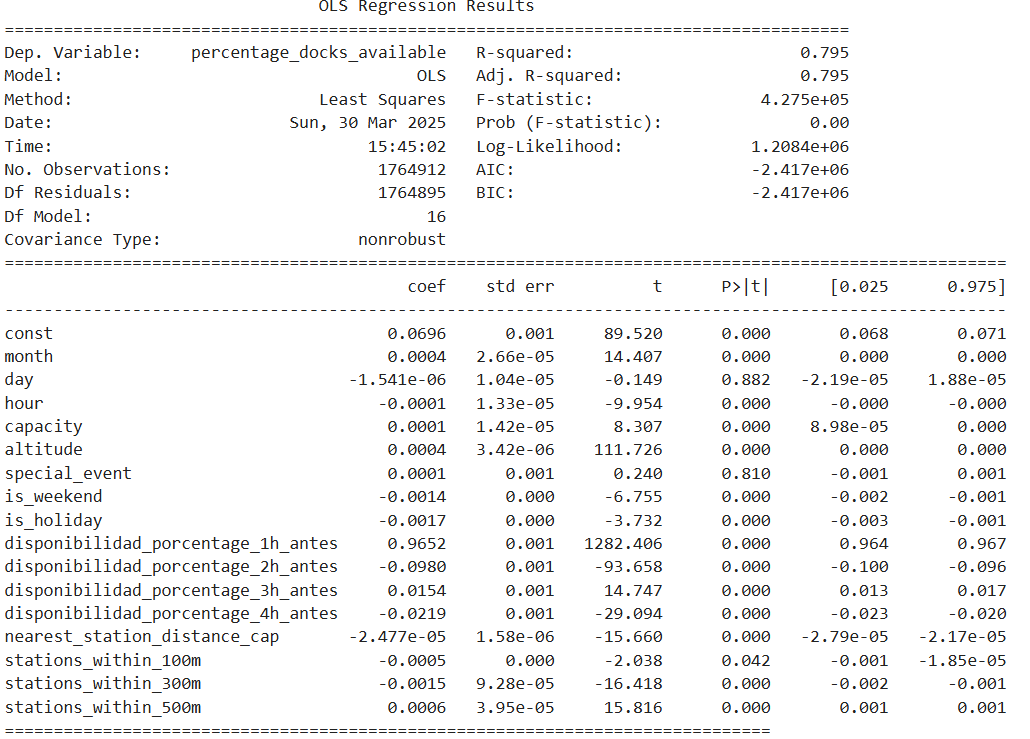
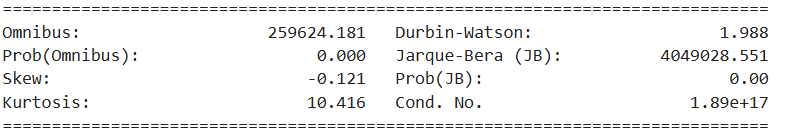
Modelo explicación

Hacer un **OLS (Ordinary Least Squares)**, o regresión lineal ordinaria, al principio de un análisis de datos es importante por varias razones clave. Aunque **XGBoost** u otros modelos complejos pueden ser muy potentes, **OLS** proporciona una base sólida para comprender la relación entre las variables y ayuda a identificar posibles problemas con los datos o el modelo.

Al realizar **OLS**, obtenemos pruebas estadísticas (como el valor p) para evaluar si las variables son **significativas** para predecir el resultado. Esto nos permite identificar cuáles variables aportan valor al modelo.

Si algunas variables no son significativas (por ejemplo, valor p > 0.05), podríamos considerar eliminar esas variables o realizar transformaciones.





**1. Estadísticas generales del modelo:**

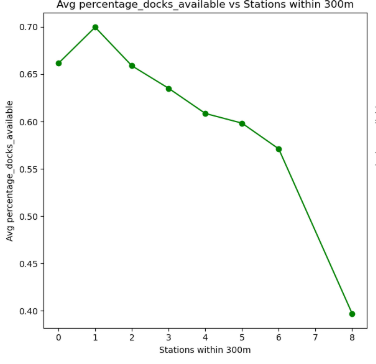
* **R-squared (0.795)**: Indica que aproximadamente el 79.5% de la variabilidad en el porcentaje de muelles disponibles puede ser explicado por las variables independientes del modelo.
* **Adj. R-squared (0.795)**: Similar al R-squared, pero ajustado por el número de predictores. Un valor cercano al R-squared indica que la inclusión de más variables no ha causado una sobreajuste del modelo.
* **F-statistic (4.275e+05) y Prob (F-statistic) (0.00)**: La prueba F evalúa si al menos una de las variables independientes tiene una relación significativa con la variable dependiente. Un valor p (Prob) de 0.00 indica que el modelo es altamente significativo.
* **Log-Likelihood (1.2084e+06)**: Mide la bondad de ajuste del modelo, con valores más altos indicando un mejor ajuste.
* **AIC (-2.417e+06) y BIC (-2.417e+06)**: Son criterios para la selección de modelos. Valores más bajos indican un buen ajuste.

