【小刻也能看懂的R数据可视化】优雅的平铺热力图指南（翻译）

> 原文链接：https://www.royfrancis.com/a-guide-to-elegant-tiled-heatmaps-in-r-2019/
>
> 包含数据集来源与下载。
>
> PDF或md文件下载：https://github.com/Kaede0614/typoranotebook

- 数据集：~~源石病~~ 麻疹一级发病率（每10万人）
- 目的：绘制简洁、干净、优雅的热力图
- 用到的包：ggplot2、dplyr、tidyr、stringr

```
# install packages
# install.packages(pkgs =
c("ggplot2","dplyr","tidyr","stringr","gplots","plotrix"),dependencies = T)

# load packages
library(ggplot2) # ggplot() for plotting
library(dplyr) # data reformatting
library(tidyr) # data reformatting
library(stringr) # string manipulation
```

1. 数据准备

首先导入csv文件并检查数据，跳过csv表格的前两行。

```
# read csv file
m <- read.csv("measles_lev1.csv",header=T,stringsAsFactors=F,skip=2)

# inspect data
head(m)
str(m)
table(m$YEAR)
table(m$WEEK)
```

- head()查看前6行数据。
- str()检查数据结构。发现年和月是int，发生率是chr。
- table()检查年和周是否有缺失数据。

目前数据是"宽"格式，而ggplot2需要"长"格式，因此需要转换。年和月保持原样，所有发生率折叠到一列。为了方便将列名改为小写。年和月变量转换为factor，值转换为数字。

```
m2 <- m %>%
  # convert data to long format
  gather(key="state",value="value",-YEAR,-WEEK) %>%
  # rename columns
  setNames(c("year","week","state","value")) %>%
  # convert year to factor
  mutate(year=factor(year)) %>%
  # convert week to factor
  mutate(week=factor(week)) %>%
  # convert value to numeric (also converts '-' to NA, gives a warning)
  mutate(value=as.numeric(value))
```

这样处理后数据集变成适合ggplot2处理的格式。

```
> head(m2)
  year week   state value
1 1928    1 Alabama  3.67
2 1928    2 Alabama  6.25
3 1928    3 Alabama  7.95
4 1928    4 Alabama 12.58
5 1928    5 Alabama  8.03
6 1928    6 Alabama  7.27
```

该m2数据集如果导出为csv，格式如下。

	year	week	state	value
33325	1960	45	DISTRICT.OF.COLUMBIA	NA
33326	1960	46	DISTRICT.OF.COLUMBIA	0.13
33327	1960	47	DISTRICT.OF.COLUMBIA	0.13
33328	1960	48	DISTRICT.OF.COLUMBIA	0.26
33329	1960	49	DISTRICT.OF.COLUMBIA	NA
33330	1960	50	DISTRICT.OF.COLUMBIA	2.09

可以看出各州名称是全大写，并且单词间用点分隔。美观考虑修改为首字母大写（Title Case），空格分隔。考虑用str_to_title()处理单词再将他们粘贴到一起。

```
# removes . and change states to title case using custom function
fn_tc <- function(x) paste(str_to_title(unlist(strsplit(x,"[.]"))),collapse=" ")
m2$state <- sapply(m2$state,fn_tc)
```

- strsplit()：按分隔符拆分字符串。注意由于“.”在R中有特殊含义，需要用正则表达式表示（链接）。
- unlist()：将列表转换为向量，以便运算。
- str_to_title：将英文字符串中的单词首字母大写。
- paste：将字符串连接起来，分隔符为空格。注意collpase负责的是字符串内部的连接，两组字符串连接用sep。
- R中匿名函数最常用于*apply()类函数。
- sapply()：返回处理后的向量，替代m2的州名。

接下来是绘图。考虑X轴为年，Y轴为各州名称，这意味着要先处理周这个变量。我们将每年所有星期的发病率相加，再删除周变量。`dplyr`的方法是用函数`group_by()`和`summarise()`。

