Minería de Datos

Estudio sobre energía renovable en los hogares españoles

Entregable 2

Integrantes y porcentaje de trabajo:

Integrante	Porcentaje de trabajo
Elena Ballesteros Morallón	21.5 %
Francisco Javier Luna Ortiz	21.5 %
Antonio Gómez Jimeno	21.5 %
Sergio Herreros Fernández	21.5 %
Pedro Sánchez Martín	14 %

Tabla de contenidos.

1. Introducción

En este segundo entregable, se preprocesarán los datasets que se han definido en la primera entrega y se obtendrán como resultado, las tarjetas de datos para cada una de las hipótesis.

El procedimiento que se ha seguido es:

- Procesar los datos de la capa *RAW*, eliminando la información que no sea necesaria y tratando los valores nulos. Estos nuevos datasets se guardan en la carpeta *SILVER*.
- A partir de la capa *SILVER*, se seleccionan las columnas y datos necesarios para validar cada hipótesis, se normalizan los valores, se hace un *profiling* y se crea la capa *GOLD*.

Se ha decidido crear cinco tarjetas de datos, una por hipótesis, ya que la mayoría de ellas no comparten columnas ni datos. También, esta decisión facilita la distribución de trabajo entre los miembros del grupo.

1.1. Cambios sobre el primer entregable

En el dataset de *consumo eléctrico anual* el dato del consumo de Cataluña está muy lejos respecto al de otras comunidades como Madrid. A raíz de esto, se consultó de nuevo el origen de los datos y se decidió descartarlo, ya que representa el consumo eléctrico industrial y este estudio está centrado en los hogares.

Se ha sustituido por otro dataset con el consumo eléctrico en los hogares de Castilla-La Mancha, desglosado por provincias y años desde el 2000 al 2023. El objetivo sigue siendo el mismo, analizar la tendencia en un futuro y ver qué provincias consumen más. Este dataset solo se usará en la hipótesis 5, por lo que no es necesario buscar información sobre otras comunidades autónomas.

2. Limpieza y transformación

En este apartado se explica el proceso de limpieza y transformación de cada dataset de la capa *RAW*, para obtener la capa *SILVER*.

En general, los pasos seguidos han sido:

- Cargar los datos en un dataframe con la librería pandas. Asegurándose de que los tipos de datos son los correctos.
- Eliminar las columnas o filas de datos totales, ya que en caso de ser necesarios para las tarjetas se podrán calcular.
- Eliminar los datos a nivel nacional. El objetivo es hacer un estudio a nivel provincial y una vez analizados las conclusiones, se podrían extrapolar a nivel nacional.
- Eliminar las columnas con el nombre de la comunidad autónoma para evitar duplicidad.
- Eliminar las provincias que no están en la península. Como ya se explicó en el anterior entregable, el alcance se limitará a las provincias de la península ya que están conectadas a la misma red eléctrica y tienen características similares de población. Por ejemplo, las ciudades autónomas no se pueden comparar con Soria o Madrid.
- Guardar el dataframe en un archivo csv. Los datasets procesados se guardan dentro de la carpeta 'data/silver'.

A continuación, se mencionan algunos de los pasos más significativos en el procesamiento de dataset:

Verificación de tipos

Por un lado, al tratar con datos de la INE que están en español, ciertos valores numéricos expresan sus unidades en miles con puntos. Esto supone un problema a la hora de preprocesar los datos. Pues el método *read_csv* de pandas trata estos números como si fuesen de tipo *float*, al leerlos como si fuesen números decimales (en el sistema decimal inglés se usa el punto).

Para solventar este problema, se ha utilizado el parámetro *dtype* de *read_csv* para leer los datos en tipo *string*. Y posteriormente, con el método replace se ha sustituido el punto del número por cadena vacía. Pudiendo transformar definitivamente los números (en *string*) a enteros con el método *astype*. Los datasets tratados son:

  - Dispositivos de energía renovable.
  - Distribución edad.
  - Intensidad de uso en viviendas.
  - Renta media.
  - Tipo núcleo familiar.
  - Viviendas según el número de personas.

Tratamiento de nulos

En general, ningún dataset tenía valores nulos o han sido eliminados al filtrar las columnas para quedarnos solo con los datos necesarios. Sin embargo, en el caso del dataset de "Renta media por hogar" faltaba el valor de renta media por hogar total para la provincia de Navarra en el año 2020. Para evitar tener que eliminar todos los valores de un año entero o hacer media de valores, se ha añadido el valor correspondiente (37.802). Siendo consultado en el Instituto de Estadística de Navarra, Nastat.

Dispositivos de energía renovable

Hay varios tipos de dispositivos, pero solo nos interesa saber si un hogar dispone o no de un dispositivo. Por ello, se han restado al número del total de viviendas el valor de las viviendas que no tienen dispositivos y así conseguimos el dato de las que sí disponen, independientemente del tipo o si combinan varios.

3. Líneas de trabajo

Para cada línea de trabajo, es decir, para cada hipótesis, se ha creado una tarjeta de datos. Esta decisión facilita la distribución de trabajo entre los miembros del grupo y facilita la identificación y compresión de los datos necesarios para validar cada hipótesis. Después de esta transformación de la capa *SILVER*, se obtendrá la capa *GOLDEN*.

3.1. Tarjeta Hipótesis 1

I. Hipótesis

"Las provincias con una edad media menor y una renta media por hogar mayor a la nacional, suelen estar más concienciadas con el uso de energías renovables y utilizan más dispositivos que aprovechan este tipo de energía."

II. Justificación y explicación

La justificación de esta hipótesis es que las personas jóvenes suelen estar más concienciadas con el medio ambiente, además de que tienen más tiempo para amortizar la inversión de este tipo de tecnología a lo largo de su vida.

Para el desarrollo de esta tarjeta de datos, se ha recopilado información de varios conjuntos de datos. Todos ellos han sido procesados (capa silver).

- distribucion_edad.csv
- renta_media_hogar.csv
- dispositivos_renovable.csv

Para responder a la hipótesis necesitamos una tarjeta de datos que debe centrarse en las provincias de España, incluyendo la población total de cada provincia. También debe contener la edad y la renta media por hogar de cada provincia, variables clave para validar la hipótesis. Finalmente, se debe incluir el porcentaje de hogares que utilizan dispositivos de energía renovable, para analizar la relación entre juventud, poder adquisitivo y adopción de la energía renovable.

