全球各國自殺率分析

資料集

- 欄位名稱:

[1] "country" 國家名稱 "year" 年份 "sex" 性別 "age" 年齡層 "suicides_no" 自殺人數

[6] "population" 人口數 "gdp_for_year" GDP總值 "gdp_per_capita" 人均收入 "generation" 世代

country	year	sex	age	suicides_no	population	suicides/100k pop	country-yea	HDI for yea	gdp_for_ye	gdp_per_ca	generation
Albania	1987	male	15-24 years	21	312900	6.71	Albania1987		#########	796	Generation X
Albania	1987	male	35-54 years	16	308000	5.19	Albania1987		#########	796	Silent
Albania	1987	female	15-24 years	14	289700	4.83	Albania1987		#########	796	Generation X
Albania	1987	male	75+ years	1	21800	4.59	Albania1987		#########	796	G.I. Generation
Albania	1987	male	25-34 years	9	274300	3.28	Albania1987		#########	796	Boomers
Albania	1987	female	75+ years	1	35600	2.81	Albania1987		#########	796	G.I. Generation
Albania	1987	female	35-54 years	6	278800	2.15	Albania1987		#########	796	Silent
Albania	1987	female	25-34 years	4	257200	1.56	Albania1987		#########	796	Boomers
Albania	1987	male	55-74 years	1	137500	0.73	Albania1987		#########	796	G.I. Generation
Albania	1987	female	5-14 years	0	311000	0	Albania1987		#########	796	Generation X
Albania	1987	female	55-74 years	0	144600	0	Albania1987		#########	796	G.I. Generation
Albania	1987	male	5-14 years	0	338200	0	Albania1987		#########	796	Generation X
Albania	1988	female	75+ years	2	36400	5.49	Albania1988		#########	769	G.I. Generation
Albania	1988	male	15-24 years	17	319200	5.33	Albania1988		#########	769	Generation X
Albania	1988	male	75+ years	1	22300	4.48	Albania1988		#########	769	G.I. Generation
Albania	1988	male	35-54 years	14	314100	4.46	Albania1988		#########	769	Silent
Albania	1988	male	55-74 years	4	140200	2.85	Albania1988		#########	769	G.I. Generation
Albania	1988	female	15-24 years	8	295600	2.71	Albania1988		#########	769	Generation X
Albania	1988	female	55-74 years	3	147500	2.03	Albania1988		#########	769	G.I. Generation
Albania	1988	female	25-34 years	5	262400	1.91	Albania1988		#########	769	Boomers
Albania	1988	male	25-34 years	5	279900	1.79	Albania1988		#########	769	Boomers

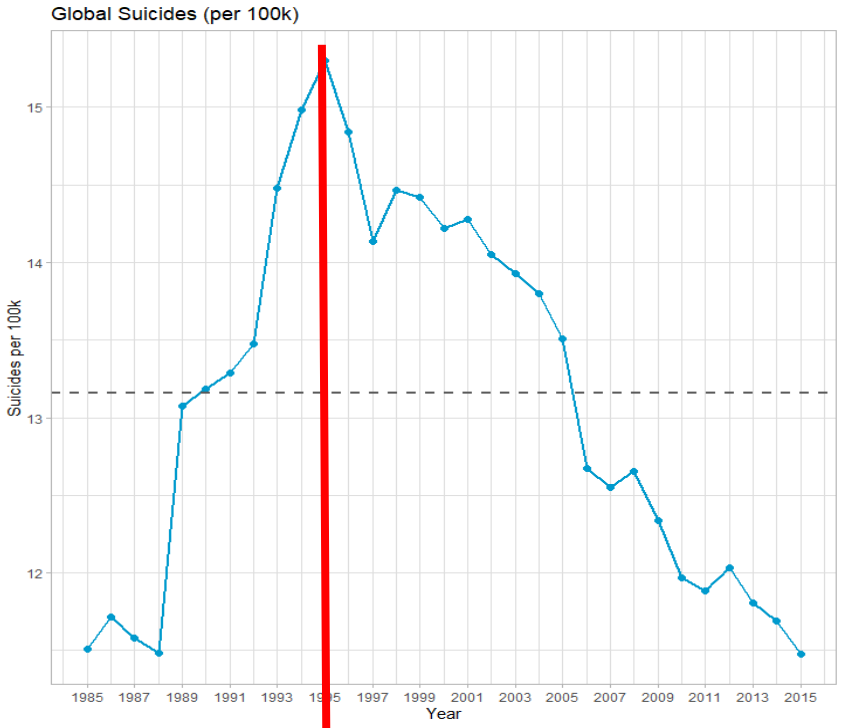
資料前處理

- 刪除 7 個國家: 刪除資料集中<=3年的國家資料
- 刪除 2016 年的數據: 絕大部分國家都缺少2016年的資料，為避免影響整體數據分析，刪除所有2016年的資料
- 刪除 HDI 欄位: 因為有2/3的數據缺失
- 刪除 country-year: 僅將國家和年份合併，為不需要使用到的分析資料
- 重整欄位名稱

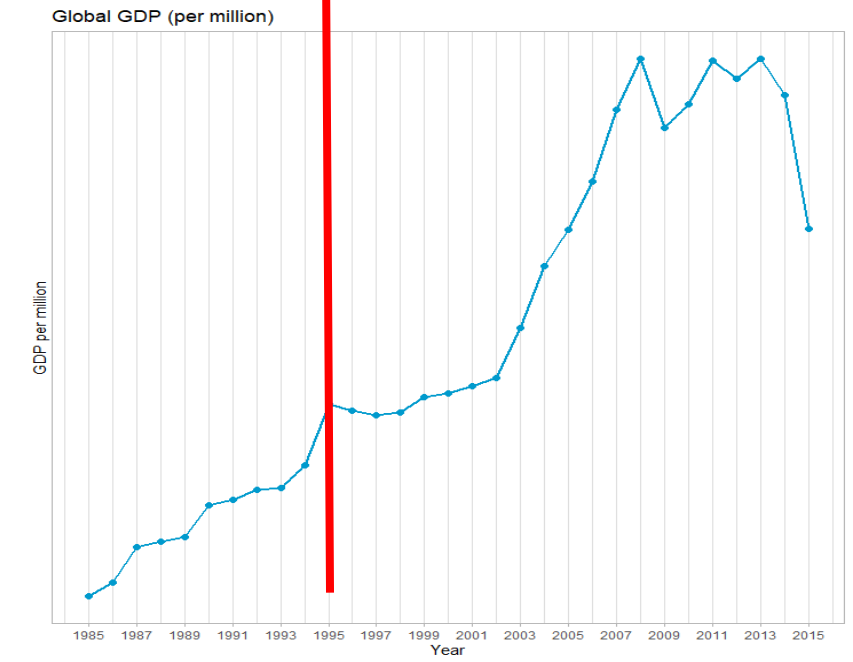
```
> glimpse(data)
Rows: 27,492
Columns: 11
$ country        <fct> Albania, Albania, Albania, Albania, Albania, Albania, Albania, Albania~
$ year           <int> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987~
$ sex            <fct> Male, Male, Female, Male, Male, Female, Female, Female, Male, Female, ~
$ age            <ord> 15-24, 35-54, 15-24, 75+, 25-34, 75+, 35-54, 25-34, 55-74, 5-14, 55-74~
$ suicides_no    <int> 21, 16, 14, 1, 9, 1, 6, 4, 1, 0, 0, 0, 2, 17, 1, 14, 4, 8, 3, 5, 5, 4,~
$ population     <int> 312900, 308000, 289700, 21800, 274300, 35600, 278800, 257200, 137500, ~
$ gdp_for_year   <chr> "2,156,624,900", "2,156,624,900", "2,156,624,900", "2,156,624,900", "2~
$ gdp_per_capita <int> 796, 796, 796, 796, 796, 796, 796, 796, 796, 796, 796, 796, 769, 769, ~
$ generation     <ord> Generation X, Silent, Generation X, G.I. Generation, Boomers, G.I. Gen~
$ continent      <fct> Europe, Europe, Europe, Europe, Europe, Europe, Europe, Europe, Europe~
$ suicides_100k  <dbl> 6.7114094, 5.1948052, 4.8325854, 4.5871560, 3.2810791, 2.8089888, 2.15~
```