`sum()`函数处理`NA`的方式比较奇怪，默认下，如果输入向量中有`NA`就会返回`NA`。如果设置参数`na.rm=TRUE`，那么`NA`会被移除并加总剩余数字。但是如果所有元素均为`NA`，那么会返回0。这不适合本例，因此选择自定义加总函数`na_sum()`来移除`NA`，并且如果所有元素为`NA`时返回`NA`。接着在`summarise()`函数中使用该自定义函数，按年份和州汇总数去，同时去掉周。

```r
# custom sum function returns NA when all values in set are NA,
# in a set mixed with NAs, NAs are removed and remaining summed.
na_sum <- function(x)
{
  if(all(is.na(x))) val <- sum(x,na.rm=F)
  if(!all(is.na(x))) val <- sum(x,na.rm=T)
  return(val)
}

# sum incidences for all weeks into one year
m3 <- m2 %>%
  group_by(year,state) %>%
  summarise(count=na_sum(value)) %>%
  as.data.frame()
```

- `na.rm`参数为T时移除所有`NA`再加总，为F时返回`NA`。
- `dplyr`分两步完成汇总。先`group_by()`定义分组变量，再`summarise()`描述如何汇总。

处理后m3中没有周变量，如下。

```
> head(m3)

  year      state  count
1 1928    Alabama 334.99
2 1928     Alaska   0.00
3 1928    Arizona 200.75
4 1928   Arkansas 481.77
5 1928 California  69.22
6 1928   Colorado 206.98
```

现在数据准备工作基本结束。数据为“长数据”格式，三变量分别为factor、factor、numeric型。

2. 绘图

2.1 ggplot2

```r
#basic ggplot
p <- ggplot(m3,aes(x=year,y=state,fill=count))+
      geom_tile()

#save plot to working directory
ggsave(p,filename="measles-basic.png")
```

