



Alstom Proyect

By: Javier Jiménez



Objetivo

El objetivo de este análisis exploratorio es identificar y analizar problemas técnicos que afectan la operación de los trenes, tales como: ruido eléctrico en las redes internas de comunicación, fallos de conexión a la red, interferencias en los sensores, y reinicios periódicos de variables debido a desbordamientos. Además, buscamos detectar las posibles causas de estos errores para proponer soluciones efectivas.

Sobre la Data



Número de trenes

El registro cuenta con una flota de 20 trenes.



Fecha de los registros

La data se extiende desde el 15 de febrero a 15 de marzo del 2024.

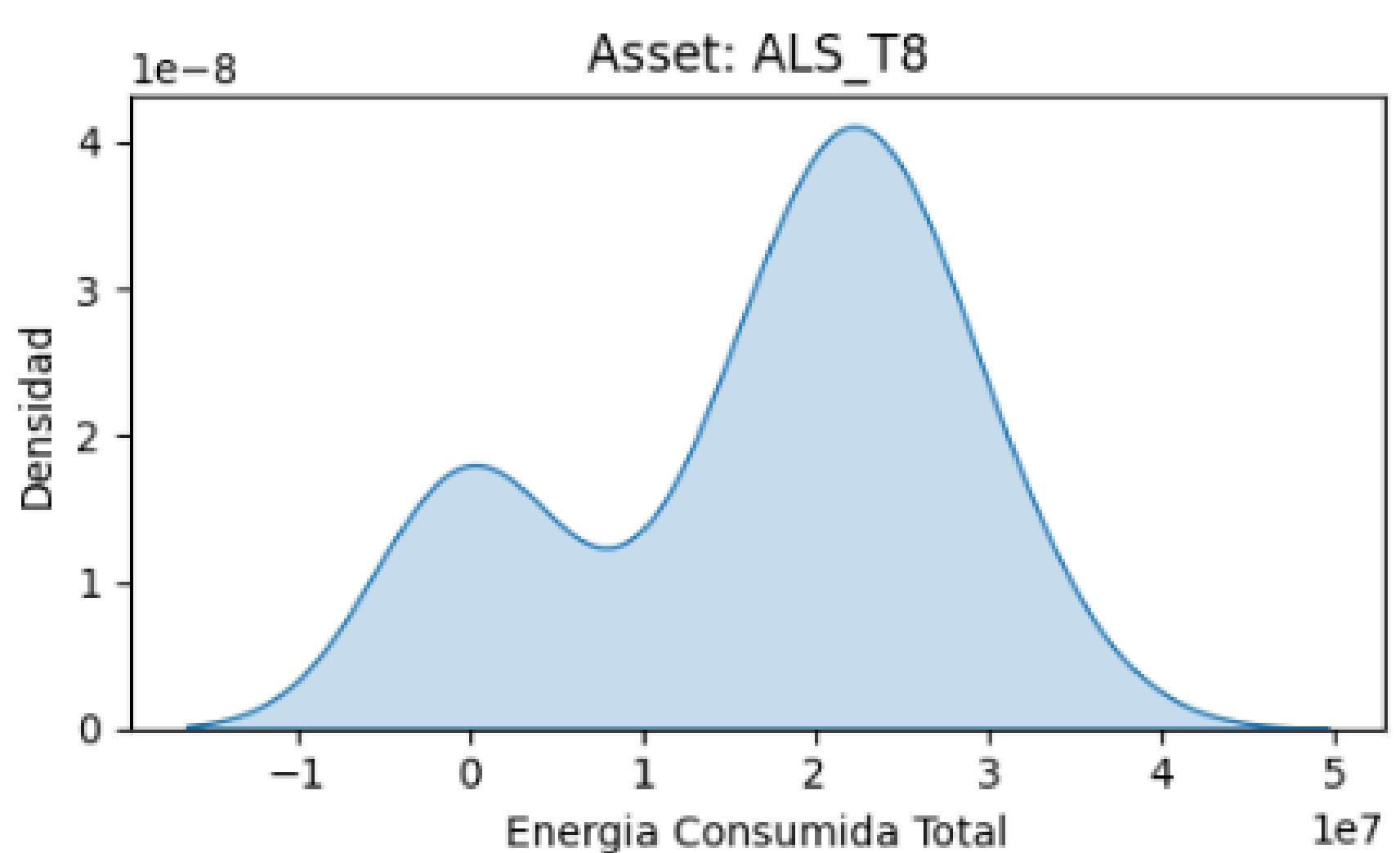


Data en formato acumulativo

La data de las variables numéricas se encuentra en un formato acumulativo.



Insights

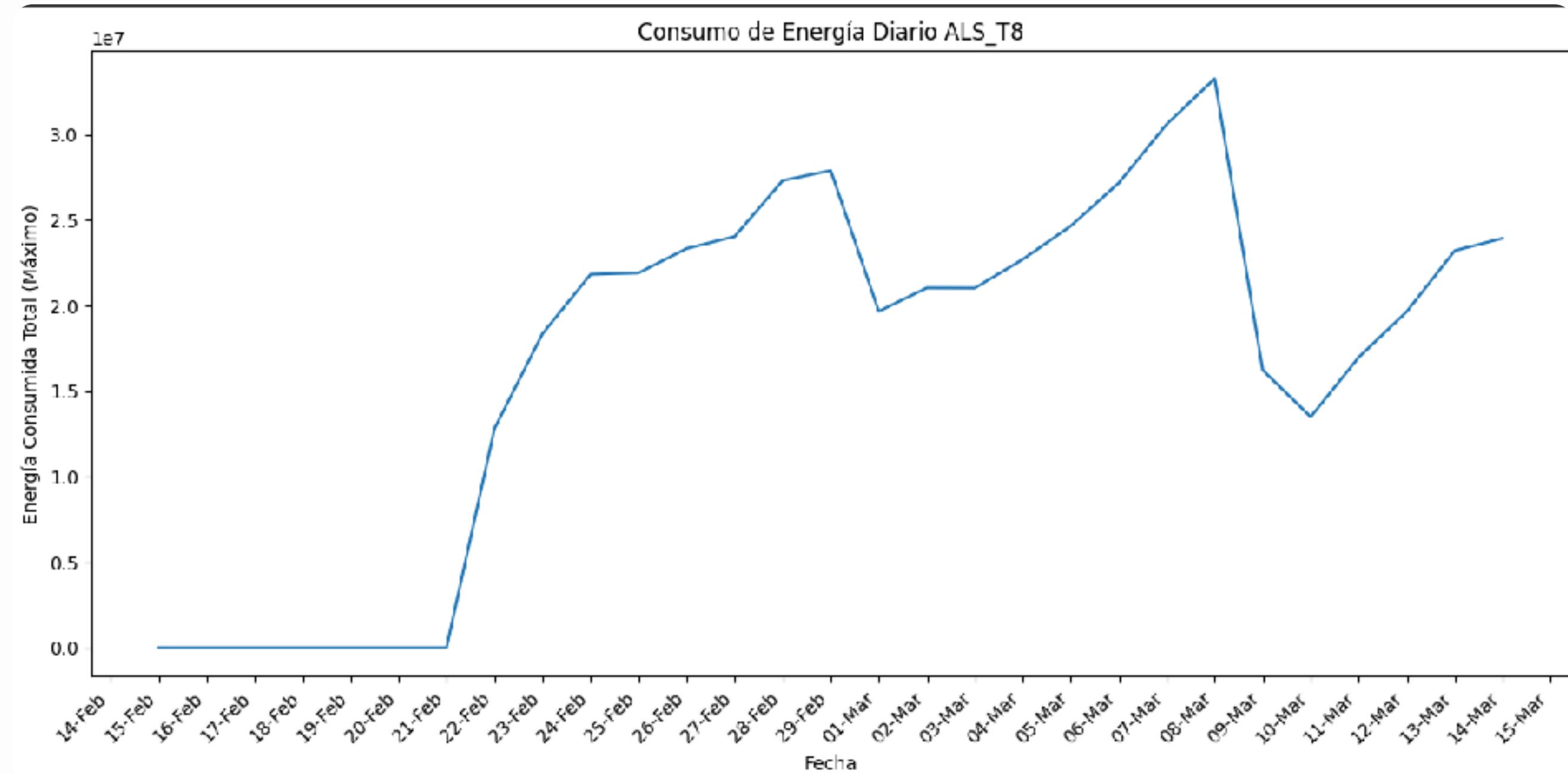


Encontramos una gran cantidad de valores en cero, sobre el consumo energético en este tren.

¿Estarán distribuidos de una manera aleatoria a lo largo de los registros de este tren?

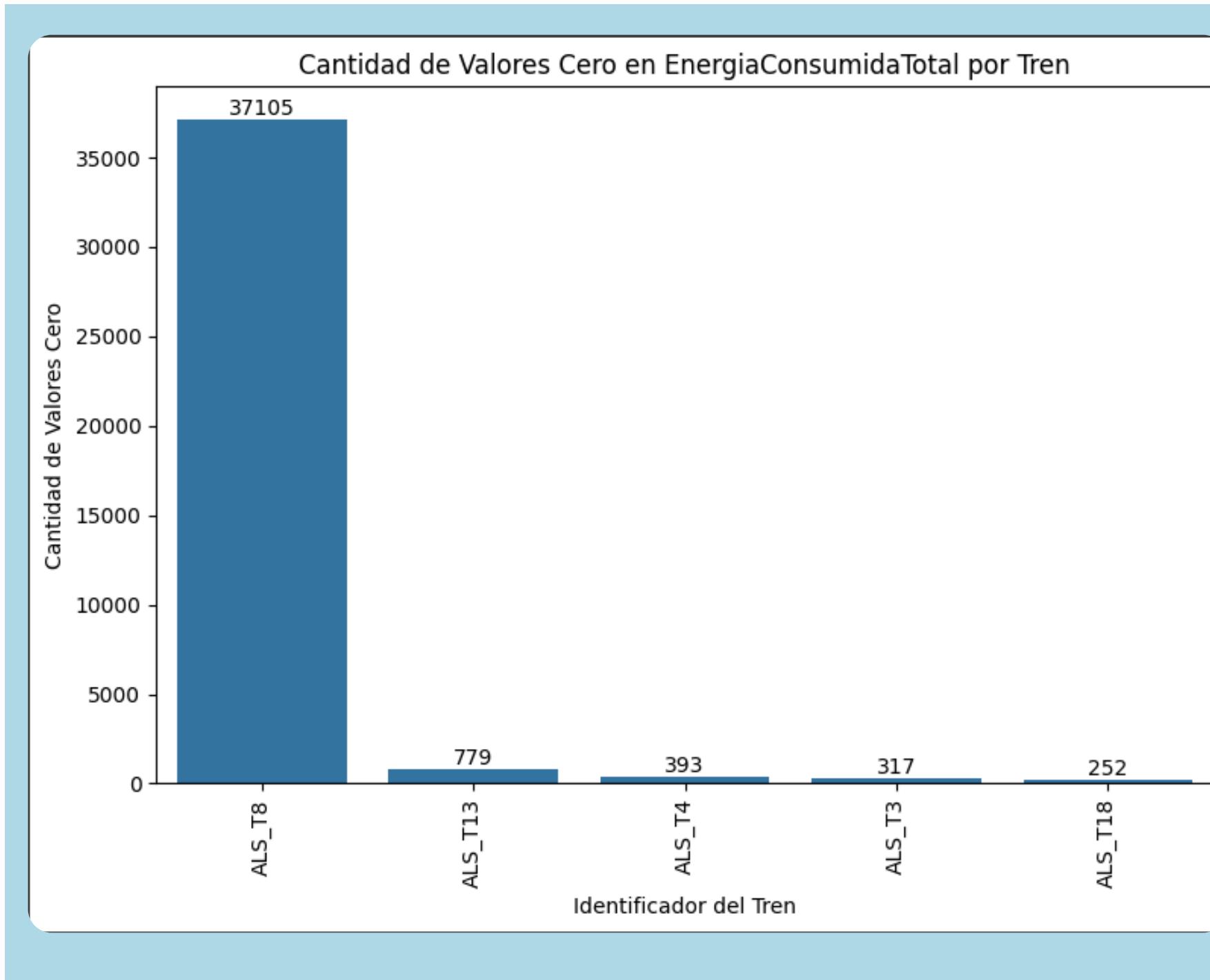
En caso de existencia de valores faltantes, ¿Estarán distribuidos de manera uniforme a lo largo del dataset?