III. Diccionario de datos

Nombre del campo	Tipo de dato	Descripción
Provincias	String	Indica el nombre de la provincia.
Población total	Int	Cantidad total de personas que habitan en la provincia.
Renta media por hogar	Int	Ingreso promedio anual de los hogares en la provincia, expresado en euros.
Edad media	Float	Edad media de los habitantes en la provincia.
Porcentaje de hogares con dispositivos de energía renovable	Float	Proporción de hogares en la provincia que utilizan dispositivos de energía renovable, expresado como un porcentaje (0%-100%).

El dataset asociado a esta tarjeta de datos está en la capa gold con el nombre "data_card_1_df.csv".

3.2. Tarjeta Hipótesis 2

I. Hipótesis

"*Existe una relación entre el porcentaje de viviendas de bajo consumo o esporádicas, de viviendas de consumo medio y de viviendas de alto consumo de una provincia y su tendencia a adoptar dispositivos de energía renovable"*.

II. Justificación y explicación

Las viviendas de bajo consumo o esporádicas suelen corresponder a residencias secundarias, utilizadas principalmente durante vacaciones o de forma ocasional. Debido a esto, las provincias con mayor porcentaje de este tipo de residencias suelen tener menor tendencia a adquirir dispositivos de energías renovables debido al poco tiempo que sus propietarios pasan en ellas.

Por otro lado, en las viviendas de alto consumo, sus propietarios suelen tener un poder adquisitivo mayor, y generalmente, prefieren contratar la electricidad a lidiar con los posibles problemas técnicos de este tipo de instalaciones.

Por último, las viviendas de consumo medio representan un equilibrio entre los costes y beneficios de las energías renovables. Por ello, la instalación de dispositivos de energía renovable puede ser una opción atractiva para los propietarios, ya que permite reducir gastos energéticos de manera significativa a medio y largo plazo.

Para esta segunda hipótesis, se quiere realizar una triple comparación entre cada uno de los tres grupos según la intensidad de uso eléctrico. Estos grupos son los siguientes:

- **Viviendas de bajo consumo**: Pertenecerán a este grupo todas las viviendas con consumo mayor a **0 kwh** (para evitar contabilizar las viviendas vacías) y menor o igual a **750 kwh**.
- **Viviendas de consumo medio**: Este grupo está compuesto por todas aquellas viviendas cuyo consumo eléctrico se sitúa entre **1001 kwh** y **4000 kwh**.
- **Viviendas de alto consumo**: Se consideran viviendas de alto consumo aquellas que superen un umbral de consumo de **6000 kwh**.

En cuanto a la creación de la tarjeta de datos, se utilizarán datasets previamente procesados y almacenados en la capa *silver*. Estos conjuntos de datos contienen información relacionada con la intensidad de uso en viviendas y el uso de dispositivos

renovables, desglosada por provincias de España. Los datasets utilizados se encuentran en la carpeta *silver* y son los siguientes:

- dispositivos_renovable.csv
- intensidad_de_uso_en_viviendas.csv

III. Diccionario de datos

Nombre del campo	Tipo de dato	Descripción
Provincias	String	Indica el nombre de la provincia.
Índice de viviendas de bajo consumo	Float	Proporción de hogares en la provincia que tienen un consumo entre 0 y 750 kwh. Está expresado como porcentaje (0 – 100%).
Índice de viviendas de medio consumo	Float	Proporción de hogares en la provincia que tienen un consumo entre 1001 y 4000 kwh. Está expresado como porcentaje (0 – 100%).
Índice de viviendas de alto consumo	Float	Proporción de hogares en la provincia que tienen un consumo superior a 6000 kwh. Está expresado como porcentaje (0 – 100%).
Índice de viviendas con dispositivos renovables	Float	Proporción de hogares en la provincia que utilizan dispositivos de energía renovable. Está expresado como porcentaje (0 – 100%).

Se creó un conjunto de datos asociado a esta tarjeta de datos en la capa *gold* con el nombre "*data_card_2_df.csv*". Se utilizaron los datasets mencionados anteriormente, encontrados en la capa *silver*.

3.3. Tarjeta Hipótesis 3

I. Hipótesis

"*A mayor número de personas en un hogar es más probable que se invierta en energía renovable"*.

II. Justificación y explicación

En un hogar en el que haya más personas, optar por dispositivos de energía renovable puede ser una decisión lógica y justificada. Debido a que el consumo energético es mayor y, por lo tanto, las facturas de electricidad son más elevadas. Aunque pueden ser significativamente más bajas al contar con energía renovable.

Por otro lado, con dispositivos de energía renovable, el hogar no depende de la red eléctrica ni de compañías eléctricas, además de disminuir el riesgo de apagones. Además, el Plan de Recuperación Transformación y Resiliencia del Gobierno de España (financiado por la Unión Europea) ofrece incentivos por instalar tecnologías renovables, haciendo más accesible la inversión inicial para familias grandes.

Para sacar conclusiones sobre esta hipótesis y poder llegar a validarla, se tendrá en cuenta el número de dispositivos renovables y el número de tipos de hogares en cada provincia. Se han utilizado los siguientes conjuntos de datos:

- tipo_nucleo_familar.csv
- dispositivos_renovable.csv

Se tendrá en cuenta el número de diferentes tipos de hogar en España para ver si hay correlación entre el tipo de hogar en cuanto a personas y el número de dispositivos electrónicos. Se observará distinciones como que si una familia nuclear es más propensa a tener dispositivos de energía renovable que una familia monoparental.

III. Diccionario de datos

Nombre del campo	Tipo de dato	Descripción
Provincias	String	Indica el nombre de la provincia.
Índice dispositivos energía renovable	Float	Índice de dispositivos de energía renovable utilizados en viviendas en cada provincia.
Familia monoparental con 0 hijos	Float	Índice de familias monoparentales (no se distingue entre padre y madre) con 0 hijos.
Familia monoparental con 1 hijo	Float	Índice de familias monoparentales (no se distingue entre padre y madre) con 1 hijo.

