Clustering and PCA

Mathias Bourel

20/5/2019

En la situación en la que se tiene un gran conjunto de datos multivariados que contiene varias variables continuas, el Análisis de componentes principales (PCA) se puede usar para reducir el tamaño de los datos a algunas variables continuas que contienen la información más importante de los datos. A continuación, puede realizar el método de clustering sobre los resultados del ACP.

El paso de ACP se puede considerar como un paso que reduce el ruido de fondo en los datos, lo que puede llevar a una clasificación más estable.

Resumen:

1- Hacer un Análisis de Componentes Principales (ACP)

2- Aplicar un método de clustering sobre el resultado del ACP (HCPC: Hierarchical Clustering on Principal Components)

```
#HCPC (res, nb.clust = 3, graph = TRUE)
```

Argumentos:

res: resultado de un ACP o data frame nb.clust: cantidad de clusters graph: si TRUE, aparecen los gráficos

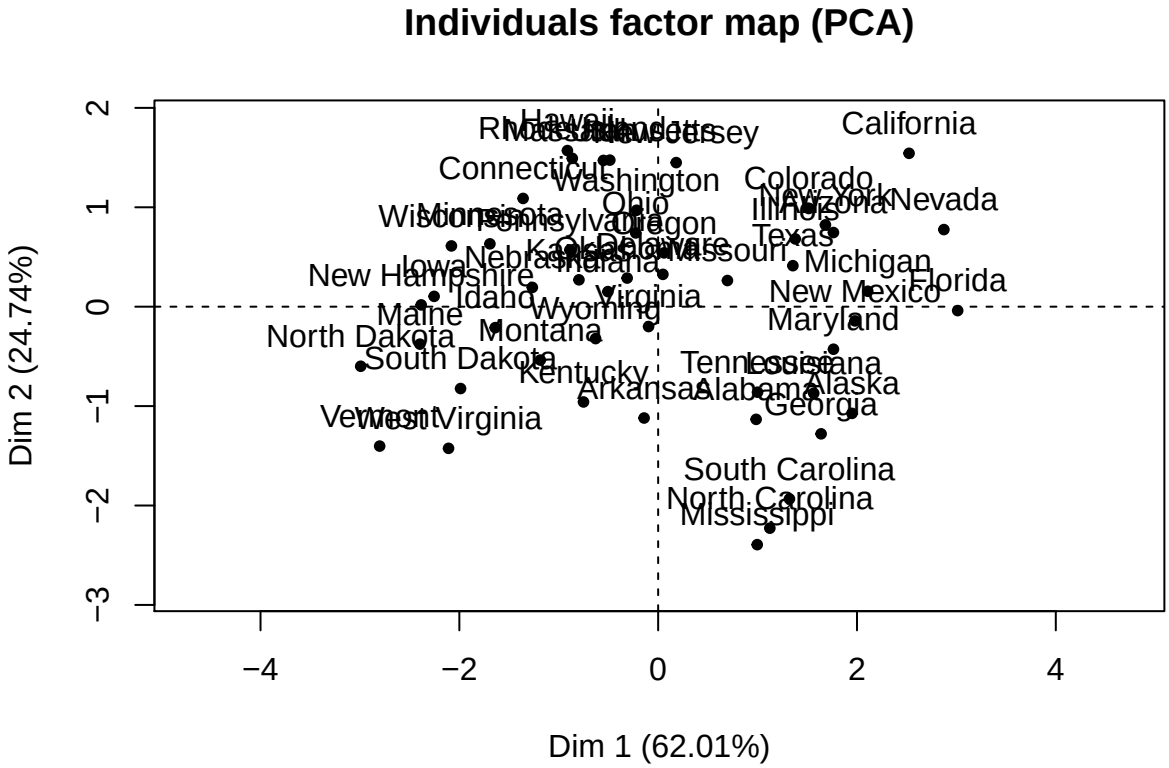
Objetos devueltos por HCPC: data.clust: Datos de origen con una última columna con el número de cluster. desc.var: las variables que describen los grupos desc.ind: los individuos que describen mejor al grupo desc.axes: los ejes que describen al grupo

