Análisis Cluster y Discriminante

## Trabajo 2

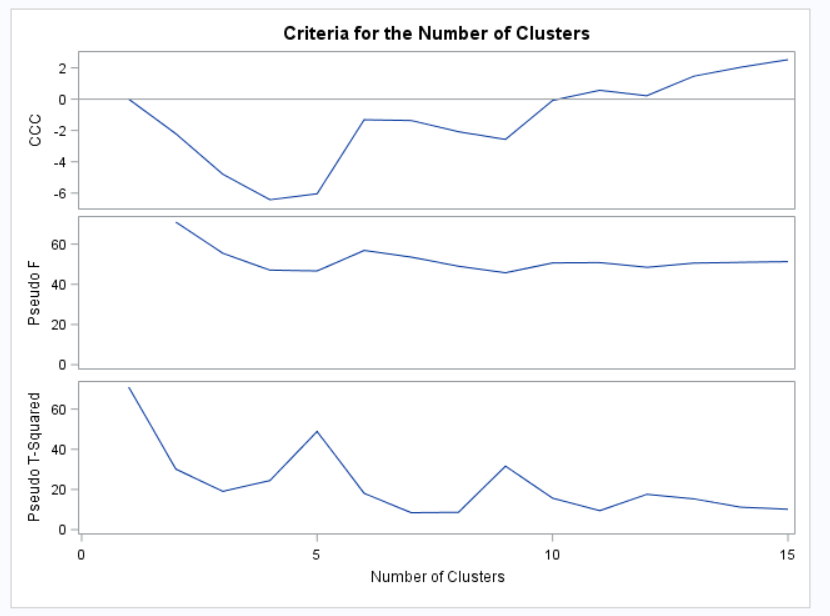
# **Análisis Cluster**

Para hacer el análisis cluster hay algunas condiciones fundamentales que se tienen que tener en cuenta. Principalmente que las variables con cuales hacemos el estudio sean independientes, que estén medidas en las mismas unidades (normales/ estandarizadas) y que en caso de que haya atípicos que se eliminen del estudio.

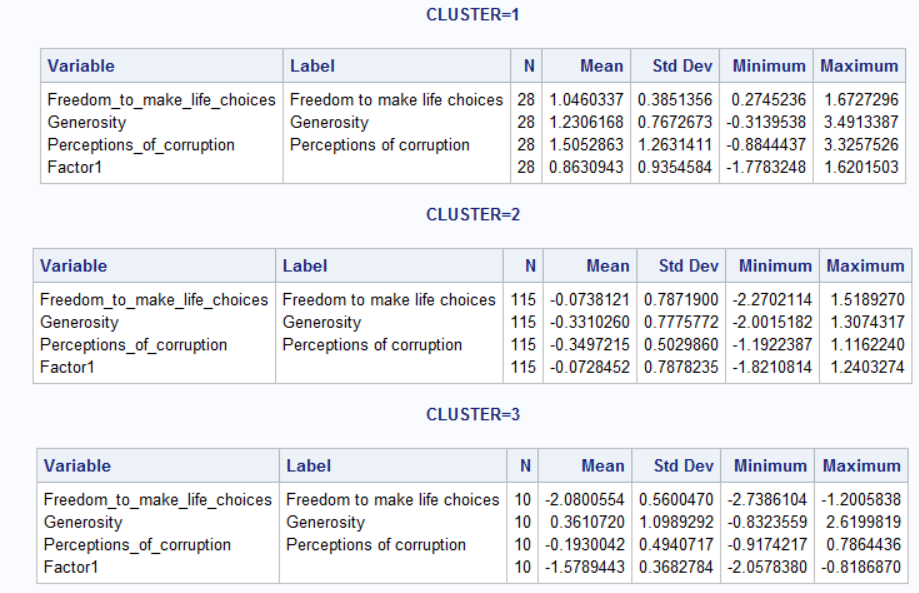
Tras hacer el análisis factorial hemos conseguido explicar parte de las variables correladas con un factor, que hemos llamado “calidad de vida”. El resto de las variables, por ser independientes no han sido agrupados en ningún factor. Por otra parte, sabemos que hay 3 observaciones atípicas que no hemos eliminado para el análisis factorial, pero hará falta en este análisis cluster. Entonces usaremos los datos de **multi.happyscore**, donde no forman parte los atípicos. Finalmente, los datos se estandarizan con cual nos quedamos con los datos de **happystd**.

Empezaremos por el Análisis Cluster Jerárquico, para ver en cuantos conglomerados podemos agrupar nuestros datos según las variables que tenemos. También basta recordar que obtenemos los mismos resultados normalizando los datos y posteriormente haciendo análisis factorial (con covarianza), que hacer análisis factorial con los datos originales (y varianzas distintas) con matriz de correlaciones, así que el Factor1 de happystd es equivalente al Factor1 de los datos originales.

El primer método que vamos a aplicar es del vecino lejano.

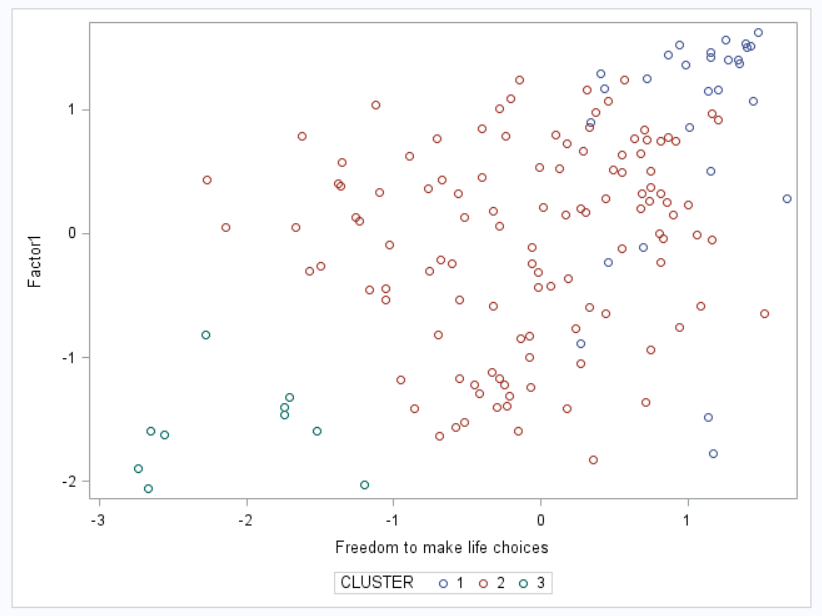


Vemos que para 3 clusters tenemos un mínimo local para el estadístico t (t= 19,1) y F=55,4. La siguiente posible agrupación será con 7 clusters que también vamos a considerar como oportunidad.

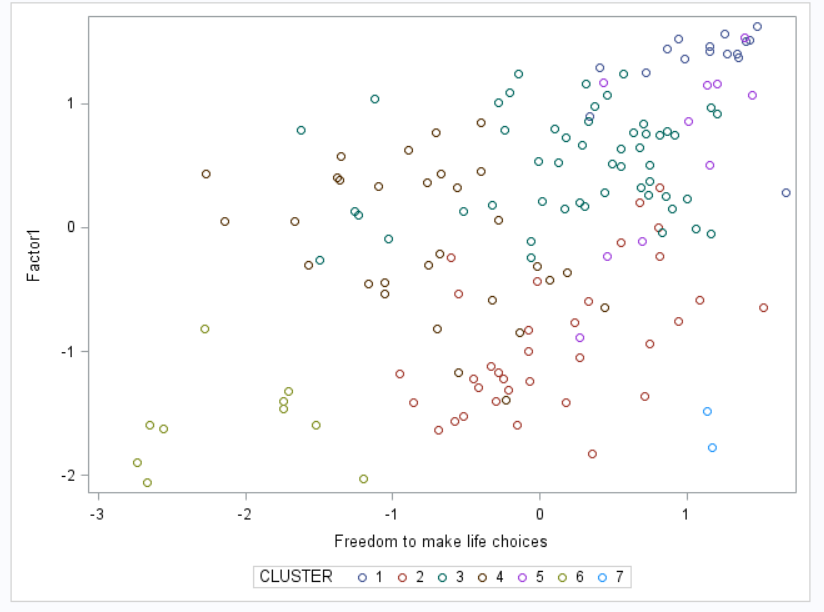


El siguiente paso es intentar encontrar las dos variables que parecen que discriminan más para poder hacer el grafico de dispersión, que nos va a orientar cual es el mejor método de formar clusters y lo más importante – con cuantos clusters nos quedamos.

