DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico III

Cuadrados Mínimos Lineales

Métodos Numéricos

Segundo Cuatrimestre de 2021

Integrante	LU	Correo electrónico
Alliani Federico	[redacted]	[redacted]
Nores Manuel	[redacted]	[redacted]
Raposeiras Lucas	[redacted]	[redacted]
Oca Mariano	[redacted]	[redacted]

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2160 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (54 11) 4576-3359

http://www.fcen.uba.ar

Resumen

El objetivo de este Trabajo Práctico es analizar los datos públicos de la Organización Mundial de la Salud con el objetivo de entender y formular relaciones teóricas entre la expectativa de vida de los países y sus características. El método central que se utilizará para elaborar dichas conjeturas es el de **Cuadrados Mínimos Lineales**.

Palabras Clave Cuadrados Mínimos Lineales, Expectativa de Vida, Correlación, OMS.

Índice

1. Introducción

A nivel individual, entender la expectativa de vida resulta vital para los humanos que como especie heredamos evolutivamente un inherente miedo a la muerte. El análisis estadístico de la expectativa de vida y el interés que esta genera podría interpretarse como una nueva forma de racionalizar el fin de nuestra existencia como individuos.

A nivel global y macroeconómico la expectativa de vida impulsa políticas tanto públicas como privadas. Un ejemplo de esto es la inversión en salud por parte de los Estados que permitirían mejorar la salud de individuos de bajos recursos (y por ende, el aumento del promedio de expectativa de vida de esta población) o la crisis que se espera llegar al sistema de pensiones mundial en los próximos años: a nivel global la expectativa de vida aumentó mientras que se redujo la taza nacimientos, provocando una pirámide demográfica invertida. Esto sentenciaría que cuando se jubilen los trabajadores actuales no va a alcanzar la población económicamente activa para sostenerlos. Especialistas aseguran que el actual modelo es insostenible aunque se aumentara la edad jubilatoria [1]. Notar que esta problemática es inherente a la expectativa de vida debido a que es mucho más costoso mantener a quien se jubila a los 60 y vive hasta los 100 que a quien se jubila a los 60 y vive hasta los 80. La tendencia mundial apunta al primer caso.

Sospechando una relación entre la expectativa de vida y la calidad de vida (sea cierta o no), esto podría tomar importancia en asuntos como la inmigración, siendo países con mayor expectativa de vida más atractivos para los inmigrantes. También se alega que una alta expectativa de vida es un problema para el desarrollo y crecimiento económico, dando el ejemplo de países europeos como Italia de cuyo estancamiento se lo culpa a su pirámide demográfica invertida y la falta de personas económicamente activas [2].

Con tantas variables involucradas y tantos intereses superpuestos resulta de particular importancia el estudio de la expectativa de vida y su relación con dichas variables.

1.1. Objetivo

El objetivo de este trabajo es entender la relación entre las características de un país y su expectativa de vida.

Para el abordaje de este problema mediremos la correlación entre datos de público acceso proporcionados por la Organización Mundial de la Salud.

1.2. Primeras consideraciones

Lo primero a considerar es que si dos variables están relacionadas de alguna manera no implica necesariamente que haya causalidad.

Para permitir evidenciar esto, traemos los siguientes ejemplos triviales donde se ve claramente que una variable, por más que parezca estar relacionada con otra, no necesariamente lo están.

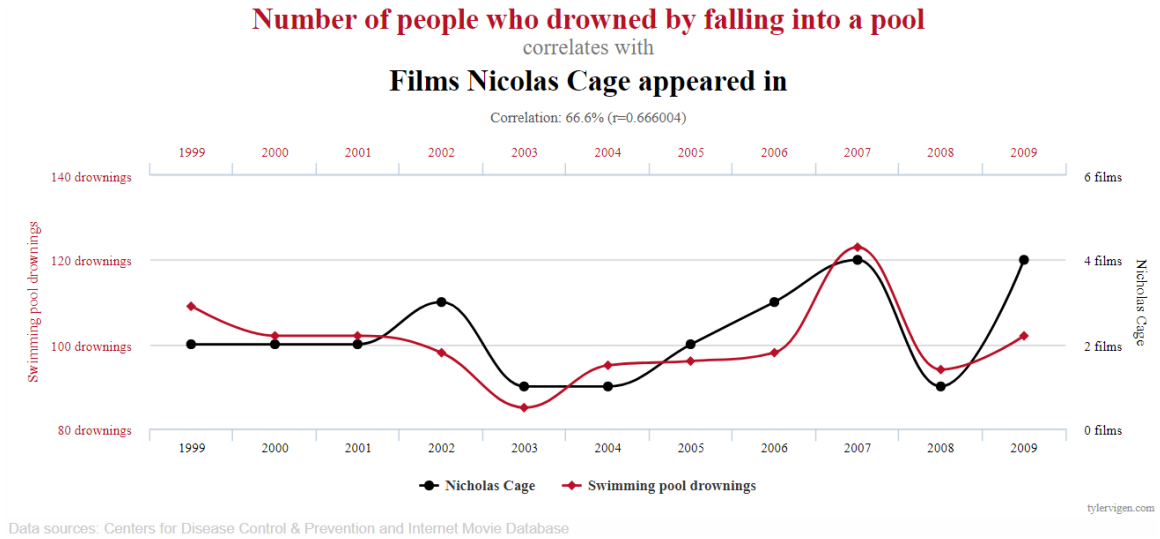


Figura 1: Correlación entre películas en las que aparece Nicolas Cage y cantidad de personas que se ahogaron por caer a una pileta [7].

En la Figura 1 se puede observar una correlación del `66,6%` entre la cantidad de películas en las que aparece

Nicolas Cage y la cantidad de personas que se ahogaron por caer a una pileta.

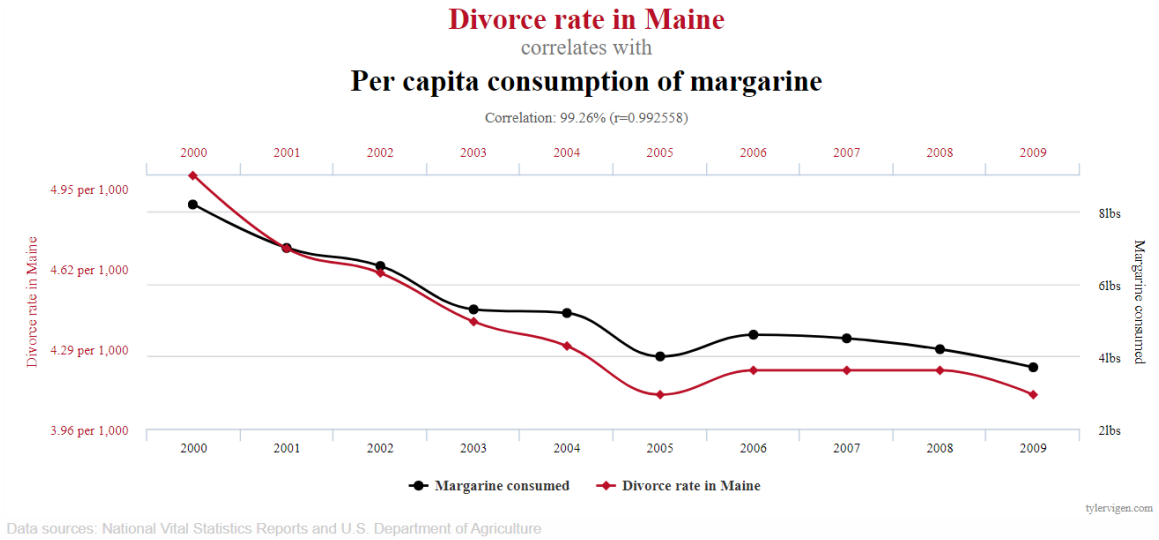


Figura 2: Correlación entre consumo per cápita de margarina y ratio de divorcios en Maine [7].

En la Figura 2 se puede apreciar una impresionante correlación del $99,26\,\%$ entre el consumo per cápita de margarina y el ratio de divorcios en el Estado de Maine.

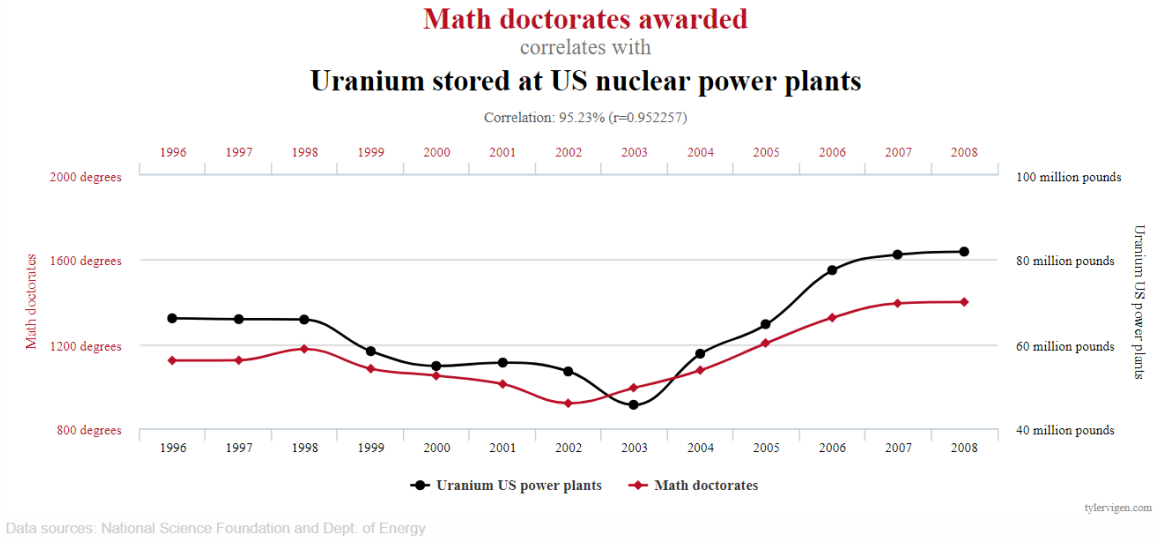


Figura 3: Correlación entre doctorados de matemática y uranio almacenado en plantas nucleares en USA [7].

Finalmente, en honor al departamento de matemática, en la Figura 3 podemos apreciar una correlación del $95,23\,\%$ entre la cantidad de doctorados en matemática y la cantidad de uranio en las plantas nucleares en los Estados Unidos. Adicionalmente invitamos al lector a visitar el sitio de Spurious Correlations [7] y formular sus propias teorías.

Queda entonces evidenciado que por más que encontremos correlaciones extremadamente altas entre las características de los países con su expectativa de vida, estas servirán únicamente para un análisis teórico y expositivo debido a que no pueden considerarse causales una de otras sin un análisis particular y puntual, el cual no es el objetivo de este trabajo práctico.

2. Análisis Exploratorio de los Datos

2.1. El dataset

El dataset utilizado contiene un promedio de los datos de distintos países entre los años 2000 y 2015, cuya fuente proviene del sitio web oficial de la **Organización Mundial de la Salud (WHO)** [8]. Dicho dataset lo provee la cátedra, ya que el mismo ha sido preprocesado a fin de facilitar la lectura y el entendimiento de los datos en sí. El dataset en cuestión está en formato `csv`, lo que facilita su procesamiento en bibliotecas como *Pandas*.

El dataset contiene una fila por país y diecinueve indicadores, entre los cuales se encuentra el indicador de la *Calidad de vida* (*Life expectancy*), que será el principal sujeto de estudio en este trabajo. Se cuenta con la información de 183 países.

En la siguiente sección exhibiremos un análisis de algunos de estos *features*.

2.2. Lista de features

Feature
Estado
Esperanza de vida
Mortalidad de adultos
Mortalidad infantil
Alcohol
Gasto porcentual en salud
Hepatitis B
Sarampión
Índice de masa corporal
Polio
Gasto total
Difteria
VIH
PBI
Población
Delgadez de 1 a 19 años
Delgadez de 5 a 9 años
Índice de desarrollo humano
Escolaridad

2.3. Análisis de los features

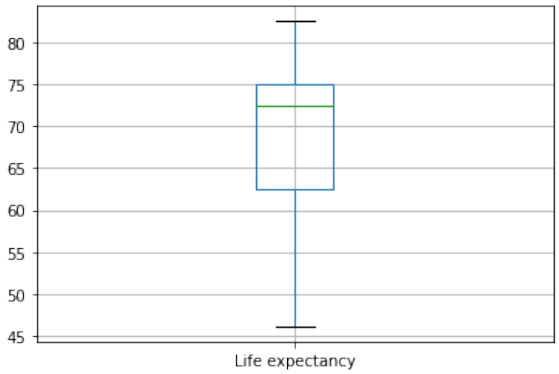
Atención Debido a que no todos los países generan estos indicadores de la misma forma, hay algunos países en donde no se tienen algunos datos.
Por ejemplo, el indicador correspondiente a la cantidad de habitantes de Antigua y Barbuda está vacío.

2.3.1. Expectativa de vida

La expectativa de vida es el número promedio de años que un recién nacido podría esperar vivir, si pasara por la vida expuesto a las tasas de muerte específicas por sexo y edad que prevalecen en el momento de su nacimiento, durante un año específico, en un período determinado, en un determinado país, territorio o área geográfica.

La expectativa de vida es el principal indicador para nuestro análisis. Compararemos los datos de esta variable con el resto de los indicadores para formar hipótesis que luego revisaremos en la sección de experimentación.

A continuación analizaremos algunos datos interesantes sobre la expectativa de vida.



	Life expectancy
count	183.000000
mean	69.224932
std	9.186432
min	46.112500
25%	62.475000
50%	72.487500
75%	75.106250
max	82.537500

Figura 4: Estadísticas del indicador de Expectativa de vida

En la figura 4 se puede observar que el promedio de la expectativa de vida es de 69 años, sin embargo existen países como *Sierra Leona*, que tienen una expectativa de vida de 46 años y países como *Japón* que tienen una expectativa de vida de 82 años. Considerando estos casos extremos podemos evidenciar que hay una diferencia de 1.78 veces entre el país con menos expectativa de vida y el país con más expectativa de vida. Tendremos estos dos países presentes durante el análisis exploratorio de algunos features.

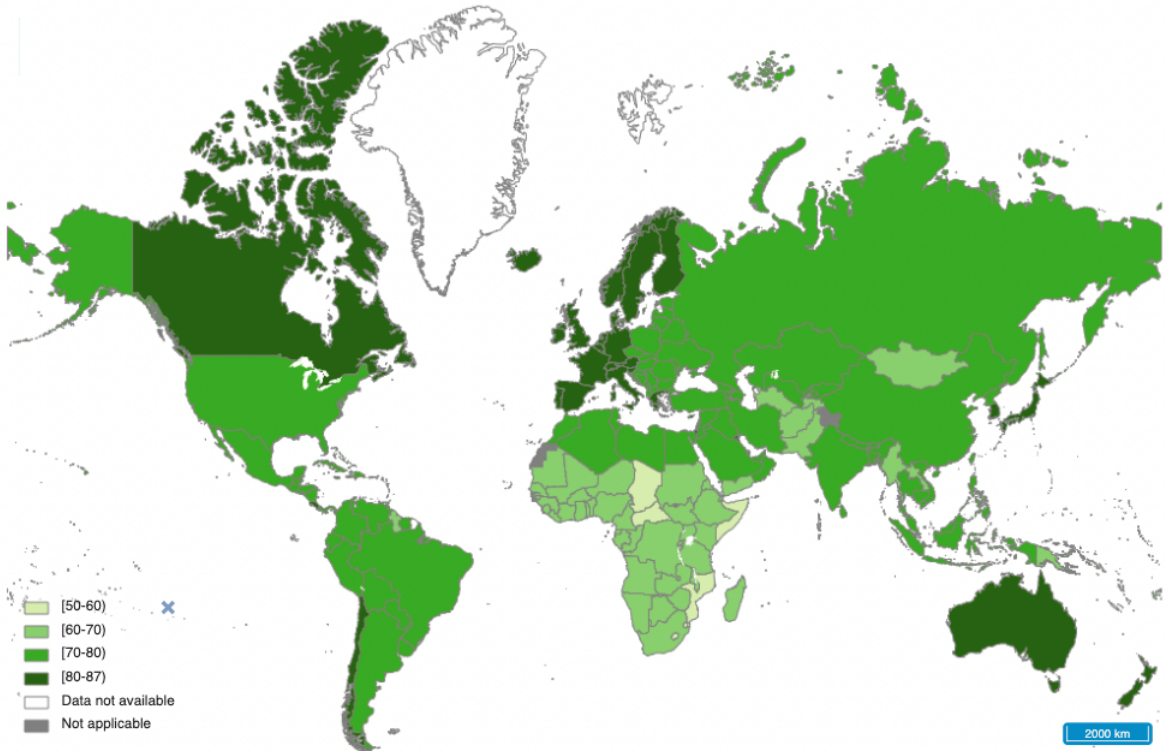
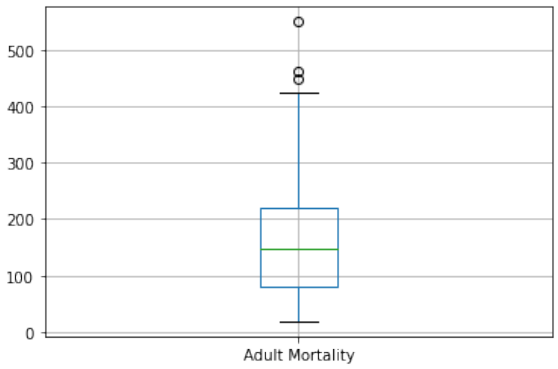


Figura 5: Mapa de la expectativa de vida de cada país en años

La figura 5 muestra un mapa de la expectativa de vida de cada país en años. En ella podemos ver que hay países con alta expectativa de vida como Canadá, Australia, Nueva Zelanda y varios países de Europa y Chile representando a América Latina.

