Primer examen parcial - Reconocimiento estadístico de patrones

Erick Salvador Alvarez Valencia

CIMAT A.C.,
erick.alvarez@cimat.mx

1. Introducción

En el presente reporte se mostrarán los resultados obtenidos al analizar el conjunto de datos referentes a un estudio de felicidad realizado a varios países del mundo en los años 2016 y 2017 en el cual se añaden varias características importantes que afectan directamente a la felicidad de las personas, tales como: Libertad de tomar decisiones, esperanza de vida, corrupción, entre otras.
Se mostrará la información arrojada por diferentes métodos aplicados a los datos, así como una comparación final realizada entre los conjuntos del 2016 y 2017.

2. Desarrollo

Lo primero que se hizo al leer los datos fue realizar un filtrado, ya que en la tabla viene información sobre diferentes años y en nuestro caso sólo nos interesa la referente a 2016 y 2017, por lo que generamos dos tablas, la primera conteniendo la información con respecto a 2016 y la segunda a 2017. Una vez que se realizó dicho filtrado se notó que varias columnas de las tablas tenían mucha ausencia de datos, incluso algunas columnas estaban vacías, por lo cual se tomó la decisión de realizar un segundo filtrado donde se eliminaron las columnas que no aportaban información alguna, en general el filtro consistía en borrar las columnas que tuvieran un 90 % o más de ausencia de datos.

Como siguiente paso fue realizar un paso de imputación con respecto a las celdas con datos faltantes, hay que notar que este puede ser un paso crítico ya que pese a que los métodos de imputación tratan de llenar los espacios faltantes con datos que correspondan con la información actual, esto nos puede producir un sesgo y posteriormente alterar nuestros resultados al aplicar los diferentes métodos de clasificación y agrupamiento. La decisión de imputamiento se tomó cuando al aplicar PCA a los datos, este nos regresó proyecciones donde varias files se quedaron vacías, y mediante una pequeña revisión se notó que la mayor parte de dichas filas correspondían a países de Africa, y el hecho de no tener esa información podría ser crítico para los resultados finales ya que estos países podrían representar la gran parte de un grupo que es o no felíz.
El proceso de imputación de datos se realizó a través de una librería llamada

missMDA la cual tiene varios métodos de imputación para diferentes fines, en el caso actual fue aplicar una función de imputación para posteriormente hacer PCA, esta función rellena los espacios usando probabilidades Gaussianas con respecto a los datos de la columna actual.

