**Redes eléctricas y balances energéticos**

**Proyectos II, integración y preparación de datos**

# 

# 

Álvaro Prado Expósito

Óscar García Martínez

Víctor Mañez Poveda

Jorge Muedra Vela

Aleixandre Tarrasó Sorní

**Índice**

* **1. Alcance del Proyecto**

- 1.1 Objetivos del Proyecto

- 1.2 Utilidad del Estudio

* **2. Configuración del Proyecto**

- 2.1 Fuentes de Datos

- 2.2 Integración y Transformación de los Datos

* **3. Resultados Obtenidos**

- 3.1 Clasificación de Países de la Unión Europea

- 3.2 Clustering de Comunidades Autónomas de España

- 3.3 Visualización de Datos Económicos y Energéticos

- 3.4 Regresión Múltiple para Predicción de Precios Energéticos

- 3.5 Evolución del Uso de Fuentes de Energía en Comunidades Autónomas

* **4. Lecciones aprendidas para mis futuros proyectos en Ciencia de Datos**
* **5. Reflexiones sobre la Privacidad y Legalidad de los Datos**
* **6. Anexos**
* **7. Lista de técnicas que se deberán explicar en detalle si se hubieran utilizado**

# 1. Alcance del proyecto

El objetivo principal de este proyecto es analizar y comprender la naturaleza del funcionamiento de los sistemas energéticos; comprendiendo cómo se relaciona con la situación económica, estudiando las diferencias entre distintas regiones y utilizando este conocimiento para realizar modelos predictivos.

**1.1.** **Objetivos del proyecto**

### 1. Categorización de Países de la Unión Europea

Este objetivo se centra en categorizar algunos de los países de la Unión Europea según los datos de sus balances energéticos anuales de 1990 a 2020, proporcionados por Eurostat. Se elaborará un diagrama de Sankey para visualizar la distribución de la energía en cada país, distinguiendo entre energía propia, importada, exportada y consumida internamente, así como las fuentes de energía utilizadas. Posteriormente se analizará la variabilidad de estas variables en el tiempo mediante gráficos de tendencias, clasificando a los países según estas tendencias. La meta es obtener conclusiones fiables sobre la situación energética de cada país y las razones detrás de esta.

### 2. Clustering de Comunidades Autónomas de España

Este objetivo consiste en realizar un análisis de clustering sobre las Comunidades Autónomas (CCAA) de España utilizando diversos datos energéticos, específicamente relacionados con la generación y la demanda de energía. Aunque los datos originales tienen una granularidad mensual, se agruparán de manera trimestral para realizar el análisis. El propósito es identificar patrones similares entre las CCAA y determinar cuáles tienen una generación de energía superior a su demanda y cuáles necesitan apoyo energético de otras regiones. Para llevar a cabo este análisis, se emplearán diversas herramientas y técnicas, incluyendo distancias euclídeas, distancias de Manhattan y gráficos de Scores. Estas técnicas nos permitirán comprender mejor los comportamientos energéticos de las diferentes CCAA y obtener información valiosa sobre su interdependencia energética.

### 3. Visualización de Datos Económicos y Energéticos

Para apoyar el análisis de clustering, generamos una visualización mediante un mapa de España por CCAA, coloreadas según el indicador que se quiera seleccionar, a elegir entre Empleo Industrial, IPI, IPC per Cápita, además de Generación y Demanda energética, así como la diferencia entre estas dos últimas. Esta visualización nos permite comprobar si las CCAA con mayor actividad económica tienden a tener un déficit energético y qué regiones son las que suplen este déficit. La visualización es interactiva, permitiendo elegir entre diferentes indicadores que hemos descrito antes, de modo que permite ver la evolución y la posible estacionalidad de las Comunidades Autónomas al igual que identificar patrones como son los de generación y demanda.

### 4. Regresión Múltiple para Predicción de Precios Energéticos

Se desarrollará un modelo de regresión múltiple para predecir el precio de la energía en España del día siguiente, utilizando datos diarios desde el 01-01-2011 al 20-03-2019. Las variables predictoras incluirán la demanda energética y la generación por fuente de energía, entre otras. El modelo se evaluará comparando los resultados tanto con los datos inmediatamente posteriores al entrenamiento como con un período de validación reciente (01-01-2024 al 31-03-2024), el cual no ha resultado significativo. El objetivo es minimizar el error de predicción y determinar cuáles estructuras de generación energética influyen más en la variación del precio en España.

### 5. Evolución del Uso de Fuentes de Energía en Comunidades Autónomas

Este objetivo se propone estudiar la evolución del uso de cada fuente de energía a lo largo del tiempo en cada Comunidad Autónoma de España. Se realizarán gráficos de tendencias por fuente de energía y se explicarán estas tendencias en el contexto de los eventos y políticas energéticas de la última década. El análisis permitirá entender cómo han cambiado los patrones de consumo y generación de energía en las diferentes regiones del país.

Estos objetivos combinan técnicas de visualización, análisis estadístico y modelado predictivo para proporcionar una comprensión integral de la situación energética y económica en Europa y España, con un enfoque especial en las tendencias y patrones que afectan a cada región.

**1.2. Utilidad del estudio**

Este estudio puede resultar de interés para los gobiernos de los países analizados ya que abarca distintos aspectos del funcionamiento de los sistemas energéticos, con análisis que ayudan a comprender la situación de los mismos.

Las compañías eléctricas también pueden beneficiarse de nuestro estudio, usando nuestras técnicas de visualización y análisis como el primer paso para estudiar la situación energética de un país.

Y por último, también interesaría a las agencias de medioambiente, ONG de renovables, y empresas petrolíferas, ya que pueden ver la evolución durante los años de los tipos de energía más utilizados, los cambios entre países, y hacer posibles predicciones sobre estos.

El interés que tiene este estudio respecto a otros estudios similares es que no solo analiza la estructura energética anualmente, sino que aumenta el nivel de granularidad para conocer cómo afecta la estacionalidad a los sistemas energéticos.

# 2. Configuración del proyecto

**2.1. Fuentes de datos**

Se han recopilado datos de múltiples fuentes, incluyendo redes eléctricas nacionales e institutos de estadística, para obtener información relevante sobre consumo energético, producción y otros indicadores económicos. Las fuentes se integraron y transformaron para garantizar coherencia y calidad en los datos.

Se encuentra detallado en el Hito 1 [[Anexo 0.1](#bookmark=id.1t17hyeeivwm)]

**2.2. Integración y transformación de los datos**

Se llevaron a cabo actividades exhaustivas de integración y transformación de datos utilizando principalmente herramientas como Python, Excel y, sobre todo, R. Estos archivos de origen, a menudo extensos y complejos, fueron procesados para destilar información relevante y concisa. Este proceso, aunque tedioso, fue crucial para preparar los datos para análisis posteriores. Consulte el Anexo 2 para obtener más detalles sobre estas actividades.[[Anexo 6.1](#bookmark=id.fvcba794clqu)]

Se encuentra detallado en los R-Markdowns, Notebooks y archivos correspondientes como el Hito 2 [[Anexo 0.2](#bookmark=id.df7bkwjr066)] [[Anexo 6.1](#bookmark=id.fvcba794clqu)] [[Anexo 6.2](#bookmark=id.d62c0pelrm7i)] [[Anexo 6.3](#bookmark=id.5oy8h3bf3p8)]

