Análisis de Happiness

Trabajo 1

VARIABLES DE LA BASE DE DATOS

En esta base de datos tenemos 8 variables:

Country\_or\_region = país

GDP\_per\_capita = PIB per cápita

Score = índice de felicidad de los residentes

Social\_support = apoyo social

Healthy\_life\_expectancy = esperanza de vida sana

Freedom\_to\_make\_life\_choices = libertad de elegir libremente

Generosity = generosidad

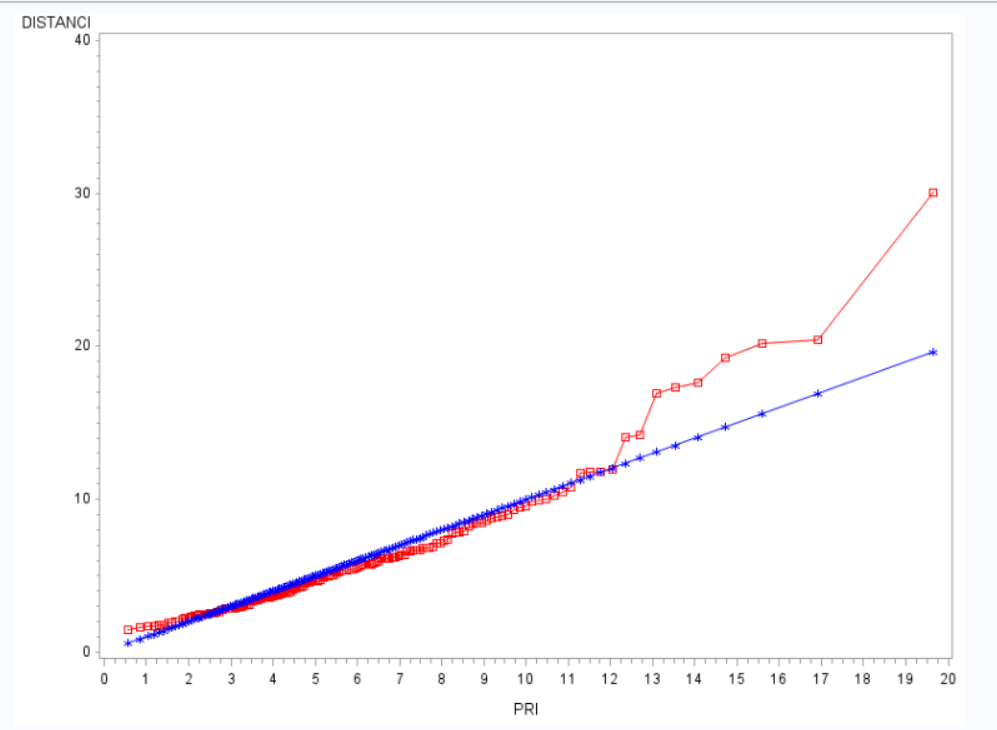
Perceptions\_of\_corruption = percepción de corrupción

ANALISIS DE BONDAD DE LOS DATOS

Empezamos por comprobar que la base de datos es válida para hacer un estudio de Componentes Principales, Análisis Factorial y últimamente análisis de Correspondencias.

El primer paso es ver si hay valores perdidos y si hay, tomar decisión como rellenarlos. En esta base de datos no hay valores perdidos. Luego, comprobamos los valores atípicos. En la base de datos tenemos tres: *Singapore, Myanmar y Central African Republic.* Creamos una nueva base de datos, copia de la original, donde quitamos las observaciones atípicas para hacer un estudio ambos con y sin outliers.

Finalmente comprobamos si los valores siguen una distribución Normal por separado y multivariante. Ejecutando la macro de normalidad podemos observar una discrepancia importante entre la distribución normal y nuestras variables para valores más grandes. También no se puede aceptar la hipótesis 0 que el apuntamiento y asimetría son de una distribución normal. Sin embargo, se acepta la normalidad multivariante. Por separado, solo Generosity y Score se distribuyen normalmente.





