Máster en Ciencia de Datos

REFOMAD:

Análisis de la relación entre la Formación y la Renta en los distritos de Madrid

Ciclo de Vida de los Datos (M1968)

REFO★MAD

Renta y Formación en Madrid

11 de febrero de 2023

Miembros del equipo:

Mónica Alcantar Martínez
Sergio Bolívar Gómez
Samuel Laso Saro
María Peña Fernández
Ignacio de la Torre Cubillo
Juan José Velasco Horcajada

UC UNIVERSIDAD DE CANTABRIA

Resumen

El proyecto *REFOMAD* es una iniciativa sin fines de lucro con el objetivo de investigar la relación entre la formación académica y la renta neta media de los habitantes en cada distrito de Madrid. La meta es comprender cómo la educación influye en la economía de la ciudad y cómo se relaciona con la distribución de la riqueza. La investigación da comienzo el 1 de febrero de 2023 y se llevará a cabo en colaboración con el Instituto Nacional de Estadística (INE) a través del cuál se recopilarán parte de los datos del estudio.

Resumen Ejecutivo

El proyecto *REFOMAD* tiene como objetivo principal el estudio de la relación existente entre la renta y el nivel de formación alcanzado por los habitantes mayores de 25 años de los 21 distritos del Ayuntamiento de Madrid. El interés último del proyecto es analizar las tendencias y desigualdades observadas en las variables anteriores con el fin de comprender la economía de la educación en la ciudad.

El proyecto tiene una duración estimada de 4 meses (desde el 01/02/2023 hasta el 24/05/2023). Está estructurado en 7 paquetes de trabajo, cada uno de los cuales tiene asignado un coordinador que se encarga de la dirección de las tareas y garantizar su cumplimiento dentro de los plazos establecidos. El proyecto está dividido en seis grandes fases: la planificación (se elaboran el *Data Management Plan* y el plan de preservación), la adquisición de los datos, el procesado de los datos (curación y limpieza de datos), el análisis de los datos (obtención de valor añadido), la presentación del proyecto (comunicar resultados) y finalmente la difusión (dar a conocer el estudio debido al componente social del proyecto). Cada una de estas fases constituye un paquete de trabajo, dirigidos todos ellos por el de Coordinación y Gestión Económica.

En lo que se refiere al origen de los datos, se utilizan dos fuentes heterogéneas. Por un lado, los datos sobre el nivel de formación alcanzado por los residentes en los 21 distritos de Madrid se recaban por el equipo *REFOMAD* mediante la elaboración y difusión de encuestas, tanto online como en papel. Por otro lado, los datos acerca de la renta media en cada uno de los distritos se obtienen mediante una solicitud a medida al Instituto Nacional de Estadística (INE). Una vez adquiridos los datos, se utilizan herramientas de software libre, como Python, para su preprocesado, curación, visualización, análisis y publicación.

El presupuesto total del proyecto asciende a 42.260 €, estando destinados el 67 % a recursos humanos, el 20 % a recursos materiales y el 13 % a las reservas de contingencia. El carácter no lucrativo del proyecto permite la concurrencia a diferentes ayudas y subvenciones públicas para obtener financiación, a saber: la "*ayuda para la realización de proyectos de interés general*" de la Comunidad de Madrid, las "*subvenciones para proyectos vinculados a la colaboración con entidades del tercer sector*" del Ayuntamiento de Madrid, y las ayudas asociadas al programa "*Horizon2020*" de la Comisión Europea.

Finalmente, se espera gran interés por parte de las instituciones políticas que operan en la ciudad y una fuerte respuesta por parte de la población madrileña debido al fuerte componente social del proyecto. En efecto, las conclusiones del proyecto *REFOMAD* podrían proporcionar información muy valiosa y esclarecedora que pueda ser utilizada para tomar decisiones de índole política en materia educativa y/o económica en beneficio de toda la población.

Índice

1. Introducción

Los individuos en la sociedad demandan educación con el objetivo de mejorar su productividad en el mercado laboral, lo que les permitirá obtener mayores ganancias futuras, es decir, aspirar a tener mejor calidad de vida.

En la actualidad no es de extrañar que aumente el interés de los jóvenes por seguir estudios universitarios, y esto se deben principalmente a un par de factores. Por un lado, la educación superior no obligatoria, permite a un individuo tener mejor productividad en el mercado laboral y así aspirar una mejor remuneración o poder adquisitivo. Por otro lado, aspirar a un nivel de educación superior profesional, amplia las oportunidades en el mercado laboral, por lo que genera a su vez, estabilidad laboral. El proyecto *REFOMAD* se enfoca en encontrar la influencia que tiene el nivel de formación educativa en la distribución de la renta en un ámbito geográfico reducido, más concretamente en los 21 distritos del Ayuntamiento de Madrid.

El conocimiento de esta información es muy valioso, pues además de poder ser utilizado por gobiernos para fundamentar decisiones políticas en materia educativa, como la construcción de más centros educativos o la elaboración de campañas de fomento de la educación, puede ser de gran interés para comprender la dinámica económico-educativa de la ciudad de Madrid.

1.1. Objetivos. Resultados esperados.

Se analizará la relación existente entre la renta media de las personas y el nivel de formación educativa de los habitantes mayores de 25 años en los 21 distritos de Madrid. Se espera que en los distritos con mayor renta per cápita, el nivel de formación alcanzado sea más alto, y viceversa, que en los distritos donde los niveles de renta son inferiores también lo sean los niveles de formación. El análisis de estos datos generará como valor añadido conocimiento nuevo acerca de la economía de la educación (aspectos económicos relacionados con la educación) en la ciudad de Madrid, información valiosa que podrá ser utilizada para mejorar y/o reforzar la dinámica educativa y el éxito económico del área metropolitana mediante la toma de decisiones políticas.

1.2. Cobertura espacio-temporal

La población objeto de estudio en el proyecto *REFOMAD* son los habitantes mayores de 25 años de los 21 distritos del Ayuntamiento de Madrid. En consecuencia, el ámbito geográfico del proyecto comprende estos veintiún distritos, a saber: Puente de Vallecas, Villaverde, Usera, Carabanchel, Vicálvaro, Latina, Villa de Vallecas, Moratalaz, San Blas-Canillejas, Tetuán, Ciudad Lineal, Fuencarral-El Pardo, Hortaleza, Arganzuela, Centro, Moncloa-Aravaca, Barajas, Chamartín, Chamberí, Retiro y Salamanca.

En cuanto a la cobertura temporal del proyecto, aunque los datos se recopilan a principios del año 2023 a través de encuestas y solicitudes a los órganos competentes, éstos se refieren a la situación educativa y económica de los habitantes de la ciudad de Madrid durante el periodo 2016 al 2019.

1.3. Requisitos. Requerimientos técnicos

El proyecto *REFOMAD* requiere de la existencia de los siguientes datos para su ejecución:

- Datos de la renta media de las personas (renta per cápita) residentes en el ayuntamiento de Madrid para el periodo 2016-2019, desagregado por distritos.

- Datos del nivel de formación alcanzado por los residentes mayores de 25 del ayuntamiento de Madrid para el periodo 2016-2019, desagregado por distritos.

Como se comentará más adelante en la Subsección 3.2, los datos relativos al nivel del formación se obtendrán mediante la difusión de encuestas entre la población madrileña, mientras que los datos acerca de la renta media de las personas serán solicitados al Instituto Nacional de Estadística.

En cuanto a los requerimientos técnicos, además de una plataforma para la difusión online de las encuestas, serán necesarios recursos informáticos para poder llevar a cabo el preprocesado, curación, análisis y preservación de los datos. En particular, será necesario disponer de equipos informáticos que puedan utilizar los miembros del equipo, un servicio de almacenamiento en la nube y otro físico, además de la instalación del software necesario para el tratamiento de los datos (principalmente software libre, más concretamente Python y los paquetes `pandas` y `NumPy`).

1.4. Presentación del proyecto en formato LogFrame.

	Objetivos	**Indicador**	**Verificación**	**Condiciones, Riesgos**
Meta	Arrojar nuevo conocimiento acerca de la economía en el ámbito educativo, que pueda utilizarse para la toma de decisiones políticas en beneficio de la población madrileña	El valor añadido obtenido permite que puedan tomarse medidas políticas que favorezcan la dinámica educativa y el éxito económico	Informe de conclusiones del análisis llevado a cabo, con el valor añadido obtenido, que será entregado a las entidades interesadas	- El valor añadido obtenido no tiene relevancia para las entidades interesadas - Las conclusiones del proyecto no aportan nuevos conocimientos
Propósito	Encontrar la relación entre renta y nivel formativo global, y también con el número de centros de enseñanza	Se mantiene la relación a lo largo de los diferentes años de estudio	- Informe acerca del análisis global - Comparativa de las diferentes gráficas de renta y nivel formativo para cada año - Gráfica de centros de enseñanza y renta	- Errores en los datos recogidos del número de centros educativos - Diferencias entre los datasets de los diferentes años utilizados
Resultados	Establecer la relación entre renta y nivel formativo con los datos de 2016	Existe cierta relación, no hay aleatoriedad entre ambas variables	- Informe acerca del análisis realizado para los datos de 2016 - Gráficas que muestren la relación hallada	- No contar con suficientes datos al respecto - Fallos en el programa utilizado para el análisis
Componentes	- Elaboración del DMP y el plan de preservación. - Adquisición de datos sobre las rentas y el nivel de formación de la población madrileña, desagregado por los 21 distritos de la ciudad. - Contratación de los recursos humanos y materiales necesarios para la ejecución del proyecto. - Preprocesamiento de los datos - Análisis de los datos	**Presupuesto:** - Coste RRMM - Coste RRHH - Reservas de contingencia **Calendario:** - Febrero de 2023: Planificación y gestión económica - Marzo de 2023: Adquisición de datos - Abril de 2023: Procesado y análisis de datos - Mayo de 2023: Presentación	- Data Mangement Plan - Plan de Preservación - Resultados de las encuestas acerca del nivel de formación. - Conjunto de datos publicado en Zenodo. - Informe detallado del análisis realizado (memoria del proyecto). - Presupuesto final detallado.	- Adquisición de los datos en un tiempo razonable - Personal capacitado para su procesamiento y análisis

Tabla 1: Presentación del proyecto *REFOMAD* en formato LogFrame

2. Planificación, implementación y gestión del proyecto

2.1. Plan de trabajo

En este apartado se muestra la estructura y funciones de cada uno de los paquetes de trabajo del proyecto *REFOMAD*. Se listarán todos ellos en primer lugar, de manera que pueda tenerse un resumen de su duración, y seguidamente se procederá con una exposición del *Work Breakdown Structure*, donde se muestran las tareas dentro de cada uno de los paquetes de trabajo. Se mostrarán además los hitos que componen el proyecto, y un diagrama de Gantt donde aparece desarrollada la organización temporal del mismo.

2.1.1. Resumen de los paquetes de trabajo

Se muestra en la Tabla 2 un resumen de los 7 paquetes de trabajo en los que se divide el proyecto para garantizar el resultado final que se espera.

Paquete de Trabajo	Título	Coordinador	Fecha Inicio	Fecha Finalización	Duración
WP1	Coordinación y gestión económica	Ignacio	01/02/23	24/05/23	81 días
WP2	Planificación	Sergio	17/02/23	27/02/23	7 días
WP3	Adquisición de datos	Juan José	22/02/23	24/05/23	66 días
WP4	Procesado de datos	Juan José	10/04/23	24/04/23	11 días
WP5	Análisis de los datos y conclusiones	María	25/04/23	18/05/23	18 días
WP6	Presentación del proyecto y preservación	Mónica	19/05/23	24/05/23	4 días
WP7	Difusión	Samuel	27/02/23	10/04/23	31 días

Tabla 2: Resumen de los paquetes de trabajo y del cronograma del proyecto *REFOMAD*.

2.1.2. Estructura de Gestión del Proyecto: Hitos

En la Tabla 3 se muestran los hitos que han de alcanzarse durante el desarrollo del proyecto, con el fin de verificar que se están siguiendo los tiempos establecidos. Se especifica además el paquete de trabajo que se encarga en mayor medida de cada hito, y la fecha en la que se espera que este se alcance.

Número de Hito	WP involucrado	Identificación del Hito	Fecha
H1	1	Inicio del proyecto	01/02/23
H2	1	Finalización del proceso de registro del proyecto	10/02/23
H3	2	Entrega del Data Management Plan	22/02/23
H4	3	Fin del diseño de las encuestas	27/02/23
H5	3	Recopilación de la totalidad de los datos	10/04/23
H6	4	Fin del procesado de los datos	24/04/23
H7	5	Fin del análisis principal	08/05/23
H8	6	Comunicado a las entidades interesadas	23/05/23
H9	6	Subida de los datos y fin de proyecto	24/05/23

Tabla 3: Hitos del proyecto *REFOMAD*.

2.1.3. *Work Breakdown Structure* (WBS)

La Figura 9 muestra la estructura del proyecto con respecto a los paquetes de trabajo y las tareas que tienen a su cargo en forma de WBS.

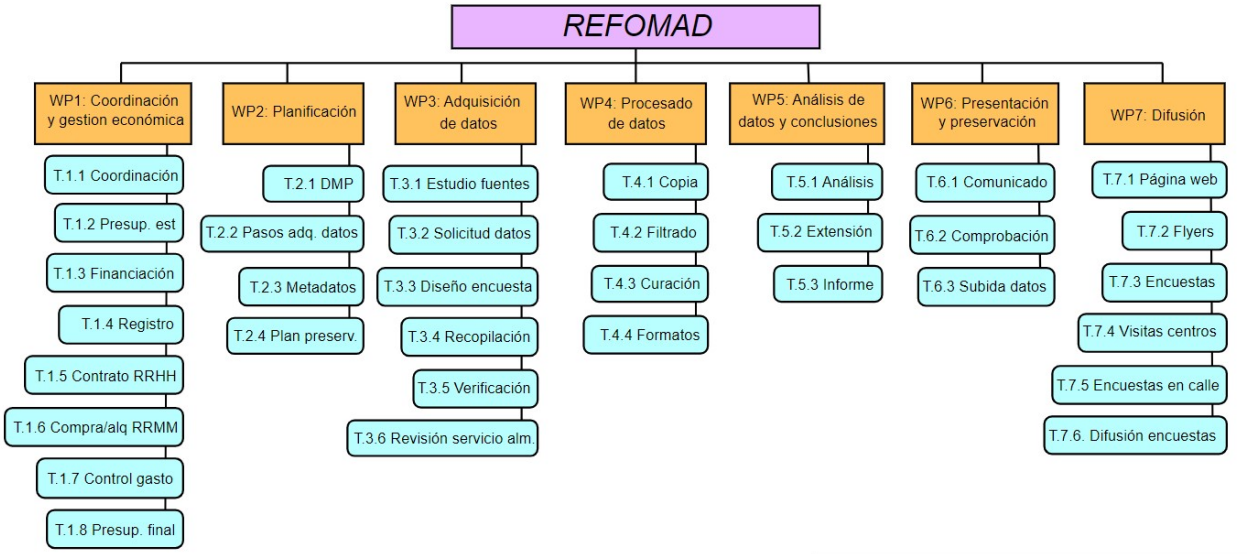


Figura 1: Organigrama del proyecto *REFOMAD* por paquetes de trabajo y tareas correspondientes.

2.1.4. Diagrama de Gantt

Se ha representado el diagrama de Gantt del proyecto *REFOMAD*, especificando los diferentes paquetes de trabajo involucrados y las fechas de inicio y fin. Se muestran también las diferentes relaciones entre cada uno de ellos. Aparece además detallado la duración y distribución de cada tarea dentro de cada uno de los paquetes de trabajo, al igual que las relaciones que se dan entre cada una de ellas. Cabe mencionar que el nombre que aparece junto a cada tarea corresponde con el responsable asociado de la misma, pero otros participantes del mismo WP podrán trabajar en ella.

Debido a las dimensiones del diagrama de Gantt, y para favorecer una mejor visualización del mismo, este ha sido adjuntado en el Apéndice A.

2.1.5. Paquetes de trabajo

A lo largo de este apartado se detalla la estructura y funciones de cada uno de los paquetes de trabajo mencionados en la Tabla 2. Se incluye tanto la fecha de inicio y fin de cada WP como el personal que trabajará en él, sus objetivos, tareas y los entregables que han de aportar para garantizar la correcta ejecución del proyecto.

Work Package	1	Coordinador del WP	Ignacio
Nombre del WP		**Coordinación y gestión económica**	
Participantes		Ignacio María	
Comienzo del WP	01/02/23	Fin del WP	24/05/23

Objetivos

1. Coordinar los distintos paquetes de trabajo.
2. Planificar el proyecto que se lleva a cabo.
3. Controlar que las tareas se realizan dentro de los plazos estipulados y con la calidad que se espera.
4. Elaborar el presupuesto del proyecto, controlar el gasto, y buscar posibles financiadores.
5. Gestionar la contratación de los recursos humanos y la compra de los recursos materiales necesarios para llevar a cabo el proyecto.