Valores faltantes



Los registros de cero del tren N°8 no están distribuidos de manera aleatoria, ya que todos se encuentran presentes en el rango del 15 a 21 de febrero. Debido a que todos estos registros marcaron cero en estas fechas, podemos concluir que hubó un error debido a un posible desbordamiento de información o algún error debido al reemplazo de algún componente del tren .

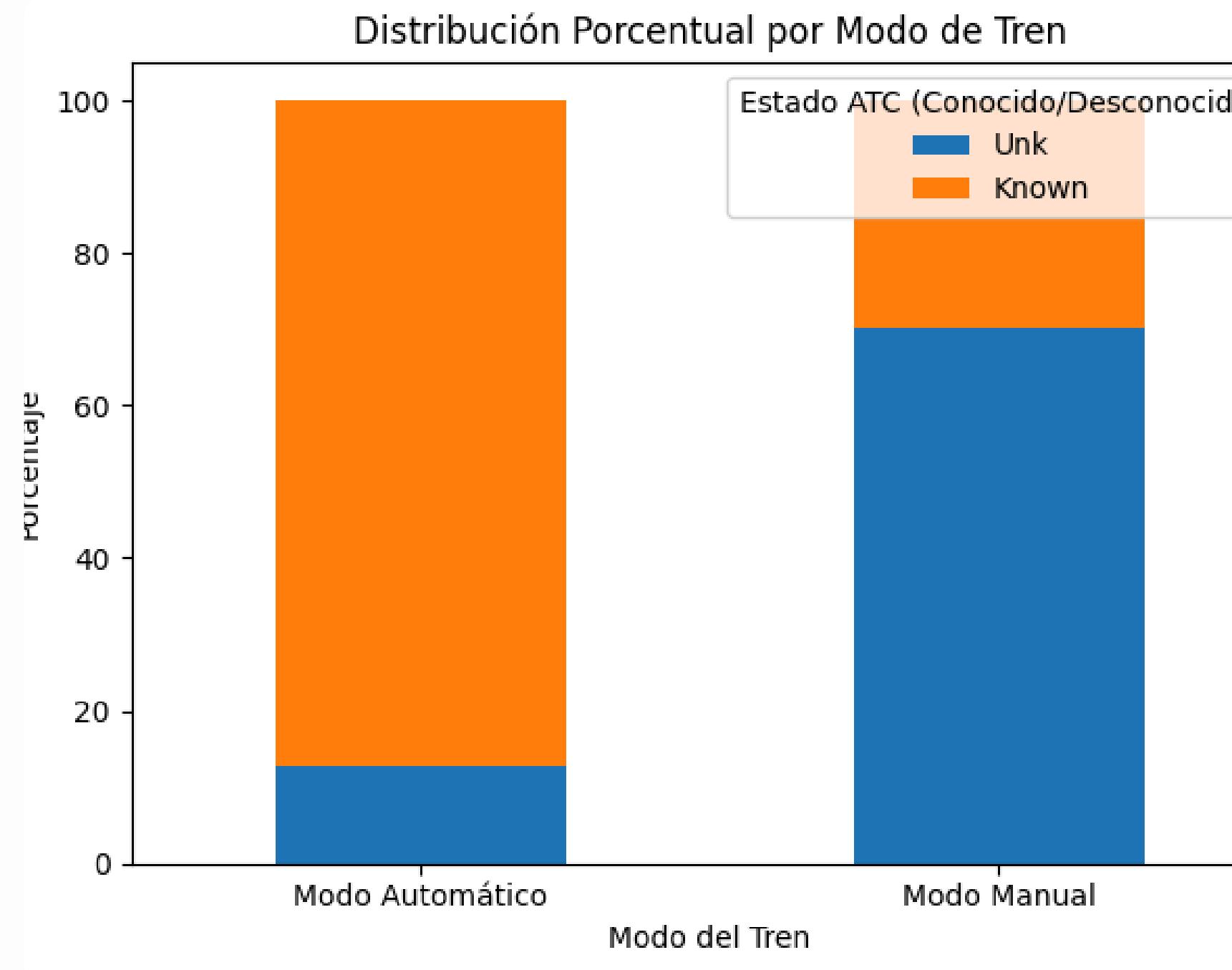
Valores faltantes en energía Consumida por tren



Los valores no parecen estar distribuidos de manera uniforme a lo largo del dataset, ya que vemos una variabilidad entre cada tren. Esto confirma que los valores no son nulos de manera aleatoria, por lo que existe un patrón.



RUTAS DESCONOCIDAS EN LOS TRENES

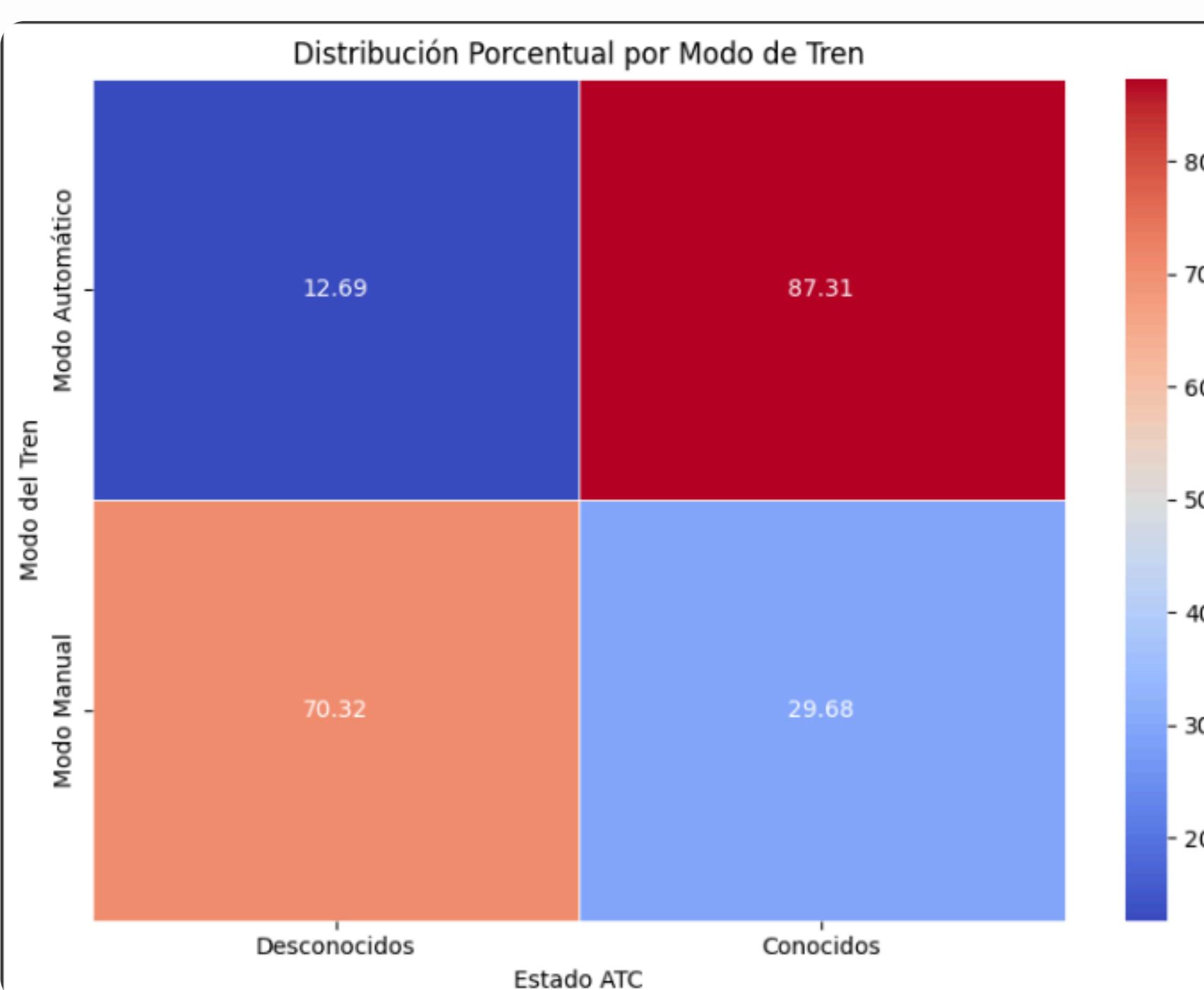


¿Cómo influye el modo de conducción del tren con el conocimiento de la ruta de origen y destino?

Cuando el tren está en modo manual, los datos pueden verse afectados por la intervención humana o por las variaciones en la operación. Esto puede generar inconsistencias en cómo la computadora recopila y transmite la información debido a una mayor variabilidad en la operación del tren

El tren tiene múltiples sistemas y sensores que registran datos como el consumo energético, la distancia, y otros parámetros operacionales.

En el modo automático, esos datos pueden estar mejor organizados y priorizados para ser transmitidos de forma más eficiente, mientras que en el modo manual, el sistema puede estar enfocado en otras funciones críticas de operación, lo que podría influir en la transmisión de datos.

