

Ejercicio 1 Ahorros por países

Alondra Sánchez Molina

Introducción

Según la hipótesis de ahorro del ciclo de vida desarrollada por Franco Modigliani, la tasa de ahorro (ahorro personal agregado dividido por la renta disponible) se explica por la renta disponible per cápita, la tasa de variación porcentual de la renta disponible per cápita y dos variables demográficas: el porcentaje de población menor de 15 años y el porcentaje de población mayor de 75 años.

Los datos se promedian durante la década de 1960 a 1970 para eliminar el ciclo económico u otras fluctuaciones a corto plazo. Los siguientes datos se obtuvieron de Belsley, D. A., Kuh, E. y Welsch, R. E. (1980). Regression Diagnostics. John Wiley & Sons, Nueva York. A su vez, obtuvieron los datos de Sterling, A. (1977) Tesis de licenciatura no publicada del Instituto de Tecnología de Massachusetts.

El conjunto de datos contiene 50 observaciones con cinco variables.

- I. Sr: numérico, ahorros personales agregados
- II. pop15: numérico, porcentaje de población menor de 15 años
- III. pop75: numérico, porcentaje de población mayor de 75 años
- IV. dpi: numérico, ingreso disponible per cápita real
- V. ddpi: numérico, porcentaje de tasa de crecimiento de dpi

Utiliza un algoritmo de agrupamiento para identificar los países similares.

Preparación de los datos

Primeramente, se carga el archivo; y se visualiza con el fin de tener un primer acercamiento a estos.

Posteriormente, se utiliza la función `sum()` de R, para observar si existen datos sin valor en el dataset, en este caso, se obtiene que no.

```
> #Preprocesamiento de datos
> sum(is.na(dt_lifeCycles))
[1] 0
```

En el dataset existe una columna que cuenta con los nombres de los países, pero el análisis a implementar requiere eliminar dicha columna.

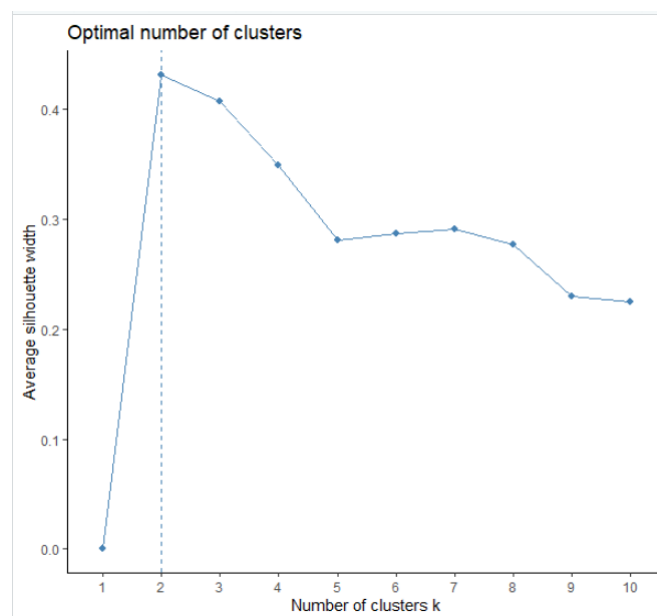
```
> rownames(dt_lifeCycles) <- dt_lifeCycles$Contry
> dt_lifeCycles <- dt_lifeCycles[, -c(colnames(dt_lifeCycles) %in% ("Contry"))]
> view(dt_lifeCycles)
```

En este caso, en lugar de eliminarla, se cambiaron a que el número de fila sea el nombre de la columna, esto, además nos beneficiara en la visualización gráfica de los resultados. Finalmente, para terminar este preprocesamiento se escalan los datos con el fin de evitar que el algoritmo dependa de una unidad de variable arbitraria.

```
data <- scale(dt_lifeCycleS)
```

Estimación de k

En la creación de clusters, es necesario elegir el número adecuado de ellos, esto es, nuestro valor k, por ello, se utiliza la función *fviz_nbclust()*, para estimar el valor k más óptimo. El método a utilizar fue silhouette; el cual dibuja la silueta de los grupos promedio de acuerdo con el número de grupos.



