Tipología y ciclo de vida de los datos - Práctica 2

Gabriel Quintairos Rial

11 de junio de 2019

Contents

---

1 Descripción del dataset

El conjunto de datos con el que se trabajará en esta práctica se ha obtenido a través del repositorio Kaggle: https://www.kaggle.com/PromptCloudHQ/world-happiness-report-2019

En concreto, este dataset está basado en el World Happiness Report del año 2019: https://worldhappiness.report/ed/2019/

Como los datos que había en el repostorio de Kaggle habían sido preparados y se había eliminado bastante información, se ha optado por trabajar con el conjunto de datos original, disponible de forma libre aquí: https://s3.amazonaws.com/happiness-report/2019/Chapter2OnlineData.xls

En este informe se miden una serie de indicadores en donde se clasifican 156 países en función de los felices que se sienten sus habitantes. Para calcular estos indicadores se establece un país imaginario llamado "Distopía" en el cual sus habitantes son los menos felices del mundo. Dicho país será la referencia para cada uno de los indicadores, siendo todos sus valores igual a cero en este caso de máxima infelicidad. Por tanto, los indicadores para los países que forman parte del informe serán siempre cero o positivos.

1.1 Variables

El conjunto de datos está constituido por 26 indicadores (columnas) de un total de 156 países y para varios años, teniendo un total de 1704 filas. Los indicadores son los siguientes:

- **Country name:** nombre del país.
- **Year:** año al que pertenece el dato.
- **Life Ladder:** valoración del nivel de felicidad.

- **Log GDP per capita:** logaritmo decimal de la renta media per cápita de cada país.
- **Social support:** medida en la que el apoyo social contribuye a la valoración de la felicidad.
- **Healthy life expectancy at birth:** esperanza de vida al nacimiento (años).
- **Freedom to make live choices:** medida en la que las libertades individuales y colectivas contribuyen a la valoración de la felicidad.
- **Generosity:** medida en la que la generosidad contribuye a la valoración de la felicidad.
- **Perceptions of corruption:** medida en la que la percepción de la corrupción contribuye a la valoración de la felicidad.
- **Positive affect:** medida de las emociones positivas.
- **Negative affect:** medida de las emociones negativas.
- **Confidence in national government:** nivel de confianza en el gobierno nacional.
- **Democratic quality:** valoración de la calidad de la democracia en términos de estabilidad política y ausencia de violencia.
- **Delivery quality:** valor medio de las valoraciones dadas a la eficacia del gobierno nacional, calidad de las leyes existentes en el país, aplicación de las leyes y control de la corrupción.
- **Standard deviation of ladder by country-year:** desviación estándar del nivel de felicidad.
- **Standard deviation/Mean of ladder by country-year:** desviación estándar / media del nivel de felicidad.
- **GINI index (World Bank estimate):** desigualdad de los ingresos dentro del país.
- **GINI index (World Bank estimate), average 2000-16:** media de la desigualdad de los ingresos dentro del país en el periodo 2000-2016.
- **gini of household income reported in Gallup, by wp5-year:** renta media obtenida por hogar
- **Most people can be trusted, Gallup:** cantidad de personas en las que podrías confiar.
- **Most people can be trusted, WVS round 1981-1984:** cantidad de personas en las que podrías confiar (media del periodo 1981-1984).
- **Most people can be trusted, WVS round 1989-1993:** cantidad de personas en las que podrías confiar (media del periodo 1989-1993).
- **Most people can be trusted, WVS round 1994-1998:** cantidad de personas en las que podrías confiar (media del periodo 1994-1998).
- **Most people can be trusted, WVS round 1999-2004:** cantidad de personas en las que podrías confiar (media del periodo 1999-2004).
- **Most people can be trusted, WVS round 2005-2009:** cantidad de personas en las que podrías confiar (media del periodo 2005-2009).
- **Most people can be trusted, WVS round 2010-2014:** cantidad de personas en las que podrías confiar (media del periodo 2010-2014).