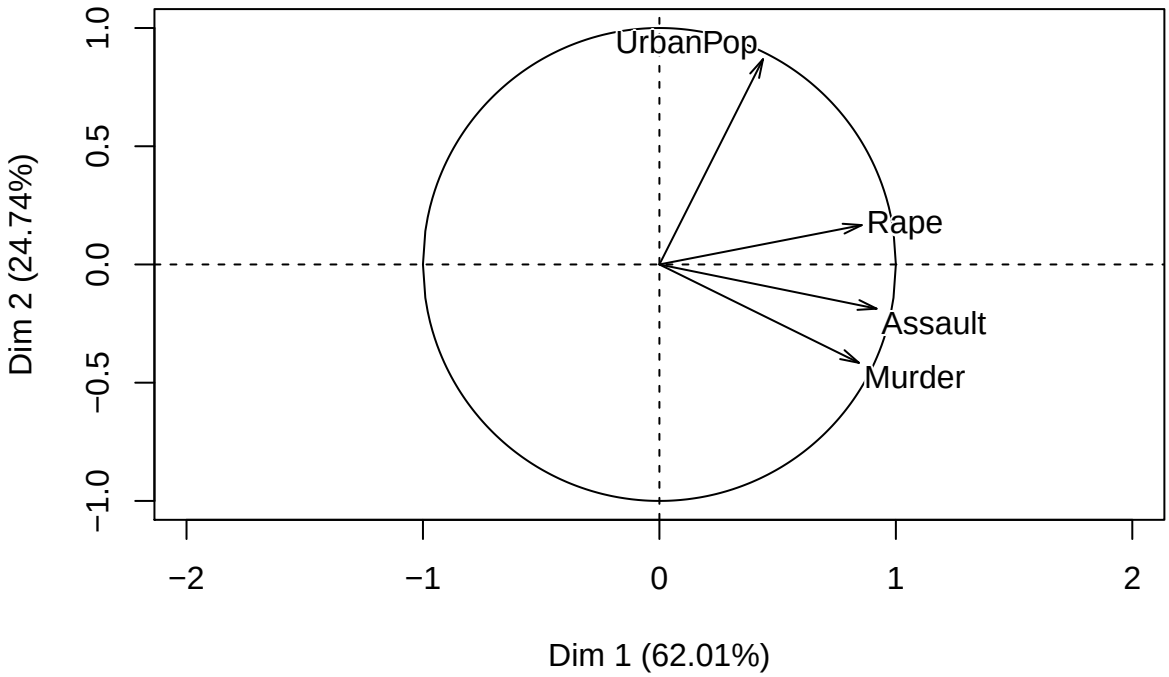
```
library(FactoMineR)
library(factoextra)
```

```
## Loading required package: ggplot2
```

