期中報告

美國凶殺案和離婚率的關係

B082040005 高念慈

時間：2022

目錄

CONTENTS

01 資料來源&介紹

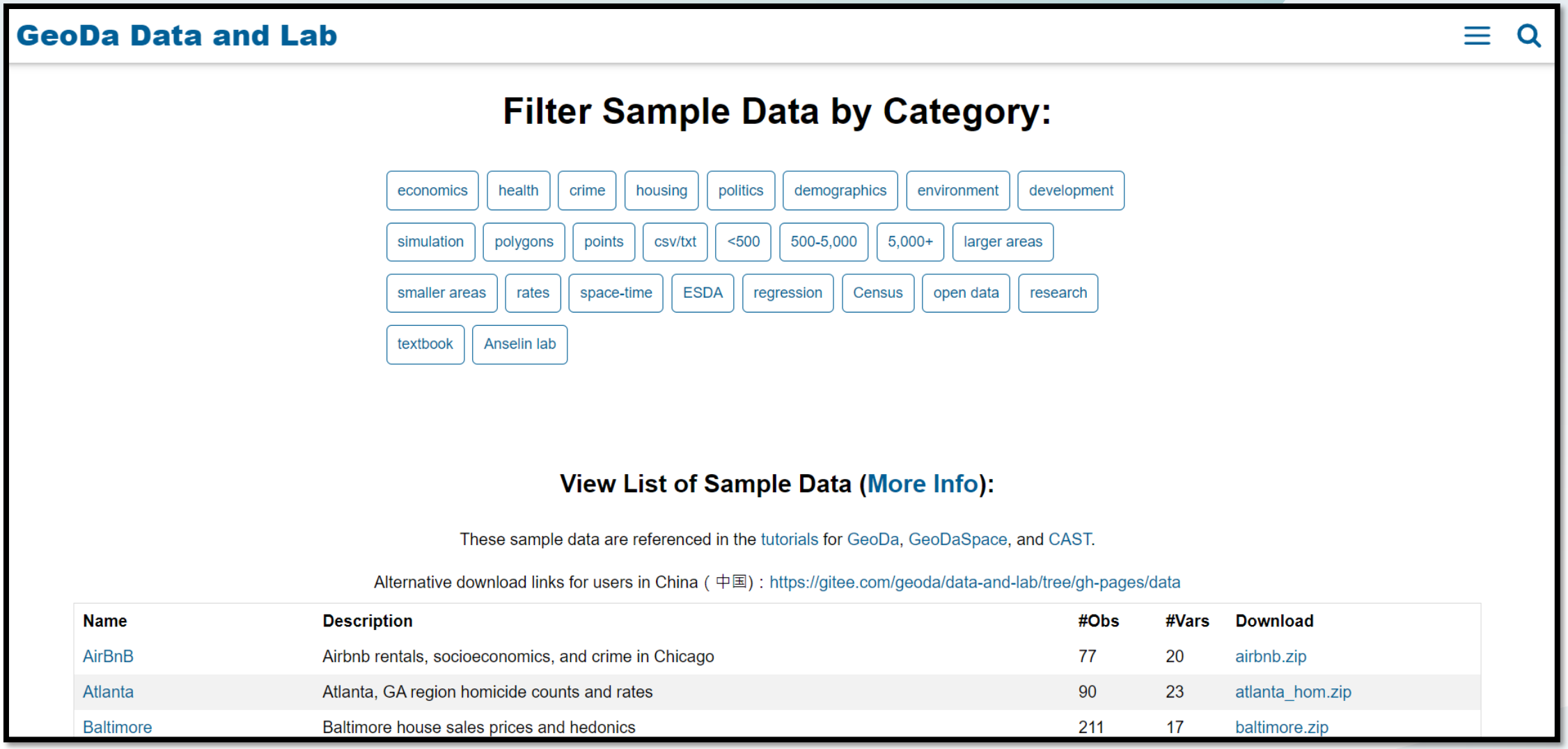
02 雙&單變數地圖

03 操作步驟

04 參考資料

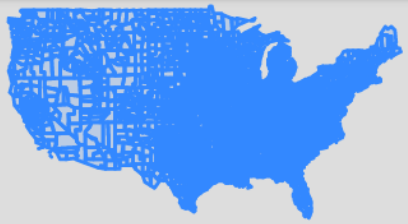
01

資料來源&介紹



資料網站：https://geodacenter.github.io/data-and-lab/

GeoDa Data and Lab



Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

下載資料

DOWNLOAD DATA

Homicides and selected socio-economic characteristics for continental U.S. counties. Data for four decennial census years: 1960, 1970, 1980 and 1990.

- Observations = 3,085
- Variables = 69
- Years = 1960s-90s

3085筆資料
69個變數
1960~1990年

Source: S. Messner, L. Anselin, D. Hawkins, G. Deane, S. Tolnay, R. Baller (2000). An Atlas of the Spatial Patterning of County-Level Homicide, 1960-1990. Pittsburgh, PA, National Consortium on Violence Research (NCOVR).

