**Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial**

**Práctica profesionalizante II**

**Modelo de predicción del consumo de energía eléctrica en un shopping de Córdoba**

**Docente:** Narciso Pérez

**Equipo 3:**

Karina Alem

María Gimena Binaghi

Florencia Bravo Corvalán

Viviana Farabollini

Oscar Ferreira

Laura Martínez Quijano

Emmanuel Reynoso

## **Tabla de contenido**

* + **[Sección I — Introducción](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "h3Introduction)**
  + **[Sección II —](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "h3Related) Antecedentes**
  + **[Sección III — Metodología](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "h3Methodology)**
  + **[Secciones IV — Experimentos y Análisis](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "h3Experiments)**
  + **[Sección V — Conclusiones y trabajo futuro](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "h3Conclusions)**

### **[Sección I — Introducción](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "TOC)**

Hoy en día, el consumo de energía en las organizaciones constituye entre el 10% y el 50% de sus costos operativos. Este consumo de energía se suele percibir como algo ineludible, como algo necesario para la realización de las tareas que son la misión de la organización. Sin embargo, el consumo de energía puede ser optimizado si se comprende cuáles son las variables que lo afectan y cómo lo hacen.

Edificios con instalaciones complejas como los shoppings son grandes consumidores de energía eléctrica. Muchos hacen esfuerzos para reducir este consumo, incorporando tecnologías más eficientes, pero se hace difícil comparar situaciones antes y después por la gran cantidad de variables que participan. El modelado de una línea de base que permita incluir las variables de temperatura exterior y público asistente es necesario para normalizar los datos en función de estas variables y entender si se está efectivamente reduciendo el consumo de energía para las mismas condiciones, antes y después de la implementación de mejoras. Este trabajo es relevante para todos los shoppings y edificios donde la temperatura exterior y la cantidad de público sean variables que afecten a sus consumos de energía.

El desafío de comprender de qué depende el consumo de energía resulta interesante porque hay variables que se pueden controlar y de este modo controlar el consumo de energía. Las consecuencias de este control del consumo permiten al usuario reducir sus costos debidos al consumo de energía. Adicionalmente, la optimización del consumo permite beneficios adicionales a la sociedad toda: menores emisiones de gases de efecto invernadero a la atmósfera por menor consumo, mejor aprovechamiento de la capacidad instalada de generación eléctrica, menor consumo de combustibles fósiles.

Los algoritmos más utilizados en los últimos cinco años para resolver este problema son aquellos de regresión lineal, tanto múltiple como logarítmica, soporte vector y gradiente ascendente; ocasionalmente también se utilizan redes neuronales. Encontramos un ejemplo de aplicación con series de tiempo que nos pareció más interesante para resolver el problema que teníamos entre manos, y desarrollamos además un modelo de regresión lineal múltiple para comparar los resultados.

Nuestra motivación para abordar y dar una posible respuesta a este problema es poner a disposición un modelo que pueda predecir el consumo en un caso real y específico. Contamos con siete personas en el equipo, con diferentes y complementarias competencias que nos permiten abordar este problema en forma multidisciplinaria.

El estudio alcanzará a un shopping de Córdoba, con posibilidad de extender los resultados a otros shoppings y edificios complejos de Córdoba y del país.

Como consecuencia del estudio, edificios complejos podrán comparar su consumo de energía en cualquier momento en situaciones comparables, y determinar si están reduciendo efectivamente su consumo.

El aporte de este dato es relevante para organizaciones que quieren demostrar una mejora de su desempeño ambiental y energético, además de demostrar la reducción de sus emisiones de gases de efecto invernadero para cumplir con compromisos de lucha contra el cambio climático.

Se sabe que el consumo de energía depende de muchas variables, algunas más relevantes que otras. En el caso de edificios complejos, la climatización constituye 50-60% del consumo total, por lo que la temperatura exterior y la afluencia de público aparecen como las variables más relevantes.

Lo nuevo que se aportaría es cómo y cuánto las variables identificadas pueden predecir el consumo, en primer lugar. En segundo lugar, si el consumo en un determinado momento/día es el esperado o menor (indicando mayor eficiencia) o mayor (indicando oportunidades de mejora para la reducción del consumo).

El enfoque que adoptamos es el de identificar en un primer momento a las variables de temperatura exterior y afluencia de público como las variables más relevantes que determinan el consumo de energía. En conversaciones con el cliente, identificamos que la afluencia de público varía según el día de la semana, y según la categoría de hábil o feriado de ese día, impactando así en el consumo de energía del shopping. Decidimos entonces ampliar el rango de variables para incluir al día de la semana, y al tipo de día (feriado o no).

### **[Sección II — Trabajo relacionado](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "TOC)**

* Predicción (forecasting) de la demanda eléctrica con Python by Joaquín Amat Rodrigo and Javier Escobar Ortiz, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/py29-forecasting-demanda-energia-electrica-python.html

Multiple linear regression, logarithmic multiple linear regression methods, and nonlinear autoregressive with exogenous input artificial neural networks <https://www.researchgate.net/publication/344604260_Machine_Learning_Modeling_for_Energy_Consumption_of_Residential_and_Commercial_Sectors>

* Linear Regression and Support Vector Regression

<https://ieeexplore.ieee.org/abstract/document/8769508>

A. González-Briones, G. Hernández, J. M. Corchado, S. Omatu and M. S. Mohamad, "Machine Learning Models for Electricity Consumption Forecasting: A Review," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019, pp. 1-6, doi: 10.1109/CAIS.2019.8769508.