全球自殺趨勢

- 實際使用xtabs()分別將全球自殺率和GDP走勢呈現
- 平均自殺率為13.15622（per 100k）
- 從1990年-2005年這段期間的自殺率一直皆高於總平均自殺率
- 其中1995年為自殺率最高峰，為每10萬人中有15.3人死亡
- 隨著GDP走勢提升，全球自殺率也逐步降低，至2015年為每10萬人11.5人（下降約 25%）
- 1990年以前的資料數據有限，因此比率不一定和走勢圖中一樣低



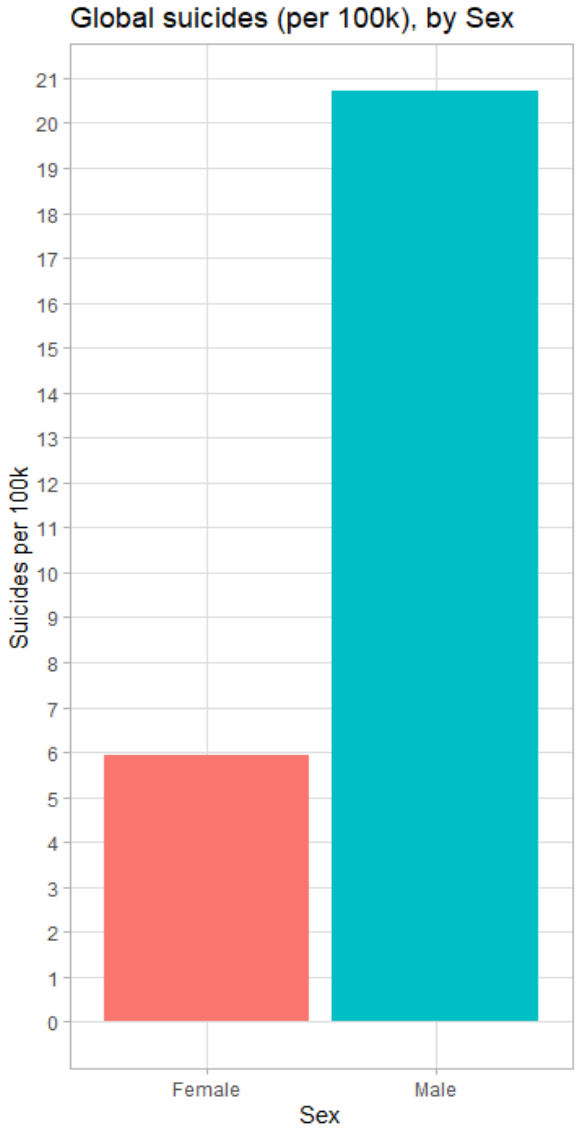
全球自殺率走勢



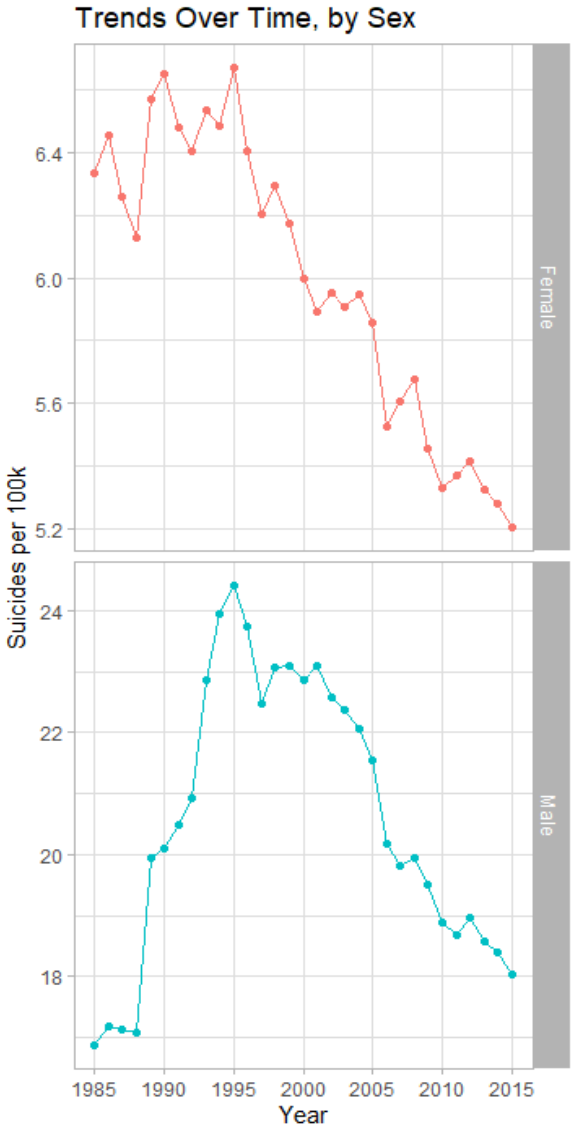
全球GDP值走勢

全球性別自殺比率

- 使用xtabs()分別將男女各自自殺比率和走勢呈現
- 左圖顯示全球的男性自殺人數比女性高出約3.5倍之多
- 右圖顯示男性與女性的自殺高峰期皆為1995年，且隨後逐年下降
- 從右圖發現在90年代中期，開始出現男女自殺人數比1:3.5的現象
- 可以推測男性通常為家庭經濟支柱，因此在生活上有較多的壓力，也是造成自殺率較高的原因