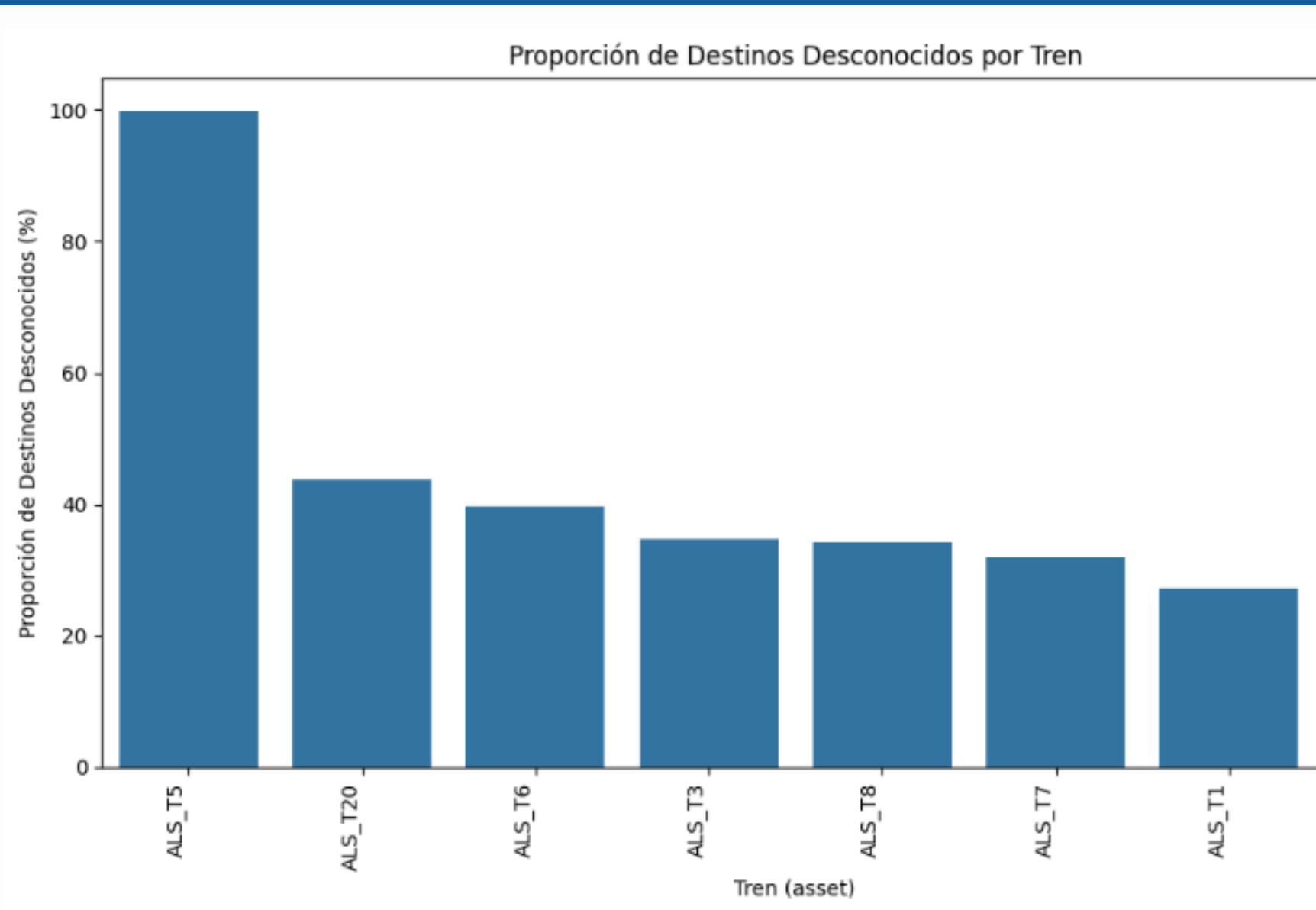


Podemos observar una diferencia significativa en la cantidad de rutas de destino y salida desconocidas cuando el tren opera en modo manual. Para investigar si esta diferencia es estadísticamente relevante, aplicamos una prueba de Chi cuadrado con las siguientes hipótesis:

Hipótesis Nula: No existe una relación estadísticamente significativa entre el modo de operación del tren y el conocimiento de los valores ATC (rutas de destino y salida).

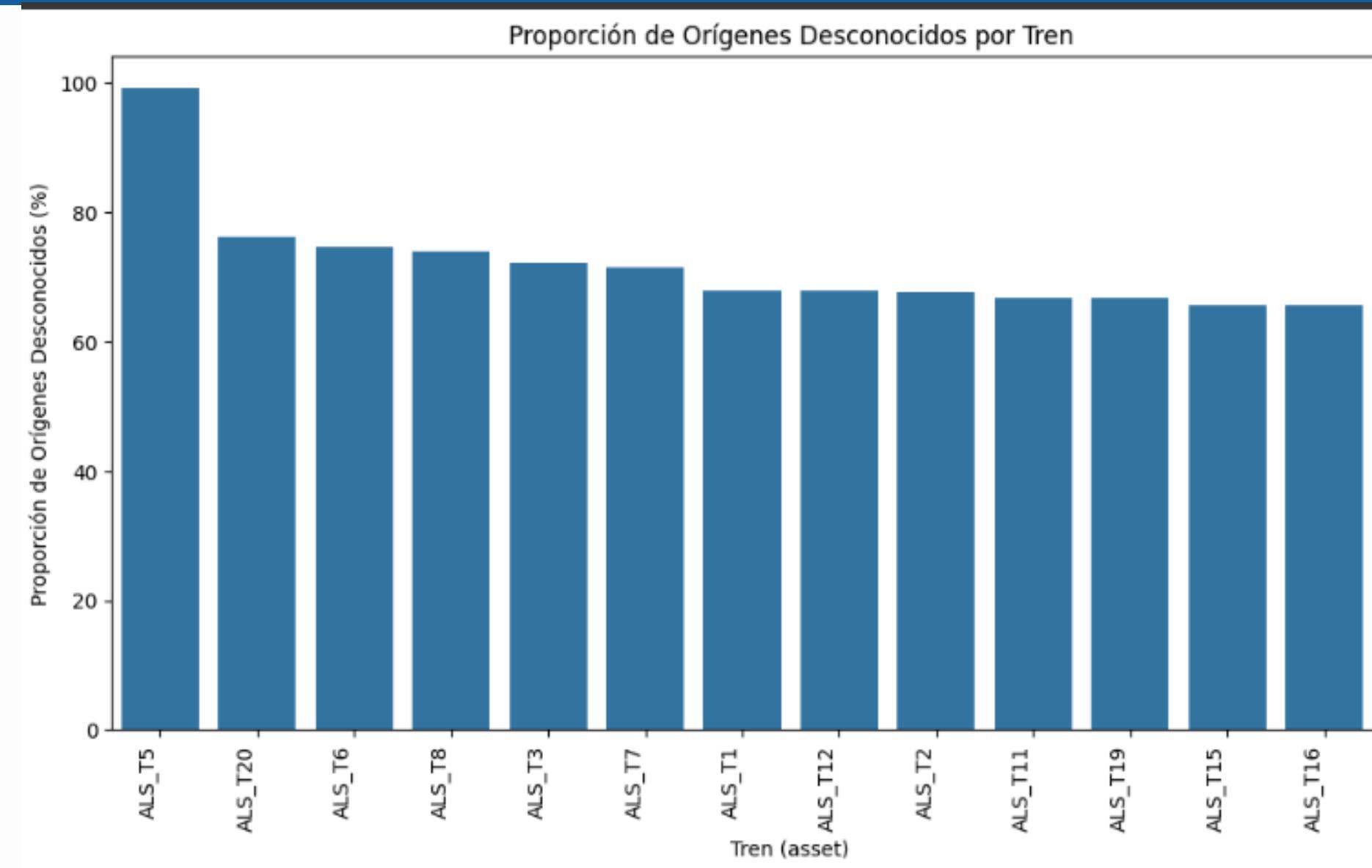
Hipótesis Alternativa: Existe una relación estadísticamente significativa entre el modo de operación del tren y el conocimiento de los valores ATC.

Resultado: El p-valor obtenido fue menor al nivel de significancia de 0.01, lo que nos lleva a rechazar la hipótesis nula. Esto indica que existe una relación estadísticamente significativa entre el modo de operación del tren y el conocimiento de los valores ATC. Además, el valor de Crámer obtenido fue de 0.58, lo que sugiere una relación moderadamente fuerte entre estas variables. Esto respalda la idea de que el modo de conducción (manual o automático) influye en la precisión de los datos de origen y destino.



¿La cantidad de ATC desconocidos es proporcional a la cantidad de registros de cada tren?

La cantidad de registros con ATC desconocidos tiende a ser proporcional al número total de registros de cada tren, con la excepción del tren N°5, que presenta un porcentaje de ATC desconocidos cercano al 100%.

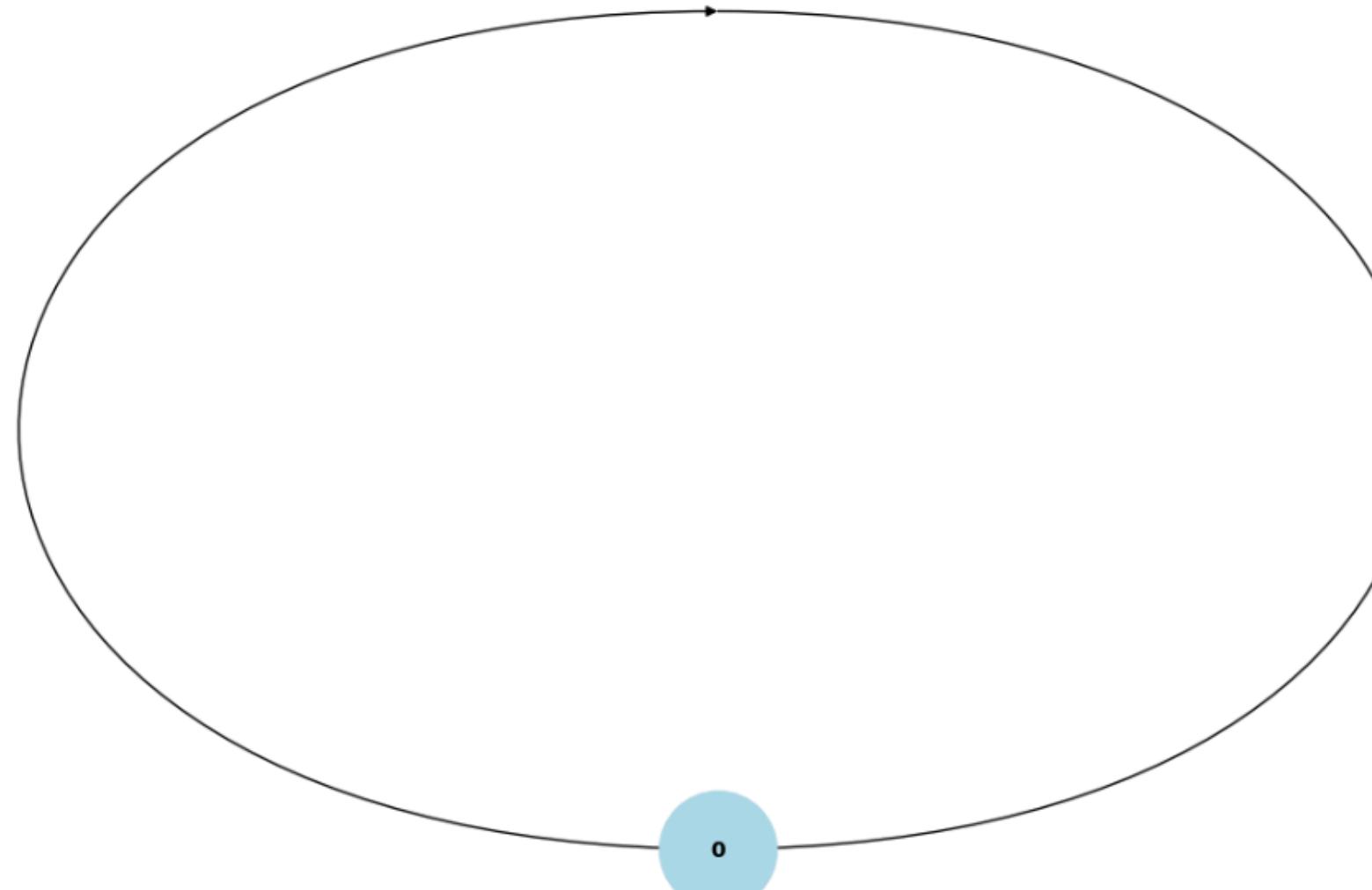


Porqué existe una mayor cantidad de orígenes desconocidos vs destinos desconocidos