**2. Coeficientes de las variables independientes:**

Cada coeficiente representa la relación de esa variable independiente con la variable dependiente (percentage\_docks\_available). Los coeficientes significativos son aquellos con un valor p (P>|t|) menor a 0.05. A continuación, se describen algunos de los resultados más relevantes:

* **month (0.0004)**: Cada aumento en el mes incrementa el porcentaje de muelles disponibles en 0.0004 unidades. Este coeficiente es muy pequeño pero significativo (valor p < 0.001).
* **day (-1.552e-06)**: El día tiene un efecto negativo muy pequeño, pero no es significativo (P>|t| = 0.881), lo que sugiere que no hay una relación clara entre el día y la disponibilidad de muelles.
* **hour (-0.0001)**: La hora del día tiene una relación negativa con la disponibilidad de muelles. Cada aumento en la hora disminuye ligeramente el porcentaje de muelles disponibles.
* **capacity (0.0001)**: La capacidad tiene un efecto positivo, lo que sugiere que a mayor capacidad, hay más muelles disponibles.
* **altitude (0.0004)**: El altitud tiene un coeficiente positivo y es altamente significativo, lo que implica que el aumento de la altitud está asociado con un aumento en la disponibilidad de muelles.
* **special\_event (0.0001)**: Los eventos especiales no parecen tener un impacto significativo en la disponibilidad de muelles (valor p > 0.05).
* **is\_weekend (-0.0014)**: Los fines de semana están asociados con una pequeña disminución en la disponibilidad de muelles.
* **is\_holiday (-0.0017)**: Similar al fin de semana, los días festivos también tienen una relación negativa con la disponibilidad de muelles.
* **is\_charging\_station (6.856e-08)**: La presencia de una estación de carga está altamente asociada con un aumento en la disponibilidad de muelles.
* **disponibilidad\_porcentage\_1h\_antes (0.9651)**: La disponibilidad de muelles una hora antes tiene un impacto muy fuerte en la disponibilidad actual, con un coeficiente cercano a 1, lo que indica una relación muy fuerte y positiva.
* **disponibilidad\_porcentage\_2h\_antes (-0.0980)**: A mayor disponibilidad de muelles dos horas antes, menor es la disponibilidad de muelles en el momento actual, lo que sugiere un patrón cíclico.
* **nearest\_station\_distance (-1.953e-05)**: A medida que aumenta la distancia hasta la estación más cercana, disminuye ligeramente la disponibilidad de muelles.
* **stations\_within\_100m (-8.15e-05)**: La proximidad a otras estaciones dentro de los 100 metros tiene un efecto negativo en la disponibilidad de muelles, pero este efecto no es significativo.
* **stations\_within\_300m (-0.0014)**: A más estaciones cercanas a 300 metros también muestran una relación negativa significativa con la disponibilidad de muelles.

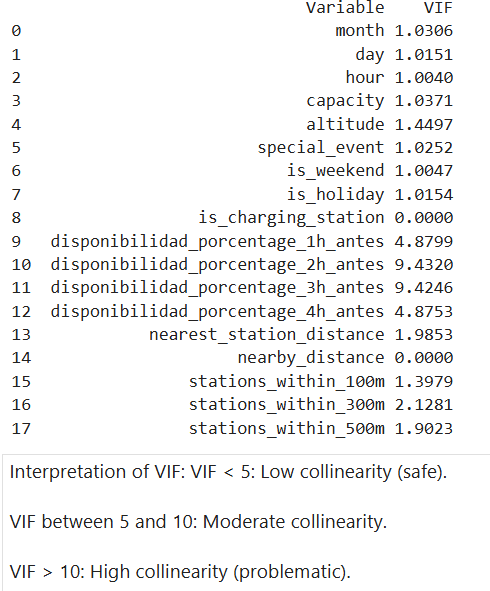
Como también sugiere la grafica del análisis exploratorio, la variable 300 metros parece tener una relación lineal negativa con la disponibilidad de muelles.