2.3.2. Mortalidad adulta

Siguiendo en linea con la expectativa de vida, analicemos los datos de la *mortalidad adulta*. La mortalidad adulta es la probabilidad de que una persona de 15 años de edad o más muera antes de cumplir 60 años (cada 1000 habitantes). Sin embargo, analizando los datos, entendemos que lo que realmente se está midiendo es la cantidad de adultos muertos entre 15 y 60 años cada 1000 personas.



	Adult Mortality
count	183.000000
mean	164.796448
std	96.185159
min	18.750000
25%	80.406250
50%	146.562500
75%	219.593750
max	550.062500

Figura 6: Estadísticas del indicador de Mortalidad adulta

En la figura 6 podemos ver que el indicador está medianamente concentrado en 150 muertes cada 1000 personas y que hay algunos *outliers* principalmente en la parte superior del gráfico. Si analizamos estos *outliers* podemos ver que son países como Lesoto (`550.06`), Zimbabue (`462.37`) y Botsuana (`448.125`). También podemos ver que hay un outlier en la parte inferior del boxplot, ya que si ordenamos los datos de menor a mayor vemos una gran diferencia entre el país con menos mortalidad adulta y el segundo país con menos mortalidad adulta. Dicha diferencia es de `2.45` veces. Investigando un poco más en profundidad y consultando otras fuentes llegamos a la conclusión de que este valor puede ser erróneo. Mirando las figuras 7 y 8 del sitio web del Banco de Datos Libre [3] nos muestran que nunca tuvo un valor inferior a 70. Esto nos hace sospechar que es un valor incorrecto.

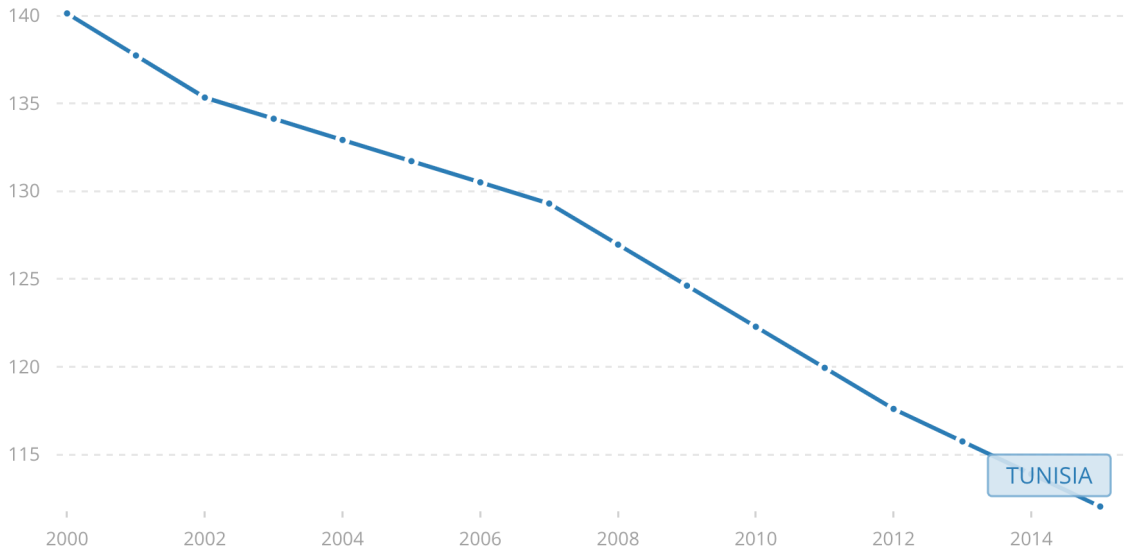


Figura 7: Mortalidad adulta masculina en Túnez

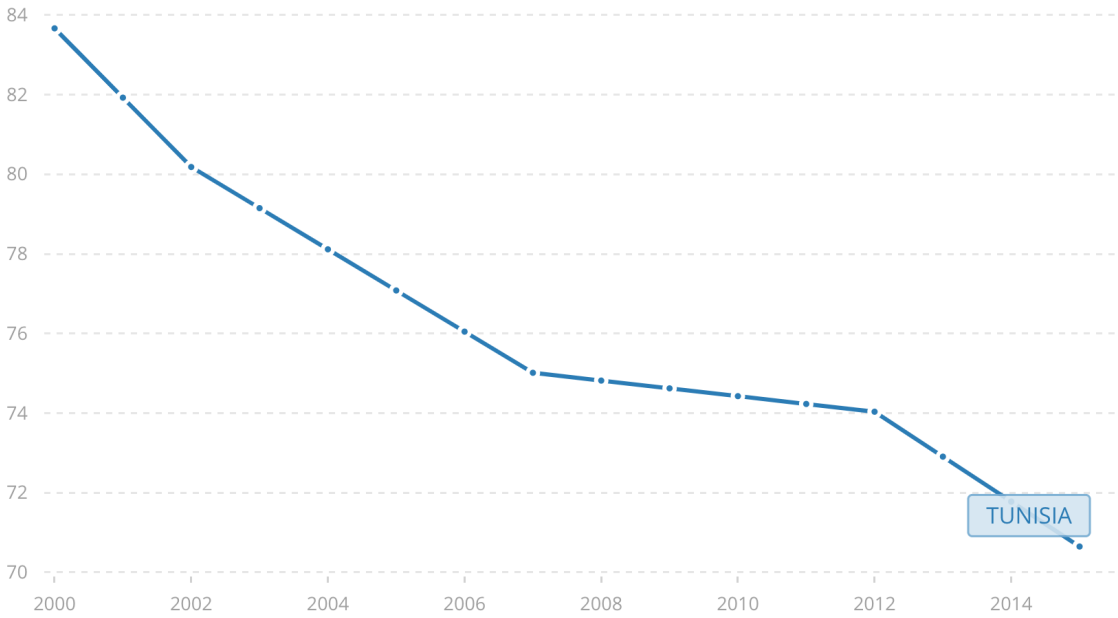


Figura 8: Mortalidad adulta femenina en Túnez

Por otra parte, analicemos las posiciones de Japón y Sierra Leona que, como vimos anteriormente, son los países con mayor y menor expectativa de vida respectivamente. En la figura 9 podemos ver a Sierra Leona en el puesto número 7 (ordenando de mayor a menor) con `357,81` muertes adultas cada 1000 habitantes, y a Japón en la posición 8 (ordenando de menor a mayor) con un valor de mortalidad adulta de `57,12`. Parece que hay alguna relación inversa entre estas dos variables ya que si sacamos al posible error de Túnez tenemos a ambos países en la posición 7 ordenando de mayor a menor y de menor a mayor.

Country	Life expectancy	Adult Mortality
Tunisia	74.35625	18.7500
Albania	75.15625	45.0625
Iceland	82.44375	49.3750
Saudi Arabia	73.46875	52.1250
Cyprus	79.67500	54.1250
Italy	82.18750	54.1875
Switzerland	82.33125	55.7500
Japan	82.53750	57.1250
Cuba	77.97500	57.5625
United States of America	78.06250	58.1875
...	...	...
South Sudan	53.87500	346.3125
Kenya	57.48125	348.5625
Zambia	53.90625	354.3125
Sierra Leone	46.11250	357.8125
South Africa	57.50000	412.7500
Côte d'Ivoire	50.38750	417.3125
Malawi	49.89375	424.4375
Botswana	56.05000	448.1250
Zimbabwe	50.48750	462.3750
Lesotho	48.78125	550.0625

Figura 9: Tabla de países ordenada según la mortalidad adulta de menor a mayor

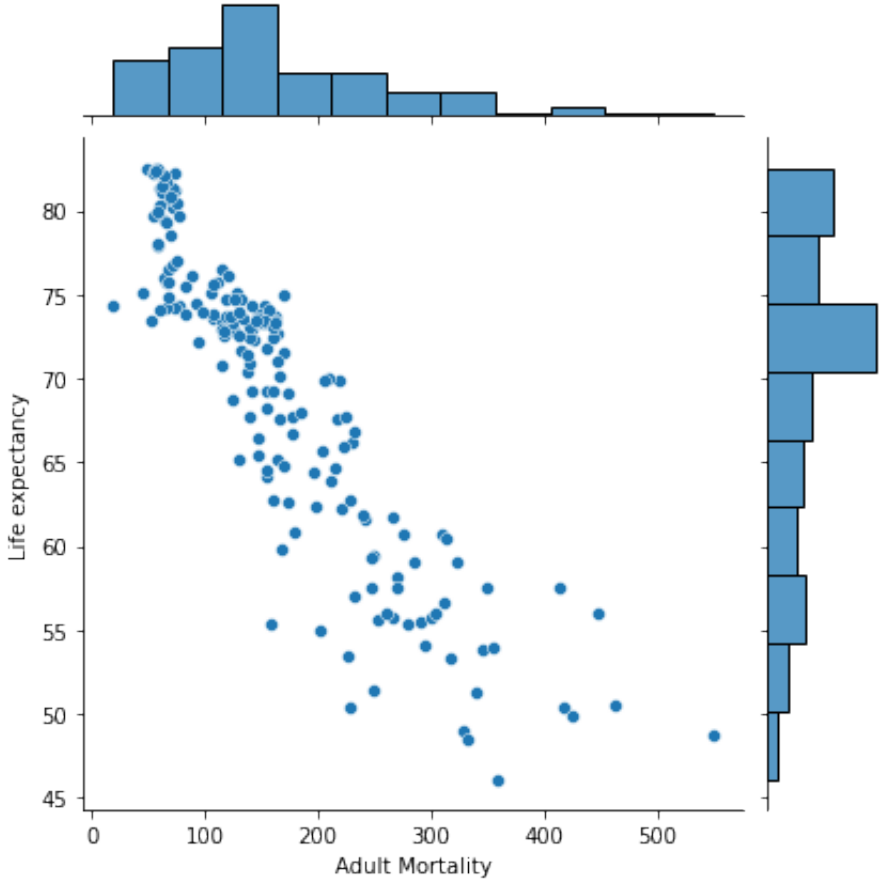


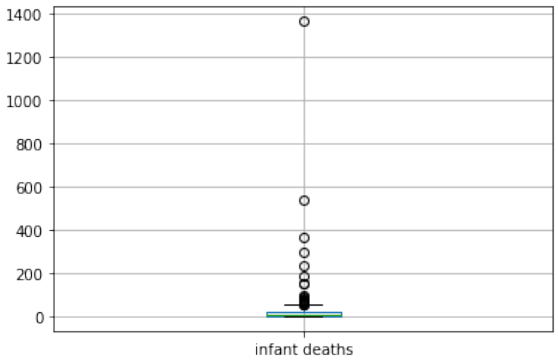
Figura 10: Gráfico de la expectativa de vida de cada país según la mortalidad adulta

En la figura 10 podemos ver el gráfico de la expectativa de vida de cada país según la mortalidad adulta. Observando detenidamente podemos ver que efectivamente parece haber una cierta relación inversa entre estos dos indicadores.

2.3.3. Mortalidad infantil

Luego de analizar el indicador de mortalidad en adultos, vamos a analizar la mortalidad infantil. La tasa de mortalidad infantil es la probabilidad de que un niño nacido en un año o período específico muera antes de cumplir un año. La tasa de mortalidad infantil, estrictamente hablando, no es una tasa (es decir, el número de muertes dividido por el número de población en riesgo durante un cierto período de tiempo), sino una probabilidad de muerte derivada de una tabla de vida y expresada como tasa por 1000 nacidos vivos.

Analizando la figura 11 podemos ver a simple vista que el dataset provisto tiene el valor máximo incorrecto. El valor máximo es 1366 que corresponde al país India. Sin embargo, este valor es incorrecto ya que la métrica nos indica la cantidad de muertes de infantes cada 1000 personas y este valor supera el número 1000, lo cual es imposible, por lo tanto incorrecto.

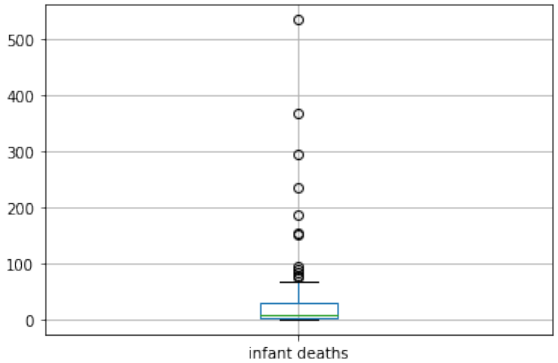


	infant deaths
count	183.000000
mean	30.407445
std	115.995720
min	0.000000
25%	0.000000
50%	2.937500
75%	22.125000
max	1366.687500

Figura 11: Estadísticas del indicador de Mortalidad infantil

Por otro lado, también podemos ver que tanto el mínimo como el 25-percentil es 0, esto es así debido a que hay 47 países que tienen un valor de 0 muertes de infantes cada 1000 personas. Sospechamos que este valor indica que no se conoce la métrica en estos países, ya que es virtualmente imposible que la tasa de mortalidad infantil de un país sea 0.

Luego de eliminar los países que tienen el valor 0 y a India que tiene un valor incorrecto, podemos ver en la figura 12 que siguen existiendo varios *outliers*, pero esta vez los valores parecen ser correctos, ya que están en un rango coherente. Por ejemplo, Nigeria, con 535 muertes infantiles cada 1000, Pakistán con 367 y China con 294 son algunos de estos *outliers*. Estos tres países están muy alejados de la media y del 75-percentil.



	infant deaths
count	135.000000
mean	31.095370
std	68.001658
min	0.187500
25%	2.062500
50%	8.000000
75%	30.000000
max	535.687500

Figura 12: Estadísticas del indicador de Mortalidad infantil con los valores incorrectos filtrados

Por otra parte, analicemos las posiciones de Japón y Sierra Leona que, como vimos anteriormente, son los países con mayor y menor expectativa de vida respectivamente. Vamos analizar las posiciones luego de hacer la limpieza del dataset. A Sierra Leona lo podemos encontrar en el puesto numero 35 (ordenando de mayor a menor) con un valor de 27.56 y a Japón en la posición 44 con un valor de 2.87. Nuevamente parece que hay alguna relación inversa entre estas dos variables ya que tenemos a ambos países casi a la misma distancia del máximo y mínimo respectivamente.

