

INTRODUCCIÓN A R

DICIEMBRE 2020



WORLD HAPPINESS REPORT

SILVIA GOÑI
MARIA PÉREZ



Introducción	2
Preparación de datos	4
Variables	4
Observaciones	5
Gráficos	6
Variables	6
Relaciones entre las variables	10
Correlaciones	14
Mapa	16
Mapas por años	17
Mapa evolución 2015-2019	19
Geoposicionamiento	24



Introducción

Nuestra intención inicial con este trabajo es medir y comparar los niveles de felicidad entre los países en base a una serie de variables tales como la economía, salud y libertad. Lo haremos a partir del World Happiness Report, índice global de felicidad publicado anualmente por las Naciones Unidas y que mide la felicidad en 157 países. Para nuestro proyecto, hemos obtenido los datos de la página web [Kaggle](#) y analizaremos los datos provistos de 2015 a 2019, ambos inclusive, de seis factores principales: economic production, social support, life expectancy, freedom, absence of corruption, and generosity. Adicionalmente, contaremos con Dystopia como un hipotético país extra cuyos valores serán equivalentes a la media de los resultados más bajos del resto del mundo para cada factor. El autor del contenido nos advierte de antemano que este hipotético país no tiene ningún tipo de impacto real sobre el marcaje/puntuación total atribuida a cada país, pero que sí que explicaría por qué ciertos países se clasifican por encima de otros.

Para la recabación de los datos que componen nuestras bases, las puntuaciones se fundamentan en las respuestas a una principal pregunta de evaluación de vida. La pregunta, también conocida con *Cantril ladder*, pide a los participantes imaginarse una escalera en la que la mejor vida posible para ellos representaría un 10 y la peor vida posible supondría un 0. En este escenario, tienen que calificar sus vidas actualmente en la escala.

A partir de estos datos nos gustaría estudiar cuáles son aquellos países que se clasifican como los más felices del mundo frente a los más infelices, comprobar si existe algún tipo de correlación significativa con alguno de los factores de la evaluación, ver si ha habido algún tipo de cambio drástico de alguno de los países...

- ❑ Country – nombre del país.
- ❑ Region – región a la que pertenece el país.
- ❑ Happiness Rank – puesto del país en función de su puntuación de felicidad.
- ❑ Happiness Score – métrica medida en X año, realizando la pregunta a los entrevistados.
- ❑ Standard Error – error estándar del Happiness Score.
- ❑ Economy – la medida en la que el PIB contribuye al cálculo del Happiness Score.
- ❑ Family – la medida en la que la familia contribuye al cálculo del Happiness Score.
- ❑ Health – la medida en la que la esperanza de vida contribuye al cálculo del Happiness Score.
- ❑ Freedom – la medida en la que la libertad contribuye al cálculo del Happiness Score.
- ❑ Trust – la medida en la que la percepción de la corrupción contribuye al cálculo del Happiness Score.

Las siguientes columnas: PIB per cápita, Familia, Esperanza de vida, Libertad, Generosidad, Confianza Corrupción del gobierno describen en qué medida estos factores contribuyen a evaluar la felicidad en cada país. La métrica de distopía residual en realidad es la puntuación de felicidad de distopía (1,85) + el valor residual o el valor no explicado para cada país.

Preparación de datos

Variables

En primer lugar, hemos empezado por limpiar los datos. Como son 5 csv –de 5 años consecutivos– con distinto número de columnas –indicadores– y filas –países–, hemos pensado que es mejor que todos contengan la misma información, lo que hará más fácil compararla.

data15	data16	data17	data18	data19
country	country	country	overall rank	overall rank
region	region	happiness rank	country or region	country or region
happiness rank	happiness rank	happiness score	score	score
happiness score	happiness score	whisker high	gdp capita	gdp capita
std error	lower confidence interval	whisker low	social support	social support
economy gdp capita	upper confidence interval	economy gdp capita	healthy life expectancy	healthy life expectancy
family	economy gdp capita	family	freedom to make life choices	freedom to make life choices
health life expectancy	family	health life expectancy	generosity	generosity
freedom	health life expectancy	freedom	perceptions of corruption	perceptions of corruption
trust gov corruption	freedom	generosity		
generosity	trust gov corruption	trust gov corruption		
dystopia residual	generosity	dystopia residual		
	dystopia residual			

Viendo las variables, quisimos primero que los nombres de estas coincidieran, y luego las ordenamos para que estuvieran en el mismo orden, además de eliminar las variables que solamente estén en una de las tablas. Después de hacer el cambio de observaciones, que explicaremos a continuación, también decidimos añadir la región en las tablas en las que no aparecía, reordenando los países de forma alfabética para poder

añadir esa variable, y calculamos el "dystopia residual" de *data18* y *data19*. Por último, creamos una variable nueva con el año para después juntar todas las tablas en una.

Por tanto, así quedaron las variables después de todos esos cambios:

data15	data16	data17	data18	data19
Rank	Rank	Rank	Rank	Rank
Country	Country	Country	Country	Country
Region	Region	Region	Region	Region
Score	Score	Score	Score	Score
Economy	Economy	Economy	Economy	Economy
Family	Family	Family	Family	Family
Health	Health	Health	Health	Health
Freedom	Freedom	Freedom	Freedom	Freedom
Corruption	Corruption	Corruption	Corruption	Corruption
Generosity	Generosity	Generosity	Generosity	Generosity
Dystopia	Dystopia	Dystopia	Dystopia	Dystopia
Year	Year	Year	Year	Year

Observaciones

Al subir los datos, nos fijamos que cada documento tenía un número de observaciones, países en este caso, así que decidimos eliminar todos los países que no estaban en las 5 tablas para que la información fuera más comparable.

	data15	data16	data17	data18	data19
Nº países	158	157	155	156	156

Para hacer esto, utilizamos *data15* y la comparamos con el resto de tablas una a una, para que esta fuera la lista final de países comunes. Después, volvimos a usar *data15* para que las demás tablas tuvieran solamente esos datos. Finalmente, quedaron solamente 141 observaciones y 12 variables.

Gráficos

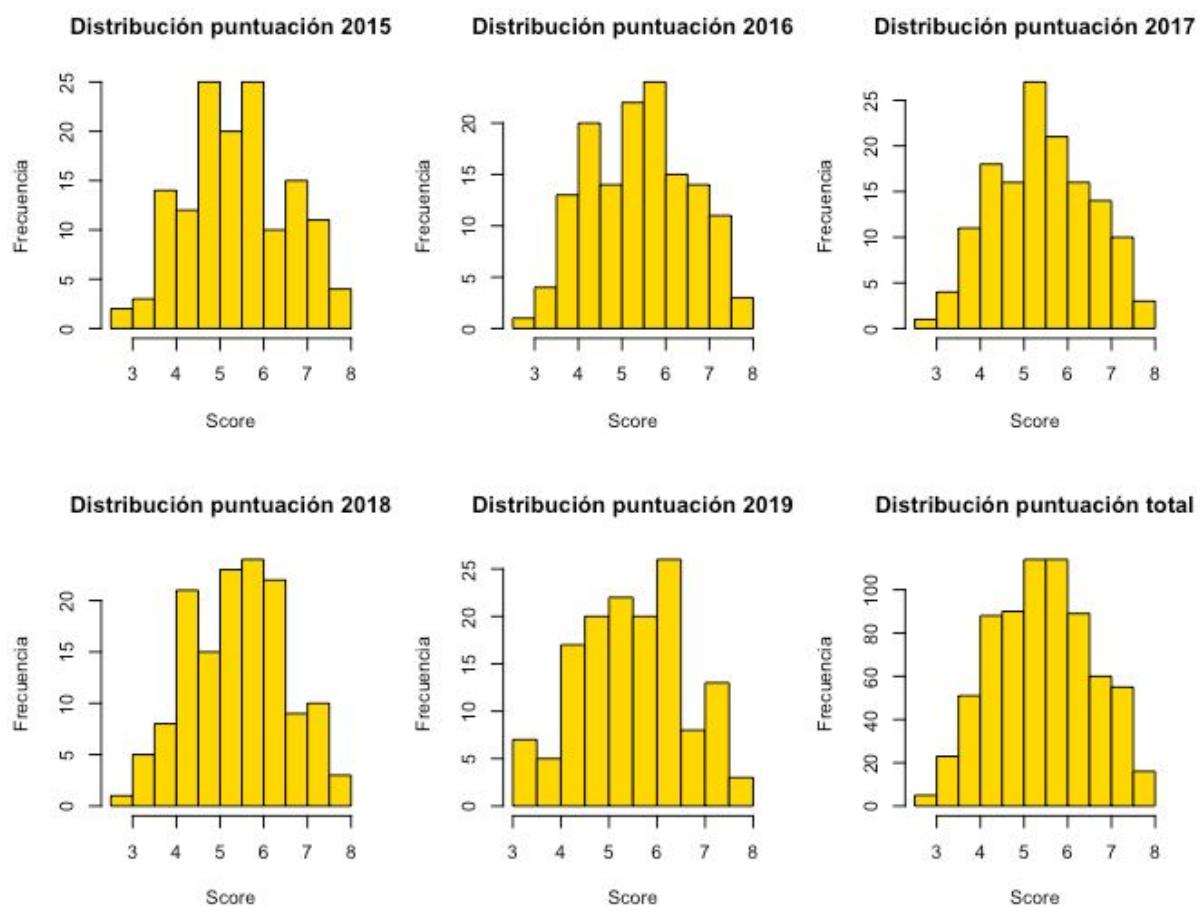
Para nuestro estudio, salvo aquellos gráficos que expresamente señalamos que estamos utilizando para comparar los años, vamos a realizar nuestra investigación y análisis sobre los datos recopilados del año 2015.

