可视化作业

中央财经大学
司徒雪颖

一、 数据变量介绍

Dailyprice 数据集包含 19 个变量，5,322,012 条样本。

变量名	变量名解释	单位
datetime	时间	
trade_code	证券代码	
open	开盘价	元
high	最高价	元
low	最低价	元
close	收盘价	元
volume	成交量	手
amt	成交额	元
chg	涨跌	元
pct_chg	涨跌幅	百分比
adjfactor	复权因子	
turn	换手率	百分比
free_turn	换手率（基准，自由流通股本）	百分比
total_shares	总股本	亿股
mkt_cap	总市值	元
free_float_shares	自由流通股本	亿股
annualstdevr_100w	年化波动率（最近 100 周）	百分比
dividendyield2	股息率（近 100 周）	百分比
trade_status	交易状态	0-1 变量

二、 数据读入

```
1. library(data.table)  #使用 data.table 可以高效读入大型数据
2. setwd("E:/graduate/class/visualization/1 数据集一：Wind 数据/")
3. dailyprice = fread("dailyprice.txt")
```

三、多变量数据的展示

1. 二维变量展示：等高线图

图 3-1 展示了大部分股票都集中在以换手率为 3，涨跌幅为 2%为圆心的附近。

```
1. library(ggplot2)
2. corplotData = dailyprice[datetime == "2015-4-
   29",.(close,pct_chg,adjfactor,turn)]
3. p<-ggplot(na.omit(corplotData),aes(x=pct_chg,y=turn))
4. p = p+geom_jitter(size = 1)+stat_density2d(h = 10,size = 1)
```

```
5.  p+coord_cartesian(ylim=c(0,20))+xlab("涨跌幅")+ylab("换手率")
```

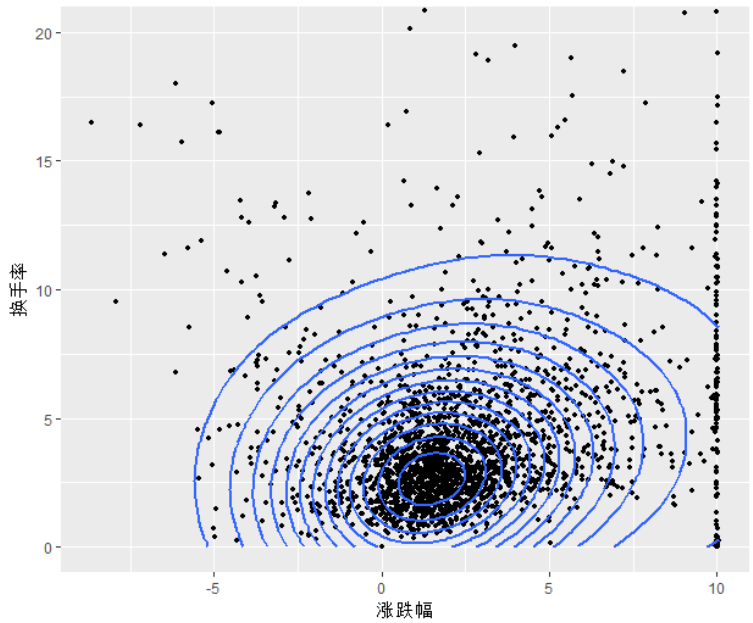

图 3-1 换手率、涨跌幅的二维密度图

2. 三维数据展示：气泡图

图 3-2 展示了股息率和年华波动率之间可能存在负相关关系，它们二者与收盘价，即气泡大小无明显关系

```
1.  p = ggplot(dailyprice[datetime == "2015-4-
    29"][1:100],aes(x = annualstdevr_100w,y = dividendyield2,size =close))
2.  p+geom_point(shape=21,colour="black",fill="lightblue")+scale_size_area(max_s
    ize= 10)+xlab("年化波动率")+ylab("股息率")
```