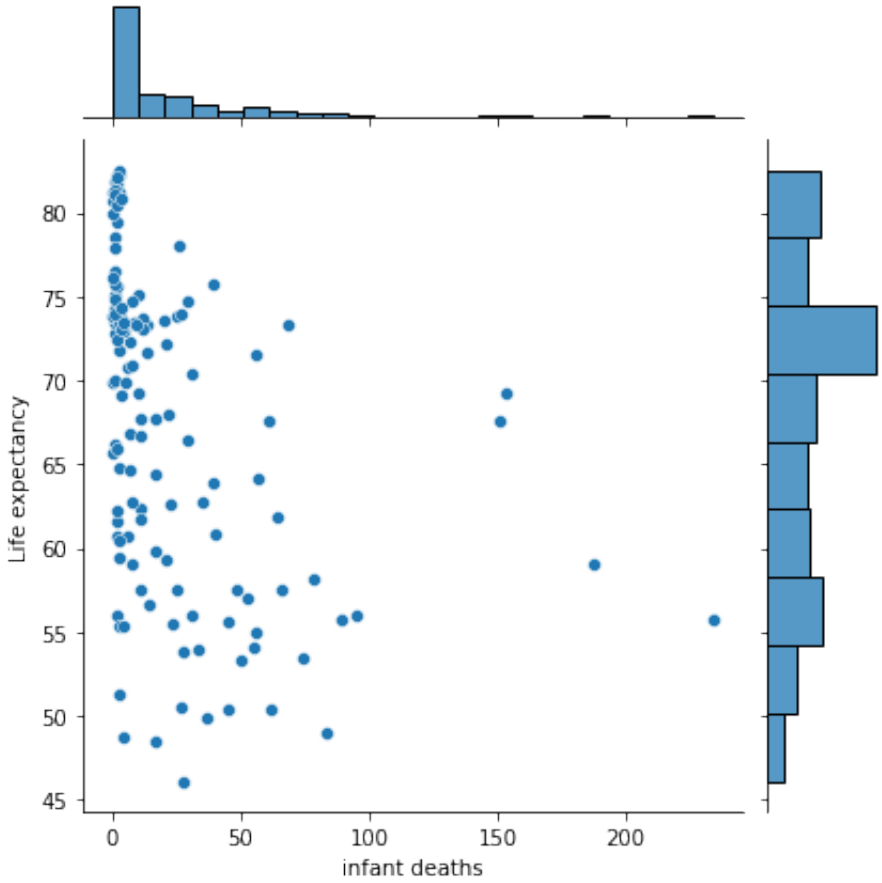


Figura 13: Gráfico de la expectativa de vida de cada país según la mortalidad infantil

Por ultimo, podemos ver la figura 13 que nos muestra un gráfico de la expectativa de vida de cada país según la mortalidad infantil. Observando dicho gráfico, parece que los datos tienen una distribución bastante uniforme sobre el eje Y. Esto parecería decir que no hay correlación entre las variables, pero esta forma de analizar los datos no es lo suficientemente fuerte como para confirmar que no hay correlación.

2.3.4. Consumo de alcohol per cápita

El consumo de alcohol per cápita registrado se define como la cantidad registrada de alcohol puro en litros consumido per cápita (más de 15 años) durante un año calendario en un país, en litros de alcohol puro. El indicador sólo tiene en cuenta el consumo que se registra a partir de los datos de producción, importación, exportación y ventas.

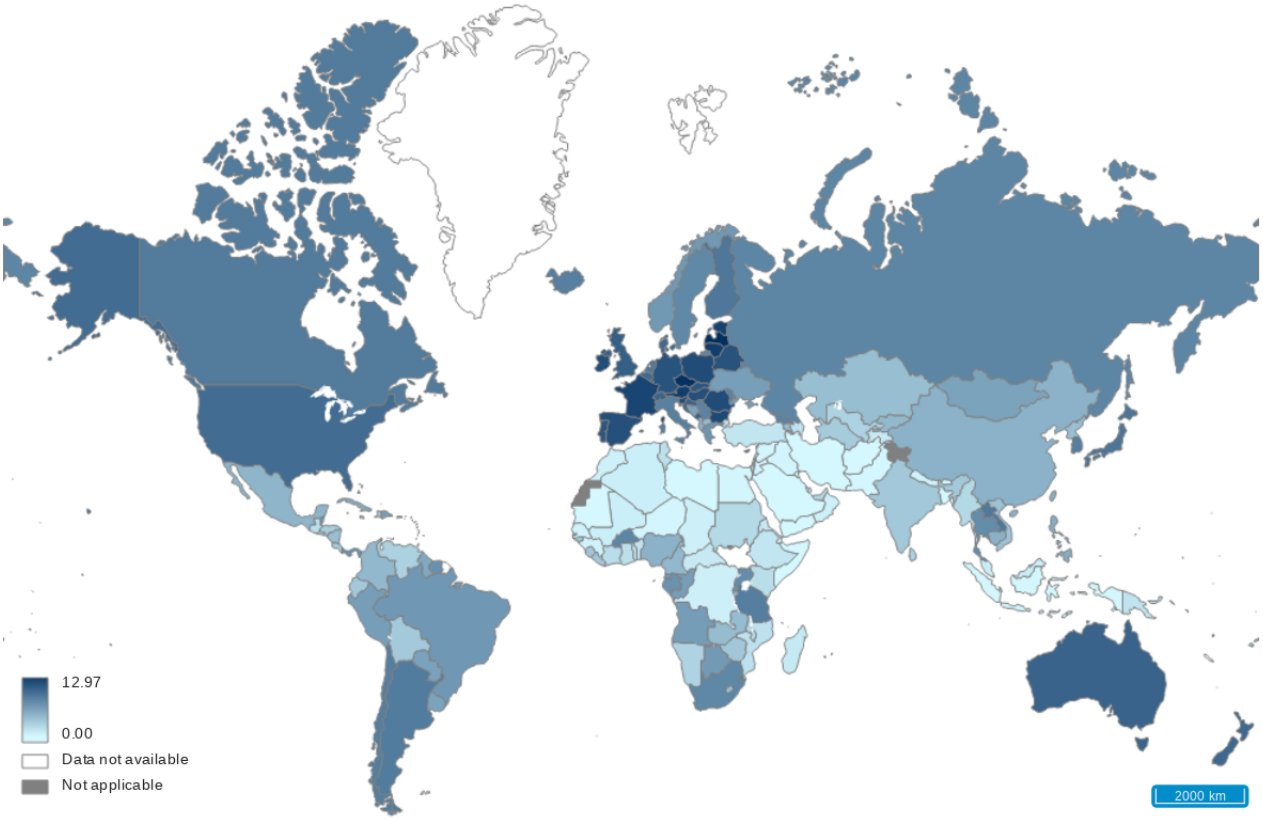
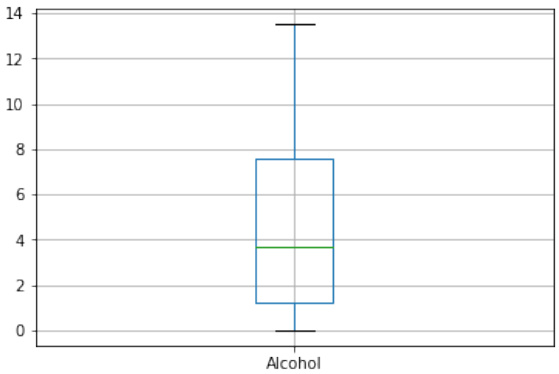


Figura 14: Mapa del consumo de alcohol per cápita registrado (15+) (en litros de alcohol puro) en el último año



	Alcohol
count	182.000000
mean	4.612109
std	3.848417
min	0.010000
25%	1.240000
50%	3.663333
75%	7.558000
max	13.497333

Figura 15: Estadísticas del indicador de Consumo de alcohol per cápita

En la figura 15 podemos ver algunas medidas del indicador de consumo de alcohol per cápita. Este indicador parece no tener *outliers*. También notamos que falta el dato de consumo de alcohol de un país. Revisando el dataset vemos que faltan los datos de Sudán del Sur. También tenemos en consideración que en países como Bangladesh, Libia y Somalia (todos con un consumo de alcohol registrado de 0,01 según el dataset), entre otros, el consumo de

alcohol está prohibido por lo que la mayoría de este se da de forma clandestina, por lo que no impacta en los registros oficiales [4].

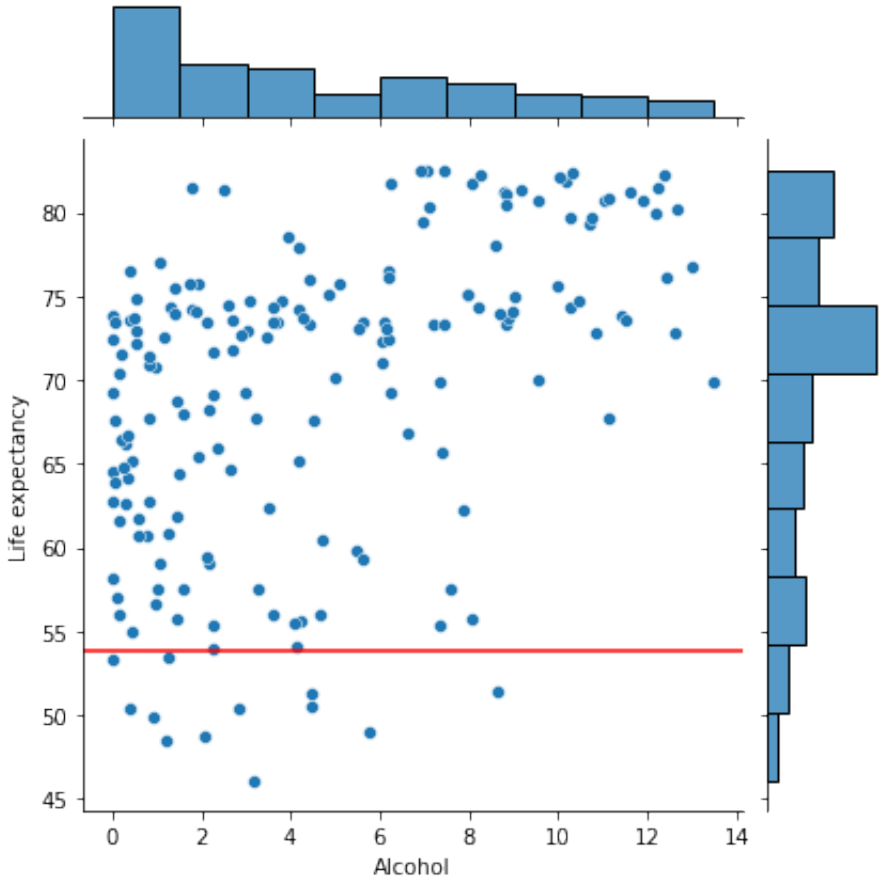


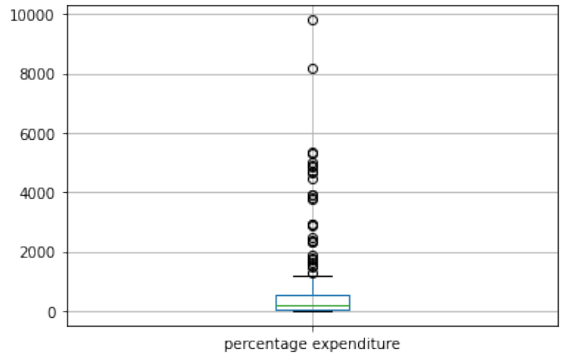
Figura 16: Gráfico de la expectativa de vida de cada país según el consumo de alcohol per cápita

En la figura 16 podemos ver un gráfico de la expectativa de vida de cada país según el consumo de alcohol per cápita. La línea roja horizontal representa la expectativa de vida de *Sudán del Sur*. Decidimos que, ya que no tenemos el dato de consumo de alcohol de dicho país, al menos podemos tener un vistazo de en qué posición se ubicaría en este gráfico. Volviendo al gráfico en si, a simple vista no parece que exista una relación directa entre el consumo de alcohol per cápita y la esperanza de vida. Aunque no haya relación directa es llamativo que en los países donde se consume mucho alcohol la esperanza de vida es alta y donde la esperanza de vida es baja el consumo de alcohol varia entre cantidades relativamente bajas.

2.3.5. Porcentaje de gasto en salud

En esta sección analizaremos el porcentaje de gasto en salud. Este indicador se define como el gasto total en salud como porcentaje del Producto Interno Bruto (PIB), es decir, es el porcentaje del gasto total del gobierno que se gasta en salud.

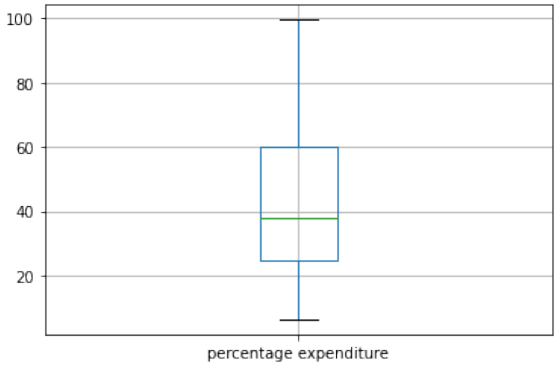
Analizando las figura 17 podemos ver a simple vista que el dataset provisto tiene bastantes valores incorrectos. El indicador debería medir el porcentaje gastado en salud sobre el porcentaje total de gasto del gobierno, sin embargo hay muchísimos valores que son superiores a 100. Un claro ejemplo es el máximo, que tiene un valor de 9801, que corresponde al país Suiza.



	percentage expenditure
count	183.000000
mean	740.321185
std	1480.930026
min	0.000000
25%	32.401540
50%	183.662908
75%	523.622686
max	9801.810377

Figura 17: Estadísticas del indicador de porcentaje de salud gastado

Luego de quitar los valores que son iguales a 0 y los valores que son mayores a 100 , obtuvimos como resultado la figura 18. El cual ahora tiene valores un poco más razonables que antes, ya que todos los valores se encuentran entre 0 y 100 como un porcentaje. Decimos un poco más razonables ya que aún tenemos sospechas de que haya países como Georgia, Djibouti y Sudán que inviertan arriba del 95 % de su presupuesto en salud.



	percentage expenditure
count	52.000000
mean	44.824009
std	26.029784
min	6.436069
25%	24.880181
50%	38.166819
75%	60.229067
max	99.525714

Figura 18: Estadísticas del indicador de porcentaje de salud gastado luego de sacar los valores incorrectos

Luego de investigar con otra fuente (WorldBank) pudimos concluir que ni Georgia ha invertido tanto presupuesto en Salud, ni que la media ni la mediana de los países invierten alrededor del 38/44 % de su presupuesto en salud tal como se ven en las figura 19. En el mapa podemos ver que la mayoría de los países invierten menos del 10 % de su presupuesto en salud.