# 3. Resultados obtenidos

**3.1. Categorización de los balances energéticos europeos**

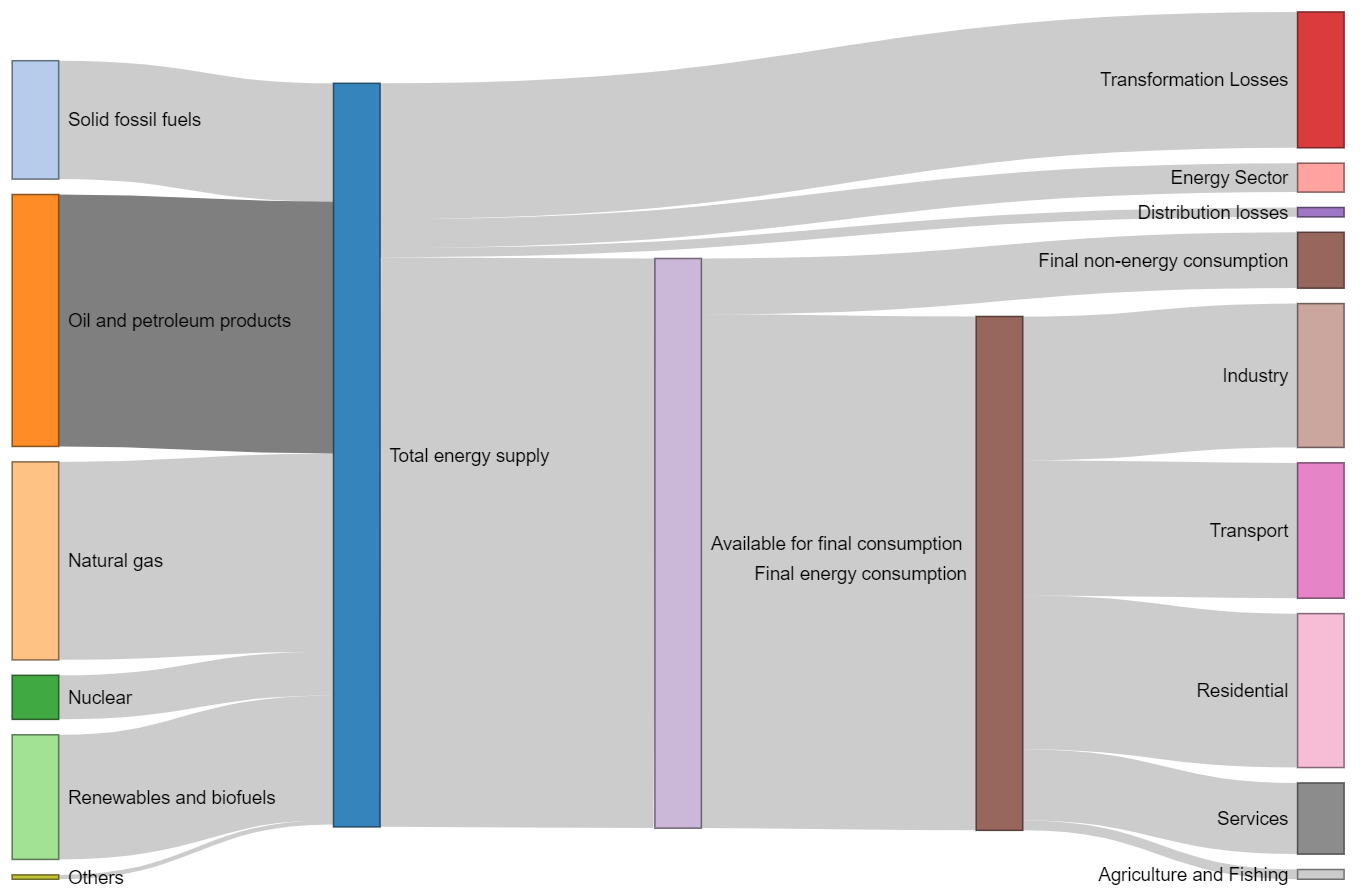
En el análisis exploratorio mostrado en el Hito 2 [[Anexo 0.2](#bookmark=id.df7bkwjr066)] queda comprobado que los datos de los balances energéticos europeos toman valores adecuados para un análisis categórico; puesto que la evolución de las variables y las diferencias entre países son suficientes cómo para que este objetivo tenga interés.

La meta de este objetivo es conseguir un informe en R Markdown donde se realice un análisis cualitativo de los balances energéticos de varios países europeos sirviéndose de varias herramientas de visualización. Para ello se utiliza el diagrama de Sankey, un gráfico que muestra el flujo transmitido entre nodos. En este estudio el diagrama de Sankey ha de mostrar cómo se distribuye en cada país la oferta energética, y de qué fuentes de energía procede. Además, para ciertas variables estudiadas se crean gráficos de líneas que muestran su evolución.

Puesto que los diagramas de Sankey no muestran evolución, y agrupando los datos de 1990 a 2020 o mostrando únicamente un año se pierde demasiada información, decidí crear una aplicación en shiny donde se pueda elegir un año y un país y mostrar su diagrama de Sankey. Está aplicación se incrusta en el documento HTML resultante al ejecutar el R Markdown, y se puede utilizar sin problemas.

En el archivo hay dos versiones de esta aplicación: una en la que se definen server y ui de manera que se elija país y año, mostrando el diagrama de Sankey correspondiente; y una función de nombre sankey\_graf a la que se le introduce el código de país y al ejecutarla pide al usuario el año deseado. La primera versión de la aplicación se utiliza en la introducción del informe de modo que el usuario pueda comprender los datos con los que trabajamos y comprobar la distribución energética de cada país. La segunda versión se utiliza en los análisis especializados por países, y da la opción de mostrar los diagramas de Sankey para cada año en el país analizado.

Ambas versiones están basadas en el mismo código, el proceso que sigue es el siguiente. En primer lugar transforma la base de datos eliminando todas las filas (variables) que no vamos a analizar. Posteriormente escoge el año elegido por el usuario y transpone la tabla, de manera que las columnas sean las variables. Esta tabla contiene la información necesaria para generar el diagrama, pero no se encuentra en el formato adecuado. Por lo tanto transformamos la tabla de manera que se especifique qué nodos (variables) están conectados entre sí y cuál es el flujo entre ellos (el menor de los valores de ambas variables). Los diagramas de Sankey han sido creados mediante la función sankeyNetwork de la librería de R networkD3.



**Fig. 1. Gráfico Sankey España 2020**

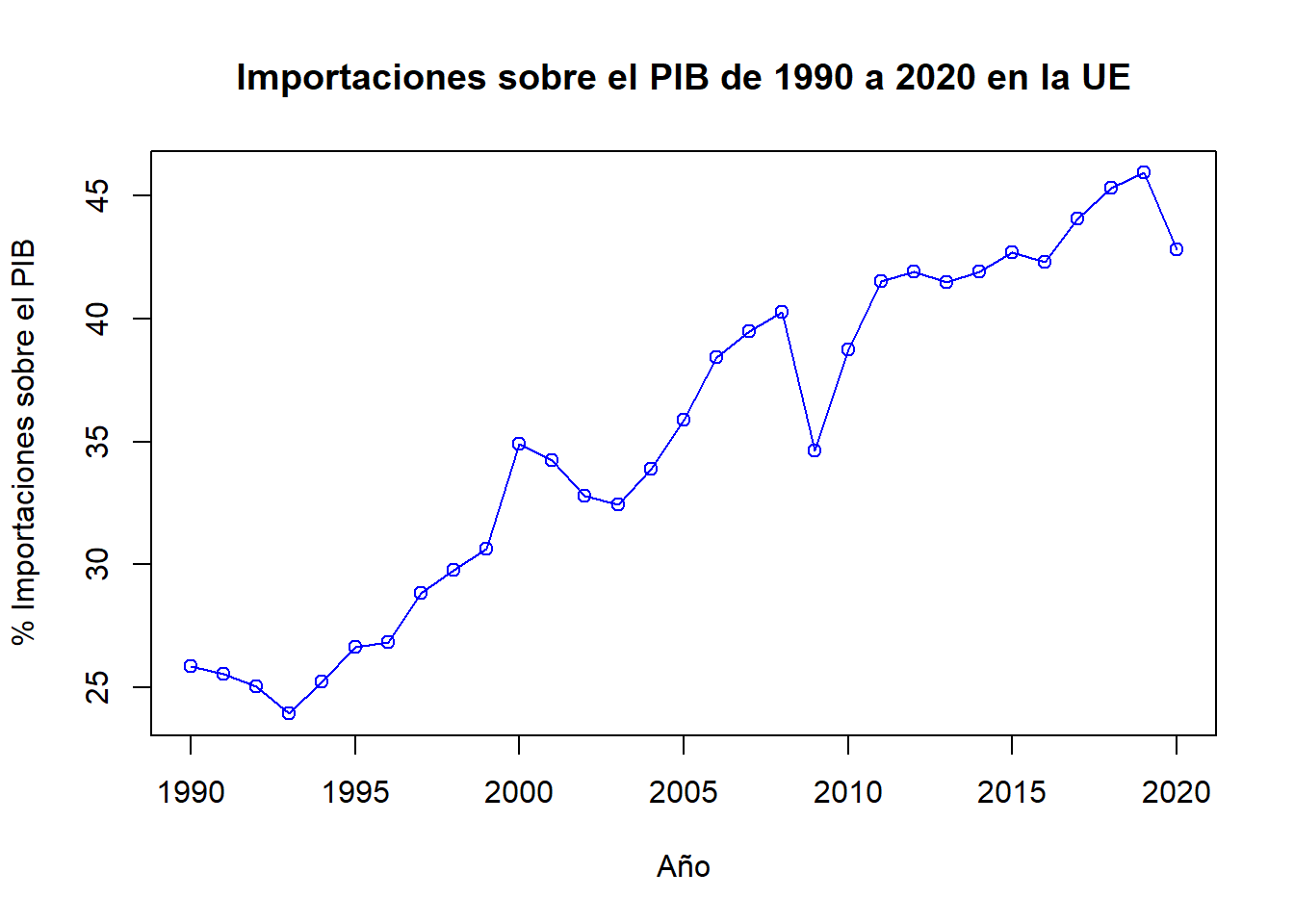
En cuánto a los gráficos de líneas, están pensados para mostrar la evolución que han tenido ciertas variables a lo largo de los treinta años estudiados. Para generarlos correctamente en el documento HTML hemos creado tres funciones, una para cada grupo de variables. La primera función es tend\_fuente que muestra la evolución de las fuentes de energía; la segunda es tend\_perd, que estudia la tendencia de las pérdidas energéticas; y la tercera, tend\_sectores, estudia las tendencias de consumo de los sectores económicos.