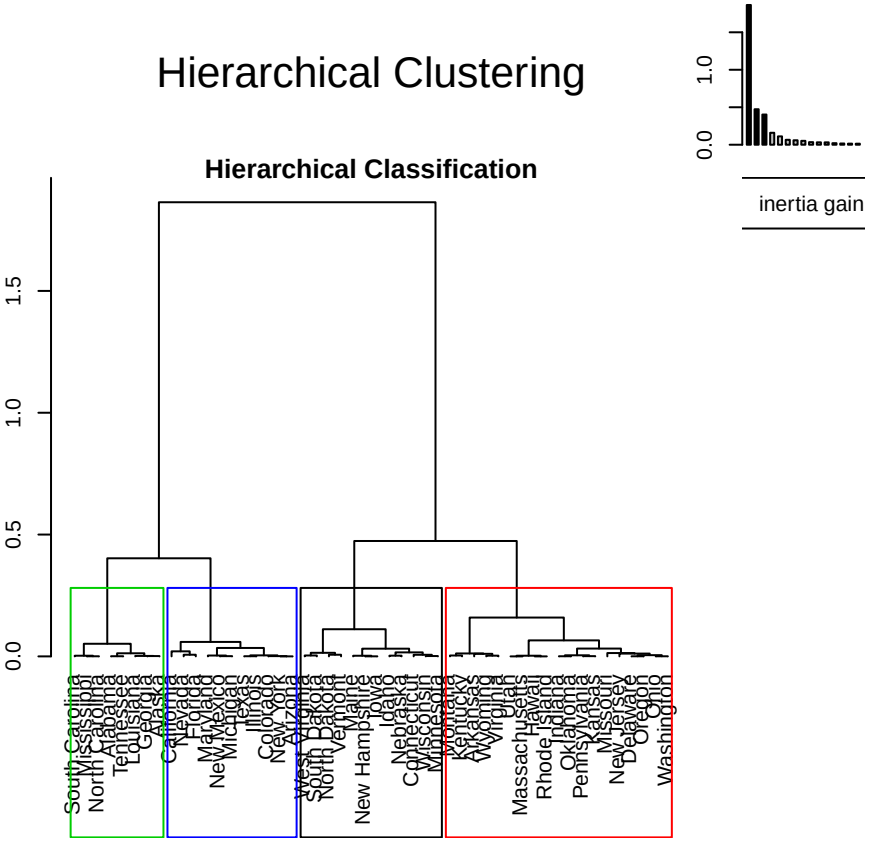
```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

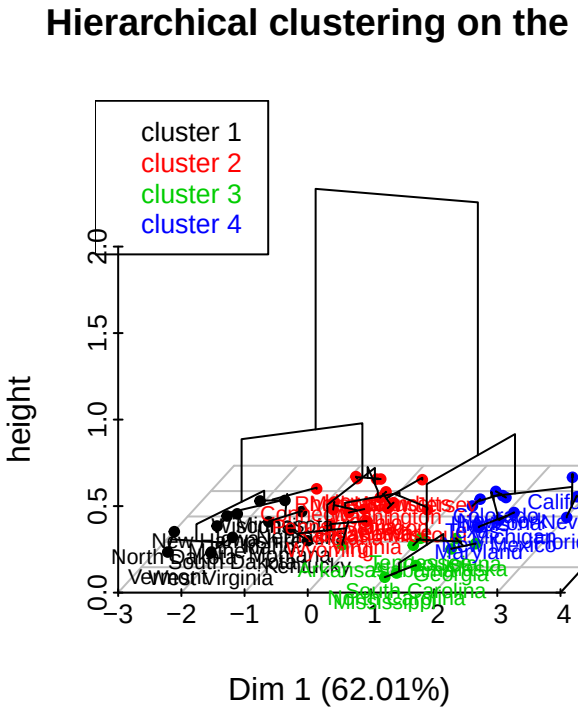
```
# 1. ACP
res.pca <- PCA(USArrests, ncp = 2, graph = TRUE)
```

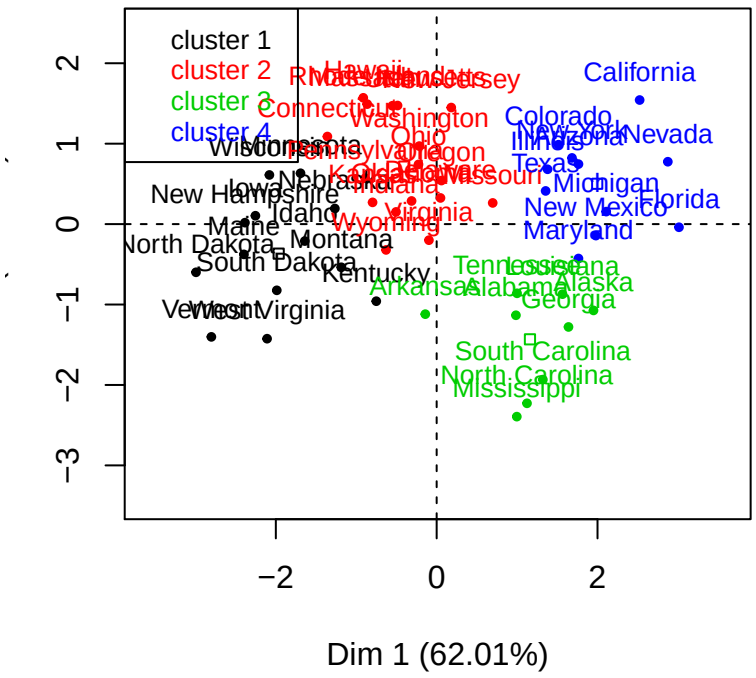