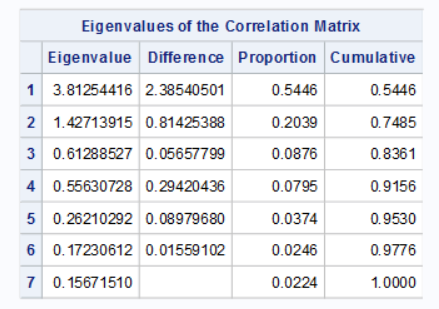
ANALISIS DE MULTICOLINEALIDAD

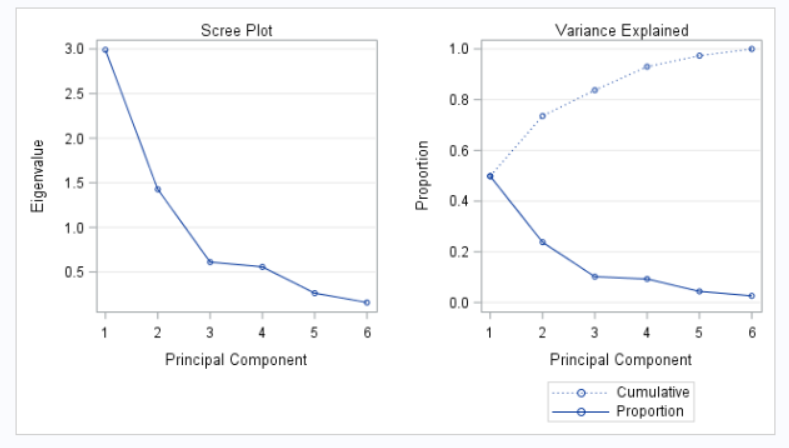
Haciendo un PROC CORR podemos observar que hay dependencia entre variables, se puede hacer estudio de CP y Análisis Factorial. Otra observación es que las variables tienen varianzas diferentes. Para no cometer el error de que las variables con mayor varianza pesen mas en el estudio, es decir que la mayoría de la varianza total está comprendida en el primer componente, tenemos que tipificar.

COMPONENTES PRINCIPALES

La idea de este estudio es poder explicar la máxima varianza posible (inercia) con menos variables posible. Como se ha mencionado arriba, se hará con correlación, por la diferencia en varianzas de las variables.

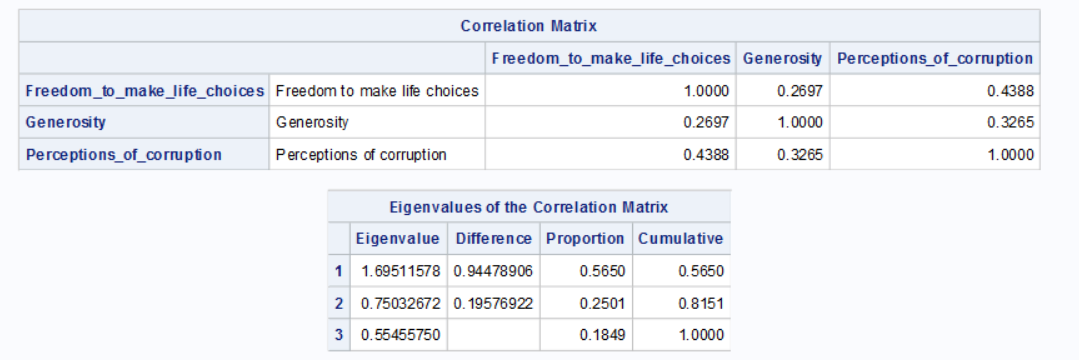
Del estudio obtenemos:



