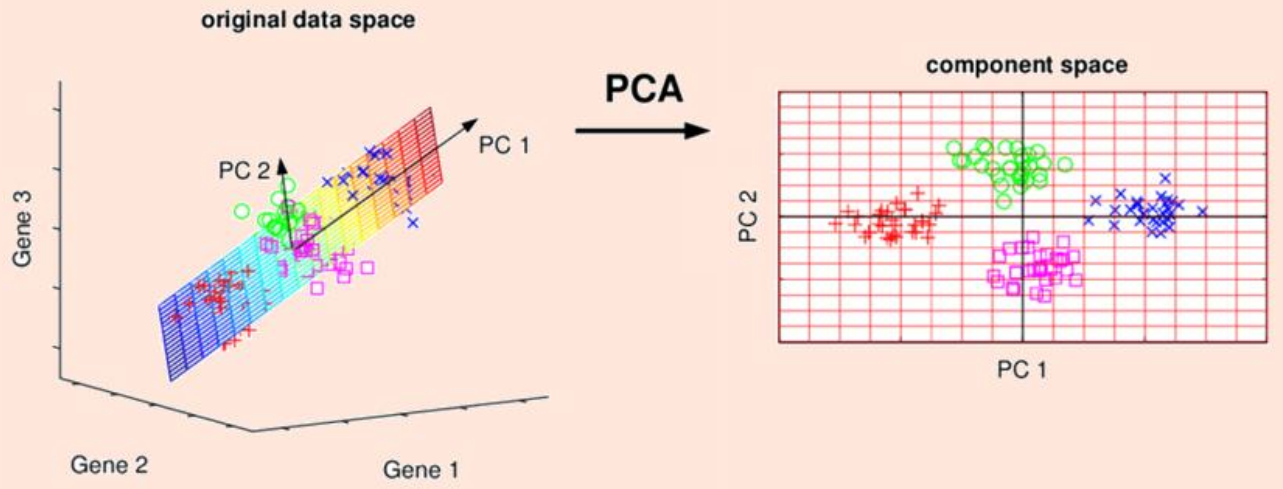
主成分分析

Principal Components Analysis

統計111 陳柔漪

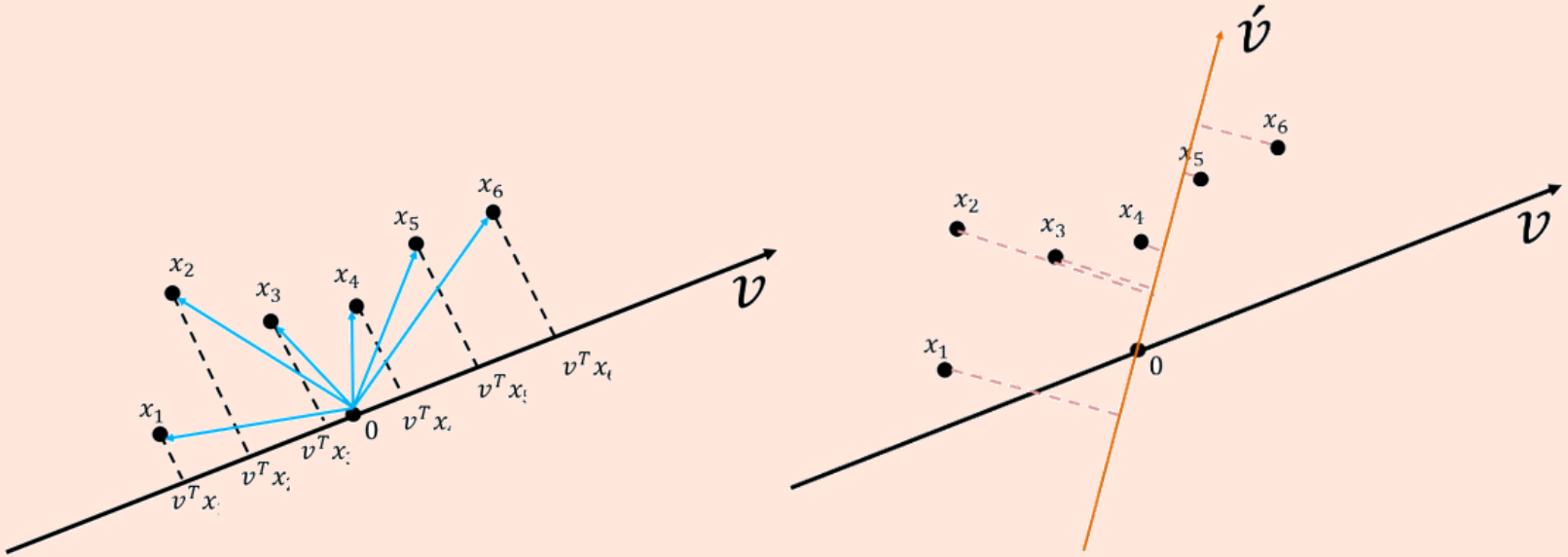
主成分分析的「目的」

- 資料降維
- 用少數變數來描述資料



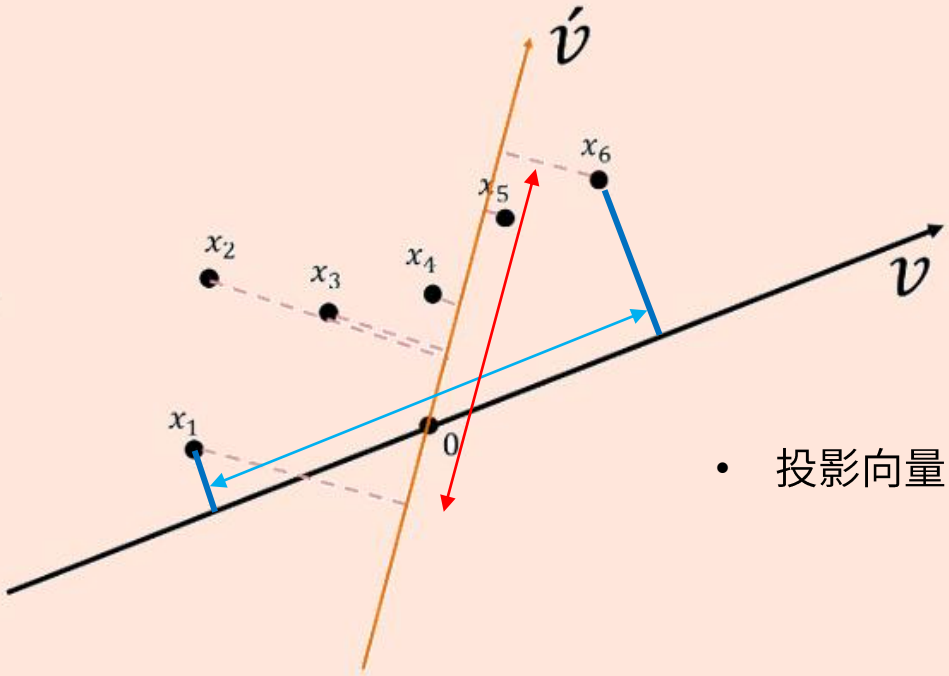
基本思想

- 尋找投影向量，讓資料在投影過後可以維持最大的變異



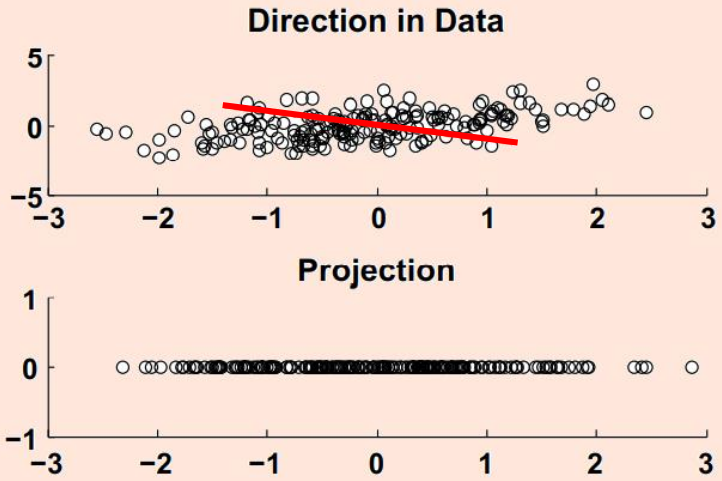
基本思想

- 尋找投影向量，讓資料在投影過後可以維持最大的變異

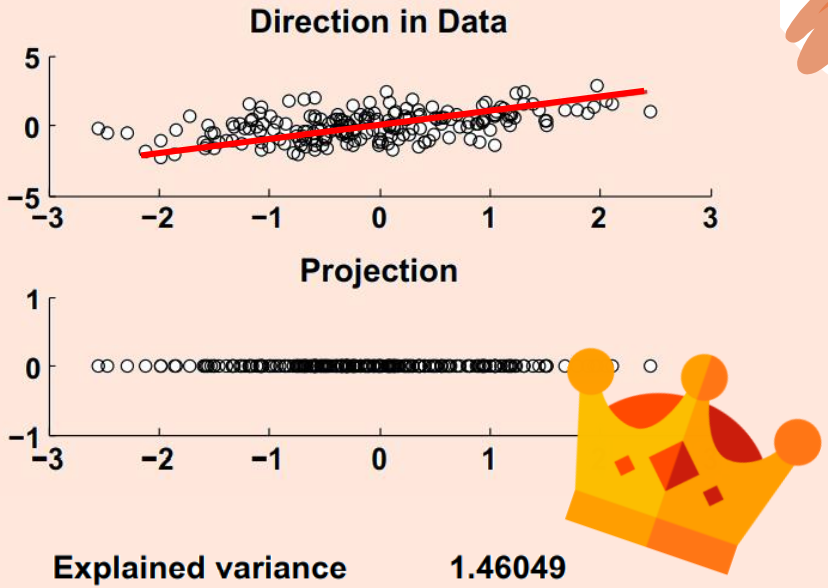


- 投影向量 $\boldsymbol{v}$ (藍色) 變異較大

回顧一下多變量！



Explained variance	0.50520
Explained percentage	0.25701