De la tabla se observa que la variable Freedom\_to\_make\_life\_choices tiene un rango bastante grande entre los tres clusters (mínimo en cluster 3 con valor -2,08 y máximo en cluster 1 con valor 1,046) y lo mismo se observa para la variable Factor1. Hacemos el grafico con estas dos:

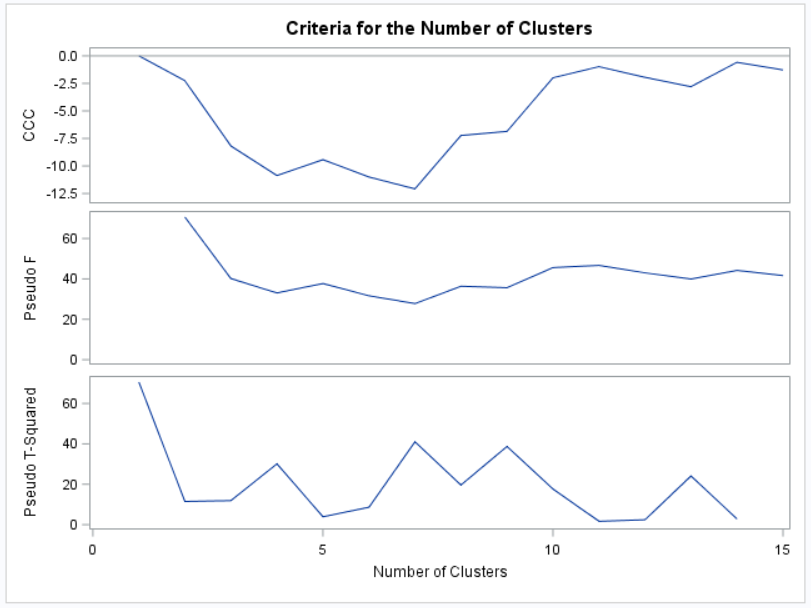


Parece que este método forma 3 clusters muy coherentes. También, como he mencionado, no rechazaremos la opción de mas clusters ya que hay muchas observaciones en grupo 2 resultando en que sea bastante disperso.

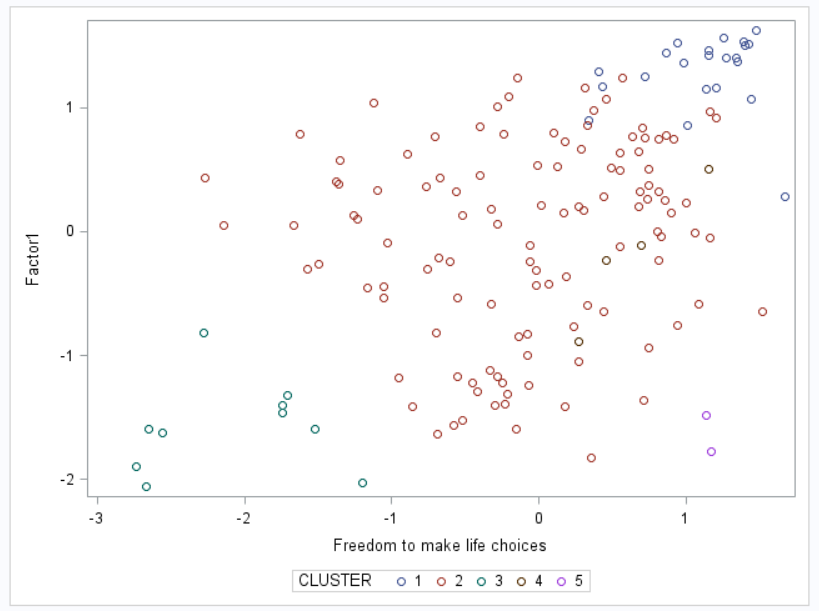


Vemos que clusters 3, 4 y 5 no son muy distinguibles en el grafico para este método para estas dos variables.

Segundo método que vamos a intentar es el método average.

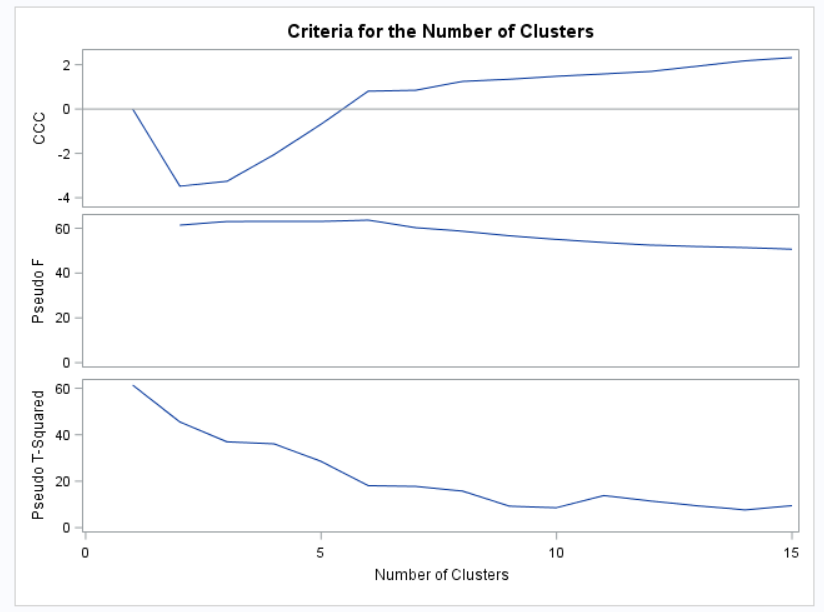


Tenemos posibles agrupaciones en 2 y en 5 clusters. Habiendo visto el grafico de dispersión no seguiría con la opción de 2 clusters porque un grupo se queda demasiado grande y disperso, pero intentaremos con 5.

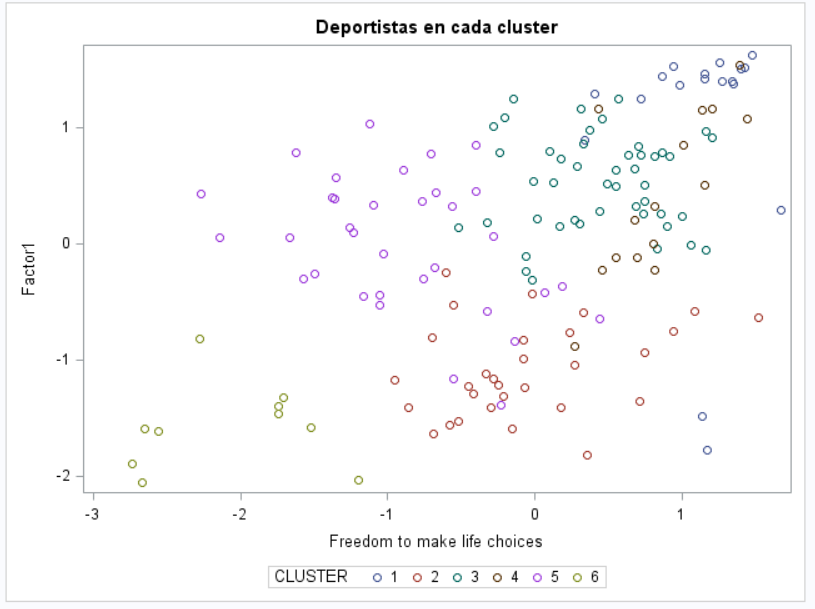


Las observaciones de cluster 4 no son distinguibles, por otra parte, tenemos un nuevo cluster (el 5) para las dos observaciones que se desvían bastante del resto.

