UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Máster en Big Data y Data Science: ciencia e ingeniería de datos

TRABAJO FIN DE MÁSTER

Análisis y Predicción del Consumo de Productos Petrolíferos en España: Identificación de Patrones Regionales y Oportunidades en la Transición Energética

German Anibal Velasquez Zelaya
Tutor: Sergio Galan Martin

Febrero 2025

Análisis y Predicción del Consumo de Productos Petrolíferos en España: Identificación de Patrones Regionales y Oportunidades en la Transición Energética

AUTOR: German Anibal Velasquez Zelaya
TUTOR: Sergio Galan Martin

Escuela Politécnica Superior
Universidad Autónoma de Madrid
febrero de 2025

Resumen

Este Trabajo Fin de Máster se centra en el análisis del consumo de productos petrolíferos en las comunidades autónomas de España entre 2020 y 2024, con el objetivo de identificar patrones de consumo y proyectar tendencias hasta 2026. En el marco de la transición energética y las políticas públicas, se ha utilizado una combinación de técnicas de análisis de datos, segmentación y modelado predictivo para explorar cómo podría evolucionar la demanda de combustibles fósiles en los próximos años.

Para ello, se ha desarrollado un pipeline automatizado que permite gestionar la recolección, limpieza y normalización de los datos mediante herramientas como PySpark y pandas. El estudio incluye un análisis exploratorio detallado, donde se destacan las correlaciones entre los distintos tipos de combustibles y las características particulares de cada comunidad autónoma. Posteriormente, se aplicó el algoritmo de clustering K-means para agrupar regiones con patrones de consumo similares, identificando el Cluster 0 como el grupo con un alto potencial para la transición energética. Este grupo presenta oportunidades para implementar políticas de energías renovables compartidas y estrategias colaborativas entre comunidades autónomas.

A partir de esta segmentación, se han identificado varias oportunidades de transición para las comunidades con alta dependencia de Gasóleo A. Entre las estrategias más destacadas se encuentran el uso de biocombustibles, la electrificación del transporte y la mejora de la eficiencia energética. El Cluster 0 mostró un panorama positivo, con comunidades como Castilla-La Mancha y Galicia, que tienen un buen potencial para integrar fuentes de energía renovable como la solar y la eólica. Este grupo presenta una oportunidad para diseñar políticas conjuntas que impulsen la descarbonización y el uso de energías limpias.

Por otro lado, el Cluster 2 se destacó como una excepción, mostrando discrepancias en las correlaciones y un comportamiento de consumo más complejo. Este grupo, en particular, requiere un análisis más profundo, especialmente en las comunidades autónomas de Ceuta y Melilla, donde las dinámicas energéticas son muy diferentes al resto del país.

Para la proyección de tendencias, se compararon varios modelos predictivos, eligiendo SARIMAX por su capacidad para manejar estacionalidades e incluir factores externos. Sin embargo, se detectó que la ausencia de variables exógenas, como la temperatura, los precios de los combustibles y las políticas energéticas, limita la capacidad predictiva del modelo. Por esta razón, se propone como línea de investigación futura la integración de estos factores para mejorar la precisión de las predicciones.

En resumen, este estudio ofrece una visión contextual del consumo de combustibles fósiles en España, establecieno una base metodológica para futuras investigaciones sobre la transición energética. La combinación de K-means y SARIMAX ha permitido segmentar el consumo en grupos con características comunes, proyectar tendencias y evaluar el impacto potencial de políticas para reducir las emisiones y promover las energías renovables en cada región.

Agradecimientos

En primer lugar, quiero expresar mi agradecimiento a mi tutor de TFM, Sergio Galán. Su orientación y apoyo han sido fundamentales durante el desarrollo de este trabajo. Aprecio profundamente la forma en que valoró mis propuestas y sus valiosas aportaciones, que me ayudaron a establecer una dirección clara para alcanzar mis objetivos. Sus correcciones y sugerencias han enriquecido mi trabajo y me permitieron avanzar con confianza. Además, su disponibilidad y disposición para resolver cualquier duda siempre fueron una constante que agradezco enormemente.

A continuación, quiero agradecer a Jenaro Gallego, uno de mis profesores durante el máster, cuyo apoyo también ha sido especial. Su disposición para enseñar y su capacidad para aclarar conceptos me ayudaron a dar forma a la idea inicial de este TFM. Su compromiso con la enseñanza y su profesionalismo marcaron una diferencia significativa en mi formación académica, definitivamente, el nacio para ser profesor.

Mi agradecimiento más sincero va a mi mamá, Luz. Desde siempre, ha sido mi mayor apoyo y ejemplo, brindándome su confianza y respaldo en cada etapa de mi vida. Gracias a su motivación constante, pude superar los momentos más difíciles y seguir adelante con mis metas. Este logro también es, en gran medida, gracias a su apoyo incondicional.

Quiero dedicar también unas palabras a dos amigos muy importantes, Josué y Ale, cuya pérdida durante el tiempo que realice este TFM fue una de las experiencias más difíciles de superar. Sin embargo, este trabajo me es una marca que me hará mantener vivo su recuerdo y, aunque su ausencia siempre estará presente, el saber que estarían orgullosos de este logro me da fuerzas para seguir celebrando la vida y agradecer el tiempo que estuvieron conmigo.

Finalmente, a mi prometido, José Ángel, le agradezco por su continuo apoyo y comprensión. Su presencia y confianza en mí han sido esenciales para que pudiera completar este proyecto. Gracias por estar a mi lado en todo momento y ayudarme a creer en mis sueños.

INDICE DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Mi principal interés en este estudio es explorar el consumo de combustibles fósiles en España y buscar formas de acelerar la transición hacia energías renovables. Mi interés en este análisis es comprender cómo ha cambiado el consumo de petróleo y sus derivados dentro del marco de las políticas energéticas actuales, así como entender el impacto de los avances tecnológicos en este comportamiento. Este enfoque responde tanto a mi curiosidad académica como a mi compromiso personal con la búsqueda de soluciones sostenibles para los desafíos energéticos globales.

Mi interés por el tema creció al observar la situación energética en mi país natal, Honduras, donde la transición hacia fuentes de energía más limpias avanza con lentitud. En contraste, al vivir en España, he podido ver de cerca los esfuerzos en curso, como la introducción de autobuses y coches eléctricos en las ciudades, y la mejora de la eficiencia energética en edificios. A pesar de estos avances, España aún enfrenta grandes desafíos, especialmente en sectores como el transporte, que sigue dependiendo en gran medida del petróleo. Este contraste entre ambos contextos me motivó a investigar cómo los países, con diferentes niveles de desarrollo energético, están abordando la transición hacia energías más sostenibles.

El estudio se centrará en el consumo de productos petrolíferos en España entre 2020 y 2024, con el objetivo de identificar patrones regionales de consumo utilizando el algoritmo K-means. Este análisis de agrupamientos permitirá ver cómo distintas comunidades autónomas están afrontando la transición energética. Luego, se aplicará el modelo SARIMAX para proyectar las tendencias hasta 2026, con el fin de entender mejor cómo se relacionan el consumo de combustibles fósiles, las políticas energéticas y el avance de las energías renovables.

Este trabajo no solo refleja mi interés por el análisis energético, sino también mi deseo de encontrar oportunidades que ayuden a acelerar el paso hacia un modelo energético más sostenible. A través de este estudio, espero proporcionar una visión más profunda de los desafíos y avances en España, y que lo aprendido pueda contribuir al diseño de estrategias más eficaces en otros contextos similares, como el de mi país natal, Honduras.

1.2 Objetivos

1.2.1 Objetivos principales

1. **Segmentación regional del consumo de petróleo**
   Utilizando el algoritmo K-means, se segmentarán las comunidades autónomas de España para identificar patrones en el consumo de petróleo y la adopción de energías renovables. Esto permitirá comparar cómo las regiones enfrentan la transición energética y los factores que impulsan la reducción del consumo de petróleo.
2. **Análisis de la evolución del consumo de petróleo**
   Mediante el modelo SARIMAX, se analizarán las tendencias históricas del consumo de petróleo en España entre 2020 y 2024, considerando factores socioeconómicos, políticas energéticas y fluctuaciones del mercado. Se proyectarán las tendencias hasta 2026 para informar estrategias energéticas futuras.

3. **Relación entre el consumo de petróleo y la transición energética**
   Se investigará cómo el crecimiento de las energías renovables ha impactado en el consumo de petróleo, especialmente en el transporte y la generación de energía, y cómo la transición hacia fuentes de energía más limpias contribuye a reducir la dependencia del petróleo.

1.2.2 Objetivos Secundarios

1. **Impacto de las políticas energéticas en el consumo de petróleo**
   Se evaluarán las políticas públicas centradas en la eficiencia energética y la promoción de energías renovables, analizando su influencia en la reducción del consumo de petróleo y cómo estas iniciativas han avanzado hacia un modelo energético más sostenible.

1.3 Organización de la memoria

Capitulo 2: Estado del Arte
Revisión de estudios previos sobre el consumo de productos petrolíferos en España, centrados en el periodo 2020-2024, así como las investigaciones relacionadas con la transición energética, las políticas públicas y el impacto de las energías renovables. También se abordan las metodologías aplicadas, como el uso de modelos predictivos y la segmentación mediante K-means.

Capitulo 3: Diseño
Explicación de los objetivos del estudio y la justificación de su relevancia en el contexto de la transición energética. Descripción del conjunto de datos utilizado, las variables analizadas y la metodología empleada para el análisis exploratorio, la segmentación geográfica con K-means y la predicción del consumo con el modelo SARIMAX.

Capitulo 4: Desarrollo
Presentación del análisis exploratorio de datos, incluyendo el procesamiento, limpieza y visualización. Descripción del proceso de segmentación geográfica con el algoritmo K-means y los resultados obtenidos, seguidos del análisis de la proyección de consumo futuro con el modelo SARIMAX.

Capitulo 5: Resultados
Interpretación de los clusters generados, mostrando patrones de consumo y su relación con las oportunidades de transición energética. Análisis de las proyecciones de consumo a futuro y la identificación de áreas con mayor potencial para la implementación de energías renovables, basado en los resultados del modelo SARIMAX.

Capitulo 6: Conclusiones y Trabajo Futuro
Resumen de los hallazgos más importantes relacionados con el consumo de combustibles fósiles y las energías renovables, con recomendaciones sobre políticas públicas. Propuestas para futuras líneas de investigación sobre eficiencia energética, tecnologías emergentes y su impacto en la transición energética.

2 Estado del arte

2.1 Exploración del Uso de Modelos SARIMAX en el Análisis de Series Temporales del Consumo de Petróleo

El análisis del consumo de petróleo requiere herramientas robustas que permitan identificar patrones en datos complejos y dinámicos. En este contexto, el modelo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) se presenta como una solución ideal gracias a su capacidad para modelar tendencias y estacionalidades en series temporales. Aunque SARIMAX también permite la inclusión de factores externos, este estudio se centra exclusivamente en las propiedades internas del sistema analizado, dadas las características del dataset disponible (Ciencia de Datos, 2023).

El modelo ha sido seleccionado para este proyecto debido a su flexibilidad y precisión, cualidades que lo hacen adecuado para trabajar con un conjunto de datos que incluye información detallada sobre el consumo de productos derivados del petróleo, desglosada por provincia, comunidad autónoma y tipo de producto, desde enero de 2020 hasta octubre de 2026. Este dataset es una base sólida para explorar cómo evolucionan los patrones de consumo a lo largo del tiempo y analizar sus tendencias estacionales y generales.

2.1.1 Ventajas del Modelo SARIMAX

2.1.1.1 Captura de estacionalidades y tendencias:

Una de las fortalezas clave de SARIMAX es su capacidad para modelar datos estacionales, lo cual resulta como un eje especial en este estudio. En el caso del consumo de petróleo, es común observar fluctuaciones estacionales relacionadas con factores como la actividad económica, los ciclos climáticos o los patrones de movilidad. Ejemplo, la demanda de productos como el gasóleo B puede aumentar en épocas de mayor actividad agrícola (Ministerio de Agricultura, Pesca y Alimentación, 2024), mientras que el consumo de queroseno de aviación podría estar vinculado a la temporada turística (Segovia, 2024).

2.1.1.2 Predicción y toma de decisiones:

Al combinar componentes autorregresivos, integrados y de media móvil, SARIMAX ofrece un marco robusto para realizar predicciones precisas. Estas predicciones son necesarias para anticipar tendencias y diseñar estrategias informadas en el sector energético. Por ejemplo, prever la evolución del consumo de gasolina o GLP en los próximos años podría contribuir a desarrollar políticas públicas que optimicen la transición energética (Ministerio para la Transición Ecológica y el Reto Demográfico, 2023).

2.1.2 Limitaciones del Modelo SARIMAX

El modelo SARIMAX, aunque es una herramienta robusta y versátil, presenta ciertos desafíos que se hicieron evidentes durante su implementación en este proyecto. A continuación, se detallan las principales limitaciones durante el proyecto y cómo influyeron en el desarrollo del análisis:

2.1.2.1 Complejidad en la implementación y ajuste:

Ajustar los parámetros del modelo SARIMAX fue un desafío técnico debido a la necesidad de comprender la serie temporal en profundidad. A pesar de las dificultades iniciales en sus

ajustes, se consideró el modelo como el más adecuado para capturar las tendencias y estacionalidades del consumo energético.

2.1.2.2 Dependencia de datos de alta calidad:

SARIMAX requiere datos de calidad para obtener resultados concretos. Durante el análisis del dataset, se realizaron pruebas exploratorias y de limpieza. Los cambios realizados incluyeron la conversión de símbolos a mayúsculas y la eliminación de una fila con valores vacíos. Sin embargo, no se encontraron variables exógenas que pudieran ser utilizadas para mejorar el modelo en otras pruebas.

2.1.2.3 Riesgo de sobreajuste:

El modelo SARIMAX tiene una gran flexibilidad, lo que puede llevar a un sobreajuste a los datos históricos, perdiendo capacidad de generalización. Para mitigar este riesgo, se utilizaron validaciones cruzadas y ajustes iterativos para mejorar la precisión de las predicciones.

2.2 Análisis del Consumo de Petróleo Mediante Algoritmos de Clustering

El algoritmo K-means es una herramienta útil para agrupar datos en distintos conjuntos o "clusters" que comparten características similares (Han, 2006). En este estudio, se emplea para analizar los patrones de consumo de petróleo en diversas regiones de España, agrupando comunidades con comportamientos similares. Es por medio de este enfoque que se busca identificar las diferencias significativas entre comunidades autónomas, con el fin de tener una visión más precisa de cómo cada región consume distintos derivados del petróleo, como gasóleo o gasolina.

2.2.1 Relación con SARIMAX

Aunque K-means y SARIMAX se utilizan por separado, ambos métodos complementan y enriquecen el análisis. K-means se emplea para segmentar las regiones de España en función de sus patrones de consumo de petróleo, lo que permite identificar grupos con características similares. Una vez que los clusters están definidos, SARIMAX se utiliza para modelar las tendencias temporales de consumo de petróleo dentro de cada uno de estos grupos, permitiendo prever cómo evolucionará el consumo en el futuro. Esta combinación ofrece un enfoque integral, ya que K-means identifica qué regiones tienen comportamientos similares, mientras que SARIMAX permite proyectar esos comportamientos a futuro, lo que es útil para identificar oportunidades de mejora o para diseñar políticas específicas, como la reducción de emisiones o la implementación de nuevas infraestructuras energéticas (por ejemplo, plantas eléctricas o políticas de eficiencia). En conjunto, estas herramientas ofrecen una visión clara de cómo se puede intervenir en diferentes regiones para optimizar el consumo y fomentar la transición energética.

2.3 Políticas Energéticas y Marcos Regulatorios en España

Analizar cómo las políticas clave, como el PNIEC, impactan en la reducción del consumo de petróleo, destacando la importancia de la transición energética.

2.3.1 PNIEC 2021-2030

El Plan Nacional Integrado de Energía y Clima (PNIEC) tiene metas claras para reducir las emisiones de gases de efecto invernadero y fomentar el uso de energías renovables en España. Establece un objetivo claro: alcanzar un 42% de energía renovable en la producción

de electricidad para 2030. Esto, junto con la electrificación del transporte y la eficiencia energética, implica una disminución del consumo de combustibles fósiles, incluido el petróleo (Ministerio para la Transición Ecológica y el Reto Demográfico, 2023).

2.3.2 Impacto del PNIEC en el Consumo de Petróleo

Aunque el objetivo del PNIEC es reducir la dependencia de los combustibles fósiles, la reciente subida del 7,1% en el consumo de productos petrolíferos, como se observa en el informe CORE de octubre de 2024, refleja el desafío que aún enfrenta España. Este aumento está relacionado con la demanda de productos como gasóleo A y gasolina, especialmente en un contexto de crecimiento económico y aumento del transporte (Corporación de Derecho Público, 2024).

2.3.3 Desafíos para la Transición Energética

Aunque las energías renovables están ganando terreno, la transición hacia un modelo energético más sostenible es compleja y debe acelerarse para cumplir con los compromisos climáticos. A pesar de la disminución del consumo de carbón y gas natural, el país sigue siendo muy dependiente de las importaciones de crudo, lo que limita la efectividad de las políticas energéticas en términos de autosuficiencia y reducción de emisiones (Grupo Forma-T, 2024). Este panorama subraya la necesidad de redoblar los esfuerzos para impulsar la electrificación, las energías renovables y las políticas que promuevan la reducción del consumo de petróleo, alineándose con los objetivos del PNIEC para asegurar un futuro energético más sostenible.

2.4 Uso de Datos Abiertos en Energía

Destacar la importancia de los repositorios públicos de datos como recurso impresindible para el análisis energético y la formulación de estrategias basadas en información confiable.

Los repositorios públicos de datos juegan un papel clave en el análisis energético y la creación de estrategias basadas en información precisa y confiable. Para este proyecto, se utilizaron cinco conjuntos de datos de la estadística de petróleo, disponibles en datos.gob.es, correspondientes al periodo 2020-2024. Estos datos incluyen detalles como la fecha, comunidad autónoma, provincia, tipo de producto y consumo en toneladas métricas. La calidad de la información está asegurada por la Dirección General de Política Energética y Minas, y su desagregación por ubicación y tipo de producto ofrece una visión profunda sobre los patrones de consumo energético en España. Esta granularidad es crucial para identificar diferencias regionales, evaluar políticas energéticas y explorar tendencias hacia la sostenibilidad (Gobierno de España, 2024). Con datos sobre productos clave como gasóleo A, B y C, gasolina y queroseno, se pueden analizar sectores específicos como el transporte y la aviación. Además, la accesibilidad de estos datos, a través de plataformas como datos.gob.es, refleja un esfuerzo por fomentar la transparencia y proporciona a investigadores, empresas y responsables políticos las herramientas necesarias para tomar decisiones informadas en la transición hacia modelos energéticos más sostenibles.

3 Diseño

3.1 Diseño de Análisis y Predicción del Consumo de Productos Petrolíferos en España: Identificación de Patrones Regionales y Oportunidades en la Transición Energética

3.1.1 Planificacion

1. **Obtención de Datos y Limpieza de datos**
   - Obtención de datos de las comunidades autónomas y provincias.
   - Justificacion y muestra inicial de los datos
2. **Preprocesamiendo de Datos**
   - Limpieza y normalización de los datos.
   - Programacion de pipelines
   - Almacenaje y clasificación de datasets limpios
3. **Integración de Datos**
   - Unión de los cinco conjuntos de datos en un DataFrame consolidado.
   - Validación de la consistencia de los datos.
4. **Análisis Exploratorio de Datos**
   - Análisis descriptivo.
   - Agrupación y agregación del consumo por CCAA y tipo de producto.
   - Visualización de patrones de consumo y comparaciones interanuales.
   - Primer vistazo a la correlación entre CCAA.
5. **Análisis de Correlación**
   - Creación de una matriz de correlación del consumo anual por CCAA.
   - Visualización de la matriz con mapas de calor.
6. **Análisis de Clusters con K-means.**
   - Uso del método del codo para determinar el número óptimo de clusters.
   - Aplicación del algoritmo K-means para agrupar las CCAA.
   - Visualización de los clusters.
7. **Evaluación de Clusters**
   - Cálculo de correlaciones dentro de cada cluster.
   - Comparación de los comportamientos de consumo en cada cluster durante los últimos 5 años.
   - Relación entre los clusters y la predicción futura del consumo.
8. **Selección y Optimización del Modelo SARIMAX**
   - Modelos considerados (ARIMA, Prophet, SARIMAX) y la razón de selección de SARIMAX.
   - Optimización del modelo mediante búsqueda en cuadrícula.
   - Resultados de la optimización (parámetros óptimos, AIC).
9. **Predicción del Consumo con SARIMAX**
   - Proyección del consumo anual para los próximos dos años (2025-2026).
   - Visualización de las tendencias esperadas para cada CCAA y producto.
10. **Interpretación de Resultados: Tendencias Generales**
    - Cambios en el consumo debido a la pandemia de COVID-19 u otros acontecimientos hisoricos.
    - Influencia de políticas energéticas y cambios económicos.

11. Interpretación de Resultados: Variación Regional

- Diferencias significativas entre regiones, relacionados con la actividad económica, turismo y estacionalidad.

12. Conclusiones

- Resumen de los hallazgos clave, identificación de patrones y factores clave en el consumo de productos petrolíferos.
- Recomendaciones para la planificación energética y estrategias económicas regionales.

3.1.2 Esquema

Ilustración 1: Estructura del proyecto; Analisis y predicción del consumo de Productos Petroliferos en España

4 Desarrollo

4.1 Obtención de Datos y Limpieza de Datos

4.1.1 Recoleccion de Datos

4.1.1.1 Justificación de la fuente de datos:

Se eligió el portal datos.gob.es como fuente principal por varios motivos académicamente necesarios:

- **Fiabilidad y veracidad:** Es un portal oficial del gobierno de España, que garantiza que los datos sean de alta calidad y auténticos.
- **Actualización constante:** Ofrece información reciente y de importancia para el período de estudio (2020-2024), permitiendo un análisis actualizado.
- **Formato estructurado:** Los datos están en formato CSV, facilitando su análisis y manipulación con herramientas de ciencia de datos.
- **Transparencia y respaldo normativo:** Los datos se recopilan según la Resolución de la Dirección General de Política Energética y Minas de 29 de mayo de 2007, asegurando su acceso público y transparencia.