Total variance	1.96569



Explained variance	1.46049
Explained percentage	0.74299

基本假設

假設我們有一份 n 個樣本、p 個變數的資料 $\boldsymbol{X}$

且此資料的平均數為 $\boldsymbol{EX = \mu}$，變異數為 $\boldsymbol{Var(X) = \Sigma = \Gamma \Lambda \Gamma^T}$

特徵分解
Γ = eigenvector
Λ = eigenvalue

那麼，主成分 Y 的轉換公式就寫做：

$$\boldsymbol{Y = \Gamma^T (X - \mu)}$$

→ 主成分可以看做一種線性轉換

$\boldsymbol{\Gamma}$ (eigenvector) 就是讓資料投影下去會有最大變異量的投影軸。

數學推導

$$\boldsymbol{Y = \Gamma^T (X - \mu)}$$

$$\boldsymbol{EY} = E\left(\Gamma^T (X - \mu)\right) = \Gamma^T E(X - \mu) = \Gamma^T (EX - \mu) = 0$$

$$\boldsymbol{Var(Y)} = Var\left(\Gamma^T (X - \mu)\right) = \Gamma\Lambda\Gamma^T = \Gamma^T\Gamma\Lambda\Gamma^T\Gamma = \Lambda$$

$$\boldsymbol{Cov(Y_i, Y_j)} = \gamma_i^T\, Var(X - \mu)\, \gamma_j = \gamma_i^T\, Var(X)\, \gamma_j = \gamma_i^T\, \Gamma\Lambda\Gamma^T\, \gamma_j = \begin{cases} 0, & i \neq j \\ \lambda_i, & i = j \end{cases}$$

