Anteproyecto: Guía 4

Según la información de la documentación inicial del caso, se observó que la empresa dispone de información comercial y técnica reportada por el medidor para cada uno de sus clientes no regulados. En cuanto a la información comercial, Electro Dunas cuenta con una base de datos que incluye el número de identificación del cliente, el departamento, la provincia, el distrito, la subestación, las coordenadas de localización, entre otros detalles (base de datos 1). Además, la información técnica abarca los valores históricos de la energía activa entregada (kWh), la energía reactiva entregada (kVarh) y el voltaje reportado cada 15 minutos por el medidor para cada cliente (base de datos 2).

Los datos almacenados en el repositorio https://github.com/Pacheco-Carvajal/GPA-Data-ElectroDunas han sido corroborados. Según lo planteado en el inicio del problema, se dispone de una primera base en el repositorio denominada "sector_economico_clientes.xlsx", la cual enumera a los clientes y sus respectivos sectores económicos. Sin embargo, a diferencia de lo mencionado en la documentación inicial, no se encuentra evidencia del departamento, distrito, subestación, entre otros elementos.

En relación con la información técnica, se han identificado 30 archivos en formato CSV (archivos separados por comas) que contienen la información técnica correspondiente a cada cliente. Estos archivos contienen la fecha de la medición, la energía activa, la energía reactiva, el voltaje FA y el voltaje FC. A pesar de que las bases de datos parecen coincidir con la información inicialmente reportada por la empresa, es importante señalar que los intervalos de tiempo no concuerdan con reportes realizados cada 15 minutos. También se cuenta con un archivo en formato de Excel con la información de los sectores económicos asociados a cada cliente. Para la consolidación de esta información, se utilizó Excel y la función XLOOKUP para agrupar todas estas bases de datos en un solo archivo. Este archivo consolidado será utilizado en el desarrollo del proyecto.

En resumen, la información inicialmente descrita por la empresa se ha verificado y consolidado, pero se han identificado discrepancias en los intervalos de tiempo de la información técnica y la ausencia de algunos elementos en la base de datos de sectores económicos.

Características de Calidad de Datos

- Totalidad

La exploración de la base de datos comenzó con la dimensión de totalidad para verificar la integridad de los registros. El dataframe resultante, denominado "df", contiene 463,425 filas y 9 columnas. Se llevó a cabo un análisis de nulidad que reveló la ausencia de valores faltantes en todas las columnas, indicando una base de datos completa.

- Formato

Se examinaron los tipos de datos de cada columna para asegurar coherencia. Las fechas están en formato datetime64, mientras que las variables numéricas como energía activa, reactiva, voltajes FA y FC están en formato float64. La ausencia de elementos de cadena vacía fortalece la calidad de los datos, proporcionando una base sólida para futuras exploraciones y análisis.

- Consistencia

Se verificó que no haya fechas repetidas dentro de un mismo cliente, evitando duplicidades en los conteos de energía. Sin embargo, se identificaron 505 registros con valores negativos en la energía activa, señalando inconsistencias en la recopilación de estos datos. No se encontraron valores negativos en las mediciones de voltaje, lo que sugiere consistencia en esta dimensión. Se realizó una exploración de los valores únicos en las variables de texto, no revelando inconsistencias ortográficas o errores en la captura de datos. Además, se revisaron los valores mínimos y máximos de las fechas, encontrando rangos que parecen razonables, abarcando desde 2021 hasta 2023.

- Claridad

La claridad de los datos se evaluó mediante estadísticas descriptivas básicas y visualizaciones, buscando comprender la distribución y las relaciones entre las variables clave. Se utilizaron diversas técnicas para proporcionar una visión holística de la calidad de los datos. Las estadísticas descriptivas básicas, como el uso de la función.describe() en la librería pandas, revelaron información valiosa sobre la variabilidad y distribución de las variables. Por ejemplo, se observaron valores negativos en la energía activa, con un mínimo de -1.329018, confirmando la inconsistencia identificada anteriormente.

Las visualizaciones, como los gráficos de series de tiempo y de dispersión, permitieron identificar patrones y anomalías en los datos. Se notaron irregularidades en los valores de energía activa al finalizar 2021 y en febrero y mayo de 2022. Estas observaciones resaltan la importancia de abordar específicamente estos períodos para garantizar la precisión de los análisis. La matriz de correlación y el correlograma proporcionaron información sobre las relaciones lineales entre las variables. Se identificaron fuertes correlaciones positivas, como la relación entre los voltajes FA y FC (0.95) y entre la energía reactiva y activa (0.64). Estos resultados, aunque indican relaciones, también podrían sugerir posibles multicolinealidades que deben considerarse en futuros análisis.

