MINERÍA DE DATOS

Estudio sobre las energías renovables en España

Elena Ballesteros
Francisco Javier Luna
Antonio Gómez
Sergio Herreros
Pedro Sánchez

CONTENIDOS

- Introducción
- Metodología
- Entregables
- Hipótesis
- Conclusiones

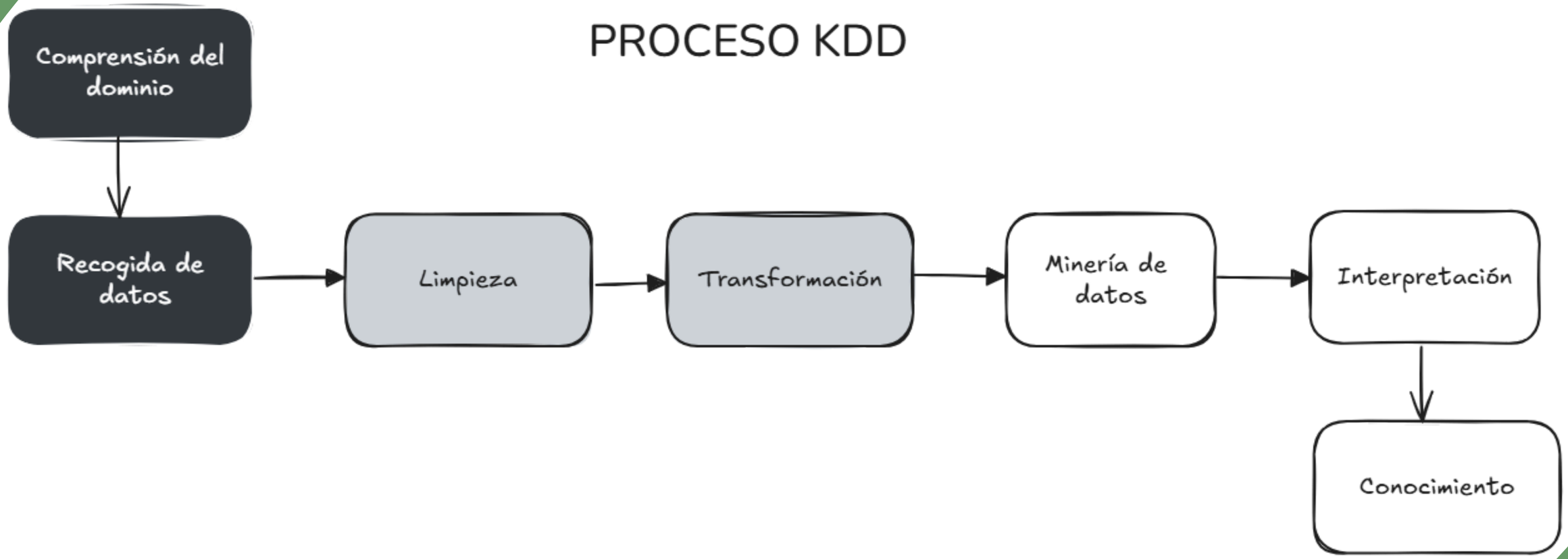
INTRODUCCIÓN

Problema: Aumento del consumo y precio de la electricidad.

Solución: Búsqueda de alternativas sostenibles, como el autoconsumo energético mediante energías renovables.

Objetivo: Identificar características demográficas y geográficas de clientes potenciales en España.

METODOLOGÍA



- Iterativo/incremental
- 3 entregables

ENTREGABLES

ENTREGABLE 1

- Comprensión del dominio.
- Recogida de datos.
- Hipótesis y objetivos del estudio.

ENTREGABLE 2

- Arquitectura Medallion.

- Procesamiento.
- Limpieza.
- Transformación

ENTREGABLE 3

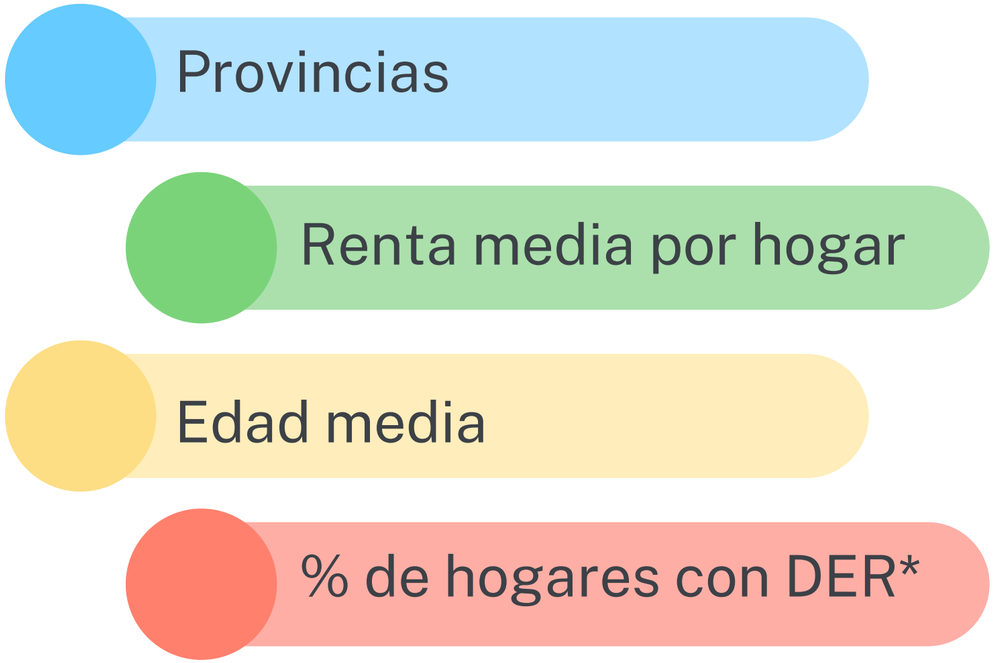
- Minería de datos.
- Interpretación y conocimiento.

Hipótesis 1:

LAS PROVINCIAS CON UNA EDAD MEDIA MENOR Y UNA RENTA MEDIA POR HOGAR MAYOR, SUELEN ESTAR MÁS CONCIENCIADAS CON EL USO DE ENERGÍAS RENOVABLES Y UTILIZAN MÁS DISPOSITIVOS QUE APROVECHAN ESTE TIPO DE ENERGÍA.

¿POR QUÉ PENSAMOS QUE ESTO ES ASÍ?

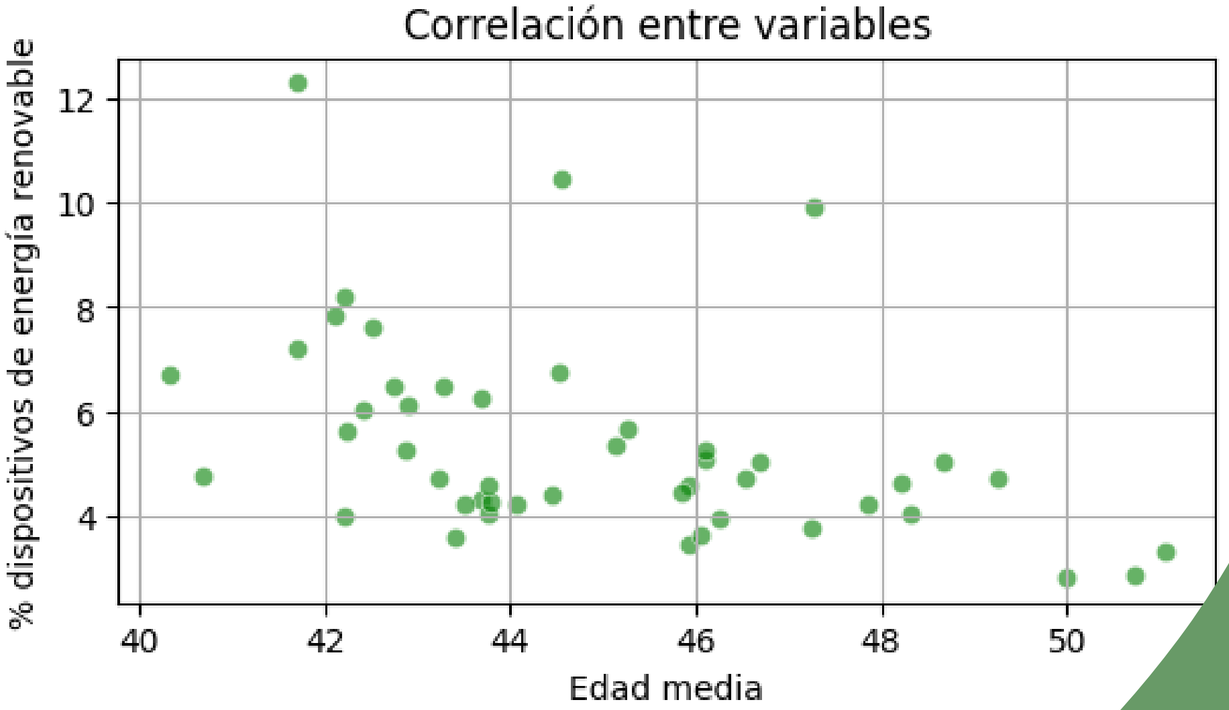
- Las generaciones más jóvenes, han crecido con una mayor conciencia sobre el cambio climático y la sostenibilidad.
- Los jóvenes tienen más tiempo para amortizar la inversión.
- Se necesita de una inversión inicial considerable.



