# Autor: Toni de Andrés Mascaró

# Cuestiones

## Ejercicio 1

El objetivo de esta actividad será la creación de un dataset a partir de los datos

contenidos al web. Tenéis que indicar las siguientes características del dataset

general:

1. Título del dataset. Poned un título que sea descriptivo.

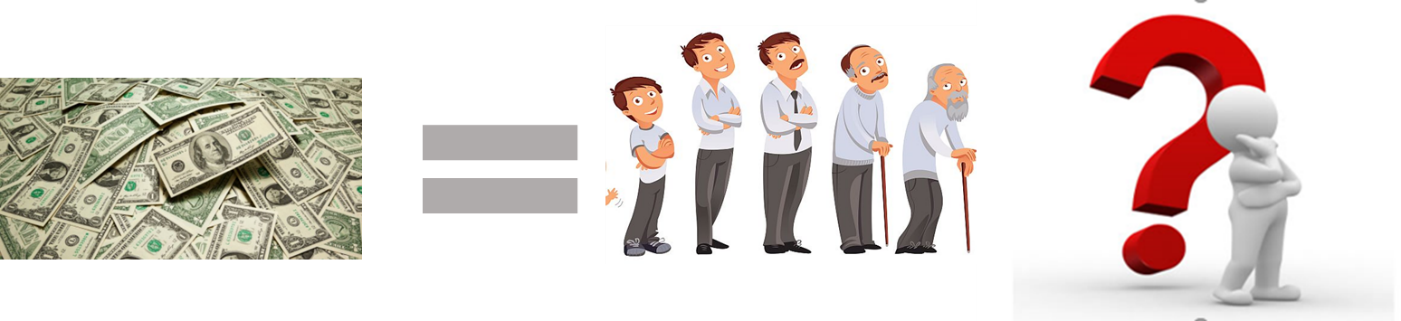
Riqueza vs. Esperanza de vida

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de

datos por vuestro subtítulo.

Estudio para entender como la riqueza de un país influye en la esperanza de vida de las personas

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente



4. Contexto. ¿Cuál es la materia del conjunto de datos?

El dataset que se ha definido es una primera aproximación a la relación que hay entre la riqueza de un país y la esperanza de vida de sus habitantes al nacer.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y

cómo se ha recogido?

El dataset está compuesto de tres campos, un primer campo que indica el país, un segundo campo que indica el la renta per cápita del país a valores de paridad de poder adquisitivo y el tercero indicando la esperanza de vida al nacer de las personas del país. Todos los datos son del año 2015 y se han extraído de Wikipedia bajo licencia [Creative Commons Atribución Compartir Igual 3.0](https://es.wikipedia.org/wiki/Wikipedia:Texto_de_la_Licencia_Creative_Commons_Atribuci%C3%B3n-CompartirIgual_3.0_Unported) la cual permite:

* ***Compartir****— copiar y redistribuir el material en cualquier medio o formato*
* ***Adaptar****— remezclar, transformar y crear a partir del material para cualquier finalidad, incluso comercial.*
* *El licenciador no puede revocar estas libertades mientras cumpla con los términos de la licencia.  
  Bajo las condiciones siguientes:*
* ***Reconocimiento****— Debe****reconocer adecuadamente****la autoría, proporcionar un enlace a la licencia e****indicar si se han realizado cambios****. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.*
* ***CompartirIgual****— Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la****misma licencia que el original****.*

***No hay restricciones adicionales****— No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.*

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de

investigación o análisis anteriores.

Los datos están extraídos de Wikipedia que a su vez se ha basado en los datos de la Organización Mundial de la Salud por lo que respecta a la esperanza de vida y del Fondo Monetario Internacional en lo relativo a la renta per cápita.

Ambos son organismos globales, públicos y mundialmente reconocidos que periódicamente publican datos relativos a temas de salud y economía respectivamente. La licencia bajo la cual el FMI internacional permite usar sus datos puede consultarse aquí: <http://www.imf.org/external/terms.htm>

La licencia bajo la que la OMS permite usar sus datos puede consultarse aquí: <http://www.who.int/about/copyright/en/>

Básicamente indican que, a efectos de estudios e investigación, hay libertad de uso de la información, pero se ha de hacer mención a la fuente de la misma.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas

le gustaría responder la comunidad?