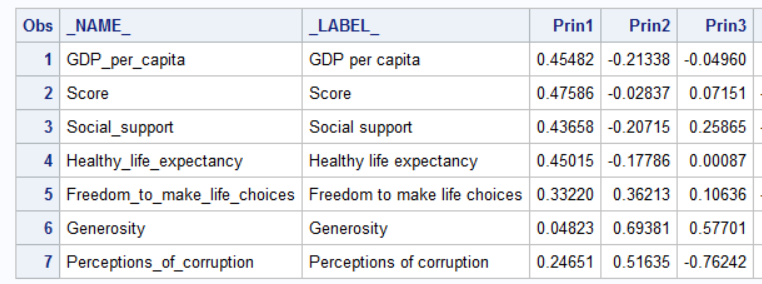


Podemos ver que con 2 componentes podemos explicar 74,85% de la variabilidad y con 3 – 83,61%. Con la regla de Cattel también se pueden retener 3 o 4 componentes y si buscamos retener los componentes cuyos autovalores son mayores que 1, retenemos 2. Como el estudio de CP es solo para darnos una idea sobre la base de datos, en el AF posterior vamos a considerar las opciones de 2 y también de 3 factores.

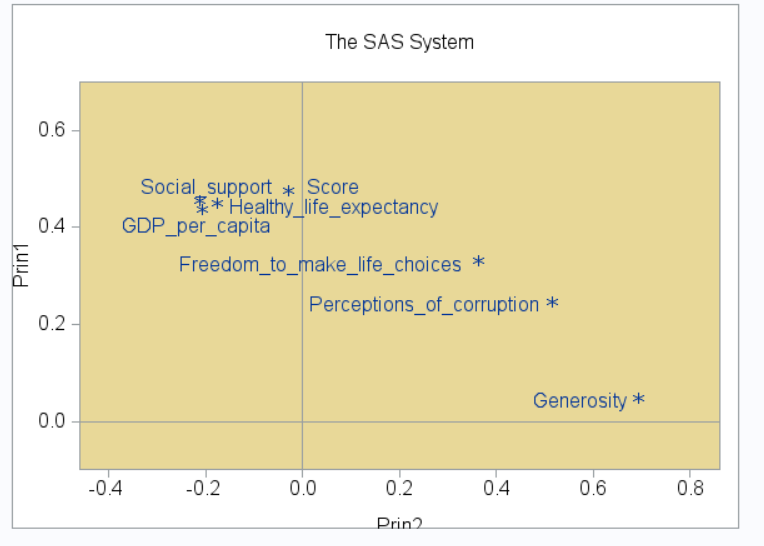
También una observación que se ha hecho sobre las variables de la base de datos - los tres variables - libertad, generosidad y corrupción están poco correlacionados con el resto de las variables y también entre sí. Por esta razon simpre se quedan poco explicadas por los factores que retenemos en el análisis factorial.

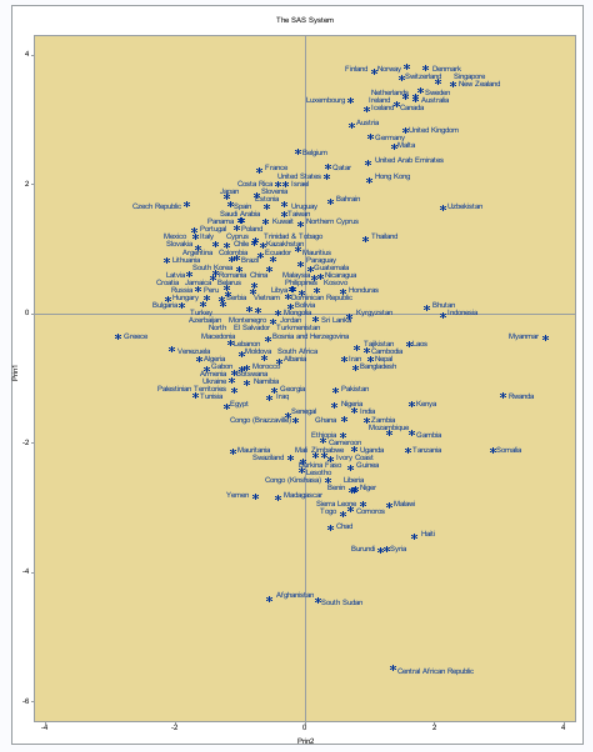


Volviendo al CP, podemos ver que Prin1 está explicado por el índice de felicidad, el PIB per cápita y las ayudas sociales, Prin2 por la Generosidad y la percepción de la corrupción y Prin3 de la Generosidad y la corrupción, pero con una relación negativa.



