

ESTUDIO DEL NIVEL EDUCATIVO ALCANZADO EN FUNCIÓN DE VARIABLES ECONÓMICAS, INDUSTRIALES Y SOCIALES.

CICLO DE VIDA DE LOS DATOS.

Curso 2021/2022



FACULTAD DE CIENCIAS
MÁSTER EN CIENCIA DE DATOS

Andrés Álvarez Carou
Miguel Arbea Gómez
Daria Ágelica Escobar Galaburda
Julia Fáfrega Torrano
Denisse Gómez González
Jesús Octavio Raboso

Santander, 05 de febrero de 2022

Índice general

1. Introducción	1
2. Planificación e implementación	3
2.1. Logframe	3
2.2. Recolección de datos	3
2.3. Diagram de Gantt	4
2.4. Paquetes de trabajo	5
2.5. Hitos	9
2.6. Entregables	9
2.7. Riesgos y contingencias	10
2.8. Stakeholders	10
2.9. Presupuesto	11
2.9.1. Recursos humanos	11
2.10. Preservación de los datos	11
2.10.1. Almacenamiento y copia de seguridad durante el proceso de investigación . .	11
2.10.2. Requisitos legales y éticos	11
2.10.3. Intercambio de datos y preservación a largo plazo	12
3. Nivel educativo alcanzado.	13
3.1. Education at a glance.	13
3.2. Contexto.	14
3.3. Estructura del dataset.	14
3.4. Metadatos.	16
3.5. Análisis preliminar.	18
4. Gastos en educación	23
4.1. Contexto	23
4.2. Estructura del dataset	23
4.3. Metadatos	25
5. Emisiones de CO2 per cápita.	27
5.1. Contexto.	27
5.2. Estructura del dataset.	27
5.3. Metadatos.	28
5.4. Análisis preliminar.	29
6. Vacunas	33
6.1. Contexto	33
6.2. Estructura del dataset	33
6.3. Metadatos	34

7. Curación de datos y unificación de datasets.	37
7.1. Descripción del proceso de curación.	37
7.1.1. Países en común e identificación.	38
7.1.2. Años con datos disponibles y formatos de fechas.	40
7.1.3. Datos vacíos y funciones de agregado.	41
7.1.4. Transformación de variables.	42
7.1.5. Formatos de tablas no adecuados.	42
7.1.6. Datos y nombres de variables duplicados o diferentes.	45
7.2. Resultado de la curación. Dataset final.	45
8. Análisis	47
8.1. Objetivos del trabajo	47
8.2. Ajuste del modelo de regresión lineal múltiple	48
8.3. Validación de hipótesis	50
9. Conclusiones y trabajo futuro.	55
Anexo	57
Bibliografía	57

Capítulo 1

Introducción

La educación es una de las cuestiones más importantes de un país. Gracias a ella, las personas adquieren conocimientos, no solo sobre las distintas materias que se imparten, sino también valores y cultura. Además, un país que cuenta con un buen nivel educativo reduce la desigualdad económica y social porque permite a todos los individuos tener las mismas oportunidades de formarse y tener más oportunidades de conseguir un empleo que permita tener un buen nivel de vida. El Gobierno de España es consciente de esta situación y quiere mejorar el nivel educativo del país. Quiere conocer cuáles son las causas principales que provocan que un país tenga mejor o peor nivel educativo. Por esta razón, el Gobierno de España ha contratado a la empresa AnalysisEsp. Nuestra empresa es una consultora que se encarga de llevar a cabo análisis estadísticos sobre las distintas cuestiones que se solicitan. En este proyecto, la variable objetivo es el nivel educativo de distintos países repartidos a lo largo del globo y se llevará a cabo un análisis que cuenta con distintas variables predictoras para averiguar cuáles de ellas tienen mayor impacto en la variable objetivo.

De este estudio se espera encontrar cierta relación entre tres variables predictoras (gasto en educación, confianza en vacunas y emisiones de CO₂) con la variable objetivo, el nivel educativo. Por nuestra parte, es de esperar que el gasto educativo tenga una gran relevancia, donde creemos que cuando mayor sea la inversión en educación mayor será el nivel educativo de la población. Por otra parte, nos resulta de interés estudiar el impacto de las variables confianza en vacunas y emisiones de CO₂ con el fin de determinar si la salud y el entorno de la población también afecta a nuestra variable objetivo.

Capítulo 2

Planificación e implementación

2.1. Logframe

En la tabla 2.1 se muestra el LogFrame del proyecto. Éste plasma los objetivos y resultados junto con sus fuentes de verificación y los posibles riesgos asociados, además de las actividades y los plazos generales.

2.2. Recolección de datos

Los conjuntos de datos asociados a cada variable del estudio han sido recabados de distintos modos.

Por un lado, para analizar la confianza en las vacunas hemos realizado una encuesta a la población de los diferentes países tenidos en cuenta en el estudio, vía online, telefónica o incluso presencial en los países menos accesibles. En ella recogimos datos como su posicionamiento sobre las vacunas, su país de residencia, sexo, edad o la fecha de realización de la encuesta. (Realmente hemos empleado los datos recogidos en el siguiente artículo de la revista The Lancet: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(20\)31558-0/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31558-0/fulltext))

Por otro lado, los datos de las otras tres variables implicadas (nivel educativo, gasto en educación y emisiones de CO₂) han sido obtenidos de fuentes fiables de cada país. En el caso de no haber encontrado fuentes de confianza o no tener una cantidad de datos suficiente o legible, nos hemos puesto en contacto con entidades del país con el fin de obtener tales datos. (Realmente, para analizar el nivel educativo alcanzado hemos empleado los datos recogidos por la OCDE en https://www.oecd-ilibrary.org/education/adult-education-level/indicator/english_36bce3fe-en . Desglosa el nivel educativo alcanzado (below upper secondary, tertiary upper secondary) en ciertos países y permite distinguir por año, sexo etc. Mientras que para los datos del gasto educativo utilizamos los recogidos por la OCDE en <https://data.oecd.org/eduresource/education-spending.htm#indicator-chart>. Nuevamente, permite distinguir el gasto por países y por etapa/nivel educativo. Por último, para analizar las emisiones de CO₂ por persona hemos usado los datos recogidos en <https://ourworldindata.org/co2-emissions#per-capita-co2-emissions>.

	Objetivos	Indicadores de éxito	Fuentes de verificación	Suposiciones, condiciones, riesgos
Meta	Que el Gobierno de España tenga en conocimiento las variables que hacen que el país tenga un cierto nivel Educativo.	- Puesta en marcha de las medidas necesarias para la mejora del nivel educativo en el país	- Informes publicados por el Gobierno con datos acerca del nivel educativo.	- Interés del Gobierno de España
Propósito	Encontrar las variables que causan un mayor o menor nivel educativo en un país.	- Obtención de conclusiones fiables.	- Informe con las conclusiones del análisis.	- Variables utilizadas son explicativas.
Resultados	Conseguir todos los datos necesarios y haber efectuado el análisis.	- Obtención de las variables necesarias para llevar a cabo el análisis. - Exposición del análisis al resto el equipo.	- Datasets con todas las variables - Ficheros con los análisis realizados	- No encontrar fuentes de datos. fiables - Fallo en la web donde se suba la encuesta - Que la encuesta no sea respondida por la población de un país. - Retraso en el tiempo de respuesta de la población.
Componentes	- Recopilación de datos - Análisis - Exposición de resultados	Presupuesto: - Coste de personal. - Plan de contingencia. Calendario 2022: - Abr-Ago: Recogida datos. - Jun-Ago: Curado. - Ago-Oct: Análisis. - Anualmente: Actualización.	- Presupuesto - Bases de datos. - Informes	- Contratación por parte del Gobierno. - Financiación adecuada - Participación de la población.

Tabla 2.1: Logframe del proyecto.

2.3. Diagram de Gantt

Con el fin de planificar y programar tareas del proyecto, presentamos un diagrama de Gantt. En él exponemos el tiempo de dedicación previsto para las diferentes tareas o actividades involucradas a lo largo del tiempo.

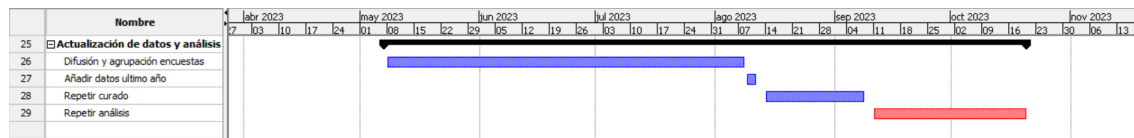
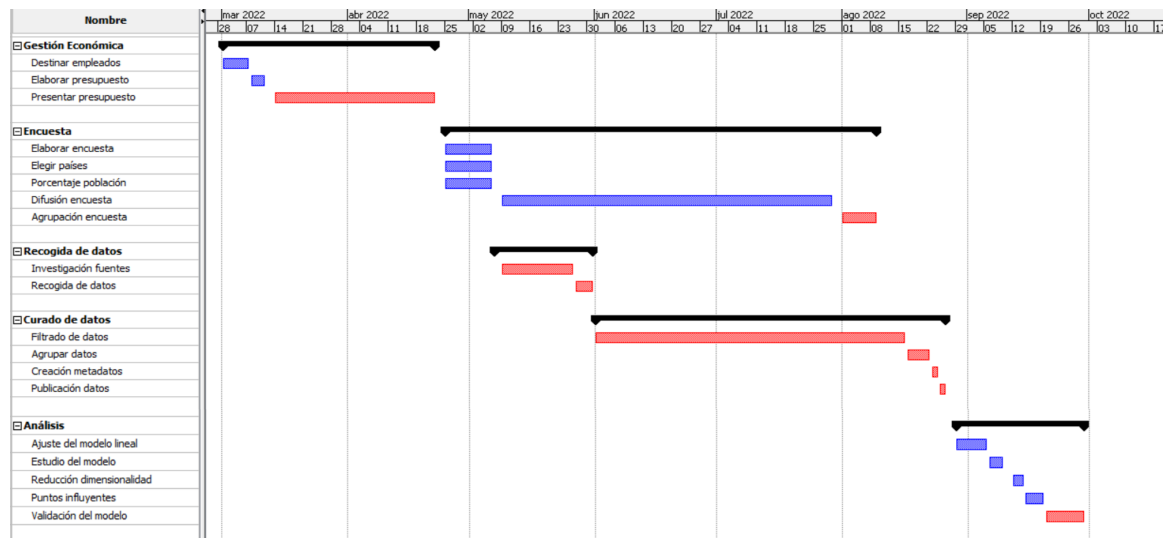


Figura 2.1: Diagrama de Gantt del proyecto.

2.4. Paquetes de trabajo

A continuación, se muestran los paquetes de trabajo en los que se divide el proyecto. Cada uno de ellos muestra el periodo de tiempo en el que se prevé llevar a cabo, el personal necesario, los objetivos, las tareas y los entregables.

Work Package 1	
GESTIÓN ECONÓMICA	
Comienzo del WP: 01 Marzo 2022	Finalización del WP: 22 Abril 2022
Participantes Graduado de relaciones laborales	
Objetivos 1. Personal necesario para llevar a cabo el proyecto. 2. Tener un presupuesto para el proyecto.	
Descripción del trabajo Tarea 1.1. Destinar a los empleados que se van a encargar del proyecto. Se asignarán unos trabajadores de la empresa para realizar el proyecto . Tarea 1.2. Elaborar el presupuesto del proyecto. Estimación de los gastos de personal e infraestructura. Tarea 1.3. Presentar el presupuesto al Gobierno.	
Entregables Entregable 1.1. Presupuesto.	

Tabla 2.2: Descripción de la duración, personal, objetivos, tareas y entregables correspondientes al paquete de trabajo WP1: Gestión económica.

Work Package 2	
ENCUESTA	
Comienzo del WP: 25 Abril 2022	Finalización del WP: 9 Agosto 2022
Participantes Enfermero Ingeniero informático 1 Data Scientist 1	
Objetivos 1. Tener la encuesta elaborada. 2. Tener respondida la encuesta.	
Descripción del trabajo Tarea 2.1. Elaboración de la encuesta. Pensar las preguntas de la encuesta de forma que se pueda analizar la confianza en las vacunas por parte de la población de distintos países de forma anónima. Tarea 2.2. Decidir los países a los que realizar la encuesta. Tarea 2.3. Decidir a qué porcentaje de la población realizar la encuesta. Tarea 2.4. Difusión de la encuesta. Difusión a través de internet. En caso de no recibir las respuestas necesarias, contactar con algún organismo del país que se encargue de hacer las preguntas a la población directamente. Tarea 2.5. Agrupación de encuestas. Agrupar en la misma base de datos todas las encuestas.	
Entregables Entregable 2.1. Encuesta elaborada. Entregable 2.2. Base de datos con las respuestas de la encuesta.	

Tabla 2.3: Descripción de la duración, personal, objetivos, tareas y entregables correspondientes al paquete de trabajo WP2: Encuesta.

Work Package 3	
RECOGIDA DE DATOS	
Comienzo del WP: 7 Mayo 2022	Finalización del WP: 31 Mayo 2022
Participantes Ingeniero informático 2	
Objetivos 1. Recopilación de los datos anuales relativos al nivel educativo, gasto en educación por niveles y emisiones de CO2 por persona .	
Descripción del trabajo Tarea 3.1. Investigación de las fuentes de datos. Buscar información en distintas bases de datos de los países para confirmar que se recogen los datos oficiales. Tarea 3.2. Recogida de los datos. Descargar los datasets de las fuentes encontradas de cada país.	
Entregables Entregable 3.1. Bases de datos de los niveles educativos de cada país por separado. Entregable 3.2. Bases de datos del gasto en educación de cada país por separado. Entregable 3.3. Bases de datos de las emisiones de CO2 por persona de cada país por separado.	

Tabla 2.4: Descripción de la duración, personal, objetivos, tareas y entregables correspondientes al paquete de trabajo WP3: Recogida de datos.