Familia monoparental con 2 hijos o más	Float	Índice de familias monoparentales (no se distingue entre padre y madre) con 2 hijos.
Pareja casada con 0 hijos	Float	Índice de familias compuestas por pareja casada con 0 hijos.
Pareja casada con 1 hijo	Float	Índice de familias compuestas por pareja casada con 1 hijo.
Pareja casada con 2 hijos o más	Float	Índice de familias compuestas por pareja casada con 2 hijos.
Pareja no casada con 0 hijos	Float	Índice de familias compuestas por pareja no casada con 0 hijos.
Pareja no casada con 1 hijo	Float	Índice de familias compuestas por pareja no casada con 1 hijo.
Pareja no casada con 2 hijos o más	Float	Índice de familias compuestas por pareja no casada con 2 hijos.

Se creó un conjunto de datos asociado a esta tarjeta de datos en la capa gold con el nombre "data_card_3_df.csv" (capa *gold*). Se utilizaron los datasets mencionados anteriormente, encontrados en la capa *silver*.

3.4. Tarjeta Hipótesis 4

I. Hipótesis

- *Las provincias situadas en el sur de España (aquellas al sur de Madrid) tienden a utilizar menos dispositivos de aprovechamiento de energías renovables que las del norte.*

II. Justificación y explicación

Las diferencias climáticas y socioeconómicas entre el norte y sur de España pueden influir significativamente en la adopción de dispositivos de energía renovable. Las provincias del norte tradicionalmente han experimentado un desarrollo industrial más temprano y sostenido, lo que podría traducirse en una mayor propensión a adoptar nuevas tecnologías energéticas. Además, las condiciones meteorológicas del norte, con

menor exposición solar pero más recursos eólicos, podrían favorecer la diversificación de fuentes de energía renovable.

Por otro lado, las provincias del sur, a pesar de contar con mayor potencial solar, pueden enfrentar barreras económicas y estructurales que dificulten la adopción de estas tecnologías. El menor poder adquisitivo medio en algunas provincias del sur podría limitar la capacidad de inversión inicial necesaria para la instalación de estos dispositivos, a pesar de los beneficios a largo plazo.

Se analizará no solo el porcentaje de dispositivos renovables, sino también factores que podrían influir en su adopción como la renta normalizada, la mediana de edad y la estructura familiar. Esto permitirá determinar si las diferencias observadas se deben realmente a la ubicación geográfica o están más relacionadas con otros factores socioeconómicos.

Para el análisis de esta hipótesis, se utilizarán datos procesados que incluyen información sobre dispositivos renovables y características sociodemográficas por provincia. Los conjuntos de datos empleados son:

- dispositivos_renovable.csv
- renta_media_hogar.csv
- tipo_nucleo_familiar.csv
- distribucion_edad.csv
- produccion_lugar.csv

Para el próximo entregable, se planea utilizar clustering jerárquico para comprobar si los grupos de provincias coinciden con zonas geográficas próximas y analizar cómo se agrupan.

III. Diccionario de datos

Nombre del campo	Tipo de dato	Descripción
Provincias	String	Indica el nombre de la provincia.
Porcentaje con dispositivo	Float	Porcentaje de hogares que tienen instalado algún dispositivo de aprovechamiento de energías renovables.

Renta normalizada	Float	Nivel de renta medio.
Mediana edad	Float	Edad mediana de la población.
Producción media	Float	Producción media de energía renovable por hogar.
Familia de 1 padres y 0 hijos	Float	Familias compuestas por 1 padre con 0 hijos.
Familia de 1 padres y 1 hijos	Float	Familias compuestas por 1 padre con 1 hijos.
Familia de 1 padres y 2 hijos	Float	Familias compuestas por 1 padre con 2 hijos.
Familia de 2 padres y 0 hijos	Float	Familias compuestas por 2 padres con 0 hijos.
Familia de 2 padres y 1 hijos	Float	Familias compuestas por 2 padres con 1 hijos.
Familia de 2 padres y 2 hijos	Float	Familias compuestas por 2 padres con 2 hijos.

3.5. Tarjeta Hipótesis 5

I. Hipótesis

"*¿Qué cantidad de placas solares fotovoltaicas se necesitarían instalar para abastecer el consumo eléctrico de Castilla-La Mancha en el año 2028?"*

II. Justificación y explicación

Para responder a la pregunta de esta hipótesis, se necesitan los siguientes datos: consumo eléctrico de Castilla-La Mancha en 2028, producción de energía eléctrica fotovoltaica en esta región y superficie de un panel fotovoltaico. Se ha recopilado información de los siguientes datasets:

- consumo_electrico_clm.csv
- produccion_lugar.csv

El primer dato, sobre el consumo eléctrico, se ha obtenido del dataset de *consumo_electrico_clm,* que contiene un histórico con el consumo por mes desde el año 2000 al 2023. Sin embargo, al ver la distribución de los datos, hay mucha variabilidad entre meses y hay dos tendencias: entre el año 2000 y al 2013 hay un pico y descenso de consumo y a partir de 2014 se estabiliza en la mayoría de las provincias (Ver Fig. 1). Por esta razón, se ha decidido agrupar los datos en años y elegir solo el rango desde 2014 a 2023 para realizar la predicción.