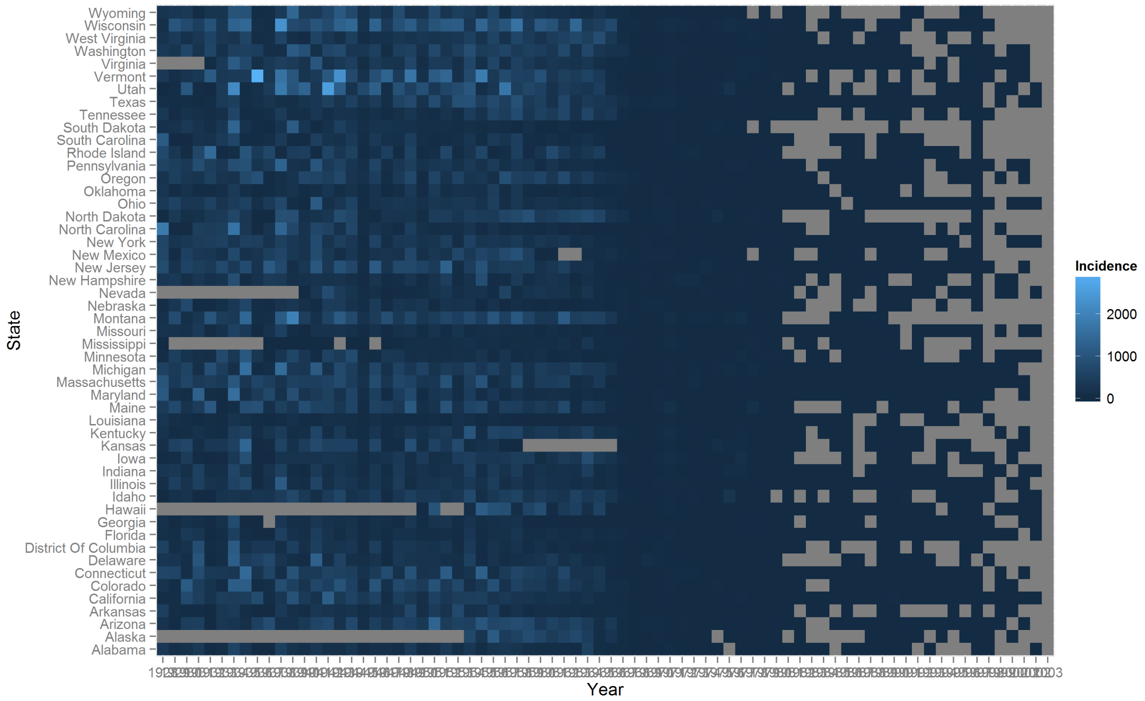

图1. ggplot2基础图像

该图存在几个问题。例如x轴粘在一起，y轴字体太大没有间距，热力图瓦片间没有分隔导致看上去很丑。因此需要添加瓦片边界、自定义x轴分隔和自定义文本大小。调整后的图像如下。

```
#modified ggplot
p <- ggplot(m3,aes(x=year,y=state,fill=count))+
  #add border white colour of line thickness 0.25
  geom_tile(colour="white",size=0.25)+
  #remove x and y axis labels
  labs(x="",y="")+
  #remove extra space
  scale_y_discrete(expand=c(0,0))+
  #define new breaks on x-axis
  scale_x_discrete(expand=c(0,0),

breaks=c("1930","1940","1950","1960","1970","1980","1990","2000"))+
  #set a base size for all fonts
  theme_grey(base_size=8)+
  #theme options
  theme(
    #bold font for legend text
    legend.text=element_text(face="bold"),
    #set thickness of axis ticks
    axis.ticks=element_line(size=0.4),
    #remove plot background
    plot.background=element_blank(),
    #remove plot border
    panel.border=element_blank())

#save with dpi 200
ggsave(p,filename="measles-mod1.png",height=5.5,width=8.8,units="in",dpi=200)
```

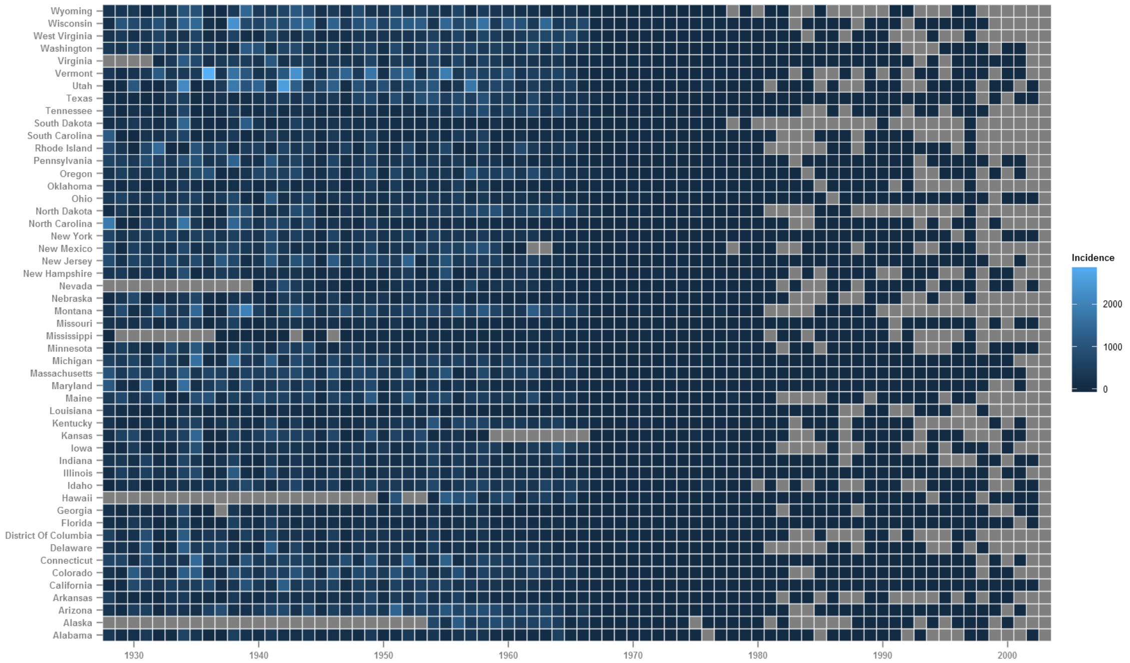

图2. 定制后的热力图

该图的问题是y轴没有按照字母排序，而且颜色选择不当使得热力图肉眼无法很好分辨。前一个问题，需要回到开始的“长数据”，重构**州**这个变量。后一个问题，由于填充变量**count**是连续变量，因此ggplot会默认用蓝色色带表示。需要将连续变量分成若干等级，每个等级用离散颜色表示。cut()函数能够分割并标记一个连续变量。

本例将**count**变量分为7个等级，并保存为一个新的countfactor变量。NA依旧保留为NA。取多少个断点取决于数据类型，可以试错决定，不要过多。擅用summary()或boxplot()能够揭示很多数据特征。

```r
m4 <- m3 %>%
      # convert state to factor and reverse order of levels
      mutate(state=factor(state,levels=rev(sort(unique(state))))) %>%
      # create a new variable from count
      mutate(countfactor=cut(count,breaks=c(-1,0,1,10,100,500,1000,max(count,na.rm=T)),
                             labels=c("0","0-1","1-10","10-100","100-500","500-1000",">1000"))) %>%
      # change level order
      mutate(countfactor=factor(as.character(countfactor),levels=rev(levels(countfactor))))
```