DER -> Dispositivos de Energía Renovable

ESTUDIO DE CORRELACIÓN

RELACIÓN NEGATIVA MODERADA ENTRE LA EDAD MEDIA Y EL PORCENTAJE DE HOGARES CON DISPOSITIVOS DE ENERGÍA RENOVABLE.



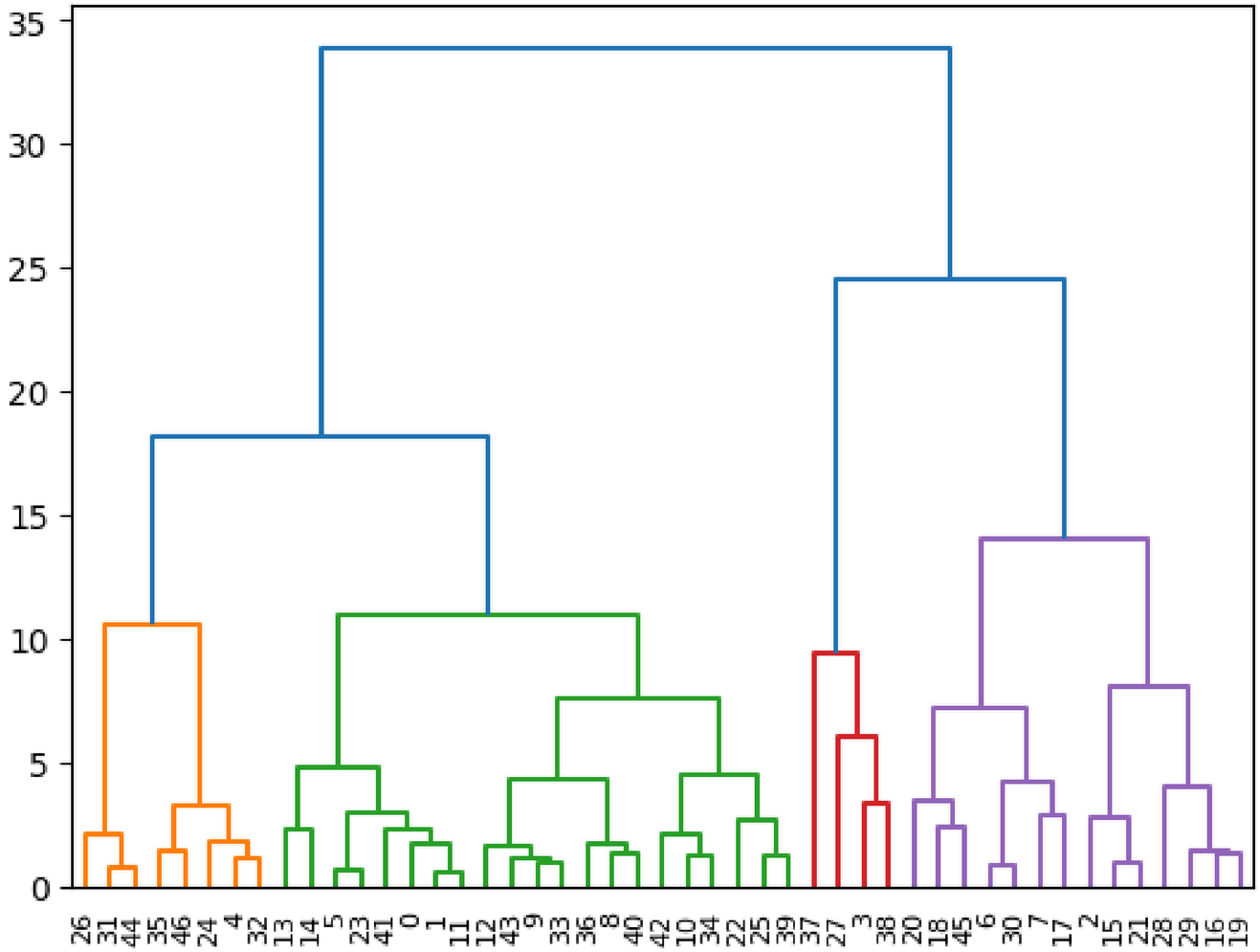
	Feature 1	Feature 2	Pearson Correlation	Pearson p-value	Spearman Correlation	Spearman p-value
0	Renta media por hogar	Edad media	-0.164277	0.269846	-0.113205	0.448667
1	Renta media por hogar	Porcentaje de hogares con dispositivos de ener...	0.254432	0.084373	0.312905	0.032234
2	Edad media	Porcentaje de hogares con dispositivos de ener...	-0.444344	0.001754	-0.511101	0.000241

CLUSTERING JERÁRQUICO

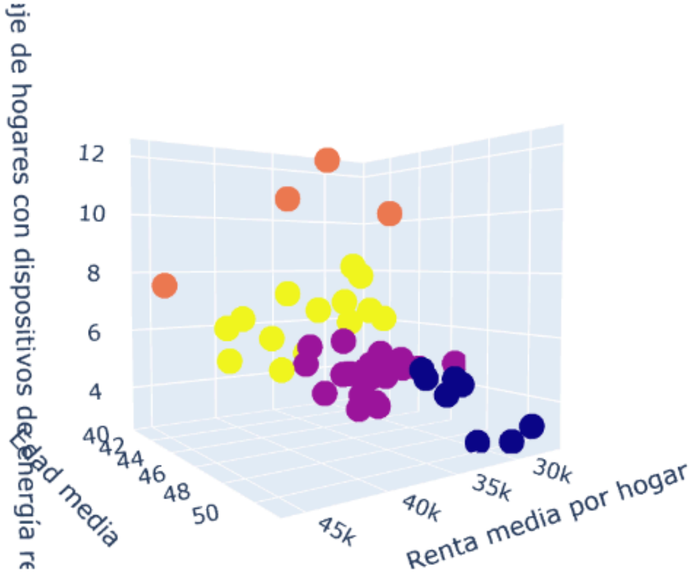
El clustering jerárquico agrupa los datos en un dendrograma que muestra la relación y jerarquía entre los clusters, desde la raíz (representa todos los elementos) hasta las hojas (corresponden a los grupos más similares).