1.2 Importancia del dataset y objetivos del análisis

Este conjunto de datos es muy importante pues nos permite medir las diferencias y similitudes existentes entre un gran número de países. Además, también nos permitirá ver una evolución temporal dentro de algunos de los indicadores al disponer de datos de años anteriores.

El World Happiness Report es un estudio realizado anualmente por la ONU desde el 2012 y que tiene un gran valor para muchas organizaciones, ya que permite detectar carencias y puntos a mejorar en varios lugares del mundo.

---

2 Integración y selección de los datos de interés

2.1 Integración

No ha sido necesario realizar un proceso de integración en esta práctica. Esto es debido a que a la hora de seleccionar el conjunto de datos, se optó por utilizar un dataset más completo en lugar del repositorio de Kaggle. Haciendo esto conseguimos el mismo resultado que haciendo una integración vertical, ya que conseguimos tener un mayor número de registros y de variables dentro la información que será objeto de análisis.

2.2 Selección de los datos de interés

Para proceder a la selección de los datos de interés, cargaremos el fichero Excel que contiene los datos en un data table de R:

```
library(xlsx)
data <- read.xlsx("C:/Users/user/Documents/Máster Big Data/Tipología y ciclo de vida de los datos/Práct
```

Descartaremos todas las variables correspondientes a índices de GINI. En primer lugar porque varias de ellas son nulas en un gran número de países, siendo por tanto poco útiles en un estudio. En segundo lugar porque otras de ellas son datos calculados (medias de años anteriores) que no nos aportan nada, ya que durante el propio análisis podríamos ser capaces de realizar dichos cálculos si fuesen necesarios. Por tanto, reduciremos la dimensionalidad de los datos, lo cual simplifica el dataset y mejora el rendimiento.

```
data <- data[,-17:-26]
```

Ahora veremos cómo trata R los datos que tenemos en nuestro dataset, por si fuese necesario realizar algún ajuste:

```
str(data)
```

```
## 'data.frame':    1704 obs. of  16 variables:
##  $ Country.name                                : chr  "Afghanistan" "Afghanistan" "Afghanistan"
##  $ Year                                        : num  2008 2009 2010 2011 2012 ...
##  $ Life.Ladder                                 : num  3.72 4.4 4.76 3.83 3.78 ...
##  $ Log.GDP.per.capita                          : num  7.17 7.33 7.39 7.42 7.52 ...
##  $ Social.support                              : num  0.451 0.552 0.539 0.521 0.521 ...
##  $ Healthy.life.expectancy.at.birth            : num  50.8 51.2 51.6 51.9 52.2 ...
##  $ Freedom.to.make.life.choices                : num  0.718 0.679 0.6 0.496 0.531 ...
##  $ Generosity                                  : num  0.178 0.2 0.134 0.172 0.244 ...
##  $ Perceptions.of.corruption                   : num  0.882 0.85 0.707 0.731 0.776 ...
##  $ Positive.affect                             : num  0.518 0.584 0.618 0.611 0.71 ...
##  $ Negative.affect                             : num  0.258 0.237 0.275 0.267 0.268 ...
##  $ Confidence.in.national.government           : num  0.612 0.612 0.299 0.307 0.435 ...
##  $ Democratic.Quality                          : num  -1.93 -2.04 -1.99 -1.92 -1.84 ...
##  $ Delivery.Quality                            : num  -1.66 -1.64 -1.62 -1.62 -1.4 ...
```

```
##  $ Standard.deviation.of.ladder.by.country.year     : num  1.77 1.72 1.88 1.79 1.8 ...
##  $ Standard.deviation.Mean.of.ladder.by.country.year: num  0.477 0.391 0.395 0.466 0.475 ...
```

Podemos observar que los tipos de datos son los correctos. Gracias a haber utilizado el parámetro "stringsAsFactors" en la carga del fichero Excel, el nombre de los países se ha cargado correctamente. Por tanto no será necesario realizar ninguna transformación más en esta parte.