Una vez que se tenía la información lista el primer método aplicado fue PCA, esto para ayudar a ver qué tanta variabilidad aportaba cada componente y de la misma forma obtener alguna información de los componentes.
A continuación se muestran los resultados de aplicar el método de PCA a los dos conjuntos de datos normalizados y centrados (usando matríz de correlación).

```
Loadings:
                                                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
Life.Ladder                                     -0.347         0.166        -0.224  0.248  0.128 -0.262  0.288          0.206   0.466            0.548
Log.GDP.per.capita                              -0.343         0.148  0.225 -0.187         -0.121  0.293 -0.245 -0.106  -0.208          -0.736
Social.support                                  -0.328         0.228 -0.119                        0.457  0.517 -0.457          -0.307   0.207
Healthy.life.expectancy.at.birth                -0.340  0.113  0.114  0.149 -0.199         -0.178  0.262 -0.462  0.215  -0.200   0.110   0.618
Freedom.to.make.life.choices                    -0.202 -0.437  0.153         0.238  0.400                -0.265  0.196   0.460  -0.451
Generosity                                             -0.336 -0.404 -0.629 -0.395         0.212  0.290 -0.102  0.114
Perceptions.of.corruption                        0.221  0.292  0.314 -0.524  0.194  0.190                -0.393 -0.392   0.174   0.224            0.145
Positive.affect                                 -0.183 -0.394  0.316 -0.271        -0.136 -0.175 -0.514         -0.155  -0.527  -0.130
Negative.affect                                  0.309                0.183 -0.479  0.580  0.301 -0.101         -0.297  -0.216  -0.240
Confidence.in.national.government                       -0.468 -0.354  0.224  0.423  0.206          0.192         -0.375  -0.199   0.389
Standard.deviation.of.ladder.by.country.year     0.295 -0.178  0.327 -0.103         0.333 -0.357  0.198  0.320  0.332            0.308          -0.407
Standard.deviation.Mean.of.ladder.by.country.year 0.368                                      -0.271  0.266         0.226  -0.274  -0.271           0.711
GINI.index..World.Bank.estimate...average.2000.15 0.181 -0.244  0.508  0.126        -0.278  0.671  0.244                        0.125
gini.of.household.income.reported.in.Gallup..by.wp5.year 0.239 -0.334         0.243 -0.444 -0.380 -0.330        -0.145 -0.316   0.428

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var  0.071  0.071  0.071  0.071  0.071  0.071  0.071  0.071  0.071   0.071   0.071   0.071   0.071   0.071
Cumulative Var  0.071  0.143  0.214  0.286  0.357  0.429  0.500  0.571  0.643   0.714   0.786   0.857   0.929   1.000
> summary(PCA)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12   Comp.13
Standard deviation     2.5425560 1.6072962 1.2365822 0.80730038 0.78779243 0.64858626 0.60280917 0.59978071 0.52123423 0.47662911 0.45744979 0.39300798 0.3469689
Proportion of Variance 0.4617565 0.1845287 0.1092240 0.04655242 0.04432978 0.03004744 0.02588679 0.02569549 0.01940608 0.01622681 0.01494717 0.01103252 0.0085991
Cumulative Proportion  0.4617565 0.6462852 0.7555091 0.80206157 0.84639135 0.87643878 0.90232557 0.92802107 0.94742715 0.96365395 0.97860112 0.98963364 0.9982327
                          Comp.14
Standard deviation     0.157294837
Proportion of Variance 0.001767262
Cumulative Proportion  1.000000000
```

(a) Figura 1. Resultados de PCA al conjunto de datos del 2017.

En la Figura 1. podemos apreciar los resultados arrojados por PCA con respecto a los datos de 2017. Lo primero que podemos notar es que la mayor parte de la variabilidad de los datos (85 %) se logra en los primeros 6 componentes, esto tiene mucho sentido ya que se esperaría que varias características de la tabla sean las que influyan en la felicidad de las personas. Otra cosa que podemos notar es en relación al primer componente principal, si analizamos los signos de los valores, podremos ver que los números negativos hacen referencia a características que afectarían a la felicidad de forma positiva, tales como: Afecto positivo, soporte social, liberta para tomar decisiones, etc. Y por el otro lado tenemos que los valores positivos aluden a características que afectan de forma negativa a la felicidad tales como: Percepción de corrupción, afecto negativo, etc.
Ahora se muestran los resultados referentes a PCA con los datos de 2016.

```
> loadings(PCA_2)

Loadings:
                                                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
Life.Ladder                                       -0.335        -0.162        -0.165 -0.184  0.237        -0.178                0.115   0.548   0.277
Log.GDP.per.capita                                -0.320        -0.172 -0.204        -0.242  0.106 -0.258  0.115         -0.196 -0.306  0.117  -0.669
Social.support                                    -0.295        -0.177  0.228 -0.130 -0.195 -0.249 -0.371  0.132 -0.139   0.658         -0.296
Healthy.life.expectancy.at.birth                  -0.324  0.109 -0.129 -0.190        -0.142        -0.117  0.182  0.122 -0.409  -0.191  -0.493   0.541
Freedom.to.make.life.choices                      -0.194 -0.413                      -0.386 -0.121  0.341 -0.336 -0.411 -0.226   0.203  -0.292  -0.180
Generosity                                              -0.336  0.291 -0.184 -0.801  0.133        -0.132  0.222 -0.116 -0.127
Perceptions.of.corruption                          0.191  0.341 -0.286  0.153 -0.356        -0.371  0.262 -0.176 -0.125 -0.105  -0.565
Positive.affect                                   -0.176 -0.364 -0.273  0.355 -0.175  0.149  0.123  0.172 -0.159  0.694
Negative.affect                                    0.242        -0.126 -0.530 -0.160 -0.307  0.451  0.306                0.407         -0.183
Confidence.in.national.government                       -0.454  0.368         0.248 -0.355 -0.219         0.252  0.122   0.180  -0.461   0.200   0.205
Democratic.Quality                                -0.275        -0.212 -0.354         0.363 -0.379  0.388                0.219
Delivery.Quality                                  -0.322                -0.352  0.157  0.261 -0.136                              -0.105   0.191
Standard.deviation.of.ladder.by.country.year       0.250 -0.131 -0.419 -0.181        -0.353 -0.312 -0.218         0.179  -0.146   0.342   0.289   0.172
Standard.deviation.Mean.of.ladder.by.country.year  0.337 -0.103 -0.102 -0.190                -0.308 -0.113  0.123  0.197           0.145  -0.228  -0.210
GINI.index..World.Bank.estimate...average.2000.15  0.172 -0.255 -0.485  0.233         0.168  0.292  0.112  0.582 -0.365
gini.of.household.income.reported.in.Gallup..by.wp5.year  0.204 -0.368 -0.181 -0.157  0.102  0.281        -0.469 -0.508 -0.196          -0.331           0.122
                                                  Comp.15 Comp.16
Life.Ladder                                       -0.104   0.542
Log.GDP.per.capita                                -0.240
Social.support                                     0.110
Healthy.life.expectancy.at.birth
Freedom.to.make.life.choices
Generosity
Perceptions.of.corruption
Positive.affect                                    0.110
Negative.affect                                    0.103
Confidence.in.national.government
Democratic.Quality                                -0.509
Delivery.Quality                                   0.771
Standard.deviation.of.ladder.by.country.year      -0.387
Standard.deviation.Mean.of.ladder.by.country.year  0.729
GINI.index..World.Bank.estimate...average.2000.15
gini.of.household.income.reported.in.Gallup..by.wp5.year -0.126

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var  0.063  0.062  0.062  0.062  0.063  0.062  0.062  0.062  0.062   0.063   0.062   0.062   0.062   0.062   0.062   0.063
Cumulative Var  0.063  0.125  0.188  0.250  0.312  0.375  0.438  0.500  0.562   0.625   0.688   0.750   0.812   0.875   0.937   1.000
```