* Our results show that gradient boosting regression models perform the best at predicting commercial building energy consumption, and can make predictions that are on average within a factor of 2 from the true energy consumption values (with an r2 score of 0.82).

<https://www.sciencedirect.com/science/article/abs/pii/S0306261917313429>

* The results show that using the gradient boosting machine model improved the R‐squared prediction accuracy and the CV(RMSE) in more than 80 percent of the cases, when compared to an industry best practice model that is based on piecewise linear regression, and to a random forest algorithm.

https://www.sciencedirect.com/science/article/abs/pii/S0378778817320844

### **[Sección III — Metodología](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "TOC)**

Hoy en día muchas organizaciones asumen compromisos de reducción del consumo de energía. Como el consumo de energía suele depender de muchas variables, es difícil comparar dos situaciones y entender si el consumo efectivamente se redujo o si aumentó por motivos esperables (por ejemplo, el consumo de energía en climatización en un mes otoñal será habitualmente más bajo que el de un mes con temperaturas más extremas). Esta complejidad puede abordarse con un modelo que involucre a todas las variables exógenas, permitiendo así predecir el consumo en un momento futuro incorporando a todas ellas en la predicción. La comparación del consumo predicho con el consumo de energía realmente consumido permitirá concluir si se está consumiendo mejor o peor, con las consecuencias mencionadas sobre los costos y las emisiones de gases de efecto invernadero.

Con el objetivo final de construir una serie temporal, utilizamos un modelo de predicción autoregresivo recursivo, con optimización de hiperparámetros y forecasting con variables exógenas.

Para lograrlo, seguimos los siguientes pasos:

* obtuvimos del cliente los datos de consumo de energía de los últimos 12 meses, con intervalos de 15 minutos
* pedimos al SMN los datos de temperatura exterior para el mismo período, que nos llegaron con intervalos horarios
* recibimos del cliente los datos de afluencia de público para los últimos 10 meses, los cuales eran diarios

Con esta información nos dedicamos de lleno a la limpieza de datos, encontrando lo siguiente:

* faltaban algunos datos puntuales de consumo de energía, decidimos reemplazar a los valores faltantes con el promedio de los datos anterior y posterior
* en algunos días particulares, faltaban datos durante muchas horas: decidimos construir los valores faltantes con el promedio de consumo de ese día

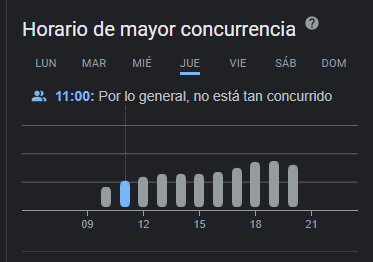
Una vez resuelta la falta de datos, nos encontramos con que teníamos datos de energía cada 15 minutos, datos de temperatura por hora, y datos de público por día.

Decidimos llevar todo a datos por hora, para lo cual:

* agrupamos los datos de consumo de energía cada 15 minutos para construir un solo dato promedio por hora
* hicimos un webscrapping durante toda una semana en octubre de 2022 para entender la distribución de la afluencia de público durante cada día de esa semana (incluyendo un lunes feriado), para luego distribuir el dato diario de público en un dato por hora, siguiendo la metodología que se describe a continuación.

# **Scrapping para público**

Necesitábamos obtener la cantidad de público que el negocio tenía por hora y por día, para analizar cuánta injerencia tenía este dato en el consumo. El negocio no nos brindó esos datos entonces optamos por traer los datos directamente del servicio de “**Google Mi Negocio**” que es la ficha que aparece a la derecha a la hora de buscar algún término de búsqueda, en este caso el negocio. En uno de los zócalos tenemos la concurrencia del público, por día y por hora.

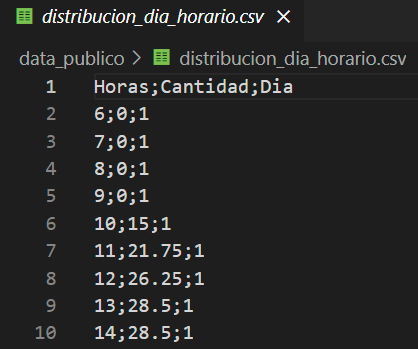


Buscamos si alguien había hecho algo similar y encontramos una persona que lo obtuvo pero usaba librerías que ya no están disponibles. Entonces nos tocó hacer el script desde cero.

Como librerías usamos **Selenium** para el Web Scraping y **Pandas** para el manejo de los datos.

En una primera instancia lo creamos buscando que la funcionalidad haga lo que necesitábamos y luego optimizarla para que funcione lo más rápido posible, con la menor cantidad de líneas de código. En el proceso de ver las mejores opciones para crearlo, probamos varias librerías que luego depuramos.