---

3 Limpieza de los datos

3.1 Ceros y elementos vacíos

Comenzaremos viendo cuántos ceros hay por cada atributo. En caso de haber alguno, deberíamos analizarlo ya que no necesariamente deberá ser un error. Esto es debido a que la puntuación dada a alguno de los indicadores podría correctamente corresponderse con el valor cero.

```
sapply(data, function(x) sum(x==0, na.rm = TRUE))
```

```
##                                      Country.name
##                                                 0
##                                              Year
##                                                 0
##                                       Life.Ladder
##                                                 0
##                                  Log.GDP.per.capita
##                                                 0
##                                     Social.support
##                                                 0
##                     Healthy.life.expectancy.at.birth
##                                                 0
##                         Freedom.to.make.life.choices
##                                                 0
##                                        Generosity
##                                                 0
##                            Perceptions.of.corruption
##                                                 0
##                                    Positive.affect
##                                                 0
##                                    Negative.affect
##                                                 0
##                   Confidence.in.national.government
##                                                 0
##                                  Democratic.Quality
##                                                 0
##                                    Delivery.Quality
##                                                 0
##      Standard.deviation.of.ladder.by.country.year
##                                                 0
## Standard.deviation.Mean.of.ladder.by.country.year
##                                                 0
```

Como podemos ver, no hay ningún cero en ninguna de nuestras variables. Por tanto, ahora veremos cuántos valores nulos hay para cada indicador:

```
sapply(data, function(x) sum(is.na(x)))
```

```
##                                       Country.name
##                                                  0
##                                               Year
##                                                  0
##                                        Life.Ladder
##                                                  0
##                                   Log.GDP.per.capita
##                                                 28
##                                     Social.support
##                                                 13
##                     Healthy.life.expectancy.at.birth
##                                                 28
##                         Freedom.to.make.life.choices
##                                                 29
##                                         Generosity
##                                                 82
##                            Perceptions.of.corruption
##                                                 96
##                                      Positive.affect
##                                                 19
##                                      Negative.affect
##                                                 13
##                    Confidence.in.national.government
##                                                174
##                                   Democratic.Quality
##                                                146
##                                     Delivery.Quality
##                                                145
##      Standard.deviation.of.ladder.by.country.year
##                                                  0
## Standard.deviation.Mean.of.ladder.by.country.year
##                                                  0
```

Vemos en el resultado obtenido que son numerosos los campos que tienen valores nulos o 'NA'. En concreto para alguna variable como la confianza en el gobierno nacional, dicha cantidad de valores supera el 10% de los registros. Este gran número de valores nulos podrá causar problemas a la hora de aplicar algoritmos de análisis de datos o de intentar representar la información. Una posible solución podría ser eliminar aquellos registros que contienen indicadores nulos pero, de esta forma, estaríamos perdiendo mucha información.

Para resolverlo, utilizaremos la técnica de los k vecinos más próximos (kNN-imputation). Esta técnica está basada en la similitud o la diferencia entre los registros y trata de asignar un valor aproximado a aquellos valores nulos. Se utiliza bajo la hipótesis de que nuestros registros guardan una relación entre ellos, lo cual es realmente cierto ya que cada registro representa una evolución temporal de varios países del mundo y se espera que tengan algunas similitudes. Como es mejor trabajar con valores aproximados que con datos nulos, realizamos la aproximación:

```
library(VIM)
data$Log.GDP.per.capita <- kNN(data)$Log.GDP.per.capita
data$Social.support <- kNN(data)$Social.support
data$Healthy.life.expectancy.at.birth <- kNN(data)$Healthy.life.expectancy.at.birth
data$Freedom.to.make.life.choices <- kNN(data)$Freedom.to.make.life.choices
```

```
data$Generosity <- kNN(data)$Generosity
data$Perceptions.of.corruption <- kNN(data)$Perceptions.of.corruption
data$Positive.affect <- kNN(data)$Positive.affect
data$Negative.affect <- kNN(data)$Negative.affect
data$Confidence.in.national.government <- kNN(data)$Confidence.in.national.government
data$Democratic.Quality <- kNN(data)$Democratic.Quality
data$Delivery.Quality <- kNN(data)$Delivery.Quality
```