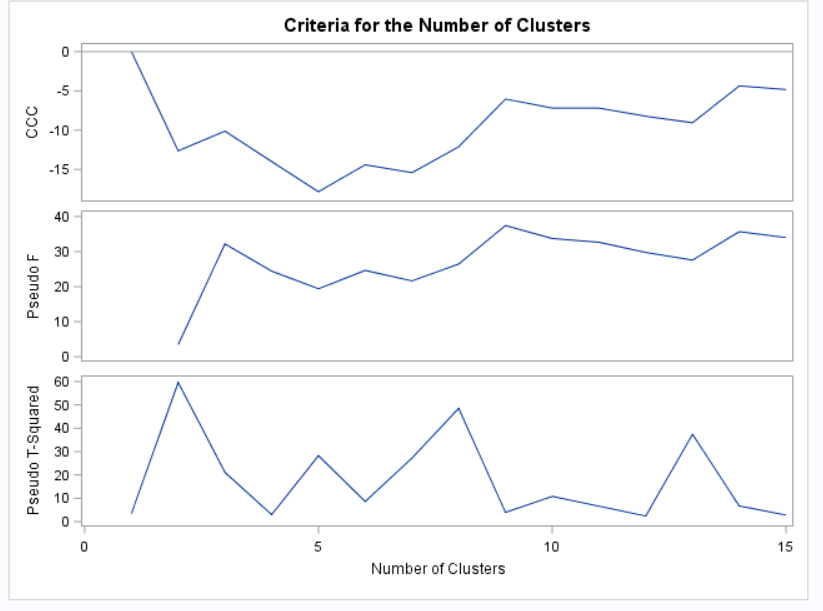
Seguimos con el método de Ward.

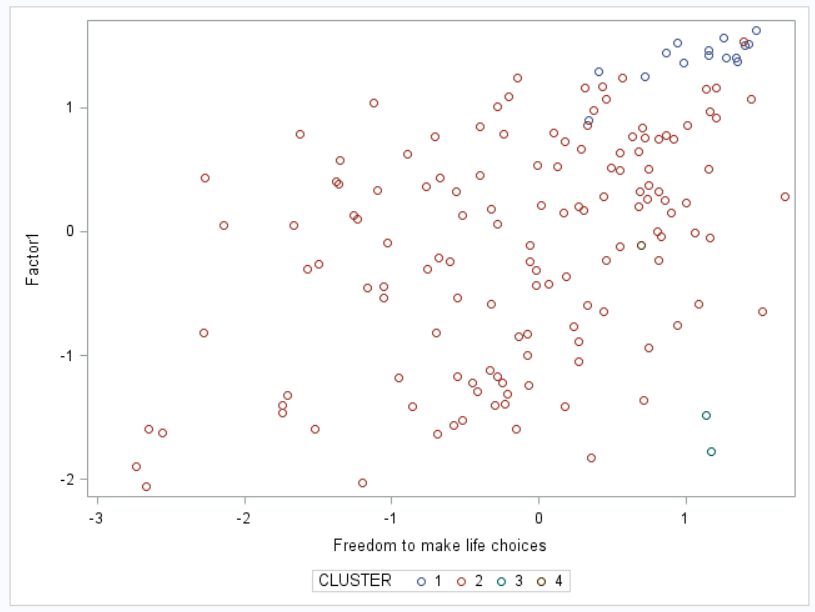


Se puede agrupar en 6 clusters o en 9, pero en 9 ya se dificulta la interpretación y por esto nos quedaremos con 6. Observamos que el grupo en el centro con este método se ha dividido en 3 clusters, disminuyendo la dispersión de cada cluster (el objetivo de este método) aunque los clusters 3 y 4 nos pueden causar problemas.



Con el método del centroide podemos quedarnos con 4 clusters, aunque la F es comparablemente pequeña.



2

El hecho de que los individuos de cluster 4 son indistinguibles y cluster 2 tiene una dispersión tan grande nos indica que esto no es el mejor método de clustering para este estudio. El método del vecino cercano tampoco encuentra una manera mejor para agrupas los individuos en clusters.

Podemos seguir con:

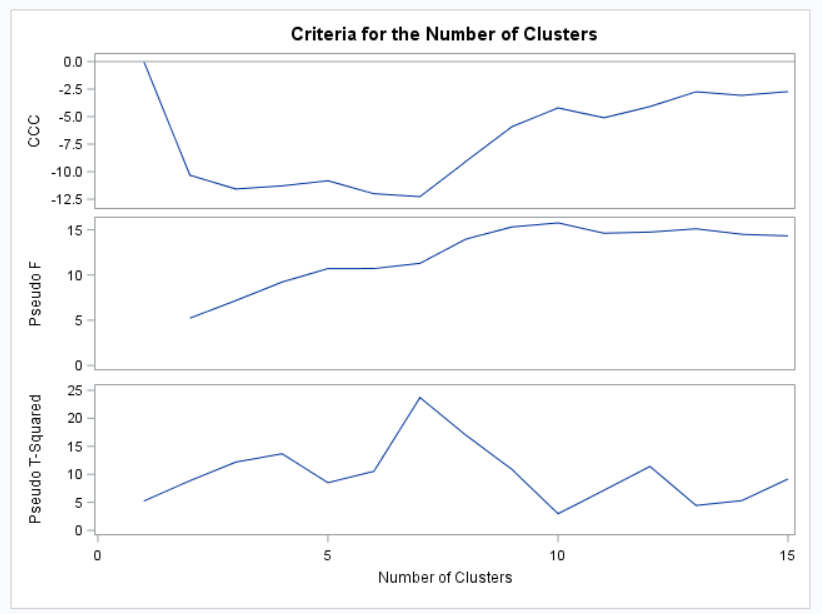
- método del vecino lejano clusters=3

- método de Ward clusters=6

- método average clusters=5

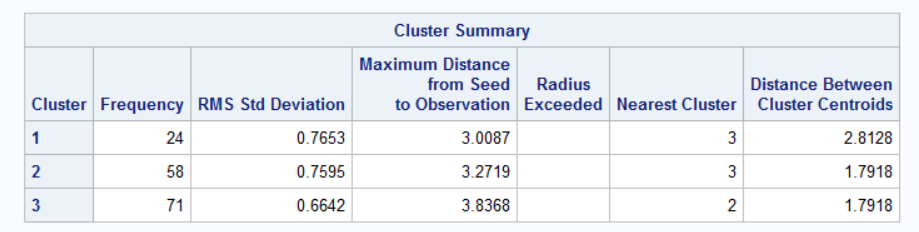
Por último, intentaremos encontrar 7 variables canónicas independientes con cuales ver si no mejoraría la agrupación con los tres métodos que hemos elegido. Es decir, no hacemos la agrupación a base de las variables originales, sino a base de estas nuevas.

Para los tres métodos observamos que se producen valores del estadístico F muy bajos y valores de t muy altos (los dos estadísticos toman valores casi iguales) por lo que no vamos a avanzar con este procedimiento y volveremos a nuestros datos originales.