Reference: Baller, R., L. Anselin, S. Messner, G. Deane and D. Hawkins (2001). Structural covariates of US county homicide rates: incorporating spatial effects. Criminology 39, 561-590.

資料來源：https://geodacenter.github.io/data-and-lab/ncovr/

主要變數介紹：

Variable	Description
NAME	County name
STATE_NAME	State name
STATE_FIPS	State fips code (character)
CNTY_FIPS	County fips code (character)
FIPS	Combined state and county fips code (character)
STFIPS	State fips code (numeric)
COFIPS	County fips code (numeric)
FIPSNO	Fips code as numeric variable
SOUTH	Dummy variable for Southern counties (South = 1)
HR**	Homicide rate per 100,000 (1960, 1970, 1980, 1990)
HC**	Homicide count, three year average centered on 1960, 1970, 1980, 1990
PO**	County population the years 1960, 1970, 1980, 1990
RD**	Resource deprivation the years 1960, 1970, 1980, 1990 (principal component, see Codebook for details)
PS**	Population structure 1960, 1970, 1980, 1990 (principal component, see Codebook for details)
UE**	Unemployment rate the years 1960, 1970, 1980, 1990
DV**	Divorce rate the years 1960, 1970, 1980, 1990 (percentage of males over 14 divorced)
MA**	Median age the years 1960, 1970, 1980, 1990
POL**	Log of population the years 1960, 1970, 1980, 1990
DNL**	Log of population density the years 1960, 1970, 1980, 1990
MFIL**	Log of median family income the years1960, 1970, 1980, 1990
FP**	Percentage of families below poverty the years 1960, 1970, 1980, 1990 (see Codebook for details)
BLK**	Percentage of black population for the years 1960, 1970, 1980, 1990
GI**	Gini index of family income inequality 1960, 1970, 1980, 1990
FH**	Percentage of female headed households 1960, 1970, 1980, 1990

縣名

州名

每10萬人凶殺率(n起事件)