→ 主成分之間彼此正交，且 $\boldsymbol{\Lambda}$ (eigenvalue) 是主成分的變異數

實作

World Happiness Report 2021

世界幸福報告

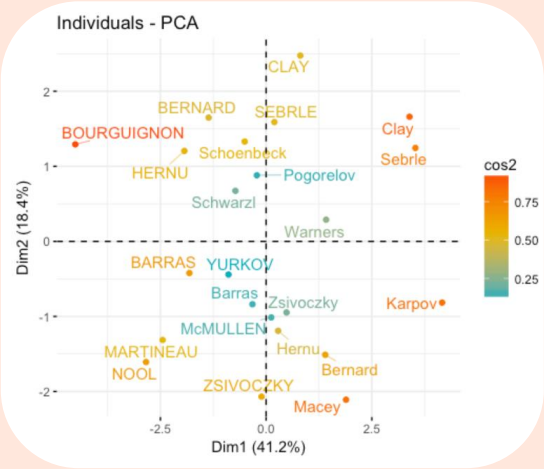
The World Happiness Report

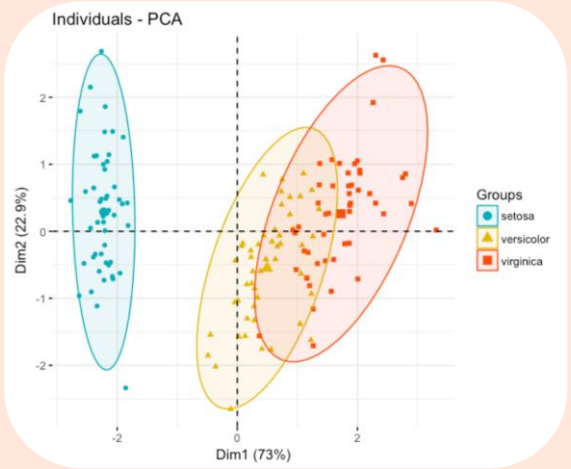
資料變數

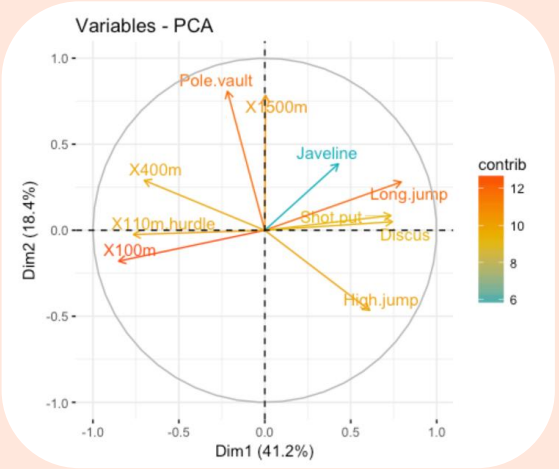
1. Social support
   社會支持 / 救助
2. Generosity
   慷慨程度
3. Log of GDP per capita
   人均 GDP
4. Freedom to make life choices
   人生自由選擇
5. Healthy life expectancy
   期望壽命
6. Perceptions of corruption
   貪汙程度

R語言實作

- Package　FactoMineR (計算 PCA)
  factoextra (Visualization)







更多參考 | factoextra : Extract and Visualize the Results of Multivariate Data Analyses

R語言實作

- 主成分計算（scale.unit 默認自動標準化)

```r
library(FactoMineR)
pca <- PCA(df_x, ncp = 6, graph = FALSE, scale.unit = TRUE)
```