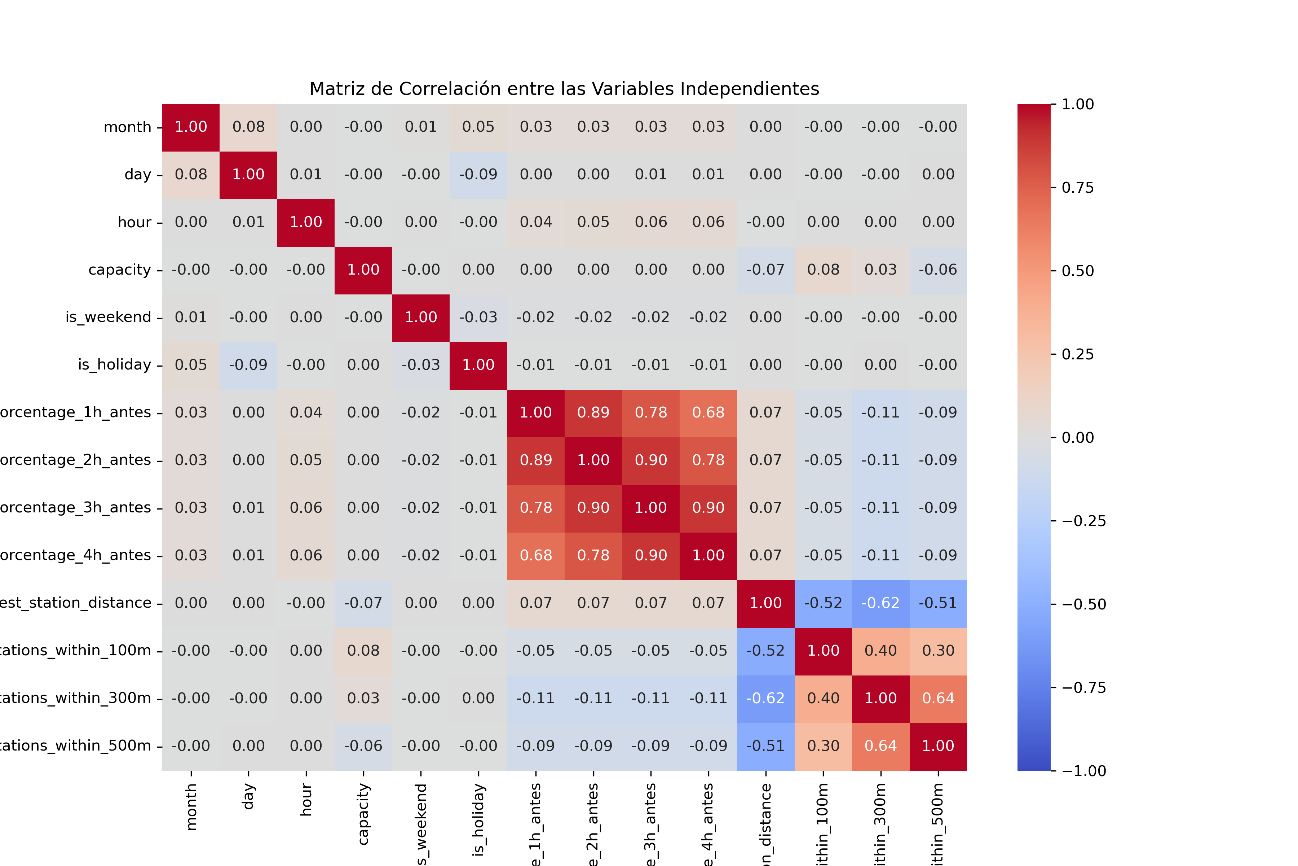
* **stations\_within\_500m (0.0006)**: Sin embargo, las estaciones dentro de los 500 metros están positivamente relacionadas con la disponibilidad de muelles. Es posible debido al efecto de autocorrelación entre estas variables, para contrarrestar el efecto de la variable anterior.