(b) Figura 2. Resultados de PCA al conjunto de datos del 2017.

```
> summary(PCA_2)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
Standard deviation     2.6772381 1.6154399 1.23101089 0.99522361 0.87522224 0.78278893 0.70896840 0.68309313 0.54470580 0.50909304 0.50370236 0.45344664
Proportion of Variance 0.4479752 0.1631029 0.09471174 0.06190438 0.04787587 0.03829741 0.03141476 0.02916351 0.01854403 0.01619848 0.01585725 0.01285087
Cumulative Proportion  0.4479752 0.6110781 0.70578987 0.76769425 0.81557012 0.85386753 0.88528229 0.91444580 0.93298983 0.94918831 0.96504556 0.97789643
                          Comp.13     Comp.14     Comp.15     Comp.16
Standard deviation     0.399153931 0.330372119 0.241245236 0.164280995
Proportion of Variance 0.009957741 0.006821609 0.003637454 0.001686765
Cumulative Proportion  0.987854172 0.994675781 0.998313235 1.000000000
```

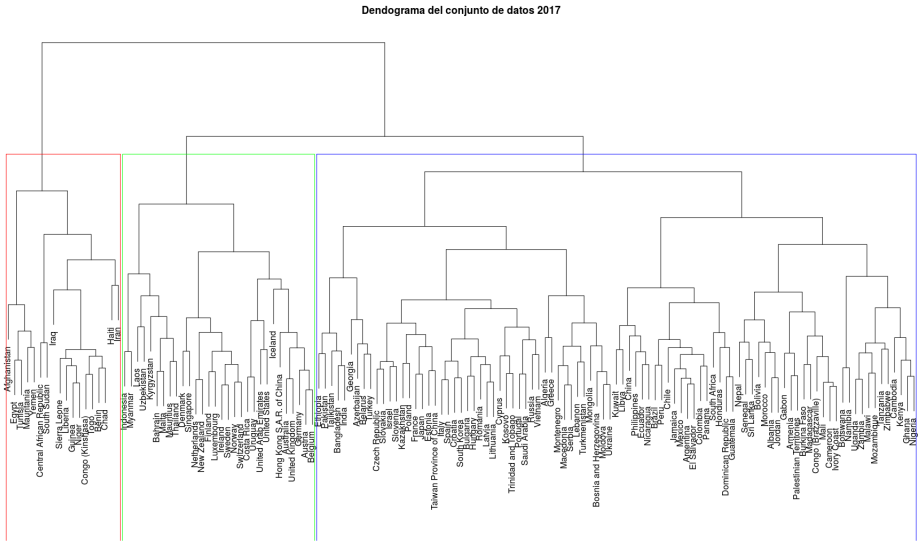
(c) Figura 3. Resultados de PCA al conjunto de datos del 2017.