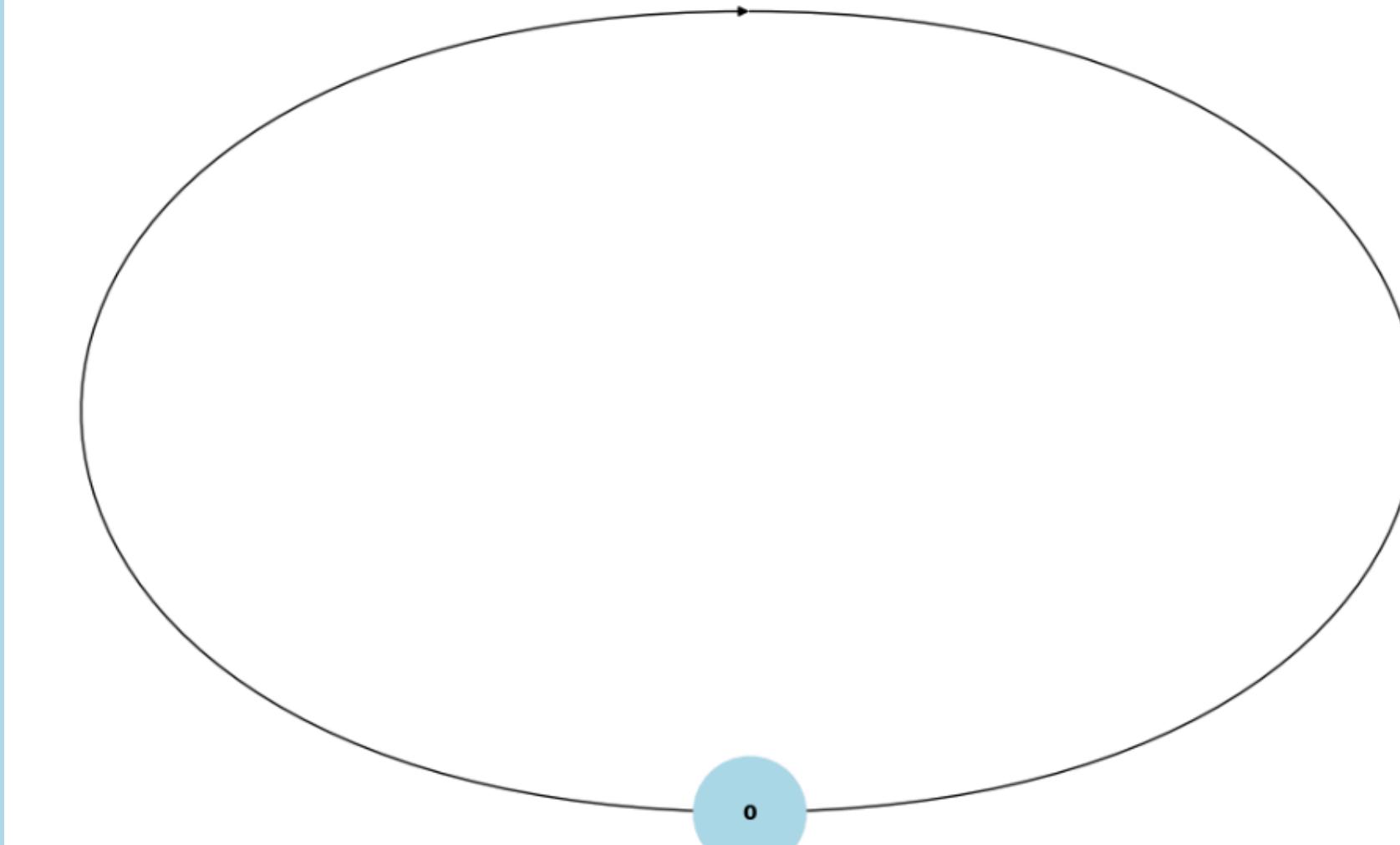
La hipótesis es que los trenes podrían tener mayor cantidad de fallos en la comunicación debido a una combinación de fallos técnicos como: fallos de comunicación más constantes al inicio del trayecto u condiciones operativas como las maneras de registrar eventos.

Solución propuesta a este problema

Árbol de recorrido del Tren ALS_T8 el 2024-02-16



Árbol de recorrido del Tren ALS_T5 el 2024-03-12

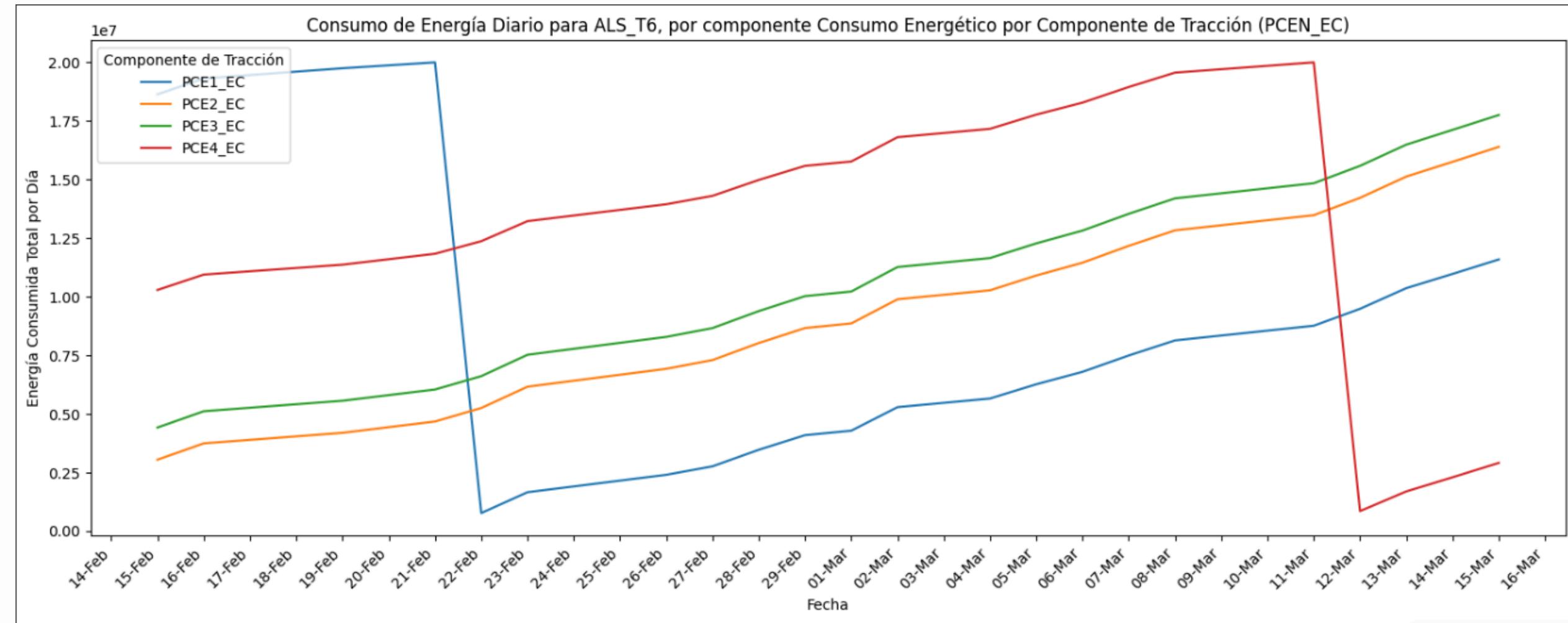


Para eliminar problemas de inconsistencia de la data como el ejemplo propuesto, donde los trenes solo tienen un único trayecto, eliminaremos todas las salidas desconocidas.

COMPONENTES ENERGÉTICOS DEL TREN



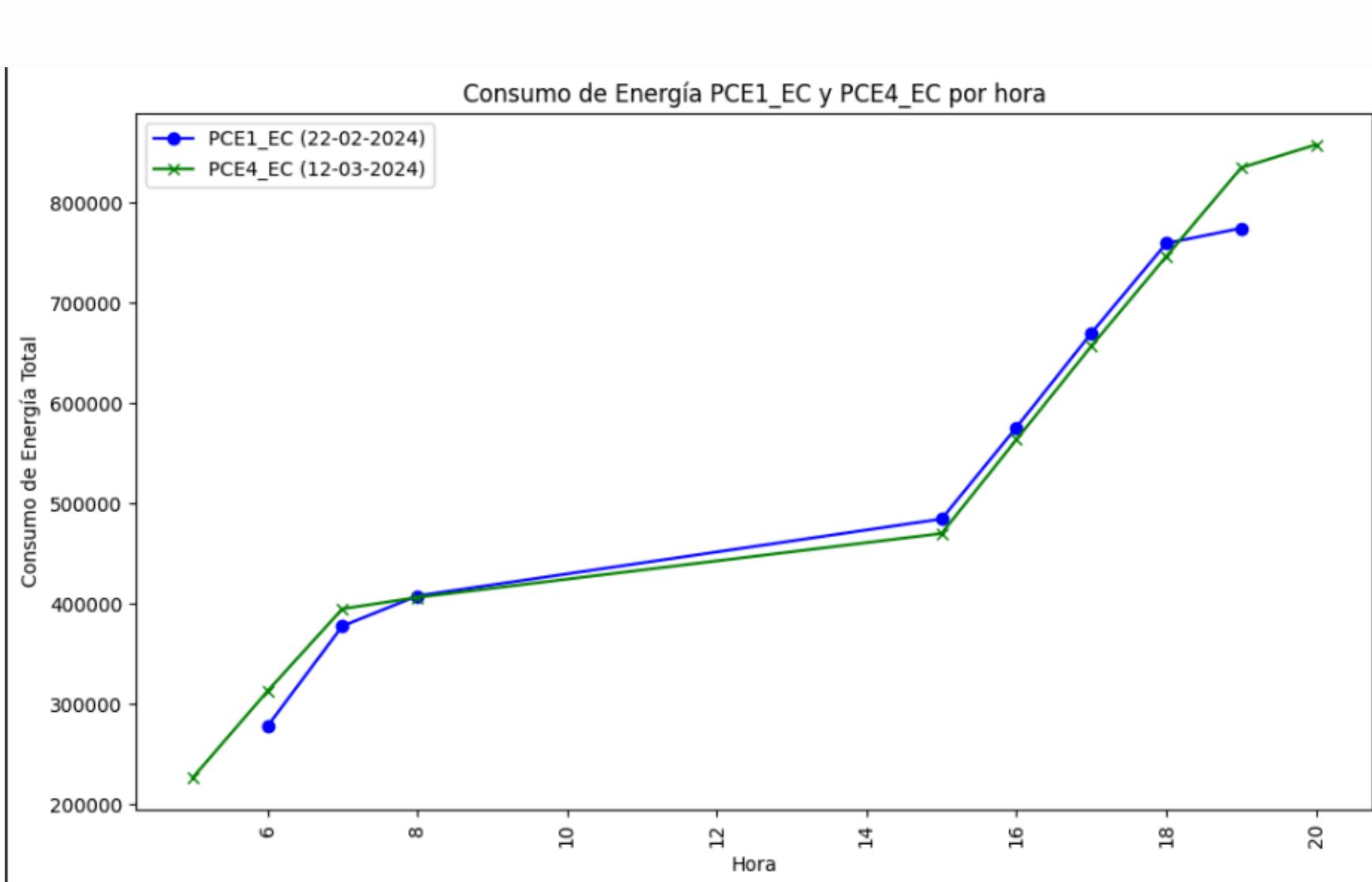
Componentes de Tracción (PCEN_EC)



Se seleccionó un tren de manera aleatoria, para estudiar el comportamiento de sus componentes energéticos.