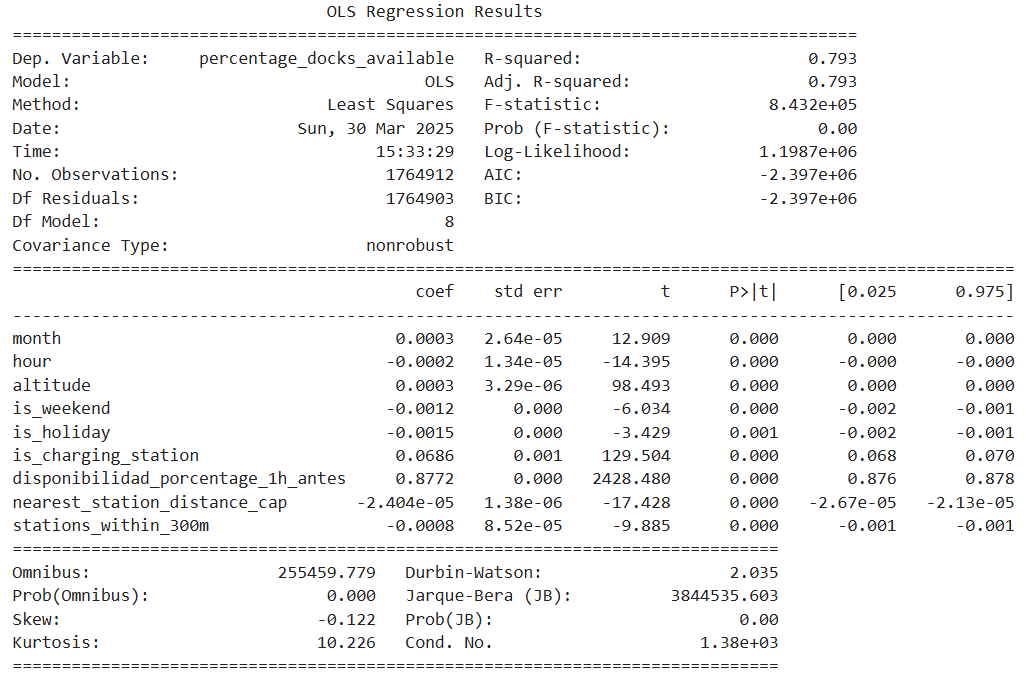
Por eso se toma la decisión de utilizar solo la variable **stations\_within\_300m** en el modelo final.

Se procede al análisis de factores de inflación de la varianza (FIV) para evaluar la multicolinealidad entre las variables independientes en el modelo. Observamos dos o más variables independientes que están altamente correlacionadas, lo que puede hacer que el modelo sea menos confiable.



Tambien observamos en las matrices de correlación fuerte correlación entre estas variables:





**3. Pruebas adicionales:**

* **Omnibus (259624.181)**: Es una prueba de normalidad de los residuos. Un valor p de 0.00 indica que los residuos no siguen una distribución normal.
* **Durbin-Watson (1.988)**: Este estadístico evalúa la autocorrelación de los residuos. Un valor cercano a 2 indica que no hay autocorrelación significativa.
* **Jarque-Bera (JB) y Prob (JB) (0.00)**: Indican que los residuos no siguen una distribución normal (esto se confirma con el valor p muy bajo).
* **Skew (-0.121) y Kurtosis (10.416)**: La asimetría es baja, pero la curtosis indica que los residuos tienen colas más gruesas que una distribución normal. Esto indica que los residuos pueden tener más valores extremos (outliers) que una distribución normal, lo que podría implicar que el modelo no está capturando algunas variaciones importantes de los datos.

Hemos intentado tratar los outliers, pero el modelo no ha mejorado significativamente, aunque Curtosis haya disminuido un poco, después de quitar el outlier.