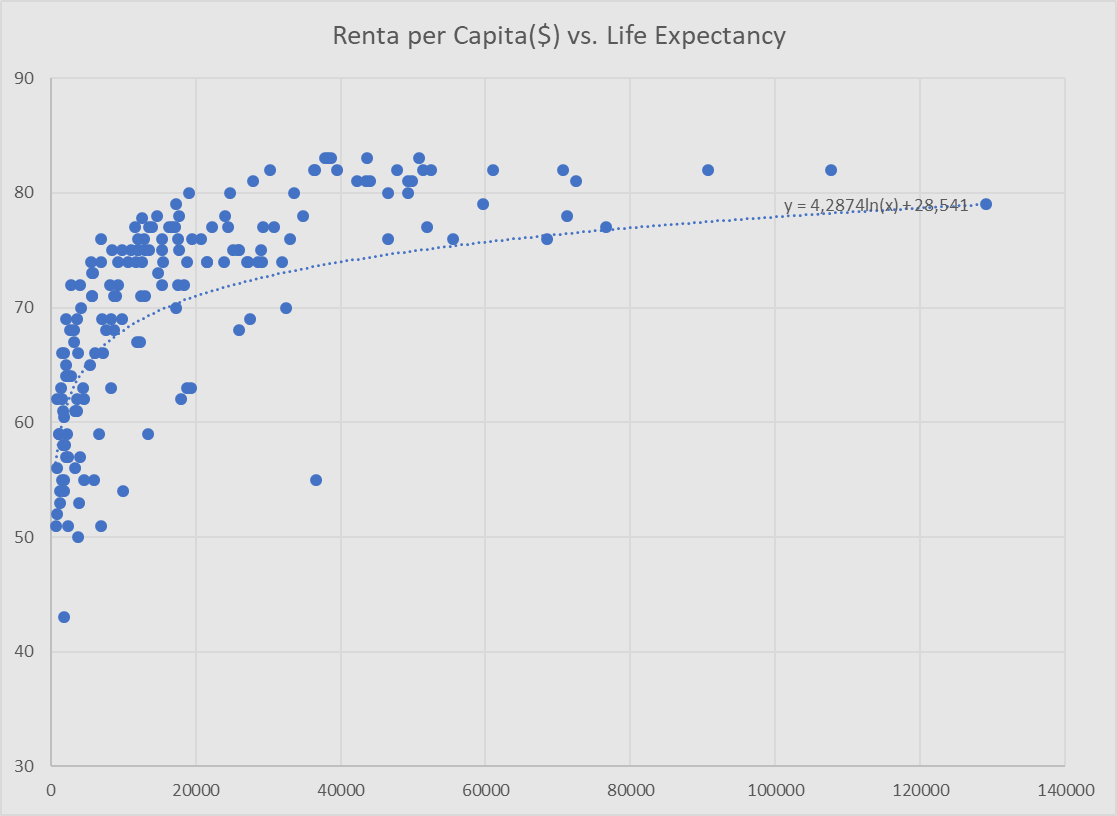
Un primer pensamiento podría llevarnos a pensar que cuanto más rico es un país más esperanza de vida se tiene y, esto, no necesariamente tiene que ser cierto. Pero, en cualquier caso, toda hipótesis tiene que tener por detrás unos datos que la soporten, de ahí el plantear un ejercicio que da una primera aproximación en base a dos variables (renta per cápita y esperanza de vida) que permite obtener una respuesta a la relación inmediata que hay, precisamente, entre la riqueza de un país y la esperanza de vida de su población al nacer.

Las personas o entidades que podrían estar interesados en aprovechar el dataset son aquellas que busquen una respuesta directa a la relación entre la riqueza y la esperanza de vida. Podemos pensar que tanto a economistas, que tengan como campo de estudio la relación entre la economía y la salud, y el sector médico/farmacéutico son grandes grupos que podrían encontrar interesante este dataset como medio para encontrar respuestas o para añadirlo a y aquellas que busquen respuestas añadiendo más variables al dataset, por lo que sería necesario integrarlo en otro dataset mayor como por ejemplo uno que integre factores de obesidad, polución, suicidio, horas de Sol, consumo de alcohol, etc. Que ayudaran a que el dataset pueda ser más potente a la hora de establecer relaciones entre las variables y a ayudar a explicar el porqué de las respuestas del modelo.