Variables

Histogramas distribución puntuaciones 2015-2019

Buscar en R: # HISTOGRAMA PUNTUACIONES POR AÑO

Función: `hist()`

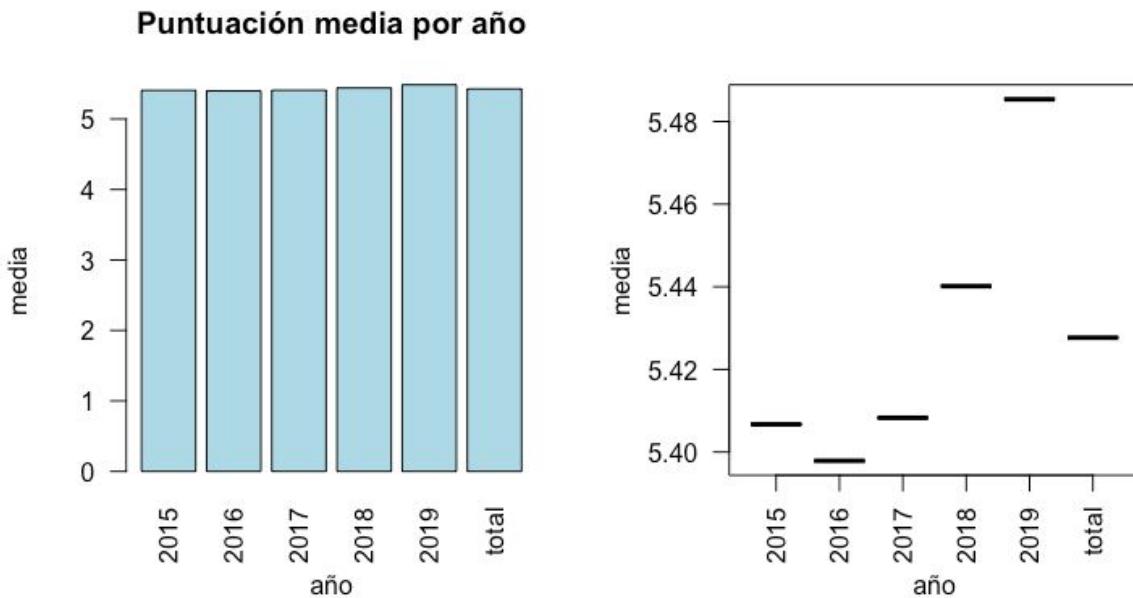


En primera instancia, hemos intentado analizar la distribución de los países según su Puntuación de Felicidad (Happiness Score). Por lo que podemos apreciar, parecen seguir una distribución relativamente normal. De forma generalizada, los países reportan con mayor frecuencia una puntuación de valores medios, comprendidos entre 4/10 y 7/10. Este hecho es algo lógico, pues si lo pensamos el país distópico ya parte de una base de una nota de al menos 1.85/10 y raros serán aquellos países que reporten notas excesivamente altas o bajas. De hecho, de acuerdo con estos gráficos ninguno de los países evaluados tiene una puntuación menor a 2.5/10, ni tampoco mayor a 8/10.

Diagrama de barras puntuación media por año

Buscar en R: # MEDIA SCORE POR AÑO

Funciones: barplot() y plot()

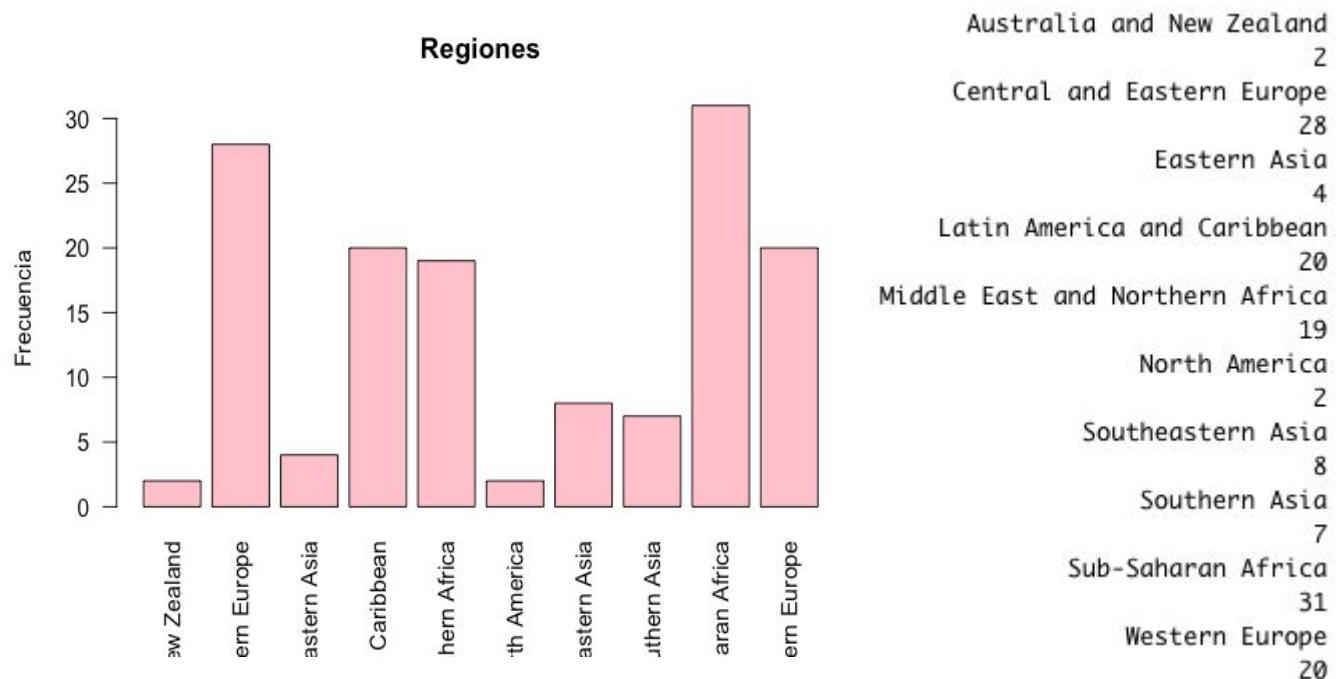


Seguidamente, hemos querido comprobar si por casualidad habría grandes diferencias en la media de la felicidad reportada por los países a lo largo de los años. En el primer diagrama de barras no logramos apreciar ninguna variación significativa. No obstante, con el segundo gráfico, somos capaces de ver un poco mejor en qué consisten verdaderamente las variaciones. Constatamos que el año que "mayor subida" atestigua en nuestra muestra ha sido 2019.

Número de países por región

Buscar en R: # GRÁFICO NÚMERO DE PAÍSES POR REGIÓN

Funciones: `plot()` y `summary()`

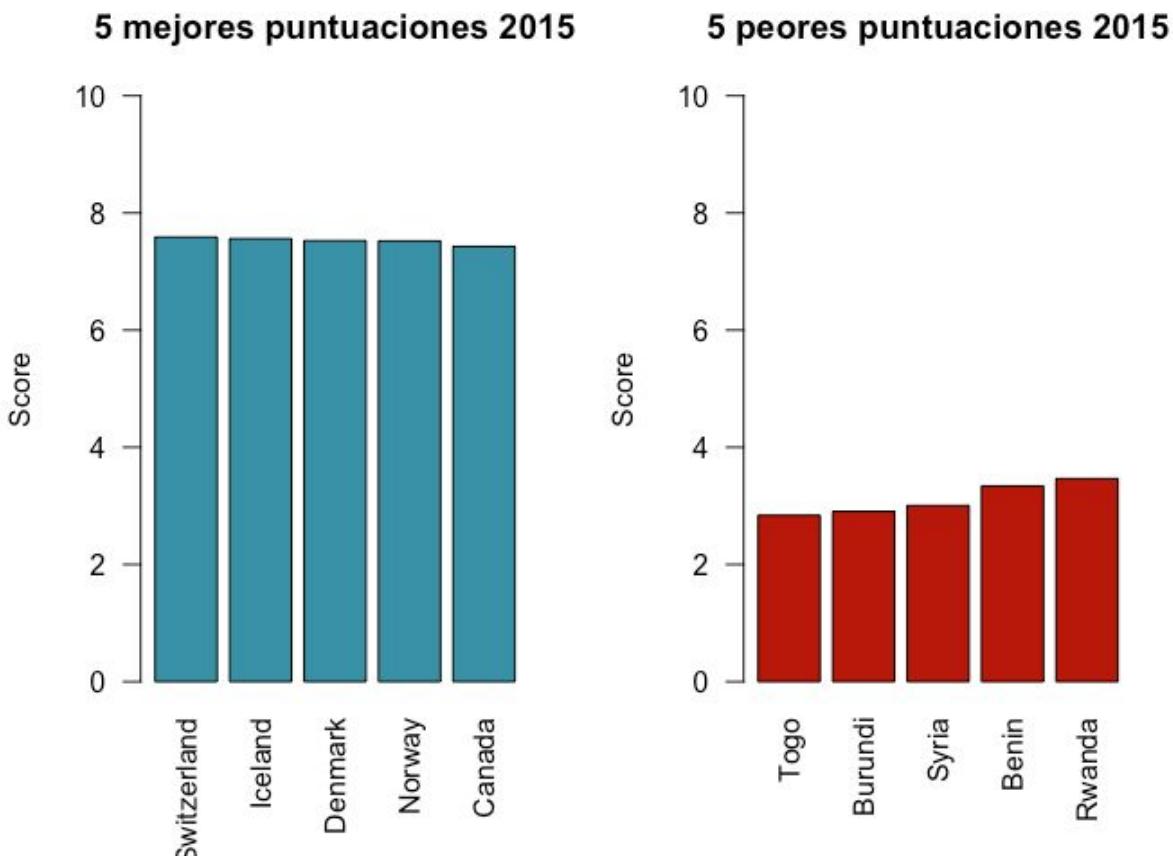


Con intención de entender un poco mejor nuestros datos y con qué estamos trabajando, hemos hecho un gráfico midiendo la frecuencia de los países que estudiamos según la región a la que pertenecen. Así es como vemos que la gran mayoría de los países que estudiamos son de África Subsahariana, Europa, Latinoamérica y el Caribe, África del Norte y Oriente Medio. Trabajamos con una totalidad de 141 países –después de nuestra criba y limpieza para que coincidiesen todos en los distintos documentos–.