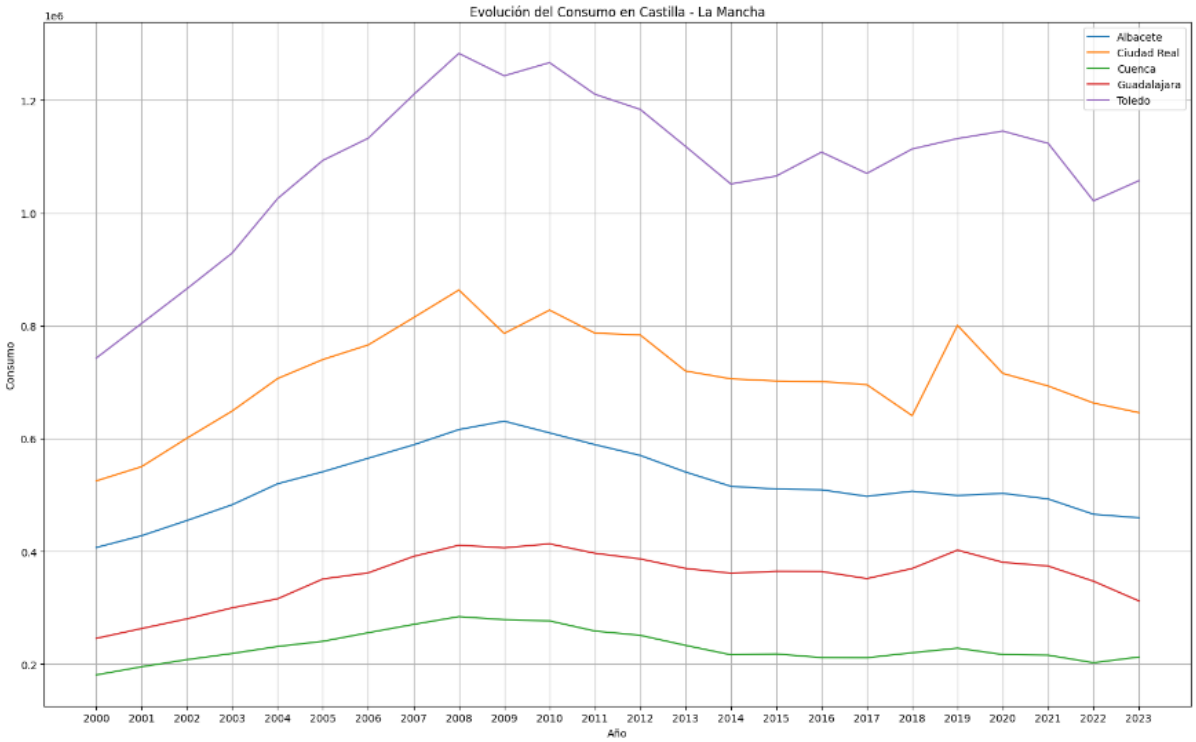

Fig. 1: Gráfico sobre la evolución del consumo en las provincias de Castilla-La Mancha entre los años 2000 y 2023

El segundo dato, sobre la producción de energía, se ha obtenido del dataset produccion_lugar. Para unificarlo con los datos anteriores, se han filtrado las filas con las provincias de Castilla-La Mancha y se han sumado los datos de cada mes. De esta forma, tenemos la producción de un metro cuadrado de panel solar en un año en cada una de las provincias de la región. Como las unidades de este dato son kWh, se han modificado para obtener MWh.

Respecto al tercer dato, dependerá del modelo de placa. Para la potencia generada, se ha utilizado una placa genérica, con los parámetros de 1kWp de potencia y un 14% de pérdidas del sistema. Estos datos son importantes para la estimación que se hará

respecto al número de placas que se deberían instalar ya que varían según diferentes modelos.

Así, en el próximo paso será crear un modelo para hacer la predicción de consumo en 2028 y calcular la estimación del número de placas solares necesarias.

III. Diccionario de datos

Nombre del campo	Tipo de dato	Descripción
Provincia	String	Nombre de las 5 provincias de Castilla-La Mancha
Potencia MWh	Float	Potencia eléctrica generada anualmente por 1 m^2 de una placa fotovoltaica
X	Float	Consumo eléctrico en MWh para el año X (de 2014 a 2023)

4. Profiling de las tarjetas

El profiling de datos es el proceso de analizar y examinar un conjunto de datos para obtener información detallada sobre su calidad, estructura y características. Este proceso ayuda a identificar patrones, inconsistencias, valores faltantes, duplicados, anomalías y otros aspectos importantes de los datos.

Todos los profiling de las tarjetas de datos se pueden encontrar en la carpeta

4.1. Hipótesis 1

Se han observado varias variables con alta correlación negativa.

Variable 1	Variable 2	Correlación
Edad media	Población total	-0.535
Edad media	Porcentaje de hogares con dispositivos de energía renovable	-0.511

La variable '*Edad media'* tiene una alta correlación negativa con '*Población total'* y '*Porcentaje de hogares con dispositivos de energía renovable'*.

Estas relaciones entre variables se pueden explicar de la siguiente manera:

- 'Edad media' y 'Población total' (-0.535): Una posible explicación para esta correlación negativa podría ser que, en áreas con una población total más alta la proporción de personas jóvenes puede ser mayor, lo que reduce la edad media de la población. En cambio, en zonas con una población total más baja puede haber una mayor concentración de personas mayores, lo que eleva la edad media. Este patrón sugiere que, en lugares con menos habitantes, la población envejece más rápidamente, mientras que en lugares más poblados hay una mayor presencia de jóvenes.

- 'Edad media' y 'Porcentaje de hogares con dispositivos de energía renovable' (-0.511): Puede ser que las zonas con una mayor edad media tienden a tener un menor porcentaje de hogares con energía renovable, ya que las personas mayores podrían estar menos inclinadas a adoptar nuevas tecnologías o a realizar inversiones en dispositivos de energía renovable. Por el contrario, las zonas con una población más joven pueden estar más abiertas a la adopción de tecnologías sostenibles.

4.2. Hipótesis 2

Durante el análisis exploratorio de los datos (EDA) mediante un profiling, se han identificado correlaciones significativas entre algunas de las variables.

Variable 1	Variable 2	Correlación
Índice de viviendas renovables	Índice de viviendas de medio consumo	0.433
Índice de viviendas de alto consumo	Índice de viviendas de bajo consumo	-0.695

A continuación, se presentan las justificaciones para las principales correlaciones observadas en los datos.

- 'Índice de viviendas renovables' y 'Índice de viviendas de medio consumo' (0.433): Esta correlación positiva muestra una relación moderada entre el índice de viviendas renovables y el índice de viviendas de medio consumo por provincia.

Este valor indica que, a medida que aumenta la proporción de viviendas de medio consumo, también tiende a aumentar el porcentaje de viviendas con al menos un dispositivo de aprovechamiento de energía renovable. Esta relación podría explicarse por el hecho de que las viviendas de consumo medio representan un equilibrio entre los costes y beneficios de realizar una instalación doméstica de dispositivos de aprovechamiento de energías renovables, lo que puede resultar ser una opción interesante para los propietarios.