```
# 2. HCPC
res.hcpc <- HCPC(res.pca,nb.clust=4)
```

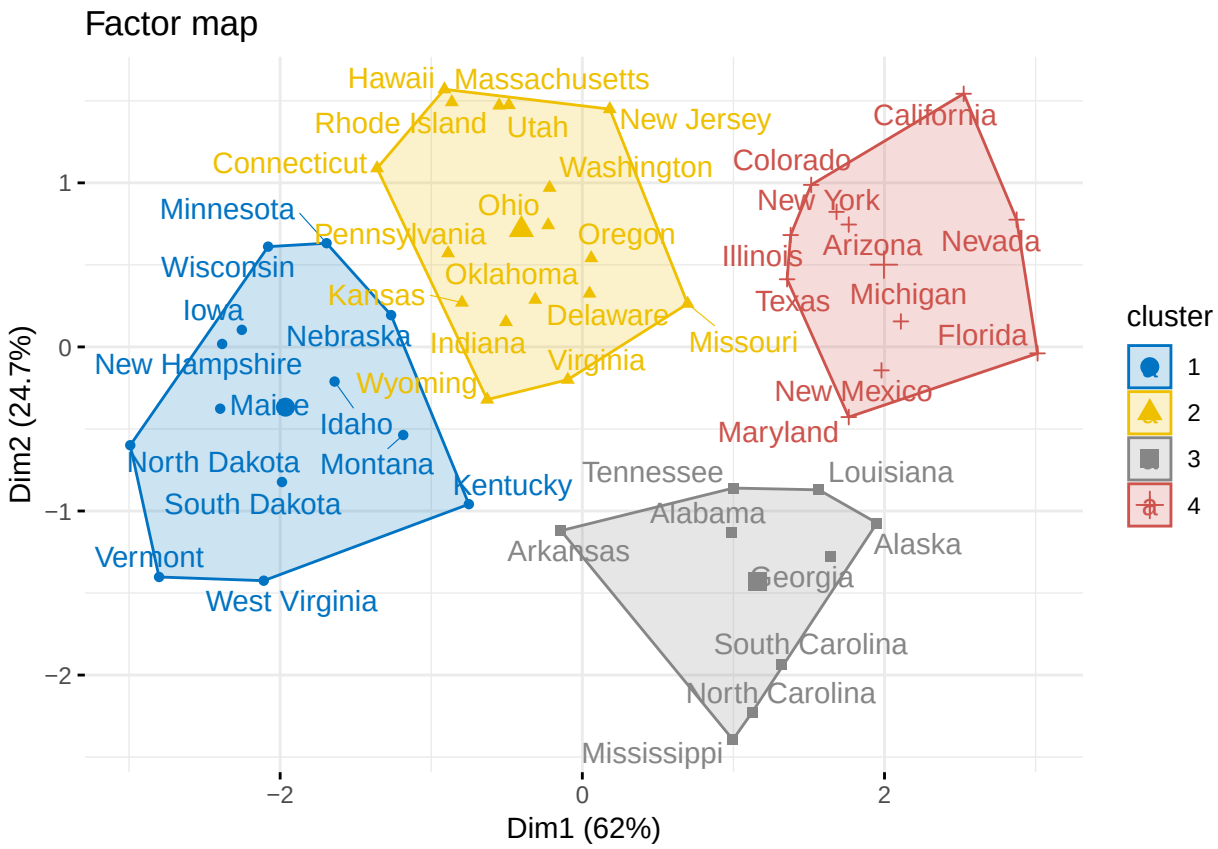




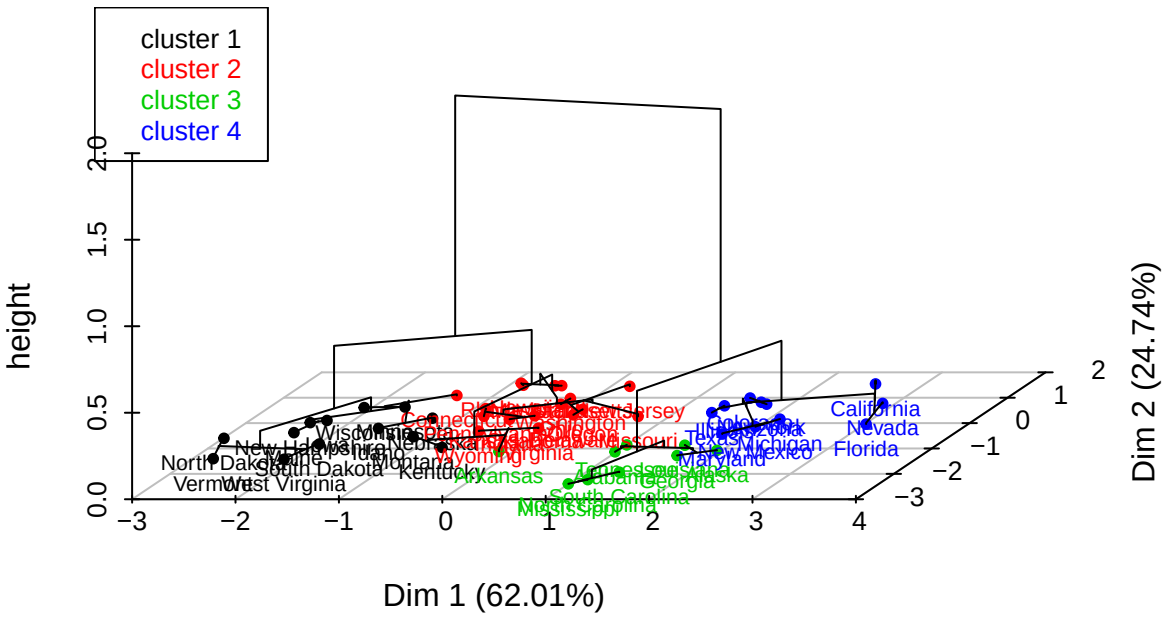