Finalmente el script abre Chrome, busca el negocio en el buscador de Google y busca el zócalo de la concurrencia del público. Al llegar a esa fecha va al primer día, recorre las horas de ese día para tomar el porcentaje del público y luego pasar al siguiente día. Cuando tenemos todos los datos en el **DataFrame**, exportamos en un **CSV** para cruzarlo con los demás datos.



Ahora ya contábamos con los datos necesarios (reducidos a 10 meses por la limitación del dato de público), pero nos dimos cuenta de que la afluencia del público estaba directamente relacionada con el día de la semana y con la categoría de hábil o feriado. Decidimos entonces incorporar estas variables al dataset.

Paralelamente a la realización de este trabajo, hicimos una búsqueda bibliográfica para identificar posibles modelos para predecir el consumo de energía eléctrica, y encontramos que una serie temporal parecía ajustarse bien a nuestro problema.

Una serie temporal es una sucesión de datos ordenados cronológicamente, espaciados a intervalos iguales o desiguales. El proceso consiste en predecir el valor futuro de una serie temporal, modelando la serie únicamente en función de su comportamiento pasado (autorregresivo) empleando otras variables externas.

Para validar el modelo y testearlo, no podíamos dividir al dataset en el clásico 70-30 train-test ni tampoco en 70-20-10 train-test-validation, porque al tratarse de datos con fuerte correlación con la época del año en que ocurren, si entrenábamos al modelo en invierno no iba a dar buenos resultados prediciendo el consumo en verano (ni viceversa). Por eso decidimos utilizar la metodología cross validation para la validación del modelo.

Por otra parte, para comparar los resultados de la serie de tiempo con otro algoritmo, elegimos una red neuronal que por falta de datos derivó en una regresión lineal múltiple.

### **[Secciones IV y V — Experimentos y Análisis](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "TOC)**

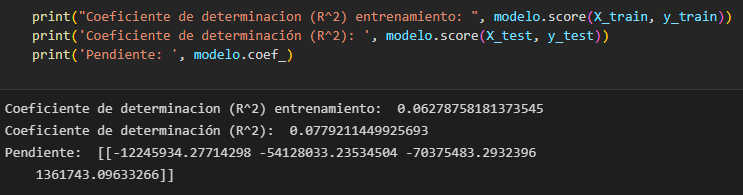
# **Experimento 1**: Red neuronal que derivó en Regresión lineal

Como alternativa a la serie de tiempo intentamos hacer una red neuronal, esta la creamos usando la librería **TensorFlow - Keras**. Probamos varias veces pero en las pruebas siempre teníamos malos resultados, eso se debió a que teníamos pocos datos para entrenar el modelo.

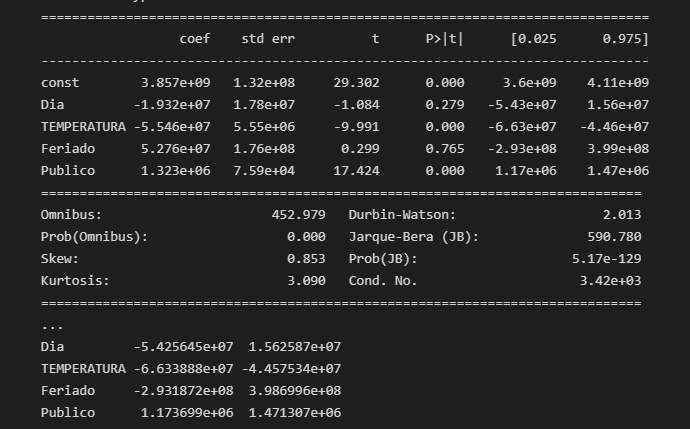
Surgió la idea de buscar un modelo ya entrenado pero por los tiempos pensamos continuar con una regresión lineal. Cuando lo construimos, en el **EDA** obtuvimos la siguiente matriz de correlaciones:

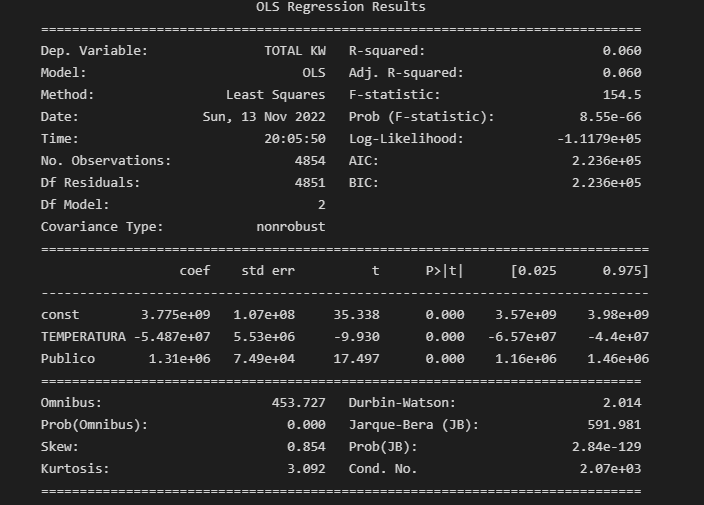


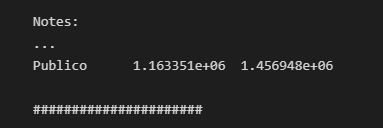
Ya nos podíamos imaginar el resultado ya que, después de entrenar el modelo, obtuvimos el siguiente resultado:

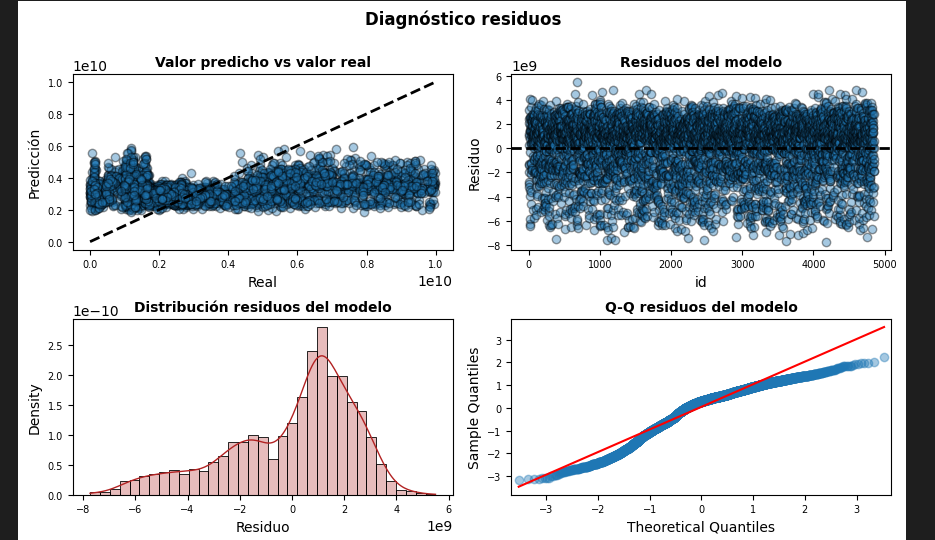


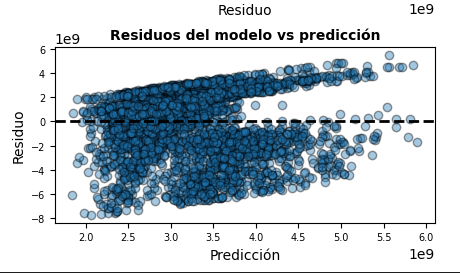
Se intentó quitar columnas (Feriado y Día) para intentar lograr cambios pero tampoco se logró modelar satisfactoriamente.











Estos resultados no permiten predecir con una precisión aceptable los consumos futuros: el coeficiente de correlación ajustado (r2 ajustado) nos dio 0.06

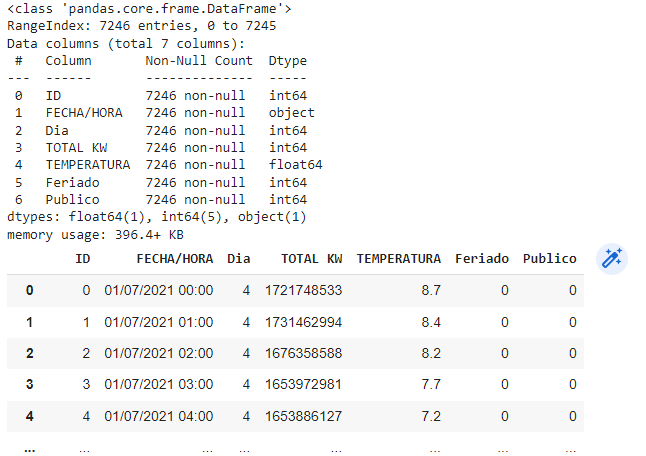
Con estos resultados, avanzamos a continuación con el modelo de serie de tiempo.

**Experimento 2: Serie temporal auto-recursiva regresiva**

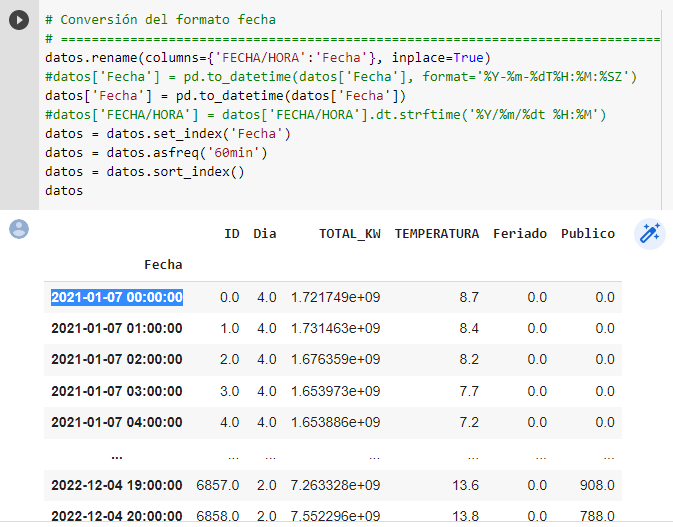
El estudio de las series de tiempo nos demandó el esfuerzo de entender como son tratados los datos referidos a esta variable.