En las dos Figuras anteriores podemos notar resultados muy parecidos a los de los datos de 2017, notamos que el 85-88 % de variabilidad está siendo aportada por los primeros 6-7 componentes y el contraste con los signos de los valores del primer componente principal. En esta parte hay que notar que los datos de 2016 conservan dos características extra que fueron eliminadas en los datos de 2017 por asusencia de información.

Como se viene mencionando en esta parte, PCA nos dice que gran parte de la variabilidad está siendo aportada por los primeros 6 componentes así que se tomó la decisión de trabajar con las proyecciones de los datos asociadas a estos componentes, por lo que a partir de aquí los métodos que se mencionarán fueron aplicados a esta reducción de la información.

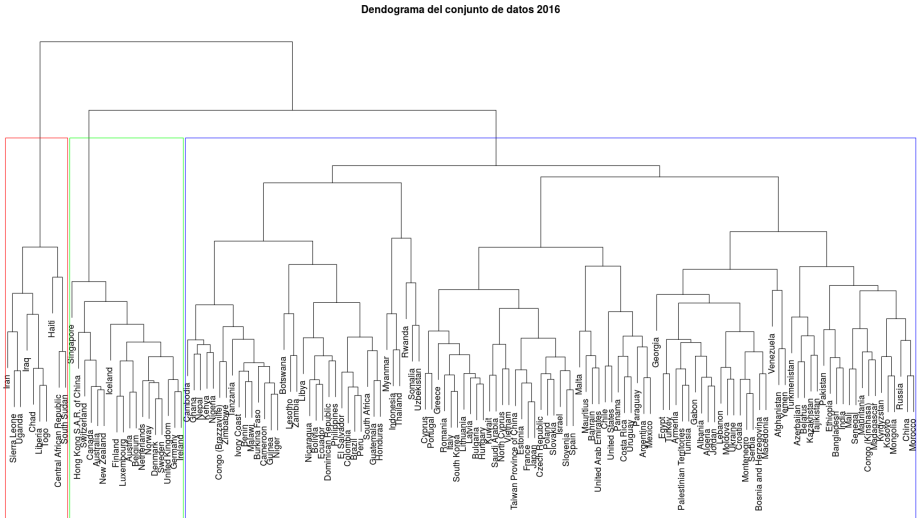
Ahora, la intención es la de buscar grupos de pasíses con ciertas similitudes, y para ello se usaron dos métodos de agrupamiento *duro*, KMeans y hclust, para los cuales necesitamos especificar la cantidad de grupos a formar, una buena idea

sería formar agrupamientos con 2 o 3 clases, ya que se esperaría ver en general si encontramos grupos de países con mucha felicidad, poca felicidad y un tercer grupo nos serviría para denotar alguna característica adicional.
A continuación se muestran los resultados de los métodos KMeans y hclust para los dos conjuntos de datos usando la distnacia euclideana en ambos casos.



