Predicción del uso del servicio bicicletas compartidas de la ciudad de Barcelona: Bicing

Francisco Javier Regalado López

Introducción

Los vehículos de combustión interna son una fuente importante de contaminación atmosférica en las zonas urbanas, para reducir su impacto los responsables de las tomas de decisiones en los gobiernos han intentado limitar el uso del automóvil en años recientes. Las restricciones impuestas vienen con el fomento de otras alternativas de transporte entre ellas los sistemas de bicicleta compartidas en la que las bicicletas están disponibles para el uso compartido de particulares durante un corto periodo de tiempo a bajo coste.

Bicing: El sistema de bicicletas compartido de la ciudad de Barcelona

El sistema de bicing (Bicing, 2022) se implanto en el 2007, promovido por el ayuntamiento de la ciudad de Barcelona con el objetivo de cubrir pequeños y medianos recorridos diarios dentro de la ciudad. El sistema bicing es compuesto por 6.000 bicicletas mecánicas y 300 eléctricas que cuenta con estaciones integradas al sistema de transporte público de la ciudad (Ajuntament de Barcelona, 2018).

El ayuntamiento de Barcelona en su portal de Open Data BCN ha anunciado que los datos relacionados con el sistema bicing dejarán de ser publicados debido a un cambio en la gestión del servicio (Ajuntament de Barcelona, 2018), dejando dos datasets del servicio para su acceso.

La intención de este proyecto es realizar una predicción de la demanda del servicio bicing en sus dos modalidades, mecánica y eléctrica con la información publicada.

Estado del arte

La previsión consiste en predecir el futuro buscando la mayor precisión posible, considerando cualquier información importante que pueda afectar a la misma, puede haber previsiones a corto, mediano y largo plazo. Cuando se tiene información cuantitativa histórica puede utilizarse técnicas de series temporales donde se previene el futuro mediante patrones pasados que pueden repetirse en el futuro (Hyndman & Athanasopoulos, 2022).

Cuando solamente se utilizan los valores históricos de la serie temporal para predecir sus valores futuros, se denomina Previsión de Series Temporales Univariantes y si se utilizan predictores distintos de la serie (es decir, variables exógenas) para prevenir, se denomina Previsión de Series Temporales Multivariantes. Las series temporales pueden tener una tendencia al mostrar que existe un aumento o una disminución a largo plazo de los datos; temporada al ser afectada por factores estacionales; cíclica al existir fluctuaciones que no tienen una frecuencia fija (Peixeiro, 2022).

El modelo para previsión de series temporales en sus siglas en ingles ARIMA (Media Móvil Integrada Autorregresiva) se basa en los valores históricos. Para ARIMA(p,d,q) ; p es la orden de la parte autorregresiva, d es el grado de la primera diferenciación implicada y d es el grado de la primera diferenciación implicada (Peixeiro, 2022).

$$(1 - \phi_1 B - \cdots - \phi_1 B^p)(1 - B)^d y_t = c + \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t \quad \text{eq (1)}$$

El modelo SARIMA(p,d,q)(P,D,Q)m amplía el modelo ARIMA(p,d,q) añadiendo parámetros estacionales, el cual tiene cuatro nuevos parámetros en el modelo: P, D, Q tienen el mismo significado que en el modelo ARIMA(p,d,q), pero son sus homólogos estacionales, el parámetro m representa la frecuencia estacional.

Existe un tercer modelo que engloba los anteriores, SARIMAX amplía el modelo SARIMA(p,d,q)(P,D,Q)m añadiendo el efecto de las variables exógenas, ver eq (2), donde el segundo termino es la parte exógena:

$$y_t = SARIMAX(p,d,q)(P,D,Q)m + \sum_{i=1}^{n} \beta_i X_t^i \quad \text{eq (2)}$$

Metodología

En la figura se muestra la secuencia de trabajo que se realizo para este proyecto.

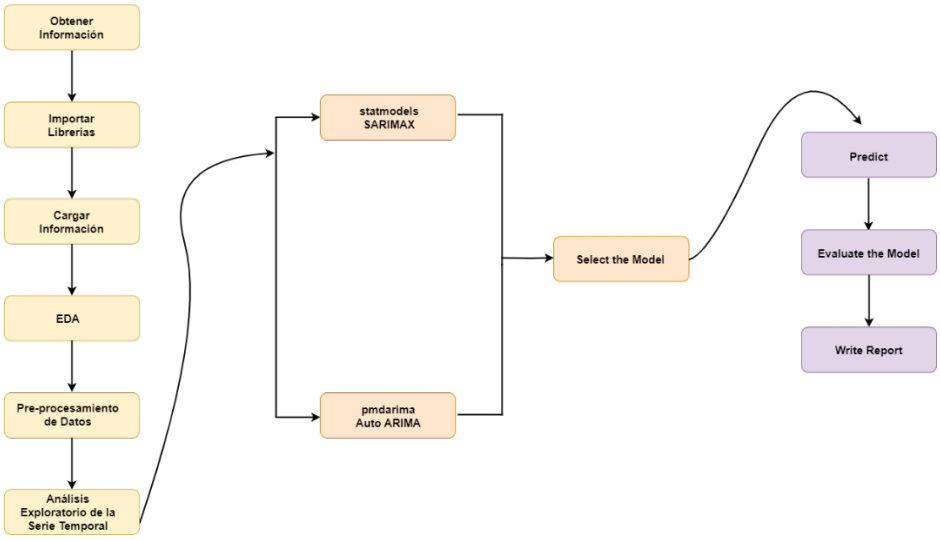

Fig. 1 Data Pipeline

Datos

El ayuntamiento de Barcelona al interrumpir la publicación de los datos del bicing proporcionó el equivalente de 6 meses de uso del servicio con algunas limitaciones y tampoco existe un glosario de términos pero que pueden inferirse, ver fig. (2). Los datos públicos mostrados en la Tabla 1 son:

Característica	Descripción
error	Error del servicio
bikesInUsage	Total de bicicletas en uso en el sistema
electricalBikesInUsage	Total de bicicletas eléctricas en uso en el sistema
mechanicalBikesInUsage	Total de bicicletas ordinarias en uso en el sistema
dateTime	Fecha y tiempo en el que los datos fueron recopilados.