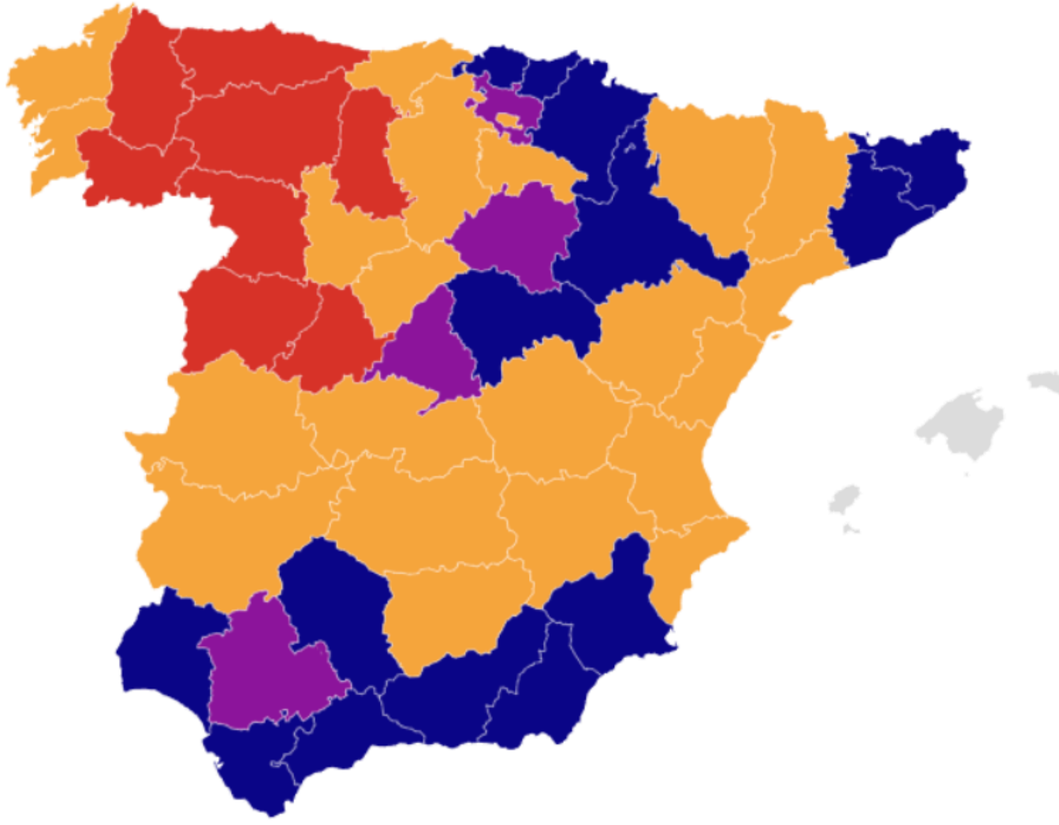
METHOD: WARD

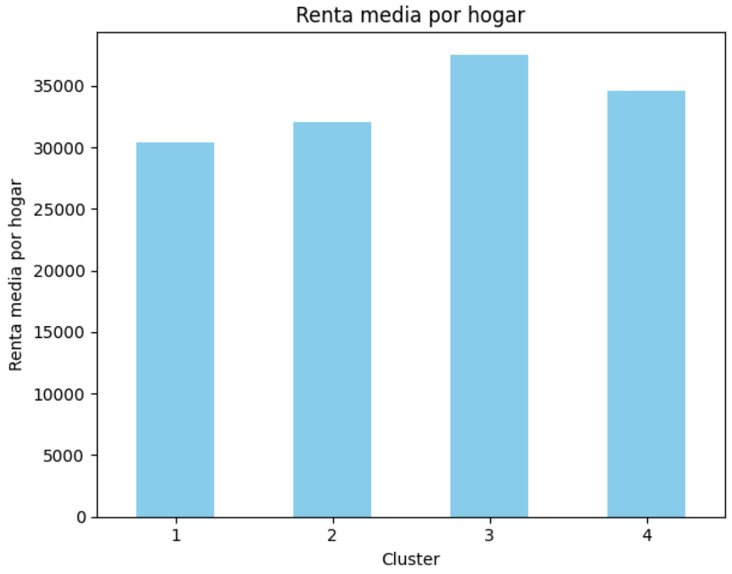
SILHOUETTE COEFFICIENT: 0.254

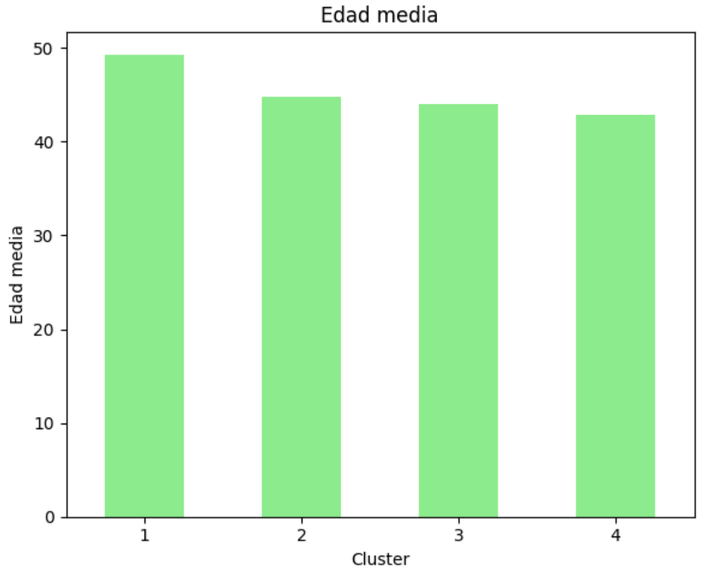


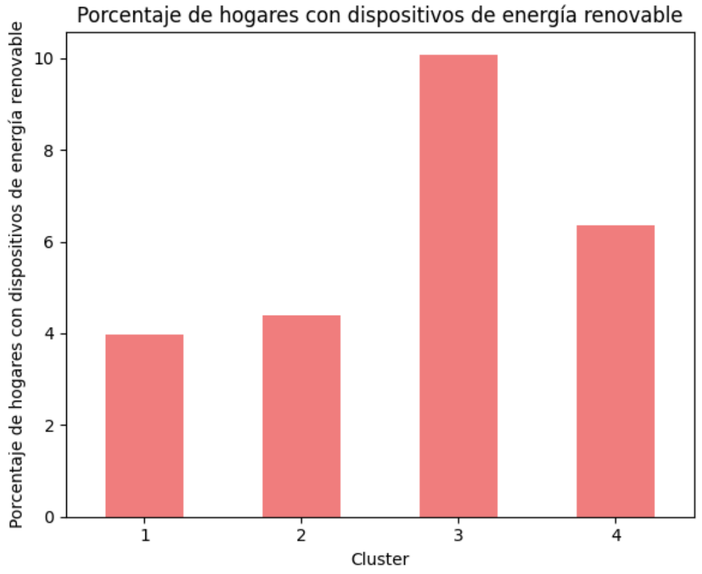
Resultados


12
10
8
6
4
40
42
44
46
48
50
Edad media
45k
40k
35k
30k
Renta media por hogar




Renta media por hogar
Renta media por hogar
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
Cluster


Edad media
Edad media
50
40
30
20
10
0
1
2
3
4
Cluster


Porcentaje de hogares con dispositivos de energía renovable
Porcentaje de hogares con dispositivos de energía renovable
10
8
6
4
2
0
1
2
3
4
Cluster

¡HIPÓTESIS 1 VALIDADA!

Hipótesis 2:

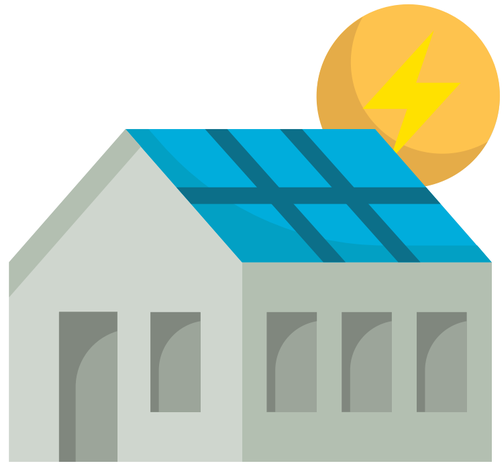
EXISTE UNA RELACIÓN ENTRE EL PORCENTAJE DE VIVIENDAS DE BAJO, MEDIO Y ALTO CONSUMO DE UNA PROVINCIA Y SU TENDENCIA A ADOPTAR DISPOSITIVOS RENOVABLES

¿POR QUÉ PENSAMOS QUE ESTO ES ASÍ?

- Las viviendas de bajo consumo o esporádicas suelen ser viviendas secundarias

- Propietarios de viviendas de alto consumo suelen tener mayor poder adquisitivo

- Las viviendas de medio consumo suponen un equilibrio coste-beneficio



Tarjeta de datos

Variables	Tipo de dato
Provincias	String
Índice de viviendas renovables	Float
Índice de viviendas de bajo consumo	Float
Índice de viviendas de medio consumo	Float
Índice de viviendas de alto consumo	Float

ESTUDIO DE CORRELACIÓN

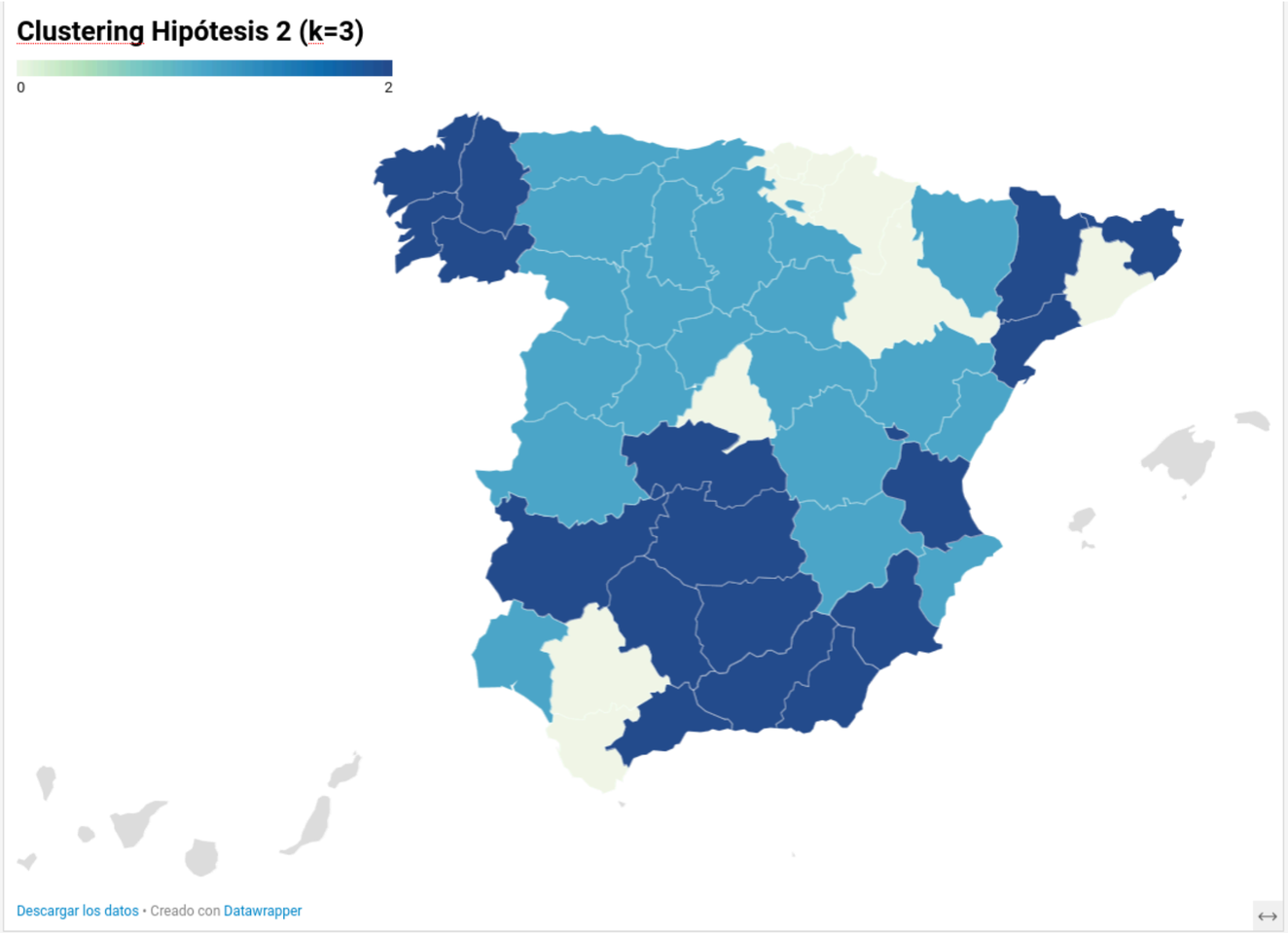
RELACIÓN POSITIVA MODERADA ENTRE EL ÍNDICE DE VIVIENDAS RENOVABLES Y EL DE VIVIENDAS DE MEDIO CONSUMO