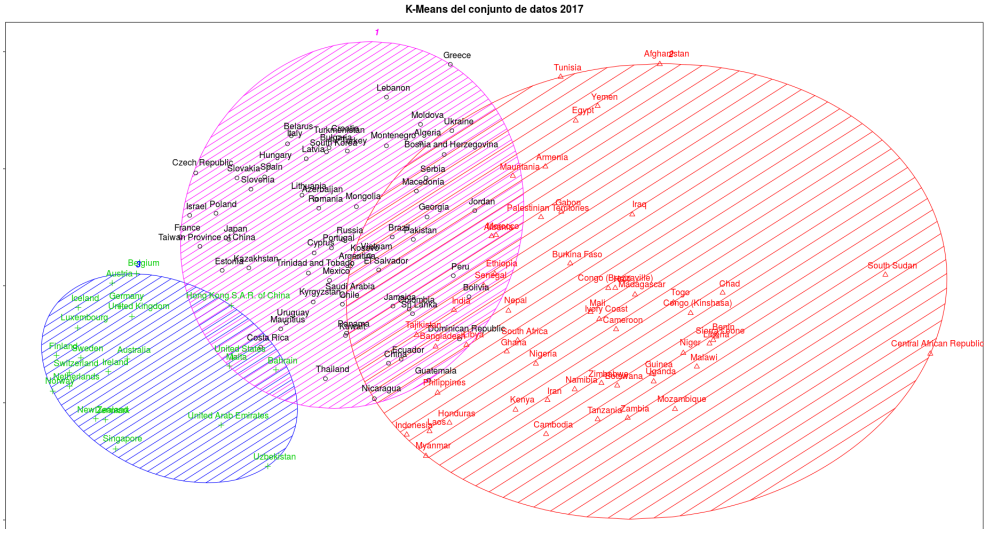
(d) Figura 4. Dendograma del conjunto de datos de 2017.

En la Figura 4. podemos el dendograma generado por el método de hclust con los datos de 2017, dicho dendograma contiene tres grupos en general, si analizamos los países contenidos en dichos grupos podemos notar que países como Estados Unidos, Reino Unido, Los Emiratos Árabes Unidos, Holanda, entre otros se encuentran en uno de los tres grupos. Por otro lado, vemos que países como: Haití, Afghanistan, Iraq, etc. se encuentran en otro grupo, esto es interesante ya que los países del primer grupo mencionado comparten características positivas como buen desarrollo económico, liberad, seguridad social, etc. mientras que en el segundo grupo vemos países considerados como tercermundistas los cuales comparten muchas características negativas. Finalmente en el tercer grupo podemos apreciar el resto de los países.

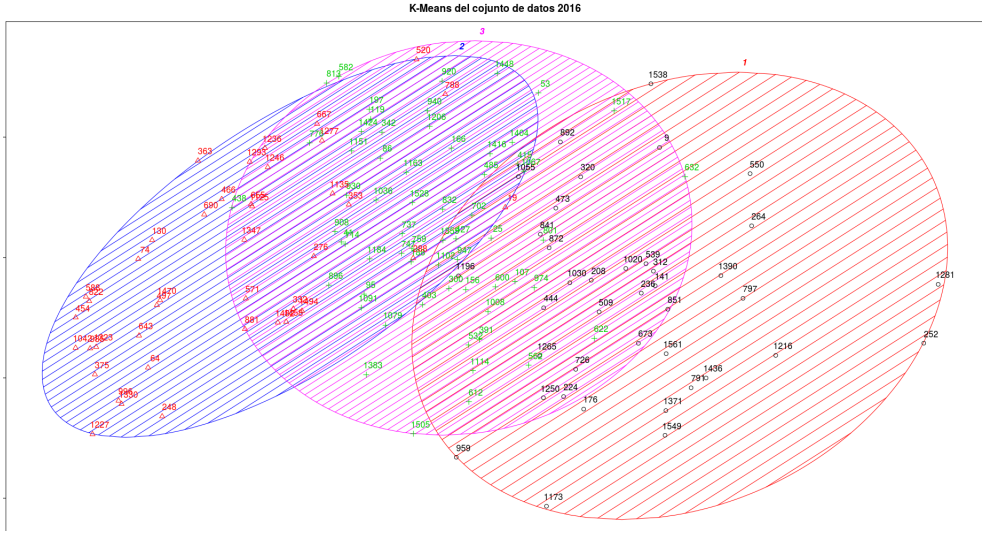


(e) Figura 5. Dendograma del conjunto de datos de 2017.

En la Figura 5. referente al dendograma asociado a los datos del 2016 podemos apreciar una estructura muy similar al dendograma anterior, un grupo con países muy desarrollados, otro con países tercermundistas y un tercero con el resto. Una diferencia clara notada en ambos dendogramas es el hecho que los dos grupos mencionados anteriormente eran más pequeños en el 2016 y en el 2017 contienen más países. Podríamos decir que en conjunto aumentaron en un 10 %.