4.1.1.2 Relevancia del período de tiempo (2020-2024):

Este período permite analizar las variaciones en el consumo de productos derivados del petróleo bajo eventos globales muy marcados:

- **Pandemia de COVID-19:** La crisis sanitaria afectó drásticamente la economía y la movilidad, lo que alteró profundamente los patrones de consumo energético. Según el Banco Mundial (Banco Mundial, 2020), la reducción de la demanda de petróleo durante las primeras etapas de la pandemia provocó una caída en los precios, los cuales no llegaron a recuperar los niveles previos a la crisis. La caída en la demanda continuó más allá de 2021, impulsada principalmente por las restricciones a los viajes y al turismo, junto con una recuperación económica global que fue mucho más lenta de lo esperado.
- **Crisis energética internacional:** Conflictos como la guerra en Ucrania han afectado en los mercados globales del petróleo, no solo afectando los precios, sino también la disponibilidad y distribución de productos derivados del petróleo. Estos conflictos han llevado a una reconfiguración de las cadenas de suministro y a cambios en las políticas energéticas de muchos países, lo que a su vez ha impactado el consumo de petróleo en diferentes regiones. (Yang, 2023)

4.1.1.3 Nivel de desagregación de los datos:

El desglose detallado de la información aporta una base sólida para un análisis profundo:

- **Fecha (mensual):** Identifica tendencias y patrones estacionales.
- **Comunidad Autónoma (CCAA)y provincia:** Analiza patrones de consumo específicos por región, ajustándose a las particularidades económicas y sociales.
- **Tipo de producto(Tipo Producto):** Evalúa los usos de cada derivado del petróleo y su importancia en diferentes sectores.

4.1.1.4 Selección de variables:

Las variables en este dataset (Fecha, Comunidad Autónoma, Provincia, Tipo de Producto, Consumo en Toneladas Métricas) aportan la información suficiente para tomar estos enfoques:

- **Reflejan indicadores económicos y energéticos:** El consumo de petróleo está directamente relacionado con la actividad económica y energética de las regiones. (Caixabank Research, 2019)
- **Permiten estudiar la transición energética:** Ayudan a identificar cambios hacia combustibles más sostenibles como el biodiésel y el bioetanol. (Asociación Española del Bioetanol (BIO-E), 2020)

La estructura y clasificación de los datos siguen prácticas internacionales comunes en estadísticas energéticas, indicando que es posible la comparabilidad con otros estudios y países. Esto con el fin de fortalecer la validez y relevancia de los resultados obtenidos.

4.1.1.5 Metodología de recolección y actualización de datos:

El proceso para obtener y actualizar los datos, asegura un análisis riguroso y reproducible:

- Descarga inicial: Realizada el 7 de octubre de 2024, incluyó datos hasta julio de 2024.
- Actualización posterior: El 26 de diciembre de 2024, se incorporaron datos hasta octubre de 2024.

4.1.1.6 Descripción de los productos petrolíferos:

La clasificación de productos en el dataset permite un análisis integral de sus usos:

- **Biodiésel y bioetanol:** Indicadores de una transición hacia energías renovables.
- **Gasóleo A, B, C y otros derivados:** Usos clave en transporte, agricultura y calefacción.
- **Querosenos de aviación:** Reflejan la actividad del sector aeronáutico.
- **Gases licuados del petróleo:** Usos diversos en hogares e industrias.

4.1.2 Muestra de los datos sin manipular

Fecha	CCAA	Provincia	Tipo Producto	Consumo Tm
2020-01	AndalucÃa	AlmerÃa	BIODIESEL	0
2020-01	AndalucÃa	AlmerÃa	GASÃ“LEO A	32442
2020-01	AndalucÃa	AlmerÃa	GASÃ“LEO B	6612
2020-01	AndalucÃa	AlmerÃa	GASÃ“LEO C	547
2020-01	AndalucÃa	AlmerÃa	OTROS GASOLEOS	826
2020-01	AndalucÃa	AlmerÃa	BIOETANOL	0
2020-01	AndalucÃa	AlmerÃa	GASOLINA AUTO. S/PB 95 I.O.	3985
2020-01	AndalucÃa	AlmerÃa	GASOLINA AUTO. S/PB 98 I.O.	295

Tabla 1: Muestra de datos sin procesar

4.2 Preprocesamiento de Datos

Esta etapa es de suma importancia en este proyecto para asegurar que la información utilizada en el análisis sea precisa y confiable. En esta fase, trabajamos con las librerías PySpark y pandas, ya que nos permiten manejar grandes volúmenes de datos de manera eficiente. (Apache Software Foundation, 2023)

4.2.1 procedimiento

4.2.1.1 Carga inicial de los datos

Comenzamos cargando los datasets en formato CSV utilizando PySpark. Especificamos el delimitador (;) y habilitamos las opciones para reconocer la cabecera e inferir automáticamente los tipos de datos. Consiguiendo así un esquema claro y ágil para las tareas de limpieza posteriores.

```
# Cargar el dataset desde un archivo CSV, especificando el delimitador como `;`
df = spark.read.option("delimiter", ";").option("header", True).option("inferSchema", True).csv("ds_2966_2024.csv")
```

4.2.1.2 Limpieza y normalización de datos

Se aplicaron diversas técnicas para garantizar la consistencia de los datos en los cinco datasets, entre ellas:

- **Eliminación de caracteres especiales y acentos**: Para evitar inconsistencias en los textos, utilizamos la función unicodedata.normalize combinada con una UDF (User Defined Function) en PySpark. Utilizado para limpiar acentos y caracteres especiales en las columnas de texto.
- **Estandarización de texto:** Convertimos el contenido de las columnas (CCAA, Provincia y Tipo Producto) a mayúsculas mediante la función upper, logrando uniformidad en los valores.
- **Formato de columnas y nombres:** Los nombres de las columnas y sus valores se ajustaron eliminando espacios en blanco y otros caracteres innecesarios mediante las funciones trim y regexp_replace, aunque este segundo es mencionado solo como una opción adicional.
- **Validación del tipo de dato:** Aseguramos que la columna Consumo Tm tuviera el tipo de dato correcto (entero), esto para facilitar la lectura de los análisis posteriores.

```
# Verificar el esquema para asegurarse de que "Consumo Tm" esté en formato entero
df.printSchema()

# Función para eliminar acentos
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(c)])

# UDF (User Defined Function) para aplicar la normalización de texto
remove_accents_udf = udf(lambda x: remove_accents(x) if x is not None else None, StringType())

# Convertir texto a mayúsculas y eliminar acentos
df_cleaned = df \
    .withColumn("CCAA", upper(remove_accents_udf(col("CCAA")))) \
    .withColumn("Provincia", upper(remove_accents_udf(col("Provincia")))) \
    .withColumn("Tipo Producto", upper(remove_accents_udf(col("Tipo Producto"))))

# Asegurarse de que la columna "Consumo Tm" sea de tipo entero
df_cleaned = df_cleaned.withColumn("Consumo Tm", col("Consumo Tm").cast(IntegerType()))
```

Ilustración 2: Codigo de limpieza de datos.

```
# Normalizar posibles inconsistencias en "Tipo Producto"
df_cleaned_no_negativos = df_cleaned_no_negativos.withColumn("Tipo Producto", trim(col("Tipo Producto")))
```

Ilustración 3: Variable normaliza incosistencias

4.2.1.3 Tratamiento de valores anómalos y nulos

Para mejorar la calidad del dataset, se realizaron los siguientes pasos:

- **Eliminación de valores negativos**: Filtramos los registros donde la columna Consumo Tm tenía valores negativos, ya que no eran válidos para nuestro análisis.
- **Limpieza de valores nulos**: Eliminamos las filas con valores faltantes utilizando la función na.drop(), garantizando que los datos restantes fueran completos.
- **Ajuste del formato de fecha:** Se transforma la columna Fecha al formato estándar YYYY-MM utilizando la función date_format, dejando este formato predeterminado para el análisis temporal y la integración de datos.
- **Revisión de valores extremos:** Como paso final, se realiza un análisis en busca de posibles valores atípicos en la columna Consumo Tm mediante el método describe, identificando y gestionando cualquier inconsistencia.

```
df_cleaned_no_negativos = df_cleaned_no_negativos.na.drop()  # Eliminar cualquier fila que tenga valores nulos

# Asegurarse de que la columna "Fecha" esté en formato correcto
df_cleaned_no_negativos = df_cleaned_no_negativos.withColumn("Fecha", date_format(col("Fecha"), "yyyy-MM"))

# Normalizar posibles inconsistencias en "Tipo Producto"
df_cleaned_no_negativos = df_cleaned_no_negativos.withColumn("Tipo Producto", trim(col("Tipo Producto")))

# Eliminar posibles duplicados
df_cleaned_no_negativos = df_cleaned_no_negativos.dropDuplicates()

# Revisión de valores extremos en "Consumo Tm"
df_cleaned_no_negativos.describe("Consumo Tm").show()
```

Ilustración 4: Codigo de tratamiento para valores nulos

4.2.2 Validación Posterior al Preprocesamiento de Datos

Después de implementar las técnicas de limpieza y normalización descritas anteriormente, los datos resultantes presentan un formato homogéneo y estandarizado. Visualizando los cambios en la muestra a continuacion:

Fecha	CCAA	Provincia	Tipo Producto	Consumo Tm
2020-01	ANDALUCIA	JAEN	GLP	1853
2020-01	ANDALUCIA	MALAGA	QUEROSENO AVIACION	23359
2020-01	ANDALUCIA	SEVILLA	GASOLEO B	9405
2020-01	CANARIAS	SANTA CRUZ DE TENERIFE	BIOETANOL	0
2020-01	COMUNITAT VALENCIANA	VALENCIA/VALENCIA GASOLINA	AUTO. S/PB 95 I.O.	19416
2020-01	GALICIA	LUGO	GASOLEO C	6275
2020-02	ANDALUCIA	GRANADA	FUELOLEO BIA	1453
2020-02	ARAGON	HUESCA	BIOETANOL	0
2020-02	CASTILLA Y LEON	LEON GASOLINA	AUTO. S/PB 95 I.O.	3525
2020-02	CASTILLA Y LEON	LEON GASOLINA	AUTO. S/PB 98 I.O.	250
2020-02	EXTREMADURA	BADAJOZ	BIOETANOL	0
2020-03	ANDALUCIA	CORDOBA	GASOLEO B	8456
2020-03	ARAGON	ZARAGOZA	GASOLEO C	6889
2020-03	CASTILLA Y LEON	PALENCIA GASOLINA	AUTO. S/PB 95 I.O.	912
2020-03	CASTILLA Y LEON	ZAMORA	OTROS GASOLEOS	0
2020-03	CATALUNA	BARCELONA	OTROS GASOLEOS	30933
2020-03	COMUNITAT VALENCIANA	CASTELLON/CASTELLO	BIODIESEL	0
2020-03	MURCIA, REGION DE	MURCIA	GASOLEO C	779
2020-04	ANDALUCIA	HUELVA	BIOETANOL	0
2020-04	ARAGON	TERUEL	BIOETANOL	0

Tabla 2: Muestra de Datos limpios

En este formato, los datos muestran una clara uniformidad en campos como las fechas, las regiones (CCAA), y los nombres de las provincias y productos, simplificando el manejo y

análisis. Asimismo, tanto los valores numéricos como las fechas están listos para realizar operaciones como análisis temporales, agregaciones, o filtros personalizados.
Para complementar, se generó un esquema de los datos procesados que valida la estructura homogénea de las columnas:

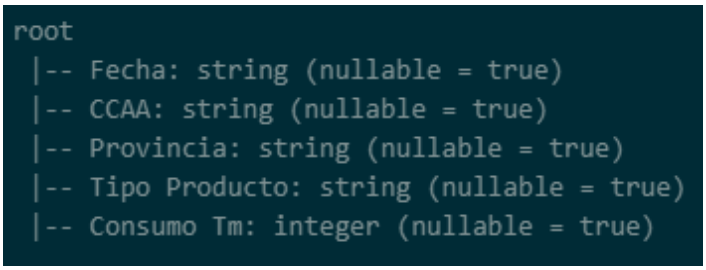


Ilustración 5: Log esquema de datos procesados

Además, se realizó un análisis exploratorio del campo Consumo Tm, obteniendo los siguientes valores estadísticos que confirman la integridad del atributo numérico:

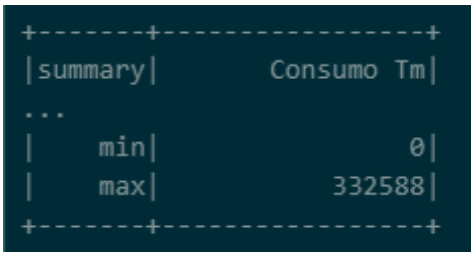


Ilustración 6: Log valores estadisticos I

4.2.3 Automatización del Pipeline de Preprocesamiento

El preprocesamiento se ha automatizado para hacer el proceso más eficiente y consistente. Con PySpark, se desarrollaron funciones modulares que se encargan de la limpieza, normalización y validación de los datos. Este pipeline permite la ejecución de manera automática en datasets futuros que tengan características similares, reduciendo así la necesidad de intervención manual. Además, se integraron registros (logs) que facilitan la detección de errores y el monitoreo en tiempo real, garantizando que el flujo de trabajo sea sólido y confiable.

4.2.4 Apertura para Datos Adicionales

Es importante destacar que el dataset fue actualizado a mediados de diciembre, una mención a su naturaleza dinámica. El pipeline se diseñó con flexibilidad, por lo que puede incorporar estos nuevos registros sin dificultad. También, si en el futuro se agregan nuevas fuentes de datos o se modifican las columnas, las funciones de limpieza y transformación se ajustarán fácilmente para integrarlas sin alterar los procesos previos.

Aunque los cinco datasets fueron limpiados y normalizados por separado, aún no se ha realizado la consolidación en un único archivo. Esta etapa se llevará a cabo más adelante, una vez que se haya validado la calidad de los datos. Al momento de la culminación de esta etapa, los datasets limpios se encuentran almacenados de manera temporal, listos para su futura integración.

```
# Guardar el dataset limpio en un único archivo CSV
df_cleaned_no_negativos.coalesce(1).write.csv("consumo2022_limpio.csv", mode="overwrite", header=True)
```

Ilustración 7: Almacenamiento de cada uno de los dataset procesados para este estudio

4.3 Integracion de Datos

Para consolidar los cinco conjuntos de datos, correspondientes a los años 2020 a 2024, en un único DataFrame llamado consumo2020_2024, se programó el código de esta forma:

1. **Carga de los Archivos CSV:** Los archivos CSV correspondientes a cada uno de los años (2020, 2021, 2022, 2023 y 2024) fueron cargados en DataFrames individuales usando PySpark. Al hacerlo, se aseguró que la primera fila de cada archivo se interpretara como los nombres de las columnas mediante el parámetro header=True, y se permitió que PySpark detectara automáticamente los tipos de datos de cada columna con inferSchema=True.
2. **Unión de los DataFrames:** Después de cargar los datos, se unieron los cinco DataFrames utilizando la función `union()`. Este paso permitió combinar todas las filas de cada conjunto de datos en un solo DataFrame consolidado, que integra la información de todos los años de manera eficiente y continua.
3. **Visualización de los Resultados:** Para asegurarse de que la unión de los datos se realizara correctamente, se visualizó una muestra del DataFrame resultante. Esto permitió confirmar que los datos de todos los años estaban combinados en un solo DataFrame con las columnas correspondientes: Fecha, CCAA, Provincia, Tipo Producto y Consumo Tm.

```
# Crear una sesión de Spark
spark = SparkSession.builder \
    .appName("Union Datasets") \
    .getOrCreate()

# Cargar cada uno de los datasets en DataFrames
consumo2020 = spark.read.csv("consumo2020.csv", header=True, inferSchema=True)
consumo2021 = spark.read.csv("consumo2021.csv", header=True, inferSchema=True)
consumo2022 = spark.read.csv("consumo2022.csv", header=True, inferSchema=True)
consumo2023 = spark.read.csv("consumo2023.csv", header=True, inferSchema=True)
consumo2024 = spark.read.csv("consumo2024.csv", header=True, inferSchema=True)

# Unir todos los DataFrames en uno solo
consumo_unido = consumo2020.union(consumo2021) \
                           .union(consumo2022) \
                           .union(consumo2023) \
                           .union(consumo2024)

# Mostrar las primeras filas del DataFrame unificado
consumo_unido.show()
```

Ilustración 8: Codigo de unificación para crear la nueva base de datos

- Guardado del DataFrame Consolidado: Finalmente, repetimos los pasos anteriores para guardar el nuevo DataFrame consolidado, se guardó en un archivo CSV único llamado consumo2020-2024.csv. Para asegurarse de que se generara un solo archivo (en lugar de múltiples archivos particionados), se utilizó el método coalesce(1). Además, el parámetro mode="overwrite" permitió sobrescribir cualquier archivo previo con el mismo nombre.

```
# Guardar el DataFrame unido como un archivo CSV
consumo_imputado.coalesce(1).write.csv("consumo2020_2024.csv", header=True, mode="overwrite")
#se cambio el nombre a consumo2020-2024
```

Ilustración 9: Codigo para guardar u actualizar la nueva base de datos

Una vez terminando la integracion de los datos, el dataset con un peso final de 1.96 mb, esta listo para la exploracion de su contenido y futuras pruebas, acentuando nuevamente la flexibilidad que tiene en sus caracteristicas para ser actualizado de forma mas automatica.

4.4 Visualizacion y exploracion del consumo petrolífero en España

En esta sección, el objetivo principal es realizar un análisis exploratorio del conjunto de datos para comprender su estructura y comportamiento antes de aplicar los modelos de segmentación y predicción. Este paso inicial será una guia importante para preparar el camino hacia el modelo de clustering K-means, que ayudará a identificar patrones de consumo entre las distintas regiones y tipos de productos. Los resultados obtenidos de esta segmentación, a su vez, facilitarán la parametrización y el ajuste del modelo SARIMAX, que se utilizará para predecir el consumo futuro.

Para este propósito, se trabajará con un notebook interactivo, que será la herramienta principal tanto para la exploración de los datos como para la creación de visualizaciones que guiarán las siguientes fases del proyecto. El análisis exploratorio tiene como propósito principal examinar las variables disponibles, identificar posibles anomalías y generar gráficos que permitan desarrollar los modelos de manera más efectiva en etapas posteriores.

El análisis exploratorio de datos es una pieza importante, ya que permite descubrir patrones, conexiones y tendencias ocultas en la información, lo cual facilita la creación de modelos más precisos (Tukey, 1977). En este sentido, es esencial comprender los datos a fondo, pues esto permitirá estructurar adecuadamente el análisis de clustering y la modelización de series temporales.

Durante esta etapa, se realizarán consultas para analizar cómo interactúan las variables entre sí. Además, se generarán visualizaciones que nos ayudarán a estudiar el comportamiento temporal del consumo, segmentado por tipo de producto, comunidad autónoma (CCAA) y provincia. Debido al alto número de CCAA y provincias, algunas visualizaciones pueden resultar demasiado cargadas, lo que dificultaría su interpretación, siendo esta dificultad superada al momento de crear los clusters. Por lo tanto, se priorizará una presentación general de las CCAA o provincias, evitando detalles excesivos para garantizar la claridad y facilitar el análisis.

4.4.1 Cargar la Nueva Base de Datos

En esta fase, los datos fueron importados desde la nueva base de datos “consumos2020_2024”. El archivo incluye un total de 41,491 registros, que detallan el consumo de productos clasificados por Comunidad Autónoma (CCAA), provincia, tipo de producto y cantidad consumida.

4.4.2 Transformación de los Datos

Una vez cargados los datos, se llevaron a cabo varias transformaciones necesarias para facilitar su manipulación:

4.4.2.1 Conversión de la columna 'Fecha'

La columna de fecha fue convertida al formato datetime, lo que simplifica la manipulación de los datos y facilita la detección de tendencias temporales.

4.4.2.2 Extracción de Año y Mes

Se añadieron dos nuevas columnas (Año y Mes) para hacer más fácil el análisis temporal. Estas nuevas variables permiten identificar patrones estacionales y tendencias a largo plazo en el consumo.

```
# Cargar los datos desde un archivo CSV
df = pd.read_csv('consumo2020_2024.csv')

# muestra que la columna 'Fecha' esté en formato datetime
df['Fecha'] = pd.to_datetime(df['Fecha'], format='%Y-%m')
df['Año'] = df['Fecha'].dt.year
df['Mes'] = df['Fecha'].dt.month

# Verifica la estructura del dataframe
print(df.head())
```

Ilustración 10: Manipulacion y converción de la columna Fecha

Este paso se llevo acabo durante el proceso de exploracion como una medida necesaria para su manipulacion en los estudios posteriores, sin embargo pudo haberse realizado al momento de la limpieza e iteracion de datos.

4.4.3 Nueva Estructura del Conjunto de Datos

El conjunto de datos final tiene la siguiente estructura:

- **Fecha:** Fecha del registro de consumo (formato datetime).
- **CCAA:** Comunidad Autónoma en la que se registró el consumo.
- **Provincia:** Provincia dentro de la CCAA correspondiente.
- **Tipo Producto:** Tipo de producto consumido (gasolina, gasóleo, GLP, etc.).
- **Consumo Tm:** Cantidad consumida en toneladas métricas.
- **Año:** Año extraído de la columna "Fecha".
- **Mes:** Mes extraído de la columna "Fecha".

4.4.4 Estadísticas Descriptivas de la columna Consumo (Tm)

Se realizó un análisis que detalla las estadísticas descriptivas, las cuales incluyen la media, la desviación estándar, los percentiles y los valores extremos. Esto para identificar la calidad de los datos:

4.4.4.1 Estadísticas Generales

- **Número de observaciones:** 41,491 registros.
- **Promedio (mean):** 5,592.29 toneladas métricas.
- **Desviación estándar (std):** 17,674.76, lo que indica una alta variabilidad.
- **Valores extremos:** El máximo (332,588) y el mínimo (0) sugieren la presencia de valores atípicos.
- **Distribución:** La mediana (168) y el primer cuartil (0) indican que la distribución es sesgada hacia valores bajos.

4.4.4.2 Rango de Fechas

- **Fecha mínima:** 1 de enero de 2020.
- **Fecha máxima**: 1 de septiembre de 2024.

Este período de análisis resulta se suma importancia para estudiar el impacto de eventos como la pandemia de COVID-19 o la guerra en Ucrania sobre el consumo de combustibles (Yang, 2023). Asimismo, ofrece la oportunidad de identificar fluctuaciones estacionales, que serán relevantes en la futura modelización de series temporales (Tiwari, 2020).

4.4.4.3 Valores nulos y duplicados

- **Valores nulos**: No se encontraron datos faltantes.
- **Filas duplicadas:** No se hallaron registros duplicados, lo que asegura la integridad del conjunto de datos.