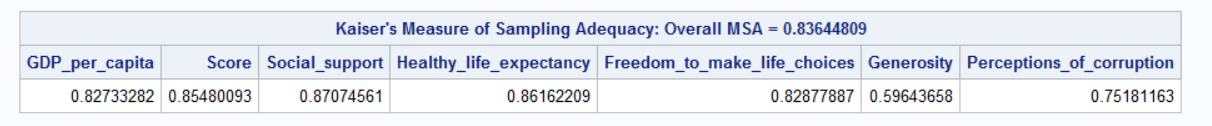
Representación gráfica de Prin1 con Prin2 de las variables y las observaciones (los países):





ANALISIS FACTORIAL

Tras hacer el análisis de componentes principales, vamos a hacer Análisis Factorial, cuyo objetivo es encontrar variables latentes en la base de datos, que explican el mayor porcentaje posible de la varianza total. Vamos a probar tres métodos - PRINCIPAL, de CP, PRIORS de Factor Principal y Máxima Verosimilitud, ya que anteriormente hemos aceptado la normalidad multivariante.

Comunalidades Iniciales:

Podemos apreciar que todas las variables están bien explicadas en más que 50% y también que el KMO es de 83.64, no hace falta quitar variables de estudio para mejorar el índice.

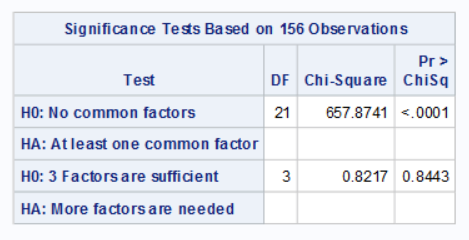
Con el método de Componentes Principales PRIORS=one, con tres factores explicamos 83,61% de la variabilidad total. Con este método conseguimos explicar todas las variables (lo vemos en las comunalidades finales) pero no nos quedamos con este, ya que la variable generosidad tiene residuo de 0.121 y 37% de su variabilidad se queda sin explicar.

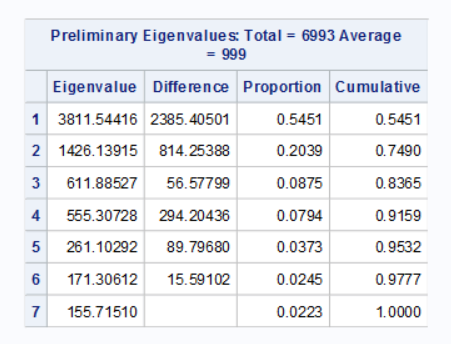
Usando el mismo método, pero con PRIORS=smc, no se consiguen explicar bien todas las variables, aunque con 2 factores se puede explicar 100% de la variabilidad total.

Haciendo el método del Factor Principal, observamos que de nuevo con 2 factores se explica 100% de la variabilidad, sin embargo, se quedan 3 variables sin explicar, así que este método no me vale (esto ambos para priors=one y smc).

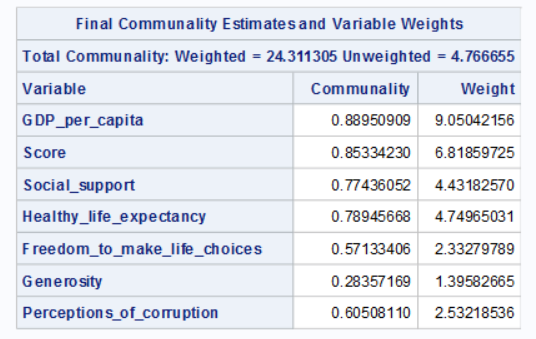
Hasta aqui veo que los métodos PRINIT con priors = (one, smc) y Principal con priors=smc explican con dos factores 100% de la variabilidad, es decir mejora el modelo, pero al final resulta que se quedan 3 variables sin explicar por estos dos factores. Por otra parte, los componentes principales explican casi 75% de la varianza, explicando todas las variables por los 2 factores.