Tabla 1 Descripción de datos.

Grid | Graph | Map | 8063 records | « | 7800 – 7899 | »

error	bikesInUsage	electricalBikesInUsage	mechanicalBikesInUsage	dateTime
0	14	1	13	2019-02-28 02:04:56
0	13	1	12	2019-02-28 02:09:57
0	12	1	11	2019-02-28 02:14:56
0	11	1	10	2019-02-28 02:19:57
0	9	0	9	2019-02-28 02:24:56
0	9	0	9	2019-02-28 02:29:56
0	9	0	9	2019-02-28 02:34:56
0	9	0	9	2019-02-28 02:39:57
0	9	0	9	2019-02-28 02:44:57
0	8	0	8	2019-02-28 02:49:56
0	8	0	8	2019-02-28 02:54:57
0	8	0	8	2019-02-28 02:59:56
0	8	0	8	2019-02-28 03:04:56
0	8	0	8	2019-02-28 03:09:56
0	8	0	8	2019-02-28 03:14:57

Fig. 2 Ejemplo de datos del uso del servicio Bicing.

Obtención de Datos

Los datos proporcionados podrían ser útiles para realizar una previsión univariante pero el Bicing es un modo movilidad activa, desplazarse por medios no motorizados o mediante la actividad física de la persona. El uso de la bicicleta puede ser afectada por lo tanto por los efectos climatológicos, considerando que Barcelona tiene un clima mediterráneo litoral. Los datos históricos en las fechas con una frecuencia por hora se obtuvieron de una base de datos de la NASA (Stackhouse, 2022): Temperatura, humedad especifica, precipitación y velocidad del viento.

Análisis Exploratorio de Datos (EDA)

Los datos del uso del sistema Bicing se analizarán según su modalidad, bicicletas eléctricas u ordinarias.

Bicicletas Ordinarias

La fig. (3) y fig (4) muestra que el uso de las bicicletas ordinarias tuvo un incremento cuando la temperatura se encontraba entre los 15° - 20° C, no había poca humedad en el aire y la velocidad del viento era bajo.

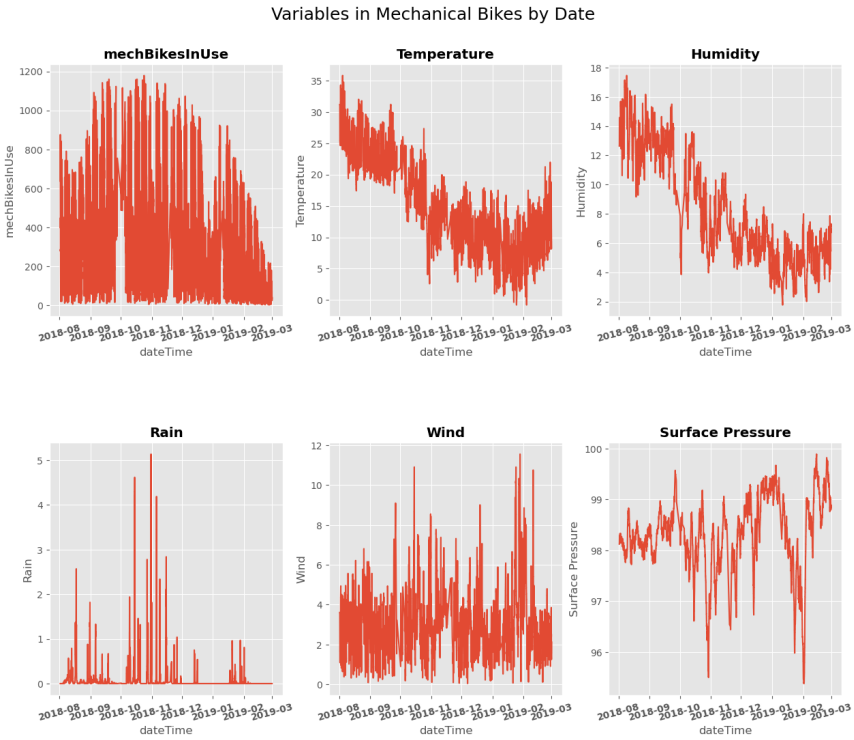

Fig. 3 Comportamiento del uso de bicicletas ordinarias de las características por fecha.

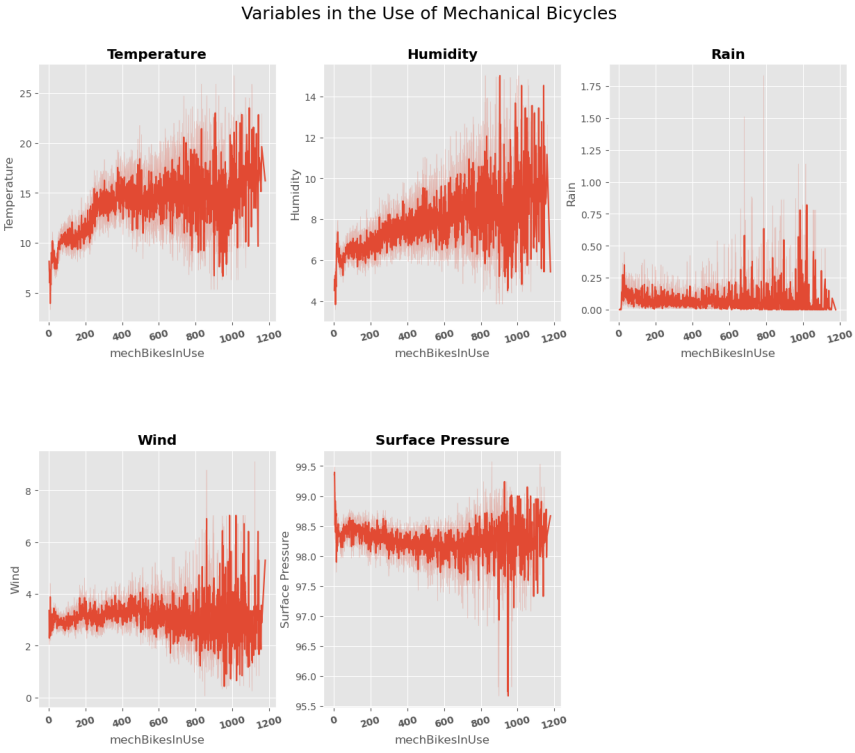

Fig. 4 Uso de las bicicletas ordinarias por características.