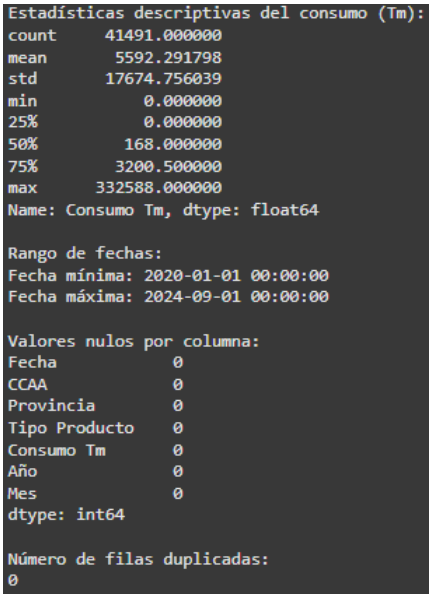


Ilustración 11: Estadisticas descriptivas del dataset consumos2020_2024

4.4.5 Visualización del Conjunto de Datos

Las visualizaciones generadas en esta etapa son necesarias para entender mejor los datos y proporcionar información que guiará tanto la segmentación con K-means como la predicción con SARIMAX en las fases posteriores:

4.4.5.1 Histograma del Consumo (Tm)

Este gráfico ayuda a identificar la distribución del consumo y detectar valores atípicos. Esta información es importante tanto para la segmentación con K-means como para evitar sesgos en los modelos predictivos.

4.4.5.2 Gráfico de Consumo Anual por Tipo de Producto

Muestra la evolución del consumo de gasolina, gasóleo y GLP entre 2020 y 2024, permitiendo identificar patrones que serán utilizados en la segmentación de consumidores con K-means.

4.4.5.3 Patrones Estacionales del Gasóleo C

El análisis de la variabilidad mensual en el consumo de gasóleo C es clave para identificar tendencias estacionales. Esta información será relevante para la identificación de grupos de

consumo homogéneos mediante K-means y, más adelante, para modelar patrones cíclicos en SARIMAX.

4.4.5.4 Distribución del Consumo por Tipo de Producto

Este gráfico circular permite visualizar la proporción de consumo entre los distintos productos. Entender este analisis es importante para una entender el comportamiento de los cluster creados por K-means, ya que facilita la identificación de segmentos de consumo con patrones similares.

4.4.5.5 Comparación de Consumo por Año y Mes

Un gráfico de líneas interactivo muestra la evolución del consumo a lo largo del tiempo. Este análisis será una guia para entender la dinámica del consumo antes de aplicar SARIMAX.

4.4.5.6 Mapa de Correlación por CCAA

Este mapa de calor revela la correlación del consumo entre las diferentes comunidades autónomas, lo que ayuda a identificar regiones con patrones de consumo similares. Esta información será utilizada en el análisis de clustering con K-means para agrupar regiones con comportamientos homogéneos.

Las visualizaciones se generaron utilizando herramientas como Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021) y Plotly (Plotly Technologies Inc., 2015), seleccionadas por su capacidad de interactividad y claridad.

4.5 Análisis de Correlación entre comunidades autónomas

En esta sección, se analiza la correlación del consumo entre las distintas comunidades autónomas mediante un mapa de calor generado en el paso anterior. Este gráfico permite identificar regiones con patrones de consumo similares, lo cual es útil para el análisis de clustering con K-means.

La correlación se calcula usando el coeficiente de Pearson, que nos indica la fuerza y dirección de la relación entre las comunidades. Un valor cercano a +1 señala una fuerte relación positiva, mientras que valores cercanos a -1 indican una relación negativa. Valores cercanos a 0 sugieren que no hay una relación clara (Hernández Lalinde, 2018).

Este análisis será de gran utilidad para entender el agrupamiento de las regiones con comportamientos de consumo similares, lo que mejorará la precisión de la segmentación y el modelo SARIMAX.

4.6 Análisis de Clusters con K-means

En esta sección se aplica el algoritmo K-means para identificar patrones de consumo de productos petrolíferos en las distintas comunidades autónomas de España. El objetivo es segmentar las regiones en clusters que compartan comportamientos similares en cuanto al consumo de estos productos, lo que facilitará la identificación de oportunidades de transición hacia energías renovables o politicas comunes.

4.6.1 Método del Codo y Silhouette Score

Para determinar el número óptimo de clusters, se utilizó el método del codo. Este enfoque evalúa la suma de los errores cuadráticos dentro de cada cluster (WCSS) para distintos números de clusters, buscando el punto en el que la reducción del error se estabiliza.

Posteriormente se verificara la calidad de la dispersion mediante el Silhouette Score. Este índice mide la calidad de la agrupación, y un valor cercano a 1 indica una buena separación entre los clusters. (Tukey, 1977).

4.6.2 Creación de los Clusters

Con el número óptimo de clusters determinado, se aplicó el algoritmo K-means para agrupar las comunidades autónomas (CCAA) en función del consumo total de productos petrolíferos ('Consumo Tm') y el porcentaje de consumo de energías renovables ('Porcentaje'). Los resultados fueron organizados en un DataFrame denominado ccaa_clusters, que muestra las CCAA agrupadas en cada uno de los clusters creados. Posteriormente, se visualizó la distribución de los clusters para evaluar su coherencia y observar posibles patrones o diferencias en el consumo de productos petrolíferos en las distintas regiones (Yang, 2023).

4.7 Evaluacion de los Cluster

Una vez generados los clusters con K-means, se procede a realizar una evaluación más detallada para analizar su comportamiento y la relación con las oportunidades de transición energética en cada región.

4.7.1 Comportamiento del Consumo por Cluster

Para estudiar la evolución del consumo en cada cluster, se agruparon los datos de consumo total por cluster y por año (2020-2024). El gráfico resultante muestra las tendencias de consumo en cada uno de los clusters, permitiendo identificar patrones a lo largo del tiempo.

4.7.2 Correlación de Consumo dentro de los Clusters

A continuación, se calculó la matriz de correlación entre los diferentes tipos de productos petrolíferos consumidos dentro de cada cluster. Este análisis permite identificar si el consumo de los productos está correlacionado o si cada tipo de producto sigue una tendencia independiente dentro del mismo cluster.

4.7.3 Consumo de Productos Petrolíferos por año en cada cluster

Para profundizar en el tipo de combustible consumido en cada cluster, se crearon gráficos de dona interactivos. Estos gráficos visualizan la distribución del consumo por tipo de producto en cada cluster, ofreciendo una representación clara de los productos predominantes en cada grupo de CCAA. Además, permiten observar cómo las regiones pueden estar alineadas o desalineadas con las tendencias de transición energética.

4.7.4 Análisis de Intensidad del Consumo

Finalmente, se generó un mapa de calor (heatmap) para evaluar la intensidad del consumo de productos petrolíferos por cluster y por tipo de producto. Este gráfico permite observar cómo los clusters varían en su consumo total de productos petrolíferos y cómo se distribuye ese consumo entre las diferentes categorías de productos en cada cluster. En términos generales, los clusters con mayor consumo de combustibles fósiles muestran una menor adopción de energías renovables, lo que resalta la oportunidad para una transición energética más acelerada en estas regiones (Waskom, 2021).

4.8 Selección y Optimizacion del Modelo Sarimax

4.8.1 Selección del Modelo de Predicción

Para el análisis y la predicción del consumo de petróleo en España, se evaluaron tres modelos principales: ARIMA, Prophet y SARIMAX. A continuación, se describen brevemente sus

características y las razones por las que se optó por SARIMAX como el modelo más adecuado.

4.8.1.1 ARIMA (AutoRegressive Integrated Moving Average)

- **Ventajas:** ARIMA es un modelo clásico ampliamente utilizado para series temporales no estacionarias. Es eficaz para predecir datos univariantes, como el consumo de petróleo, y ha sido utilizado en diversas aplicaciones económicas (Box, 2015).
- **Limitaciones:** A pesar de su utilidad en series temporales simples, ARIMA tiene dificultades para manejar la estacionalidad y no es ideal para series temporales con múltiples ciclos o fluctuaciones. Para este estudio, la serie presenta estacionalidad significativa, lo que hace que un modelo ARIMA básico no sea adecuado. Además, ARIMA no es particularmente eficiente al tratar grandes volúmenes de datos (Hyndman, 2018).

4.8.1.2 Prophet

- **Ventajas**: Desarrollado por Facebook, Prophet es un modelo robusto que maneja bien la estacionalidad diaria, semanal y anual, adecuado para series temporales con estos patrones. Además, permite ajustar efectos de días festivos o eventos especiales, una herramienta ideal si se busca incluir factores externos como políticas energéticas o fluctuaciones de mercado (Taylor, S. J., & Letham, B, 2018).
- **Limitaciones:** Aunque Prophet maneja bien la estacionalidad, tiene limitaciones cuando se enfrenta a series temporales con factores complejos o cuando se requieren ajustes más finos. Además, la capacidad para integrar variables exógenas de manera detallada es limitada, lo que podría ser un factor importante en el consumo de petróleo, ya que este está influenciado por variables externas como los precios del petróleo o las políticas energéticas (Yadav, 2022).

4.8.1.3 SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors)

- **Ventajas:** Manejo de Estacionalidad: SARIMAX extiende ARIMA para incorporar componentes estacionales, lo que lo hace perfecto para captar patrones cíclicos como los que se observan en el consumo de energía, que fluctúa con la estación (Hyndman, 2018).
- **Factores Exógenos:** Este modelo permite añadir variables externas, como cambios en políticas energéticas, precios del petróleo o avances en tecnologías renovables, datos necesarios para modelar el consumo de petróleo, ya que depende de factores tanto históricos como externos (Ade, 2023).
- **Adecuación para Series Temporales Complejas**: Dado que el análisis se realiza a nivel mensual, y el consumo de petróleo tiene tanto patrones estacionales como posibles influencias externas, SARIMAX es el modelo más adecuado para reflejar estas características complejas (Ade, 2023).

4.8.2 Modelo Seleccionado: SARIMAX

Se eligió SARIMAX por las siguientes razones:

- **Manejo de Estacionalidad:** El consumo de petróleo presenta claras fluctuaciones estacionales, y capturar esta variante es fundamental para la

predicción. SARIMAX se adapta perfectamente a esta necesidad, incorporando tanto estacionalidades como tendencias no estacionarias.

- **Capacidad para Incluir Factores Exógenos:** Aunque el conjunto de datos inicial no incluye variables externas, el consumo de petróleo está influenciado por factores externos como políticas energéticas o el crecimiento de las energías renovables. SARIMAX ofrece la flexibilidad de integrar estos factores, lo que significaría una mejora en la capacidad predictiva del modelo.
- **Adecuación para Series Complejas:** Dado que el consumo de petróleo es una serie mensual con patrones estacionales y posibles influencias externas, SARIMAX es el modelo que mejor captura la complejidad de los datos.

4.8.3 Optimización del Modelo (Búsqueda en Cuadrícula)

4.8.3.1 Evaluación del Rango de los Datos y Estadísticas Descriptivas

Como parte del análisis inicial, se exploró el rango de valores de la variable objetivo, Consumo Tm, y se calcularon estadísticas descriptivas importantes. Este análisis incluyó la generación de un histograma para visualizar la distribución de los datos y detectar posibles sesgos. Este paso tiene un gran valor debido a las siguientes razones:

- **Identificación de valores extremos:** Permite localizar outliers que podrían distorsionar el proceso de normalización o influir negativamente en el rendimiento del modelo. Por ejemplo, se detectaron valores extremos en los meses de enero y agosto, considerados para la interpretación de los resultados del modelo.
- **Comprensión de la distribución:** Ayuda a determinar si los datos siguen una distribución cercana a la normalidad o si presentan sesgos significativos. En este caso, la distribución mostró un ligero sesgo positivo, con valores concentrados alrededor de la media de 4,245,297 Tm y un rango de 3,241,264 a 4,655,847 Tm.
- **Establecimiento de expectativas:** Proporciona una referencia sobresaliente del rango de los datos, facilitando la evaluación de si los resultados del modelo son razonables y consistentes en el contexto del problema.

4.8.3.2 División de Datos

El conjunto de datos fue dividido en dos partes principales:

- **Conjunto de entrenamiento (80%):** Incluye el 80% de los datos históricos y se utilizó para ajustar y entrenar el modelo. Este porcentaje fue elegido siguiendo prácticas comunes en problemas de series temporales, ya que permite proporcionar al modelo suficiente información para identificar patrones subyacentes (Hyndman, 2018).
- **Conjunto de prueba (20%):** Se reservó el 20% restante para evaluar el desempeño del modelo con datos no utilizados durante el entrenamiento. Esta separación busca mitigar el riesgo de sobreajuste y medir la capacidad del modelo para generalizar a datos futuros (Hyndman, 2018).

4.8.3.3 Normalización de los Datos

Para garantizar la estabilidad numérica durante el proceso de optimización, se aplicó una normalización a la variable Consumo Tm utilizando el método StandardScaler (Pedregosa, 2011). Este proceso ajustó los datos para que tuvieran una media de 0 y una desviación estándar de 1.

Los parámetros de normalización se calcularon únicamente con el conjunto de entrenamiento y luego se aplicaron al conjunto de prueba. Esto asegura que el modelo no esté influenciado por información futura (Hyndman, 2018).

4.8.3.4 Búsqueda en Cuadrícula

Para encontrar la configuración óptima del modelo SARIMAX, se llevó a cabo una búsqueda exhaustiva combinando diferentes valores para los parámetros (Bergmeir, 2012). Este proceso buscó minimizar el valor del criterio de información Akaike (AIC), que evalúa la calidad del ajuste penalizando la complejidad excesiva del modelo (Akaike, 1974).

- **Exploración de múltiples configuraciones:** Se probaron combinaciones de órdenes autorregresivas, diferenciaciones y medias móviles, así como parámetros estacionales con una periodicidad mensual (Bergmeir, 2012).
- **Incorporación de estacionalidad:** El análisis sirvió para capturar patrones cíclicos mensuales presentes en los datos. Al incluir estos parámetros estacionales, el modelo logró identificar tendencias recurrentes específicas de cada mes.

El uso del AIC como métrica central permitió comparar modelos de forma eficiente, seleccionando finalmente la configuración con un AIC de 146.375. Esta configuración equilibró adecuadamente el ajuste y la simplicidad del modelo.

4.8.3.5 Validación Cruzada

Para evaluar la estabilidad y consistencia del modelo, se implementó una validación cruzada específica para series temporales utilizando TimeSeriesSplit (Bergmeir, 2012). Este método respeta la estructura temporal de los datos, evitando fugas de información y garantizando una evaluación objetiva.

- **Entrenamiento y evaluación por particiones:** El modelo fue ajustado en una porción creciente de los datos y evaluado en la siguiente partición temporal. En este procedimiento se asegura que el entrenamiento solo utilice información pasada.
- **Cálculo del RMSE promedio:** Se calculó el Error Cuadrático Medio (RMSE) para cada partición, promediando los resultados para obtener una métrica global de desempeño. Se identifican posibles variaciones en el rendimiento del modelo entre las particiones, garantizando que las predicciones fueran generalizables a nuevos datos.

4.9 Prediccion del Consumo con Sarimax para 2024-2026

En esta sección se describe el proceso de estimación del consumo de productos energéticos utilizando el modelo SARIMAX, con énfasis en las predicciones para el período de 2024 (último trimestre), 2025 e inicios de 2026.

4.9.1 Preparación de los Datos

El conjunto de datos utilizado corresponde al consumo energético en toneladas métricas (Consumo Tm), organizado mensualmente y desglosado por Comunidad Autónoma. Para garantizar la calidad de la serie temporal, se realizaron los siguientes pasos:

- Filtrado Temporal: Se seleccionaron los datos a partir de febrero de 2021, eliminando registros previos que pudieran estar influenciados por anomalías debido a la situación del COVID-19.
- Tratamiento de Valores Faltantes: A pesar de que no se detectaron valores faltantes en el dataset, se aplicó la imputación preventiva mediante el método de propagación hacia adelante.

4.9.2 Definición del Modelo SARIMAX

El modelo SARIMAX fue seleccionado por su capacidad para capturar tanto tendencias a largo plazo como fluctuaciones estacionales. Los parámetros elegidos, mediante el proceso de optimización descrito previamente, se utilizan para modelar las características complejas del consumo energético.

- **Orden del Modelo ARIMA:** Se configuró con los parámetros $(p=2, d=2, q=3)$, donde p es el número de retardos autorregresivos, d es el grado de diferenciación, y q se refiere al número de términos de medias móviles.
- **Orden Estacional:** Se definió el orden $(P=1, D=1, Q=1, s=12)$ para capturar la estacionalidad mensual en los datos, con $s=12$ representando la periodicidad anual.
- **Restricciones:** No se impusieron restricciones adicionales de estacionariedad, dado que los patrones estacionales fueron evidentes en los datos.

4.9.3 Generación de Predicciones

Se generaron predicciones mensuales para el período entre enero de 2024 y julio de 2026, con intervalos de confianza al 95%. Estos intervalos ayudan a contextualizar la incertidumbre inherente a las predicciones del modelo.

4.9.4 Visualización

Se desarrollaron las siguientes visualizaciones para ilustrar el comportamiento esperado del consumo energético:

- **Tendencia General del Consumo y Predicciones:** Se presentó un gráfico que incluye tanto los datos históricos como las proyecciones para los años 2024 a 2026, con los intervalos de confianza correspondientes.
- **Visualización Extendida del Dataset Completo:** Se graficó toda la serie temporal, integrando tanto los datos históricos como las proyecciones futuras.
- **Comparación de Consumo Real en 2024 vs. Predicciones (2024-2026):** Se compararon los valores reales de consumo en 2024 con las proyecciones del modelo, añadiendo una tabla resumen para facilitar el análisis.

5 Integración, pruebas y resultados

5.1 Integración de los Modelos y Herramientas

En este apartado, se explica cómo se integraron todas las fases del análisis para crear un flujo de trabajo coherente, desde el preprocesamiento de los datos hasta la predicción del consumo de petróleo, destacando la conexión entre K-means y SARIMAX.

5.1.1 Tamaño de la Muestra y Análisis

El dataset utilizado en este estudio, descargado de datos.gob.es, abarca el período 2020-2024 e incluye 41,491 registros mensuales sobre el consumo de productos petrolíferos en 19 comunidades autónomas y 52 provincias de España. Tras el preprocesamiento, se verificó la ausencia de valores faltantes y duplicados, garantizando la calidad del análisis.

Este conjunto de datos, con una cobertura temporal y geográfica amplia, permite analizar tendencias estacionales y diferencias regionales en el consumo, proporcionando una base sólida para la segmentación con K-means y la predicción con SARIMAX.

5.1.2 Tipo de Variables y Distribuciones

El dataset incluye variables categóricas y cuantitativas importantes para el análisis. Entre las primeras, se encuentran CCAA (19 comunidades autónomas), Provincia (52 valores únicos) y Tipo de Producto (14 categorías de productos petrolíferos, como GLP, biodiésel y gasolina). Las variables cuantitativas incluyen Consumo Tm (en toneladas métricas) y datos temporales (Año y Mes), fundamentales para el análisis de series temporales y la identificación de patrones estacionales. El consumo presenta una distribución asimétrica, con valores mayormente bajos o nulos y picos que alcanzan hasta 332,588 Tm. La media es de 5,592.29 Tm, con una mediana de 168 Tm y una alta desviación estándar (17,674.76 Tm), lo que evidencia una fuerte variabilidad entre regiones y productos. La presencia de valores nulos sugiere que algunos productos no se consumen de manera uniforme en todas las áreas o periodos, lo que es relevante para interpretar los patrones de consumo.

5.1.3 Estadísticas Resumidas

Un análisis resumido del consumo (variable Consumo Tm) indica:

- Media: Representa el consumo promedio por registro en el conjunto de datos.
- Mediana: Al ser significativamente menor que la media, esto sugiere que la distribución está sesgada hacia valores bajos, lo que puede indicar que los productos no se consumen de manera constante en todas las regiones o meses.
- Desviación Estándar: Muestra una gran variabilidad entre los registros, lo que refuerza la idea de que el consumo es muy variable, dependiendo de la región y del tipo de producto.
- Máximos y Mínimos: Estos valores identifican meses o regiones con patrones de consumo extremos, como picos de consumo durante ciertas épocas del año, lo que puede ser útil para identificar eventos atípicos o anomalías.

	Métrica	Consumo Tm
0	Recuento	41491.000000
1	Media	5592.290000
2	Mediana	168.000000
3	Desviación Estándar	17674.760000
4	Mínimo	0.000000
5	Máximo	332588.000000
6	Percentil 25	0.000000
7	Percentil 75	3200.500000

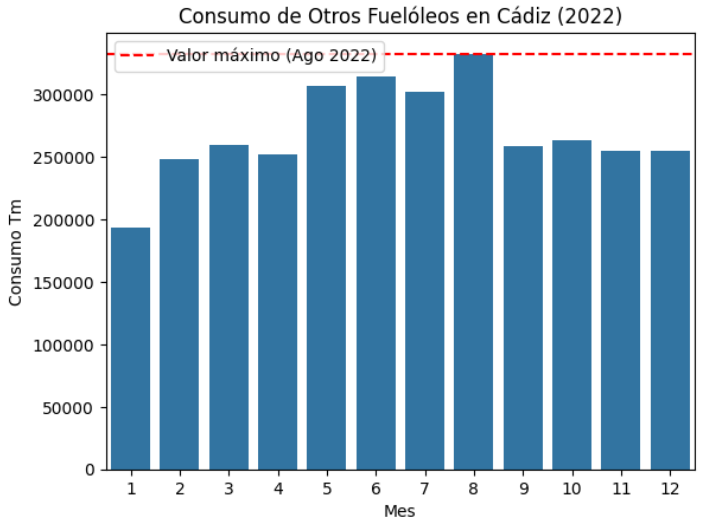
Tabla 3: Resumen Estadisticas

5.1 Visualización Inicial del Dataset

Para facilitar la comprensión del conjunto de datos y sus características, se generaron varias visualizaciones iniciales utilizando herramientas como Matplotlib, Seaborn y Plotly. Estas visualizaciones no solo permiten identificar patrones y tendencias, sino también posibles valores atípicos que podrían influir en los modelos predictivos.

5.1.1 Histograma del Consumo Tm:

La distribución del consumo muestra valores mayoritariamente bajos, con algunos picos significativos, como el valor máximo de 332,588 Tm registrado en Cádiz en 2022. Este análisis permite identificar posibles valores atípicos.