Finalmente lo intentamos con Máxima Verosimilitud. No se puede rechazar que 3 factores son suficientes, así que vamos a trabajar con 3.



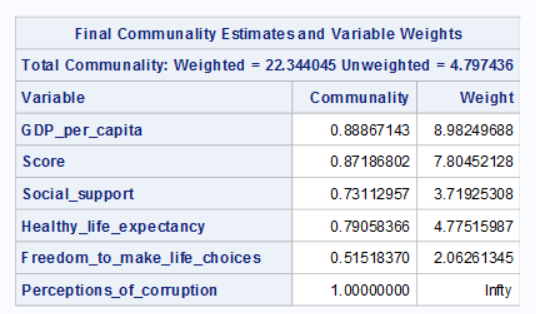


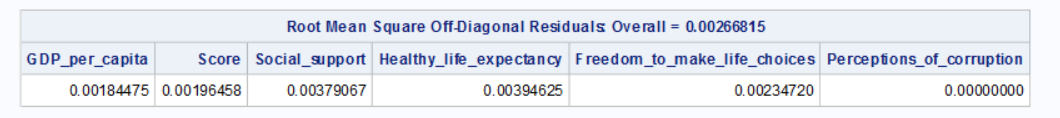
Podemos ver que, con 3 factores conseguimos explicar 83,65%, que mejora ligeramente los PC (con 3 factores se explica 83,61%). Mirando las comunalidades podemos ver que la generosidad se queda sin explicar. Por esto la vamos a quitar del estudio.



Sin Generosidad nos queda que con 3 factores se consiguen explicar todas las variables, tienen residuos muy pequeños (menores que 0,07) y el porcentaje de la variabilidad que se queda por explicar también es mínimo. Sin embargo, no se puede aceptar la hipótesis de que 3 factores son suficientes.

Probando con 2 factores tenemos un p-valor de 0.0030, con que tampoco podemos aceptar la H0 y con 4 ya tenemos los grados de libertad negativos. Con objetivos didácticos me voy a quedar con este método para poder rotar los 3 factores retenidos.







ROTAR

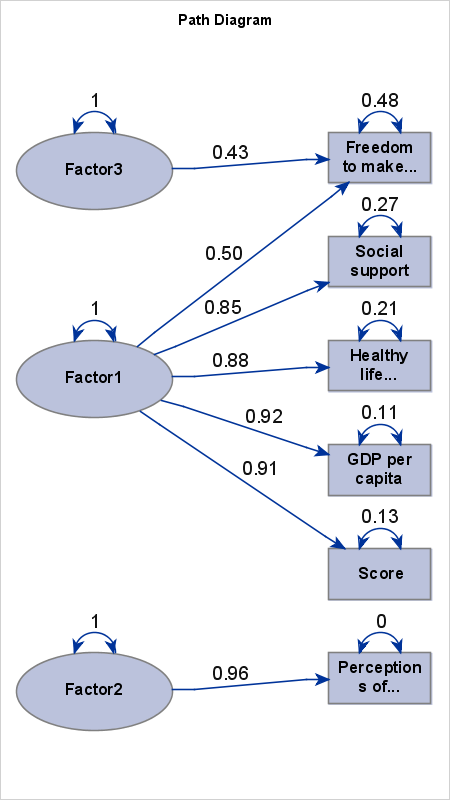
Cuando hemos elegido con que método trabajar y cuantos factores retener es tiempo para dar nombre a los factores latentes, para interpretar mejor los resultados del estudio. Vamos a usar la rotación VARIMAX y QUARTIMAX y ver cual nos satisface.