Volvemos de nuevo a ver cuántos valores nulos hay:

```
sapply(data, function(x) sum(is.na(x)))
```

```
##                                       Country.name
##                                                  0
##                                               Year
##                                                  0
##                                        Life.Ladder
##                                                  0
##                                   Log.GDP.per.capita
##                                                  0
##                                      Social.support
##                                                  0
##                     Healthy.life.expectancy.at.birth
##                                                  0
##                         Freedom.to.make.life.choices
##                                                  0
##                                         Generosity
##                                                  0
##                            Perceptions.of.corruption
##                                                  0
##                                    Positive.affect
##                                                  0
##                                    Negative.affect
##                                                  0
##                   Confidence.in.national.government
##                                                  0
##                                   Democratic.Quality
##                                                  0
##                                     Delivery.Quality
##                                                  0
##      Standard.deviation.of.ladder.by.country.year
##                                                  0
## Standard.deviation.Mean.of.ladder.by.country.year
##                                                  0
```

Y efectivamente podemos comprobar que ya no hay ninguno.

3.2 Valores extremos

Los valores extremos o outliers son aquellos que se alejan mucho de la distribución normal de una variable. Estos valores pueden ser correctos en países que estén en unas condiciones extremas, pero también podrían estar producidos por errores a la hora de recoger la muestra o de rellenar el dataset. Para poder detectarlos en nuestro conjunto de datos se podrían tomar dos caminos. El primero de ellos consistiría en hacer representaciones gráficas como histogramas o diagramas de cajas de las variables. El segundo camino, que