凶殺案數，1960~90年

各縣人口數，1960~90年

失業率，1960~90年

離婚率，1960~90年

02

單&雙變數地圖

- 非地圖變數：

  凶殺率，離婚率皆為連續型資料

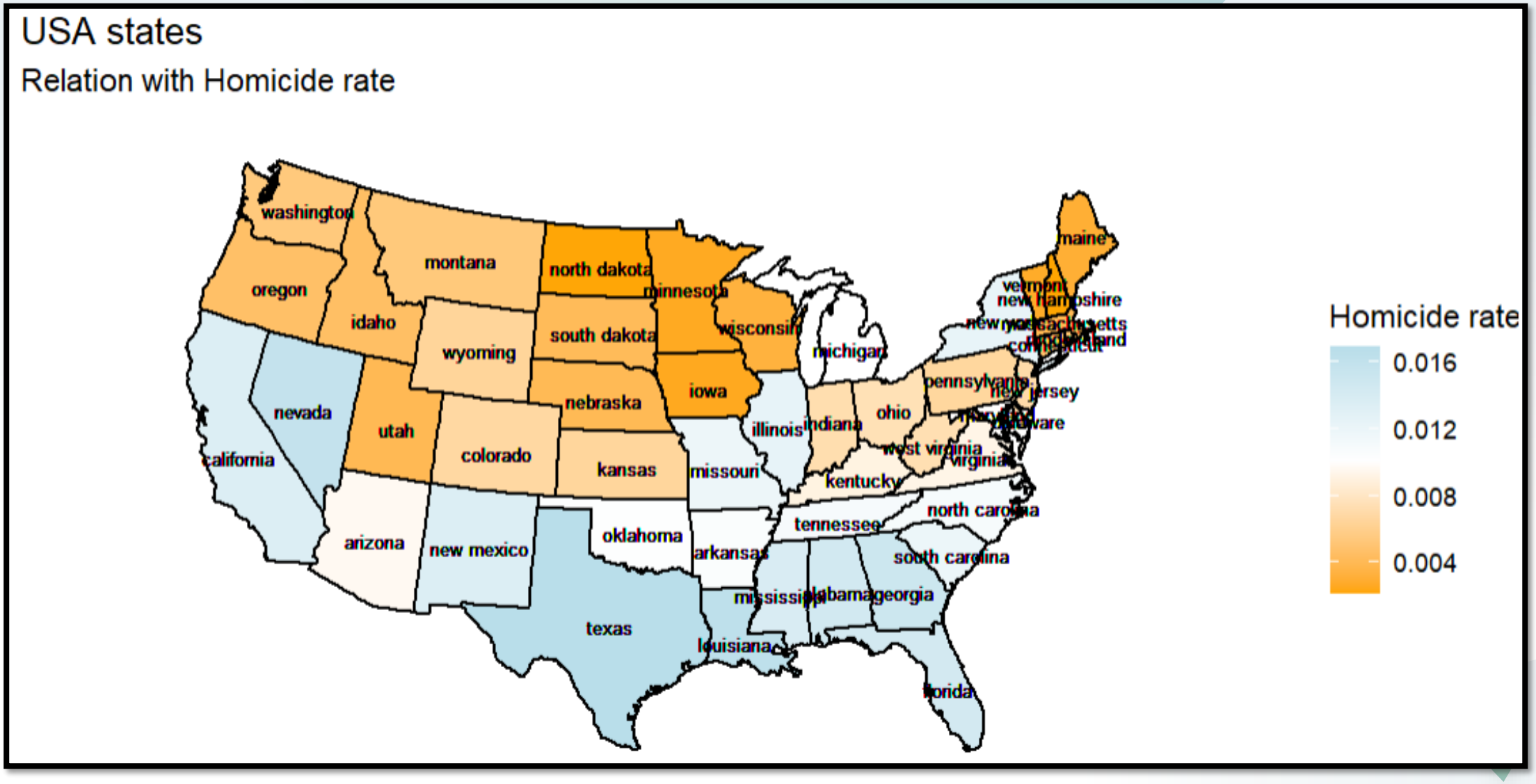
NOTE：比率是連續型資料，用漸層色是個好選擇

- 地圖格式：

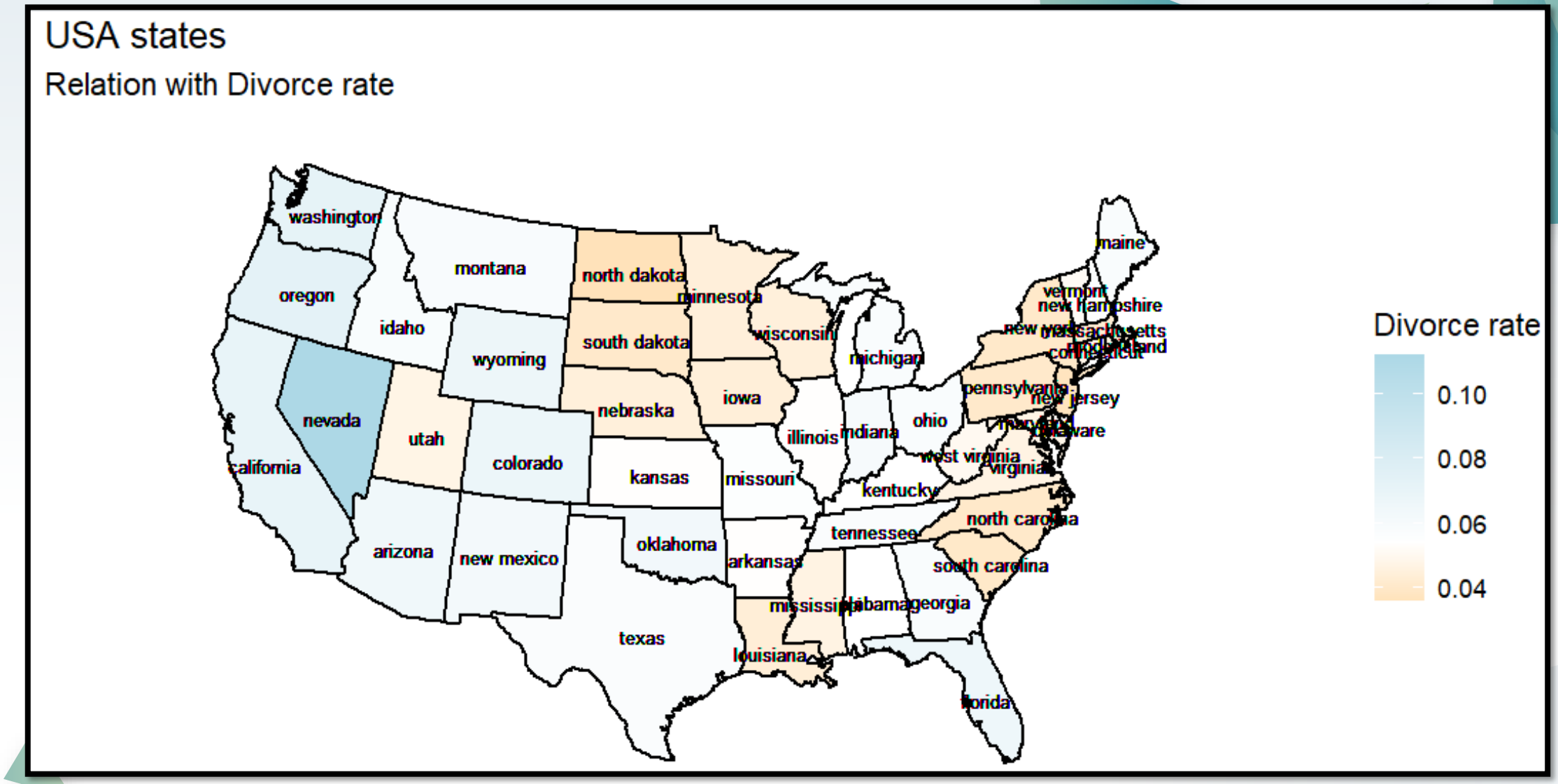
  Choropleth map，又稱等值線圖，面量圖

  把資料用顏色畫在對應地圖上的一種資料視覺化方式

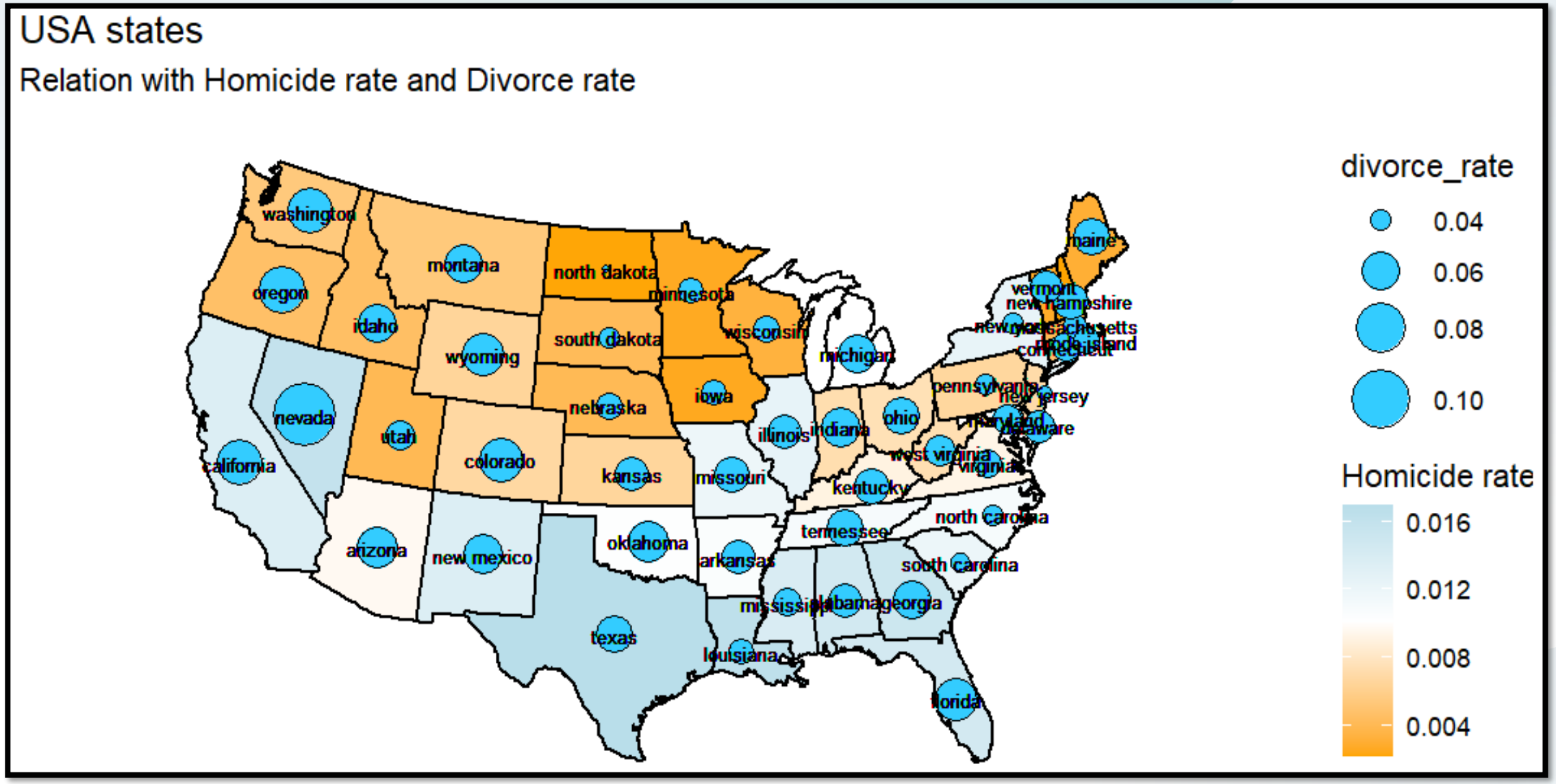
單變數：1980年各州凶殺率



單變數：1980年各州離婚率



雙變數：1980年各州凶殺率跟離婚率的關係



結論：

- 表格 < 雙變數地圖 < 兩張單變數地圖放一起
- 地圖呈現不一定比較好，不過能發現其他東西

  EX：南方凶殺率高，西部離婚率高
- 單看這兩變數關係，看相關係數最快，畫散佈圖

03

操作步驟

R

```
# install.packages('maps')
# install.packages("ggrepel")
library(maps) # For map data
library(ggplot2) # ggplot2 must be loaded to use map_data()
library(dplyr) # For arrange() function
```

```
states_map <- map_data("state") # Get map data for USA
str(states_map)
```

```
## 'data.frame':    15537 obs. of  6 variables:
##  $ long     : num  -87.5 -87.5 -87.5 -87.5 -87.6 ...
##  $ lat      : num  30.4 30.4 30.4 30.3 30.3 ...
##  $ group    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ order    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ region   : chr  "alabama" "alabama" "alabama" "alabama" ...
##  $ subregion: chr  NA NA NA NA ...
```