现在可以准备画最终的数据集了。

```r
# assign text colour
textcol <- "grey40"

# further modified ggplot
p <- ggplot(m4,aes(x=year,y=state,fill=countfactor))+
  geom_tile(colour="white",size=0.2)+
  guides(fill=guide_legend(title="Cases per\n100,000 people"))+
  labs(x="",y="",title="Incidence of Measles in the US")+
  scale_y_discrete(expand=c(0,0))+
  scale_x_discrete(expand=c(0,0),breaks=c("1930","1940","1950","1960","1970","1980","1990","2000"))+
  scale_fill_manual(values=c("#d53e4f","#f46d43","#fdae61","#fee08b","#e6f598","#abdda4","#ddf1da"),na.value = "grey90")+
  #coord_fixed()+
  theme_grey(base_size=10)+
  theme(legend.position="right",legend.direction="vertical",
        legend.title=element_text(colour=textcol),
        legend.margin=margin(grid::unit(0,"cm")),
        legend.text=element_text(colour=textcol,size=7,face="bold"),
        legend.key.height=grid::unit(0.8,"cm"),
        legend.key.width=grid::unit(0.2,"cm"),
        axis.text.x=element_text(size=10,colour=textcol),
        axis.text.y=element_text(vjust=0.2,colour=textcol),
        axis.ticks=element_line(size=0.4),
        plot.background=element_blank(),
        panel.border=element_blank(),
        plot.margin=margin(0.7,0.4,0.1,0.2,"cm"),
        plot.title=element_text(colour=textcol,hjust=0,size=14,face="bold"))

#export figure
ggsave(p,filename="measles-mod3.png",height=5.5,width=8.8,units="in",dpi=200)
```

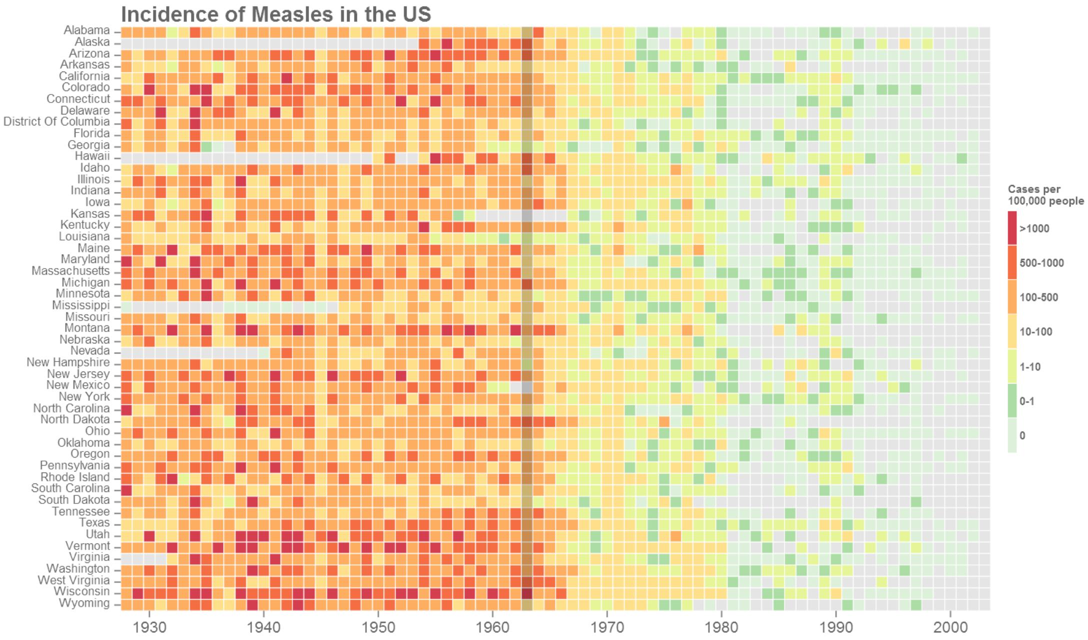

图3. 进一步定制后的热力图

图3就是最终的版本。字体颜色为grey40，缺失值NA颜色为grey90。引入疫苗接种的年份用深色垂直条表示。用R包RColorBrewer()或ggplot2函数scale_fill_brewer()可以打开所有调色板（ColorBrewer网站）。下面是个例子。

```
library(RColorBrewer)

#change the scale_fill_manual from previous code to below
scale_fill_manual(values=rev(brewer.pal(7,"YlGnBu")),na.value="grey90")+
```