	Fecha	CCAA	Provincia	Tipo Producto	Consumo Tm	Año	Mes
20961	2022-08-01	ANDALUCIA	CADIZ	OTROS FUELOLEOS	332588	2022	8

Ilustración 12:Histograma para visualizar valores atipicos

5.1.2 Gráfico de Consumo Anual por Producto:

Se observan variaciones en el consumo de productos como gasolina, gasóleo y GLP entre 2020 y 2024. Los patrones estacionales identificados serán de gran utilidad para referencias para el análisis de consumo en los posteriores cluster.

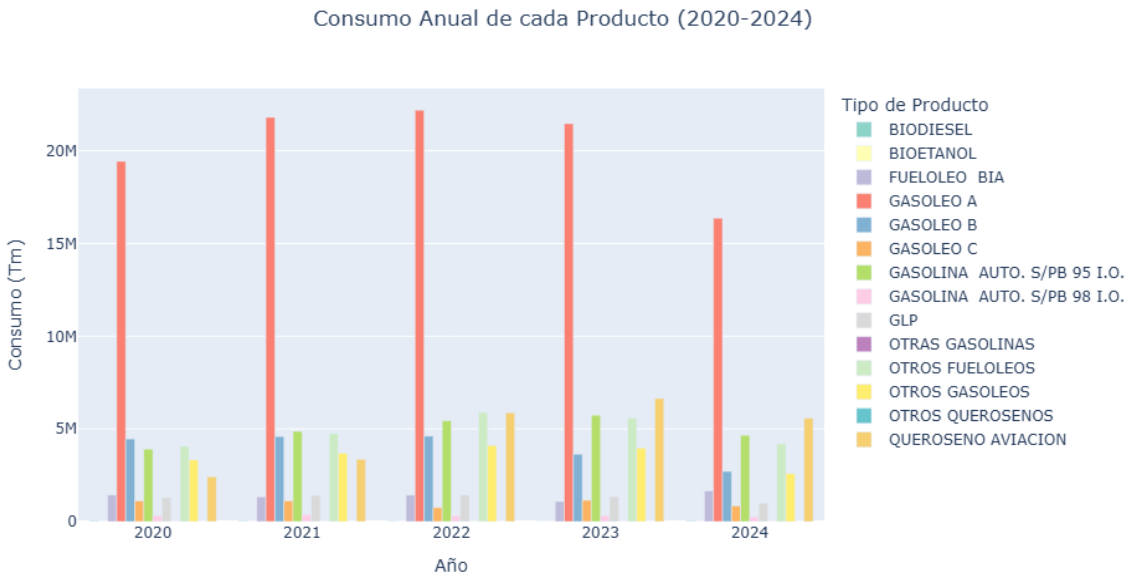


Ilustración 13: Consumo Anual de cada Producto derivado del petróleo en cinco años

El análisis del consumo energético en España entre 2020 y 2024 evidencia una alta dependencia de los combustibles fósiles, especialmente en el transporte y la industria. El Gasóleo A destaca como el principal combustible terrestre, superando los 22 millones de toneladas métricas en 2022, impulsado por la recuperación económica tras la pandemia (CORES, 2022; OBS Business School, 2025). Esta dependencia resalta la vulnerabilidad del sector ante variaciones en precios y disponibilidad.

El Gasóleo B, utilizado en sectores agrícolas y pesqueros, mantiene una demanda estable de entre 4 y 5 millones de toneladas anuales, mientras que el Gasóleo C, empleado en calefacción, muestra fluctuaciones marcadas por factores climáticos y estacionales (Ministerio de Industria, Turismo y Comercio, 2007).

Las gasolinas, en particular la Sin Plomo 95 I.O., han seguido una tendencia de crecimiento hasta 2024, cuando se observa un descenso, posiblemente ligado a la adopción de vehículos eléctricos e híbridos, cambios en hábitos de movilidad y mejoras en la eficiencia de los motores (KPMG, 2024).

El consumo de Queroseno de Aviación, tras mínimos históricos en 2020, creció hasta 2023, alcanzando 6.6 millones de toneladas métricas, pero en 2024 sufre una caída que podría estar relacionada con factores económicos o la introducción de combustibles sostenibles para la aviación (SAF), aunque aún sin impacto significativo en los datos.

Las energías renovables continúan con una adopción limitada. El biodiésel muestra un crecimiento hasta 2024, pasando de 6,017 a 34,632 toneladas métricas, aunque sigue representando una fracción menor del consumo total debido a la falta de incentivos y limitaciones en infraestructura. El bioetanol, en cambio, mantiene un consumo marginal, principalmente como aditivo en gasolinas.

Finalmente, productos como los Otros Fuelóleos y Otros Gasóleos, con aplicaciones en la industria pesada y transporte en zonas remotas, presentan un consumo estable. El repunte de

los Otros Fuelóleos en 2022 podría estar vinculado a necesidades industriales específicas en un contexto económico favorable.

5.1.3 Pruebas en Patrones Estacionales del Gasóleo C

El Gasóleo C, utilizado mayormente en calefacción, muestra una marcada variabilidad mensual a lo largo de los años, lo que refleja su alta sensibilidad a factores climáticos y estacionales. Este patrón es considerado como un ejemplo para modelar con precisión el consumo en el contexto de SARIMAX, ya que la estacionalidad es un componente clave del modelo y las pruebas sobre este combustible sirven para demostrar que los productos en este conjunto de datos tienen una alta relación con la estacionalidad.

Para ilustrar este comportamiento, a continuación, se presenta un gráfico que muestra la variabilidad mensual en el consumo de Gasóleo C durante el período 2020-2024:

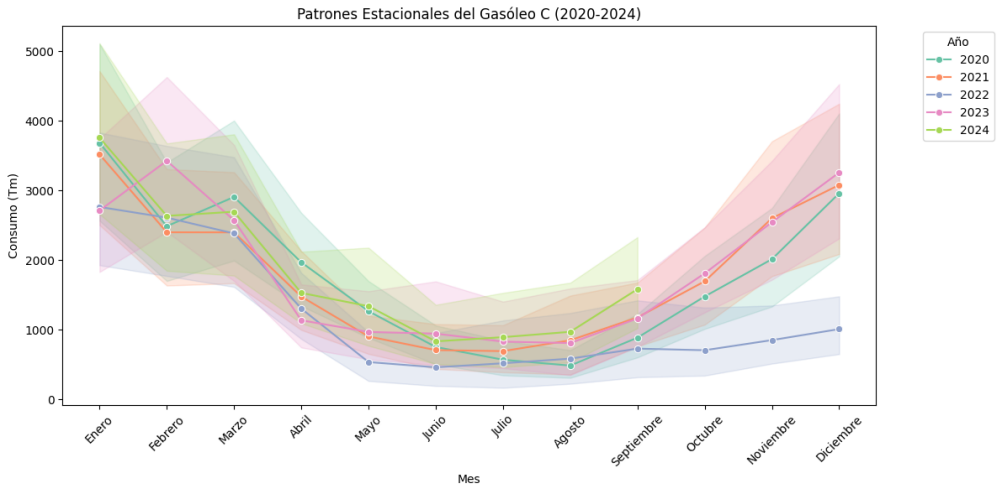


Ilustración 14: Patrones Estacionales del Gasoleo C

El consumo de Gasóleo C a lo largo de los años (2020-2024) muestra una clara estacionalidad y variabilidad de un año a otro, como se puede ver en el gráfico de líneas que presenta los datos mensuales.

5.1.3.1 Patrón Estacional

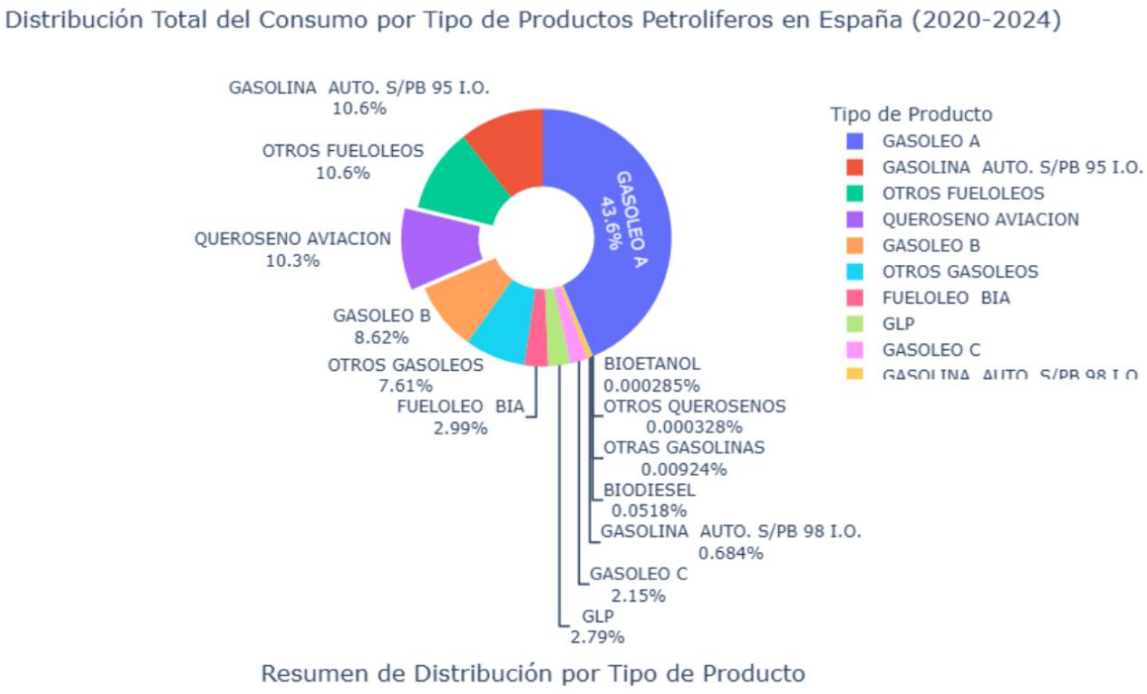
El consumo de Gasóleo C muestra un patrón estacional claro, con incrementos en invierno (enero, diciembre) y descensos en verano (junio-agosto), reflejando su uso predominante en calefacción. Esta estacionalidad es clave para la descomposición de series temporales en el modelo SARIMAX, permitiendo capturar variaciones naturales en su demanda. Durante el período analizado, el consumo presentó fluctuaciones anuales significativas. En 2020, sufrió una fuerte caída debido a las restricciones por la pandemia de COVID-19, especialmente en primavera y verano. A partir de 2021 y 2022, se estabiliza, aunque sin recuperar plenamente los niveles previos a la crisis.

En 2023, se observa una recuperación impulsada por la reactivación económica y el aumento de la demanda de calefacción en invierno. Sin embargo, en 2024, el consumo vuelve a descender, posiblemente debido a la transición hacia energías más limpias, promovida por políticas energéticas más estrictas y un mayor enfoque en eficiencia energética. Dado que el Gasóleo C está fuertemente influenciado por factores climáticos y estacionales, identificar

estos patrones con precisión en el modelo SARIMAX es un perfecto objeto de estudio para mejorar la predicción de su consumo futuro.

5.1.4 Distribución del Consumo de por Producto:

Un gráfico circular resume cómo se reparte el consumo total de petróleo y sus derivados entre distintos productos en estos últimos cinco años, ofreciendo un analisis que servirá como base para futuras segmentaciones utilizando K-means.



Resumen de Distribución por Tipo de Producto

Tipo de Producto	Consumo Total (Tm)	Porcentaje (%)
GASOLEO A	101,265,128.00	43.64%
GASOLINA AUTO. S/PB 95 I.O.	24,613,957.00	10.61%
OTROS FUELOLEOS	24,492,028.00	10.56%
QUEROSENO AVIACION	23,864,691.00	10.29%
GASOLEO B	20,001,624.00	8.62%
OTROS GASOLEOS	17,655,346.00	7.61%
FUELOLEO BIA	6,941,487.00	2.99%
GLP	6,477,094.00	2.79%
GASOLEO C	4,987,796.00	2.15%
GASOLINA AUTO. S/PB 98 I.O.	1,587,459.00	0.68%
BIODIESEL	120,296.00	0.05%
OTRAS GASOLINAS	21,451.00	0.01%
OTROS QUEROSENOS	760.00	0.00%
BIOETANOL	662.00	0.00%

Ilustración 15: Distribucion del consumo total de productos del petróleo y sus derivados

5.1.5 Comparación de Consumo por Año y Mes:

Gráfico de líneas destacan las fluctuaciones mensuales del consumo, lo que facilita la detección de tendencias y la identificación de meses con consumos inusualmente altos o bajos.

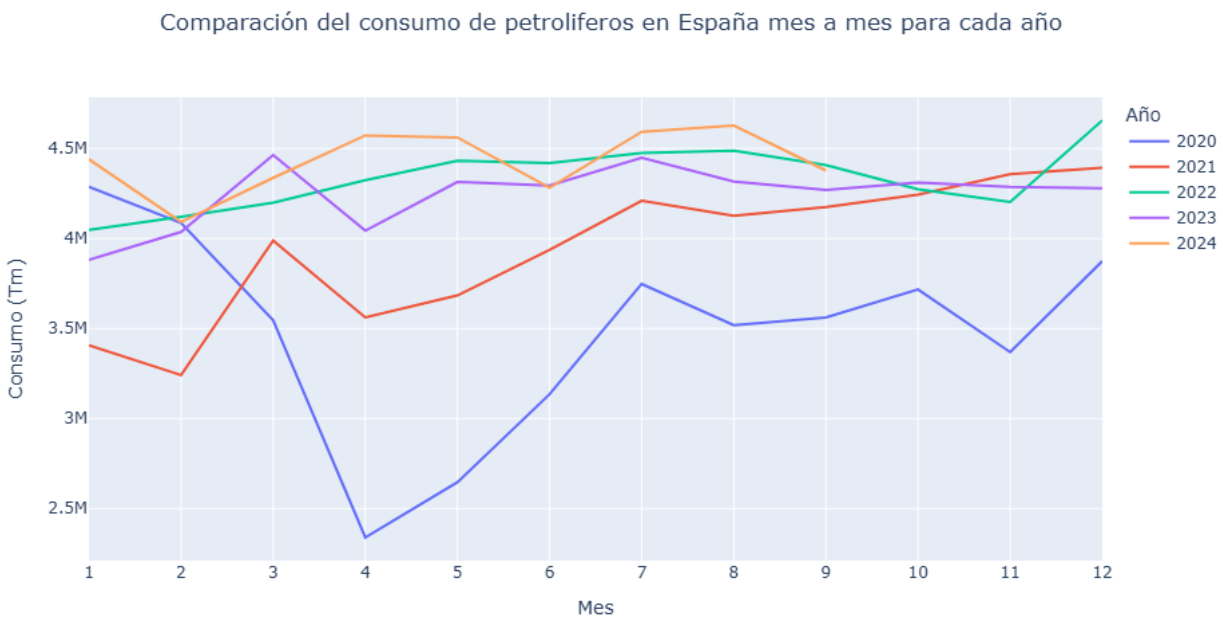


Ilustración 16:Comparacion del consumo de Petroliferos en España mes a mes cada año

Al analizar las fluctuaciones mensuales en el consumo de energía, se observa que ciertos meses han presentado picos de consumo récord en los últimos cinco años en 2024, especialmente en enero, abril, mayo, julio y agosto. Estos aumentos son fácilmente identificables en los gráficos de líneas, detectando patrones inusuales y el poder explorar posibles factores que los expliquen. Aunque no se profundiza en las causas, se puede asociar con lo investigado anteriormente en este estudio, la posibilidad de que estos meses estén relacionados con variaciones estacionales en la demanda energética, como las altas necesidades durante el invierno (enero) o el comienzo del verano (julio y agosto), lo que subraya la importancia de estos meses en el comportamiento del consumo energético.

5.1.6 Mapa de Correlación por CCAA

Este análisis de correlación proporciona una visión clara sobre las similitudes entre las comunidades autónomas en cuanto a los patrones de consumo de energía. Las regiones con correlaciones más altas, como Andalucía y Canarias (0.966359) o Castilla y León con Cantabria (0.983439), sugieren que pueden tener comportamientos de consumo similares, lo que puede ser útil para segmentar las comunidades autónomas en clusters durante el análisis con el modelo K-means.

Para obtener una representación visual de estas correlaciones, se podría generar un mapa de calor, donde las zonas con valores cercanos a 1 indican fuertes correlaciones positivas, mientras que valores cercanos a -1 indican correlaciones negativas.

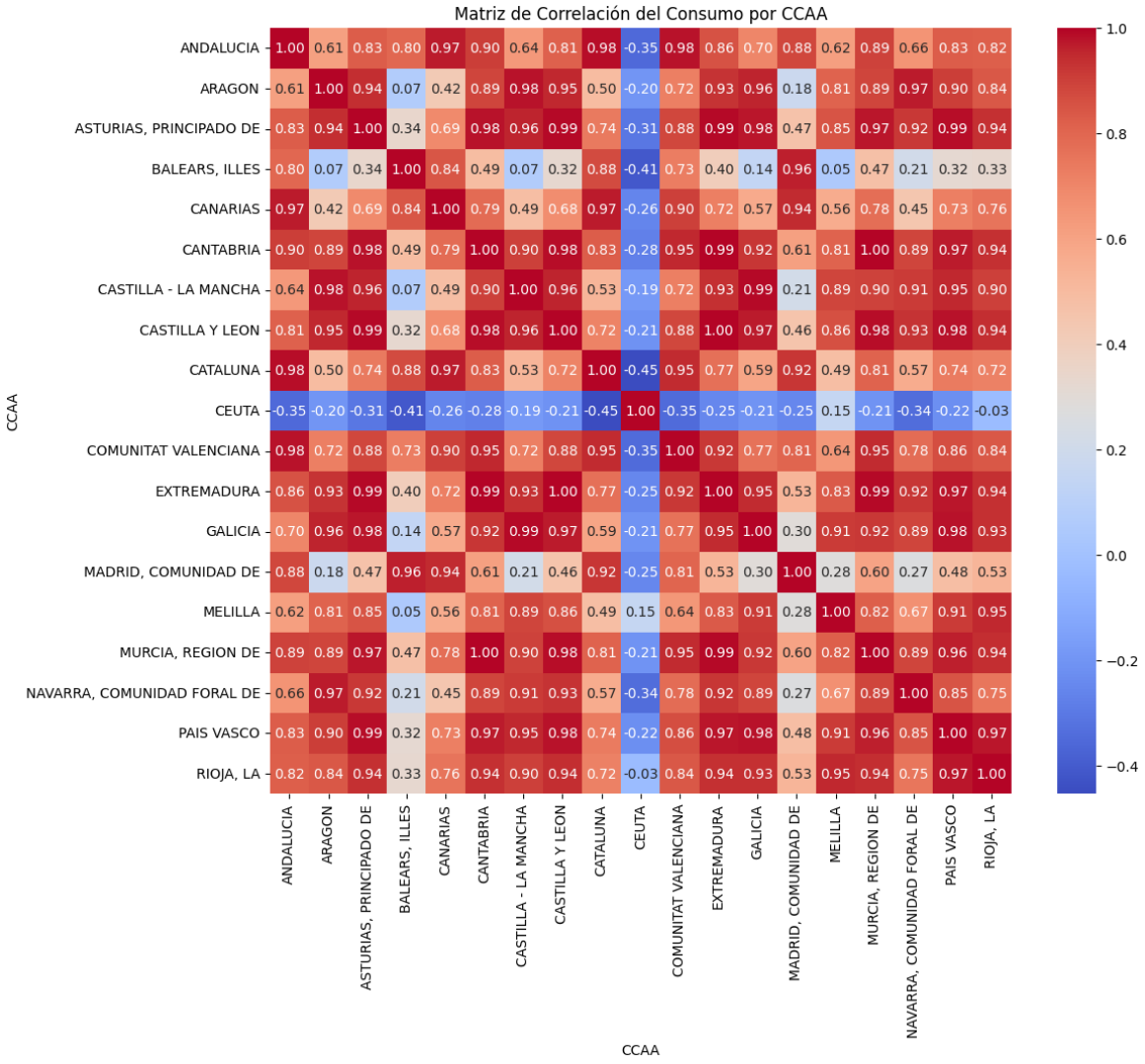


Ilustración 17: Matriz de Correlación por cada Comunidad Autónoma

5.2 K-Means como clasificador de consumo por comunidades

En esta sección se lleva a cabo un análisis detallado de los clusters generados mediante el algoritmo K-means para el consumo de productos petrolíferos en las diferentes comunidades autónomas de España. El propósito de este análisis es identificar patrones regionales en el consumo de estos productos y explorar cómo estos patrones se relacionan con las oportunidades de transición hacia energías renovables en cada región.

5.2.1 Método del Codo y Silhouette Score

Para determinar el número óptimo de clusters para segmentar las comunidades autónomas, se llevó a cabo un análisis utilizando el método del codo y el Silhouette Score. Los resultados indicaron que 10 clusters eran la mejor opción, ya que lograban un buen equilibrio entre la simplicidad del modelo y su capacidad para capturar patrones significativos. En el método del codo, se observó que la reducción de la inercia comenzaba a estabilizarse precisamente en este punto, pero para reducir el número de variables a estudiar, se consideró reducir el número de clúster a 8.