Variable 1	Variable 2	Correlación
Índice de viviendas renovables	Índice de viviendas de bajo consumo	-0.19
Índice de viviendas renovables	Índice de viviendas de medio consumo	0.4
Índice de viviendas renovables	Índice de viviendas de alto consumo	-0.029

K-MEANS

- Se ha utilizado el elbow method y número de clusters óptimo es 3

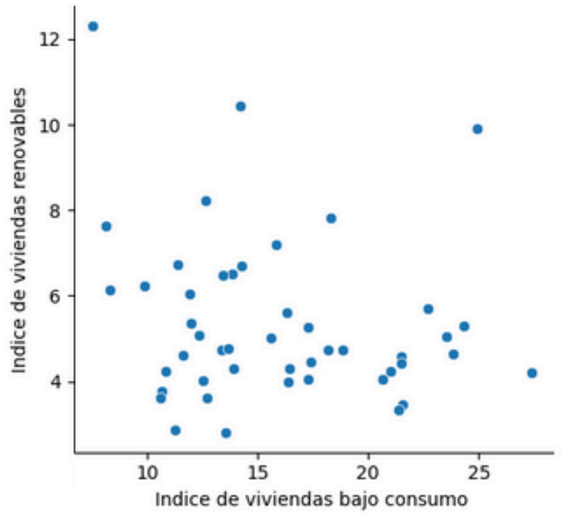
¿HAN SIDO LOS CLÚSTERES FROMADOS CON SENTIDO O HAN SIDO FORZADOS?

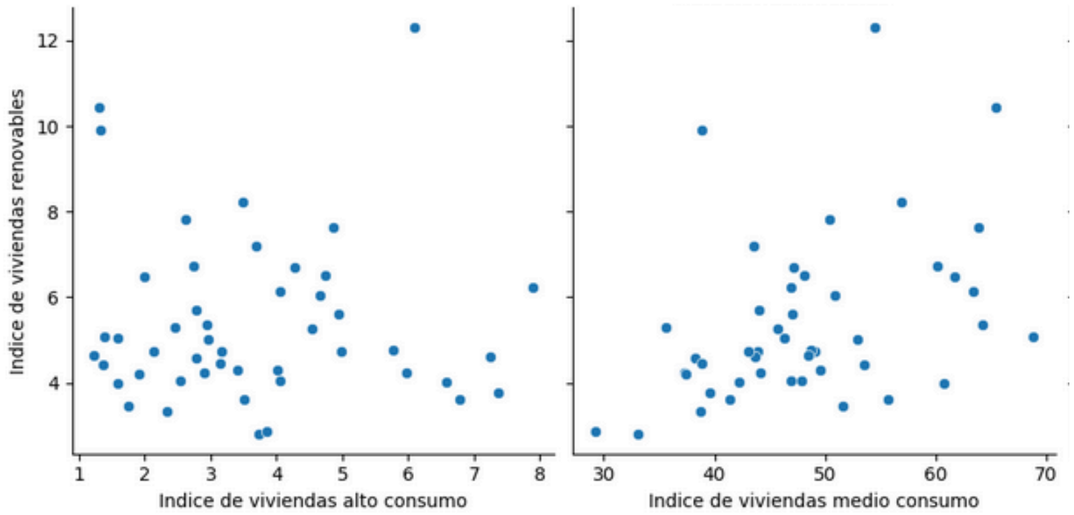


REGRESIÓN LINEAL SIMPLE

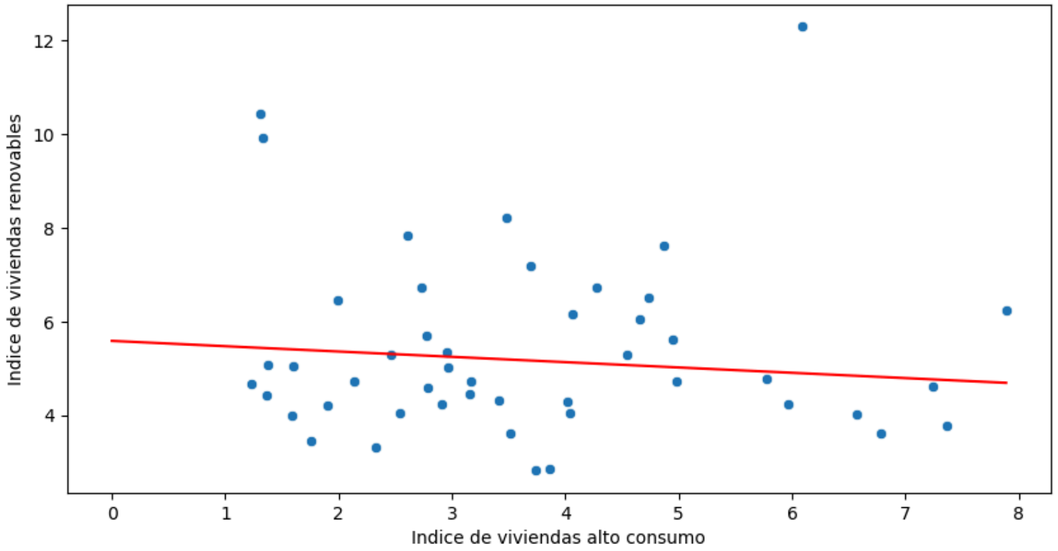
- Un modelo para cada *feature*

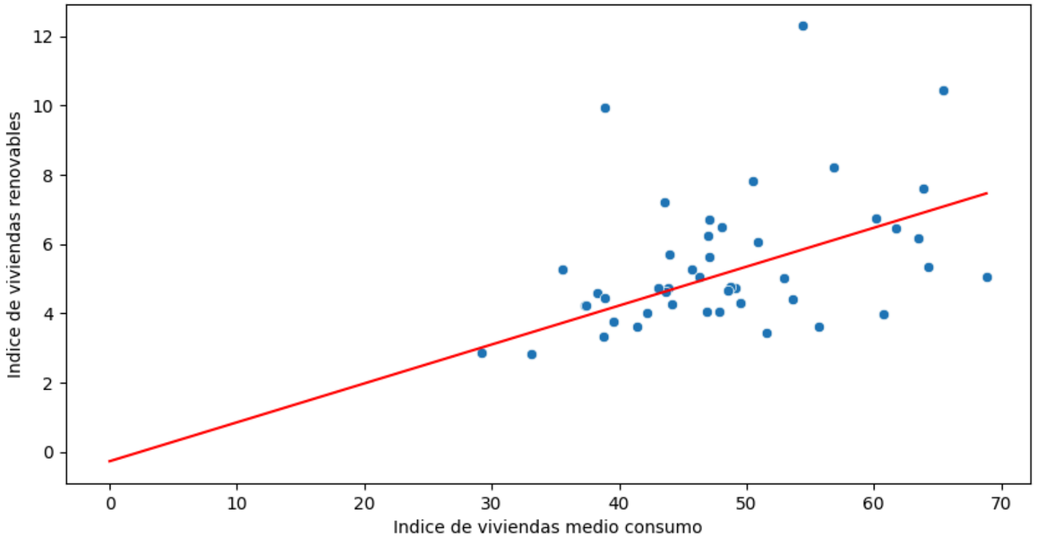
- Se reducen los sesgos en las predicciones debidos a la multicolinealidad entre variables altamente correlacionadas

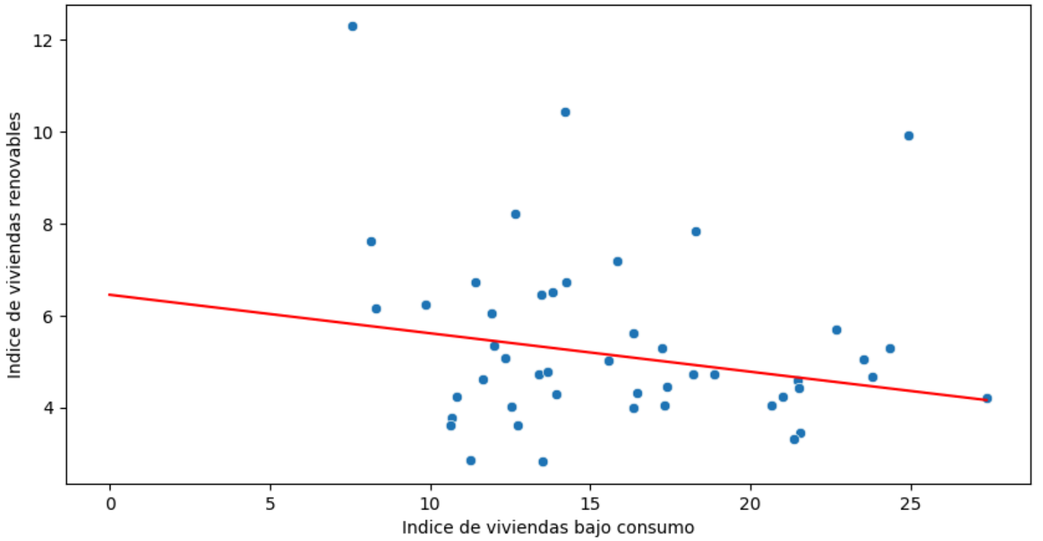




RESULTADOS


Indice de viviendas renovables
Indice de viviendas alto consumo


Indice de viviendas renovables
Indice de viviendas medio consumo


Indice de viviendas renovables
Indice de viviendas bajo consumo

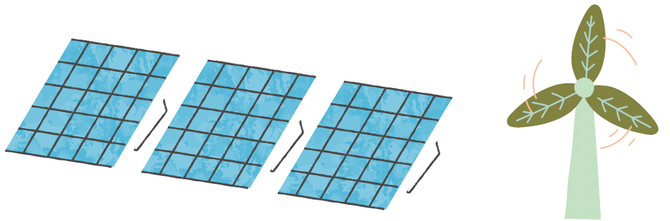
¡HIPÓTESIS 2 VALIDADA!