El código de las tres es prácticamente idéntico, únicamente diferenciándose en las variables introducidas en la función ggplot de la librería ggplot2. El proceso que siguen se basa en limpiar la base de datos de las variables que no vayan a ser utilizadas, y posteriormente transponer la tabla de manera que las variables se sitúen en las columnas. Una vez así ya se puede usar la función ggplot.

Gracias a estas herramientas de visualización, conseguimos comprender la base de datos con la que trabajamos y realizar un análisis categórico de varios países europeos adecuado. Las conclusiones obtenidas del estudio son las siguientes.

**Conclusiones**

En todos los países estudiados apreciamos un descenso del consumo en el sector industrial. Para distinguir si este descenso es atribuible a una desindustrialización generalizada en el continente o si es fruto de una mayor eficiencia energética hemos decidido estudiar las importaciones en los años estudiados. Si las importaciones descienden significa que la productividad del país aumenta y la economía tiene menor necesidad de comprar productos producidos fuera de sus fronteras. En cambio, una subida de las importaciones implica menor productividad local y mayor compra de productos extranjeros. Visualizamos la tendencia de las visualizaciones mediante un gráfico de líneas.



**Fig. 2. Gráfico importaciones sobre PIB 1990-2020 UE**

El incremento de las importaciones es evidente, manifestando el descenso de la productividad europea y por ende una desindustrialización en el continente.En los análisis de cada país esta tendencia se confirma, siendo una característica común de todas las economías europeas.

En todos los países estudiados las pérdidas energéticas han disminuido su valor a lo largo de los últimos treinta años. Podemos entender este hecho como consecuencia de una mejora de la eficiencia de las infraestructuras energéticas europeas.

La evolución de la distribución energética de cada país depende principalmente del contexto geopolítico e histórico de cada país, habiendo países que destacan por estrategias más largoplacistas y otros que carecen de ellas. Por ejemplo, Francia y España han apostado por descarbonizar su oferta energética mediante la producción nuclear y renovable respectivamente, aumentando así su soberanía energética. Mientras tanto, países como Alemania no han sabido gestionar su oferta energética de manera que se optimice su descarbonización y autonomía, deshaciéndose de la nuclear careciendo de una alternativa a su altura.

No obstante, hay tendencias que podemos generalizar a todos los países analizados. En mayor o menor medida las renovables han aumentado su proporción sobre la oferta energética en todos los casos estudiados, debido a sus ventajas ecológicas y estratégicas. También se ha extendido el uso del gas natural en detrimento de los productos petrolíferos y el carbón, gracias a ofrecer menos polución y un suministro más fiable. El uso de nuclear se suele mantener constante en el tiempo, habiendo importantes diferencias entre países respecto a su uso.

Los resultados se encuentran en [[Anexo 1.1](#bookmark=id.g7dzqhhgqldj)].

**3.2. Clustering de Comunidades Autónomas en España por Indicadores Económicos y Demográficos**

Nuestro proyecto inicialmente tenía como objetivo aplicar la técnica de clustering a todos los datos para entender los patrones que influyen en su comportamiento. En un principio, planeamos extender el análisis a diferentes países y comunidades autónomas, pero tras un análisis más detallado, decidimos concentrarnos exclusivamente en las regiones de España.

Para llevar a cabo este análisis, fue fundamental determinar qué datos utilizamos para agrupar las comunidades. Contábamos con una amplia variedad de indicadores, desde valores económicos hasta datos energéticos. Finalmente, nos decidimos por centrarnos exclusivamente en los datos de generación y demanda de energía. Esta decisión se fundamentó en nuestra intención de identificar patrones específicos en el comportamiento energético de las regiones, lo que nos permitiría evaluar su autosuficiencia y dependencia energética de manera más precisa.

Para iniciar el análisis de datos, fue necesario realizar un tratamiento inicial para facilitar su manejo. Esto implicó cambiar la granularidad de los datos de mensual a trimestral [[Anexo 2.1](#bookmark=id.5moyqxkc8cdy)][[Anexo 2.2](#bookmark=id.nq7t19xnok85)][[Anexo 2.3](#bookmark=id.sg4ouq4cjjc1)], así como la creación de una nueva base de datos para analizar la diferencia entre generación y demanda [[Anexo 2.3](#bookmark=id.sg4ouq4cjjc1)].Posteriormente, comenzamos con el proceso de clustering. Para explicar nuestras acciones y mostrar las conclusiones alcanzadas, primero describiremos los procesos y técnicas utilizadas. Aunque los análisis de clustering siguieron una metodología similar, también explicaremos las diferencias clave entre ellos para resaltar sus particularidades y resultados únicos.

Para poder aplicar el clustering fue necesario realizar un centrado en cada una de las BBDD, ya que aunque estuvieran medidas en las mismas unidades los factores y las variables que afectan a las diferentes comunidades no son los mismos, como por ejemplo la densidad de población, la riqueza, la extensión de territorio entre otros. Tras esto se realizó el cálculo del estadístico de Hopkins, estadístico que ofrece un valor entre 0 y 1, y que nos muestra como de adecuados son los datos para ser agrupados, en nuestro caso los datos oscilaron entre el 0.66 y el 0.95, por lo que en todos los casos eran adecuados para realizar su aplicación. [[Anexo 2.1](#bookmark=id.5moyqxkc8cdy)][[Anexo 2.2](#bookmark=id.nq7t19xnok85)][[Anexo 2.3](#bookmark=id.sg4ouq4cjjc1)]

Dado el éxito de los resultados obtenidos sobre este índice, avanzamos hacia un análisis de clustering más exhaustivo. En un principio, intentamos buscar patrones de similitud utilizando distancias de correlación como la de Pearson o Spearman. Sin embargo, los resultados obtenidos no fueron los esperados, ya que el valor de Silhouette fue negativo, indicando una agrupación incorrecta de los datos. Ante esta situación, decidimos modificar nuestra estrategia inicial y buscar agrupamientos en función de los valores en sí. Es decir, los clusters obtenidos ahora se definirían en base a la magnitud de los valores, en lugar de la correlación entre ellos.

Para poder aplicar estas técnicas hicimos uso de diferentes modelos y métodos, aquí nos referimos a los Modelos jerárquicos, Ward y Medias, y Modelos de partición, K-Means y K-Medoides, de los cuales acabaríamos eligiendo uno tras realizar una selección y validación de los modelos [[Anexo 2.1](#bookmark=id.5moyqxkc8cdy)][[Anexo 2.2](#bookmark=id.nq7t19xnok85)][[Anexo 2.3](#bookmark=id.sg4ouq4cjjc1)]. Para poder realizar todo esto, nos apoyamos en conocimientos obtenidos en otras materias como MDP, donde habíamos hecho uso ya de herramientas que pertenecen a librerías como NBClust, FactoMineR, cluster y factoextra, y que nos ayudaron a realizar el análisis. Las herramientas utilizadas para la selección del número de clusters fueron:

-Gráficos para Silhouette:permite evaluar la coherencia de la agrupación y la separación entre los clusters en función del valor medio de silhouette para cada nº de clusters.

-Gráficos de diferencia media intra-cluster: representan la diferencia media intra-cluster en función del número de clusters, lo que ayuda a determinar el número óptimo de clusters mediante la identificación del "codo" en la curva, indicando el punto de equilibrio entre la compacidad y la separación de los clusters.

-Dendograma: representación visual de la estructura jerárquica de los datos, mostrando cómo se agrupan las observaciones en diferentes niveles de similitud. Es útil para comprender la estructura de los clusters y tomar decisiones sobre la cantidad de clusters a seleccionar.

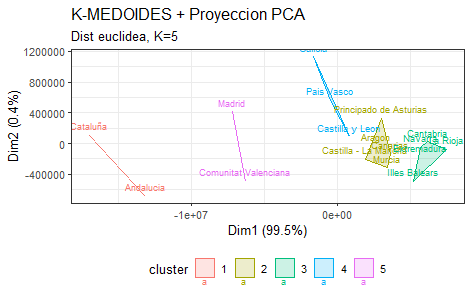
Tras realizar la selección inicial del número de clusters pasamos a representarlos en un gráfico de scores para observar cómo se distribuyen las observaciones y confirmar que se ha realizado correctamente tanto la selección, como la representación gráfica. En este momento habríamos acabado la creación de nuestro clustering para los modelos, pero nos queda la selección del mejor modelo, y esto se realiza a través de representaciones para cada modelo, donde se establece una especie de eje para el valor de silhouette y se representan los diferentes modelos en él, en este se elige el que mejor valor de silhouette y el que menos error tenga.