En un escenario ideal, todos los componentes seguirían el patrón de los componentes PCE2_EC Y PCE3_EC, donde el consumo energético sigue una tendencia bien definida a lo largo del tiempo, a medida que los sensores se desequilibran.

¿Qué genera ese desequilibrio?



Recordando que estamos comparando los componentes en diferentes fechas, con diferentes horarios, por eso un componente registra una hora antes que el anterior, pero igual podemos observar que siguen un comportamiento muy parecido en cuanto a cantidad de energía.

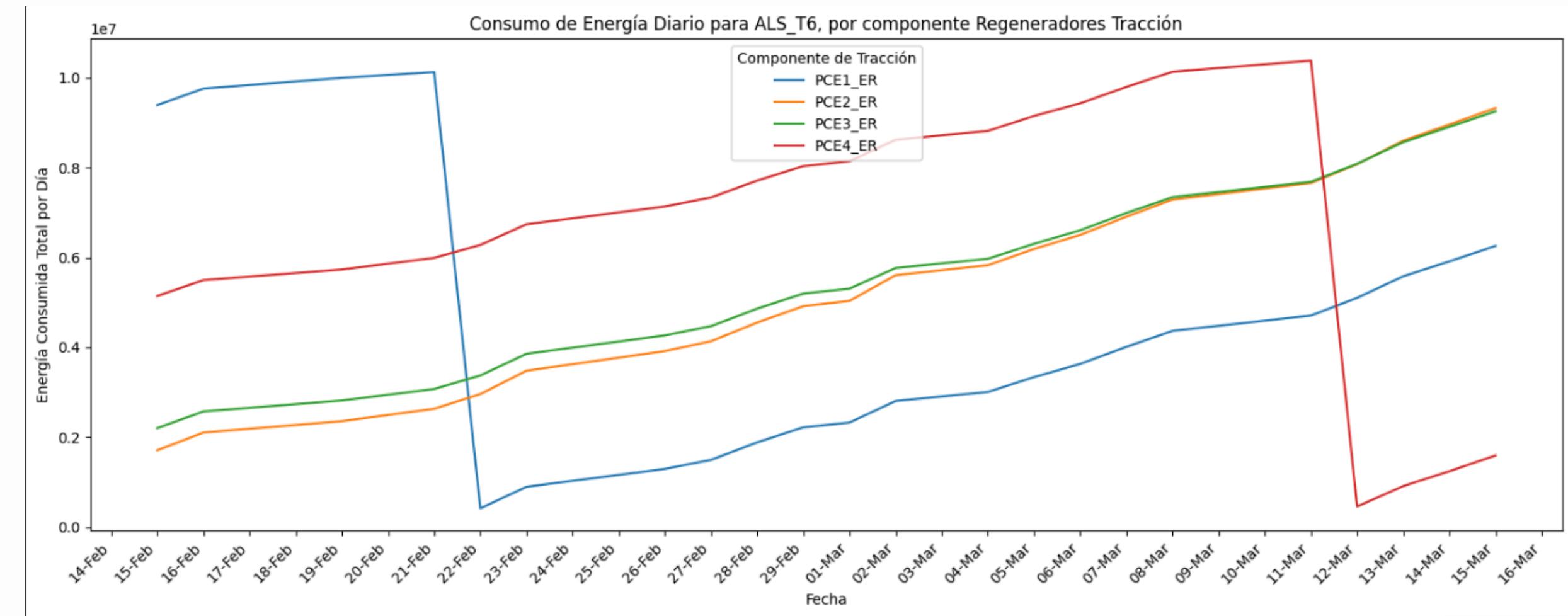
Análisis de las caídas repentinas sobre ambos componentes

Al examinar más de cerca las caídas abruptas en los componentes, observamos que estas ocurren cuando el consumo energético alcanza su punto máximo. En particular, ambas caídas coinciden con momentos en que el componente se aproxima a los 20M de consumo energético.

Con base en esta observación, proponemos la siguiente hipótesis:

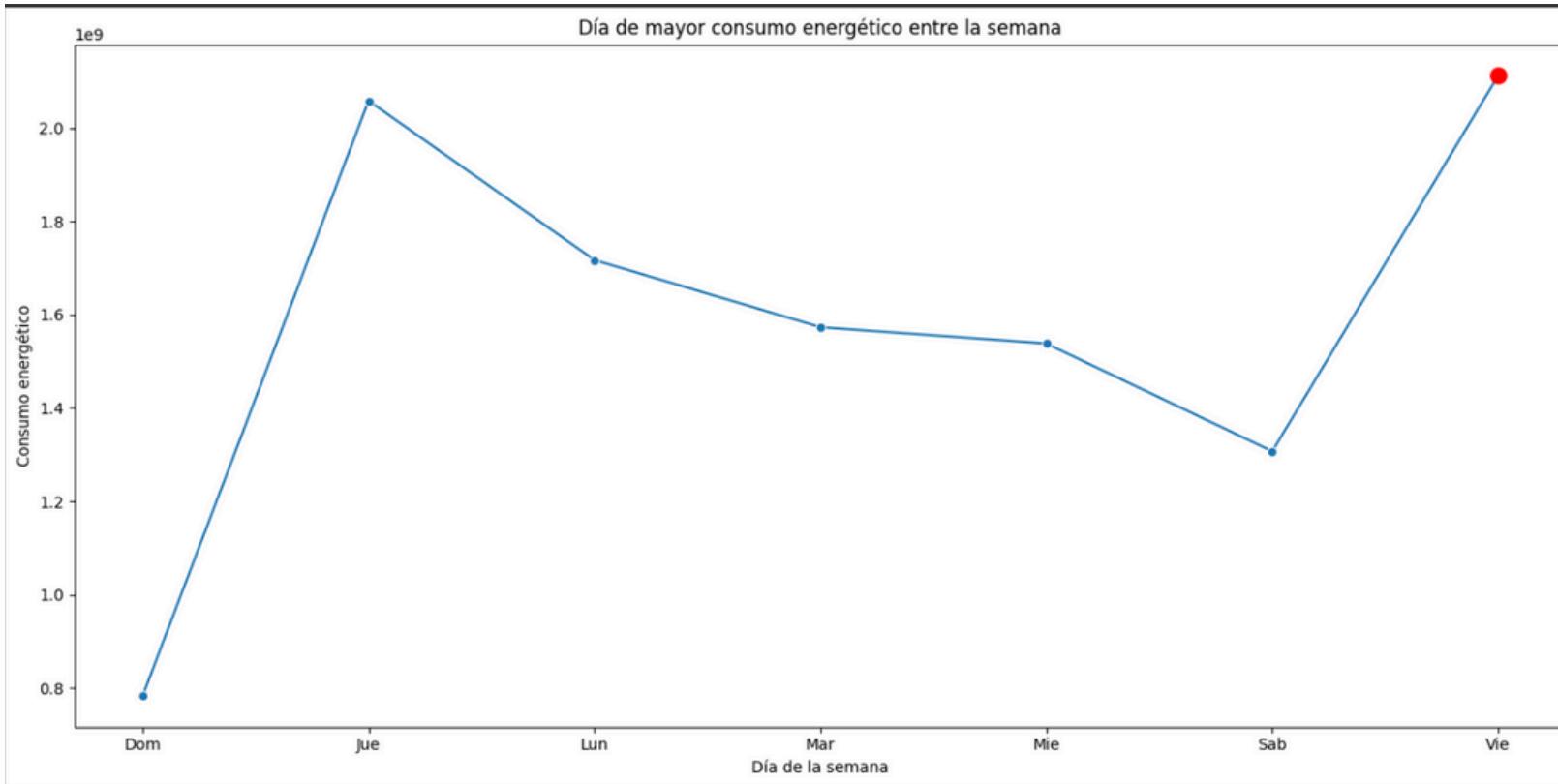
Cuando un componente de PCE1_EC alcanza aproximadamente los 20M de consumo energético, podría estar ocurriendo un desbordamiento de memoria en el sistema de control del tren, lo que genera la caída repentina de los valores registrados. Este fenómeno no se observa en los demás componentes porque no alcanzan dicho límite a lo largo del tiempo. Este patrón sugiere un posible problema de mantenimiento, ya que recalibrar solo un componente en cada intervención no es eficiente ni óptimo desde el punto de vista operativo.

Componentes de Regeneración energética (PCEN_ER)