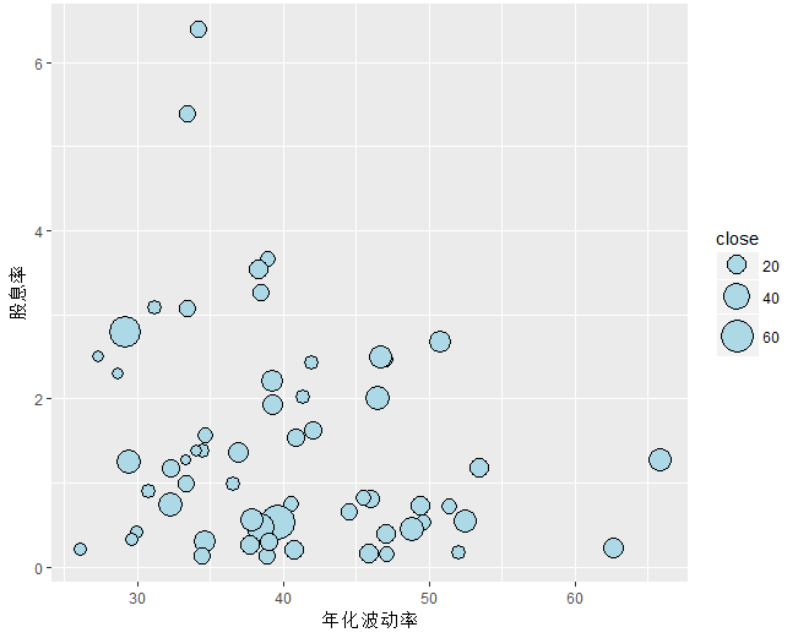

图 3- 2 股息率、年化波动率与收盘价的气泡图

3.多维变量展示

（1）散点图矩阵

图 3-3-（1）展示了收盘价，涨跌幅，复权因子，换手率之间的散点图和各自的分布，其中复权因子有明显的异常值，其他 3 个变量都是偏态分布，这 4 个变量之间看不出有什么明显的相关关系。

```
1. corplotData = dailyprice[datetime == "2015-4-
   29",.(close,pct_chg,adjfactor,turn)]
2. library(car)
3. scatterplotMatrix(na.omit(corplotData),var.labels = c("收盘价","涨跌幅","复权
   因子","换手率"))
```

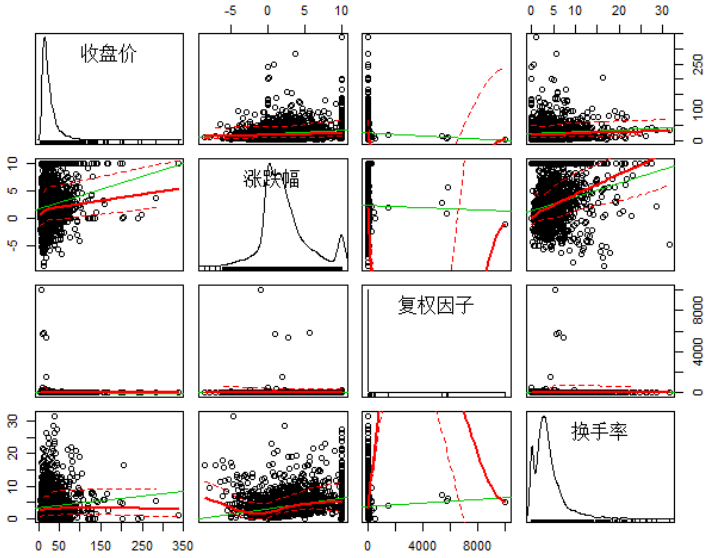

图 3- 3-(1) 收盘价，涨跌幅，复权因子，换手率的散点图矩阵

（2）相关图矩阵

图 3- 3-(2)展示了 8 个变量之间的相关关系，收盘价与涨跌额成正相关但与涨跌幅无明显相关关系，因为价格高的股票，稍微涨一点或跌一点数额都较大。涨跌幅涨跌额，换手率与基准换手率，总市值与自由流通股本本来就是衡量同一种东西，因此相关度很高。

```
1. corplotData2 = dailyprice[datetime == "2015-4-
   29",.(close,chg,pct_chg,adjfactor,turn,free_turn,mkt_cap,mkt_freeshares)]
2. library(corrplot)
3. corrplot(cor(na.omit(corplotData2)),tl.col="black")
```