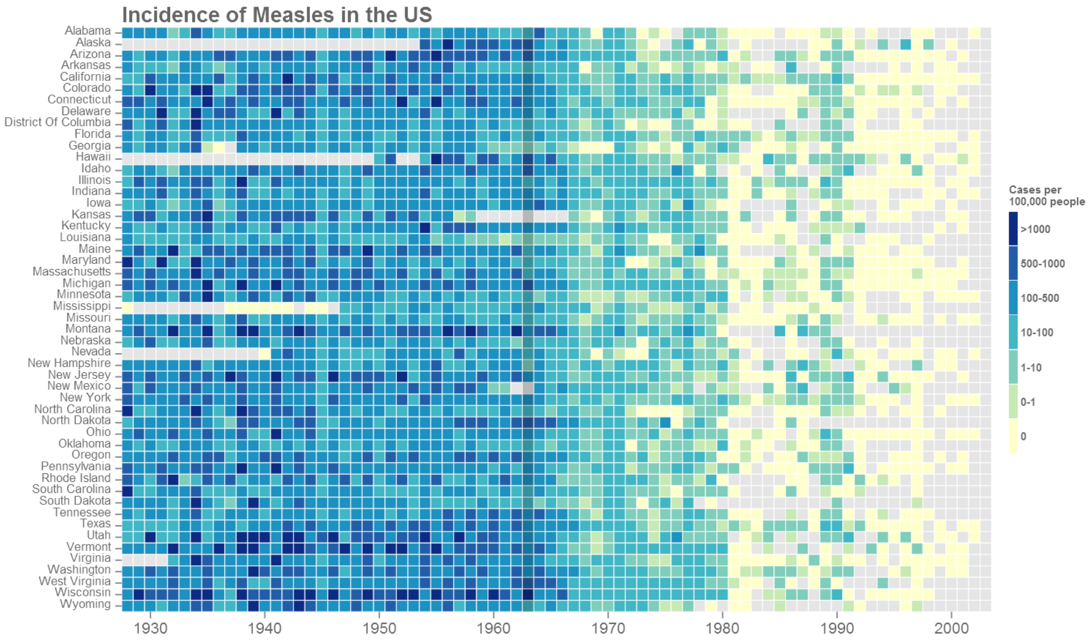

图4. Colorbrewer调色板的热力图

大部分的绘图定制选项在theme()部分。extrafont()包可以改变字体。另外可以选择矢量格式导出（svg或pdf），以便用AI等软件编辑，加入更多的文字。

2.2 基础函数绘图（选读）

可以用gplots包的heatmap()或heatmap2()函数。此时输入的数据矩阵是“宽”格式的。用**tidyr**包的spread()函数将长数据转换为宽数据。将非数值列移除从而将宽数据转换为矩阵。州名重新被分配为矩阵的行名，用作y轴文本。

```
# load package
library(gplots) # heatmap.2() function
library(plotrix) # gradient.rect() function

# convert from long format to wide format
m5 <- m3 %>% spread(key="state",value=count)
m6 <- as.matrix(m5[,-1])
rownames(m6) <- m5$year

#base heatmap
png(filename="measles-base.png",height=5.5,width=8.8,res=200,units="in")
heatmap(t(m6),Rowv=NA,Colv=NA,na.rm=T,scale="none",col=terrain.colors(100),
 xlab="",ylab="",main="Incidence of Measles in the US")
dev.off()
```