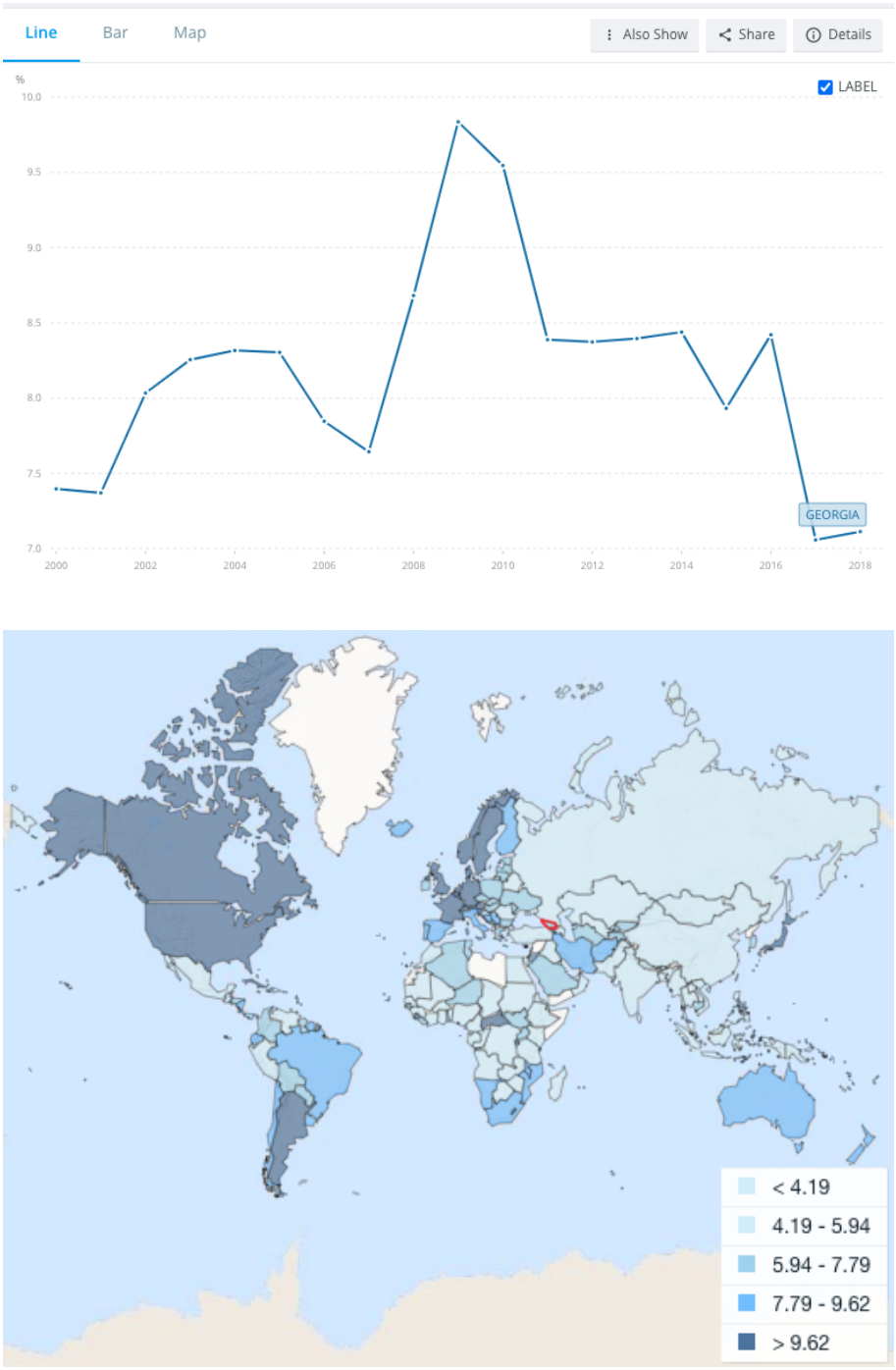
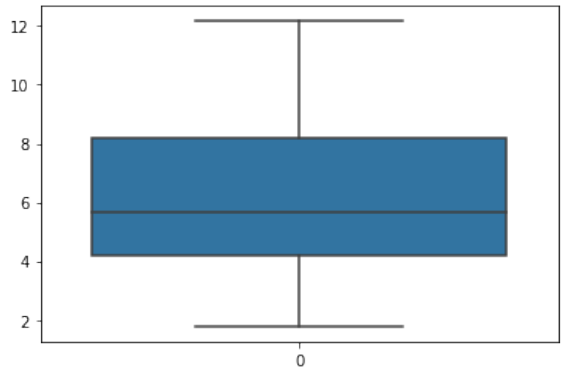


Figura 19: Gráficos que corroboran las anomalías del dataset

Debido a que hemos visto demasiadas anomalías en esta columna, decidimos volver a descargar el índice de nuevo de la página web del banco mundial [3]. Ahora podemos observar en la figura 20 que los datos son más coherentes ya que ninguno supera el 12.2 % de gasto en salud con respecto del PBI. Podemos observar también en la figura 21 que los datos parecen estar muy distribuidos con respecto de la expectativa de vida.



count	154.000000
mean	6.014616
std	2.287822
min	1.778925
25%	4.203710
50%	5.652560
75%	8.179863
max	12.166462

Figura 20: Estadísticas del indicador de porcentaje de salud gastado

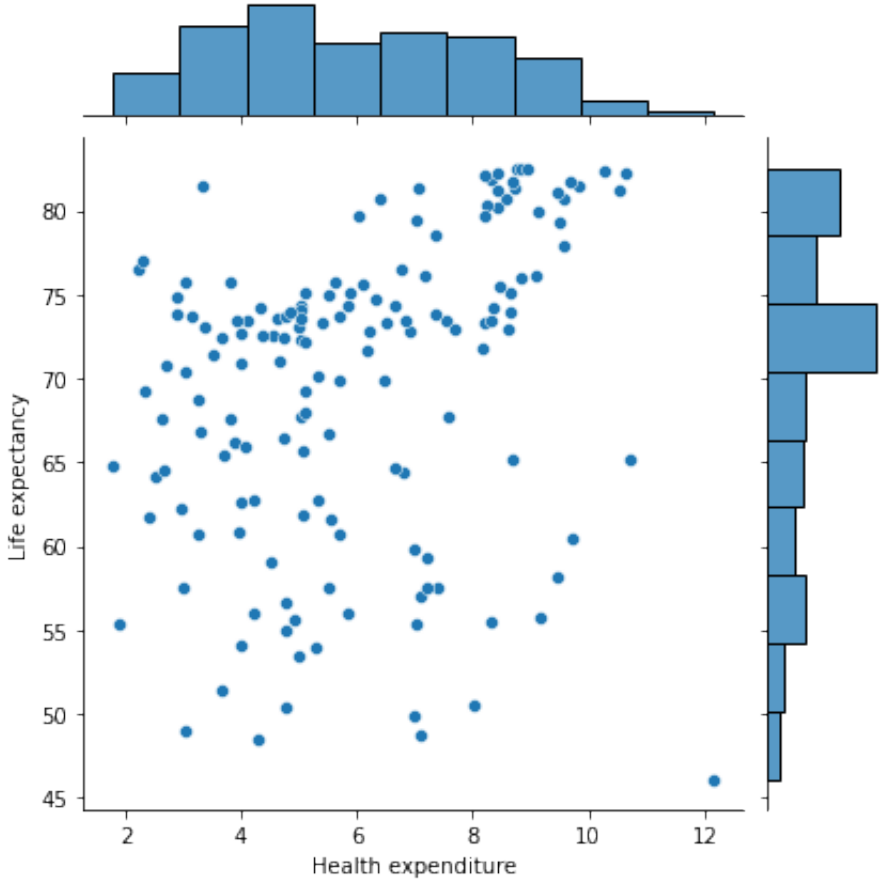


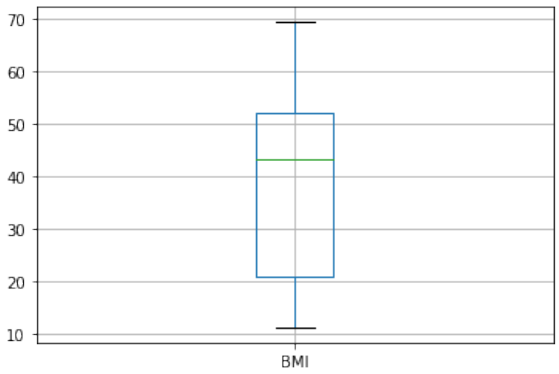
Figura 21: Gráfico de la expectativa de vida de cada país según el porcentaje de gasto en salud proporcional al PBI

2.3.6. Índice de masa corporal

En este apartado vamos a analizar la columna del dataset llamada índice de masa corporal. El sobrepeso y la obesidad se definen como una acumulación anormal o excesiva de grasa que puede perjudicar la salud. El índice de masa corporal (IMC) es un índice simple de peso para la altura que se usa comúnmente para clasificar el sobrepeso y la obesidad en adultos. Se define como el peso de una persona en kilogramos dividido por el cuadrado de su altura en metros.

Adentrándonos en el dataset, pudimos ver que de los 183 países que normalmente estamos analizando, hay dos países que no tienen el indicador medido. Estos países son Sudán y Sudán del Sur. Quitaremos estos dos países del dataset para realizar el análisis exploratorio.

Es muy interesante analizar esta columna que podemos ver en la figura 22 debido a que tanto la media como la mediana (percentil 50) están por arriba de 30, y como podemos ver en la figura 23 estar arriba de 30 indica obesidad. Esto nos muestra que hay muchísimos países que tienen una población con muchísima gente que sufre de obesidad. Por otro lado también podemos ver que casi el 25 % de los países tiene un índice de masa corporal inferior al normal. Estos países con muy bajo peso son países como Vietnam (11.18750), Bangladesh (12.87500) y Laos (14.36250).



	BMI
count	181.000000
mean	38.235394
std	15.910691
min	11.187500
25%	20.850000
50%	43.312500
75%	51.993750
max	69.431250

Figura 22: Estadísticas del indicador de índice de masa corporal

Composición corporal	Índice de masa corporal (IMC)
Peso inferior al normal	Menos de 18.5
Normal	18.5 – 24.9
Peso superior al normal	25.0 – 29.9
Obesidad	Más de 30.0

Figura 23: Tabla del índice de masa corporal

Estos datos no parecen asemejarse a la realidad ya a que el gráfico muestra que la mayoría de los países tienden a tener gente obesa y además tienen valores muy altos que nos hacen sospechar que los datos son incorrectos. Si miramos el mapa 24 de la misma fuente que el dataset podemos ver que no hay valores superiores a 32. Lo cual nos confirma la hipótesis de que el dataset provisto del índice masa corporal está erróneo.

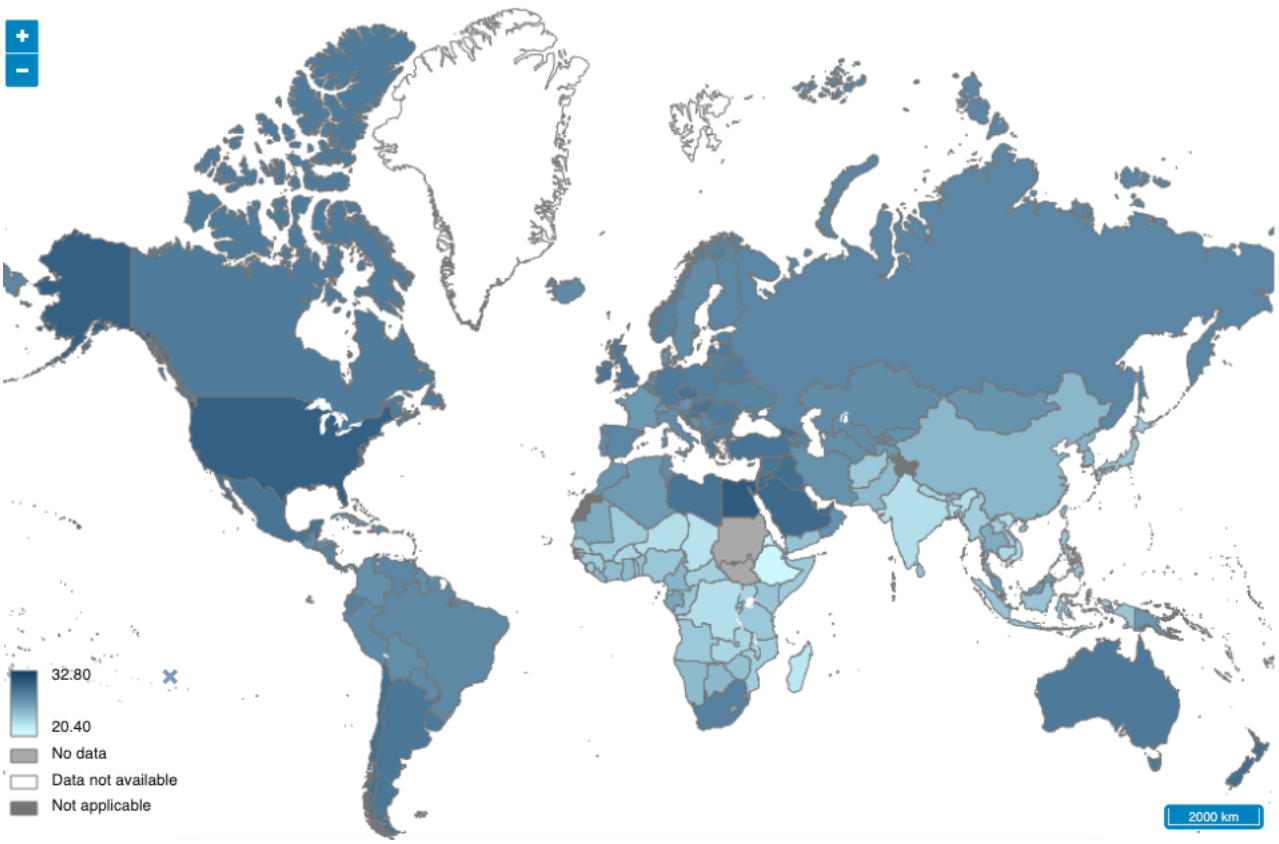
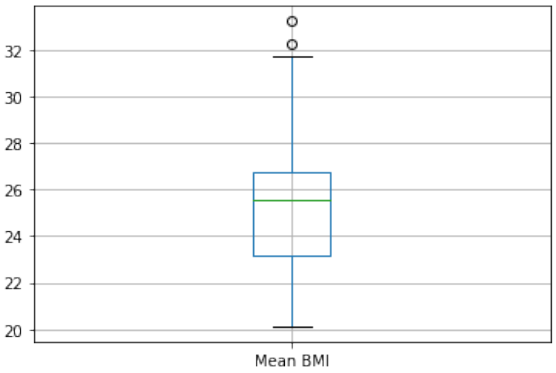


Figura 24: Mapa del índice de masa corporal

Debido a que el indicador de índice de masa corporal nos parecía interesante para analizar su relación con la esperanza de vida, decidimos darle otra oportunidad en lugar de descartarlo, con lo cual nuevamente fuimos a la fuente original de la organización mundial de la salud y descargamos de nuevo los datos desde el año 2000 al año 2015 y obtuvimos unos resultados más coherentes.

Analicemos los datos nuevamente con el dataset que contiene los valores correctos.

Introduciéndonos en el dataset correcto, pudimos ver que hay un total de 191 países, de los cuales Sudán y Sudán del sur no tienen medido este indicador.



	Mean BMI
count	191.000000
mean	25.200916
std	2.427837
min	20.118750
25%	23.159375
50%	25.575000
75%	26.715625
max	33.231250

Figura 25: Estadísticas del indicador de índice de masa corporal

Analizando la columna, podemos ver en la figura 25 que tanto la media como la mediana (50-percentil) están cercanos a 25 que es el máximo valor ”normal”según la figura 23. También se puede ver que la varianza no es tan

grande entre los diferentes datos ya que el valor mínimo es de 20 (peso normal) y el máximo es 33 (que nos indica que se tiene obesidad en dicho país). Como dato curioso podemos ver que el país con más índice de masa corporal es Nauru y el que menos índice corporal tiene es Etiopía.

Luego, si analizamos a Sierra Leona, que es país con menos expectativa de vida, vemos que tiene un indicador de masa corporal de 22.42. Por otro lado Japón (el país con más expectativa de vida) tiene 22.77. Resulta interesante ver que ambos países tienen un índice de masa corporal similar.

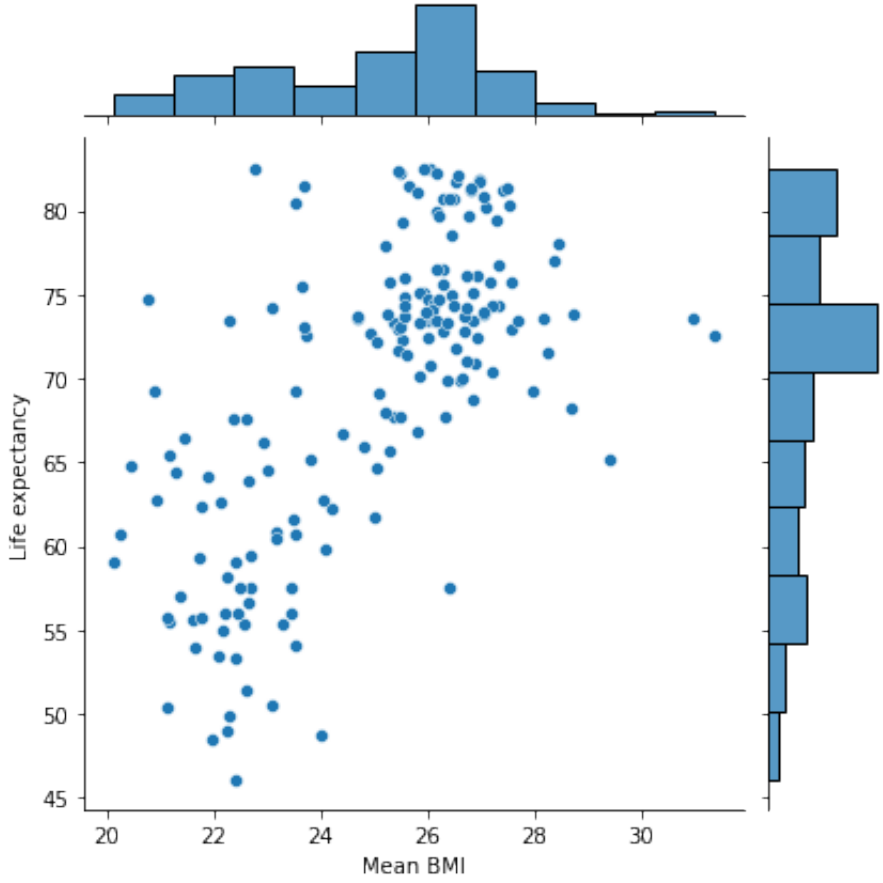


Figura 26: Gráfico de la expectativa de vida de cada país según el índice de masa corporal

Por último analicemos cómo se comporta el índice de masa corporal respecto a la expectativa de vida de los países. Parecería verse en la figura 26 que no hay una relación, a simple vista, entre la expectativa de vida y el índice de masa corporal.