Work Package 4	
CURADO DE DATOS	
Comienzo del WP: 1 Junio 2022	Finalización del WP: 28 Agosto 2022
Participantes Ingeniero informático 2 Data Scientist 2	
Objetivos 1. Curación de los datos.	
Descripción del trabajo Tarea 4.1. Filtrado de datos. Eliminar los datos dudosos y los innecesarios. Cambiar el formato de los datos que se considere necesario. Tarea 4.2. Agrupar los datos en la misma base de datos. Crear un único dataset que contenga los datos de todas las variables empleadas. Tarea 4.3. Creación de metadatos. Tras crear un único dataset, crear los metadatos que lo expliquen. Tarea 4.4. Publicación de los datos . Subir los datos a la nube.	
Entregables Entregable 4.1. Bases de datos correctamente agrupadas y unificadas. Entregable 4.2. Identificador persistente o dirección de los datos publicados.	

Tabla 2.5: Descripción de la duración, personal, objetivos, tareas y entregables correspondientes al paquete de trabajo WP4: Curado de datos.

Work Package 5	
ANÁLISIS DE DATOS	
Comienzo del WP: 29 Agosto 2022	Finalización del WP: 29 Septiembre 2022
Participantes Data Scientist 1 Data Scientist 2	
Objetivos 1. Obtener el análisis de los datos y las conclusiones.	
Descripción del trabajo Tarea 5.1. Ajuste del modelo lineal. Tarea 5.2. Estudio del modelo. Análisis del ajuste. Tarea 5.3. Reducción de la dimensionalidad. Supresión de las variables mediante el criterio global AIC. Tarea 5.4. Búsqueda de puntos influyentes. Búsqueda de aquellos puntos que pueden ejercer más influencia que otros en el ajuste. Tarea 5.5. Validación del modelo. Es la validación del modelo con las hipótesis de homocedasticidad, normalidad y linealidad.	
Entregables Entregable 5.1. Informe en el que se incluyen gráficas y resultados de código, además de las conclusiones del proyecto.	

Tabla 2.6: Descripción de la duración, personal, objetivos, tareas y entregables correspondientes al paquete de trabajo WP5: Análisis de datos.

Work Package 6	
ACTUALIZACIÓN DE DATOS Y ANÁLISIS	
Fechas: Anualmente durante 4 meses hasta la finalización del proyecto	
Participantes Data Scientist 3	
Objetivos 1. Comprobar que las conclusiones obtenidas con anterioridad siguen vigentes en el tiempo.	
Descripción del trabajo Tarea 6.1. Difusión y agrupación de encuestas Difundir la encuesta para obtener los datos del último año y agrupar las respuestas en la misma base de datos. Tarea 6.2. Añadir los datos correspondientes al último año. Tarea 6.3. Repetir las tareas de curado de datos. En el caso de los metadatos únicamente habría que actualizarlos. Tarea 6.3. Repetir las tareas de análisis. Volver a realizar un análisis de los resultados obtenidos con el fin de ver si son los mismos que años anteriores.	
Entregables Entregable 6.1. Nuevo informe con las conclusiones del proyecto actualizadas.	

Tabla 2.7: Descripción de la duración, personal, objetivos, tareas y entregables correspondientes al paquete de trabajo WP6: Actualización de datos y análisis.

2.5. Hitos

Hito	Descripción	WPs	Fecha de entrega
Datos	Disponer de todos los datos	2,3	9/08/22
Datos curados	Datos filtrados y agrupados	4	26/08/22
Conclusiones	Conclusiones extraídas del análisis	5	29/09/2022

Tabla 2.8: Hitos del proyecto.

2.6. Entregables

Entregable	WP	Título	Clase	Tipo	Fecha de entrega
E1.1	1	Presupuesto	Informe	Interno	11/03/2022
E2.1	2	Encuesta elaborada	Encuesta	Interno	06/05/2022
E2.2	2	Base de datos con las respuestas de la encuesta	Base de datos	Interno	09/08/2022
E3.1	3	Base de datos de los niveles educativos por países	Base de datos	Interno	31/05/2022
E3.2	3	Base de datos del gasto en educación por países	Base de datos	Interno	31/05/2022
E3.3	3	Base de datos de las emisiones de CO2 por persona en cada país	Base de datos	Interno	31/05/2022
E4.1	4	Base de datos unificada	Base de datos	Público	22/08/2022
E4.2	4	Identificador persistente	Identificador	Público	24/08/2022
E5.1	5	Informe con las conclusiones del proyecto	Informe	Público	29/09/2022
E6.1	6	Nuevo informe con las conclusiones del proyecto actualizadas	Informe	Público	Anual

Tabla 2.9: Resumen de entregables de los paquetes de trabajo.

2.7. Riesgos y contingencias

Riesgos	Probabilidad	Impacto	WPs	Plan de contingencia
Fallo en la web donde se suba la encuesta	Baja	Alto	2	El informático soluciona el problema
Que la encuesta no sea respondida por la población de algún país	Baja	Alto	2	Contratar a alguien del país para que realice la encuesta presencialmente.
Solo encontrar fuentes de datos no fiables o no encontrar suficientes datos	Media	Alto	3	Contactar con el país del que no se encuentren datos fiables para conseguir buenos datos. De no ser posible, excluir al país del estudio.
Metadatos no adecuados o incompletos	Media	Bajo	4	Completar aquello que falta y en el peor de los casos rehacerlos
Formatos de variables en los datasets diferentes	Alta	Bajo	4	Unificación de los formatos
Que el tiempo de respuesta de las encuestas por parte de población sea mayor del esperado	Media	Medio	Todos	Posponer el resto de tareas que dependen de la consecución de esta.
Cancelación del proyecto por parte del Gobierno	Baja	Alto	Todos	Proponer el proyecto a otros países del estudio.

Tabla 2.10: Riesgos de los paquetes de trabajo.

2.8. Stakeholders

Stakeholder	Interés para el proyecto	Apoyo necesario	Apoyo esperado
Población encuestada	Aporta información sobre el nivel de confianza en las vacunas en su país	Alto	Medio
Gobierno de España	Es la entidad que contrata a la empresa	Alto	Alto
Instituto Nacional de Estadística de otros países	Aportan datos acerca de las variables estudiadas en caso de que sea necesario	Bajo	Medio
AnaylisisEsp	Empresa contratada para llevar a cabo el proyecto	Alto	Alto
Público interesado	Público que puede verse afectado por los resultados del proyecto	Medio	Medio

Tabla 2.11: Diferentes stakeholders del proyecto.

2.9. Presupuesto

2.9.1. Recursos humanos

Teniendo en cuenta los diferentes profesionales involucrados en cada paquete de trabajo, sus horas invertidas en el proyecto y su salario/hora, hemos calculado el presupuesto del estudio (suponemos un trabajo de 8 horas diarias).

Profesión	Horas	Pago por hora (€)	Coste total (€)
Graduado en relaciones laborales (RRL)	312	12	3.744
Enfermero (E)	80	13,5	1.080
Ingeniero Informático 1 (IN1)	536	14	7.504
Ingeniero Informático 2 (IN2)	184	14	2.576
Data Scientist 1 (DS1)	176	16	2.816
Data Scientist 2 (DS2)	586	16	9.376
Data Scientist 3 (DS3)	960	16	15.360
Coste Total RRHH del proyecto			42.456

Tabla 2.12: Costes humanos de los paquete de trabajo.

Además del presupuesto obtenido, hemos previsto un fondo económico para solucionar las posibles contingencias del proyecto. Este fondo es de unos 5.000€, y algunas de las posibles contingencias planteadas son la solución de problemas técnicos de la web en la que se encuentra las encuestas de las vacunas por parte de un ingeniero informático, o la contratación de personal en ciertos países para la realización presencial de estas encuestas en caso de ser necesario.

2.10. Preservación de los datos

2.10.1. Almacenamiento y copia de seguridad durante el proceso de investigación

Durante las actividades de investigación los datos se almacenarán en GitHub y la copia de seguridad se realizará cada semana. En caso de incidente los datos se recuperarán de la copia de seguridad efectuada en el disco duro. Durante la investigación, únicamente tendrá acceso a los datos el personal que lleva a cabo el proyecto. Para garantizar este acceso, el repositorio de GitHub donde se almacenan los datos es privado

2.10.2. Requisitos legales y éticos

Una vez finalizado el estudio, en aquellos datasets en los que se procesan datos personales (por ejemplo, los datos recogidos durante las encuestas de las vacunas), se garantizará la anonimidad de las personas con el fin de proteger y salvaguardar estos datos personales de terceros. Así, una vez estos sean publicados y se tenga un acceso abierto a ellos, estos podrán ser usados sin restricciones asegurando su confidencialidad.

Para su acceso se considerara una licencia Creative Commons, con el fin de permite a los usuarios la reutilización de los datos sin solicitar el permiso del autor de la obra.

2.10.3. Intercambio de datos y preservación a largo plazo

Los datos serán compartidos públicamente una vez finalizado el proyecto. No tendrán restricciones de uso y estarán sujetos bajo la licencia Creative Commons.

Se podrá acceder a ellos a partir de dos repositorios públicos: uno en el portal de Github y el otro en la plataforma Zenodo. El formato de los datasets es csv, por lo tanto los usuarios que estén interesados en su reutilización no necesitarán herramientas específicas para acceder a ellos.

Capítulo 3

Nivel educativo alcanzado.

La Organización para la Cooperación y el Desarrollo Económico (OCDE) es un Organismo Internacional de carácter intergubernamental del que forman parte 38 países miembros. La OCDE fue creada en 1960 para dar continuidad y consolidar el trabajo realizado por la antigua Organización Europea de Cooperación Económica (OECE). La OCDE asumió la tarea de impulsar la reconstrucción del continente tras la Segunda Guerra Mundial.

Se trata de un foro en el que los Gobiernos de los Estados miembros (democracias con una economía de mercado) trabajan juntos con el fin de afrontar los desafíos económicos, sociales y de buenas prácticas de gobierno para, de ese modo, aprovechar con más eficiencia las nuevas oportunidades y coordinar políticas locales e internacionales.

Los objetivos marcados en su fundación siguen aún hoy vigentes:

- (I) Lograr la máxima expansión posible de la economía y del empleo y aumentar la calidad de vida en los países miembros, manteniendo la estabilidad financiera y contribuyendo así al desarrollo de la economía mundial.
- (II) Contribuir a la sana expansión económica de los países miembros y en los países no miembros en vías de desarrollo.
- (III) Contribuir a la expansión del comercio mundial sobre una base multilateral y no discriminatoria conforme a las obligaciones internacionales.

Por tanto, el trabajo de la OCDE se centra tanto en el análisis del Desarrollo Económico y Social como en las políticas que influyen en el mismo. Es por esto que su ámbito de actividad no abarca sólo la actividad económica sino que aborda las cuestiones sociales, medioambientales, energéticas, sanitarias... Por supuesto, también trata el tema que nos atañe a nosotros: **la educación**.

Desde un punto de vista práctico, los trabajos que realiza la OCDE se traducen en informes y recomendaciones para los Gobiernos de los Estados miembros así como para otros países no miembros que siguen con interés el trabajo de la institución.

Para más información sobre la OCDE, véase [\[Asu\]](#), [\[EOa\]](#).

3.1. Education at a glance.

En el marco educativo, el informe *Education at a glance* es la fuente de información de la OCDE sobre el estado de la educación en todo el mundo [\[EOc\]](#).

Se trata de una publicación anual que proporciona una amplia colección de datos estadísticos y análisis comparando el funcionamiento de los sistemas educativos de los países de la OCDE y

países asociados. Proporciona datos sobre la estructura, la situación financiera y el desempeño de los sistemas educativos

Entre otras variables, se analiza el acceso, participación y progreso de los integrantes del sistema educativo; los recursos económicos invertidos; el entorno de aprendizaje; la organización de los centros educativos... Del mismo modo, se aborda cómo la educación influye en el mercado laboral y cómo este se ve afectado por el género, el estado socioeconómico y la ubicación regional.

Dentro del informe *Education at a glance*, [OEC21] uno de los indicadores que se recoge es el *Adult education level*. Este constituirá nuestro *dataset objetivo*.

El indicador *Adult education level* se define como el nivel de educación más alto alcanzado por la población entre 25 y 64 años. Existen tres niveles: *below upper-secondary* [EOd]; *upper secondary* [EOe]; *tertiary education* [EOf]. Posteriormente, se analizarán con detalle.

Este indicador se mide el porcentaje del total de la población adulta entre 25 y 64 años que ha alcanzado cierto nivel educativo. El nivel educativo alcanzado se define como el último ciclo educativo completado. Los datos también están disgregados por género.

3.2. Contexto.

La igualdad de oportunidades a la hora de obtener una educación de calidad debe formar parte del contrato social. Dos factores críticos que suponen desigualdades bastante acentuadas son la situación socio-económica y la capacidad de movilidad.

El nivel educativo alcanzado se utiliza habitualmente como una medida indirecta del capital humano y como medidor de las habilidades o capacidades de una persona. Es decir, sirve para medir las habilidades o capacidades asociadas a un nivel educativo y disponibles en la población activa.

Generalmente, un alto nivel educativo se asocia con indicadores económicos y sociales positivos. Habitualmente se piensa que las personas con alto nivel educativo están más comprometidas socialmente. Del mismo modo, parece que suelen tener menor tasa de desempleo y mayores ingresos. También se relaciona con más implicación en la educación reglada y no reglada para con sus iguales.

Es por ello que las personas parecen tener incentivos para alcanzar un mayor nivel educativo y los gobiernos, alientes para proporcionar inversión, política e infraestructura que apoye niveles más altos de educación en toda la población.

Durante las últimas décadas, casi todos los países de la OCDE han experimentado un aumento significativo del nivel educativo, especialmente entre los jóvenes y las mujeres,.

3.3. Estructura del dataset.

El dataset [EOb] que recoge el indicador *Adult education level* se encuentra disponible en [ocde.dataset.attainment].

