

Tipología y ciclo de vida de los datos - Práctica 2

Gabriel Quintairos Rial

11 de junio de 2019

Contents

1 Descripción del dataset	1
1.1 Variables	2
1.2 Importancia del dataset y objetivos del análisis	3
2 Integración y selección de los datos de interés	3
2.1 Integración	3
2.2 Selección de los datos de interés	3
3 Limpieza de los datos	4
3.1 Ceros y elementos vacíos	4
3.2 Valores extremos	7
3.3 Exportación de los datos	8
4 Análisis de los datos	9
4.1 Análisis descriptivo	9
4.2 Análisis inferencial	10
4.2.1 Selección de los grupos de datos que se quieren analizar	10
4.2.2 Comprobación de la normalidad y homogeneidad de la varianza	11
4.2.3 Pruebas estadísticas	12
5 Representación de los resultados a partir de tablas y gráficas	16
6 Conclusiones	19
7 Contribuciones	20

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 repositorio 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:

```
apply(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.

3.3 Exportación de los datos

Con la limpieza de datos que se ha realizado, crearemos un nuevo fichero Excel donde se recoja el nuevo dataset creado. Estará en “data/WorldHappinessReport_trasLimpieza.xls”.

```
write.xlsx(data, "C:/Users/user/Documents/Máster Big Data/Tipología y ciclo de vida de los datos/Práctica 3/WorldHappinessReport_trasLimpieza.xls")
```


4 Análisis de los datos

El objetivo del análisis de datos será tratar de obtener conocimiento a partir de la información que tenemos y tratar de responder a posibles preguntas planteadas.

4.1 Análisis descriptivo

El análisis descriptivo comprende el cálculo de valores a través de la propia muestra. Dentro de este análisis se recogen las medidas de tendencia central y las medidas de dispersión. Para ver un resumen del análisis estadístico, utilizaremos la función `summary` de R:

```
summary(data)
```

```
## Country.name      Year      Life.Ladder      Log.GDP.per.capita
## Length:1704      Min.      :2005      Min.      :2.662      Min.      : 6.457
## Class :character  1st Qu.:2009      1st Qu.:4.611      1st Qu.: 8.306
## Mode  :character  Median :2012      Median :5.340      Median : 9.400
##                      Mean  :2012      Mean  :5.437      Mean  : 9.222
##                      3rd Qu.:2015      3rd Qu.:6.274      3rd Qu.:10.195
##                      Max.   :2018      Max.   :8.019      Max.   :11.770
## Social.support    Healthy.life.expectancy.at.birth
## Min.      :0.2902      Min.      :32.30
## 1st Qu.:0.7477      1st Qu.:58.50
## Median :0.8331      Median :65.03
## Mean  :0.8107      Mean  :63.17
## 3rd Qu.:0.9043      3rd Qu.:68.39
## Max.   :0.9873      Max.   :76.80
## Freedom.to.make.life.choices    Generosity
## Min.      :0.2575              Min.      : -0.3363845
## 1st Qu.:0.6399              1st Qu.: -0.1131292
## Median :0.7534              Median : -0.0247297
## Mean  :0.7346              Mean  : -0.0006653
## 3rd Qu.:0.8480              3rd Qu.: 0.0920538
## Max.   :0.9852              Max.   : 0.6777426
## Perceptions.of.corruption    Positive.affect    Negative.affect
## Min.      :0.0352              Min.      :0.3625      Min.      :0.08343
## 1st Qu.:0.6901              1st Qu.:0.6224      1st Qu.:0.20583
## Median :0.7995              Median :0.7179      Median :0.25532
## Mean  :0.7472              Mean  :0.7091      Mean  :0.26590
## 3rd Qu.:0.8708              3rd Qu.:0.8001      3rd Qu.:0.31422
## Max.   :0.9833              Max.   :0.9436      Max.   :0.70459
## Confidence.in.national.government    Democratic.Quality    Delivery.Quality
## Min.      :0.06877              Min.      : -2.4482      Min.      : -2.14497
## 1st Qu.:0.33665              1st Qu.: -0.7437      1st Qu.: -0.69957
## Median :0.46150              Median : -0.1935      Median : -0.20325
## Mean  :0.47961              Mean  : -0.1114      Mean  : 0.01235
## 3rd Qu.:0.60772              3rd Qu.: 0.6753      3rd Qu.: 0.72410
## Max.   :0.99360              Max.   : 1.5750      Max.   : 2.18472
## Standard.deviation.of.ladder.by.country.year
## Min.      :0.863
## 1st Qu.:1.743
## Median :1.973
```