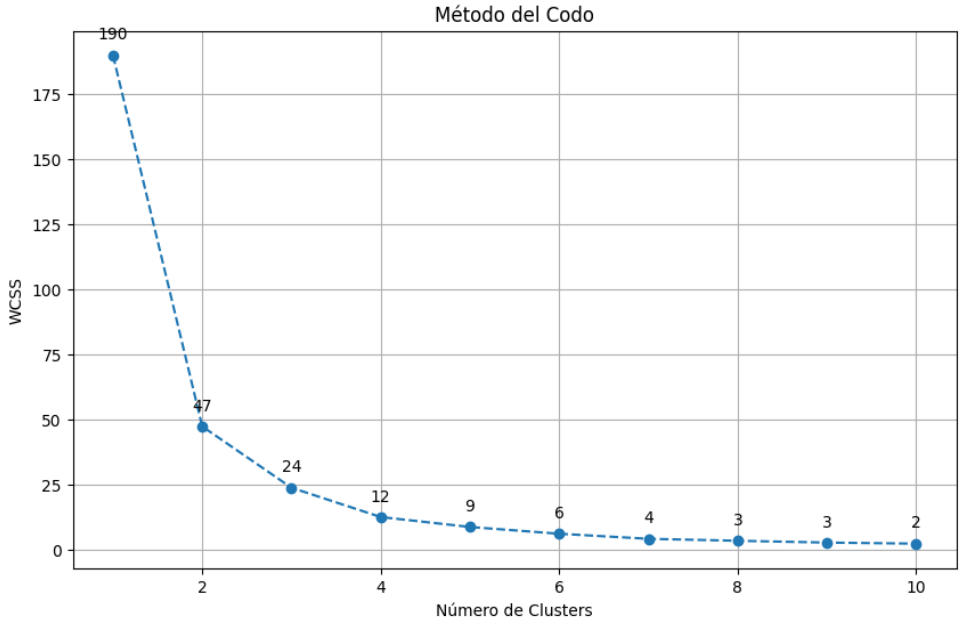


Ilustración 18: Metodo del codo

Para asegurar que esta decisión de reducir el cluster a 8, no afecatria los resultados, se procedio a aplicar el metodo Silhouette Score, este obtuvo un valor de 0.68, lo que refleja una buena calidad en la separación de los clusters. Esto significa que las comunidades autónomas agrupadas dentro de un mismo cluster comparten características similares, mientras que los distintos clusters presentan diferencias claras entre ellos. Este resultado valida la efectividad del modelo, no solo para la segmentación regional, sino también para desarrollar estrategias específicas de transición energética basadas en los patrones identificados.

```
Silhouette Score para el número de clusters 8: 0.5666129650073602
```

Ilustración 19: Resultado de Silhouette Score

5.2.2 Creación de los Clusters

Una vez determinado que el número óptimo de clusters era 8, se procedió a aplicar el algoritmo K-means para segmentar las comunidades autónomas (CCAA) según el consumo total de productos petrolíferos ("Consumo Tm"). Para preparar los datos, se organizó la información utilizando una tabla dinámica, lo que permitió estructurar los datos de manera adecuada para su análisis. Con los datos ya escalados, el modelo de K-means fue entrenado para formar 8 clusters. El algoritmo asignó a cada comunidad autónoma uno de estos grupos, y los resultados se guardaron en un DataFrame llamado "ccaa_clusters", donde se relacionan las comunidades autónomas con su respectivo cluster.

A continuación, se presenta la distribución de las comunidades autónomas según los clusters obtenidos:

```
# Mostrar las regiones agrupadas
print(ccaa_clusters.sort_values(by="Cluster"))
```

	CCAA	Cluster
12	GALICIA	0
17	PAIS VASCO	0
6	CASTILLA - LA MANCHA	0
7	CASTILLA Y LEON	0
8	CATALUNA	1
9	CEUTA	2
16	NAVARRA, COMUNIDAD FORAL DE	2
14	MELILLA	2
18	RIOJA, LA	2
2	ASTURIAS, PRINCIPADO DE	2
5	CANTABRIA	2
13	MADRID, COMUNIDAD DE	3
0	ANDALUCIA	4
4	CANARIAS	5
11	EXTREMADURA	6
3	BALEARS, ILLES	6
15	MURCIA, REGION DE	6
1	ARAGON	6
10	COMUNITAT VALENCIANA	7

Tabla 4: Distribucion de cada Comunidad Autónoma con K-means

5.2.3 Mapa de Dispersion de los Cluster

Se creó un mapa de dispersión para visualizar cómo se agrupan las comunidades autónomas según su consumo de productos petrolíferos. En este gráfico, cada punto representa una comunidad autónoma, y los colores indican el cluster al que pertenece. También se marcaron los centroides de cada grupo, lo que ayuda a identificar los centros de los clusters. Este gráfico ofrece una visión clara de las relaciones entre las regiones y resalta las características centrales de cada grupo, facilitando la interpretación de los patrones regionales.

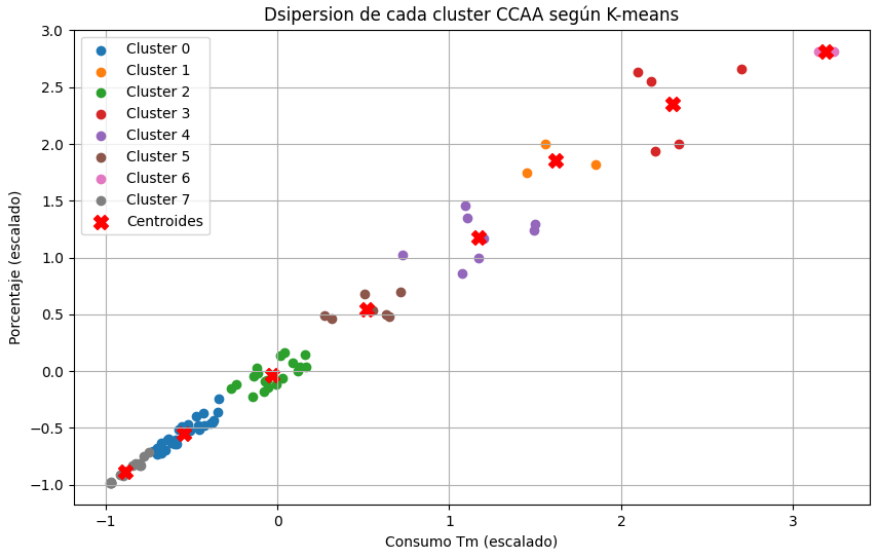


Ilustración 20: Mapa de Dispersión de cada Cluster según K-means

5.2.4 Diferencia de consumo en cada Cluster

Después de agrupar las comunidades autónomas en distintos clusters según sus patrones de consumo, este gráfico de barras permite visualizar de manera clara el consumo total de productos petrolíferos en cada grupo. Su importancia radica en que muestra cómo varía la demanda energética entre los clusters, facilitando la identificación de las diferencias de consumo en cada uno de ellos. En la visualización, los clusters aparecen en el siguiente orden:

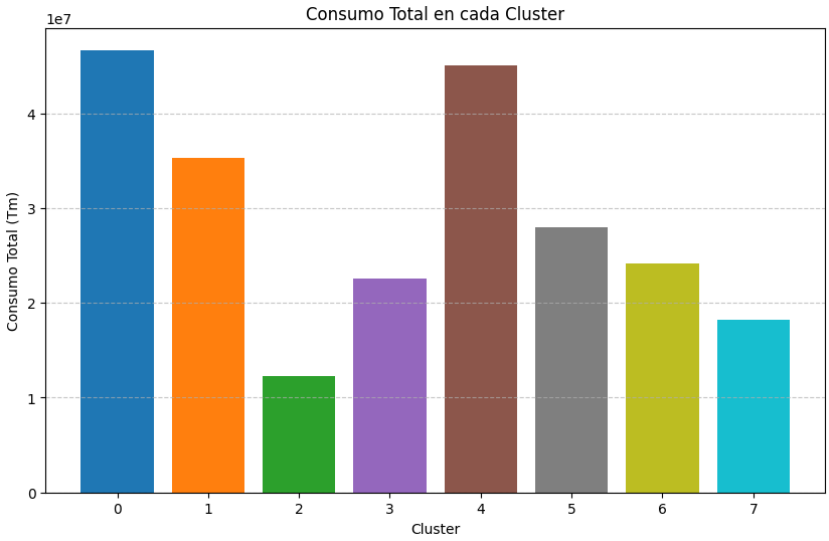


Ilustración 21: Consumo total de Petroliferos en cada Cluster en los últimos 5 años

5.2.5 Mapa de Calor del Consumo por Cluster

Para entender mejor cómo se distribuye el consumo de productos petrolíferos en cada cluster, se ha elaborado un mapa de calor. Esta herramienta es ideal para visualizar de manera clara qué combustibles se usan con mayor intensidad en cada grupo, ayudándonos a identificar patrones de consumo regionales y a evaluar la dependencia de ciertos combustibles fósiles. Con esta información, podemos sentar las bases para futuras estrategias de transición hacia energías más sostenibles.

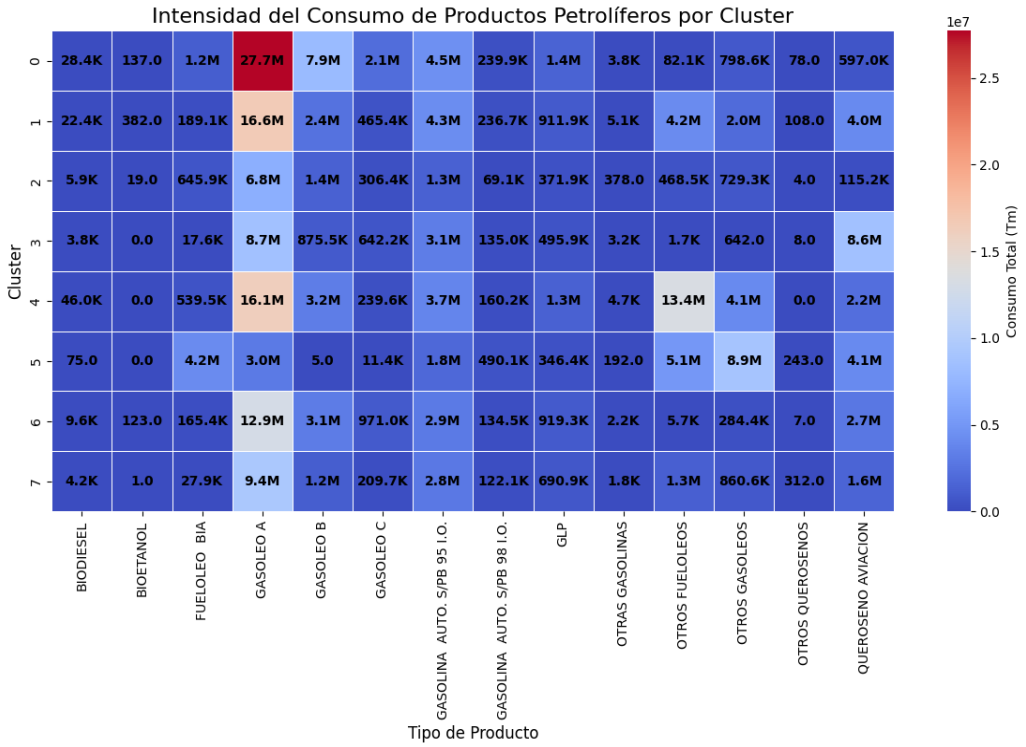


Ilustración 22: Mapa de Calor entre los Cluster y los Productos derivados del Petroleo

5.2.5.1 ¿Qué nos dice el Mapa de Calor?

El análisis revela diferencias significativas en el consumo de productos petrolíferos entre los distintos clusters. Algunos grupos muestran una alta dependencia de ciertos combustibles, mientras que otros presentan un uso más equilibrado, un ejemplo a estudiar es el cluster 0.

Cluster 0: Es el de mayor consumo de Gasóleo A (27,720,036 Tm) y también registra un uso considerable de Gasóleo B y C. Además, muestra un consumo elevado de Fueloóleo BIA (1,188,840 Tm) y, aunque en menor medida, presencia de biocombustibles como Biodiésel (28,354 Tm) y Bioetanol (137 Tm). Su fuerte dependencia de los gasóleos indica que podría beneficiarse de estrategias para reducir su uso y fomentar alternativas más sostenibles. Puede notarse como otros clusters presentan patrones de consumo diferenciados:

- **Cluster 1:** Alto consumo de Gasóleo A (16,577,438 Tm), pero con menor dependencia de Gasóleo B y C en comparación con el Cluster 0. Destaca su uso de Queroseno de Aviación (4,020,610 Tm).
- **Cluster 2:** Consumo moderado de Gasóleo A (6,837,276 Tm) y menor uso de biocombustibles. También muestra un consumo considerable de Fueloóleo BIA (645,937 Tm).
- **Cluster 3:** Se distingue por su alto consumo de Queroseno de Aviación (8,555,405 Tm), lo que sugiere una fuerte vinculación con aeropuertos. Su uso de otros productos petrolíferos es relativamente bajo.
- **Cluster 4:** Elevado consumo de Gasóleo A (16,149,540 Tm) y una importante presencia de Otros Gasóleos (13,352,372 Tm), lo que indica una mayor diversificación en los combustibles utilizados.
- **Cluster 5:** Es el grupo con el mayor consumo de Fueloóleo BIA (4,167,230 Tm) y Otros Fueloóleos (5,078,744 Tm), lo que sugiere una fuerte dependencia industrial de estos combustibles.
- **Cluster 6:** Consumo significativo de Gasóleo A (12,942,562 Tm), con una distribución más equilibrada de otros productos petrolíferos.
- **Cluster 7:** Notable consumo de Gasóleo A (9,400,070 Tm) y un uso moderado de Otros Gasóleos (1,287,937 Tm), lo que indica una dependencia energética considerable.

5.2.5.2 Implicaciones y Oportunidades para la Transición Energética

El análisis del mapa de calor nos da una visión detallada del consumo de combustibles fósiles en cada cluster, esto permite el poder identificar oportunidades para avanzar hacia una matriz energética más sostenible. Algunas de esas oportunidades podrian ser:

- **Cluster 0:** Dado su alto consumo de gasóleos, es prioritario analizar estrategias para reducir su impacto ambiental.
- **Clusters con alta dependencia de gasóleos:** Podrían beneficiarse de incentivos para la adopción de combustibles alternativos, como el biodiésel, o de medidas para fomentar la electrificación de ciertos sectores.
- **Cluster 3 (Queroseno de Aviación):** Destaca la necesidad de desarrollar combustibles más sostenibles para la aviación.
- **Clusters con elevado uso de Fueloóleos (como el Cluster 5):** Podrían explorar alternativas como biocombustibles industriales o electrificación de procesos productivos.

5.2.6 Análisis del Cluster 0: Un Caso Representativo

Para comprender mejor el comportamiento del consumo de productos petrolíferos en los distintos clusters, hemos tomado el Cluster 0 como caso de estudio. Este grupo destaca por su elevado uso de Gasóleo A, B y C, lo que lo convierte en un referente para analizar la dependencia de estos combustibles y evaluar posibles estrategias de transición hacia fuentes de energía más sostenibles.

Además, para enriquecer el análisis, se ha integrado información del Boletín Estadístico de Hidrocarburos de la CORE y del Informe OBS: El sector energético en España, hacia una descarbonización sostenible. Esto permite situar los resultados en un contexto más amplio y comprender mejor las tendencias de consumo dentro del panorama energético nacional.

El estudio del Cluster 0 se centrará en tres aspectos importantes:

5.2.6.1 Relación entre los distintos combustibles

Para comprender mejor cómo se comporta el consumo de productos petrolíferos dentro del Cluster 0, se ha analizado la correlación entre las comunidades autónomas que lo conforman: Castilla-La Mancha, Castilla y León, Galicia y el País Vasco.

Correlación de Consumo: Se identifican las relaciones entre los distintos productos petrolíferos en las comunidades autónomas agrupadas en el Cluster 0. A través de la matriz de correlación, se observa que existen fuertes correlaciones entre los consumos de productos en diversas comunidades, como se evidencia en la alta correlación entre Castilla-La Mancha y Galicia (0.99), lo que sugiere patrones de consumo similares entre estas regiones.

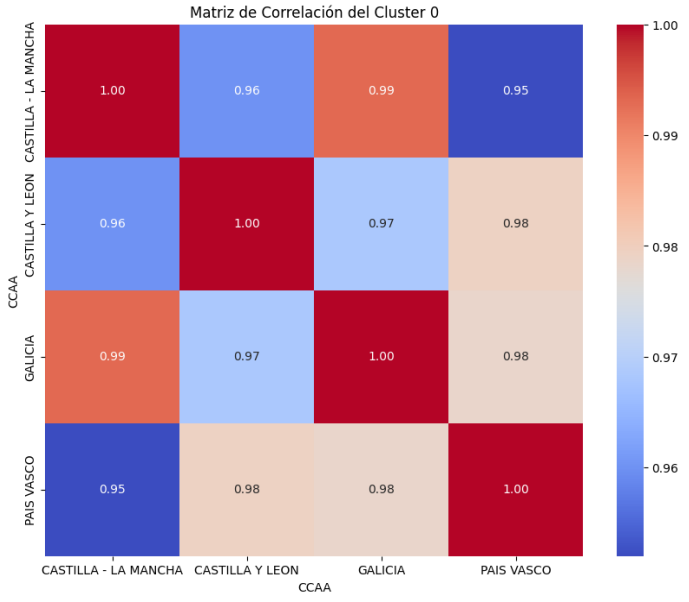


Ilustración 23: Matriz Correlación del Cluster 0

Evolución del Consumo Mes a Mes cada Año: Se analizan las tendencias de consumo mes a mes a lo largo de los años dentro del Cluster 0. Los datos muestran fluctuaciones significativas en los consumos, con un aumento general en los meses de invierno, alcanzando picos en diciembre. En los años 2020 y 2021, los consumos presentan una tendencia al alza, mientras que en 2023 y 2024, se observa una ligera disminución en los primeros meses del año, con un repunte hacia el final del periodo. Esta evolución resalta la estacionalidad del

consumo, que parece estar influenciada por factores climáticos y otras variaciones regionales.

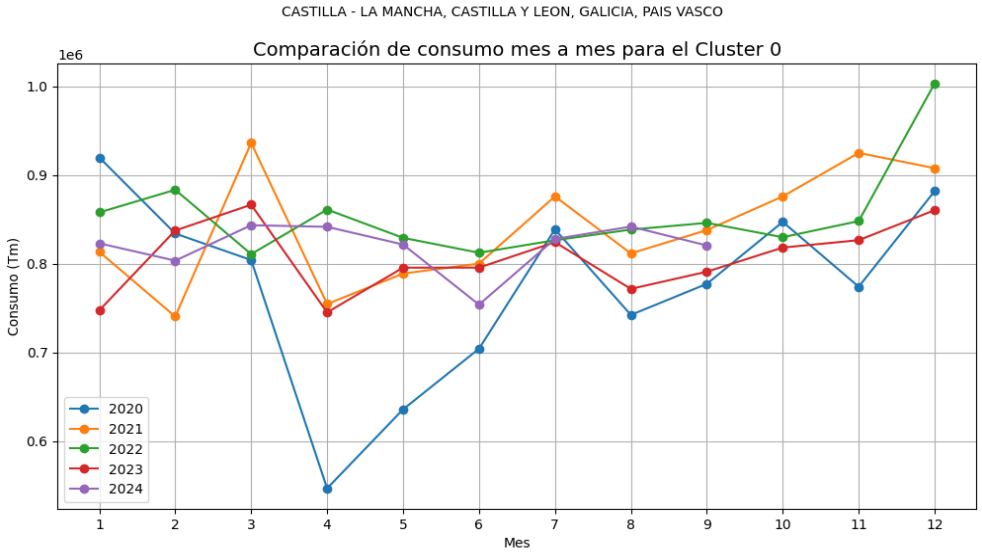


Ilustración 24: Comparacion de consumo total de petróleo mes a mes cada año

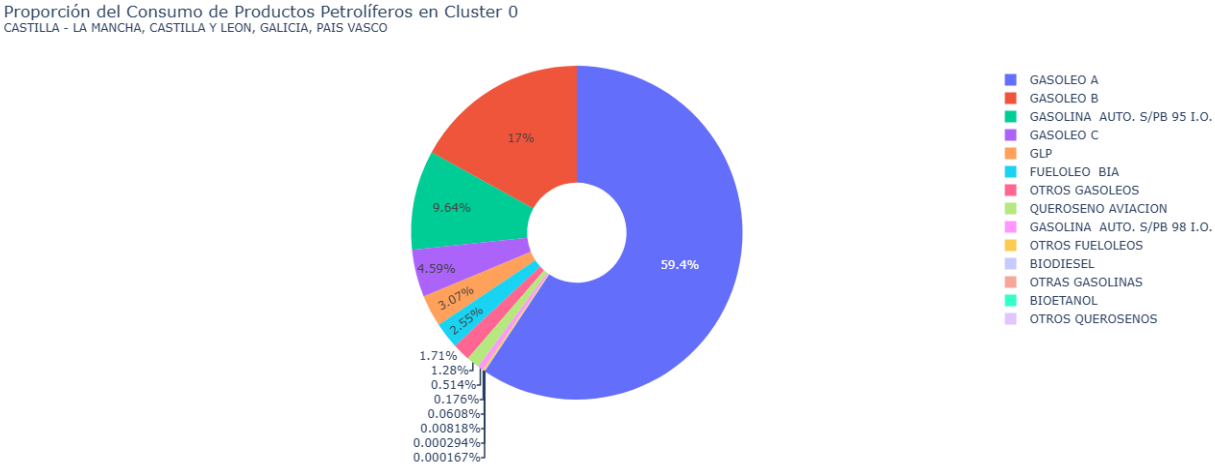
```
# Mostrar la tabla resultante
print(cluster_0_pivot)
```

```
Mes          1         2         3         4         5         6         7  \
Año
2020  919655.0  833936.0  804035.0  546657.0  635705.0  704066.0  838682.0
2021  813267.0  740382.0  936180.0  754490.0  788784.0  800024.0  875752.0
2022  857721.0  883262.0  810725.0  860971.0  829089.0  812295.0  826355.0
2023  747349.0  837353.0  866415.0  745042.0  795464.0  795520.0  824202.0
2024  822904.0  803362.0  843328.0  841496.0  821569.0  753774.0  827830.0

Mes          8         9        10        11        12
Año
2020  742210.0  777180.0  847068.0  773980.0   881762.0
2021  811526.0  837791.0  875761.0  924804.0   907639.0
2022  838379.0  845971.0  829791.0  847949.0  1003096.0
2023  771563.0  790907.0  818103.0  826454.0   860255.0
2024  841748.0  820485.0       NaN       NaN        NaN
```

Tabla 5: Distribucion de consumo total mes a mes cada año del cluster 0

Distribución del Consumo por Tipo de Producto: Se identifica cómo el consumo de diferentes tipos de productos petrolíferos varía dentro del Cluster 0. El Gasóleo A es el combustible más utilizado, con un aumento constante a lo largo de los años, especialmente en 2022, mientras que otros productos como el Biodiesel y el Gasóleo C muestran menores volúmenes de consumo. Este patrón indica que los productos más consumidos tienen un impacto significativo en la transición energética, dado su uso predominante en sectores como la calefacción y el transporte. Los productos alternativos como el Biodiesel y el Bioetanol presentan un crecimiento modesto, lo que podría reflejar el inicio de un cambio hacia energías más limpias, aunque aún limitado en comparación con los combustibles fósiles tradicionales.