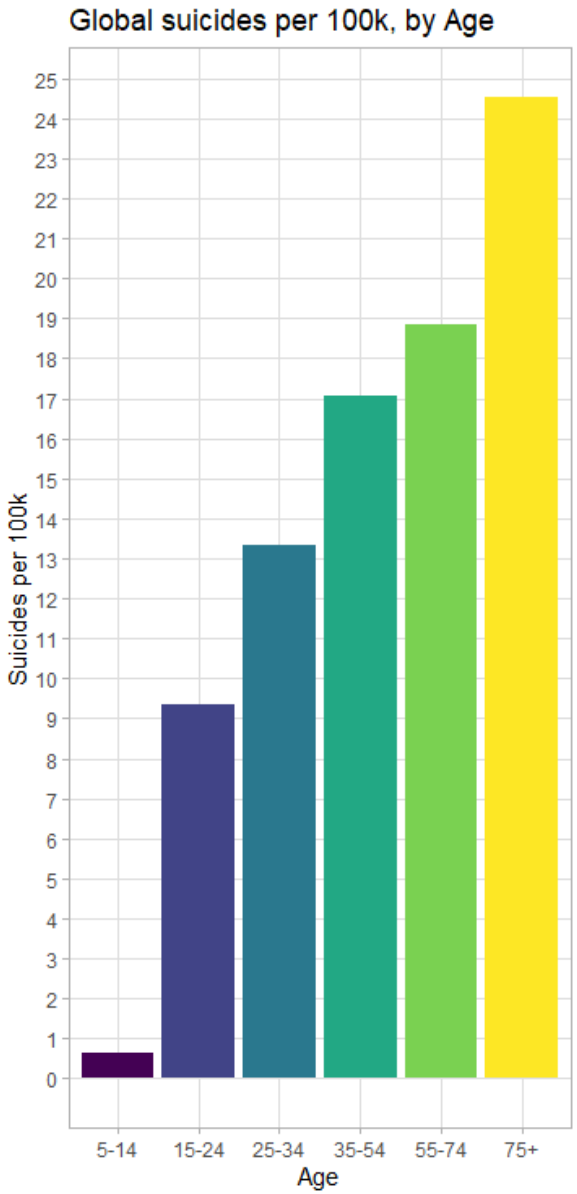
男女自殺比率



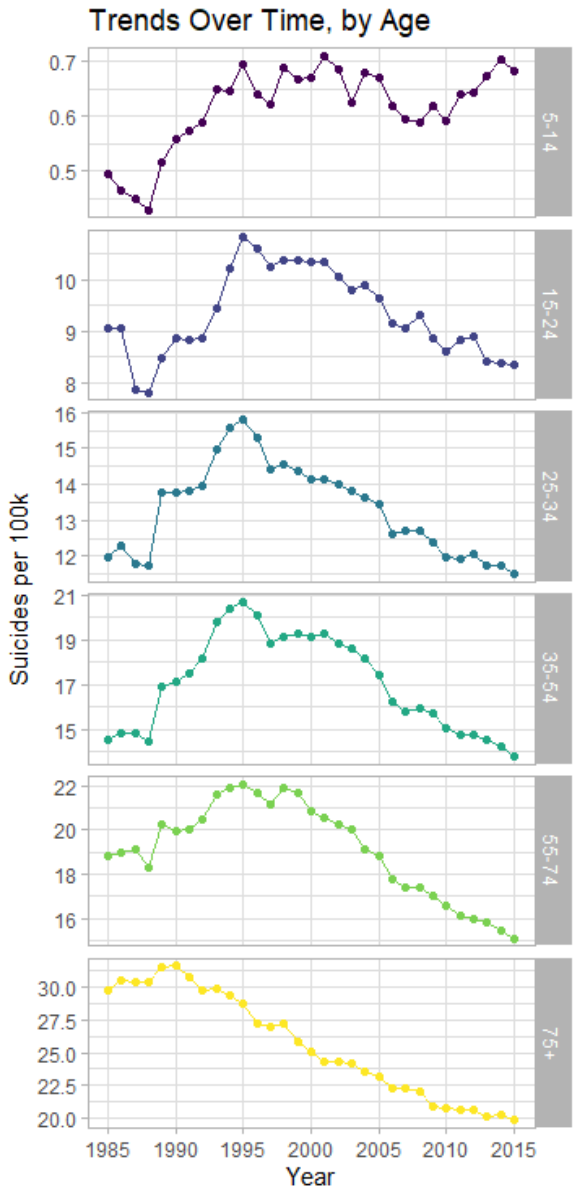
男女自殺比率走勢

全球各年齡層自殺比率

- 使用xtabs()分別將各年齡層比率和走勢呈現
- 從左圖可看出自殺比率隨著年齡層的上升越來越高
- 從右圖可觀察到大於15歲的年齡層，皆在1995年之後有逐年下降的趨勢，可能與社會型態的改變有關，使自殺年齡層降低
- 在高年齡層中75歲以上從1990年開始到2015年間自殺人數下降了將近50%
- 75歲以上自殺率最高，可能和人口數相對較低的原因有關