- 綜合報告

```r
summary(pca)
```

```
Call:
PCA(X = df_x, ncp = 6, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               3.114   1.287   0.703   0.518   0.250   0.127
% of var.             51.906  21.445  11.716   8.640   4.173   2.119
Cumulative % of var.  51.906  73.351  85.068  93.708  97.881 100.000

Individuals (the 10 first)
                        Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
Finland                | 3.918 | 3.160  2.152  0.651 | 1.025  0.548  0.068 | -1.920  3.519  0.240 |
Denmark                | 3.983 | 3.234  2.254  0.659 | 1.631  1.388  0.168 | -1.551  2.297  0.152 |
Switzerland            | 3.549 | 3.124  2.103  0.775 | 1.171  0.715  0.109 | -1.155  1.273  0.106 |
Iceland                | 2.948 | 2.539  1.389  0.742 | 0.972  0.493  0.109 |  1.132  1.222  0.147 |
Netherlands            | 3.412 | 2.732  1.608  0.641 | 1.850  1.786  0.294 | -0.492  0.231  0.021 |
Norway                 | 3.778 | 3.208  2.218  0.721 | 1.723  1.548  0.208 | -0.925  0.817  0.060 |
Sweden                 | 3.708 | 2.996  1.934  0.653 | 1.790  1.671  0.233 | -1.184  1.337  0.102 |
Luxembourg             | 3.212 | 2.914  1.830  0.823 | 0.586  0.179  0.033 | -1.032  1.018  0.103 |
New Zealand            | 3.724 | 2.929  1.849  0.619 | 1.969  2.021  0.279 | -1.004  0.962  0.073 |
Austria                | 2.708 | 2.565  1.417  0.897 | 0.780  0.317  0.083 | -0.313  0.094  0.013 |

Variables
                          Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
Social support           |  0.858 23.645  0.736 | -0.168  2.189  0.028 |  0.293 12.235  0.086 |
Generosity               | -0.110  0.387  0.012 |  0.851 56.321  0.725 |  0.394 22.135  0.156 |
Healthy life expectancy  |  0.907 26.441  0.823 | -0.146  1.658  0.021 | -0.004  0.002  0.000 |
Logged GDP per capita    |  0.916 26.969  0.840 | -0.204  3.242  0.042 |  0.020  0.055  0.000 |
Freedom to make life choices |  0.669 14.351  0.447 |  0.428 14.220  0.183 |  0.164  3.845  0.027 |
Perceptions of corruption |  -0.506  8.206  0.256 | -0.536 22.370  0.288 |  0.659 61.729  0.434 |
```

R語言實作

- Eigenvector ＝ 主成分的係數

```
# eigen vectors (the coefficients of each principal components)
pca$svd$V
```

		y_1	y_2	y_3	y_4	y_5	y_6
		[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
x_1	[1,]	0.48626433	-0.1479644	0.3497787	0.09204782	-0.7392970	-0.25352628
x_2	[2,]	-0.06219858	0.7504715	0.4704751	0.45233669	0.0758332	0.03489756
x_3	[3,]	0.51421159	-0.1287739	-0.0043476	0.23444807	0.5854848	-0.56676845
x_4	[4,]	0.51931565	-0.1800458	0.0234936	0.25979073	0.1540898	0.77852933
x_5	[5,]	0.37883063	0.3770996	0.1960861	-0.80670747	0.1421436	0.06965440
x_6	[6,]	-0.28646708	-0.4729648	0.7856764	-0.11698141	0.2468772	0.04817104

變數 x

主成分 y

$$y_1 = \ 0.49x_1 - 0.06x_2 + 0.51x_3 + 0.52x_4 + 0.38x_5 - 0.29x_6$$

$$y_2 = -0.15x_1 + 0.75x_2 - 0.13x_3 - 0.18x_4 + 0.37x_5 - 0.47x_6$$

$$\vdots$$

$$y_6 = -0.25x_1 + 0.03x_2 - 0.57x_3 - 0.78x_4 + 0.07x_5 - 0.05x_6$$

R語言實作

- Eigenvalue = 主成分的變異數

```
library(factoextra)
## eigenvalue (variances of each principal components)
eig.val <- get_eigenvalue(pca)
eig.val
```