Ejemplos de preguntas que podrían responderse con el dataset actual son:

¿Qué relación hay entre la riqueza de un país y su esperanza de vida?

¿Hay algún nivel de renta a partir del cual la esperanza de vida aumenta o disminuye?



Las abscisas representan la renta per cápita a valores de paridad de poder adquisitivo en dólares y las ordenadas la esperanza de vida. Se puede observar a partir de cierto nivel de renta de unos 20.000$ la esperanza de vida no tiene una tendencia a mejorar. El gráfico está apoyado por una curva logarítmica que ayuda a visualizar la tendencia de como evoluciona la esperanza de vida con la renta.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

○ Released Under CC0: Public Domain License

○ Released Under CC BY-NC-SA 4.0 License

○ Released Under CC BY-SA 4.0 License

○ Database released under Open Database License, individual contents

under Database Contents License

○ Other (specified above)

○ Unknown License

La licencia que se selecciona es la [Released Under CC BY-NC-SA 4.0 License](https://creativecommons.org/licenses/by-nc-sa/4.0/). El motivo que subyace tras esta decisión está fundamentado en el hecho de que para poder llevar a cabo el ejercicio he podido aprovecharme de publicaciones de datos si tener que abonar ningún importe económico y no hay intención por mi parte de obtener un beneficio económico. Bajo la licencia descrita se puede hacer uso del dataset bajo las siguientes condiciones principales:

* Compartición. Se puede copiar y distribuir el material en cualquier medio o format.
* Adaptación. Se puede remezclar, transformar y construir sobre el dataset.
* Atribución. Debe darse el crédito apropiado, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerse de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante es respaldado por el autor.
* No se puede hacer uso comercial del dataset.
* Si remezcla, transforma o construye sobre el material, debe distribuir sus contribuciones bajo la misma licencia que el original.
* Sin restricciones adicionales: no puede aplicar términos legales o medidas tecnológicas que restrinjan legalmente que otros hagan cualquier cosa que la licencia permita.

Finalmente, los datos se publicarán en formato CSV en la web y se informará convenientemente en la CKAN para que sean fácilmente localizables por la comunidad.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset,

preferiblemente con R o Python, que os ha ayudado a generar el dataset

Ver Anexo I

10. Dataset: Dataset en formato CSV

Ver Anexo II

## Anexo I. Código R

#Fuentes de datos:

#Esperanza de vida: https://es.wikipedia.org/wiki/Anexo:Pa%C3%ADses\_por\_esperanza\_de\_vida

#Renta per capita a valores de paridad de poder adquisitivo: <https://es.wikipedia.org/wiki/Anexo:Pa%C3%ADses_por_PIB_(PPA)_per_c%C3%A1pita>

#Pasos a dar:

#------------

#1- Buscar y capturar los datos de interés para el estudio a realizar. Asegurarse de que la licencia que hay sobre los mismos permite capturarlos.

#2- Capturar los datos con Web Scraping.

#3- Transformar los datos para poder hacer el estudio deseado

#4- Realizar los análisis sobre los datos transformados

#5- Publicar los resultados. Elegir la licencia de publicación

#1- Capturar los datos:

#Comandos R

#------------

#Objetivo: relacionar la renta per capita la esperanza de vida

#Carga de las librerias necesarias

library(stringr)

library(XML)

#library(maps)

library(RCurl)

#Captura de los datos en wikipedia

#Esperanza de Vida

urlev="https://es.wikipedia.org/wiki/Anexo:Pa%C3%ADses\_por\_esperanza\_de\_vida" #se guarda la url en una variable

EV=getURL(urlev) #se descarga la página web

tablesEV<-readHTMLTable(EV,stringSAsFactors=FALSE) #se extraen las tablas

tableEV<-tablesEV[[2]] #se filtra sólo la tabla 2 que contiene la info buscada

tableEV<-tableEV[,c(2,3)] #se seleccionan sólo las columnas de la tabla que interesan

names(tableEV)<-c("Pais", "EV") #cambiamos el nombre de las etiquetas

#Se repite proceso para la tabla PIB

urlpib="https://es.wikipedia.org/wiki/Anexo:Pa%C3%ADses\_por\_PIB\_(PPA)\_per\_c%C3%A1pita" #se guarda la url en una variable