En el transcurso del desarrollo del trabajo nos encontramos que si dejamos que Python reclasifique los datos se produce un desfasaje de fechas, lo cual provocaba que el entrenamiento del modelo resultase en errores, ya que su interpretación de las fechas provocaba espacios en blanco del modelo, con lo cual los entrenamientos fallaban.

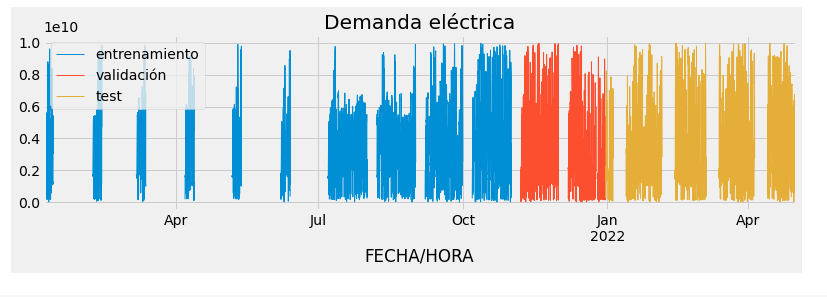
Los datos en crudo muestran lo siguiente:



Intentamos convertir las fechas sin tratamiento y nos encontramos con lo siguiente:



Lo que nos daba como resultado lo siguiente:



Estos espacios en blanco, provocaron un error del modelado de predicciones en la serie de tiempo, como puede verse a continuación:



Esto se resuelve con un pequeño cambio en el modelo de datos, y un tratamiento correcto de las fechas en el set de datos:

ID;Fecha;Dia;Demanda;Temperatura;Feriado;Publico;

0;2021/07/01 00:00;4;1721748533;8.70;0;0;

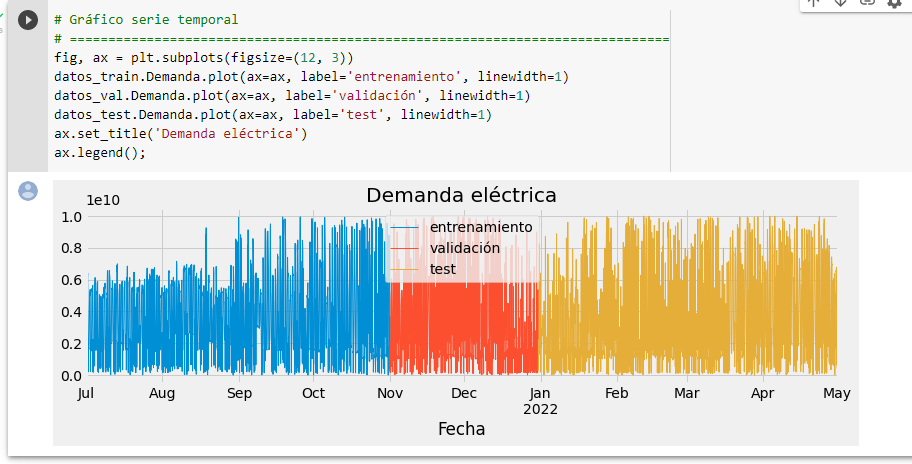
1;2021/07/01 01:00;4;1731462994;8.40;0;0;

2;2021/07/01 02:00;4;1676358588;8.20;0;0;

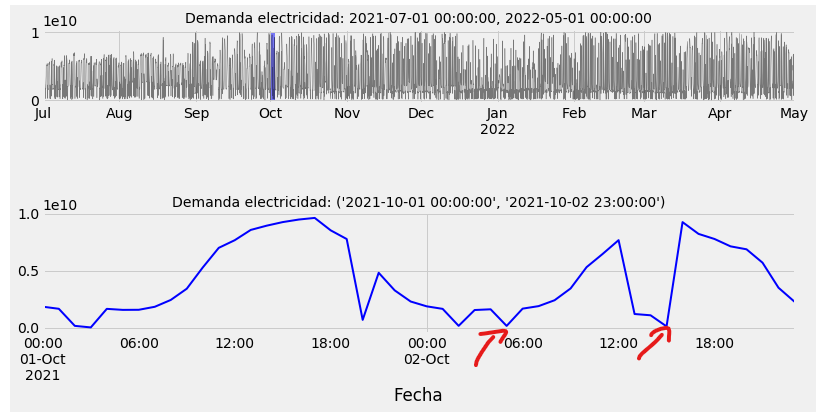
3;2021/07/01 03:00;4;1653972981;7.70;0;0;

4;2021/07/01 04:00;4;1653886127;7.20;0;0;

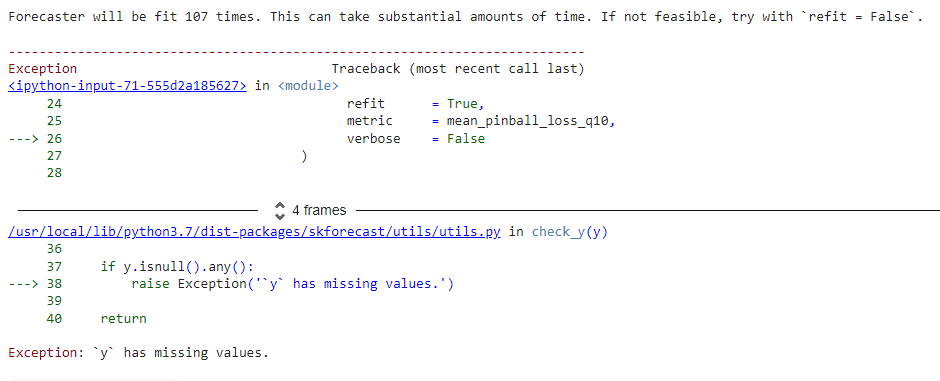
Lo que nos permitió limpiar los datos, como puede verse en la siguiente figura:



Al momento de iniciar los modelos predictivos, nos encontramos con problemas relacionados con la continuidad de los datos, como puede verse en el set de datos más abajo, donde algunos valores caen abruptamente.