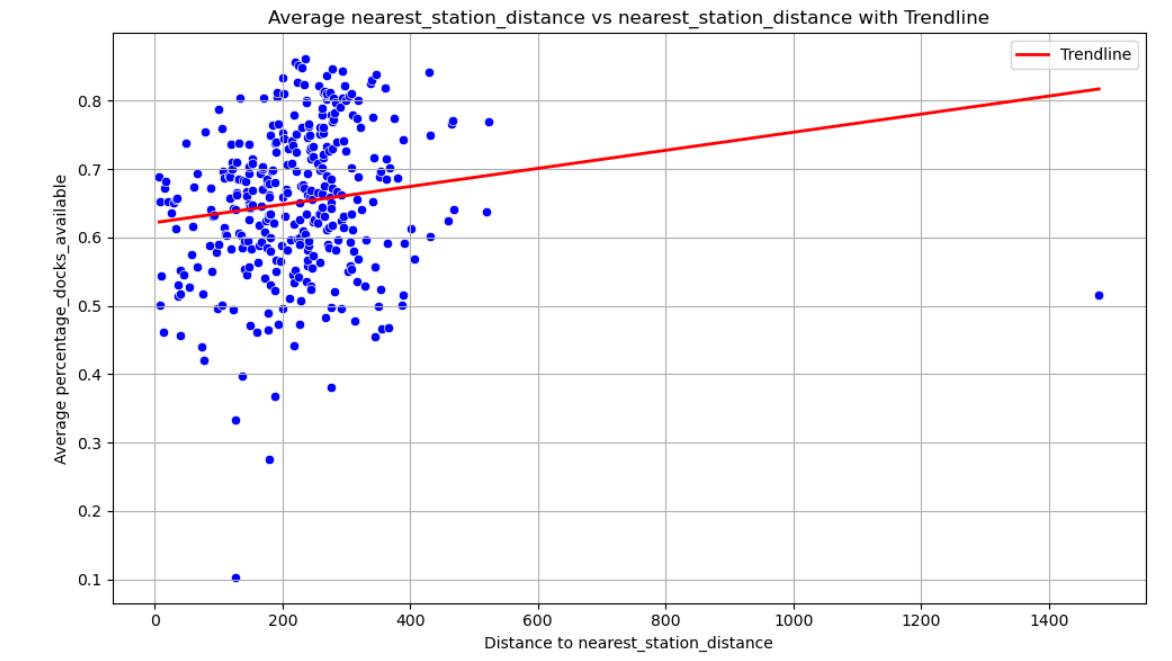
Por ejemplo hemos capado la variable “nearest\_station\_distance” :

percentil\_1 = df\_prediction\_task['nearest\_station\_distance'].quantile(0.01)

percentil\_99 = df\_prediction\_task['nearest\_station\_distance'].quantile(0.99)

df\_prediction\_task['nearest\_station\_distance\_cap']=df\_prediction\_task['nearest\_station\_distance'].clip(lower=percentil\_1, upper=percentil\_99)

Con esto quitamos el outlier que es la distancia de 1400 metros en la grafica, y lo reemplazamos con el quantil (0,99)



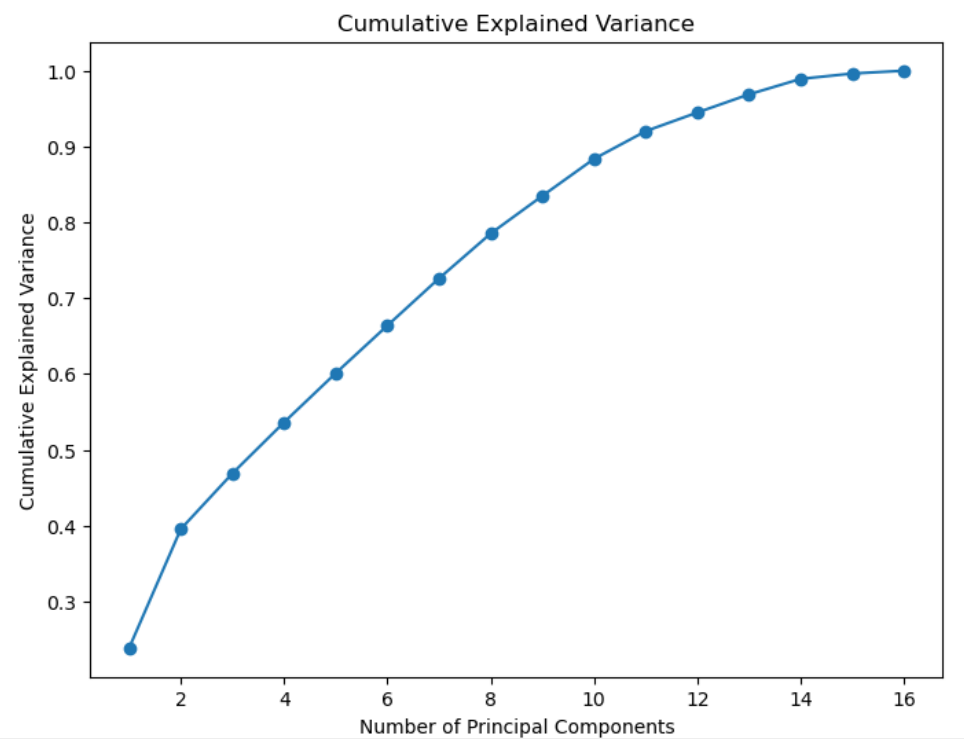
En resumen, el modelo OLS muestra que varias variables, como la capacidad, la altitud, y la disponibilidad 1 hora pasada de muelles tienen una relación significativa con el porcentaje de muelles disponibles, mientras que algunas otras variables como los eventos especiales y el día de la semana no tienen un impacto significativo.