Mejores y peores 5 puntuaciones de 2015

Buscar en R: # 5 MEJORES Y PEORES PUNTUACIONES POR PAÍSES

Función: `barplot()`



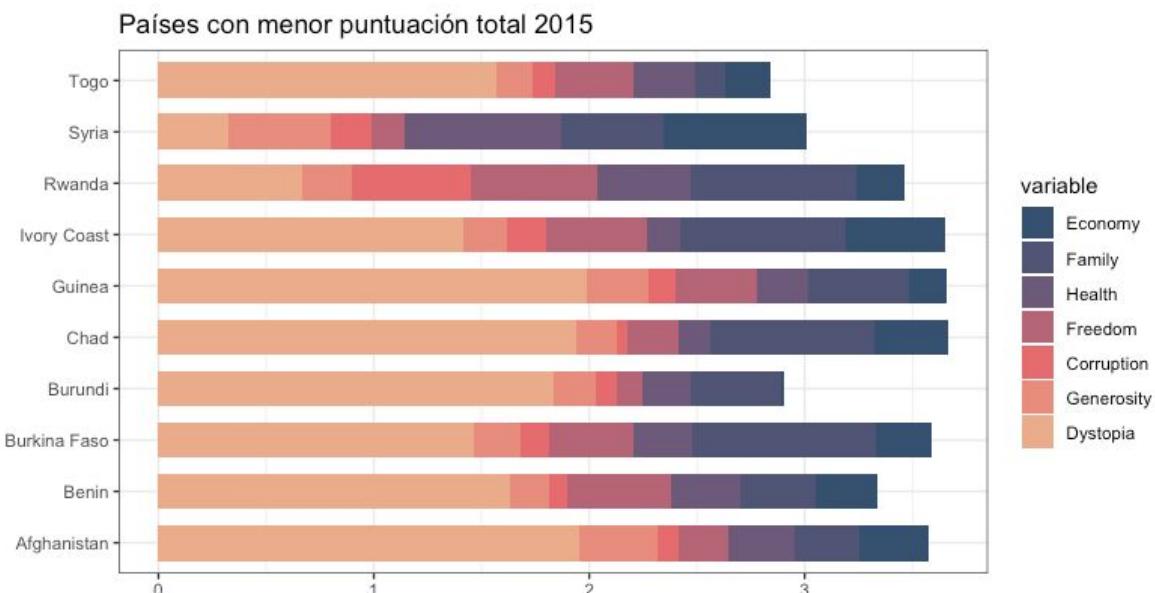
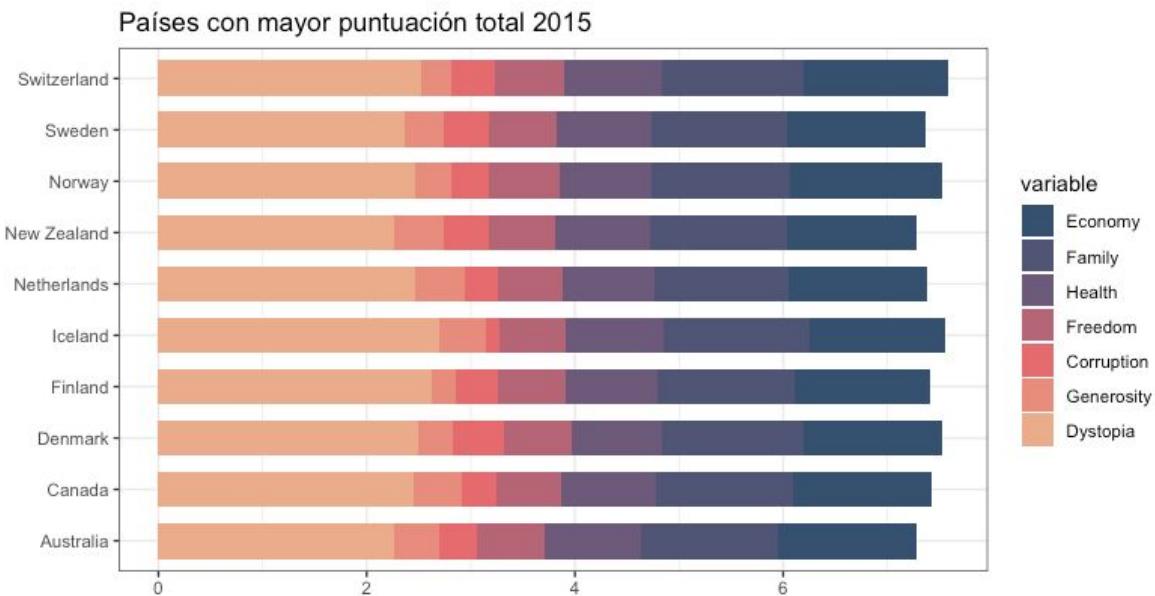
Por lo que parece, en 2015, los cinco países autodeclarados más felices del mundo, en base a las opiniones de su población, eran Suecia, Islandia, Dinamarca, Noruega y Canadá. En contraposición, en ese mismo año, los cinco países menos felices –por ende, más infelices–, eran Ruanda, Benín, Siria, Burundi y Togo. Se puede ver cómo la diferencia entre el país con mayor y peor puntuación –Suiza y Togo–, es de casi 5 puntos.

Relaciones entre las variables

Países con mayor y menor puntuación, dividido por las variables.

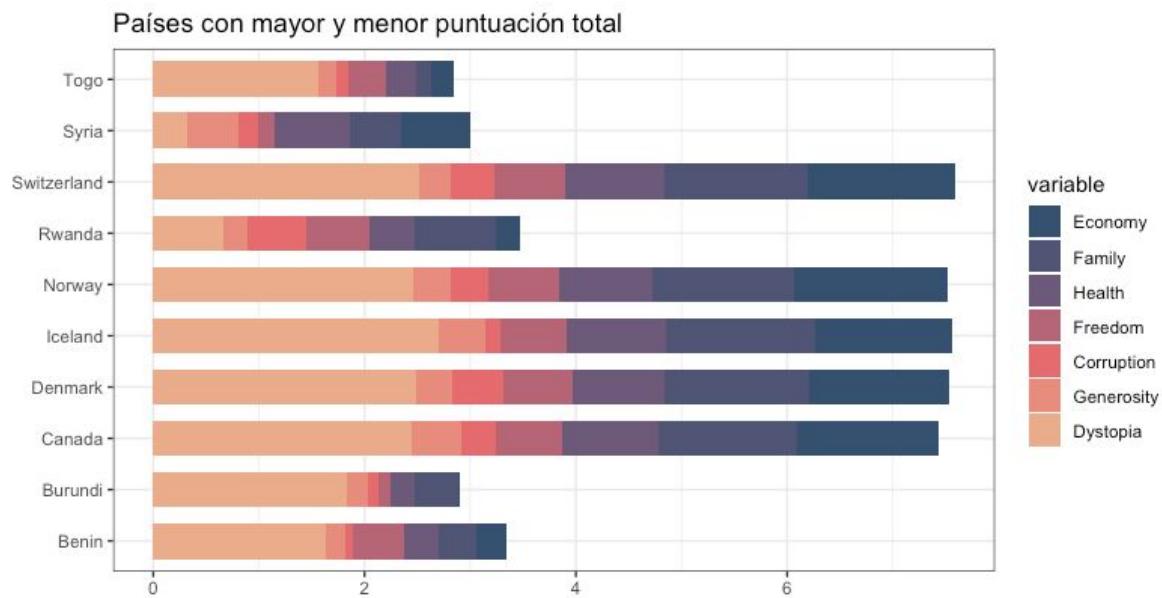
Buscar en R: # STACKED BARPLOT RANKING

Función: `ggplot()`



Queriendo ir más a fondo en el análisis de impacto de nuestras variables sobre la felicidad, hemos hecho un ranking del efecto de dichas variables sobre los diez países más felices y los diez más infelices. Asimismo, hemos hecho otro gráfico mezclando ambos estudios, con los cinco países más felices y los cinco más infelices dentro de un mismo gráfico. En primer lugar, nos puede llamar la atención el hecho de que dystopia

tenga tanto peso sobre la nota total de la felicidad, pero ese se debe a la manera en la que los datos han sido tratados para el cálculo del Happiness Score, con el mínimo anteriormente mencionado en el que dystopia supone de por sí 1.85 puntos. Dejando ese de lado, Economy y Health son aquellas que más peso parecen tener sobre la población entrevistada en cuanto al Happiness Score.