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.1143759	51.906266	51.90627
Dim.2	1.2867076	21.445126	73.35139
Dim.3	0.7029828	11.716380	85.06777
Dim.4	0.5184066	8.640110	93.70788
Dim.5	0.2503944	4.173240	97.88112
Dim.6	0.1271327	2.118878	100.00000

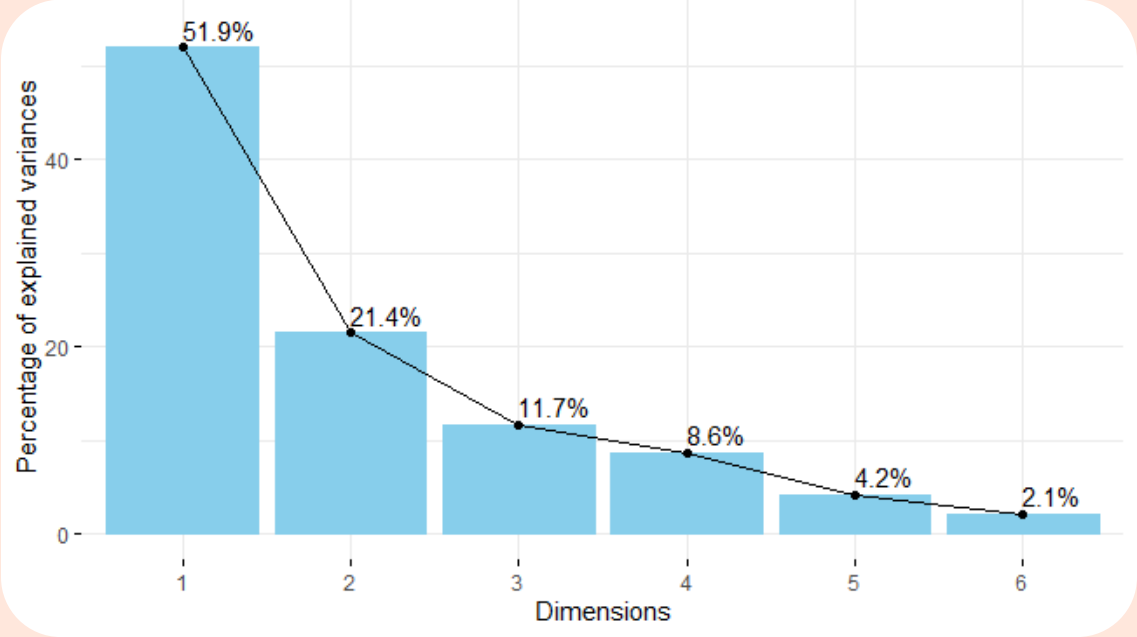
Q: 選擇幾個主成分？

A: 通常考慮累積變異在約 70% 以上 或 eigenvalue > 1 的前幾個主成分

R語言實作

fviz_screeplot

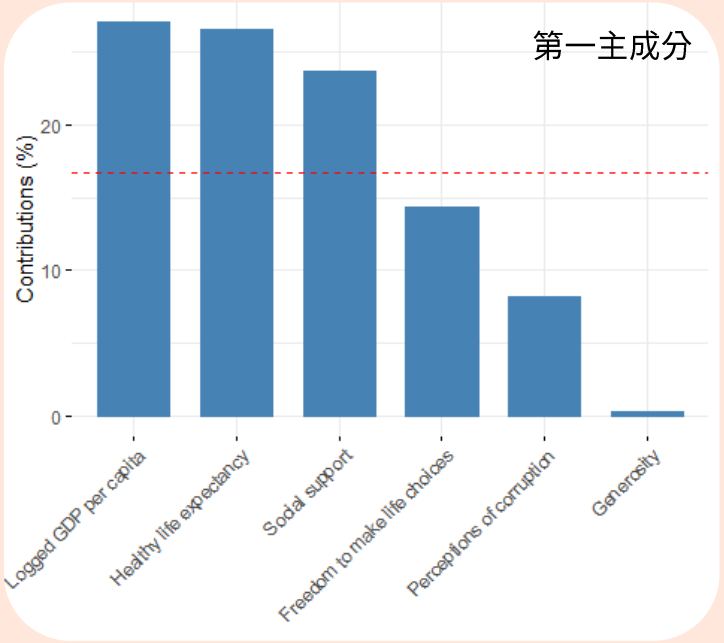
- 將eigenvalue視覺化，可以看見每個主成分解釋了數據多少的變異(百分比)

