# UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

## FACULTAD DE INGENIERÍAS

### MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

**TEMA**:

---

**ESTUDIO DE UBICACIÓN DE VEHÍCULOS ELÉCTRICOS EN EL ÁREA DE CONCESIÓN DE LA EEQ A TRAVÉS DE MINERÍA DE DATOS DE CONSUMO RESIDENCIAL**

---

Trabajo de Titulación previo a la obtención del título de Magíster en Big Data y Ciencia de Datos


**Autor(a)**

Ing. Luis Andres Mera Maldonado

**Tutor(a)**

Ing. Víctor Emilio Calle García, MBA



AMBATO – ECUADOR

2025

## AUTORIZACIÓN POR PARTE DEL AUTOR PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE TITULACIÓN

Yo, Luis Andres Mera Maldonado, declaro ser autor del Trabajo de Titulación con el nombre "ESTUDIO DE UBICACIÓN DE VEHÍCULOS ELÉCTRICOS EN EL ÁREA DE CONCESIÓN DE LA EEQ A TRAVÉS DE MINERÍA DE DATOS DE CONSUMO RESIDENCIAL", como requisito para optar al grado de Magíster en Big Data y Ciencia de Datos y autorizo al Sistema de Bibliotecas de la Universidad Indoamérica, para que con fines netamente académicos divulgue esta obra a través del Repositorio Digital Institucional (RDI-UTI).

Los usuarios del RDI-UTI podrán consultar el contenido de este trabajo en las redes de información del país y del exterior, con las cuales la Universidad tenga convenios. La Universidad Indoamérica no se hace responsable por el plagio o copia del contenido parcial o total de este trabajo.

Del mismo modo, acepto que los Derechos de Autor, Morales y Patrimoniales, sobre esta obra, serán compartidos entre mi persona y la Universidad Indoamérica, y que no tramitaré la publicación de esta obra en ningún otro medio, sin autorización expresa de la misma. En caso de que exista el potencial de generación de beneficios económicos o patentes, producto de este trabajo, acepto que se deberán firmar convenios específicos adicionales, donde se acuerden los términos de adjudicación de dichos beneficios.

Para constancia de esta autorización, en la ciudad de Ambato a los 29 días del mes de julio de 2025, firmo conforme:

Autor: Luis Andres Mera Maldonado

Firma:

Número de Cédula: 1723471403

Dirección: Pichincha, Quito, Cochapamba, La Pulida.

Correo Electrónico: luandy1220@gmail.com

Teléfono: (+593) 0963875731

## APROBACIÓN DEL DIRECTOR

En mi calidad de Director del Trabajo de Titulación "ESTUDIO DE UBICACIÓN DE VEHÍCULOS ELÉCTRICOS EN EL ÁREA DE CONCESIÓN DE LA EEQ A TRAVÉS DE MINERÍA DE DATOS DE CONSUMO RESIDENCIAL" presentado por Luis Andres Mera Maldonado, para optar por el Título de Magíster en Big Data y Ciencia de Datos.

CERTIFICO

Que dicho Trabajo de Titulación ha sido revisado en todas sus partes y considero que reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte los Examinadores que se designe.

Ambato, 29 de julio del 2025

Firmado electrónicamente por:
VICTOR EMILIO CALLE GARCIA
Validar únicamente con FirmaEC

.....

Ing. Víctor Emilio Calle García, MBA

DIRECTOR

## DECLARACIÓN DE AUTENTICIDAD

Quien suscribe, declaro que los contenidos y los resultados obtenidos en el presente Trabajo de Titulación, como requerimiento previo para la obtención del Título de Magíster en Big Data y Ciencia de Datos, son absolutamente originales, auténticos y personales y de exclusiva responsabilidad legal y académica del autor

Ambato, 29 de julio del 2025

……

Luis Andres Mera Maldonado

1723471403

## APROBACIÓN DE EXAMINADORES

El Trabajo de Titulación ha sido revisado, aprobado y autorizada su impresión y empastado, sobre el Tema: ESTUDIO DE UBICACIÓN DE VEHÍCULOS ELÉCTRICOS EN EL ÁREA DE CONCESIÓN DE LA EEQ A TRAVÉS DE MINERÍA DE DATOS DE CONSUMO RESIDENCIAL, previo a la obtención del Título de Magíster en Big Data y Ciencia de Datos, reúne los requisitos de fondo y forma para que el estudiante pueda presentarse a la sustentación del Trabajo de Titulación.

Ambato, 1 de agosto del 2026

Firmado electrónicamente por:
ALBA DE LOS CIELOS MIRANDA VILLACIS
Validar únicamente con FirmaEC

……..

Mg. ALBA MIRANDA VILLACÍS
EXAMINADOR

Firmado electrónicamente por:
JOSE LUIS VARELA ALDAS
Validar únicamente con FirmaEC

……

Ph.D. JOSÉ VARELA ALDÁS
EXAMINADOR

**DEDICATORIA**

Con profunda gratitud, dedico este logro a mis padres,
cuyo sacrificio y dedicación me permitieron llegar hasta aquí.
A mis hermanas, fuente constante de alegría y apoyo.
A Pablo y Paula, mis amados sobrinos.
Y en memoria de mi abuelo Carlos (+),
cuyo ejemplo de vida me guía en cada paso.

**AGRADECIMIENTO**

Deseo expresar mi profundo agradecimiento a Emilio
por su invaluable apoyo y guía durante el desarrollo de este proyecto.
Su amistad y sus enseñanzas constantes fueron fundamentales
para alcanzar esta meta.
Agradezco también a mis padres por su amor incondicional y
respaldo inquebrantable a lo largo de mi camino.
Extiendo mi gratitud a mis amigos y a todas aquellas personas
que me brindaron su apoyo y compañía en este recorrido.
A los profesionales de Arconel, cuyo apoyo fue esencial para
la realización de este proyecto,
les agradezco su colaboración y dedicación.
Finalmente, agradezco a la Universidad Indoamérica
por brindarme la oportunidad de formarme
y crecer como profesional.

## ÍNDICE DE CONTENIDOS

PORTADA ..... i
AUTORIZACIÓN PARA EL REPOSITORIO DIGITAL ..... ii
APROBACIÓN DEL DIRECTOR ..... iii
DECLARACIÓN DE AUTENTICIDAD ..... iv
APROBACIÓN DE EXAMINADORES ..... v
DEDICATORIA ..... vi
AGRADECIMIENTO ..... vii
RESUMEN EJECUTIVO ..... xvi
ABSTRACT ..... xvii

**CAPÍTULO I**

**INTRODUCCIÓN**

INTRODUCCIÓN ..... 1
CONTEXTO ..... 1
PROBLEMÁTICA ..... 2
ANTECEDENTES ..... 3
CONTEXTO Y PROYECCIONES DE LA ELECTROMOVILIDAD EN ECUADOR ..... 4
ENFOQUES METODOLÓGICOS PARA LA DETECCIÓN DE PATRONES DE CONSUMO ..... 5
REFLEXIÓN Y BRECHA INVESTIGATIVA ..... 6
JUSTIFICACIÓN ..... 7
IMPORTANCIA ..... 7
IMPACTO ..... 7
UTILIDAD ..... 8
FACTIBILIDAD ..... 8
OBJETIVO GENERAL ..... 8
OBJETIVOS ESPECÍFICOS ..... 8

## CAPÍTULO II

## METODOLOGÍA

ÁREA DE ESTUDIO ..... 10
ENFOQUE ..... 10
ENFOQUE CUANTITATIVO-EXPLORATORIO ..... 10
DESCRIPCIÓN DE LA METODOLOGÍA ..... 11
TIPOS DE INVESTIGACIÓN ..... 11
Investigación de Asociación de Variables ..... 11
Investigación Explicativa ..... 11
MÉTODOS DE INVESTIGACIÓN ..... 12
Método Bibliográfico-Documental ..... 12
Método de Campo ..... 12
DISEÑO DEL TRABAJO ..... 13
OPERACIONALIZACIÓN DE LAS VARIABLES ..... 13
DEFINICIÓN DE LAS VARIABLES DE ESTUDIO ..... 15
Variable independiente (Causal) ..... 15
Variable Dependiente (Efecto) ..... 16
Variables de Caracterización ..... 16
PROCEDIMIENTO PARA OBTENCIÓN Y ANÁLISIS DE DATOS ..... 16
MÉTODOS ..... 16
Preparación, limpieza y filtrado inicial de las series de tiempo de consumo eléctrico. ..... 16
Análisis exploratorio inicial (EDA) y detección de anomalías ..... 16
Tratamiento de valores atípicos y faltantes ..... 17
Detección de aumento en el consumo de energía ..... 17
Clustering con K-Means ..... 17
Visualización y análisis de clústeres ..... 17
Creación y aplicación del perfil de referencia No-VE ..... 18
Estratificación de clientes con VE ..... 18
Validación y refinamiento ..... 18
Mapeo geográfico ..... 18

TÉCNICAS ..... 19
Lectura de datos desde archivos CSV y conversión al formato DataFrame ..... 19
Transformación de datos ..... 19
Pivot ..... 19
Estadísticas descriptivas ..... 20
Visualizaciones para inspección gráfica de series de tiempo ..... 20
Imputación para valores faltantes ..... 20
Detección de aumento de consumo ..... 20
Aplicación del algoritmo K-Means para agrupar series de tiempo ..... 20
Determinación del número óptimo de clústeres (k) con el método del codo ..... 21
Iteración de K-Means para diferentes valores de k y cálculo de la inercia ..... 21
Visualización de la curva de codo para identificar el número óptimo de clústeres ..... 21
Reducción de dimensionalidad mediante PCA para proyectar las series de tiempo en 2D ..... 21
Cuantificación de los clústeres refinados ..... 21
Comparación de umbrales de similitud para la optimización de filtros ..... 22
Estratificación de Clientes VE ..... 22
Cálculo del Índice de Davies-Bouldin para evaluar la calidad del clustering ..... 22
Implementación de llamadas a la API con la biblioteca googlemaps de Python ..... 22
Generación de mapas interactivos con la biblioteca Folium ..... 22
INSTRUMENTOS ..... 23
RStudio ..... 23
Tidyverse ..... 23
Janitor ..... 23
Python ..... 23
Pandas ..... 24
NumPy ..... 24
Scikit-learn ..... 24
Matplotlib y Seaborn ..... 24
Yellowbrick ..... 25
Scipy ..... 25

Folium .....25
Google Maps .....25
Google Colab.....25
Multiprocessing.....26
POBLACIÓN Y MUESTRA.....26
POBLACIÓN.....26
MUESTRA .....26
HIPÓTESIS .....27

**CAPÍTULO III**

**DESARROLLO DE LA INVESTIGACIÓN**

METODOLOGÍA DE MINERÍA DE DATOS APLICADA: CRISP-DM.....28
FASE 1: PROCESAMIENTO Y PREPARACIÓN DE BASE DE DATOS .....30
PREPARACIÓN DE LA BASE DE DATOS .....30
ANÁLISIS EXPLORATORIO INICIAL (EDA) Y DETECCIÓN DE ANOMALÍAS .....35
TRATAMIENTO DE VALORES ATÍPICOS Y FALTANTES .....36
FILTRAR CONSTANTES Y TARIFA DIGNIDAD .....37
FASE 2: ALGORITMO DE CLASIFICACIÓN K-MEANS.....38
DETECCIÓN DE AUMENTO EN EL CONSUMO.....38
CLUSTERING INICIAL CON K-MEANS .....39
VISUALIZACIÓN Y ANÁLISIS DE CLÚSTER INICIALES .....40
CREACIÓN Y APLICACIÓN DEL PERFIL DE REFERENCIA NO-VE .....41
ANÁLISIS Y ESTRATIFICACIÓN DE CLIENTES CON VEHÍCULOS ELÉCTRICOS .....44
FASE 3: VALIDACIÓN Y OBTENCIÓN DE POTENCIALES CLIENTES CON VE.46
FILTRO HACIENDO USO DE LOS ESTRATOS DE VE'S .....46
CLUSTERIZACIÓN FINAL DE CLIENTES CON POSIBLES VE'S.....47
FASE 4: MAPEO GEOGRÁFICO.....48
GEOCODIFICACIÓN DE CLIENTES CON POTENCIALES VE'S .....49
VISUALIZACIÓN INTERACTIVA .....49

**CAPÍTULO IV**

**RESULTADOS Y DISCUSIÓN**

INTERPRETACIÓN DE RESULTADOS....52
FASE 1: PROCESAMIENTO Y PREPARACIÓN DE LA BASE DE DATOS....52
FASE 2: ALGORITMO DE CLASIFICACIÓN K-MEANS....53
Detección de aumento en el consumo....53
Clustering inicial con K-Means. ....53
Visualización y análisis de clústeres iniciales....54
Creación y aplicación del perfil de referencia No-VE....58
Análisis y estratificación de clientes con vehículos eléctricos....59
FASE 3: VALIDACIÓN Y OBTENCIÓN DE POTENCIALES CLIENTES....61
Clusterización final de clientes con posibles VE's....63
FASE 4: MAPEO GEOGRÁFICO....65
CONTRASTE CON OTRAS INVESTIGACIONES....67
VERIFICACIÓN DE LA HIPÓTESIS....68
DISCUSIÓN....71
IMPLICACIONES....71
LIMITACIONES E IMPLICACIONES DEL TRABAJO REALIZADO....71
RESUMEN COMPARATIVO CON TRABAJOS SIMILARES....72
FUTUROS TRABAJOS....73

**CAPÍTULO V**

**CONCLUSIONES Y RECOMENDACIONES**

CONCLUSIONES....74
RECOMENDACIONES....74
BIBLIOGRAFÍA....76

## ÍNDICE DE TABLAS

Tabla 1 ..... 14
Tabla 2 ..... 29
Tabla 3 ..... 31
Tabla 4 ..... 61

# ÍNDICE DE GRÁFICOS

Gráfico 1 ..... 40
Gráfico 2 ..... 41
Gráfico 3 ..... 42
Gráfico 4 ..... 43
Gráfico 5 ..... 44
Gráfico 6 ..... 45
Gráfico 7 ..... 54
Gráfico 8 ..... 56
Gráfico 9 ..... 57
Gráfico 10 ..... 59
Gráfico 11 ..... 60
Gráfico 12 ..... 62
Gráfico 13 ..... 63
Gráfico 14 ..... 64
Gráfico 15 ..... 65

## ÍNDICE DE IMÁGENES

Imagen 1 ..... 31
Imagen 2 ..... 35
Imagen 3 ..... 36
Imagen 4 ..... 37
Imagen 5 ..... 39
Imagen 6 ..... 47
Imagen 7 ..... 51
Imagen 8 ..... 58
Imagen 9 ..... 66

**UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA**

**FACULTAD DE INGENIERÍAS**

**MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS**

**TEMA:** ESTUDIO DE UBICACIÓN DE VEHÍCULOS ELÉCTRICOS EN EL ÁREA DE CONCESIÓN DE LA EEQ A TRAVÉS DE MINERÍA DE DATOS DE CONSUMO RESIDENCIAL

**AUTOR(A):** Ing. Luis Andres Mera Maldonado

**TUTOR (A):** Ing. Víctor Emilio Calle García, MBA.

## RESUMEN EJECUTIVO

Este estudio aborda la problemática para la identificación de usuarios de vehículos eléctricos (VE's) que realizan cargas en sus residencias sin la utilización de medidores de energía dedicados, en la ciudad de Quito, según lo establece la normativa vigente, lo que dificulta la medición y el monitoreo de la demanda energética asociada. El presente trabajo, propone una metodología basada en técnicas de minería de datos, el cual permite analizar patrones de consumo eléctrico residencial e identificar anomalías compatibles con la carga de VE. Se aplican algoritmos de agrupamiento (K-Means) para dividir a los usuarios en grupos según similitudes en el consumo mensual. Adicionalmente, en un mapa de ubicación de los usuarios identificados se presentan los resultados de forma visual la distribución geográfica y con ello facilitar la planificación de la infraestructura de carga. El análisis se basa en datos de consumo eléctrico mensual de hogares en Quito. La metodología incluye la preparación de la base de datos, la aplicación del algoritmo K-Means, la validación de los resultados y la generación de un mapa geográfico. Se utilizan herramientas como RStudio y Python con bibliotecas como Pandas, NumPy, Scikit-learn y Folium, así como la API de Google Maps. La aplicación de la metodología permitió identificar 109 clientes residenciales como potenciales usuarios de vehículos eléctricos sin medidor dedicado, cuyos patrones de consumo mostraron un incremento sostenido y una alta similitud con los perfiles de carga conocidos. Geográficamente, estos usuarios se concentran principalmente en la zona norte de Quito y en los valles de Cumbayá y Tumbaco. Se concluye que, a pesar de las limitaciones por la granularidad mensual de los datos, la minería de datos es una herramienta viable y eficaz para que las empresas eléctricas puedan estimar la penetración de la electromovilidad en el sector residencial, facilitando así una mejor planificación de la infraestructura y gestión de la demanda energética.

**DESCRIPTORES**: Consumo de energía, Estratificación, K-Means, vehículos eléctricos.

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTY OF ENGINEERING

MASTER'S IN BIG DATA AND DATA SCIENCE

**AUTHOR:** MERA MALDONADO LUIS ANDRES

**TUTOR:** null CALLE GARCIA VICTOR EMILIO

## ABSTRACT

ELECTRIC VEHICLE SITING STUDY IN THE EEQ CONCESSION AREA USING RESIDENTIAL CONSUMPTION DATA MINING

This research addresses the issue of identifying electric vehicle (EV) users who charge their cars at home without using dedicated energy meters in Quito, as required by current regulations, which makes it difficult to measure and monitor the associated energy demand. This research proposes a methodology based on data mining techniques, which allows for the analysis of residential electricity consumption patterns and the identification of anomalies compatible with EV charging. Clustering algorithms (K-Means) are applied to divide users into groups according to similarities in monthly consumption. Additionally, the results are presented visually on a map showing the location of the identified users, thereby facilitating the planning of charging infrastructure. The analysis is based on monthly electricity consumption data from households in Quito. The methodology includes preparing the database, applying the K-Means algorithm, validating the results, and generating a geographic map. Tools such as RStudio and Python are used in conjunction with libraries like Pandas, NumPy, Scikit-learn, and Folium, as well as the Google Maps API. The application of the methodology identified 109 residential customers as potential users of electric vehicles without a dedicated meter, whose consumption patterns showed a sustained increase and a high similarity to known load profiles. Geographically, these users are mainly concentrated in the northern part of Quito and the valleys of Cumbayá and Tumbaco. In conclusion, despite the limitations imposed by the monthly granularity of the data, data mining is a viable and effective tool for electricity companies to estimate the penetration of electromobility in the

**KEYWORDS:**
Electric vehicles, energy consumption, K-Means, stratification

residential sector, thereby facilitating better infrastructure planning and energy demand management.

**KEYWORDS:**
Electric vehicles, energy consumption, K-Means, stratification

# CAPÍTULO I

# INTRODUCCIÓN

## INTRODUCCIÓN

### CONTEXTO

La preocupante crisis climática es uno de los retos más grandes del mundo, el sector del transporte, por su parte, tiene una participación del 24% en las emisiones globales de CO2 relacionadas con energía (Pathak et al., 2022). Al respecto, los vehículos eléctricos han surgido como una solución que ayudara a mitigar este problema: como afirma la Agencia Internacional de Energía (IEA) (Global EV Outlook 2023 – Analysis, 2023), estos pueden eliminar potencialmente hasta 150 millones de toneladas métricas de emisiones de CO2 para el año 2030, siempre que alcancen el 30% de las ventas globales. Sin embargo, un aumento tan rápido plantea desafíos técnicos, para la red eléctrica. La adopción generalizada de este tipo de cargas aumentará la demanda de electricidad en los edificios de las ciudades en hasta un 50%, lo que creará cargas pesadas que pueden poner en peligro la sostenibilidad del sistema de distribución. (Clustering Analysis for Active and Reactive Energy Consumption Data Based on AMI Measurements, s. f.).

En América Latina se avanza en la electrificación del transporte; sin embargo, no se ha llegado a las metas esperadas. Chile se destaca como uno de los pioneros en la región, actualmente cuenta con 1,500 autobuses eléctricos en Santiago y persigue alcanzar cero emisiones en el transporte público para el año 2040(Kerrigan, s. f.) . Asimismo, Colombia ha conseguido que su flota vehicular eléctrica alcance el 6%; este logro se obtuvo al eliminar el Impuesto al Valor Agregado (IVA) y los gravámenes aduaneros (Escorcia Herrera et al., 2023) . No obstante, la región presenta varios obstáculos importantes y que son estructurales: el 70% de la producción de electricidad proviene de fuentes renovables, predominantemente hidroeléctricas, las redes de distribución carecen de la inteligencia operacional necesaria para manejar la carga distribuida(Clustering Analysis for Active and Reactive Energy Consumption Data Based on AMI Measurements, s. f.).

En el caso específico de Ecuador, se cuenta con la estrategia nacional de movilidad eléctrica 2021-2030, cuyo objetivo principal es disminuir las emisiones de CO2 en 7.5 millones de toneladas mediante incentivos tributarios y la construcción de 200 estaciones

de carga públicas para 2025 (MTOP y BID preparan Estrategia Nacional de Movilidad Segura en Ecuador 2021 – 2030 – Ministerio de Transporte y Obras Públicas, s. f.). La crisis climática y energética en el país han impedido un avance significativo.

En Quito, según datos de la AMT (Agencia Metropolitana de Tránsito) se ha dado un crecimiento en la adopción de vehículos eléctricos del 40% respecto del total registrados en el país se registra la mayor concentración en la capital.

A nivel de infraestructura eléctrica, los resultados de los estudios realizados por la Empresa Eléctrica Quito muestran que el 15% de los transformadores ubicados en áreas residenciales con un estrato económico alto, como Cumbayá y el Valle de los Chillos, operan al 90% de su capacidad (Rendición de Cuentas - Empresa Eléctrica Quito - Empresa Eléctrica Quito, s. f.).

Se ha identificado que el 80% de los VE's en Quito que se cargan en hogares que carecen de medidores dedicados o exclusivos. Esta falta de trazabilidad dificulta la identificación precisa de la demanda energética asociada a la carga de VE's, así como la cuantificación de su impacto real en la red (Lascano et al., 2023). Esta invisibilidad operativa limita la implementación de estrategias de gestión de demanda, como la implementación de tarifas dinámicas o la gestión activa, herramientas fundamentales para el desarrollo de redes inteligentes y la optimización del suministro eléctrico (Yang et al., 2020).

Además el 85% de los usuarios de VE's en Quito cargan en el periodo de 18:00 - 22:00, mismo que coincide con la demanda máxima residencial(Rendición de Cuentas - Empresa Eléctrica Quito - Empresa Electrica Quito, s. f.).

La falta de información detallada sobre la carga de VE's impide a los operadores de la red eléctrica anticipar y gestionar los picos de demanda, lo que puede ocasionar sobrecargas, interrupciones en el suministro y un aumento en los costos operativos. Por lo tanto, es fundamental implementar soluciones que permitan la medición y el monitoreo individualizado del consumo de energía de los VE's, con el fin de garantizar la estabilidad y eficiencia de la red eléctrica en el contexto de la creciente electromovilidad.

## PROBLEMÁTICA

La adopción de vehículos eléctricos (VE's) en Quito presenta un desafío complejo: a pesar de sus beneficios ambientales en la reducción de emisiones locales, su carga si no es gestionada, a largo plazo puede amenazar la confiabilidad del sistema eléctrico.

Para abordar esta problemática, se identifican tres dimensiones críticas:

1. Falta de claridad en el consumo: La falta de medidores exclusivos impide identificar el consumo de los VE's, ya que se mezclan o se confunden entre otros electrodomésticos de alto consumo (p. ej., aires acondicionados), imposibilitando su diferenciación mediante métodos tradicionales (Yang et al., 2020).
2. Saturación de la red: Modelos de simulación indican que un incremento del 10% en la penetración de VE's en Quito sobrecargaría el 30% de los transformadores de distribución aumentando el riesgo de apagones en épocas de alta demanda (Lascano et al., 2023).
3. Inversiones mal direccionadas: La EEQ ha destinado USD 5 millones anuales a ampliación de redes desde 2020, pero el 40% de estas inversiones se localizan en zonas de baja adopción de VE, evidenciando una carencia de planificación con visión en la electromovilidad. (Rendición de Cuentas - Empresa Eléctrica Quito - Empresa Eléctrica Quito, s. f.).

Para abordar esta problemática, la investigación se centra en la evidencia empírica y la brecha existente. Estudios en contextos análogos demuestran que la minería de datos de consumo residencial puede identificar VE's con una precisión de entre el 85-92% mediante técnicas como clustering y detección de anomalías (Koohfar et al., 2023). No obstante, estos métodos no han sido adaptados a las particularidades propias de la ciudad de Quito. Factores como la altitud (2,850 msnm) y la estacionalidad climática modifican patrones de consumo. Además, el 95% de los hogares quiteños tiene tarifas planas, es decir es la misma sin importar el horario, a diferencia de ciudades europeas o asiáticas con tarifas horarias que facilitan la detección de VE's (Koohfar et al., 2023).