Boxplots de puntuación de felicidad por regiones

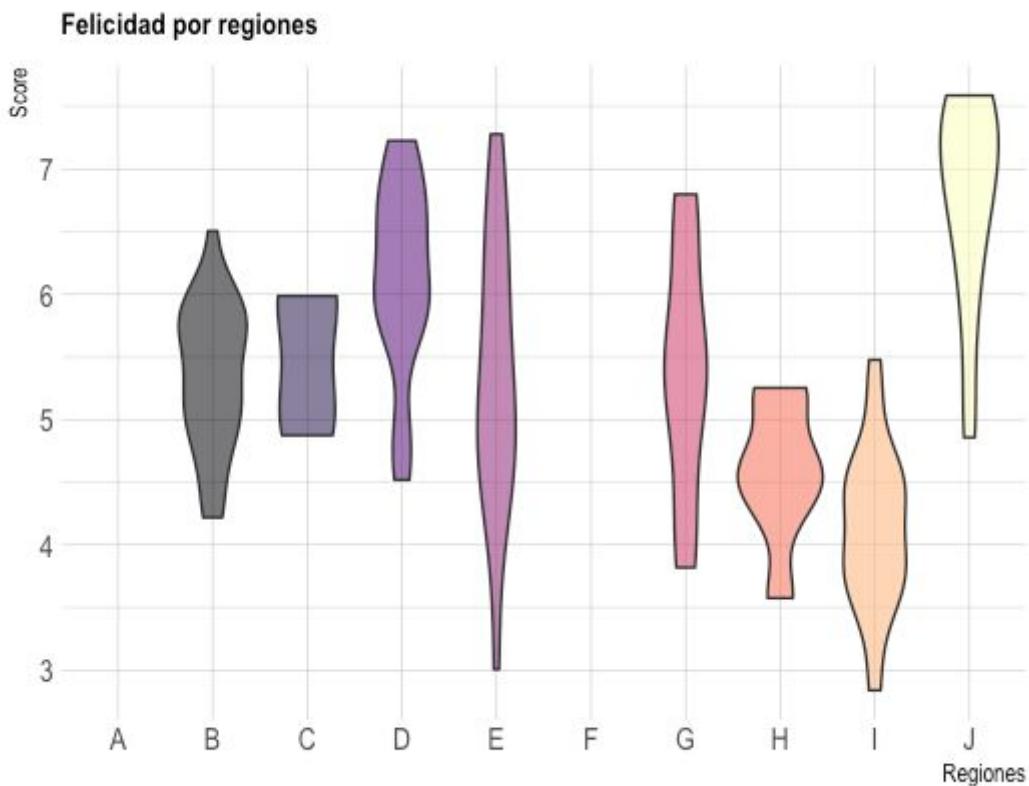
Buscar en R: # BOXPLOT HAPPINESS POR REGIÓN

Función: `ggplot()`

Leyenda regiones:

A	Australia and New Zealand	F	North America
B	Central and Eastern Europe	G	Southeastern Asia
C	Eastern Asia	H	Southern Asia
D	Latin America and Caribbean	I	Sub-Saharan Africa
E	Middle East and Northern Africa	J	Western Europe

Violin chart: en "Australia and New Zealand" y "North America" no hay nada porque sólo hay dos países en cada uno y tienen valores muy similares, por lo que no crea ningún gráfico.



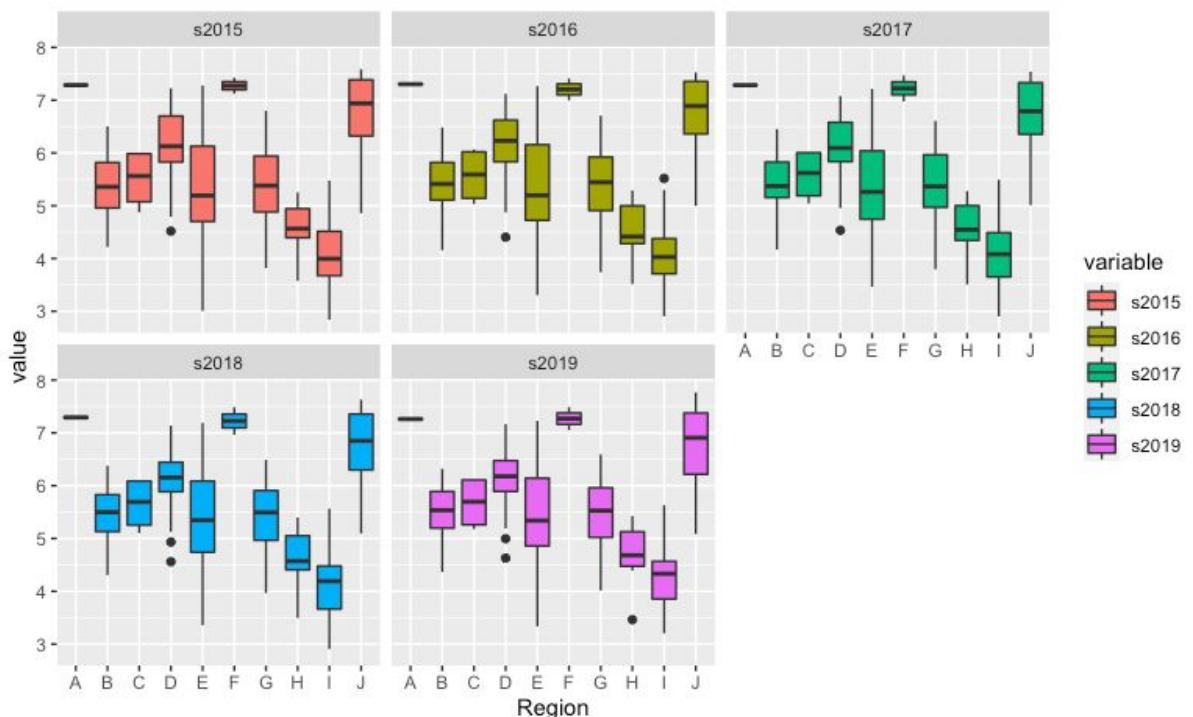
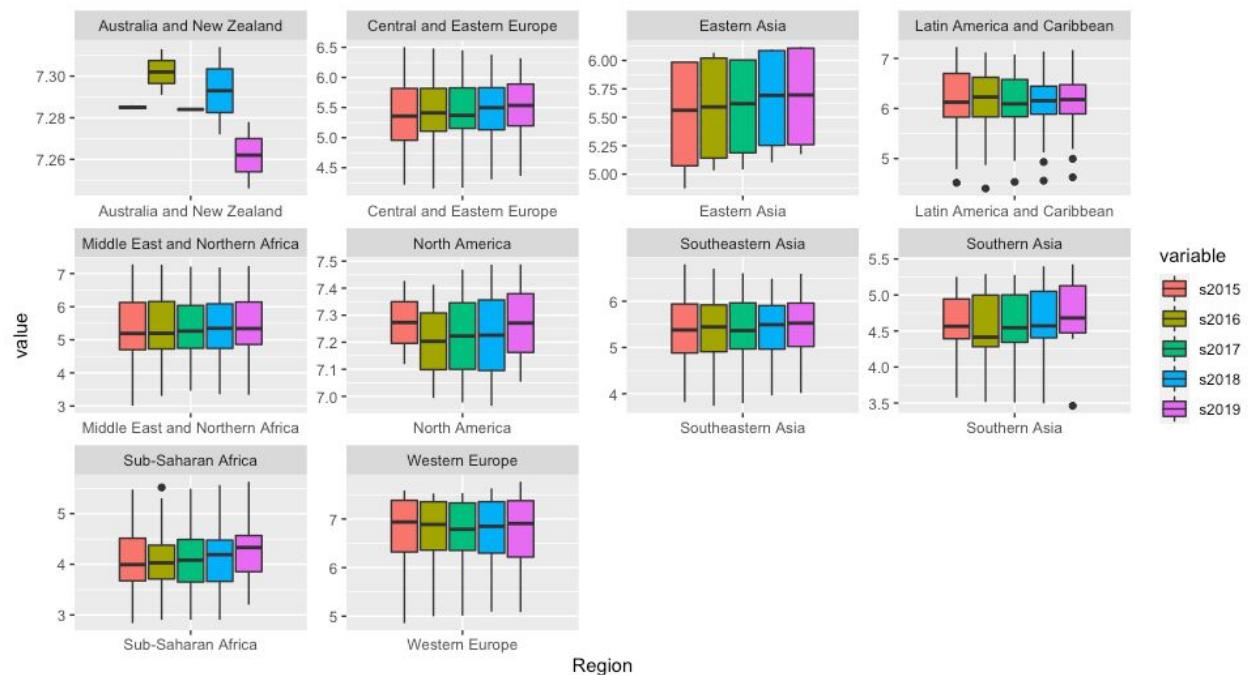
Hemos creído que podía resultar interesante intentar graficar de manera un poco distinta la distribución de los países, midiendo su felicidad de acuerdo con sus regiones. Por lo que parece, Oriente Medio y Norte de África es la región que mayor diversidad encuentra en términos de variación de su felicidad dentro de una misma región. El Oeste

de Europa es la zona que alcanza las puntuaciones de felicidad más altas y África Subsahariana la que tiene las puntuaciones de felicidad más bajas.

A continuación adjuntamos otras maneras de medir esto mismo, que puedan resultarnos más o menos útiles.

Buscar en R: # BOXPLOT HAPPINESS POR REGIÓN Y AÑOS

Función: `ggplot()`



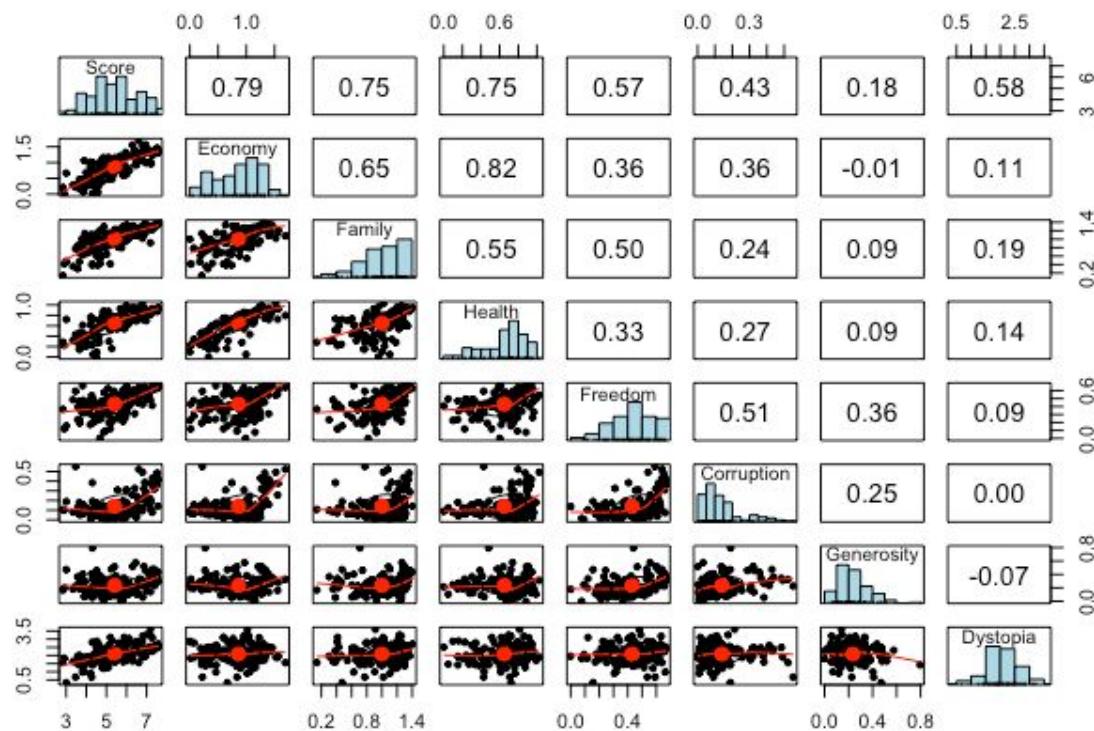
Correlaciones

En este caso usaremos solamente las variables numéricas.