será el que tomaremos, será utilizar la función boxplot.stats de R que nos devolverá los outliers de cada variable:

```
boxplot.stats(data$Life.Ladder)$out
```

```
## numeric(0)
```

```
boxplot.stats(data$Log.GDP.per.capita)$out
```

```
## numeric(0)
```

```
boxplot.stats(data$Social.support)$out
```

```
##  [1] 0.4506623 0.4835519 0.4908801 0.5075158 0.4665535 0.4447812 0.3823735
##  [8] 0.4774944 0.5060913 0.4343885 0.4928159 0.4358790 0.5035440 0.2909338
## [15] 0.3256925 0.4222400 0.4847152 0.4833339 0.3873909 0.2901842 0.3195891
## [22] 0.4789509 0.5029374 0.5105746 0.5116161 0.4943816 0.4788874 0.3729079
## [29] 0.5098841 0.4856810 0.4639129 0.4354136 0.2913337 0.3029551 0.4443390
## [36] 0.4785934 0.5094410 0.5078052
```

```
boxplot.stats(data$Healthy.life.expectancy.at.birth)$out
```

```
##  [1] 40.90 42.70 43.08 43.18 43.66 40.38 32.30 36.86 41.42 40.30 41.20
## [12] 42.10 41.58 42.86
```

```
boxplot.stats(data$Freedom.to.make.life.choices)$out
```

```
## [1] 0.2575338 0.2600693 0.3061319 0.2946118 0.2814579 0.3035404 0.3145646
## [8] 0.2868144
```

```
boxplot.stats(data$Generosity)$out
```

```
##  [1] 0.4174387 0.4209864 0.4363546 0.4266045 0.4592676 0.4875419 0.4754607
##  [8] 0.4993778 0.4422074 0.4808939 0.4188714 0.4604585 0.4577791 0.4041515
## [15] 0.6249432 0.6691009 0.6777426 0.6671360 0.6589326 0.6295767 0.4699646
## [22] 0.4190406 0.5185034 0.5293711 0.4493156 0.5455707
```

```
boxplot.stats(data$Perceptions.of.corruption)$out
```

```
##   [1] 0.39041594 0.36612740 0.38177174 0.36825174 0.35655439 0.39854512
##   [7] 0.41134652 0.40464750 0.40560842 0.36958781 0.41262212 0.41265959
##  [13] 0.40623614 0.38509044 0.36203432 0.37174085 0.23652171 0.20600568
##  [19] 0.24750531 0.20576976 0.17489609 0.22004308 0.18740761 0.17004217
##  [25] 0.23721834 0.19101639 0.20989338 0.18114756 0.15060744 0.13243018
##  [31] 0.21656753 0.41251570 0.31959319 0.36073396 0.30577046 0.26547989
##  [37] 0.22336966 0.24965957 0.19241278 0.19860484 0.35334641 0.32088760
##  [43] 0.34871361 0.41552565 0.41216829 0.41402119 0.35598481 0.27394506
##  [49] 0.27212471 0.25577542 0.24488659 0.37978315 0.40281257 0.41581020
##  [55] 0.40603626 0.40875691 0.39854431 0.33708474 0.36221024 0.32815811
##  [61] 0.38817060 0.40275320 0.30081177 0.36628678 0.37539047 0.35633633
##  [67] 0.33017358 0.38514611 0.39859185 0.35939589 0.41182211 0.36313364
##  [73] 0.37055779 0.22422023 0.29461622 0.33375087 0.32074818 0.26933020
##  [79] 0.28929794 0.31223580 0.27260861 0.18588871 0.27827078 0.22188748
##  [85] 0.20658022 0.39715013 0.36804268 0.40482584 0.29881436 0.40966612
##  [91] 0.24971138 0.26820144 0.18379813 0.35511589 0.35511589 0.29864353
##  [97] 0.28640723 0.40970287 0.16147466 0.08132490 0.11716541 0.07800018
## [103] 0.09460447 0.15860139 0.21375722 0.16380996 0.06361488 0.06577528
## [109] 0.03519799 0.06028207 0.09892445 0.24239805 0.13260315 0.09894388
## [115] 0.04731115 0.16179068 0.09656293 0.41023576 0.35734090 0.33383173
```

```
## [121] 0.39845687 0.28933215 0.31396121 0.29211217 0.25308666 0.26851302
## [127] 0.25354311 0.32448155 0.25038999 0.23196414 0.24618244 0.23936692
## [133] 0.26279658 0.40793142 0.34242702 0.32324079 0.28308958 0.20953351
## [139] 0.30156296 0.31618348 0.30125996 0.20335877 0.33887646 0.35511589
## [145] 0.35511589 0.35511589 0.33887646 0.39845687 0.41861135 0.40427601
```

```
boxplot.stats(data$Positive.affect)$out
```

```
## numeric(0)
```

```
boxplot.stats(data$Negative.affect)$out
```

```
##  [1] 0.5137192 0.4819340 0.4942681 0.5993355 0.5382454 0.5438362 0.4833790
##  [8] 0.4821832 0.5249687 0.5518397 0.5115691 0.5198582 0.5258768 0.4931489
## [15] 0.5570987 0.5542787 0.5636311 0.5812669 0.5697581 0.5905387 0.5090467
## [22] 0.5025545 0.4950400 0.5492569 0.5173638 0.4955055 0.7045897 0.6222299
## [29] 0.6425887 0.4828859
```

```
boxplot.stats(data$Confidence.in.national.government)$out
```

```
## numeric(0)
```

```
boxplot.stats(data$Democratic.Quality)$out
```

```
## numeric(0)
```

```
boxplot.stats(data$Delivery.Quality)$out
```

```
## numeric(0)
```

Podemos ver que hay bastantes variables que tienen valores extremos, como por ejemplo la esperanza de vida, la percepción de la corrupción o las emociones negativas. No obstante, estos valores en su apariencia podrían ser perfectamente reales, ya que proceden de países subdesarrollados donde la esperanza de vida es muy baja, de países bajo dictaduras o gobiernos que no tienen la confianza de sus ciudadanos o donde hay guerras o enfrentamientos violentos. Por ello, la decisión con respecto a estos valores es dejarlos tal y como están, finalizando de esta forma la parte de limpieza de los datos.

---

4 Análisis de los datos