各年齡層自殺比率



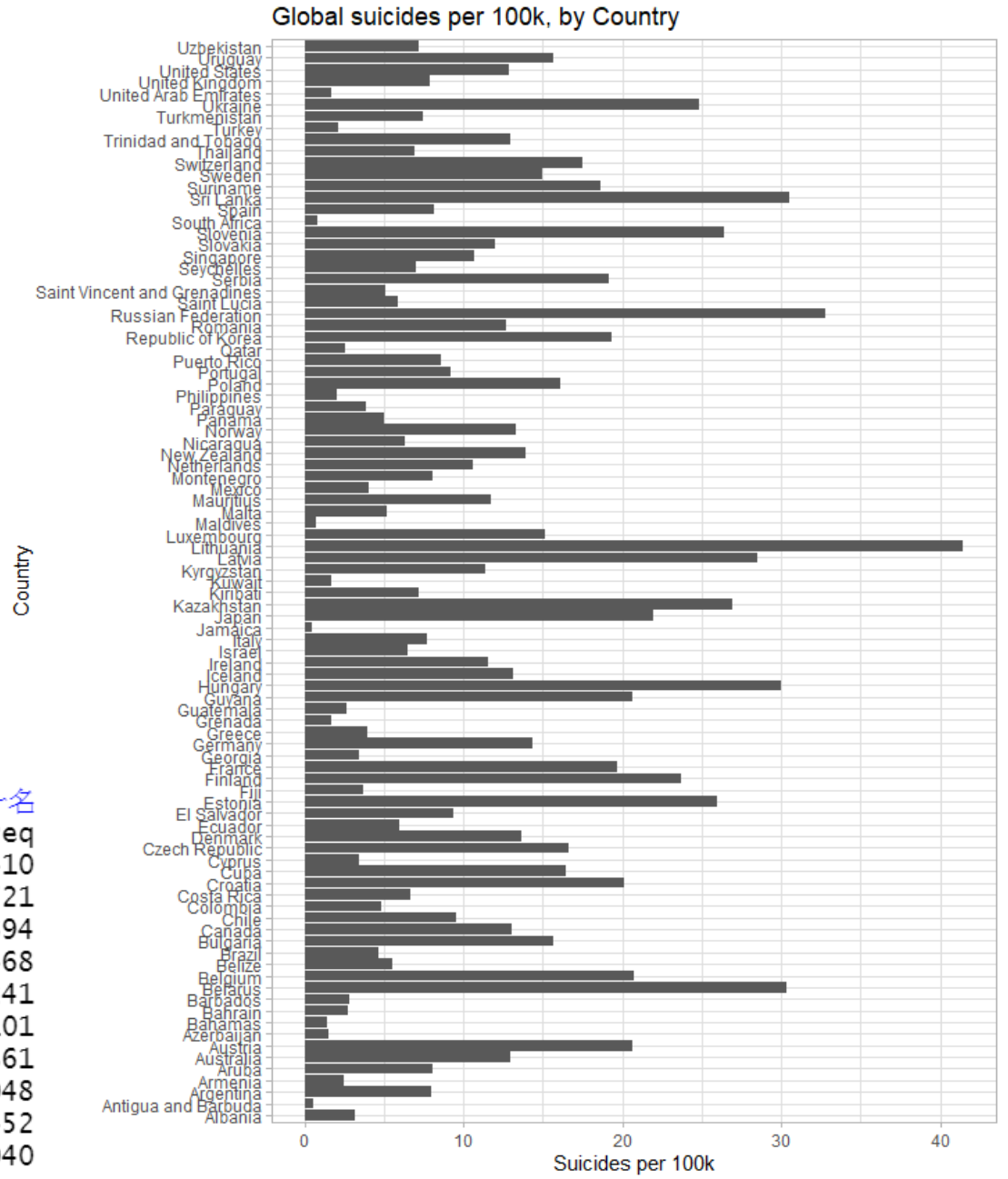
各年齡層自殺比率走勢

各國自殺比率

- 將各國家自殺率進行排名，並顯示前十名國家
- Lithuania相較於其他國家來說高非常多，受環境因素，因日照長年不足，使人患抑鬱症的比例較高
- 前十名高自殺率的國家多為歐洲國家和少許的亞洲國家，多和當地氣候、社會型態有很大的關係
- 斯里蘭卡排名第三，其原因為它是世界最貧窮的國家之一，並有長期的內戰，造成窮人走上自殺這條路

```
> head(country_list,10) #前十名
                country     Freq
1             Lithuania 41.46410
2  Russian Federation 32.77721
3             Sri Lanka 30.48394
4               Belarus 30.34468
5               Hungary 30.02241
6                Latvia 28.47101
7            Kazakhstan 26.89861
8              Slovenia 26.36048
9               Estonia 25.96452
10              Ukraine 24.87040
```

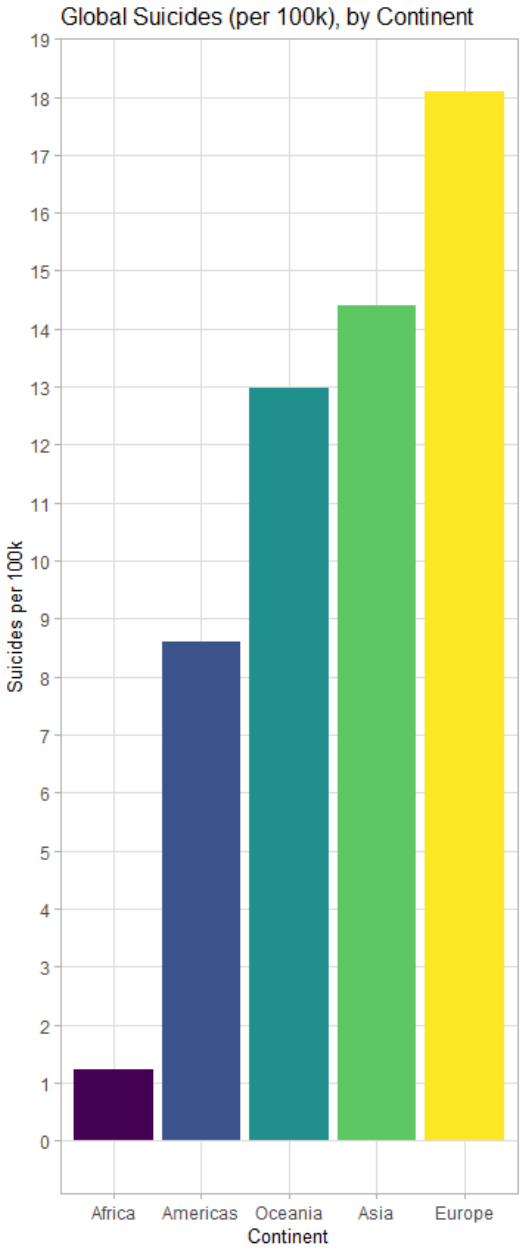
自殺率排名前十名國家



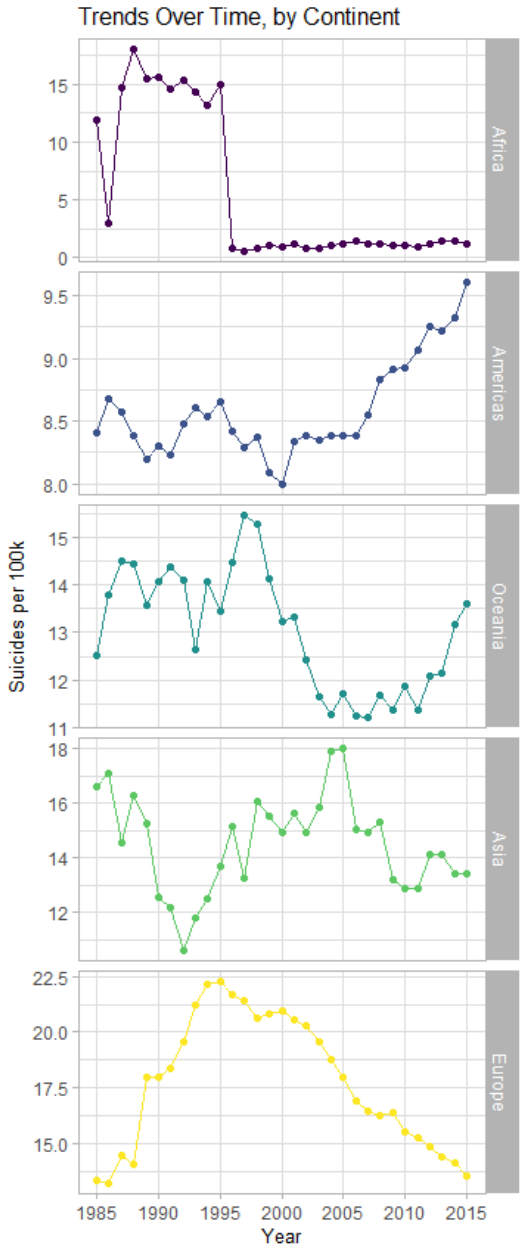
各國自殺比率

各洲自殺比率

- 使用library(countrycode)找到各國家所屬的洲別
- 歐洲佔比最高,且可從右圖得知歐洲從1995年開始自殺人數穩定下降約40%
- 到了2015年時，亞洲和大洋洲的自殺比率逐漸攀升至與歐洲
- 在1995年後非洲國家多不提供自殺人數，因此在之後的趨勢圖較為不準確
- 美洲在1999到2010年之間，經歷了經濟衰退和次貸危機，因此間接造成自殺人數持續飆升