- Concordancia con el Problema de Negocio

El problema de negocio de Electro Dunas se centra en mejorar la eficiencia operativa, especialmente en el consumo de energía eléctrica de los clientes no regulados. La calidad de los datos es esencial para lograr estos objetivos. En general, las características de calidad de los datos son favorables, destacando la integridad y la coherencia en la mayoría de las dimensiones. Sin embargo, la presencia de valores negativos en la energía activa representa una inconsistencia crítica que debe abordarse para garantizar la validez de los resultados analíticos. La exploración visual de los datos reveló patrones temporales y anomalías que deben considerarse en el análisis posterior. Por ejemplo, las irregularidades en los valores de energía activa en ciertos períodos podrían indicar problemas en la medición o cambios significativos en el consumo de energía.

Proceso de limpieza de datos

- Eliminación de Registros con Valores Negativos de Energía

El primer paso consistió en abordar la presencia de valores negativos en la variable de energía activa. Se eliminaron 505 registros que contenían valores de energía negativos. La eliminación de estos

registros radica en que los valores negativos en la energía activa no tienen sentido en el contexto de consumo eléctrico. Podrían deberse a errores en la medición o problemas técnicos. Al remover estos registros, se mejora la integridad de la información y se garantiza la consistencia de los datos utilizados en el análisis posterior.

- Eliminación de Columnas Redundantes

Se identificaron columnas redundantes, específicamente 'Source.Name', 'CLIENTE', y 'Proper', todas referentes al identificador del cliente. En este paso, se decidió conservar solo la columna 'Proper' debido a que presenta una representación visual y gramatical más adecuada. Además, se eliminaron las columnas 'Source.Name' y 'CLIENTE'. Esta acción simplifica la estructura de la base de datos y reduce la redundancia, facilitando así futuros análisis.

- Renombrado de Etiquetas de Columnas

Para mejorar la coherencia y comprensión de la base de datos, se optó por renombrar las etiquetas de las columnas. Algunas etiquetas estaban en español y otras en inglés, y se buscó unificarlas en español. Se creó un diccionario de mapeo con los nuevos nombres de las columnas y se utilizó el método rename para aplicar estos cambios. Este paso contribuye a una mayor armonización y comprensión de la base de datos, facilitando su interpretación por parte de los analistas y usuarios finales.

Las técnicas de limpieza de datos implementadas se eligieron de manera coherente con el propósito de mejorar la calidad y utilidad de la base de datos de Electro Dunas. La eliminación de registros con valores negativos aborda inconsistencias críticas en los datos de consumo de energía. Simultáneamente, la eliminación de columnas redundantes simplifica la estructura de la base de datos, y el renombrado de etiquetas de columnas contribuye a mejorar la coherencia y comprensión de la información. Asimismo, no se aplican técnicas de imputación, dado que los datos presentan completitud, es decir, no hay registros ausentes. Además, se decide preservar los valores atípicos, ya que el objetivo del proyecto está asociado a identificar comportamientos anómalos que estarán determinados por estos valores atípicos.

Proceso de entendimiento de datos: justificación proceso de entendimiento datos desde el problema de negocio

Para la realización del proceso de entendimiento inicial de los datos, se ha decidido implementar diversas técnicas de análisis estadístico que se realizaron posterior al proceso de limpieza y transformación de los datos, esto con el fin de asegurar que se pueden alcanzar los objetivos establecidos al momento de definir la problemática de negocio que buscamos resolver. Seleccionamos y justificamos las siguientes técnicas para lograr un entendimiento efectivo de los datos:

Agregar Nueva información a los Datos:

- **Identificador de Cliente y Sector:**

Durante el proceso de consolidación de información, se consideró que sería muy útil agregar las características principales de los clientes asociadas a cada uno de los registros de consumo correspondiente, esto con el fin de contar una fácil identificación de los mismos al momento de realizar los procesos de exploración y segmentación de los datos, esta decisión del equipo conllevó que tuviésemos dos columnas extras resultantes '**Cliente**' y '**Sector Económico**' donde nos indican el identificador del cliente y al sector económico al que pertenece respectivamente.

- **Eficiencia Energética:**

La eficiencia energética es una variable importante que debemos calcular y añadir a nuestro proyecto porque nos permite evaluar el desempeño energético de los clientes e identificar oportunidades de mejora.

La eficiencia energética mide la relación entre la energía activa consumida y la potencia aparente consumida. La energía activa es la energía que se utiliza para realizar un trabajo útil, mientras que la potencia aparente es la potencia total que se consume.