K-Medoids con PAM

Justificación

Se utilizó dicho algoritmo pues usa medoides para la creación de clusters, los puntos más centrales del grupo. K-means, es más sensible a valores atípicos, y PAM es más robusto. Se podría usar CLARA, pero este es más recomendado para datasets de mayor tamaño, y puesto que el dataset con el cual se trabaja en esta actividad es prácticamente de pequeño tamaño, es que se inicializó K-Medoids con PAM para generar clusters.

PAM

Se utiliza la función *pam()*, pasando nuestros datos y el número de clusters, en este caso 2. Dicha función al imprimirla nos retorna la información de los medoids de cada cluster. Así como el vector de cluster generado.

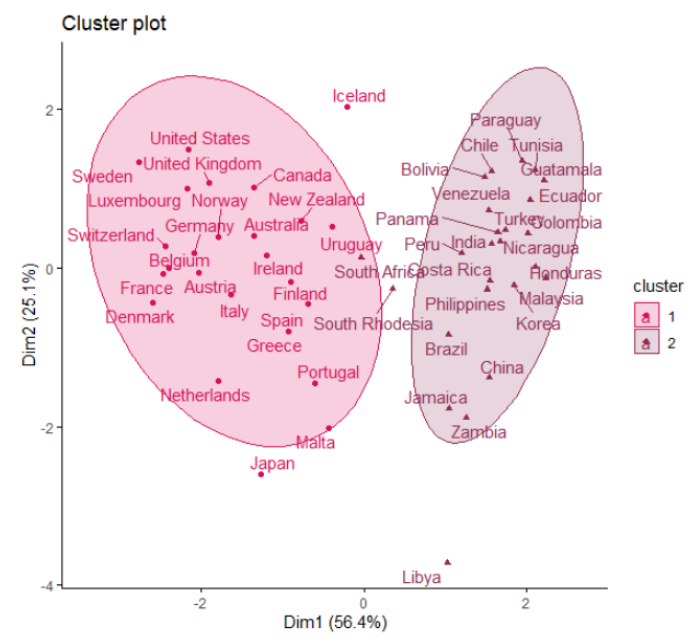
```

> pam_resul <- pam(data, 2)
> print(pam_resul)
Medoids:
      ID      sr      pop15      pop75      dpi      ddpi
Norway  27  0.1292293 -0.9986749  1.0668039  1.1346321 -0.04794641
Nicaragua 30 -0.5291930  1.0872702 -0.8390332 -0.7884175 -0.44517686
Clustering vector:
      Australia      Austria      Belgium      Bolivia      Brazil      Canada      Chile      China
      1             1             1             2             2             1             2             2
      Colombia      Costa Rica      Denmark      Ecuador      Finland      France      Germany      Greece
      2             2             1             2             1             1             1             1
      Guatemala      Honduras      Iceland      India      Ireland      Italy      Japan      Korea
      2             2             1             2             1             1             1             2
      Luxembourg      Malta      Norway      Netherlands      New Zealand      Nicaragua      Panama      Paraguay
      1             1             1             1             1             2             2             2
      Peru      Philippines      Portugal      South Africa      South Rhodesia      Spain      Sweden      Switzerland
      2             2             1             2             2             1             1             1
      Turkey      Tunisia      United Kingdom      United States      Venezuela      Zambia      Jamaica      Uruguay
      2             2             1             1             2             2             2             1
      Libya      Malaysia
      2             2
Objective function:
      build      swap
1.740398  1.441066
Available components:
[1] "medoids"      "id.med"      "clustering" "objective" "isolation" "clusinfo" "silinfo" "diss" "call"
[10] "data"

```

Visualización

Con el plotear de los datos, podremos analizar de una manera visual cuáles países pertenecen a cada grupo.