各年齡層自殺比率



各年齡層自殺比率走勢

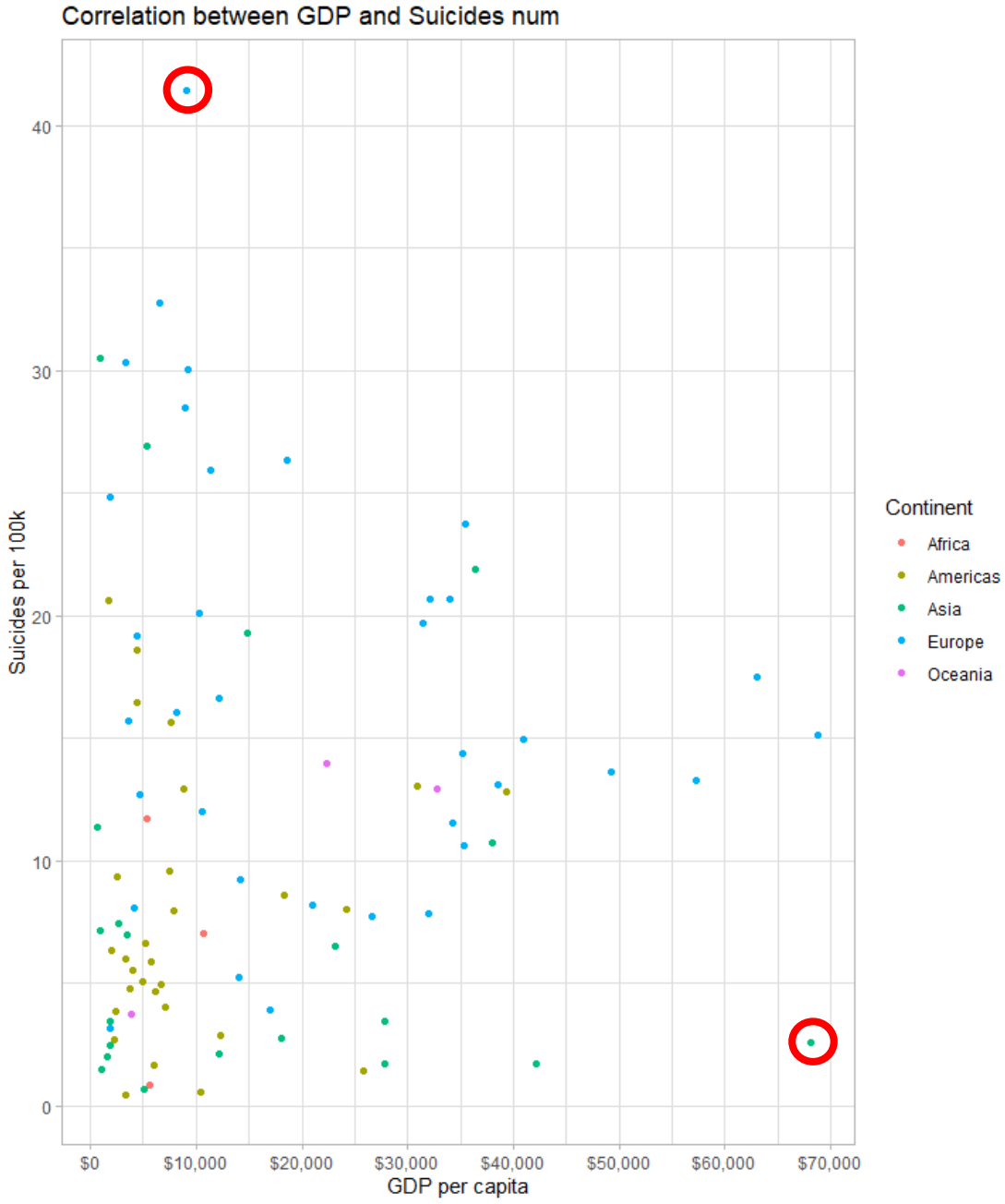
Correlation

- 透過圖表的方式，展現各國家之GDP和自殺比率之間的相關性
- 其中有些許 high leverag 和 residual countries可能會對回歸線產生影響（例如左上角的Lithuania）
- 使用 Cook's Distance 去除異常值，Cook's Distance代表單個樣本對整個回歸模型的影響程度，而Cook's Distance越大，說明此樣本對整體影響越大

- model1 <-

  lm(suicide_per_100k ~ gdp_per_capita,

  data = country_mean_gdp)