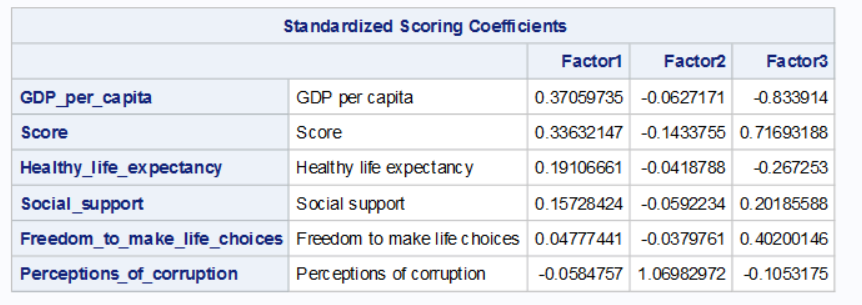
Al final, hemos elegido QUARTIMAX, ya que se pueden dar nombres más coherentes a los tres factores. Visto el gráfico, podemos decir que

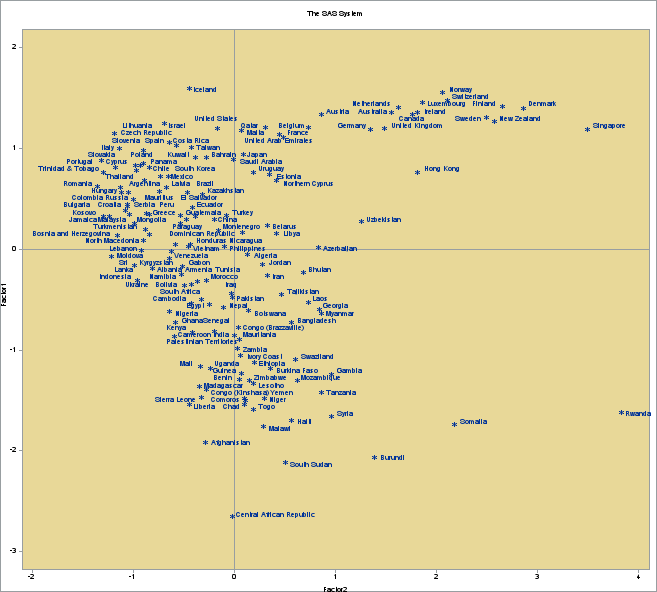
FACTOR 1 Calidad de Vida - el dinero que ganan, esperanza de la vida sana, ayudas sociales

FACTOR 2 Percepción de la corrupción

FACTOR 3 Libertad







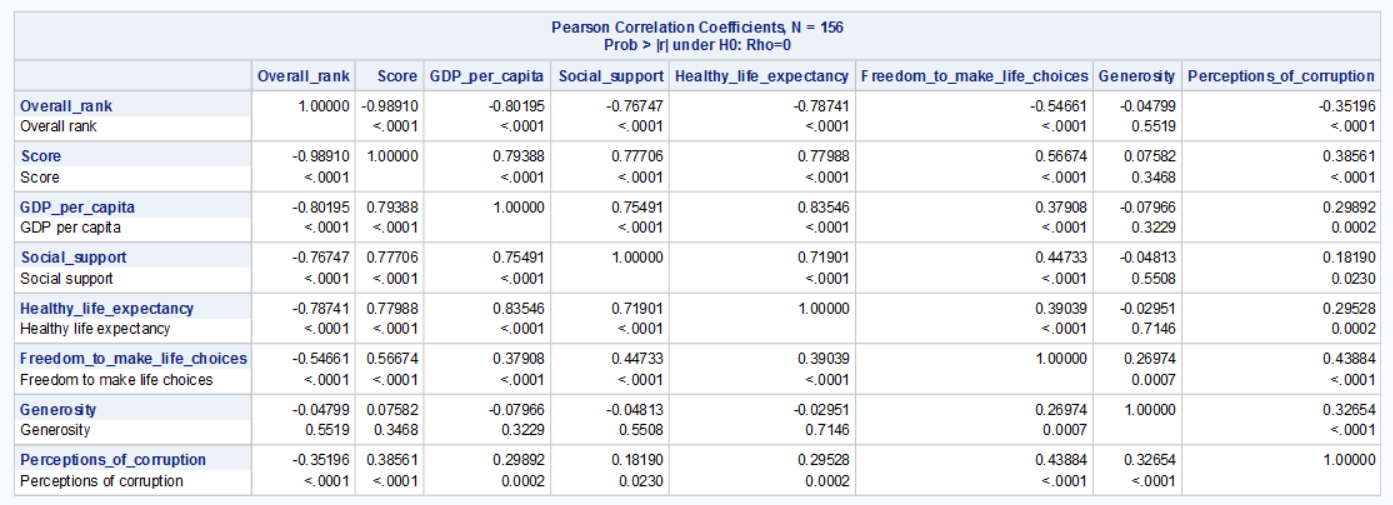
Finalmente, haciendo el estudio sin los valores atípicos, no obtenemos resultados muy diferentes. Todos los resultados comentados arriba han sido obtenidos para la base de datos sin observaciones atípicas.