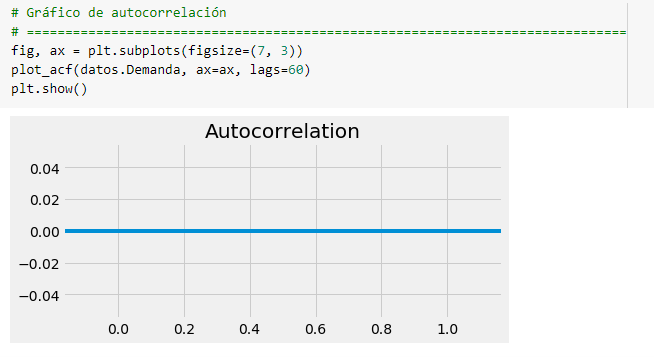
Estos valles nos producen problemas en los modelos de Predicción (backtest) al generar (paradójicamente) valores faltantes.



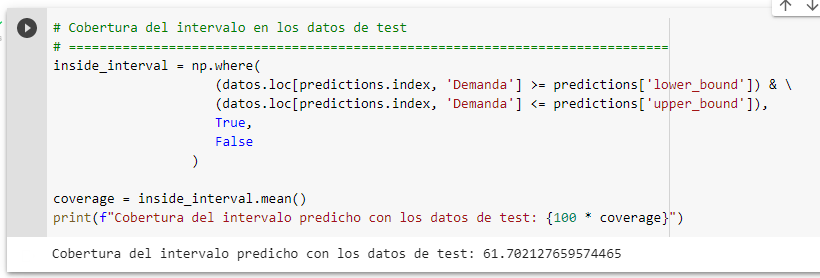
Investigando el tema, se llega a la conclusión de que la escasez de datos produce este tipo de errores, ya que la característica de los modelos forecast es completar los espacios vacíos con valores nulos, sobre todo cuando llega a los extremos de la serie de datos.

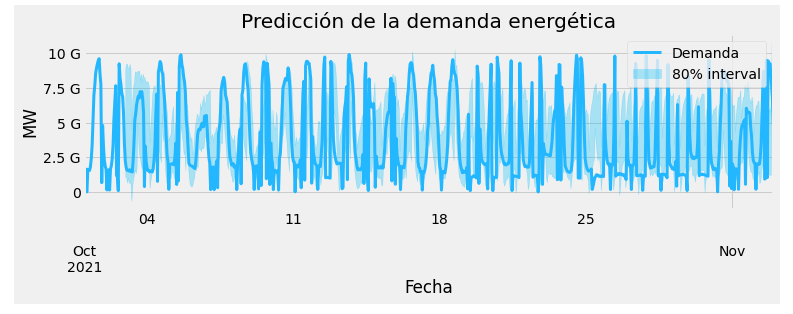
Como conclusión, podemos decir que si bien el modelo teórico funciona, se debe tener un gran trabajo en la preparación de los datos y en el volumen del mismo, ya que gran parte del éxito de la predicción radica en obtener la mayor cantidad y calidad de datos y su preparación.

En nuestro caso, no llegamos a un resultado positivo,por diversos problemas, pero el mayor de todos se dio en que se disponía SOLO de 10 meses de datos y eso es un volumen muy escaso para realizar predicciones del tipo forward – back forecasting, en particular con nuestra distribución de datos que dado su volumen no alcanzó a generar correlaciones apreciables, lo que fue un primer indicio de que el volumen de datos no alcanzaba a impactar en el modelo de predicción, como puede verse a continuación.



Con los datos existentes y dado que el periodo de entrenamiento es muy pequeño se ha logrado un factor de cobertura de 61.7%