Otra alternativa que hemos utilizado para mitigar las correlaciones entre variables independientes es utilizar **PCA (Análisis de Componentes Principales)** para reducir la **multicolinealidad** en nuestro modelo y luego usar las componentes principales como entradas en un modelo de **XGBoost** para mejorar la eficiencia y la estabilidad del modelo, especialmente cuando las variables de entrada están altamente correlacionadas, como es nuestro caso.

**explained\_variance\_ratio\_** nos da el porcentaje de varianza explicada por cada componente principal. Si sumas los valores de explained\_variance\_ratio\_, obtenemos la cantidad total de varianza explicada por todas las componentes principales.

Ahora obtenemos las componentes principales (que no están correlacionadas), y podríamos usar estos nuevos datos como entradas en tu modelo de **XGBoost**.

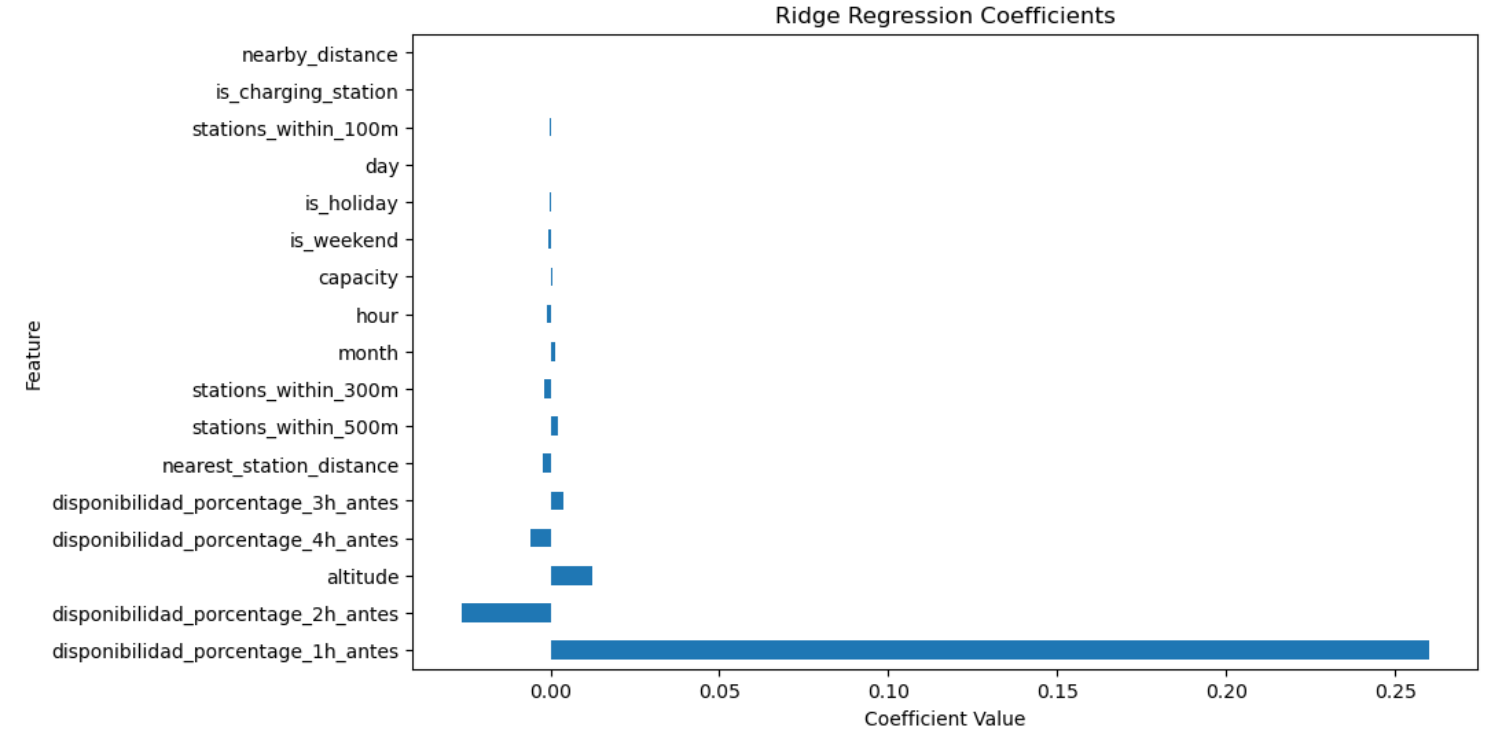
Nos ha salido el R2 muy parecido que en el modelo de regression 0,78