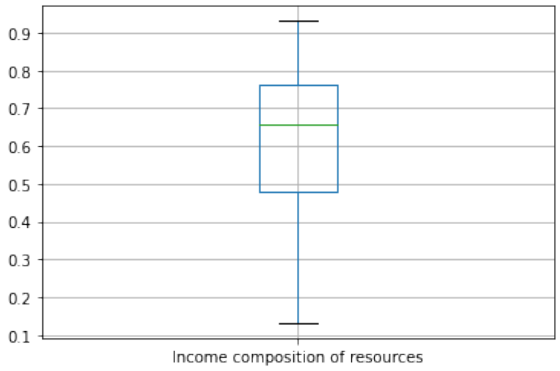
2.3.7. Índice de desarrollo humano

En este apartado vamos a analizar la columna del dataset llamada índice de desarrollo humano(índice que va de 0 a 1).

Adentrándonos en el dataset, pudimos ver que de los 183 países que normalmente estamos analizando, hay diez países que no tienen el indicador medido. Estos países son:

- República Checa
- Costa de Marfil
- Corea del Norte
- República del Congo
- Corea del Sur
- Moldavia
- Somalia
- Reino Unido
- Tanzania
- Estados Unidos

Estos países los quitaremos del dataset para realizar el análisis exploratorio.



	Income composition of resources
count	173.000000
mean	0.627419
std	0.183704
min	0.131688
25%	0.480250
50%	0.657250
75%	0.762625
max	0.931437

Figura 27: Estadísticas del indicador de índice de desarrollo humano

Luego de remover estos países, podemos ver en la figura 27 que la varianza es muy baja, a pesar de tener valores mínimos en 0.13 (Sudán del Sur) y máximos en 0.93 (Noruega). En la figura no parece haber *outliers* y esto se explica como dijimos recién por la baja varianza que tienen los datos y además porque están muy acumulados cerca de la media.

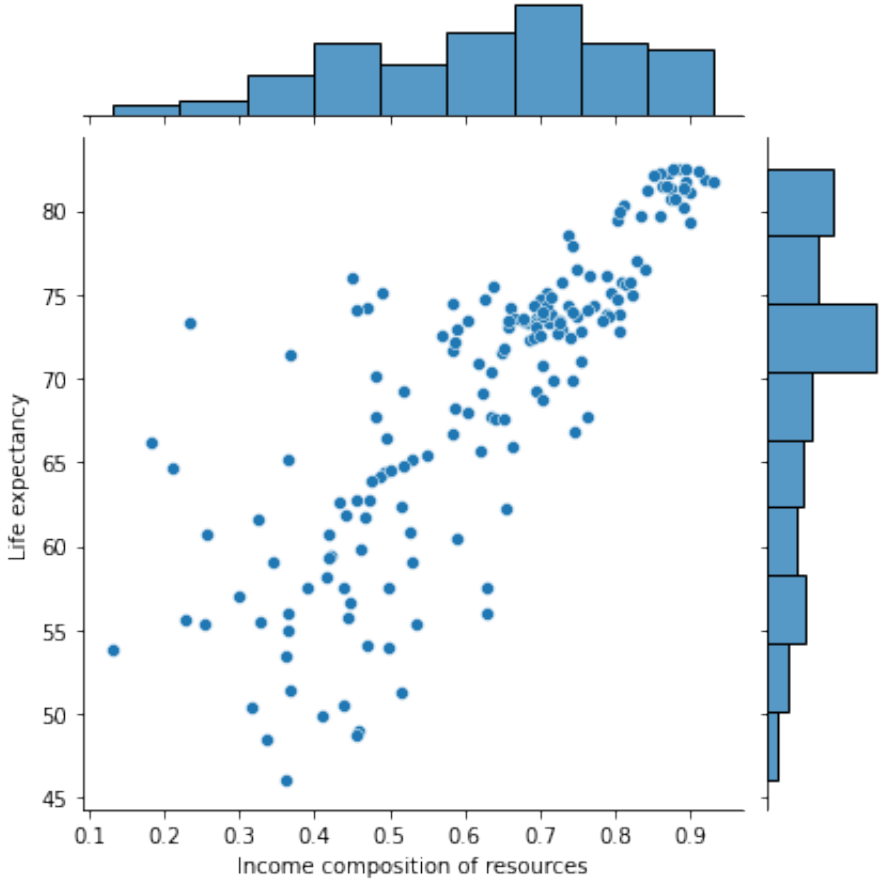


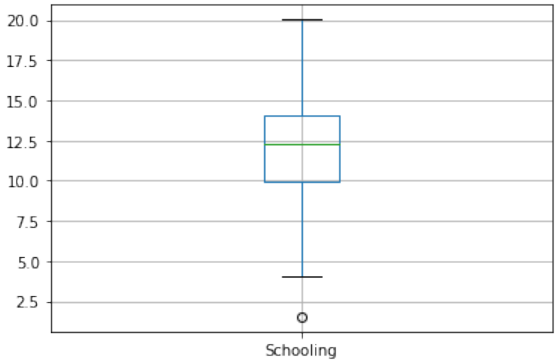
Figura 28: Gráfico de la expectativa de vida de cada país según el índice de desarrollo humano

Por último, analicemos cómo se comporta el índice de desarrollo humano respecto a la expectativa de vida de los países. Pareciera verse a ojo en la figura 28 que hay una especie de relación lineal entre la expectativa de vida y el índice de desarrollo humano. Pareciera ser que a mayor índice de desarrollo humano en la población , mayor expectativa de vida. Esto suena bastante razonable, debido a que cuantos más ingresos tiene una población, mas acceso a las ultimas tecnologías de la salud tiene y por lo tanto mas expectativa de vida. Analizaremos esto en la parte de experimentación.

2.3.8. Índice de escolaridad

En este apartado vamos a analizar la columna del dataset llamada índice de escolaridad. El índice de escolaridad es el número promedio de años de educación completados de la población de un país, excluidos los años dedicados a la repetición de grados individuales.

Adentrándonos en el dataset, pudimos ver que de los 183 países que normalmente estamos analizando, hay diez países que no tienen el indicador medido. Estos países son los mismos que listamos en el índice de desarrollo humano. Estos países los quitaremos del dataset para realizar el análisis exploratorio.



	Schooling
count	173.000000
mean	11.999639
std	3.128800
min	1.531250
25%	9.968750
50%	12.331250
75%	14.056250
max	20.037500

Figura 29: Estadísticas del indicador de índice de escolaridad

Analizando el dataset podemos ver en la figura 29 que el promedio de años de escolaridad está entre 10 y 15 aproximadamente. Sin embargo, el máximo índice es de 20.03 (el cual corresponde a Australia) y el mínimo es de 1.53 (el cual corresponde a Sudán del Sur). Pareciera mirando la figura que el valor mínimo es un *outlier*. Además, si miramos el siguiente valor ordenando de menor a mayor, obtenemos al país Nigeria con un valor de 4.01, es decir, 2.62 veces más grande que el mínimo. Por esta última razón, vamos a realizar una poda del valor mínimo (outlier), para trabajar con datos más representativos.

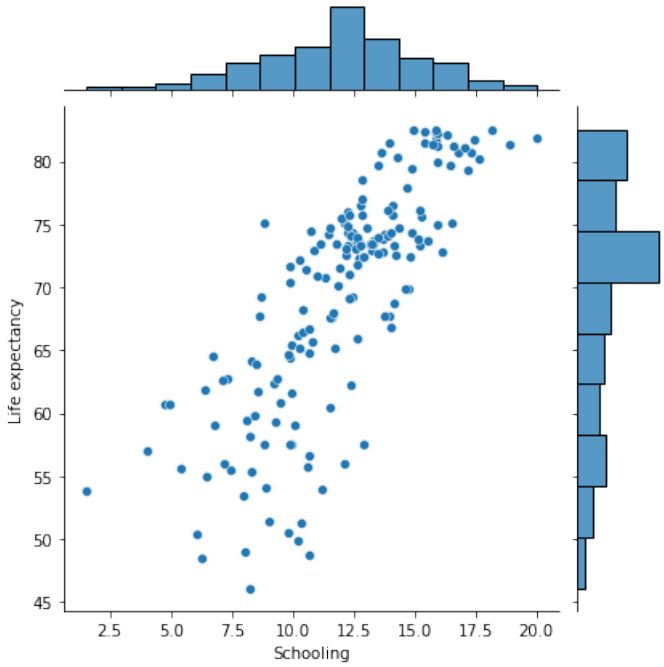


Figura 30: Gráfico de la expectativa de vida de cada país según el índice de escolaridad índice de escolaridad

Por último, analicemos cómo se comporta el índice de escolaridad respecto a la expectativa de vida de los países. Pareciera verse a ojo en la figura 30 que hay una especie de relación lineal entre la expectativa de vida y el índice de escolaridad. Pareciera ser que a mayor índice de escolaridad en la población, mayor expectativa de vida. Esto suena bastante razonable, debido a que los países más desarrollados permiten a los niños y adolescentes mantenerse más tiempo estudiando a diferencia de los países que necesitan que los niños trabajen para aportar dinero a la casa. Y además, nuevamente, los países más desarrollados tienen acceso a las últimas tecnologías de la salud, y por lo tanto, más expectativa de vida. Analizaremos esto en la sección de experimentación.

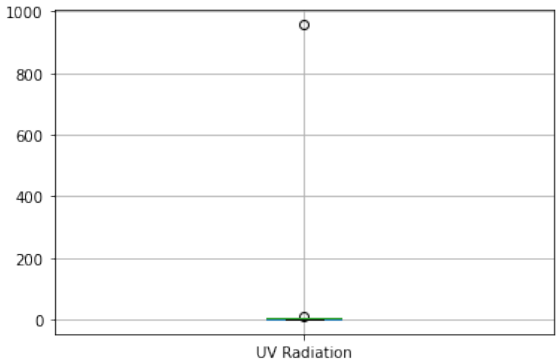
2.3.9. Índice de radiación UV

En este apartado vamos a analizar una columna que no se encuentra en el dataset llamada Índice de radiación UV. El índice UV es un indicador de la intensidad de radiación ultravioleta proveniente del Sol en la superficie terrestre en una escala que comienza en 0 y no está acotado superiormente. El índice UV también señala la capacidad de la radiación UV solar de producir lesiones en la piel. Podemos ver en la figura 31 qué valores de UV son riesgosos para la salud. [5]

Color	Riesgo	Índice UV
Verde	**Bajo**	<2
Amarillo	**Moderado**	3-5
Naranja	**Alto**	6-7
Rojo	**Muy Alto**	8-10
Morado	**Extremadamente alto**	> 11

Figura 31: Tabla de riesgo relacionada con el índice de radiación UV

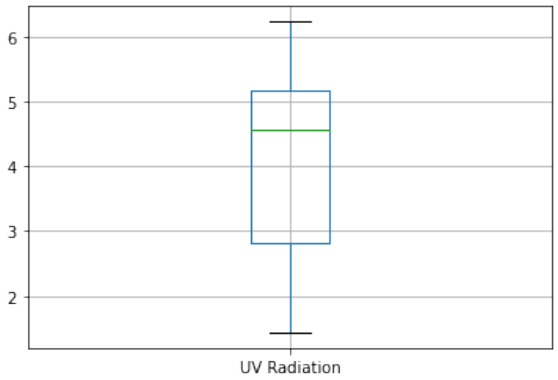
Adentrándonos en el dataset, podemos ver claramente en la figura 32 que hay un claro *outlier* que corresponde a Islandia, con un máximo de 957 en su índice de radiación UV. Viendo la tabla 31 podemos concluir que no sólo es un *outlier*, sino que es algún error en el valor del dataset, ya que si Islandia tuviera dicha radiación UV, no existiría ningún ser humano viviendo en dicho país. También podemos ver en la parte inferior de la figura que hay otro *outlier* que es Maldivas, que parece ser un *outlier* ya que también está a 1.6 veces del siguiente país. Debido a esto, eliminaremos a Islandia y Maldivas del dataset y calcularemos de nuevo los valores.



	Life expectancy	UV Radiation
count	177.000000	177.000000
mean	69.435064	9.480876
std	9.034745	71.639065
min	46.112500	1.439000
25%	62.743750	2.899000
50%	72.487500	4.555000
75%	75.156250	5.203000
max	82.537500	957.000000

Figura 32: Estadísticas del indicador de índice de radiación UV

Adentrándonos en el dataset luego de eliminar a Islandia y Maldivas, podemos ver en la figura 33 que no hay mucha varianza entre los valores de UV en los países. Por ultimo podemos ver que la media y el p50 son muy similares indicando que los países están muy concentrados en valores alrededor de 4.



	UV Radiation
count	175.000000
mean	4.062234
std	1.373102
min	1.439000
25%	2.826000
50%	4.552000
75%	5.176500
max	6.229000

Figura 33: Estadísticas del indicador de índice de radiación UV sin Islandia ni Maldivas

Por último, si analizamos a Sierra Leona, que es el país con menor expectativa de vida, vemos que tiene un índice de radiación UV de 5.08. Japón, por otro lado, tiene 2.52. Veamos qué sucede si comparamos la expectativa de vida de todos los países con el índice de radiación UV.

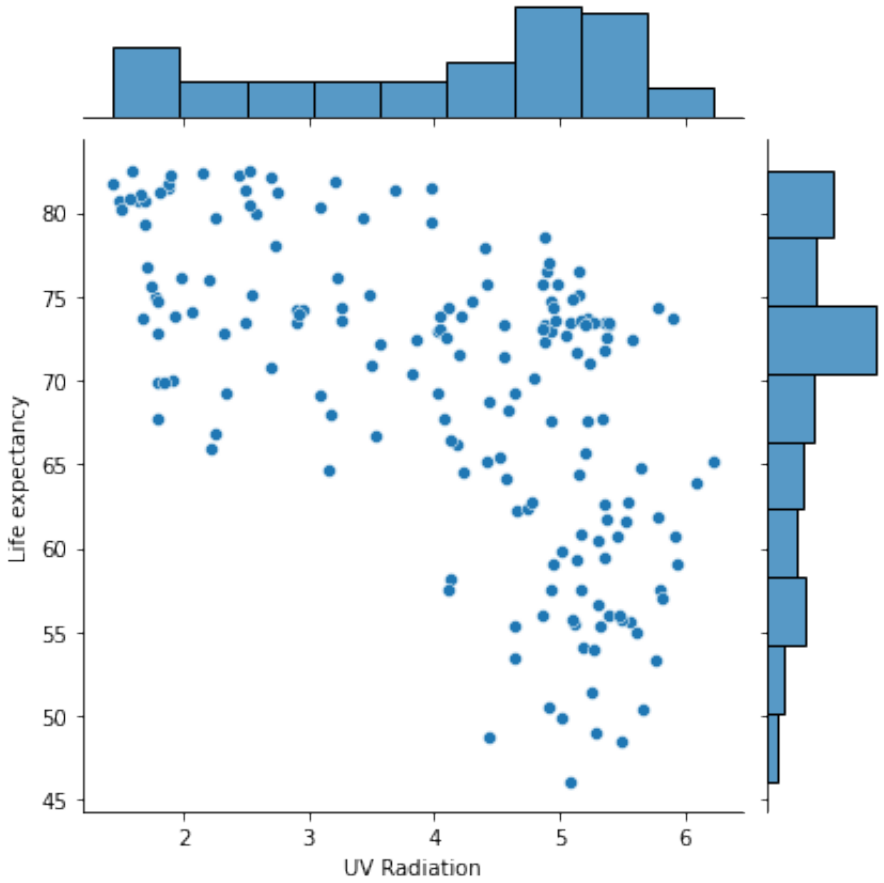


Figura 34: Gráfico de la expectativa de vida de cada país según el índice de radiación UV

Pareciera verse en la figura 34 a ojo que hay una especie de relación exponencial negativa entre la expectativa de vida y el índice de radiación UV. Pareciera ser que a mayor índice de radiación UV en la población, se reduce exponencialmente la expectativa de vida. Esto parece tener sentido debido a que vimos en la tabla 31 que a mayor radiación UV mayor riesgo hay.