Bicicletas Eléctricas

La fig. (5) y fig. (6) muestra que el uso de las bicicletas eléctricas es afectado en menor medida que las bicicletas ordinarias por las condiciones atmosféricas, esto se aprecia mejor en el gráfico de correlaciones, ver fig. (7).

Fig. 5 Comportamiento del uso de bicicletas eléctricas de las características por fecha.

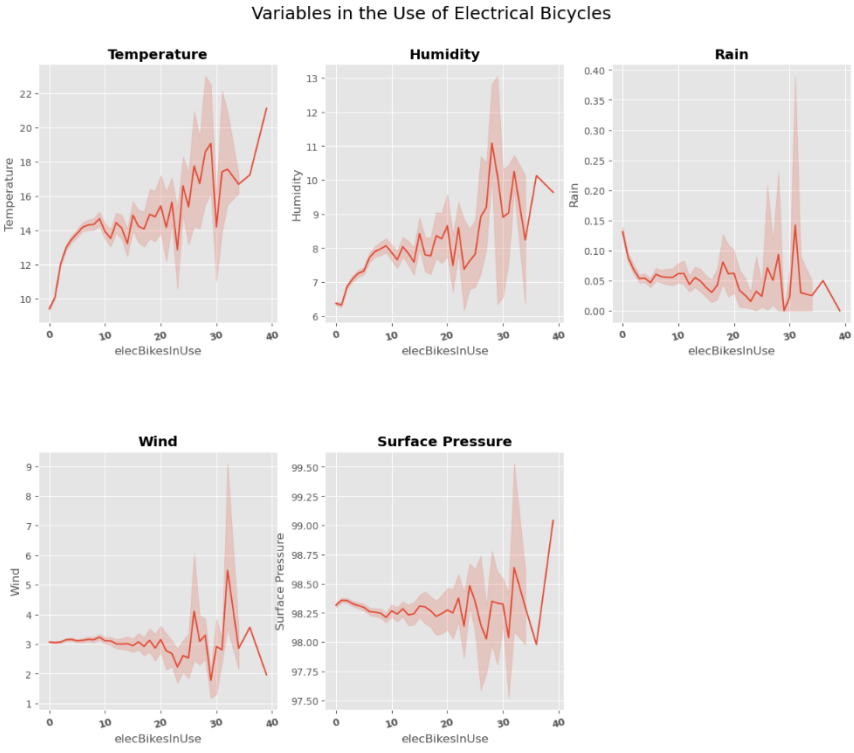

Fig. 6 Uso de las bicicletas eléctricas por características.

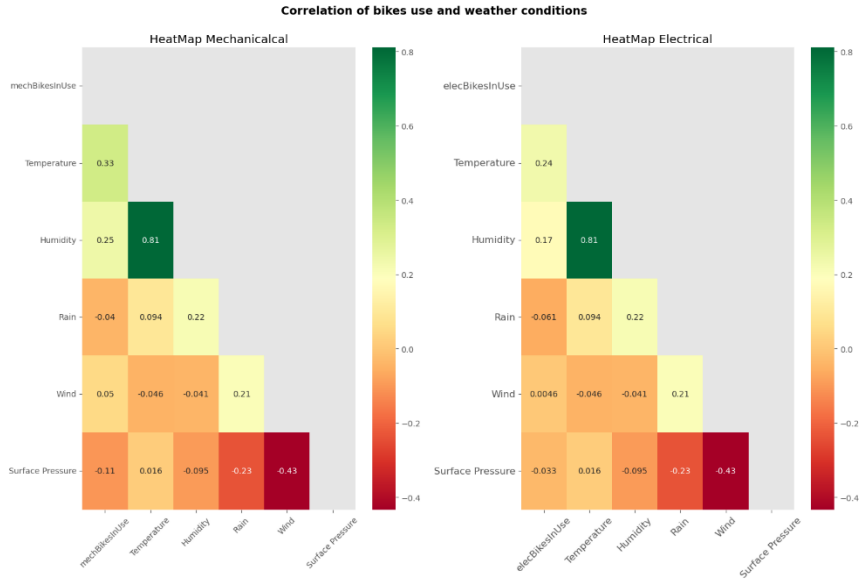

Fig. 7 Gráfico de correlación entre las diferentes características de los datos.

Pre-procesado de Datos

Dentro del pre-proceso de datos se realizo el agrupamiento del uso de las bicicletas por hora, los datos originales fueron recopilados entre un lapso de 3-5 min siguiendo el formato de YYYY:MM:DD HH:MM:SS lo que incrementaba el tamaño del dataset.

Unos datos "ocultos" dentro de la serie temporal es conocer si el día es laborable o fin de semana, también si el día es festivo.

Una particularidad de este dataset a pesar de no tener datos nulos, tenia una gran cantidad de datos con valor cero, lo especial viene en que son datos consecutivos. Además, con una gran variabilidad, se buscarán estos datos y se desecharán, esto con la conclusión que es una caída del sistema por eso hay cero bicicletas operativas o bien es un error en el sistema.