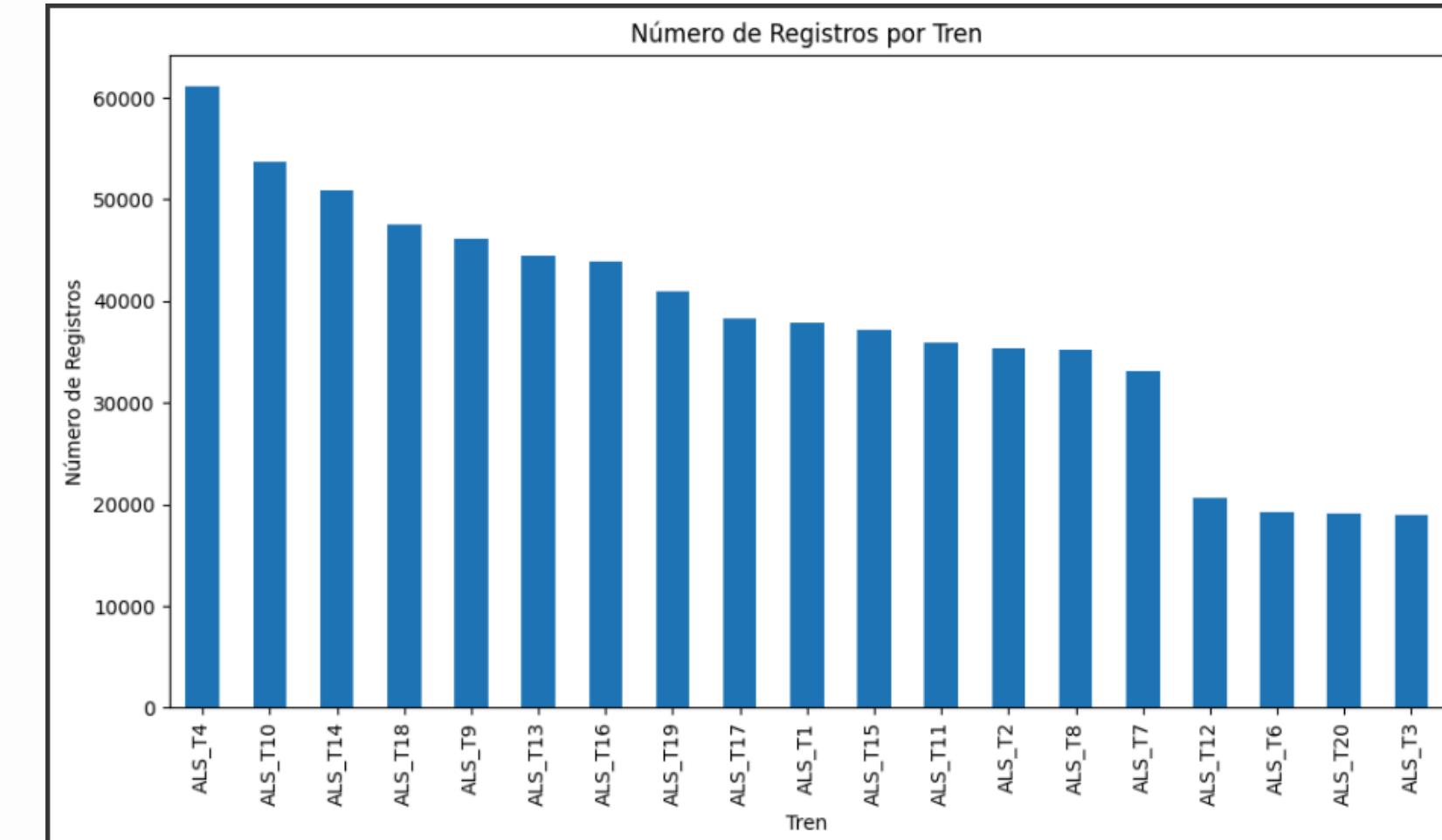
Esta gráfica para los componentes regenerativos presenta el mismo patrón para PCE1_ER y PCE4_ER, por lo que parece que el problema no es desbordamiento de memoria, sino que el mantenimiento se hace cuando un componente llega a una cantidad de energía registrada.

Día de la semana con mayor consumo



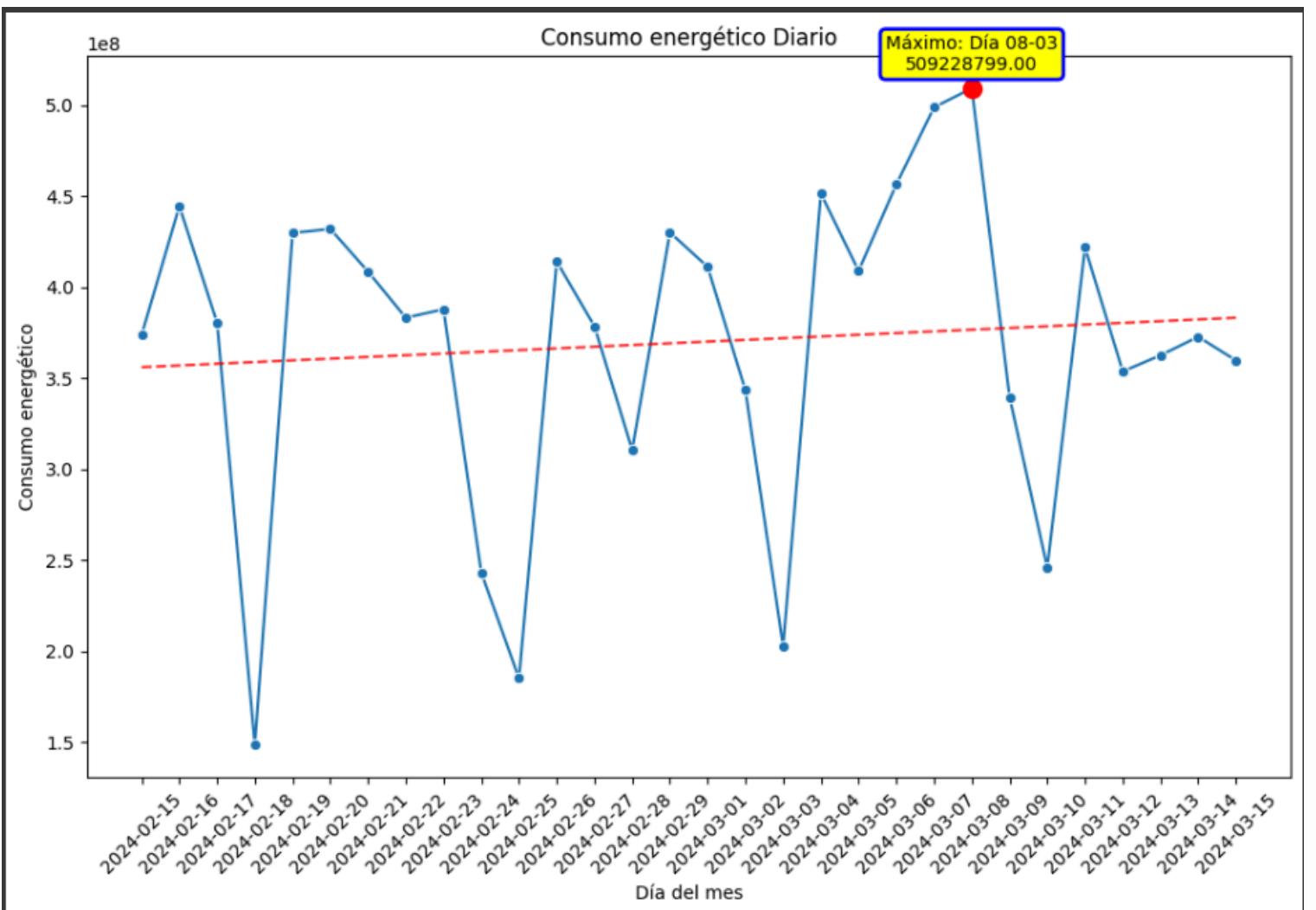
El día de la semana de mayor consumo energético fue el viernes, esto fue porque viernes y jueves son los únicos días de la semana que tienen una frecuencia de 5 días entre las fechas 15 de febrero a 15 de marzo. Quitando el sesgo o utilizando la media el claro ganador es el Lunes.

Registros por Tren



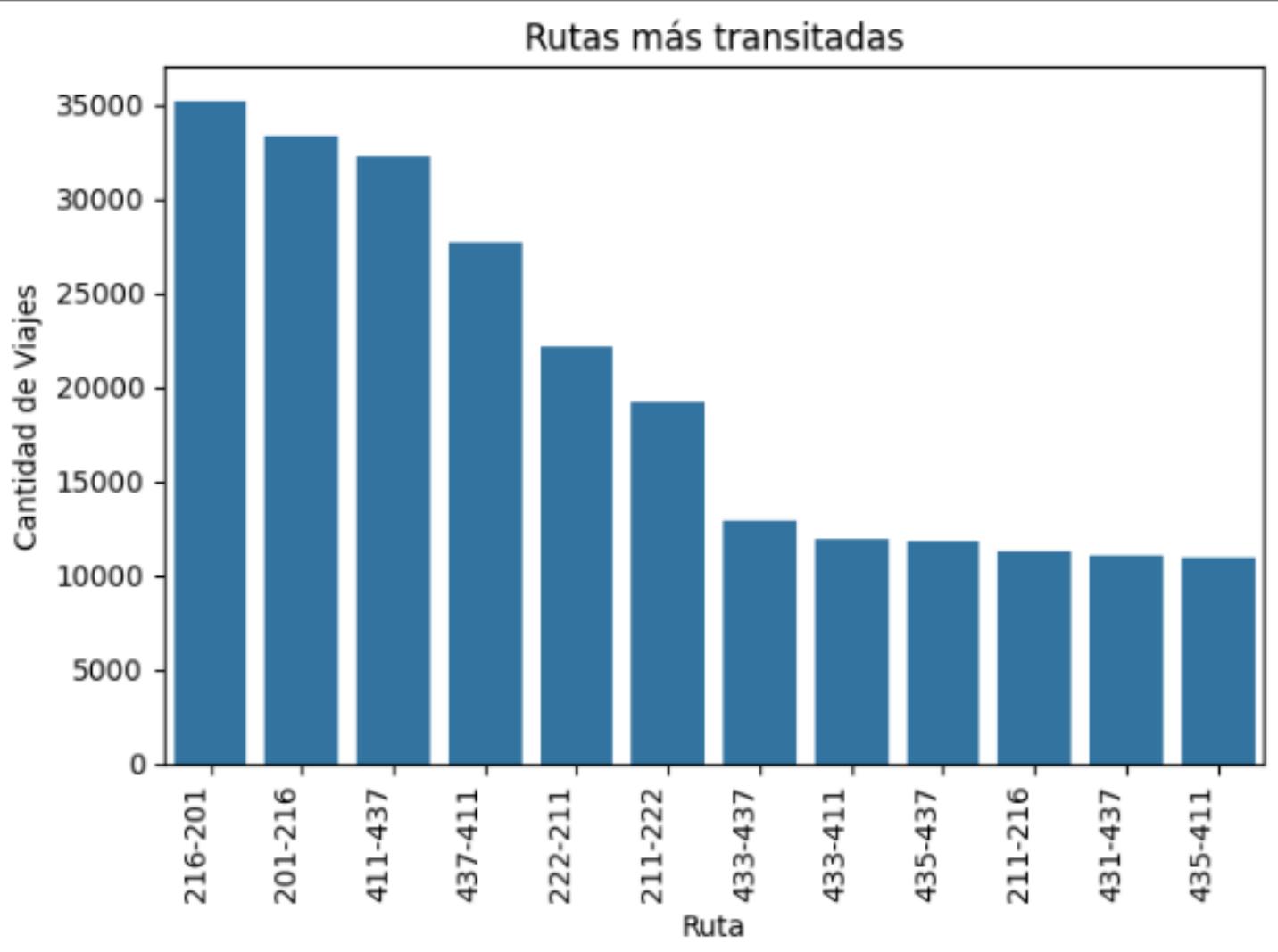
El tren con mejor atributo de conectividad, es el tren N°4, ya que tuvo registros 28 días del mes y acumuló más de 61,000 registros, con una diferencia de casi 10,000 registros al tren N°10