**Conclusiones**

Tras la explicación de cómo hemos desarrollado el objetivo, pasamos a la muestra de conclusiones:

* El análisis de los datos de demanda utilizando el método de K-medoides reveló una clara diferenciación entre los clusters, lo que sugiere que las comunidades tienen necesidades energéticas distintas y se agrupan en función de estas. Al crear el gráfico de scores para las dimensiones 1 y 2, se observó una diferenciación significativa entre los clusters, especialmente en la dimensión 1, que capturó el 99.5% de la variabilidad. Esta dimensión resultó ser suficiente para representar efectivamente los clusters.

En particular, los clusters ubicados a la izquierda (rojo, morado y azul) corresponden a comunidades autónomas con alta densidad poblacional, una industria robusta y una importante actividad turística. Destaca especialmente Cataluña, cuya demanda energética es significativamente superior debido a su base industrial considerable, que requiere un suministro energético elevado para su funcionamiento adecuado.

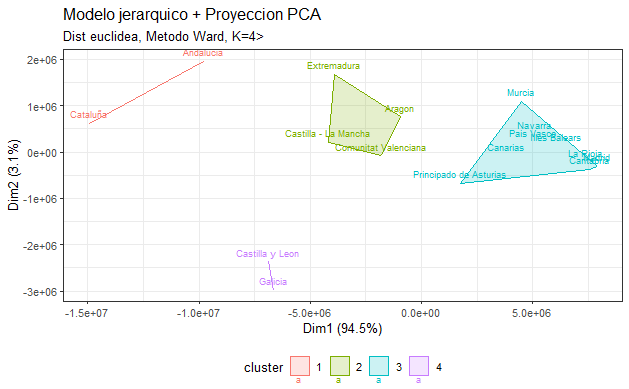


**Fig. 3. Gráfico scores demanda k-Medoides**

* Para la generación obtuvimos este gráfico de Scores haciendo uso del método de Ward.Tras crear el gráfico de scores para las dimensiones 1 y 2, se observa una clara diferenciación entre los clusters, especialmente en la dimensión 1, que abarca el 94.5% de la variabilidad, siendo suficiente para representar efectivamente los clusters.

Al analizar los clústeres, el cluster número 3 destaca por tener un gran número de observaciones concentradas, indicando una producción energética más baja en las CCAA correspondientes. El clúster 2 muestra una producción superior a la media, pero alejada de los valores más altos, lo que puede relacionarse con la extensión del territorio y el tipo de energía predominante. El clúster número 4 incluye observaciones de Castilla y León y Galicia, destacando por su alta producción en energías renovables, como la fotovoltaica, eólica, impulsada por factores geográficos y políticas favorables [[Anexo 3.2](#bookmark=id.31zo8bvqqdzr)][[Anexo 5.1](#bookmark=id.ofcj8ahkw80a)]. Por otra parte, tenemos a Cataluña y Andalucía, líderes en producción, se benefician de una amplia diversidad energética y políticas de incentivo a las energías renovables, satisfaciendo así la demanda tanto local como nacional, además de ello también cuentan con una gran producción no renovables como la nuclear, siendo la más importante en la zona de Cataluña[[Anexo 5.1](#bookmark=id.ofcj8ahkw80a)].

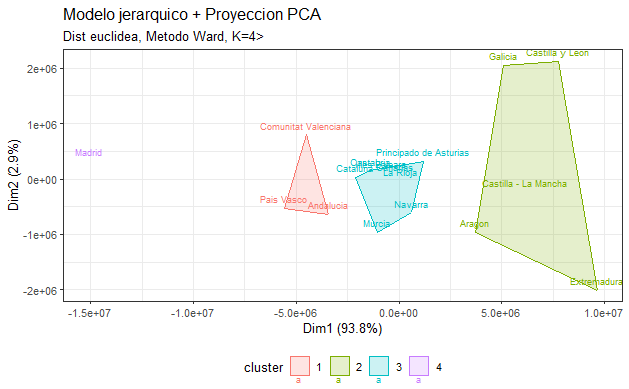
En conclusión, el análisis de clústeres revela patrones distintivos en la producción energética de las regiones españolas, destacando la importancia de factores geográficos, políticos y económicos en la configuración de sus perfiles energéticos. [[Anexo 2.2](#bookmark=id.nq7t19xnok85)]



**Fig. 4. Gráfico scores generación Ward**

* Para la generación-demanda obtuvimos este gráfico de Scores haciendo uso del método de Ward.Tras crear el gráfico de scores para las dimensiones 1 y 2, se observa una clara diferenciación entre los clusters, especialmente en la dimensión 1, que abarca el 93.8% de la variabilidad, siendo suficiente para representar efectivamente los clusters.

Ahora pasaremos al análisis de clústeres, en primer lugar el cluster 3 se destaca por su gran número de observaciones cercanas al valor central, sugiriendo un equilibrio entre la demanda y generación de energía en estas comunidades autónomas. Por otro lado, el cluster 4, que incluye únicamente a Madrid, muestra un valor muy negativo en comparación con el resto, indicando un caso atípico posiblemente debido a su alta demanda de energía y baja capacidad de producción, hecho que se produce por su pequeña extensión y su gran concentración de industria, población y otros motivos. El cluster 2 exhibe un alto valor en la dimensión 1, reflejando una alta producción energética. Sin embargo, estas regiones tienen una densidad de población menor, lo que sugiere una posible sobreproducción de energía en comparación con sus necesidades reales, además esto combinado con los factores favorables a la producción como una gran extensión y cantidad de materias primas y factores que favorecen a que este valor sea mayor a la demanda . En resumen, el análisis de los clusters revela patrones significativos en la relación entre la demanda y la generación de energía en las comunidades autónomas, destacando la importancia de considerar tanto la generación como la demanda de energía al planificar políticas y estrategias energéticas a nivel regional.[[Anexo 2.3](#bookmark=id.sg4ouq4cjjc1)]



**Fig. 5. Gráfico scores generación-demanda Ward**

**3.3. Mapa de España por Índice Económico Mensual y Datos Energéticos de las CCAA**

Para hacer este análisis tuvimos que tratar los datos de distintas fuentes de información y de este modo poder juntar todos los indicadores en un único conjunto de datos. Esto supuso la mayor carga de trabajo porque la estructura de cada archivo era distinta, esto se debe a que para indicadores económicos como son el PIB per cápita, el IPI o el porcentaje de empleo industrial fueron extraídos del INE, mientras que los referidos a la generación y demanda industrial los obtuvimos de la Red Eléctrica Española (REE). Después de trasponer algunos de los ficheros y de realizar el tratamiento pudimos juntarlos todos en un archivo csv llamado *datos\_mapas*. Una vez hecho esto usamos la librería de Shiny para generar la aplicación donde se muestra el mapa.

Mostraremos ahora algunos de los mapas que produce por la aplicación [[Anexo 3.1](#bookmark=id.jvu7dgnduh6i)], todos ellos referidos al último trimestre de 2023:

| Para el IPI: (Incluye sector Servicios) | Para el PIB per cápita: |
| --- | --- |
| **Fig. 6. IPI T42023** | **Fig. 7. PIB per cápita T42023** |
| Para la generación energética: | Para la demanda energética: |
| **Fig. 8 .Generación energética T42023** | **Fig. 9. Demanda energética T42023** |

| Para la diferencia generación-demanda: |
| --- |
|  |

**Fig. 10. Diferencia generación-demanda energética T42023**

**Conclusiones**

Por desgracia no podemos incluir en este informe todos los gráficos que puede generar la aplicación y que nos ayudan a interpretar los patrones sobre todo de generación y demanda energética, pero también serán abordados en estas líneas.

En primer lugar, para el IPI vemos como Las Islas Canarias tienen un valor muy alto, indicado con un color azul oscuro, esto se debe a que como está indicado más arriba, el sector de Servicios también ha sido incluido para el cálculo de este indicador, lo que hace que se dispare para las islas. Seguidamente una parte que llama la atención a simple vista es que aquellas comunidades con un valor muy alto en el PIB per cápita tienen un valor bajo en la generación. Esto viene dado porque también estos territorios (exceptuando Cataluña) tienen una superficie muy pequeña, lo que frena la proliferación de métodos para la obtención de energía. En cuanto a la demanda vemos cómo estas comunidades se comportan de manera “normal” por estar sus valores en la mitad del gradiente. Las comunidades que más energía demandan son, curiosamente, unas de las que más producen, Andalucía y Cataluña parecen seguir un modelo muy distinto a las demás Comunidades Autónomas (tal y como podemos verificar en el objetivo anterior). Finalmente, las comunidades que menos destacan en todos los valores vistos son las que mejor balance generación-demanda tienen, precisamente las referentes a la España vaciada. Esto indica que la producción en estos territorios es mayor que lo que demandan, por lo que podrán ayudar a las comunidades con mayor déficit.