Análisis de la Serie Temporal

En primer lugar, se revisará que tipo de serie temporal se va analizar, esto se hará descomponiéndola en tres componentes; tendencia, estacionalidad y residuales. En la fig (8) y fig (9) muestran la descomposición estacional de la serie de bicicletas ordinarias y eléctricas, ambas series no muestran una tendencia, pero si estacionalidad y los residuales que muestran la oscilación de los valores alrededor de cero bastante uniforme al eliminar de la serie, la estacionalidad y tendencia, muestran algunos puntos de alta variación.

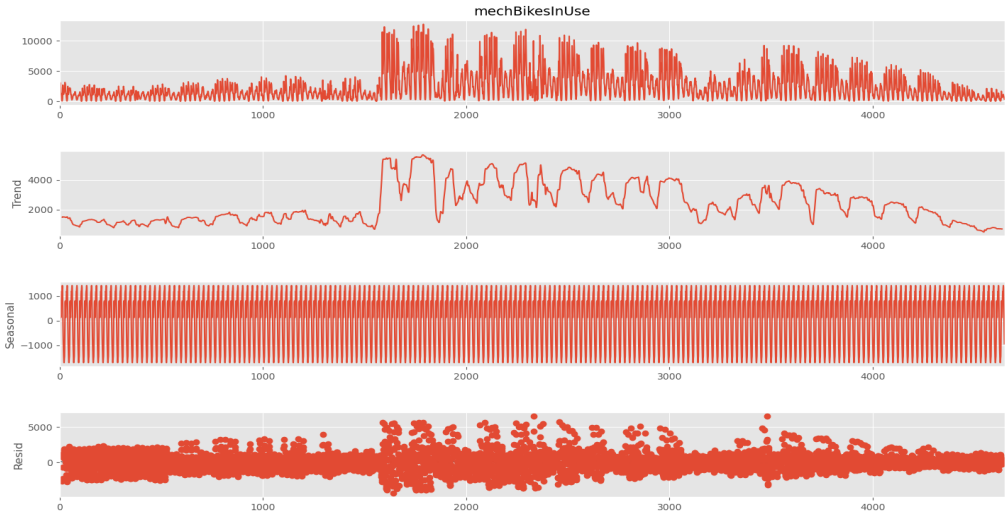

Fig. 8 Descomposición estacional bicicletas mecánicas.

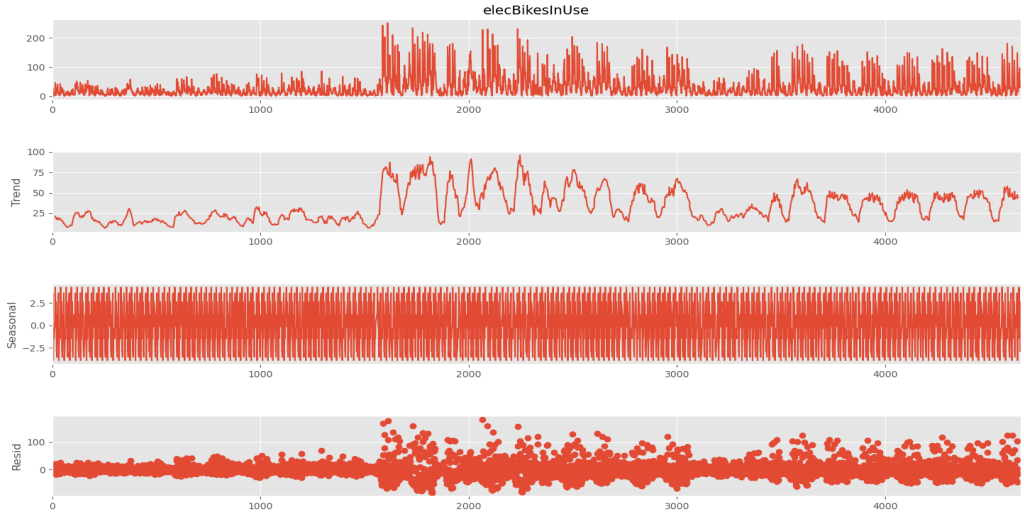

Fig. 9 Descomposición estacional bicicletas eléctricas

En segundo lugar, se debe de comprobar la comprobar si la serie es estacionaria utilizando la prueba de Dickey Fuller Aumentado, una serie temporal estacionaria es aquella cuyas propiedades estadísticas no cambian con el tiempo, muchos modelos de previsión asumen la estacionariedad y sólo pueden utilizarse si se comprueba que los datos son realmente estacionarios.

Los resultados para la serie de bicicletas ordinarias son:

- ADF : -4.9650
- p-value: $2.6023e^{-5}$

con un valor ADF negative y un p-value menor a .05 así que podemos decir que es estacionaria, se realizará un gráfico de autocorrelación, fig. (7) para

visualizarlo, a la izquierda tenemos la gráfica de autocorrelación no tiene el comportamiento esperado tal vez no sea del todo estacionaria, pero se comporta como una, esto se comprueba aplicando la diferenciación. En el gráfico de la derecha a pesar que se tiene un comportamiento esperado tiene muchos valores negativos, así que se desecha la idea de transformar la serie y se considera estacionaria como la estadística la describe.