## ANTECEDENTES

La identificación de usuarios de vehículos eléctricos (VE) a partir de datos de consumo residencial es un desafío central para la gestión moderna de las redes eléctricas. El estado del arte demuestra que las técnicas de minería de datos y aprendizaje automático son herramientas fundamentales para segmentar clientes y detectar patrones anómalos que indiquen la presencia de una nueva carga significativa, como la de un VE. El presente análisis se enmarca en un contexto ecuatoriano con proyecciones y datos oficiales que subrayan la urgencia de adoptar estas metodologías para una transición energética ordenada.

## CONTEXTO Y PROYECCIONES DE LA ELECTROMOVILIDAD EN ECUADOR

El panorama en Ecuador ha sido definido por una serie de estudios estratégicos y datos oficiales que confirman la tendencia creciente de la electromovilidad y su potencial impacto. Estos documentos, tanto a nivel nacional como local, establecen el marco contextual y normativo de la investigación, pero también revelan la necesidad de herramientas técnicas para gestionar el impacto a nivel de distribución.

A nivel nacional, el "Estudio de Prospectiva de Electromovilidad en Ecuador" (IIGE, 2021-2023) sentó las bases al analizar tres escenarios de adopción de VE (tendencial, conservador y de máximo esfuerzo), proyectando que, en el mejor de los casos, el 85% de los autobuses públicos podrían ser eléctricos para 2050. Este estudio destacó la importancia de una hoja de ruta clara para la transición. En el plano normativo, el análisis de la Ley Orgánica de Eficiencia Energética (Paredes et al., 2020) estableció el marco legal, señalando que la sustitución de autobuses diésel por eléctricos podría reducir las emisiones de $CO_2$ en un 30% y destacando que el sector transporte consume el 52.39% de la energía del país.

Datos más recientes del Balance Energético Nacional (Ministerio de Energía y Minas, 2023) (Las dos herramientas estadísticas energéticas más importantes del país se presentaron en Quito – Ministerio de Energía y Minas, s. f.) y las Estadísticas del Sector Eléctrico (ARCERNNR, 2021-2022) reforzaron esta visión (Las dos herramientas estadísticas energéticas más importantes del país se presentaron en Quito – Ministerio de Energía y Minas, s. f.). El primero mostró un crecimiento del 18% en el consumo energético del transporte en 2022, mientras que el segundo confirmó una cobertura eléctrica nacional del 97.3%, indicando una capacidad base del sistema para absorber nuevas cargas. A nivel local, el *"Análisis de Escenarios de Penetración de VE's en Quito"* (Mora, s. f.)proporcionó un modelo predictivo específico para la capital, estimando que la demanda por VE podría alcanzar los 970 GWh para 2050 (aproximadamente el 20% del consumo actual de la ciudad). Finalmente, proyectos prácticos como *"SolutionsPlus en Quito"* (2020-2023) demostraron la viabilidad de la electromovilidad ligera, logrando reducir 491.74 kg de $CO_2$ en 1,071 km recorridos con solo 10 bicicletas eléctricas de carga.

## ENFOQUES METODOLÓGICOS PARA LA DETECCIÓN DE PATRONES DE CONSUMO

Para abordar el desafío técnico de identificar y gestionar la carga de VE, la literatura científica ha validado consistentemente el uso de algoritmos de clustering y otras técnicas de aprendizaje automático para segmentar a los consumidores y analizar sus patrones.

La técnica más extendida es el clustering, que agrupa a los usuarios según la similitud de sus perfiles de consumo. Investigaciones como las de Castro (Castro et al., 2023) y Ramírez (Ramírez, 2022) demostraron la eficacia de algoritmos como K-Means y DBSCAN. El primero validó la partición de redes de distribución estándar (IEEE 34 y 123 nodos) en grupos coherentes, mientras que el segundo aplicó clustering en el contexto chileno para segmentar clientes residenciales y optimizar programas de respuesta a la demanda. De manera más específica, trabajos como los de Miyazaki (Miyazaki et al., 2020) y Serrano Guerrero (Serrano Guerrero, 2020) utilizaron el clustering para identificar patrones directamente asociados a la carga de VE. Miyazaki analizo datos del Estado de Carga (SOC) para definir perfiles como *"carga rápida frecuente"* o *"carga nocturna"*, información crucial para la planificación de infraestructura. Por su parte, Serrano Guerrero utilizó un modelado de mezcla Gaussiana (GMM) para segmentar más de 5,000 hogares en Eslovenia, identificando un perfil de *"consumo nocturno alto"* asociado a los VE. Estos enfoques en patrones específicos, en lugar de una búsqueda genérica de anomalías, han resultado clave para una detección más dirigida.

Paralelamente, se han explorado metodologías más avanzadas que complementan el clustering. Jahangir propusieron un enfoque híbrido que combinaba redes neuronales profundas (LSTM) con K-Means para capturar las complejas dinámicas temporales del comportamiento de los usuarios de VE en mercados energéticos (Jahangir et al., 2020). Otros estudios se centraron en la detección de anomalías ((Fieni, s. f.); (Aparicio Wallis et al., 2019)). Aunque sus objetivos eran distintos (gestión energética en universidades o en la industria), su relevancia es crucial, ya que la incorporación de un VE al consumo de un hogar puede ser tratada, en efecto, como una anomalía positiva y sostenida en un patrón residencial típico. Estos trabajos validan el uso de redes neuronales y autoencoders para automatizar la detección de patrones inusuales, reduciendo la necesidad de verificaciones manuales.

Finalmente, la aplicación de estas técnicas se ha extendido a la planificación de infraestructura y la optimización de redes. El estudio *"Discovering Electric Vehicle Charging Locations"* aplicó clustering a datos de movilidad urbana (trayectorias de taxis) para identificar zonas óptimas de instalación de cargadores y reducir la "ansiedad de autonomía" de los usuarios. En una línea similar, Zhao propuso un algoritmo de clustering adaptativo para redes vehiculares (VANETs) que lograba reducir el consumo energético en la comunicación en un 30%, demostrando que los enfoques adaptativos pueden optimizar el uso de la energía en sistemas dinámicos (Zhao et al., 2022).

**REFLEXIÓN Y BRECHA INVESTIGATIVA**

A partir de esta revisión, se concluye que las técnicas de minería de datos, y en particular el clustering, son herramientas robustas y validadas para el análisis de perfiles de consumo eléctrico. La comunidad científica ha demostrado su eficacia para segmentar clientes, identificar patrones de carga de VE, detectar anomalías y planificar la infraestructura necesaria.

Sin embargo, se identifica una brecha investigativa clara y significativa: la gran mayoría de los estudios mencionados se han desarrollado en contextos con datos de alta granularidad (horaria o sub-horaria) y, a menudo, con estructuras de tarifas dinámicas que incentivan patrones de carga fácilmente detectables (por ejemplo, cargar durante la noche para aprovechar precios más bajos). La realidad del área de concesión de la EEQ en Quito presentaba un desafío distinto y más complejo, caracterizado por dos limitaciones fundamentales: la disponibilidad de datos de consumo únicamente a nivel mensual, lo que diluye las señales de carga a corto plazo; y la existencia de tarifas planas, que no generan patrones de comportamiento predecibles.

Este trabajo de titulación se propuso, por tanto, abordar directamente esta brecha, adaptando y validando una metodología basada en minería de datos que, a pesar de estas limitaciones, fuera capaz de identificar patrones de consumo consistentes con la carga de VE. El objetivo fue aportar una solución práctica, robusta y contextualizada a un problema real, proporcionando una herramienta que pudiera ser implementada por la empresa eléctrica local para mejorar su planificación y gestión en el avance hacia la electromovilidad.

En adelante se muestra una base investigativa relevante relacionada con la electromovilidad, la gestión de redes eléctricas y la minería de datos aplicada al consumo energético en Ecuador, con énfasis en el caso específico de Quito.

## JUSTIFICACIÓN

### IMPORTANCIA

La transición hacia la electromovilidad es un pilar estratégico para cumplir con los Objetivos de Desarrollo Sostenible (ODS 7 y 11) y los compromisos del Acuerdo de París, especialmente en ciudades intermedias como Quito, donde el transporte contribuye al 35% de las emisiones locales de $CO_2$ (Proyecto de Electromovilidad impulsado por la Empresa Eléctrica Quito contará con apoyo del PNUD | Programa De Las Naciones Unidas Para El Desarrollo, s. f.). La presente investigación se centra en abordar una brecha crítica identificada por la Agencia Internacional de Energía: el 60% de las redes eléctricas urbanas en países en desarrollo carecen de sistemas para monitorear la carga residencial de VE's. Esta falta de monitoreo limita su capacidad para evitar colapsos.

En Ecuador, la penetración de VE crece a un ritmo del 22% anual (Descargas – AEADE, 2025).

La ausencia de metodologías validadas localmente para identificar estos consumidores (VE's sin medidor dedicado) representa un riesgo operativo el cual radica en la sobrecarga de la red eléctrica, esto podría llevar a colapsos en la red, afectando la calidad del servicio para otros usuarios y generando costos adicionales para la Empresa Eléctrica Quito (EEQ) por la necesidad de realizar inversiones no planificadas en infraestructura.

### IMPACTO

**Impacto ambiental:** Al optimizar la gestión de la carga de VE, se reduciría uso de combustibles fósiles como resultado del cambio de gasolina a electricidad en el transporte. Esto evitaría hasta 1,200 toneladas de $CO_2$ anuales en Quito (simulaciones basadas en datos de la EEQ).

**Impacto económico:** La detección precisa de usuarios de VE permitiría a la EEQ evitar USD 2.5 millones en pérdidas anuales por sobreinversión en infraestructura mal planificada (Rendición de Cuentas - Empresa Eléctrica Quito - Empresa Eléctrica Quito, s. f.).

**Impacto social:** Mejoraría la calidad del servicio eléctrico para 250,000 hogares en zonas de alta penetración de VE's (p. ej., Valle de los Chillos).

**UTILIDAD**

**Utilidad para la Empresa Eléctrica Quito (EEQ):** Se proporcionará un sistema de detección automatizada de VE's basado en minería de datos.

**Utilidad para tomadores de decisiones:** Se ofrecerá evidencia para actualizar la Estrategia Nacional de Movilidad Eléctrica. Esta actualización incorporará criterios técnicos de gestión de redes.

**Utilidad para la academia:** Se validarán métodos de clustering en contextos con datos limitados.

**FACTIBILIDAD**

**Factibilidad técnica:** Se cuenta con acceso a 5 años de datos históricos de consumo residencial (2018-2023) de la EEQ, con granularidad mensual. Además, herramientas como Python y plataformas cloud (Google Colab) permiten procesar grandes volúmenes de datos.

**Factibilidad metodológica:** Estudios previos validan la efectividad del clustering (K-Means + PCA) para identificar patrones de VE's, incluso sin medidores exclusivos.

**Factibilidad legal:** Los datos son anónimos y cumplen con la Ley Orgánica de Protección de Datos Personales del Ecuador.

## OBJETIVO GENERAL

Desarrollar una metodología basada en minería de datos para identificar patrones de consumo eléctrico mensual que evidencien la carga residencial de vehículos eléctricos (VE's) sin medidores exclusivos en la ciudad de Quito, con el fin de proporcionar a la Empresa Eléctrica Quito (EEQ) una herramienta para la planificación de la infraestructura y la gestión optimizada de la demanda energética.

## OBJETIVOS ESPECÍFICOS

- Fundamentar el marco metodológico de la investigación mediante el análisis del estado del arte sobre técnicas de minería de datos y patrones de consumo, para adaptarlos a la problemática de la detección de VE con datos de baja granularidad.

- Aplicar un modelo de clustering (K-Means) y análisis de series temporales sobre la base de datos de consumo de la EEQ para segmentar a los usuarios e identificar un subconjunto de clientes con patrones compatibles con la carga de un VE.
- Validar los clústeres identificados y diseñar una visualización georreferenciada de los usuarios potenciales, con el propósito de entregar una propuesta de valor que facilite la toma de decisiones estratégicas para la expansión de la red eléctrica.

# CAPÍTULO II

# METODOLOGÍA

## ÁREA DE ESTUDIO

Esta investigación se desarrolla en el ámbito del Big Data y la Ciencia de Datos, con un enfoque particular en su aplicación a la ingeniería eléctrica. Se justifica esta área de estudio por las siguientes razones:

**Naturaleza del problema:** La carga masiva de vehículos eléctricos (VE) genera enormes volúmenes de datos, lo que plantea desafíos en su procesamiento, análisis y modelado. Para abordar estos retos, se requieren técnicas avanzadas de Big Data, como el análisis de series temporales y el clustering de demanda.

**Caso de Aplicación:** El área de concesión de la EEQ abarca una extensión de 15.155 kilómetros cuadrados, proporcionando servicio en diversas provincias. En Pichincha, su cobertura incluye los cantones de Quito, Rumiñahui, Mejía, Pedro Moncayo, Pedro Vicente Maldonado y San Miguel de Los Bancos, además de sectores de Puerto Quito y Cayambe. En Napo, opera en los cantones de Quijos y El Chaco. Asimismo, su alcance se extiende a pequeñas zonas de las provincias de Cotopaxi, Imbabura y Santo Domingo de los Tsáchilas.

## ENFOQUE

## ENFOQUE CUANTITATIVO-EXPLORATORIO

Esta investigación adopto un enfoque cuantitativo-exploratorio, dado el carácter del problema y la brecha metodológica existente. La identificación de usuarios de vehículos eléctricos (VE) que cargan en sus residencias sin medidores exclusivos requiere el análisis de grandes volúmenes de datos sobre el consumo eléctrico mensual, los cuales no han sido recolectados ni estructurados con este propósito específico.

Ante la ausencia de técnicas validadas localmente para la detección de VE en estas condiciones, se propuso un enfoque basado en la minería de datos. Este enfoque permitió descubrir patrones subyacentes en los datos de consumo eléctrico que puedan estar relacionados con la carga de VE, y segmentar a los usuarios en grupos con características similares. Las principales técnicas cuantitativas que se aplicaron fueron:

- **Clustering no supervisado:** Se utilizaron algoritmos como K-Means para segmentar a los usuarios en grupos con patrones de consumo similares.
- **Análisis comparativo:** Se realizaron análisis comparativos entre los clústeres identificados con el fin de caracterizar los grupos de usuarios y determinar su relación con la posible presencia de VE.
- **Análisis de series temporales:** Se analizaron las series temporales de consumo eléctrico mensual para identificar tendencias y patrones estacionales que pudieran estar asociados a la carga de VE's.
- **Análisis de clientes con medidores exclusivos: S**e analizaron las series temporales de consumo de clientes que sí utilizaban un medidor exclusivo para la carga de sus VE, con el fin de establecer un perfil de referencia

La justificación de un análisis cuantitativo-exploratorio se justifica por su capacidad para:

- **Manejar grandes volúmenes de datos:** Permite procesar y analizar eficientemente grandes volúmenes de datos de consumo eléctrico.
- **Identificar patrones ocultos:** Facilita la detección de patrones y relaciones no evidentes a simple vista.
- **Generar resultados escalables:** Los resultados obtenidos podrán ser replicados en otras ciudades con características similares, contribuyendo a la generalización del conocimiento.

## DESCRIPCIÓN DE LA METODOLOGÍA

### TIPOS DE INVESTIGACIÓN

Esta investigación integró dos enfoques complementarios:

#### Investigación de Asociación de Variables

Este enfoque se utilizó para identificar y analizar la relación entre los patrones de consumo eléctrico y la posible presencia de vehículos eléctricos. A través del análisis de clustering, se buscó determinar qué características del consumo se vinculaban con la probabilidad de que un cliente poseyera un VE.

#### Investigación Explicativa

Este componente se concentró en comprender por qué ciertos patrones de consumo indicaban la carga de VE. Se buscó explicar cómo el aumento sostenido en el consumo

se reflejaba en los datos de series de tiempo y la lógica subyacente en la aplicación de técnicas de clustering para aislar y caracterizar este fenómeno.

## MÉTODOS DE INVESTIGACIÓN

Se fundamentó en la aplicación de dos métodos principales: Bibliográfico-Documental y de Campo. Estos métodos se complementaron para construir una base sólida de conocimiento y aseguro la validez y aplicabilidad de los resultados.

### Método Bibliográfico-Documental

Se aplicó este método para construir el marco teórico y metodológico del análisis. Se realizó una revisión de literatura científica sobre la identificación de clientes con VE, análisis de series de tiempo, técnicas de clustering y perfiles de carga. Esta revisión permitió fundamentar teóricamente la investigación, conocer el estado del arte y justificar la elección de las técnicas y algoritmos utilizados. Esta revisión permitió:

- **Fundamentar teóricamente la investigación:** Se ha establecido el marco conceptual y las bases teóricas para la identificación de patrones de consumo asociados a la carga de VE's.
- **Conocer el estado del arte:** Se han identificado las metodologías existentes, los algoritmos utilizados y los resultados obtenidos en investigaciones previas sobre la temática.
- **Justificar la metodología propuesta:** La elección de las técnicas y algoritmos utilizados (K-Means, distancias euclidianas, perfiles de referencia) se basó en la evidencia científica y las mejores prácticas identificadas en la literatura.
- **Interpretar y discutir los resultados:** Se contrastaron los hallazgos de la tesis con los resultados de investigaciones previas, identificando similitudes, diferencias y contribuciones originales.

### Método de Campo

El método de campo fue esencial en esta investigación, ya que se basó en el análisis de datos reales de consumo eléctrico de clientes. La fuente primaria de información fueron los datos de series de tiempo proporcionados por la Empresa Eléctrica Quito (EEQ). La recolección y el análisis de estos datos fueron cruciales para la validez y la aplicabilidad de los resultados.

El análisis se centró en patrones de consumo reales, reflejando la variabilidad y la complejidad de los datos del mundo real. El método de campo permitió contrastar la metodología propuesta con datos empíricos y evaluar su desempeño en un contexto práctico y relevante para las empresas distribuidoras de energía eléctrica.

## DISEÑO DEL TRABAJO

### OPERACIONALIZACIÓN DE LAS VARIABLES

El diseño de este proyecto se fundamentó en la operacionalización de variables clave que describen el consumo energético de clientes residenciales y su posible adopción de VE's. El estudio se enmarco en un enfoque cuantitativo-exploratorio, utilizando series de tiempo mensuales de consumo eléctrico como principal fuente de información.

El resultado principal de este estudio, la clasificación de un *"Cliente Potencial con VE"*, se operacionalizó como una categoría binaria (Sí/No). Esta clasificación final fue el producto de un proceso metodológico de varias etapas que se aplicó sobre la variable dependiente principal: el patrón de consumo eléctrico. Este proceso incluyó:

- **Identificación de aumento sostenido:** Se analizo la tendencia del consumo energético para determinar si existe un aumento sostenido que pueda ser indicativo de la incorporación de un VE. Esta variable se definió como categórica binaria.
- **Clustering con K-Means:** Se aplico el algoritmo K-Means para agrupar a los clientes según patrones de consumo similares. La variable Clúster resultante fue categórica nominal.
- **Filtrado y estratificación:** Se refinaron los clústeres mediante la comparación con perfiles de referencia de clientes con y sin VE. Se estratificaron los clientes con VE confirmados según su consumo total, generando la variable Estrato de Consumo, que fue categórica ordinal.
- **Evaluación de Similitud:** Se calculo la similitud entre los patrones de consumo de los clientes y los perfiles de referencia de clientes sin VE. La similitud al perfil sin VE se operacionalizo como una variable continua, utilizando la distancia euclidiana como medida de similitud.

En la Tabla 1 se detalla cada una de las variables de las diferentes etapas desarrolladas en este proyecto.

**Tabla 1**

Matriz de Variables.

| Variable | Tipo | Operacionalización | Descripción |
|---|---|---|---|
| Consumo Energético | Continua | kWh consumidos por mes | Cantidad de energía eléctrica consumida por un cliente residencial en un mes. Se obtiene de las series temporales de consumo. |
| Frecuencia de Datos | Categórica | Mensual | Resolución temporal de los datos de consumo eléctrico. |
| Valores Atípicos | Categórica | Presencia | Indica si se detectaron valores atípicos en las series de tiempo durante el análisis exploratorio. |
| Valores Faltantes | Categórica | Presencia | Indica si se detectaron valores faltantes en las series de tiempo. |
| Aumento Sostenido | Categórica | Presencia o ausencia | Indica si una serie de tiempo presenta un aumento sostenido en el consumo energético, potencialmente indicativo de la incorporación de un VE. |
| Umbral de similitud | Continua | Valor umbral definido | Valor numérico utilizado como criterio para determinar la similitud entre series temporales. |
| Número de Clústeres (k) | Entera | Valor de 'k' en K-Means | Cantidad de clústeres utilizados en K-Means. |
| Índice de Davies-Bouldin Inicial | Continua | Valor del Índice de Davies-Bouldin para el clustering inicial | Medida de la calidad de la agrupación en el clustering inicial. |

| Variable | Tipo | Operacionalización | Descripción |
|---|---|---|---|
| Centroides Clústeres Inicial | Vector de Series de Tiempo | Series de tiempo promedio de cada clúster inicial | Representan los patrones de consumo promedio para cada clúster resultante del clustering inicial. |
| Perfil de Referencia No-VE | Vector de Series de Tiempo | Patrón de consumo promedio de clientes sin VE | Perfil de consumo típico de clientes residenciales que no poseen vehículos eléctricos, utilizado como base para el filtrado. |
| Estrato de Consumo | Categórica | Nivel de consumo estratificado. | Categorización de clientes con VE confirmados en diferentes niveles de consumo energético total. |
| Consumo Promedio por Estrato | Continua | Consumo energético promedio (kWh) en cada estrato | Valor representativo del nivel de consumo energético típico de cada estrato de clientes con VE. |
| Dirección Postal Cliente Potencial | Categórica | Dirección obtenida de datos del cliente | Dirección postal asociada a cada cliente potencial identificado, utilizada para la geocodificación. |

## DEFINICIÓN DE LAS VARIABLES DE ESTUDIO

Para abordar la problemática descrita y alcanzar los objetivos de la investigación, fue imperativo definir formalmente las variables de estudio que guiaron el análisis. La relación causal postulada en este trabajo se centró en las siguientes variables principales:

### Variable independiente (Causal)

La variable independiente, o la causa que se presume genera un efecto observable, fue la presencia de carga de un Vehículo Eléctrico (VE) en una residencia sin un medidor dedicado. Esta es una variable de naturaleza categórica y binaria (presencia/ausencia), la

cual no se midió directamente, sino que se infirió a través de los métodos de análisis de datos. La identificación de esta variable fue el objetivo central del estudio.

**Variable Dependiente (Efecto)**

La variable dependiente, o el efecto medido, fue el patrón de consumo eléctrico mensual de los clientes residenciales. Esta variable se representó como una serie de tiempo continua, compuesta por los registros históricos de consumo en kilovatios-hora (kWh) proporcionados por la EEQ. Las alteraciones, tendencias y anomalías en esta variable fueron el principal objeto de análisis para refutar o comprobar la hipótesis.

**Variables de Caracterización**

Adicionalmente, se utilizaron variables secundarias para la caracterización, segmentación y georreferenciación de los datos, entre las que destacan la ubicación geográfica del cliente (provincia, cantón, parroquia) y el tipo de tarifa o contrato (residencial, con medidor exclusivo para VE).

## PROCEDIMIENTO PARA OBTENCIÓN Y ANÁLISIS DE DATOS

## MÉTODOS

El procedimiento para la obtención y análisis de los datos se estructuró en las siguientes fases, donde se aplicaron métodos, técnicas e instrumentos específicos.

**Preparación, limpieza y filtrado inicial de las series de tiempo de consumo eléctrico.**

En esta etapa se preprocesaron los datos de consumo eléctrico. El objetivo fue optimizar la calidad y coherencia de los datos para la aplicación posterior de técnicas de minería

**Análisis exploratorio inicial (EDA) y detección de anomalías**

El análisis exploratorio de datos (EDA) se empleó para obtener una comprensión profunda de los datos. Se aplicaron estadísticas descriptivas como medias, medianas y desviaciones estándar para resumir las características clave de los datos, definir rangos de valores esperados y detectar posibles errores o inconsistencias. Las visualizaciones jugaron un papel fundamental, ya que a través de la inspección gráfica de las series de tiempo se lograron identificar visualmente valores atípicos, patrones inusuales, tendencias y estacionalidad.

**Tratamiento de valores atípicos y faltantes**

Para el manejo de valores faltantes, se aplicó la imputación por la media, un método que consistió en sustituir dichos valores por el promedio de la columna. Esta técnica fue útil al presumir que los datos ausentes eran aleatorios. Por otro lado, la interpolación lineal se aplicó como una técnica más avanzada, especialmente en las series de tiempo, para estimar valores faltantes considerando la tendencia lineal entre los puntos adyacentes, asumiendo una continuidad temporal en la información.