Tipo de Producto	Consumo Total (Tm)
GASOLEO A	27720036
GASOLEO B	7936259
GASOLINA AUTO. S/PB 95 I.O.	4500417
GASOLEO C	2142153
GLP	1434322
FUELOLEO BIA	1188840
OTROS GASOLEOS	798608
QUEROSENO AVIACION	597044
GASOLINA AUTO. S/PB 98 I.O.	239875
OTROS FUELOLEOS	82124
BIODIESEL	28354
OTRAS GASOLINAS	3816
BIOETANOL	137
OTROS QUEROSENOS	78

Ilustración 25: Proporcion del consumo de Productos petrolíferos en el cluster 0

Consumo de cada combustible a través de los años en el cluster 0

Año	2020	2021	2022	2023	2024
Tipo Producto					
BIODIESEL	12257	4341	9	2246	9501
BIOETANOL	25	38	20	21	33
FUELOLEO BIA	358750	299068	301518	149136	80368
GASOLEO A	5355554	5954912	6085031	5832835	4491704
GASOLEO B	1794672	1836154	1855833	1414981	1034619
GASOLEO C	496700	464925	292174	512717	375637
GASOLINA AUTO. S/PB 95 I.O.	718137	888845	979175	1046612	867648
GASOLINA AUTO. S/PB 98 I.O.	50084	56320	45635	47760	40076
GLP	297612	310649	317385	300473	208203
OTRAS GASOLINAS	584	826	802	885	719
OTROS FUELOLEOS	2472	16470	19094	30231	13857
OTROS GASOLEOS	159174	151712	196703	169817	121202
OTROS QUEROSENOS	13	18	18	19	10
QUEROSENO AVIACION	58902	82122	152207	170894	132919

Tabla 6: Consumo de cada combustible a través de los años en el cluster 0

5.2.7 Oportunidades de Transición a Energías Renovables en el Cluster 0

5.2.7.1 Sustitución progresiva del Gasóleo A por Biocombustibles

Dado que el Gasóleo A es el producto más consumido en este cluster, una estrategia clave sería incrementar el uso de biocombustibles (biodiésel y bioetanol), los cuales presentan un crecimiento modesto. Se podrían diseñar incentivos para el uso de mezclas con mayor

contenido de biodiésel en transporte y calefacción, reduciendo progresivamente el uso de combustibles fósiles.

¿Cómo contribuye SARIMAX?
El modelo puede simular escenarios en los que se incrementa el uso de biodiésel y analizar su impacto en la reducción del consumo de gasóleo, permitiendo estimar en qué medida la transición es factible y en qué plazos.

5.2.7.2 Estrategias de eficiencia energética en calefacción y transporte

La tendencia estacional del consumo, con picos en invierno, sugiere que una gran parte del consumo de gasóleo está relacionada con calefacción y transporte. Algunas alternativas Alternativas:

- Mayor impulso a sistemas de calefacción basados en bombas de calor y energía geotérmica en viviendas y edificios públicos.
- Sustitución de flotas de transporte diésel por vehículos eléctricos o impulsados por hidrógeno verde.

¿Cómo contribuye SARIMAX?
Se pueden modelar escenarios con una reducción progresiva del consumo de gasóleo debido a la adopción de calefacción más eficiente y evaluar su impacto en la demanda total de productos petrolíferos.

5.2.7.3 Desarrollo de infraestructura renovable en zonas rurales

Castilla-La Mancha y Castilla y León poseen un gran potencial para la energía solar y eólica, lo que permitiría reemplazar parte del consumo de gasóleo con electrificación renovable en sectores como el agrícola y el transporte.

Oportunidades regionales:
- Galicia se destaca como una de las comunidades autónomas líderes en generación de energía renovable, con un alto potencial en energía eólica y marina.
- El País Vasco, junto con Cantabria, Madrid, Baleares, La Rioja y Canarias, contribuye con solo el 3,7% de la producción renovable nacional, lo que sugiere oportunidades de mejora.

¿Cómo contribuye SARIMAX?
Permite evaluar si una mayor electrificación impactaría en la reducción del consumo de gasóleo en el cluster y predecir cómo estos cambios afectarían la demanda energética regional.

5.2.7.4 Políticas de transición basadas en el consumo regional

Distribución Regional:
- Castilla y León, Castilla-La Mancha, Andalucía, Aragón, Galicia y Extremadura generan el 82% de la energía renovable en España.
- Madrid es la región más deficitaria en términos de producción de energía renovable.

Contexto: La alta correlación entre Castilla-La Mancha y Galicia indica que estas regiones podrían beneficiarse de políticas conjuntas para acelerar la transición.

Propuestas: Programas como incentivos fiscales para energía renovable o restricciones progresivas al uso de combustibles fósiles pueden diseñarse en función de la tendencia de consumo observada.

¿Cómo contribuye SARIMAX?
Puede simular escenarios con diferentes niveles de intervención política (subsidios a renovables, impuestos al diésel) y proyectar cómo cambiaría el consumo de productos petrolíferos a futuro.

5.2.7.5 Potencial de Adaptación del País Vasco a Estrategias de Transición Energética

A pesar de que el País Vasco no lidera la generación de energías renovables como otras comunidades del Cluster 0, su correlación con regiones con un alto consumo de hidrocarburos sugiere que podría beneficiarse de estrategias similares en la transición energética.

Factores clave:

- Su industria y transporte aún dependen en gran medida de combustibles fósiles, lo que lo convierte en un candidato para medidas de reducción progresiva.
- Su capacidad de innovación tecnológica y su infraestructura industrial podrían facilitar la adopción de soluciones como el hidrógeno verde, eficiencia energética en procesos industriales y electrificación del transporte.

Estrategias de aplicación:

- Incentivos a la electrificación del transporte y la logística industrial.
- Desarrollo de parques eólicos offshore en colaboración con comunidades como Galicia.
- Políticas de eficiencia energética adaptadas a su perfil industrial y urbano.

¿Cómo contribuye SARIMAX?
El modelo puede evaluar en qué medida la implementación de estas estrategias impactaría en la reducción del consumo de hidrocarburos en el País Vasco y proyectar escenarios de transición más alineados con el resto del Cluster 0.

5.2.8 Oportunidades de Transición a Energías Renovables en el Cluster 2

Como un ejemplo contrarestante al Cluster 0 y que se puede visualizar gracias a la segmentacion, el Cluster 2 presenta una evolución distinta en sus patrones de consumo energético, reflejada en su matriz de correlación y en las características de cada comunidad. Si bien algunas regiones, como Ceuta y Melilla, presentan patrones de consumo algo desvinculados del resto, existen oportunidades de transición hacia energías renovables que pueden abordarse de manera específica según el contexto de cada territorio.

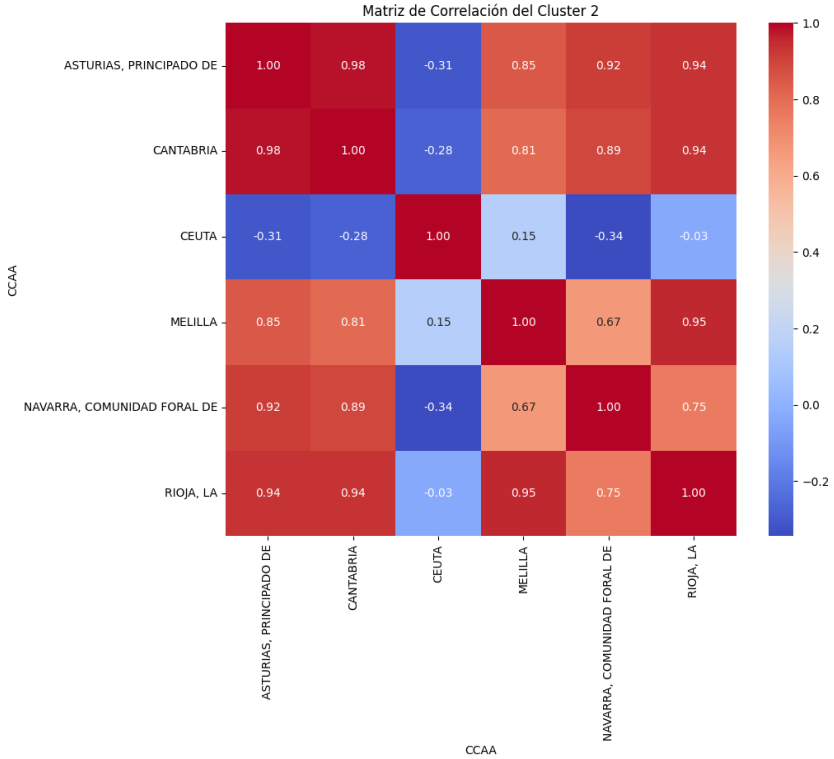
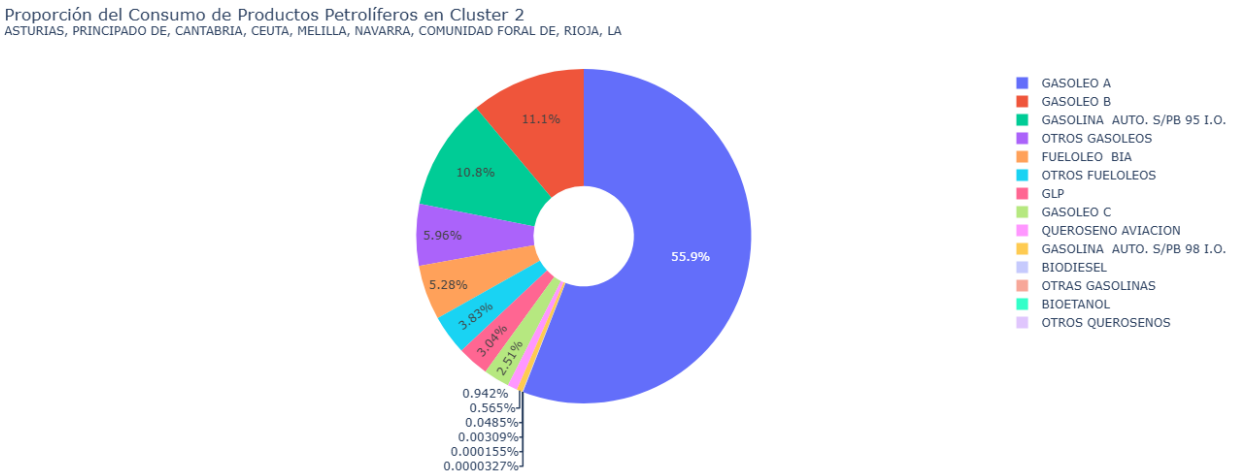


Ilustración 26: Matriz de correlación del Cluster 2

5.2.8.1 Sustitución de combustibles fósiles por alternativas renovables

A pesar de la prevalencia del Gasóleo A, algunas regiones dentro del cluster tienen el potencial de avanzar en la sustitución de este combustible por opciones renovables. En Navarra y La Rioja, por ejemplo, la conexión más estrecha con otras comunidades permite implementar con mayor facilidad la adopción de biocombustibles, así como explorar soluciones de electrificación renovable en áreas como el transporte y la calefacción. Estas regiones ya están avanzando en la integración de energía solar y biomasa, lo que facilita una transición hacia alternativas más limpias. En el caso de Ceuta y Melilla, la transición debe ser más personalizada, considerando su dependencia de infraestructuras limitadas.



Tipo de Producto	Consumo Total (Tm)
GASOLEO A	6837276
GASOLEO B	1357966
GASOLINA AUTO. S/PB 95 I.O.	1319572
OTROS GASOLEOS	729339
FUELOLEO BIA	645937
OTROS FUELOLEOS	468464
GLP	371888
GASOLEO C	306432
QUEROSENO AVIACION	115175
GASOLINA AUTO. S/PB 98 I.O.	69082
BIODIESEL	5933
OTRAS GASOLINAS	378
BIOETANOL	19
OTROS QUEROSENOS	4

Ilustración 27: Proporción del Consumo de Productos Petroliferos en el Cluster 2

Consumo de cada combustible a través de los años en el cluster 2

Año	2020	2021	2022	2023	2024
Tipo Producto					
BIODIESEL	412	0	6	300	5215
BIOETANOL	0	10	0	9	0
FUELOLEO BIA	140091	136113	174905	121383	73445
GASOLEO A	1295363	1536118	1462743	1442564	1100488
GASOLEO B	294628	313260	317816	243561	188701
GASOLEO C	75516	77610	45553	62274	45479
GASOLINA AUTO. S/PB 95 I.O.	207761	274904	286370	302772	247765
GASOLINA AUTO. S/PB 98 I.O.	14892	16589	13032	13469	11100
GLP	69659	77582	96812	76349	51486
OTRAS GASOLINAS	57	77	80	72	92
OTROS FUELOLEOS	115321	91316	90783	94709	76335
OTROS GASOLEOS	130083	145681	127994	156768	168813
OTROS QUEROSENOS	3	0	0	1	0
QUEROSENO AVIACION	12150	15405	30834	33202	23584

Ilustración 28: Comparacion de consumo de combustibles a trabes de los años en el cluster 2

¿Cómo contribuye SARIMAX?
El modelo SARIMAX puede analizar escenarios de transición, proyectando el impacto del uso de biocombustibles y otras energías renovables en el consumo energético, lo cual es un dato especifico y necesario para evaluar el tiempo y la viabilidad de estos cambios.

5.2.8.2 Potencial de eficiencia energética y electrificación en transporte y calefacción

La electrificación y la mejora de la eficiencia energética son claves en regiones con una dependencia significativa del Gasóleo A. En Navarra y La Rioja, donde existe un perfil industrial y de transporte destacado, iniciativas como la electrificación del transporte o el uso de tecnologías de calefacción eficientes pueden acelerar la reducción del consumo de combustibles fósiles. Mientras tanto, Ceuta y Melilla requieren soluciones más específicas debido a su infraestructura limitada.

¿Cómo contribuye SARIMAX?
SARIMAX permite proyectar cómo la implementación de tecnologías eficientes podría modificar la demanda energética a corto y largo plazo, evaluando el impacto de políticas de electrificación y eficiencia.

5.2.8.3 Infraestructura renovable: áreas de mejora y diversificación

En regiones como Navarra y La Rioja, con buenas condiciones para la energía eólica y solar, el desarrollo de infraestructuras renovables es una oportunidad clara para reducir el consumo de combustibles fósiles. Navarra, con su potencial para proyectos eólicos, y La Rioja, con el

creciente uso de instalaciones fotovoltaicas, están bien posicionadas para diversificar su matriz energética. Por otro lado, Ceuta y Melilla enfrentan retos de interconexión con la península, pero el desarrollo de microrredes renovables podría ser una alternativa efectiva para integrar energías renovables locales, principalmente solares.

¿Cómo contribuye SARIMAX?
SARIMAX puede simular los efectos de proyectos renovables en el consumo de combustibles fósiles, proporcionando predicciones sobre cómo la infraestructura renovable influirá en la reducción de la dependencia de los combustibles fósiles.

5.2.8.4 Políticas de transición adaptadas al consumo regional

La transición energética en el Cluster 2 se puede hacer más eficiente al diseñar políticas adaptadas a las necesidades particulares de cada comunidad. Navarra y La Rioja pueden beneficiarse de incentivos fiscales y subsidios para la adopción de energías renovables, impulsando aún más su liderazgo en energías limpias. En regiones como Ceuta y Melilla, políticas que favorezcan la interconexión eléctrica o el desarrollo de microrredes podrían ser una vía viable para incorporar energías renovables y mejorar la sostenibilidad local.

¿Cómo contribuye SARIMAX?
El modelo SARIMAX es útil para evaluar el impacto de diversas políticas en las comunidades del Cluster 2, simulando diferentes intervenciones y proyectando cómo estas pueden modificar los patrones de consumo energético, especialmente en lo relativo a la reducción del uso de productos petrolíferos.
El análisis de clusters ha concluido en la importancia de entender los distintos patrones de consumo energético en las comunidades autónomas de España. Al agrupar las regiones según sus características, hemos podido identificar tanto sus similitudes como sus particularidades, lo que nos proporciona una base sólida para desarrollar estrategias de transición energética más personalizadas y efectivas. Ahora que hemos terminado este análisis, el siguiente paso es avanzar con el modelo SARIMAX, que nos permitirá simular cómo diferentes políticas y escenarios de transición hacia energías renovables impactarán en el consumo energético en el futuro, tanto a corto como a largo plazo.

5.3 Decisión en el uso de Sarimax

5.3.1 Consideracion del Modelo ARIMA

Durante el proceso de selección de modelos, se consideró ARIMA como una opción inicial para predecir el consumo mensual de petróleo. Este modelo, reconocido por su capacidad para modelar series temporales univariantes no estacionarias, fue evaluado con el objetivo de identificar patrones y realizar predicciones a corto plazo. Sin embargo, los resultados obtenidos mostraron limitaciones significativas para este caso de estudio.

5.3.1.1 División de los Datos

Para evaluar el desempeño del modelo ARIMA, se dividieron los datos en dos subconjuntos:

- Entrenamiento: Incluye todas las observaciones disponibles, excepto los últimos 12 meses.
- Prueba: Contiene los últimos 12 meses del periodo analizado.
- Esta división permitió entrenar el modelo con datos históricos y validar su capacidad predictiva en un horizonte de un año.

5.3.1.2 Configuración y Entrenamiento del Modelo

El modelo inicial fue ajustado utilizando los parámetros (p,d,q) = (1,1,1), donde:

- **p=1:** Número de retardos autorregresivos.
- **d=1:** Diferenciación necesaria para convertir la serie en estacionaria.
- **q=1:** Número de términos de promedio móvil.

El entrenamiento se realizó empleando el conjunto de datos de entrenamiento, y se generaron predicciones para los últimos 12 meses.

5.3.1.3 Visualización de Resultados

Las predicciones del modelo fueron comparadas con los datos reales de prueba a través de un gráfico que incluía:

- La serie de datos históricos de entrenamiento.
- Los valores reales de prueba.
- Las predicciones del modelo ARIMA, junto con sus intervalos de confianza al 95%.
- Aunque las predicciones mostraron una tendencia general que seguía parcialmente el comportamiento histórico, el modelo no fue capaz de capturar adecuadamente las fluctuaciones mensuales observadas en los datos reales.

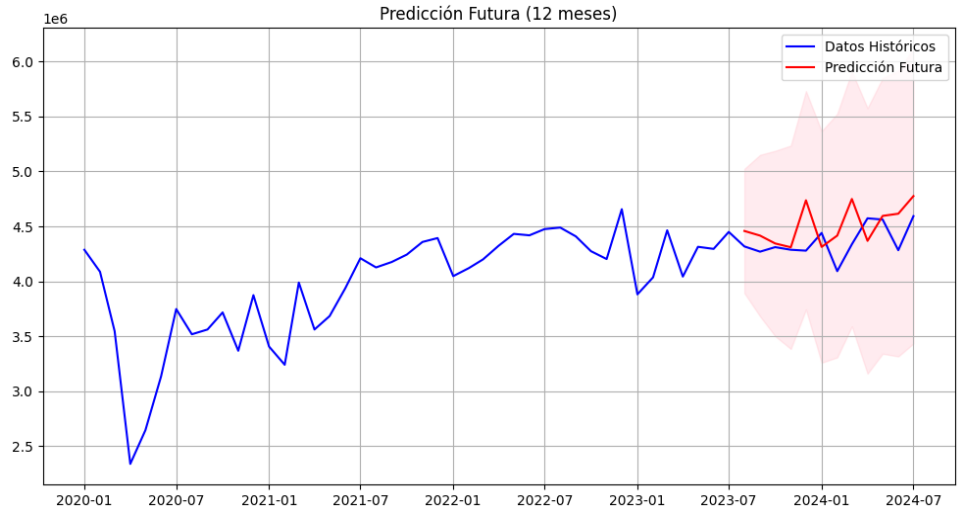


Ilustración 29: Prediccion con el modelo ARIMA

4. Evaluación del Desempeño
Para cuantificar la precisión del modelo, se utilizó la métrica RMSE (Root Mean Squared Error), obteniendo un valor de 246,600.66. Este error refleja una desviación significativa entre las predicciones y los valores reales, lo que indica que el modelo no es adecuado para realizar pronósticos precisos en este caso.

```
# Calcular error
rmse = np.sqrt(mean_squared_error(test, predicted_mean))
print(f"RMSE: {rmse}")

RMSE: 246600.65879383724
```

Ilustración 30: Resultados de RMSE para ARIMA

5.3.1.4 Limitaciones Encontradas

El modelo ARIMA presentó las siguientes limitaciones durante su implementación:

- Incapacidad para Capturar la Estacionalidad: A pesar de ser adecuado para datos univariantes no estacionarios, ARIMA no maneja de manera eficiente patrones estacionales, los cuales son una característica clave del consumo de petróleo en este análisis.
- Altas Fluctuaciones en los Datos: El consumo mensual presenta variaciones significativas que ARIMA no logró modelar adecuadamente, incluso después de probar diferentes configuraciones de hiperparámetros (p,d,q).
- Error Significativo: El alto valor de RMSE refleja la incapacidad del modelo para proporcionar predicciones útiles para fines prácticos.

5.3.1.5 Conclusión sobre ARIMA

Aunque ARIMA es un modelo robusto para ciertos tipos de series temporales, no se adaptó correctamente al comportamiento estacional y complejo del consumo mensual de petróleo en este estudio. Por estas razones, se descartó su uso en favor de modelos más adecuados, como SARIMAX, que ofrecen una mejor capacidad para capturar patrones estacionales y, potencialmente, incorporar variables exógenas.

5.3.2 Reflexión Personal sobre la Selección del Modelo

Al principio, pensé que ARIMA y Prophet serían las opciones más adecuadas, ya que el conjunto de datos contenía solo cinco columnas y no tenía variables externas. Sin embargo, al probar ARIMA en otros proyectos, me di cuenta de que no era ideal para trabajar con series temporales complejas, como las que tienen estacionalidad. Aunque ARIMA es útil en series simples, no detectaba patrones significativos en este caso, lo que me llevó a descartar este modelo.

Por otro lado, Prophet fue recomendado por varios compañeros, pero como no estaba familiarizado con la programación de este modelo, tuve problemas al implementarlo. Aunque parecía una herramienta interesante, las pruebas realizadas resultaron en errores y no pude obtener un modelo funcional. Por esta razón, decidí dejar de lado Prophet y buscar una alternativa más accesible.

Finalmente, al probar SARIMAX, encontré más documentación y apoyo de materia disponible otorgado por el Master. Este modelo resultó ser más robusto y eficiente, y su programación era más comprensible para mí. Aunque mi principal desafío fue interpretar los resultados y considerar la falta de variables exógenas, el modelo permitió identificar patrones estacionales claros en el consumo de petróleo, lo que me dio una base sólida para continuar con el análisis.