Descripción del trabajo

Tarea 1.1. Coordinación.
Se planificará el proyecto y se coordinarán los distintos paquetes de trabajo.
Tarea 1.2. Previsión del presupuesto.
Se estimará el presupuesto total del proyecto, incluyendo recursos humanos y materiales, licencias y tasas.
Tarea 1.3. Búsqueda de financiación.
Se buscarán ayudas y subvenciones para financiar al menos cierta parte del proyecto.
Tarea 1.4. Registro del proyecto.
Se registrará el proyecto en el Ayuntamiento de Madrid.
Tarea 1.5. Contratación de recursos humanos.
Se contratará a los RRHH del proyecto y se fijarán sus salarios.
Tarea 1.6. Compra/alquiler de recursos materiales.
Se buscarán las mejores alternativas para los recursos materiales necesarios.
Tarea 1.7. Control del gasto.
Se llevará a cabo un control exhaustivo de los gastos que acarrea la ejecución del proyecto, ajustando el presupuesto estimado inicialmente.
Tarea 1.8. Elaboración del presupuesto final.
Se elaborará el presupuesto final detallado teniendo en cuenta los gastos imprevistos durante la ejecución del proyecto.

Entregables

Entregable 1.1. Calendario de ejecución del proyecto y organigrama de responsabilidades.
Entregable 1.2. Previsión del presupuesto con los convenios de colaboración alcanzados.
Entregable 1.3. Resguardo del registro del proyecto.
Entregable 1.4. Informes de fiscalización (gastos imprevistos, facturas).
Entregable 1.5. Copia de los contratos de trabajo de los RRHH.
Entregable 1.6. Presupuesto final detallado.
Entregable 1.7. Documento final detallado de la evolución del proyecto a fecha de finalización.

Tabla 4: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Coordinación y gestión económica (WP1)".

Work Package	2	Coordinador del WP	Sergio
Nombre del WP		**Planificación**	
Participantes		Sergio Mónica	
Comienzo del WP	17/02/23	Fin del WP	27/02/23
Objetivos			
1. Preparar el Data Management Plan. 2. Planificar como se hará la adquisición de los datos. 3. Definir los metadatos y el plan de preservación de los datos, asegurando que cumpla los principios FAIR.			
Descripción del trabajo			
Tarea 2.1. Preparación del Data Management Plan. Se formalizará el DMP del proyecto. **Tarea 2.2. Estudio de los pasos que han de seguirse para la adquisición de los datos.** Se realizará un análisis sobre las diferentes alternativas que se tienen para obtener los datos necesarios para el estudio. **Tarea 2.3. Definición de los metadatos de los datos.** Se establecerán los metadatos que acompañarán a los datos que se generen. **Tarea 2.4. Definición del plan de preservación.** Se estudiarán las diferentes opciones y plataformas donde poder publicar los datos y los análisis realizados para su acceso posterior.			
Entregables			
Entregable 2.1. Data Management Plan. **Entregable 2.2.** Documento detallado con las diferentes alternativas para la adquisición de los datos. **Entregable 2.3.** Documento con los metadatos Dublin Core creados. **Entregable 2.4.** Informe del plan de preservación que se ha definido para los datos y el análisis realizado sobre ellos.			

Tabla 5: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Planificación (WP2)".

Work Package	3	Coordinador del WP	Juan José
Nombre del WP		**Adquisición de datos**	
Participantes		Juan José María	
Comienzo del WP	22/02/23	Fin del WP	24/05/23

Objetivos

1. Estudiar las fuentes de datos donde puedan adquirirse los datos necesarios de la renta.
2. Tramitar la solicitud de los datos a la fuente de datos seleccionada.
3. Generar una encuesta para la recogida de datos educativos, y recoger los resultados de las mismas.
4. Organizar el servicio de almacenamiento de los datos.

Descripción del trabajo

Tarea 3.1. Estudio de fuentes de datos de la renta per cápita en los distritos madrileños.
Se buscarán entre las posibilidades existentes aquellas fuentes a las que puedan solicitarse los datos necesarios para el estudio.
Tarea 3.2. Tramitación de la solicitud necesaria.
Se llevarán a cabo los trámites administrativos necesarios para adquirir los datos.
Tarea 3.3. Diseño de una encuesta.
Se generará una encuesta con las preguntas precisas que permitan obtener los datos educativos de la población madrileña que necesitamos para nuestro estudio.
Tarea 3.4. Recopilación de resultados.
Se recogerán los resultados de las encuestas y se generará un dataset con los datos correspondientes.
Tarea 3.5. Verificación de la calidad de los datos adquiridos.
Se comprobará que los datos que han sido recogidos de las fuentes son de la calidad necesaria para llevar a cabo el estudio, correspondiendo a los años requeridos y apareciendo toda la información solicitada.
Tarea 3.6. Revisión regular del servicio de almacenamiento de los datos.
Se realizarán revisiones periódicas en el servicio de almacenamiento, comprobando que no haya fallos que puedan perjudicar la ejecución del proyecto.

Entregables

Entregable 3.1. Informe con las conclusiones obtenidas acerca de las diferentes fuentes de datos.
Entregable 3.2. Resguardo de la solicitud realizada al INE.
Entregable 3.3. Informe de evaluación de la calidad de los datos adquiridos.
Entregable 3.4. Encuesta realizada a la población madrileña que ha de estar disponible online para ser respondida, y también en papel.
Entregable 3.5. Dataset donde se recogen los datos de los niveles educativos alcanzados por los madrileños obtenidos a partir de las encuestas.
Entregable 3.6. Tabla de las revisiones realizadas del servicio de almacenamiento, con la fecha de las mismas.

Tabla 6: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Adquisición de datos (WP3)".

Work Package	4	Coordinador del WP	Juan José
Nombre del WP		**Procesado de datos**	
Participantes		Juan José Samuel	
Comienzo del WP	10/04/23	Fin del WP	24/04/23

Objetivos

1. Realizar un filtrado de los datos a partir de los adquiridos, y la curación de los mismos una vez han sido filtrados.
2. Adecuar los datos provenientes de ambas fuentes para poder realizar el análisis posterior.

Descripción del trabajo

Tarea 4.1. Generación de una copia de los datos adquiridos.
Se copiarán los datos recopilados para tener un backup en caso de fallos posteriores.
Tarea 4.2. Filtración de los datos.
Se eliminarán de los datos adquiridos aquellos que carezcan de importancia para el estudio que se realiza.
Tarea 4.3. Curación de los datos.
Se revisará la presencia de erratas, inconsistencias y la falta de valores, y se corregirán cuando corresponda tras un estudio de las posibles causas de esos errores.
Tarea 4.4. Cambio de formatos.
Se adecuarán los datos provenientes de las dos fuentes de datos con el fin de facilitar el trabajo con los mismos.

Entregables

Entregable 4.1. Backup de los datos adquiridos de las fuentes.
Entregable 4.2. Ficheros de datos únicamente con la información necesaria para el análisis, y tras haber realizado la curación y el cambio de formato de los mismos.

Tabla 7: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Procesado de datos (WP4)".

Work Package	5	Coordinador del WP	María
Nombre del WP		**Análisis de los datos y conclusiones**	
Participantes		María Juan José Sergio	
Comienzo del WP	25/04/23	Fin del WP	18/05/23
Objetivos			
1. Estudiar las relaciones existentes entre las diferentes variables que interesa analizar a través de diferentes análisis. 2. Obtener un valor añadido a partir del estudio realizado.			
Descripción del trabajo			
Tarea 5.1. Estudio de las correlaciones entre variables. Se realizarán diversos análisis sobre las variables bajo estudio para comprobar la existencia de correlaciones entre las mismas. **Tarea 5.2. Extensión de los resultados obtenidos a otros años para los que se tengan datos, y el estudio de la variabilidad de los mismos.** Se estudiará la variabilidad de los datos para otros años y se comprobará si las correlaciones encontradas inicialmente se conservan. **Tarea 5.3. Realización de un informe de resultados.** Se realizará un informe detallado de todos los análisis realizados sobre las variables, el valor añadido obtenido a partir de estos y las conclusiones alcanzadas.			
Entregables			
Entregable 5.1. Archivo .ipynb con el código utilizado en los análisis realizados. **Entregable 5.2.** Informe detallado con los análisis realizados sobre las variables de interés y los estudios extra generados a partir de ellos. **Entregable 5.3.** Informe de resultados que será entregado a las entidades interesadas.			

Tabla 8: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Análisis de los datos y conclusiones (WP5)".

Work Package	6	Coordinador del WP	Mónica
Nombre del WP		**Presentación del proyecto y preservación**	
Participantes		Mónica Ignacio	
Comienzo del WP	19/05/23	Fin del WP	24/05/23
Objetivos			
1. Comunicar a las entidades interesadas los resultados obtenidos en el estudio. 2. Seguir el plan de preservación establecido por el WP 2, que asegure que se cumplan los principios FAIR.			
Descripción del trabajo			
Tarea 6.1. Comunicación a las entidades interesadas. Se pondrá en contacto con las entidades interesadas en el estudio y se les adjuntará el informe realizado por el WP 5 donde se exponen las conclusiones obtenidas. **Tarea 6.2. Comprobación de la información a subir a repositorios.** Se comprobará que puede mantenerse la interoperabilidad, y que además la información es correcta y con metadatos detallados para poder permitir la reutilización de los datos. **Tarea 6.3. Subida al repositorio especificado en el plan de preservación.** Se subirán a Zenodo los diferentes datasets utilizados junto con los metadatos, además del propio análisis realizado para el estudio.			
Entregables			
Entregable 6.1. Documento de las comunicaciones mantenidas con las entidades interesadas. **Entregable 6.2.** Informe de valoración de la calidad de la información a subir al repositorio.			

Tabla 9: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Presentación del proyecto y preservación (WP6)".

Work Package	7	Coordinador del WP	Samuel
Nombre del WP		**Difusión**	
Participantes		Samuel Sergio	
Comienzo del WP	27/02/23	Fin del WP	10/04/23
Objetivos			
1. Dar a conocer el estudio que se va a llevar a cabo con el fin de animar a la gente a realizar la encuesta para recabar datos sobre educación. 2. Repartir las encuestas, tanto en papel como digitalmente			
Descripción del trabajo			
Tarea 7.1. Creación y mantenimiento de una página web que dé acceso a la encuesta. Se creará y mantendrá una página web donde los encuestados puedan encontrar información acerca de los diferentes aspectos sobre los que se les pregunta, los fines de las encuestas, y el acceso a las mismas para su realización online. **Tarea 7.2. Diseño de flyers.** Se diseñarán panfletos para repartir en los centros educativos y en las calles donde aparece el link para acceder a la encuesta. **Tarea 7.3. Impresión de encuestas.** Se estimará el número de encuestas que puedan ser respondidas en papel y se imprimirán para ser repartidas junto con los flyers. **Tarea 7.4. Visitas a centros.** Se organizarán reuniones en distintos centros educativos, culturales y sociales de Madrid donde se divulgue entre los ciudadanos acerca de los análisis de datos y la importancia de participar en encuestas como la que diseña el WP3. **Tarea 7.5. Encuestas en la calle.** Se realizarán jornadas para encuestar en calles transitadas para que aquellos que estén interesados la realicen en ese momento. **Tarea 7.6. Difusión de las encuestas entre los ciudadanos de Madrid.** Se mandará por correo electrónico a los madrileños el link de acceso a la encuesta para ser respondida online.			
Entregables			
Entregable 7.1. Página web con acceso a la encuesta. **Entregable 7.2.** Encuestas impresas para ser repartidas. **Entregable 7.3.** Panfletos con información acerca de las encuestas.			

Tabla 10: Descripción de los datos, objetivos, tareas y entregables correspondientes al paquete de trabajo "Difusión (WP7)".

2.1.6. Riesgos y seguridad

En todo proyecto existen ciertos riesgos que pueden hacer peligrar la consecución de los objetivos. Con el fin de minimizar las consecuencias que podrían llevar asociados esos problemas, han de trazarse planes de contingencia. En la Tabla 11 se listan los riesgos que pueden afectar a las actividades que se llevan a cabo en la duración de todo el proyecto. Se muestran también tanto la probabilidad de que ocurra, como el impacto en el desarrollo del proyecto y el plan de contingencia asociado.

Riesgo	Probabilidad	Impacto	Plan de contingencia
Problemas de comunicación entre los empleados.	3	6	Contratar un mayor número de coordinadores y revisar la red de comunicación.
Problemas para el registro del proyecto.	2	9	Revisar de los documentos, cambiar aquello que no se ha aceptado y reenviar la documentación.
Tareas retrasadas respecto al plan establecido.	6	6	Estudiar el motivo del retraso, planificar nuevos calendarios y comunicar a las empresas el cambio.
Gastos muy superiores a los esperados.	3	7	Estudiar a qué se debe la anomalía, establecer nuevos presupuestos y buscar más financiación.
Difusión del proyecto poco efectiva.	3	9	Estudiar nuevas medidas de difusión del proyecto y puesta en marcha de las mismas.
Poca participación en las encuestas.	3	10	Estudiar nuevas medidas para ampliar el alcance de las mismas.
Fallos en la página web y del resto de RRSS.	5	5	Contratar una persona de refuerzo que se asegure de su correcto funcionamiento.
La calidad de los datos recogidos no es buena.	3	8	Solicitar de nuevo los datos a la fuente.
Problemas con el programa Python.	4	5	Estudiar los fallos del sistema. Explorar otros como R.
Problemas para la subida de datos.	3	3	Estudiar los fallos que se estén dando y buscar nuevas alternativas.

Tabla 11: Riesgos del proyecto *REFOMAD*, junto con los planes de contingencia.

2.2. Gestión económica

El presupuesto del proyecto *REFOMAD* se divide en dos grandes bloques: los recursos materiales (que incluyen subcontrataciones de servicios) y los recursos humanos. Además, se ha incluido una reserva de contingencia que supone el 15 % del total presupuestado para los recursos humanos y materiales, de acuerdo con las leyes aplicables, para cubrir eventos y/o riesgos previstos pero que no se tiene una certeza de que vayan a ocurrir, como horas extra, retrasos en la ejecución del proyecto, deterioro de algunos de los recursos materiales o aumento de los precios de los servicios contratados, entre otros.

2.2.1. Recursos materiales

Los recursos materiales del proyecto se adquirirán a través de distintos proveedores. Por un lado, los ordenadores y el disco duro externo SSD se comprarán en la empresa *MediaMarkt* [1, 2]. El material de oficina, los trípticos informativos y las copias de las encuestas en papel serán suministrados por la empresa *Lyreco* [3]. El servicio de almacenamiento en la nube se subcontratará a *Google Cloud* [4], mientras que *WIX* [5] será la empresa encargada del hosting de la página web. Finalmente, el informe con los datos sobre la renta media de las personas en los veintiún distritos del Ayuntamiento de Madrid se encargará al INE (las tarifas por este servicio están recogidas en la *Resolución de 1 de septiembre de 2021, del Instituto Nacional de Estadística, por la que se regulan los precios privados de los productos de difusión del organismo* [6]).

En la Tabla 12 se incluye el presupuesto de los recursos materiales y de las subcontrataciones

del proyecto REFOMAD. Se indica, para cada uno de los conceptos, el precio por unidad, el número de unidades y el precio total en euros.

Concepto	**Precio unitario** (€)	**Unidades**	**Precio total** (€)
Material de oficina	200	6	1200
Ordenadores portátiles	600	6	3600
Encuestas en papel [7]	0,02	20000	400
Disco duro SSD externo (1TB)	90	1	90
Servicio de almacenamiento *cloud* (1TB)	23	4	92
Informe del INE	912	1	912
Página web	17	4	68
Trípticos Informativos	0,05	40000	2000
COSTE TOTAL RRMM PROYECTO			**8.362**

Tabla 12: Presupuesto detallado de los recursos materiales (incluidas subcontrataciones) necesarios para ejecutar el proyecto *REFOMAD*.

2.2.2. Recursos humanos

La Tabla 13 recoge los días de trabajo y la nómina total de los miembros del equipo. Se ha considerado que la semana tiene cinco 5 laborables (lunes a viernes) y que la jornada completa es de 8 horas diarias con turno partido: de 8:00 h de la mañana a las 17:00 h de la tarde, con un descanso de 13:00 h a 14:00 h para la comida.

Nombre	**Categoría**	**Días de trabajo**	**Salario/día (€)**	**Nómina (€)**
María	Gestora de Proyectos	81	95	7.695
Ignacio	Economista	81	61	4.941
Juan José	Data Scientist	66	90	5.940
Sergio	Data Scientist	56	90	5.040
Samuel	Data Scientist	42	90	3.780
Mónica	Data Scientist	11	90	990
COSTE TOTAL RRHH PROYECTO				**28.386**

Tabla 13: Resumen de los recursos humanos del proyecto *REFOMAD*, junto con su nómina total desglosada. El salario de las distintas profesiones se ha consultado en el portal *Indeed* [8].