```
#'usa' (美國的輪廓)
#'state' (美國的每個州)
#'county' (美國的每個縣)
```

```
mapdata = read.csv("C:/Users/user/Desktop/southorAll/ncovr/NAT.csv")
#str(mapdata)

#來源:https://geodacenter.github.io/data-and-lab/ncovr/
```

```r
new_mapdata = mapdata[c('STATE_NAME','HR80','DV80','HC80','PO80','UE80')]

# STATE_NAME 改 region 再轉小寫方便合併
names(new_mapdata)[1] = "region"
new_mapdata$region =  tolower(new_mapdata$region)

str(new_mapdata)
```

```
## 'data.frame':    3085 obs. of  6 variables:
##  $ region: chr  "minnesota" "washington" "washington" "washington" ...
##  $ HR80  : num  8.86 17.21 3.45 3.26 7.77 ...
##  $ DV80  : num  3.75 6.63 5.45 7.12 5.29 ...
##  $ HC80  : num  0.333 1 1 1 0.667 ...
##  $ PO80  : int  3764 5811 28979 30639 8580 7289 17752 51966 10628 5559 ...
##  $ UE80  : num  5.9 15.4 13.6 12.7 18.1 ...
```

```r
summary(new_mapdata)
```

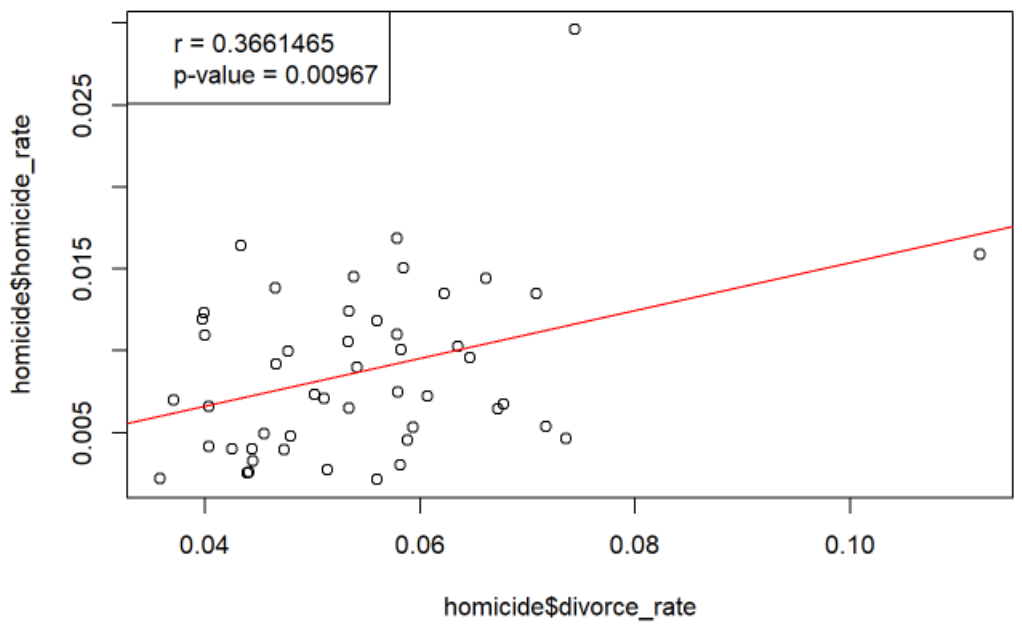
```
##     region              HR80             DV80             HC80        
##  Length:3085        Min.   : 0.000   Min.   : 0.7174   Min.   :   0.0000  
##  Class :character   1st Qu.: 1.933   1st Qu.: 3.7032   1st Qu.:   0.3333  
##  Mode  :character   Median : 5.226   Median : 4.4807   Median :   1.3333  
##                     Mean   : 6.928   Mean   : 4.6074   Mean   :   7.4786  
##                     3rd Qu.:10.229   3rd Qu.: 5.3274   3rd Qu.:   3.3333  
##                     Max.   :59.134   Max.   :18.3673   Max.   :1756.6667  
##       PO80              UE80        
##  Min.   :     91   Min.   : 0.000  
##  1st Qu.:  10543   1st Qu.: 4.500  
##  Median :  21763   Median : 6.415  
##  Mean   :  72985   Mean   : 6.782  
##  3rd Qu.:  51029   3rd Qu.: 8.520  
##  Max.   :7477503   Max.   :27.534
```

```r
homicide =
  new_mapdata %>% group_by(region) %>% summarise(homicide_rate = 100*sum(HC80)/sum(PO80),divorce_rate =
 sum(DV80*PO80/100)/sum(PO80),unemployment_rate = sum(UE80*PO80/100)/sum(PO80))
```

```r
# 美國凶殺率和離婚率的相關係數
# 0.3 ~ 0.7:中等相關

cor.test(homicide$divorce_rate,homicide$homicide_rate)
```

```
##
##  Pearson's product-moment correlation
##
## data:  homicide$divorce_rate and homicide$homicide_rate
## t = 2.6975, df = 47, p-value = 0.00967
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09470016 0.58691483
## sample estimates:
##       cor
## 0.3661465
```