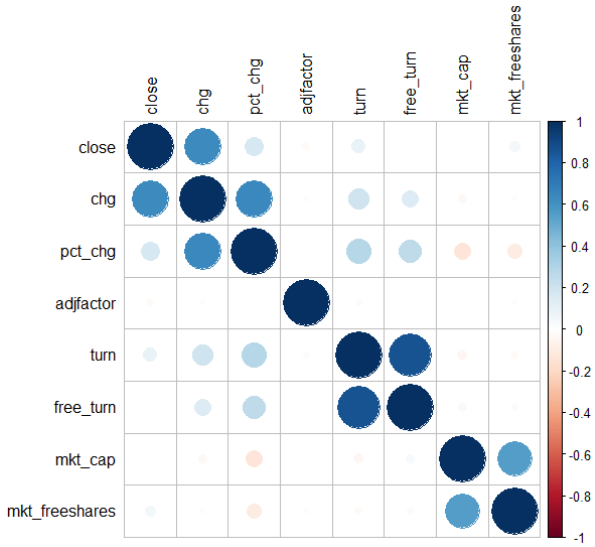

图 3- 3-(2) 相关图矩阵

(3) 平行坐标图

构造了新的一列，upordown 指示股票涨跌与否，涨跌幅为正数则为 1，为负数则为-1，为 0 则为 0。可以看到涨跌与股价、复权因子无明显相关关系。图 3- 3-(3)展示了 2015 年 4 月 29 日四个变量之间的平行坐标图。

```
1. library(lattice)
2. pData = dailyprice[datetime == "2015-4-
   29",.(close,chg,pct_chg,adjfactor,turn)]
3. upordown = data.frame(upordown = rep(0,dim(pData)[1]))
4. upordown[pData$chg>0,]=1 #涨
5. upordown[pData$chg<0,]=-1 #跌
6. upordown[pData$chg==0,]=0 #持平
7. pData = cbind(pData,upordown)#添加定性变量
8. parallel(~pData[,1:4],pData,group = upordown,horizontal.axis= FALSE)
```

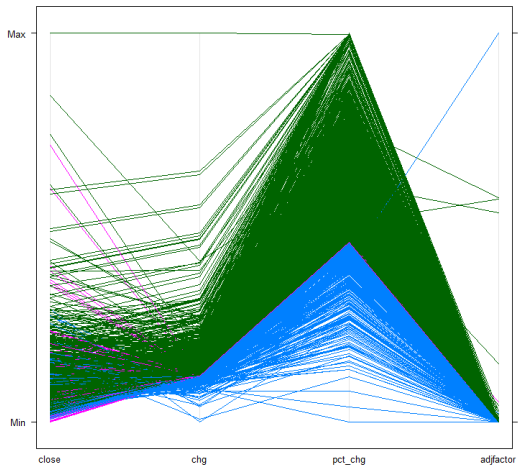

图 3- 3-(3)平行坐标图

（4）雷达图

图 3- 3-（4）展示了前 10 只股票 15 年 4 月 29 日在五个维度上的比较

```
stars(pData[1:10,1:5],locations=c(0,0),col.lines= 2:7,radius=FALSE,key.loc=c(0,0),lwd=1.5,scale = TRUE,main="Star (Spider/Radar) Plot")
```