2.2.3. Presupuesto total

En la Tabla 14 se incluye el presupuesto total del proyecto *REFOMAD* desglosado en recursos materiales, recursos humanos y reservas de contingencia.

Concepto	**Coste** (€)
Recursos materiales	8.362
Recursos humanos	28.386
Reservas de contingencia (15 % de RRMM + RRHH)	5.512
COSTE TOTAL DEL PROYECTO	**42.260**

Tabla 14: Presupuesto total del proyecto *REFOMAD*.

La Figura 2 muestra la distribución del presupuesto del proyecto *REFOMAD* en forma de diagrama de sectores.

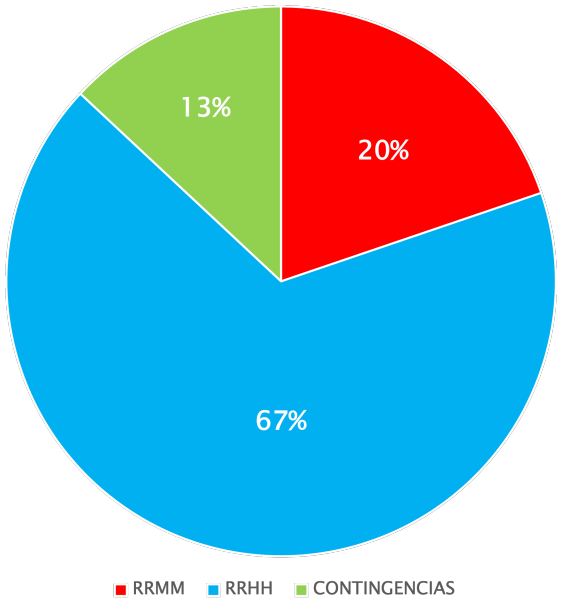


Figura 2: Presupuesto total del proyecto *REFOMAD* representado en diagrama de sectores.

2.2.4. Financiación, subvenciones y ayudas

Un aspecto importante del proyecto *REFOMAD* es que no tiene fines lucrativos, es decir, no busca obtener un beneficio económico. Se trata de un estudio sobre la economía de la educación en el Ayuntamiento de Madrid cuyo único fin es de interés social ya que pretende ilustrar la relación entre la renta media de las personas residentes en los distintos distritos de la ciudad y su nivel de formación. Por lo tanto, dado que los resultados del proyecto podrían ser de gran interés para instituciones políticas que operan en la ciudad, se espera que estos organismos puedan financiar total o parcialmente el proyecto.

Por un lado, el proyecto *REFOMAD* sería susceptible de recibir la ayuda para la realización de proyectos de interés general para atender a fines de interés social de la Comunidad de Madrid [9], con la que podría financiarse en su totalidad. Por otro lado, también podría solicitar las subvenciones para proyectos vinculados a la colaboración con entidades del tercer sector 2022 del Ayuntamiento de Madrid [10], con las que podría financiarse de manera parcial. Finalmente, se podría recurrir a los Fondos Europeos para la Investigación [11], específicamente a la subvención Horizon2020 [12], que podría cubrir parcialmente los costes del proyecto destinados a promover que los datos generados cumplan los principios FAIR (en nuestro caso, los recursos humanos porque el software utilizado y el repositorio empleado son gratuitos).

3. Ciclo de vida de los datos en el proyecto

Esta sección está dedicada en su totalidad a describir detalladamente el ciclo de vida de los datos en el proyecto *REFOMAD*; cubriendo desde la planificación de la recogida de datos hasta su publicación y preservación.

NOTA: *En la sección anterior, se ha asumido de manera ficticia que el equipo de REFOMAD era el encargado de recoger los datos, a través del paquete de trabajo "Adquisición de datos (WP3)", para dar un toque más realista al proyecto. Sin embargo, siguiendo las guías de la asignatura, para la elaboración de este análisis y de esta memoria, se han utilizado datos en abierto previamente publicados por el INE y el Ayuntamiento de Madrid, como se detalla en el Plan de Gestión de Datos. Aunque los datos son reutilizados, en la Subsección 3.2 se explica cómo ficticiamente el equipo de REFOMAD adquiría los datos.*

3.1. Plan de Gestión de Datos (DMP)

3.1.1. Información general

Nombre del proyecto: *REFOMAD: análisis de la relación entre la formación y la renta en los distritos de Madrid.*

Contribuidores:

- Mónica Alcantar Martínez
- Sergio Bolívar Gómez
- Samuel Laso Saro
- María Peña Fernández
- Ignacio de la Torre Cubillo
- Juan José Velasco Horcajada

Descripción: Este documento presenta el Plan de Gestión de Datos (DMP, por sus siglas en inglés) del proyecto "*REFOMAD: análisis de la relación entre la formación y la renta en los distritos de Madrid*". Este DMP se ha elaborado siguiendo las guías sobre la gestión de datos de los proyectos del programa Horizon 2020 de la Comisión Europea [13], asegurando así el cumplimiento de los principios FAIR (*findable*, *accesible*, *interoprable* y *reusable*, del inglés).

Instituciones involucradas: Universidad de Cantabria (UC), Universidad Internacional Menéndez Pelayo (UIMP).

Versión: v.1.0

3.1.2. Resumen

- **¿Se van a reutilizar datos existentes?**

  Sí. El proyecto reutilizará datos existentes procedentes de dos fuentes gubernamentales. Los administradores de datos del proyecto evaluarán minuciosamente la calidad y relevancia de los datos existentes que se van a reutilizar, y se asegurarán de que se atribuyan y citen adecuadamente en cualquier análisis o informe.

Por un lado, los datos sobre la renta media de las personas en los veintiún distritos de Madrid se toman del Atlas de Distribución de Renta de los Hogares (ADRH), una operación estadística elaborada por el INE que se basa en la explotación de registros administrativos con el objetivo de obtener información sobre el nivel y la distribución de renta desglosada según variables demográficas básicas de la población a nivel territorial muy detallado (todos los municipios, distritos y secciones censales en que se organiza territorialmente el Estado) [14]. Los datos están disponibles en esta página web, que fue consultada por última vez el 8 de enero de 2023 a las 19:45 horas.

Por otro lado, los niveles de formación de los adultos mayores de 25 años que residen en los distintos distritos del Ayuntamiento de Madrid y la información acerca de los centros educativos se toman del *"Panel de indicadores de distritos y barrios de Madrid"*, un estudio sociodemográfico elaborado por el Ayuntamiento de Madrid que ofrece una visión territorial de las variables socioeconómicas, de salud, demográficas, educativas, calidad de vida, vivienda, medio ambiente, equipamientos municipales, participación ciudadana y presupuesto de los distritos de Madrid [15]. Los datos están disponibles en esta página web, que fue consultada por última vez el 8 de enero de 2023 a las 19:50 horas.

- **¿Para quién podrían ser útiles los datos (fuera del proyecto)?**

  En general, la utilidad de los datos de este proyecto es alta, ya que la información obtenida del análisis de los datos podría servir de base para una serie de intervenciones e iniciativas destinadas a mejorar la educación y las condiciones económicas en los diferentes distritos de Madrid. Por ejemplo, los datos generados en este proyecto podrían ser de utilidad para muchos colectivos, incluidos investigadores, instituciones gubernamentales, responsables políticos y/o pedagogos que trabajan para mejorar la educación y la economía del Ayuntamiento de Madrid. Asimismo, los datos también podrían ser de interés para el público en general que desee comprender la dinámica de la educación y los ingresos de los distintos distritos de la ciudad.

- **¿Qué tipos y formatos de datos generará o reutilizará el proyecto?**

  Se utilizarán únicamente datos de tipo numérico, incluyendo enteros y floats, con excepción de las etiquetas que identifican a las variables de interés y a los distritos del Ayuntamiento de Madrid, que serán de tipo caracter. Los datos, tanto los originales como los generados para el análisis, tendrán un formato estructurado. En particular, los datos recopilados del Ayuntamiento de Madrid y los generados para el proyecto serán archivos Excel XSLX, mientras que los recopilados por el INE tienen formato CSV (*Comma-Separated Values*, del inglés).

- **¿Cuál es la finalidad de la generación o reutilización de datos y su relación con los objetivos del proyecto?**

  La finalidad de la reutilización de los datos en este proyecto es analizar en profundidad el nivel de formación alcanzado por los habitantes mayores de 25 años de los distintos distritos de Madrid en función de la renta media de las personas en estos distritos. Se trata de un aspecto crucial de los objetivos del proyecto, ya que comprender cómo se correlacionan estos factores puede proporcionar información muy valiosa sobre la dinámica educativa y económica del ayuntamiento. Por ejemplo, si los datos revelan que los distritos con mayores niveles de renta tienden a tener mayores niveles de formación, esto podría indicar que la educación es un factor clave para impulsar el éxito económico en estas zonas. Por otro lado, si los datos

muestran que los distritos con rentas medias más bajas tienden a tener niveles educativos más bajos, esto podría sugerir que la falta de educación está contribuyendo a los problemas económicos de estos distritos. En cualquier caso, los datos recogidos y analizados a través de este proyecto serán de gran interés para fundamentar decisiones de índole política, así como las inversiones destinadas a mejorar la educación y la economía del ayuntamiento de Madrid.

- **¿Cuál es el tamaño previsto de los datos que se pretenden generar o reutilizar?**

  Los datos que se reutilizarán en el proyecto no serán de gran tamaño. Por ejemplo, el archivo del INE que contiene la información sobre la renta neta media en los 21 distritos del Ayuntamiento de Madrid durante los años 2016-2019 tiene un tamaño de 68 kB, mientras que los archivos obtenidos del portal de datos abiertos del Ayuntamiento de Madrid con información sobre formación y centros educativos durante los años 2016-2019 tienen un tamaño total de 21 MB (incluyen variables adicionales a las de interés).

  En cuanto a los datos generados para el análisis, se espera que su tamaño sea reducido. En particular, se medirán 15 variables numéricas en 22 localizaciones (los 21 distritos y la ciudad en su conjunto). Suponiendo que cada valor ocupa 8 bytes, esta información ocuparía aproximadamente 22x15x8 bytes = 2.64 kB. Si también se consideran las variables de tipo caracter que identifican las localizaciones (22) y las variables de estudio (15), y se asume que como máximo tendrán 50 caracteres ASCII (cada uno ocupa 1 byte), entonces a la cantidad anterior deberíamos sumar aproximadamente 37x50 bytes = 1.85 kB. En total, los datos de cada año ocuparían 4.50 kB, lo que multiplicado por los 4 años que se analizan da un total de 18 kB.

3.1.3. Principios FAIR

Localización de los datos, incluidos los metadatos

- **¿Se identificarán los datos mediante un identificador persistente?**

  Sí, se utilizará un identificador persistente para los datos del proyecto, asegurando así que puedan ser encontrados, identificados y referenciados fácilmente por investigadores y/o otras partes interesadas (stakeholders). Para ello, se asignará un DOI (Digital Object Identifier) [16] al conjunto de datos, que permitirá su localización y acceso a largo plazo.

- **¿Se proporcionarán metadatos enriquecidos para facilitar el descubrimiento/localización de los datos? ¿Qué metadatos se crearán? ¿Qué estándares de metadatos se seguirán?**

  Sí, se proporcionarán metadatos para facilitar el descubrimiento y/o la localización del conjunto de datos utilizado en el proyecto. Los metadatos incluirán, como mínimo, la siguiente información sobre los datos: título, creadores, palabras clave, descripción, contribuidores, fecha, tipo, formato, identificador, idioma, cobertura espacio-temporal y derechos o licencias. Se utilizarán, al menos, los esquemas de metadatos Dublin Core [17] y DataCite [18] para garantizar que los metadatos sean fácilmente localizables, comprensibles y accesibles para un público amplio.

- **¿Se incluirán palabras clave en los metadatos para optimizar la posibilidad de descubrimiento y posterior reutilización?**

Sí, se proporcionarán palabras clave que caractericen adecuadamente el contenido y el contexto de los datos para maximizar las posibilidades de descubrimiento y el potencial de reutilización. Estas palabras clave se incluirán en los metadatos que acompañan a los datos y se seleccionarán para aumentar la facilidad con la que otros investigadores y/o stakeholders pueden encontrar los datos.

- **¿Se ofrecerán los metadatos de forma que puedan ser recolectados e indexados?**

  Sí. El conjunto de datos utilizado en el proyecto se subirá al repositorio abierto Zenodo [19], que se adhiere al protocolo OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting) [20]. Se trata de un protocolo de interoperabilidad para el intercambio y la difusión de metadatos procedentes de diversos repositorios. En concreto, este protocolo permite la recolección de metadatos utilizando peticiones HTTP.

Accesibilidad en abierto de los datos

- **¿Se depositarán los datos en un repositorio seguro?**

  Sí, como se ha indicado anteriormente, los datos se depositarán en Zenodo [19], que es un repositorio de datos abiertos que cumple las normas de conservación y acceso a los recursos digitales establecidas por la iniciativa OpenAIRE de la Unión Europea [21], una red de repositorios cuyo objetivo es proporcionar acceso abierto a la investigación financiada por la Comisión Europea.

- **¿Garantiza el repositorio anterior que se asigne un identificador persistente a los datos?**

  Sí, Zenodo garantiza que los conjuntos de datos que se depositan en su plataforma sean asignados un identificador persistente. En concreto, Zenodo asigna un DOI (Digital Object Identifier) a cada recurso digital que se deposita en el repositorio, lo que permite acceder de manera sencilla y rápida al recurso a través de un enlace. Cuando un usuario hace clic en el DOI, se le redirige a la página de destino del recurso, donde se proporciona información sobre él y se permite al usuario acceder al contenido del mismo.

- **¿Los datos estarán disponibles en abierto?**

  Sí, el conjunto de datos estará disponible en abierto en el repositorio de Zenodo, lo que facilitará la accesibilidad tanto para la comunidad científica como para el público en general.

- **Si se aplica un embargo temporal de los datos, especifique por qué y cuánto tiempo se aplicará, teniendo en cuenta que los datos de la investigación deben estar disponibles lo antes posible.**

  No aplica. Los datos que se utilizan en el proyecto están publicados en repositorios gubernamentales de datos en abierto y son visibles para el público en general sin ningún tipo de embargo temporal.

- **¿Se podrá acceder a los datos mediante un protocolo de acceso gratuito y normalizado?**

  Sí, los datos utilizados en el proyecto serán accesibles a través de protocolos de acceso gratuitos

y estandarizados, como formatos universales y estructurados como Excel XLSX. También se podrá acceder a los datos a través de APIs (Application Programming Interfaces, del inglés). En el caso del repositorio de Zenodo, existe la Zenodo REST API [22] que, entre otras cosas, permite la publicación y descarga de recursos digitales.

- **Si existen restricciones de uso, ¿cómo se facilitará el acceso a los datos, tanto durante el proyecto como una vez finalizado?**

  No aplica. Los datos serán accesibles para el público general sin ningún tipo de restricción tanto durante el proyecto como una vez finalizado.

- **¿Cómo se determinará la identidad de la persona que accede a los datos?**

  No aplica. No será necesario que el usuario se identifique para acceder a los datos.

- **¿Se pondrán los metadatos a disposición del público y bajo licencia CC0 de dominio público? ¿Contendrán los metadatos información que permita al usuario acceder a los datos?**

  Sí, los metadatos se pondrán a disposición del público general bajo la licencia Creative Commons Zero (CC0) [23], lo que significa que cualquier usuario podrá utilizarlos para cualquier fin, incluido el uso comercial, sin necesidad de obtener permiso ni pagar ningún royalty. Estos metadatos se ceden al dominio público con el objetivo de hacerlos lo más accesibles posible, reconociendo al mismo tiempo la contribución de los creadores. Además, estos metadatos incluirán el identificador persistente de tipo DOI asociado a los datos, así como un enlace que permite descargar directamente el conjunto de datos al hacer clic en él.

- **¿Durante cuánto tiempo seguirán estando disponibles y localizables los datos? ¿Se garantizará la disponibilidad de los metadatos cuando los datos dejen de estar disponibles?**

  Los datos estarán disponibles y/o localizables durante el máximo tiempo posible (idealmente, de forma permanente). En particular, el equipo del proyecto intentará que los datos estén disponibles para su reutilización, al menos, mientras se consideren útiles y relevantes para la comunidad científica. Si por alguna razón los datos dejan de ser accesibles, los metadatos seguirán estando disponibles y proporcionarán información sobre los datos, permitiendo así a los investigadores y/o stakeholders comprender el alcance y el contexto del conjunto de datos. Esto último estará garantizado gracias al "Plan de Preservación de Datos" elaborado para el proyecto (ver Sección 3.5).