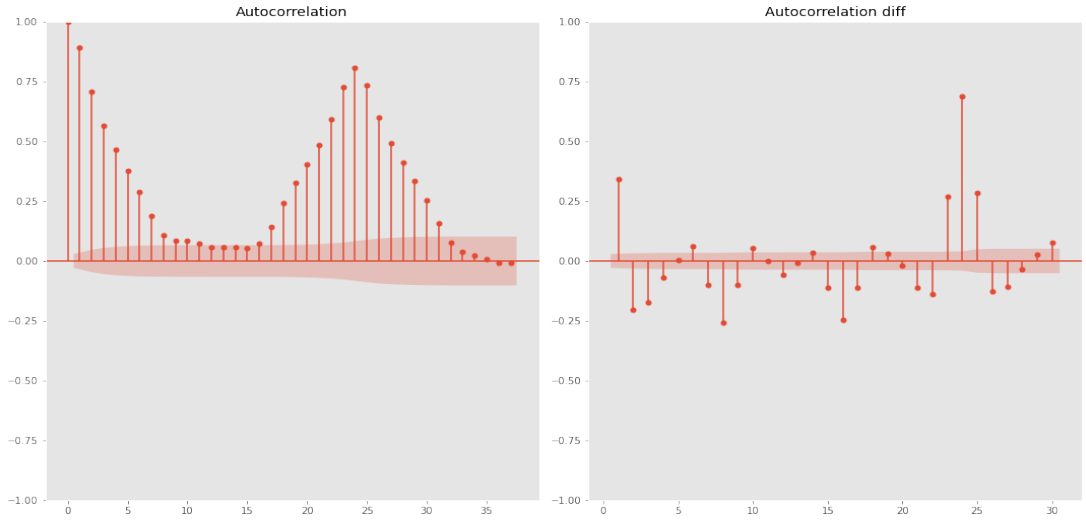

Fig. 10 a la izquierda un gráfico de autocorrelación y a la derecha una transformación diferencial de los datos de bicicletas ordinarias.

Para los datos de la bicicleta eléctrica se realiza la misma operación, los resultados para la serie de bicicletas eléctricas es:

- ADF: -6.6930
- p-value: $4.0628e^{-9}$

con estos resultados se puede decir que la serie temporal es estacionaria.

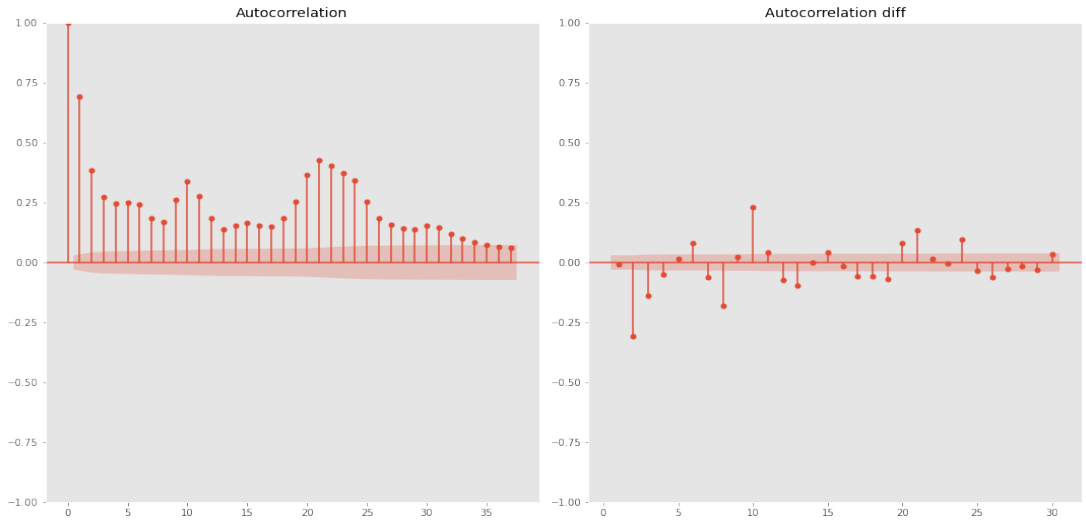

Fig. 11 a la izquierda un gráfico de autocorrelación y a la derecha una transformación diferencial de los datos de bicicletas ordinarias.

Modelos de Series Temporales

SARIMAX

Bicicletas ordinarias

Para encontrar los parámetros óptimos para el modelo SARIMAX se realizará un Grid Search, con los siguientes parámetros:

- p_mech = range(0, 3, 1)
- d_mech = range(0, 1, 1)
- q_mech = range(0, 3, 1)
- P_mech = range(0, 3, 1)
- D_mech = range(0,2,1)
- Q_mech = range(0, 3, 1)
- s_mech = 24

con los análisis anteriores se puede saber que d=0 y la frecuencia es de 24 al ser las horas por día, con esto se puede optimizar encontrando:

- p = 1
- q = 1
- d = 0
- P = 0
- Q = 1
- D = 1

con un valor mínimo de AIC 59124.71 y con los resultados mostrados en la fig. (12).

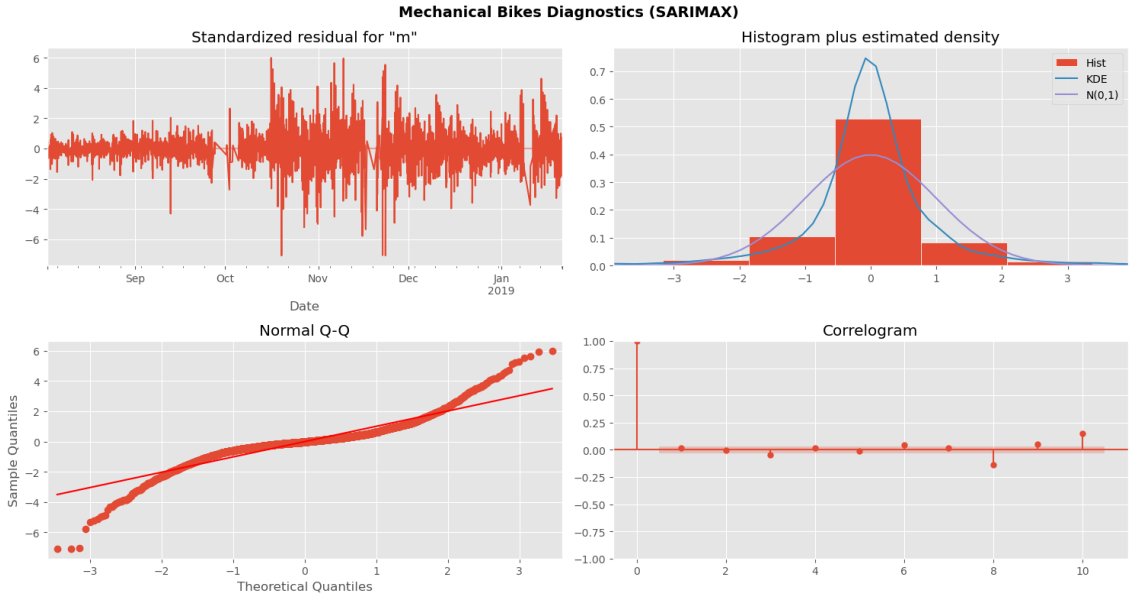

Fig. 12 Diagnostico de SARIMAX para bicicletas ordinarias.