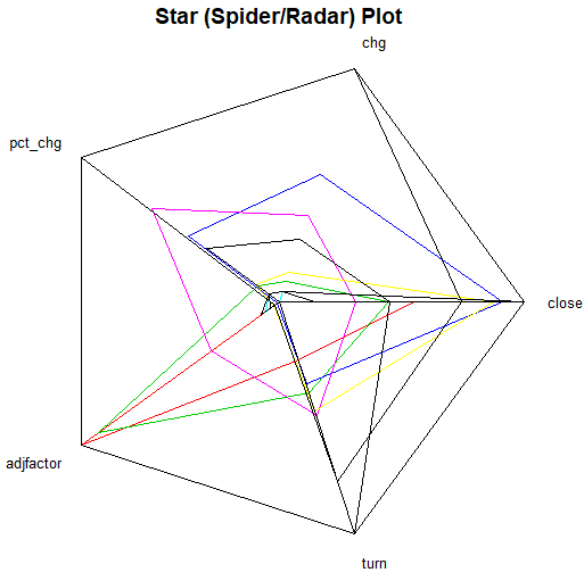

图 3- 3-（4）雷达图

（5）热图

把日期中的年和季度取出来变成新的两列，以年为横轴，季度为纵轴，一个季度的平均价格映射到热力图的颜色深浅，价格越高，颜色越浅。从图 3- 3-（5）可以看出，股票代码为 000001 的股票 07 年之前颜色较浅，07、08 年价格上涨，之后又回落。

```
heatplotData = dailyprice[trade_code == "000001.SZ",.(datetime,close)]
heatplotData$datetime = as.Date(heatplotData$datetime,format='%Y-%m-%d')
heatplotData = cbind(heatplotData,year(heatplotData$datetime),quarter(heatplotData$datetime))
colnames(heatplotData)[3:4] = c("quarter","year")
heatplot_group<-heatplotData[,.(mean_price=mean(close)),by=.(quarter,year)]
p<-ggplot(heatplot_group,aes(x=year,y=quarter,fill= mean_price))
p+ geom_tile(na.rm = T)
```

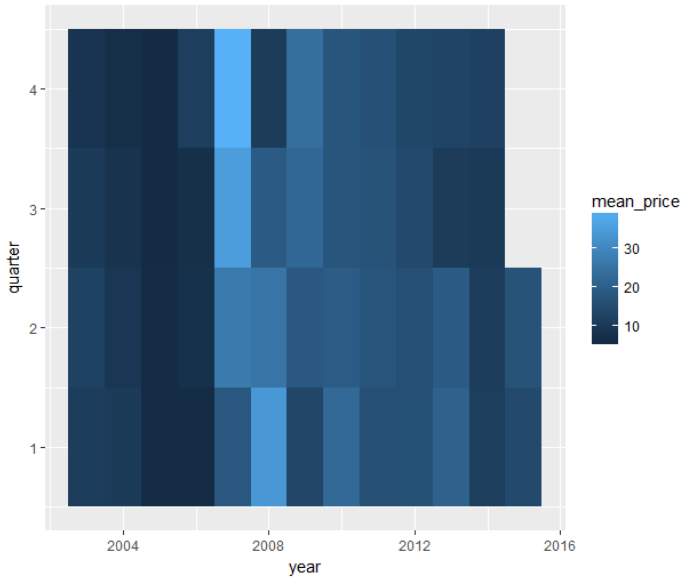

图 3- 3-（5）股票价格与时间的热力图

四、数据分布形态展示

1.直方图+密度曲线

图 4- 1 展示了 2015 年 4 月 29 日这天的所有股票收盘价的分布，明显的右偏分布。

```
1.  p = ggplot(dailyprice[datetime == "2015-4-29"],aes(x = close))
2.  p = p+geom_histogram(aes(y = ..density..),bins = 50,alpha = 0.5,fill = I("st
    eelblue"))+stat_density(geom = "line",colour = I("blue"))
3.  p+coord_cartesian(xlim=c(0,100))+xlab("收盘价")
```