```
## Mean :2.027
## 3rd Qu.:2.242
## Max. :3.719
## Standard.deviation.Mean.of.ladder.by.country.year
## Min. :0.1339
## 1st Qu.:0.3101
## Median :0.3727
## Mean :0.3921
## 3rd Qu.:0.4563
## Max. :1.0228
```

4.2 Análisis inferencial

El análisis inferencial pretende modelar los datos. Con ello se podrán realizar inferencias, predicciones, clasificaciones, agrupaciones, etc.

4.2.1 Selección de los grupos de datos que se quieren analizar

En primer lugar seleccionaremos los grupos de nuestro dataset que podría ser interesante analizar o comparar:

```
# Datos por año
data2005 <- data[data$Year == "2005",]
data2006 <- data[data$Year == "2006",]
data2007 <- data[data$Year == "2007",]
data2008 <- data[data$Year == "2008",]
data2009 <- data[data$Year == "2009",]
data2010 <- data[data$Year == "2010",]
data2011 <- data[data$Year == "2011",]
data2012 <- data[data$Year == "2012",]
data2013 <- data[data$Year == "2013",]
data2014 <- data[data$Year == "2014",]
data2015 <- data[data$Year == "2015",]
data2016 <- data[data$Year == "2016",]
data2017 <- data[data$Year == "2017",]
data2018 <- data[data$Year == "2018",]

# Países por renta per capita
paísesRicos <- data[data$Log.GDP.per.capita >= mean(data$Log.GDP.per.capita),]
paísesPobres <- data[data$Log.GDP.per.capita < mean(data$Log.GDP.per.capita),]

# Países por esperanza de vida
paísesEVAalta <- data[data$Healthy.life.expectancy.at.birth >= mean(data$Healthy.life.expectancy.at.birth),]
paísesEVBaja <- data[data$Healthy.life.expectancy.at.birth < mean(data$Healthy.life.expectancy.at.birth),]
```

Realizamos una primera agrupación por año, ya que podría ser interesante para ver una evolución histórica. Al hacerla, vemos que para los años 2005 y 2006 hay muy pocos datos, así que no se tendrán en cuenta.

La segunda agrupación la realizamos por renta per cápita, agrupando a los países con una renta por encima de la media por una parte y a los que la tienen por debajo de la media en otra. Podemos ver que hay más “países ricos” que “países pobres”.

La última agrupación se realiza por esperanza de vida, agrupando al igual que en el caso anterior. Por una parte estarán los países con una esperanza de vida superior a la media y por otra los que la tienen inferior a la media. Vemos que hay una mayoría de países con esperanza de vida superior a la media.

4.2.2 Comprobación de la normalidad y homogeneidad de la varianza

Para realizar la comprobación de la normalidad utilizaremos el Shapiro-Test, por ser uno de los métodos más robustos que permiten hacerlo.

Este test asume como hipótesis nula que la población está normalmente distribuida. Por tanto, si el p-valor es menor que un α dado, en nuestro caso $\alpha = 0.05$, se rechazará la hipótesis nula y se concluirá que los resultados no siguen una distribución normal.