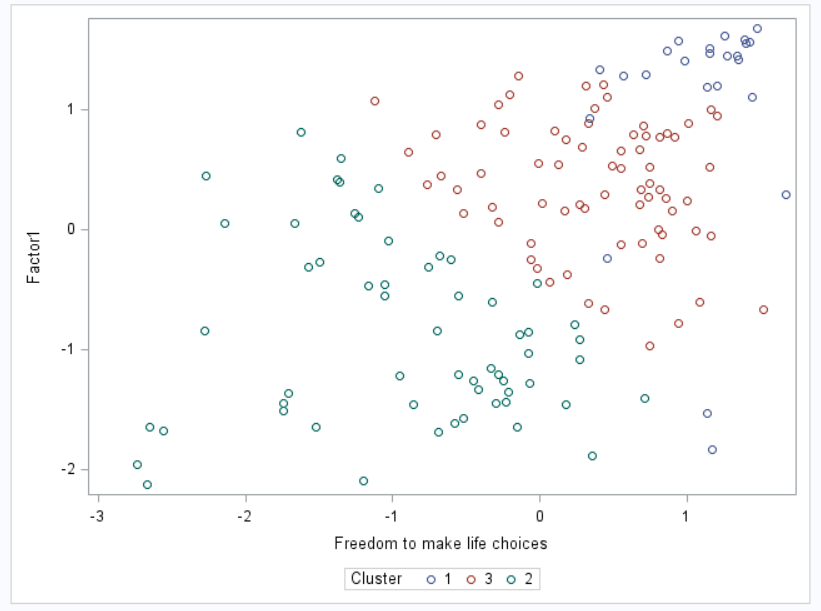


Análisis Cluster No Jerárquico

En esta parte del estudio, habiendo elegido algunos métodos aptos para dividir la población en varios clusters homogéneos vamos a ver si estos clusters no se pueden mejorar usando el análisis cluster no jerárquico. Se calcularán los centroides de los clusters formados y se reasignarán las observaciones según cual centroide le esté más cerca. Empezamos por el método del vecino lejano.



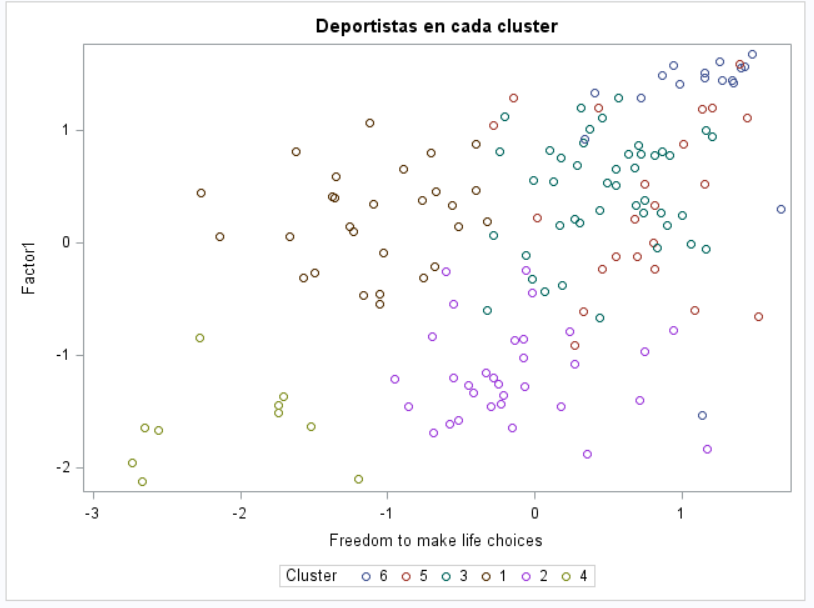
Vemos que la distancia máxima entre el centroide y la observación más lejana es mayor que la distancia entre los centroides de los clusters, que puede causar problemas. Sin embargo, observando el gráfico parece que los grupos son muy distinguibles.

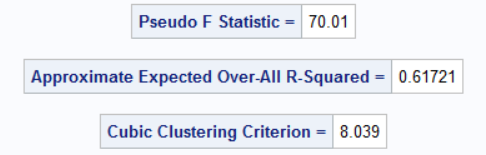


Finalmente podemos ver que los estadísticos F y CCC mejoran significativamente con la nueva asignación.



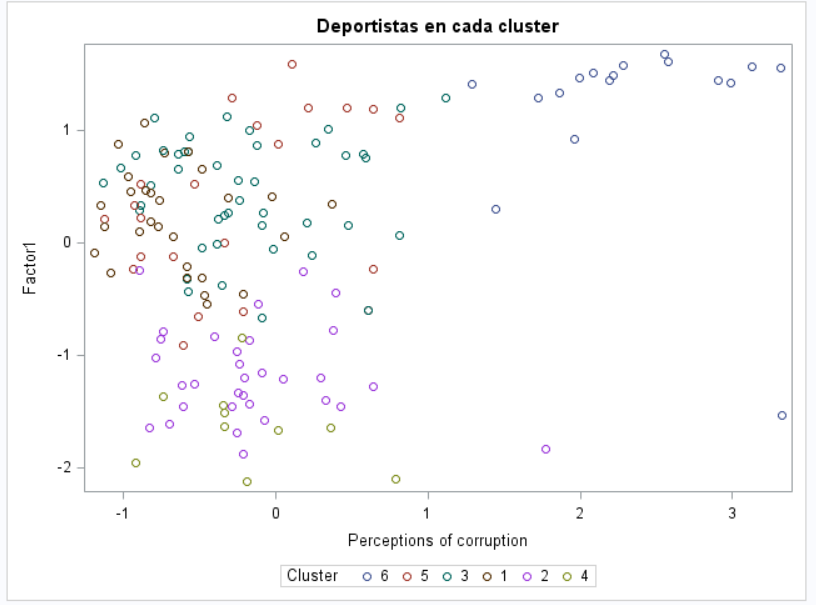
Seguimos con el mismo procedimiento para el método de Ward.





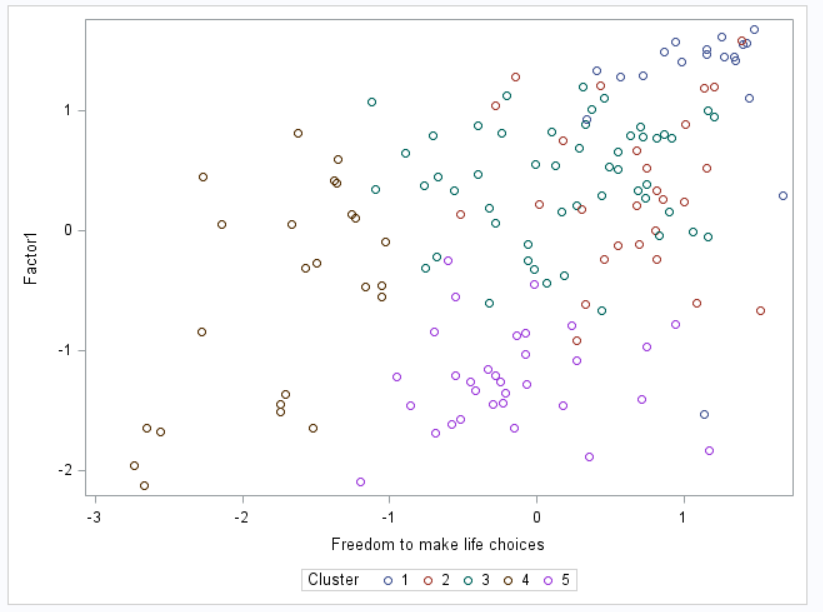
Comparando los dos gráficos de dispersión antes y después del análisis no jerárquico, solo viendo el grafico parece que la agrupación ha empeorado, ya que las observaciones de grupo 1 y 2 están muy mezcladas. Sin embargo, los estadísticos F y CCC han mejorado (un poco menos que con el método del vecino lejano) y también la R^2 que indica mayor diferencia entre clusters que por ejemplo teníamos con el método anterior.