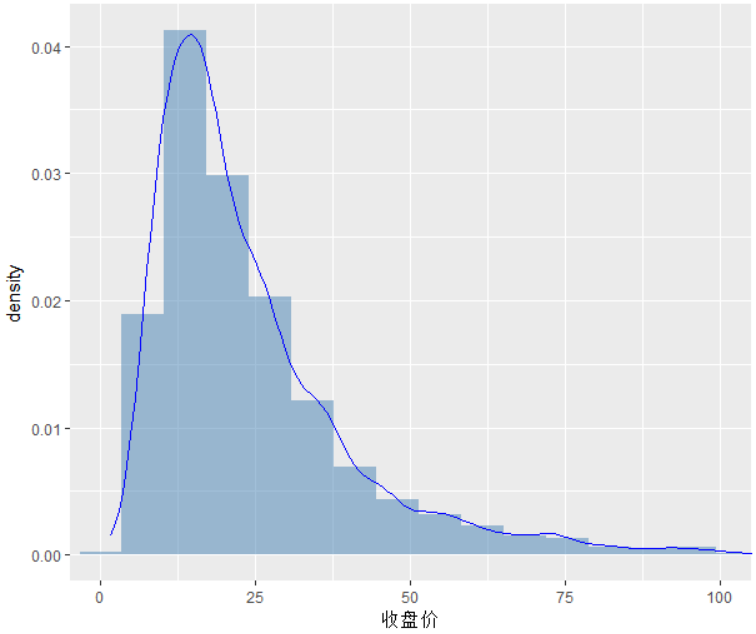

图 4- 2 收盘价密度图

2. 箱线图与小提琴图

图 4- 2-1 的左图为有异常点的箱线图，右图为隐去异常点的箱线图。图 4- 2-2 的左图为有异常点的小提琴图，右图为隐去异常点的小提琴图。4 图都是以涨跌与否为分类变量，看复权因子的分布。明显看出复权因子有很多异常值。

```
#有异常值
pData$upordown = as.character(pData$upordown)
pData$adjfactor[pData$adjfactor>100] = NA
p = ggplot(pData,aes(x=upordown,y=adjfactor))+xlab("涨或跌
")+stat_summary(fun.y="mean",geom="point",shape=23,size=3,fill="white")
p+geom_boxplot()
p +geom_violin()
#无异常值
p = ggplot(pData,aes(x=upordown,y=adjfactor))+coord_cartesian(ylim=c(0,20))+
xlab("涨或跌
")+stat_summary(fun.y="mean",geom="point",shape=23,size=3,fill="white")#有异
常值
p+geom_boxplot(outlier.shape= NA)
p +geom_violin(outlier.colour= NA,trim=FALSE)
```

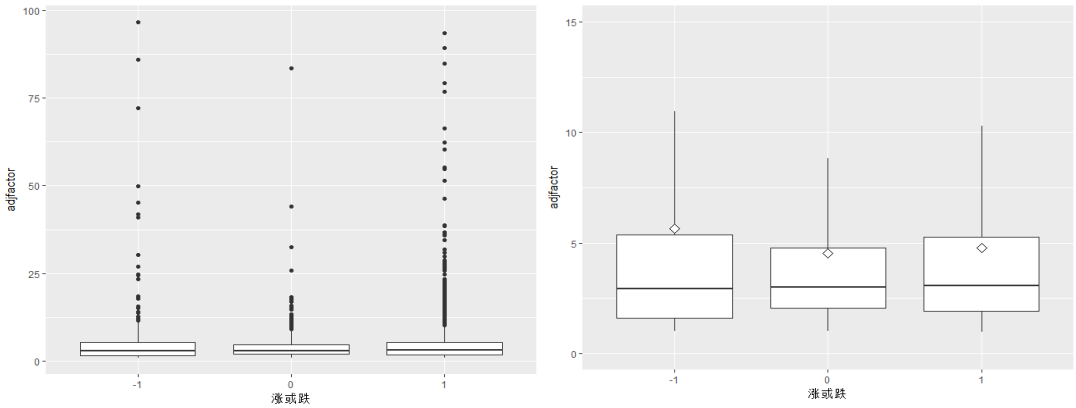

图 4- 3-1 箱线图

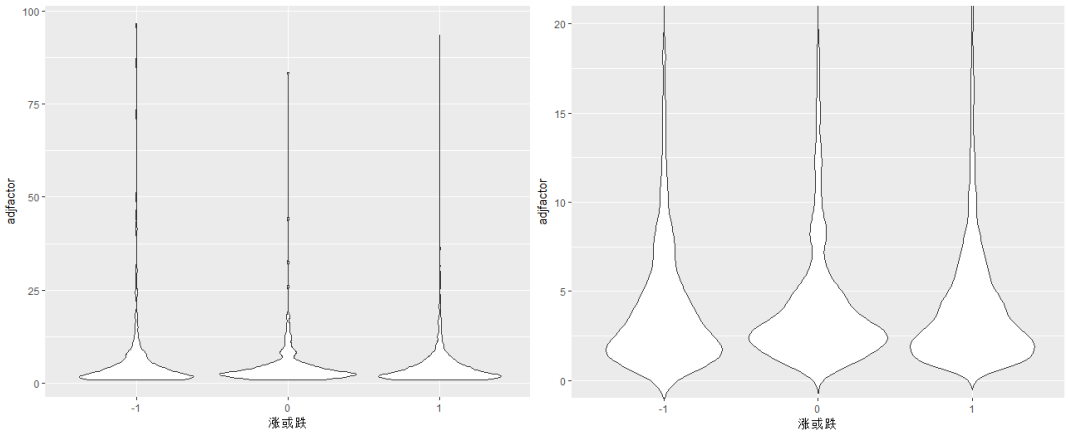

图 4- 2-2 小提琴图

五、常用分布绘制

1.F 分布

图 5- 1 所示的 F 分布的自由度为 10,5，红线为累积分布曲线，绿线为概率密度分布曲线。

```
set.seed(1)
x <-seq(0,5,length.out=100)
y <-df(x,10,5)
z = pf(x,10,5)
myrandom = data.frame(x = c(x,x),y = c(y,z),symbol = c(rep("pdf",100),rep("cdf",100)))
p = ggplot(myrandom)
p+geom_line(aes(x,y,colour = symbol))
```