Esta ultima hipótesis la analizaremos en la sección de experimentación.

2.4. Correlación de los features

En la figura 35 podemos ver la **matriz de correlación** de los features del dataset provisto por la cátedra. La matriz de correlación (también llamada matriz de covarianza) es una matriz cuadrada que contiene la covarianza entre cada par de features. Recordemos que la covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables.

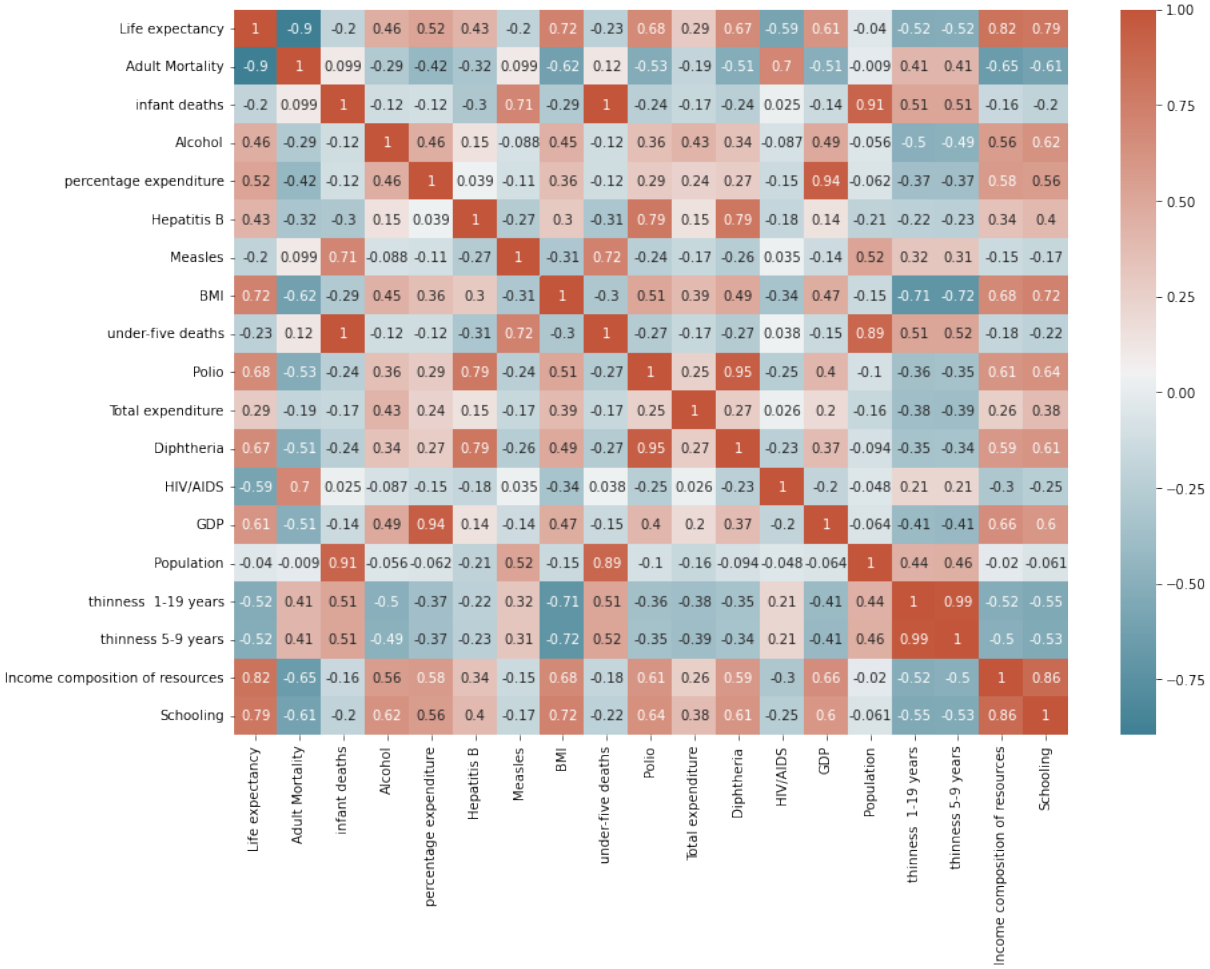


Figura 35: Matriz de correlación

Cada celda en la matriz de correlación tiene un número entre `-1` y `1`. Si la covarianza entre dos indicadores está cerca de `-1` significa que dichos indicadores se relacionan de forma inversamente proporcional. Por el contrario, si la covarianza entre dos indicadores está cerca de `1` quiere decir que dichos indicadores se relacionan de forma directamente proporcional. Una covarianza cercana a `0` indica que no hay relación entre ambos indicadores.

Por ejemplo, la covarianza entre el indicador de *Expectativa de vida* (*Life expectancy*) y el indicador de *Mortalidad adulta* (*Adult Mortality*) es de `-0,9`, con lo cual están relacionados de forma inversamente proporcional: a mayor *Mortalidad adulta*, menor *Expectativa de vida*, en promedio. Para evidenciar esto de forma más visual volvemos a la figura 10, donde se puede ver que, en promedio, los países con menos expectativa de vida también tienen mayor mortalidad adulta.

Otro fenómeno similar ocurre con el indicador del *Índice de desarrollo humano* (*Income composition of resources*). Podemos ver que la covarianza entre dicho indicador y la *Expectativa de vida* es de `0,82`, lo que significa que ambos indicadores están relacionados de forma directamente proporcional (en promedio). En la figura 28 podemos ver que los países con máxima expectativa de vida son también aquellos en los que hay mayor *Índice de desarrollo humano*.

Por último, podemos ver que la covarianza entre el indicador de *Población* (*Population*) y el indicador de *Expectativa de vida* es de `-0,04`. Esto parecería indicar que los indicadores no se relacionan para nada. Veamos si esto es así.

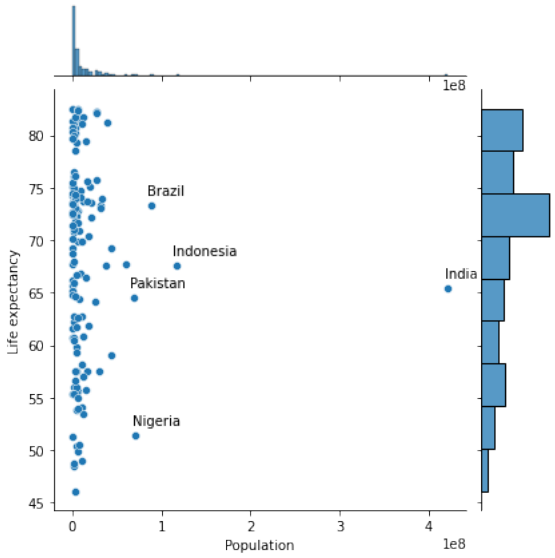


Figura 36: Gráfico de la expectativa de vida de cada país según su población

En la figura 36 podemos ver el gráfico de la relación entre la expectativa de vida de cada país y su población. Podemos ver que, si bien existen algunos países que son *outliers* en lo que respecta a la población, los mismos no parecerían ser *outliers* con respecto a la expectativa de vida.

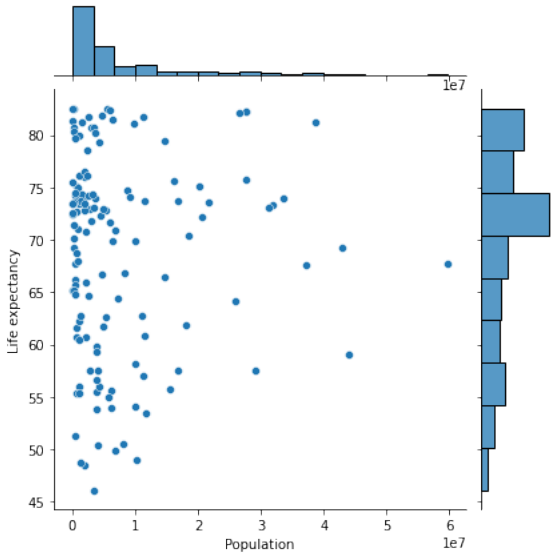


Figura 37: Gráfico de la expectativa de vida de cada país según su población con *outliers* filtrados

En la figura 37 podemos ver el mismo gráfico de la figura 36 sin los *outliers* mencionados. De esta forma queda un poco más evidente que la distribución de los países resulta aleatoria en el plano cartesiano.

3. Método Utilizado: Regresión lineal

La regresión lineal hace honor a su nombre: es un enfoque muy sencillo para predecir una respuesta cuantitativa Y sobre la base de un conjunto de variables predictoras X_i. Supone que hay una relación aproximada entre los X_i e Y. Matemáticamente, podemos escribir esta relación como

$$Y \approx \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{1}$$

Donde $\approx$ se lee como *"es modelado aproximadamente como"* y p es la cantidad de variables predictoras.

En la práctica los coeficientes β_i son desconocidos. Es por esto que antes de poder usar nuestro modelo para hacer predicciones es necesario utilizar nuestro dataset para estimar los coeficientes. Nuestro objetivo es obtener aquellos β_i tales que el modelo lineal aproxime *bien* nuestros datos disponibles. En otras palabras, la forma bilineal generada por los β_i debe estar tan cerca como sea posible de nuestros datos.

Hay muchas formas de medir *cercanía*. Sin embargo, la forma más común de hacerlo involucra minimizar el criterio de los *cuadrados mínimos*.

Sea $\hat{y_j} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ij}$ la predicción para Y basada en el *j-ésimo* dato. Entonces $e_j = y_j - \hat{y_j}$ representa el *j-ésimo* residuo entre la variable predicha y la observada. Luego podemos definir la *Suma residual de cuadrados (RSS por sus siglas en inglés)* como

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 \tag{2}$$

El enfoque de cuadrados mínimos elije los β_i tal que minimice el RSS. Por propiedades vistas en la teórica de la cátedra, el problema de cuadrados mínimos puede representarse matricialmente como encontrar $b = (b_0, ..., b_p)$ tales que minimicen

$$||Xb - Y||_2 \tag{3}$$

Donde X tiene la primer columna con todos 1 que representa la función $I : \mathbb{R} \to \mathbb{R}$ identidad , X_j en la columna $j + 1$ e Y son los valores observados. Resolver este problema de minimización es equivalente a resolver el sistema lineal de ecuaciones de la forma

$$X^t \cdot X \cdot b = X^t \cdot Y \tag{4}$$

3.1. Algoritmo

function LINEARREGRESSION($X \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^{1 \times p}$)
  $X \leftarrow$ AGREGAR1COLUMNA(X) ▷ $O(np)$
  for *i ∈ [1, . . . , n]* **do** ▷ $O(n)$
    $X[i][m+1] \leftarrow 1$ ▷ $O(m^2 \cdot n)$
  end for

  $M \leftarrow X^t \cdot X$ ▷ $O(p^2 \cdot n)$
  $b \leftarrow X^t \cdot Y$ ▷ $O(p \cdot n)$

  $x \leftarrow$ LDLT(M, b) ▷ $O(p^3)$

  $coef\_ \leftarrow x[0:p]$ ▷ $O(p)$
  $intercept\_ \leftarrow x[p+1]$ ▷ $O(1)$
end function

La función AGREGAR1COLUMNA lo que hace es crear una nueva matriz de dimensión $n \times m + 1$ donde $X[i][j] =$ AGREGAR1COLUMNA$(X)[i][j]$ $\forall i < n, j < m$. Luego se llena la nueva columna de unos, esta nueva columna representa el termino independiente de la regresión lineal. Por último se computa la matriz M, el vector b y se aplica LDLT donde la función LDTL resuelve el sistema lineal mediante factorización Cholesky. Esta función se puede usar ya que M es simétrica definida positiva por ser el producto de una matriz y su transpuesta.

4. Experimentación

Luego de realizar un análisis exploratorio del dataset, plantearemos algunas hipótesis y realizaremos experimentos con los datos del mismo.

4.1. Preliminares

Una pregunta válida para hacerse previo al análisis de datos es *¿Por qué debemos estimar una función f que aproxime a Y (en este caso la esperanza de vida de una población)?*. Hay dos razones principales por las que querríamos estimar f: para predecir los valores de Y dados X_i o para ver cómo infieren los X_i sobre Y.

Para este análisis y experimentación haremos énfasis sobre la segunda razón. Esto consiste en intentar entender la asociación entre Y y $X_1, ..., X_p$. En este caso nuestro deseo es estimar f, pero nuestro objetivo no es lograr hacer predicciones. Para hacerlo, f no puede ser tratado como una *caja negra*, porque nosotros necesitamos saber su forma exacta para poder preguntarnos:

¿Qué predictores se asocian con la respuesta?

Es usual que sólo una pequeña fracción de los predictores disponibles estén sustancialmente asociados con Y. Identificar algunos importantes puede ser extremadamente útil.

¿Qué relación hay entre la respuesta y cada predictor?

Algunos predictores pueden tener una relación positiva con Y, en el sentido que cuanto más grandes los valores del predictor más grandes los valores de Y. Otros predictores pueden tener la relación opuesta. Dependiendo de la complejidad de f, la relación entre la respuesta y un predictor dado puede también depender de los valores de los otros predictores.

¿Puede la relación entre Y y cada predictor ser adecuadamente resumida usando una ecuación lineal, o es la relación más compleja?

Históricamente la mayoría de los métodos para estimar f han tomado una forma lineal. En algunas situaciones es razonable o incluso deseado. Pero muy seguido la verdadera relación es más complicada, en ese caso un modelo lineal no proveerá una representación precisa de la relación entre el input y el output.

4.1.1. ¿Cómo estimamos *f*?

Para estimar a f tomamos el método de *regresión lineal*. En el cual asumimos que tiene la forma

$$Y \approx \beta_0 + \sum_{i=1}^{p} \beta_i T_i(X) \tag{5}$$

donde T_i es alguna transformación sobre uno o varios predictores de Y. Y como vimos elegimos los β_i tales que minimicen el RSS.

La principal desventaja de este acercamiento paramétrico es que el modelo que aproximamos usualmente no va a representar la forma de la verdadera f desconocida. Si la forma de la verdadera f está muy lejos de nuestra aproximación nuestra estimación va a ser pobre.

La principal ventaja de este enfoque es que es sumamente interpretable. Nos va a ser mucho más fácil entender la relación entre los predictores e Y, que es en fin lo que queremos entender.

4.1.2. Evaluación del accuracy del modelo

Es natural querer cuantificar en qué medida el modelo se ajusta a los datos. Para medir el accuracy de nuestras regresiones lineales vamos a utilizar dos medidas: el *residual standard error* (RSE) y el R^2 estadístico.

RSE

El RSE es una estimación del error intrínseco que tiene la f verdadera desconocida. En términos mas sencillos, es la cantidad promedio que la respuesta se desviará de la verdadera regresión lineal. Otra manera de pensarlo es que, incluso si nuestro modelo estimara a f perfectamente, cualquier predicción tendría algún tipo de error que sería aproximado al RSE.

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{6}$$

R²

El RSE provee una medición absoluta de la falta de ajuste del modelo a los datos. Pero como es medido en unidades de Y, no es siempre claro qué constituye un buen RSE. El R^2 provee una alternativa para medirlo. Toma la forma de una proporción. Una proporción formada por la suma total de cuadrados (TSS) de la muestra Y y el RSS, que mide la cantidad de variación que no puede ser explicada por el modelo.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{7}$$

4.2. Experimento 1

El primer experimento que vamos a plantear busca encontrar si existe algún tipo de relación entre la expectativa de vida y el *índice de desarrollo humano*. El indicador del *índice de desarrollo humano* es un indicador incluido en el dataset original de la cátedra, así que para realizar el análisis exploratorio y la experimentación tan solo tuvimos que preprocesarlos.

4.2.1. Pre-procesamiento de datos

El pre-procesamiento de esta feature fue bastante simple, únicamente tuvimos que eliminar aquellos países que no tenían este indicador. Estos países están listados en la sección de análisis exploratorio.

4.2.2. Hipótesis

El índice de desarrollo humano es un indicador que nos muestra qué tan avanzada está una sociedad con respecto a varios factores como educación, calidad de vida, ingresos, etc.

Luego de hacer el análisis exploratorio en la sección anterior, nos planteamos la siguiente hipótesis como guía para realizar la experimentación.

> El índice de desarrollo humano es directamente proporcional con la expectativa de vida.