**Detección de aumento en el consumo de energía**

Se utilizó la regresión lineal simple como técnica estadística para modelar la relación entre la variable dependiente (consumo eléctrico) y la variable independiente (tiempo). Su propósito fue analizar la tendencia del consumo a lo largo del tiempo mediante una línea recta, cuya pendiente representó la tasa de cambio. Está pendiente actuó como un indicador clave y, al establecer un umbral positivo, fue posible identificar de manera objetiva aquellas series de tiempo que reflejaron un incremento sostenido, lo que podría señalar la adopción de un vehículo eléctrico.

**Clustering con K-Means**

El algoritmo K-Means fue una herramienta poderosa que se utilizó para descubrir patrones en los datos no etiquetados. Su aplicación inicial permitió realizar una aproximación efectiva para identificar grupos de consumidores con comportamientos similares. Este algoritmo de clustering no supervisado se usó para dividir el conjunto de datos en $k$ grupos (clústeres). Su funcionamiento se basó en asignar iterativamente cada punto de datos al clúster cuyo centroide (media) era el más cercano. Para determinar el número óptimo de clústeres (k), se empleó el método del codo, una heurística con la que se evaluó la curva de inercia, sugiriendo el número más adecuado de clústeres.

**Visualización y análisis de clústeres**

La visualización de clústeres mediante el Análisis de Componentes Principales (PCA) optimizó la interpretación de los resultados del clustering. El PCA fue una técnica de reducción de dimensionalidad que se empleó para transformar los datos de alta dimensión a un espacio de menor dimensión (2D), manteniendo la mayor parte de la varianza original. Al proyectar las series temporales en un espacio bidimensional, se facilitó la

interpretación visual de la estructura de los clústeres, lo que permitió observar cómo se agruparon las series.

**Creación y aplicación del perfil de referencia No-VE**

La creación de un perfil de referencia No-VE constituyó una estrategia eficaz que se implementó para optimizar la detección de usuarios de vehículos eléctricos. Este proceso comenzó con el análisis de los clústeres obtenidos para identificar aquellos que reflejaban patrones de consumo de usuarios sin VE. Una vez identificado un clúster representativo, se calculó el patrón de consumo promedio de sus series de tiempo. Este perfil promedio No-VE se utilizó como referencia para filtrar otras series, eliminando aquellas que coincidían excesivamente con él.

**Estratificación de clientes con VE**

La estratificación de clientes con VE confirmados permitió segmentar a dichos usuarios según sus niveles de consumo, lo que resultó clave para comprender la variabilidad del consumo. Esta técnica dividió a la población de usuarios de VE en grupos (estratos) basados en su consumo total de energía. Para ello, se calculó el consumo total de cada cliente y se definieron los estratos utilizando percentiles. A continuación, se asignó cada cliente a un estrato, lo que proporcionó valiosas perspectivas sobre las características de consumo de cada segmento.

**Validación y refinamiento**

La validación de los resultados del clustering fue un paso crucial para garantizar la calidad y robustez de estos. El Índice de Davies-Bouldin se presentó como una métrica clave para este fin, ya que ofreció una medida cuantitativa de la calidad del agrupamiento. Este índice evaluó la similitud promedio entre cada clúster y el más cercano, proporcionando una forma objetiva de medir la cohesión y separación de los grupos. Un valor más bajo del índice indicó un clustering de mayor calidad, por lo que esta métrica fue útil para seleccionar la configuración que minimizó el índice.

**Mapeo geográfico**

El mapeo geográfico fue una herramienta para visualizar la distribución espacial de los clientes potenciales de VE. Al utilizar Folium y la API de Google Maps, se pudo crear un mapa interactivo que permitió explorar la distribución espacial de los clientes identificados y reconocer patrones geográficos relevantes. La geocodificación se realizó

convirtiendo las direcciones postales en coordenadas geográficas (latitud y longitud), facilitando así la interacción y el análisis visual.

## TÉCNICAS

Las técnicas fueron las herramientas y procedimientos específicos que se utilizaron para implementar los métodos de investigación. Representaron las acciones concretas y detalladas que se llevaron a cabo en cada fase del estudio con el fin de obtener información, analizar los datos y ejecutar las tareas particulares del proyecto.

### Lectura de datos desde archivos CSV y conversión al formato DataFrame

El formato CSV, ampliamente utilizado para almacenar datos tabulares, fue el formato en que se recibieron los datos iniciales. Para su procesamiento, se leyeron estos archivos y se convirtieron a un DataFrame, una estructura bidimensional proporcionada por la biblioteca Pandas en Python. Esta conversión fue un paso fundamental que facilitó la manipulación y el análisis posterior, ya que los DataFrames ofrecieron una gran flexibilidad para realizar operaciones como la selección, el filtrado, la agrupación y la transformación de los datos, convirtiéndose en una herramienta esencial en el flujo de trabajo.

### Transformación de datos

La transformación de datos fue un proceso crucial que se aplicó para adaptar los datos a un formato adecuado para el análisis. Los datos recopilados originalmente contenían columnas almacenadas como cadenas de texto que, para poder realizar operaciones numéricas o estadísticas, debieron ser convertidas a un formato reconocido por el software, como números enteros o decimales. Esta conversión aseguró la precisión y la eficiencia de los cálculos posteriores y facilitó la aplicación de los modelos de análisis.

### Pivot

El pivoteo de datos fue una operación clave que se ejecutó para la manipulación y transformación de los datos. Esta técnica se aplicó para reorganizar la tabla de datos, transformando las filas (que representaban observaciones mensuales por cliente) en columnas (donde cada columna representaba un mes específico). Su objetivo principal fue reestructurar la información de un formato largo a un formato ancho para facilitar el análisis de las series de tiempo de cada cliente como una entidad única.

**Estadísticas descriptivas**

Se calcularon y analizaron diversas estadísticas descriptivas para obtener una visión general de las características principales de los datos. El cálculo de medidas como la media, mediana, desviación estándar, valores mínimos, máximos y percentiles permitió comprender la distribución de los datos, identificar los rangos de valores típicos y detectar posibles valores atípicos. Estas estadísticas fueron cruciales para identificar errores en la calidad de los datos, como valores fuera de rango o distribuciones inusuales que podrían haber afectado la validez de los análisis.

**Visualizaciones para inspección gráfica de series de tiempo**

La visualización de las series de tiempo mediante gráficos y representaciones visuales fue una herramienta poderosa que se utilizó para identificar patrones y tendencias a lo largo del tiempo. La inspección gráfica permitió detectar visualmente tendencias, estacionalidad, ciclos y valores atípicos que no eran evidentes solo con las estadísticas descriptivas. Además, facilitó la identificación de patrones inusuales o inesperados que requirieron una investigación más profunda.

**Imputación para valores faltantes**

La imputación de valores faltantes fue un paso esencial para manejar los datos incompletos. Se aplicó la imputación por la media, técnica que reemplazó los valores faltantes con el valor promedio de la columna correspondiente. Este método se consideró útil bajo el supuesto de que los valores faltantes eran aleatorios. Sin embargo, se tuvo en cuenta que esta técnica podía subestimar la variabilidad de los datos.

**Detección de aumento de consumo**

Se aplicó esta técnica para identificar patrones en las series de tiempo de consumo eléctrico que mostraban un incremento significativo y persistente. El proceso consistió en detectar aumentos en el consumo, aislar los incrementos que eran significativos y validar su sostenibilidad, comparando el consumo mínimo posterior con el nivel del incremento. Se consideró un aumento solo si el consumo no caía por debajo de este nivel.

**Aplicación del algoritmo K-Means para agrupar series de tiempo**

Se aplicó el algoritmo K-Means como técnica clave en el análisis de las series de tiempo, especialmente por su utilidad para identificar patrones similares en grandes volúmenes de datos. En esta fase, el algoritmo agrupó las series de tiempo con características de

consumo similares en *k* clústeres. Este proceso asignó cada serie al clúster cuyo centroide (el promedio del clúster) era más cercano, minimizando la varianza intra-clúster.

**Determinación del número óptimo de clústeres (k) con el método del codo**

Determinar el número adecuado de clústeres fue un paso crucial al aplicar K-Means. Se utilizó el método del codo, una técnica heurística que consistió en ejecutar el algoritmo K-Means para diferentes valores de *k* y graficar la inercia (suma de distancias cuadradas dentro de los clústeres) respecto a *k*. La curva resultante, conocida como la curva de codo, mostró un punto de inflexión que se consideró como el valor óptimo para *k*.

**Iteración de K-Means para diferentes valores de k y cálculo de la inercia**

La implementación del método del codo requirió ejecutar repetidamente el algoritmo K-Means con diferentes valores de *k*. Para cada valor, se calculó la inercia del clustering resultante, sumando las distancias cuadradas entre cada serie de tiempo y el centroide de su clúster. Estos valores de inercia, junto con sus correspondientes valores de *k*, se utilizaron para crear la curva de codo.

**Visualización de la curva de codo para identificar el número óptimo de clústeres**

Una vez calculada la inercia para varios valores de *k*, se graficó la curva de codo, donde la inercia se colocó en el eje vertical y *k* en el eje horizontal. La inspección visual de esta curva permitió identificar el *"codo"*, el punto donde la curva comenzaba a aplanarse. Este punto indicó el equilibrio entre la reducción de la inercia y la complejidad del modelo, lo que se consideró el número óptimo de clústeres.

**Reducción de dimensionalidad mediante PCA para proyectar las series de tiempo en 2D**

Visualizar los clústeres obtenidos con K-Means resultó difícil dado que las series de tiempo se encontraban en un espacio de alta dimensión. Para facilitar la visualización, se utilizó el Análisis de Componentes Principales (PCA), que proyectó las series de tiempo en un espacio bidimensional, conservando la mayor parte de la varianza de los datos originales. Esta proyección facilitó la interpretación y comparación de los clústeres.

**Cuantificación de los clústeres refinados**

Una vez refinado el clustering, se analizaron y describieron cuantitativamente las características de cada clúster. Esto incluyó el cálculo de estadísticas descriptivas como el tamaño del clúster (número de series de tiempo), el consumo promedio, la variabilidad

del consumo y los centroides (patrones de consumo promedio). La cuantificación ayudó a comprender las diferencias entre los clústeres y a caracterizar los tipos de patrones de consumo.

**Comparación de umbrales de similitud para la optimización de filtros**

Se aplicó esta técnica iterativa para optimizar el rendimiento de los filtros basados en la similitud entre series de tiempo y patrones de referencia. Consistió en variar sistemáticamente el umbral de similitud y evaluar su impacto en el filtrado y el clustering. Al probar diferentes umbrales, se buscó encontrar el valor óptimo que minimizara los falsos positivos y los falsos negativos.

**Estratificación de Clientes VE**

Se definieron categorías o estratos de consumo dividiendo el rango de valores de consumo total en segmentos basados en rangos predefinidos (bajo, medio, alto consumo). Estos estratos permitieron segmentar a los clientes de VE en grupos con características de consumo similares, facilitando un análisis más granular.

**Cálculo del Índice de Davies-Bouldin para evaluar la calidad del clustering**

El Índice de Davies-Bouldin fue la métrica que se utilizó para evaluar la calidad del clustering. Se calculó para las diferentes configuraciones de clústeres, donde un valor bajo indicó un buen clustering, con menor varianza intra-clúster y mayor separación entre clústeres.

**Implementación de llamadas a la API con la biblioteca googlemaps de Python**

Para interactuar con la API de Geocoding de Google Maps desde Python, se usó la biblioteca googlemaps. Esta biblioteca simplificó la autenticación, el envío de solicitudes y la recepción de respuestas, lo que permitió automatizar la geocodificación de las direcciones de los clientes identificados.

**Generación de mapas interactivos con la biblioteca Folium**

Para visualizar geográficamente a los clientes potenciales, se empleó la biblioteca Folium. Esta herramienta permitió crear mapas interactivos, sobreponiendo marcadores y otros elementos geográficos sobre mapas base, lo que enriqueció la presentación de los resultados.

## INSTRUMENTOS

Los instrumentos utilizados en esta investigación fueron las herramientas de software y las bibliotecas de programación que facilitaron la aplicación de las técnicas de análisis de datos. Estos recursos fueron esenciales para llevar a cabo la investigación de manera eficiente y reproducible, permitiendo la manipulación, el análisis y la visualización de los grandes volúmenes de datos de consumo eléctrico.

### RStudio

Es un entorno de desarrollo integrado (IDE) de código abierto ampliamente utilizado para el lenguaje de programación R, especialmente popular en el análisis de datos, la estadística y la ciencia de datos. RStudio proporciona una interfaz gráfica intuitiva y completa que facilita la escritura, ejecución, depuración y gestión de código R(Grolemund, s. f.). En este proyecto RStudio se utilizó para:

- Importación de datos
- Manipulación de datos
- Transformar datos

### Tidyverse

Dentro de RStudio, se utilizó el conjunto de paquetes Tidyverse. Específicamente, bibliotecas como dplyr, tidyr y lubridate fueron empleadas para realizar tareas de manipulación, transformación y limpieza de datos. Tidyverse facilitó la reestructuración de los datos de un formato largo a uno ancho (pivot) y la gestión de las fechas, lo que fue un paso crucial para preparar los datos para el análisis de series temporales.

### Janitor

Para complementar la limpieza de datos en R, se utilizó la biblioteca Janitor. Esta herramienta simplificó tareas comunes como la estandarización de los nombres de las columnas y la eliminación de filas duplicadas, asegurando un conjunto de datos más limpio y consistente antes de su exportación para el análisis en Python.

### Python

Python es un lenguaje de programación de alto nivel, versátil y de código abierto, ampliamente utilizado en ciencia de datos, ingeniería y muchas otras áreas. Su sintaxis clara y legible, su extensa biblioteca estándar y la gran cantidad de bibliotecas de terceros disponibles para análisis de datos, aprendizaje automático y visualización lo convierten

en una herramienta ideal para esta investigación. Python permite escribir código de manera eficiente, realizar análisis complejos y automatizar tareas, lo cual es fundamental para el procesamiento de grandes volúmenes de datos de consumo eléctrico (Python Data Science Handbook | Python Data Science Handbook, s. f.).

En este proyecto, Python fue el lenguaje de programación de alto nivel que se utilizó como pilar para la mayor parte del análisis, se utilizó para escribir los scripts que ejecutaron todo el flujo de trabajo, desde la carga de los datos preprocesados hasta la clusterización, validación y visualización final.

**Pandas**

Fue la biblioteca de Python fundamental para la manipulación y el análisis de los datos. Proporcionó las estructuras de datos, principalmente el DataFrame, que permitieron almacenar y operar sobre los datos de consumo de manera eficiente. Se utilizó Pandas para cargar los datos desde los archivos CSV, realizar el análisis exploratorio, limpiar y preprocesar los datos aplicando métodos de imputación, y para manipular y transformar las series temporales para su posterior análisis.

**NumPy**

Numerical Python (NumPy) fue la biblioteca que proporcionó el soporte para la computación numérica. Se utilizó para realizar operaciones matemáticas y estadísticas sobre los datos de consumo eléctrico de manera optimizada. Sus estructuras de datos de tipo array fueron esenciales para la implementación de los algoritmos de aprendizaje automático y para el cálculo de las distancias euclidianas entre las series de tiempo.

**Scikit-learn**

Esta fue la biblioteca de Python líder que se empleó para implementar las técnicas de aprendizaje automático. A través de scikit-learn, se aplicó el algoritmo de clustering K-Means para segmentar a los clientes, se utilizó la técnica de reducción de dimensionalidad PCA para visualizar los clústeres, y se evaluó la calidad de la agrupación calculando el Índice de Davies-Bouldin. También se usó para realizar la regresión lineal en la detección de tendencias de consumo.

**Matplotlib y Seaborn**

Ambas bibliotecas se utilizaron para la visualización de datos en Python. Matplotlib fue la herramienta base para crear gráficos estáticos, mientras que Seaborn, construida sobre Matplotlib, se empleó para crear gráficos estadísticos más complejos y visualmente atractivos. Con ellas se

generaron los histogramas, gráficos de caja, gráficos de series de tiempo y los gráficos de dispersión que permitieron visualizar los clústeres resultantes.

### Yellowbrick

Yellowbrick fue una biblioteca de visualización que extendió las capacidades de scikit-learn. Se utilizó específicamente para una tarea crucial: visualizar el método del codo. Esta herramienta generó un gráfico claro que ayudó a determinar el número óptimo de clústeres ($k$) a utilizar en el algoritmo K-Means, facilitando la interpretación y justificación de esta elección.

### Scipy

Esta biblioteca se utilizó para la computación científica y técnica. En este proyecto, su función principal fue complementar a NumPy en tareas específicas, como el cálculo de la distancia euclidiana entre series de tiempo, una métrica que fue fundamental tanto para el algoritmo K-Means como para medir la similitud entre diferentes patrones de consumo.

### Folium

Folium fue la biblioteca de Python que se empleó para la creación de los mapas interactivos. Permitió superponer los datos geográficos de los clientes identificados sobre mapas base de Leaflet.js. Con Folium se añadieron marcadores, polígonos y otros elementos interactivos, lo que resultó en una visualización geográfica rica y fácil de explorar.

### Google Maps

Se utilizó la API de Google Maps a través de la biblioteca googlemaps de Python. Su función en este trabajo fue la de geocodificación; es decir, se empleó para convertir las direcciones postales de los clientes, obtenidas de la base de datos, en coordenadas geográficas (latitud y longitud), un paso indispensable para poder ubicarlos en el mapa interactivo.

### Google Colab

Google Colab es un entorno de desarrollo integrado (IDE) basado en la nube que permite escribir y ejecutar código Python. Ofrece acceso gratuito a recursos de computación, incluyendo GPUs, y facilita la colaboración y el uso compartido de cuadernos de código (Google Colab, s. f.).

Google Colab se utilizó como:

- Plataforma principal para el desarrollo, ejecución y documentación del código de análisis de datos.

### Multiprocessing

Se utilizó la biblioteca multiprocessing de Python, y en particular su módulo Pool, como una herramienta para la computación paralela. Se empleó para paralelizar la ejecución de funciones que identificaban patrones en las series de tiempo, lo que permitió distribuir el trabajo entre múltiples núcleos de CPU y reducir significativamente el tiempo total de ejecución del análisis.

## POBLACIÓN Y MUESTRA

### POBLACIÓN

La presente investigación considero dos grupos de clientes de la Empresa Eléctrica Quito (EEQ) como población de estudio:

1. **Clientes residenciales:** Este grupo constituyo la población principal del estudio, estando compuesto por 1.102.770 clientes residenciales que recibían servicio eléctrico de la EEQ hasta su corte. Se caracterizo por ser el segmento de consumidores con mayor número de usuarios y donde está previsto una creciente adopción de vehículos eléctricos (VE) para uso particular.
2. **Clientes con medidor exclusivo (No Residenciales):** Este grupo secundario, que abarco 178 clientes, se caracterizó por poseer un medidor exclusivo para la carga de VE, separado de su consumo residencial. Si bien represento un porcentaje menor, su inclusión en el análisis fue relevante, ya que exhibió patrones de consumo diferenciados o ser indicativos de casos particulares, como clientes con una segunda vivienda o con un uso intensivo de VE.

### MUESTRA

Una muestra se describe como un subconjunto finito y representativo que se extrae de una población con el fin de examinarlo y obtener conclusiones generalizables a dicha población.

Sin embargo, para esta investigación se tomó la decisión metodológica de analizar la población completa de clientes de la EEQ, en lugar de seleccionar y trabajar sobre una muestra. Esta determinación se fundamentó en la disponibilidad de los datos de consumo para la totalidad de la población de estudio y en las ventajas que ofrecía este enfoque:

- **Análisis exhaustivo:** Permitió un análisis completo de todos los patrones de consumo presentes, sin el riesgo de omitir información relevante que podría ocurrir al trabajar con un subconjunto.

- **Precisión y robustez:** Se incrementó la precisión en la identificación de los patrones de consumo asociados a la carga de VE, lo cual fue especialmente importante dado que la proporción de estos clientes en la población general era desconocida y potencialmente baja.
- **Eliminación del error de muestreo:** Al utilizar todos los datos disponibles, se aseguró que los resultados reflejaran las características reales de la población, eliminando la variabilidad y el sesgo que pueden ser introducidos por la selección de una muestra.

Si bien el análisis de la población completa implicó procesar un volumen de datos considerablemente mayor, los instrumentos y la capacidad computacional disponibles permitieron llevar a cabo esta tarea de manera eficiente.

## HIPÓTESIS

A continuación, se presenta la hipótesis fundamental que guio esta investigación. Su formulación establece una relación precisa y verificable entre las variables de estudio, la cual fue sistemáticamente evaluada a través del análisis empírico de los datos.

Considerando la variable independiente como la presencia de carga de un Vehículo Eléctrico (VE) sin medidor dedicado y la variable dependiente como el patrón de consumo eléctrico mensual, se formuló la siguiente hipótesis:

***"Se postula que la presencia de la variable independiente (carga de un VE) induce una alteración estadísticamente significativa y observable en la variable dependiente (patrón de consumo eléctrico). Específicamente, se hipotetiza que esta alteración se manifiesta como un incremento abrupto y persistente en la serie temporal del consumo del cliente. La comprobación de esta relación causal permitirá validar un método indirecto para la identificación y clasificación fiable de usuarios potenciales de VE dentro del área de estudio."***

# CAPÍTULO III

# DESARROLLO DE LA INVESTIGACIÓN

## METODOLOGÍA DE MINERÍA DE DATOS APLICADA: CRISP-DM

Para el desarrollo del presente estudio y la consecución de los objetivos planteados, se ha adoptado como marco de trabajo la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta es la metodología estándar de facto en la industria para llevar a cabo proyectos de minería de datos, ya que proporciona un enfoque estructurado y probado que guía el proceso desde la comprensión del problema de negocio hasta el despliegue de la solución.

La metodología CRISP-DM se compone de seis fases iterativas, las cuales se han adaptado para guiar la estructura de esta investigación:

1. **Comprensión del negocio (Business Understanding):** Esta fase inicial se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva de negocio. En este estudio, esta fase se abordó en el Capítulo I, donde se definió la problemática de la carga no registrada de vehículos eléctricos (VE), su impacto en la red de la EEQ y se establecieron los objetivos de la investigación.
2. **Comprensión de los datos (Data Understanding):** Consiste en la recolección inicial de datos y su análisis exploratorio para familiarizarse con la información, identificar problemas de calidad y descubrir primeras ideas. Esta fase se correspondió con la primera parte de la FASE 1: PROCESAMIENTO Y PREPARACIÓN DE BASE DE DATOS, donde se realizó el Análisis Exploratorio Inicial (EDA).
3. **Preparación de los datos (Data Preparation):** Cubre todas las actividades para construir el conjunto de datos final que será utilizado en la fase de modelado. Incluye la selección, limpieza, construcción y formato de los datos. Esta fase se correspondió directamente con el resto de la FASE 1 de este trabajo, incluyendo el tratamiento de valores atípicos, el filtrado y la transformación de las series temporales.

4. **Modelado (Modeling):** En esta fase se seleccionan y aplican diversas técnicas de modelado (algoritmos de minería de datos) y se calibraron sus parámetros para obtener resultados óptimos. Esta fase se alineo con la FASE 2: ALGORITMO DE CLASIFICACIÓN K-MEANS, donde se aplicó el clustering, se determinó el número de clústeres y se crearon los perfiles de referencia.
5. **Evaluación (Evaluation):** Antes de desplegar el modelo, es crucial evaluarlo rigurosamente para asegurar que cumple con los objetivos de negocio. En esta investigación, la evaluación se llevó a cabo en la FASE 3: VALIDACIÓN Y OBTENCIÓN DE POTENCIALES CLIENTES CON VE, donde se utilizaron los estratos, se aplicaron filtros de validación y se evaluó la calidad del clustering final mediante el índice de Davies-Bouldin.
6. **Despliegue (Deployment):** La fase final consiste en poner el modelo en producción o presentarlo de una manera útil para los tomadores de decisiones. En este caso, el despliegue se materializo en la FASE 4: MAPEO GEOGRÁFICO, donde los resultados se presentan en un mapa interactivo georreferenciado, una herramienta de alto valor para la planificación estratégica de la EEQ.

La Tabla 2 resume la correspondencia entre las fases de la metodología CRISP-DM y las fases desarrolladas en este documento:

**Tabla 2**

Aplicación del marco CRISP-DM a las fases de la investigación.

| Fase CRISP-DM | Fase desarrollada en el documento | Actividades principales |
|---|---|---|
| 1. Comprensión del negocio | Capítulo I: Introducción | Definición del problema, justificación y objetivos. |
| 2. Comprensión de los datos | FASE 1 (Parte A) | Análisis Exploratorio de Datos (EDA) inicial. |
| 3. Preparación de los Datos | FASE 1 (Parte B) | Limpieza, tratamiento de atípicos, filtrado. |

| Fase CRISP-DM | Fase desarrollada en el documento | Actividades principales |
|---|---|---|
| 4. Modelado | FASE 2 | Aplicación de K-Means, creación de perfiles No-VE. |
| 5. Evaluación | FASE 3 | Validación con estratos, clustering final, índice Davies-Bouldin. |
| 6. Despliegue | FASE 4 | Geocodificación y generación de mapa interactivo |

**FASE 1: PROCESAMIENTO Y PREPARACIÓN DE BASE DE DATOS**

**Método principal:** Preparación, limpieza y filtrado inicial de las series de tiempo de consumo eléctrico. El objetivo fue acondicionar los datos para el análisis posterior, eliminando ruido y preseleccionando series relevantes mediante un criterio de aumento sostenido

**PREPARACIÓN DE LA BASE DE DATOS**

Dada la magnitud de la base de datos (aproximadamente 65 GB) y las limitaciones de la capacidad computacional en el entorno inicial, se determinó que R sería la herramienta más eficiente para la fase inicial de preparación de datos. La elección de R se fundamenta en su robustez y eficacia para el procesamiento de grandes volúmenes de información en este contexto específico.