Un cliente con una eficiencia energética alta es aquel que consume menos energía para realizar un trabajo útil. Esto significa que el cliente está utilizando su energía de manera más eficiente, lo que puede reducir sus costes energéticos y su impacto ambiental.

- Energía Total Consumida:

La energía total consumida se calcula sumando la energía activa consumida y la energía reactiva consumida. La energía activa es la energía que se utiliza para realizar un trabajo útil, mientras que la energía reactiva es la energía que se utiliza para crear un campo magnético.

- La energía total consumida es una variable importante que debemos calcular y añadir a nuestro proyecto porque nos permite:

- Comparar el consumo de energía de diferentes clientes. Esto nos puede ayudar a identificar clientes con un consumo de energía alto, que podrían ser candidatos para programas de eficiencia energética.

- Identificar tendencias en el consumo de energía. Esto nos puede ayudar a planificar la capacidad de la red eléctrica y a desarrollar políticas energéticas.

- Evaluar el impacto ambiental de la empresa. El consumo de energía es una de las principales fuentes de emisiones de gases de efecto invernadero.

- Factor de potencia y desviación de voltaje:

Los cálculos de la calidad de la energía son importantes para efectos del entendimiento de los datos y para el cumplimiento de los objetivos de proyecto por las siguientes razones:

Para entender los datos: Los indicadores de calidad de la energía proporcionan información sobre el estado de la red eléctrica y los equipos eléctricos que la utilizan. Esta información puede utilizarse para identificar problemas potenciales en la red eléctrica o en los equipos eléctricos.

Para cumplir los objetivos de proyecto: El proyecto de Electro Dunas tiene como objetivo mejorar la eficiencia energética y la sostenibilidad de la empresa. Los cálculos de la calidad de la energía pueden utilizarse para identificar oportunidades de mejora en estas áreas.

Por ejemplo, si el cálculo del voltaje promedio muestra que el voltaje es inestable, esto puede indicar que la red eléctrica está sobrecargada o que hay problemas con los equipos eléctricos. Esta información puede utilizarse para tomar medidas para mejorar la estabilidad del voltaje.

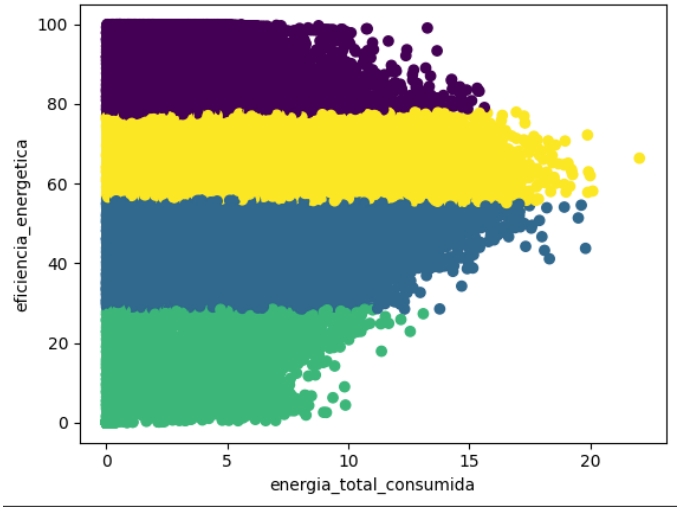
Si el cálculo del factor de potencia promedio muestra que el factor de potencia es bajo, esto puede indicar que la energía eléctrica no se está utilizando de manera eficiente. Esta información puede utilizarse para tomar medidas para mejorar el factor de potencia.

Segmentar los datos y encontrar agrupaciones:

Para llevar a cabo la segmentación de datos, se ha optado por la implementación de un modelo de clusterización conocido como KMeans. Este modelo posibilita la agrupación de datos según sus similitudes. En este escenario, se emplearán dos variables previamente calculadas: la energía total consumida y la eficiencia energética. La elección de estas variables apunta a identificar agrupaciones que compartan niveles similares de consumo de energía y eficiencia.

La determinación del número de grupos se basará en la magnitud de los datos y en los objetivos específicos del análisis. En el contexto del proyecto de Electro Dunas, el propósito central radica en identificar oportunidades de mejora en la eficiencia energética de los clientes. Tras evaluar los resultados de las posibles agrupaciones, se ha establecido que el número óptimo de clústers es 4, lo que representa segmentos que contienen aproximadamente el 25% de los registros totales.