PIB=getURL(urlpib) #se descarga la página web

tablesPIB<-readHTMLTable(PIB,stringSAsFactors=FALSE) #se estraen las tablas

tablePIB<-tablesPIB[[3]] #se filtra sólo la tabla 2 que contiene la info buscada

tablePIB<-tablePIB[,c(2,3)] #se seleccionan sólo las columnas de la tabla que interesan

names(tablePIB)<-c("Pais", "PIBDolares") #cambiamos el nombre de las etiquetas

#Ahora se fusionan las tablas y tiene que hacer un match de los países para determinar la Esperanza de vida ligada al PIB

tabla<-list(pais=tablePIB$Pais,pib=tablePIB$PIBDolares) #se monta la tabla con los datos

#Se limpian los datos quitando espacios en blancos y comas mal puestas de los importes

tabla$pib<-gsub(",","",tabla$pib)

tabla$pib<-gsub(" ","",tabla$pib)

# Match de paises para tener la esperanza de vida

for (i in 1:length(tabla$pais)){

for (x in 1:length(tableEV$Pais)){

if (str\_detect(as.character(tableEV$Pais[x]), as.character(tabla$pais[i]))){

tabla$ev[i]<-as.character(tableEV$EV[x])

x<-length(tableEV$Pais)+1}

}

}

#Grabación de la tabla en un fichero CSV

write.csv(tabla, "PIBvsLE.csv")

## Anexo II. Dataset

"","pais","rpc","ev"

"1","Catar","129112","79"

"2","Luxemburgo","107736","82"

"3","Singapur","90724","82"

"4","Brunéi","76567","77"

"5","Irlanda","72529","81"

"6","Kuwait","71306","78"

"7","Noruega","70665","82"

"8","Emiratos Árabes Unidos","68424","76"

"9","Suiza","61014","82"

"10","Hong Kong","60553",NA

"11","Estados Unidos","59609","79"

"12","Arabia Saudita","55477","76"

"13","Países Bajos","53139",NA

"14","Islandia","52496","82"

"15","Baréin","51956","77"

"16","Suecia","51377","82"

"17","Australia","50817","83"

"18","Taiwán","49901",NA

"19","Alemania","49814","81"

"20","Austria","49370","81"

"21","Dinamarca","49364","80"

"22","Canadá","47771","82"

"23","Bélgica","46551","80"

"24","Omán","46475","76"

"25","Reino Unido","44001","81"

"26","Francia","43652","83"

"27","Finlandia","43545","81"

"28","Japón","42860",NA

"29","Malta","42239","81"

"30","Corea del Sur","39446","82"

"31","Nueva Zelanda","38706","83"

"32","España","38238","83"

"33","Italia","37905","83"

"34","Guinea Ecuatorial","36556","55"

"35","Chipre","36442","82"

"36","Israel","36378","82"

"37","República Checa","34849","78"

"38","Eslovenia","33579","80"

"39","Eslovaquia","33054","76"

"40","Trinidad y Tobago","32520","70"

"41","Lituania","31848","74"

"42","Estonia","30764","77"

"43","Portugal","30192","82"

"44","Polonia","29349","77"

"45","Seychelles","29155","74"

"46","Hungría","28965","75"

"47","Malasia","28636","74"

"48","Grecia","27861","81"

"49","Rusia","27466","69"

"50","Letonia","27189","74"

"51","San Cristóbal y Nieves","27130","74"

"52","Antigua y Barbuda","25998","75"

"53","Kazajistán","25942","68"

"54","Turquía","25776","75"

"55","Bahamas","25173","75"

"56","Chile","24796","80"

"57","Panamá","24526","77"

"58","Croacia","24052","78"

"59","Rumania","23957","74"

"60","Uruguay","22271","77"

"61","Mauricio","21607","74"

"62","Bulgaria","21498","74"

"63","Argentina","20707","76"

"64","México","19480","76"

"65","Gabón","19396","63"

"66","Líbano","19128","80"

"67","Irán","18858","74"

"68","Turkmenistán","18771","63"

"69","Bielorrusia","18348","72"

"70","Botsuana","17918","62"