Bicicletas eléctricas

Para las bicicletas eléctricas los parámetros para la búsqueda optima son:

- p_elec = range(0, 2, 1)
- d_elec = range(0,1,1)
- q_elec = range(0, 2, 1)
- P_elec = range(0, 2, 1)
- D_elec = range(0,2,1)
- Q_elec = range(0, 2, 1)
- s_elec = 24

los parámetros optimizados son:

- p = 1
- q = 1
- d = 0
- P = 1
- Q = 1
- D = 1

con un valor mínimo de AIC 11778.50 y con los resultados mostrados en la fig. (13).

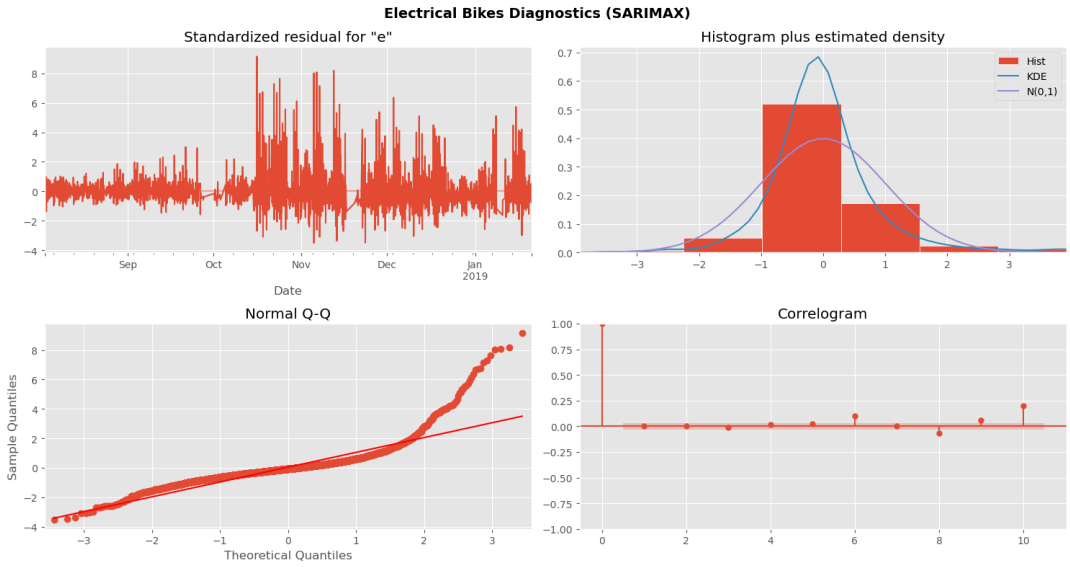

Fig. 13 Diagnostico de SARIMAX para bicicletas eléctricas.

Auto-ARIMA

Bicicletas ordinarias

La ventaja de utilizar la librería pmdarima es que no se tiene que escribir una función para optimizar los paráméros, quedando:

- p = 1
- q = 0
- d = 1
- P = 2
- Q = 1
- D = 1

El resultado no menciona el AIC mínimo solamente los parámetros, resultando como se muestra en la fig. (14):

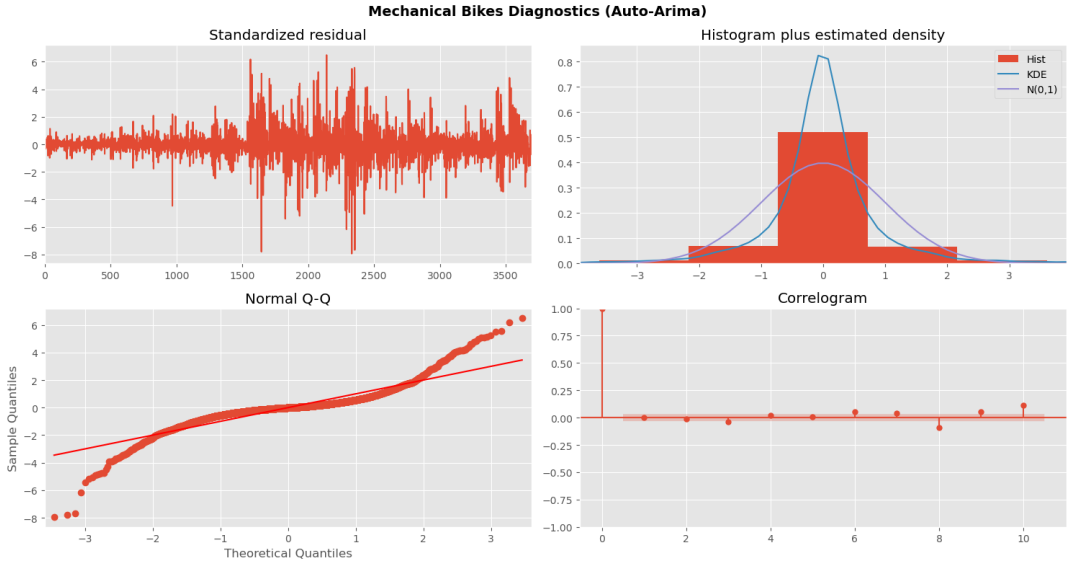

Fig. 14 Diagnostico de Auto-Arima para bicicletas ordinarias.

Bicicletas eléctricas

Los parámetros para las bicicletas eléctricas son:

- p = 2
- q = 1
- d = 0
- P = 2
- Q = 0
- D = 1

Los resultados del modelo se muestran en la fig (15).