De todas formas, volvemos al asunto de que no hemos encontrado un método realístico para hacer el análisis factorial. Para terminar el estudio vamos a usar la base de datos sin observaciones atípicas.

Recordamos que el método de CP consigue explicar todas las variables, pero deja residuos mayores que 0,12 para la generosidad. Quitando esta variable no mejora los resultados.

El método del Factor Principal no consigue explicar la generosidad, libertad y corrupción y quitando la que menos se explica por el factor retenido no mejora los resultados. Tampoco se consigue convergencia de las comunalidades cuando se requiere que se retengan 2 factores y se aumenta el máximo número de iteraciones hasta 100.

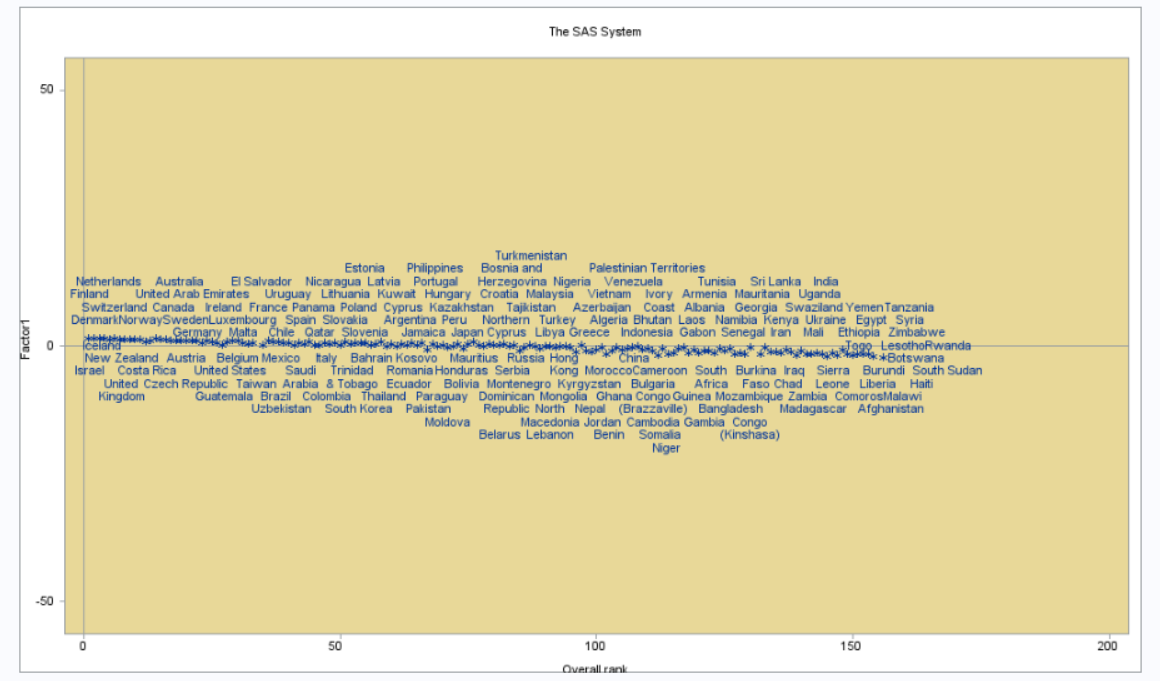
El método de máxima verosimilitud ha sido detalladamente explicado más arriba. El problema general viene de que la variable generosidad no está relacionada con casi ninguna otra variable, salvo la libertad y la corrupción. Con esto, ningún método consigue explicarla, pero cuando la quitamos, las variables libertad y corrupción se quedan sin explicar.