Probamos graficar con otras dos variables (al final del estudio se explica mejor porque elegimos estas dos):

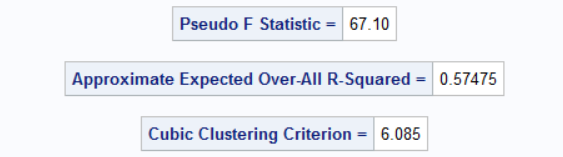


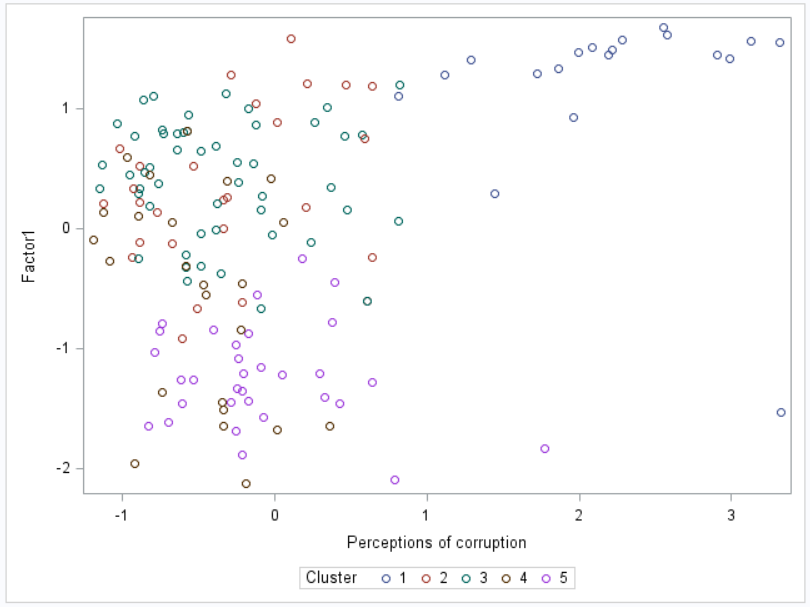
Tampoco mejora la separación de los individuos de grupo 1, 2 y 5.

Por último, estudiaremos para el método del enlace medio si mejora la agrupación.



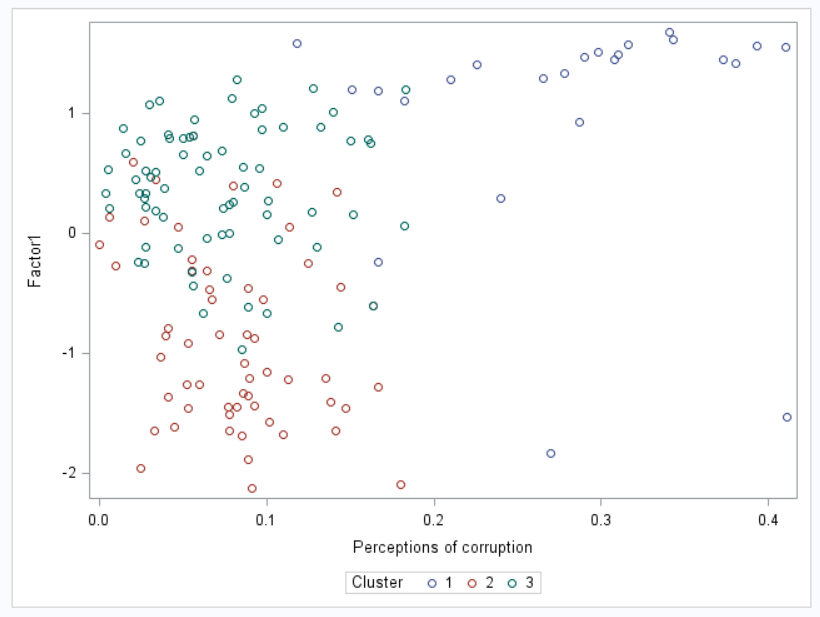
De nuevo, clusters 2 y 3 son muy poco distinguibles. Podemos cometer muchos errores clasificar individuos a uno de los dos grupos.





Finalmente, viendo la asignación para cada método y los estadísticos lo lógico será quedarme con los 3 clusters del método del vecino lejano después de la reasignación, ya que tiene el F y CCC mayor. Aunque la R^2 es la más pequeña, los otros estadísticos son más de fiar. Aunque parecía en el grafico que grupo 2 es demasiado disperso, graficando con la variable percepción de corrupción (que luego vemos que discrimina mejor) que con libertad podemos ver que esto cambia.

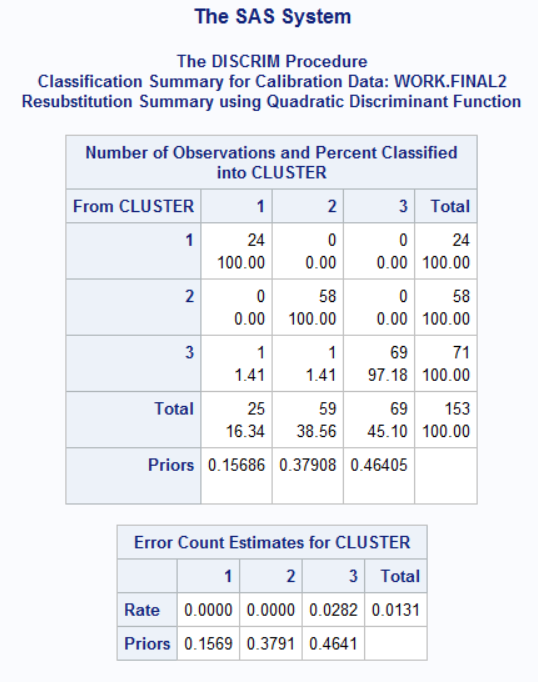
Parece que los grupos son mucho más compactos, salvo grupo 1 que combina observaciones con la tasa más baja de corrupción (cuanto mayor el valor tanto menos es la corrupción dentro del país), pero con distintos niveles de calidad de vida (Factor1).

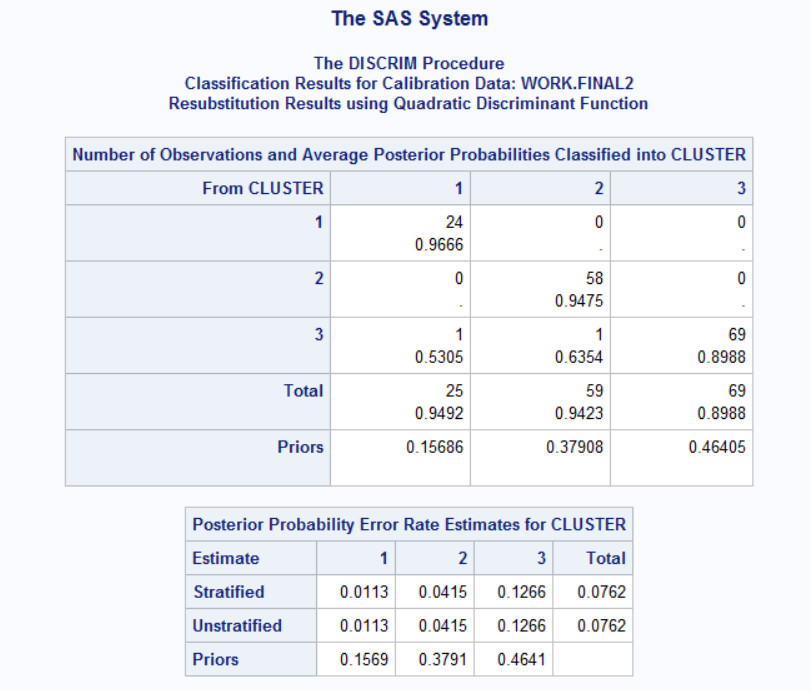


# **Análisis Discriminante**

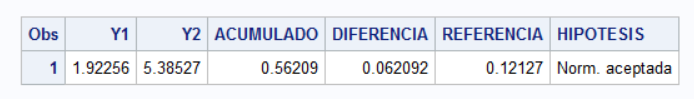
El siguiente parte de este estudio es el análisis discriminante. Ya tenemos 3 clusters (grupos) formados a base de las 4 variables independientes Freedom\_to\_make\_life\_choices, Generosity, Perceptions\_of\_corruption y Factor1. El siguiente paso es estudiar estas 4 variables para determinar cuáles son mejores para discriminar y clasificar las observaciones cometiendo el mínimo error posible. También comprobar si podemos encontrar funciones con cuales clasificar bien nuestras observaciones en los 3 clusters que tenemos.

Empezaremos por el análisis discriminante para las variables originales que he mencionado previamente. Vamos a ver si con ellas se puede discriminar entre grupos con un error muy pequeño.



