早餐时间，这三者可能是很多上班族的早餐选择。

4.下午三四点烤肠和甜不辣销售最多，猜测这个超市可能在中小学附近，下午三四点是大部分中小学休息时间，而且这两类商品应该以学生消费为主。

5.如果他夹心棒棒糖下午六点时销售最好，加强了上一条的猜想，猜测放学时间，很多女学生会选择买个棒棒糖回家。

3.关联规则挖掘

载入关联规则分析包 arules 及关联规则可视化分析包 arulesViz。 将(单据号，商品名称)single 数据类型转换成 arules 包可以识别的 transaction 数据格式。 删除无效的购物袋类产品。

```
library(arules)

library(arulesViz)

transdata<-valdata%>%
  select(单据号，商品名称)%>%
  .[!duplicated(.)，]
transdata<-filter(transdata，!grepl("购物袋"，商品名称))
transdata$商品名称<-as.character(transdata$商品名称)
#只有转成 char 才可以重新 items 显示为字符型，因为过滤掉 4 个商品名，但原来是 factor 编码未变
transrule<-as(split(transdata[，2]， transdata[，1])， "transactions
")
```

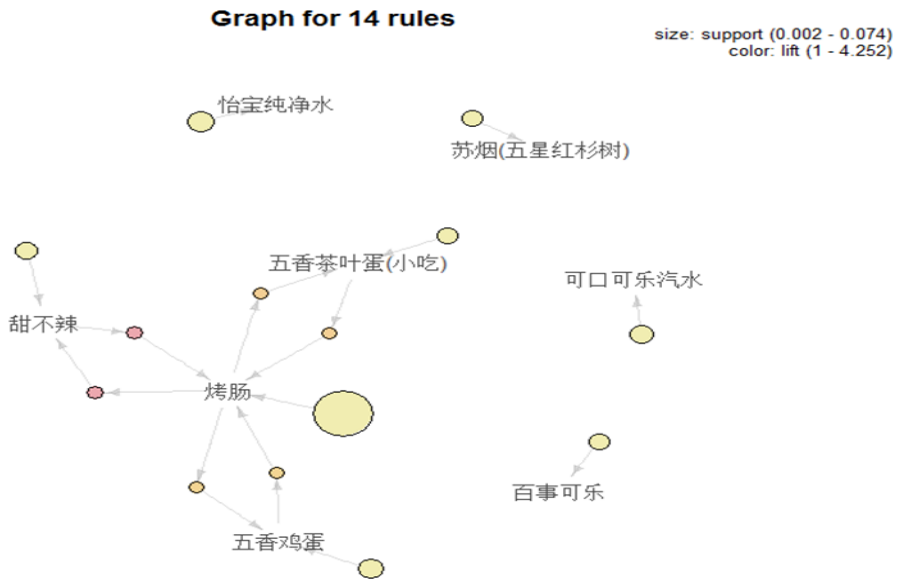
查看前 10 条交易，通过 apriori 算法计算关联规则，并按照 Lift 降序排列查看规则。

```
inspect(transrule[1:10])

##    items

## 1  {绿箭口香糖}

## 2  {百乐宝奶昔 巧克力口味雪糕}

## 3  {多力葵花籽油}

## 4  {海天酱油(金标生抽王)}

## 5  {雀巢丝滑拿铁咖啡饮料}

## 6  {卫岗原味低脂酸奶，五香茶叶蛋(小吃)}

## 7  {白猫洗洁精，雕牌加香透明皂，恒顺白醋，洁云福瑞 200 抽塑包面纸-单包装（大幅）}
## 8  {金龙鱼黄金比例食用调和油}

## 9  {奥妙净蓝全效水清莲香洗衣粉}
```

```
## 10 {海飞丝丝质柔滑型去屑洗发露 0025，红石洁厕灵}      （ID 略）

rules<-apriori(transrule, parameter=list(support=0.001, confidence=0.
01))

inspect(head(sort(rules,  by = "lift"),  10))

##    lhs                    rhs                 support     confidence
## 11 {甜不辣}           => {烤肠}              0.004702986 0.31567797
## 12 {烤肠}             => {甜不辣}            0.004702986 0.06335034
## 9  {五香茶叶蛋(小吃)} => {烤肠}              0.002146329 0.18230563
## 10 {烤肠}             => {五香茶叶蛋(小吃)} 0.002146329 0.02891156
## 13 {五香鸡蛋}         => {烤肠}              0.003030112 0.16161616
## 14 {烤肠}             => {五香鸡蛋}          0.003030112 0.04081633
## 1  {}                 => {苏烟(五星红杉树)} 0.013351430 0.01335143
## 2  {}                 => {百事可乐}          0.012467647 0.01246765
## 3  {}                 => {五香茶叶蛋(小吃)} 0.011773247 0.01177325
## 4  {}                 => {可口可乐汽水}      0.016602487 0.01660249
##    lift
## 11 4.252257
## 12 4.252257
## 9  2.455700
## 10 2.455700
## 13 2.177008
## 14 2.177008
## 1  1.000000
## 2  1.000000
## 3  1.000000
## 4  1.000000
```

反复实验，发现当 supp 和 conf 均取很小时才有合适的规则，表明数据关联性不强。这可能是因为数据据量少和商品命名过于详细有关，需要对商品名称进行一定集中后可能更有利于关联规则挖掘。由于汇总较为复杂，这里不进行过多无效的关联规则挖掘了。仅对 support=0.001 和 confidence=0.01 时做些可视化分析。 作出关联规则可视化图：

```
subrules<-head(sort(rules,  by = "lift"),  10)
plot(rules, method = "graph")
```



该图呈现的信息与上表相同。可以从箭头关系比较清晰的发现热销商品及对应提升关系。

R 经验总结:

1.dcast 返回的 data.fame 格式的数据，会自动为 cast 的变量“byweek”增加了新变量名称 v1-v7
2.reshape2-dplyr 包中一些函数反复运用可以解决绝大部分数据整理问题
3.折线图，x 为因子型变量要增加参数 group=1，使得系统识别
4. data.frame 格式会默认整个数据是同一种类型，如果因为转置的原因导致另一种类型数据混入，就可能会改变所有数据的类型
5. knit,Rmakdown:文件读取路径和 R 的工作路径不同，所以要使用绝对路径

待优化：

1.多次复用的代码函数化，做成超市数据分析包，结合 shiny 做成超市数据在线分析 app
2.时间序列分析中可以加入区域天气信息等额外信息等，研究大雨等天气对销售影响
3.报告撰写时间较急，没能完善代码解释，及更多功能

参考资料

[1].数据整型，reshape2 使用：http://seananderson.ca/2013/10/19/reshape.html
[2].数据集分组，dplyr 包使用：
http://blog.csdn.net/sinat_26917383/article/details/506884[3].Tidy data.Hadley Wickham