```
##Comprobacion de la normalidad y homogeneidad de la varianza
alpha = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
    p_val = shapiro.test(data[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(data) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## Year, Life.Ladder,
## Log.GDP.per.capita, Social.support, Healthy.life.expectancy.at.birth,
## Freedom.to.make.life.choices, Generosity, Perceptions.of.corruption,
## Positive.affect, Negative.affect, Confidence.in.national.government,
## Democratic.Quality, Delivery.Quality, Standard.deviation.of.ladder.by.country.year
## Standard.deviation.Mean.of.ladder.by.country.year
```

A la vista de la aplicación del test a las variables numéricas, podemos ver que se rechaza en todas la hipótesis nula y se concluye que los resultados no siguen una distribución normal. Una posición conservadora nos obligaría a rechazar la normalidad de los datos. No obstante, aplicando el Teorema Central del Límite y al tener un número de registros lo suficientemente grande, podremos asumir la normalidad.

Para realizar la comprobación de la homogeneidad de la varianza, también llamada homocedasticidad, utilizaremos el Test de Fligner-Killeen. En este test la hipótesis nula se corresponde con que ambas varianzas son iguales.

```
##Comprobacion de la homocedasticidad
fligner.test(x = list(paísesRicos$Life.Ladder,paísesPobres$Life.Ladder))

##
## Fligner-Killeen test of homogeneity of variances
##
## data: list(paísesRicos$Life.Ladder, paísesPobres$Life.Ladder)
## Fligner-Killeen:med chi-squared = 43.682, df = 1, p-value =
## 3.863e-11

fligner.test(x = list(paísesEVALta$Life.Ladder,paísesEVBaja$Life.Ladder))

##
## Fligner-Killeen test of homogeneity of variances
##
## data: list(paísesEVALta$Life.Ladder, paísesEVBaja$Life.Ladder)
```

```
## Fligner-Killeen:med chi-squared = 58.634, df = 1, p-value =  
## 1.899e-14
```

A la vista de que ambos p-valores son bajísimos, rechazaremos la hipótesis nula y no podremos considerar que exista homocedasticidad.

4.2.3 Pruebas estadísticas

4.2.3.1 ¿El dinero da la felicidad?

La primera prueba que vamos a realizar será un contraste de hipótesis entre los grupos correspondientes a los países con rentas per cápita por encima de la media y a aquellos que tienen las rentas per cápita por debajo de la media.

Dado que no se pudo comprobar la homocedasticidad, utilizaremos para dicho contraste el test paramétrico de Mann-Whitney. En este test la hipótesis nula asume que las distribuciones de ambos grupos de datos son las mismas. Por lo tanto, solo si el p-valor es menor al nivel $\alpha = 0.05$ se rechazará la hipótesis nula y se concluirá que existen diferencias estadísticamente significativas entre el nivel de felicidad en función de la renta per cápita.

```
##Contraste de hipotesis entre rentas altas y bajas para ver si siguen la misma distribución  
wilcox.test(paísesRicos$Life.Ladder, paísesPobres$Life.Ladder)
```

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data: paísesRicos$Life.Ladder and paísesPobres$Life.Ladder  
## W = 625650, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0
```

Dado que el p-valor es menor al nivel de significación, se rechaza la hipótesis nula y se puede concluir que existen diferencias entre la felicidad en los países con una renta per cápita por encima de la media y aquellos que están por debajo.

```
mean(paísesRicos$Life.Ladder)
```

```
## [1] 6.073536
```

```
mean(paísesPobres$Life.Ladder)
```

```
## [1] 4.659725
```

Observando ambas medias, se puede ver que el dinero sí da la felicidad (al menos estadísticamente hablando).

4.2.3.2 Predicción del nivel de felicidad

Un objetivo muy interesante sería poder predecir el nivel de felicidad a partir de los indicadores recogidos en este estudio. Para ello utilizaremos un modelo de regresión lineal que nos permita realizar predicciones sobre esta variable.