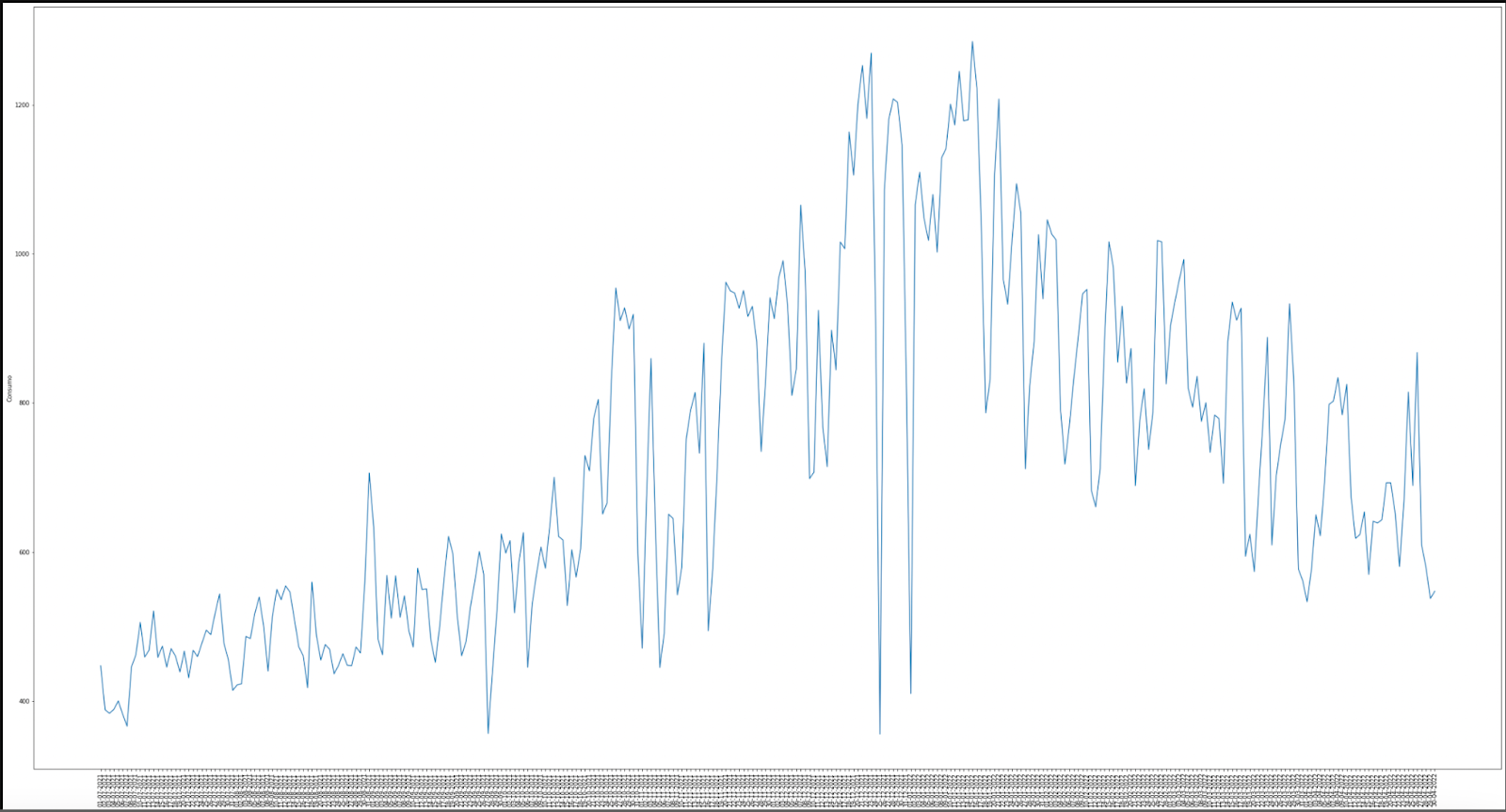
Antes de adentrarnos en un mayor análisis de los resultados, descubrimos que no conocíamos a fondo a nuestros datos, por lo que iniciamos (al final del trabajo!) un EDA (análisis exploratorio de datos) para entenderlos mejor.

**Experimento 3: Análisis exploratorio de datos**

En vista de que los experimentos anteriores no estaban llegando a buenas métricas, se toma un set de datos “crudos”. Sin prácticamente limpieza de datos se encuentran dos hallazgos importantes en el EDA para validar con el cliente:

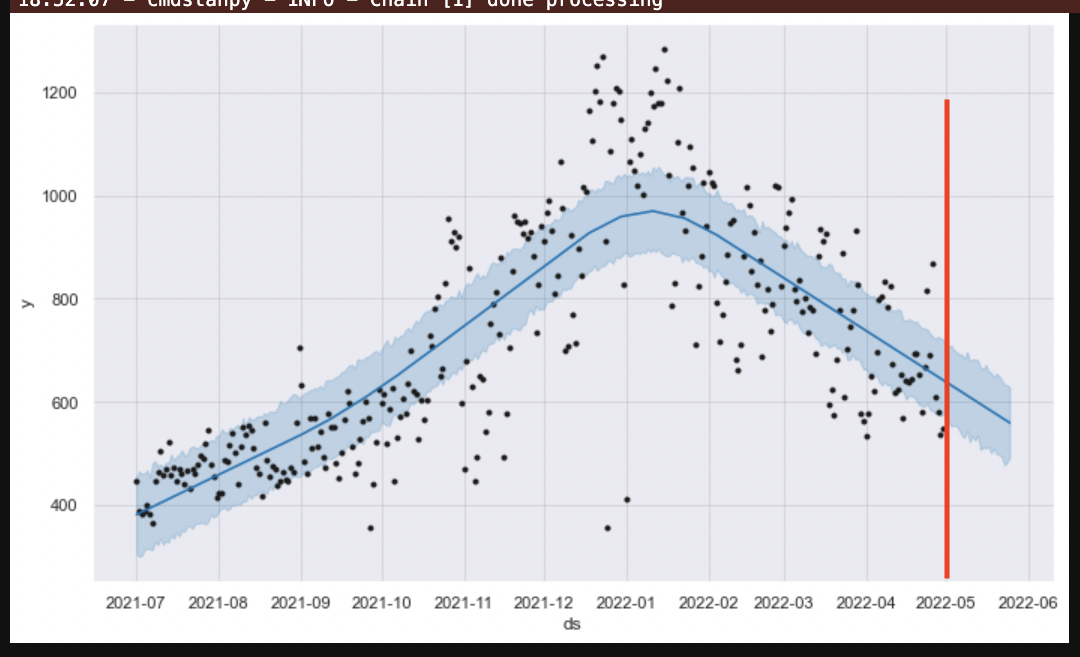
* Hay picos a finales de diciembre y enero. Y se nota el cambio de consumo en 2021 con 2022. En la distribución por día los días que más se consumen no necesariamente son los días de los que más datos hay. Se esperaba que los días con más consumo fueran el fin de semana. ##hallazgo1
* Hay mediciones muy bajas en fechas esperables. Navidad | Año nuevo. Pero qué pasó el 27/09/21, 31/10/21, 26/01/22, 211/02/22. ##hallazgo2