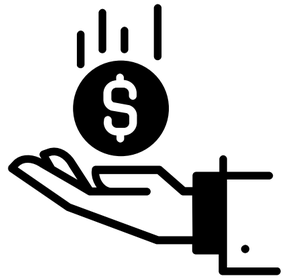
Hipótesis 3:

EN UN NÚCLEO FAMILIAR COMPUESTOS POR MÁS MIEMBROS ES MÁS PROBABLE QUE SE INVIERTA EN ENERGÍA RENOVABLE

¿POR QUÉ PENSAMOS QUE ESTO ES ASÍ?

- Un núcleo familiar compuesto por más miembros tiene un consumo energético más elevado.
- No se depende de la red eléctrica ni de compañías eléctricas
- Incentivos por instalar tecnología renovable





DER -> Dispositivos de Energía Renovable

TARJETA DE DATOS

NOMBRE DEL CAMPO		TIPO DE DATO
Provincias		String
Índice DER		Float
Tipos de núcleos familiares	• Familia monoparental con 0 hijos • ... (9 tipos) • Pareja casada con 2 hijos o más	Float

DER -> Dispositivos de Energía Renovable

ESTUDIO DE CORRELACIÓN

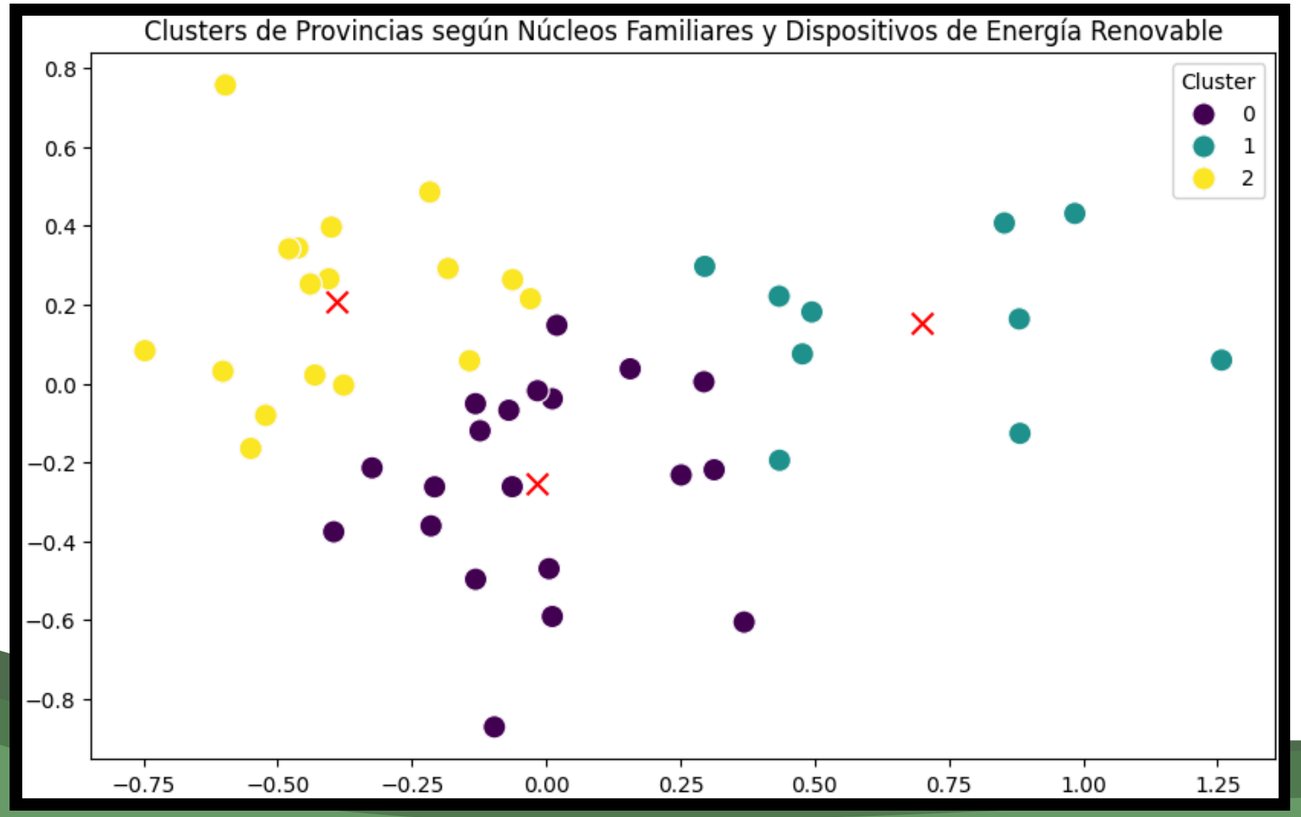
Variable 1	Variable 2	Correlación
ÍNDICE DER	Familia monoparental con 0 hijos	-0,464
ÍNDICE DER	Pareja casada con 0 hijos	-0,534
ÍNDICE DER	Pareja casada con 2 hijos o más	0,366
ÍNDICE DER	Pareja no casada con 2 hijos o omás	0,548

Correlación negativa

Correlación positiva

K-MEANS

- **Análisis de Componentes Principales (componentes = 2).**
- **¿Mejor número de clústeres? Elbow Method ➡ Codo = 3**



SILHOUETTE COEFFICIENT: 0,366

¿HAN SIDO LOS CLÚSTERES FROMADOS CON SENTIDO O HAN SIDO FORZADOS?

ANÁLISIS INTRACLÚSTER

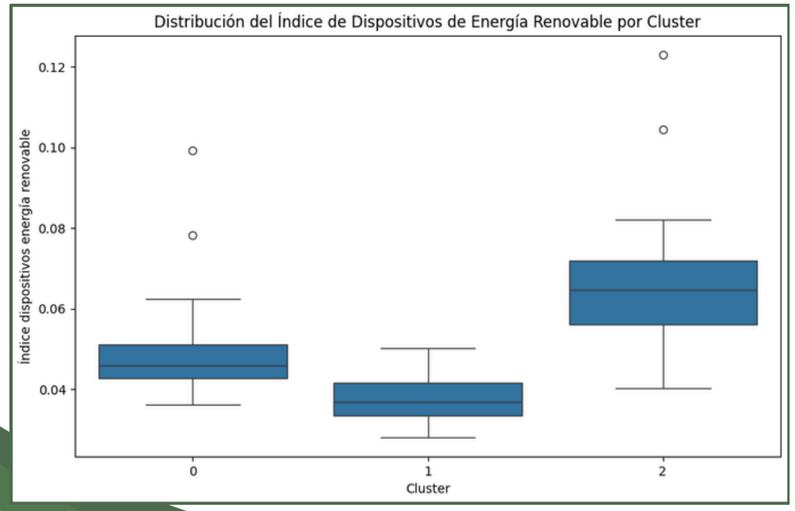
- Valores medios por clúster

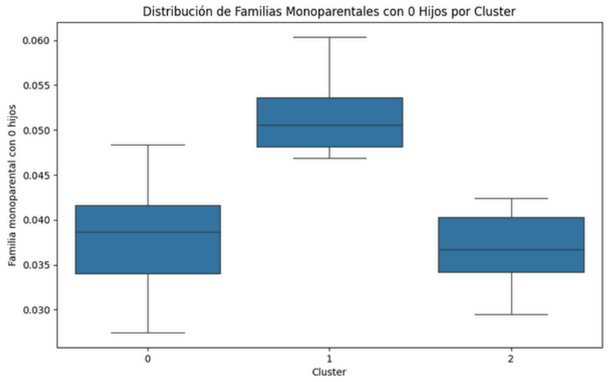
	Clúster 0	Clúster 1	Clúster 2
Índice DER	5,04%	3,79%	6,74%
Familia monoparental con 0 hijos	3,8%	5,17%	3,7%
Familia monoparental con 2 hijos o más	3,86%	3,46%	4,55%
Pareja no casada con 2 hijos o más	2,53%	1,95%	2,89%