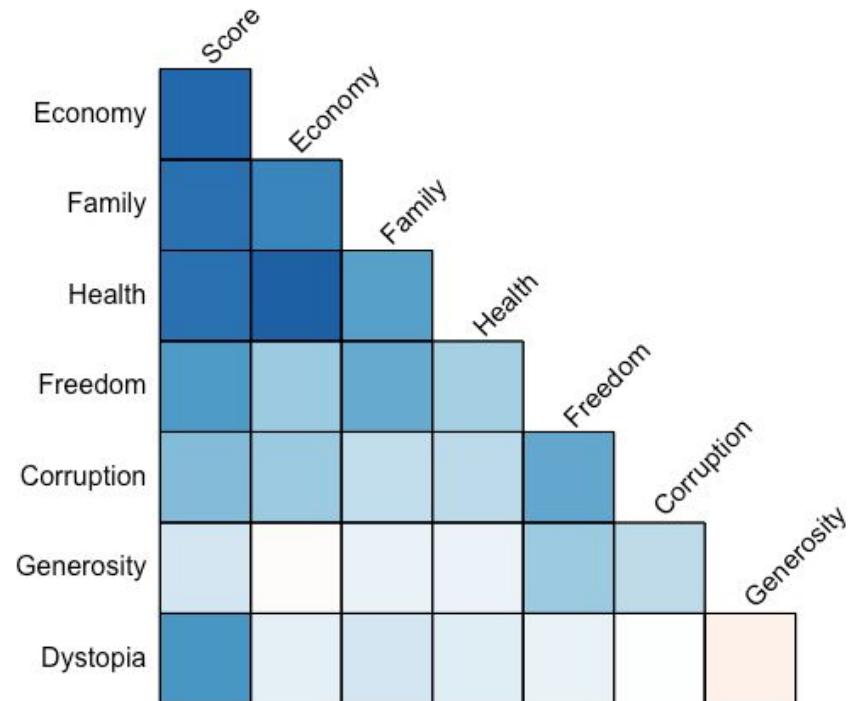
Buscar en R: # RELACIONES Y CORRELACIONES ENTRE VARIABLES NUMÉRICAS

Funciones: `pairs.panels()` y `corrplot()`

En esta ocasión, tratamos de sacar otro tipo de evidencia numérica o visual de la posible correlación existente entre las distintas variables.



En el siguiente gráfico, a mayor correlación, más oscura será la casilla. Al igual que en el anterior análisis, nos salta a la vista que aquellas variables que mayor correlación parecen demostrar con Happiness Score son las de *Health*, *Family* y *Economy*. Además, *Economy* y *Health* presentan una muy alta correlación entre ellas, de 0,82. Del mismo modo, nos facilita ver que una de las variables con menor correlación, si no la que menos, es la de *Generosity* –esto podría deberse a que como factor más cultural, se dé por sentado o se tenga visiones muy distintas–.



Mapa

Buscar en R: # MAPAS

Funciones: `fortify()` y `ggplot()`

Con la idea de visualizar mejor las distintas puntuaciones de felicidad -Score-, decidimos hacer unos mapas.

Para poder hacer esto utilizamos la ayuda de un tutorial de youtube (<https://www.youtube.com/watch?v=sLiNAsmpXPo>) y descargamos los datos correspondientes de: https://thematicmapping.org/downloads/world_borders.php.

Utilizando esa página web, descargamos los archivos y los subimos a R. Después, creamos un csv con los nombres de los países (paises.csv) para luego compararlos con el documento que ya tenemos.

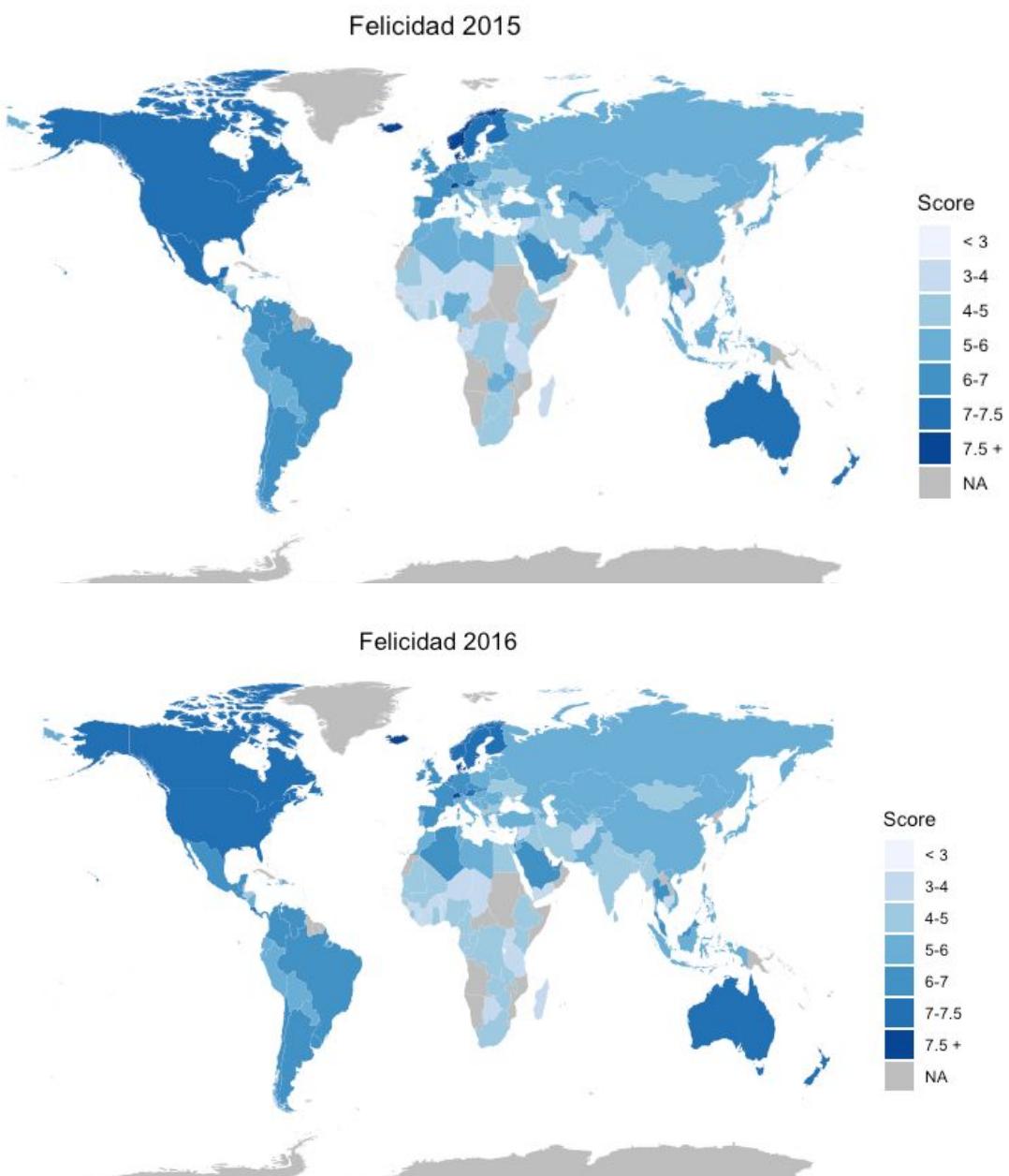
Al comprobar que los nombres de los países estuvieran escritos de la misma manera en ambos archivos, aparecieron varios distintos, así que utilizando `gsub` los cambiamos:

Nombre	Nombre nuevo
South Korea	Korea, Republic of
Moldova	Republic of Moldova
Libya	Libyan Arab Jamahiriya
Kosovo	NA → esta fila desaparece
Vietnam	Viet Nam
Palestinian Territories	Palestine
Iran	Iran (Islamic Republic of)
Congo (Kinshasa)	Democratic Republic of the Congo
Myanmar	Burma
Congo (Brazzaville)	Congo
Tanzania	United Republic of Tanzania
Ivory Coast	Cote d'Ivoire
Syria	Syrian Arab Republic