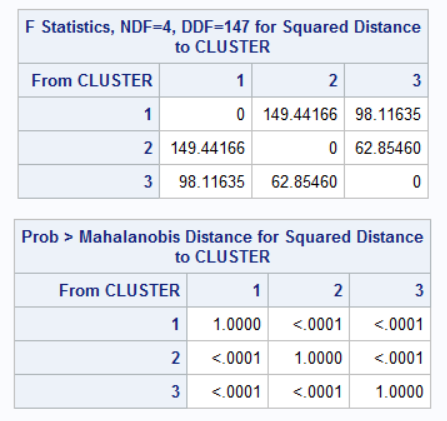


Se puede apreciar que el error a priori es 0.0131 y a posteriori 0.0762, es decir bastante pequeño. Casi no se cometen errores clasificando los individuos en los grupos, salvo en grupos 3 que clasifican erróneamente en total 2 observaciones. Aquí utilizaremos una función cuadrática, ya que no se acepta H0 de igualdad de varianzas, pero la hipótesis de normalidad sí.



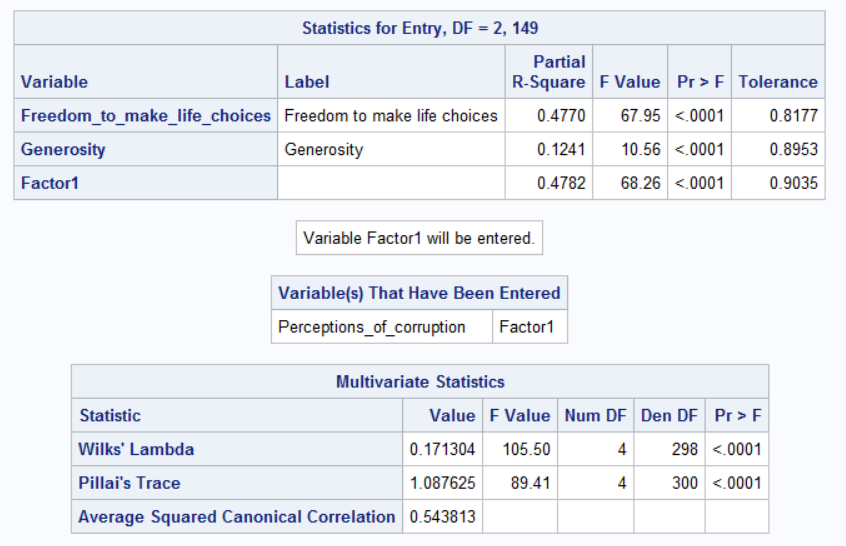
Antes de crear la función definitiva vamos a explorar otros métodos discriminatorios para ver si no encontraremos otra función discriminatoria con que se comete aún menos error.

Lo que podemos hacer es comprobar si en un principio las variables que estábamos utilizando hasta ahora discriminan bien. Luego estudiaremos cuales de las variables con que nos hemos quedado discriminan mejor, basándonos en el criterio de que la media para la variable es muy distinta entre grupos y la dispersión interna es muy pequeña.



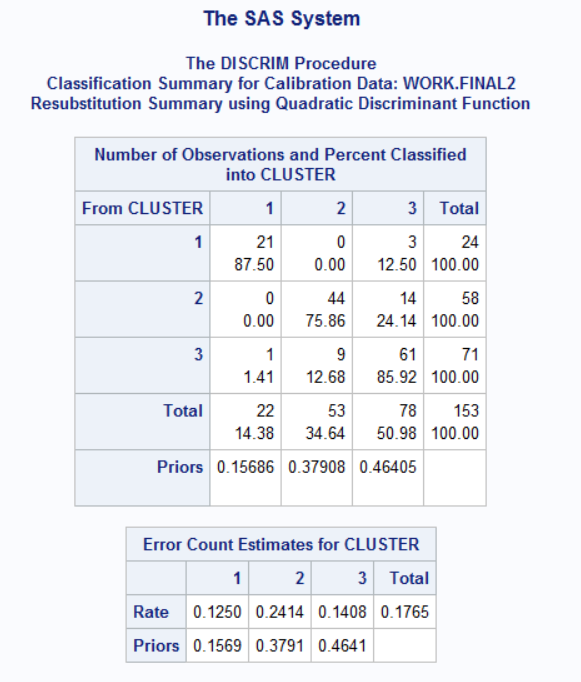
De la información en estas dos tablas podemos confirmar que la distancia entre clusters es significativa para todos los clusters, es decir las variables originales discriminan bien. De todas formas, vamos a ver si las que discriminan mejor son las que he elegido inicialmente.

El resultado de PROC STEPDISC nos enseña que hemos acertado para Factor1, que es una de las dos variables que discriminan mejor, pero la segunda no es la libertad, pero la percepción de corrupción. Aunque hemos visto que la libertad tenía mayor rango de la media entre clusters, la percepción de corrupción tiene menor rango, pero también menor dispersión, que es otra condición importante.

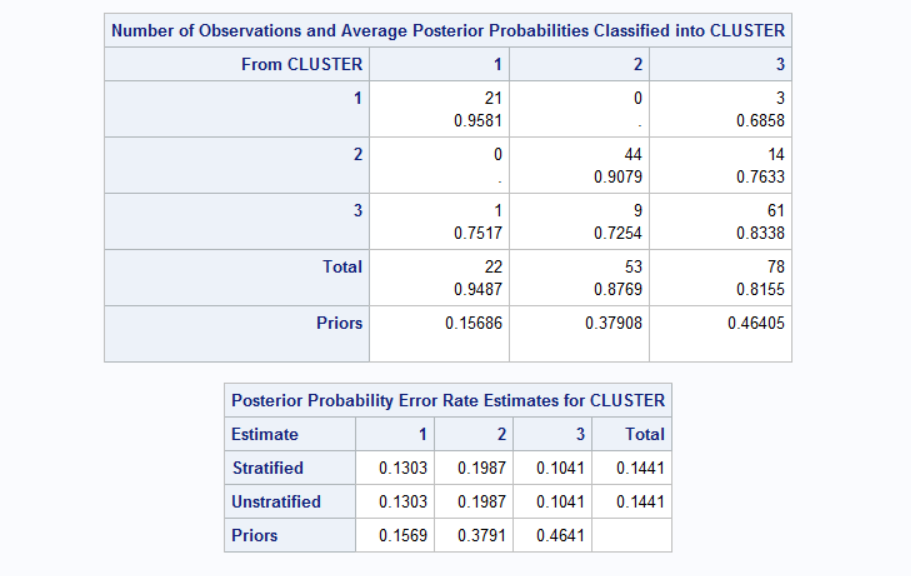


Vamos a ver cuál será el error cometido si intentamos discriminar solo a base de las primeras dos variables elegidas.

A priori:



A posteriori:

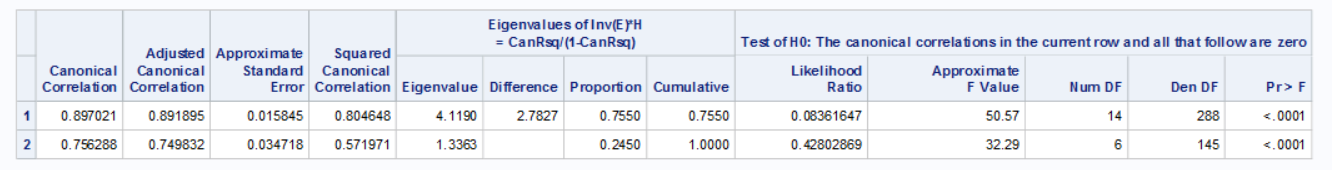


El error para ambos a priori y a posteriori es mayor que el error que cometemos usando las 4 variables que nos indica que solo las 2 no son suficientes para discriminar con mínimo error de asignación.