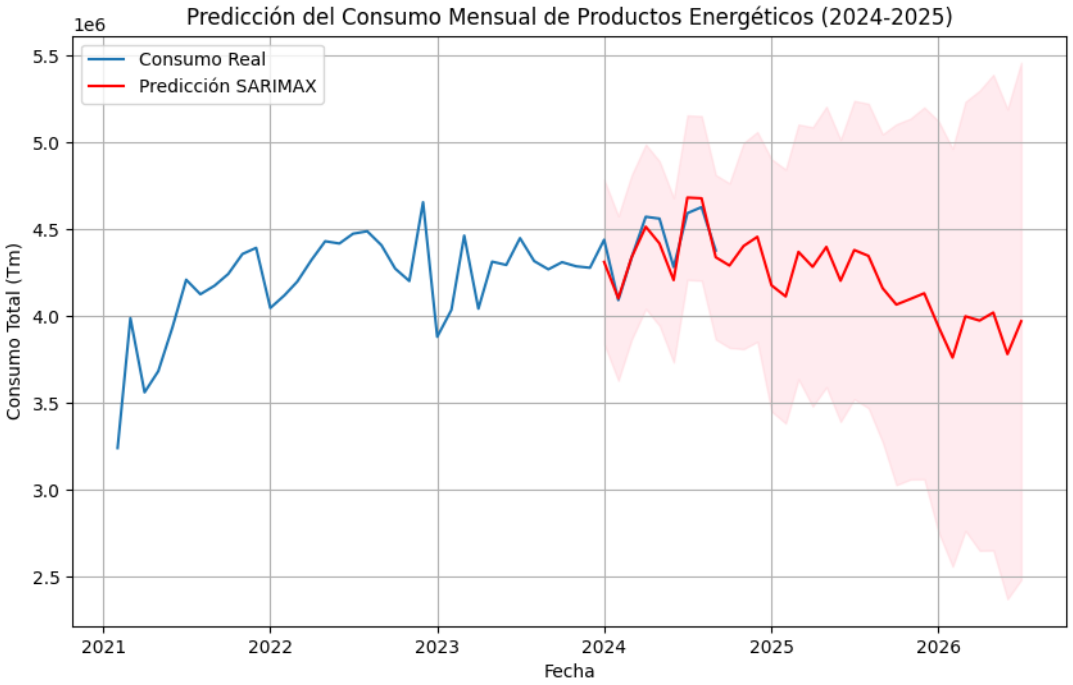


Ilustración 31: Prediccion del consumo mensual de Productos Petrolíferos

5.3.2.1 Lecciones Aprendidas y Mejoras Futuras

Una de las lecciones más importantes de este proceso fue la necesidad de considerar las características específicas del conjunto de datos y los desafíos asociados a trabajar con grandes volúmenes de información. En el futuro, sería interesante mejorar el modelo SARIMAX incorporando variables exógenas, como políticas energéticas, fluctuaciones en los precios del petróleo o avances en tecnologías renovables. Esto enriquecería las predicciones y ofrecería una interpretación más precisa de los resultados.

5.4 Modelo SARIMAX

5.4.1 Preparación

5.4.1.1 Filtrado de los Datos:

Se filtraron los datos para excluir las fechas anteriores a febrero de 2021, debido a que la pandemia de COVID-19 alteró significativamente los patrones de consumo en ese período. Esta medida garantiza que el análisis se base en datos más representativos de la situación actual, eliminando distorsiones causadas por comportamientos atípicos durante el confinamiento. Excluir los datos de 2020 y principios de 2021 fue una decisión acertada, ya que estos datos podrían haber afectado la precisión de las predicciones.

5.4.1.2 Agrupación y Visualización:

Acción: Los datos se agruparon por fecha, sumando el consumo mensual. Luego, se visualizó la serie temporal en un gráfico de líneas para identificar patrones y fluctuaciones a lo largo del tiempo.

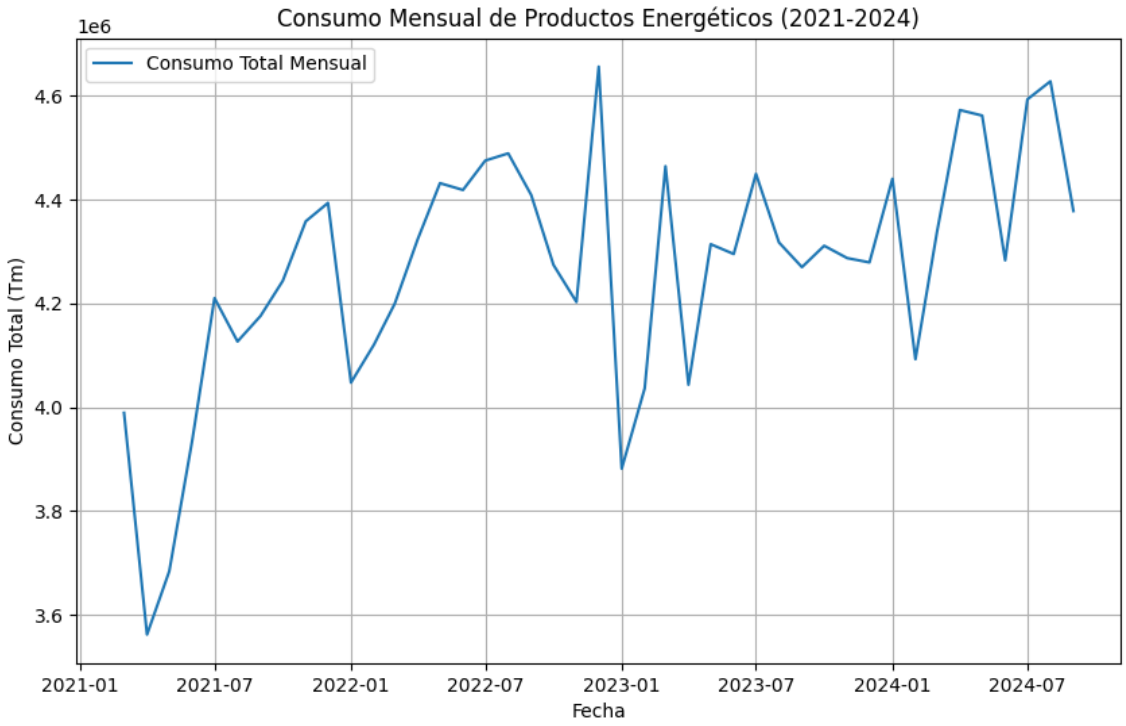


Ilustración 32: Fluctuaciones provocadas por los eventos historicos

Este gráfico de líneas es ideal para observar las tendencias y variaciones estacionales del consumo. Esto permitió identificar picos en meses clave, como durante el invierno, pero también evidenció fluctuaciones atípicas, no solo de las fluctuaciones provocadas por la pandemia COVID-19 sino también relacionadas con eventos globales:

- Febrero 2022 - Impacto de la guerra en Ucrania: La invasión de Ucrania en febrero de 2022 causó inestabilidad en los mercados energéticos, lo que llevó a una mayor volatilidad en los precios de los combustibles fósiles. Aunque el consumo de energía no reflejó directamente los precios, la incertidumbre y los altos costos probablemente influyeron en el comportamiento de los consumidores, que ajustaron su consumo en respuesta. (Yang, 2023)
- Marzo a junio 2022 - Aumento de precios y reconfiguración del mercado energético: Entre marzo y junio de 2022, los precios de la energía aumentaron considerablemente, lo que llevó a una disminución en el consumo en algunos sectores y cambios en las estrategias de ahorro energético. A pesar de la mayor demanda durante los meses de verano, la inestabilidad económica generó fluctuaciones significativas en los patrones de consumo (Yang, 2023).
- 2023 - Continuación de la guerra y crisis energética: En 2023, las fluctuaciones en el consumo continuaron debido a la prolongación del conflicto en Ucrania y la crisis energética mundial. Los precios inestables de la energía generaron picos y valles en el consumo, especialmente en sectores como la industria y el residencial. Aunque los patrones seguían siendo inestables, se empezó a notar un ajuste gradual en respuesta a los precios elevados y las políticas energéticas adoptadas (Yang, 2023).

Estos picos fueron esenciales para ajustar los parámetros del modelo y minimizar su impacto en las predicciones.

5.4.1.3 Distribución del consumo:

Se calcularon estadísticas descriptivas, como la media, mediana, desviación estándar y cuartiles, para entender mejor el comportamiento del consumo de energía.

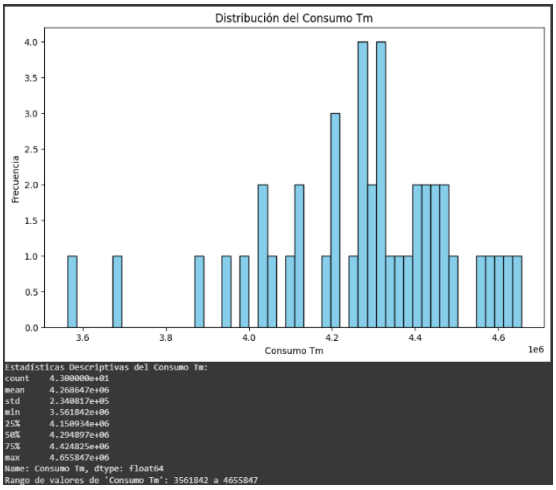


Ilustración 33: Distribucion del Consumo total (Tm)

Las estadísticas descriptivas proporcionaron una visión clara de la variabilidad del consumo. Con una desviación estándar alta, de aproximadamente 234,081 Tm, se observó una considerable fluctuación en los datos. Este análisis ayudó a identificar posibles valores atípicos que podrían haber influido en las predicciones, lo que llevó a ajustar el modelo para reflejar mejor esa variabilidad.

5.4.1.4 Normalización:

Se utilizó el StandardScaler para normalizar los datos de entrenamiento y prueba, ya que modelos como SARIMAX pueden ser sensibles a la escala de los datos.

```
# Normalización de los datos de 'Consumo Tm'
scaler = StandardScaler()

# Ajustar el scaler solo en los datos de entrenamiento y luego transformar ambos conjuntos de datos
train['Consumo Tm'] = scaler.fit_transform(train[['Consumo Tm']])
test['Consumo Tm'] = scaler.transform(test[['Consumo Tm']])
```

Ilustración 34: Normalizacion de datos para el modelo Sarimax

La normalización es esencial para evitar que las diferencias en la escala de las variables afecten el rendimiento del modelo. Al aplicar esta técnica, se garantiza que el modelo reciba los datos en el formato adecuado, facilitando la optimización de los parámetros.

5.4.1.5 Búsqueda en Cuadrícula y Evaluación del Modelo:

Comienza una búsqueda exhaustiva de parámetros para el modelo SARIMAX, evaluando diversas combinaciones de los parámetros (p, d, q) y (P, D, Q, s), con el objetivo de minimizar el AIC y encontrar la mejor configuración. La búsqueda en cuadrícula es una técnica eficaz para optimizar modelos y encontrar estos parámetros. Evaluar el AIC permite encontrar el modelo que mejor se ajusta a los datos sin sobre ajustarse. La configuración de parámetros ajustada tuvo en cuenta las fluctuaciones estacionales y los eventos

excepcionales, como la pandemia y la guerra en Ucrania, lo que permitió al modelo ser flexible ante cambios en los patrones de consumo.

El modelo SARIMAX con la configuración (2, 2, 3)x(1, 1, 1, 12) alcanzó un AIC de -18.62 y un RMSE de 1,278,357.94 en los datos de prueba. Un AIC negativo sugiere un buen ajuste del modelo a los datos. Aunque el RMSE es relativamente alto, esto es comprensible debido a la naturaleza de los datos. Este valor es un buen punto de partida, pero aún queda espacio para mejorar la precisión, especialmente al incorporar más variables exógenas en el futuro, lo que podrían explicar mejor las fluctuaciones del consumo.

```
# Mostrar el mejor modelo encontrado
if best_model:
    print(f"Mejor modelo encontrado: SARIMAX{best_order}x{best_seasonal_order} - AIC: {best_aic}")

    # Predicciones con el mejor modelo y calcular RMSE
    predictions = best_model.predict(start=test.index[0], end=test.index[-1])

    # Convertir las predicciones de pandas.Series a numpy.array y desescalar
    predictions = predictions.to_numpy().reshape(-1, 1)
    predictions = scaler.inverse_transform(predictions)

    actual_values = test['Consumo Tm'].values.reshape(-1, 1)
    actual_values = scaler.inverse_transform(actual_values)

    # Calcular el RMSE
    rmse = np.sqrt(mean_squared_error(actual_values, predictions))
    print(f"RMSE en los datos de prueba: {rmse}")
else:
    print("No se encontró un modelo adecuado")
```

```
Mejor modelo encontrado: SARIMAX(2, 2, 3)x(1, 1, 1, 12) - AIC: -18.619997425636356
RMSE en los datos de prueba: 1278357.9404594726
```

Ilustración 35: Busqueda de los mejores Parametros.

5.4.1.6 Validación Cruzada:

Se implementó validación cruzada utilizando TimeSeriesSplit, dividiendo los datos en 5 subconjuntos. El RMSE promedio fue de 1,080,858.12.

```
# Definir el número de splits (subconjuntos)
tscv = TimeSeriesSplit(n_splits=5)
```

Ilustración 36: Time Series Split

La validación cruzada sirve para comprobar la estabilidad del modelo. El RMSE promedio obtenido fue algo más bajo que el RMSE en los datos de prueba, lo que sugiere que el modelo es robusto y no está sobre ajustado. Las mejoras realizadas en la búsqueda de parámetros y la exclusión de valores atípicos ayudaron a reforzar esta robustez.

```
RMSE promedio en validación cruzada: 1080858.122061879
```

Ilustración 37: Resultados de Validación Cruzada

5.4.2 Pruebas del Modelo

Se comparó el consumo real con las predicciones obtenidas a través del modelo SARIMAX. La visualización mostró cómo el modelo sigue las tendencias estacionales de consumo, con una línea que refleja los picos y caídas típicas de cada año.

5.4.2.1 Visualización de la Predicción y Consumo Real:

En el gráfico, se observa claramente cómo la predicción del modelo sigue la variabilidad del consumo real. Sin embargo, algunas desviaciones pueden notarse en los puntos donde hay datos extremos, como los picos de consumo de energía en meses importantes como lo es enero, abril, mayo, julio y agosto, todos estos del 2024, que fueron meses records en los últimos 5 años.

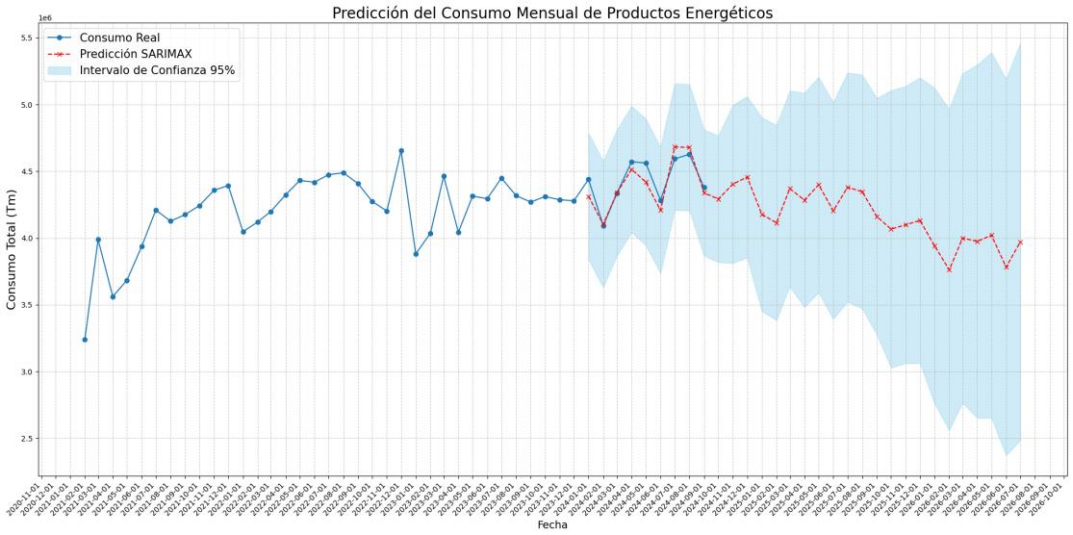


Ilustración 38: Visualizacion histórica y furura con Sarimax

El intervalo de confianza del 95% proporcionó una visualización útil para comprender la incertidumbre de las predicciones. Tanto las líneas de consumo real como las predicciones se mantienen dentro del área sombreada del intervalo, esta área crece conforme avanza el tiempo, lo que indica un aumento en la incertidumbre de las predicciones a largo plazo.

5.4.2.2 Comparación entre Consumo Real y Predicción para 2024-2025

En esta segunda visualización, se compararon las predicciones de consumo para 2024 y los primeros meses de 2025 con los valores reales de consumo. El gráfico mostró un seguimiento cercano entre las predicciones y los valores reales, aunque se observó una ligera desviación en ciertos puntos antes mencionados.

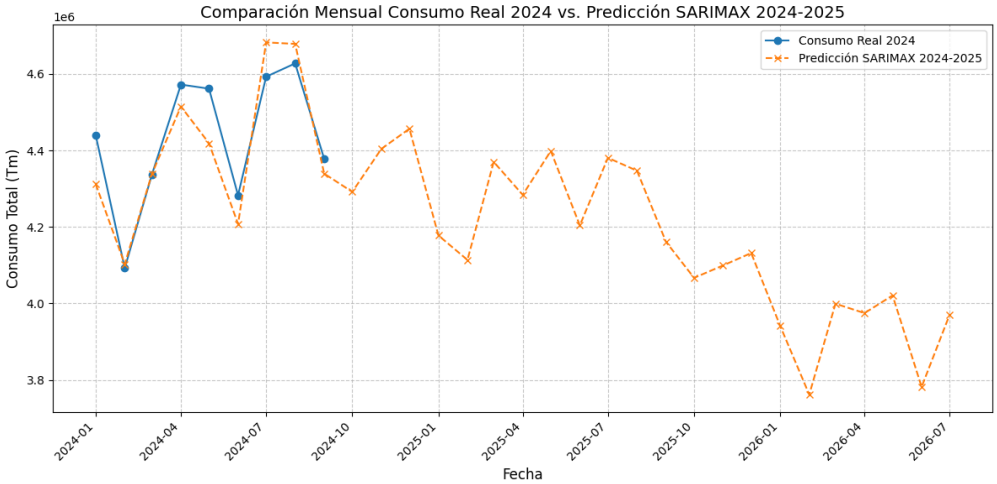


Ilustración 39: Comparacion de predicción Sarimax con Datos reales

5.4.2.3 Predicciones a combustibles especificos

Si bien esta sección estaba preestablecida como uno de los objetivos principales de este estudio, en el que se buscaba la optimización del uso de combustibles a futuro, buscando posibles patrones predictivos que ayudaran a cumplir este fin, los resultados de esta prueba demostraron que:

El modelo SARIMAX no logró una predicción óptima del consumo de Gasóleo C (combustible que tomamos como referencia por su fácil análisis de consumo estacional) debido a la ausencia de variables exógenas que pudieran explicar mejor las fluctuaciones del consumo. Aunque el modelo capturó patrones estacionales generales en los gráficos anteriores, su precisión se vio limitada al no considerar factores importantes como las temperaturas invernales, los cambios en los precios de los combustibles o las políticas energéticas, que tienen un impacto directo en el uso de este combustible. Estas limitaciones hicieron que, si bien el modelo estructuralmente era adecuado, no pudiera adaptarse completamente a las particularidades del comportamiento del Gasóleo C, especialmente en picos y caídas del consumo.

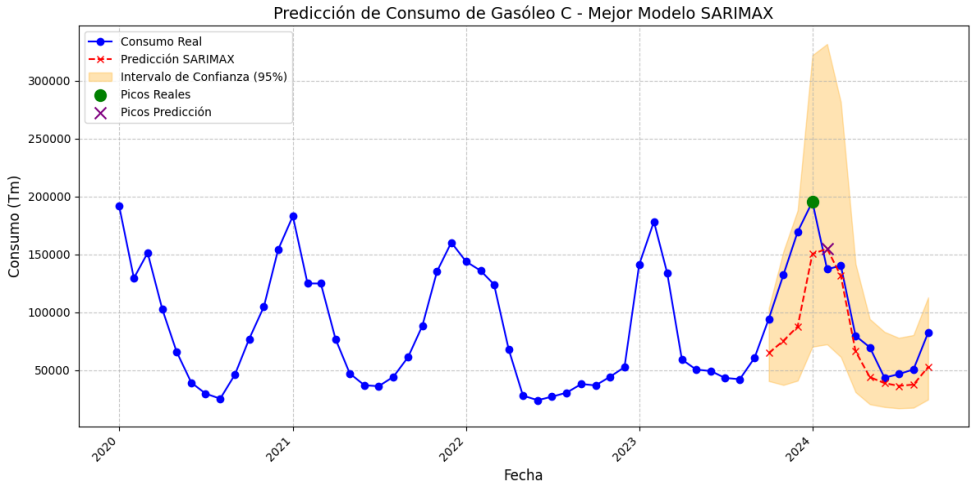


Ilustración 40: Pruebas de Sarimax con Gasoleo C

5.4.3 Resultados:

1. El modelo SARIMAX proporciona predicciones razonables para los datos mensuales, con un seguimiento cercano entre las predicciones y los consumos reales. Sin embargo, el modelo muestra algunas desviaciones en los meses de mayor variabilidad, lo que podría deberse a la falta de variables exógenas como el precio o datos de producción y compra, que ayuden a capturar factores específicos no reflejados en los datos de consumo.

2. La inclusión del intervalo de confianza en las visualizaciones aporta una capa adicional de análisis, permitiendo evaluar la precisión de las predicciones del modelo. En general, los intervalos de confianza se ajustan bien a la variabilidad observada en los datos, aunque se observa un aumento en la incertidumbre conforme se proyecta más lejos en el tiempo.

3. Las predicciones son robustas, especialmente en las primeras partes del periodo de análisis (2024). Sin embargo, la ausencia de valores reales para los meses posteriores

en el año 2025 muestra la necesidad de una mayor disponibilidad de datos para validar completamente las predicciones a largo plazo.

4. Tomando el Gasóleo C como ejemplo, el modelo SARIMAX se vio limitado por la falta de variables exógenas, como temperaturas, precios de combustibles o políticas energéticas, que habrían mejorado la precisión de las predicciones y facilitado la planificación de acciones preventivas para optimizar su consumo.