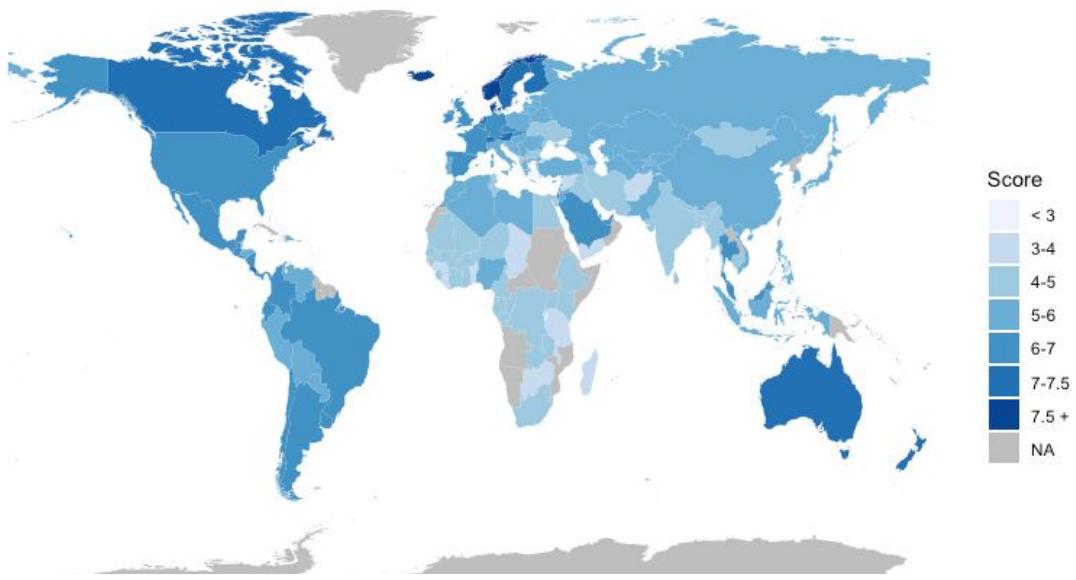
Mapas por años

Cuando ya tenemos los datos cambiados, tenemos que crear una variable nueva categórica para los datos de *Score*, que en este caso hemos llamado *Felicidad*. Al ser un dato que está entre 1 y 10, lo hemos dividido en tramos de un punto cada uno, es decir, de 1 a 2, de 2 a 3, etc. excepto en los últimos tramos, que decidimos dividirlo en medio punto para distinguir mejor los países –de 7 a 7.5– y por último, mayores de 7.5 –7.5+-. En el mapa también aparecen los países de los que no tenemos datos, NA.

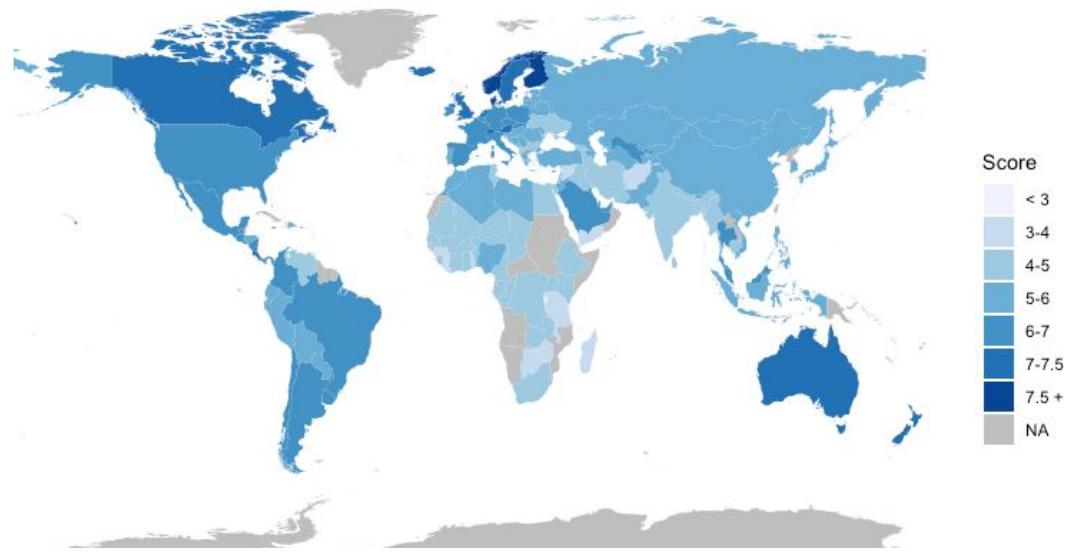
A continuación, creamos un nuevo data frame para cada año y que solamente contenga las variables *Country* –que pasa a llamarse *id*–, *Score* y *Felicidad*. Cambiamos el nombre a *id* para que ahora podamos juntar este data frame con el que creamos anteriormente ([paises.csv](#)) por esa variable. Con esto ya tenemos todo para poder hacer un mapa con la función `ggplot`.



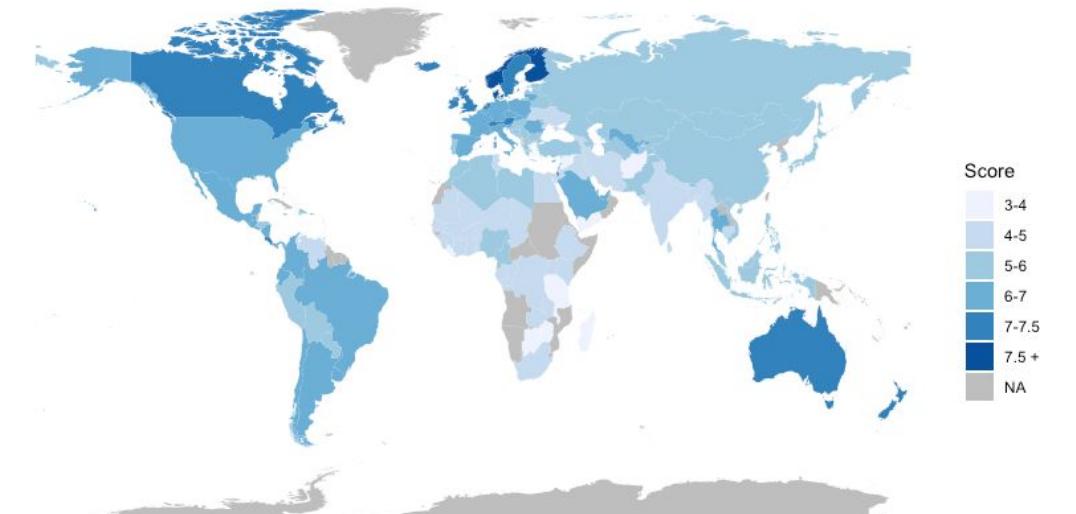
Felicidad 2017



Felicidad 2018



Felicidad 2019



Mapa evolución 2015-2019

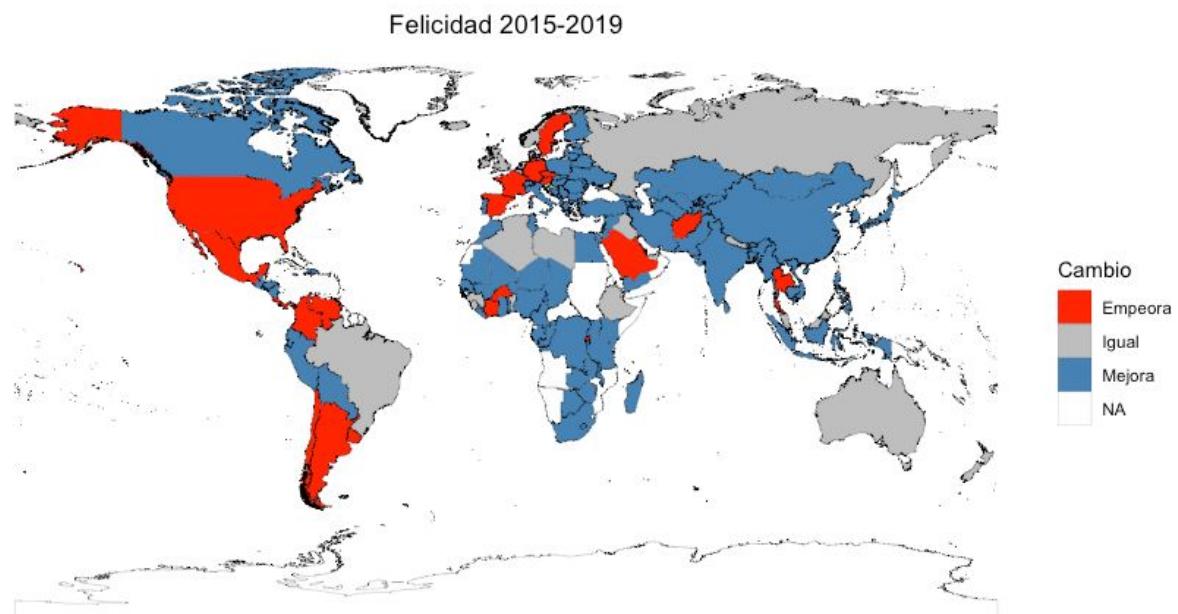
Buscar en R: # MAPA EVOLUCIÓN

Funciones: `fortify()` y `ggplot()`

Para ver más claramente el cambio, hemos hecho un mapa donde se ve si la puntuación -score- ha mejorado o empeorado entre 2015 y 2019. Para eso, juntamos ambas puntuaciones en un dataframe nuevo y las restamos. Convertimos el resultado en una variable discreta dividida en 3:

- Mejora, si la diferencia es mayor que 0,05
- Empeora, si la diferencia es menor que -0,05
- Igual, si la diferencia está entre -0,05 y 0,05

Decidimos dividirlo así y no a partir de 0 porque creemos que una diferencia tan pequeña no es un cambio suficiente como para señalarlo, así que lo marcamos como si no hubiera cambiado.