Día de mayor consumo Energético



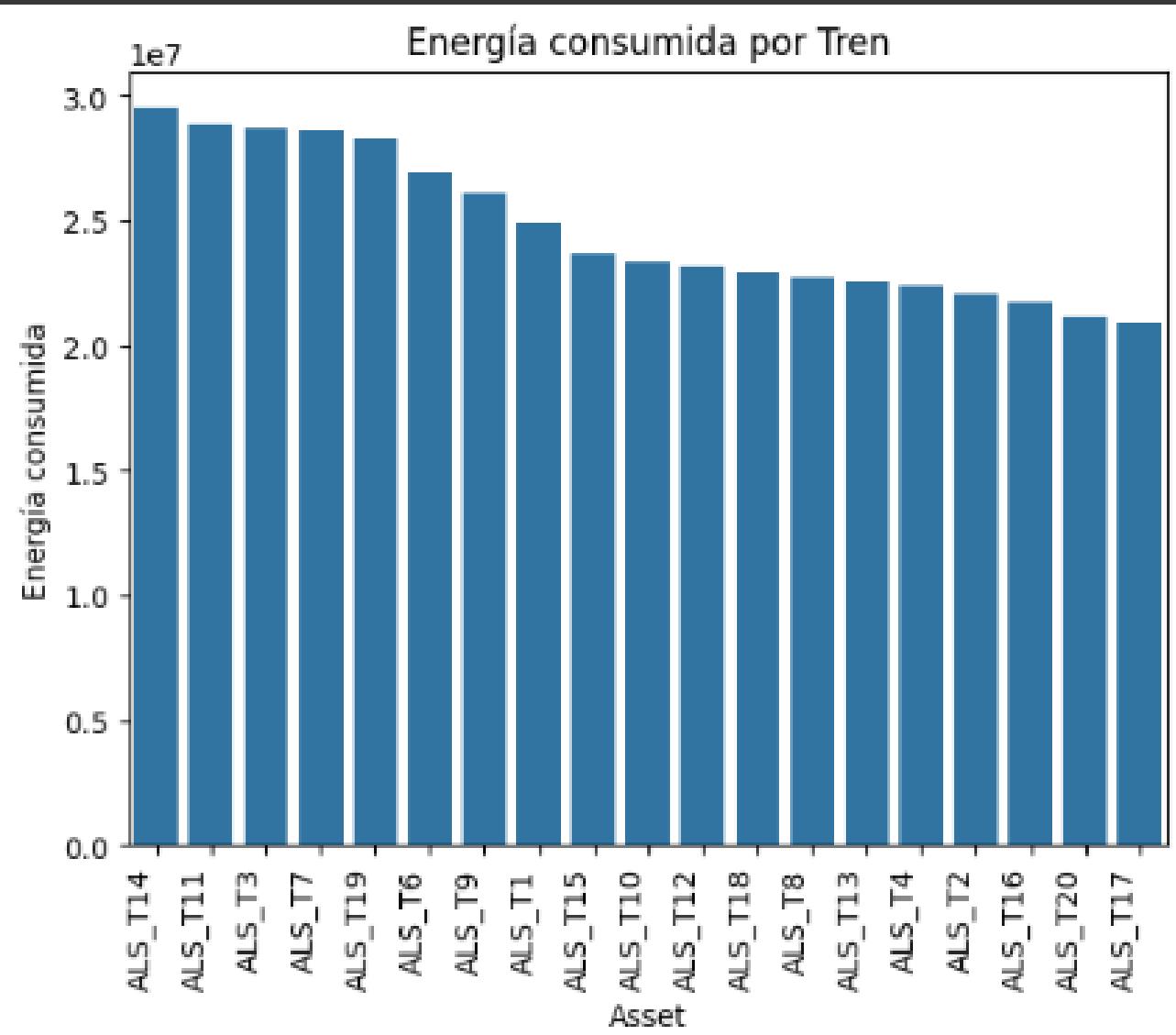
El 8 de marzo fue el día con mayor consumo energético, y se observa que el día anterior también registró un pico elevado en el consumo. Dado que estas fechas coinciden con la celebración del Día Internacional de la Mujer, surge la pregunta de si podría existir alguna relación entre ambos eventos.

Rutas más transitadas



Las rutas más transitadas se caracterizan por ser altamente recurrentes, un patrón que se mantiene a lo largo de todo el top. Es notable la marcada diferencia en el volumen de tránsito entre las cuatro rutas principales y las demás, lo que sugiere una concentración de tráfico en estas rutas clave.

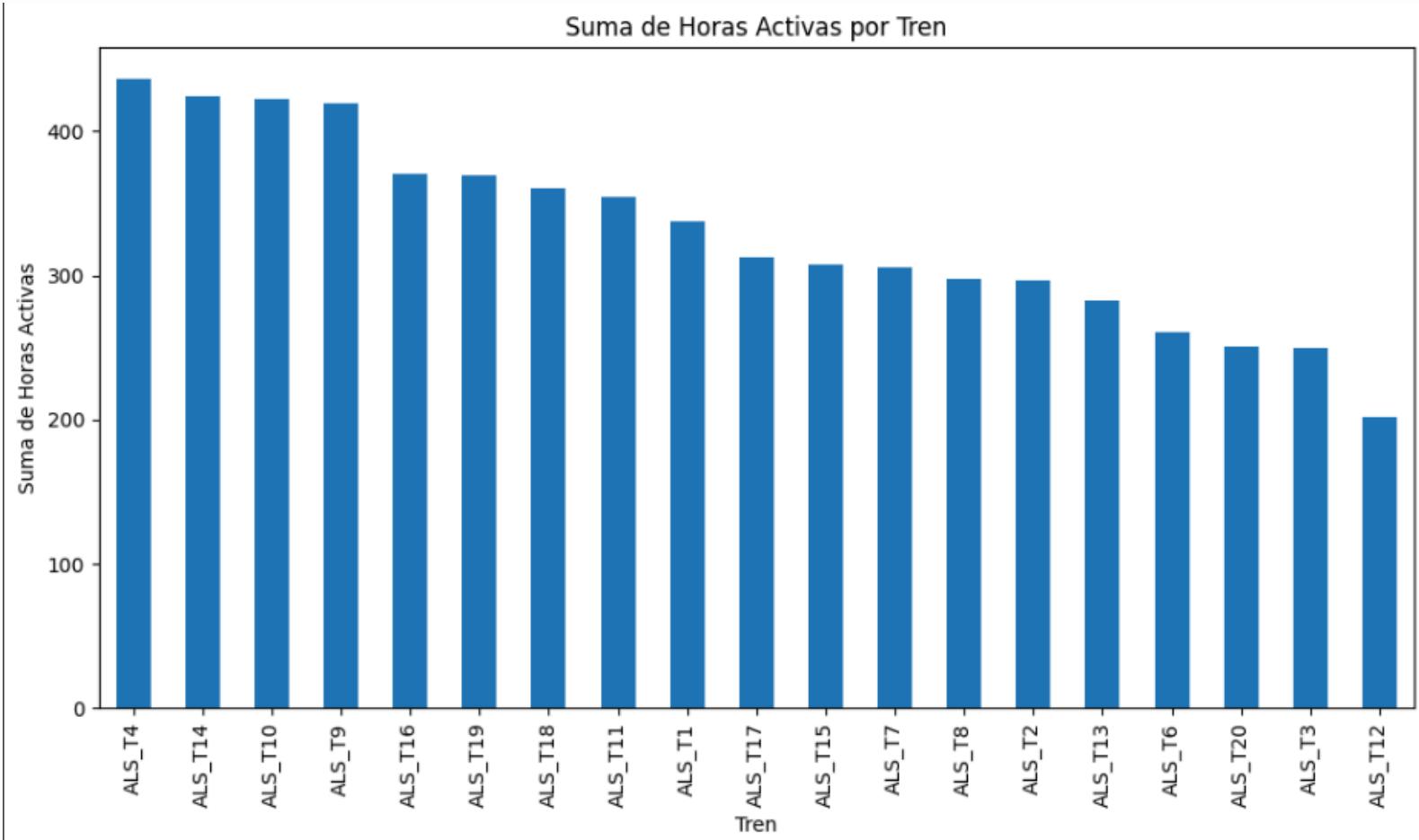
Tren más costoso energéticamente



Se utilizó la media para identificar cuál fue el tren con mayor consumo promedio. Podemos observar que los cuatro primeros trenes presentan valores similares entre sí, mientras que difieren significativamente de los trenes con menor consumo.

¿Existe una diferencia de medias significativa?

Tren más activo en el mes



El tren N°4 fue el más activo durante todo el mes, seguido de otros tres trenes cuyos valores de actividad fueron similares. Es interesante destacar que, aunque el tren N°3 fue uno de los menos activos, también resultó ser uno de los más costosos en términos de consumo energético

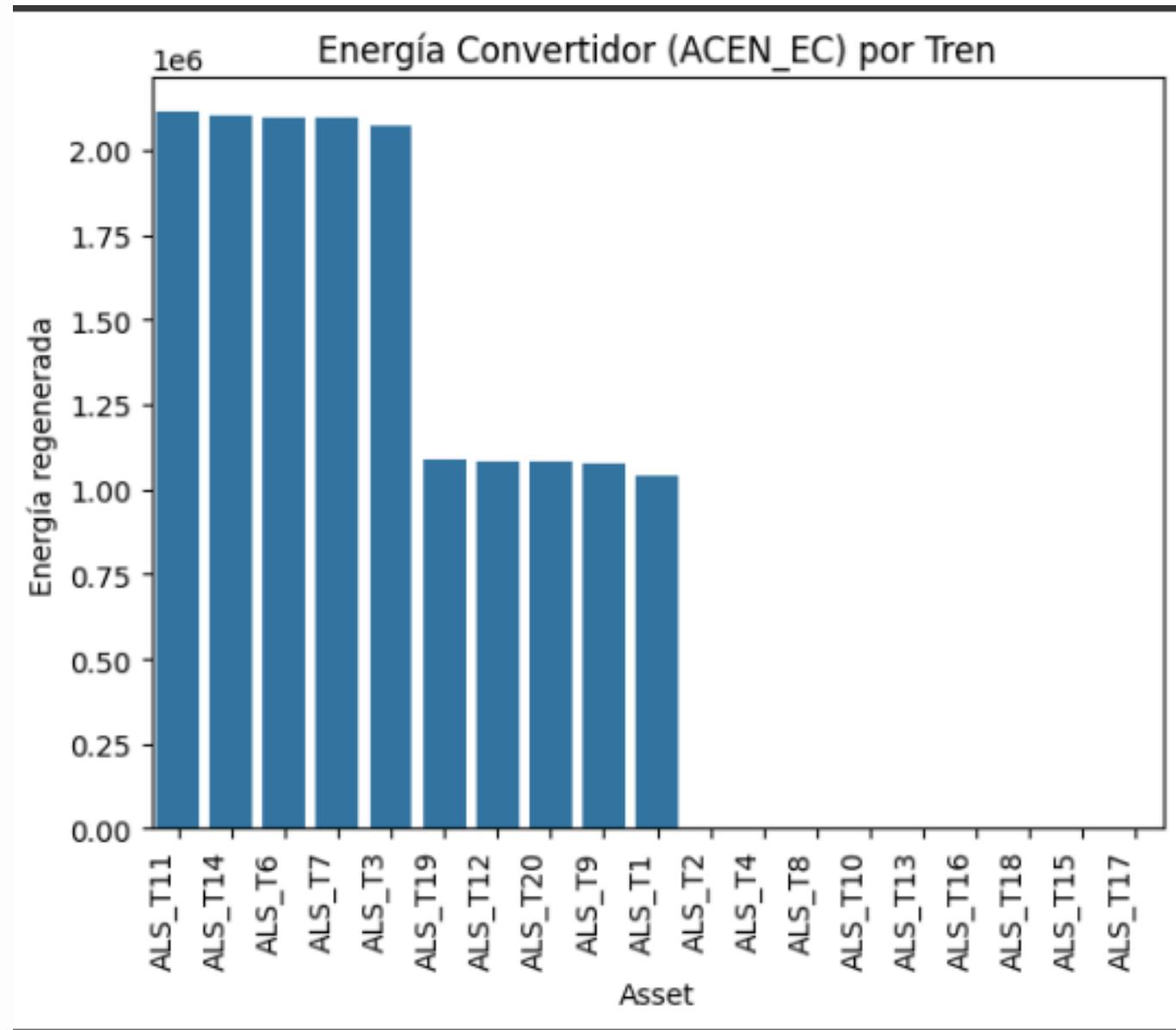
¿Existe una diferencia de medias significativa?