Por último, otra conclusión que podemos extraer de este análisis es que, junto con los informes de la REE [[Anexo 3.2](#bookmark=id.31zo8bvqqdzr)] se puede descubrir el patrón de generación que tiene nuestro país.

Gracias al gráfico referido a la generación energética podemos analizar la estacionalidad de cada una de las variables en la distinta geografía de España. Donde vemos que las Comunidades Autónomas que predominan son Cataluña, Andalucía y Castilla y León. Mantienen este orden prácticamente invariable desde el 1r trimestre del año 2013 -el primer año del que tenemos datos- hasta 2016, donde, para los tres primeros meses del año, la comunidad gallega pasa a ocupar el 2do lugar por delante de Castilla y León y Andalucía. A continuación, en los trimestres posteriores, Castilla y la Región andaluza sobrepasan a Galicia y se mantienen en un equilibrio muy igualado entre las dos.

Sobre todo vemos como los datos siguen este mismo patrón para casi todos los años posteriores. En estaciones con meses más fríos como son el primero y último trimestre vemos como las Comunidades que priman en cuanto a la generación son Castilla y León y Galicia, mientras que en los meses más cálidos es Andalucía la que abandera en la generación. Todo esto claro está por detrás de la Comunidad Catalana que parece tener una estructura de generación muy robusta ya que sus datos predominan frente a todas los demás.

Las CCAA que predominan en la obtención de energía por el viento son Castilla y León , Aragón, Castilla - La Mancha y Galicia.

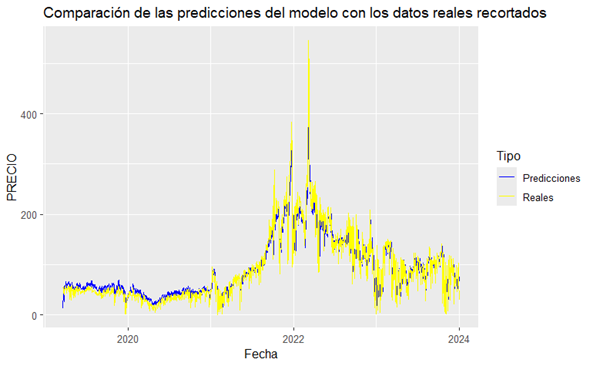
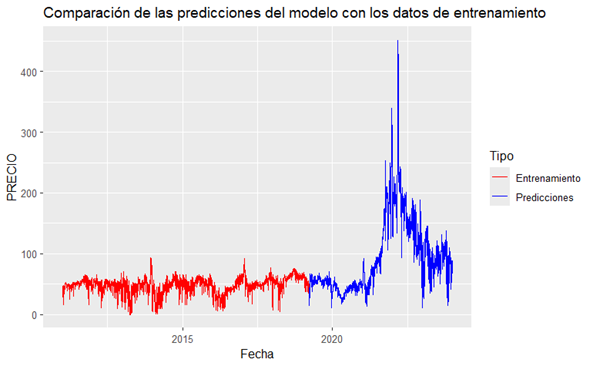
En cuanto a la energía hidráulica la Comunidad Autónoma que predomina es Galicia junto con Castilla y León

El territorio español en el que se impone la fotovoltaica frente al resto es, sin lugar a dudas, Extremadura, Andalucía y Castilla La Mancha, de forma muy parecida lo hace la energía solar térmica.

Como vemos esto encaja perfectamente con los resultados obtenidos en la generación, excepto para el caso concreto de Cataluña, del cual hablaremos más adelante.

Si tuviésemos que encontrar un hilo conductor entre todo lo comentado anteriormente sería que las comunidades que hemos comentado son las que aparecen con un color más oscuro en el mapa que hemos creado para el apartado de Generación-Demanda, lo que nos indica que efectivamente son todas estas comunidades autónomas son las que nos ayudan un poco a suplir ese déficit energético que tanto nos acecha. Este análisis aporta ese punto extra de valor a los gráficos de la REE porque nos ayuda a poder encontrar patrones que son estacionarios que haciendo el cómputo general por año -cómo hacen en la web del sistema eléctrico- no se pueden apreciar. De este modo podemos prever y anticiparnos a lo que se espera en cada temporada porque gracias a este estudio hemos hallado esos patrones.

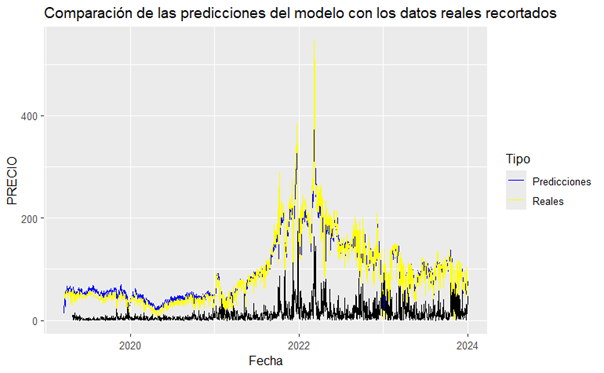
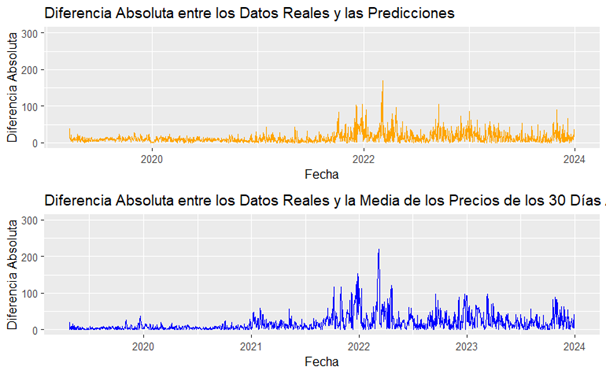
**3.4. Análisis del Precio del MWH en Relación con la Estructura de Generación Energética**

Después del análisis estadístico descriptivo para este objetivo [[Anexo 4.1](#bookmark=id.94tzo42w2wvj)], he separado los datos de España en entrenamiento y regresión [[Anexo 4.2](#bookmark=id.iqxqsd8k76sl)], gracias al análisis previo ya comprobé que variables eran explicativas para este modelo de regresión{ turbinacion.bombeo\_ant, carbon\_ant, ciclo.combinado\_ant, eólica\_ant, solar.fotovoltaica\_ant, precio\_ant, demanda\_ant, precio\_antS, demanda\_antS y demanda\_ant2; donde las que tienen el sufijo ant es que son los datos del día anterior, demanda\_ant2 es la demanda de hace 2 dias y demanda\_antS y precio\_antS son la demanda y el precio de la semana anterior}, por lo que nos quedaremos solo con esas. Hemos entrenado el modelo con unas 3000 observaciones, aproximadamente 2/3 del modelo.

**Fig 11. Gráfico datos entrenamiento con datos test. Fig 12. Gráfico comparando las predicciones con los reales.**

Como podemos observar nuestro modelo se ajusta bastante bien a los datos reales, sobre todo después de un tiempo, ya que se apoya en las series temporales.

Hemos hecho un modelo trivial para demostrar la eficacia del nuestro, el valor del día es simplemente la media de los valores del mes anterior (30 días), ya que se ve bastante mejor que con otras instancias temporales.

El error cuadrático medio de ambos modelos es 881.8547 para el trivial y 400.7717 para el modelo de predicciones.

**Fig 13. Gráfico errores del modelo con uno trivial. Fig 14. Gráfico con las predicciones, los reales y el error.**

En este gráfico podemos comparar los precios con sus predicciones, y la diferencia entre estos, con valores normalmente muy cercanos a 0, pero si que se nota que en ciertas subidas nuestro modelo intenta ajustarlo, pero alguna vez no lo alcanza, ya que se producen por la existencia de factores externos no controlados.

**Conclusiones y resultados:**

**1. Modelo Predictivo Preciso:** Se ha desarrollado un modelo predictivo con alta precisión para predecir el precio del MWH. Este modelo se basa en la demanda y la generación por fuente de energía, utilizando datos de días anteriores.

**2. Relevancia de la Generación Hidráulica:** Aunque España cuenta con numerosos embalses, la generación hidráulica tiene un impacto menor de lo esperado en el precio del MWH. Esto se debe a la estabilidad relativa de esta fuente de generación, la cual no presenta variaciones significativas que afecten directamente al precio.

**3. Comparación con Modelos ARIMA:** El modelo desarrollado es comparable a una serie temporal ARIMA, pero con mayor flexibilidad. A diferencia de los modelos ARIMA tradicionales, que se basan estrictamente en datos de días anteriores, nuestro modelo incorpora múltiples instantes de tiempo pasados, permitiendo capturar una mayor variabilidad en los datos y ofrecer predicciones más ajustadas.

**4. Capacidad de Aprendizaje del Modelo:** El modelo no solo ha sido entrenado con datos históricos, sino que también tiene la capacidad de mejorar con el tiempo. A medida que se incorporan más datos, el modelo aprende y se ajusta mejor a las nuevas condiciones del mercado, mostrando una evolución continua y una mejora en su capacidad predictiva.