- **¿Se incluirá documentación sobre algún software necesario para acceder a los datos o leerlos? ¿Será posible incluir el software pertinente (por ejemplo, en código abierto)?**

  El campo de descripción de los metadatos especificará que el conjunto de datos utilizado en el proyecto se publica en formato Excel XLSX y mencionará las herramientas utilizadas para leer, curar y analizar los datos, que son Python y los paquetes `NumPy` y `pandas`.

Interoperabilidad de los datos

- **¿Qué vocabularios de datos y metadatos, normas, formatos o metodologías se seguirán para que los datos sean interoperables y permitan su intercambio y reutilización dentro de cada disciplina y entre ellas? ¿Seguirá las mejores prácticas de interoperabilidad aprobadas por la comunidad? ¿Cuáles?**

  El conjunto de datos utilizado en el proyecto se procesará y publicará de manera que sea lo más interoperable posible para que el resto de la comunidad pueda reutilizarlo fácilmente. Se utilizarán herramientas de software libre, como Python, que promuevan la interoperabilidad de los datos y se seguirán buenas prácticas referentes a formatos de datos y metadatos. Por ejemplo, los datos se publicarán en el formato estructurado XLSX y se garantizará que los metadatos estén, al menos, en el formato Dublin Core y DataCite. Estas prácticas buscan garantizar que los datos estén organizados y formateados de manera que sean fáciles de integrar en otras fuentes de datos.

- **¿Los datos incluirán referencias cualificadas a otros datos?**

  Sí. Como se ha comentado anteriormente, el conjunto de datos empleado en el proyecto reutiliza datos de dos repositorios gubernamentales, a saber: el INE y el portal de datos en abierto del Ayuntamiento de Madrid. Estas fuentes de datos originales se referenciarán de la manera oportuna en esta memoria, en el campo de metadatos de descripción y también se incluirán en el campo de metadatos de contribuidores.

Reutilización de los datos

- **¿Cómo se proporcionará la documentación necesaria para validar el análisis de los datos y facilitar su reutilización (por ejemplo, archivos README con información sobre metodología, libros de códigos, limpieza de datos, análisis, definiciones de variables, unidades de medida, etc.)?**

  Con el objetivo de garantizar la transparencia y reproducibilidad del análisis, adicionalmente al conjunto de datos se subirá al repositorio de Zenodo el código de Python utilizado para leer, limpiar y analizar los datos. La memoria del proyecto también será una fuente valiosa de información sobre los datos y su estructura, en especial porque detalla las variables estudiadas y su significado.

- **¿Estarán los datos disponibles gratuitamente en el dominio público para permitir una reutilización lo más amplia posible? ¿Se concederán licencias para la reutilización de los datos conforme a las obligaciones establecidas en el acuerdo de subvención?**

  Los datos curados serán publicados en abierto bajo la Atribución 4.0 Internacional de Creative Commons (CC BY 4.0) [24], lo que permite a cualquier usuario distribuir, adaptar y utilizar los datos libremente siempre y cuando se conceda el debido crédito al autor original y se indique cualquier modificación realizada sobre los mismos.

- **¿Los datos producidos y/o utilizados en el proyecto son utilizables por terceros, en particular después del final del proyecto?**

Los datos utilizados en el proyecto podrán ser utilizados por terceros, tanto durante el proyecto como una vez finalizado, sin restricción alguna.

- **Describir los procesos seguidos en el proyecto que garanticen la calidad de los datos.**

  Los datos en crudo se obtienen directamente de los repositorios del INE y del portal de datos abiertos del Ayuntamiento de Madrid, por lo que este proyecto no tiene control sobre los criterios de calidad utilizados por estas instituciones para recopilar los datos. Sin embargo, tanto el INE como el Ayuntamiento de Madrid están sujetos al Código de Buenas Prácticas de las Estadísticas Europeas [25], lo que garantiza que las estadísticas que publican sean independientes, fiables y de alta calidad.

  Por otro lado, en la fase de preprocesado y curación de los datos, los administradores de datos del proyecto llevarán a cabo una fase exhaustiva de depuración de los datos para garantizar su consistencia. Esto incluirá la detección y corrección de valores anómalos, la imputación de valores faltantes, la eliminación de valores duplicados, la corrección de errores tipográficos y cualquier otra tarea necesaria para asegurar la calidad de los datos.

3.1.4. Asignación de recursos

- **¿Cuáles son los costes para hacer que los datos cumplan los principios FAIR en el proyecto?**

  Los costes para cumplir con los principios FAIR en el proyecto son principalmente los relacionados con los recursos humanos, ya que los datos reutilizados son gratuitos y se analizarán con herramientas de software libre y se publicarán en un repositorio también gratuito. Los costes incluyen las remuneraciones de los miembros del equipo que se encargarán de limpiar, curar y publicar los datos y metadatos para que sean fácilmente localizables, accesibles, interoperables y reutilizables.

- **¿Cómo serán cubiertos estos gastos?**

  Los costes destinados a garantizar que los datos empleados en el proyecto sean FAIR quedarán cubiertos por el presupuesto del proyecto, las subvenciones y el dinero procedente de fuentes externas. En particular, los gastos relacionados con el acceso abierto a los datos podrán ser subvencionables por la Comisión Europea, en el marco de los proyectos Horizon 2020, si se ajustan a las directrices establecidas en el acuerdo de subvención.

- **¿Quién será responsable de la gestión de datos en tu proyecto?**

  El proyecto *REFOMAD* tiene dos administradores de datos principales, que son Sergio Bolívar Gómez y Juan José Velasco Horcajada. No obstante, las responsabilidades individuales de cada uno de los miembros del equipo en materia de gestión de datos pueden consultarse en la división de paquetes de trabajo de la Sección 2.1.5.

3.1.5. Seguridad de los datos

- **¿Se almacenan los datos de manera segura en repositorios certificados para su conservación a largo plazo y curación?**

Sí. Como se ha mencionado anteriormente, los datos utilizados en el proyecto serán publicados en Zenodo, un repositorio seguro mantenido por el CERN (Conseil Européen pour la Recherche Nucléaire) que garantiza la conservación a largo plazo de los datos y cumple con los principios de la iniciativa OpenAIRE.

- **¿Qué medidas se han adoptado o se tomarán para garantizar la seguridad de los datos (incluyendo la recuperación de datos y el almacenamiento y traslado seguros de datos sensibles)?**

  Además del almacenamiento de los datos en un repositorio seguro, se tomarán otras medidas para garantizar la seguridad de los mismos. Entre ellas, destacan: crear copias de seguridad tanto en la nube como en un disco duro externo de forma frecuente, utilizar un sistema de control de versiones y guardar todo el material generado en el proyecto en un disco duro externo una vez finalizado el mismo. Para una información más detallada, se puede consultar el plan de preservación de datos elaborado para el proyecto (ver Sección 3.5).

3.1.6. Aspectos éticos

- **¿Hay alguna cuestión ética o legal que pueda tener un impacto a la hora de compartir los datos?**

  Teniendo en cuenta que los datos son de dominio público y que no se manejan datos sensibles, no hay ninguna cuestión ética y/o legal que pueda tener un impacto negativo a la hora de compartir los datos.

3.2. Recolección: fuentes de datos

A continuación, se describen las dos fuentes de datos utilizados en este proyecto, indicando en cada caso el procedimiento seguido para la recopilación de los datos y el formato final de los mismos.

3.2.1. Datos del nivel de formación educativa en los distritos de Madrid

Esta fuente de datos contiene información sobre el nivel de formación de los ciudadanos de Madrid en los años 2016, 2017, 2018 y 2019. Los ciudadanos han sido divididos en 21 distritos de acuerdo a su residencia en la ciudad.

Para la recopilación de los datos se han utilizado una encuesta en dos formatos, tanto en papel como en formato online (ver Apéndice B). La encuesta en papel se distribuyó en distintos distritos de la ciudad de Madrid y se recopilaron las respuestas mediante un proceso de digitalización en un archivo de Excel. Por otro lado, también se realizó una encuesta online, con el objetivo de alcanzar una mayor cantidad de participantes y poder obtener una muestra más amplia de la población. Además, esto nos permite automatizar el formato de los datos y exportarlos de forma sencilla en un archivo de Excel.

La encuesta online se distribuyó a través de diversas plataformas (redes sociales, página web del proyecto...) y se utilizó *Google Forms* para recopilar las respuestas. Al igual que en la encuesta en papel se preguntó sobre el nivel de formación de los ciudadanos en los años 2016, 2017, 2018 y

2019, con el objetivo de analizar si ha habido una evolución en el nivel educativo de los ciudadanos de Madrid.

De esta manera, utilizando ambos métodos de recolección de datos, se obtiene una muestra más amplia y representativa de la población, para así poder analizar las posibles diferencias en el nivel educativo entre los distintos distritos de Madrid.

Además de dividir las respuestas de la encuesta en función del distrito de residencia de los ciudadanos encuestados, también se ha incluido la población de cada distrito y el número de colegios en cada uno de ellos

Formato

La información contenida en esta fuente de datos se ha recopilado en varios archivos de Excel, uno para cada año: 2016, 2017, 2018 y 2019. Además, en cada archivo, se han separado los 21 distritos en hojas de cálculo diferentes para facilitar su consulta. Cada hoja consta de 14 filas que representan los indicadores que hemos tomado y podemos ver descritos a continuación descritos:

3.2.2. Datos de la renta neta media por persona

Esta fuente de datos contiene información acerca de la renta neta media por persona para cada distrito de la ciudad de Madrid.

Para la recopilación de los datos se ha llevado a cabo una solicitud [6] al Instituto Nacional de Estadística (INE) ya que cuenta con las competencias necesarias para recopilar, procesar y publicar información estadística sobre la población y los hogares en España. Legalmente, por la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales [26] el equipo de *REFOMAD* no tiene la capacidad de llevar a cabo esta acción y recoger datos sobre la renta neta media de la población madrileña. El INE, como organismo oficial encargado de la recopilación y difusión de estadísticas en España, es la institución más adecuada para obtener datos confiables y actualizados sobre este tema [27]. Además, al ser una entidad pública, el INE garantiza el acceso público y la transparencia de la información que recopila, lo que es esencial para el desarrollo de investigaciones y estudios. Por estas razones, se han solicitado al INE los datos de la renta neta media por persona en cada distrito de la cuidad de Madrid durante los años 2015, 2016, 2017, 2018, 2019 y 2020.

Formato

La información contenida en esta fuente de datos se ha recopilado en un archivo de Excel, donde se presenta el valor de la renta media neta por persona para cada distrito de la ciudad de Madrid desde el año 2015 hasta el 2020. El archivo consta de 3 columnas: la primera columna indica el distrito, la segunda columna indica el año en el que se ha obtenido el valor del indicador y, por último, en la tercera columna el valor de la renta neta media por persona.

3.2.3. Difusión para la adquisición de los datos

Para lograr obtener la cantidad de datos suficiente (consideraremos éxito el que las encuestas sean rellenadas por un 5 % de la población por distrito) como para proceder con el análisis que pretende realizarse, hay una clara necesidad de difundir el proyecto. Es, por tanto, parte esencial

Indicadores	Tipo de dato	Descripción
Superficie (Ha.)	Coma flotante decimal	Superficie del distrito en hectáreas
Densidad (hab./Ha.)	Coma flotante decimal	Cantidad de ciudadanos que viven en el distrito
No sabe leer ni escribir, sin estudios o primaria incompleta	Entero	Número de ciudadanos encuestados que no saben leer ni escribir o no tienen estudios.
Bachiller Elemental, Graduado Escolar, ESO, Formación profesional primer grado	Entero	Número de ciudadanos encuestados con graduado escolar o formación profesional básica.
Formación profesional 2º grado, Bachiller Superior o BUP	Entero	Cantidad de ciudadanos encuestados con el título de bachillerato o una formación profesional media.
Titulados medios, Diplomados, Arquitecto o Ingeniero Técnico	Entero	Cantidad de ciudadanos encuestados con una formación profesional superior.
Estudios superiores, licenciado, Arquitecto o Ingeniero, estudios superiores no universitarios, doctorado, estudios postgraduados	Entero	Cantidad de ciudadanos encuestados con una formación universitaria o superior.
Nivel de estudios / Desconocido y No consta	Entero	Cantidad de ciudadanos encuestados que no conocen su nivel de estudios.
Escuelas Infantiles Municipales	Entero	Número de escuelas infantiles municipales en el distrito.
Escuelas Infantiles Públicas CAM	Entero	Número de escuelas infantiles públicas en el distrito.
Escuelas Infantiles Privadas	Entero	Número de escuelas infantiles privadas.
Colegios Públicos Infantil y Primaria	Entero	Número de colegios públicos de Educación Infantil y Primaria en el distrito.
Institutos Públicos de Educación Secundaria	Entero	Número de institutos públicos de Educación Secundaria en el distrito.
Colegios Privados Inf. o Pri. o Inf. y Pri.	Entero	Número de colegios privados de Educación Infantil, Primaria en el distrito.

Tabla 15: Indicadores recogidos a través de la encuesta.

Indicadores	Tipo de dato	Descripción
Renta neta media por persona (€)	Entero	Ingreso promedio que recibe un ciudadano de un distrito determinado en un año.

Tabla 16: Indicadores recogidos a través del INE.

en el proyecto el planificar de manera adecuada la difusión que ha de llevarse a cabo, ya que es primordial para poder hacer el estudio.

El grupo de personas de mayor interés para el fin de nuestro proyecto son aquellas mayores de 25 años que residen en Madrid, ya que son aquellos que por su edad han podido alcanzar niveles

superiores (por ejemplo, haber terminado una carrera universitaria o similar). Por tanto, serán sujeto de estudio todas aquellas personas que cumplan esos dos requisitos y estén dispuestos a ceder sus datos para la realización del mismo.

Como ya se ha mencionado, habrá dos formatos para las encuestas. Para la distribución en formato físico, se realizará, por un lado, un trabajo de campo mediante la recopilación de datos en la calle, es decir, una persona se encargará de distribuir las encuestas en distintos distritos de la ciudad. Junto con ellas, se distribuirán flyers que se diseñan con el objetivo de poder informar a los encuestados acerca del proyecto que se está llevando a cabo. Se indicará en ellos el link de acceso a la página web de la que se hablará posteriormente. Habrá además visitas a diferentes centros, tanto educativos como culturales y sociales para la realización de ponencias en las cuales tratar la importancia de realizar encuestas para poder hacer estudios que pueden tener un gran impacto en la sociedad, y animar a que rellenen la que se ha diseñado en este proyecto. Los centros en los que se realizaran serán elegidos de entre todos aquellos que estén interesados siguiendo un proceso aleatorio, y también a conveniencia de los intereses del proyecto.

Por otro lado, para el formato de encuestas online, se diseñará una página web, que no implicará la concesión de permisos especiales para su utilización. Esta servirá de soporte para varias tareas de la difusión. En primer lugar, cuenta con el acceso a la encuesta que ha de rellenarse. Aparecerá plasmada además la información correspondiente al proyecto que se lleva a cabo, y a los fines con los que se está recogiendo la información acerca de los niveles de formación entre los ciudadanos madrileños. También, aparecerán pequeñas indicaciones sobre el estado y correcto avance del proyecto. Por último, se habilitará un buzón de sugerencias dentro de la web en el que se podrán realizar las indicaciones que se consideren oportunas y serán utilizadas aquellas que sean consideradas como positivas. Estará destinada a todas las personas que desean visitarla sin la necesidad de ningún requerimiento.

Junto con la página web se crearán perfiles en las principales redes sociales del país como son Facebook, Instagram o Twitter. En ellas, se irá informando del estado del proyecto, y también se habilitarán enlaces para poder realizar las encuestas.

Finalmente, se optará por buscar la difusión en medios externos (es decir, aquellos que son ajenos al proyecto) y gratuitos, para reducir costes. Los medios de comunicación escritos y online serán importantes a la hora de dar a conocer y dar visibilidad al proyecto puesto que sirven para colocar el estudio en un lugar que sea del conocimiento de todos. En particular, se busca aparecer en periódicos y revistas de la propia ciudad de Madrid. En cuanto a la televisión, se buscará contactar con cadenas regionales madrileñas que permitan que se difunda una noticia acerca del proyecto y la importancia de la recogida de datos. De esta manera, se logrará llegar a un publico heterogéneo y abaratar los costes que podría suponer la realización de un anuncio o alguna campaña similar.

Combinando tanto el formato digital como el de papel, se pretende llegar al mayor número de personas posibles, puesto que cuantos más datos se obtengan y más variable sea la población que rellene las encuestas, mejor será el estudio que pueda realizarse.