Interpretación

Parte del análisis, es definir las características que hacen que cada grupo pertenezca a ese. Es por ello que se utiliza el método `aggregate()`, para obtener los valores promedios de cada cluster.

```
> aggregate(dt_lifeCycleS, by=list(cluster=pam_resul$cluster), mean)
  cluster      sr  pop15  pop75      dpi  ddpi
1       1 11.5464 26.8644 3.4224 1868.269 3.9228
2       2  7.7956 43.3148 1.1636  345.248 3.5924
> |
```

Otra manera de visualizar información sería imprimiendo los datos de cada país pero solo del cluster al cual pertenecen:

```
> new_lifeCycleS[new_lifeCycleS$cluster==1,]
  sr pop15 pop75      dpi  ddpi cluster
Australia 11.43 29.35  2.87 2329.68 2.87      1
Austria   12.07 23.32  4.41 1507.99 3.93      1
Belgium    13.17 23.80  4.43 2108.47 3.82      1
Canada     8.79 31.72  2.85 2982.88 2.43      1
Denmark    16.85 24.42  3.93 2496.53 3.99      1
Finland    11.24 27.84  2.37 1681.25 4.32      1
France     12.64 25.06  4.70 2213.82 4.52      1
Germany    12.55 23.31  3.35 2457.12 3.44      1
Greece     10.67 25.62  3.10  870.85 6.28      1
Iceland     1.27 34.03  3.08 1900.10 1.12      1
Ireland    11.34 31.16  4.19 1139.95 2.99      1
Italy       14.28 24.52  3.48 1390.00 3.54      1
Japan       21.10 27.01  1.91 1257.28 8.21      1
Luxembourg 10.35 21.80  3.73 2449.39 1.57      1
Malta       15.48 32.54  2.47  601.05 8.12      1
Norway      10.25 25.95  3.67 2231.03 3.62      1
Netherlands 14.65 24.71  3.25 1740.70 7.66      1
New Zealand 10.67 32.61  3.17 1487.52 1.76      1
Portugal    12.49 28.96  2.85  579.51 7.48      1
Spain       11.77 27.74  2.87  768.79 4.35      1
Sweden       6.86 21.44  4.54 3299.49 3.01      1
Switzerland 14.13 23.49  3.73 2630.96 2.70      1
United Kingdom 7.81 23.27  4.46 1813.93 2.01      1
United States 7.56 29.81  3.43 4001.89 2.45      1
Uruguay      9.24 28.13  2.72  766.54 1.88      1
. |

> new_lifeCycleS[new_lifeCycleS$cluster==2,]
  sr pop15 pop75      dpi  ddpi cluster
Bolivia     5.75 41.89  1.67 189.13 0.22      2
Brazil      12.88 42.19  0.83 728.47 4.56      2
Chile        0.60 39.74  1.34 662.86 2.67      2
China       11.90 44.75  0.67 289.52 6.51      2
Colombia     4.98 46.64  1.06 276.65 3.08      2
Costa Rica  10.78 47.64  1.14 471.24 2.80      2
Ecuador      3.59 46.31  1.19 287.77 2.19      2
Guatemala    3.01 46.05  0.87 289.71 1.48      2
Honduras     7.70 47.32  0.58 232.44 3.19      2
India         9.00 41.31  0.96  88.94 1.54      2
Korea        3.98 41.74  0.91 207.68 5.81      2
Nicaragua    7.30 45.04  1.21 325.54 2.48      2
Panama        4.44 43.56  1.20 568.56 3.61      2
Paraguay     2.02 41.18  1.05 220.56 1.03      2
Peru         12.70 44.19  1.28 400.06 0.67      2
Philippines  12.78 46.26  1.12 152.01 2.00      2
South Africa 11.14 31.94  2.28 651.11 2.19      2
South Rhodesia 13.30 31.92  1.52 250.96 2.00      2
Turkey       5.13 43.42  1.08 389.66 2.96      2
Tunisia       2.81 46.12  1.21 249.87 1.13      2
Venezuela    9.22 46.40  0.90 813.39 0.53      2
Zambia       18.56 45.25  0.56 138.33 5.14      2
Jamaica       7.72 41.12  1.73 380.47 10.23      2
Libya        8.89 43.69  2.07 123.58 16.71      2
Malaysia     4.71 47.20  0.66 242.69 5.08      2
```

Conclusiones

Se observa que los países miembros del cluster 1, cuentan con un mayor valor de ahorros personales agregado, el porcentaje de menores de 15 años es menor, pero cuentan con un mayor porcentaje de personas mayores de 75 años. Además, los países agrupados en el cluster 2, tienen un menor ingreso disponible per cápita real a comparación de los países pertenecientes al cluster 1.