| tren_comparado | tren_base | media_tren_base | media_tren_comparado | diferencia_medias | estadistico_t | p_valor | significativo |
|----------------|-----------|-----------------|----------------------|-------------------|---------------|---------|---------------|
| ALS_T1 | ALS_T6 | 2.236224e+07 | 2.130468e+07 | -1.057565e+06 | 2.4677 | 0.0139 | True |
| ALS_T10 | ALS_T6 | 2.236224e+07 | 1.731150e+07 | -5.050736e+06 | 10.5891 | 0.0000 | True |
| ALS_T11 | ALS_T6 | 2.236224e+07 | 2.385361e+07 | 1.491372e+06 | -2.1784 | 0.0298 | True |
| ALS_T12 | ALS_T6 | 2.236224e+07 | 2.013005e+07 | -2.232196e+06 | 4.4355 | 0.0000 | True |
| ALS_T13 | ALS_T6 | 2.236224e+07 | 1.770485e+07 | -4.657396e+06 | 7.0305 | 0.0000 | True |
| ALS_T14 | ALS_T6 | 2.236224e+07 | 2.363993e+07 | 1.277687e+06 | -2.3797 | 0.0176 | True |
| ALS_T15 | ALS_T6 | 2.236224e+07 | 1.756772e+07 | -4.794517e+06 | 8.1739 | 0.0000 | True |
| ALS_T16 | ALS_T6 | 2.236224e+07 | 1.486240e+07 | -7.499843e+06 | 14.4265 | 0.0000 | True |
| ALS_T17 | ALS_T6 | 2.236224e+07 | 1.616655e+07 | -6.195691e+06 | 13.6376 | 0.0000 | True |
| ALS_T18 | ALS_T6 | 2.236224e+07 | 1.722155e+07 | -5.140690e+06 | 9.5561 | 0.0000 | True |
| ALS_T19 | ALS_T6 | 2.236224e+07 | 2.348961e+07 | 1.127368e+06 | -2.3002 | 0.0218 | True |
| ALS_T2 | ALS_T6 | 2.236224e+07 | 1.696013e+07 | -5.402113e+06 | 11.3991 | 0.0000 | True |
| ALS_T20 | ALS_T6 | 2.236224e+07 | 1.796397e+07 | -4.398267e+06 | 7.6688 | 0.0000 | True |
| ALS_T3 | ALS_T6 | 2.236224e+07 | 2.167546e+07 | -6.867768e+05 | 0.8714 | 0.3841 | False |
| ALS_T4 | ALS_T6 | 2.236224e+07 | 1.631441e+07 | -6.047827e+06 | 13.1173 | 0.0000 | True |
| ALS_T7 | ALS_T6 | 2.236224e+07 | 2.274519e+07 | 3.829524e+05 | -0.7124 | 0.4765 | False |
| ALS_T8 | ALS_T6 | 2.236224e+07 | 1.710670e+07 | -5.255539e+06 | 10.9548 | 0.0000 | True |
| ALS_T9 | ALS_T6 | 2.236224e+07 | 2.054966e+07 | -1.812579e+06 | 3.7733 | 0.0002 | True |

Resultado

Se utilizó el tren N°6, seleccionado de manera aleatoria, para aplicar un T-test que comparara las medias del consumo energético con las de cada uno de los otros trenes. Los resultados mostraron diferencias significativas en el consumo energético entre el tren N°6 y todos los demás trenes, a excepción del tren N°3 y el tren N°7.

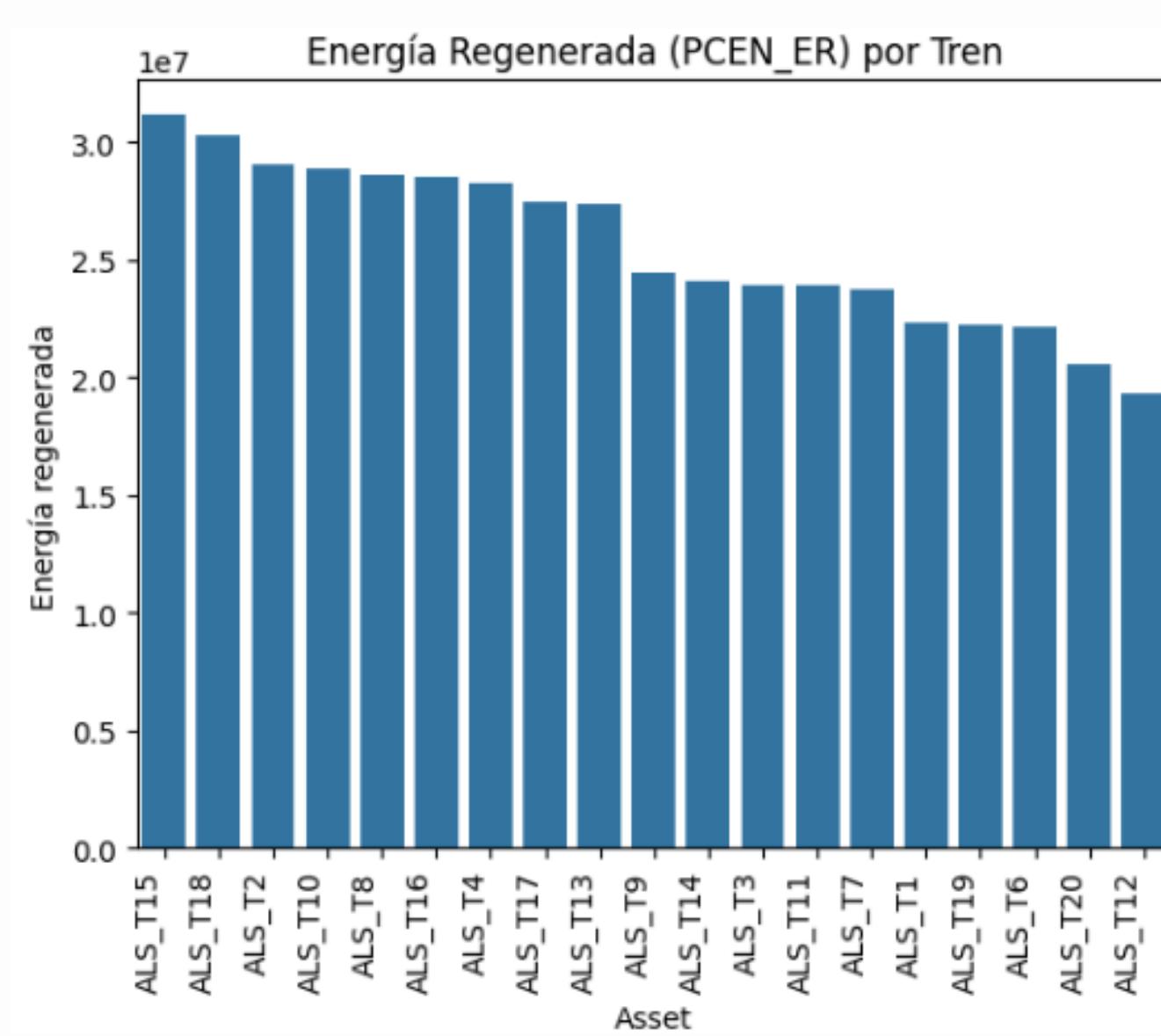
Energía Consumida por los Convertidores



Una gran cantidad de trenes presenta valores notablemente bajos en los convertidores. No se trata de valores nulos, sino de una escala considerablemente menor en comparación con otros trenes.

¿Qué rutas son las visitadas por los trenes con valores bajos?

Energía Consumida por los Regeneradores



El tren N°15 fue el más destacado en cuanto a regeneración de energía. Sería interesante analizar sus rutas para investigar si existe alguna relación entre los trayectos recorridos y el consumo total de energía.

Recomendaciones

1. Incorporación de información adicional de los trenes: Para mejorar el análisis, se recomienda agregar datos sobre condiciones ambientales y mantenimiento de los trenes. Esta información ayudaría a explicar los picos y caídas observados en los distintos componentes y trenes.
2. Revisión del modo de conducción del tren: Dado que se encontraron diferencias significativas en los valores desconocidos cuando el tren opera en modo automático, se sugiere revisar el software de las computadoras a bordo. Esto permitiría identificar posibles fallos de comunicación, problemas de compatibilidad del software o posibles errores de memoria que hacen que haya una pérdida de información.
3. Relleno de Rutas Desconocidas: Propongo las siguientes estrategias para abordar el problema de los valores desconocidos en las rutas:
 - I. Implementar un algoritmo de Sliding Window, de modo que cada vez que un registro presente una entrada o salida con valor 0, se complete la información utilizando los registros vecinos más cercanos.
 - II. Reemplazar los valores 0 en las entradas con NaN y luego aplicar un algoritmo de imputación basado en cadenas de Markov o K-Nearest Neighbors (KNN) para estimar los valores correctos.
 - III. Crear una nueva variable que identifique el inicio y el final de un viaje, es decir, el momento en que el tren cambia de ATC_CS a ATC_DS. Con esta información, calcular la diferencia de tiempo entre los registros inicial y final, estimando la duración promedio de la ruta. Los registros faltantes dentro de ese intervalo podrían imputarse utilizando estos valores promedio.

Recomendaciones

Modelos Predictivos: En caso de intentar predecir el consumo energético al día de la flota recomiendo lo siguiente:

1. Ampliar las fechas, recopilar la información de 1 o 2 años sobre un tren en específico para ser capaz de capturar las tendencias y estacionalidades de los consumos energéticos de cada tren.
2. Enfocar el modelo en un solo tren con intervalos de tiempo constantes, además de definir el intervalo de tiempo en el cual se quiere hacer la predicción del consumo de la flota (1 mes, 1 día, etc...).
3. Empezar por un modelo AR (Autoregresivo), luego con un MA (medias móviles), después un ARIMA (AutoRegresivo, Diferencial de Medias Móviles) y por último una red neuronal basada en la arquitectura de una LSTM (Long-Short Term Memory). Luego hacer un model selection mediante pruebas de AIC (Akaike Information Creteria) o RMSE (Root Mean Squared Error) para ver que tan bien es capaz de capturar la variabilidad de la estacionalidad y los patrones de la data.