```
fviz_cluster(res.hcpc,
             repel = TRUE,            # Para no solapar los testos
             show.clust.cent = TRUE, # Muestra el centro de cada cluster
```

```
          palette = "jco",          # Colores, ver ?ggpubr::ggpar
          ggtheme = theme_minimal(),
          main = "Factor map"
          )
```




```
# Principal components + tree
plot(res.hcpc, choice = "3D.map")
```



Para ver los datos de origen junto con el cluster asignado:

```
head(res.hcpc$data.clust, 10)
```

```
##             Murder Assault UrbanPop Rape clust
## Alabama       13.2     236       58 21.2     3
## Alaska        10.0     263       48 44.5     3
## Arizona        8.1     294       80 31.0     4
## Arkansas       8.8     190       50 19.5     3
## California     9.0     276       91 40.6     4
## Colorado       7.9     204       78 38.7     4
## Connecticut    3.3     110       77 11.1     2
## Delaware       5.9     238       72 15.8     2
## Florida       15.4     335       80 31.9     4
## Georgia       17.4     211       60 25.8     3
```

Variables cuantitativas que contribuyen más al cluster:

```
res.hcpc$desc.var$quanti
```

```
## $`1`
##             v.test Mean in category Overall mean sd in category Overall sd
## UrbanPop -3.898420         52.07692       65.540       9.691087  14.329285
## Murder   -4.030171          3.60000        7.788       2.269870   4.311735
## Rape     -4.052061         12.17692       21.232       3.130779   9.272248
## Assault  -4.638172         78.53846      170.760      24.700095  82.500075
##                p.value
## UrbanPop 9.682222e-05
## Murder   5.573624e-05
## Rape     5.076842e-05
## Assault  3.515038e-06
##
## $`2`
```