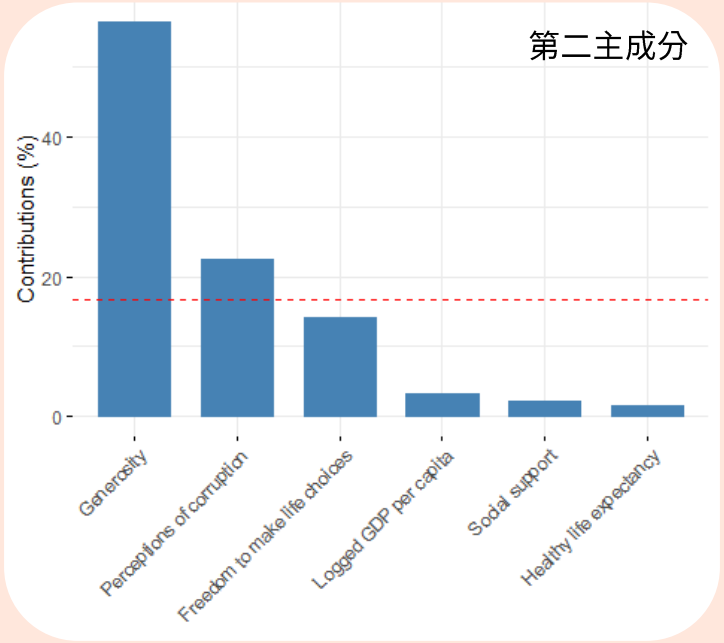


R語言實作

fviz_contrib

- 每個變數在前兩個主成分的貢獻比例 (超過平均貢獻視為重要貢獻)