A simple vista, el dataset es de la forma:

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
0	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1989	44.650639	NaN
1	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1991	44.127056	NaN
2	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1993	47.159046	NaN
3	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1994	49.802025	NaN
4	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1995	44.935852	NaN

Figura 3.1: Cabecera del dataset [EOb].

Observamos que el dataset contiene 8 variables:

- (I) **LOCATION**: Se corresponde con los códigos ISO 3166-1 alpha-3. Son los códigos de tres letras para los país definidos en la ISO 3166-1, parte de la norma ISO 3166 pulicada por la Organización Internacional de Normalización (ISO). Sirven para representar países, territorios dependientes y zonas especiales de interés geográfico. Existen 249 códigos ISO 3166-1 alfa-3 asignados oficialmente en la actualidad. En este dataset se recogen datos de 46 países más los datos del G20 (relativo a los países del G20) y OAVG (media de la OCDE) Algunos ejemplos son:
- AUS: Australia
 - LUX: Luxemburgo
 - G20: Media de los países del G20
 - OAVG: Media de los países de la OCDE
- (II) **INDICATOR**. Muestra el indicador evaluado en el dataset. En este caso, toma un único valor ('EDUADULT') que se corresponde con el nivel educativo alcanzado por la población adulta entre 25 y 64 años.
- (III) **SUBJECT**. Toma 7 valores relacionados con el nivel educativo alcanzado y el género:
- BUPPSRY: Below upper secondary. Se corresponde con el nivel Lower secondary education [EOd].
 - TRY: Tertiary. Se corresponde con el nivel universitario [EOf]
 - UPPSRY: Upper secondary. Corresponde a la etapa final de la educación secundaria en la mayoría de los países de la OCDE. La edad de inicio de sitúa en los 15 o 16 años [EOe].
 - TRY_MEN: Tertiary education for men
 - TRY_WOMEN: Tertiary education for women
 - UPPSRY_MEN: uppersecondary education for men
 - UPPSRY_WOMEN: Upper secondary education for women
- (IV) **MEASURE**. Población sobre la que se están midiendo los resultados. Toma un único valor (PC_25_64) que se corresponde con la población adulta ente 25 y 64 años.
- (V) **FREQUENCY**. Frecuencia de medición. Toma un único valor (A) que indica que los datos son recogidos anualmente.
- (VI) **TIME**. Años en los que se han recogido datos. Se abarcan 40 años: desde 1981 hasta 2020.
- (VII) **Value**. Para cada fila, porcentaje de población adulta entre 25 y 64 años que ha alcanzado el nivel educativo descrito en el valor SUBJECT de su misma fila en el año indicado por la columna TIME en su misma fila.
- (VIII) **Flag Codes**. Un *flag* se corresponde con un atributo que representa información cualitativa sobre el valor de la propia celda. Sólo toma valores nulos. No nos aporta información.

En cuanto a los tipos de cada una de las variables:

```

: LOCATION      object
  INDICATOR      object
  SUBJECT        object
  MEASURE        object
  FREQUENCY      object
  TIME           int64
  Value          float64
  Flag Codes     float64
dtype: object

```

Figura 3.2: Tipos de las columnas dataset [EOB].

3.4. Metadatos.

Dublin Core, también conocido como *Dublin Core Metadata Element Set* es un estándar de metadatos elaborado y promovido por la DCMI (Dublin Core Metadata Initiative), una organización dedicada a fomentar la adopción de los estándares interoperables de los metadatos y a promover el desarrollo de los vocabularios especializados para describir recursos y así permitir su difusión e interoperabilidad [Pow+07].

El fin de los estándares de metadatos es promover los principios FAIR [dat]:

- (I) **FINDABLE**. Los datos y metadatos pueden ser encontrados por la comunidad después de su publicación mediante herramientas de búsqueda estándares. Para ello, algunas de las acciones más adecuadas pasan por asignarles un identificador único y persistente tanto a los datos como a los metadatos, describir los datos con metadatos de manera adecuada, registrar los datos y metadatos en un recurso de búsqueda.
- (II) **ACCESIBLE**. Los datos y metadatos deben ser accesibles mediante sus identificadores y protocolos estándar que sean abiertos y de implementación universal.
- (III) **INTEROPERABLE**. Tanto los datos como los metadatos deben estar descritos según las reglas de la comunidad y empleando estándares abiertos para favorecer su intercambio y reutilización.
- (IV) **REUSABLE**. Los datos y metadatos han de poder ser reutilizados por otras personas al dejar clara su fuente y las condiciones de reutilización, pues almacenan atributos relevantes. Su licencia y derechos deben ser concisos.

Algunos de los estándares de metadatos más conocidos son DataCite o DublinCore. Nos centraremos en el segundo.

DublinCore es un sistema de 15 definiciones semánticas descriptivas que abarcan elementos relacionados con el contenido, elementos relacionados con la propiedad intelectual y elementos relacionados con la instanciación del recurso que describen.

Las 15 etiquetas que forman el formato Dublin Core son:

- **Title**. Nombre dado a un recurso.
- **Creator**. Persona u organización responsable de la creación del contenido intelectual del recurso.
- **Subject**. Palabras clave o frases que describen el título o el contenido del recurso.
- **Description**. Descripción textual del recurso.
- **Contributor**. Persona u organización que haya tenido una contribución intelectual significativa.

- **Date.** Fecha en la cual el recurso se puso a disposición del usuario en su forma actual.
- **Type.** Categoría del recurso.
- **Format.** Formato de datos de un recurso, usado para identificar el software y, posiblemente, el hardware que se necesitaría para mostrar el recurso.
- **Identifier.** Secuencia de caracteres utilizados para identificar unívocamente un recurso.
- **Source.** Secuencia de caracteres usados para identificar unívocamente un trabajo a partir del cual proviene el recurso actual.
- **Language.** Idioma/s del contenido intelectual del recurso.
- **Relation.** Identificador de un segundo recurso y su relación con el recurso actual. Este elemento permite enlazar los recursos relacionados y las descripciones de los recursos.
- **Coverage.** Cobertura espacial y/o temporal del contenido intelectual del recurso.
- **Rights.** Referencia sobre los derechos de autor y propiedad intelectual.

Para la obtención de metadatos con Python, empleamos el módulo **request**

Para la obtención de los metadatos del propio dataset [\[EOB\]](#), podemos emplear dos resolvedores ya que el identificador DOI se basa en HANDLE. Pueden consultarse los notebooks.

En ambas ocasiones, la respuesta nos es dada en formato JSON, de modo que es fácil convertirla en un diccionario y acceder a sus campo. Si bien, esto no siempre ocurre [5.3](#) Para más información, consúltese el notebook *education_attainment.ipynb*

Por tanto, los metadatos en formato DublinCore resultan:

- **Title.** Education attainment // Adult Education Level (Indicator).
- **Creator.** Organisation for Economic Cooperation and Development (OECD)
- **Subject.** Education.
- **Description.** This indicator looks at adult education level as defined by the highest level of education completed by the 25-64 year-old population. There are three levels: below upper-secondary, upper secondary and tertiary education. Upper secondary education typically follows completion of lower secondary schooling. Lower secondary education completes provision of basic education, usually in a more subject-oriented way and with more specialised teachers. The indicator is measured as a percentage of same age population; for tertiary and upper secondary, data are also broken down by gender.
- **Contributor.** Countries considered in the study.
- **Date.** 25-11-2020.
- **Type.** Dataset.
- **Format.** CSV.
- **Identifier.** DOI: 10.1787/36bce3fe-en
- **Source.** Crossref. OCDE Education at a Glance.
- **Language.** English.
- **Relation.** [\[OEC21\]](#)

- **Coverage.** Countries (ISO 3166-1 alpha-3) AUS, AUT, BEL, CAN, CZE, DNK, FIN, FRA, DEU, GRC, HUN, ISL, IRL, ITA, JPN, KOR, LUX, MEX, NLD, NZL, NOR, POL, PRT, SVK, ESP, SWE, CHE, TUR, GBR, USA, BRA, CHL, CHN, COL, EST, IDN, ISR, LVA, RUS, SVN, ZAF, ARG, SAU, IND, LTU, CRI' thorough years 1981-2020.
- **Rights.** <https://www.oecd-ilibrary.org/oecd/copyright>

3.5. Análisis preliminar.

Analicemos algunos datos que pueden resultar de interés.

Seleccionamos los datos de 2018.

En primer lugar, estudiamos qué porcentaje de la población ha alcanzado cada uno de los niveles educativos según el país.

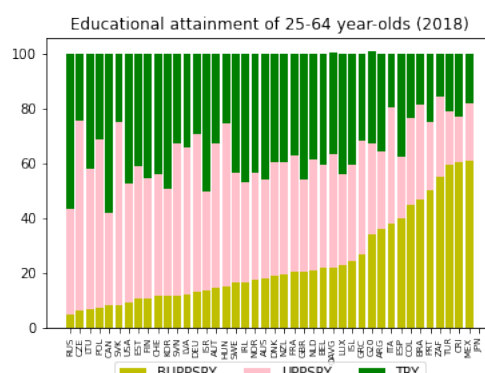


Figura 3.3: Porcentaje de población adulta según nivel educativo alcanzado y país.

En la figura 3.3 llama la atención que, en países como Rusia, sean mayoría aquellos que superan la educación universitaria, algo que contrasta con países como Japón o España. Parece ir en contra de la imagen que se tiene de estos países a priori más tecnológicos o modernos.

A continuación, exploramos cómo ha ido variando el porcentaje de hombres y mujeres con educación universitaria por país y año. Emplearemos heatmaps. En el eje horizontal, situamos la dimensión temporal medida en años. En el eje vertical, situamos la dimensión espacial, es decir, los países de estudio. Cada casilla se corresponde con el porcentaje de población que ha superado educación universitaria.

En primer lugar, analizamos el porcentaje de mujeres en la figura 3.4. Observamos el aumento progresivo con el paso del tiempo.

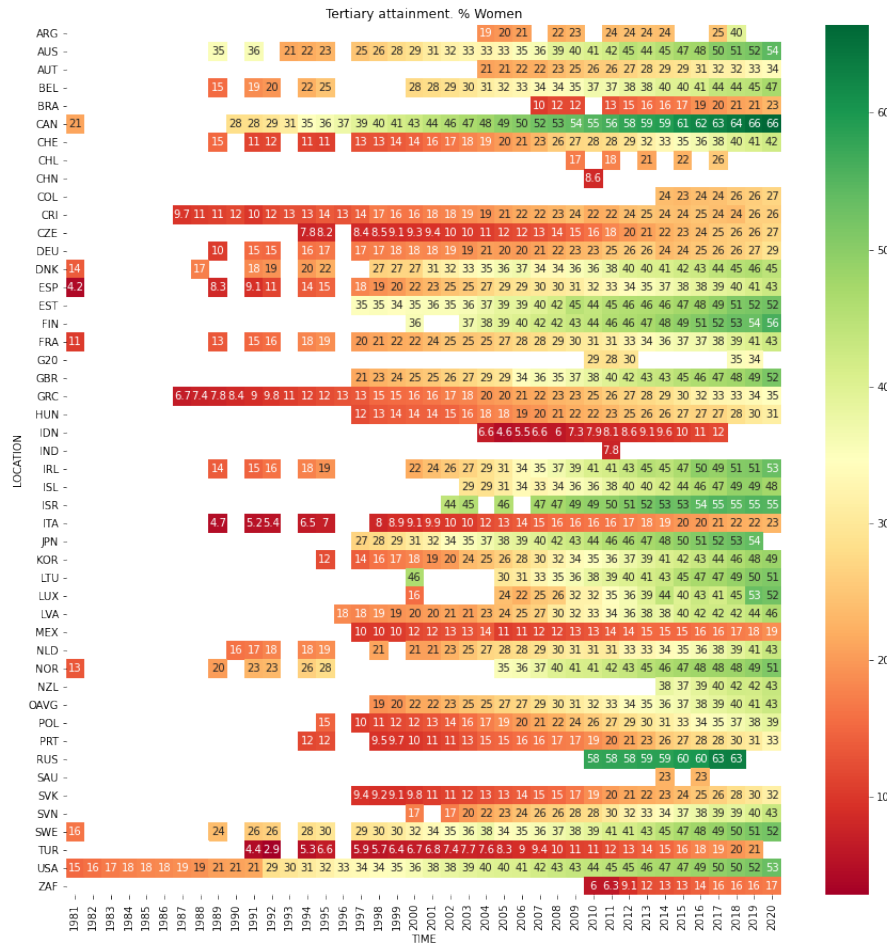


Figura 3.4: *Porcentaje de mujeres que han superado la educación universitaria por país y año.*

En segundo lugar, analizamos el porcentaje de hombres en la figura 3.5. Nuevamente, observamos el aumento progresivo con el paso del tiempo.



Por último, analizamos la diferencia restando el porcentaje de hombres menos el porcentaje de mujeres en la figura 3.6. Observamos que, en la mayoría de países, son ellas quienes, en los últimos tiempos, han superado la educación universitaria. Con el paso de los años, las mujeres han accedido a la educación superior.