En este modelo incluiremos como variable dependiente la felicidad y como variables independientes todas las demás (excepto el año, el nombre del país y las desviaciones, ya que poco importan). Posteriormente tendremos en cuenta el AIC o Criterio de Información de Akaike para escoger el mejor modelo. El mejor modelo será aquel que tenga un valor más bajo del AIC, ya que tendrá las variables óptimas para explicar la variable dependiente sin que sobre ni falte información. Para ello utilizaremos la función `step` de R, la cual irá eliminando variables hasta quedarnos con el modelo que tenga el mejor AIC.

##regresión lineal para predecir el nivel de felicidad

```
modelo <- lm(Life.Ladder ~ . - Year - Country.name - Standard.deviation.of.ladder.by.country.year - Standard.deviation.Mean.of.ladder.by.country.year)
step(modelo)
```

Start: AIC=-2042.72

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

	Df	Sum of Sq	RSS	AIC
## - Negative.affect	1	0.000	506.67	-2044.7
## - Democratic.Quality	1	0.017	506.69	-2044.7
## - Delivery.Quality	1	0.142	506.82	-2044.2
## <none>			506.67	-2042.7
## - Generosity	1	6.526	513.20	-2022.9
## - Freedom.to.make.life.choices	1	6.862	513.54	-2021.8
## - Healthy.life.expectancy.at.birth	1	12.206	518.88	-2004.2
## - Confidence.in.national.government	1	13.212	519.89	-2000.9
## - Perceptions.of.corruption	1	16.722	523.40	-1989.4
## - Social.support	1	31.909	538.58	-1940.7
## - Positive.affect	1	39.797	546.47	-1915.9
## - Log.GDP.per.capita	1	54.003	560.68	-1872.2

Step: AIC=-2044.72

```
## Life.Ladder ~ Log.GDP.per.capita + Social.support + Healthy.life.expectancy.at.birth +
##   Freedom.to.make.life.choices + Generosity + Perceptions.of.corruption +
##   Positive.affect + Confidence.in.national.government + Democratic.Quality +
##   Delivery.Quality
```

	Df	Sum of Sq	RSS	AIC
## - Democratic.Quality	1	0.017	506.69	-2046.7
## - Delivery.Quality	1	0.143	506.82	-2046.2
## <none>			506.67	-2044.7
## - Generosity	1	6.597	513.27	-2024.7
## - Freedom.to.make.life.choices	1	6.927	513.60	-2023.6
## - Healthy.life.expectancy.at.birth	1	12.368	519.04	-2005.6
## - Confidence.in.national.government	1	13.423	520.10	-2002.2
## - Perceptions.of.corruption	1	16.855	523.53	-1991.0
## - Social.support	1	35.238	541.91	-1932.2
## - Positive.affect	1	42.010	548.68	-1911.0
## - Log.GDP.per.capita	1	54.704	561.38	-1872.0

Step: AIC=-2046.66

```
## Life.Ladder ~ Log.GDP.per.capita + Social.support + Healthy.life.expectancy.at.birth +
##   Freedom.to.make.life.choices + Generosity + Perceptions.of.corruption +
##   Positive.affect + Confidence.in.national.government + Delivery.Quality
```