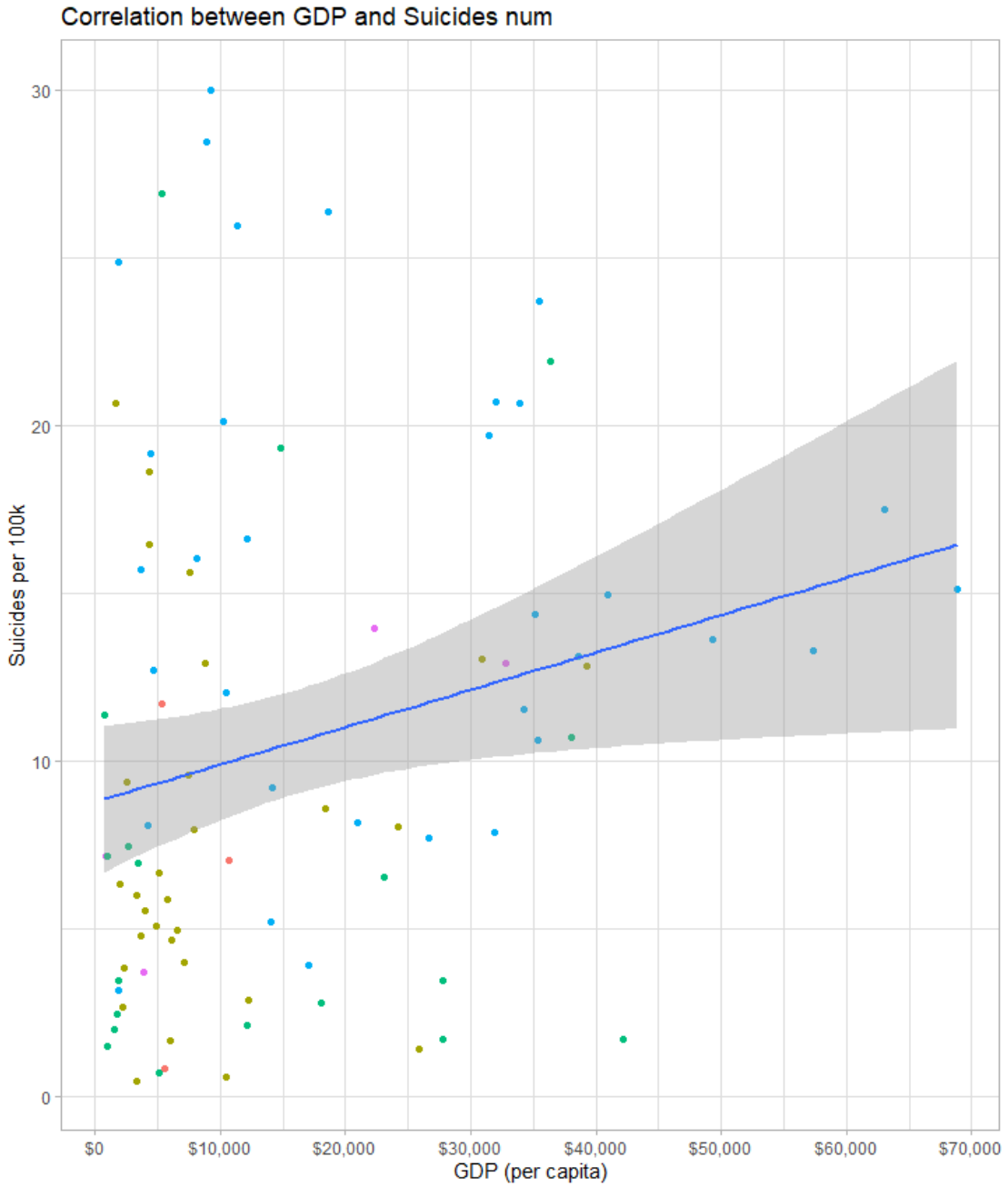


Correlation

- 在國家層面和分析的時間範圍內（1985 年至 2015 年），GDP（人均）增加與每年自殺人數增加有關
- 我們實際去觀察相近的國家，發現有一個微弱但顯著的正線性關係，富裕國家可能造成有更高的自殺率
- 從圖中可以發現回歸線其實是位於平均的位置，仍有許多國家不在灰色區域內，與回歸線距較遠
- model2 <-

  lm(suicide_per_100k ~ gdp_per_capita,

  data = gdp_suicide_no_outliers)



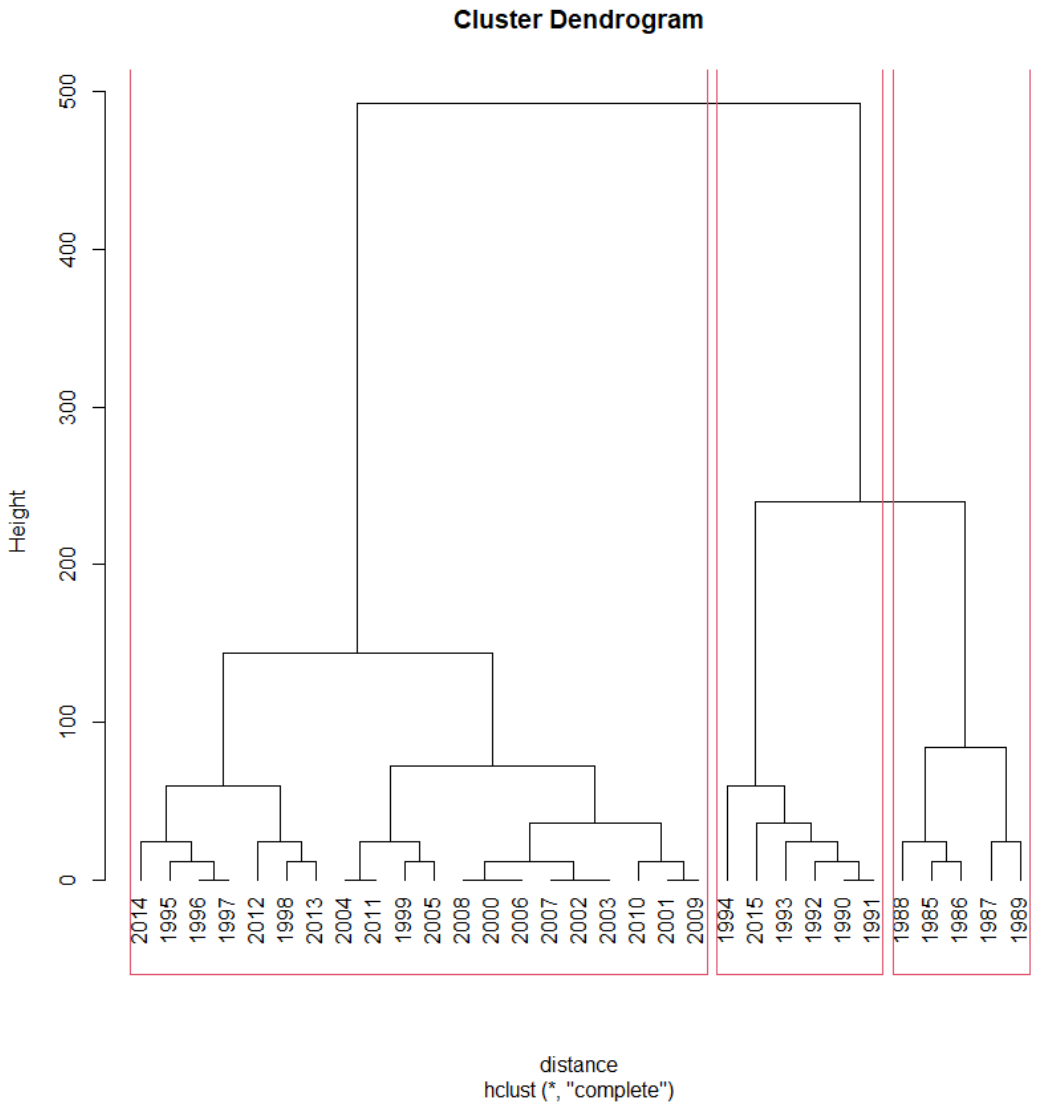
Correlation

- 模型中 p value為 0.0288 < 0.05。這代表可證明一個國家的 GDP（人均）與其自殺率（每 10 萬人）的假設是成立的。
- R-square為 0.05436，因此 GDP（人均）對總體自殺率的變化影響很小

```
> summary(model2)

Call:
lm(formula = suicide_per_100k ~ gdp_per_capita, data = gdp_suicide_no_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-11.769  -5.145  -1.724   3.227  20.221

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.772e+00  1.119e+00   7.839 1.12e-11 ***
gdp_per_capita 1.115e-04  5.015e-05   2.223   0.0288 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.331 on 86 degrees of freedom
Multiple R-squared:  0.05436,   Adjusted R-squared:  0.04337
F-statistic: 4.944 on 1 and 86 DF,  p-value: 0.02881
```

Clustering-Wards Method

- 反覆比較每對資料合併後的群內總變異數的增量，並找增量最小的組別優先合併，越早合併的子集表示其間的相似度越高
- 其中1986-1989年間的自殺人數是穩定的，在GDP走勢圖中也沒有明顯的差異，代表這幾年的條件內容相似，因此歸在同一子集
- 2015年被分在與1991-1994年同一子集，代表2015年在自殺人數和年齡分布等等相關條件與這幾年份有較高的相似度
- 左邊子集裡多為1995年後的年份，和我們先前在分析各趨勢時，多在1995年後有很大的改變，使這些年份被歸為一類