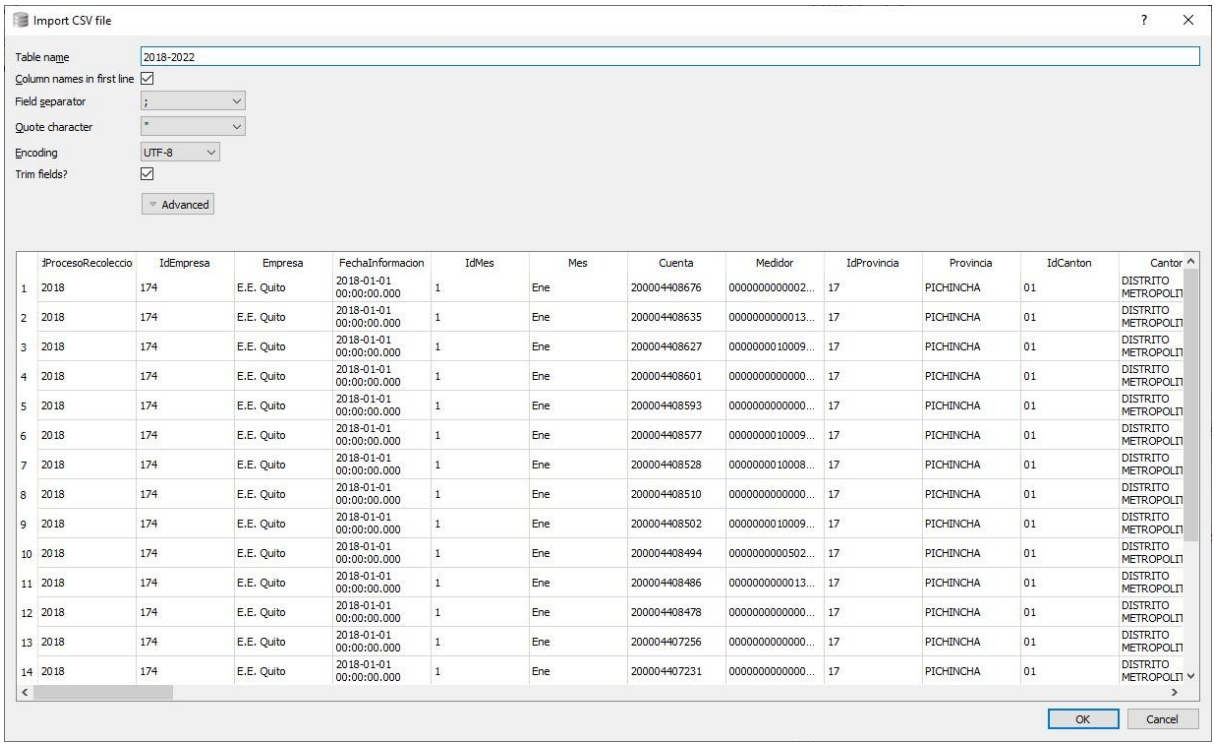
La base de datos se encontraba estructurada en formato largo con más de 80 millones de registros, validos desde enero de 2018 hasta abril de 2019, como se puede apreciar en la Imagen 1. Este formato se caracteriza por presentar múltiples filas para una misma entidad, donde cada fila contiene una observación o registro individual. La estructura de formato largo es común en la recopilación de datos a lo largo del tiempo, ya que permite registrar múltiples mediciones para cada entidad sin necesidad de crear una columna separada para cada punto temporal.

Si bien el formato largo es eficiente para el almacenamiento y la gestión de grandes volúmenes de datos, puede presentar desafíos para el análisis y la visualización. Por esta

razón, en etapas posteriores del análisis, se transformó la base de datos a un formato ancho, donde cada entidad (cliente) se representó en una única fila, y las mediciones de consumo se organizarán en columnas, lo que facilitó la aplicación de técnicas de análisis y la generación de visualizaciones informativas.

**Imagen 1**

Base de datos clientes residenciales EEQ.




| | IProcesoRecoleccio | IdEmpresa | Empresa | FechaInformacion | IdMes | Mes | Cuenta | Medidor | IdProvincia | Provincia | IdCanton | Cantor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408676 | 0000000000002... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 2 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408635 | 0000000000013... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 3 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408627 | 0000000010009... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 4 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408601 | 0000000000000... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 5 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408593 | 0000000000000... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 6 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408577 | 0000000010009... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 7 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408528 | 0000000010008... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 8 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408510 | 0000000000000... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 9 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408502 | 0000000010009... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 10 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408494 | 0000000000502... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 11 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408486 | 0000000000013... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 12 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004408478 | 0000000000000... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 13 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004407256 | 0000000000000... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |
| 14 | 2018 | 174 | E.E. Quito | 2018-01-01 00:00:00.000 | 1 | Ene | 200004407231 | 0000000000000... | 17 | PICHINCHA | 01 | DISTRITO METROPOLI |

La base datos de la EEQ tiene en total 87 variables las cuales se presentan en la Tabla 3.

**Tabla 3**

Variables de la base de datos EEQ.

| **Variable** | **Tipo de Variable** | **Variable** | **Tipo de Variable** |
|---|---|---|---|
| IdProcesoRecoleccion | Numérico (Entero) | Parroquia | Texto |
| IdEmpresa | Numérico (Entero) | NombreAbonado | Texto |
| Empresa | Texto | UbicacionAbonado | Texto |

| Variable | Tipo de Variable | Variable | Tipo de Variable |
|---|---|---|---|
| fechaInformacion | Fecha | FechaNacimientoAbonado | Fecha |
| IdMes | Numérico (Entero) | EsTercera Edad | Texto (Categórico) |
| Mes | Texto | CedulaAbonado | Texto |
| Cuenta | Texto | NivelVoltaje | Texto (Categórico) |
| Medidor | Texto | GrupoConsumo | Texto (Categórico) |
| IdProvincia | Numérico (Entero) | CodigodGrupoConsumo | Texto (Código) |
| Provincia | Texto | Tarifa | Texto (Código) |
| IdCanton | Numérico (Entero) | Desc_Tarifa | Texto |
| Canton | Texto | TipoCliente | Texto (Categórico) |
| IdParroquia | Numérico (Entero) | Energia | Numérico (Decimal) |
| DemandaHorasPico | Numérico (Decimal) | SubsidioCruzadoCalcula | Numérico (Decimal) |
| DemandaFacturada | Numérico (Decimal) | SubcidioSolidarioCalcula | Numérico (Decimal) |
| EnergiaReactiva | Numérico (Decimal) | ValorComercial | Numérico (Decimal) |
| ValorComercializacion | Numérico (Decimal) | FacturacionServicio | Numérico (Decimal) |

| Variable | Tipo de Variable | Variable | Tipo de Variable |
|---|---|---|---|
| DtoAlumbradoP | Numérico (Decimal) | RecibeSubsidio | Texto (Categórico) |
| TotalRecoleccion | Numérico (Decimal) | Subsidio | Texto (Categórico) |
| InpuestoBomberos | Numérico (Decimal) | SubsidioTerceraEd | Numérico (Decimal) |
| FacturaServicioElect | Numérico (Decimal) | SubsidioCruzado | Numérico (Decimal) |
| FacturaElectrica | Numérico (Decimal) | emaSubsidioCruz | Numérico (Decimal) |
| FacturaTotal | Numérico (Decimal) | TotalSubsidioCruzado | Numérico (Decimal) |
| ValorRefacturacion | Numérico (Decimal) | TotalFacturaSinTarD | Numérico (Decimal) |
| EnergiaRefacturacion | Numérico (Decimal) | TotalFacturaConTarD | Numérico (Decimal) |
| MesesMoraRefacturacion | Numérico (Decimal) | SubsidioTarifaD | Numérico (Decimal) |
| MesesAdeudados | Numérico (Decimal) | SistemaTotalFacturacion | Texto (Código) |
| RecaudacionServicioEle | Numérico (Decimal) | Cliente | Numérico (Entero) |
| RecaudacionTotal | Numérico (Decimal) | TipoMedidor | Texto (Categórico) |

| Variable | Tipo de Variable | Variable | Tipo de Variable |
|---|---|---|---|
| SistemaValorEnergia | Texto | Transformador | Texto (Código) |
| ConsumoResidencial | Numérico (Decimal) | SubsidioLeyDisca | Numérico (Decimal) |
| IngresoSisComi | Fecha | LineaBase | Numérico (Decimal) |
| Energía Electrica | Numérico (Decimal) | KwhSubsidio | Numérico (Decimal) |
| FactorCorreccion | Numérico (Decimal) | Equipamiento | Texto (Código) |
| CargoDemanda | Numérico (Decimal) | ValorSubsidiadoPe | Numérico (Decimal) |
| InfraccionBajoFactorf | Numérico (Decimal) | FacturadoDamnificadoVolcanTungurahua | Texto (Categórico) |
| FactorPotencia | Numérico (Decimal) | RefacturadaSubsido | Texto (Código) |
| FinanciamientoCocinaInduccion | Numérico (Decimal) | CuentaContrato | Texto |
| InstalacionCircuitoExpreso | Numérico (Decimal) | SubsidioTarifaElectrica | Numérico (Decimal) |

La aplicación para preparar la base de datos consistió en dos técnicas clave:

- **Reducción de la base de datos:** Se llevó a cabo una selección y extracción de las variables esenciales para el análisis, descartando aquellas columnas que no aportaban información relevante. Esta estrategia permitió disminuir el tamaño del dataset, optimizando así el procesamiento en las fases subsiguientes.

- **Exportación de datos procesados:** La base de datos, una vez pre-procesada, fue almacenada en un formato eficiente (CSV) para facilitar su importación en Python, donde se llevarían a cabo las siguientes fases del análisis.

Se aprovecharon las capacidades de R para la gestión eficiente de grandes conjuntos de datos, así como para la ejecución de operaciones de filtrado y selección de variables. Se utilizaron tanto las funcionalidades nativas de R como bibliotecas especializadas como tidyverse y lubridate, en la Imagen 2 se muestra lo descrito.

**Imagen 2**

Base de datos clientes residenciales EEQ.

```r
library(tidyverse)
library(lubridate)

# Lectura y preprocesamiento inicial (ejemplo)
datos <- read_csv("datos/2018-2022.csv", show_col_types = FALSE) %>%
  select(-ConsumoResidencial, -EnergiaElectrica) %>%
  mutate(Energia = replace_na(Energia, 0),
         Mes_1 = paste(Mes, IdProcesoRecoleccion, sep = " ")) %>%
  select(-IdProcesoRecoleccion, -Mes, -CuentaContrato)

# Funcion para segmentar y pivotar ( importante !)
procesar_segmento <- function(datos, meses) {
  datos %>%
    filter(Mes_1 %in% meses) %>%
    pivot_wider(names_from = Mes_1, values_from = Energia, values_fill = 0)
}

# ... (procesamiento intermedio, como limpieza de columnas) ...

# Union de DataFrames
lista_df <- list(df_2018_2019, df_2020_2021, df_2022, df_2023_24)
df_final <- Reduce(function(x, y) merge(x, y, by = c("Cuenta", "Medidor"), all =
   TRUE), lista_df)

# Renombrado de Columnas
nombres_meses <- grep("[A-Za-z]{3} \\d{4}", names(df_final), value = TRUE)
nombres_nuevos <- format(parse_date_time(nombres_meses, orders = "b Y"), "%B-%Y")
nombres_nuevos <- str_to_title(nombres_nuevos)
names(df_final)[names(df_final) %in% nombres_meses] <- nombres_nuevos

# Eliminacion de columnas redundantes
df_final <- df_final %>% select(-matches("GrupoConsumo\\.[xy]"))

# Guardar el dataframe final
write.csv(df_final, "datos/df_18_24.csv", row.names = FALSE)
```

## ANÁLISIS EXPLORATORIO INICIAL (EDA) Y DETECCIÓN DE ANOMALÍAS

En la fase de análisis exploratorio de datos, se aplicaron técnicas de estadística descriptiva utilizando la biblioteca Pandas (funciones describe (), mean (), std (), quantile ()) para identificar rangos de valores típicos, distribuciones y posibles errores o inconsistencias en las series de tiempo. Adicionalmente, se emplearon las bibliotecas Matplotlib y Seaborn para generar visualizaciones (histogramas, boxplots, gráficos de líneas) que

permitieron inspeccionar gráficamente las series de tiempo, detectar visualmente valores atípicos y patrones inusuales. La Imagen 3 ilustra esta técnica.

**Imagen 3**

Fragmento de código Python: Análisis exploratorio inicial (EDA) y detección de anomalías.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import drive

# Montar Google Drive
drive.mount('/content/drive')

# Cargar datos
datos = pd.read_csv("/content/drive/MyDrive/df_series_final.
    csv")

# Seleccionar columnas de consumo
columnas_consumo = ['Agosto-2022', 'Septiembre-2022', '
    Octubre-2022']
datos_consumo = datos[columnas_consumo]

# Estad sticas descriptivas
print("Estad sticas descriptivas:\n", datos_consumo.describe
    ())

# Visualizaciones
fig, axes = plt.subplots(1, 2, figsize=(14, 6))

# Histograma
sns.histplot(datos_consumo['Agosto-2022'], kde=True, ax=axes
    [0])
axes[0].set_title('Distribuci n del Consumo (Agosto-2022)')
axes[0].set_xlabel('Consumo Energ tico (kWh/mes)')
axes[0].set_ylabel('Frecuencia')

# Detecci n de outliers
Q1 = datos_consumo['Agosto-2022'].quantile(0.25)
Q3 = datos_consumo['Agosto-2022'].quantile(0.75)
IQR = Q3 - Q1
limite_inferior = Q1 - 1.5 * IQR
limite_superior = Q3 + 1.5 * IQR

# Contar outliers
num_outliers = datos_consumo[(datos_consumo['Agosto-2022'] <
    limite_inferior) | (datos_consumo['Agosto-2022'] >
    limite_superior)].shape[0]
print(f"\nN mero de outliers en Agosto-2022: {num_outliers}"
    )
```

## TRATAMIENTO DE VALORES ATÍPICOS Y FALTANTES

Para abordar los valores faltantes en las series de tiempo durante la aplicación de la metodología, se emplearon dos técnicas de imputación:

- Sustitución por la media del conjunto de datos para valores faltantes generales;
- Imputación por interpolación lineal para valores atípicos extremos o errores evidentes, estimando el valor faltante a partir de una interpolación lineal entre los valores de los meses anterior y posterior.

Se optimizó el código mediante funciones específicas y se implementó un esquema de paralelización para reducir el tiempo de cómputo, utilizando la biblioteca Pandas de Python. La Imagen 4 ilustra este proceso.

**Imagen 4**

Fragmento de código Python: Tratamiento de valores atípicos y faltantes.

```
import pandas as pd
import numpy as np
from google.colab import drive

# Cargar datos
datos = pd.read_csv("/content/drive/My Drive/df_18_24_final.
    csv")

# Convertir 'Cuenta' a entero y las columnas de meses a
    num ricas
datos['Cuenta'] = pd.to_numeric(datos['Cuenta'], downcast='
    integer')
meses = datos.columns[2:]
datos[meses] = datos[meses].apply(pd.to_numeric, errors='
    coerce')

def filtrar_usuarios(df, umbral_consumo=110,
    umbral_desviacion=0.1):
    """Filtra usuarios basados en umbrales."""
    df['Consumo_Promedio'] = df[meses].mean(axis=1)
    df['Desviacion_Relativa'] = df[meses].std(axis=1) / df['
        Consumo_Promedio']
    return df[(df['Consumo_Promedio'] > umbral_consumo) &
              (df['Desviacion_Relativa'] > umbral_desviacion)
                 ].drop(['Consumo_Promedio', '
                 Desviacion_Relativa'], axis=1)

# Limpieza de datos
datos = filtrar_usuarios(datos.copy())
datos = datos.drop('Medidor', axis=1, errors='ignore')
datos = datos.set_index('Cuenta')

def corregir_atipicos(serie):
    """Corrige valores at picos usando el m todo IQR."""
    Q1 = serie.quantile(0.25)
    Q3 = serie.quantile(0.75)
    IQR = Q3 - Q1
    limite_inferior = Q1 - 1.5 * IQR
    limite_superior = Q3 + 1.5 * IQR
    promedio = serie[(serie >= limite_inferior) & (serie <=
        limite_superior)].mean()
    serie[(serie < limite_inferior) | (serie >
        limite_superior)] = promedio
    serie[serie > 1e6] = promedio
    return serie

# Aplicar correcci n de at picos
df_final = datos.copy()
for index, row in df_final.iterrows():
    df_final.loc[index] = corregir_atipicos(row)

df_final = filtrar_usuarios(df_final.copy()) # Segunda
    depuraci n
df_final.to_csv('/content/drive/My Drive/df_series_final.csv'
    , index=True) #Guardar datos.
```

## FILTRAR CONSTANTES Y TARIFA DIGNIDAD

Una vez corregidos los valores faltantes y erróneos en las series de tiempo, se aplicó un primer filtro con el objetivo de eliminar aquellas series que no cumplían con ciertos criterios, considerados relevantes para el análisis. Este proceso de filtrado inicial se llevó a cabo en dos etapas:

- **Filtrado por variabilidad**: Se eliminaron todas aquellas series de tiempo cuyo consumo no presentaba una desviación estándar superior al 1%. Este criterio se

basa en la premisa de que las series de tiempo con muy poca variabilidad en el consumo no son de interés para el análisis, ya que no reflejan patrones de consumo dinámicos y característicos.

- **Filtrado por tarifa dignidad:** Se eliminaron todas aquellas series de tiempo cuyo consumo promedio a lo largo del tiempo era menor o igual a 110 kWh. Este criterio se fundamenta en la observación de que los clientes con un consumo tan bajo no se ajustan al perfil de usuarios típicos de vehículos eléctricos, cuyo consumo suele ser significativamente mayor.

La aplicación de estos criterios de filtrado permitió descartar series de tiempo que no aportaban información relevante para el análisis, optimizando así el procesamiento en las etapas subsiguientes y mejorando la calidad de los resultados obtenidos.

**FASE 2: ALGORITMO DE CLASIFICACIÓN K-MEANS**

**Método Principal:** Agrupamiento de las series de tiempo preseleccionadas (con aumento de consumo) mediante el algoritmo K-Means.

**DETECCIÓN DE AUMENTO EN EL CONSUMO**

Para identificar series de tiempo con aumentos sostenidos en el consumo eléctrico, se implementó tres etapas principales:

- Cálculo de diferencias entre valores consecutivos para identificar puntos de aumento;
- Aislamiento de incrementos significativos mediante la selección de diferencias positivas;
- Validación de la sostenibilidad del aumento, comparando el valor mínimo del consumo posterior al incremento con el valor del consumo en el momento del incremento.

Se consideró un aumento si el consumo no disminuye por debajo del nivel previo al incremento. Las bibliotecas Pandas, Matplotlib y Scikit-learn fueron utilizadas en este análisis, cuyo desarrollo se presenta en la Imagen 5.

**Imagen 5**

Fragmento de código Python: Identificación de series temporales con aumento de consumo.

```
import pandas as pd
from google.colab import drive

# Se asume que 'series_tiempo_final' es una lista de pandas
    Series.
# Ejemplo de carga (reemplazar con la carga real):
drive.mount('/content/drive')
datos = pd.read_csv("/content/drive/MyDrive/df_series_final.
    csv")
series_tiempo_final = []
for index, row in datos.iterrows():
    nombre_serie = row['cuenta']
    serie = pd.Series(row[1:], name=nombre_serie)
    series_tiempo_final.append(serie)

def procesar_serie(serie):
    """Verifica si una serie tiene un aumento sostenido."""
    diferencias = serie.diff()
    aumentos = diferencias[diferencias > 0]
    aumento_sostenido = False
    if len(aumentos) > 0:
        indice_aumento = aumentos.index[0]
        consumo_despues_aumento = serie[indice_aumento:]
        if consumo_despues_aumento.min() >= serie[
            indice_aumento]:
            aumento_sostenido = True
    return aumento_sostenido

# Aplicar la funci n a cada serie
series_con_aumento = []
for serie in series_tiempo_final:
    if procesar_serie(serie):
        series_con_aumento.append(serie)