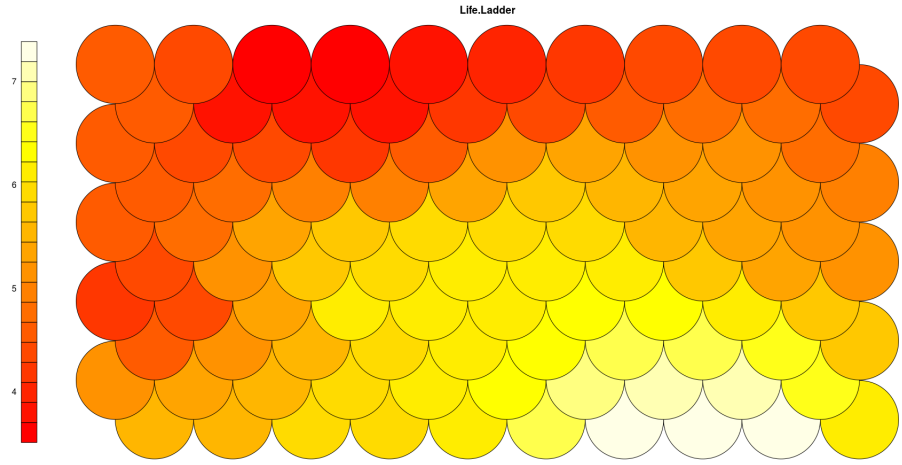
(f) Figura 6. KMeans del conjunto de datos de 2017.



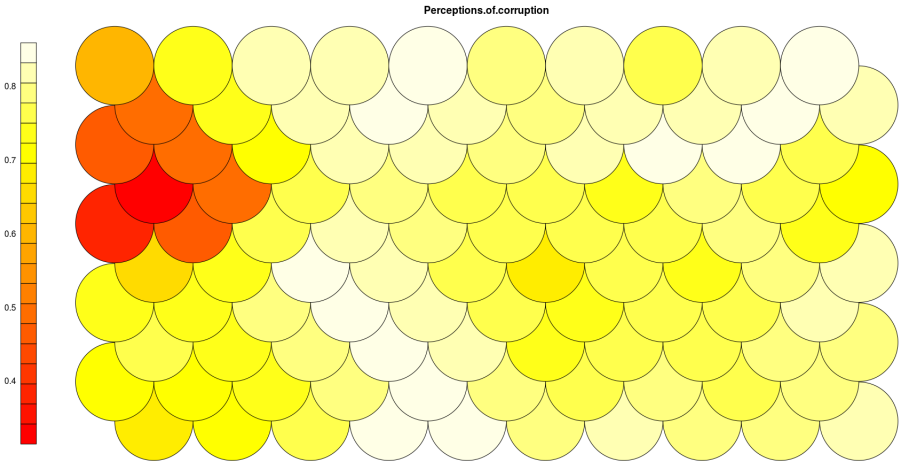
(g) Figura 7. KMeans del conjunto de datos de 2016.

En las Figuras 6 y 7 se muestran los resultados de los KMeans aplicados a los datos, en ellos notamos una estructura un poco similar a la del dendograma, con países asociados por su desarrollo económico y social, y países con características negativas. Hay que notar que al igual que los dendogramas, los conjuntos de países antes mencionados crecieron considerablemente en el transcurso del 2016-2017.
Una característica que se notó al generar los diagramas anteriores es que los resultados se vieron muy afectados por la característica de "Life Ladder", la cual consiste en preguntar a las personas el qué tan felíz creen que son y la respuesta es almacenada en una escala. Aparentemente esta característica influyó en gran medida ya que los países más altos en ella son los que están en el grupo antes comentado de países desarrollados y por el contrario los que están hasta abajo son los países tercermundistas. Esto será importante más adelante.

Otro método de clusterizado aplicado a los datos fue el Self-Organizing Maps (SOM) el cual genera un mapeo de los datos pero intentando preservar cierta estructura denotada por alguna función, en este caso se usó la distancia euclideana. El uso de este método sólo fue con el afán de ver la estructura que generaba el SOM con HeatMaps para una característica positiva y una negativa, cabe destacar que sólo se generaron dos HeatMaps para el conjunto de datos de 2017 los cuales se muestran a continuación.