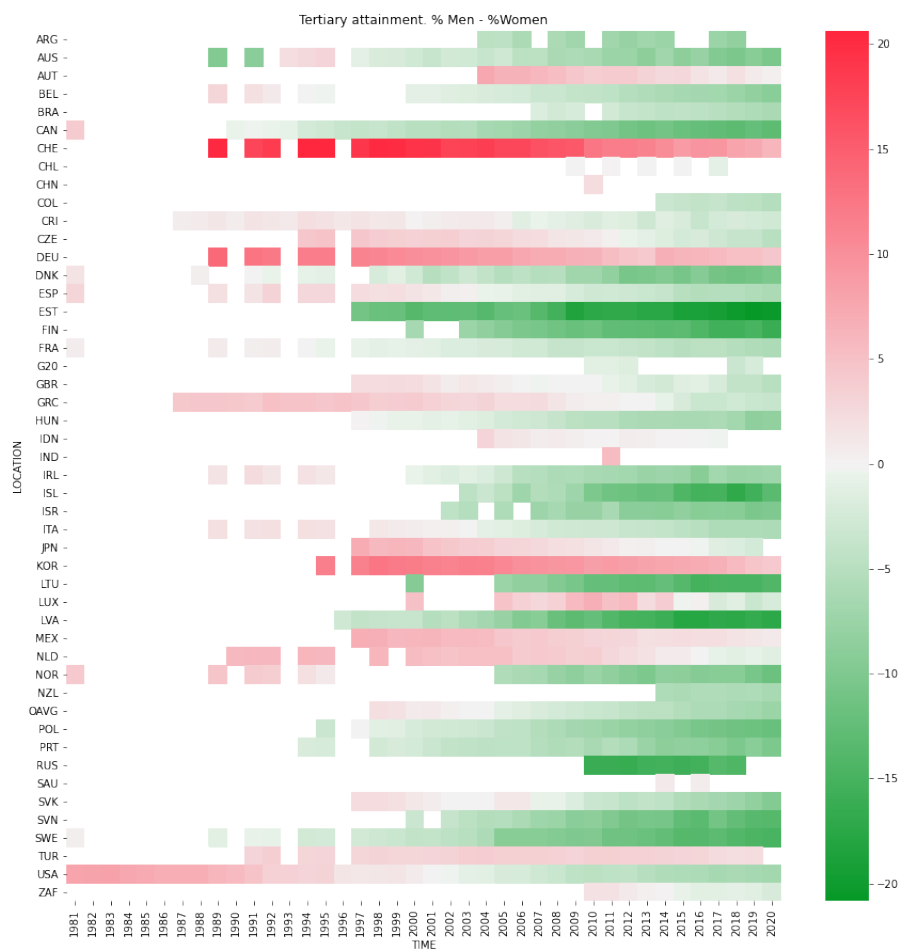


Figura 3.6: *Diferencia (en porcentaje) entre hombres y mujeres que han superado la educación universitaria por país y año*

Capítulo 4

Gastos en educación

Dentro del informe *Education at a glance*, [OEC21] otro de los indicadores que se recoge es el *Education Spending*. Este presenta datos sobre el gasto en educación, público y privado. El gasto público incluye tanto el gasto directo en instituciones educativas como los subsidios educativos a los hogares administrados por instituciones educativas. El gasto privado se registra neto de las subvenciones públicas que puedan recibir las instituciones educativas. El gasto en educación cubre el gasto en escuelas, universidades y otras instituciones educativas públicas y privadas. El gasto incluye instrucción y servicios auxiliares para estudiantes y familias proporcionados a través de instituciones educativas. El gasto se muestra en USD por alumno y como porcentaje del PIB.

4.1. Contexto

Como se había comentado previamente, el capital humano esta relacionado de manera indirecta con el nivel educativo alcanzado, es decir que generalmente un alto nivel educativo se asocia con indicadores económicos positivos. Es por esta razón que se toma muy en cuenta los valores de los gastos que se tienen en la educación.

4.2. Estructura del dataset

El dataset que recoge el indicador *Education spending* se encuentra disponible en [. A simple vista, el dataset es de la forma:](#)

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
0	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1989	44.650639	NaN
1	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1991	44.127056	NaN
2	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1993	47.159046	NaN
3	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1994	49.802025	NaN
4	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1995	44.935852	NaN

Figura 4.1: *Cabecera del dataset.*

Observamos que el dataset contiene 8 variables:

- (1) **LOCATION**: Se corresponde con los códigos ISO 3166-1 alpha-3. Son los códigos de tres letras para los país definidos en la ISO 3166-1, parte de la norma ISO 3166 publicada por la Organización Internacional de Normalización (ISO). Sirven para representar países, territorios

dependientes y zonas especiales de interés geográfico. Existen 249 códigos ISO 3166-1 alfa-3 asignados oficialmente en la actualidad. En este dataset se recogen datos de 46 países más los datos del G20 (relativo a los países del G20) y OAVG (media de la OCDE) Algunos ejemplos son:

- AUS: Australia
 - LUX: Luxemburgo
 - G20: Media de los países del G20
 - OAVG: Media de los países de la OCDE
- (II) **INDICATOR.** Muestra el indicador evaluado en el dataset. En este caso, toma un único valor ('EDUADULT') que se corresponde con el nivel educativo alcanzado por la población adulta entre 25 y 64 años.
- (III) **SUBJECT.** Toma 7 valores relacionados con el nivel educativo alcanzado y el género:
- BUPPSRY: Below upper secondary. Se corresponde con el nivel Lower secondary education.
 - TRY: Tertiary. Se corresponde con el nivel universitario.
 - UPPSRY: Upper secondary. Corresponde a la etapa final de la educación secundaria en la mayoría de los países de la OCDE. La edad de inicio de sitúa en los 15 o 16 años.
 - TRY_MEN: Tertiary education for men
 - TRY_WOMEN: Tertiary education for women
 - UPPSRY_MEN: uppersecondary education for men
 - UPPSRY_WOMEN: Upper secondary education for women
- (IV) **MEASURE.** Población sobre la que se están midiendo los resultados. Toma un único valor (PC_25.64) que se corresponde con la población adulta ente 25 y 64 años.
- (V) **FREQUENCY.** Frecuencia de medición. Toma un único valor (A) que indica que los datos son recogidos anualmente.
- (VI) **TIME.** Años en los que se han recogido datos. Se abarcan 40 años: desde 1981 hasta 2020.
- (VII) **Value.** Gasto en USD por alumno como porcentaje del PIB, para la población adulta entre 25 y 64 años haciendo referencia al nivel educativo descrito en el valor SUBJECT de su misma fila en el año indicado por la columna TIME.
- (VIII) **Flag Codes.** Un *flag* se corresponde con un atributo que representa información cualitativa sobre el valor de la propia celda. Sólo toma valores nulos. No nos aporta información.

En cuanto a los tipos de cada una de las variables:

```

: LOCATION      object
  INDICATOR      object
  SUBJECT        object
  MEASURE        object
  FREQUENCY      object
  TIME           int64
  Value          float64
  Flag Codes     float64
dtype: object

```

Figura 4.2: Tipos de las columnas dataset.

4.3. Metadatos

Los metadatos en formato DublinCore resultan:

- **Title.** Education spending.
- **Creator.** Organisation for Economic Cooperation and Development (OECD)
- **Subject.** Education.
- **Description.** Education spending covers expenditure on schools, universities and other public and private educational institutions. Spending includes instruction and ancillary services for students and families provided through educational institutions. Spending is shown in USD per student and as a percentage of GDP.
- **Contributor.** Countries considered in the study.
- **Date.** 25-12-2020.
- **Type.** Dataset.
- **Format.** CSV.
- **Identifier.** DOI: 10.1787/ca274bac-en
- **Source.** Crossref. OCDE Education at a Glance.
- **Language.** English.
- **Relation.**
- **Coverage.** Countries (ISO 3166-1 alpha-3) AUS, AUT, BEL, CAN, CZE, DNK, FIN, FRA, DEU, GRC, HUN, ISL, IRL, ITA, JPN, KOR, LUX, MEX, NLD, NZL, NOR, POL, PRT, SVK, ESP, SWE, CHE, TUR, GBR, USA, BRA, CHL, CHN, COL, EST, IDN, ISR, LVA, RUS, SVN, ZAF, ARG, SAU, IND, LTU, CRI' thorough years 1981-2020.
- **Rights.** <https://www.oecd-ilibrary.org/oecd/copyright>

Capítulo 5

Emisiones de CO2 per cápita.

Hasta ahora, uno de los indicadores de industrialización de un país eran las emisiones de CO2. Si bien, con la necesaria aparición de las energías renovables, este indicador puede dejar de ser tan robusto.

Dado este cambio de paradigma, nos preguntamos si, a partir de esta variable, podemos inferir sobre el nivel educativo alcanzado. Por ello, nos plantamos algunas preguntas: ¿un alto nivel de emisiones de co2 per cápita es sinónimo de un mayor nivel educativo?, ¿o es justo al contrario? ¿Hay una relación directa entre ingresos y emisiones de CO2? ¿Hay una relación directa entre inversión en educación y emisiones de CO2? ¿Hay una relación directa entre el nivel educativo alcanzado y las emisiones de CO2 per cápita?

5.1. Contexto.

Las emisiones de CO2 son el principal motor del cambio climático. Para mitigar los peores efectos del cambio climático, nuestra sociedad necesitar reducir sus emisiones de CO2 urgentemente. Pero, esta es una responsabilidad compartida tanto por los gobiernos como por los individuos de la sociedad.

Existen varias formas de comparar las emisiones de CO2: comparando las emisiones anuales por país, por persona; comparando el histórico de emisiones; comparando si se ajustan a los bienes y servicios comercializados por el país... Todos estos criterios son válidos pero nos pueden llevar a resultados diferentes.

Nos centraremos en las emisiones de CO2 per cápita. La contribución de cada ciudadano es calculada dividiendo las emisiones totales del país entre el total de su población. En el dataset se analizan las emisiones basadas en la producción, es decir, en las emisiones producidas dentro de las fronteras de un país sin tener en cuenta cómo se distribuyen los bienes.

5.2. Estructura del dataset.

El dataset [EOB] que recoge el indicador *Adult education level* se encuentra disponible en [ocde.dataset.attainment].

A simple vista, el dataset es de la forma:

	Entity	Code	Year	Annual CO2 emissions (per capita)
0	Afghanistan	AFG	1949	0.0019
1	Afghanistan	AFG	1950	0.0109
2	Afghanistan	AFG	1951	0.0117
3	Afghanistan	AFG	1952	0.0115
4	Afghanistan	AFG	1953	0.0132

Figura 5.1: Cabecera del dataset [Fri+21]

Observamos que el dataset contiene 4 variables:

- (I) **Entity**: Se corresponde con el nombre completo del país que identifica.
- (II) **Code**: Se corresponde con los códigos ISO 3166-1 alpha-3. Son los códigos de tres letras para los país definidos en la ISO 3166-1, parte de la norma ISO 3166 publicada por la Organización Internacional de Normalización (ISO). Sirven para representar países, territorios dependientes y zonas especiales de interés geográfico.
- (III) **Year**: Se corresponde con el año para el cuál se recogen los datos.
- (IV) **Annual CO2 emissions (per capita)**. Se corresponde con las emisiones de CO2 per cápita medidas en toneladas (t).

En cuanto a los tipos de cada una de las variables:

```
Entity      object
Code        object
Year        int64
Annual CO2 emissions (per capita) float64
dtype: object
```

Figura 5.2: Tipos de las columna del dataset [Fri+21].

5.3. Metadatos.

Al igual que con el dataset que recogía el indicador *Adult education level*, seguiremos el formato Dublin Core.

Para la obtención de metadatos con Python, empleamos el módulo **request**. Si bien, aún empleando diferentes resolvedores y aún explicitando respuestas de tipo XML o JSON, no hemos obtenido respuestas adecuadas y no hemos podido parsearlas como nos hubiese gustado. Aún así, se ha intentado parsear lo recibido. Para consultar errores y los diferentes intentos, consultar *co2.ipynb*.

Las 15 etiquetas que forman el formato Dublin Core son:

- **Title**. Data supplement to the Global Carbon Budget
- **Creator**. Multiple authorrrs. Visit <https://doi.org/10.5194/essd-2021-386>
- **Subject**. Climate. CO2 emissions.
- **Description**. The Global Carbon Project (GCP) integrates knowledge of greenhouse gases for human activities and the Earth system. Projects include global budgets for three dominant greenhouse gases — carbon dioxide, methane, and nitrous oxide — and complementary efforts in urban, regional, cumulative, and negative emissions.

- **Contributor.** -
- **Date.** 2021.
- **Type.** Datsaet. Article.
- **Format.** CSV
- **Identifier.** DOI: 10.18160/gcp-2021.
- **Source.** Our World in Data <https://ourworldindata.org/co2-emissions>, The Global Carbon Project <https://www.globalcarbonproject.org/>
- **Language.** English.
- **Relation.** <https://doi.org/10.5194/essd-2021-386> <https://doi.org/10.5281/zenodo.5569235>
- **Coverage.** Worldwide. Years 1750 - 2020.
- **Rights.** <https://www.icos-cp.eu/data-services/about-data-portal/data-license>, <https://www.icos-cp.eu/privacy> Contact info: info@globalcarbonproject.org

5.4. Análisis preliminar.

Seleccionamos los datos de 2018.

Los países con mayor emisión de CO₂ per capita se corresponden con los principales productores de petróleo. De hecho, es especialmente llamativo que países con relativamente poca población como Qatar (38.4t), Trinidad y Tobago (29.1t), Curazao (23.5t), Kuwait (23.1t) y Brunei (22.3t) sean los principales emisores:

	Entity	Code	Year	Annual CO ₂ emissions (per capita)
17134	Qatar	QAT	2018	38.4397
21033	Trinidad and Tobago	TTO	2018	29.1223
5141	Curacao	CUW	2018	23.5257
11552	Kuwait	KWT	2018	23.1008
3381	Brunei	BRN	2018	22.3619

Figura 5.3: Máximas emisiones de CO₂ per cápita (t) en 2018

Sin embargo, otros productores de petróleo tienen poca población en comparación con el total de emisiones. Los países más poblados con mayores emisiones de CO₂ per capita son Estados Unidos (16.4t), Australia (16.7t) y Canadá (15.6t).

	Entity	Code	Year	Annual CO ₂ emissions (per capita)
1471	Australia	AUS	2018	16.7081
4014	Canada	CAN	2018	15.6299
22262	United States	USA	2018	16.4340

Figura 5.4: Emisiones de CO₂ per cápita (t) en 2018 en determinados países.

Esto es más del triple de la media de emisiones de CO₂ en 2018, que se sitúa en 5.033989519650653 t per cápita.

Puesto que parece haber una relación directa entre ingresos y emisiones de CO₂ per cápita, es de esperar que los resultados sean los anteriores: los países con mayores estándares de vida tienen más huella de CO₂. Pero es claro que puede haber grandes diferencias incluso entre países con unos estándares de vida similares. Por ejemplo, las emisiones de muchos países europeos son significativamente menores que las de Australia, Canadá y Estados Unidos.