4.2.3. Análisis

Comenzamos el análisis donde lo habíamos dejado en la sección de análisis exploratorio. En aquella sección habíamos concluido que, mirando la figura 27 a ojo, teníamos la intuición que existía una relación lineal entre ambos indicadores. Ahora, con la hipótesis planteada, vamos a ver si esto es cierto o simplemente es una ilusión óptica.

Para ver si la función generada por las variables expectativa de vida e índice de desarrollo humano están relacionadas linealmente utilizaremos la técnica de regresión lineal. Utilizaremos la técnica de regresión lineal para generar las funciones de tipo lineal, logarítmica y cuadrática que mejor aproximen a los puntos. Estas familias de funciones no son todas, pero nos servirán para comparar con la hipótesis propuesta.

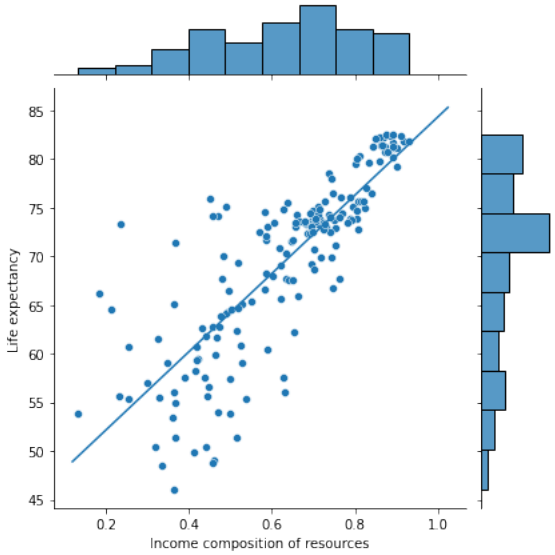


Figura 38: Función lineal que mejor aproxima a los puntos

En la figura 38 podemos ver los puntos reales de la función de expectativa de vida según el indicador de desarrollo humano y también podemos observar una linea recta que es la aproximación lineal de dicha función utilizando regresión lineal. Nuevamente a ojo podemos ver que la función lineal que genera la regresión lineal aproxima muy bien los puntos originales. Sin embargo, necesitamos un indicador que nos confirme la variación que hay entre los puntos reales y la función lineal generada por la regresión lineal. En particular, utilizaremos el R^2, indicador que explicamos anteriormente en la sección regresión lineal. El calculo del R^2 nos dio $0{,}5$, el cual es un número que indica que la aproximación no es perfecta, pero sí sugiere que hay una relación entre los puntos y la función lineal generada.

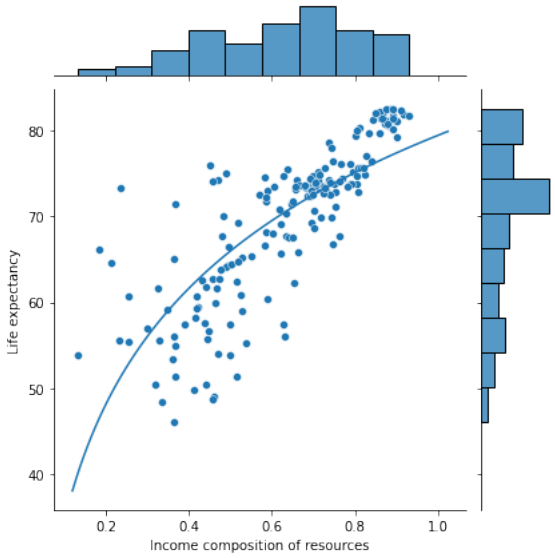


Figura 39: Función logarítmica que mejor aproxima a los puntos

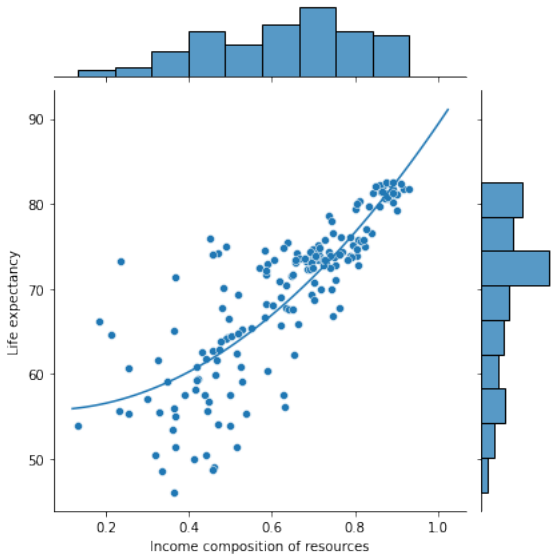


Figura 40: Función cuadrática que mejor aproxima a los puntos

Por otra parte, podemos ver en las figuras 39 y 40 que las funciones logarítmicas y cuadráticas que mejor aproximan a los puntos son bastante malas. Además, como hicimos con la aproximación lineal, calculamos el R^2 de ambas y nos dio $-41{,}49$, $-658{,}80$ respectivamente. Ambos valores demuestran que tienen muy malas aproximaciones.

4.2.4. Conclusiones

La hipótesis planteada nos parece que es correcta, ya que el indicador R^2 es positivo y dio $0{,}5$ (1 es la aproximación perfecta), además los gráficos acompañan la hipótesis planteada.

Tenemos algunas excepciones de países como Granada, que tiene buena expectativa de vida (73.29) pero muy bajo índice de desarrollo humano, y por otro lado tenemos a Rusia que tiene una expectativa de vida 67.76 por debajo de la media (69.22), pero tiene un buen índice de desarrollo humano (0.76) superando la media del mismo (0.62). Este error se puede deber a la complejidad de la realidad que el modelo lineal no alcanza a explicar. Para trabajo futuro se podría volver a hacer el mismo experimento pero con más muestras. Esto se puede hacer agregando

países o tomando poblaciones más pequeñas y así ver qué tanto aumenta la cantidad de valores alejados del modelo.

4.3. Experimento 2

El segundo experimento que vamos a plantear busca encontrar si existe algún tipo de relación entre la expectativa de vida y el *índice de escolaridad*. Además sumaremos la columna que indica si un país es desarrollado o no, para luego analizar la intuición que tuvimos en el análisis exploratorio de que un país desarrollado tiene mayor escolaridad y mayor expectativa de vida.

El indicador del *índice de escolaridad* es una métrica incluida en el dataset original de la cátedra, así que para realizar el análisis exploratorio y la experimentación tan solo tuvimos que preprocesarlos.

4.3.1. Pre-procesamiento de datos

En el pre-procesamiento de esta feature tuvimos que eliminar aquellos países que no tenían este indicador y eliminamos el mínimo ya que creíamos que era un *outlier* que nos podía aumentar el error en el modelo, lo cual bajaría la precisión de nuestro análisis para el conjunto más grande de los datos. Los países eliminados se pueden observar en el análisis exploratorio del indicador del índice de escolaridad.

Por otro lado, aprovechamos la columna del dataset provisto por la cátedra donde se indica si un país es desarrollado o en vías de desarrollo. Esto nos permite visualizar en el análisis los datos de escolaridad de los países desarrollados y en vías de desarrollo respectivamente. Estos datos se utilizaran para analizar la segunda parte de la hipótesis.

4.3.2. Hipótesis

El índice de escolaridad es el número promedio de años de educación completados de la población de un país. Nos resultó interesante realizar una experimentación sobre el índice de escolaridad, la expectativa de vida y si un país es desarrollado o no.

Luego de hacer el análisis exploratorio en la sección anterior, nos planteamos la siguiente hipótesis como guía para realizar la experimentación.

> El índice de escolaridad es directamente proporcional con la expectativa de vida y además los países con alto índice de escolaridad son países desarrollados.

4.3.3. Análisis

Nuevamente comenzamos el análisis donde lo habíamos dejado en la sección de análisis exploratorio. En aquella sección habíamos concluido que, mirando la figura 29 a ojo, teníamos la intuición de que tenían una relación lineal con la expectativa de vida. Ahora con la hipótesis planteada, vamos a ver si esto es cierto o simplemente es una falacia visual. Además sumamos la figura 41 para comparar la segunda parte de la hipótesis que enuncia lo siguiente: un país con alto índice de escolaridad suele ser un país desarrollado. En aquella figura, a simple vista también se ve que los países con alto índice de escolaridad son desarrollados.

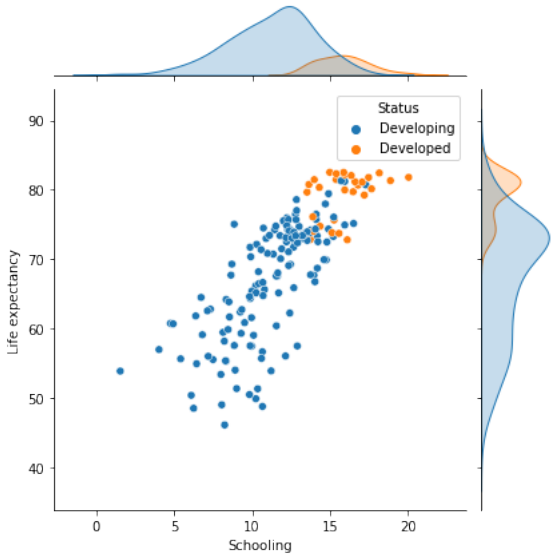


Figura 41: Función comparativa entre la expectativa de vida y el índice de escolaridad, distinguiendo los países desarrollados de los no desarrollados

Para ver si la función generada por las variables expectativa de vida e índice de escolaridad están relacionadas linealmente utilizaremos la técnica de regresión lineal. Utilizaremos la técnica de regresión lineal para generar las funciones de tipo lineal, logarítmica y cuadrática que mejor aproximen a los puntos. Estas familias de funciones no son todas, pero nos servirá para comparar con la hipótesis propuesta.

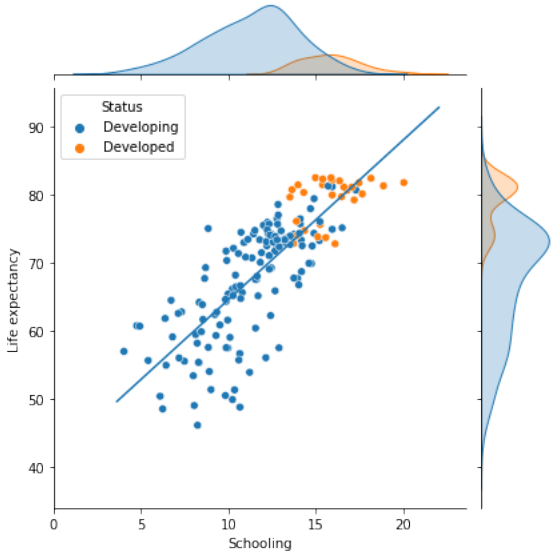


Figura 42: Función lineal que mejor aproxima a los puntos

En la figura 42 podemos ver los puntos reales y también podemos observar una linea recta que es la aproximación lineal de dicha función utilizando regresión lineal. Nuevamente a ojo podemos ver que la función lineal que se genera aproxima muy bien los puntos reales, pero nuevamente, necesitamos un indicador que nos confirme la variación que hay entre los puntos reales y la función lineal generada. En particular, utilizaremos el R^2, indicador que explicamos anteriormente en la sección Regresión lineal. El cálculo del R^2 nos dio 0.4189, el cual es un numero que esta indicando que la aproximación no es perfecta, e incluso es menor que en el del experimento anterior, pero sin embargo, a nuestro criterio, es un valor que sugiere que efectivamente existe una relación entre los puntos y la función lineal generada.

Por otra parte, podemos ver en las figuras 43 y 44 las funciones logarítmicas y cuadráticas que mejor aproximan a los puntos. Si analizamos la cuadrática, vemos que tomó el coeficiente principal tan cercano a cero que parece a ojo que es una función lineal y por otro lado, la logarítmica tiene muy mala aproximación con un R cuadrado de -9.91. Con estos resultados, podemos ver que la función lineal es la más indicada para modelar la relación.

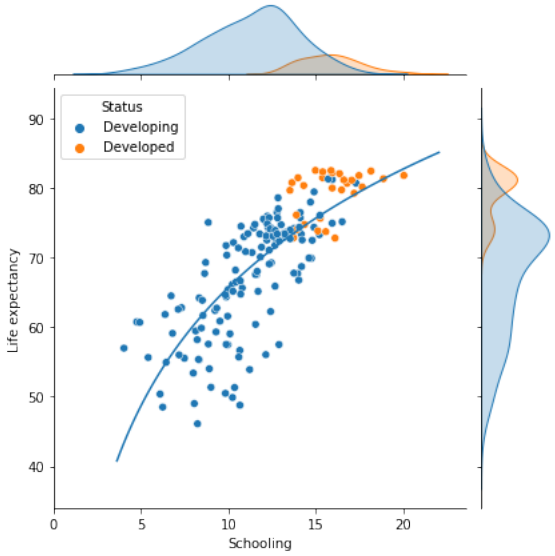


Figura 43: Función logarítmica que mejor aproxima a los puntos

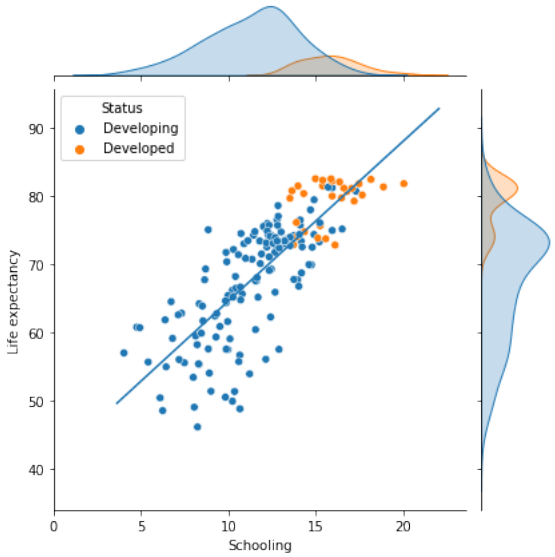


Figura 44: Función cuadrática que mejor aproxima a los puntos

Por último, veamos qué sucede con la segunda parte de la hipótesis que enunciaba lo siguiente: los países con alto índice de escolaridad son países desarrollados.

Para analizar esto utilizamos los dos nuevos datasets creados y calculamos sus estadísticas para realizar una comparación. En la figura 45 podemos ver el resultado de estas comparaciones. Analizando estas estadísticas, a simple vista podemos concluir que los países con altos índices de escolaridad son desarrollados. Esto se puede ver principalmente ya que el 75 % de los países no desarrollados tienen un índice de escolaridad (12.86) menor al mínimo índice de escolaridad de los países desarrollados (13.51). Por otro lado también se puede ver que los países desarrollados tienen en promedio un índice de escolaridad de 44 % mayor a los países no desarrollados.

	Schooling Developed
count	29.000000
mean	15.845474
std	1.647126
min	13.518750
25%	14.350000
50%	15.868750
75%	16.787500
max	20.037500

	Schooling Not Developed
count	144.000000
mean	11.225130
std	2.764039
min	1.531250
25%	9.746875
50%	11.668750
75%	12.860938
max	17.293750

Figura 45: Estadísticas del indicador de escolaridad diferenciado por si es un país desarrollado o no

4.3.4. Conclusiones

La hipótesis planteada nos parece que es correcta ya que el indicador R^2 es positivo y dio $0{,}41$ (1 es la aproximación perfecta). Además los gráficos acompañan la hipótesis planteada. Por otra parte, en cuanto a la segunda parte de la hipótesis, también creemos que es cierta, ya que los estadísticos acompañan fuertemente la hipótesis.

Como en el experimento anterior, tenemos algunas excepciones de países como Botsuana que tiene baja expectativa de vida (con respecto al promedio), pero un índice de escolaridad superior al promedio. Esto se debe a que Botsuana hizo grandes avances en el desarrollo educativo después de la independencia en 1966. Con el descubrimiento de diamantes justo después de la independencia y el aumento de los ingresos gubernamentales que esto trajo, hubo un enorme aumento en la oferta educativa en el país. A todos los estudiantes se les garantizó diez años de educación básica y aproximadamente la mitad de la población escolar asiste dos años más a la escuela secundaria [6].

A pesar de esta excepción que investigamos, podemos concluir que mirando el dataset completo, la hipótesis parece ser cierta.

4.4. Experimento 3

Vimos en un análisis previo cómo se relaciona la *escolarización* con la esperanza de vida por si sola. La idea de este experimento es agregar el *gasto en salud como porcentaje del PBI* de un país y analizar si de esta forma se puede explicar mejor la esperanza de vida. También, en la experimentación, planteamos el mismo experimento para el *índice de desarrollo humano*. Es decir, vimos cómo interactúa con el *gasto en salud como porcentaje del PBI* para explicar la esperanza de vida.