- 'Índice de viviendas de alto consumo' e 'Índice de viviendas de bajo consumo' (-0.695): Se trata de una correlación negativa fuerte entre la proporción de viviendas de alto consumo y las de bajo consumo. Este valor sugiere que a medida que aumenta el porcentaje de viviendas de alto consumo, disminuye proporcionalmente el de viviendas de bajo consumo. Esta relación es lógica y no nos ofrece información útil para probar la veracidad de la hipótesis.

4.3. Hipótesis 3

Con el profiling realizado para la tarjeta de datos de la hipótesis 3, se puede ver correlaciones interesantes:

Variable 1	Variable 2	Número	Correlación
Índice dispositivos energía renovable	Familia monoparental con 0 hijos	1	-0.464
Índice dispositivos energía renovable	Familia monoparental con 2 hijos o más	3 o más	0.351
Índice dispositivos energía renovable	Pareja casada con 0 hijos	2	-0.534
Índice dispositivos energía renovable	Pareja casada con 2 hijos o más	4 o más	0.366
Índice dispositivos energía renovable	Pareja no casada con 2 hijos o más	4 o más	0.548

- Por un lado, la variable índice dispositivos energía renovable tiene **correlación negativa** con Familia monoparental con 0 hijos y con Pareja casada con 0 hijos.

- Por otro lado, la variable índice dispositivos energía renovable tiene **correlación positiva** con Familia monoparental con 2 hijos o más, Pareja casada con 0 hijos y con Pareja no casada con 2 hijos o más.

- Para explicar estas correlaciones, se ha añadido a la tabla una columna aclarativa de número de miembros que tiene cada tipo de familia. De esta manera, se deduce rápidamente que los núcleos familiares que tienen menos miembros presentan la correlación negativa. Esto da a entender que cuanto menor sea el número de miembros en una familia, menor será el número de dispositivos de energía renovable. En contaste, con las familias de más miembros, la correlación positiva explica que optan más por dispositivos de energía renovable.

- Cuando una familia cuenta con más miembros, los gastos como la vivienda y los servicios (luz, agua, etc.) son mayores, por lo que optar por invertir en dispositivos de energía renovable, es una buena opción para generar ahorros y disminuir los costos producidos por la alta demanda energética.

- Asimismo, las familias más numerosas tienen una mayor fuente de ingresos e inclusos varias. Esto les da una mayor capacidad económica para invertir en energía renovable, pues requiere gastos iniciales elevados.
- En cambio, en las familias menos numerosas, factores como tener menos ingresos, tener una vivienda más pequeña y no tener hijos son razones por las cuales no se optaría por energía renovable, pues la inversión es muy costosa para ellos y no hay tanta demanda energética, por lo que no se ve reflejado un gasto elevado en las facturas.

4.4. Hipótesis 4

Podemos observar alguna correlación interesante:

Variable 1	Variable 2	Correlación
Mediana edad	familias de 1 padres y 2 hijos	0.406
Mediana edad	familias de 2 padres y 0 hijos	-0.622

Mediana edad	familias de 2 padres y 1 hijos	-0.507
Mediana edad	familias de 2 padres y 1 hijos	-0.670

- La relación de las familias con la mediana de edad es la más fuerte, siendo negativa en las familias con 2 padres y positiva monoparentales. Esto sugiere que las zonas con población más joven tienden a tener más familias tradicionales con dos padres y dos hijos, mientras que las áreas con población de mayor edad tienden a tener más familias monoparentales con dos hijos.

4.5. Hipótesis 5

El profiling en esta tarjeta no ofrece mucha información, ya que solo hay 5 instancias. Sin embargo, hay algunos puntos que comentar:

- Las variables que representan el consumo de cada uno de los años aparecen todas con correlación igual a 1. Esto se debe a que las tendencias de consumo se mantienen estables entre años y es probable que si un año es alto al siguiente también lo sea.
- La variable 'Potencia MWh' tiene una correlación de 0.6 con el resto de las variables numéricas. Es posible que esta relación no sea causal ya que influirán otros factores.

5. Metodología seguida

Este segundo entregable, se corresponde con las etapas de limpieza y transformación del proceso KDD. Aunque, también ha sido necesario volver a la etapa inicial de recogida de datos para la búsqueda del nuevo dataset. Esto se explica en la sección 1.1 (Cambios sobre el primer entregable).

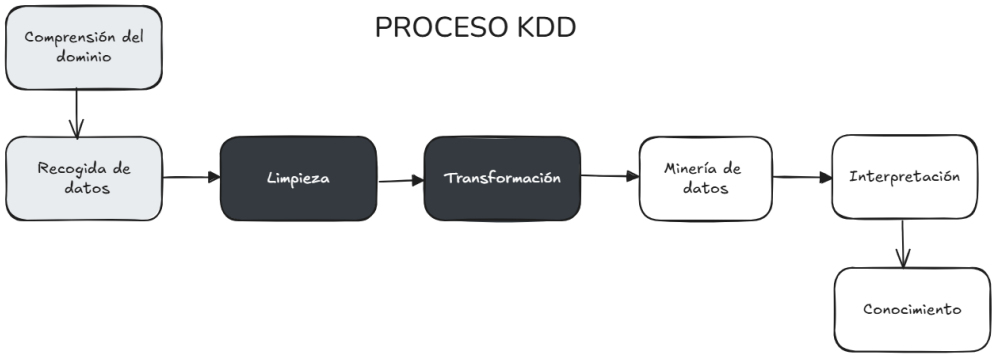

Fig. 2: Esquema de las fases del proceso KDD de este entregable.

Para procesar los todos los datasets se ha seguido la arquitectura de Medalion, que estructura los datos por capas:

- Raw o Bronze: almacena los datasets descargados, sin ningún tipo de procesamiento.
- Silver: los datos se procesan con técnicas de filtrado, validación y normalización.
- Gold: esta capa es la equivalente a las tarjetas de datos, que contienen los datos agregados y preparados para ser usados en la validación de las hipótesis.