	Df	Sum of Sq	RSS	AIC
## - Delivery.Quality	1	0.548	507.24	-2046.8

```
## <none> 506.69 -2046.7
## - Generosity 1 6.631 513.32 -2026.5
## - Freedom.to.make.life.choices 1 7.197 513.89 -2024.6
## - Healthy.life.expectancy.at.birth 1 12.354 519.05 -2007.6
## - Confidence.in.national.government 1 13.675 520.37 -2003.3
## - Perceptions.of.corruption 1 17.823 524.51 -1989.8
## - Social.support 1 35.964 542.66 -1931.8
## - Positive.affect 1 42.734 549.43 -1910.7
## - Log.GDP.per.capita 1 55.773 562.46 -1870.7
##
## Step: AIC=-2046.82
## Life.Ladder ~ Log.GDP.per.capita + Social.support + Healthy.life.expectancy.at.birth +
## Freedom.to.make.life.choices + Generosity + Perceptions.of.corruption +
## Positive.affect + Confidence.in.national.government
##
## Df Sum of Sq RSS AIC
## <none> 507.24 -2046.8
## - Generosity 1 7.515 514.75 -2023.8
## - Freedom.to.make.life.choices 1 8.038 515.28 -2022.0
## - Healthy.life.expectancy.at.birth 1 13.789 521.03 -2003.1
## - Confidence.in.national.government 1 14.778 522.02 -1999.9
## - Perceptions.of.corruption 1 22.058 529.30 -1976.3
## - Social.support 1 36.096 543.34 -1931.7
## - Positive.affect 1 42.519 549.76 -1911.7
## - Log.GDP.per.capita 1 71.357 578.60 -1824.5
##
## Call:
## lm(formula = Life.Ladder ~ Log.GDP.per.capita + Social.support +
## Healthy.life.expectancy.at.birth + Freedom.to.make.life.choices +
## Generosity + Perceptions.of.corruption + Positive.affect +
## Confidence.in.national.government, data = data)
##
## Coefficients:
## (Intercept) -1.65766
## Log.GDP.per.capita 0.34806
## Social.support 1.80022
## Healthy.life.expectancy.at.birth 0.02213
## Freedom.to.make.life.choices 0.73033
## Generosity 0.46472
## Perceptions.of.corruption -0.80960
## Positive.affect 1.99086
## Confidence.in.national.government -0.65774
```

Vemos que el mejor modelo se obtiene quitando las variables Negative.affect, Democratic.Quality y Delivery.Quality. Por tanto, creamos dicho modelo para hacer predicciones.

```
modeloOptimo <- lm(formula = Life.Ladder ~ Log.GDP.per.capita + Social.support + Healthy.life.expectancy
```

Con este nuevo modelo, nos inventaremos dos países y realizaremos predicciones sobre sus niveles de felicidad:

```
paisInventadoFeliz <- data.frame(
  Log.GDP.per.capita = 13,
  Social.support = 0.99,
  Healthy.life.expectancy.at.birth = 100,
```

```

Freedom.to.make.life.choices = 0.99,
Generosity = 0.8,
Perceptions.of.corruption = 0.01,
Positive.affect = 0.99,
Confidence.in.national.government = 0.95
)

paisInventadoInfeliz <- data.frame(
  Log.GDP.per.capita = 2.5,
  Social.support = 0.05,
  Healthy.life.expectancy.at.birth = 38,
  Freedom.to.make.life.choices = 0.1,
  Generosity = 0.01,
  Perceptions.of.corruption = 0.95,
  Positive.affect = 0.04,
  Confidence.in.national.government = 0.05
)
predict(modeloOptimo, paisInventadoFeliz)

##          1
## 9.295365

predict(modeloOptimo, paisInventadoInfeliz)

##          1
## -0.5011555

```

Como podemos ver, el modelo funciona bien para estos dos casos. En el país que tiene los buenos indicadores predice un nivel de felicidad superior a 9. En cambio, en el país infeliz llega incluso a producir un valor negativo de felicidad.

4.2.3.3 ¿Qué es lo que nos hace más felices?

Considerando todos los indicadores de los que consta este estudio, queremos saber cuáles son los más importantes a la hora de calcular nuestro nivel de felicidad. Para esto, realizaremos un análisis de la correlación entre las variables. Dado que nuestros datos no cumplen la hipótesis de homocedasticidad, utilizaremos el Coeficiente de Correlación de Spearman.

```

matrizCorrelaciones <- matrix(nc = 2, nr = 0)
colnames(matrizCorrelaciones) <- c("coeficiente", "p-valor")
# Coeficiente de correlacion de las variables independientes con respecto a la felicidad
for (i in 4:(ncol(data)-2)) {
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
    spearman_test = cor.test(data[,i], data$Life.Ladder, method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    matrizCorrelaciones <- rbind(matrizCorrelaciones, pair)
    rownames(matrizCorrelaciones)[nrow(matrizCorrelaciones)] <- colnames(data)[i]
  }
}
print(matrizCorrelaciones)