El *índice de desarrollo humano* y la *escolarización* fueron datos provistos por la cátedra y previamente analizados en los experimentos anteriores.

4.4.1. Pre-procesamiento de datos

El pre-procesamiento de esta feature fue bastante simple, únicamente tuvimos que eliminar aquellos países que no tenían alguno de los indicadores usados. Estos países están listados en la sección de análisis exploratorio.

4.4.2. Hipótesis 1

Luego de hacer el análisis exploratorio en la sección anterior, nos planteamos la siguiente hipótesis como guía para realizar la experimentación.

> Al tener un mayor gasto en salud, tanto el índice de desarrollo humano como la escolarización se van a potenciar al momento de aumentar la calidad de vida.

4.4.3. Análisis 1

Para este análisis hemos planteado dos regresiones multilineales para predecir la esperanza de vida de una población. Para esto utilizamos los métodos vistos anteriormente.

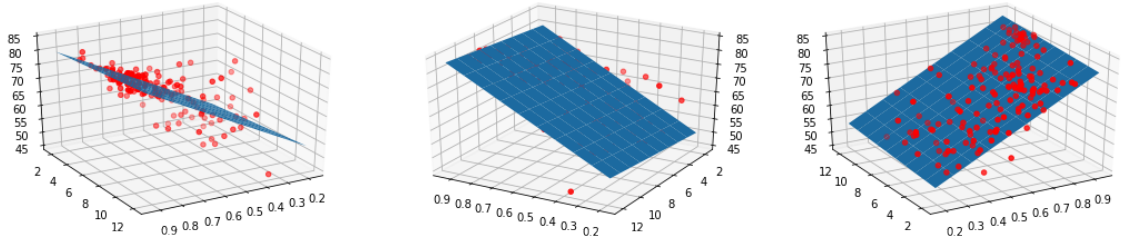


Figura 46: Regresión multilineal entre índice de desarrollo humano y gasto en salud como porcentaje del PBI para predecir la esperanza de vida.

En la figura 46 podemos ver gráficamente el plano generado por la regresión lineal y los puntos que representan cada elemento de nuestro dataset.El R^2 para este modelo fue de $0{,}44$.

Esta estimación nos sugiere que el gasto en salud en proporción al PBI no es significativo a la hora de construir la esperanza de vida, ya que los valores del gasto en salud de la muestra varían entre 0.2 y 0.12, es decir, que lo que aporta a la esperanza de vida el porcentaje de gasto en salud varía entre -0.05 y -0.03 aproximadamente. Además, nuestro nuevo R^2 es menor al anterior visto en el modelo donde sólo se tomaba como feature al índice de desarrollo humano y el R^2 fue de 0.5.

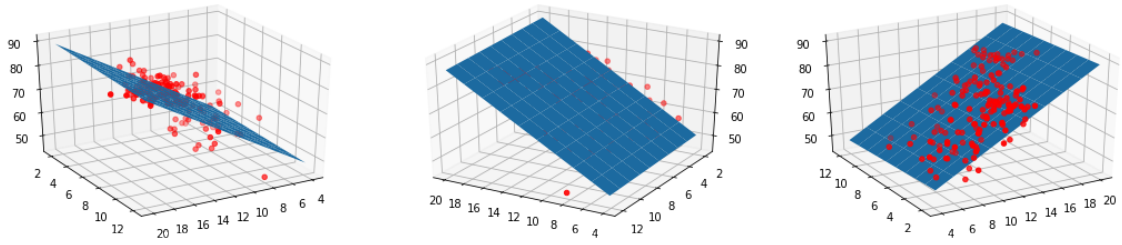


Figura 47: Regresión multilineal entre la escolarización y el gasto en salud como porcentaje del PBI para predecir la esperanza de vida.

En al figura 47 podemos ver gráficamente el plano generado por la regresión lineal y los puntos que representan cada elemento de nuestro dataset. El R^2 para este modelo fue de $0{,}54$.

Nuevamente la estimación nos sugiere que el gasto en salud en proporción al PBI no explica significativamente la esperanza de vida, ya que los valores del gasto en salud de la muestra varían entre 0.2 y 0.12, es decir, que lo que aporta a la esperanza de vida el porcentaje de gasto en salud varía entre 0.066 y 0.04 aproximadamente. Sin embargo nuestro nuevo R^2 es mayor al anterior visto en el modelo donde solo se tomaba como feature a la escolarización. En nuestro anterior experimento el R^2 fue $0{,}41$. Esto nos sugiere que los datos se ajustan mejor que en nuestro anterior ejemplo.

4.4.4. Hipótesis 2

Luego de ver cómo afecta el gasto público a la esperanza de vida junto con la escolarización y el índice de desarrollo humano, nos propusimos investigar cómo los dos últimos features pueden predecir juntos a la esperanza de vida. Esperamos un mejor ajuste ya que ambas parecen ser significativas al explicar la esperanza de vida.

> Hacer una regresión lineal entre el escolarización y el índice de desarrollo humano ajustará mejor a la expectativa de vida que ambas por separado

4.4.5. Análisis 2

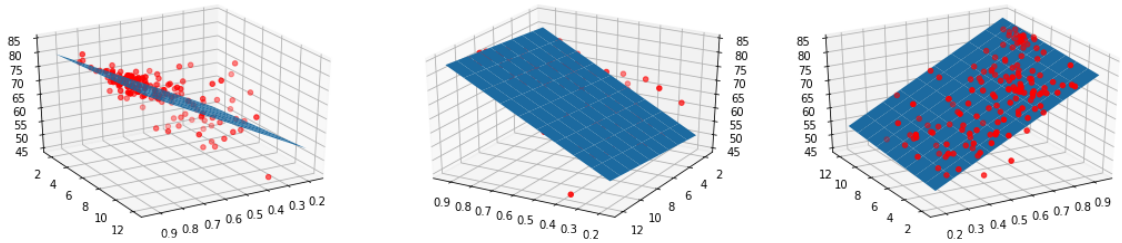


Figura 48: Regresión multilineal entre índice de desarrollo humano y la escolarización para predecir la esperanza de vida.

En al figura 48 podemos ver gráficamente el plano generado por la regresión lineal y los puntos que representan cada elemento de nuestro dataset. El R^2 para este modelo fue de $0{,}58$.

En este nuevo modelo logramos obtener un mejor ajuste que analizando los features por separado. Es interesante notar que el coeficiente de la escolarización al ser 1 hace que la escolarización aporte entre 2 y 20 años a la esperanza de vida. Esto significa que es sumamente significativo a la hora de predecir la esperanza de vida. Más aún si tenemos en cuenta que el índice de desarrollo humano se construye también en base a la escolarización promedio de un país.

4.4.6. Conclusiones

El gasto en salud como porcentaje del PBI no parece ser significativo para la expectativa de vida de un país. Con esto no podemos decir que no sea relevante hacerlo, sino que puede ser más complejo analizar este índice. Podríamos tener gasto en salud en áreas que no aumente la esperanza de vida pero sí reduzca la mortalidad infantil, o gasto en salud destinado a cuadros clínicos menos frecuentes en la población. También se puede pensar que el PBI per cápita y el PBI en sí varía de país a país, por lo que no es lo mismo un 5 % del PBI destinado a salud en un país con un alto PBI que el 5 % del PBI destinado a salud en un país con un bajo PBI.

También vimos cómo una alta escolarización y un alto índice de desarrollo humano parecen estar correlacionados con una alta esperanza de vida.

4.5. Experimento 4

El cuarto experimento que vamos a plantear busca encontrar si existe algún tipo de relación entre la expectativa de vida y el *índice de radiación UV*. El indicador de el *índice de radiación UV* es un indicador que no está incluido en el dataset original de la cátedra, así que para realizar el análisis exploratorio y la experimentación tuvimos que descargarlo, realizar un procesamiento y luego *mergearlo* con el dataset original.

4.5.1. Pre-procesamiento de datos

En el pre-procesamiento de esta feature en principio tuvimos que descargar el dataset de la misma fuente del dataset original. Luego de realizar la descarga, eliminamos varias columnas del dataset quedándonos únicamente con el país y el valor del índice de radiación UV. Luego de tener preprocesado el dataset, hicimos el análisis exploratorio que se puede ver en la sección de análisis exploratorio.

Una vez filtrado aquellos países que comentamos en el análisis exploratorio, procedimos a realizar un *merge* con el dataset original, teniendo como columna de unión el país. Dejamos afuera aquellos países que sólo tenían el índice de radiación UV o solo tenían la expectativa de vida para realizar un análisis más preciso.

4.5.2. Hipótesis

El índice de radiación UV es un índice que permite saber la exposición a la luz ultravioleta del sol en un determinado lugar. La exposición a la radiación solar puede producir, en el ser humano, efectos agudos y crónicos en la salud de la piel, los ojos y el sistema inmunitario. Es frecuente la creencia, equivocada, de que sólo las personas de piel clara deben preocuparse por la sobre exposición al sol. Las pieles más oscuras contienen más melanina protectora y la incidencia de cáncer de piel es menor en personas con este tipo de piel. Sin embargo, se producen casos de cáncer de piel en estas personas y, por desgracia, estos cánceres a menudo se detectan en estadios más avanzados y más peligrosos. El riesgo de efectos sobre la salud ocular y del sistema inmunitario relacionados con la radiación UV es independiente del tipo de piel.

Debido a todos estos factores negativos que tiene la exposición a altos índices de radiación UV, quisimos realizar una experimentación para ver si los países con altos valores de radiación UV tienen menor expectativa de vida.

Luego de hacer el análisis exploratorio en la sección anterior, nos planteamos la siguiente hipótesis como guía para realizar la experimentación.

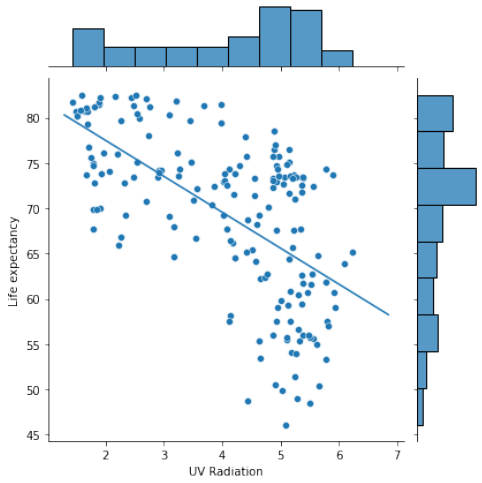
> Los países con alto índice de radiación UV poseen una baja expectativa de vida.

4.5.3. Análisis

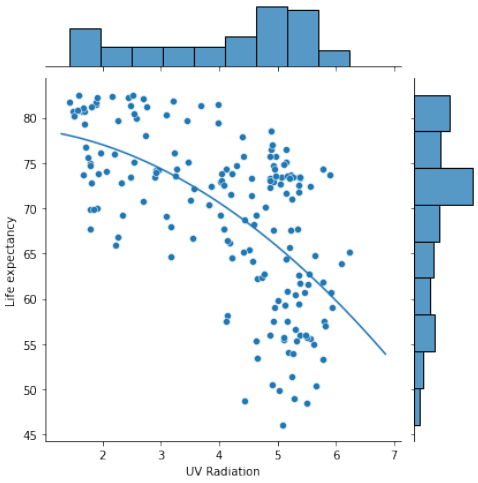
Nuevamente comenzamos el análisis donde lo habíamos dejado en la sección de análisis exploratorio. En aquella sección habíamos concluido que, mirando la figura 33 a ojo, teníamos la intuición de que tenían una relación exponencial negativa con la expectativa de vida. Ahora con la hipótesis planteada, vamos a ver si esto es cierto o simplemente es un fraude visual.

En la figura 49 podemos ver 4 gráficos que representan las funciones de distintos tipos que mejor aproximan los puntos utilizando la técnica de regresión lineal. Con una mirada por arriba de los gráficos podemos ver que no hay ninguna función que parezca representar fielmente los puntos originales. Sin embargo, la función lineal y la

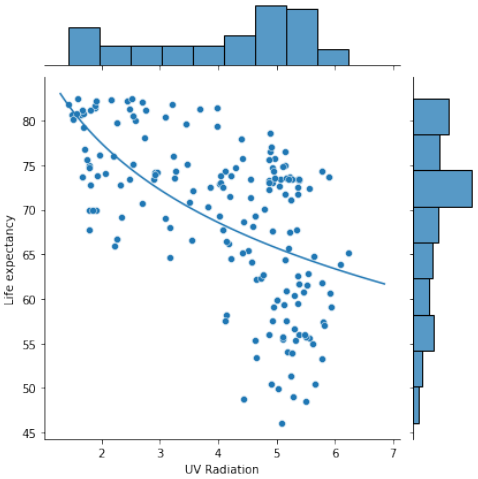
cuadrática parecen ser las que mejor representarían a los puntos originales, pero nuevamente, necesitamos un indicador que nos confirme la variación que hay entre los puntos reales y las funciones generadas por la regresión lineal.



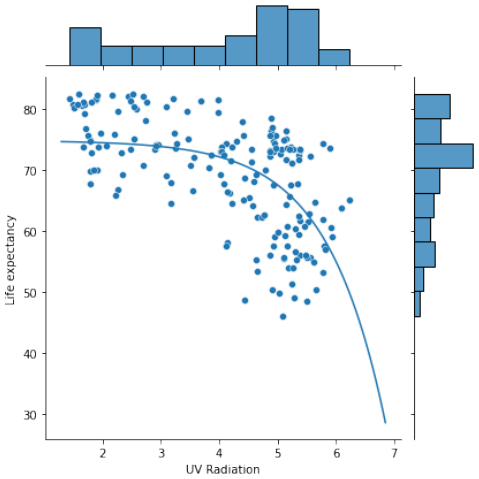
(a) Función lineal que mejor aproxima a los puntos



(b) Función cuadrática que mejor aproxima a los puntos



(c) Función logarítmica que mejor aproxima a los puntos



(d) Función exponencial que mejor aproxima a los puntos

Figura 49: Funciones que mejor aproximan los puntos de la expectativa de vida y el índice de radiación UV

Para ello, utilizaremos el R^2, indicador que explicamos anteriormente en la sección regresión lineal. El calculo del R^2 lo podemos observar en el cuadro 1, el cual nos muestra que ninguna función aproxima bien a los puntos. Por lo tanto, lo que veíamos a ojo y la hipótesis no se condicen con la información que nos dan los datos.

	R^2
Función Lineal	-0.73
Función Cuadrática	-10544
Función Logarítmica	-3.62
Función exponencial	-24212

Cuadro 1: Tabla que muestra el R^2 de las funciones que mejor aproximan a los puntos.

4.5.4. Conclusiones

Debido a los resultados obtenidos, creemos que hay 3 posibilidades:

- La hipótesis es falsa.
- Existe algún tipo de función que aproxima mejor a los puntos.
- El dataset no tenia suficiente información para experimentar.

Para responder esta duda se debería realizar más experimentos, como por ejemplo, generar muchísimos mas modelos de regresión lineal para tener una familia de funciones mucho más grande y tratar de encontrar un tipo de función que aproxime mejor a los puntos. También se podría concluir con algún otro experimento de que la hipótesis es falsa. Debido a los efectos de este trabajo practico, dejaremos para una próxima investigación el desarrollo de estos experimentos.

Referencias

[1] Técnicos del FMI aconsejan subir 5 años la edad jubilatoria y reducir el haber inicial - Ámbito
https://www.ambito.com/economia/fmi/tecnicos-del-aconsejan-subir-5-anos-la-edad-jubilatoria-y-reducir-el-haber-inicial-n5011304

[2] Italia, el país que devoró a su juventud - El Confidencial
https://www.elconfidencial.com/mundo/2017-12-13/el-pais-que-acabo-con-sus-jovenes_1491382/

[3] Banco de datos libre
https://data.worldbank.org/

[4] Países donde está prohibido el alcohol
https://intriper.com/estos-son-los-paises-en-los-que-el-alcohol-esta-prohibido-y-aun-es-un-tabu/

[5] Índice UV
https://es.wikipedia.org/wiki/Indice_UV

[6] Sistema de educación de Botsuana - Wikipedia
https://es.wikipedia.org/wiki/Botsuana

[7] Spurious Correlations
https://tylervigen.com/spurious-correlations

[8] World Health Organization - Global Health Observatory data repository
https://apps.who.int/gho/data/node.home