3.3. Preprocesado: limpieza y curación

Para realizar la limpieza y curación de los datos se ha optado por utilizar el lenguaje Python y hacer uso de las librerías Pandas y Numpy. El proceso de curación para ambos dataset ha sido el siguiente:

El primer paso ha sido cargar los archivos, lo que ha planteado el primer problema. El dataset

recopilado por el INE es un fichero '.csv' fácil de importar, sin embargo, los ficheros correspondientes al Ayuntamiento de Madrid tienen formatos diferentes dependiendo del año: los años anteriores al 2020 se encuentran en formato '.xls', una extensión de Excel 2007, mientras que los posteriores a 2020 tienen una extensión '.xlsx', un formato más actoual. Además, los ficheros del Ayuntamiento de Madrid tienen asignado un hoja de Excel a cada distrito de Madrid, mientras que el fichero del INE contiene toda la información en una única hoja.

Durante la carga de los datos, en espacial con el dataset del Ayuntamiento de Madrid, se presta especial a cargar únicamente las hojas que contienen datos, pues algunos archivos tienen hojas extra donde se puede observar una portada o una introducción acerca de los indicadores utilizados.

Una vez cargados todos los ficheros, se observa que el dataset del INE nombra a cada distrito con un número en lugar de utilizar su nombre como hace el Ayuntamiento de Madrid. Se estudia como se relacionan ambos criterios y se modifica el dataset del INE, estableciendo el nombre real de cada distrito en lugar del valor numérico.

Posteriormente, se unifican los datos del dataset del Ayuntamiento de Madrid. Al encontrarse los datos de cada distrito en diferentes hojas, se busca agruparlos en un dataframe donde cada columna corresponda a un distrito. En este proceso se comprueba los formatos de los datos, si hay valores no establecidos (NaN), datos en formato 'string' en lugar de 'float' o que el paso de la coma decimal (criterio de Excel por defecto) al punto decimal (criterio de Python por defecto) en la carga de los datos no haya supuesto un cambio en las magnitudes de los mismos. Si fuera así, se corrigen de forma automática. Además, en este proceso se seleccionan los indicadores de interés del dataset, obteniendo así un dataframe final cuyas filas corresponden a los indicadores de estudio y cuyas columnas hacen referencia a los distritos de Madrid. También se ha añadido una columna extra con los datos de la ciudad de Madrid, por si en un futuro se quiera hacer referencia o una comparación con la ciudad. Realizamos esta acción para cada año, obteniendo un dataframe para cada archivo cargado.

Con los datos del INE se lleva a cabo un proceso similar, teniendo en cuanta que de este dataset solo se quiere un indicador (renta neta media por persona). Como resultado se obtiene un dataframe cuyas filas corresponden a los distritos de la ciudad de Madrid y las columnas hacen referencia al año en el que se ha recogido el valor de dicho indicador.

Una vez que nos hemos asegurado de que todos los datos se encuentran en un formato correcto, unimos los dataframes de ambas fuentes, generando un dataframe para cada año donde las filas corresponden con todos los indicadores de interés para este trabajo y las columnas con los distritos de la ciudad de Madrid. Estos dataframes se exportan como hojas de un archivo '.xlsx' (cada hoja será un año) que recoge los datos limpios y curados. Este archivo a su vez se comparte a través de Zenodo para cumplir los principios FAIR, teniendo el siguiente DOI:

Figura 3: DOI de Zenodo de los datos limpios y curados.

Por último indicar que los años que se han tenido en consideración en este trabajo han sido del 2016 al 2019, ambos incluido. Esta decisión se ha tomado debido a que el INE nos ha cedido datos del 2015 hasta el 2020, mientras que el Ayuntamiento de Madrid del 2016 al 2022, con la peculiaridad de que los años 2020 y 2021 se encuentran juntos. Por motivos de la pandemia del COVID-19 vivida en 2020, no se llevó a cabo un estudio anual de los indicadores como se estaba

haciendo en años anteriores y se agrupó con el siguiente año, el 2021. Es por ello que hemos escogido utilizar los años de los que tenemos datos en común en ambas fuentes de datos.

3.4. Análisis de los datos

Una vez realizado el preprocesado de los datos, es posible realizar ciertos análisis que permitan obtener un valor añadido. Para ello, primero se parte describiendo la metodología seguida, y después se mostrarán los resultados obtenidos.

3.4.1. Metodología

Con el fin de analizar las relaciones existentes entre los niveles de formación de los ciudadanos madrileños y la renta del distrito en el que residen, se ha procedido a visualizar los datos. Para ello, se ha realizado una gráfica que combina la información acerca de la renta de los diferentes distritos (aparece como un diagrama de barras), con los datos acerca de la formación, divididos por colores según el nivel correspondiente (pueden verse como puntos unidos a través de líneas discontinuas). Este primer análisis se realiza para los datos del año 2016.

Una vez realizado, se compararon los datos de la renta de todos los años que se tenían (de 2016 a 2019) a través de boxplots, de manera que era posible conocer no solo la media de la renta para cada distrito, sino también la variabilidad de esos valores con los años. Gracias a ello, fue posible determinar la importancia de realizar el análisis desglosado según los años ya que, como podrá verse más adelante, existen diferencias en esta variable notables en algunos de los distritos, que podrían afectar a las tendencias en los niveles de formación.

Finalmente, se ha repetido el mismo procedimiento que para el primer análisis, pero utilizando los datos que corresponden a los años 2017, 2018 y 2019, con el fin de comprobar si los resultados obtenidos para el año 2016 siguen manteniéndose en los sucesivos.

Todo este tratamiento de los datos para el análisis ha sido realizado, al igual que ocurría en el apartado previo, a través del lenguaje de programación Python, con las librerías Pandas, Numpy y Plotly (esta última permite la visualización de los datos).

3.4.2. Resultados

Se muestra en la Figura 4 la gráfica correspondiente a los datos de los distintos niveles de formación y a la renta para el año 2016, ordenados según esta última variable mencionada.

Como puede apreciarse, existe una tendencia notable hacia la disminución del nivel de formación de Bachiller Elemental, Graduado Escolar, ESO y Formación profesional de primer grado (línea discontinua azul) a medida que se aumenta la renta. Ocurre lo mismo para el porcentaje de personas que no sabe leer ni escribir, que no tiene estudios o tiene primaria incompleta (línea discontinua roja). Para este último caso, la disminución es más irregular que en el anterior, ya que existen varios picos que se dan para algunos de los distritos, como son el de La Latina, o el de Hortaleza.

Además, observamos que el nivel de estudios superiores, licenciados, arquitectos o ingenieros, estudios superiores no universitarios, doctorados y estudios postgraduados (en negro) aumenta a medida que lo hace la renta. De nuevo, para esta medida existen ciertas irregularidades en ese ascenso. En particular, cabe destacar el pico existente para el distrito Centro. Con las variables que se están manejando en este estudio no es posible establecer las razones por las que se da esta forma. Sin embargo, se cree que puede deberse a la gran oferta cultural, científica y de otras índoles que

hay en este distrito, y puede resultar atractiva para aquellas personas que cuentan con cierto nivel de formación, y que por tanto decidan residir ahí para satisfacer sus intereses.

Respecto al resto de niveles, que son los intermedios y aquel que tiene en cuenta datos que se desconocen, se mantienen relativamente constantes para todos los distritos.

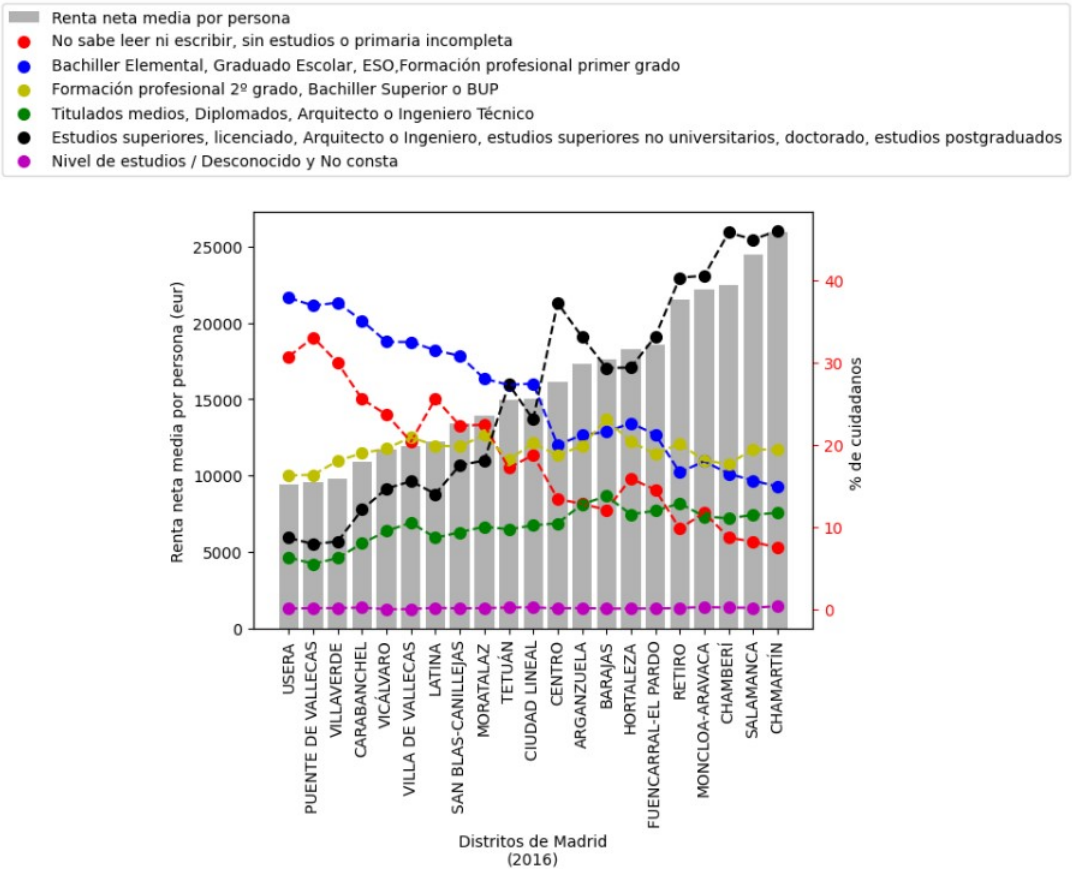


Figura 4: Distribución de los niveles de formación bajo estudio mediante líneas discontinuas, y de la renta de cada distrito de Madrid en un diagrama de barras, para el año 2016.

Por otro lado, en la Figura 5 se representan boxplots de los datos de la renta de los años 2016 a 2019 para cada distrito de Madrid. Están ordenados de la misma forma que en la Figura 4, pese a que para algunos años posteriores hay cambios en el orden de los distritos según la renta.

Puede apreciarse cómo para los distritos con renta más baja no hay mucha variabilidad en los datos, como puede ser el de Usera y el de Puente de Vallecas. En el caso opuesto, para los distritos de mayor renta la varían más los datos. Vemos que, por ejemplo, en el caso de Chamartín y Chamberí, los valores llegan a distar hasta en unos 3000€. Ocurre lo mismo para el distrito Centro.

Además, vemos que algunos de los boxplots, al ser más largos que otros, pueden solapar sus valores respecto al eje vertical. Esto podría llegar a afectar la distribución de los niveles de formación, así que por ese motivo se ha decidido comprobar lo visto con la Figura 4 para los años entre 2017 y 2019.

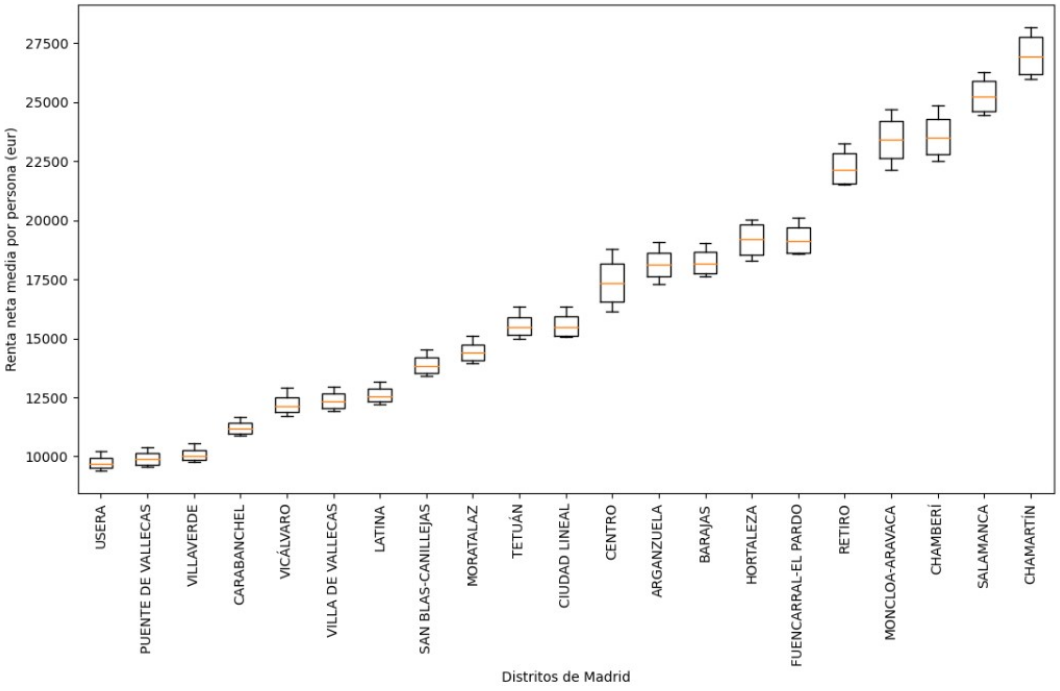


Figura 5: Boxplots de los datos de la renta neta media por persona para los diferentes distritos de Madrid, entre los años 2016 y 2019.

La Figura 6 muestra las gráficas correspondientes a los datos de los años 2017 a 2019 por separado, representando los distintos niveles de formación y la renta por distritos, ordenados ascendentemente según la renta.

En primer lugar, es destacable mencionar que el eje de la derecha ha cambiado su dimensión respecto a la de la correspondiente al año 2016, llegando a alcanzar ahora el 50 % de los ciudadanos.

Si se observa la línea correspondiente al nivel para el cual no se sabe leer ni escribir, o no tiene estudios (en rojo), se aprecia que esta decrece a medida que aumenta la renta. Ocurre de forma similar para los tres años que se tienen. Sin embargo, es notable el hecho de que su valor máximo, dado en el Puente de Vallecas tanto en estos años como en 2016, pasa de ser superior al 30 % en este último año mencionado, a alrededor del 15 % en los demás. Esta disminución en este porcentaje se traduce en un incremento generalizado del porcentaje de ciudadanos que alcanzan el siguiente nivel de formación, el correspondiente a Bachiller Elemental, Graduado Escolar, ESO y Formación profesional de primer grado.

Para el último nivel de formación mencionado, se sigue observando la disminución del porcentaje de ciudadanos que lo han alcanzado, a medida que aumenta la renta neta media por persona.

El caso opuesto lo vemos de nuevo en el nivel de formación más alto, que puede observarse con la línea negra discontinua, que crece con la renta. La excepción más clara de esta tendencia es, como ocurría antes, el distrito Centro. Finalmente, mencionar que los porcentajes de ciudadanos que han alcanzado niveles intermedios de formación permanecen estables a pesar de aumentar la renta.