series_tiempo_final = series_con_aumento

print(f"N mero de series con aumento sostenido: {len(
    series_tiempo_final)}")
```

## CLUSTERING INICIAL CON K-MEANS

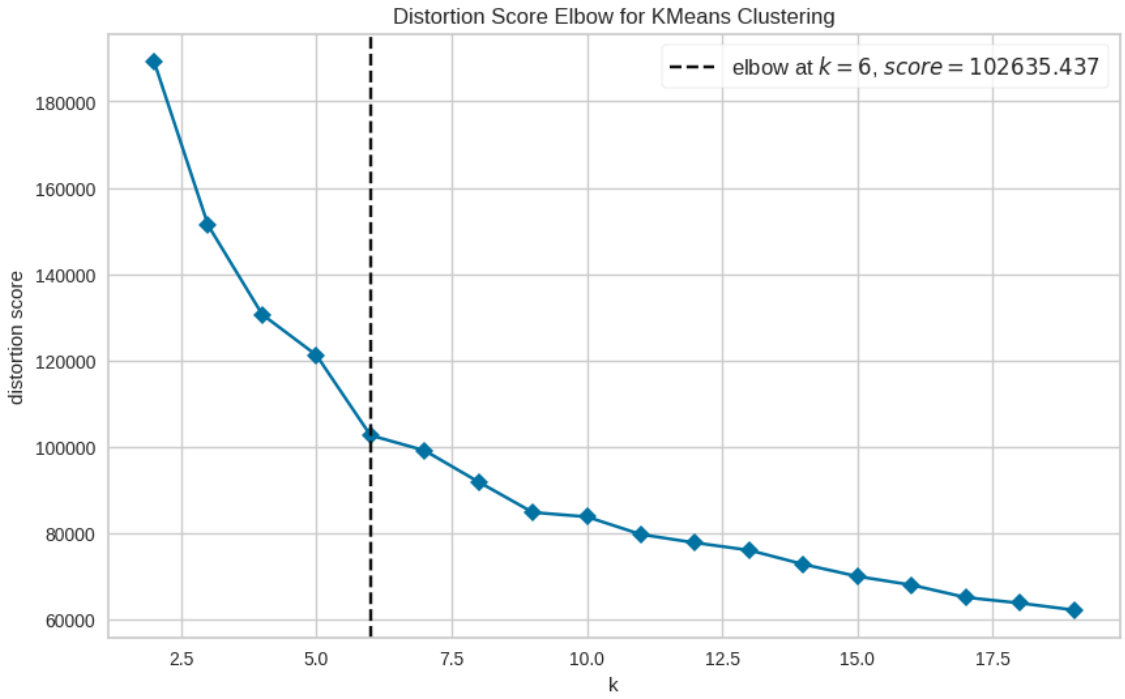
Una vez identificadas las series temporales con aumento de consumo energético se aplicó el algoritmo de clustering K-Means. Un aspecto fundamental en la aplicación de K-Means es la determinación del número óptimo de clústeres (k). Para ello, se implementó el método del codo, una técnica ampliamente utilizada que consiste en iterar el algoritmo K-Means para un rango de valores de k y calcular la inercia para cada iteración. La inercia representa la suma de las distancias al cuadrado de cada punto al centroide de su clúster asignado, por lo que un valor bajo de inercia indica una mayor cohesión interna en los clústeres.

La visualización de la curva del codo, obtenida al graficar la inercia en función de k, permite identificar el punto en el que la disminución de la inercia se vuelve menos pronunciada, asemejándose a la forma de un codo. Este punto se considera el valor óptimo

de k, ya que, a partir de él, aumentar el número de clústeres aporta una mejora marginal en la reducción de la inercia, pudiendo incurrir en un sobreajuste de los datos. La implementación de este proceso se llevó a cabo utilizando las bibliotecas Scikit-learn y Matplotlib/Seaborn de Python. Scikit-learn. El Gráfico 1 ilustra los resultados de este análisis, mostrando la curva del codo y el valor de k.

**Gráfico 1**

Aplicación del método del codo (inicial).




## VISUALIZACIÓN Y ANÁLISIS DE CLÚSTER INICIALES

Con el fin de visualizar la distribución de los clústeres de series de tiempo en un espacio de menor dimensionalidad, se aplicó la técnica de reducción de dimensionalidad mediante PCA. Esta técnica permitió proyectar las series de tiempo desde su espacio original de alta dimensionalidad a un espacio bidimensional, facilitando su representación gráfica y análisis visual.

La aplicación de PCA se realizó utilizando las bibliotecas Scikit-learn, Pandas y Matplotlib/Seaborn de Python. En primer lugar, se prepararon los datos para el análisis, lo que implicó la estandarización de las series de tiempo para asegurar que todas las variables contribuyeran de manera equitativa al análisis. Posteriormente, se aplicó la

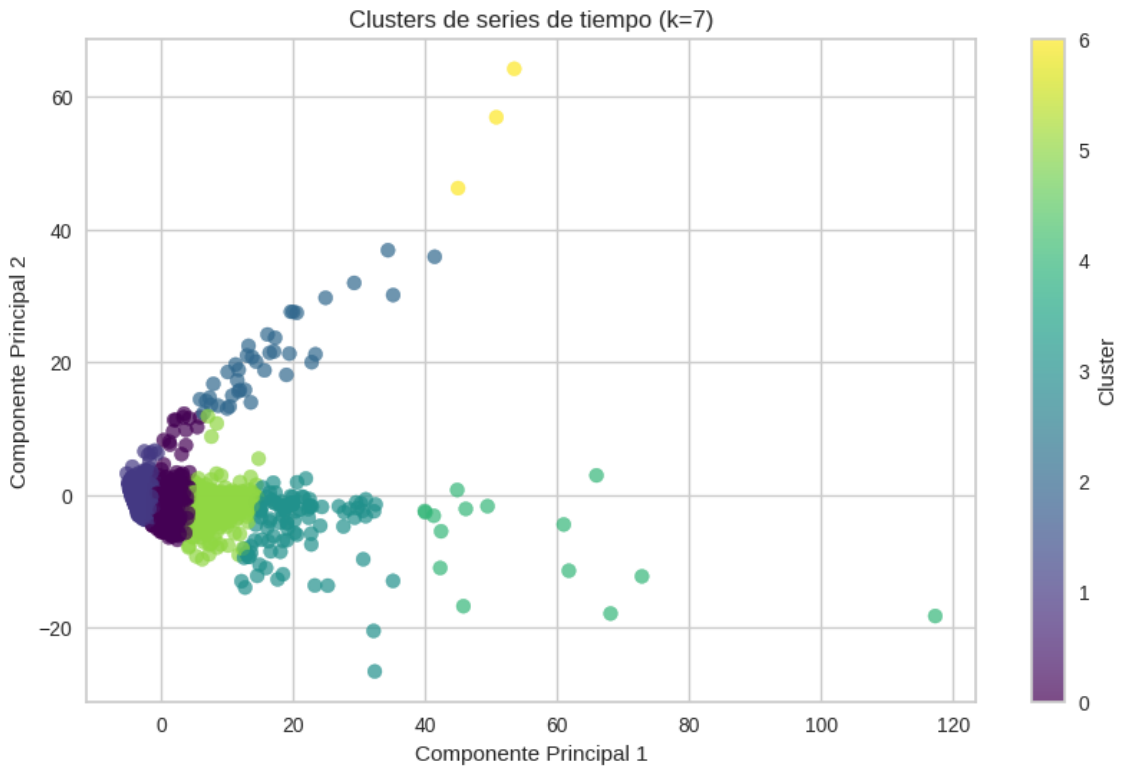
técnica de PCA para obtener las dos componentes principales que mejor capturan la variabilidad de los datos originales.

- **Componente principal 1**: Es el componente que explica la mayor parte de la varianza en los datos originales. Representa el patrón general de consumo energético a lo largo del tiempo.
- **Componente principal 2:** Representa variaciones en el patrón de consumo, como diferencias en la estacionalidad o en los picos de consumo.

El Gráfico 2 muestra la distribución de los clústeres de series de tiempo en el espacio bidimensional resultante tras la aplicación de PCA. Cada punto en el gráfico representa una serie de tiempo individual, y su color indica el clúster al que pertenece.

**Gráfico 2**

Reducción PCA.




## CREACIÓN Y APLICACIÓN DEL PERFIL DE REFERENCIA NO-VE

Tras la aplicación del algoritmo K-Means a las series de tiempo con aumentos sostenidos en el consumo eléctrico, se llevó a cabo un análisis de los clústeres resultantes con el objetivo de identificar aquellos que no se ajustaban al perfil de usuarios de vehículos eléctricos (VE). Para ello, se examinaron las características de cada clúster, incluyendo

su consumo promedio, variabilidad y patrones temporales, buscando desviaciones significativas respecto a los patrones

Para discriminar series de tiempo que no se ajustan al perfil de usuarios de VE's, se implementó una metodología basada en la comparación con un perfil de referencia.

Inicialmente, se crea un perfil de referencia de clientes sin VE, asumiendo que un clúster o clústeres específico que representan a dichos clientes. Se calcula el consumo promedio de energía de todos los clientes dentro de este clúster, convirtiéndolo en el *"perfil de referencia"* para clientes sin VE. Posteriormente, se calculó la distancia euclidiana entre el patrón de consumo de cada cliente y el perfil de referencia, donde una distancia mayor indica una mayor diferencia con respecto al perfil de referencia. El Gráfico 3 representa un ejemplo de cómo se realiza este proceso.

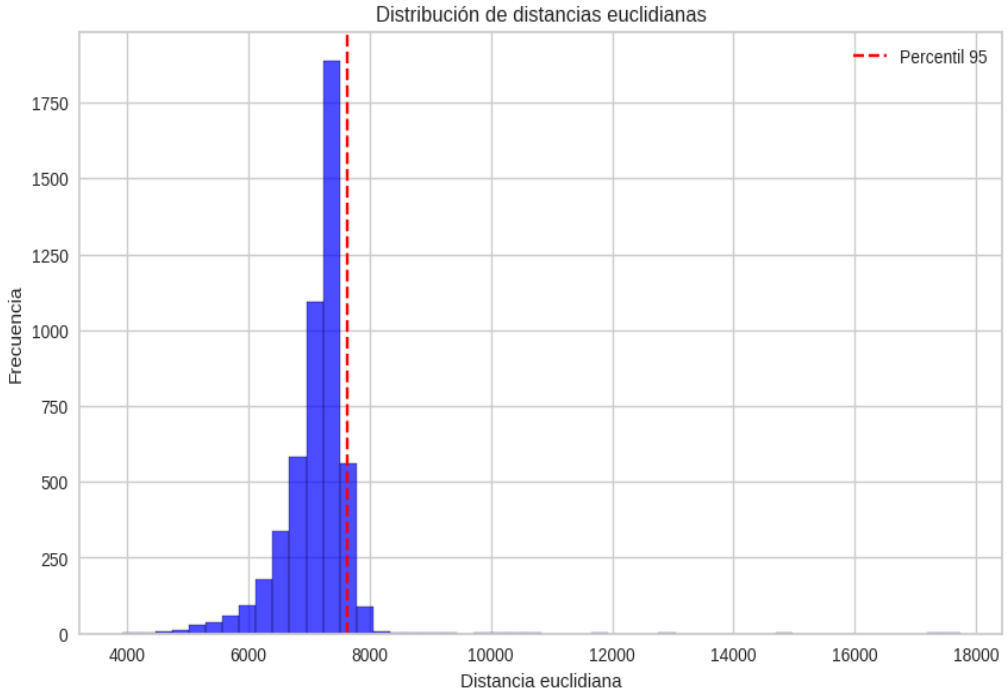
**Gráfico 3**

Series de tiempo en clústeres.

Posteriormente se visualiza la distribución de estas distancias mediante un histograma, lo que permitió comprender cómo de diferentes son los patrones de consumo de los clientes con respecto al perfil de referencia. Se estableció un umbral de similitud (en este caso, el percentil 95 de las distancias) para identificar clientes que se desvían significativamente del perfil de referencia, considerando a los clientes con distancias por encima de este umbral como potenciales propietarios de VE, tal como ilustra en el Gráfico 4. Finalmente, se filtraron los datos para seleccionar solo a los clientes que no pertenecen al clúster de referencia y que tienen una distancia al perfil de referencia mayor o igual al umbral de similitud, obteniendo un nuevo conjunto de datos que contiene información solo de los clientes que, según su patrón de consumo, podrían tener VE'S. Este nuevo conjunto de datos se utilizó en análisis posteriores para estudiar más a fondo a estos clientes potenciales. Adicional se calcula el índice Davies-Boulding para su posterior comparación con los resultados finales. Se utilizan bibliotecas como Scikit-learn, Matplotlib/Seaborn y Pandas.

**Gráfico 4**

Distribución de distancias euclidianas.



## ANÁLISIS Y ESTRATIFICACIÓN DE CLIENTES CON VEHÍCULOS ELÉCTRICOS

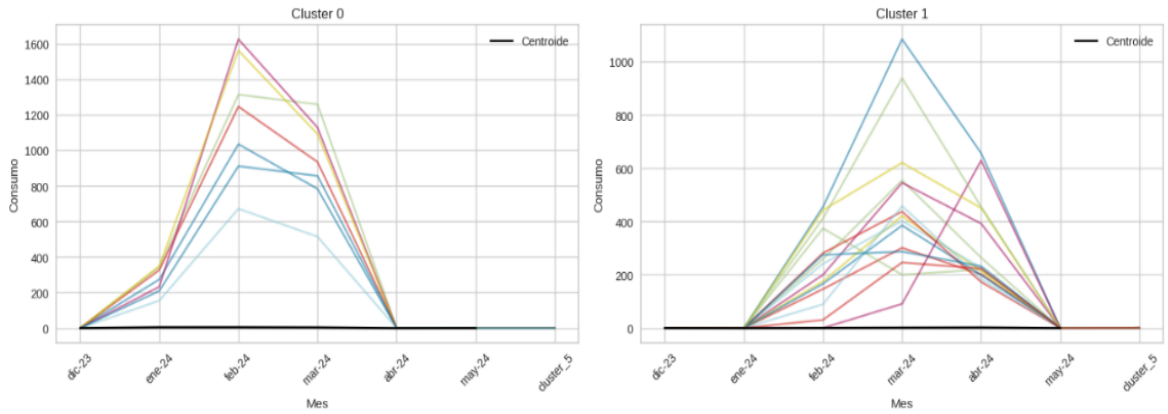
Para el análisis y estratificación de clientes con vehículos eléctricos se estructura en tres fases principales:

**Recolección y preparación de datos:** En esta fase inicial, se procedió a la obtención de los datos de consumo eléctrico de los clientes con VE, provenientes de medidores dedicados a la carga del vehículo. Los datos son organizados en un formato de series de tiempo, donde cada serie representa el consumo de un cliente específico a lo largo del tiempo. Posteriormente, se realiza una limpieza de los datos, corrigiendo valores de consumo nulos o atípicos, y se transforman a una escala común para evitar que las diferencias de magnitud afecten el análisis. Se emplea la librería Pandas y NumPy para realizar cálculos numéricos y manipulación de datos.

**Clustering:** Los clústeres generados son visualizados para comprender las características de cada grupo, mediante gráficos que muestran la evolución del consumo a lo largo del tiempo para los clientes de cada clúster. El Gráfico 5 muestra que los diferentes clústeres exhiben patrones de consumo diferenciados, lo que sugiere la existencia de distintos perfiles de usuarios de vehículos eléctricos Se denota que existen series que no aportan información relevante para el estudio y no se ajustaba a los patrones de interés para el análisis. Se utiliza Scikit-learn de Python, la cual para el aprendizaje automático y la clase StandardScaler para escalar los datos. Adicionalmente, se emplea la librería Yellowbrick de Python, que extiende Scikit-learn y proporciona visualizaciones para facilitar la comprensión de los resultados del clustering.

**Gráfico 5**

Clúster de clientes con VE’s.

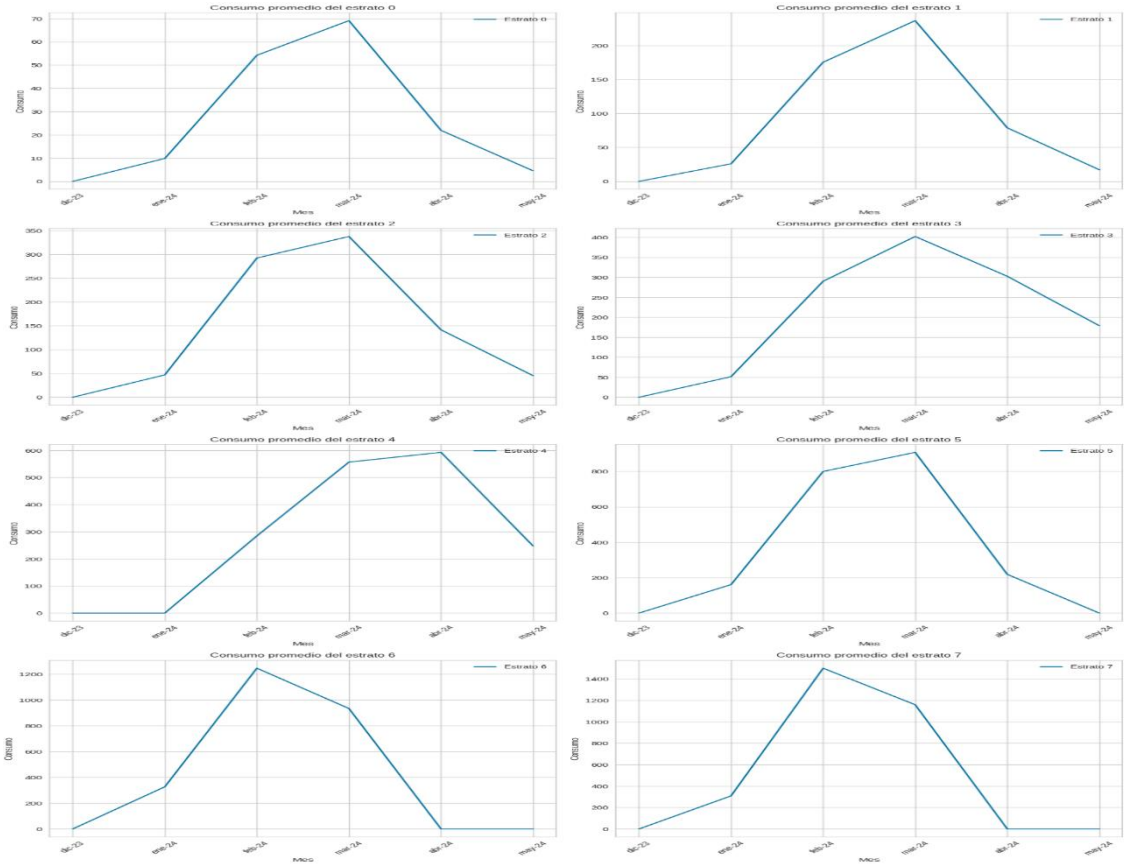


**Estratificación por consumo:** En esta fase, se calculó el consumo total de energía para cada cliente con VE y se agrupan en diferentes estratos (niveles) de consumo, dividiendo el rango de consumo total en un número determinado de intervalos. La distribución de los clientes en los diferentes estratos es visualizada mediante gráficos, como histogramas. Se muestra, además, el consumo promedio de cada estrato para entender las diferencias entre ellos. Finalmente, se genera un resumen que muestra la cantidad de clientes en cada estrato, el consumo máximo y promedio de cada uno, lo que permite caracterizar los diferentes niveles de consumo entre los clientes con VE. En esta fase, se utiliza nuevamente la librería Pandas para calcular el consumo total de cada cliente y agruparlos en estratos utilizando la función pd.cut.

El Gráfico 6 muestra el consumo promedio de energía para cada estrato a lo largo del tiempo. Se puede observar que existen diferencias significativas en los patrones de consumo entre los distintos estratos, lo que sugiere que cada grupo de usuarios tiene características y necesidades energéticas particulares.

**Gráfico 6**

Estratos de VE's.



## FASE 3: VALIDACIÓN Y OBTENCIÓN DE POTENCIALES CLIENTES CON VE FILTRO HACIENDO USO DE LOS ESTRATOS DE VE'S

A continuación, se comparó a los potenciales clientes con VE's con el consumo promedio de diferentes estratos de clientes con vehículos eléctricos, buscando incrementos significativos y persistentes.

Se ha hecho uso de una función denominada detectar_aumento_por_estrato es el núcleo de este análisis. Esta función toma como entrada la serie de tiempo de consumo de un cliente individual y la tabla de estratos estratos_df. Su objetivo es determinar si el cliente ha experimentado un aumento en su consumo que se pueda asociar a la presencia de un VE.

La función opero de la siguiente manera:

- Itero a través de los diferentes estratos de consumo.
- Para cada estrato, compara el cambio en el consumo del cliente con el consumo promedio del estrato.
- Si encuentra un aumento significativo y sostenido en el consumo del cliente que se asemeje al consumo promedio de un estrato, la función considera que el cliente tiene un patrón de consumo compatible con la presencia de un VE y devuelve True. En caso contrario, devuelve False.
- La función detectar_aumento_por_estrato se aplica a cada serie de tiempo de consumo en el DataFrame df_noe, que contiene información sobre clientes potenciales con VE.
- Los resultados se almacenan en una nueva columna llamada aumento_sostenido, que indica si se detectó un aumento sostenido en el consumo para cada cliente.
- Posteriormente, se filtra el DataFrame para seleccionar solo aquellos clientes que tienen un valor True en la columna aumento_sostenido. Estos clientes son considerados como aquellos que tienen un patrón de consumo compatible con la presencia de un VE y se almacenan en un nuevo DataFrame.

Se visualizaron las series de tiempo de los clientes con aumento sostenido para inspeccionar visualmente sus patrones de consumo. Esto se realizó mediante gráficos de

líneas que muestran la evolución del consumo a lo largo del tiempo para cada cliente. Se ha utilizado Pandas y Matplotlib. El detalle de este proceso se encuentra ilustrado en la Imagen 6.

**Imagen 6**

Fragmento de código Python: Filtro haciendo uso de estrato.

```
import pandas as pd
import numpy as np

# Datos de estratos
estratos_df
# Funci n para detectar aumento sostenido relativo al
    estrato
def detectar_aumento_sostenido_estrato(serie, estrato,
    estratos_df, porcentaje_aumento=0.10, ventana=10):
    """Detecta un aumento sostenido relativo al estrato."""
    consumo_promedio_estrato = estratos_df.loc[estratos_df['
        Estrato'] == estrato, 'Consumo Promedio (kWh)'].values
        [0]
    umbral_superior = consumo_promedio_estrato * (1 +
        porcentaje_aumento)

    aumento_sostenido = False
    for i in range(len(serie) - ventana + 1):
        if np.all(serie[i:i + ventana] > umbral_superior):
            aumento_sostenido = True
            break
    return aumento_sostenido