決策樹-rpart

- suicide.rpart <- rpart(suicides_100k ~gdp_per_capita+age+sex+continent , data=train)
- 使用suicides_100k作為分類預測結果，參數分別有GDP值、年齡層、性別、洲別
- 由於原先suicides_100k為連續值，因此我們依summary結果將值分為六區段
- 我們有嘗試在參數中增加其他欄位，但跑超過四小時還沒有結果，所以先針對這些欄位做預測

```
> summary(data)
      country           year          sex           age         suicides_no
 Argentina:  372   Min.   :1985   Female:13746   5-14 :4582   Min.   :    0.0
 Austria  :  372   1st Qu.:1995   Male  :13746   15-24:4582   1st Qu.:    3.0
 Belgium  :  372   Median :2002                  25-34:4582   Median :   25.0
 Brazil   :  372   Mean   :2001                  35-54:4582   Mean   :  244.9
 Chile    :  372   3rd Qu.:2008                  55-74:4582   3rd Qu.:  133.0
 Colombia :  372   Max.   :2015                  75+  :4582   Max.   :22338.0
 (Other)  :25260
   population        gdp_for_year       gdp_per_capita              generation
 Min.   :      278   Length:27492       Min.   :   251   G.I. Generation:2726
 1st Qu.:   99298   Class :character   1st Qu.:  3418   Silent         :6298
 Median :  436562   Mode  :character   Median :  9283   Boomers        :4926
 Mean   : 1861366                      Mean   : 16799   Generation X   :6338
 3rd Qu.: 1503556                      3rd Qu.: 24870   Millenials     :5746
 Max.   :43805214                      Max.   :126352   Generation Z   :1458

    continent     suicides_100k
 Africa  :  828   Min.   :  0.000
 Americas: 9156   1st Qu.:  0.947
 Asia    : 5268   Median :  6.037
 Europe  :11268   Mean   : 12.869
 Oceania :  972   3rd Qu.: 16.678
                  Max.   :224.972
```

```
temp$suicides_100k = cut(temp$suicides_100k,

breaks = c(-1, 3, 6, 12, 15, 50, 225),

 c("< 3", "3-6", "6-12", "12-15","15-50",
"50-225") )
```

決策樹-rpart

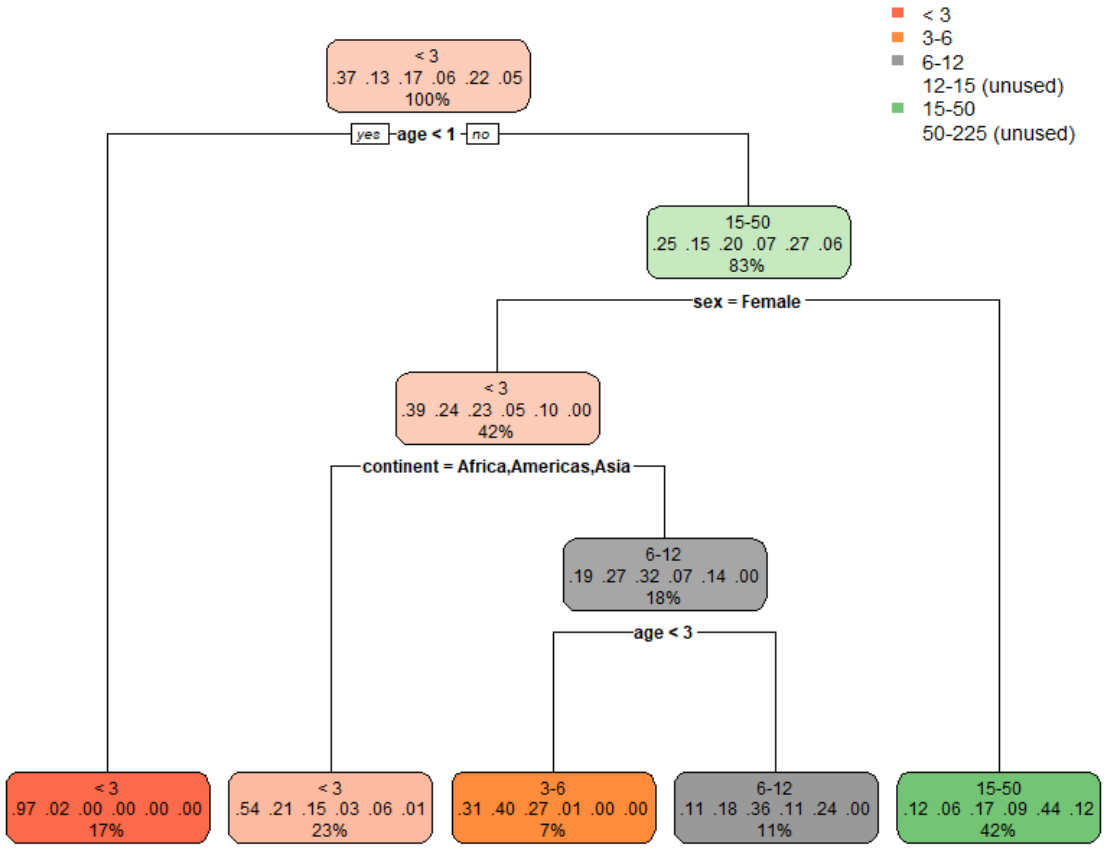
- suicide.rpart <- rpart(suicides_100k ~gdp_per_capita+age+sex+continent , data=train)
- 使用suicides_100k作為分類預測結果，參數分別有GDP值、年齡層、性別、洲別
- 由於原先suicides_100k為連續值，因此我們依summary結果將值分為六區段，作為預測分類結果
- 我們有嘗試在參數中增加其他欄位，但跑超過四小時還沒有結果，所以先針對這些欄位做預測

```
> summary(data)
     country           year          sex           age         suicides_no
 Argentina:  372   Min.   :1985   Female:13746   5-14 :4582   Min.   :    0.0
 Austria  :  372   1st Qu.:1995   Male  :13746   15-24:4582   1st Qu.:    3.0
 Belgium  :  372   Median :2002                 25-34:4582   Median :   25.0
 Brazil   :  372   Mean   :2001                 35-54:4582   Mean   :  244.9
 Chile    :  372   3rd Qu.:2008                 55-74:4582   3rd Qu.:  133.0
 Colombia :  372   Max.   :2015                 75+  :4582   Max.   :22338.0
 (Other)  :25260
   population       gdp_for_year       gdp_per_capita              generation
 Min.   :     278   Length:27492       Min.   :   251   G.I. Generation:2726
 1st Qu.:   99298   Class :character   1st Qu.:  3418   Silent         :6298
 Median :  436562   Mode  :character   Median :  9283   Boomers        :4926
 Mean   : 1861366                      Mean   : 16799   Generation X   :6338
 3rd Qu.: 1503556                      3rd Qu.: 24870   Millenials     :5746
 Max.   :43805214                      Max.   :126352   Generation Z   :1458

    continent     suicides_100k
 Africa  :  828   Min.   :  0.000
 Americas: 9156   1st Qu.:  0.947
 Asia    : 5268   Median :  6.037
 Europe  :11268   Mean   : 12.869
 Oceania :  972   3rd Qu.: 16.678
                  Max.   :224.972
```

```
temp$suicides_100k = cut(temp$suicides_100k,

breaks = c(-1, 3, 6, 12, 15, 50, 225),

 c("< 3", "3-6", "6-12", "12-15","15-50",
"50-225") )
```