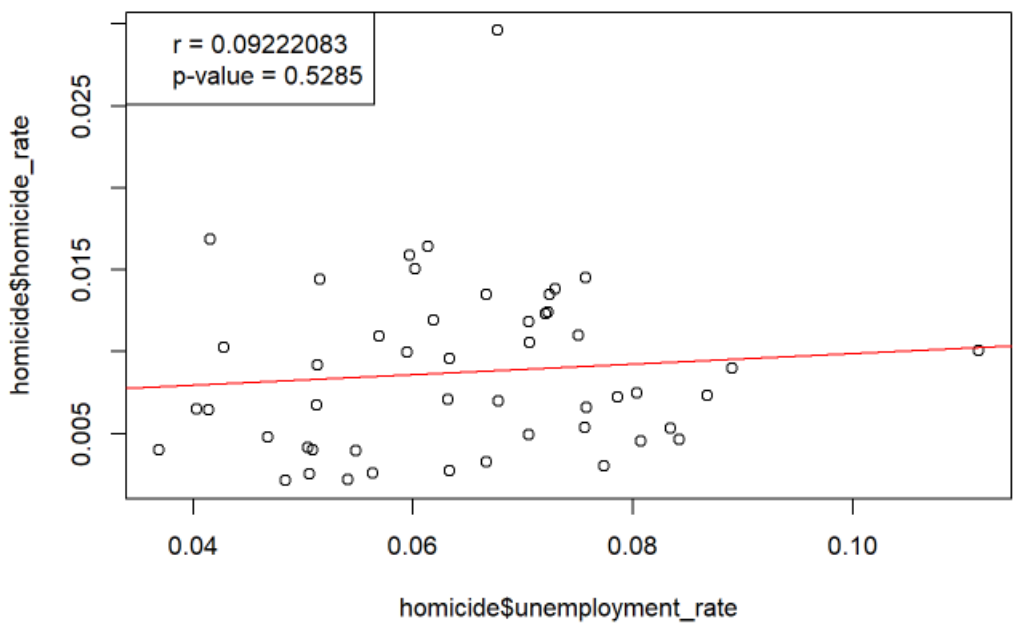
```r
plot(homicide$divorce_rate,homicide$homicide_rate)
abline(lm(homicide$homicide_rate~homicide$divorce_rate),col='red')
legend('topleft', legend = c('r = 0.3661465', 'p-value = 0.00967'))
```



```r
# 美國凶殺率和失業率的相關係數
# p-value = 0.5285，沒顯著相關

cor.test(homicide$unemployment_rate,homicide$homicide_rate)
```

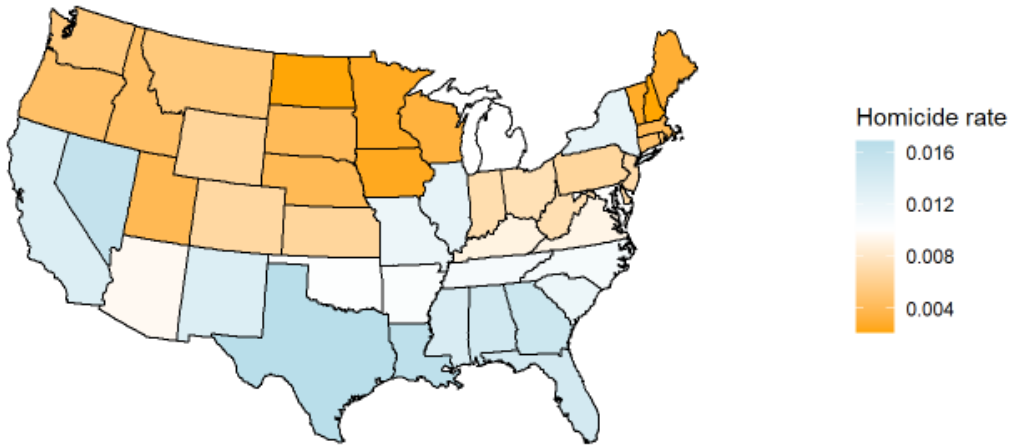
```
##
##  Pearson's product-moment correlation
##
## data:  homicide$unemployment_rate and homicide$homicide_rate
## t = 0.63494, df = 47, p-value = 0.5285
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1940068  0.3639787
## sample estimates:
##       cor
## 0.09222083
```

```r
plot(homicide$unemployment_rate,homicide$homicide_rate)
abline(lm(homicide$homicide_rate~homicide$unemployment_rate),col='red')
legend('topleft', legend = c('r = 0.09222083', 'p-value = 0.5285'))
```