```
##             v.test Mean in category Overall mean sd in category Overall sd
## UrbanPop  2.842525        73.647059       65.540       8.443175  14.329285
## Murder   -2.254798         5.852941        7.788       1.735822   4.311735
##              p.value
## UrbanPop 0.004475767
## Murder   0.024146000
##
## $`3`
##             v.test Mean in category Overall mean sd in category Overall sd
## Murder    4.344737         13.50000        7.788       2.606829   4.311735
## Assault   2.982203        245.77778      170.760      44.298928  82.500075
## UrbanPop -2.844691         53.11111       65.540       7.324911  14.329285
##               p.value
## Murder   1.394424e-05
## Assault  2.861821e-03
## UrbanPop 4.445447e-03
##
## $`4`
##             v.test Mean in category Overall mean sd in category Overall sd
## Rape     4.565151         32.61818       21.232       6.605620   9.272248
## Assault  4.205651        264.09091      170.760      38.068012  82.500075
## UrbanPop 3.515653         79.09091       65.540       6.515081  14.329285
## Murder   2.816429         11.05455        7.788       2.078183   4.311735
##               p.value
## Rape     4.991343e-06
## Assault  2.603317e-05
## UrbanPop 4.386740e-04
## Murder   4.856083e-03
```

Del resultado anterior deducimos que las variables UrbanPop, Murder, Rape y Assault son las más significativas para el cluster 1. Por ejemplo,el valor medio de la variable Assault en el cluster 1 es de 78.53, lo cual es inferior a la media global (170.76) en todos los clusters. Concluimos que el cluster 1 se caracteriza por un bajo nivel de la variable Assault con respecto a los demás grupos.Las variables UrbanPop y Murder son las más significativas asociadas al cluster 2.

Ejes principales asociados a los clusters:

```
res.hcpc$desc.axes$quanti
```

```
## $`1`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1 -5.175764        -1.964502 -5.639933e-16      0.6192556   1.574878
##            p.value
## Dim.1 2.269806e-07
##
## $`2`
##         v.test Mean in category  Overall mean sd in category Overall sd
## Dim.2 3.609122        0.7146644 -5.369316e-16      0.6060621  0.9948694
##            p.value
## Dim.2 0.0003072349
##
## $`3`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1  2.415379         1.159858 -5.639933e-16      0.5599226  1.5748783
## Dim.2 -4.722089        -1.432428 -5.369316e-16      0.5568189  0.9948694
##            p.value
```

```
## Dim.1 1.571886e-02
## Dim.2 2.334343e-06
##
## $`4`
##         v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1 4.712022         1.996146 -5.639933e-16      0.5499968   1.574878
##            p.value
## Dim.1 2.452704e-06
```

Individuos representativos de cada cluster:

```
res.hcpc$desc.ind$para
```

```
## Cluster: 1
##         Idaho         Maine  South Dakota          Iowa New Hampshire
##     0.3602999     0.4324263     0.4568091     0.5528255     0.5694679
## ------------------------------------------------------------
## Cluster: 2
##         Ohio   Washington     Oklahoma       Oregon Pennsylvania
##    0.1795640    0.3162919    0.4365324    0.4939058    0.5058174
## ------------------------------------------------------------
## Cluster: 3
##        Alabama        Georgia South Carolina      Tennessee      Louisiana
##      0.3461218      0.5033947      0.5261702      0.5941581      0.6921943
## ------------------------------------------------------------
## Cluster: 4
##    Arizona   Michigan   New York   Illinois New Mexico
##  0.3377824  0.3637539  0.4493116  0.6430740  0.6444338
```

Descripción de los ejes principales

```
res.hcpc$desc.axes
```

```
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##            Eta2      P-value
## Dim.1 0.8779383 5.065328e-21
## Dim.2 0.6398838 2.796000e-10
##
## Description of each cluster by quantitative variables
## ======================================================
## $`1`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1 -5.175764        -1.964502 -5.639933e-16      0.6192556   1.574878
##            p.value
## Dim.1 2.269806e-07
##
## $`2`
##         v.test Mean in category  Overall mean sd in category Overall sd
## Dim.2 3.609122        0.7146644 -5.369316e-16      0.6060621  0.9948694
##            p.value
## Dim.2 0.0003072349
##
## $`3`
##          v.test Mean in category  Overall mean sd in category Overall sd
```

```
## Dim.1  2.415379         1.159858 -5.639933e-16      0.5599226  1.5748783
## Dim.2 -4.722089        -1.432428 -5.369316e-16      0.5568189  0.9948694
##            p.value
## Dim.1 1.571886e-02
## Dim.2 2.334343e-06
##
## $`4`
##         v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1 4.712022         1.996146 -5.639933e-16      0.5499968   1.574878
##            p.value
## Dim.1 2.452704e-06
```

Video showing how to perform clustering with FactoMineR

https://www.youtube.com/watch?v=4XrgWmN9erg&list=PLnZgp6epRBbTsZEFXi_p6W48HhNyqwxIu&index=7