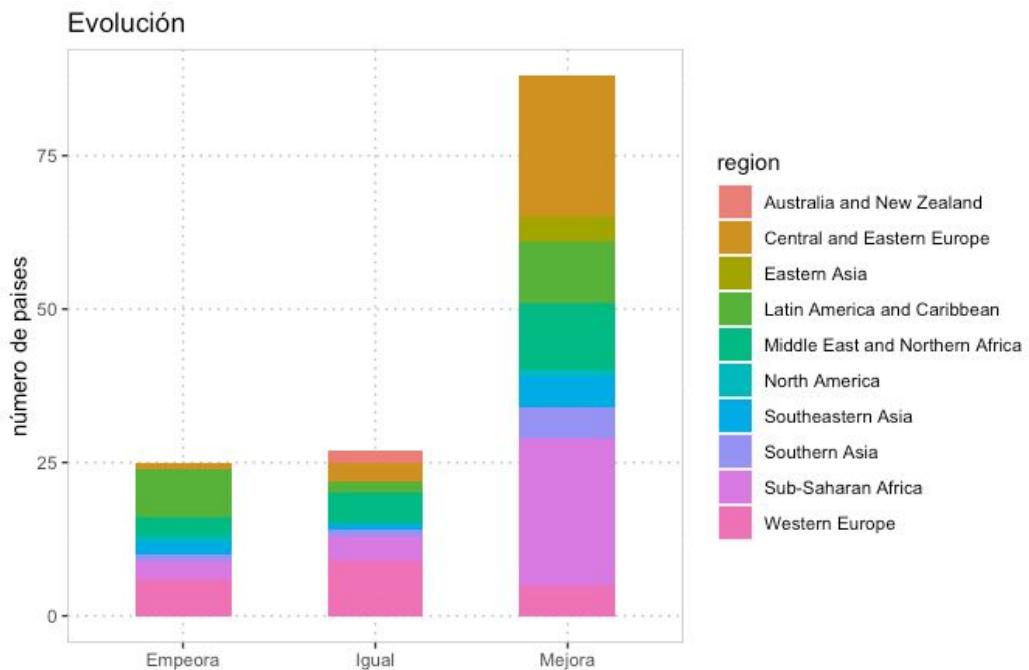


Como se puede observar a primera vista, hay muchos más países que mejoran que países que empeoran con respecto a 2015. Veremos esto con más profundidad en los siguientes gráficos.

Buscar en R: # GRÁFICOS EVOLUCIÓN

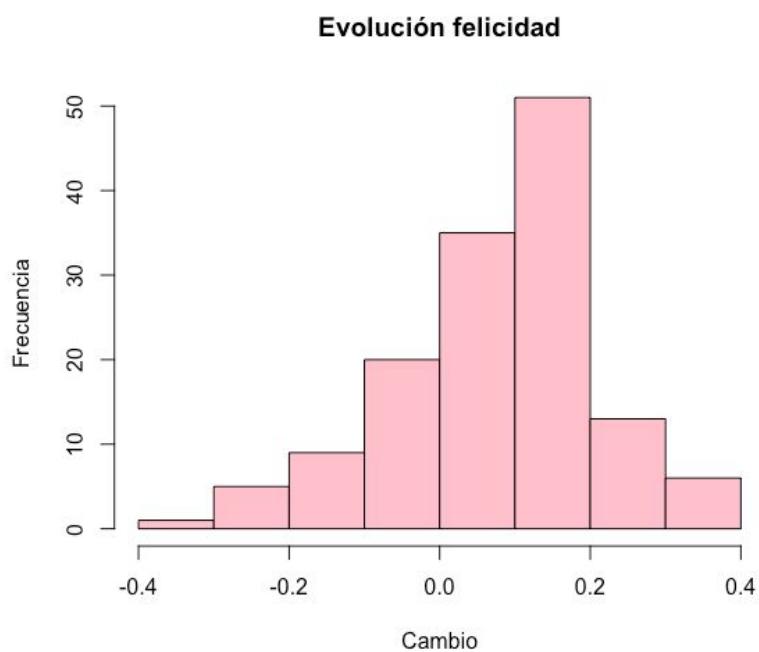
Funciones: `ggplot()`, `hist()` y `barplot()`

Número de países que mejoran y empeoran por región



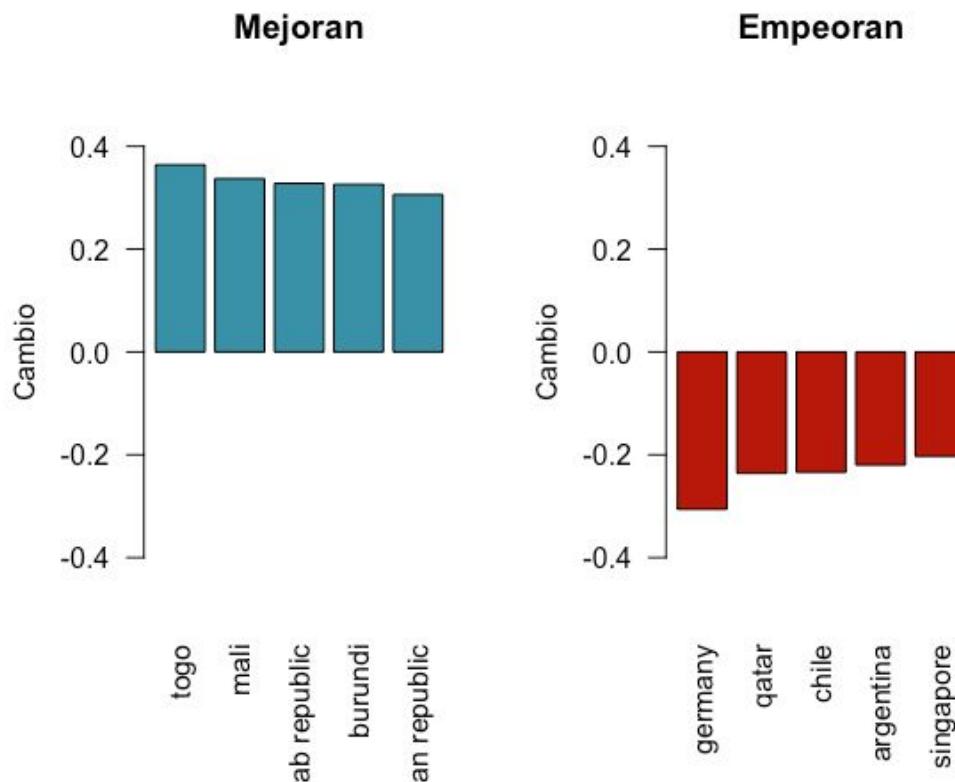
En este gráfico de barras podemos ver cómo el número de países que mejora su puntuación con respecto a 2015 es más grande que la suma de los que empeoran o se mantienen igual. Además, se ve cómo la mayor parte de países que mejoran son de Europa Central y del Este, y del África Subsahariana.

Frecuencia de la evolución



En este histograma se observa la frecuencia del cambio en la puntuación total de felicidad en los países entre 2015 y 2019, donde queda claro que la mayor parte de estos mejoran entre 0,0 y 0,2, sumando entre estos dos casi 90 países.

Países que más mejoran y empeoran entre 2015-2019



En estos dos gráficos podemos apreciar qué países han mejorado y empeorado más entre 2015 y 2019. Entre los que más han mejorado están Togo, Mali, Siria, Burundi y la República Dominicana. Viendo la clasificación de 2015, 3 de estos países, Togo, Siria y Burundi, ocupaban los puestos más bajos del ranking. En cuanto a los países que más empeoran, tenemos Alemania, Catar, Chile, Argentina y Singapur. Al contrario que los países que más mejoran, estos no estaban en los puestos más altos de la tabla, sino que curiosamente, se encontraban todos entre los puestos 24 y 30.

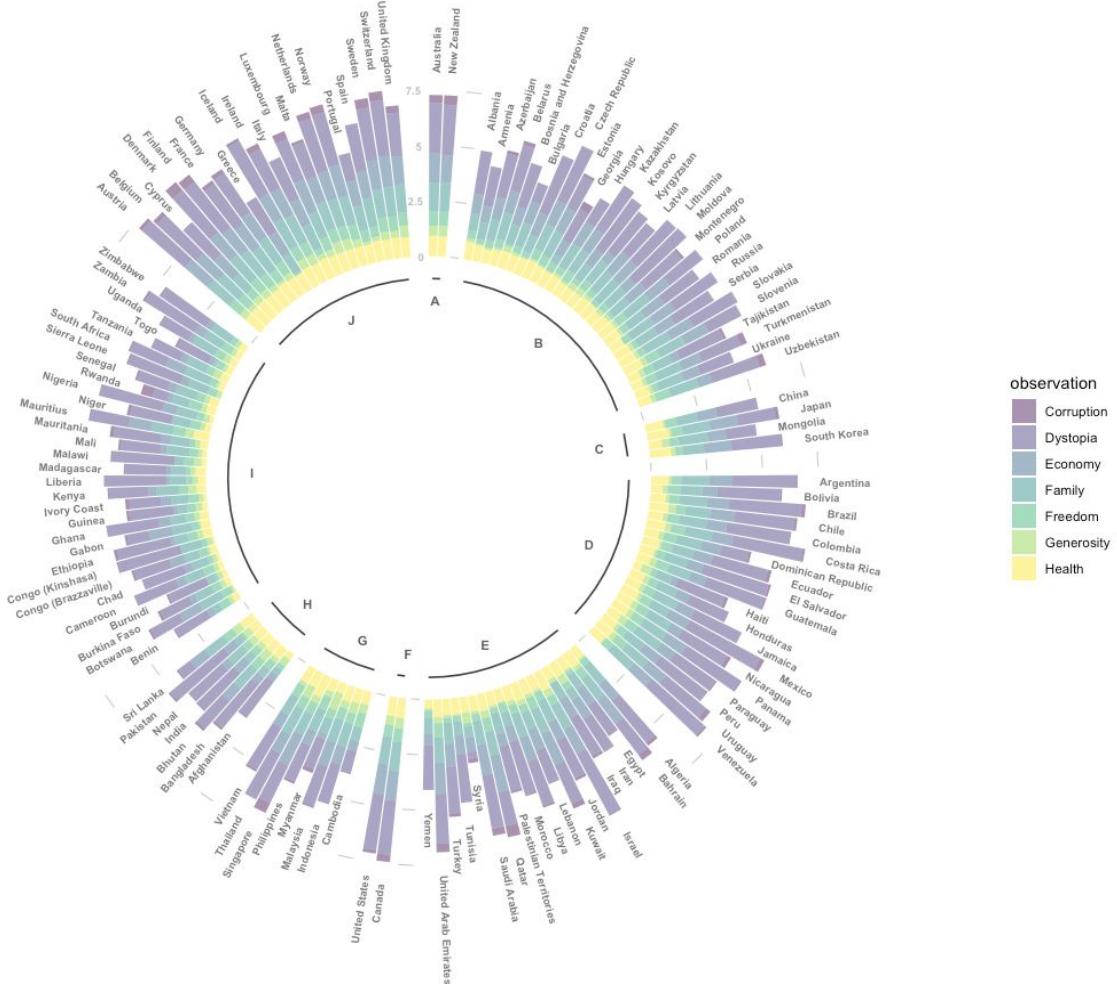
Buscar en R: # STACKED BARPLOT TODOS

Función: `ggplot()`

A fin de variar un poco nuestros modelos y en busca de inspiración, conseguimos sacar los siguientes gráficos de esta página: <https://www.r-graph-gallery.com/index.html>