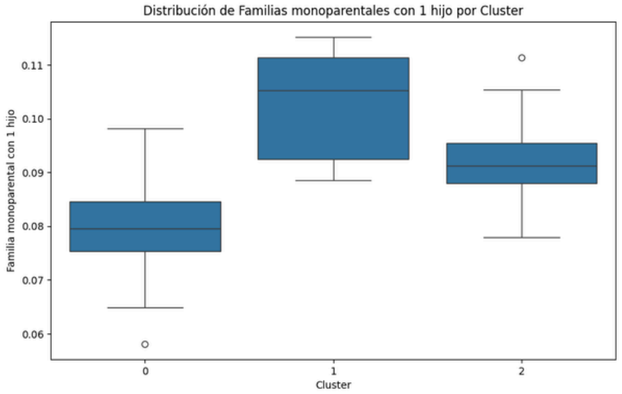
ANÁLISIS INTERCLÚSTER

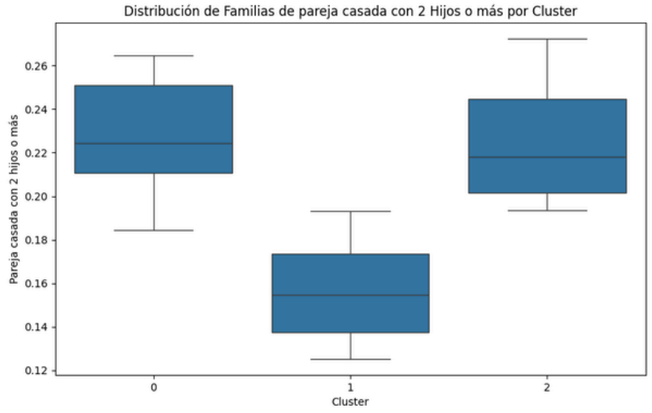
- **Boxplots**

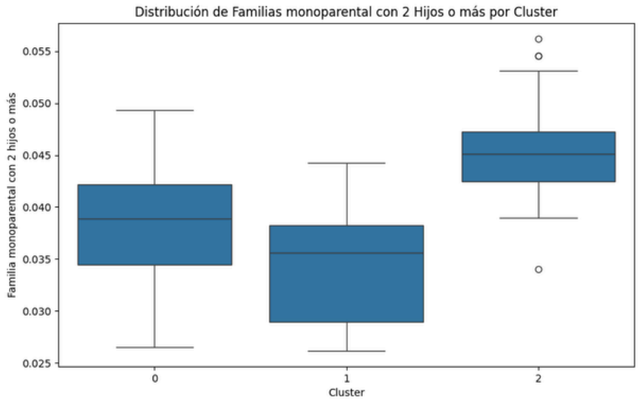
  - **Outliers del clúster 0**: **Soria y Huelva**.
  - Soria tiene un alto porcentaje familias monoparentales con 0 hijos.
  - Huelva es de las provincias del clúster 0 con mayor porcentaje de familias compuestas por pocos miembros.





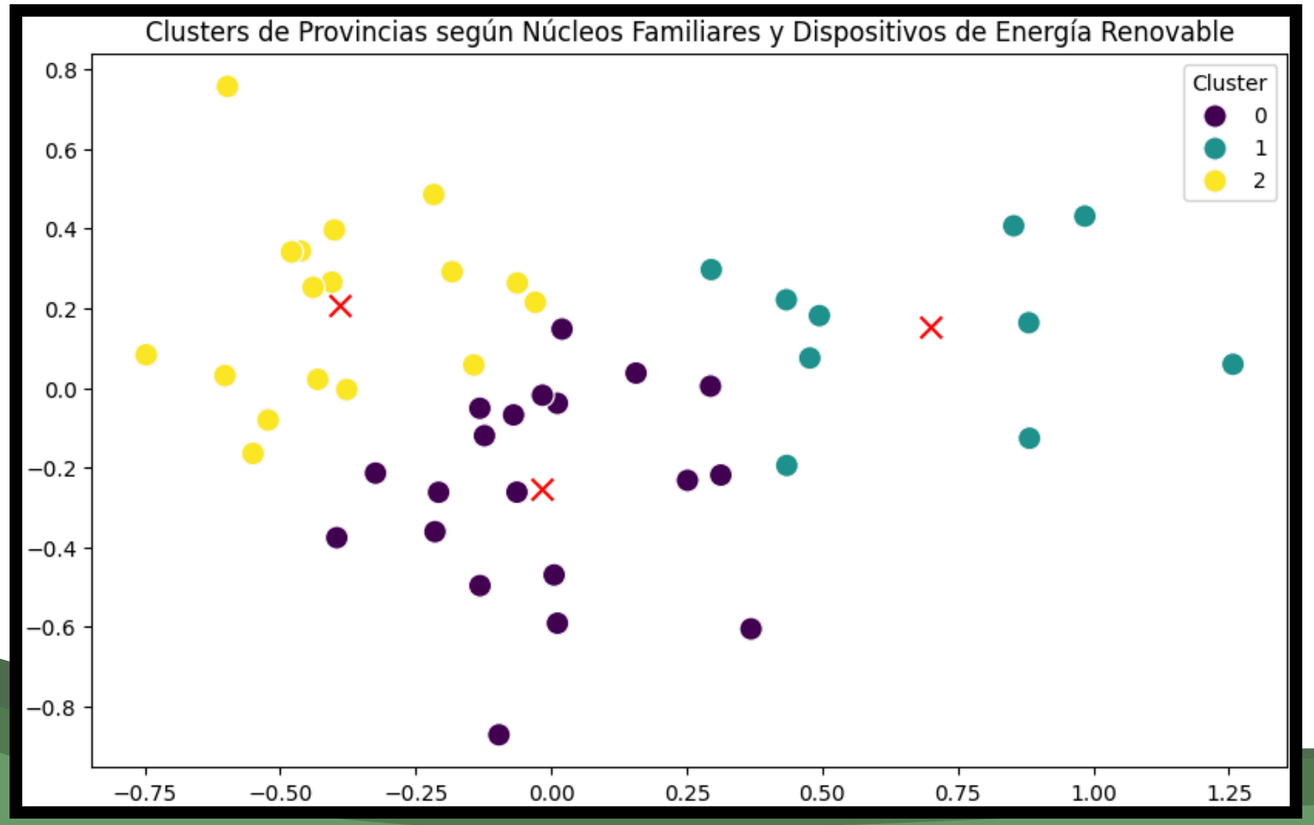






ANÁLISIS INTERCLÚSTER

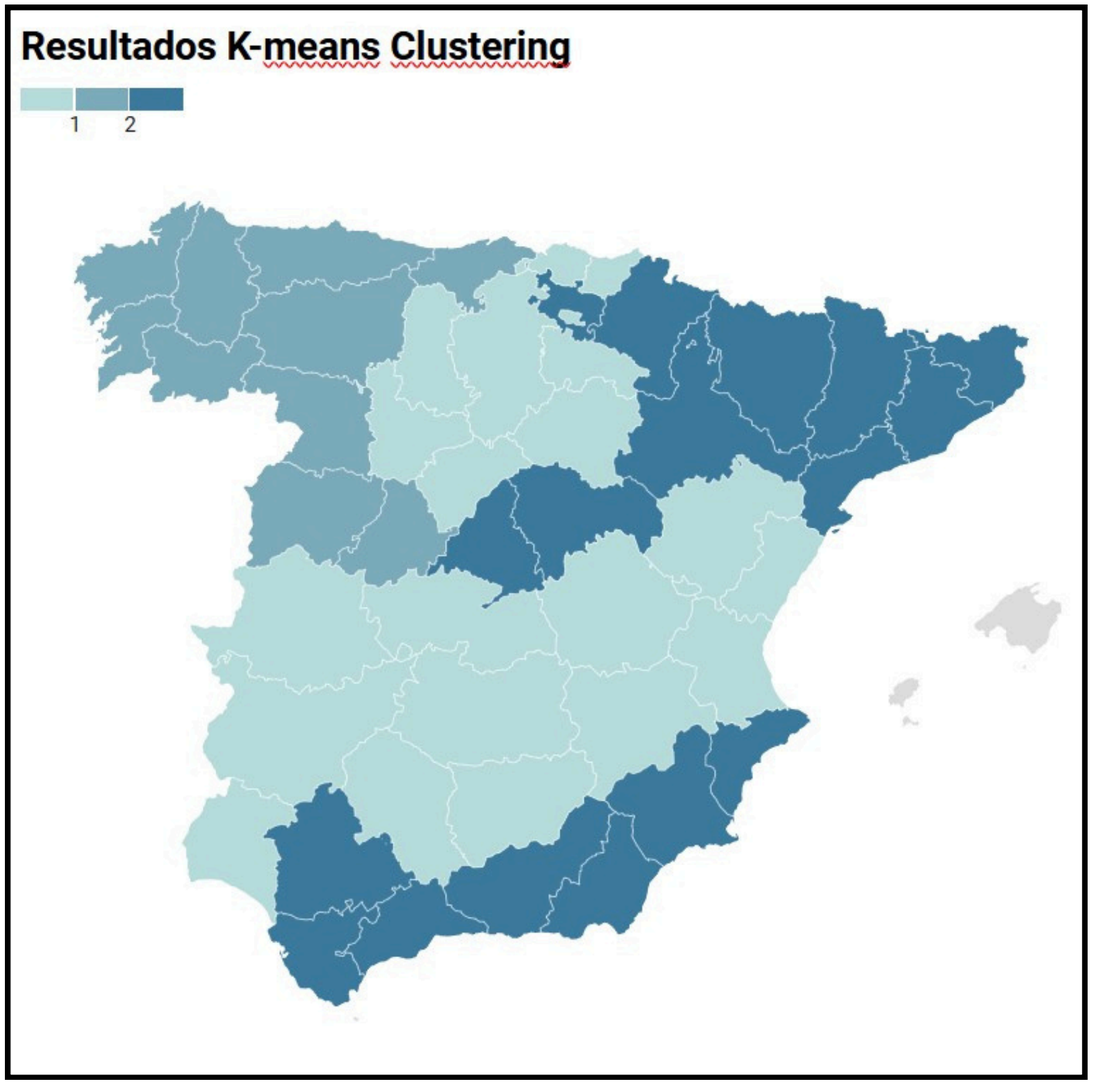
- **Análisis de Componentes Principales (componentes = 2).**
- **¿Mejor número de clústeres? Elbow Method ➡ Codo = 3**



SILHOUETTE COEFFICIENT: 0,366

¿HAN SIDO LOS CLÚSTERES FROMADOS CON SENTIDO O HAN SIDO FORZADOS?