df_noe = pd.DataFrame(data)
df_noe['serie'] = df_noe['serie'].apply(lambda x: pd.Series(x
    ))
# Aplicar funciones y filtrar
series_con_aumento = []
for idx, row in df_noe.iterrows():
    serie = row['serie']
    estrato = asignar_estrato(np.mean(serie), estratos_df)
    if detectar_aumento_sostenido_estrato(serie, estrato,
        estratos_df):
        series_con_aumento.append(serie)
print(f"N mero de series con aumento sostenido: {len(
    series_con_aumento)}")
```

## CLUSTERIZACIÓN FINAL DE CLIENTES CON POSIBLES VE'S

Tras la identificación de clientes con patrones de consumo compatibles con la presencia de vehículos eléctricos (VEs) mediante el análisis de estrato, se aplicó clúster k-means para agruparlos según similitudes en sus patrones de consumo.

Se utilizó el algoritmo K-Means, una técnica de aprendizaje automático no supervisado, para agrupar a los clientes en clústeres basados en la similitud de sus patrones de consumo. La elección del número óptimo de clústeres se llevó a cabo mediante el método del codo, una técnica que evalúa la inercia para diferentes valores de k y selecciona aquel valor a partir del cual la disminución de la inercia se vuelve menos significativa.

Para visualizar los clústeres resultantes, se empleó la técnica de reducción de dimensionalidad. PCA proyecta los datos a un espacio bidimensional, lo que permite representar cada cliente como un punto en un gráfico de dispersión, donde el color indica el clúster al que pertenece.

Adicionalmente, se creó un DataFrame que contiene información de cada cliente y el clúster asignado. Se calculó la cantidad de clientes por clúster para analizar la distribución de los clientes entre los diferentes grupos. Se utilizó el índice de Davies-Bouldin para evaluar la calidad de la Clusterización. Un valor bajo en este índice indica una buena separación entre los clústeres.

Para un análisis más profundo de los patrones de consumo característicos de cada grupo, se visualizaron las series de tiempo de los clientes dentro de cada clúster. Esta visualización permitió observar las particularidades de consumo de cada grupo y comprender mejor las diferencias entre ellos.

Finalmente, se creó una tabla que contiene los identificadores de los clientes y el clúster al que pertenecen. Esta tabla resume los resultados de la Clusterización y facilita la identificación de los clientes pertenecientes a cada grupo. La tabla final se guardó en un archivo para su posterior análisis o uso en otras aplicaciones.

Las herramientas utilizadas fueron:

- Algoritmo K-Means (Scikit-learn)
- Análisis de Componentes Principales (PCA) (Scikit-learn)
- Índice Davies-Bouldin (Scikit-learn)
- DataFrames (Pandas)
- Gráficos de dispersión y series de tiempo (Matplotlib)
- Archivos CSV (Pandas)

**FASE 4: MAPEO GEOGRÁFICO**

La etapa de geocodificación se centra en la obtención de las coordenadas geográficas (latitud y longitud) de los clientes previamente identificados como potenciales usuarios de vehículos eléctricos.

## GEOCODIFICACIÓN DE CLIENTES CON POTENCIALES VE'S

Inicialmente, se importan las librerías necesarias para la geocodificación y visualización de mapas, incluyendo Google Maps, pandas, Folium y branca. Se configuró la API de Google Maps con la clave de API proporcionada y se procedió a cargar dos conjuntos de datos esenciales para el análisis:

- DB_Ubicacion_Limpio.csv: Este archivo contiene información detallada sobre la ubicación de los clientes, incluyendo dirección, parroquia, cantón y provincia.
- nombres_series_finales.csv: Este archivo contiene los nombres o identificadores únicos de los clientes identificados como potenciales usuarios de VEs en las etapas de clusterización y análisis de aumento sostenido.
- Se agregan dos nuevas columnas, *"latitud"* y *"longitud"*
- Se inicia el proceso de geocodificación, que consiste en convertir las direcciones de los clientes en coordenadas geográficas.

Para cada cliente, se construye una dirección completa concatenando la información de ubicación disponible, incluyendo dirección, parroquia, cantón, provincia y el país *"Ecuador"*. Se utiliza la API de Google Maps para obtener las coordenadas geográficas de cada dirección. En caso de que se encontraran las coordenadas, se almacenaron en las columnas "latitud" y "longitud".

## VISUALIZACIÓN INTERACTIVA

Para facilitar la comprensión y el análisis de la distribución geográfica de los clientes identificados como potenciales usuarios de vehículos eléctricos, se desarrolla una visualización interactiva utilizando un mapa. Esta herramienta permite ubicar espacialmente a los clientes, proporcionando información adicional sobre cada uno de ellos y ofreciendo la posibilidad de seleccionar y visualizar clientes específicos.

Inicialmente, se eliminan del conjunto de datos aquellas filas que no contenían coordenadas geográficas válidas (latitud y longitud).

Posteriormente, se crea un mapa base utilizando la librería Folium, el cual se centró en la ubicación promedio de los clientes y se configuró con un nivel de zoom inicial que

permitiera una visualización adecuada de la distribución general. Se optó por el estilo de mapa *"OpenStreetMap"* para una mejor presentación visual.

Para evitar la superposición de marcadores en áreas con alta densidad de clientes, se emplea la técnica de agrupación de marcadores (MarkerCluster). Los marcadores de cada cliente se agregaron a un grupo de marcadores, lo que facilita la visualización y evita la saturación visual del mapa.

Se itera a través de cada cliente con coordenadas geográficas válidas y, para cada uno de ellos, se creó una ventana emergente (popup) que se despliega al hacer clic en el marcador. Esta ventana emergente muestra información relevante sobre el cliente, como su nombre (cuenta contrato), ubicación, provincia y cantón.

Para permitir la selección y visualización de clientes específicos, se implementa un menú interactivo en el mapa.

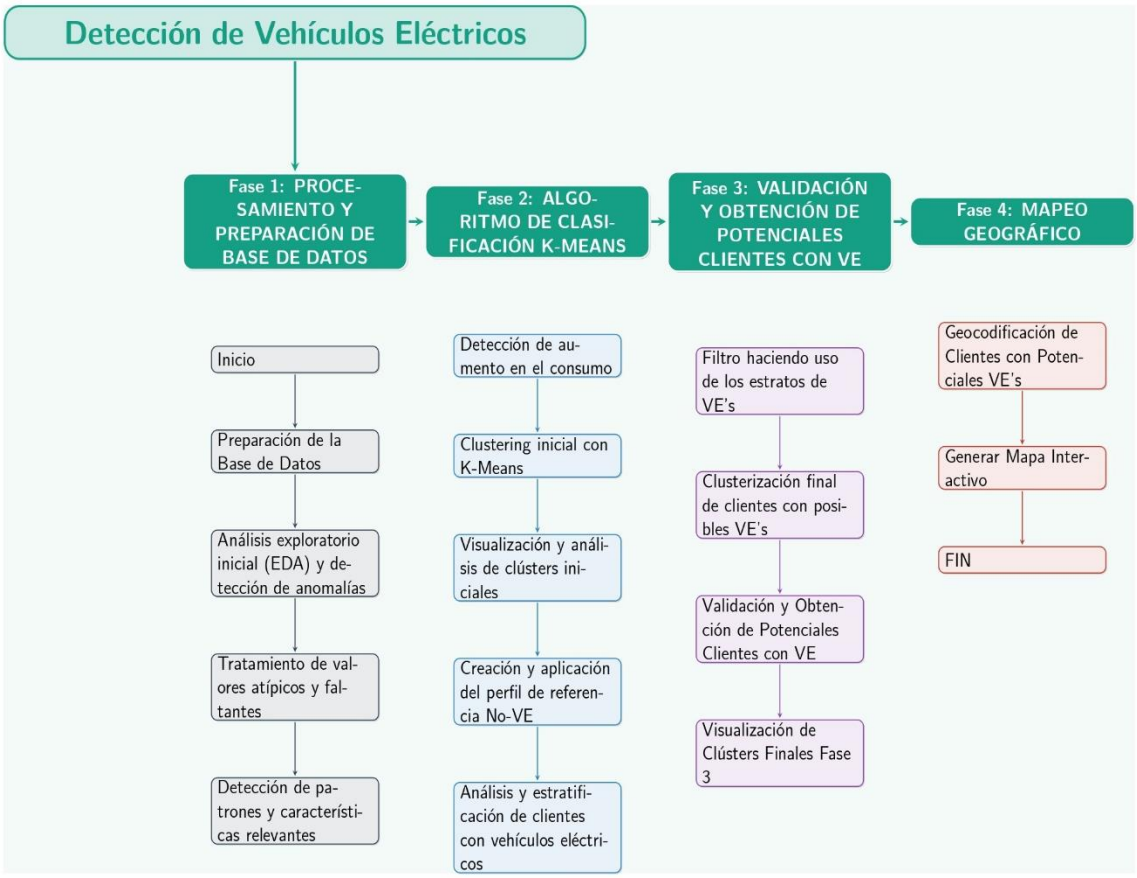
Se define una plantilla HTML utilizando el motor de plantillas Jinja2, la cual se utilizó para crear el menú desplegable en el mapa. Los datos preparados se incorporaron a la plantilla HTML y se agregó el menú desplegable al mapa.

Finalmente, el mapa interactivo generado se guardó en un archivo HTML para que pueda ser abierto y visualizado en un navegador web.

La Imagen 7 ilustra cada uno de los ítems descritos en este capítulo.

**Imagen 7**

Diagrama de la metodología.



# CAPÍTULO IV

# RESULTADOS Y DISCUSIÓN

## INTERPRETACIÓN DE RESULTADOS

El presente estudio ha logrado, a través de la aplicación sistemática de métodos, técnicas e instrumentos de minería de datos, generar resultados significativos en la identificación de patrones de consumo residencial asociados a la posible presencia de vehículos eléctricos en el área de concesión de la EEQ. A continuación, se resumen e interpretan los hallazgos principales, vinculándolos directamente con el cumplimiento de los objetivos planteados inicialmente.

## FASE 1: PROCESAMIENTO Y PREPARACIÓN DE LA BASE DE DATOS

La fase inicial de preparación de la base de datos, llevada a cabo en el entorno de programación R, se centró en optimizar el manejo del elevado volumen de información disponible. Este proceso, crucial para el éxito de la investigación, permitió reducir y estructurar el conjunto de datos original, eliminando variables irrelevantes para el análisis de series de tiempo de consumo eléctrico y organizando la información esencial en dos sub-bases diferenciadas, cada una con un propósito específico dentro del estudio.

La primera sub-base, destinada al análisis de series temporales, fue construida a partir de la selección de las siguientes variables clave, consideradas fundamentales para la identificación de patrones de consumo:

- **Cuenta:** Identificador único de cada cliente, que permite el seguimiento individualizado de su consumo a lo largo del tiempo.
- **Energía:** Consumo mensual de energía eléctrica expresado en kWh, que constituye la variable central del análisis de series de tiempo.
- **Mes:** Indicador temporal del período de consumo, que permite ordenar cronológicamente los datos y analizar su evolución a lo largo del tiempo.

La segunda sub-base, orientada a la georreferenciación de los clientes, incluyó las siguientes variables, que posibilitan la ubicación geográfica de los usuarios:

- **Dirección:** Dirección postal del cliente, que permite ubicarlo geográficamente.
- **Cantón:** Demarcación territorial cantonal a la que pertenece el cliente, que facilita el análisis de patrones de consumo a nivel cantonal.

- **Parroquia:** Demarcación territorial parroquial a la que pertenece el cliente, que permite un análisis más detallado a nivel parroquial.
- **Provincia:** Demarcación territorial provincial a la que pertenece el cliente, que posibilita el análisis de patrones de consumo a nivel provincial.

La lectura de datos en el entorno Pandas, seguida de la conversión a series temporales, allanó el camino para el análisis exploratorio.

Durante esta etapa, se identificaron valores de ceros y valores extremos en el rango de e^6, los cuales fueron objeto de un análisis individualizado para determinar su impacto en la integridad de los datos. Adicionalmente, se aplicaron criterios de filtrado basados en la identificación de series constantes y en el uso del criterio de tarifa dignidad. Esta estrategia de filtrado no solo optimizó los tiempos de procesamiento en las etapas subsiguientes, sino que también contribuyó a mejorar la precisión y la interpretabilidad de los resultados al reducir sustancialmente el volumen de datos, pasando de 1.102.770 series temporales originales a un conjunto depurado de 485.000 series.

Al eliminar series constantes y valores atípicos, se fortaleció la base para la identificación de patrones de consumo relevantes, permitiendo un enfoque más nítido en las señales que podrían indicar la presencia de usuarios de vehículos eléctricos (VE) sin medidor dedicado.

## FASE 2: ALGORITMO DE CLASIFICACIÓN K-MEANS

### Detección de aumento en el consumo

El objetivo principal de este fragmento código radicaba en la identificación y selección de aquellas series de tiempo que exhiben un patrón de consumo caracterizado por un aumento en el consumo de energía.

En este punto, se aplica un filtro a la lista original de series de tiempo, seleccionando exclusivamente aquellas que fueron marcadas como aumento de energía. De esta manera las series temporales se ven reducidas a 5100.

### Clustering inicial con K-Means.

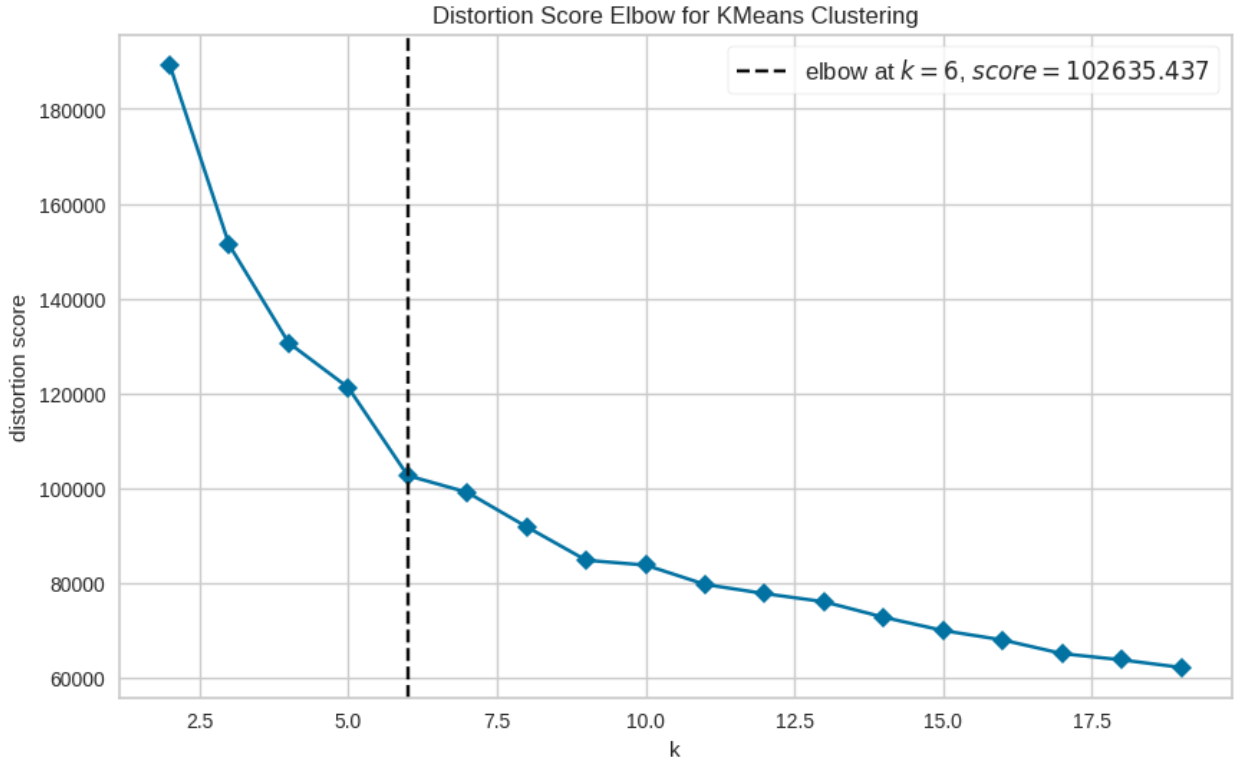
El Gráfico 7 ilustra la curva del codo resultante del análisis, donde se representa en el eje horizontal el número de clústeres (k), con valores que varían entre 2 y 20, y en el eje vertical el *"distortion score"* asociado a cada valor de k.

El análisis visual revela una disminución notable en el *"distortion score"* a medida que k se incrementa en las primeras iteraciones. No obstante, se observó un punto de inflexión a partir del cual la curva exhibe una tendencia de aplanamiento, lo que sugirió que el beneficio marginal derivado de aumentar el número de clústeres en términos de reducción de la inercia se reduce sustancialmente.

El codo de la curva, claramente distinguible en k = 6, señalo el punto en el que la disminución de la inercia se torna menos pronunciada. En este contexto específico, el análisis gráfico del método del codo sugiere que seis clústeres constituyen un número óptimo para segmentar el conjunto de series de tiempo bajo estudio. El *"distortion score"* correspondiente a este punto de inflexión se situó en 102635.437, lo que indica el nivel de inercia inherente a un clustering configurado con seis grupos.

**Gráfico 7**

Método del codo resultado clúster inicial.




**Visualización y análisis de clústeres iniciales**

En el Gráfico 8 visualiza la distribución de los clústeres resultantes en un espacio bidimensional generado a partir del Análisis de Componentes Principales. Cada punto en

este diagrama representa una serie de tiempo individual, y su coloración codificada revela el clúster específico al que fue asignada mediante el algoritmo K-Means.

La inspección visual del gráfico revelo una estructura de clustering discernible. Se observo una tendencia de los puntos a agruparse en regiones caracterizadas por una homogeneidad cromática relativa. Esta observación sugirió que el algoritmo K-Means ha logrado identificar grupos de series de tiempo que comparten características intrínsecas similares dentro del espacio bidimensional generado por PCA.

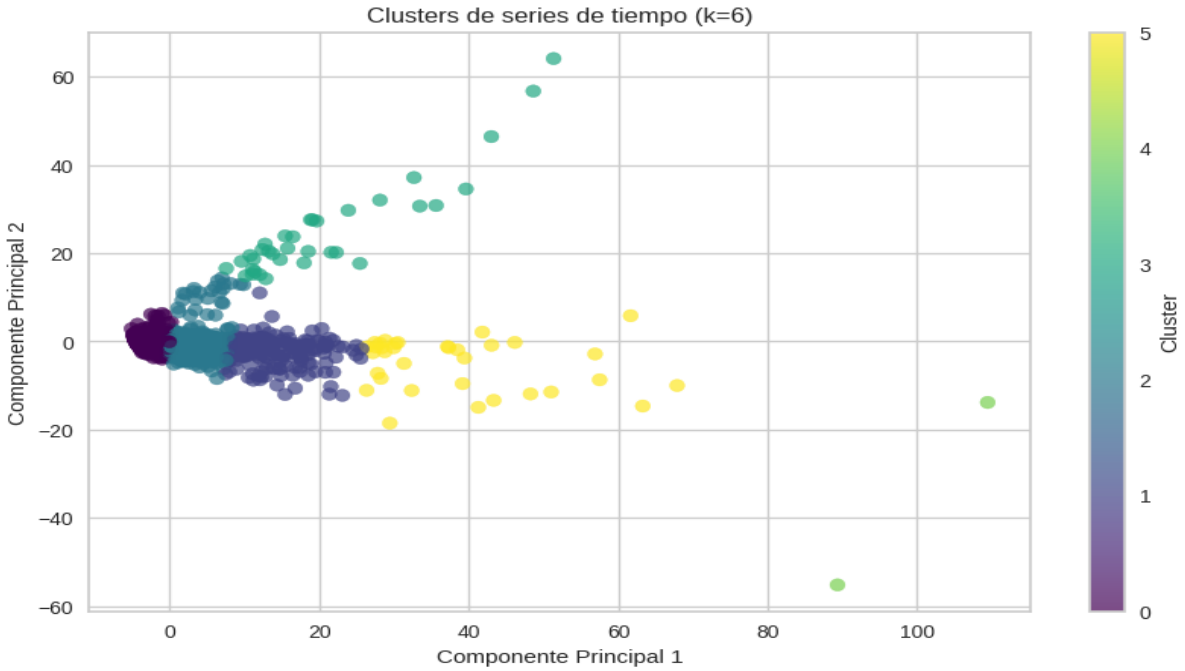
El Clúster 0, representado en morado oscuro, se distinguió por su alta densidad y ubicación central, donde concentra una proporción considerable de las series de tiempo analizadas. Esta concentración sugiere la presencia de un patrón de consumo predominante o basal en el conjunto de datos, el cual podría representar el comportamiento típico de un segmento mayoritario de usuarios.

En contraste, los clústeres 1 a 5, representados por una gama de colores que varían del azul al amarillo claro, se distribuyen de manera más dispersa alrededor del Clúster 0, extendiéndose hacia diversas direcciones dentro del espacio definido por las componentes principales. Esta dispersión podría reflejar variaciones significativas en los patrones de consumo, posiblemente asociadas a diferentes comportamientos de carga o a la influencia de factores externos.

Es importante señalar que se observó cierta superposición entre algunos clústeres, especialmente en las zonas de transición entre colores. Esta superposición, esperable en datos empíricos, sugiere que, si bien K-Means se logró identificar grupos principales, las fronteras que separan algunos patrones de consumo pueden no ser completamente definidas.

**Gráfico 8**

Reducción PCA clúster inicial.




El Gráfico 9 presenta la distribución de las series de tiempo en los seis clústeres identificados mediante el algoritmo K-Means. Cada subplot corresponde a un clúster específico, mostrando la evolución temporal de las series de tiempo que lo componen.

El análisis visual de los subplots revelo patrones de consumo diferenciados entre los clústeres.

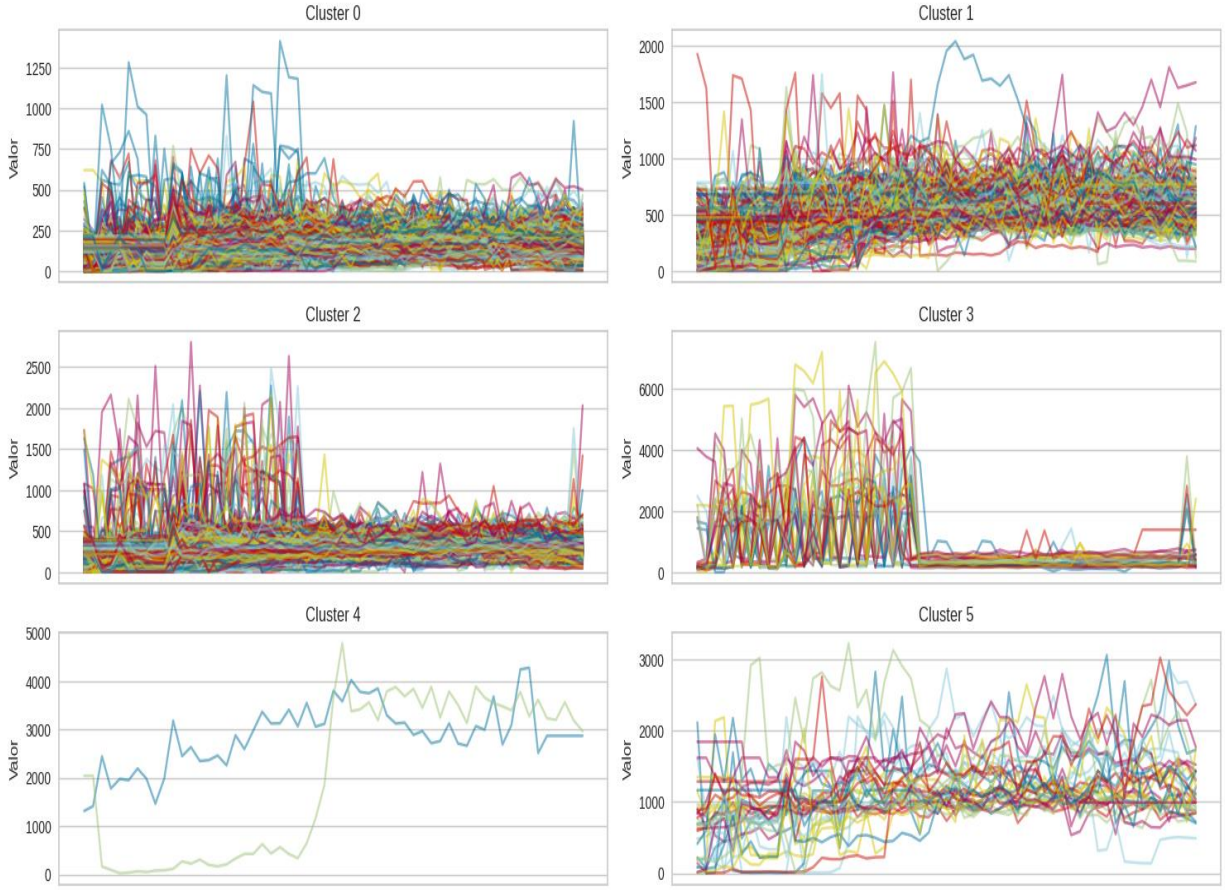
- **Clúster 0:** Exhibió una variabilidad considerable en los niveles de consumo, con series de tiempo que fluctúan significativamente a lo largo del tiempo.
- **Clúster 1:** Se caracterizo por un consumo relativamente bajo y estable, con algunas fluctuaciones puntuales.
- **Clúster 2:** Muestro un patrón de consumo más elevado que el clúster 1, con picos de consumo más pronunciados.
- **Clúster 3:** A diferencia de los demás clústeres, no exhibió un patrón de consumo claramente asociado a la carga de vehículos eléctricos. Las series de tiempo en este clúster muestran fluctuaciones irregulares y no presentan el incremento sostenido característico de la carga de VE.

- **Clúster 4:** Se distinguió por un incremento gradual en el consumo a lo largo del tiempo, lo que podría sugerir un aumento progresivo en la demanda energética.
- **Clúster 5:** Presento un patrón de consumo variable, con algunos picos de consumo que podrían estar asociados a eventos específicos o actividades particulares.

El análisis visual revelo que el Clúster 3 no se alineo con el perfil de cliente con VE buscado en este estudio. A diferencia de otros clústeres que muestran patrones de consumo compatibles con la carga de vehículos eléctricos, el clúster 3 se caracterizó por fluctuaciones irregulares y la ausencia de un incremento sostenido en el consumo. Esta observación sugirió que las series de tiempo agrupadas en este clúster podrían corresponder a patrones de consumo residencial típicos no relacionados con la carga de VE. El índice Davies-Bouldin fue de: 1.1095384962663644.

**Gráfico 9**

Series temporales obtenidas del clúster inicial.



**Creación y aplicación del perfil de referencia No-VE**

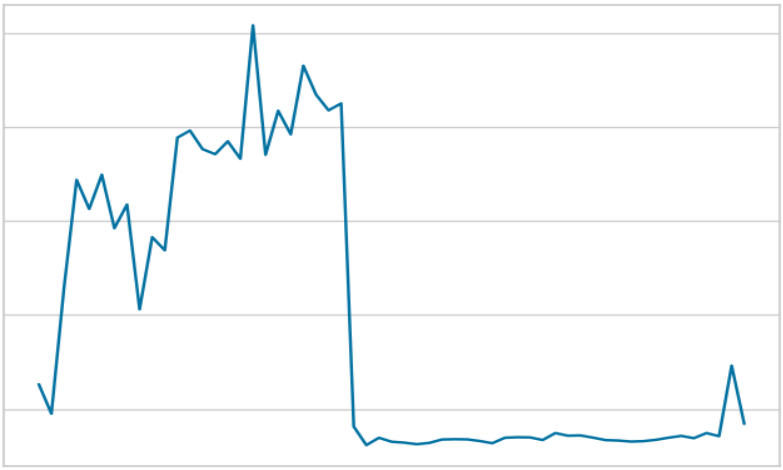
La Imagen 8 presenta un ejemplo representativo del perfil de consumo eléctrico considerado como "No-VE". Este perfil, derivado del análisis de clústeres y la selección de un clúster representativo de patrones de consumo típicos de clientes sin vehículos eléctricos.

Este perfil mostró una tendencia general al descenso en el consumo a lo largo del tiempo. Se inicia con valores superiores disminuye progresivamente.

Tras el pico inicial, el perfil mostró un descenso relativamente continuo en el consumo, con algunas fluctuaciones menores, pero manteniendo la dirección general descendente. Este descenso puede ser indicativo de una disminución gradual en la actividad residencial a lo largo del tiempo, o de factores estacionales o comportamentales que influyen en la demanda de energía.

**Imagen 8**

Perfil de referencia para clientes sin VE.



El Gráfico 10 presenta un histograma que visualiza la distribución de las distancias euclidianas calculadas. El eje horizontal representa la *"Distancia euclidiana"*, mientras que el eje vertical indica la *"Frecuencia"* de cada rango de distancia. Adicionalmente, una línea vertical discontinua de color rojo marca la posición del "Percentil 95" de esta distribución de distancias.

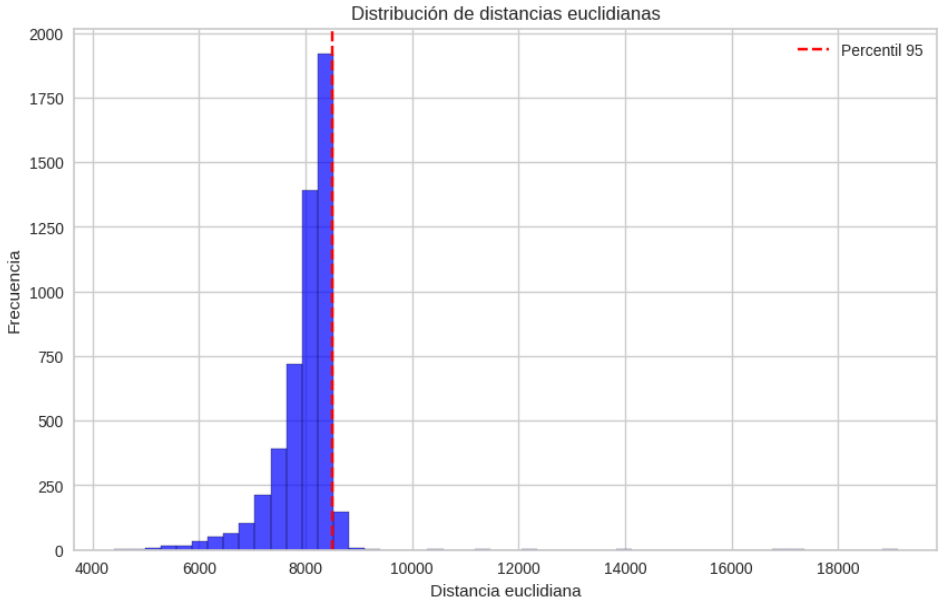
El histograma mostró una distribución claramente sesgada hacia la derecha. La mayoría de las barras se concentró en los rangos de distancia más bajos, desplegándose hacia la derecha con una cola más larga y menos frecuente. Esto indicó que la mayoría de las

series de tiempo presentan distancias euclidianas relativamente pequeñas, mientras que un número menor de series se aleja significativamente en términos de distancia.