Es decir, cada capa reduce la dimensionalidad del dataset y aumenta la calidad de los datos.

6. Referencias

Consumo eléctrico en Castilla – La Mancha: https://estadistica.castillalamancha.es/datos-sobre-consumo-electrico

Consumo eléctrico anual (sustituido): https://www.ine.es/jaxi/Tabla.htm?tpx=31411

Viviendas según número personas: https://www.ine.es/jaxi/Tabla.htm?path=/t20/p276/2020-2035/l0/&file=01001.px&L=0

Dispositivos energía renovable: https://www.ine.es/jaxi/Tabla.htm?tpx=56927&L=0

Intensidad uso viviendas: https://www.ine.es/jaxi/Tabla.htm?tpx=59531

Distribución edad población: https://www.ine.es/jaxiT3/Tabla.htm?t=36781&L=0

Salario medio población: https://www.ine.es/jaxiT3/Tabla.htm?t=13930

Producción de energía por lugar: https://re.jrc.ec.europa.eu/pvg_tools/es/

Tipo núcleo familiar: https://www.ine.es/dynt3/inebase/es/index.htm?padre=9544&capsel=9548

Renta Hogar INE: https://www.ine.es/jaxiT3/Tabla.htm?t=9949

Renta Hogar Nastat: https://nastat.navarra.es/es/tablas_powerbi/-/tag/estadistica-renta

7. Anexo

I. Visualizaciones de gráficas importantes

a. Índice de viviendas por provincia.

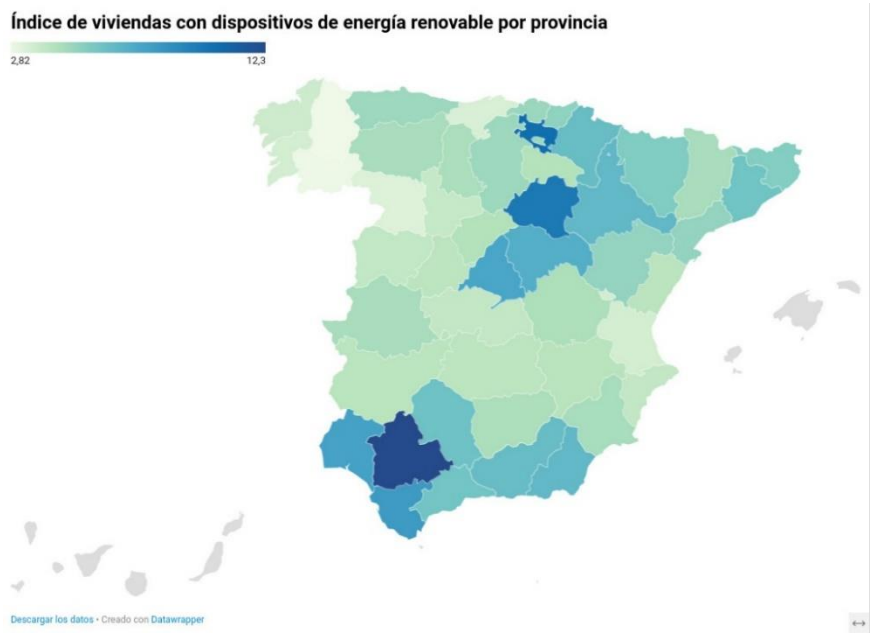

Fig. 3: Índice de viviendas con dispositivos de energía renovable por provincia.

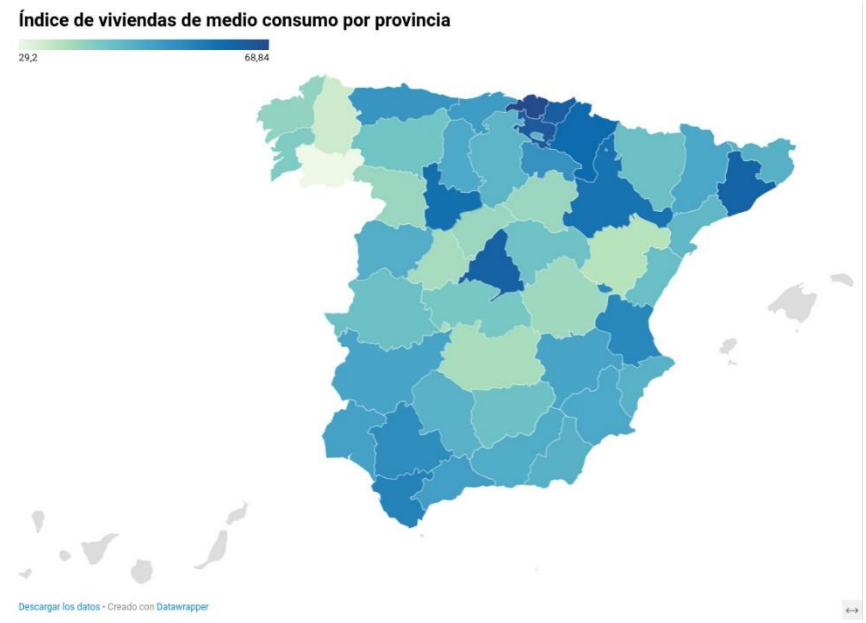

Fig. 4 Índice de viviendas de medio consumo por provincia.