```

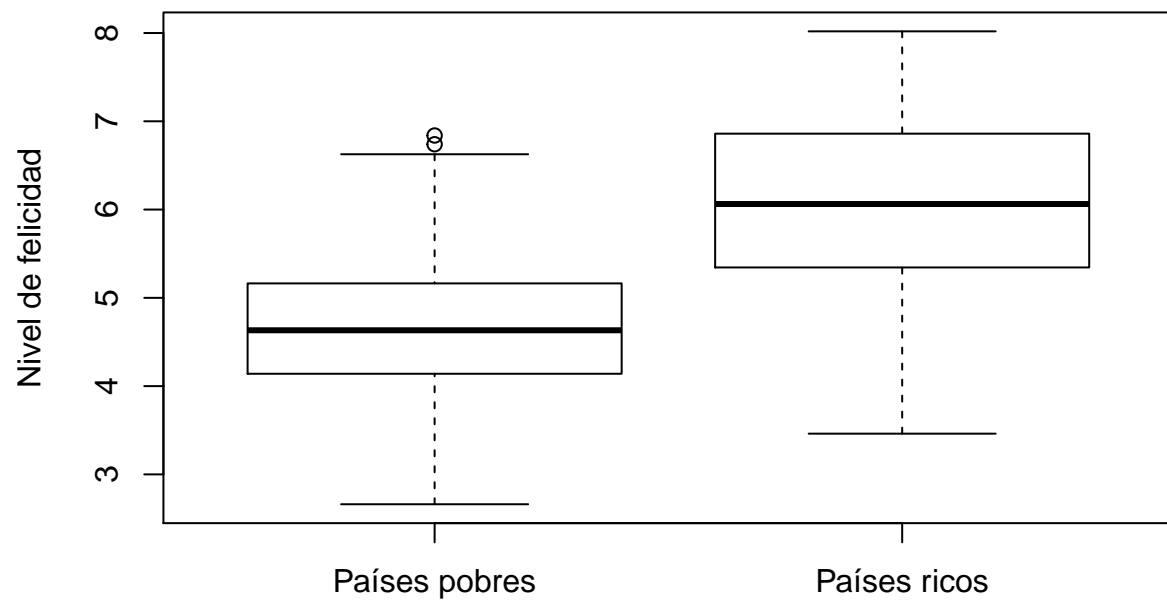
##	coeficiente	p-valor
## Log.GDP.per.capita	0.7927167	0.000000e+00
## Social.support	0.7488933	1.917603e-306
## Healthy.life.expectancy.at.birth	0.7686560	0.000000e+00
## Freedom.to.make.life.choices	0.5201765	9.156201e-119
## Generosity	0.1638708	1.006295e-11
## Perceptions.of.corruption	-0.3275129	6.825021e-44
## Positive.affect	0.5423818	6.180923e-131
## Negative.affect	-0.2348233	8.878472e-23
## Confidence.in.national.government	-0.1106234	4.712753e-06
## Democratic.Quality	0.5968013	5.934618e-165
## Delivery.Quality	0.6674902	3.112323e-220

A la vista de los resultados podemos ver que las variables con más peso estadístico con respecto a la felicidad son el apoyo social, la calidad de la democracia y la eficacia del gobierno. También podemos ver como es lógico que los coeficientes de la corrupción o de las emociones negativas son negativos (su ausencia aumenta la felicidad) y que el de la renta per cápita o la esperanza de vida son los más positivos (su presencia aumenta la felicidad).