**1. Impacto de las Fuentes de Energía en el Precio:**

**- Turbinación/Bombeo Anterior:** tiene un impacto inverso en el precio de la energía. A medida que aumenta, el precio disminuye debido a la eficiencia en la generación.

**- Carbón Anterior:** El uso de carbón incrementa el precio de la energía debido a sus altos costos ambientales y de salud, además de ser una materia bastante escasa.

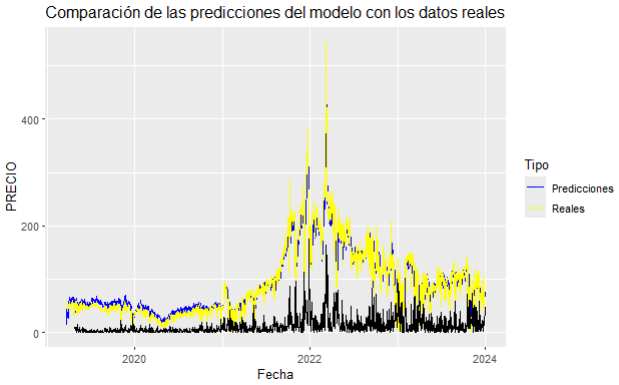
**-Ciclo Combinado Anterior:** aunque eficientes, son costosos de implementar y mantener, lo que lleva a un aumento en el precio de la energía.

**- Eólica Anterior y Solar Fotovoltáica Anterior:** debido a la variabilidad del viento y el sol, causan un aumento en el precio de la energía al incrementar su generación debido a sus costos variables.

**2. Estabilidad de la Energía Hidráulica:** La energía hidroeléctrica, pese a lo que podría parecer por la cantidad de pantanos de España, tiende a ser más estable en términos de precio en comparación con otras fuentes más variables, esto es por la gran cantidad que hay de estos, y porque está basada en una fuente relativamente estable como el agua.

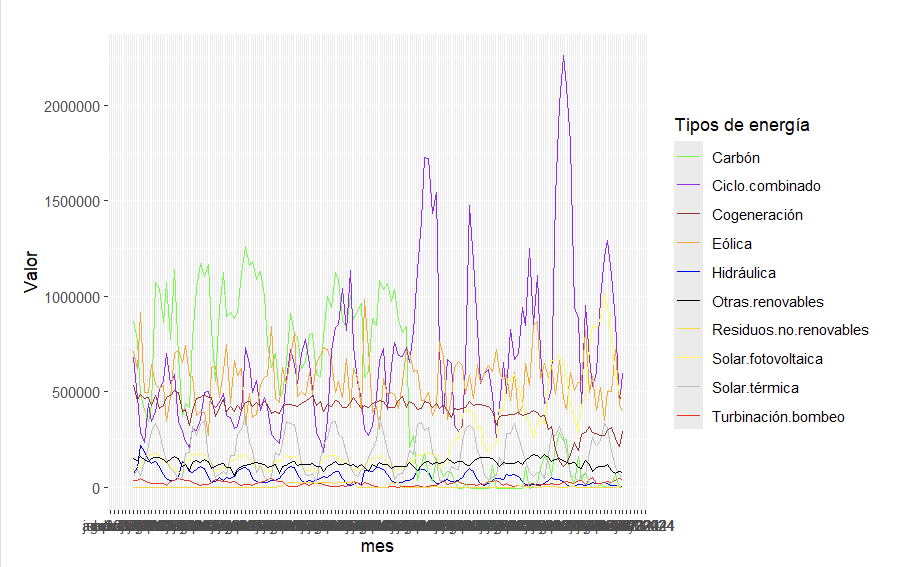
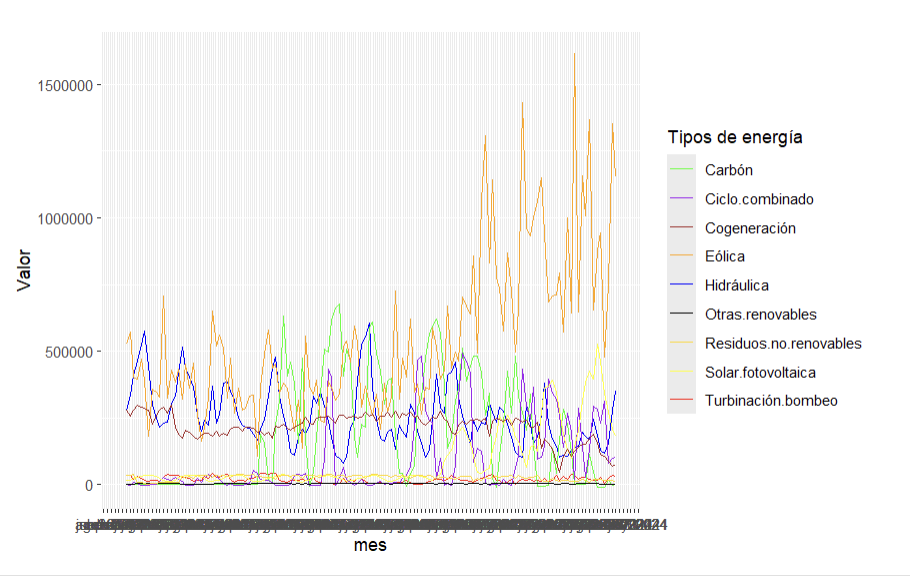
**3. Influencia de Precios y Demanda de Días Anteriores:** El precio de la energía de un día está fuertemente influenciado por el precio y la demanda de días anteriores. Los análisis muestran que tanto el precio como la demanda del día anterior, de la semana y de hace dos días tienen una fuerte correlación (p-valor cercano a 0), lo que los convierte en estimadores perfectos. Esto sugiere que considerar la información de días anteriores mejora significativamente la predicción del precio de la energía.

**4. Modelo en 2024:** Al intentar emplear el modelo en 2024 [[Anexo 4.3](#bookmark=id.xxqsylflxqtc)], no nos ajustaba bastante bien, ya que se entrenaba con datos bastante inestables por lo que veíamos que no recogía tan bien la información como con los datos del test anterior [[Anexo 4.2](#bookmark=id.iqxqsd8k76sl)]. Tras investigar hemos llegado a la conclusión de que esto se debe a la fuerte inestabilidad energética mundial que hay desde 2022 (post-pandemia, guerra, conflictos de intereses mundiales, etc), que se acentuó este año, aún así, intentamos emplear los datos estables del modelo anterior, y en vez de predecir solo los datos de 2024, usamos como test los del modelo anterior y los de 2024 conjuntamente.



**Fig. 15. Gráfico completo de 2024.**

**3.5. Evolución del Uso de Fuentes de Energía en Comunidades Autónomas**

A lo largo de este proyecto, he llevado a cabo un exhaustivo análisis cualitativo de la producción de energía en las distintas comunidades autónomas de España. Mi objetivo principal ha sido identificar patrones, tendencias y variaciones en la producción energética de cada región. Para facilitar este análisis, he desarrollado gráficas detalladas para cada comunidad autónoma que proporcionan una representación visual clara de los diferentes tipos de energía producidos.

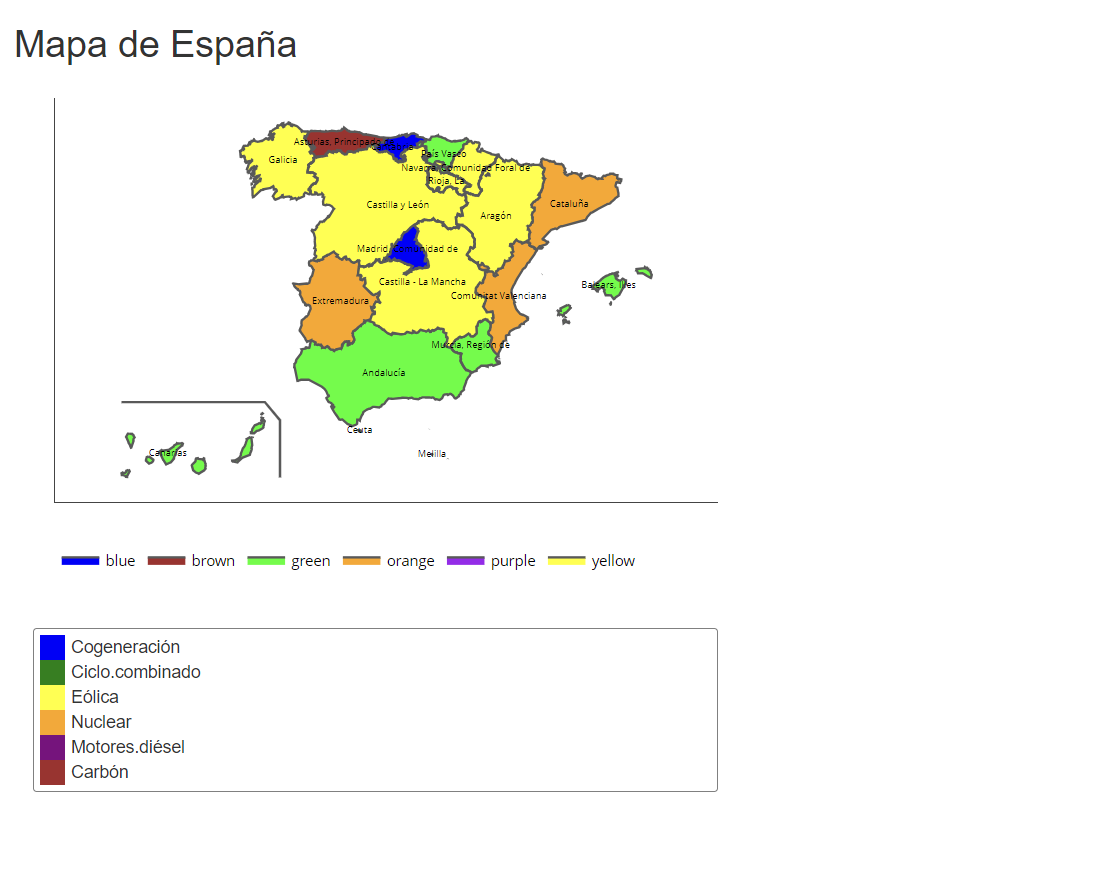
**Fig. 16. Gráfico de tipos de energía 1. Fig. 17. Gráfico de tipos de energía 2.**