La decisión más lógica será quitar estas tres variables del estudio y explicar el resto de las variables con 1 factor, que tendrá el mismo nombre, que hemos elegido antes (“calidad de la vida”). En este caso, el mejor método será el Factor Principal ya que con un factor conseguimos explicar 99,98% de la variabilidad total, todas las variables se consiguen explicar, el RMSR=0.0234 y lo que se queda sin explicar es muy poco.

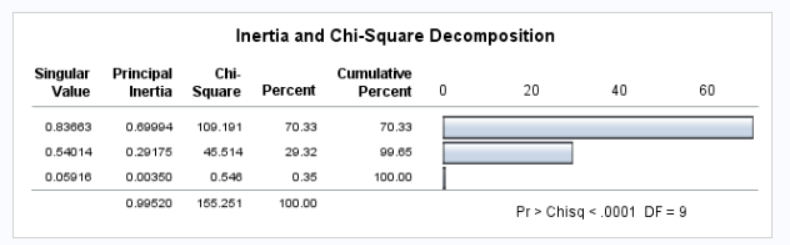
El método de ML explica más de la variabilidad con un factor, pero la H0, de que un factor es suficiente se tiene que aceptar con nivel de significación 0,01 (p-valor=0.0125). Finalmente, CP explica menos de la variabilidad con 1 factor (82,7%).

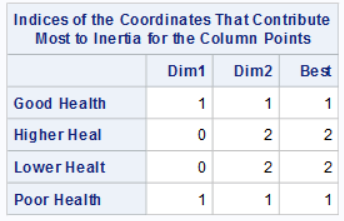
Con esta decisión no hace falta rotar, ya que hemos retenido un único factor.

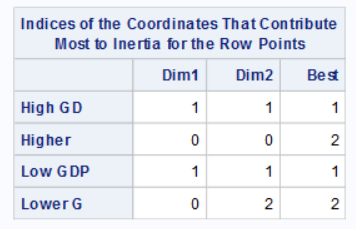


ANALISIS DE CORRESPONDENCIAS

Originalmente en la base de datos no hay valores cualitativos, sin embargo, haciendo un analisis de correlacion se puede ver que las variables PIB per capita y Esperanza de vida son las más relacionadas, así que vamos a discretizar ellas y ver como se relacionan con un análisis de correspondencias.







Por filas:

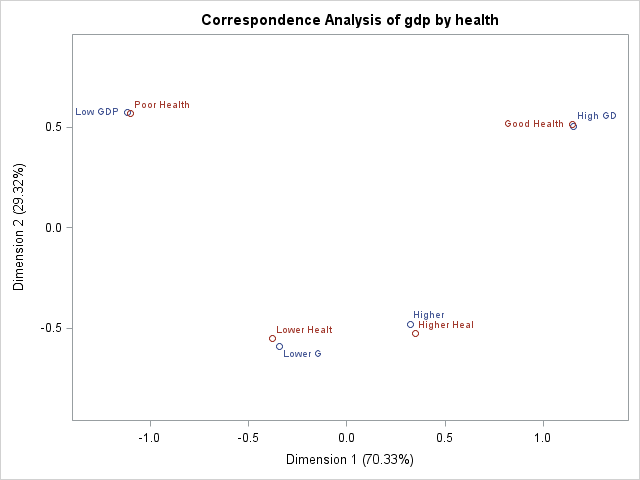
-Dim1: Las categorías que mejor explican Dim1 son los dos extremos - PIB per capita más alto y el más bajo

-Dim2: La categoria PIB per capita medio alto y la PIB per cápita medio bajo

Por columnas:

-Dim1: Las categorias que mejor explican Dim1 son tambien los extremos - esperanza de vida alta y baja

-Dim2: La speranza de vida medio alta y la medio baja

Como era de esperar (y como sucede realmente) las categorías que están en mismo cuartil estan más relacionadas. Es decir, al PIB per cápita alto le corresponde una esperanza de vida alta. Para PIB medio alto - esperanza medio alta y así las 4 categorías.