La cola derecha del histograma, aunque con frecuencias decrecientes, se extiende hasta valores de distancia superiores a 18000 unidades. Esto indicó la presencia de series de tiempo con distancias euclidianas significativamente mayores, representando patrones de consumo que se desvían más del patrón de referencia.

**Gráfico 10**

Resultado de distancias euclidianas.




**Análisis y estratificación de clientes con vehículos eléctricos**

El Gráfico 11 presenta una visualización detallada de los cinco clústeres identificados en el análisis de series de tiempo para clientes con VE's.

**Clúster 0:** El consumo se elevó abruptamente a valores superiores a 1200 unidades, para luego descender con la misma rapidez en los meses subsiguientes.

**Clúster 1:** Este clúster exhibió un pico de consumo, aunque menos pronunciado y más extendido en el tiempo, abarcando desde enero hasta marzo de 2024 ("ene-24" a "mar-24").

**Clúster 2:** Este clúster se distinguió por un consumo general muy bajo, manteniéndose por debajo de 25 unidades en la mayoría de los meses. Se observa un ligero pico al inicio

del período ("dic-23"), pero de magnitud significativamente menor en comparación con los clústeres 0 y 1.

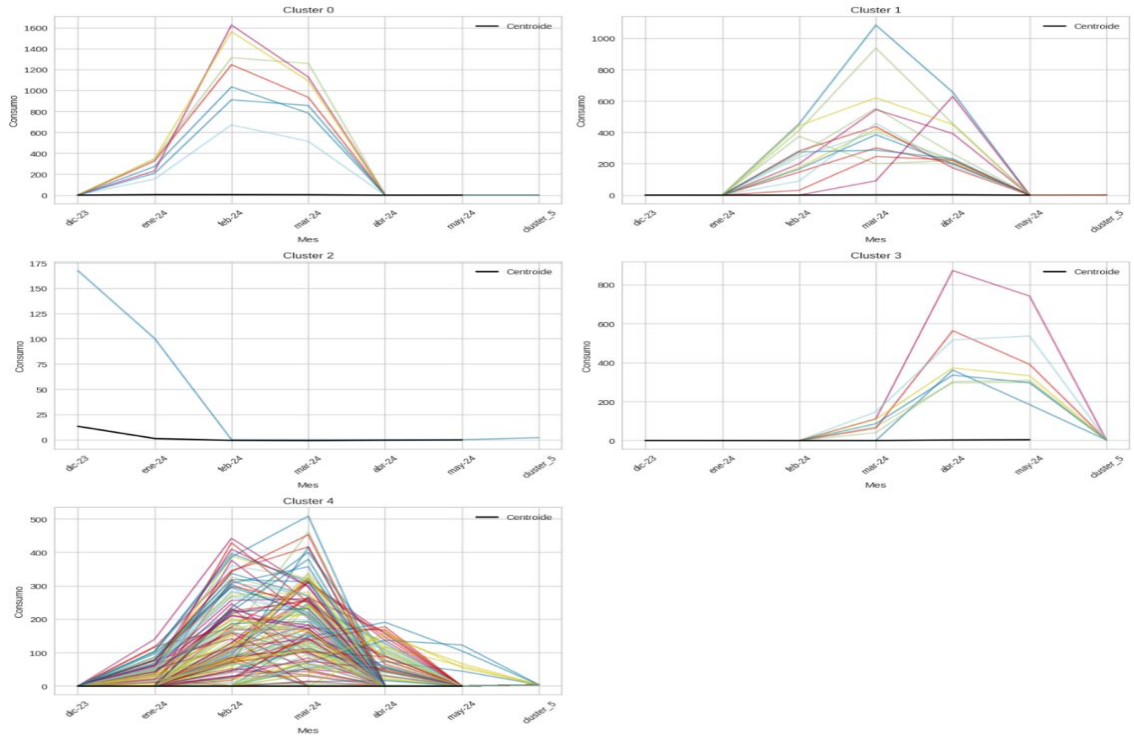
**Clúster 3:** Este clúster presento un patrón de consumo más complejo y variable, con múltiples picos y valles a lo largo del período.

**Clúster 4:** Este clúster se caracterizó por un pico de consumo más tardío, que ocurre principalmente en abril y mayo de 2024 ("abr-24" y "may-24"). El pico es pronunciado y extenso, alcanzando valores alrededor de 400 a 500 unidades. El consumo en los meses previos es relativamente bajo, para luego elevarse en los últimos meses del período.

Esta etapa ha servido como un filtro esencial para el posterior proceso de estratificación de clientes. En este contexto, se procedió a eliminar todas aquellas series de tiempo que se encontraban agrupadas en el clúster 2, dado que su comportamiento, caracterizado por un consumo extremadamente bajo o nulo, no aporta información relevante para el análisis principal del estudio.

**Gráfico 11**

Clúster de clientes VE's.




La Tabla 4 muestra un resumen de los estratos de consumo de vehículos eléctricos, incluyendo información sobre la cantidad de vehículos en cada estrato, el consumo máximo registrado y el consumo promedio. Se pueden observar las siguientes tendencias:

- **Distribución de vehículos:** El estrato 0 concentro la mayor cantidad de vehículos (77), mientras que los estratos 3, 4, 5 y 6 tienen una menor representación (entre 1 y 3 vehículos cada uno).
- **Consumo máximo:** El consumo máximo aumento progresivamente a medida que se avanza hacia estratos superiores. El estrato 7 registra el consumo máximo más alto (2999.87 kWh), mientras que el estrato 0 presenta el consumo máximo más bajo (374.08 kWh).
- **Consumo promedio:** De manera similar al consumo máximo, el consumo promedio también se incrementó a medida que se asciende en los estratos. El estrato 7 tiene el consumo promedio más alto (2966.10 kWh) y el estrato 0 el más bajo (159.62 kWh).

**Tabla 4**

Estratos de clientes con VE's.

| Estrato | Número de Vehículos | Consumo Máximo (kWh) | Consumo Promedio (kWh) |
|---|---|---|---|
| 0 | 77 | 374.0788 | 159.616227 |
| 1 | 70 | 749.8037 | 533.376422 |
| 2 | 16 | 1076.7083 | 863.335866 |
| 3 | 3 | 1339.0696 | 1225.419567 |
| 4 | 3 | 1803.9180 | 1681.023833 |
| 5 | 3 | 2194.2904 | 2086.790667 |
| 6 | 1 | 2505.6835 | 2505.683500 |
| 7 | 3 | 2999.8690 | 2966.098600 |

## FASE 3: VALIDACIÓN Y OBTENCIÓN DE POTENCIALES CLIENTES

El Gráfico 12 presenta las series de tiempo correspondientes a los clientes identificados como "potenciales VE" después de aplicar el filtro basado en estratos de consumo de clientes VE's. Este filtro, como se mencionó anteriormente, buscaba refinar la detección

de aumentos sostenidos, considerando las particularidades de diferentes segmentos de clientes VE ya identificados.

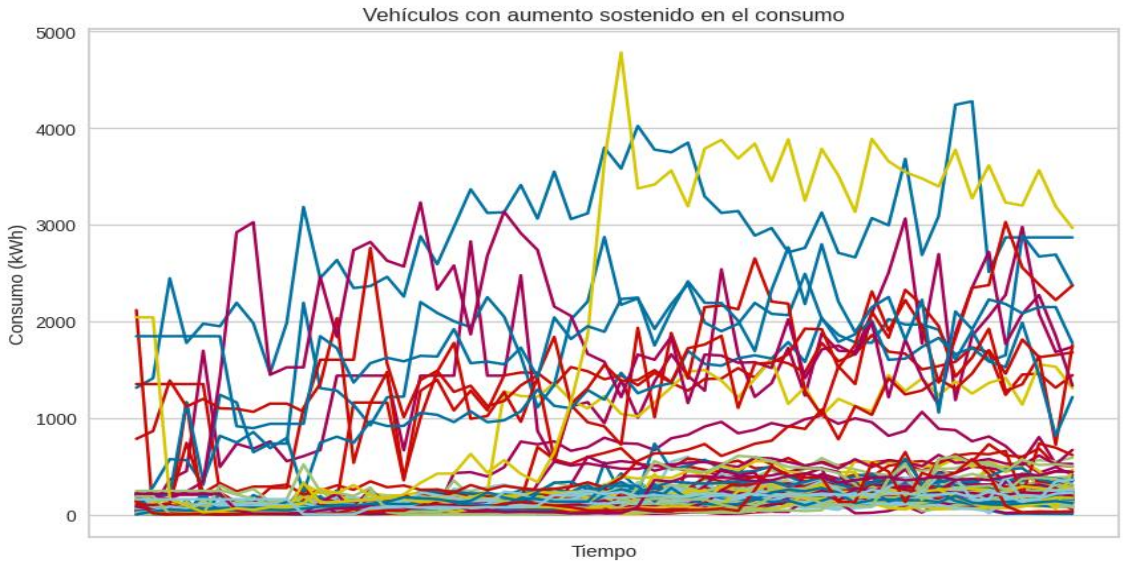
A diferencia del perfil "No-VE" previamente analizado, la característica dominante en este gráfico es la presencia de una tendencia ascendente en la mayoría de las series de tiempo. Muchas líneas mostraron un incremento notable en el consumo a lo largo del período representado. Este aumento sostenido reforzó la hipótesis de que estas series de tiempo podrían corresponder a clientes que han incorporado una nueva carga significativa y persistente, como la carga de un vehículo eléctrico.

Además, se tuvo una tendencia ascendente general, algunas series de tiempo muestran picos y fluctuaciones adicionales, superpuestos a la tendencia principal. Estos picos podrían representar eventos de consumo puntual (ej. uso intensivo de electrodomésticos, eventos climáticos) o variaciones en los patrones de carga (ej. carga más intensiva en ciertos períodos del mes).

Concentración en rangos de consumo elevados: En general, las series de tiempo en este gráfico se concentraron en rangos de consumo más elevados en comparación con el perfil "No-VE". Muchas líneas alcanzan valores superiores a 2000 kWh, y algunas incluso superaron los 4000 kWh. Estos niveles de consumo más altos fueron consistentes con la hipótesis de la adición de la carga de un VE al consumo residencial base.

**Gráfico 12**

Posibles clientes con VE's.

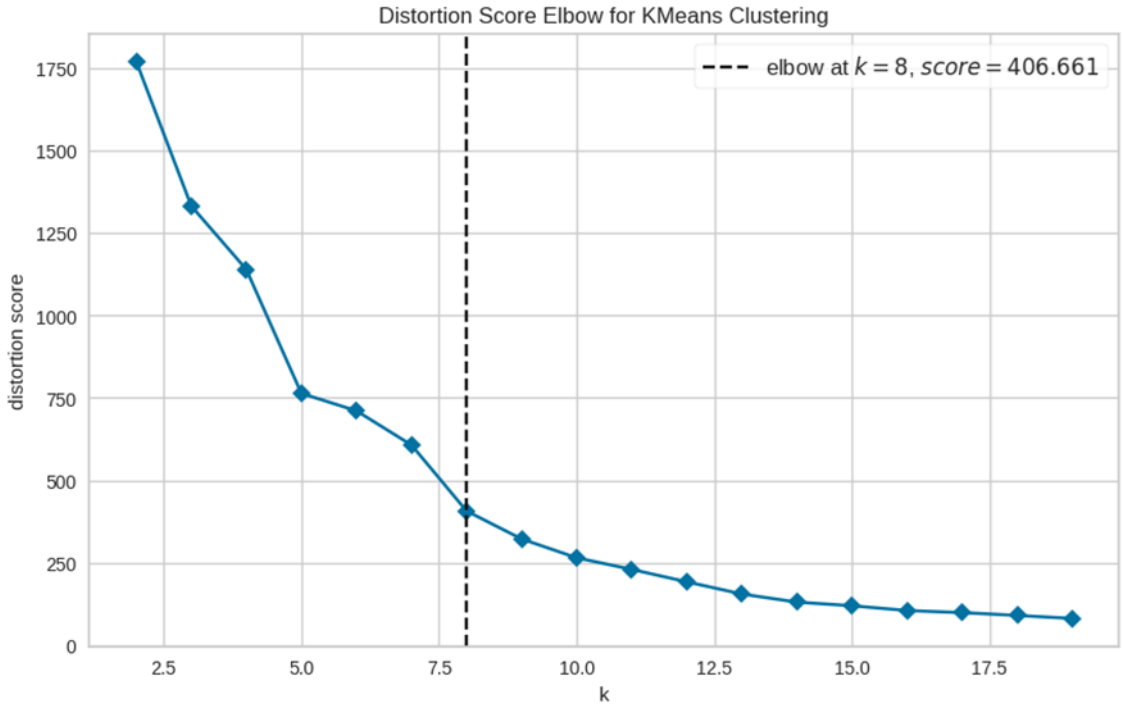


**Clusterización final de clientes con posibles VE's**

El Gráfico 13 presenta un nuevo resultado del método del codo, generado después de aplicar filtros adicionales y refinar el conjunto de datos analizado. Se observo una tendencia general descendente en el *"distortion score"* a medida que k aumenta. La inercia del modelo K-Means disminuyo al aumentar el número de clústeres. En comparación con el Gráfico 7 , el codo en esta curva fue más pronunciado y visualmente más claro, donde se sugirió que el número óptimo de clústeres para este conjunto de datos refinado fue de k = 8. Este valor difiere del número óptimo obtenido en el análisis anterior (k = 6).

**Gráfico 13**

Método del codo para clúster final.




El Gráfico 14 presenta la visualización de los clústeres obtenidos tras aplicar el algoritmo K-Means con el número óptimo de clústeres reevaluado como k=8.

En comparación con la visualización anterior (Gráfico 8), el Gráfico 14 mostro una estructura de clustering más detallada y compleja. Con 8 clústeres, se observó una mayor diferenciación de grupos en el espacio de PCA.

Si bien persistió cierta superposición, algunos clústeres parecen estar más definidos y mejor separados que en la visualización previa. Por ejemplo:

**Clúster 7 (color amarillo brillante):** Se destaco como un clúster muy aislado y separado del resto, ubicado en la parte superior del gráfico. Este aislamiento podría indicar un patrón de consumo muy distintivo y específico.

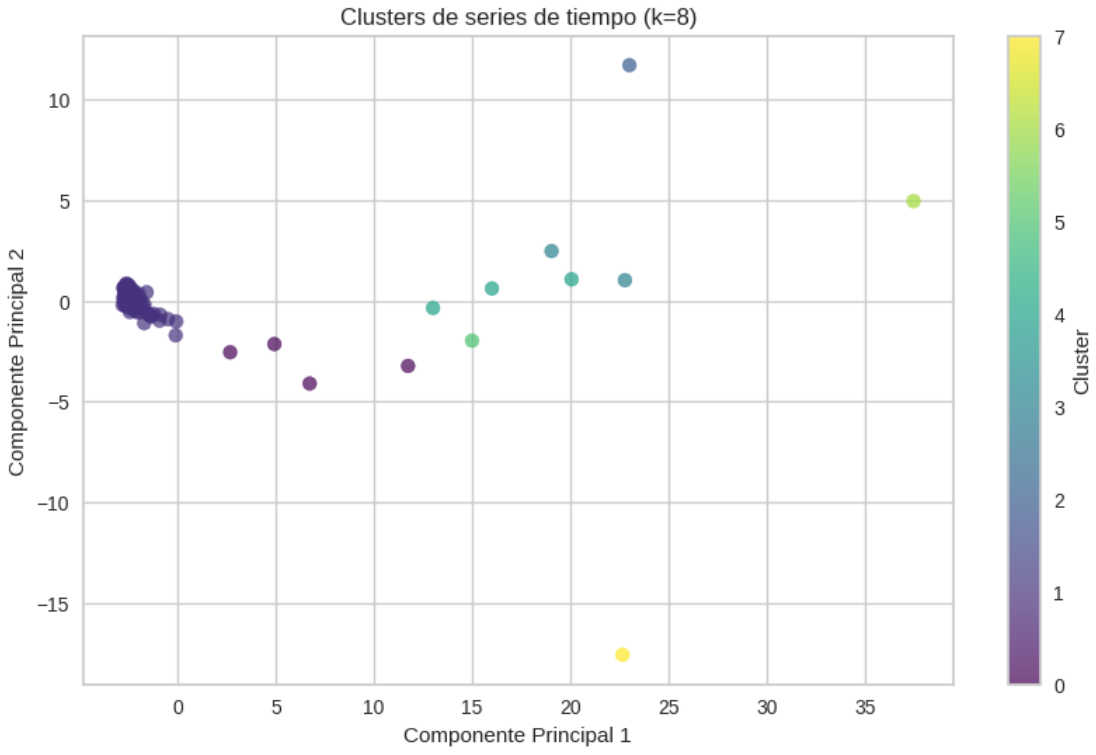
**Clúster 6 (color verde claro):** También mostro una cierta separación, ubicándose en la parte derecha del gráfico, aunque menos aislado que el clúster 7.

**Clústeres 0, 1, 2, 3 (colores morado oscuro a verde oscuro):** Estos clústeres permanecieron más agrupados y superpuestos en la región central del gráfico, pero incluso dentro de esta región se pudo apreciar subdivisiones y cierta diferenciación de colores, sugiriendo la identificación de sub-patrones dentro del grupo central.

**Clúster 4 y 5 (colores amarillo verdoso y amarillo):** Se ubicaron en una posición intermedia, mostrando cierta conexión con el grupo central, pero también cierta tendencia a separarse hacia la derecha.

**Gráfico 14**

Reducción PCA clúster final.



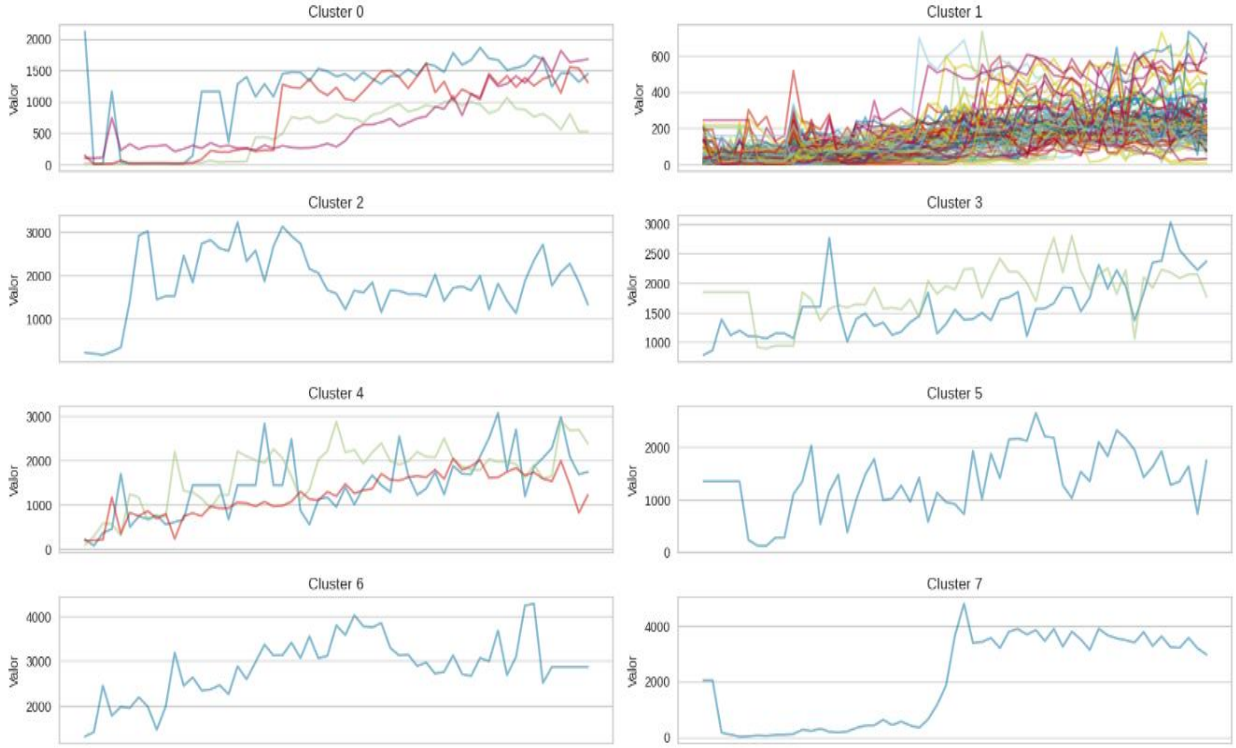

El Gráfico 15 presenta una visualización detallada de las series de tiempo pertenecientes a cada uno de los 8 clústeres obtenidos mediante el algoritmo K-Means con k=8. Se

desplegaron ocho subplots, cada uno dedicado a un clúster específico (Clúster 0 a Clúster 7).

Los diferentes clústeres mostraron un aumento gradual del consumo con el tiempo. Comenzando en niveles más bajos, las series aumentan constantemente a lo largo del período, alcanzando valores más altos hacia el final. La tendencia dominante fu ascendente, lo que indico un consumo creciente con el tiempo. Se obtuvo 109 series y el índice final de Davies-Bouldin fue de 0.6767006573684351.

**Gráfico 15**

Resultados clúster final.




## FASE 4: MAPEO GEOGRÁFICO

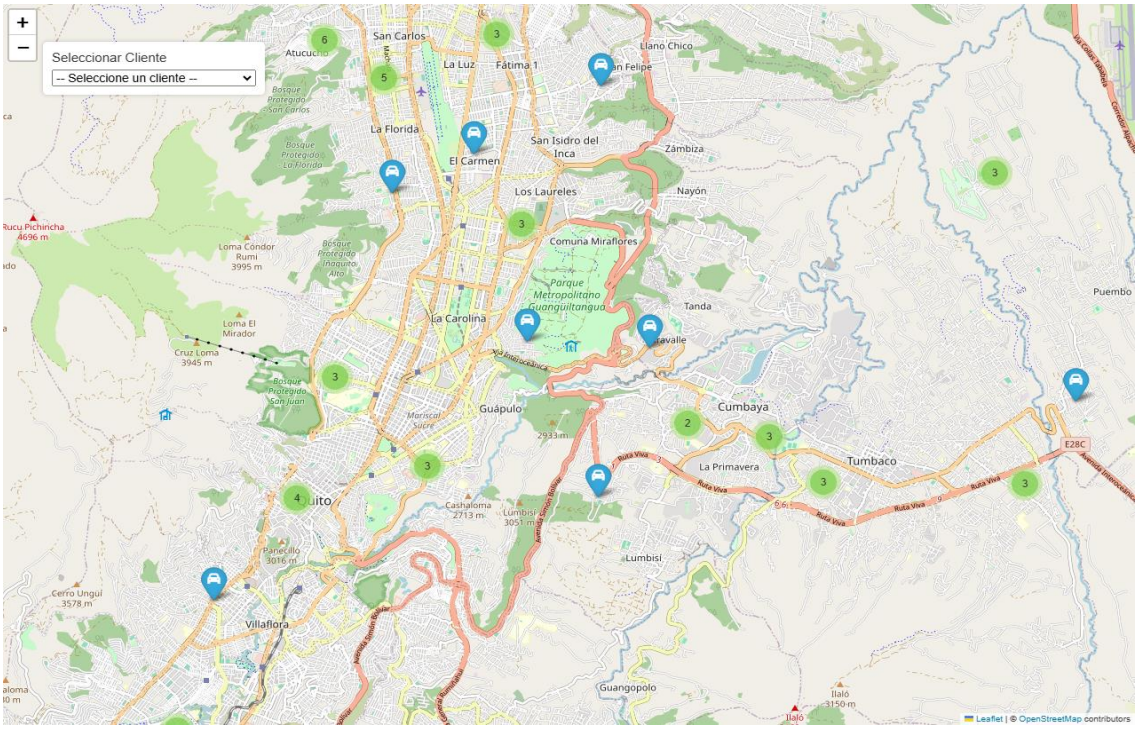
El análisis de georreferenciación ha permitido determinar las coordenadas geográficas (latitud y longitud) de 109 clientes potenciales de vehículos eléctricos, lo que ha posibilitado su ubicación precisa en un mapa. La distribución espacial de estos clientes revelo una concentración notable en la zona norte de Quito, particularmente en sectores como La Carolina, El Condado y Carcelén, los cuales se caracterizan por un nivel socioeconómico medio-alto y una mayor receptividad hacia nuevas tecnologías.

Es importante destacar que, si bien la zona norte de Quito presenta una concentración significativa de clientes potenciales de VE, otros sectores de la ciudad, como el Valle de los Chillos, Cumbayá y Tumbaco, también albergan una cantidad considerable de usuarios de VE, tal como se aprecia en la Imagen 9 . Estos sectores, en su mayoría ubicados en valles aledaños a la ciudad, se caracterizan por una densidad poblacional menor, mayor poder adquisitivo y un entorno más residencial, lo que podría estar asociado a una mayor preferencia por vehículos eléctricos como una alternativa de movilidad más sostenible y amigable con el medio ambiente.

La visualización de la distribución espacial de los clientes potenciales de VE en el mapa no solo permitió identificar las zonas de mayor concentración, sino que también ofreció información valiosa sobre las características socioeconómicas y geográficas de los usuarios de VE en la ciudad de Quito. Este análisis puede ser utilizado para orientar estrategias de marketing y promoción de VE, así como para planificar la infraestructura de carga necesaria para satisfacer la creciente demanda de este tipo de vehículos.

**Imagen 9**

Ubicación geográfica VE’s.



## CONTRASTE CON OTRAS INVESTIGACIONES

Al contrastar los resultados de esta investigación con los antecedentes mencionados, se observan las siguientes convergencias y divergencias:

**Similitud metodológica general:** Existe una convergencia metodológica general con la mayoría de los estudios revisados en el uso de técnicas de clustering (especialmente K-Means) y el análisis de series de tiempo como herramientas centrales para la segmentación y caracterización de patrones de consumo eléctrico, tanto en el contexto general residencial como específicamente en la detección de VE. Estudios como (Miyazaki et al., 2020), (Castro et al., 2023), (Ramírez, 2022)utilizan K-Means o variantes de clustering para lograr objetivos similares de segmentación y detección de patrones.

**Énfasis en la detección de patrones específicos:** Al igual que en (Miyazaki et al., 2020) que se enfoca en patrones de *"carga rápida frecuente"* y *"carga nocturna"* de VE, y (Jahangir et al., 2020)que identifica un perfil de *"consumo nocturno alto"* asociado a VE's, esta tesis también se centra en la detección de patrones específicos de consumo que se correlacionan con la hipótesis de la carga de VE, principalmente aquellos clústeres que exhiben tendencias ascendentes y consumo consistentemente. Este enfoque en patrones específicos permite una detección más dirigida y con mayor interpretabilidad que la búsqueda genérica de anomalías.

**Complementariedad con técnicas de aprendizaje profundo:** El estudio de (Koohfar et al., 2023) combina clustering con redes neuronales profundas (LSTM), lo cual sugiere una posible vía de extensión para esta tesis. Si bien el presente trabajo se centra en clustering clásico, la incorporación de técnicas de aprendizaje profundo para la extracción de características más complejas de las series de tiempo o para la clasificación posterior de los clústeres podría mejorar aún más la precisión de la detección de clientes VE potenciales.

**Aplicación a diferentes contextos y datos:** Mientras que (*Discovering Electric Vehicle Charging Locations Based on Clustering Techniques Applied to Vehicular Mobility Datasets*, s. f.) aplica clustering a datos de movilidad urbana para la planificación de infraestructura de carga, y (Zhao et al., 2022) se enfoca en redes vehiculares para optimizar el consumo energético en comunicaciones, esta tesis se centra específicamente

en el contexto residencial y en datos de consumo eléctrico de la EEQ. Esta especificidad contextual permite adaptar la metodología y los hallazgos a las particularidades del área de estudio y a las necesidades de la empresa eléctrica. La metodología propuesta podría ser transferible a otros contextos residenciales y empresas eléctricas, pero requeriría ajustes y validaciones específicas para cada caso.

**Innovación en la estratificación y visualización geográfica:** Como se mencionó anteriormente, la estratificación de clientes VE confirmados y la visualización geográfica interactiva representan contribuciones metodológicas originales de esta tesis. Estos elementos no se encuentran explícitamente en los antecedentes revisados y añaden valor al enfoque propuesto, permitiendo una detección más precisa y una mejor comprensión espacial de los resultados.

**Consideración de la gestión de la demanda:** El estudio de (Lascano et al., 2023) enfatiza la aplicación del clustering para programas de respuesta a la demanda. Si bien la presente tesis se centra en la detección de clientes potenciales de VE, los resultados obtenidos podrían ser utilizados en el futuro para estrategias de gestión de la demanda específicas para usuarios de VE, como la implementación de tarifas diferenciadas o incentivos para la carga en horarios valle. La segmentación lograda proporciona una base para la focalización de este tipo de programas.

**Detección de anomalías como perspectiva complementaria:** Aunque el enfoque principal de esta tesis es la detección de patrones específicos de VE, el trabajo de (Aparicio Wallis et al., 2019) sobre detección de anomalías mediante autoencoders variacionales sugiere una perspectiva complementaria. La carga de VE, en ciertos contextos, podría ser considerada como una *"anomalía"* en el patrón de consumo residencial típico. La combinación de técnicas de clustering con métodos de detección de anomalías podría ser una línea de investigación futura para refinar aún más la detección de usuarios de VE y distinguir entre diferentes tipos de patrones inusuales de consumo.

## VERIFICACIÓN DE LA HIPÓTESIS

La prueba t de Student realizada para comparar la variable *"consumo final"* la cual NO es una variable preexistente, es simplemente el último valor de cada serie de tiempo. Es el valor de consumo que se registra al final del período que estás analizando para cada cliente o punto de datos.

La comparación de los grupos de *"Potenciales VE's" y "NO VE's",* con la variable consumo final, no reveló diferencias estadísticamente significativas ($p = 0.5591 > 0.05$). Esto implica que, al utilizar la media de ambos casos y un nivel de significancia del 5%, no se encontró evidencia suficiente para afirmar que existe una diferencia en el consumo final promedio entre ambos grupos. Por lo tanto, la prueba t, en su forma actual, no respalda estadísticamente la hipótesis planteada.

La observación de los perfiles de los clústeres revelo exhiben patrones visuales que guardan similitud con la curva de carga característica de un vehículo eléctrico (VE). Estos clústeres se caracterizan por:

- Un período de base relativamente estable al inicio del período analizado.
- Un incremento sostenido y notable en el consumo energético hacia el final del período.
- Formas de curva que, en algunos casos podrían interpretarse como similares a una curva de carga de VE, con un aumento gradual o repentino y un mantenimiento de un nivel de consumo más alto.

Esta similitud visual entre los perfiles de los clústeres y la curva de carga de VE proporciona un cierto grado de soporte cualitativo a la hipótesis planteada. A pesar de que la prueba t no arrojó diferencias estadísticamente significativas en el consumo final, la inspección visual sugiere que existen patrones de consumo en estos clústeres que son consistentes con lo que se esperaría observar en usuarios de VE.

Para reforzar la verificación de la hipótesis se hizo uso de las distancias DTW (Dynamic Time Warping), las cuales son una herramienta poderosa para cuantificar la similitud entre las formas de las series de tiempo. A diferencia de otros métodos que comparan solo valores puntuales, DTW considera la curva completa de consumo a lo largo del tiempo. Esto permite capturar mejor las similitudes y diferencias en los patrones de consumo, incluso si existen desplazamientos temporales o variaciones en la escala.

En esencia, DTW mide la distancia entre dos series de tiempo al encontrar la deformación óptima de una serie con respecto a la otra. Esta deformación permite alinear los puntos correspondientes de ambas series, incluso si no coinciden temporalmente. La distancia

DTW resultante refleja cuánto se deben deformar las series para que coincidan, lo que proporciona una medida de su similitud.