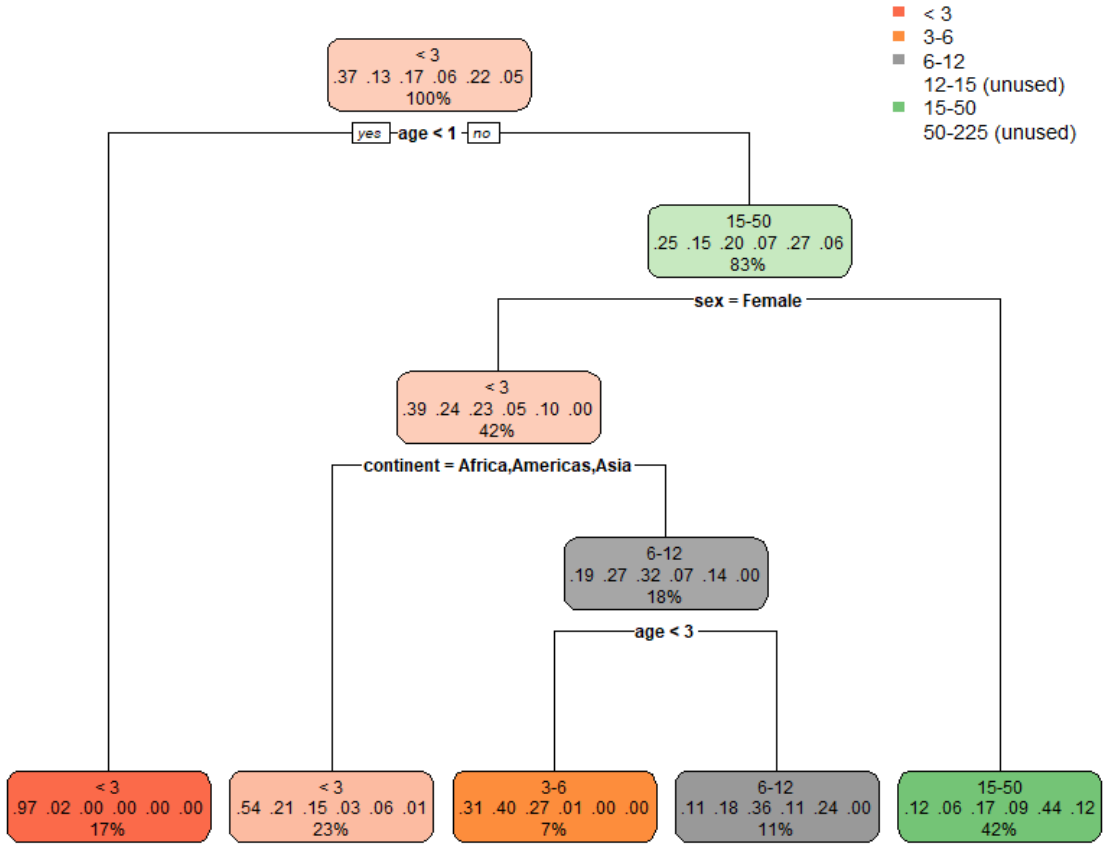
決策樹-rpart

- 年齡層在5-14歲的階段自殺人數佔了將近兩成，其中<3佔0.97，代表介於5-14歲的數據其suicides_100k值會被歸類在<3，也代表成年後的自殺率超過八成，合理推測是因為社會化及經濟等因素造成的
- 年齡超過14歲且是男性的自殺比率超過一半，合理推測男性在生活經濟方面壓力較大，在42%中佔約0.44，所以超過14歲的男性數據多半會分類在15-50
- 在非洲、美洲、及亞洲年齡層在14歲以上的女性自殺率高於大洋洲和歐洲，在23%中又佔0.54，所以年齡大於14歲的非洲、美洲、及亞洲女性suicides_100k值會被歸類在<3



決策樹-rpart

- 其中大於14歲並位於大洋洲和歐洲的女性又可分為介於15-34和35-75+年齡層，各佔約7%(其中3-6佔0.4)和11%(其中6-12佔0.36)
- 可以推測出15-34歲年齡層並位於大洋洲和歐洲的女性其suicides_100k值會被歸類在3-6
- 而35-75+歲年齡層並位於大洋洲和歐洲的女性其suicides_100k值會被歸類在6-12
- 其中suicides_100k值為12-15和50-225的數據，可能因決策樹在分類時其他佔比較高而會歸類在其他類別，造成未出現在分類結果的狀況



決策樹-rpart

- 我們實際使用predict(suicide.rpart, newdata = test, type = "class")測是我們訓練的決策樹模型，但得到的準確率並沒有很高，約0.5406
- 其預測準確率並不高的原因與我們選擇的訓練參數有關，由於選擇非數值的參數需要跑4個小時以上的訓練時間，但也因此我們可以得知在year和country可能有較大的影響，加入整體的訓練參數可以使我們的準確率提高，使分類更加準確
- 而在使用test進行預測的結果中，可以發現在12-15和50-225類別中是完全零的情況，可能的原因為在整體數據中就佔不多的數據，而在亂數分類train和test資料時，test就沒有分到這兩類的測試資料

```
n= 21994

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 21994 13773 < 3 (0.37 0.13 0.17 0.056 0.22 0.051)
   2) age< 0.5 3664   104 < 3 (0.97 0.024 0.0044 0.00027 0 0) *
   3) age>=0.5 18330 13411 15-50 (0.25 0.15 0.2 0.067 0.27 0.061)
     6) sex=Female 9202  5639 < 3 (0.39 0.24 0.23 0.049 0.096 0.0041)
      12) continent=Africa,Americas,Asia 5150  2357 < 3 (0.54 0.21 0.15 0.031 0.058 0.0054)
*
      13) continent=Europe,Oceania 4052  2736 6-12 (0.19 0.27 0.32 0.072 0.14 0.0025)
        26) age< 2.5 1619   971 3-6 (0.31 0.4 0.27 0.014 0.0037 0) *
        27) age>=2.5 2433  1552 6-12 (0.11 0.18 0.36 0.11 0.24 0.0041) *
     7) sex=Male 9128  5090 15-50 (0.12 0.062 0.17 0.085 0.44 0.12) *
```

	預測					
實際	< 3	3-6	6-12	12-15	15-50	50-225
< 3	1526	153	68	0	280	0
3-6	261	147	99	0	135	0
6-12	203	114	253	0	405	0
12-15	47	5	68	0	197	0
15-50	79	2	137	0	1046	0
50-225	7	0	2	0	264	0

```
> accuracy <- sum(diag(cm)) / sum(cm)
> accuracy
[1] 0.5405602
```