Leyenda:

A	Australia and New Zealand	F	North America
B	Central and Eastern Europe	G	Southeastern Asia
C	Eastern Asia	H	Southern Asia
D	Latin America and Caribbean	I	Sub-Saharan Africa
E	Middle East and Northern Africa	J	Western Europe



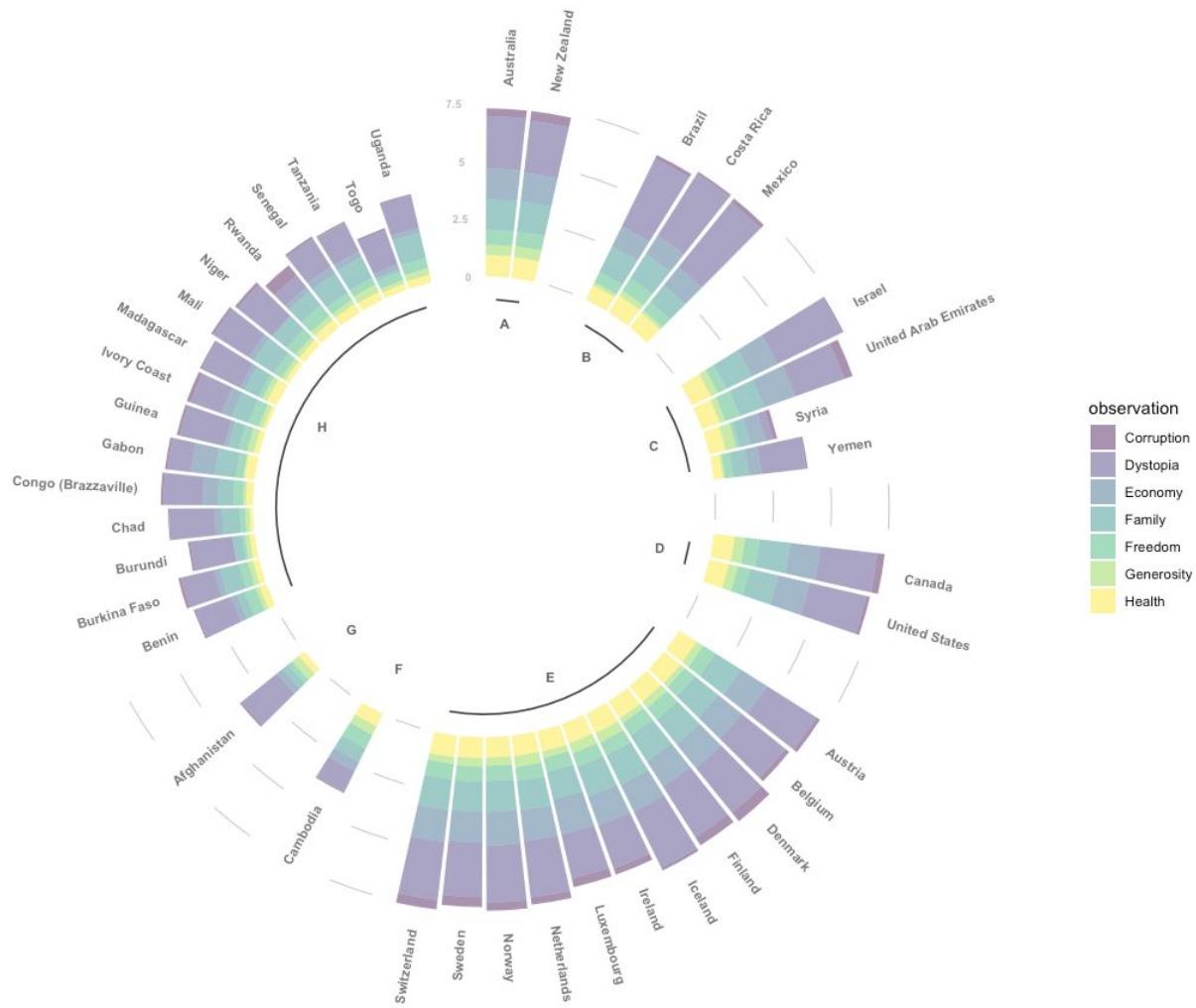
Inicialmente queríamos comprobar qué se distinguía o si éramos capaces siquiera de distinguir algo que incluyese la totalidad de países, con las variables que los afectan sobre el total de su Happiness Score. No siendo el caso, decidimos a continuación reducirlo a los veinte países más felices y los veinte menos felices.

Buscar en R: # STACKED BARPLOT TOP

Función: `ggplot()`

Leyenda:

A	Australia and New Zealand	E	Western Europe
B	Latin America and Caribbean	F	Southeastern Asia
C	Middle East and Northern Africa	G	Southern Asia
D	North America	H	Sub-Saharan Africa



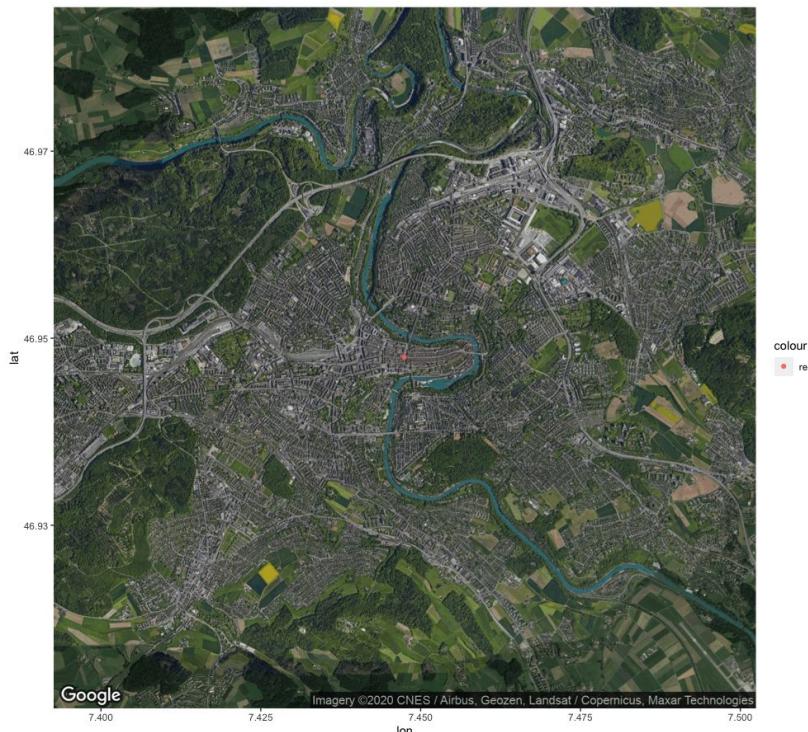
Algo que destaca, además de todo lo dicho anteriormente, es que aunque en la mayor parte de casos todos los países de una misma región tengan puntuaciones similares, cabe destacar el caso de Oriente Medio y Norte de África, donde tenemos dos países en el top 20 –Israel y EAU–, y dos países entre los últimos del ranking –Siria y Yemen–.

Geoposicionamiento

Buscar en R: # API GOOGLE

Funciones: `tryCatch()`, `get_map()`, `ggmap()` y `plot()`

Decidimos utilizar la API de Google para geoposicionar el primer y el último país del ranking. El primero es Suiza, en 2015, por lo que sacamos una foto de su capital, Berna.



En último lugar del ranking se encuentra Togo, tenemos foto de su capital, Lomé.



Conclusiones

En resumidas cuentas, podemos decir que nuestro estudio parecería señalar que las variables de *health* y *economy* sostienen un peso importante sobre el *happiness score*, pero podría haber otros factores que no estamos teniendo en cuenta en nuestro estudio.

Observando la evolución de los países en los últimos años, queda claro que, en general, mejoran su puntuación, especialmente algunas regiones como África Subsahariana, aunque todavía se mantiene muy por debajo del resto. En efecto, las regiones muestran ciertas similitudes en sus puntuaciones, a excepción de Oriente Medio y Norte de África, donde hay grandes diferencias entre ellos. A pesar de todo, este hecho podría indicar que el Happiness Score de nuestros países tal vez dependiese de su localización.

Por tanto, a grandes rasgos, aquellas regiones o países con un Happiness Score relativamente bajo, podrían ver su puntuación aumentar durante los próximos años como consecuencia del desarrollo económico del país. Sin embargo, aquellos países ya desarrollados o dotados de cierta estabilidad económica, pero interdependientes, podrían ver su felicidad más fácilmente mermada por shocks externos, tales como una crisis económica o los efectos del populismo en la democracia, entre otros.