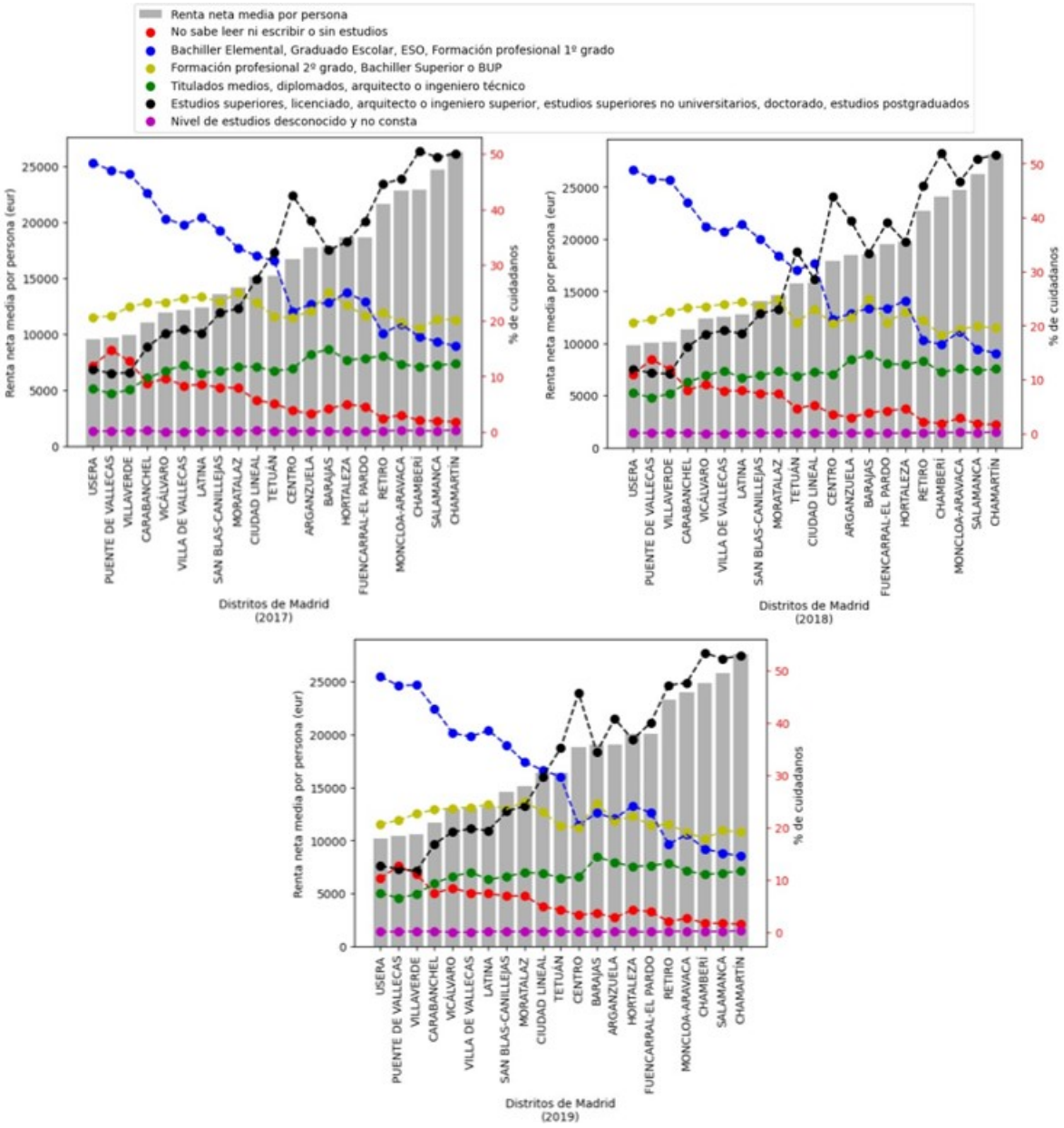


Figura 6: Distribución de los niveles de formación bajo estudio mediante líneas discontinuas, y de la renta de cada distrito de Madrid en un diagrama de barras, para los años 2017, 2018 y 2019.

Finalmente, se ha analizado si el nivel de renta neta media por persona está también relacionado con el número de centros de enseñanza (incluyendo colegios, institutos y universidades) que hay por cada 100.000 habitantes. Esto podría corresponderse a su vez con el nivel de formación de las personas: si estas disponen de una gran oferta de centros educativos en su distrito, podría ser un reclamo para que mejoren su nivel formativo.

La Figura 7 representa los datos correspondientes al número de centros de enseñanza por cada 100000 habitantes, en forma de puntos rojos, y aquellos correspondientes a la renta neta media por persona en un diagrama de barras azul (de igual forma que lo teníamos en las figuras previas), para cada uno de los distritos de Madrid. La información que aparece es la recogida del año 2016.

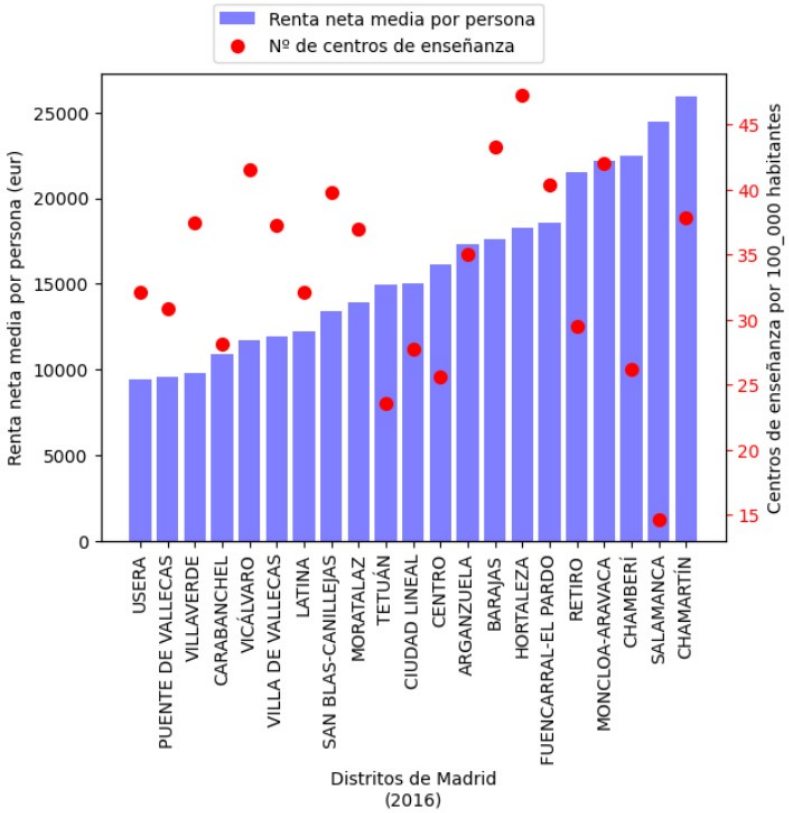


Figura 7: Distribución de los niveles de formación bajo estudio mediante líneas discontinuas, y de la renta de cada distrito de Madrid en un diagrama de barras, para los años 2017, 2018 y 2019.

Es fácil apreciar a simple vista que no existe relación entre ambas variables. El caso más destacado podría ser el del distrito de Salamanca, que cuenta con la segunda renta neta media más alta, y sin embargo cuenta con menos de 15 centros de enseñanza en su área. A través de la Figura 4 vimos que no estaba mal posicionado en cuanto al nivel de formación de los habitantes (más de un 40 % de estos contaban con el nivel más alto). En vista de los resultados, tampoco parece que el número de centros educativos que hay en cada distrito no es una variable que tenga una clara relación con el nivel de formación de los ciudadanos.

3.4.3. Conclusiones

Lo primero que puede observarse a partir de las figuras 4 y 6 es la clara relación directa existente entre el nivel más alto de formación, el correspondiente a Estudios superiores, licenciados, arquitectos o ingenieros superiores, estudios superiores no universitarios, doctorados y estudios postgraduados, y la renta media neta por persona para cada uno de los distritos madrileños. Esto está en consonancia con lo que cabría pensar, ya que un mayor nivel adquisitivo podría indicar el tener acceso a un mayor catálogo de opciones, sobre todo en niveles altos de formación.

La relación inversa se observa en el caso de los niveles más bajos de formación, cuando no se sabe leer ni escribir, o no se tienen estudios, y cuando se tiene Bachiller Elemental, Graduado Escolar, ESO y Formación profesional de primer grado. De nuevo, la explicación a este hecho puede relacionarse con lo argumentado en el punto anterior.

Respecto a estos dos niveles mencionados, cabe destacar además la diferencia entre los porcentajes observados el primer año analizado (2016) comparado con los años posteriores. Se observa cómo el nivel más bajo disminuye su valor en, aproximadamente, un 15 % para los distritos con renta más baja (Usera, Puente de Vallecas,...), mientras que el del nivel de Bachiller Elemental aumenta alrededor de un 12 %. Sin embargo, los valores de la renta, como pueden verse en la Figura 5, no varían en gran medida para esos distritos. Por lo tanto, para encontrar la razón por la que ese elevado porcentaje de ciudadanos haya ascendido de nivel formativo, habría que analizar otras variables además de las que se han tenido en cuenta en este estudio. Esta cuestión quedaría, por tanto, fuera del alcance de este proyecto.

Finalmente, se ha visto que los niveles de formación intermedios, que corresponden con Formación profesional de segundo grado, Bachiller Superior o BUP, y Titulados medios, diplomados, arquitectos o ingenieros superiores, no tienen una relación tan evidente con la renta neta media, ya que se observa una estabilidad en el porcentaje de ciudadanos que los han alcanzado a pesar de la variación en la renta.

3.5. Plan de preservación

En este apartado se detalla el plan de preservación de datos del proyecto *REFOMAD*, que incluye las medidas adoptadas para garantizar la integridad, seguridad y accesibilidad de los datos a largo plazo. La elaboración de este plan es esencial para la confiabilidad y reproducibilidad del análisis de datos efectuado en el proyecto y, además, es una herramienta crucial para el cumplimiento de los principios FAIR. El plan puede dividirse en las siguientes acciones de preservación:

Cumplimiento de los principios FAIR

La accesibilidad de los datos se asegurará mediante su publicación en el repositorio abierto Zenodo, que cumple con las normas de conservación y acceso a los recursos digitales establecidas por OpenAIRE. En concreto, el recurso será identificado con un DOI (Digital Object Identifier), a través del cual se podrá acceder a él de manera gratuita y sin necesidad de autenticación. Por otro lado, se garantizará que el conjunto de datos sea fácilmente localizable documentándolo adecuadamente. En particular, se elaborarán metadatos, tanto en formato DublinCore como DataCite, que se publicarán con licencia Creative Commons Zero (CC0) [23] y que incluirán, como mínimo: título, creadores, palabras clave, una descripción de los datos, la procedencia, el formato y la localización de los mismos. Esta información será fácilmente visible en la dirección URL de Zenodo con el recurso. Los datos estarán disponibles en el repositorio de manera permanente, garantizando la disponibilidad de los metadatos en caso de que los datos dejen de estar disponibles (por motivos ajenos al proyecto) o el formato empleado quedase en desuso (por ejemplo, por el paso del tiempo).

En cuanto a las medidas para fomentar la interoperabilidad, el conjunto de datos curado empleado en el análisis será publicado con licencia Atribución 4.0 Internacional de Creative Commons (CC BY 4.0) [24] y en un formato estructurado universal y gratuito (en particular, en XLSX). Así, se garantizará que cualquier usuario pueda acceder y leer los datos sin necesidad de instalar software específico. Además, junto con los metadatos y el conjunto de datos curados, se incluirán los datos originales y el script utilizado para la lectura, procesamiento, curación y análisis de los datos. Se espera que esto, conjuntamente con esta memoria, facilite la reutilización de los datos, ya que el usuario dispondrá de absolutamente toda la información empleada para la ejecución del proyecto.

Almacenamiento de los datos y copias de seguridad

En el transcurso del proyecto, todos los datos generados (incluyendo el conjunto de datos para el análisis y los datos originales) serán almacenados en un servicio de almacenamiento en la nube seguro y confiable contratado para el proyecto. Además, una vez finalizado el proyecto se hará una copia de los datos y se guardará en un disco duro externo para asegurar su conservación a largo plazo. Por otro lado, dado que el volumen de datos generados en el proyecto no es muy grande, los administradores de los datos también guardarán una copia en sus nubes personales. Esto garantizará que, en caso de que Zenodo pierda la información, el proyecto pueda ser recuperado y publicado de nuevo.

Asimismo, durante el proyecto se realizarán copias de seguridad regulares para garantizar que se pueda recuperar la mayor cantidad de información posible en caso de fallo del sistema. Estas copias de seguridad se guardarán tanto en la nube como en un disco duro externo. El calendario de copias de seguridad es el siguiente: el primer domingo de cada mes se hará una copia de seguridad completa (full backup), semanalmente se hará una copia diferencial (que guarda los cambios desde la última full backup) y, finalmente, se hará una copia de seguridad incremental diaria. Los miembros del equipo se encargarán de que el sistema de copias de seguridad funcione correctamente y se realizarán pruebas para asegurar que la información puede ser recuperada en caso de fallo.

Sistema de Control de versiones

El equipo del proyecto *REFOMAD* utilizará un sistema de control de versiones, Git en particular, que gestionará a través de la plataforma GitHub. El repositorio público del proyecto puede encontrarse en https://github.com/bolivars/proyecto-refomad. Esto permitirá a los miembros del equipo hacer un seguimiento de los cambios en los datos y la documentación relativa al proyecto a lo largo del tiempo, y volver fácilmente a versiones anteriores si fuese necesario. Esta práctica también fomentará la accesibilidad de los datos, ya que otros usuarios podrán acceder fácilmente a ellos desde el repositorio de GitHub del proyecto.

4. Conclusiones

A lo largo de esta memoria se ha expuesto la forma de proceder ante un proyecto que tiene interés tanto social como político. El objetivo del mismo era el de establecer si existía algún tipo de relación entre el nivel de formación de los madrileños y la renta neta media, para cada uno de los distintos distritos de Madrid.

En primer lugar, se ha visto en el Capítulo 2 la planificación que se ha diseñado para llevar a cabo el proyecto *REFOMAD* y lograr alcanzar los objetivos que se esperaban. Como puede apreciarse en la Tabla 2, se cuenta con 7 paquetes de trabajo, siendo el WP1 el correspondiente a la coordinación interna y gestión económica. Este realiza sus tareas a lo largo de toda la duración del proyecto, que es de 3 meses y 24 días, comenzando el 1 de febrero de 2023 y terminando el 24 de mayo de ese mismo año. Cinco de los restantes coordinan cada una de las etapas del ciclo de vida de los datos: planificación, adquisición, procesado y análisis de los datos, y preservación (junto con presentación del proyecto). Hay finalmente un paquete de trabajo encargado de la difusión, que para este estudio, donde de manera ficticia se ha considerado que parte de los datos utilizados se toman a través de encuestas sociales, tiene gran importancia. A través de la Figura 9, y de los apartados 2.1.4 y 2.1.5 pueden observarse las diferentes tareas y entregables asignados a cada uno de los paquetes ya mencionados.

Además de los aspectos logísticos, el segundo capítulo expone también en la Subsección 2.1.6 los diferentes riesgos que pueden darse, junto con los planes de contingencia correspondientes. Entre ellos se encuentran problemas con la toma (ficticia) de los datos y con el software de programación, que podrían hacer peligrar la consecución de los objetivos en caso de no lograr subsanarlos. Finalmente, se muestra también el estudio del presupuesto del proyecto, que se ha dividido en dos secciones diferentes: recursos materiales y humanos. Para el primero de ellos se estima una cantidad de 8.362 €, mientras que el segundo triplica este valor, siendo de 28.386 €. Se añade además un 15 % a la suma de ambas cantidades, que queda reservado para contingencias, de forma que finalmente el presupuesto total es de 42.260 €.

En cuanto al Capítulo 3, este está dedicado a la descripción de las diferentes etapas que conforman el ciclo de vida de los datos en el proyecto *REFOMAD*. Primeramente, se cubre en la Sección 3.1 el Plan de Gestión de Datos, siguiendo para ello las guías sobre la gestión de datos de proyectos del programa Horizon 2020, de manera que con él pueda asegurarse el cumplimiento de los principios FAIR. Posteriormente, se expone la recolección de los datos, detallando el modo a partir del cual se han obtenido los datos (la forma ficticia en la que fueron recogidos), que son una encuesta tanto en formato en papel como online con la cual se obtuvo la información acerca del nivel formativo, y una solicitud al INE para los datos de las rentas.

Las dos siguientes secciones incluyen los pasos seguidos en el procesado de los datos, con el fin de adecuar el formato de los mismos para facilitar su análisis posterior, y el propio análisis. Con este último, se ha podido establecer la relación directa entre el nivel de formación más alto y la renta, y la relación inversa entre los dos niveles más bajos y la renta (ver Figuras 4 y 6), dentro de los distritos madrileños. Se ha visto además que no existe relación aparente entre el número de centros de enseñanza en cada distrito y las dos variables previamente mencionadas. Finalmente, se ha detallado en la Sección 3.5 el plan de preservación de datos, que incluye las medidas adoptadas para asegurar el cumplimiento de los principios FAIR, además de la descripción acerca del almacenamiento y copias de seguridad de los datos, y del sistema de Control de versiones.

Una vez finalizado el proyecto, son las entidades interesadas en el mismo las que han de tomar

medidas, a partir de la información que este aporta, con el fin de garantizar que los ciudadanos alcancen el nivel de formación educativa que deseen, independientemente del distrito en el que residan.

Referencias

[1] MediaMarkt. *MediaMarkt - Ordenadores LENOVO*. Última consulta el 14 de enero de 2023. URL: https://www.mediamarkt.es/es/product/_portatil-lenovo-ideapad-3-15itl6-156-full-hd-intelr-coretm-i5-1155g7-8gb-ram-512gb-ssd-irisr-xe-graphics-windows-11-home-1543514.html.

[2] MediaMarkt. *MediaMarkt - Disco Duro Externo SSD*. Última consulta el 14 de enero de 2023. URL: https://www.mediamarkt.es/es/product/_disco-duro-externo-1-tb-wd-elements-se-ssd-portatil-lectura-400-mbs-usb-30-para-windows-y-mac-negro-1520380.html.

[3] Lyreco. *Lyreco - Tienda Online*. Última consulta el 14 de enero de 2023. URL: https://www.lyreco.com/webshop/SPSP/index.html?lc=SPSP.