5 Representación de los resultados a partir de tablas y gráficas

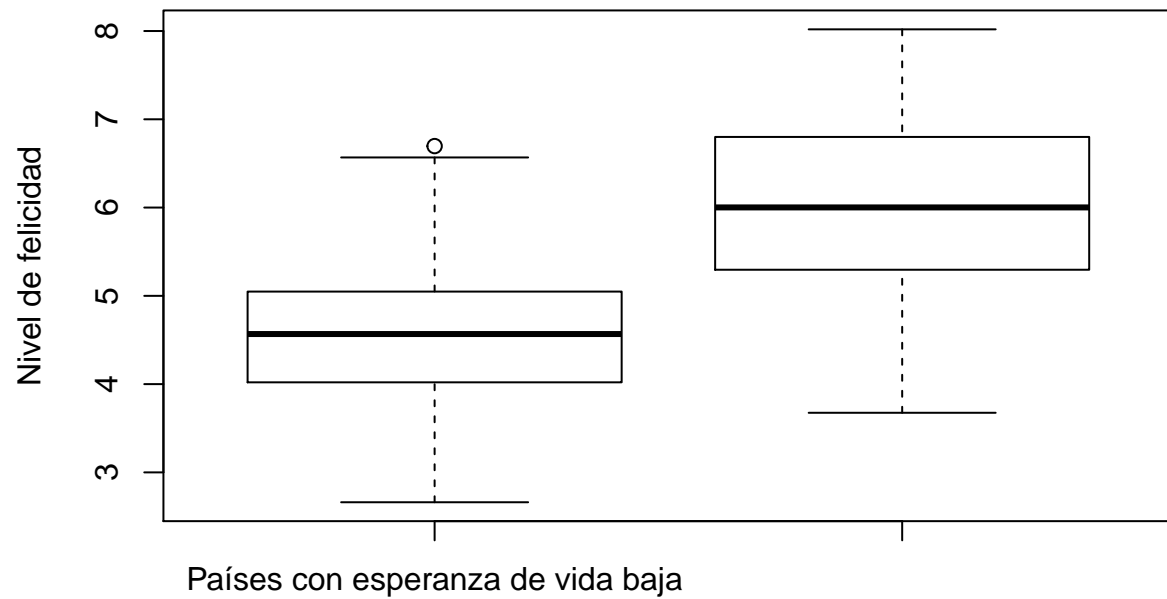
En primer lugar veremos gráficamente la representación del nivel de felicidad entre los países con una renta per cápita por debajo de la media y los que la tienen por encima de la media.

```
boxplot(paisesPobres$Life.Ladder, paisesRicos$Life.Ladder, names = c("Países pobres", "Países ricos"), y
```

Se puede ver en este diagrama de cajas como el nivel de felicidad de los países ricos es claramente superior al nivel de felicidad de los países pobres. Haremos lo mismo para la esperanza de vida:

```
boxplot(paísesEVBaja$Life.Ladder, paísesEVALta$Life.Ladder, names = c("Países con esperanza de vida baja", "Países con esperanza de vida alta"))
```