De hecho, las emisiones de países europeos como Portugal (5.0t) , Francia (4.9t) y Reino Unido (5.6) se sitúan en la media global.

	Entity	Code	Year	Annual CO2 emissions (per capita)
8176	France	FRA	2018	4.9603
17061	Portugal	PRT	2018	5.0206
22041	United Kingdom	GBR	2018	5.6878

Figura 5.5: Emisiones de CO₂ per cápita (t) en 2018 en determinados países

Sin embargo, algunos vecinos como Alemania (9t), Holanda (9.3t) y Bélgica (8.7t) disparan sus emisiones de CO₂. Si bien, esto debe leerse adecuadamente, pues en Francia, Reino Unido y Portugal gran parte de la electricidad es producida con energía nuclear y renovables.

	Entity	Code	Year	Annual CO2 emissions (per capita)
2520	Belgium	BEL	2018	8.7289
8836	Germany	DEU	2018	9.0721
14454	Netherlands	NLD	2018	9.3215

Figura 5.6: Emisiones de CO₂ per cápita (t) en 2018 en determinados países.

Históricamente, un alto nivel de emisiones de CO₂ se ha entendido como sinónimo de prosperidad, algo que aún puede comprobarse en países como Chad, Niger:

	Entity	Code	Year	Annual CO2 emissions (per capita)
4209	Chad	TCD	2018	0.0675
14809	Niger	NER	2018	0.0830

Figura 5.7: Emisiones de CO₂ per cápita (t) en 2018 en determinados países.

En el siguiente gráfico mostramos la evolución histórica de emisiones de CO₂ para ciertos países:

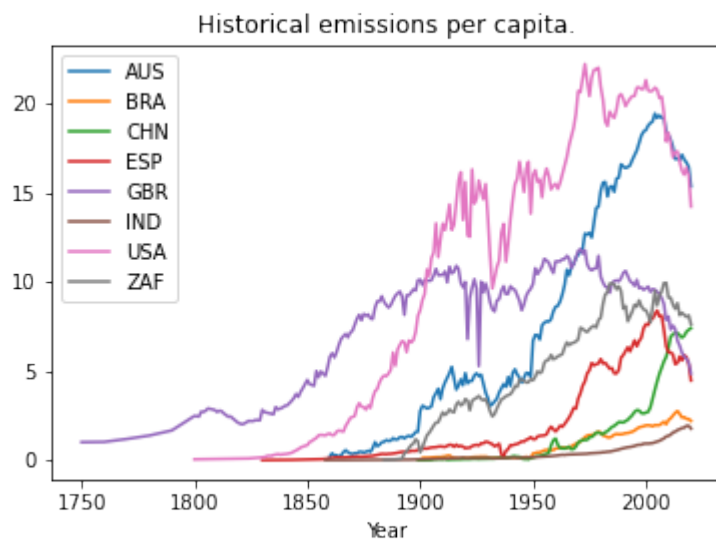


Figura 5.8: Evolución histórica de emisiones de CO₂ per cápita (t) en determinados países.

En el notebook *co2.ipynb* también puede consultarse un mapa interactivo en el que, para cada país, se muestra las emisiones per cápita:

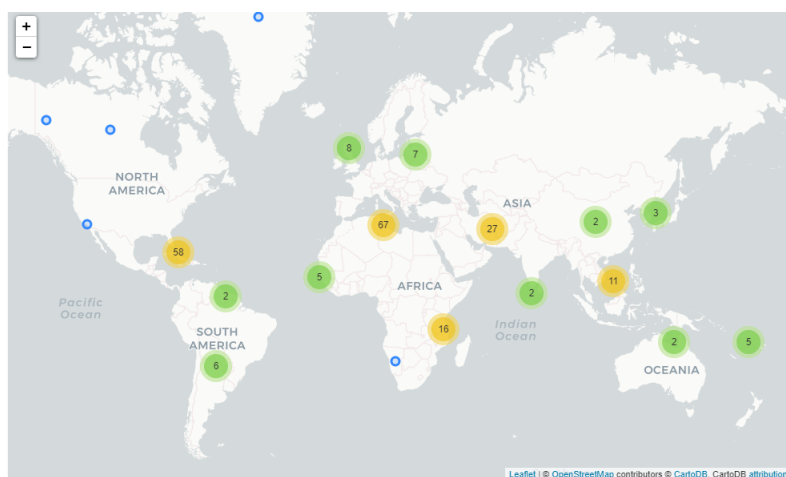


Figura 5.9: Imágen estática del mapa interactivo

Capítulo 6

Vacunas

En este análisis retrospectivo a gran escala, se examinan las tendencias globales en la confianza en las vacunas utilizando datos de 290 encuestas realizadas entre septiembre de 2015 y diciembre de 2019, en 149 países e incluyendo 284×381 personas. Se utiliza un modelo de proceso logit gaussiano multinomial bayesiano para producir estimaciones de las percepciones públicas sobre la seguridad, la importancia y la eficacia de las vacunas. Mediante regresiones logísticas bayesianas univariadas se determinaron las asociaciones entre la aceptación de la vacuna y una amplia gama de supuestos impulsores de la aceptación, incluida la confianza en la vacuna, el nivel socioeconómico y las fuentes de confianza.

A priori podemos pensar que, en el contexto actual de la situación sanitaria en la que vivimos del COVID-19, los países en los que existe una aceptación de las vacunas serán aquellos que tengan un mayor nivel educativo.

6.1. Contexto

Cada vez hay más pruebas de retrasos o rechazos de vacunas debido a la falta de confianza en la importancia, la seguridad o la eficacia de las vacunas, junto con problemas persistentes de acceso. Aunque la cobertura de inmunización se informa administrativamente en todo el mundo, no existe un sistema de seguimiento igualmente sólido para la confianza en la vacuna. En este estudio, se cartografió la confianza en las vacunas en 149 países entre 2015 y 2019.

6.2. Estructura del dataset

Los datos de las vacunas se encuentran en el archivo **vaccine.xlsx** que contiene varias hojas con datos, entre ellas: *raw_data*, *rank_safe* o *rank_effective*. Nosotros elegiremos el primero de estos, el *raw_data* para realizar nuestros cálculos.

Esta hoja de cálculo contiene datos sobre las encuestas realizadas a las diferentes personas de cada territorio. En estas encuestas se realizarán 3 diferentes preguntas: Efectividad, seguridad e importancia de las preguntas. En el dataset también aparecerán reflejadas las respuestas a cada pregunta. Estas respuestas serán de tipo binario, es decir, totalmente de acuerdo o totalmente en desacuerdo.

country or territory	who_region	count	sagree	sdisagree	question	time
Algeria	AFR	397	363	2	important	2015,83333
Algeria	AFR	397	302	13	safe	2015,83333
Algeria	AFR	397	295	19	effective	2015,83333
Algeria	AFR	928	603	17	important	2018,76666
Algeria	AFR	928	397	33	safe	2018,76666

Figura 6.1: *Cabecera del dataset.*

Nuestro dataset contiene 7 variables diferentes

- (I) **COUNTRY OR TERRITORY**: Corresponde al nombre completo del país donde se tomaron los datos.
- (II) **who_region**. Siglas de la región a la que pertenece cada país anterior. Tendremos las siguientes:
 - **AFR** Territorio africano
 - **AMR** Región de las americas
 - **EMR** Región mediterránea del este
 - **EUR** Territorio europeo
 - **SEAR** Región de Asia del este
 - **WPR** Región del pacífico del oeste
- (III) **count**. Número total de individuos entrevistados en cada país o territorio.
- (IV) **sagree**. Contador de personas que están totalmente de acuerdo con la efectividad, importancia o seguridad de las vacunas.
- (V) **sdisagree**. Número de personas que están totalmente en desacuerdo con la efectividad, importancia o seguridad de las vacunas.
- (VI) **question**. Pregunta sobre las vacunas que se le encuentra a la población.
 - a) **important** Importancia de las vacunas.
 - b) **safe** Creencia sobre si las vacunas serán seguras o no.
 - c) **effective** Pregunta sobre si la vacuna será efectiva.
- (VII) **date**. Fecha en la que se realiza la encuesta. Tendrá el siguiente formato: YYYY + (MM-1)/12

6.3. Metadatos

Los metadatos en formato DublinCore resultan:

- **Title**. Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: a large-scale retrospective temporal modelling study.
- **Creator**. Alexandre de Figueiredo, Clarissa de Figueiredo, Emilie de Figueiredo, Pauline de Figueiredo, Heidi J de Figueiredo
- **Subject**. General Medicine.

- **Description.** There is growing evidence of vaccine delays or refusals due to a lack of trust in the importance, safety, or effectiveness of vaccines, alongside persisting access issues. Although immunisation coverage is reported administratively across the world, no similarly robust monitoring system exists for vaccine confidence. In this study, vaccine confidence was mapped across 149 countries between 2015 and 2019.
- **Contributor.** Countries considered in the study.
- **Date.** 10-9-2020
- **Type.** Dataset.
- **Format.** XLSX.
- **Identifier.** DOI: 10.1016/s0140-6736(20)31558-0
- **Source.** Crossref.
- **Language.** English.
- **Relation.**
- **Coverage.** Countries 'Algeria' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi' 'Cameroon' 'Chad' 'Comoros' 'Democratic Republic of the Congo' 'Ethiopia' 'Gabon' 'Gambia' 'Ghana' 'Guinea' 'Ivory Coast' 'Kenya' 'Liberia' 'Madagascar' 'Malawi' 'Mali' 'Mauritania' 'Mauritius' 'Mozambique' 'Namibia' 'Niger' 'Nigeria' 'Republic of Congo' 'Rwanda' 'Senegal' 'Sierra Leone' 'South Africa' 'Swaziland' 'Tanzania' 'Togo' 'Uganda' 'Zambia' 'Zimbabwe' 'Argentina' 'Bolivia' 'Brazil' 'Canada' 'Chile' 'Colombia' 'Costa Rica' 'Dominican Republic' 'Ecuador' 'El Salvador' 'Guatemala' 'Haiti' 'Honduras' 'Mexico' 'Nicaragua' 'Panama' 'Paraguay' 'Peru' 'Uruguay' 'USA' 'Venezuela' 'Afghanistan' 'Egypt' 'Iran' 'Iraq' 'Jordan' 'Kuwait' 'Lebanon' 'Libya' 'Morocco' 'Pakistan' 'Palestine' 'Saudi Arabia' 'Syria' 'Tunisia' 'United Arab Emirates' 'Yemen' 'Albania' 'Armenia' 'Austria' 'Azerbaijan' 'Belarus' 'Belgium' 'Bosnia and Herzegovina' 'Bulgaria' 'Croatia' 'Cyprus' 'Czech Republic' 'Denmark' 'Estonia' 'Finland' 'France' 'Georgia' 'Germany' 'Greece' 'Hungary' 'Iceland' 'Ireland' 'Israel' 'Italy' 'Kazakhstan' 'Kosovo' 'Kyrgyzstan' 'Latvia' 'Lithuania' 'Luxembourg' 'Macedonia' 'Malta' 'Moldova' 'Montenegro' 'Netherlands' 'Northern Cyprus' 'Norway' 'Poland' 'Portugal' 'Romania' 'Russia' 'Serbia' 'Slovakia' 'Slovenia' 'Spain' 'Sweden' 'Switzerland' 'Tajikistan' 'Turkey' 'Turkmenistan' 'UK' 'Ukraine' 'Uzbekistan' 'Bangladesh' 'India' 'Indonesia' 'Myanmar' 'Nepal' 'Sri Lanka' 'Thailand' 'Australia' 'Cambodia' 'China' 'Fiji' 'Hong Kong' 'Japan' 'Laos' 'Malaysia' 'Mongolia' 'New Zealand' 'Papua New Guinea' 'Philippines' 'Singapore' 'South Korea' 'Taiwan' 'Vietnam'
- **Rights.** [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(20\)31558-0/fulltext#seccestitle10](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31558-0/fulltext#seccestitle10)

Capítulo 7

Curación de datos y unificación de datasets.

Según [TechRepublic](#), podemos definir la *curación de datos* como los procesos llevados a cabo para obtener datos de diversas fuentes e integrarlos o unificarlos en repositorios o datasets que adquieren más útiles que las diferentes partes de manera independiente.

Se trata de un proceso enfocado al mantenimiento de los datos y, a su vez, se compone de sub-tareas que garantizan que los datos estén listos para ser utilizados en las tareas para los que se requieran. Es decir, que los datos sean correctos, no tengan errores, estén completos y actualizados.

El curado de datos es un proceso continuo está presente en todas las fases del ciclo de vida de los datos, desde que se recolectan los datos iniciales hasta que se almacenan para su uso futuro.

La limpieza y el preprocesado de datos forman parte del proceso de curación. Se definen como los procesos de detección y corrección o eliminación de errores e inconsistencias en los datos con el objetivo de mejorar su calidad.

Mediante una curación y limpieza de datos adecuada, se evita llegar a conclusiones erróneas y favorece el cumplimiento los principios FAIR

Considerando los 4 datasets descritos anteriormente, nuestro objetivo es unificar toda su información en un único dataset.

7.1. Descripción del proceso de curación.

Considerando los 4 datasets descritos anteriormente, queremos seleccionar los datos para un año particular (2018) y para la mayor cantidad posible de países. La procedencia de los datasets es importante, pues cada institución suele seguir su propio esquema y estos no tienen por qué coincidir. Teniendo esto en cuenta, algunos de los principales problemas que hemos encontrado son:

- (I) Países en común. Cada dataset, recoge datos para ciertos países. No todos contienen datos de todos los países y no se garantiza que haya datos para dichos países en todos los años que abarca el dataset.
- (II) Cómo se identifica a los países. Cada dataset, recoge de manera distinta los nombres de los países: bien con el nombre completo, bien con el código ISO 3166-1 alpha-3.
- (III) En qué años hay datos disponibles.
- (IV) Formatos de fechas.
- (V) Datos vacíos

- (VI) Funciones de agregado
- (VII) Transformación de variables.
- (VIII) Formatos de tablas no adecuados
- (IX) Datos y nombres de variables duplicados o diferentes.