[4] Google Cloud. *Google Cloud - Cloud Storage Pricing*. Última consulta el 14 de enero de 2023. URL: https://cloud.google.com/storage/pricing#cloud-storage-pricing.

[5] Google Cloud. *WIX - Información de planes*. Última consulta el 14 de enero de 2023. URL: https://es.wix.com/upgrade/website.

[6] Boletín Oficial del Estado. *BOE - Peticiones de datos a medida al INE. Precios*. Última consulta el 14 de enero de 2023. URL: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2021-14822.

[7] Raosoft. *Survey Sample Size Calculator*. Última consulta el 14 de enero de 2023. URL: http://www.raosoft.com/samplesize.html.

[8] Indeed. *Sueldos en España*. Última consulta el 3 de abril de 2022. URL: https://es.indeed.com/career/salaries.

[9] Comunidad de Madrid. *Ayudas para programas interés social*. Última consulta el 14 de enero de 2023. URL: https://sede.comunidad.madrid/ayudas-becas-subvenciones/ayudas-programas-interes-social.

[10] Ayuntamiento de Madrid. *Subvenciones para proyectos vinculados a la colaboración con entidades del tercer sector 2022*. Última consulta el 14 de enero de 2023. URL: https://sede.madrid.es/portal/site/tramites/menuitem.62876cb64654a55e2dbd7003a8a409a0/?vgnextoid=52729698ea4a1810VgnVCM2000001f4a900aRCRD&vgnextchannel=aec8a38813180210VgnVCM100000c90da8c0RCRD&vgnextfmt=default.

[11] Agencia Estatal de Investigación. *Fondos Europeos para la Investigación*. Última consulta el 14 de enero de 2023. URL: https://www.aei.gob.es/fondos-europeos/fondos-europeos-investigacion.

[12] Comisión Europea. *Horizon 2020*. Última consulta el 14 de enero de 2023. URL: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

[13] Comisión Europea. *Horizon Europe - Data Management Plan Template (5 May 2021)*. Última consulta el 16 de enero de 2023. URL: https://web.cimne.upc.edu/groups/proposals/Newsletter/data-management-plan-template_he_en.pdf.

[14] Instituo Nacional de Estadística. *Atlas de distribución de renta de los hogares*. Última consulta el 8 de enero de 2023. URL: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177088&menu=ultiDatos&idp=1254735976608.

[15] Ayuntamiento de Madrid. *Panel de indicadores de distritos y barrios de Madrid. Estudio sociodemográfico*. Última consulta el 8 de enero de 2023. URL: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=71359583a773a510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default.

[16] DOI. *The DOI System*. Última consulta el 17 de enero de 2023. URL: https://www.doi.org/.

[17] DublinCore. *DCMI: Home*. Última consulta el 17 de enero de 2023. URL: https://www.dublincore.org/.

[18] DataCite. *Welcome to DataCite*. Última consulta el 17 de enero de 2023. URL: https://datacite.org/.

[19] Zenodo. *Página de Inicio*. Última consulta el 17 de enero de 2023. URL: https://zenodo.org/.

[20] Open Archives Initiative. *Protocol for Metadata Harvesting*. Última consulta el 17 de enero de 2023. URL: https://www.openarchives.org/pmh/.

[21] OpenAIRE. *OpenAIRE Guidelines for Data Archives*. Última consulta el 17 de enero de 2023. URL: https://guidelines.openaire.eu/en/latest/data/index.html.

[22] Zenodo. *Zenodo REST API*. Última consulta el 17 de enero de 2023. URL: https://developers.zenodo.org/.

[23] Creative Commons. *CC0 1.0 Universal (CC0 1.0) - Ofrecimiento al Dominio Público*. Última consulta el 17 de enero de 2023. URL: https://creativecommons.org/publicdomain/zero/1.0/deed.es_ES.

[24] Creative Commons. *Attribution 4.0 International (CC BY 4.0)*. Última consulta el 17 de enero de 2023. URL: https://creativecommons.org/licenses/by/4.0/.

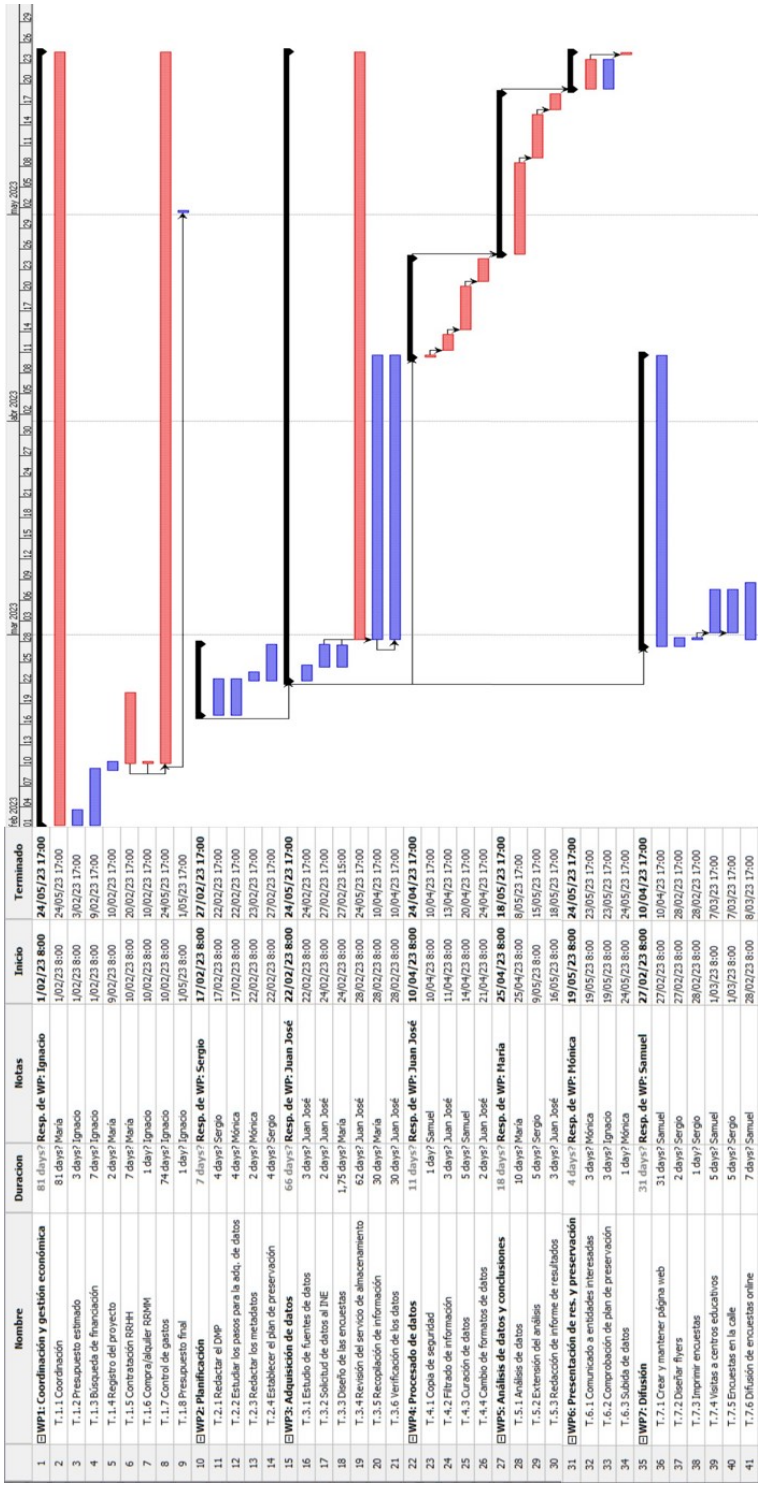
[25] EUROSTAT. *Código de Buenas Prácticas de las Estadísticas Europeas*. Última consulta el 17 de enero de 2023. URL: https://ec.europa.eu/eurostat/documents/4031688/9394048/KS-02-18-142-ES-N.pdf/e792b761-6f09-42a9-a1e0-3a3356a0de1c.

[26] Boletín Oficial del Estado. *Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales*. Última consulta el 22 de enero de 2023. URL: https://www.boe.es/eli/es/lo/2018/12/05/3/con.

[27] Boletín Oficial del Estado. *Ley de la Función Pública Estadística*. Última consulta el 22 de enero de 2023. URL: https://www.boe.es/eli/es/l/1989/05/09/12.

Appendices

A. Diagrama de Gantt



	Nombre	Duracion	Notas	Inicio	Terminado
1	**WP1: Coordinación y gestión económica**	81 days?	**Resp. de WP: Ignacio**	**1/02/23 8:00**	**24/05/23 17:00**
2	T.1.1 Coordinación	81 days?	María	1/02/23 8:00	24/05/23 17:00
3	T.1.2 Presupuesto estimado	3 days?	Ignacio	1/02/23 8:00	3/02/23 17:00
4	T.1.3 Búsqueda de financiación	7 days?	Ignacio	1/02/23 8:00	9/02/23 17:00
5	T.1.4 Registro del proyecto	2 days?	María	9/02/23 8:00	10/02/23 17:00
6	T.1.5 Contratación RRHH	7 days?	María	10/02/23 8:00	20/02/23 17:00
7	T.1.6 Comprar/alquilar RRHH	1 day?	Ignacio	10/02/23 8:00	10/02/23 17:00
8	T.1.7 Control de gastos	74 days?	Ignacio	10/02/23 8:00	24/05/23 17:00
9	T.1.8 Presupuesto final	1 day?	Ignacio	1/05/23 8:00	1/05/23 17:00
10	**WP2: Planificación**	7 days?	**Resp. de WP: Sergio**	**17/02/23 8:00**	**27/02/23 17:00**
11	T.2.1 Redactar el DMP	4 days?	Sergio	17/02/23 8:00	22/02/23 17:00
12	T.2.2 Estudiar los pasos para la adq. de datos	4 days?	Mónica	17/02/23 8:00	22/02/23 17:00
13	T.2.3 Redactar los metadatos	2 days?	Mónica	22/02/23 8:00	23/02/23 17:00
14	T.2.4 Establecer el plan de preservación	4 days?	Sergio	22/02/23 8:00	27/02/23 17:00
15	**WP3: Adquisición de datos**	66 days?	**Resp. de WP: Juan José**	**22/02/23 8:00**	**24/05/23 17:00**
16	T.3.1 Estudio de fuentes de datos	3 days?	Juan José	22/02/23 8:00	24/02/23 17:00
17	T.3.2 Solicitud de datos al INE	2 days?	Juan José	24/02/23 8:00	27/02/23 17:00
18	T.3.3 Diseño de las encuestas	1,75 days?	María	24/02/23 8:00	27/02/23 15:00
19	T.3.4 Revisión del servicio de almacenamiento	62 days?	Juan José	28/02/23 8:00	24/05/23 17:00
20	T.3.5 Recopilación de información	30 days?	María	28/02/23 8:00	10/04/23 17:00
21	T.3.6 Verificación de los datos	30 days?	Juan José	28/02/23 8:00	10/04/23 17:00
22	**WP4: Procesado de datos**	11 days?	**Resp. de WP: Juan José**	**10/04/23 8:00**	**24/04/23 17:00**
23	T.4.1 Copia de seguridad	1 day?	Samuel	10/04/23 8:00	10/04/23 17:00
24	T.4.2 Filtrado de información	3 days?	Juan José	11/04/23 8:00	13/04/23 17:00
25	T.4.3 Curación de datos	5 days?	Samuel	14/04/23 8:00	20/04/23 17:00
26	T.4.4 Cambio de formatos de datos	2 days?	Juan José	21/04/23 8:00	24/04/23 17:00
27	**WP5: Análisis de datos y conclusiones**	18 days?	**Resp. de WP: María**	**25/04/23 8:00**	**18/05/23 17:00**
28	T.5.1 Análisis de datos	10 days?	María	25/04/23 8:00	8/05/23 17:00
29	T.5.2 Extensión del análisis	5 days?	Sergio	9/05/23 8:00	15/05/23 17:00
30	T.5.3 Redacción de informe de resultados	3 days?	Juan José	16/05/23 8:00	18/05/23 17:00
31	**WP6: Presentación de res. y preservación**	4 days?	**Resp. de WP: Mónica**	**19/05/23 8:00**	**24/05/23 17:00**
32	T.6.1 Comunicado a entidades interesadas	3 days?	Mónica	19/05/23 8:00	23/05/23 17:00
33	T.6.2 Comprobación de plan de preservación	3 days?	Ignacio	19/05/23 8:00	23/05/23 17:00
34	T.6.3 Subida de datos	1 day?	Mónica	24/05/23 8:00	24/05/23 17:00
35	**WP7: Difusión**	31 days?	**Resp. de WP: Samuel**	**27/02/23 8:00**	**10/04/23 17:00**
36	T.7.1 Crear y mantener página web	31 days?	Samuel	27/02/23 8:00	10/04/23 17:00
37	T.7.2 Diseñar flyers	2 days?	Sergio	27/02/23 8:00	28/02/23 17:00
38	T.7.3 Imprimir encuestas	1 day?	Sergio	28/02/23 8:00	28/02/23 17:00
39	T.7.4 Visitas a centros educativos	5 days?	Sergio	1/03/23 8:00	7/03/23 17:00
40	T.7.5 Encuestas en la calle	5 days?	Sergio	1/03/23 8:00	7/03/23 17:00
41	T.7.6 Difusión de encuestas online	7 days?	Samuel	28/02/23 8:00	8/03/23 17:00

Figura 8: Diagrama de Gantt del proyecto *REFOMAD* por paquetes de trabajo y tareas correspondientes.

B. Encuesta

ENCUESTA DE NIVEL DE FORMACIÓN POR DISTRITO DE MADRID

Buenos días/tardes

El equipo del proyecto *REFOMAD* está realizando un estudio de acerca del nivel de formación de la población y la relación existente con la renta en los distintos distritos de Madrid. Por ello, solicitamos su colaboración y le agradecemos de antemano su atención. La selección de las personas a las que se solicita la colaboración en el estudio es estrictamente aleatoria, y podrá haberle llegado por vía online o en papel. Toda la información que nos facilite está sujeta a las especificaciones del Reglamento General de Protección de Datos ("RGPD") de 25 de mayo de 2018. Los datos que le solicitamos se tratarán de forma totalmente ANÓNIMA, sin grabar sus datos personales.

El tiempo estimado de duración de la encuesta es, aproximadamente, 5 minutos.

GRACIAS ANTICIPADAS POR SU COLABORACIÓN

0. ¿Podría decirme si tiene 25 años cumplidos y vive en la ciudad de Madrid desde hace más de seis meses?

- ○ Sí
- ○ No

1. **¿En qué distrito reside actualmente? (Escribir en mayúsculas)**

_____

2. ¿Cuáles son los estudios de más alto nivel oficial que usted finalizó en 2016?

- ○ No sabe leer ni escribir, sin estudios o primaria incompleta
- ○ Bachiller Elemental, Graduado Escolar, ESO, Formación profesional primer grado
- ○ Formación profesional 2º grado, Bachiller Superior o BUP
- ○ Titulados medios, Diplomados, Arquitecto o Ingeniero Técnico
- ○ Estudios superiores, licenciado, Arquitecto o Ingeniero, estudios superiores no universitarios, doctorado, estudios postgraduados
- ○ No contesta

3. Si ha habido algún cambio en ese estado en los años posteriores, ¿qué nivel ha sido alcanzado?

- ○ No sabe leer ni escribir, sin estudios o primaria incompleta
- ○ Bachiller Elemental, Graduado Escolar, ESO, Formación profesional primer grado
- ○ Formación profesional 2º grado, Bachiller Superior o BUP
- ○ Titulados medios, Diplomados, Arquitecto o Ingeniero Técnico
- ○ Estudios superiores, licenciado, Arquitecto o Ingeniero, estudios superiores no universitarios, doctorado, estudios postgraduados
- ○ No contesta

4. ¿En qué año finalizó esa última etapa?

- ○ 2017
- ○ 2018
- ○ 2019

GRACIAS POR SU COLABORACIÓN

Figura 9: Encuesta distribuida a los ciudadanos de Madrid.