|  |  |
| --- | --- |

Las tablas generadas para cada comunidad autónoma muestran la información de cada tipo de energía utilizando colores distintos. Esta visualización por colores permite una comprensión más rápida y efectiva de los datos.

Sin embargo, el análisis detallado del valor diario presentó ciertos desafíos, ya que los datos están divididos por meses en el archivo CSV, lo que complicó la extracción clara de cada mes con su respectivo año. A pesar de estos desafíos, los datos han sido analizados de manera rigurosa abarcando un período desde 2013 hasta 2023.

La recopilación de datos de todas las comunidades autónomas de España me permitió crear un mapa que analiza y visualiza, mediante colores, las diferentes energías predominantes en cada región. Para lograr esto, sumé los datos de cada tipo de energía, integrándose en un único dataframe, lo que facilitó la comparación de la energía más predominante en cada comunidad autónoma.

Podemos llegar a obtener este mapa de España:

**Fig. 18. Gráfico del tipo de energía más usado para cada CCAA.**

El mapa resultante ofrece una visión clara de la distribución de la producción energética en España. Cada comunidad autónoma está representada por un color específico que indica la energía predominante en esa región. Los distintos colores para cada tipo de energía permiten una diferenciación visual efectiva, haciendo el mapa fácil de interpretar y destacando las variaciones regionales en la producción de energía.

**Conclusión**

En primer lugar, mi análisis concluye que las diferentes comunidades autónomas de España se destacan en la producción de seis tipos principales de energía. Para la energía eólica, las regiones prioritarias son Galicia, Castilla y León, Aragón, Navarra, La Rioja y Castilla-La Mancha. En cuanto a la energía de ciclo combinado, Andalucía, Murcia, Canarias y Baleares son las comunidades que predominan. Madrid y Cantabria sobresalen en la producción de energía mediante cogeneración, mientras que Melilla y Ceuta son destacadas por el uso de motores diésel. En el ámbito de la energía nuclear, las regiones principales son la Comunidad Valenciana, Cataluña y Extremadura. Finalmente, Asturias es la comunidad destacada en la producción de energía a partir de carbón.

El mapa de producción y demanda que he creado refleja claramente estas prioridades, con cada comunidad autónoma pintada de un color oscuro que representa la energía predominante en esa región. Este patrón coincide con los resultados de los informes de la Red Eléctrica de España (REE), aunque mi análisis añade un valor adicional al revelar patrones estacionales específicos que no se pueden apreciar en los informes anuales convencionales.

Para las comunidades que predominan en la generación eólica, los factores geográficos y climáticos juegan un papel crucial, proporcionando condiciones ideales para la instalación de parques eólicos.

En las regiones de ciclo combinado, la infraestructura existente y la demanda energética son factores determinantes.

Madrid y Cantabria se destacan en cogeneración debido a su desarrollo industrial y necesidades específicas de energía.

Melilla y Ceuta utilizan motores diésel principalmente por su aislamiento geográfico y falta de conexión a la red principal, mientras que la energía nuclear es prominente en la Comunidad Valenciana, Cataluña y Extremadura debido a las instalaciones nucleares existentes y las políticas energéticas regionales.

Asturias, con su tradición industrial y recursos de carbón, sigue siendo una región clave para la producción de energía a partir de este recurso.

En resumen, cada comunidad autónoma tiene un rol específico y prioritario en la matriz energética de España, contribuyendo de manera significativa a satisfacer la demanda energética y ayudando a equilibrar el déficit energético del país. Este análisis permite prever y anticipar las tendencias estacionales en la generación de energía, proporcionando una herramienta valiosa para la planificación y gestión energética a nivel regional y nacional.

# 4. Lecciones aprendidas para mis futuros proyectos en Ciencia de Datos

**1. Planteamiento de Objetivos Claros y Alcanzables:**

- Desde el inicio, definir claramente los objetivos del proyecto es crucial. Tener una visión precisa de lo que se quiere lograr nos permitió enfocar nuestros esfuerzos y recursos de manera eficiente.

**2. Selección y Discriminación de Fuentes de Datos:**

- Hemos mejorado significativamente en la identificación de fuentes de datos relevantes y fiables. Por ejemplo, hemos aprendido a discriminar entre datos oficiales y no oficiales, y a evaluar la calidad y relevancia de cada fuente antes de su utilización.

**3. Manejo y Limpieza de Datos:**

- El manejo y limpieza de datos se ha convertido en una habilidad fundamental. Utilizando R, Python y Excel, hemos refinado nuestros métodos para seleccionar, descargar y limpiar datos. Por ejemplo, empleamos librerías como `dplyr` y `tidyr` en R para la manipulación y limpieza de datos, lo que nos permitió estructurar grandes volúmenes de datos de manera eficiente.

- En Python, usamos `pandas` para la manipulación de datos y `openpyxl` para interactuar con archivos Excel. Estas herramientas fueron esenciales para asegurar que los datos estuvieran en el formato adecuado para el análisis.

**4. Dominio de Herramientas y Librerías de R:**

Nuestra habilidad en R ha mejorado considerablemente. Hemos utilizado librerías como `ggplot2` para la visualización de datos, `forecast` para el análisis de series temporales, y `caret` para la creación y evaluación de modelos predictivos. Estas herramientas nos han permitido realizar análisis complejos y presentar los resultados de manera clara y efectiva.

Para crear los diagramas de Sankey hemos usado ‘networkD3’. Es una librería con un funcionamiento muy intuitivo y flexible que nos ha ayudado a mejorar la visualización de nuestro trabajo.

Para la extracción de datos de la API hemos usado ‘httr’ y ‘jsonlite’. La base de ambas librerías es muy parecida a requests de Python por lo que no ha sido muy complejo familiarizarse con ellas.

Para poder realizar el clustering hemos hecho uso de diferentes librerías como ‘cluster’, ‘dbscan’, NbClust’, nos han permitido aplicar las diferentes técnicas y realizar agrupamiento k-means, agrupamiento jerárquico, agrupamiento basado en densidad, etc., y evaluar los resultados del agrupamiento. Además también hemos hecho uso de otras librerías como ‘FactoMineR’, ‘factorextra’ para desarrollar el PCA, para poder observar cómo se distribuían las observaciones y qué direcciones tenían las variables.

En cuanto a la creación de los mapas interactivos, hemos utilizado la librería `mapSpain`donde se recoge la información necesaria para hacer el contorno de España, nos dejaba seleccionar la opción de hacerlo por Comunidad Autónoma, así que decidimos escogerla para nutrir aún más nuestro estudio y la visualización. La librería de `plotly` nos ayudó a hacer el mapa interactivo, además `shiny` nos permitió crear la aplicación web necesaria para mostrar los resultados. Esta combinación de herramientas nos ha permitido desarrollar visualizaciones geográficas interactivas y detalladas. Finalmente también hemos hecho uso de `sf` para controlar el puntero e indicar los datos de cada región al pasar el cursor sobre ellos.

**5. Integración de Lenguajes de Programación:**

- La combinación de R y Python ha sido particularmente poderosa. Mientras que R se destacó en el análisis estadístico y la visualización de datos, Python nos ayudó en tareas de preprocesamiento de datos y en la integración con otras aplicaciones.

- Además, utilizamos RStudio como nuestro entorno principal de desarrollo para R, lo que facilitó la gestión de proyectos y la colaboración en equipo.