En lo que sigue, explicaremos las soluciones a estos problemas. Si bien, es aconsejable la lectura de esta sección de manera paralela a la lectura del notebook *combine_datasets.ipynb*.

7.1.1. Países en común e identificación.

Los datasets que recogen los datos del nivel educativo alcanzado [EO22a] y de gasto público en educación per cápita [EO22b] provienen de la misma fuente: la OCDE. De hecho, ambos están elaborados para el informe *Education at a glance*, de modo que, como ya hemos visto anteriormente, su estructura es la misma.

Ambos recogen los países participantes en el estudio en la columna "LOCATION". Cada país está identificado mediante su código ISO 3166-1 alpha-3 y en ambos dataset aparecen los mismos:

```
Column: LOCATION
Type: object
Unique values 48:
['AUS' 'AUT' 'BEL' 'CAN' 'CZE' 'DNK' 'FIN' 'FRA' 'DEU' 'GRC' 'HUN' 'ISL'
 'IRL' 'ITA' 'JPN' 'KOR' 'LUX' 'MEX' 'NLD' 'NZL' 'NOR' 'POL' 'PRT' 'SVK'
 'ESP' 'SWE' 'CHE' 'TUR' 'GBR' 'USA' 'BRA' 'CHL' 'CHN' 'COL' 'EST' 'IDN'
 'ISR' 'LVA' 'RUS' 'SVN' 'ZAF' 'OAVG' 'ARG' 'SAU' 'IND' 'LTU' 'CRI' 'G20']
```

Figura 7.1: Países identificados con ISO 3166-1 alpha en los dataset [EO22b], [EO22a]

En total, recogen datos para 48 países. Para cerciorarnos de que ambos dataset contienen los mismos países, hemos comprobado que los valores únicos la columna "LOCATION" del dataset del nivel educativo alcanzado sean un subconjunto de los valores únicos de la columna "LOCATION" del dataset de gasto educativo y viceversa.

Llama la atención que dos de estos países son G20 (grupo político del G20) y OAVG (media de la OCDE), luego es esperable que estos no estén contenidos en el resto de datasets.

En el dataset que recoge los datos de las emisiones de CO2 per cápita [Fri+21], los países son identificados de dos formas. En la columna ".Entity", se recoge el nombre completo del país; en la columna "Code" se recoge el código ISO 3166-1 alpha-3:

```
Column: Entity
Type: object
Unique values 230:
['Afghanistan' 'Africa' 'Albania' 'Algeria' 'Andorra' 'Angola' 'Anguilla'
 'Antigua and Barbuda' 'Argentina' 'Armenia' 'Aruba' 'Asia'
 'Asia (excl. China & India)' 'Australia' 'Austria' 'Azerbaijan' 'Bahamas'
 'Bahrain' 'Bangladesh' 'Barbados' 'Belarus' 'Belgium' 'Belize' 'Benin'
 'Bermuda' 'Bhutan' 'Bolivia' 'Bonaire Sint Eustatius and Saba'
 'Bosnia and Herzegovina' 'Botswana' 'Brazil' 'British Virgin Islands'
 'Brunei' 'Bulgaria' 'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon'
 'Canada' 'Cape Verde' 'Central African Republic' 'Chad' 'Chile' 'China'
 'Colombia' 'Comoros' 'Congo' 'Cook Islands' 'Costa Rica' 'Cote d'Ivoire'
 'Croatia' 'Cuba' 'Curacao' 'Cyprus' 'Czechia'
 'Democratic Republic of Congo' 'Denmark' 'Djibouti' 'Dominica']
```

Figura 7.2: Extracto de los países identificados en [Fri+21] mediante nombre completo.

```

Column: Code
Type: object
Unique values 219:
['AFG' nan 'ALB' 'DZA' 'AND' 'AGO' 'AIA' 'ATG' 'ARG' 'ARN' 'ABW' 'ALS'
 'AUT' 'AZE' 'BHS' 'BHR' 'BGD' 'BRB' 'BLR' 'BEL' 'BLZ' 'EEN' 'BMU' 'BTN'
 'BOL' 'BES' 'BIH' 'BWA' 'BRA' 'VGB' 'BRN' 'BGR' 'BFA' 'EDI' 'KHM' 'CMR'
 'CAN' 'CPV' 'CAF' 'TCD' 'CHL' 'CHN' 'COL' 'COM' 'COG' 'COK' 'CRI' 'CIV'
 'HRV' 'CUB' 'CUW' 'CYP' 'CZE' 'COD' 'DNK' 'DJI' 'DMA' 'DOM' 'ECU' 'EGY'
 'SLV' 'GNQ' 'ERI' 'EST' 'SWZ' 'ETH' 'FRO' 'FJI' 'FIN' 'FRA' 'GUF' 'PYF'
 'GAB' 'GMB' 'GEO' 'DEU' 'GHA' 'GRC' 'GRL' 'GRD' 'GLP' 'GTM' 'GIN' 'GNB'
 'GUY' 'HTI' 'HND' 'HKG' 'HUN' 'ISL' 'IND' 'IDN' 'IRN' 'IRQ' 'IRL' 'ISR'
 'ITA' 'JAM' 'JPN' 'JOR' 'KAZ' 'KEN' 'KIR' 'KMT' 'KGZ' 'LAO' 'LVA' 'LBN'
 'LSO' 'LBR' 'LBV' 'LIE' 'LTU' 'LUX' 'MAC' 'MDG' 'MMI' 'MYS' 'MDV' 'MLI'
 'MIT' 'MHI' 'MTO' 'MRT' 'MUS' 'MYT' 'MFX' 'MNA' 'MNG' 'MNF' 'MSR' 'MAR'

```

Figura 7.3: Extracto de los países identificados en [Fri+21] mediante código ISO 3166-1 alpha-3

Observamos que hay 230 valores únicos en la columna `.Entity`: 219 valores únicos en la columna `Code`. Esto se debe a que no existe código ISO 3166-1 alpha-3 para los siguientes valores de `.Entity`:

```

array(['Africa', 'Asia', 'Asia (excl. China & India)', 'EU-27', 'EU-28',
 'Europe', 'Europe (excl. EU-27)', 'Europe (excl. EU-28)',
 'North America', 'North America (excl. USA)', 'Oceania',
 'South America'], dtype=object)

```

Figura 7.4: Valores de la columna `.Entity` del dataset [Fri+21] para los que no existe código ISO 3166-1 alpha-3.

Que los países estén identificados de dos maneras distintas, quizá pueda asumirse como un problema. Sin embargo, en nuestro caso es toda una ventaja, pues nos permite hacer una función para identificar los países del dataset que recoge los datos de las vacunas [De +20].

Puede comprobarse que los únicos países de los datasets [EO22b], [EO22a] no contenidos en [Fri+21] son OAVG y G20. Luego, a priori, sólo tendremos 46 países en común entre estos tres dataset. Y, recordemos, aún no hemos filtrado por año.

En el dataset [De +20], los países están identificados mediante su nombre completo en la columna `country or territory`:

```

Column: country or territory
Type: object
Unique values 149:
['Algeria' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi' 'Cameroon' 'Chad'
 'Comoros' 'Democratic Republic of the Congo' 'Ethiopia' 'Gabon' 'Gambia'
 'Ghana' 'Guinea' 'Ivory Coast' 'Kenya' 'Liberia' 'Madagascar' 'Malawi'
 'Mali' 'Mauritania' 'Mauritius' 'Mozambique' 'Namibia' 'Niger' 'Nigeria'
 'Republic of Congo' 'Rwanda' 'Senegal' 'Sierra Leone' 'South Africa'
 'Swaziland' 'Tanzania' 'Togo' 'Uganda' 'Zambia' 'Zimbabwe' 'Argentina'

```

Figura 7.5: Extracto de la columna `country or territory` del dataset [De +20]

Es por ello que la doble identificación en el dataset del CO2 nos es beneficiosa. Con dicho dataset, podemos identificar los países en tuplas del tipo:

```

[('ABW', 'Aruba'),
 ('AFG', 'Afghanistan'),
 ('AGO', 'Angola'),
 ('AIA', 'Anguilla'),
 ('ALB', 'Albania'),
 ('AND', 'Andorra'),
 ('ARE', 'United Arab Emirates'),

```

Figura 7.6: Tuplas identificativas para los países

La primera componente contiene el código ISO 3166-1 alpha-3 y la segunda componente contiene el nombre completo. Gracias a este mapeado, podemos comprobar que los países coincidentes entre

este último dataset y el resto son 43. Hemos perdido, además de OAVG y G20, los países Czechia, United Kingdom, United States.

7.1.2. Años con datos disponibles y formatos de fechas.

Como ya hemos apuntado, nada nos garantiza que todos los dataset recojan datos para los mismo años. De hecho, para cada uno de los dataset, no está garantizado que recojan datos para los mismos países cada año.

En primer lugar, estudiamos los formatos de fechas. Los dataset de nivel educativo [EO22a] y gasto público en educación [EO22b] al estar creados por la misma organización y con el mismo objetivo, recogen datos para los mismos 40 años y estos están recogidos como números enteros en las respectivas columnas "TIME":

```
Column: TIME
Type: int64
Unique values 40:
[1989 1991 1993 1994 1995 1997 1998 1999 2000 2001 2002 2003 2004 2005
 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
 2020 1992 1981 1990 1996 1988 1987 1982 1983 1984 1985 1986]
```

Figura 7.7: Fechas recogidas en [EO22b], [EO22a]

Análogamente, el dataset [Fri+21] recoge los años como enteros. En esta ocasión, la serie abarca 226 años en la columna "Year":

```
Column: Year
Type: int64
Unique values 226:
[1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962
 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976
 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990
 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
 2019 2020 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895]
```

Figura 7.8: Extracto de las fechas recogidas en [Fri+21]

Si bien, el formato de las fechas en el dataset de las vacunas es distinto a los anteriores. Almacenadas en la columna "time", las fechas siguen la fórmula:

$$time = YYYY + (MM - 1)/12,$$

done YYYY es el año en el que se realizó la encuesta y MM el mes. Por ejemplo, el valor 2018,8333 se corresponde con el mes de noviembre del año 2018:

$$2018,833 = 2018 + \frac{11 - 1}{12}.$$

```
Column: time
Type: object
Unique values 143:
['2015,833333' '2018,766667' '2018,566667' '2018,802778' '2016,33'
 '2018,497222' '2018,713889' '2018,341667' '2018,780556' '2018,777778'
 '2018,477778' '2018,611111' '2018,502778' '2018,9' '2018,669444'
 '2018,755556' '2018,416667' '2018,405556' '2018,619444' '2018,366667'
 '2018,730556' '2018,322222' '2018,688889' '2018,655556' '2018,527778'
 '2018,394444' '2019,846575' '2018,538889' '2018,5' '2018,547222']
```

Figura 7.9: Extracto de las fechas recogidas en el dataset [De +20]

Puesto que nos interesa únicamente el año y están almacenadas como string, el proceso de conversión que hemos llevado a cabo consta de las siguientes partes:

- (I) Manteniendo el tipo string, reemplazar la "," por un punto "."
- (II) Convertir a float
- (III) Redondear hacia abajo
- (IV) Puesto que pueden haber llevado a cabo múltiples encuestas en un mismo país durante el mismo año, emplear la función de agregado `np.mean` (posteriormente se explicará)

Con todas las fechas bien formateadas, seleccionamos el año 2018, que es nuestro año de estudio, y seleccionamos los países en común entre todos los datasets.

Al filtrar por 2018 en los dataset [EO22a], [EO22b] sólo obtenemos datos para 43 de los 48 países iniciales. Incluso obtenemos datos incompletos para algunos países, cuestión que se abordará a continuación. Con respecto al total de países de estos dataset, hemos perdido CHL, CHN, IDN, IND y SAU.

Originalmente, sin filtrar por año, coincidían 46 países entre los dataset de la OCDE y el de CO2 de un total de 48 países (todos salvo G20 y OAVG). Filtrando por 2018 en los dataset de la OCDE, tenemos 43 países. Como ninguno de los países que hemos perdido es G20 u OAVG, entonces, como mucho, habrá 41 países en CO2 para el año 2018.

De hecho, sólo coinciden 30 países entre estos 3 dataset para el año 2018. Hemos perdido, como era de esperar G20 y OAVG y, además, CZE (Czechia), GBR (United Kingdom) y USA (United States). Esto concuerda con lo que hemos observado anteriormente.

Al filtrar por 2018 en el dataset de las vacunas mantenemos estos 38 países.

7.1.3. Datos vacíos y funciones de agregado.

Al filtrar por el año 2018 en los dataset [EO22b], [EO22a], obtenemos datos vacíos para JPN:

JPN	NaN	51.928062	51.028938	52.841282	NaN	NaN	NaN
-----	-----	-----------	-----------	-----------	-----	-----	-----

Figura 7.10: *Missing values para JPN*

En estos datasets, puesto que no existe más de una observación de cada variable para cada año, en particular para 2018, podremos usar `aggfunc='first'` en el comando `pivot_table`.

Sin embargo, en el dataset de las vacunas esto no ocurre. Puede haber (o no) varias observaciones de las variables para un mismo año, de modo que si usásemos `aggfunc='first'`, estaríamos perdiendo información. Por ejemplo, en el año 2018 en el país ZAF se realiza una única encuesta:

	country or territory	count	sagree	sdisagree	question
126	ZAF	421	346	14	important
127	ZAF	421	287	16	safe
128	ZAF	421	279	24	effective

Figura 7.11: *Una única encuesta de confianza en las vacunas realizada en 2018 en ZAF*

Sin embargo, en ESP se realizan dos encuestas:

	country or territory	count	sagree	sdisagree	question
678	ESP	1005	823	2	important
679	ESP	1005	642	6	safe
680	ESP	1005	703	8	effective
681	ESP	981	735	22	important
682	ESP	981	590	24	safe
683	ESP	981	587	31	effective

Figura 7.12: Dos encuestas de confianza en las vacunas realizadas en 2018 en ESP

Es por ello que al usar `aggfunc='first'` en el comando `pivot_table` y hemos optado por emplear `aggfunc='np.mean'`.