C. XML con metadatos

C.1. Metadatos en formato Dublin Core

```
<?xml version='1.0' encoding='utf-8'?>
<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
 ↪  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
 ↪  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 ↪  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
 ↪  http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:contributor>Sergio Bolívar Gómez</dc:contributor>
  <dc:contributor>Juan José Velasco Horcajada</dc:contributor>
  <dc:contributor>María Peña Fernández</dc:contributor>
  <dc:contributor>Samuel Laso Saro</dc:contributor>
  <dc:contributor>Ignacio de la Torre Cubillo</dc:contributor>
  <dc:contributor>Mónica Alcantar Martínez</dc:contributor>
  <dc:contributor>Datos Abiertos del INE</dc:contributor>
  <dc:contributor>Portal de Datos Abiertos del Ayuntamiento de
   ↪  Madrid</dc:contributor>
  <dc:coverage>name=Centro; east=-3.7038; north=40.4168</dc:coverage>
  <dc:coverage>name=Arganzuela; east=-3.6922; north=40.3964</dc:coverage>
  <dc:coverage>name=Retiro; east=-3.7167; north=40.3719</dc:coverage>
  <dc:coverage>name=Salamanca; east=-3.6167; north=40.3818</dc:coverage>
  <dc:coverage>name=Chamartín; east=-3.6743; north=40.4096</dc:coverage>
  <dc:coverage>name=Tetuán; east=-3.6856; north=40.4313</dc:coverage>
  <dc:coverage>name=Chamberí; east=-3.6944; north=40.4531</dc:coverage>
  <dc:coverage>name=Ciudad Lineal; east=-3.7167; north=40.4168</dc:coverage>
  <dc:coverage>name=Fuencarral-El Pardo; east=-3.7236; north=40.4148</dc:coverage>
  <dc:coverage>name=Moncloa-Aravaca; east=-3.6471; north=40.5497</dc:coverage>
  <dc:coverage>name=Latina; east=-3.7245; north=40.4386</dc:coverage>
  <dc:coverage>name=Carabanchel; east=-3.6169; north=40.3969</dc:coverage>
  <dc:coverage>name=Usera; east=-3.7813; north=40.3573</dc:coverage>
  <dc:coverage>name=Villaverde; east=-3.7155; north=40.3538</dc:coverage>
  <dc:coverage>name=Villa de Vallecas; east=-3.7339; north=40.3095</dc:coverage>
  <dc:coverage>name=Vicálvaro; east=-3.8667; north=40.3322</dc:coverage>
  <dc:coverage>name=San Blas-Canillejas; east=-3.5667; north=40.3831</dc:coverage>
  <dc:coverage>name=Barajas; east=-3.5667; north=40.4667</dc:coverage>
  <dc:coverage>name=Hortaleza; east=-3.8567; north=40.3742</dc:coverage>
  <dc:coverage>name=Moratalaz; east=-3.7567; north=40.4313</dc:coverage>
  <dc:coverage>name=Puente de Vallecas; east=-3.6285; north=40.4457</dc:coverage>
  <dc:creator>Sergio Bolívar Gómez</dc:creator>
  <dc:creator>Juan José Velasco Horcajada</dc:creator>
  <dc:creator>María Peña Fernández</dc:creator>
  <dc:creator>Samuel Laso Saro</dc:creator>
  <dc:creator>Ignacio de la Torre Cubillo</dc:creator>
  <dc:creator>Mónica Alcantar Martínez</dc:creator>
  <dc:date>2023-01-17</dc:date>
```

```
  <dc:description>El conjunto de datos "refomad-data.xlsx" ha sido utilizado en el
  ↪  proyecto REFOMAD. Incluye información sobre la renta media por persona, el
  ↪  nivel de formación educativa de los residentes mayores de 25 años y el
  ↪  número de centros educativos en cada uno de los 21 distritos del
  ↪  Ayuntamiento de Madrid durante los años 2016-2019. El nivel de formación
  ↪  tiene 6 niveles: no sabe leer ni escribir, bachiller elemental, formación
  ↪  profesional, titulados medios, estudios universitarios y nivel de estudios
  ↪  no consta. Los centros educativos se clasifican también en 6 niveles:
  ↪  escuelas infantiles municipales, escuelas infantiles públicas de la
  ↪  Comunidad Autónoma de Madrid, escuelas infantiles privadas, colegios
  ↪  públicos infantil/primaria, institutos públicos de educación secundaria y
  ↪  colegios privados infantil/primaria. Los datos de renta fueron
  ↪  proporcionados por el INE, los datos sobre la formación fueron obtenidos
  ↪  mediante encuestas propias realizadas por el equipo REFOMAD y los datos de
  ↪  los centros educativos fueron solicitados al portal de datos abiertos del
  ↪  Ayuntamiento de Madrid. Los datos están en formato XLSX y tienen un tamaño
  ↪  de 17 kB. Los datos fueron preprocesados y curados utilizando herramientas
  ↪  de software libre como NumPy y pandas en Python versión 3.10.6.

Se incluyen adicionalmente los datos originales ("DATOS ORIGINALES.zip") y el
 ↪  script de Python empleado para la lectura, preprocesado, curación y análisis
 ↪  de los datos ("REFOMAD-ETL-CURATION.ipynb").</dc:description>
  <dc:identifier>https://zenodo.org/record/7542519</dc:identifier>
  <dc:identifier>10.5281/zenodo.7542519</dc:identifier>
  <dc:identifier>oai:zenodo.org:7542519</dc:identifier>
  <dc:language>spa</dc:language>
  <dc:relation>doi:10.5281/zenodo.7542518</dc:relation>
  <dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
  <dc:rights>https://creativecommons.org/licenses/by/4.0/legalcode</dc:rights>
  <dc:subject>renta</dc:subject>
  <dc:subject>formación</dc:subject>
  <dc:subject>educación</dc:subject>
  <dc:subject>Madrid</dc:subject>
  <dc:subject>distrito</dc:subject>
  <dc:subject>datos</dc:subject>
  <dc:subject>análisis de datos</dc:subject>
  <dc:subject>política</dc:subject>
  <dc:subject>nivel de estudios</dc:subject>
  <dc:subject>estudios universitarios</dc:subject>
  <dc:subject>bachillerato</dc:subject>
  <dc:subject>formación profesional</dc:subject>
  <dc:subject>ESO</dc:subject>
  <dc:subject>titulados medios</dc:subject>
  <dc:subject>escuelas infantiles</dc:subject>
  <dc:subject>colegios</dc:subject>
  <dc:subject>institutos</dc:subject>
  <dc:subject>colegios privados</dc:subject>
```

```
  <dc:title>Renta media por persona y nivel de formación en los 21 distritos del
  ↪  Ayuntamiento de Madrid (2016-2019)</dc:title>
  <dc:type>info:eu-repo/semantics/other</dc:type>
  <dc:type>dataset</dc:type>
</oai_dc:dc>
```

C.2. Metadatos en formato DataCite

```
<?xml version='1.0' encoding='utf-8'?>
<resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
↪  xmlns="http://datacite.org/schema/kernel-4"
↪  xsi:schemaLocation="http://datacite.org/schema/kernel-4
↪  http://schema.datacite.org/meta/kernel-4.1/metadata.xsd">
  <identifier identifierType="DOI">10.5281/zenodo.7542519</identifier>
  <creators>
    <creator>
      <creatorName>Sergio Bolívar Gómez</creatorName>
      <affiliation>Universidad de Cantabria</affiliation>
    </creator>
    <creator>
      <creatorName>Juan José Velasco Horcajada</creatorName>
      <affiliation>Universidad de Cantabria</affiliation>
    </creator>
    <creator>
      <creatorName>María Peña Fernández</creatorName>
      <affiliation>Universidad de Cantabria</affiliation>
    </creator>
    <creator>
      <creatorName>Samuel Laso Saro</creatorName>
      <affiliation>Universidad de Cantabria</affiliation>
    </creator>
    <creator>
      <creatorName>Ignacio de la Torre Cubillo</creatorName>
      <affiliation>Universidad Internacional Menénedez Pelayo</affiliation>
    </creator>
    <creator>
      <creatorName>Mónica Alcantar Martínez</creatorName>
      <affiliation>Universidad de Cantabria</affiliation>
    </creator>
  </creators>
  <titles>
    <title>Renta media por persona y nivel de formación en los 21 distritos del
    ↪  Ayuntamiento de Madrid (2016-2019)</title>
  </titles>
  <publisher>Zenodo</publisher>
  <publicationYear>2023</publicationYear>
  <subjects>
    <subject>renta</subject>
```

```
    <subject>formación</subject>
    <subject>educación</subject>
    <subject>Madrid</subject>
    <subject>distrito</subject>
    <subject>datos</subject>
    <subject>análisis de datos</subject>
    <subject>política</subject>
    <subject>nivel de estudios</subject>
    <subject>estudios universitarios</subject>
    <subject>bachillerato</subject>
    <subject>formación profesional</subject>
    <subject>ESO</subject>
    <subject>titulados medios</subject>
    <subject>escuelas infantiles</subject>
    <subject>colegios</subject>
    <subject>institutos</subject>
    <subject>colegios privados</subject>
  </subjects>
  <contributors>
    <contributor contributorType="DataManager">
      <contributorName>Sergio Bolívar Gómez</contributorName>
      <affiliation>Proyecto REFOMAD</affiliation>
    </contributor>
    <contributor contributorType="DataCurator">
      <contributorName>Juan José Velasco Horcajada</contributorName>
      <affiliation>Proyecto REFOMAD</affiliation>
    </contributor>
    <contributor contributorType="ProjectManager">
      <contributorName>María Peña Fernández</contributorName>
      <affiliation>Proyecto REFOMAD</affiliation>
    </contributor>
    <contributor contributorType="ProjectMember">
      <contributorName>Samuel Laso Saro</contributorName>
      <affiliation>Proyecto REFOMAD</affiliation>
    </contributor>
    <contributor contributorType="ProjectMember">
      <contributorName>Ignacio de la Torre Cubillo</contributorName>
      <affiliation>Proyecto REFOMAD</affiliation>
    </contributor>
    <contributor contributorType="ProjectMember">
      <contributorName>Mónica Alcantar Martínez</contributorName>
      <affiliation>Proyecto REFOMAD</affiliation>
    </contributor>
    <contributor contributorType="DataCollector">
      <contributorName>Datos Abiertos del INE</contributorName>
      <affiliation>INE</affiliation>
    </contributor>
    <contributor contributorType="DataCollector">
```

```
    <contributorName>Portal de Datos Abiertos del Ayuntamiento de
    ↪ Madrid</contributorName>
    <affiliation>Ayuntamiento de Madrid</affiliation>
  </contributor>
</contributors>
<dates>
  <date dateType="Issued">2023-01-17</date>
</dates>
<language>es</language>
<resourceType resourceTypeGeneral="Dataset"/>
<alternateIdentifiers>
  <alternateIdentifier
  ↪ alternateIdentifierType="url">https://zenodo.org/record/7542519</alternateIdentifier>
</alternateIdentifiers>
<relatedIdentifiers>
  <relatedIdentifier relatedIdentifierType="DOI"
  ↪ relationType="IsVersionOf">10.5281/zenodo.7542518</relatedIdentifier>
</relatedIdentifiers>
<version>1.0.0</version>
<rightsList>
  <rights
  ↪ rightsURI="https://creativecommons.org/licenses/by/4.0/legalcode">Creative
  ↪ Commons Attribution 4.0 International</rights>
  <rights rightsURI="info:eu-repo/semantics/openAccess">Open Access</rights>
</rightsList>
<descriptions>
  <description descriptionType="Abstract">&lt;p&gt;El conjunto de datos
  ↪ &amp;quot;refomad-data.xlsx&amp;quot; ha sido utilizado en el proyecto
  ↪ REFOMAD. Incluye informaci&amp;oacute;n sobre la renta media por persona,
  ↪ el nivel de formaci&amp;oacute;n educativa de los residentes mayores de 25
  ↪ a&amp;ntilde;os y el n&amp;uacute;mero de centros educativos en cada uno
  ↪ de los 21 distritos del Ayuntamiento de Madrid durante los a&amp;ntilde;os
  ↪ 2016-2019. El nivel de formaci&amp;oacute;n tiene 6 niveles: no sabe leer
  ↪ ni escribir, bachiller elemental, formaci&amp;oacute;n profesional,
  ↪ titulados medios, estudios universitarios y nivel de estudios no consta.
  ↪ Los centros educativos se clasifican tambi&amp;eacute;n en 6 niveles:
  ↪ escuelas infantiles municipales, escuelas infantiles p&amp;uacute;blicas
  ↪ de la Comunidad Aut&amp;oacute;noma de Madrid, escuelas infantiles
  ↪ privadas, colegios p&amp;uacute;blicos infantil/primaria, institutos
  ↪ p&amp;uacute;blicos de educaci&amp;oacute;n secundaria y colegios privados
  ↪ infantil/primaria. Los datos de renta fueron proporcionados por el INE,
  ↪ los datos sobre la formaci&amp;oacute;n fueron obtenidos mediante
  ↪ encuestas propias realizadas por el equipo REFOMAD y los datos de los
  ↪ centros educativos fueron solicitados al portal de datos abiertos del
  ↪ Ayuntamiento de Madrid. Los datos est&amp;aacute;n en formato XLSX y
  ↪ tienen un tama&amp;ntilde;o de 17 kB. Los datos fueron preprocesados y
  ↪ curados utilizando herramientas de software libre como NumPy y pandas en
  ↪ Python versi&amp;oacute;n 3.10.6.&lt;/p&gt;
```

```
&lt;p&gt;Se incluyen adicionalmente los datos originales (&amp;quot;DATOS
↪  ORIGINALES.zip&amp;quot;) y el script de Python empleado para la lectura,
↪  preprocesado, curaci&amp;oacute;n y an&amp;aacute;lisis de los datos
↪  (&amp;quot;REFOMAD-ETL-CURATION.ipynb&amp;quot;).&lt;/p&gt;</description>
  </descriptions>
  <geoLocations>
    <geoLocation>
      <geoLocationPlace>Centro</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.7038</pointLongitude>
        <pointLatitude>40.4168</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
    <geoLocation>
      <geoLocationPlace>Arganzuela</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.6922</pointLongitude>
        <pointLatitude>40.3964</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
    <geoLocation>
      <geoLocationPlace>Retiro</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.7167</pointLongitude>
        <pointLatitude>40.3719</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
    <geoLocation>
      <geoLocationPlace>Salamanca</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.6167</pointLongitude>
        <pointLatitude>40.3818</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
    <geoLocation>
      <geoLocationPlace>Chamartín</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.6743</pointLongitude>
        <pointLatitude>40.4096</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
    <geoLocation>
      <geoLocationPlace>Tetuán</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.6856</pointLongitude>
        <pointLatitude>40.4313</pointLatitude>
      </geoLocationPoint>
```

```
</geoLocation>
<geoLocation>
  <geoLocationPlace>Chamberí</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.6944</pointLongitude>
    <pointLatitude>40.4531</pointLatitude>
  </geoLocationPoint>
</geoLocation>
<geoLocation>
  <geoLocationPlace>Ciudad Lineal</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.7167</pointLongitude>
    <pointLatitude>40.4168</pointLatitude>
  </geoLocationPoint>
</geoLocation>
<geoLocation>
  <geoLocationPlace>Fuencarral-El Pardo</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.7236</pointLongitude>
    <pointLatitude>40.4148</pointLatitude>
  </geoLocationPoint>
</geoLocation>
<geoLocation>
  <geoLocationPlace>Moncloa-Aravaca</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.6471</pointLongitude>
    <pointLatitude>40.5497</pointLatitude>
  </geoLocationPoint>
</geoLocation>
<geoLocation>
  <geoLocationPlace>Latina</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.7245</pointLongitude>
    <pointLatitude>40.4386</pointLatitude>
  </geoLocationPoint>
</geoLocation>
<geoLocation>
  <geoLocationPlace>Carabanchel</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.6169</pointLongitude>
    <pointLatitude>40.3969</pointLatitude>
  </geoLocationPoint>
</geoLocation>
<geoLocation>
  <geoLocationPlace>Usera</geoLocationPlace>
  <geoLocationPoint>
    <pointLongitude>-3.7813</pointLongitude>
    <pointLatitude>40.3573</pointLatitude>
```

```
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>Villaverde</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.7155</pointLongitude>
      <pointLatitude>40.3538</pointLatitude>
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>Villa de Vallecas</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.7339</pointLongitude>
      <pointLatitude>40.3095</pointLatitude>
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>Vicálvaro</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.8667</pointLongitude>
      <pointLatitude>40.3322</pointLatitude>
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>San Blas-Canillejas</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.5667</pointLongitude>
      <pointLatitude>40.3831</pointLatitude>
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>Barajas</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.5667</pointLongitude>
      <pointLatitude>40.4667</pointLatitude>
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>Hortaleza</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.8567</pointLongitude>
      <pointLatitude>40.3742</pointLatitude>
    </geoLocationPoint>
  </geoLocation>
  <geoLocation>
    <geoLocationPlace>Moratalaz</geoLocationPlace>
    <geoLocationPoint>
      <pointLongitude>-3.7567</pointLongitude>
```

```
        <pointLatitude>40.4313</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
    <geoLocation>
      <geoLocationPlace>Puente de Vallecas</geoLocationPlace>
      <geoLocationPoint>
        <pointLongitude>-3.6285</pointLongitude>
        <pointLatitude>40.4457</pointLatitude>
      </geoLocationPoint>
    </geoLocation>
  </geoLocations>
</resource>
```