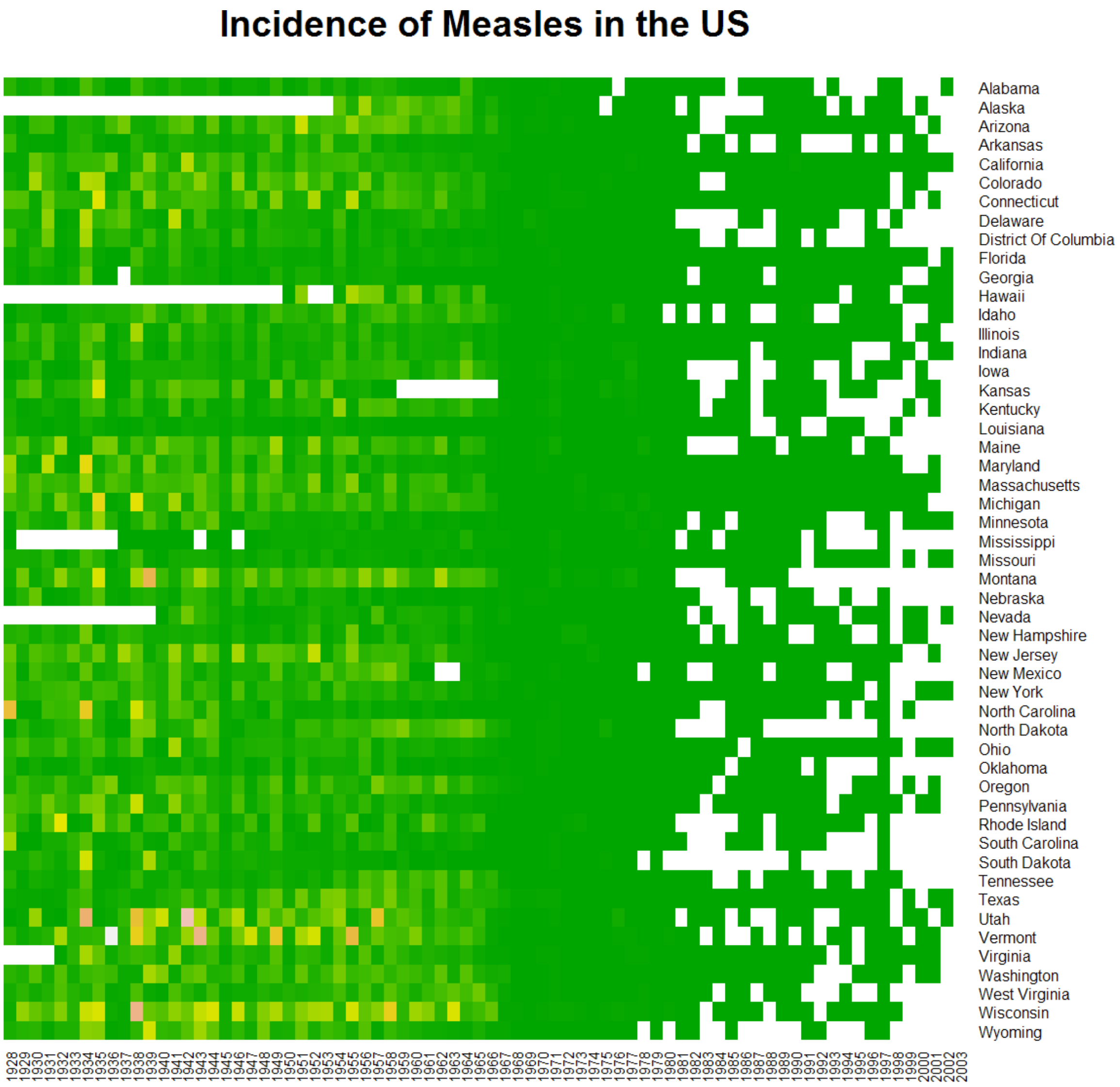

图5. heatmap()绘制的热力图

但用heatmap()的话这差不多就是全部能做的了（很丑）。heatmap2()可以做更多。用**plotrix**包中的gradient.rect()函数来定制自己的图例。多次调整和试错后得到如下图像。可以看出还是ggplot2方便又好看。

```r
# gplots heatmap.2
png(filename="measles-gplot.png",height=6,width=9,res=200,units="in")
par(mar=c(2,3,3,2))
gplots::heatmap.2(t(m6),na.rm=T,dendrogram="none",Rowv=NULL,Colv="Rowv",trace="none",scale="none",offsetRow=0.3,offsetCol=0.3,

 breaks=c(-1,0,1,10,100,500,1000,max(m4$count,na.rm=T)),colsep=which(seq(1928,2003)%%10==0),

 margin=c(3,8),col=rev(c("#d53e4f","#f46d43","#fdae61","#fee08b","#e6f598","#abdda4","#ddf1da")),
                  xlab="",ylab="",key=F,lhei=c(0.1,0.9),lwid=c(0.2,0.8))
gradient.rect(0.125,0.25,0.135,0.75,nslices=7,border=F,gradient="y",col=rev(c("#d53e4f","#f46d43","#fdae61","#fee08b","#e6f598","#abdda4","#ddf1da")))
text(x=rep(0.118,7),y=seq(0.28,0.72,by=0.07),adj=1,cex=0.8,labels=c("0","0-1","1-10","10-100","100-500","500-1000",">1000"))
text(x=0.135,y=0.82,labels="Cases per\n100,000 people",adj=1,cex=0.85)
title(main="Incidence of Measles in the US",line=1,oma=T,adj=0.21)
dev.off()
```