¡HIPÓTESIS 3 VALIDADA!



Hipótesis 4:

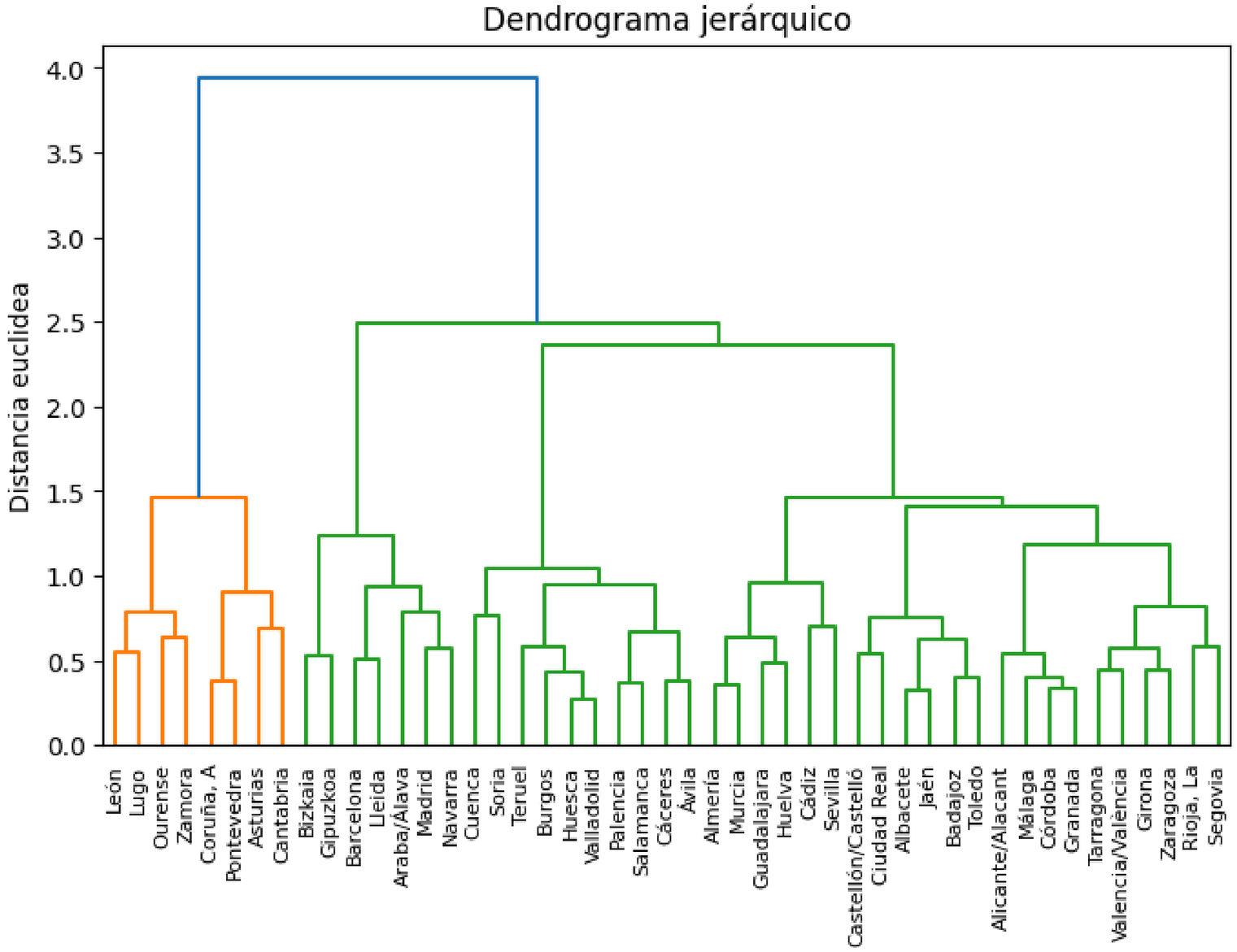
LAS PROVINCIAS AL SUR DE MADRID TIENDEN A UTILIZAR MENOS ENERGÍAS RENOVABLES

Tarjeta de datos

Variables
% Dispositivos
Renta
Edad media
Producción media
Nº familias 1 padre - 0 hijos
Nº familias 1 padre - 1 hijo
Nº familias 1 padre - 2 hijos
Nº familias 2 padres - 0 hijos
Nº familias 2 padres - 1 hijo
Nº familias 2 padres - 2 hijos

¿POR QUÉ PENSAMOS QUE ESTO ES ASÍ?

- El clima del norte es más frío.
- En el norte hay más desarrollo industrial lo que puede aumentar la renta.
- ¿Influyen las tradiciones y el tipo de familias?
- Estudiar todas las variables de nuestros datasets.