7.1.4. Transformación de variables.

En los datasets [EO22a], [EO22b], [Fri+21] se recogen los datos per cápita. En el dataset [De+20], se recogen los datos en frecuencia absoluta. Simplemente se almacena un contador que contiene a cuántas personas se ha preguntado y, para cada variable, se almacena cuantas personas han contestado esa respuesta. Si bien, nada nos garantiza que en cada encuesta realizada en cada país se haya preguntado al mismo número de personas. Tampoco nada nos garantiza que todas las personas encuestadas hayan respondido a la encuesta.

Es por ello que, en lugar de los valores en frecuencia absoluta, nos quedamos con la frecuencia relativa. Dividimos el valor de cada variable por el total de personas sondeadas en cada encuesta.

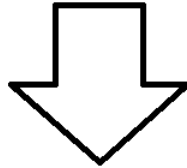
7.1.5. Formatos de tablas no adecuados.

Para llevar a cabo nuestro objetivo, los formatos originales de los dataset no nos son válidos. Es por ello que debemos transformarlos.

Veamos con detalle la transformación del dataset que recoge el nivel educativo alcanzado [EO22a].

En la imagen, se muestra la transformación desde el dataset original hasta los datos que nos interesan.

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
0	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1989	44.650639	NaN
1	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1991	44.127056	NaN
2	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1993	47.159046	NaN
3	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1994	49.802025	NaN
4	AUS	EDUADULT	BUPPSRY	PC_25_64	A	1995	44.935852	NaN



	SUBJECT	BUPPSRY-level	TRY-level	TRY_MEN-level	TRY_WOMEN-level	UPPSRY-level	UPPSRY_MEN-level	UPPSRY_WOMEN-level
LOCATION								
ARG		36.359135	35.663551	31.336042	39.529755	27.977314	28.792295	27.249208
AUS		18.108641	45.727478	40.441612	50.837093	36.163879	41.903530	30.615618
AUT		14.702526	32.711426	33.656857	31.769451	52.586048	55.005234	50.175697
BEL		21.773218	40.638546	36.949390	44.334103	37.588238	40.415058	34.756508
BRA		46.954311	18.429998	15.601707	20.977964	34.615692	34.144085	35.040554

Figura 7.13: Transformación del dataset [EO22a]

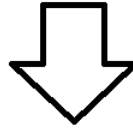
En primer lugar, se ha filtrado por año 2018 y hemos seleccionado los países en común entre todos los datasets que nos atañen. A continuación, hemos despreciado las columnas "FREQUENCY", "Flag Codes", "INDICATOR", "MEASURE" y "TIME". Esto se debe a que las 5 primeras columnas citadas toman cada una un único valor que no nos aporta nada y la columna "TIME" ha sido filtrada por el año 2018. Quizá podríamos dejarla para el futuro uso del dataset conjunto. Si bien, en los metadatos conjunto, está detallado en qué año se han recogido los datos.

En segundo lugar, para favorecer la unicón de los 4 dataset, hemos despreciado los índices naturales y los hemos sustituido por los códigos ISO 3166-1 alpha-3 que identifican a los países.

En tercer lugar, hemos cambiado las columnas. Las mediciones que nos interesan se recogen, en el dataset original, en la columna "SUBJECT". De modo que hemos establecido sus valores (toma 7 valores únicos como ya se ha explicado) como nuevas columnas. Para ello, hemos tenido que tratar con índices jerárquicos.

El proceso es análogo en el resto de datasets. Puede consultarse el notebook *combine_datasets.ipynb*. Adjuntamos imágenes que resumen la transformación del resto:

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
0	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1989	44.650639	NaN
1	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1991	44.127056	NaN
2	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1993	47.159046	NaN
3	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1994	49.802025	NaN
4	AUS	EDUADULT	BUPPSRY	PC_25_64		A 1995	44.935852	NaN



SUBJECT	BUPPSRY-spend	TRY-spend	TRY_MEN-spend	TRY_WOMEN-spend	UPPSRY-spend	UPPSRY_MEN-spend	UPPSRY_WOMEN-spend
LOCATION							
ARG	36.359135	35.663551	31.336042	39.529755	27.977314	28.792295	27.249208
AUS	18.108641	45.727478	40.441612	50.837093	36.163879	41.903530	30.615618
AUT	14.702526	32.711426	33.656857	31.769451	52.586048	55.005234	50.175697
BEL	21.773218	40.638546	36.949390	44.334103	37.588238	40.415058	34.756508
BRA	46.954311	18.429998	15.601707	20.977964	34.615692	34.144085	35.040554

Figura 7.14: Transformación del dataset [EO22b]


	Entity	Code	Year	Annual CO2 emissions (per capita)
0	Afghanistan	AFG	1949	0.0019
1	Afghanistan	AFG	1950	0.0109
2	Afghanistan	AFG	1951	0.0117
3	Afghanistan	AFG	1952	0.0115
4	Afghanistan	AFG	1953	0.0132



:	C02
LOCATION	
ARG	4.0824
AUS	16.7081
AUT	7.4865
BEL	8.7289
BRA	2.3091

Figura 7.15: Transformación del dataset [Fri+21]

	country or territory	who_region	count	sagree	sdisagree	question	time
0	Algeria	AFR	397	363	2	important	2015,833333
1	Algeria	AFR	397	302	13	safe	2015,833333
2	Algeria	AFR	397	295	19	effective	2015,833333
3	Algeria	AFR	928	603	17	important	2018,766667
4	Algeria	AFR	928	397	33	safe	2018,766667



	count	sagree-effective	sagree-important	sagree-safe	sdisagree-effective	sdisagree-important	sdisagree-safe
LOCATION							
ARG	962.0	0.729730	0.937630	0.766112	0.008316	0.006237	0.013514
AUS	988.0	0.715587	0.834008	0.609312	0.017206	0.018219	0.033401
AUT	942.0	0.449045	0.585987	0.363057	0.031847	0.054140	0.058917
BEL	970.0	0.501546	0.668557	0.393814	0.030928	0.022680	0.065979
BRA	950.0	0.574737	0.840000	0.614737	0.012632	0.002105	0.014737

Figura 7.16: Transformación del dataset [De +20]

7.1.6. Datos y nombres de variables duplicados o diferentes.

En las figuras 7.13 y 7.14, observamos que los nombres de las columnas de los dataset originales coinciden. Para diferenciarlos, hemos tenido que añadir los sufijos -level y -spend.

Del mismo modo, para aplanar índices jerárquicos y establecer los nombres adecuados, hemos tenido que modificar los nombres de las columnas como puede verse en la figura 7.16.

Para eliminar espacios en los nombres de columnas y favorecer la comodidad, hemos renombrado variables como puede verse en la figura 7.15. *Annual CO2 emissions (per capita)* ha sido renombrada, simplemente, como *CO2*.

Además, durante el proceso de transformación hemos tenido que eliminar variables (columnas) duplicadas. Por ejemplo, al transformar el dataset de las vacunas, obteníamos un dataframe intermedio tal que así:

	count-effective	count-important	count-safe	sagree-effective	sagree-important	sagree-safe	sdisagree-effective	sdisagree-important	sdisagree-safe
LOCATION									
ARG	962.0	962.0	962.0	702.0	902.0	737.0	8.0	6.0	13.0
AUS	988.0	988.0	988.0	707.0	824.0	602.0	17.0	18.0	33.0
AUT	942.0	942.0	942.0	423.0	552.0	342.0	30.0	51.0	55.5
BEL	970.0	970.0	970.0	486.5	648.5	382.0	30.0	22.0	64.0

Figura 7.17: Dataset intermedio en la transformación del dataset [De +20]

Tras cerciorarnos que los valores de las primeras columnas siempre coinciden, pues indica a cuántas personas se ha preguntado en la encuesta, hemos decidido quedarnos con una única columna "count".

7.2. Resultado de la curación. Dataset final.

Tras hacer frente a estas contingencias y a algunas otras vicisitudes que pueden consultarse en el notebook *combine_datasets.ipynb*, hemos obtenido un dataset final que contiene las siguientes columnas:

LOCATION	object
BUPPSRY-level	float64
TRY-level	float64
TRY_MEN-level	float64
TRY_WOMEN-level	float64
UPPSRY-level	float64
UPPSRY_MEN-level	float64
UPPSRY_WOMEN-level	float64
BUPPSRY-spend	float64
TRY-spend	float64
TRY_MEN-spend	float64
TRY_WOMEN-spend	float64
UPPSRY-spend	float64
UPPSRY_MEN-spend	float64
UPPSRY_WOMEN-spend	float64
CO2	float64
count	float64
sagree-effective	float64
sagree-important	float64
sagree-safe	float64
sdisagree-effective	float64
sdisagree-important	float64
sdisagree-safe	float64
dtype:	object

Figura 7.18: Columnas del dataset final

La primera se corresponde con el país estudiado. Las 7 siguientes se corresponden con nuestras variables objetivos, procedentes del dataset [EO22a]. El resto se corresponden con los predictores, provenientes del resto de datasets.

Los metadatos del dataset final en formato DublinCore resultan:

1. **Title.** Target education.
2. **Creator.** Analyses Esp.
3. **Subject.** Education.
4. **Description.** Dataset containing as target the education level and different predictors related to the CO2 emissions, vaccines acceptance and spending on education. All information collected from different datasets and preprocessed.
5. **Contributor.** Countries considered in the study.
6. **Date.** 22-08-2022.
7. **Type.** Dataset.
8. **Format.** CSV.
9. **Identifier.** Identifier not assigned.
10. **Source.** Crossref. Datasets originales de nivel educativo, gasto educativo, emisiones CO2 y confianza en las vacunas
11. **Language.** English.
12. **Relation.**
13. **Coverage.** Countries (ISO 3166-1 alpha-3) ARG, AUS, AUT, BEL, BRA, CAN, CHE, CHL, CHN, COL, CRI, DEU, DNK, ESP, EST, FIN, FRA, GRC, HUN, IDN, IND, IRL, ISL, ISR, ITA, JPN, KOR, LTU, LUX, LVA, MEX, NLD, NOR, NZL, POL, PRT, RUS, SAU, SVK, SVN, SWE, TUR, ZAF on year 2018.
14. **Rights.** Creative Commons.

Capítulo 8

Análisis

8.1. Objetivos del trabajo

Para realizar el análisis del nivel de educación vamos a ajustar un modelo de regresión lineal. Nuestra variable objetivo será **BUPPSRY**, que se corresponde con el nivel Lower secondary education.

► `Y: BUPPSRY.level`

Y las variables explicativas:

- `X1: BUPPSRY.spend`
- `X2: TRY.spend`
- `X3: TRY_MEN.spend`
- `X4: TRY_WOMEN.spend`
- `X5: UPPSRY.spend`
- `X6: UPPSRY_MEN.spend`
- `X7: UPPSRY_WOMEN.spend`
- `X8: C02`
- `X9: agree.effective`
- `X10: agree.important`
- `X11: agree.safe`
- `X12: disagree.effective`
- `X13: disagree.important`
- `X14: disagree.safe`

El primer paso será hacer el ajuste de nuestro modelo de regresión. Después estudiaremos los datos atípicos y si hay puntos con capacidad de influencia. Finalmente haremos una validación de nuestro modelo, es decir, que cumpla las siguientes hipótesis:

- **Homocedasticidad:** La varianza del error es la misma cualquiera que sea el valor de la variable explicativa:

$$Var(\varepsilon|X = x) = \sigma^2 \forall x$$

- **Normalidad:** El error tiene distribución normal.

$$\varepsilon \in N(0, \sigma^2)$$

- **Linealidad:** La función de regresión es una línea recta.
- **Independencia :** Las variables aleatorias son mutuamente independientes. No será necesario demostrar que se cumple esta hipótesis, ya que viene implícito que nuestras variables serán independientes entre ellas.

Iremos explicando cada paso a lo largo del capítulo y, finalmente, llegaremos a una serie de conclusiones.

Nuestras primeras creencias sobre cómo influyen nuestras variables explicativas sobre nuestra variable respuesta son las siguientes:

8.2. Ajuste del modelo de regresión lineal múltiple

Cargaremos los datos del dataset creado anteriormente y eliminaremos las filas que contienen valores NAN.

```
'data.frame': 37 obs. of 23 variables:
 $ LOCATION      : chr  "ARG" "AUS" "AUT" "BEL" ...
 $ BUPPSRY.level : num  36.4 18.1 14.7 21.8 47 ...
 $ TRY.level      : num  35.7 45.7 32.7 40.6 18.4 ...
 $ TRY_MEN.level  : num  31.3 40.4 33.7 36.9 15.6 ...
 $ TRY_WOMEN.level : num  39.5 50.8 31.8 44.3 21 ...
 $ UPPSRY.level   : num  28 36.2 52.6 37.6 34.6 ...
 $ UPPSRY_MEN.level : num  28.8 41.9 55 40.4 34.1 ...
 $ UPPSRY_WOMEN.level : num  27.2 30.6 50.2 34.8 35 ...
 $ BUPPSRY.spend  : num  36.4 18.1 14.7 21.8 47 ...
 $ TRY.spend      : num  35.7 45.7 32.7 40.6 18.4 ...
 $ TRY_MEN.spend  : num  31.3 40.4 33.7 36.9 15.6 ...
 $ TRY_WOMEN.spend : num  39.5 50.8 31.8 44.3 21 ...
 $ UPPSRY.spend   : num  28 36.2 52.6 37.6 34.6 ...
 $ UPPSRY_MEN.spend : num  28.8 41.9 55 40.4 34.1 ...
 $ UPPSRY_WOMEN.spend : num  27.2 30.6 50.2 34.8 35 ...
 $ C02            : num  4.08 16.71 7.49 8.73 2.31 ...
 $ count          : num  962 988 942 970 950 985 930 920 949
                   950 ...
 $ agree.effective : num  0.73 0.716 0.449 0.502 0.575 ...
 $ agree.important : num  0.938 0.834 0.586 0.669 0.84 ...
 $ agree.safe      : num  0.766 0.609 0.363 0.394 0.615 ...
 $ sdisagree.effective: num  0.00832 0.01721 0.03185 0.03093
                   0.01263 ...
 $ sdisagree.important: num  0.00624 0.01822 0.05414 0.02268
                   0.00211 ...
 $ sdisagree.safe   : num  0.0135 0.0334 0.0589 0.066 0.0147 ...
 - attr(*, "na.action")= 'omit' Named int 22
 ..- attr(*, "names")= chr "22"
```


Multiple R-squared:	1,	Adjusted R-squared:	1
F-statistic:	8.31e+31 on 13 and 23 DF,	p-value:	< 2.2e-16

Podemos ver que el p-valor de las variables BUPPSRY.spend , TRY.spend, C02 van a estar por debajo del 10%, lo cual nos hará presuponer que estas son las variables predictoras que mejor explicarán a la variable objetivo. Por otra parte, el p-valor asociado al ajuste del modelo es del orden de $1E - 17$, por lo que podemos presuponer que nuestro modelo es adecuado para explicar el nivel de educación.