6 Conclusiones y trabajo futuro

6.1 Conclusiones

6.1.1 Segmentación del consumo de productos petrolíferos con K-means

Este estudio ha demostrado que la segmentación del consumo de productos petrolíferos mediante el algoritmo K-means es una herramienta efectiva para identificar patrones diferenciados entre las comunidades autónomas en España. Los resultados obtenidos indican que algunas regiones presentan similitudes en su consumo energético, lo que sugiere que podrían beneficiarse de políticas conjuntas para facilitar la transición hacia fuentes renovables.

Un caso interesante es el del Cluster 0, donde se han detectado oportunidades aprovechables para reducir el uso de Gasóleo A a través de biocombustibles, mejorar la eficiencia energética en calefacción y transporte, desarrollar infraestructura renovable en zonas rurales y diseñar estrategias de transición alineadas con los patrones de consumo regional. En este contexto, el modelo SARIMAX resultaria una herramienta valiosa para evaluar distintos escenarios de reducción del consumo de productos petrolíferos y estimar el impacto de diversas estrategias en la demanda energética.

No obstante, el análisis también ha revelado ciertos desafíos, como los observados en el Cluster 2, donde la correlación del consumo energético entre comunidades autónomas y las ciudades autónomas de Ceuta y Melilla no ha sido consistente. Estas diferencias ponen de manifiesto que, si bien la segmentación en clusters facilita la identificación de estrategias comunes, es importante profundizar en estudios adicionales para abordar caracteristicas particulares que podrían afectar la eficacia de las políticas energéticas en algunas regiones.

6.1.2 Limitaciones del modelo SARIMAX y la relevancia de las variables exógenas

A pesar de que SARIMAX ha sido útil para identificar patrones estacionales en el consumo de combustibles fósiles, una de sus principales limitaciones es la falta de variables exógenas. Esta carencia ha afectado especialmente la predicción del consumo de Gasóleo C, cuyo comportamiento depende en gran medida de la temperatura estacional. Dado que este combustible es un claro ejemplo de consumo con estacionalidad marcada, su análisis podría servir como referencia para estudiar el comportamiento de otros productos energéticos. Sin embargo, la ausencia de variables exógenas también ha limitado la capacidad del modelo para evaluar el consumo de otros combustibles incluidos en este estudio.

Factores como las fluctuaciones en los precios de los combustibles, las políticas energéticas o las iniciativas de promoción de energías renovables tienen un impacto directo en la demanda energética. Sin esta información, el modelo solo puede basarse en patrones históricos, lo que limita su capacidad para captar fluctuaciones causadas por factores externos. Esto se evidencia, por ejemplo, en los picos y caídas del consumo, donde elementos como cambios regulatorios, incentivos gubernamentales o variaciones climáticas pueden desempeñar un papel determinante.

Esta limitación no solo afecta la precisión del modelo, sino también su utilidad en la planificación estratégica. Para diseñar políticas energéticas efectivas y anticiparse a cambios en el consumo, es fundamental que el modelo incorpore información contextual que refleje la realidad socioeconómica, climática y política del país. La inclusión de variables exógenas permitiría no solo mejorar la precisión de las predicciones, sino también evaluar escenarios futuros con mayor realismo y prever el impacto de nuevas políticas o cambios en el mercado.

Desde esta perspectiva, la combinación de K-means y SARIMAX podría haberse optimizado aún más mediante la inclusión de variables exógenas relevantes para cada cluster. De este modo, el modelo habría podido estimar con mayor precisión el impacto de la transición energética en cada segmento y facilitar el diseño de estrategias adaptadas a la dinámica de consumo de cada comunidad autónoma.

6.2 Trabajo futuro

A partir de estos hallazgos, surgen diversas líneas de investigación que podrían desarrollarse en futuros estudios:

- **Análisis más detallado de clusters específicos**
  Dado que en el Cluster 2 se han identificado discrepancias significativas, sería interesante profundizar en sus particularidades para diseñar estrategias mejor adaptadas a su contexto energético.
- **Incorporación de variables exógenas en SARIMAX**
  Integrar factores como la temperatura, los precios de los combustibles y las políticas energéticas en el modelo permitiría mejorar la precisión de las predicciones y comprender mejor los determinantes del consumo energético.
- **Exploración de modelos complementarios**
  La combinación de SARIMAX con enfoques más avanzados, como redes neuronales recurrentes (RNNs) o modelos híbridos, podría mejorar la capacidad predictiva y ofrecer simulaciones de escenarios más dinámicos.
- **Evaluación del impacto de políticas energéticas**
  Aplicar SARIMAX para modelar distintos escenarios futuros permitiría estimar cómo medidas como incentivos a la electrificación del transporte o restricciones al uso de combustibles fósiles podrían afectar el consumo en cada cluster.
- **Extensión del estudio a nivel europeo**
  Analizar la transición energética en otros países permitiría comparar estrategias y detectar oportunidades de colaboración en el proceso de descarbonización.
- **Integración de nuevas fuentes de datos**
  Incorporar información en tiempo real, como datos de sensores IoT, imágenes satelitales o informes de mercado, podría enriquecer los modelos predictivos y mejorar la capacidad de respuesta ante cambios en la demanda y oferta energética.

6.2.1 Ejemplos reales:

6.2.1.1 Profundización en el análisis de clusters específicos

Realizar un análisis más detallado de los clusters con comportamientos de consumo atípicos podría ser una oportunidad para diseñar estrategias adaptadas a sus características particulares. Un ejemplo interesante es el Cluster 2, que agrupa comunidades como Ceuta y Melilla, donde el suministro energético depende en gran medida de combustibles importados y generadores diésel. En estas ciudades, podría resultar útil estudiar alternativas como la interconexión eléctrica con la península o la implementación de microrredes renovables. De hecho, proyectos como la central hidroeléctrica Gorona del Viento en la isla de El Hierro han mostrado que soluciones como estas pueden ayudar a reducir la dependencia de los combustibles fósiles, especialmente en territorios aislados, lo que podría ser una referencia para las ciudades mencionadas (Gorona del Viento El Hierro, S.A., 2025).

6.2.1.2 Incorporación de variables exógenas en SARIMAX

El modelo SARIMAX podría mejorar notablemente su precisión si se incorporan variables exógenas que influyan en el consumo energético. Investigaciones previas, como las del IDAE, han demostrado que factores como la temperatura media tienen un impacto directo en la demanda de Gasóleo C para calefacción (Instituto para la Diversificación y Ahorro de la Energía [IDAE], 2025). Si se incluyen también datos de la AEMET (AEMET, 2025) sobre anomalías térmicas, las predicciones del modelo podrían ajustarse con mayor precisión. Algo similar se ha observado en Francia, donde la EDF utilizó registros de temperatura histórica para mejorar la precisión de las predicciones de la demanda invernal de electricidad (EDF, 2025).

6.2.1.3 Exploración de modelos complementarios

La combinación de SARIMAX con modelos de redes neuronales es una línea de investigación prometedora para mejorar la capacidad predictiva en el ámbito energético. Según un estudio de (Ma, 2023), al integrar SARIMAX con redes neuronales recurrentes (RNNs), se puede capturar tanto los patrones estacionales como los efectos no lineales, lo que mejora la precisión de las predicciones de demanda energética. Esta combinación podría aplicarse al análisis del consumo de petróleo, permitiendo una solución más robusta y flexible para predecir cómo evolucionará la demanda de combustibles en el futuro.

6.2.1.4 Análisis del impacto de políticas energéticas

SARIMAX también puede ser útil para evaluar cómo las políticas energéticas afectan al consumo de combustibles. Por ejemplo, el Plan Nacional Integrado de Energía y Clima (PNIEC 2021-2030), que establece restricciones progresivas al diésel en España, podría modelarse con SARIMAX para entender cómo la expansión de zonas de bajas emisiones (ZBE) en ciudades como Madrid y Barcelona afectará la demanda de Gasóleo A en los próximos años. Un caso similar se dio en Londres, donde la implementación de la Ultra Low Emission Zone (ULEZ) resultó en una reducción del 22% en emisiones en los primeros 6 meses del 2024 (The Guardian, 2025) .

6.2.1.5 Extensión geográfica del studio

Ampliar el análisis a nivel europeo puede dar pistas útiles para la transición energética en España. En Noruega, casi el 90% de los coches nuevos vendidos en 2024 fueron eléctricos, gracias a incentivos fiscales y penalizaciones a los vehículos de gasolina y diésel (Reuters,

2025). Sin embargo, el consumo de combustibles fósiles sigue siendo un reto, especialmente en el transporte pesado, que aún depende en gran medida del diésel.
En Alemania, el biogás ha ayudado a reducir la dependencia de los combustibles fósiles en la industria. Aunque su producción se ha mantenido estable en los últimos años, sigue siendo indispensable en la estrategia energética del país (Inversiones, 2023).

6.2.1.6 Integración de nuevas fuentes de datos

El uso de datos en tiempo real sería una forma de enriquecer el análisis del consumo energético. En países como Singapur y Japón, se han implementado sensores IoT en estaciones de servicio para monitorear el consumo de combustibles de manera instantánea (International Energy Agency (IEA), 2019). Además, integrar datos satelitales de la ESA sobre contaminación y tráfico permitiría estudiar cómo la reducción de la movilidad, como las restricciones en zonas turísticas, afecta la demanda de productos petrolíferos. Un ejemplo de esto se vio en China, donde las imágenes satelitales de NO_2 fueron utilizadas para evaluar la disminución del consumo de gasolina durante el confinamiento por la COVID-19 (Observatory, 2020).

7 Referencias

Ade, M. (2023). *The Role of Exogenous Variables in Time Series Forecasting of Economic Indicators.* Obtenido de Journal of Business Research.: https://www.researchgate.net/profile/Martins-Ade/publication/384665221_The_Role_of_Exogenous_Variables_in_Time_Series_Forecasting_of_Economic_Indicators/links/67041aadb753fa724d6479fc/The-Role-of-Exogenous-Variables-in-Time-Series-Forecasting-of-Economic-

AEMET. (2025). *Datos climatológicos*. Obtenido de https://www.aemet.es/es/serviciosclimaticos/datosclimatologicos

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 716-723.

Alconchel, M. (2019). *openwebinars*. Obtenido de ¿Qué es XML y para qué se usa?: https://openwebinars.net/blog/que-es-xml-y-para-que-se-usa/

Alvarez, M. A. (28 de 7 de 2020). *desarrolloweb*. Obtenido de Que es MVC?: https://desarrolloweb.com/articulos/que-es-mvc.html

Apache Software Foundation. (2023). *PySpark Documentation*. Obtenido de https://spark.apache.org/docs/latest/api/python/reference/pyspark.pandas/index.html

Arcos Vargas, Á. (2023). Soberanía energética: De los combustibles a las materias primas. Economía Industrial. *Economía Industrial, (427)*, 113-123.

Asociación Española del Bioetanol (BIO-E). (2020). *Biorrefinerías y transición energética. BIO-E*. Obtenido de https://bio-e.es/biorrefinerias-y-transicion-energetica

Banco Mundial. (2020). *El impacto de la COVID-19 sobre los mercados de productos básicos.* Obtenido de https://www.bancomundial.org/es/news/press-release/2020/10/22/impact-of-covid-19-on-commodity-markets-heaviest-on-energy-prices-lower-oil-demand-likely-to-persist-beyond-2021

Bergmeir, C. &. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 192-213.

Box, G. E. (2015). *Time series analysis: Forecasting and control (5th ed.). Wiley.*

Caixabank Research. (2019). *La economía española y el petróleo: una relación estrecha*. Obtenido de Caixabank: https://www.caixabankresearch.com/es/economia-y-mercados/materias-primas/economia-espanola-y-petroleo-relacion-estrecha

Corporación de Derecho Público. (2024). *Boletín Estadístico de Hidrocarburos 323: Informe de octubre 2024.* Obtenido de https://www.cores.es/sites/default/files/archivos/publicaciones/boletin-est-hidrocarburos-323-octubre-2024.pdf

Corporación de Reservas Estratégicas de Productos Petrolíferos (CORES). (2022). *Balance de producción y consumo de productos petrolíferos en España.*

EDF. (2025). *Estimación de la demanda invernal de electricidad utilizando registros históricos de temperatura*. Obtenido de https://www.edf.fr/

ENERDATA. (s.f.). *Enerdata*. Obtenido de https://datos.enerdata.net/productos-petroliferos/estadisticas-consumo-mundial-petroleo-consumo-domestico.html

Gobierno de España. (2024). *Estadística del petróleo 2024 - Consumos mensuales provincial (TM)*. Obtenido de https://datos.gob.es/es/catalogo/ea0042931-estadistica-petroleo-2024-consumos-mensuales-provincial-tm

Gorona del Viento El Hierro, S.A. (2025). *Central Hidroeólica de El Hierro*. Obtenido de https://www.goronadelviento.es/

Grupo Forma-T. (2024). *España en la transición energética: ¿Qué oportunidades y desafíos enfrenta?* Obtenido de https://www.grupoforma-t.com/2024/11/07/espana-en-la-transicion-energetica-que-oportunidades-y-desafios-enfrenta/

Han, J. &. (2006). *Data Mining: Concepts and Techniques*. Obtenido de Morgan Kaufmann: https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_425?form=MG0AV3

Hernández Lalinde, J. D. (2018). *Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones.* Obtenido de Archivos Venezolanos de Farmacología y Terapéutica: https://www.redalyc.org/journal/559/55963207025/55963207025.pdf

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 90-95. Obtenido de Computing in Science & Engineering, 9(3), 90-95.

Hyndman, R. J. (2018). *Forecasting: principles and practice (2nd ed.)*. OTexts.

Instituto para la Diversificación y Ahorro de la Energía (IDAE). (2025). *Ayudas a la climatización en sector residencial. RD 477/2021*. Obtenido de https://www.idae.es/ayudas-y-financiacion/para-energias-renovables-en-autoconsumo-almacenamiento-y-termicas-sector/rd-4772021-ayudas-la-climatizacion-en-sector-residencial

International Energy Agency (IEA). (2019). *Better energy efficiency policy with digital tools*. Obtenido de IEA: https://www.iea.org/articles/better-energy-efficiency-policy-with-digital-tools

Inversiones, I. E. (2023). *Mercado de biocombustibles y combustibles sintéticos en Alemania 2023*. Obtenido de ICEX: https://www.icex.es/content/dam/es/icex/oficinas/017/documentos/2023/12/estudios-de-mercado/RE_Mercado%20de%20biocombustibles%20y%20combustibles%20sint%C3%A9ticos%20en%20Alemania%202023_REV.pdf

Joaquín Amat Rodrigo, J. E. (2023). *Modelos ARIMA y SARIMAX en Python.* Obtenido de https://cienciadedatos.net/documentos/py51-modelos-arima-sarimax-python

KPMG. (2024). *¿Cómo será la movilidad del futuro? La movilidad del futuro será multimodal, sostenible y entrelazada.* Obtenido de https://www.tendencias.kpmg.es/2024/11/movilidad-futuro-multimodal-sostenible-entrelazada/

Luján, J. D. (2018). *EDteam*. Obtenido de ¿Cómo funcionan los hilos en programación?: https://ed.team/blog/como-funcionan-los-hilos-en-programacion

Ma, X. (2023). *Deep Learning Combinatorial Models for Intelligent Supply Chain Demand Forecasting.* Obtenido de mdpi: https://doi.org/10.3390/biomimetics8030312

Ministerio de Agricultura, Pesca y Alimentación. (2024). *Informe Semanal: Precios Gasóleo Agrario y Pesquero.*

Ministerio de Industria, Turismo y Comercio. (29 de junio de 2007). *Resolución de 29 de mayo de 2007, de la Dirección General de Política Energética y Minas, por la que se aprueban los nuevos formularios oficiales para la remisión de información a la Dirección General de Política Energética y Minas, a la Comisión Nacional.* Obtenido de Boletín Oficial del Estado, núm. 155: https://www.boe.es/buscar/doc.php?id=BOE-A-2007-12750

Ministerio para la Transición Ecológica y el Reto Demográfico. (2023). *Plan Nacional Integrado de Energía y Clima (PNIEC 2023-2030).* Obtenido de https://www.miteco.gob.es/es/energia/estrategia-normativa/pniec-23-30.html?form=MG0AV3

Moreno Villa, E. R. (2021). *Análisis comparativo entre PySpark y Pandas para el procesamiento de datos masivos de covid19*. Obtenido de Universidad de Buenos Aires: https://repositorio.ub.edu.ar/handle/123456789/9535

Mullins, C. S. (7 de 2021). *searchdatacenter*. Obtenido de Db2: https://searchdatacenter.techtarget.com/es/definicion/DB2

OBS Business School. (2025). *El sector energético en España, hacia una descarbonización sostenible.* Obtenido de https://www.obsbusiness.school/actualidad/informes-de-investigacion/informe-obs-el-sector-energetico-en-espana-hacia-una-descarbonizacion-sostenible

Observatory, N. E. (2020). *Airborne nitrogen dioxide plummets over China*. Obtenido de NASA Earth Observatory: https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china

Pedregosa, F. V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Plotly Technologies Inc. (2015). *Collaborative data science. Montreal, QC: Plotly Technologies Inc*. Obtenido de https://plot.ly

Reuters. (2025). *Norway: Nearly all new cars sold in 2024 were fully electric*. Obtenido de Reuters: https://www.reuters.com/business/autos-transportation/norway-nearly-all-new-cars-sold-2024-were-fully-electric-2025-01-02

Segovia, M. (13 de 12 de 2024). *El Independiente*. Obtenido de Descarbonizar el cielo, la gran asignatura pendiente: el uso del avión dispara el consumo de queroseno: https://www.elindependiente.com/economia/2024/12/13/descarbonizar-el-cielo-la-gran-asignatura-pendiente-el-uso-del-avion-dispara-el-consumo-de-queroseno/?form=MG0AV3

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. . *The American Statistician*, 72(1), 37-45.

The Guardian. (2025). *Ulez expansion led to significant drop in air pollutants in London, report finds*. Obtenido de https://www.theguardian.com/environment/article/2024/jul/25/ulez-expansion-led-to-significant-drop-in-air-pollutants-in-london-report-finds?utm_source=chatgpt.com

Tiwari, A. (2020). *Time series analysis: Forecasting with SARIMAX model and stationarity concept.* Obtenido de Journal of Emerging Technologies and Innovative Research, 7(12), 2349-5162: https://www.jetir.org/papers/JETIREJ06034.pdf

Tukey, J. W. (1977). *Exploratory Data Analysis. Addison-Wesley.*

Waskom, M. L. (2021). *seaborn: statistical data visualization*. Obtenido de Journal of Open Source Software, 6(60), 3021: https://doi.org/10.21105/joss.03021

Yadav, S. (. (2022). *Comparative Study of ARIMA, Prophet and LSTM for Time Series Prediction.* Obtenido de Journal of Artificial Intelligence, Machine Learning and Data Science, 1(1), 1813-1816: https://urfjournals.org/open-access/a-comparative-study-of-arima-prophet-and-lstm-for-time-series-prediction.pdf

Yang, Y.-H. L.-L.-H. (2023). *Visibility graph analysis of crude oil futures markets: Insights from the COVID-19 pandemic and Russia-Ukraine conflict*. Obtenido de arXiv: https://arxiv.org/abs/2310.18903

8 Anexos

Ilustración 41: Tendencia Historia Registrada por ENERDATA desde 1990 hasta 2023

	Fecha	CCAA	Provincia	Tipo Producto	Consumo Tm	Año	Mes
20961	2022-08-01	ANDALUCIA	CADIZ	OTROS FUELOLEOS	332588	2022	8

Ilustración 42: Busqueda de valores atípicos, en este caso el valos mas alto en 5 años.

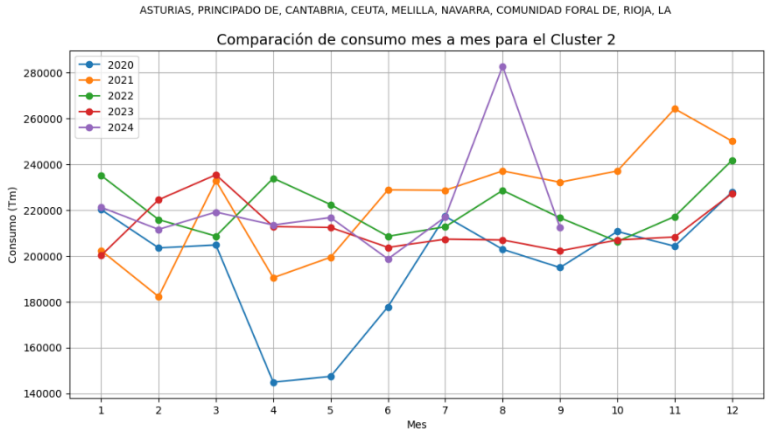


Ilustración 43: Consumo mes a mes cada año en el Cluster 2

Tipo Producto	2020	2021	2022	2023	2024
BIODIESEL	38,811.00	26,042.00	14,794.00	6,017.00	34,632.00
BIOETANOL	418.00	60.00	20.00	84.00	80.00
FUELOLEO BIA	1,436,110.00	1,334,842.00	1,429,219.00	1,087,001.00	1,654,315.00
GASOLEO A	19,434,204.00	21,808,667.00	22,189,887.00	21,463,364.00	16,369,006.00
GASOLEO B	4,458,429.00	4,589,985.00	4,611,154.00	3,635,000.00	2,707,056.00
GASOLEO C	1,116,943.00	1,119,081.00	752,674.00	1,154,081.00	845,017.00
GASOLINA AUTO. S/PB 95 I.O.	3,909,045.00	4,869,554.00	5,442,568.00	5,739,726.00	4,653,064.00
GASOLINA AUTO. S/PB 98 I.O.	329,675.00	372,993.00	308,473.00	319,130.00	257,188.00
GLP	1,292,497.00	1,412,150.00	1,428,064.00	1,348,903.00	995,480.00
OTRAS GASOLINAS	3,845.00	4,688.00	4,451.00	4,715.00	3,752.00
OTROS FUELOLEOS	4,067,229.00	4,745,950.00	5,888,361.00	5,587,761.00	4,202,727.00
OTROS GASOLEOS	3,324,656.00	3,684,018.00	4,103,717.00	3,960,327.00	2,582,628.00
OTROS QUEROSENOS	186.00	210.00	139.00	152.00	73.00
QUEROSENO AVIACION	2,417,764.00	3,357,140.00	5,870,860.00	6,639,977.00	5,578,950.00

Disminucion en el Consumo
Aumento en el consumo

Tabla 7: Comparativa en incremento o disminución de consumo de carburantes cada año