**6. Presentación de Resultados:**

- La capacidad de presentar los resultados de manera clara y comprensible ha sido fundamental. Aprendimos a utilizar herramientas como `shiny` en R para crear aplicaciones web interactivas que permiten a los usuarios explorar los datos y resultados por sí mismos.

- También mejoramos nuestras habilidades en la creación de informes y presentaciones utilizando `RMarkdown` y `knitr`, lo que nos permitió documentar y comunicar nuestros hallazgos de manera profesional.

**7. Trabajo en Equipo y Comunicación:**

- El uso de herramientas de colaboración como Google Drive para el control de versiones y Discord o Whatsapp para la comunicación interna del equipo ha sido vital. Estas herramientas nos permiten mantenernos organizados y comunicarnos de manera efectiva, asegurando que todos los miembros del equipo estuvieran alineados y al tanto de los avances del proyecto.

**8. Adaptabilidad y Aprendizaje Continuo:**

- A lo largo del proyecto, enfrentamos diversos desafíos que requirieron adaptabilidad y un enfoque de aprendizaje continuo. Nos mantenemos actualizados con las últimas tendencias y técnicas en ciencia de datos, y aprovechamos los recursos en línea y la documentación de las librerías para superar obstáculos y mejorar nuestras habilidades.

En resumen, este proyecto ha sido una experiencia enriquecedora que nos ha permitido mejorar nuestras habilidades técnicas, metodológicas y colaborativas. Hemos aprendido a manejar y analizar datos con mayor eficacia, a utilizar diversas herramientas y lenguajes de programación de manera complementaria, y a presentar nuestros resultados de manera clara y convincente. Estas lecciones serán de gran valor en nuestros futuros proyectos en ciencia de datos.

# 5. Reflexiones sobre la Privacidad y Legalidad de los Datos

Es importante destacar que la obtención de datos, aunque puede ser un desafío, no es necesariamente el obstáculo más grande en el campo de la ciencia de datos. Muchos datos son de hecho accesibles y públicos. Sin embargo, el verdadero desafío radica en saber encontrar cuáles son valiosos, cómo limpiarlos e interpretarlos.

Las Fuentes de Datos (2.1) pueden ser diversas y amplias, abarcando desde datos económicos hasta datos energéticos. La habilidad para discernir qué datos son relevantes para estos objetivos es crucial. Una vez identificados, estos datos deben ser limpiados e integrados en el proceso de Integración y Transformación de los Datos (2.2).

La interpretación de los datos es otro aspecto fundamental. En la Clasificación de Países de la Unión Europea (3.1) y el Clustering de Comunidades Autónomas de España (3.2) obtenemos resultados que no solo requieren una comprensión profunda de los datos, sino también la capacidad de aplicar técnicas de ciencia de datos para extraer significado de ellos.

Por último, pero no menos importante, la privacidad y la legalidad de los datos son aspectos que no deben ser pasados por alto. Si bien muchos datos son públicos, aunque sean datos sensibles, debemos ser conscientes de las leyes y regulaciones que rigen su uso. La privacidad de los individuos y las entidades debe ser respetada en todo momento, pese a que la mayoría de los datos son accesibles para todo el mundo en Internet. En este sentido, es esencial tener en cuenta las leyes de protección de datos y las normativas de privacidad al trabajar con datos.

# 6. ANEXOS

**# Anexos Hitos**

1. Anexo 0.1: [HITO1.docx : en este documento se encuentra el Hito 1]

2. Anexo 0.2: [HITO2.docx : en este documento se encuentra el Hito 2] (REESCRITO)

**# Objetivo 1**

1. Anexo 1.1: [categorizacion\_r.Rmd : archivo empleado para el Análisis Exploratorio de Datos del objetivo 1, así como para el diagrama de Sankey, y las conclusiones de este apartado.]

**# Objetivo 2**

1. Anexo 2.1: [generacion clustering.Rmd: archivo empleado para realizar el clustering a los datos de generación, así como para su representación y los resultados obtenidos.]

2. Anexo 2.2: [demanda clustering.Rmd: archivo empleado para realizar el clustering a los datos de demanda, así como para su representación y los resultados obtenidos.]

3. Anexo 2.3: [gen-dem clustering.Rmd: archivo utilizado para realizar el clustering de los datos de generación menos demanda, así como para su representación y los resultados obtenidos.]

**# Objetivo 3**

1. Anexo 3.1: [Mapa\_definitivo.R : código empleado para generar el mapa de CCAA por generación, PIB, IPI, demanda y empleo]

2. Anexo 3.2: [[Informe de energías renovables | Informes del sistema (sistemaelectrico-ree.es)](https://www.sistemaelectrico-ree.es/informe-de-energias-renovables)

3. Anexo 3.3: [[Informe del sistema eléctrico 2022 | Informes del sistema (sistemaelectrico-ree.es)](https://www.sistemaelectrico-ree.es/2022/informe-del-sistema-electrico)]

**# Objetivo 4**

1. Anexo 4.1: [AED Objetivo4.Rmd : se hace limpieza y preparación de datos, así como el Análisis Exploratorio de Datos y la estimación de parámetros son significativos para la regresión.]

2. Anexo 4.2: [RegresionObjetivo4.Rmd : se aplica el entrenamiento para el modelo de regresión, así como su test, y gráficos para las conclusiones]

3. Anexo 4.3: [RegresionObjetivo42024Prueba.Rmd : se prueba el modelo para datos de 2024 (no se usa ni para conclusiones ni resultados, es mera curiosidad]

**# Objetivo 5**

1. Anexo 5.1: [Proyecto.Rmd: análisis de los diferentes tipos de energía en las Comunidades Autónomas de España y la representación máxima de cada energía por comunidad en un mapa de España ]

**# Limpieza y Preparación de Datos**

1. Anexo 6.1: [dinamarcaayudaobj4r.Rmd : fichero de apoyo para recopilar los datos de diferentes dataframes de Dinamarca en uno (objetivo 4)]

2. Anexo 6.2: [Junta\_datos\_PROY(PIB).Rmd : fichero de apoyo para recopilar los datos de diferentes dataframes de CCAA en uno solo (objetivo 3)]

3. Anexo 6.3: [Limpia\_datos.Rmd : fichero de apoyo para la limpieza del dataframe resultante para la visualización (objetivo 3)]

# 7. Lista de técnicas que se deberán explicar en detalle si se hubieran utilizado:

1. **Transformación de variables numéricas en categóricas y viceversa**: Hemos utilizado esta técnica para convertir las variables numéricas, como la generación y demanda de energía, en categorías que representan rangos de valores. Esto nos ha permitido agrupar los países y las Comunidades Autónomas (CCAA) en categorías basadas en su producción y consumo de energía.
2. **Transformación de formatos (fechas, renombre de factores, etc.)**: Hemos transformado las fechas a un formato estándar para poder realizar análisis temporales. También hemos renombrado factores para facilitar la interpretación de los resultados.
3. **Ordenación de valores**: Hemos ordenado los valores de las variables para identificar tendencias y patrones. Por ejemplo, hemos ordenado los países y las CCAA según su generación y demanda de energía.
4. **Extracción de dataframes**: Hemos extraído dataframes a partir de los datos proporcionados por diversas páginas web, explicadas en el HITO 2 ([Anexo 0.2](#bookmark=id.df7bkwjr066)) para realizar nuestros análisis.
5. **Extracción de vectores y otros elementos graficables**: Hemos extraído vectores de los dataframes para realizar gráficos y visualizaciones.
6. **Implementación de funciones**: Hemos implementado funciones personalizadas para realizar análisis específicos: como el cálculo de la distancia euclídea en el análisis de clustering, la creación de gráficos en la categorización o la generación del mapa de España en la visualización.
7. **Regresión**: Hemos desarrollado un modelo de regresión múltiple para predecir el precio de la energía en España.
8. **Clustering jerárquicos y motivación del método de agrupación**: Hemos realizado análisis de clustering jerárquico y no jerárquico para agrupar a los países y las CCAA según sus valores de generación y demanda. Tras probar varias opciones, nos hemos decidido por el clustering jerárquico mediante el método de Ward y el clustering según el método de k-medoides usando la distancia euclídea.
9. **Dendrogramas**: Hemos utilizado dendrogramas para visualizar los resultados del análisis de clustering jerárquico.
10. **Uso de gráficos (hist, plot, bwplot, boxplot, beanplot, stripplot, density, abline, qqplot, mosaicplot, libreria factoextra, fviz\_cluster, etc.)**: Hemos utilizado una variedad de gráficos para visualizar nuestros datos y resultados. Por ejemplo, hemos utilizado gráficos de líneas para observar la evolución de la distribución de energía en cada país.
11. **Uso de gráficos en ggplot**: Hemos utilizado la librería ggplot para generar gráficos de alta calidad que nos han permitido visualizar de manera efectiva nuestros datos y resultados.