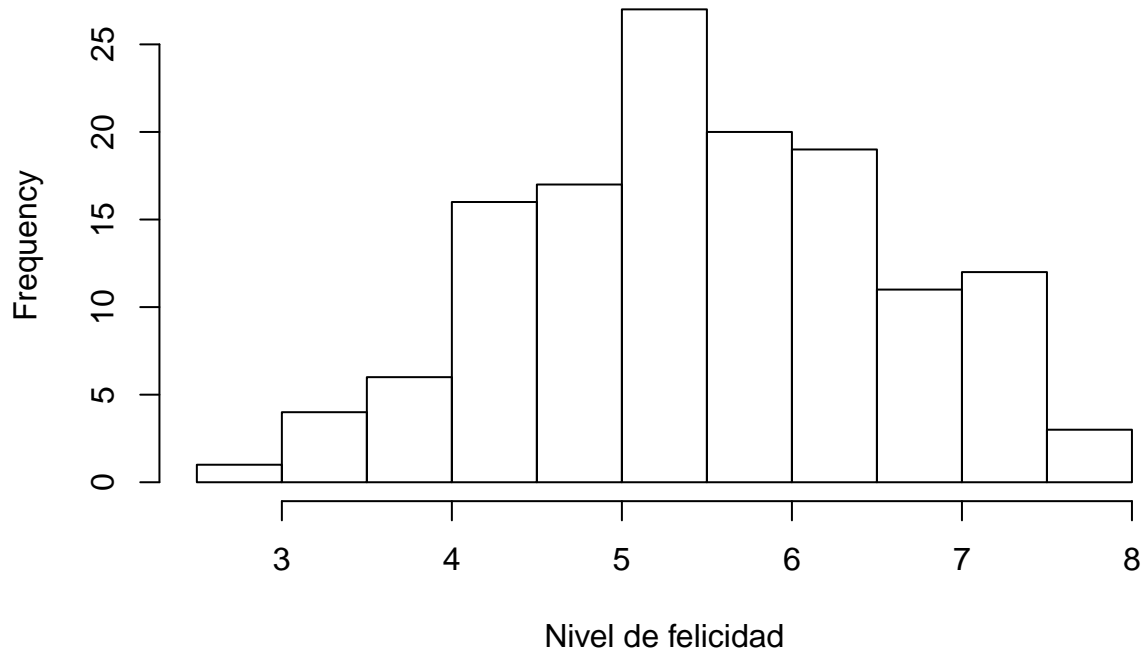


De nuevo podemos ver como la esperanza de vida influye en el nivel de felicidad, siendo claramente superior aquel correspondiente a los países con una esperanza de vida más alta.

Veremos ahora una distribución de la felicidad en los distintos países para el año 2018:

```
hist(data$Life.Ladder[data$Year=="2018"], xlab = "Nivel de felicidad", main = "Distribución de la felicidad en 2018")
```

Distribución de la felicidad en el año 2018



Podemos ver como la distribución de la felicidad parece seguir una especie de campana de Gauss, estando la mayoría de países entre el 4 y el 6.5.

6 Conclusiones

A la vista de la limpieza y del análisis de datos realizados, podemos concluir lo siguiente:

- Se ha conseguido realizar un análisis de datos sobre un conjunto que contenía muchos valores nulos. Esto ha sido posible gracias a una eficiente limpieza de los datos y a la utilización del método de los k-vecinos más próximos para rellenar los datos que faltaban.
- No se han podido comprobar las hipótesis de normalidad y homocedasticidad. No obstante, hemos supuesto que los datos eran normales utilizando el Teorema Central del Límite.
- Se ha podido comprobar que estadísticamente hablando el dinero da la felicidad. Por tanto la pregunta queda respondida. Se ha podido comprobar que la distribución de la felicidad es significativamente diferente entre los países que tienen una renta per cápita superior a la media y los que la tienen inferior. Además, observando los valores de felicidad para dichos países, se observa que los de los países con rentas superiores son mayores.
- También pudimos realizar una predicción del nivel de felicidad en función de varios de los indicadores utilizando un método de regresión lineal. El modelo generado podría ser de utilidad para evaluar países que no estén incluidos en este estudio y realizar predicciones futuras.

- Por último, hemos realizado una correlación que nos ha permitido observar que el apoyo social, la calidad de la democracia y la eficacia del gobierno son las variables más importantes a la hora de medir la felicidad. También hemos podido ver que una mayor esperanza de vida y una mayor renta per cápita influyen positivamente a la hora de tener una mayor felicidad.
-

7 Contribuciones

La práctica ha sido desarrollada de forma individual por Gabriel Quintairos Rial. Por tanto los 3 apartados requeridos en el enunciado: “Investigación previa”, “Redacción de las respuestas” y “Desarrollo código” tienen como integrante a Gabriel Quintairos Rial.