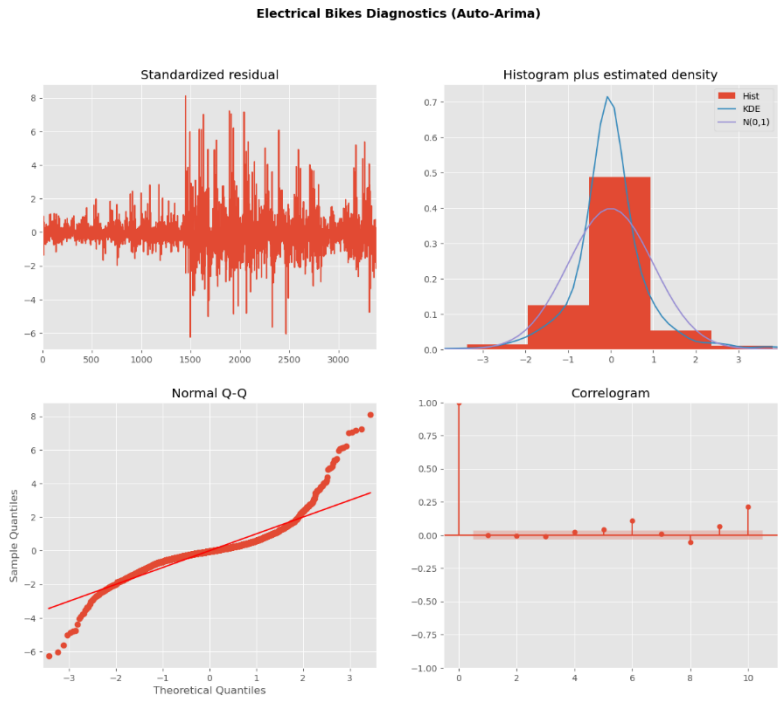

Fig. 15 Diagnostico de Auto-Arima para bicicletas eléctricas.

Predicciones

Bicicletas ordinarias

Los resultados entre ambos modelos son los mismos, realmente la diferencia es el gasto computacional, en mi caso con un equipo de prestaciones limitadas con el Grid Search puede ser lento pero se termina el procedimiento, el auto-arima no lo puede terminar. Aún así el Auto-Arima tiene un procediendo más sencillo, así que decidí realizar las predicciones con este método.

La previsión comenzará desde el 2019-01-21 06:00 con un periodo de 930, los resultados se muestran en la fig. (16), los resultados no son los mejores parece que sigue un patrón, se puede decir que tiene un overfitting.

Al evaluar los resultados de la serie temporal:

- RMSE = 1684.77
- MAPE = 191.12
- MAE = 1306.42

La evaluación de la previsión de la serie temporal es mala.

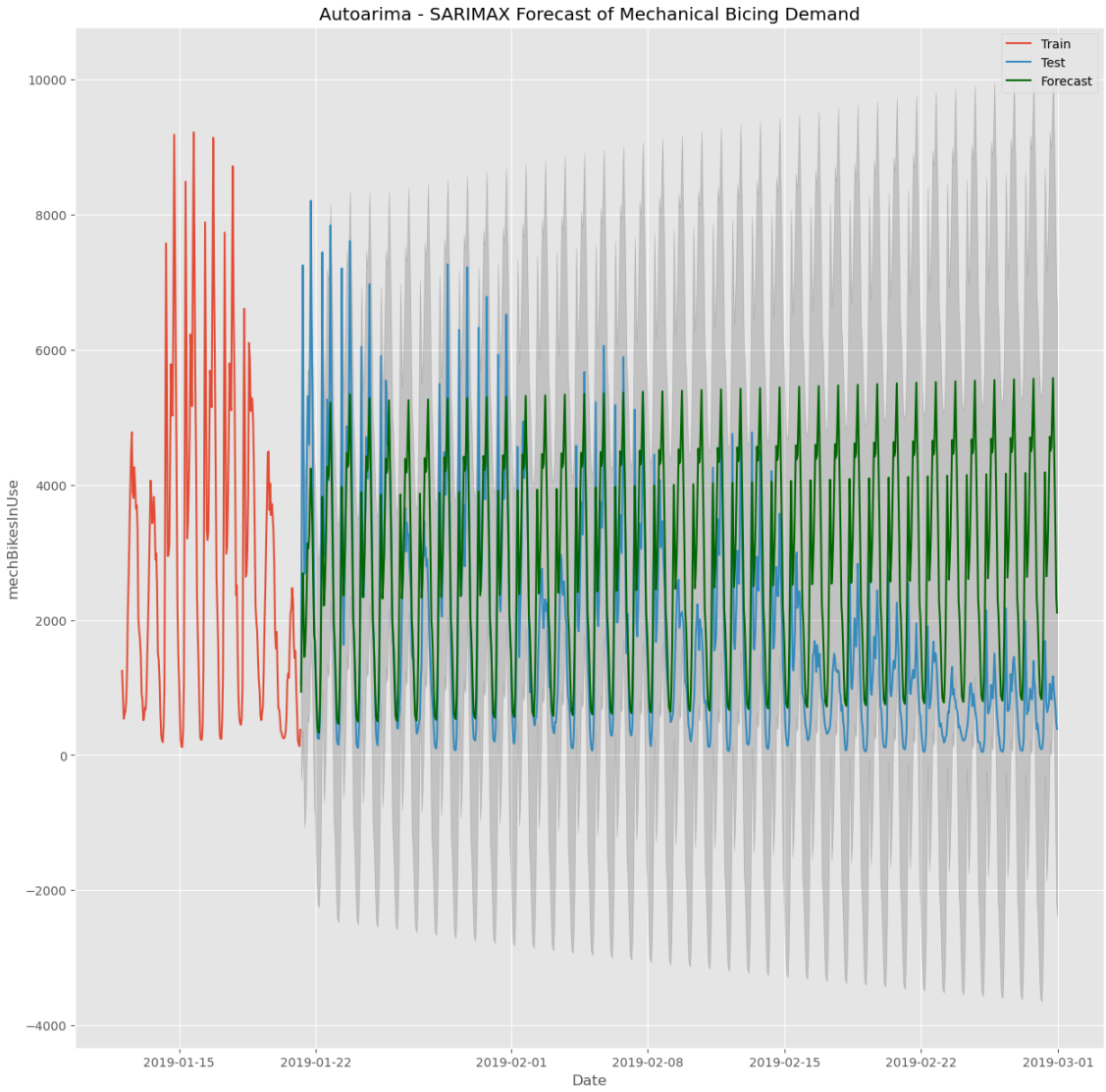

Fig. 16 Previsión del uso de bicicletas ordinarias.