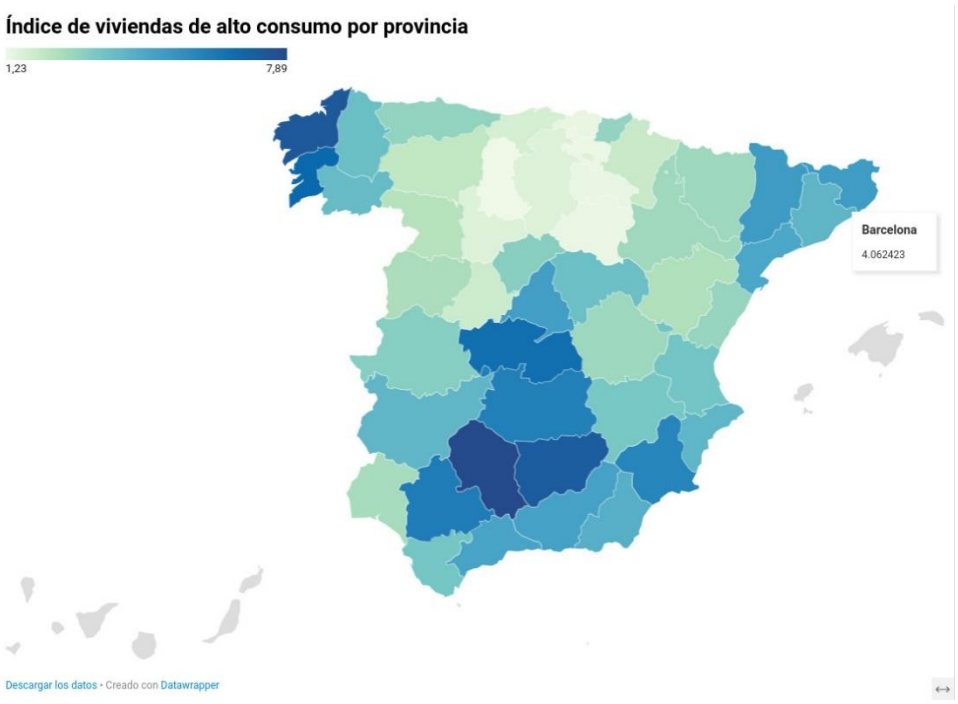

Fig. 5: Índice de viviendas de alto consumo por provincia.

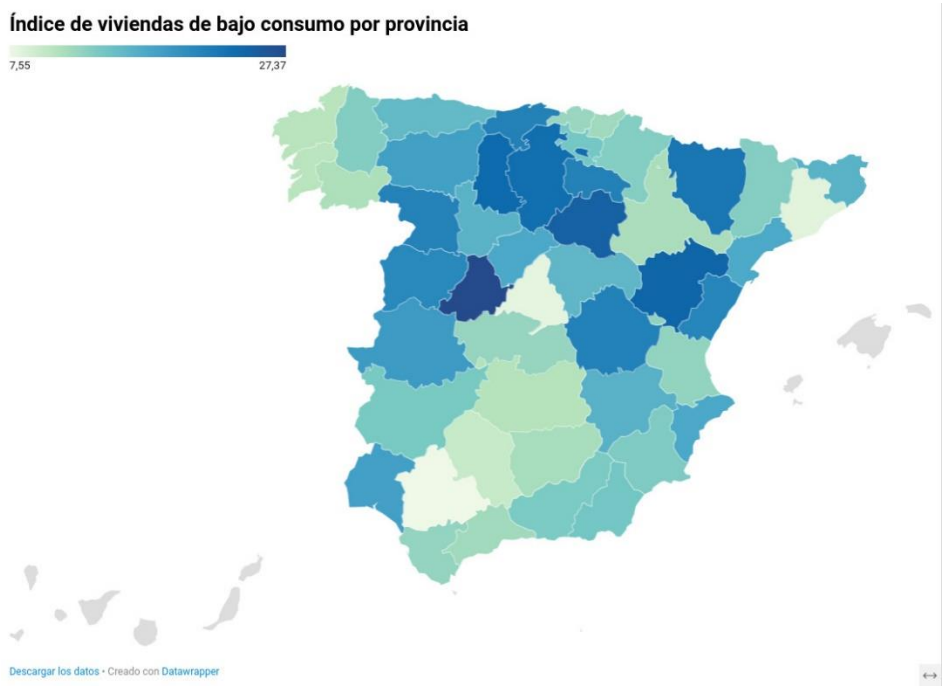

Fig. 5: Índice de viviendas de bajo consumo por provincia.

b. Población por provincia

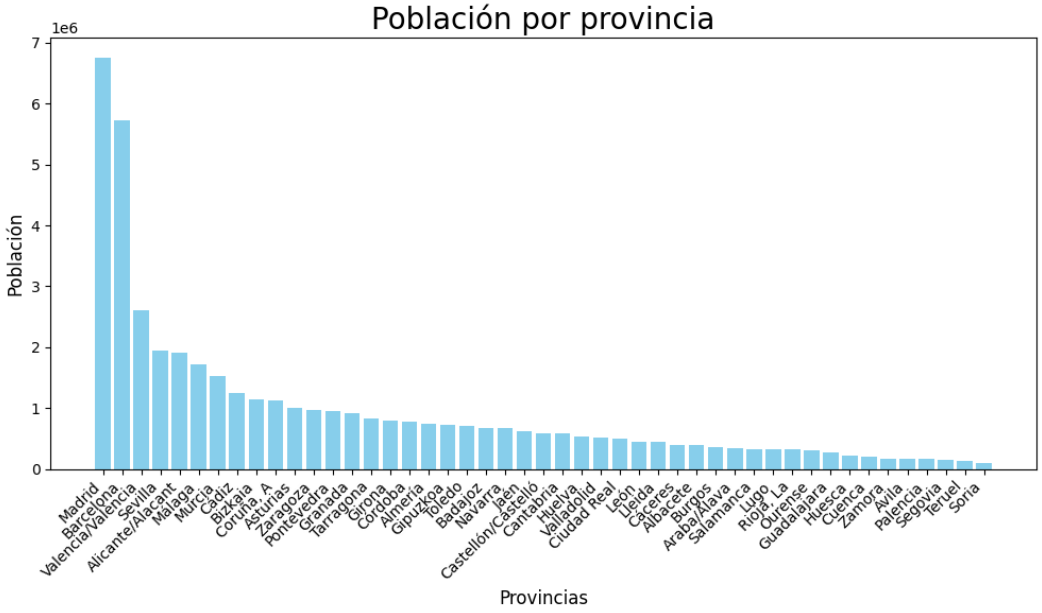

Fig. 6: Población total por provincia.