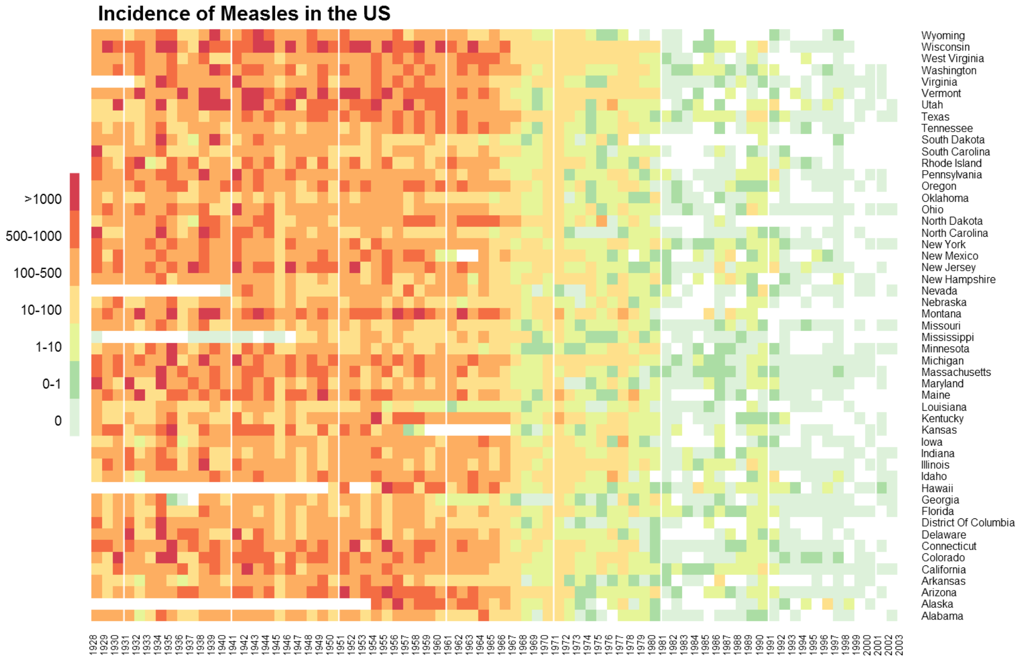

图6. heatmap2()绘制的热力图

3. 完整代码例

```r
# 2019 | Roy Mathew Francis
# Heatmap R code

#load packages
library(ggplot2) # ggplot() for plotting
library(dplyr) # data reformatting
library(tidyr) # data reformatting
library(stringr) # string manipulation

# DATA PREPARATION ------------------------------------------------------------------