Con el fin de reducir la dimensionalidad de nuestro problema, estudiaremos si podemos eliminar alguna variable de nuestro ajuste. Para la selección de variables usaremos un criterio de significación global que, en nuestro caso, será el AIC (Akaike information criterion).

(Intercept)	BUPPSRY.spend	TRY.spend	UPPSRY_MEN.spend
-1.804e-12	1.000e+00	1.783e-14	9.051e-15
UPPSRY_WOMEN.spend		C02	sagree.effective
8.871e-15	8.554e-16	1.414e-14	

El método AIC selecciona las variables BUPPSRY.spend, TRY.spend, UPPSRY_MEN.spend, UPPSRY_WOMEN.spend, C02 y agree.effective. Por tanto, pasamos de tener un modelo con 13 variables a tener un modelo con solamente 6 variables.

El ajuste del nuevo modelo quedaría del siguiente modo:

Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.630e-14	3.809e-15	-4.280e+00 0.000176 ***
sagree.effective	-4.546e-15	3.216e-17	-1.414e+02 < 2e-16 ***
UPPSRY_WOMEN.spend	3.638e-17	1.901e-17	1.914e+00 0.065218 .
TRY.spend	2.061e-16	3.799e-17	5.426e+00 7.00e-06 ***
UPPSRY_MEN.spend	1.246e-16	1.909e-17	6.526e+00 3.24e-07 ***
C02	3.922e-19	1.136e-18	3.450e-01 0.732427
BUPPSRY.spend	1.000e+00	3.807e-17	2.627e+16 < 2e-16 ***

Signif. codes:	0	***	0.001 ** 0.01 * 0.05 .
	0.1	1	
Residual standard error: 2.007e-17 on 30 degrees of freedom			
Multiple R-squared: 1, Adjusted R-squared: 1			
F-statistic: 4.225e+36 on 6 and 30 DF, p-value: < 2.2e-16			

8.3. Validación de hipótesis

Ahora procederemos a comprobar si nuestro modelo cumple las hipótesis preconcebidas, que son: homocedasticidad, normalidad y linealidad.

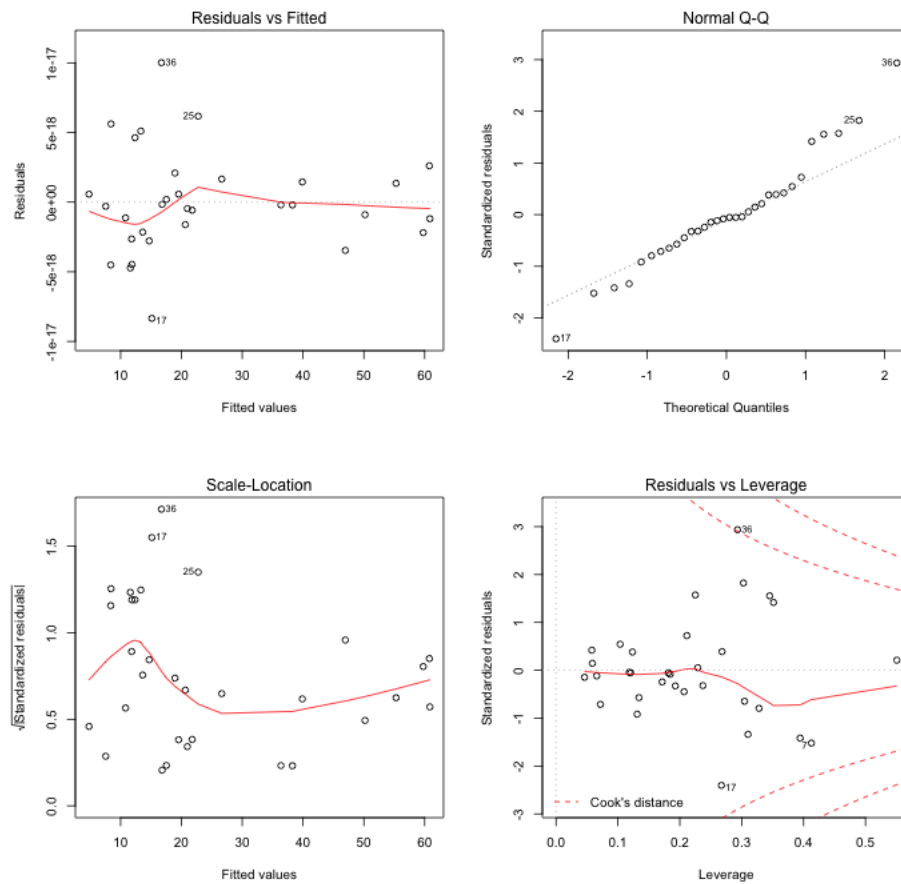


Figura 8.1: Extracto de los países identificados en *mediante nombre completo*

La gráfica anterior nos da una idea de cómo se comportarán nuestros datos ajustados. Nos centraremos en las 2 primeras imágenes de arriba.

El primer gráfico (Residual vs Fitted) nos muestra que los puntos se distribuyen de forma lineal y que la varianza será constante, ya que podemos apreciar que la escala del eje Y va a ser muy pequeño.

Viendo el segundo gráfico (Normal Q-Q) vemos que los puntos están muy próximos a la recta discontinua y, por tanto, podemos presuponer la normalidad del modelo.

Ahora pasaremos a la validación del modelo.

■ Homocedasticidad

Utilizaremos el test de Breusch-Pagan para comprobar que nuestro modelo es homocedástico

$$H_0 = \text{"Varianza igual para todo } x\text{"}$$

$$H_a = \text{"Varianza depende de cada } x\text{"}$$

```
>bptest(modelo_new)

studentized Breusch-Pagan test

data:  modelo_new
BP = 6.7452, df = 6, p-value = 0.3451
```

El nivel de significacion es bastante alto, un 0.3451, por tanto no tenemos pruebas significativas en contra de nuestra hipótesis nula. Por tanto, podemos afirmar que modelo es homocedástico.

■ Normalidad

Para estudiar la normalidad del modelo usaremos el test de Shapiro

$$H_0 = \text{"Los errores son normales"}$$

$$H_a = \text{"Los errores no son normales"}$$

```
>res_new<-rstandard(modelo_new)
>shapiro.test(res_new)

Shapiro-Wilk normality test

data:  res_new
W = 0.9623, p-value = 0.3172
```

El p-valor es lo suficientemente grande (0.3172) para aceptar la hipótesis nula H_0 frente la alternativa. Afirmamos entonces que nuestro modelo cumple la hipótesis de normalidad.

■ Linealidad

Vamos a usar el test de Harvey-Collier para comprobar si nuestro modelo es lineal.

$$H_0 = \text{"El modelo es lineal"}$$

$$H_a = \text{"Funcion no lineal y siempre concava o convexa"}$$

```
harvtest(modelo_new)

Harvey-Collier test

data:  modelo_new
HC = 1.8029, df = 24, p-value = 0.08398
```


El test nos devuelve un p-valor de 0.08, por lo que tendremos que establecer un nivel de significación para poder asegurar la linealidad del modelo. A un nivel de significación del 5 % tendremos que rechazar la hipótesis nula de que el ajuste del modelo es lineal pero, a un nivel mayor del 8 %, aceptaremos la hipótesis de linealidad. Por tanto, vemos conveniente fijar un nivel de $\alpha = 10\%$ para así aceptar las 3 hipótesis.

Capítulo 9

Conclusiones y trabajo futuro.

Uno de los puntos más importantes sobre los que reflexionar es el relativo a la viabilidad y adecuación de este proyecto. En condiciones totalmente reales ¿habríamos calculado un presupuesto adecuado?, ¿el DMP estaría ajustado a un tiempo realista?

Del mismo modo, podríamos preguntarnos si las variables predictoras escogidas son las más adecuadas, ¿existen variables predictoras con más capacidad de discriminación? Hemos escogido estas variables por ser temas de actualidad.: el cambio climático y la covid-19.

También cabe preguntarse si este proyecto influiría realmente en la toma de decisiones sobre el ámbito educativo. ¿El informe sería realmente tenido en cuenta?

Por último, sería interesante hacernos algunas preguntas sobre la svariables espaciales y temporales consideradas en los dataset. ¿Los países son una unidad espacial demasiado grande?. En lugar de países, ¿podríamos haber considerado comunidades autónomas y análogos?.Podríamos medir esas variables por zonas geográficas más ampliar. Podríamos ampliar la serie temporal y no limitarnos a un único año. Sería interesante considerar la media en las últimas décadas o años, tener en cuenta eventos con tanta capacidad de transformación como la pandemia actual.

Anexo

Notebooks disponibles.

- (I) **education_attainment.ipynb**. Desarrollo del capítulo 3.
- (II) **education_spend.ipynb**. Desarrollo del capítulo 4.
- (III) **co2.ipynb**. Desarrollo del capítulo 5.
- (IV) **vaccines.ipynb** Desarrollo del capítulo 6.
- (V) **combine_datasets.ipynb** Desarrollo del capítulo 7.
- (VI) **analysis.ipynb** Desarrollo del capítulo 8.

Datasets disponibles.

- (I) **adult_education_level.csv** Dataset sobre e educativo alcanzado [EO22a], capítulo 3.
- (II) **education_spending.csv** Dataset sobre el gasto educativo per cápita [EO22b], capítulo 4.
- (III) **co-emissions-per-capita.csv** Dataset sobre las emisiones de CO2 per cápita [Fri+21], capítulo 5.
- (IV) **vaccine.xlsx** Dataset completo sobre la confianza en las vacunas [De +20]. Capítulo 6
- (V) **vaccine.csv** Dataset seleccionado sobre la confianza en las vacunas [De +20].
- (VI) **target_education_0.csv** Dataset relultante de combinar los anteriores. Missing values como 0. Capítulo 7.
- (VII) **target_education_nan.csv** Dataset relultante de combinar los anteriores. Missing values como nan. Capítulo 7.

Bibliografía

- [Asu] Ministerio de Asuntos Exteriores. Gobierno de España. *¿Qué es la OCDE?* URL: <http://www.exteriores.gob.es/representacionespermanentes/ocde/es/quees2/paginas/default.aspx> (vid. pág. 13).
- [dat] datos.gob.es. *Principios FAIR: Buenas prácticas para la gestión y administración de datos científicos*. URL: <https://datos.gob.es/es/noticia/principios-fair-buenas-practicas-para-la-gestion-y-administracion-de-datos-cientificos> (vid. pág. 16).
- [De +20] Alexandre De Figueiredo y col. “Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: a large-scale retrospective temporal modelling study”. En: *The Lancet* 396.10255 (2020), págs. 898-908 (vid. págs. 39, 40, 42, 45, 57).
- [EOa] Organisation for Economic Cooperation y Development (OECD). *Acerca de la OCDE*. URL: <https://www.oecd.org/acerca/> (vid. pág. 13).
- [EOb] Organisation for Economic Cooperation y Development (OECD). “Adult Education Level (Indicator)”. En: (). DOI: 10.1787/36bce3fe-en (vid. págs. 14, 16, 17, 27).
- [EOc] Organisation for Economic Cooperation y Development (OECD). *Education at a Glance*. URL: <https://www.oecd.org/education/education-at-a-glance/> (vid. pág. 13).
- [EOd] Organisation for Economic Cooperation y Development (OECD). *OCDE: Glossary of Statistical Terms*. URL: <https://stats.oecd.org/glossary/detail.asp?ID=5385> (vid. págs. 14, 15).
- [EOe] Organisation for Economic Cooperation y Development (OECD). *OCDE: Glossary of Statistical Terms*. URL: <https://stats.oecd.org/glossary/detail.asp?ID=5450> (vid. págs. 14, 15).
- [EOf] Organisation for Economic Cooperation y Development (OECD). *OCDE: Glossary of Statistical Terms*. URL: <https://stats.oecd.org/glossary/detail.asp?ID=5568> (vid. págs. 14, 15).
- [EO22a] Organisation for Economic Cooperation y Development (OECD). “Adult Education Level (Indicator)”. En: (2022) (vid. págs. 38-43, 46, 57).
- [EO22b] Organisation for Economic Cooperation y Development (OECD). “Public Spending on education (Indicator)”. En: (2022) (vid. págs. 38-42, 44, 57).
- [Fri+21] Pierre Friedlingstein y col. “Global carbon budget 2021”. En: *Earth System Science Data Discussions* (2021), págs. 1-191 (vid. págs. 28, 38-40, 42, 44, 57).
- [OEC21] OECD. *Education at a Glance 2021*. 2021, pág. 474. DOI: <https://doi.org/https://doi.org/10.1787/b35a14e5-en>. URL: <https://www.oecd-ilibrary.org/content/publication/b35a14e5-en> (vid. págs. 14, 17).
- [Pow+07] Andy Powell y col. “DCMI abstract model”. En: (2007) (vid. pág. 16).