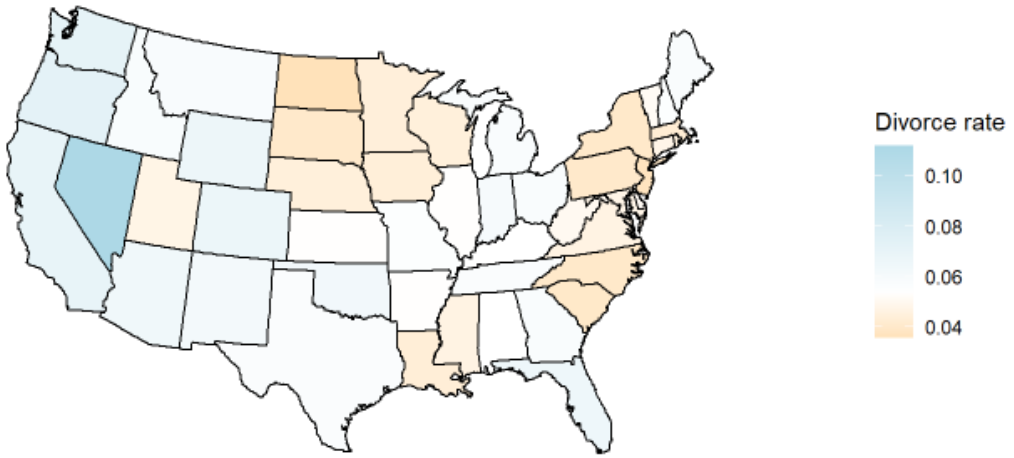
```r
# Merge the data sets together
finaldata = merge(states_map, homicide, by = "region")

# 取得美國各州中心座標資料
StateCenter = data.frame(region=tolower(state.name),lon=state.center$x,lat=state.center$y)
finaldata = merge(finaldata, StateCenter, by = 'region')

# 合併後，順序發生了變化，會導致多邊形繪製的順序不正確，所以將對數據進行排序
# Sort by group, then order
finaldata <- arrange(finaldata, group, order)
```

```r
# 各州兇殺率
ghc =
  ggplot(finaldata, aes(x = long, y = lat.x, group = group, fill = homicide_rate)) +
  geom_polygon(colour = "black") +
  coord_map("polyconic")+
  scale_fill_gradient2(low = "orange", mid = "white", high = "light blue",
                       midpoint = median(finaldata$homicide_rate))+
  labs(fill = "Homicide rate")+
  theme_void()
ghc
```




```r
# 多圓錐投影:平行線為非同心圓弧，但赤道為直線
# theme_void():刪除背景元素
# scale_colour_gradient2():三色梯度，順序為低-中-高
# 中點預設值是0，可以用引數midpoint 將其設定為任意值
```

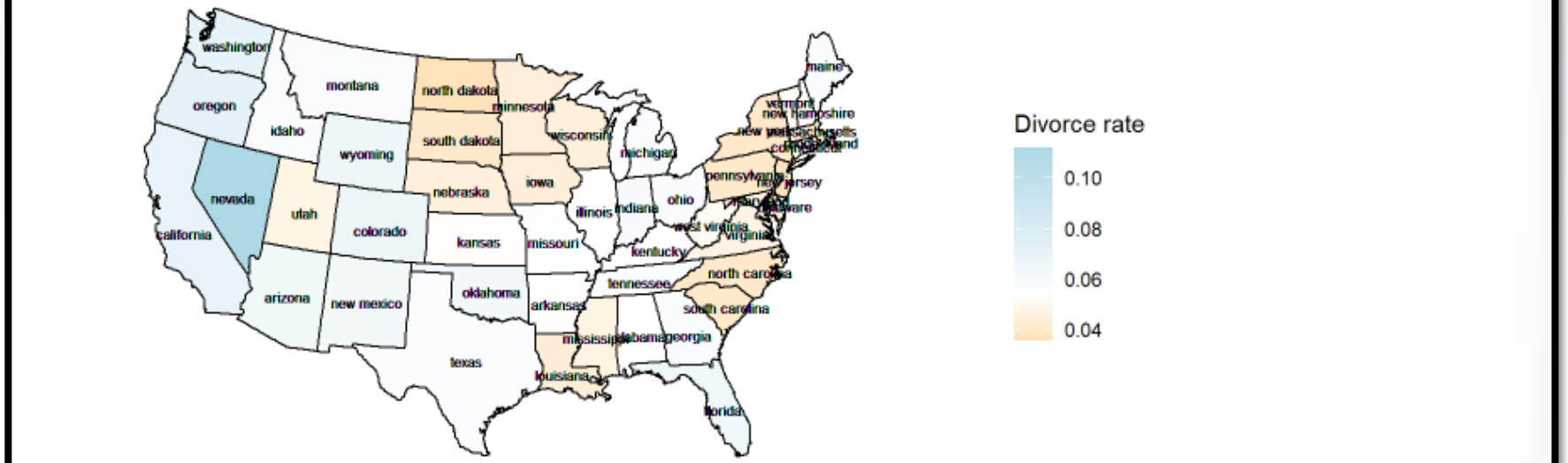
```r
#各州離婚率
gdv =
  ggplot(finaldata, aes(x = long, y = lat.x, group = group, fill = divorce_rate)) +
  geom_polygon(colour = "black") +
  coord_map("polyconic")+
  scale_fill_gradient2(low = "orange", mid = "white", high = "light blue",
                       midpoint = median(finaldata$divorce_rate))+
  labs(fill = "Divorce rate")+
  theme_void()
gdv
```