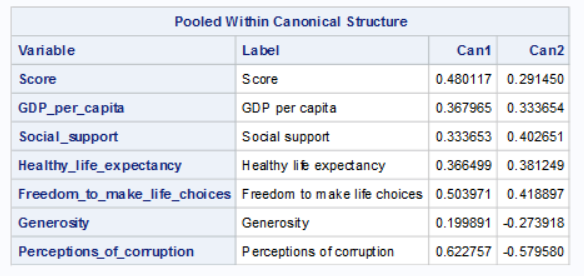
Análisis Discriminante Canónico

Otra técnica que podemos probar para conseguir clasificar bien nuestros individuos en los grupos respectivos es a través de grupos canónicos. El **análisis discriminante canónico** encuentra combinaciones lineales de las variables cuantitativas que proporcionan la máxima separación entre clases o grupos (una técnica de reducción de dimensión parecida a CP).

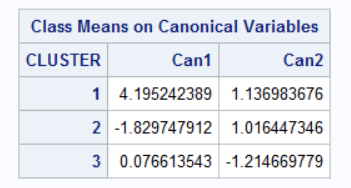
En un principio vemos que se han formado min{grupos-1;variables}=2 factores canónicos, de cuyos autovalores podemos apreciar que el primero explica casi 76% de la variabilidad.



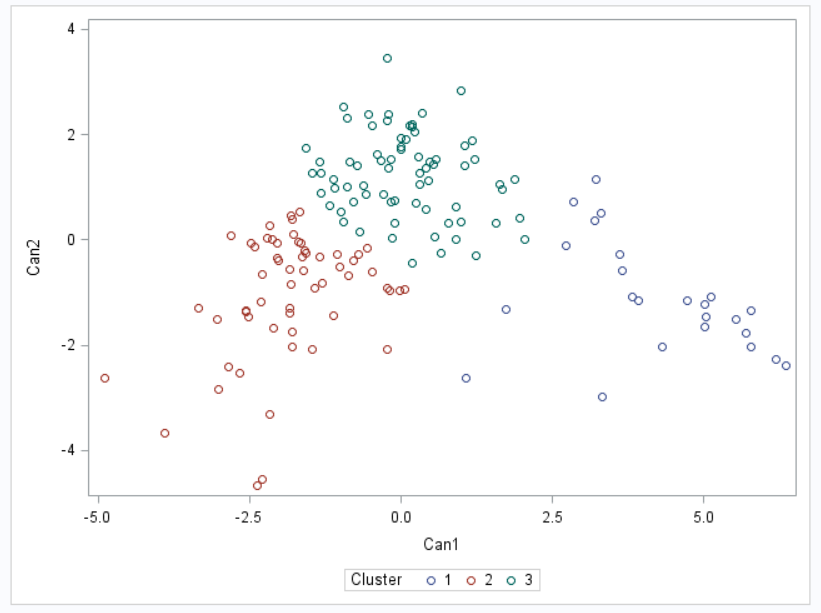
Observando la matriz de cargas podemos ver que CAN1 está explicado por la corrupción, la libertad de hacer decisiones y el índice de felicidad (score) es decir en el eje x, cuanto más a la derecha es un país, menos corrupta, más libre y más feliz es. CAN2 también está muy explicada por el nivel de corrupción, pero de forma negativa, entonces los países menos corruptos están cerca del cero en el eje y, también, aunque en menor grado CAN2 está explicado por las ayudas sociales y la esperanza de vida sana.



Por último, se puede concluir que los que tienen mejor percepción de la corrupción (tienen muy poca corrupción) son de grupo 1, seguidos por grupo 3 y finalmente grupo 2 (los más corruptos). CAN1 explica 75% de la varianza, así que nos fiamos principalmente en su capacidad discriminatoria.

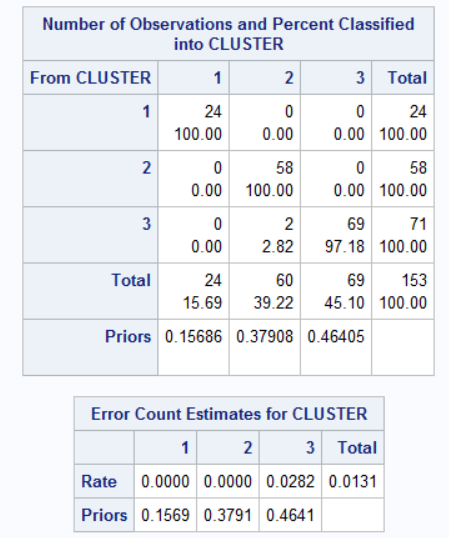


Vemos el gráfico de dispersión:

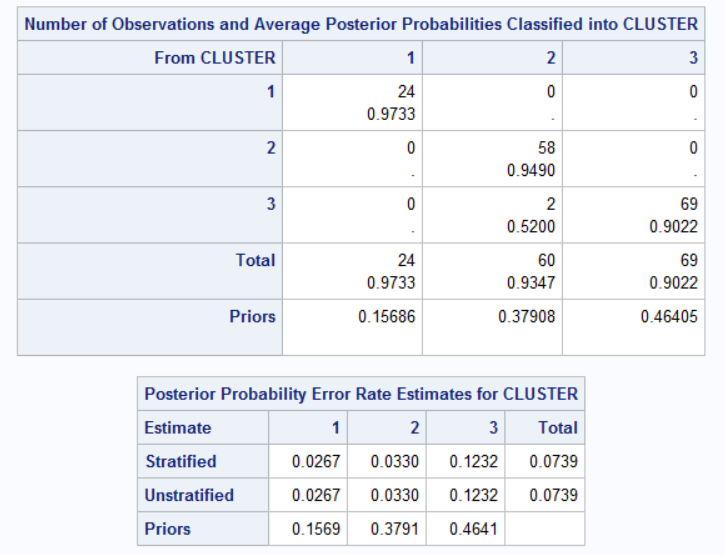


Finalmente, comprobaremos el error que se comete cuando discriminamos con las 2 variables canónicas y definiremos la función cuadrática de discriminación.

A priori:



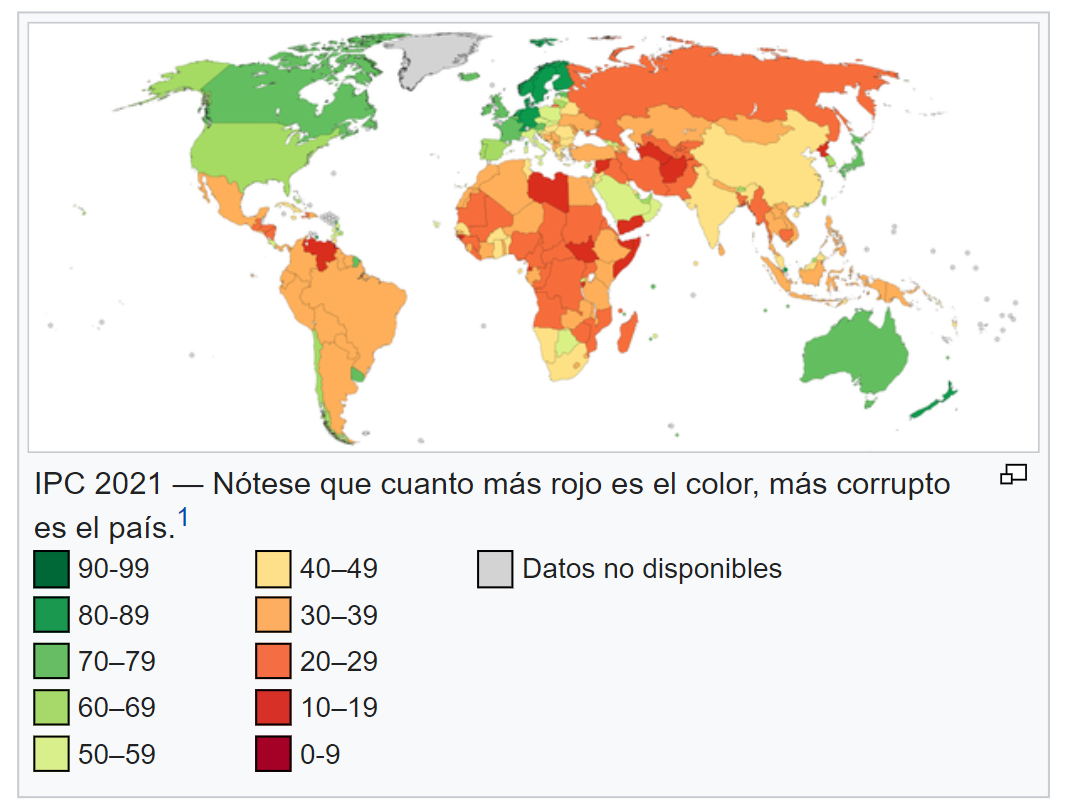
A posteriori:

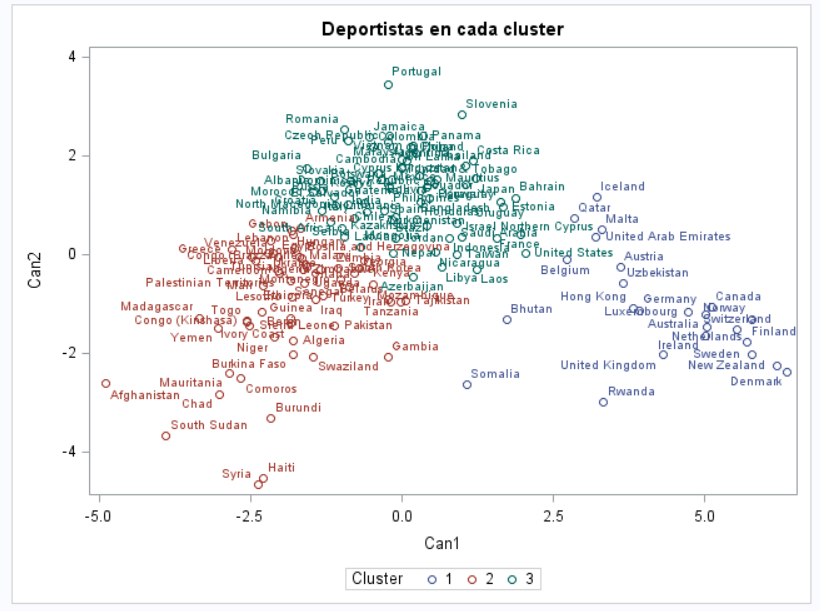


Observamos algunas cosas muy importantes – en un principio se aprecia que el error a posteriori es un poco menor que usando las variables originales (no canónicas) y el error a priori se queda igual. Segundo, mirando el gráfico parece que los clusters son mucho más homogéneos y compactos, es decir cometeremos menos error asignando nuevas observaciones a los clusters.

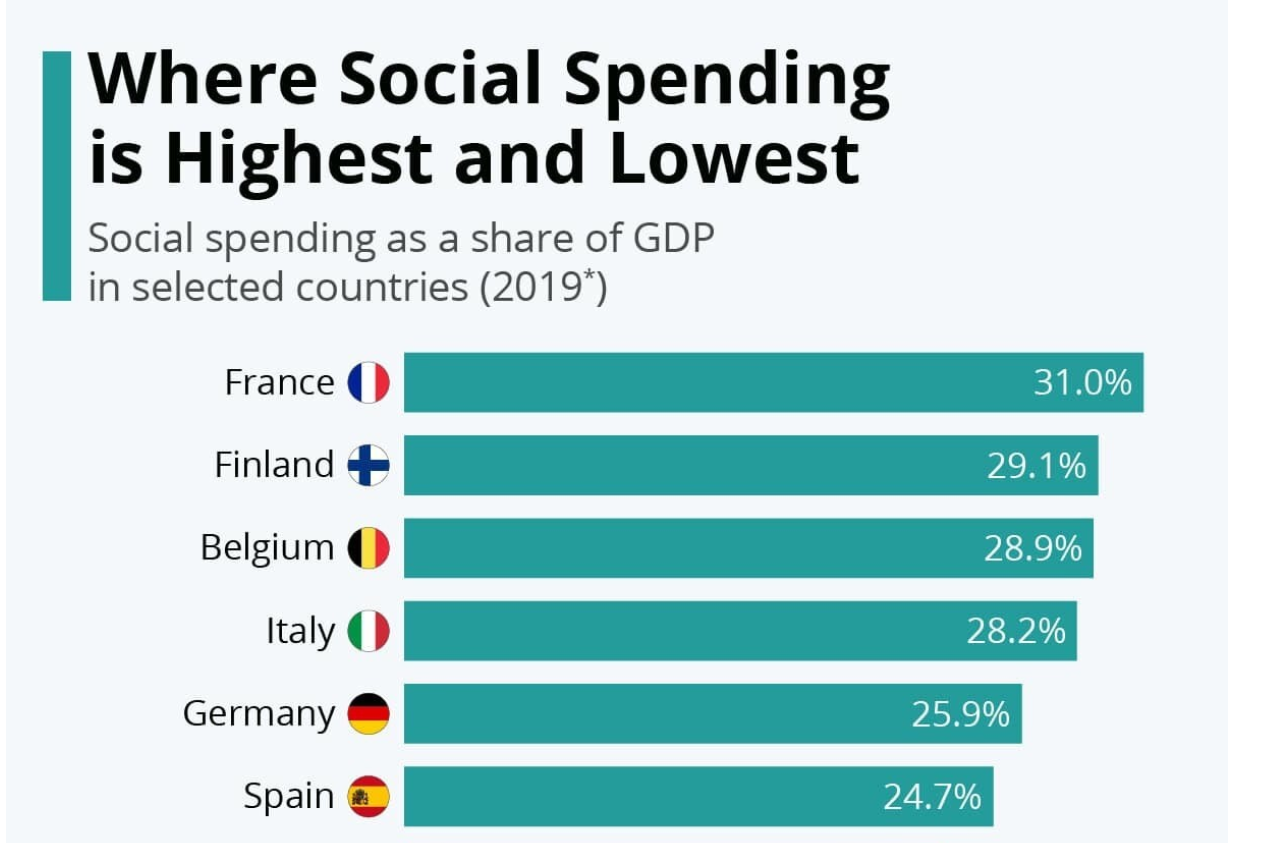
Tercero, como la variable CAN1 (altamente explicada por la corrupción) es la que más discrimina, será interesante ver si hemos podido clasificar nuestras variables entre los 3 clusters de acuerdo con la realidad (con datos oficiales). He sacado un mapa de calor de Wikipedia que enseña los países más corruptos, que lo compararé con mis clusters.

Los menos corruptos son países como Australia, Canadá que entran en cluster 1. Los más corruptos como Venezuela, Yemen y Afganistán entran en grupo 2.





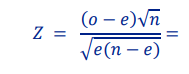
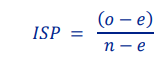
Aunque CAN2 es un poco menos interpretable y explica menos de la varianza, hemos dicho que hasta cierto punto está explicada por las ayudas sociales y la esperanza de vida sana. Entonces los países que están cerca del 0 en el eje y son los que tienen menos corrupción, más ayudas sociales y esperanza de vida sana mayor.



Podemos ver que se acerca a la realidad. Por esto, la decisión final será quedarme con 3 clusters y con las dos variables canónicas.

# **Validación del Modelo**

Según el test estadístico de idoneidad del Análisis Discriminante obtenemos un valor de 15.34. El ISP=0,97 con que podemos afirmar que esto es una buena regla discriminante.

La función discriminante para grupo (cluster) 3 será:

Grupo3 = -1,64+0,43\*Can1+1.98\*Can2+2\*(-0.14)\*Can1\*Can2-0,79\*(Can1)^2-0.66\*(Can2)^2

De la misma forma se calularán de Grupo1 y de Grupo2 y una nueva observación entaría en el grupo para cual maximice el valor de su función.