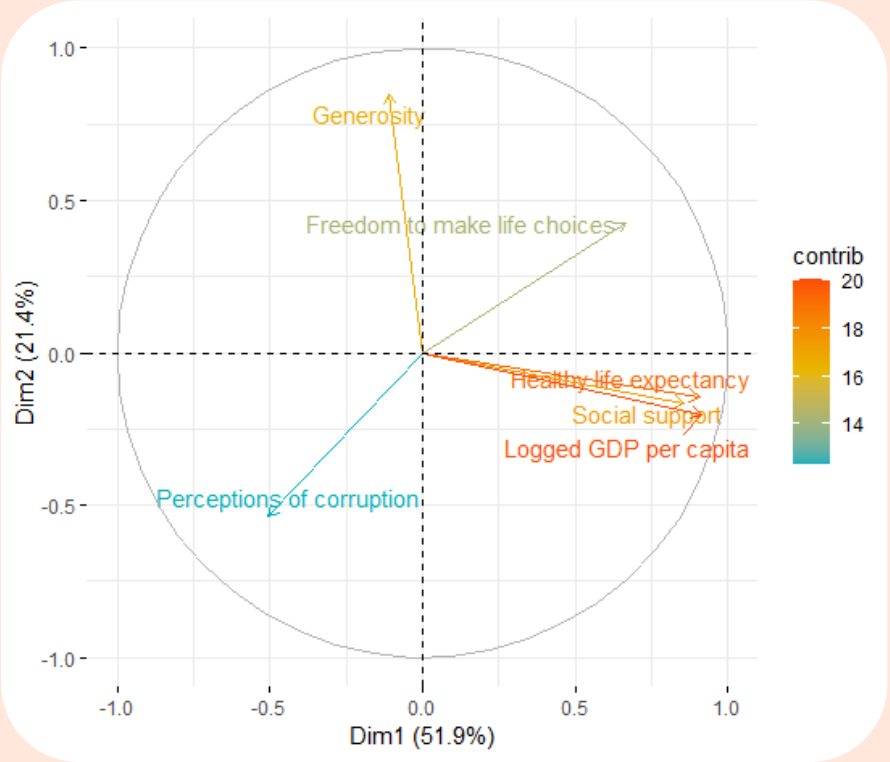




R語言實作

fviz_pca_var

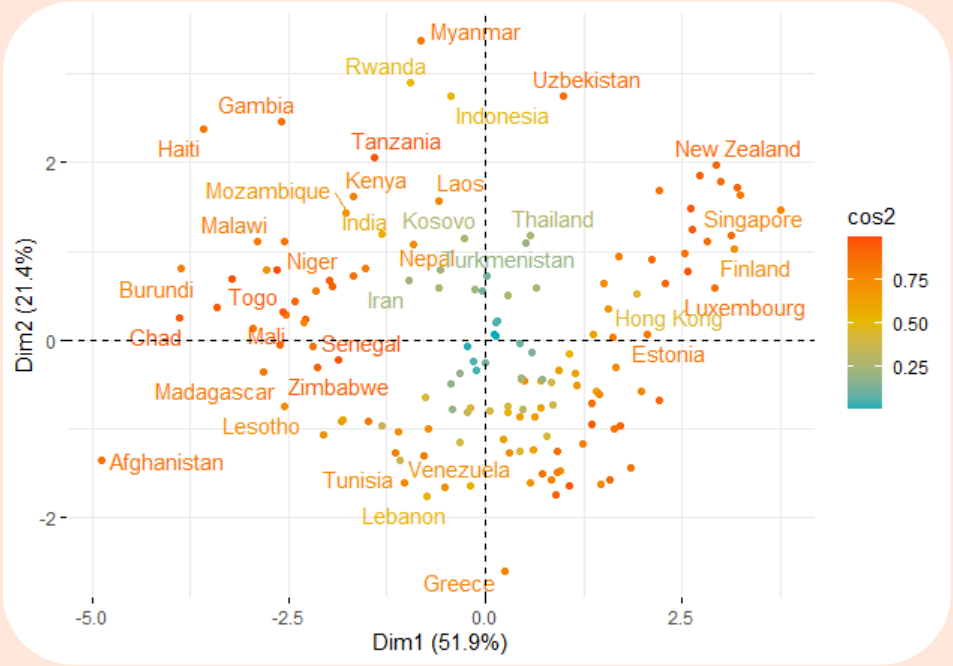
- 每個變數與前兩個主成分的相關圓盤圖
- 箭頭越接近圓框，表示此變數貢獻越大
- 第一主成分(x軸)
- 第二主成分(y軸)



R語言實作

fviz_pca_ind

- 每筆資料與主成分的關係圖
- 點越接近表示他們的變量特徵越相近



結論

主成分分析的「用途」

1 **資料探索**

將主成分分析繪圖後，能用來查看變數或資料筆之間的關係

2 **資料降維**

根據變異貢獻比例選定前幾個主成分後，可以刪去在各主成分都是低貢獻的變數

3 **變數選擇**

~~因此~~進行迴歸分析前，使用 PCA 來探索解釋變量間的關係，有助於選取合適的解釋變量

報告結束　謝謝大家

► THANKS ◄

Reference

內容參考：

1. 主成分分析的概念及應用 | by 行銷資料科學 | Marketingdatascience | Medium
2. 機器/統計學習:主成分分析 - Tommy Huang | Medium
3. 第 81 章 主成分分析 Principal Component Analysis | 醫學統計學 (wangcc.me)
4. 主成分分析 - 維基百科，自由的百科全書 (wikipedia.org)
5. PCA - Principal Component Analysis Essentials - Articles - STHDA

資料集來源：

1. World Happiness Report 2021 | The World Happiness Report
2. World Happiness Report 2021 | Kaggle