La serie se ve de la siguiente manera

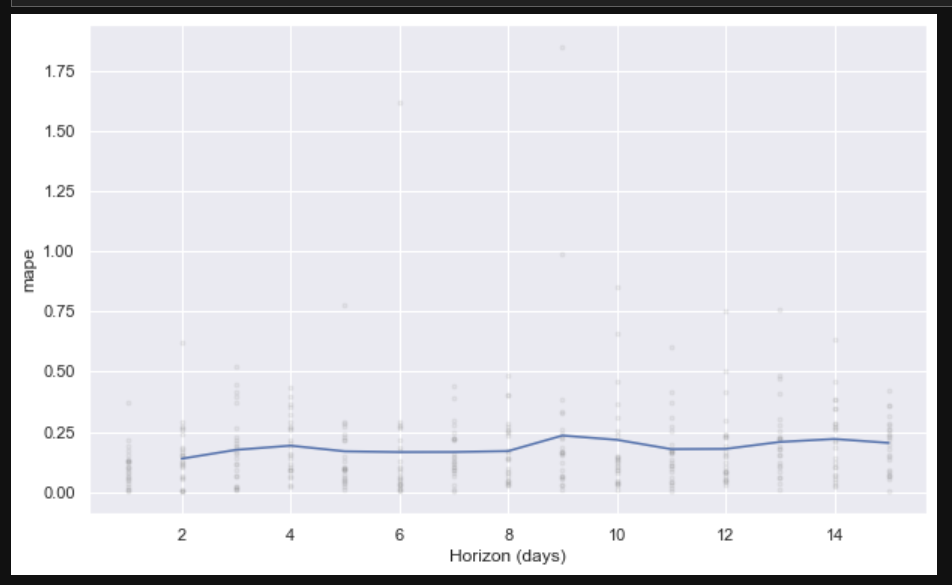


Para modelar se utiliza la librería de Prophet

Se realiza de un primer entrenamiento sin hiperparámetros y con un forecast de 25 días se consigue lo siguiente:



Luego de usar crossvalidation y tomar MAPE como métrica y conseguimos lo siguiente para los siguientes días:



Se recomienda continuar con una fuerte limpieza de datos y probar con más variables. Dado que solo se utilizó el consumo. Y realizar un EDA mucho más profundo.

### **[Sección V — Conclusiones y trabajo futuro](https://pyimagesearch.com/2022/04/13/planning-and-writing-a-research-paper/" \l "TOC)**

Queremos predecir el consumo de energía de un shopping en la ciudad de Córdoba. Creímos que con conseguir los datos (consumo de energía pasado, y variables relevantes tales como temperatura y afluencia de público) y con identificar un modelo de predicción (serie de tiempo auto-recursiva regresiva) sería suficiente.

Sin embargo, conseguir los datos no fue tan simple como creimos en un primer momento, particularmente porque la granulometría de cada uno era distinta (consumo de energía cada cuarto de hora, público por día), lo que nos obligó a manipularlos fuertemente para poder ingresarlos a un modelo de predicción. Aplicar el modelo que encontramos a nuestro set de datos tampoco fue tan sencillo: el modelo funcionaba con un set de cinco años y nosotros teníamos diez meses.

Logramos superar estas dificultades, y sólo después de reconocer que los resultados estaban lejos de ser alentadores (r2=0.06 para la regresión lineal múltiple), nos dimos cuenta de que teníamos que hacer un análisis exploratorio de los datos: primero teníamos que verlos para entenderlos, para luego hacer una buena limpieza de outliers y datos irrelevantes (consumo en días con el shopping cerrado, consumo en días con el aire acondicionado averiado, etc).

Este trabajo pretende presentar el camino recorrido, las dificultades que encontramos y los posibles caminos a seguir para construir un modelo válido para este ejemplo y para generalizarlo a cualquier usuario de energía que quiera comprender y predecir su consumo.