"71","Tailandia","17749","75"

"72","Barbados","17735","78"

"73","Montenegro","17554","76"

"74","Azerbaiyán","17498","72"

"75","Irak","17326","70"

"76","Costa Rica","17260","79"

"77","República Dominicana","17096","77"

"78","China","16676","77"

"79","Maldivas","16275","77"

"80","Brasil","15485","74"

"81","Macedonia","15380","76"

"82","Serbia","15320","75"

"83","Argelia","15281","72"

"84","Granada","14764","73"

"85","Colombia","14609","78"

"86","Surinam","13934","77"

"87","Libia","13502","75"

"88","Perú","13501","77"

"89","Sudáfrica","13409","59"

"90","Sri Lanka","13012","75"

"91","Egipto","12982","71"

"92","Venezuela","12856","76"

"93","Jordania","12558","74"

"94","Albania","12567","77,8"

"95","Indonesia","12432","71"

"96","Mongolia","12333","67"

"97","Túnez","12065","76"

"98","Santa Lucía","12022","75"

"99","Dominica","11978","75"

"100","Namibia","11838","67"

"101","San Vicente y las Granadinas","11790","74"

"102","Bosnia y Herzegovina","11564","77"

"103","Ecuador","11004","75"

"104","Georgia","10644","74"

"105","Suazilandia","9896","54"

"106","Paraguay","9779","75"

"107","Fiyi","9777","69"

"108","Jamaica","9307","74"

"109","El Salvador","9286","72"

"110","Armenia","9066","71"

"111","Marruecos","8801","71"

"112","Bután","8762","68"

"113","Ucrania","8677","71"

"114","Belice","8399","75"

"115","Guyana","8306","63"

"116","Filipinas","8270","69"

"117","Guatemala","8132","72"

"118","Bolivia","7552","68"

"119","India","7153","66"

"120","Uzbekistán","7023","69"

"121","Cabo Verde","6994","74"

"122","Vietnam","6925","76"

"123","Angola","6881","51"

"124","República del Congo","6720","59"

"125","Myanmar","6360",NA

"126","Laos","6115","66"

"127","Nigeria","5960","55"

"128","Nicaragua","5755","73"

"129","Samoa","5752","73"

"130","Tonga","5703","71"

"131","Moldavia","5697","71"

"132","Honduras","5492","74"

"133","Pakistán","5374","65"

"134","Ghana","4650","62"

"135","Sudán","4578","55"

"136","Mauritania","4488","63"

"137","Bangladés","4207","70"

"138","Camboya","4022","72"

"139","Zambia","3982","57"

"140","Costa de Marfil","3841","53"

"141","Timor Oriental","3774","66"

"142","Lesoto","3752","50"

"143","Kirguistán","3669","69"

"144","Papúa Nueva Guinea","3635","62"

"145","Yibuti","3584","61"

"146","Kenia","3516","61"

"147","Camerún","3358","56"

"148","Tanzania","3296","61"

"149","Santo Tomé y Príncipe","3220","67"

"150","Tayikistán","3146","68"

"151","Vanuatu","2747","72"

"152","Senegal","2732","64"

"153","Nepal","2642","68"

"154","Yemen","2476","64"

"155","Chad","2439","51"

"156","Malí","2360","57"

"157","Benín","2229","59"

"158","Uganda","2155","57"

"159","Etiopía","2104","64"

"160","Ruanda","2090","65"

"161","Islas Salomón","2031","69"

"162","Zimbabue","2002","58"

"163","Burkina Faso","1879","58"

"164","Kiribati","1878","66"

"165","Afganistán","1833","60,5"

"166","Haití","1819","55"

"167","Guinea-Bisáu","1816","54"

"168","Sierra Leona","1760","43"

"169","Gambia","1697","61"

"170","Togo","1619","58"

"171","Comoras","1567","62"

"172","Sudán del Sur","1554","55"

"173","Madagascar","1563","66"

"174","Eritrea","1441","63"

"175","Guinea","1315","54"

"176","Mozambique","1263","53"

"177","Malaui","1177","59"

"178","Níger","1154","59"

"179","Liberia","879","62"

"180","Burundi","812","56"

"181","República Democrática del Congo","788","52"

"182","República Centroafricana","684","51"