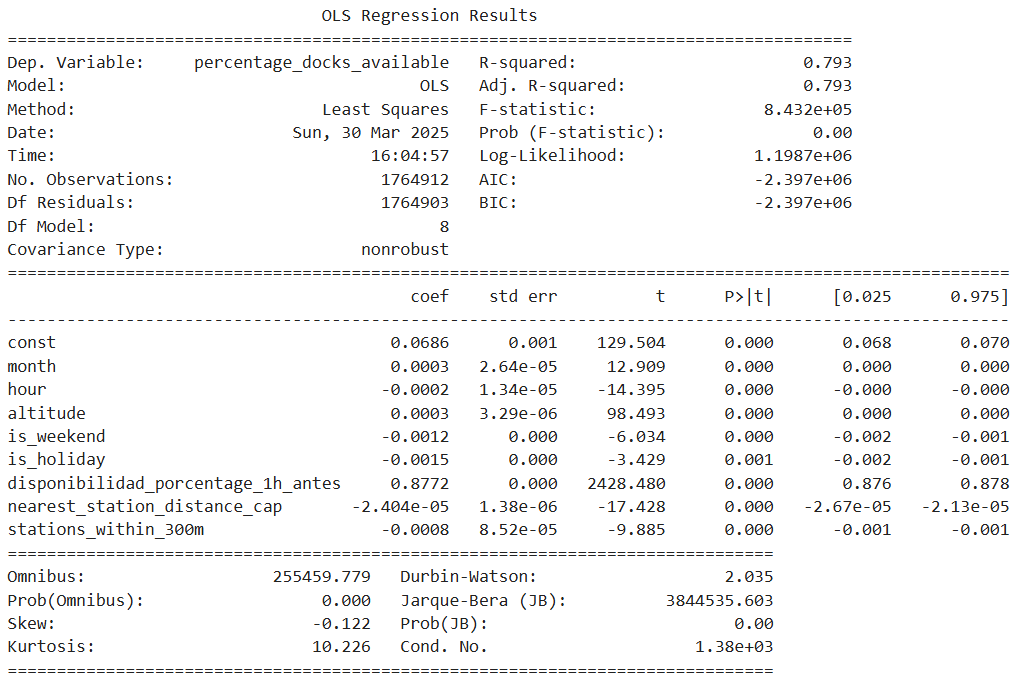
Así mismo, se ha procedido a evaluar con el modelo Ridge para minimizar el error, pero con una penalización que busca reducir su magnitud (esto se hace para controlar el sobreajuste o **overfitting**).

El código ha ordenado las características según el **valor absoluto** de los coeficientes. La idea aquí es identificar cuáles características tienen el **mayor impacto** en el modelo. Los coeficientes con valores absolutos más grandes son más importantes, independientemente de si son positivos o negativos.

Obtenemos un 0,79 con Regression de Ridge



Hemos tomado la decisión de utilizar las variables con p value < 5% y el resto de las variables que no aportan mucha variabilidad, eliminar del modelo.



Finalmente, entrenamos el modelo con las variables seleccionadas con el algoritmo XGBoost y obtenemos el mejor R2 - 0,82

**XGBoost** es un modelo basado en **árboles de decisión** que puede capturar **relaciones no lineales** entre las variables. Los árboles de decisión pueden modelar interacciones complejas entre características sin la necesidad de especificar explícitamente las relaciones.

**Regresión lineal**: En un modelo de regresión lineal (o Ridge, Lasso, etc.), se asume que la relación entre las características y la variable dependiente es lineal. Si las relaciones entre las características son no lineales, los modelos de regresión tradicional pueden tener un rendimiento limitado, ya que no pueden capturar esas complejidades.