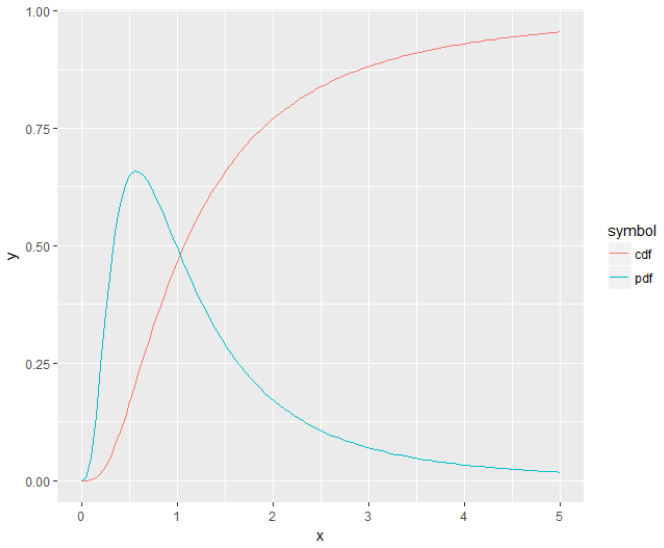

图 5- 1 F 分布

2.卡方分布

图 5- 2 所示的卡方分布的自由度为 2，红线为概率密度分布曲线，绿线为累积分布曲线。

```
set.seed(1)
x <-seq(0,5,length.out=100)
y <-dchisq(x,2)
z = pchisq(x,2)
plot(x,y,col="red",xlim=c(0,5),ylim=c(0,1),type='l',xaxs="i", yaxs="i",ylab='y',xlab='x')
lines(x,z,col="green")
legend("topleft",legend=c("概率密度分布",'累积分布'), lwd=1,col=c("red", "green"))
```

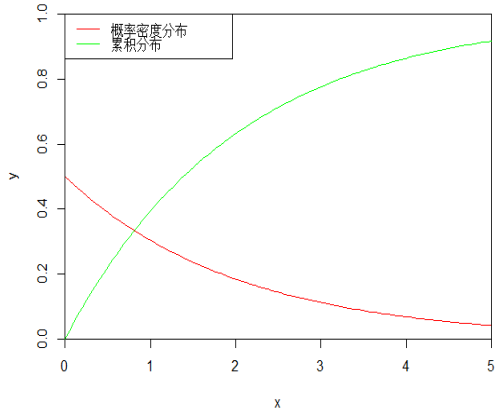

图 5- 2 卡方分布

六、交互图

```
1.  library(plotly)
2.  plot_ly(pData,x=~chg,y=~pct_chg,type="scatter")
```

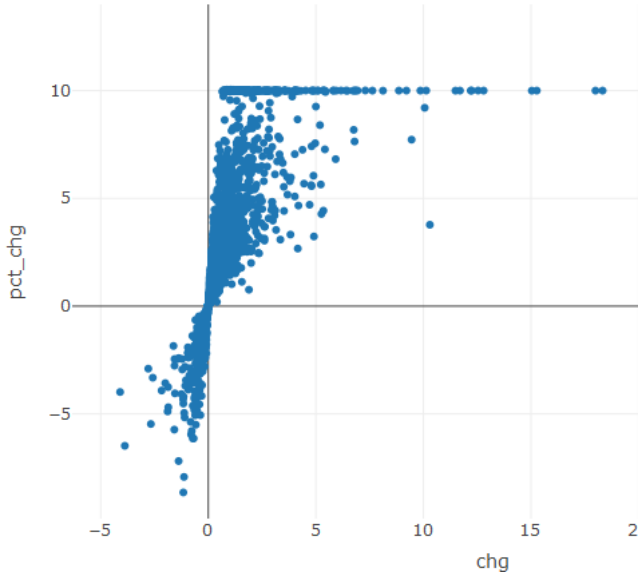

图 6 涨跌幅和涨跌额的散点图