(h) Figura 8. SOM de Life Ladder con datos 2017.

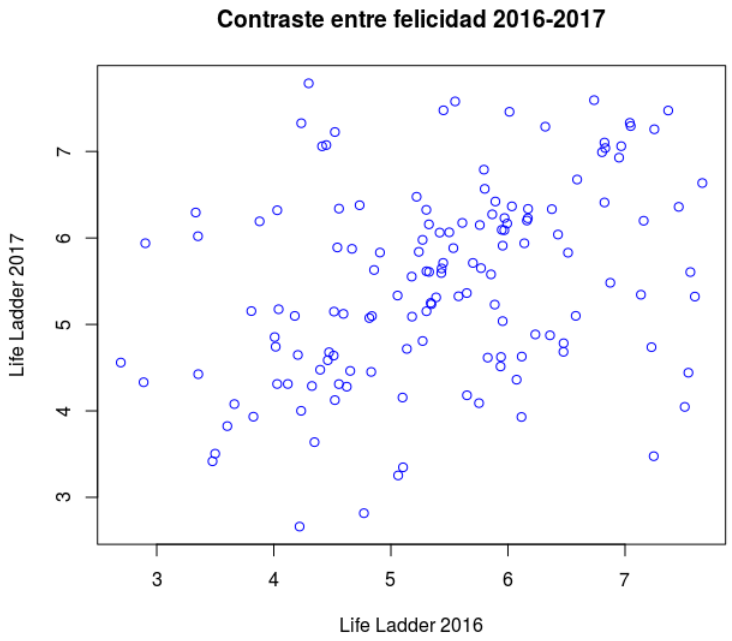


(i) Figura 9. SOM de la percepción de la corrupción con datos 2017.

EN la Figura 8. vemos un HeatMap relacionado a la característica de Life Ladder antes mencionada, se puede notar que los valores más intensos de rojo se relacionan con Life Ladder más pequeños y por el contrario los valores más claros se relacionan con Life Ladder más grandes, podemos notar que para estos dos anteriores sólo un pequeño conjunto de grupos están dentro de ellos.
La misma explicación se hace para la Figura 9 con relación a los colores, sólamente que en ese HeatMap el cual explica la característica de *Percepción de corrupción*

hay que notar que sólo un pequeño conjunto de círculos tiene un color rojo intenso, los demás círculos contienen una gama de amarillos un poco más uniforme, por lo cual podríamos decir que esta característica no cointribuiría mucho a la felicidad de las personas en los determinados países.
En general sólo se obtuvieron estos dos ejemplos ya que pese a que podemos obtener información sobre la estructura de los datos, esto no nos dice mucho acerca de qué países son los que se encuentran dentro de los grupos interesantes.

Finalmente se generó una simple gráfica de puntos en donde un eje corresponde a la característica de Life Ladder del año 2016 y el otro corresponde al Life Ladder del 2017. Esto se hace ya que como se comentó anteriormente, esta característica juega un papel importante en la descripción de la felicidad de los distintos países, y podemos usarla para ver cuánto cambio hubo desde el 2016 al 2017.



(j) Figura 10. Contraste entre Life Ladder de datos 2016-2017.

En la Figura 10. vemos la gráfica de puntos descrita anteriormente, en esa gráfica cada punto representa a un país distinto. Como conclusión queda decir que los puntos que están muy cerca de la recta identidad son los que en general mantuvieron su nivel de felicidad a lo largo de ese año, por otra parte los que están por debajo de ella su felicidad se vió decrementada y los que están por

encima, su felicidad se vió aumentada. Se puede notar que hubo gran variabilidad con relación al 2016-2017, pero al parecer el mismo porcentaje de países que aumentó su felicidad fue el que también se decrementó, por otra parte también vemos a un gran conjunto de países que se mantuvieron a lo largo del año, probablemente estos países fueron los que se comentaron en los dos grupos importantes que arrojó hclust y KMeans.

Como nota final, el código de R usado para generar todas estas gráficas se adjuntará en conjunto con el reporte.