Estos grupos permiten interpretar la energía total consumida en cada registro y su correspondiente eficiencia energética. Esta clasificación nos posibilita establecer 4 categorías que reflejan la calidad de transmisión durante el periodo definido, además de discernir los patrones de consumo energético presentes en cada cliente.



Descripción Estadística Básica:

La descripción estadística de los datos es necesaria para poder tener un mejor entendimiento de los mismos. Esta descripción nos permite conocer las características principales de los datos, como su tendencia central, su dispersión y su forma.

Mencionado lo anterior esta una herramienta fundamental en cualquier análisis de datos, y en el proyecto de Electrodunas, su importancia radica en varios aspectos:

1. Tipificación de comportamientos: Con este punto no ayuda a poder reconocer tendencias o comportamientos recurrentes en los datos. Podríamos encontrar dadas comparaciones el consumo promedio de energía eléctrica en diferentes momentos del año, se puede definir la estacionalidad del consumo, en otras palabras, si observamos que la energía activa promedio es mayor en los días de verano que en los días de invierno, podemos concluir que el consumo de energía eléctrica es mayor en los días de verano.

2. Comparación de conjuntos de datos: Nos puede ayudar a la identificación de conjuntos o subconjuntos de agrupaciones de los datos, como la energía activa consumida en distintas zonas geográficas. Esto ayuda a determinar si hay diferencias significativas en el consumo o si hay patrones similares en diferentes áreas dadas condiciones climáticas, físicas, de infraestructura, desarrollo etc.

3. Identificación y detección de outliers: Identifica valores que se están significativamente por fuera de la distribución o comportamiento general de los datos, en otras palabras, todos aquellos datos que se encuentran muy alejados del resto. Estos valores pueden indicar errores de medición o problemas generales con la prestación del servicio en puntos específicos con los clientes, lo que permite tomar medidas correctivas o validar la integridad de los datos y también implementar las estrategias para solventar el error en las mediciones.

Finalmente, algunos de las medidas estadísticas que podríamos tener en cuenta frente a la medición de estadísticos descriptivos se pueden incluir:

- Medidas de tendencia central: como la media, mediana y moda, que ofrecen una idea de dónde se concentran los datos.
- Medidas de dispersión: como la desviación estándar, el rango intercuartílico y, que muestran cuánto se alejan los datos de la media o mediana.
- Representaciones gráficas: como histogramas, diagramas de caja (boxplots) o gráficos de dispersión, que visualizan la distribución y la variabilidad de los datos.

Al analizar las variables de Energía total consumida y la eficiencia energética por cada uno de los sectores, encontramos los resultados que muestran la cantidad total de energía consumida y la eficiencia energética promedio en diferentes sectores económicos.

- Captación, tratamiento y distribución de agua: Este sector ha consumido una considerable cantidad de energía, indicando la demanda energética asociada a las operaciones de

tratamiento y distribución de agua. La eficiencia energética, aunque no es la más alta en comparación con otros sectores, aún muestra una eficiencia razonable en el uso de la energía.

- Cultivo de Hortalizas: El consumo de energía en este sector es menor en comparación con otros, lo que podría deberse a la naturaleza de las actividades agrícolas. La eficiencia energética parece estar en el rango medio.

- Cultivo de hortalizas y melones, raíces y tubérculos: Similar al cultivo de hortalizas, el consumo de energía es moderado, pero la eficiencia energética es considerablemente alta en este caso, posiblemente debido a prácticas más eficientes en el manejo de energía.

- Cultivo de otros frutos y nueces de árboles y arbustos: Este sector tiene un consumo de energía moderado con una eficiencia energética razonablemente alta, lo que sugiere prácticas más eficientes en comparación con otros sectores agrícolas.

- Cultivo de Árboles Frutales y Nueces: Muestra un alto consumo de energía, probablemente debido a las demandas energéticas asociadas con la agricultura intensiva. A pesar del alto consumo, la eficiencia energética es relativamente alta, lo que puede ser indicativo de prácticas más eficientes en el uso de la energía en este sector.

- Elaboración de cacao y chocolate y de productos de confitería: Presenta un alto consumo de energía, lo que es comprensible debido a las demandas energéticas en los procesos de elaboración. La eficiencia energética es razonablemente alta, lo que sugiere un manejo relativamente eficiente de la energía.

- Venta al por mayor de metales y minerales metalíferos: Este sector muestra un consumo de energía considerable, probablemente debido a las operaciones intensivas asociadas con la manipulación y el procesamiento de metales. La eficiencia energética es alta, lo que indica una gestión efectiva de la energía a pesar del alto consumo.