**Distancia DTW promedio DENTRO del grupo Potenciales VE's: 18873.07;** Este valor representa la distancia DTW promedio entre todas las parejas posibles de series de tiempo dentro del grupo que hemos etiquetado como *"Potenciales VE's"*. Un valor relativamente menor aquí sugirió que las series de tiempo dentro del grupo *"Potenciales VE's"* tienden a tener formas más similares entre sí. Hay cierta consistencia en los patrones de consumo dentro de este grupo.

**Distancia DTW promedio DENTRO del grupo NO VE's: 32330.57;** Este valor es la distancia DTW promedio entre todas las parejas de series de tiempo dentro del grupo *"NO VE's"*. Este valor es mayor que la distancia dentro del grupo *"Potenciales VE's"*. Esto indico que las series de tiempo dentro del grupo *"NO VE's"* son, en promedio, menos similares entre sí en forma que las series dentro del grupo *"Potenciales VE's"*. Hay más variabilidad en los patrones de consumo dentro este grupo.

**Distancia DTW promedio ENTRE grupos (VE's vs. NO VE's): 48834.98;** Este es el valor más importante. Representa la distancia DTW promedio entre todas las combinaciones posibles de series de tiempo, tomando una serie del grupo *"Potenciales VE's"* y otra del grupo *"NO VE's"*. Este valor es significativamente mayor que las distancias DTW promedio dentro de ambos grupos. Esto es un indicador clave.

Los resultados de DTW sugirieron que sí existe una diferencia en la forma de las series de tiempo entre el grupo que hemos identificado como *"Potenciales VE's"* y el grupo *"NO VE's"*.

Las series de tiempo dentro del grupo *"Potenciales VE's"* tienden a ser más similares entre sí en forma (menor distancia DTW dentro del grupo).

Las series de tiempo del grupo *"Potenciales VE's",* en promedio, tienen una forma más diferente a las series de tiempo del grupo *"NO VE's"* (mayor distancia DTW entre grupos).

Estos resultados de DTW dan un soporte numérico a la idea de que el grupo *"Potenciales VE's"* tiene patrones de consumo distintos en comparación con el grupo *"NO VE's"*. Aunque la prueba t de Student con el *"consumo final"* no encontró diferencias

significativas, DTW, que compara la forma completa de las series de tiempo, sí está detectando una diferencia.

## DISCUSIÓN

## IMPLICACIONES

### Potencial de la metodología

A pesar de las limitaciones, este trabajo demostró el potencial de las técnicas de minería de datos, especialmente DTW, para el análisis de patrones de consumo eléctrico y la identificación de comportamientos específicos, como la carga de VE's, incluso con datos de granularidad mensual.

### Planificación energética e infraestructura de carga

La metodología desarrollada, con mejoras y validaciones futuras, podría convertirse en una herramienta útil para las empresas eléctricas y los planificadores urbanos en Quito.

La identificación temprana de zonas con alta probabilidad de carga de VE's puede facilitar la planificación de la expansión de la infraestructura de carga pública y privada, la optimización de la red eléctrica y la implementación de tarifas eléctricas diferenciadas para promover la carga inteligente de VE's.

Por ejemplo, las empresas eléctricas podrían usar esta información para identificar áreas con alta demanda de energía debido a la carga de VE's y planificar la expansión de la red para evitar sobrecargas. Los planificadores urbanos podrían utilizar estos datos para identificar ubicaciones estratégicas para la instalación de estaciones de carga pública.

## LIMITACIONES E IMPLICACIONES DEL TRABAJO REALIZADO

### Granularidad de los datos mensuales

La principal limitación de este estudio FUE la granularidad mensual de los datos de consumo eléctrico. El uso de datos a nivel mensual diluye la información temporal más detallada que podría ser crucial para detectar patrones de carga de VE's, los cuales ocurren generalmente en escalas de tiempo más cortas (horaria o incluso sub-horaria). Esta limitación pudo haber reducido la sensibilidad de la prueba t de Student al analizar únicamente el *"consumo final"*, ya que esta métrica simplificada no captura la forma completa de la curva de consumo.

**Limitaciones en la calidad de los datos**

Una limitación importante fue la calidad de los datos de consumo eléctrico proporcionados por la empresa distribuidora. Durante el análisis, se identificaron valores atípicos y anomalías que no se ajustan a los patrones de consumo residencial típicos, lo que sugirió deficiencias en el control de calidad de los datos. Esta situación introducción incertidumbre y pudo afectar la fiabilidad de los resultados obtenidos mediante las técnicas de minería de datos. Por lo tanto, las conclusiones de este trabajo deben interpretarse con cautela, teniendo en cuenta esta limitación en la calidad de los datos.

**Dependencia de la calidad del clustering**

La identificación *de "potenciales usuarios de VE's"* dependió de la calidad del clustering realizado con K-Means. Aunque se exploraron diferentes configuraciones y se analizaron visualmente los perfiles de los clústeres, la asignación de usuarios a clústeres y la interpretación de los clústeres como "potenciales usuarios de VE's" introdujo un grado de subjetividad y depende de la robustez del algoritmo de clustering.

**Validación de la hipótesis**

Aunque el análisis de distancias DTW aporto evidencia favorable a la hipótesis, la validación definitiva de la identificación de usuarios de VE's requeriría información adicional, como encuestas directas a los usuarios, datos de ventas de VE's en la zona, o la instalación de medidores de carga específicos. En ausencia de esta validación externa, los resultados deben considerarse como indicativos y exploratorios.

**RESUMEN COMPARATIVO CON TRABAJOS SIMILARES**

Este trabajo se distinguió por su enfoque en la identificación de usuarios de vehículos eléctricos (VE's) a partir de datos de consumo mensual mediante técnicas de clustering. Aunque existen investigaciones previas sobre el análisis de consumo eléctrico con minería de datos y aprendizaje automático, la mayoría se centra en objetivos distintos, como la detección de anomalías en la red, la predicción de la demanda o la identificación de electrodomésticos (NILM). Si bien se emplearon técnicas similares, no abordan directamente la identificación de carga de VE's con datos mensuales.

Existen trabajos previos que utilizan clustering (como K-Means) para agrupar usuarios según sus perfiles de consumo, con el fin de gestionar la demanda o la tarificación. Sin

embargo, la aplicación específica de clustering para identificar usuarios con patrones compatibles con la carga de VE's es una contribución original de esta tesis.

## FUTUROS TRABAJOS

Este trabajo abre varias líneas para futuras investigaciones:

Este estudio abre un abanico de posibilidades para investigaciones futuras que podrían enriquecer y profundizar los hallazgos presentados. En primer lugar, es fundamental utilizar datos de consumo con mayor granularidad temporal (horaria o subhoraria) para afinar la detección de patrones de carga de vehículos eléctricos, capturando mejor las dinámicas de consumo a corto plazo.

En segundo lugar, se sugiere explorar técnicas de clustering y clasificaciones más avanzadas para series de tiempo, y la inclusión de métodos basados en deep learning podría revelar patrones complejos y no lineales en los datos, mejorando la precisión en la identificación de usuarios de vehículos eléctricos. Además, es relevante incorporar variables complementarias, como datos socioeconómicos, tipología de viviendas y características del parque vehicular, para robustecer los modelos y reducir la incertidumbre en la identificación de usuarios de vehículos eléctricos.

Un aspecto crucial para validar la metodología propuesta es la contrastación de los resultados con datos externos.

Realizar encuestas a usuarios o acceder a datos de ventas de vehículos eléctricos permitiría cuantificar la precisión de la metodología y establecer su nivel de confianza. Finalmente, se plantea como objetivo a largo plazo el desarrollo de una herramienta o plataforma visual que facilite a las empresas eléctricas y planificadores urbanos la aplicación práctica de la metodología en la toma de decisiones, promoviendo una gestión más eficiente de la infraestructura de carga y la transición hacia la movilidad eléctrica.

# CAPÍTULO V
# CONCLUSIONES Y RECOMENDACIONES

## CONCLUSIONES

Tras el desarrollo de la investigación y el análisis de los resultados, se presentan las siguientes conclusiones, las cuales responden directamente a los objetivos específicos planteados:

1. Se fundamentó un marco metodológico basado en el estándar CRISP-DM, adaptando con éxito técnicas de minería de datos para abordar el desafío de identificar vehículos eléctricos (VE) a partir de datos de consumo mensual de baja granularidad. La revisión del estado del arte confirmó la idoneidad del clustering, pero destacó la necesidad de un enfoque ajustado a las condiciones locales, carentes de tarifas horarias.
2. Se aplicó un modelo de clustering K-Means y análisis de series temporales que permitió segmentar la población de estudio y aislar un subconjunto de 109 clientes cuyos patrones de consumo son compatibles con la carga de un vehículo eléctrico. El método demostró ser eficaz para identificar incrementos sostenidos en el consumo, característicos de la adición de esta nueva carga significativa.
3. Se validó la agrupación final de clústeres, obteniendo una mejora en la calidad de la segmentación (índice de Davies-Bouldin de 0.67), y se diseñó una herramienta de visualización georreferenciada. El mapa interactivo resultante constituye una propuesta de valor tangible, al revelar una concentración geográfica de usuarios potenciales en el norte de Quito y los valles de Cumbayá y Tumbaco, facilitando así la toma de decisiones estratégicas para la EEQ.

## RECOMENDACIONES

A partir de los hallazgos y limitaciones del estudio, se proponen las siguientes acciones:

**Para la Empresa Eléctrica Quito (EEQ):**

1. Utilizar la metodología y el mapa geográfico desarrollados como una herramienta de apoyo para la planificación estratégica de la red,

priorizando el refuerzo y la expansión de la infraestructura en las zonas identificadas con alta concentración de potenciales usuarios de VE.

2. Implementar un plan de mejora en la calidad y granularidad de la recolección de datos de consumo. El acceso a datos horarios o sub-horarios aumentaría exponencialmente la precisión de futuros análisis y permitiría el desarrollo de sistemas de gestión de la demanda en tiempo real.

**Para futuras investigaciones**

1. Validar los hallazgos de este estudio mediante trabajo de campo, como encuestas directas a los clientes identificados o la instalación de medidores dedicados en una muestra de hogares, para cuantificar la precisión del modelo.
2. Explorar el uso de algoritmos de aprendizaje automático más avanzados (ej. redes neuronales LSTM, DBSCAN) que puedan capturar patrones temporales más complejos y potencialmente mejorar la exactitud en la identificación de usuarios de VE.

**BIBLIOGRAFÍA**


Aparicio Wallis, A. F., Serna Polania, J. C., & Camargo Ramirez, J. J. (2019). *Detección de anomalías en el consumo de energía eléctrica industrial*. https://doi.org/10.11144/Javeriana.10554.46925

Castro, J., Soto, P., Reategui, R., & Castillo, T. (2023). Partición de una Red Eléctrica de Distribución Aplicando Algoritmos de Agrupamiento K-means y DBSCAN. *Revista Técnica «energía»*, *20*(1), Article 1. https://doi.org/10.37116/revistaenergia.v20.n1.2023.572

*Clustering Analysis for Active and Reactive Energy Consumption Data Based on AMI Measurements*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.mdpi.com/1996-1073/18/1/221

*Descargas – AEADE*. (2025, febrero 7). https://www.aeade.net/sdm_downloads/

*Discovering Electric Vehicle Charging Locations Based on Clustering Techniques Applied to Vehicular Mobility Datasets*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.mdpi.com/2220-9964/13/10/368

*Documentación de Google Maps Platform*. (s. f.). Google for Developers. Recuperado 13 de febrero de 2025, de https://developers.google.com/maps/documentation?hl=es-419

Escorcia Herrera, Á., Vargas Arias, C., Escorcia Herrera, Á., & Vargas Arias, C. (2023). Impuesto al Valor Agregado en Colombia. Consideraciones sobre su incidencia distributiva y propuestas de reforma. *Revista Latinoamericana de Desarrollo Económico*, *39*, 9-33. https://doi.org/10.35319/lajed.202239277

*Estudio de análisis y prospectiva de la electromovilidad en Ecuador y el mix energético al 2030 – Instituto de Investigación Geológico y Energético*. (s. f.). Recuperado

12 de febrero de 2025, de https://www.geoenergia.gob.ec/estudio-de-analisis-y-prospectiva-de-la-electro-movilidad-en-ecuador-y-el-mix-energetico-al-2030/

Fieni, À. M. (s. f.). *Smart University. Detección de Anomalías en Consumo Energético*.

*Global EV Outlook 2023 – Analysis*. (2023, abril 26). IEA. https://www.iea.org/reports/global-ev-outlook-2023

*Google Colab*. (s. f.). Recuperado 13 de febrero de 2025, de https://colab.research.google.com/notebooks/intro.ipynb

Grolemund, H. W. and G. (s. f.). *Welcome | R for Data Science*. Recuperado 13 de febrero de 2025, de https://r4ds.had.co.nz/

*Innovador proyecto de movilidad eléctrica en Quito le apuesta a un Centro Histórico Cero Emisiones—Elements*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.elementsgroup.com.ec/emprendedores-integran-proyecto-de-movilidad-sostenible-en-quito/

Jahangir, H., Gougheri, S. S., Vatandoust, B., Golkar, M. A., Ahmadian, A., & Hajizadeh, A. (2020). Plug-in Electric Vehicle Behavior Modeling in Energy Market: A Novel Deep Learning-Based Approach With Clustering Technique. *IEEE Transactions on Smart Grid*, *11*(6), 4738-4748. IEEE Transactions on Smart Grid. https://doi.org/10.1109/TSG.2020.2998072

Kerrigan, G. (s. f.). *Políticas públicas relacionadas con la electromovilidad en América Latina y El Caribe*.

Koohfar, S., Woldemariam, W., & Kumar, A. (2023). Performance Comparison of Deep Learning Approaches in Predicting EV Charging Demand. *Sustainability*, *15*, 4258. https://doi.org/10.3390/su15054258

*La ARCERNNR presenta la Estadística del Sector Eléctrico Ecuatoriano 2021 – Ministerio de Energia y Minas*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.recursosyenergia.gob.ec/la-arcernnr-presenta-la-estadistica-del-sector-electrico-ecuatoriano-2021/

*Las dos herramientas estadísticas energéticas más importantes del país se presentaron en Quito – Ministerio de Energia y Minas*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.recursosyenergia.gob.ec/las-dos-herramientas-estadisticas-energeticas-mas-importantes-del-pais-se-presentaron-en-quito/

Lascano, J. S., Saraguro, R., Quinatoa, C., Tapia, J., & Chiza, L. (2023). Estimación de la Demanda de una Estación de Carga para Vehículos Eléctricos Mediante la Aplicación de Métodos Probabilísticos. *Revista Técnica energía*, *20*(1), 52-64. https://doi.org/10.37116/revistaenergia.v20.n1.2023.569

Miyazaki, K., Uchiba, T., & Tanaka, K. (2020). Clustering to Predict Electric Vehicle Behaviors using State of Charge data. *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*, 1-6. https://doi.org/10.1109/EEEIC/ICPSEurope49358.2020.9160675

Mora, K. N. Z. (s. f.). *ANÁLISIS DE ESCENARIOS DE PENETRACIÓN DE VEHÍCULOS ELÉCTRICOS EN EL PARQUE AUTOMOTOR EN LA CIUDAD DE QUITO, SU PROSPECTIVA ENERGÉTICA Y HUELLA DE CARBONO*.

*MTOP y BID preparan Estrategia Nacional de Movilidad Segura en Ecuador 2021 – 2030 – Ministerio de Transporte y Obras Públicas*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.obraspublicas.gob.ec/mtop-y-bid-preparan-estrategia-nacional-de-movilidad-segura-en-ecuador-2021-2030/

Paredes, L., Pozo, M., Paredes, L., & Pozo, M. (2020). Movilidad Eléctrica y Eficiencia Energética en el Sistema de Transporte Público del Ecuador un Mecanismo para Reducir Emisiones de CO2. *Revista Técnica energía*, *16*(2), 91-99. https://doi.org/10.37116/revistaenergia.v16.n2.2020.356

Pathak, M., Slade, R., Pichs-Madruga, R., Ürge-Vorsatz, D., Shukla, P. R., Skea, J., Abdulla, A., Al Khourdajie, A., Babiker, M., Bai, Q., Bashmakov, I. A., Bataille, C., Berndes, G., Blanco, G., Cabeza, L. F., Carraro, C., Clarke, L., de Coninck, H., Creutzig, F., … Winkler, H. (2022). Technical Summary. En P. R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, & J. Malley (Eds.), *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 49-147). Cambridge University Press. https://doi.org/10.1017/9781009157926.002

*Proyecto de Electromovilidad impulsado por la Empresa Eléctrica Quito contará con apoyo del PNUD | Programa De Las Naciones Unidas Para El Desarrollo*. (s. f.). Recuperado 12 de febrero de 2025, de https://www.undp.org/es/ecuador/comunicados-de-prensa/proyecto-de-electromovilidad-impulsado-por-la-empresa-electrica-quito-contara-con-apoyo-del-pnud

*Python Data Science Handbook | Python Data Science Handbook*. (s. f.). Recuperado 13 de febrero de 2025, de https://jakevdp.github.io/PythonDataScienceHandbook/

Ramírez, C. Y. (2022). Uso de algoritmos de aprendizaje automático para analizar datos de energía eléctrica facturada. Caso: Chile 2015 – 2021. *Revista de I+D Tecnológico*, *18*(2), 17-31.

*Rendición de Cuentas—Empresa Eléctrica Quito—Empresa Electrica Quito*. (s. f.). Empresa Eléctrica Quito. Recuperado 12 de febrero de 2025, de https://www.eeq.com.ec/rendicion-de-cuentas

Serrano Guerrero, J. X. (2020). *Caracterización de la demanda de energía mediante patrones estocásticos en las Redes Eléctricas Inteligentes* [Universitat Politècnica de València]. https://doi.org/10.4995/Thesis/10251/153810

Yang, Y., Zhang, Y., & Meng, X. (2020). A Data-Driven Approach for Optimizing the EV Charging Stations Network. *IEEE Access*, *PP*, 1-1. https://doi.org/10.1109/ACCESS.2020.3004715

Zhao, H., Tang, J., Adebisi, B., Ohtsuki, T., Gui, G., & Zhu, H. (2022). An Adaptive Vehicle Clustering Algorithm Based on Power Minimization in Vehicular Ad-Hoc Networks. *IEEE Transactions on Vehicular Technology*, *71*(3), 2939-2948. IEEE Transactions on Vehicular Technology. https://doi.org/10.1109/TVT.2021.3140085