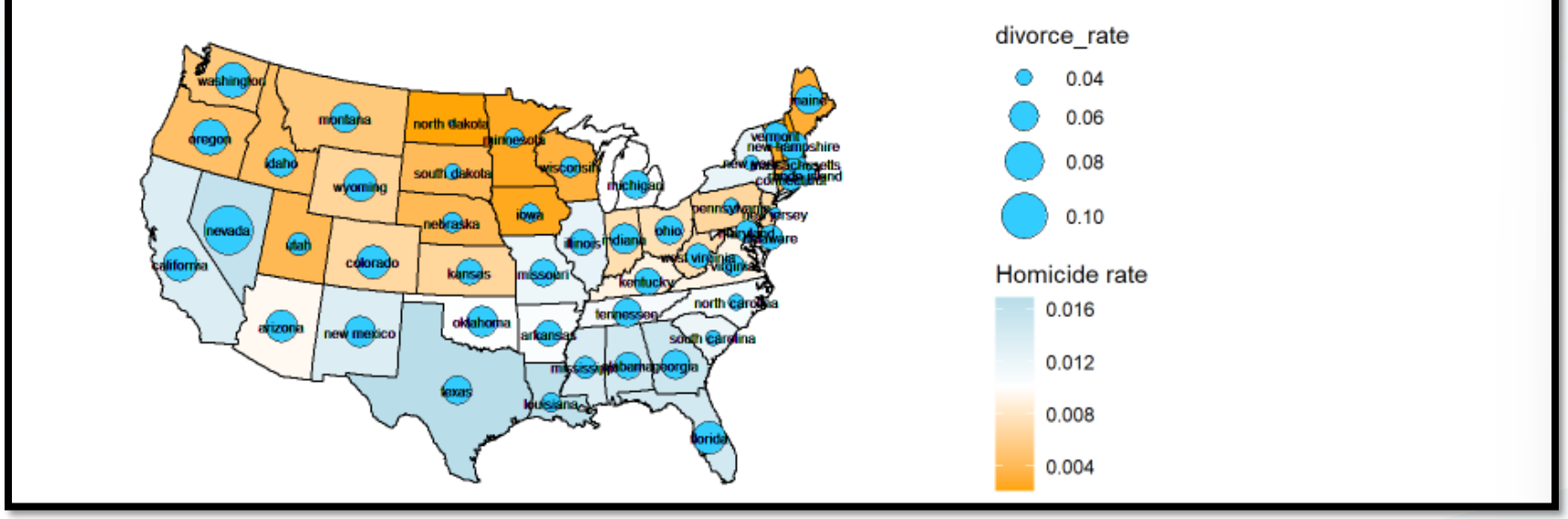

```r
# 多圓錐投影:平行線為非同心圓弧，但赤道為直線

# scale_colour_gradient2():三色梯度，順序為低-中-高
# 中點預設值是0，可以用引數midpoint 將其設定為任意值
```

```
GHC =
  ghc +
  labs(title = 'USA states',
       subtitle = "Relation with Homicide rate")+
  geom_text(data = finaldata, aes(x = lon, y = lat.y, label = region),size = 2.2)
GHC
```

```
GDV =
  gdv +
  labs(title = 'USA states',
       subtitle = "Relation with Divorce rate")+
  geom_text(data = finaldata, aes(x = lon, y = lat.y, label = region),size = 2.2)
GDV
```



```
finalgraph =
  ghc +
  labs(title = 'USA states',
       subtitle = "Relation with Homicide rate and Divorce rate")+
  geom_point(data=finaldata, aes(x=lon, y=lat.y, size=divorce_rate),shape=21,fill="#33CCFF", color = 'black')+
  scale_size(range=c(1,10))+
  geom_text(data = finaldata, aes(x = lon, y = lat.y, label = region),size = 2.2)
finalgraph
```



04

參考資料

程式主要參考資料

- https://r-graphics.org/recipe-miscgraph-map
- https://r-graphics.org/recipe-colors-palette-
- continuous#RECIPE-COLORS-PALETTE-CONTINUOUS
- https://www.rdocumentation.org/
- http://www.sthda.com/english/wiki/ggplot2-point-shapes
- https://yijutseng.github.io/DataScienceRBook/

THANK YOU

謝謝大家