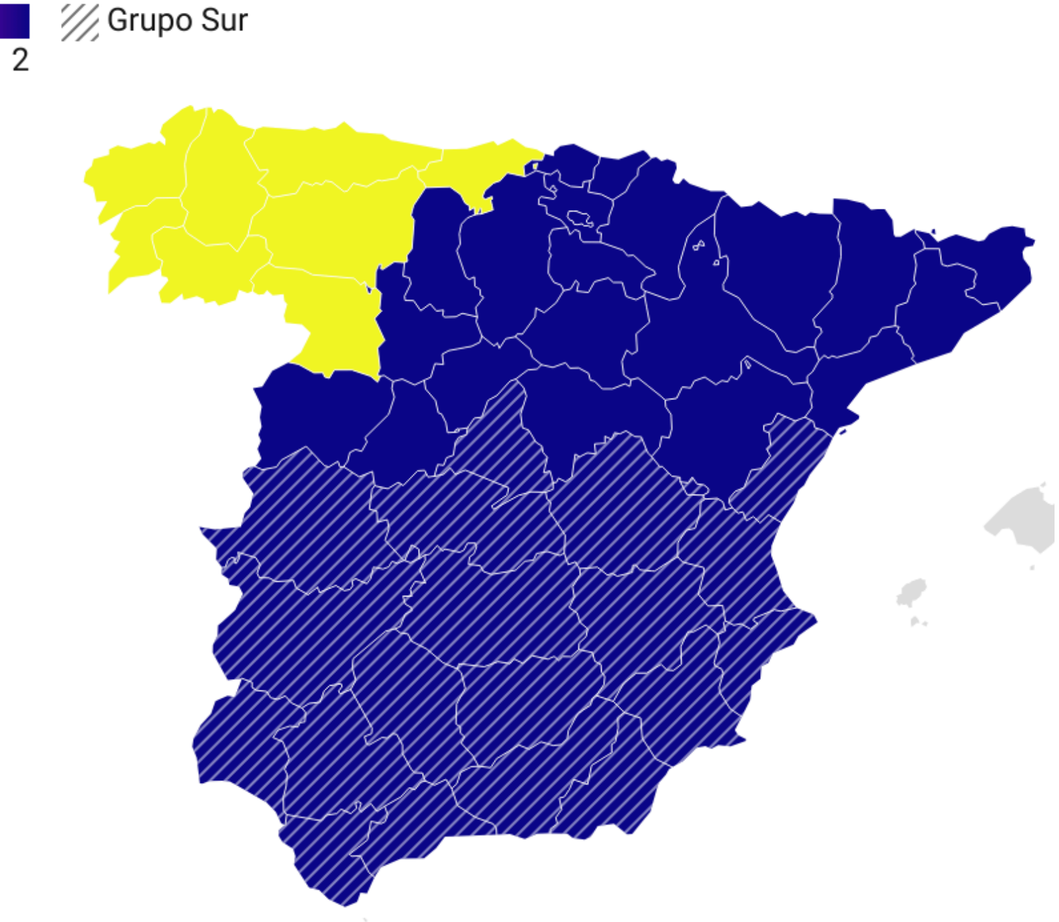


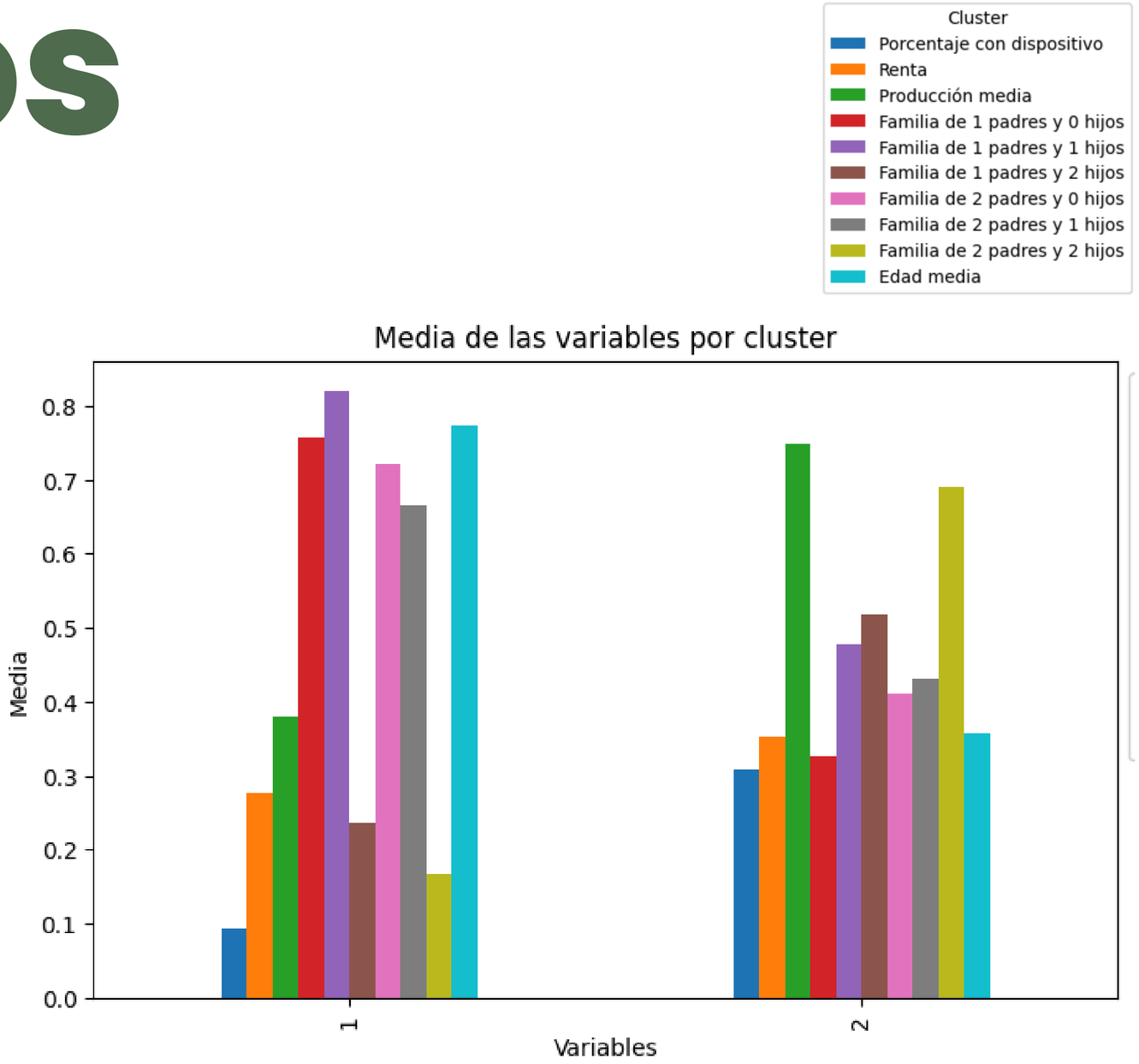
CLUSTERING JERÁRQUICO

Método: WARD

Coeficiente Silhouette: 0.329

Resultados


Grupo Sur
2


Cluster
Porcentaje con dispositivo
Renta
Producción media
Familia de 1 padres y 0 hijos
Familia de 1 padres y 1 hijos
Familia de 1 padres y 2 hijos
Familia de 2 padres y 0 hijos
Familia de 2 padres y 1 hijos
Familia de 2 padres y 2 hijos
Edad media
Media de las variables por cluster
Media
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
Variables

¡HIPÓTESIS 4 NO VALIDADA!

Hipótesis 5:

¿QUÉ CANTIDAD DE PLACAS SOLARES FOTOVOLTAICAS SE NECESITARÍAN INSTALAR PARA ABASTECER EL CONSUMO ELÉCTRICO DE LOS HOGARES DE CASTILLA-LA MANCHA EN EL AÑO 2028?

¿POR QUÉ NOS INTERESA SABER ESTO?

- El aumento del precio energético provoca una búsqueda de alternativas para abastecer esa necesidad .

- La instalación de placas puede ser un mercado en auge y una buena idea de negocio.

TARJETA DE DATOS

NOMBRE DEL CAMPO	TIPO DE DATO	DESCRIPCIÓN
Provincias	String	Las 5 provincias
Potencia MWh	Float	Potencia eléctrica anual de 1m² de placa fotovoltaica
X	Float	Consumo eléctrico en MWh para el año X (de 2014 a 2023)

SELECCIÓN DEL MODELO DE PREDICCIÓN

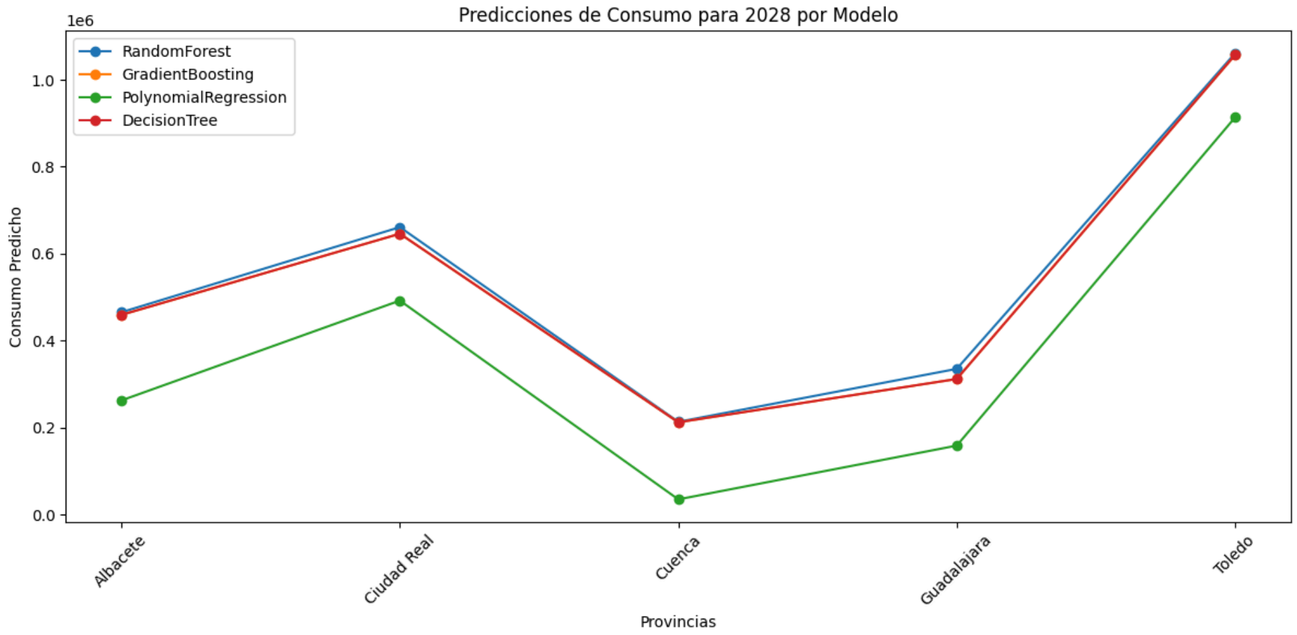
Modelos evaluados:

- RandomForest
- GradientBoosting
- Regresión Polinomial
- Decision Tree

Criterios de selección:

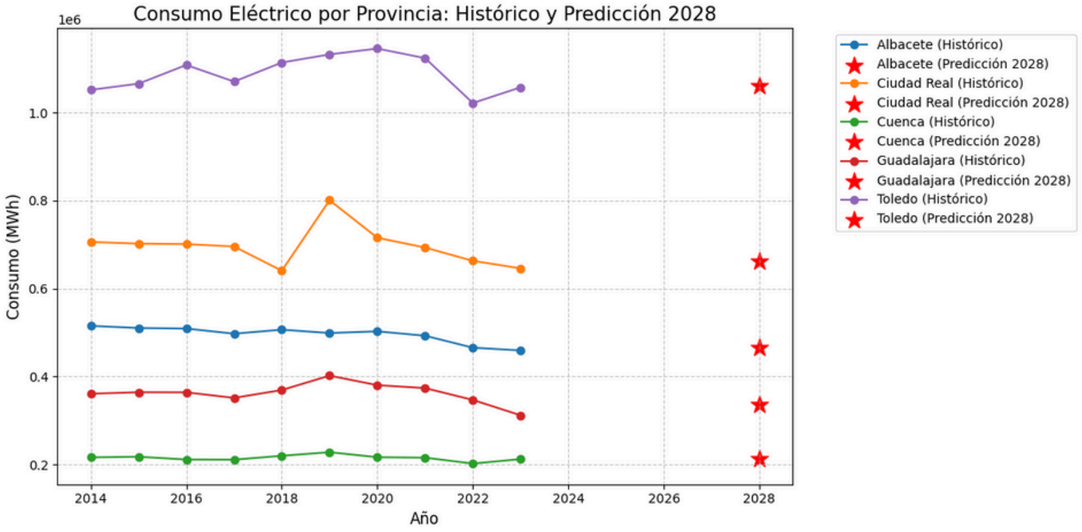
- Coeficiente de determinación (R^2)
- Error cuadrático medio (MSE)

Seleccionado: RandomForest
R^2: 0.982



Resultados Predicción 2028

- ★ **Toledo:** 3.260.288 placas
- ★ **Ciudad Real:** 2.035.302 placas
- ★ **Albacete:** 1.420.673 placas
- ★ **Guadalajara:** 1.076.719 placas
- ★ **Cuenca:** 664.122 placas



Para ser sostenible con energía solar Castilla La Mancha necesitaría 8.457.102 m^2 de placas fotovoltaicas

¡GRACIAS!

¿ALGUNA PREGUNTA?