Bicicletas eléctricas

La previsión comenzará desde el 2019-01-05 14:00 con un periodo de 1306, los resultados se muestran en la fig. (16), los resultados no son los mejores parece que sigue un patrón, se puede decir que tiene un overfitting.

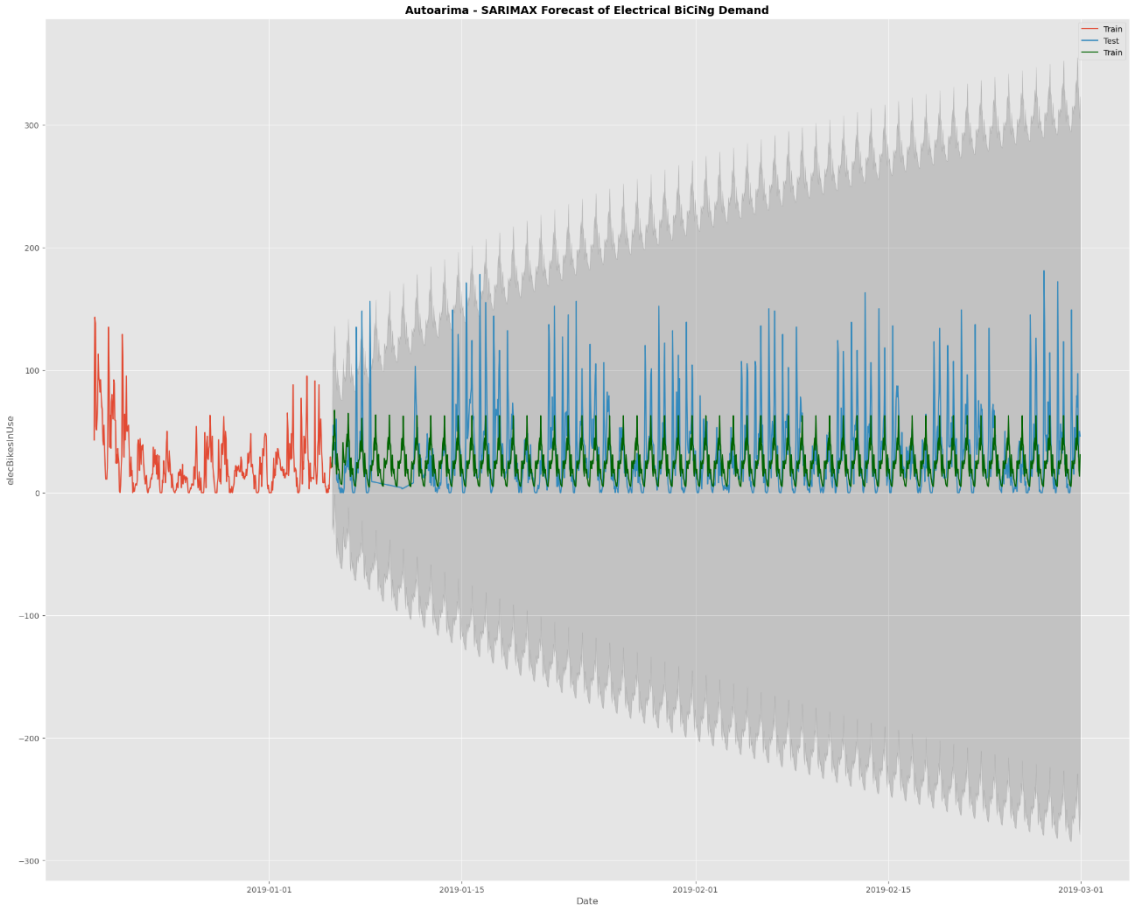

Fig. 17 Previsión del uso de bicicletas eléctricas.

Al evaluar los resultados de la serie temporal:

RMSE = 39.18

MAPE = 168.26

MAE = 27.87

La evaluación de la serie temporal es mucho mejor que los resultados para las bicicletas ordinarias, pero aun se puede mejorar el modelo.

Conclusiones

Los modelos pueden mejorarse, primero tal vez quitando variables exógenas, las variables atmosféricas tal vez estén aportando más ruido que ayuda al modelo. La frecuencia por hora puede ser muy pequeña para captar todas las características de los datos, tal vez sería mejor agrupar los datos por franjas horarias. No se realizaron otros modelos de aprendizaje supervisado, sería interesante ver los resultados con otro tipo de modelos que no sean series temporales.

Referencias

Ajuntament de Barcelona. (2018). *Uso del servicio Bicing de la ciudad de Barcelona - Conjuntos de datos - Open Data Barcelona*. https://opendata-ajuntament.barcelona.cat/data/es/dataset/us-del-servei-bicing

Bicing. (2022). *¡Muévete de forma sostenible por Barcelona! | Bicing*. https://www.bicing.barcelona/es

Hyndman, R., & Athanasopoulos, G. (2022). *Forecasting: Principles and Practice*. https://otexts.com/fpp3/

Peixeiro, M. (2022). *Time Series Forecasting in Python*. Manning.

Stackhouse, P. (2022). *POWER | Data Access Viewer*. https://power.larc.nasa.gov/data-access-viewer/