c. Renta media por hogar

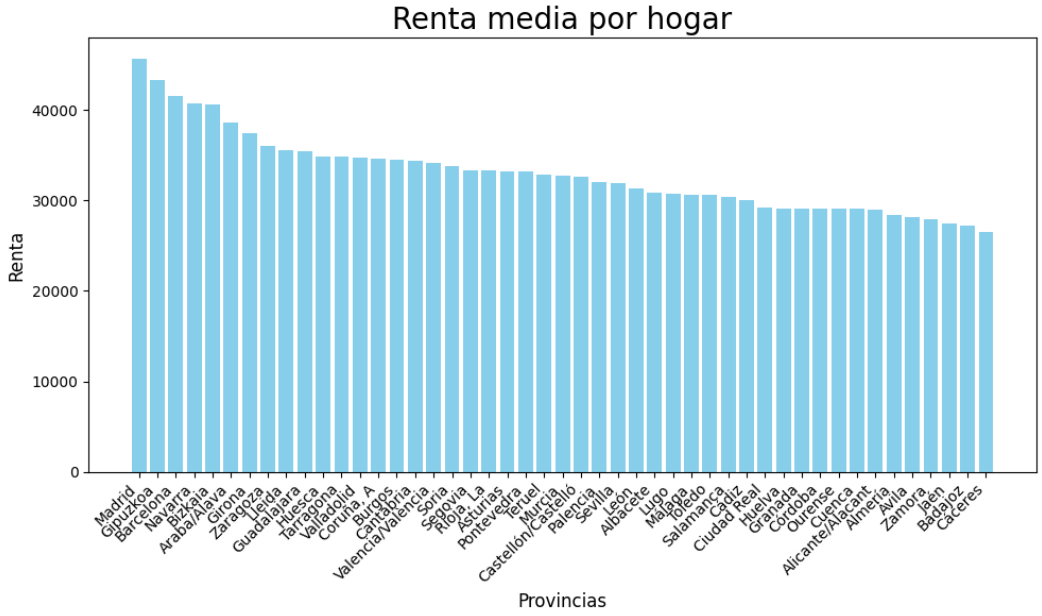

Fig. 7: Renta media por hogar.

d. Edad media

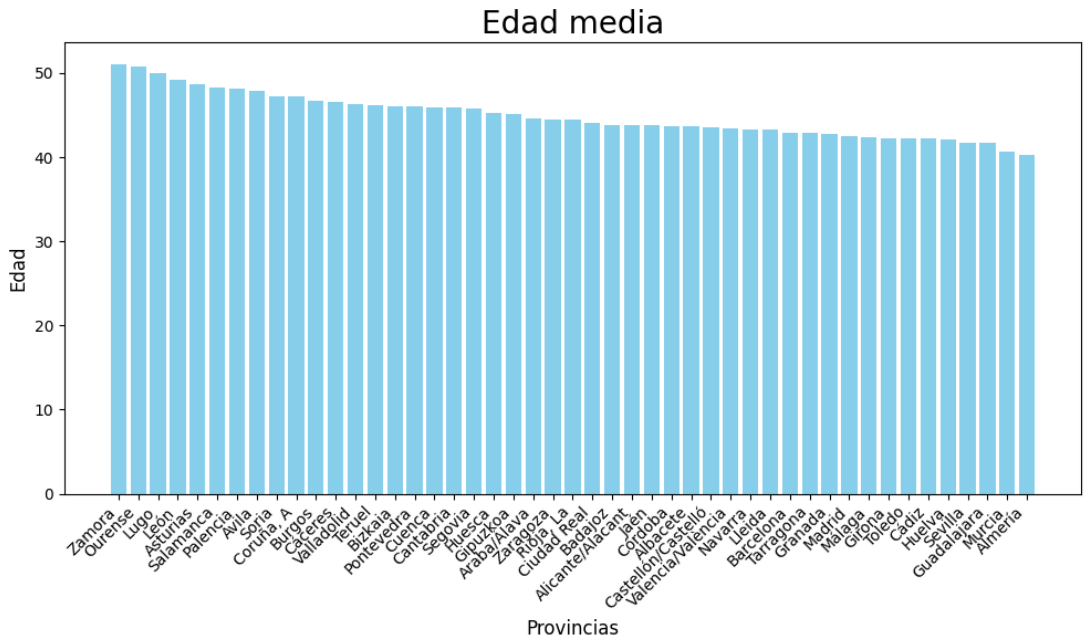

Fig. 8: Edad media por provincia.

e. Porcentaje de hogares con dispositivos de energía renovable

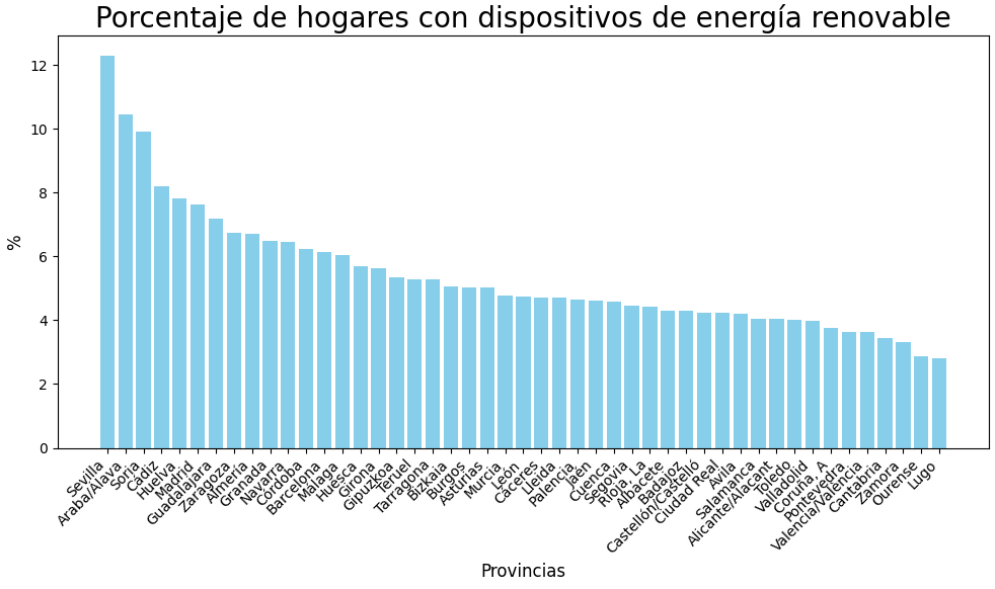

Fig. 9: Porcentaje de hogares con dispositivos de energía renovable por provincia.