```

```r
#read csv file
m <- read.csv("measles_lev1.csv",header=T,stringsAsFactors=F,skip=2)

m2 <- m %>%
  # convert data to long format
  gather(key="state",value="value",-YEAR,-WEEK) %>%
  # rename columns
  setNames(c("year","week","state","value")) %>%
  # convert year to factor
  mutate(year=factor(year)) %>%
  # convert week to factor
  mutate(week=factor(week)) %>%
  # convert value to numeric (also converts '-' to NA, gives a warning)
  mutate(value=as.numeric(value))

# removes . and change states to title case using custom function
fn_tc <- function(x) paste(str_to_title(unlist(strsplit(x,"[.]"))),collapse=" ")
m2$state <- sapply(m2$state,fn_tc)

# custom sum function returns NA when all values in set are NA,
# in a set mixed with NAs, NAs are removed and remaining summed.
na_sum <- function(x)
{
  if(all(is.na(x))) val <- sum(x,na.rm=F)
  if(!all(is.na(x))) val <- sum(x,na.rm=T)
  return(val)
}

# sum incidences for all weeks into one year
m3 <- m2 %>%
  group_by(year,state) %>%
  summarise(count=na_sum(value)) %>%
  as.data.frame()

m4 <- m3 %>%
      # convert state to factor and reverse order of levels
      mutate(state=factor(state,levels=rev(sort(unique(state))))) %>%
      # create a new variable from count
      mutate(countfactor=cut(count,breaks=c(-1,0,1,10,100,500,1000,max(count,na.rm=T)),
                             labels=c("0","0-1","1-10","10-100","100-500","500-1000",">1000"))) %>%
      # change level order
      mutate(countfactor=factor(as.character(countfactor),levels=rev(levels(countfactor))))

# GGPLOT -----------------------------------------------------------------------

# assign text colour
textcol <- "grey40"

# further modified ggplot
p <- ggplot(m4,aes(x=year,y=state,fill=countfactor))+
  geom_tile(colour="white",size=0.2)+
  guides(fill=guide_legend(title="Cases per\n100,000 people"))+
  labs(x="",y="",title="Incidence of Measles in the US")+
  scale_y_discrete(expand=c(0,0))+
  scale_x_discrete(expand=c(0,0),breaks=c("1930","1940","1950","1960","1970","1980","1990","2000"))+
  scale_fill_manual(values=c("#d53e4f","#f46d43","#fdae61","#fee08b","#e6f598","#abdda4","#ddf1da"),na.value = "grey90")+
  #coord_fixed()+
  theme_grey(base_size=10)+
  theme(legend.position="right",legend.direction="vertical",
        legend.title=element_text(colour=textcol),
        legend.margin=margin(grid::unit(0,"cm")),
        legend.text=element_text(colour=textcol,size=7,face="bold"),
        legend.key.height=grid::unit(0.8,"cm"),
        legend.key.width=grid::unit(0.2,"cm"),
        axis.text.x=element_text(size=10,colour=textcol),
        axis.text.y=element_text(vjust=0.2,colour=textcol),
        axis.ticks=element_line(size=0.4),
        plot.background=element_blank(),
        panel.border=element_blank(),
        plot.margin=margin(0.7,0.4,0.1,0.2,"cm"),
        plot.title=element_text(colour=textcol,hjust=0,size=14,face="bold"))

# export figure
ggsave(p,filename="measles-mod3.png",height=5.5,width=8.8,units="in",dpi=200)

# BASE GRAPHICS ----------------------------------------------------------------

# load package
library(gplots) # heatmap.2() function
library(plotrix) # gradient.rect() function

# convert from long format to wide format
m5 <- m3 %>% spread(key="state",value=count)
m6 <- as.matrix(m5[,-1])
rownames(m6) <- m5$year

# base heatmap
png(filename="measles-base.png",height=5.5,width=8.8,res=200,units="in")
heatmap(t(m6),Rowv=NA,Colv=NA,na.rm=T,scale="none",col=terrain.colors(100),
        xlab="",ylab="",main="Incidence of Measles in the US")
dev.off()

# gplots heatmap.2
png(filename="measles-gplot.png",height=6,width=9,res=200,units="in")
par(mar=c(2,3,3,2))
gplots::heatmap.2(t(m6),na.rm=T,dendrogram="none",Rowv=NULL,Colv="Rowv",trace="none",scale="none",offsetRow=0.3,offsetCol=0.3,
                  breaks=c(-1,0,1,10,100,500,1000,max(m4$count,na.rm=T)),colsep=which(seq(1928,2003)%%10==0),
                  margin=c(3,8),col=rev(c("#d53e4f","#f46d43","#fdae61","#fee08b","#e6f598","#abdda4","#ddf1da")),
                  xlab="",ylab="",key=F,lhei=c(0.1,0.9),lwid=c(0.2,0.8))
```

```r
gradient.rect(0.125,0.25,0.135,0.75,nslices=7,border=F,gradient="y",col=rev(c("#d53e4f","#f46d43","#fdae61","#fee08b","#e6f598","#abdda4","#ddf1da")))
text(x=rep(0.118,7),y=seq(0.28,0.72,by=0.07),adj=1,cex=0.8,labels=c("0","0-1","1-10","10-100","100-500","500-1000",">1000"))
text(x=0.135,y=0.82,labels="Cases per\n100,000 people",adj=1,cex=0.85)
title(main="Incidence of Measles in the US",line=1,oma=T,adj=0.21)
dev.off()

# End of script -----------------------------------------------------------------
```