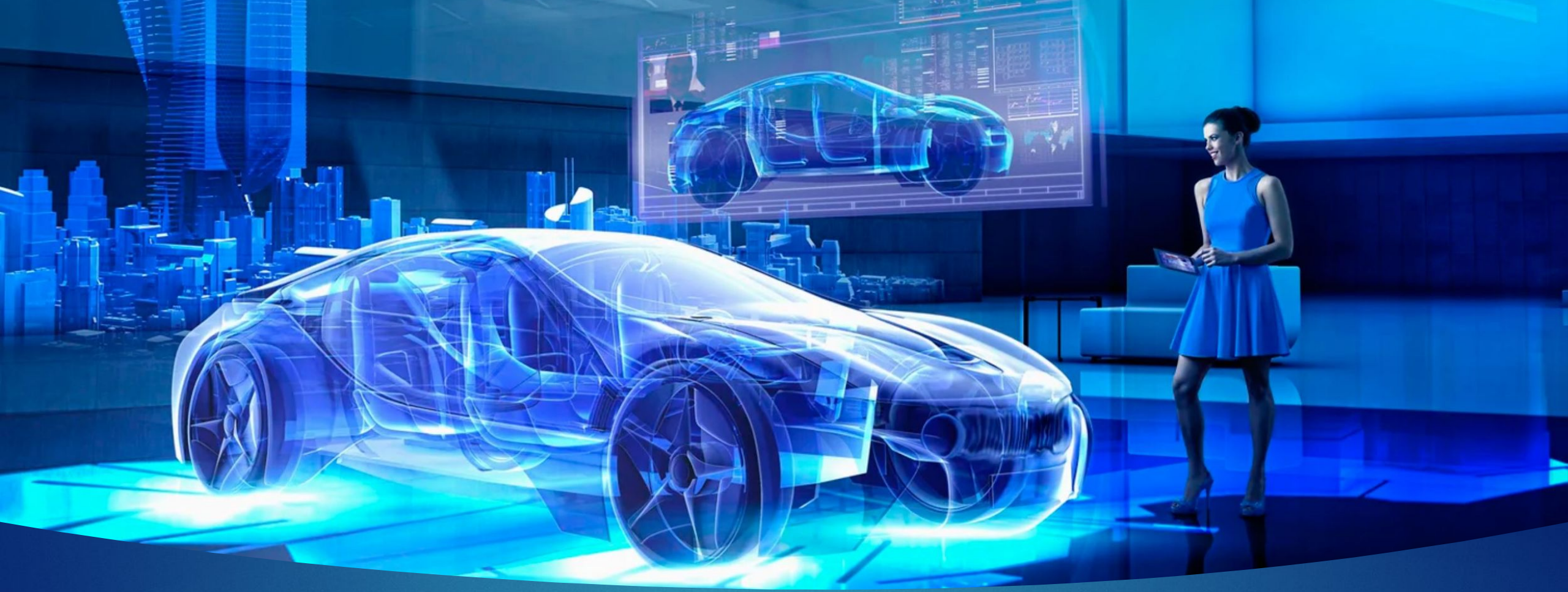


# Proyecto Data Science

Autor: Cristian Garcia Zanfardini

**Contexto Empresarial:**

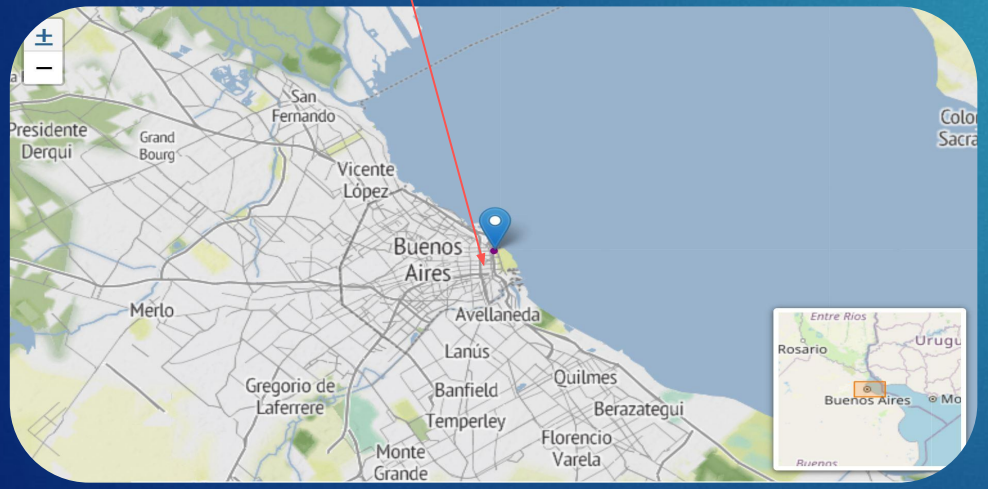
Somos una empresa joven dedicada al negocio de la Consultoria Automotriz. Nos apasiona lo que hacemos y queremos brindar la mejor experiencia a nuestros clientes. Creemos que el futuro es para aquellos que piensan más allá de los límites del presente y descubren nuevas fuentes de valor. Queremos hacer los cambios que marcan la diferencia. Tenemos un sólido objetivo, el cual se basa en ofrecer asesoría sobre las diferentes conecciones que regulan el precio de un automovil. Es entre estos aspectos, que comienza la comprensión de las variables del campo, su profundidad y posibilidades; permitiendonos asi brindar un modelo de negocio exitoso y sustentable.

Data Science (Creando valor):

En el mundo de hoy, los negocios y la tecnología cambian rápidamente, presentando oportunidades inigualables para un crecimiento rápido y transformador.

Nuestra área de Data Science puede ayudarlo a aprovechar estas oportunidades; para descubrir los secretos que existen detrás de la información.

Podemos ofrecer el asesoramiento preciso y directo, en el momento adecuado, para que puedan tomar la mejor decisión.




**Contexto Comercial:**

_Una empresa automovilística china, aspira a entrar en el mercado estadounidense estableciendo allí su unidad de fabricación. Producira automóviles para competir con sus contrapartes locales y europeas.

**Objetivo:**

Nos han contratado para comprender los factores que afectan el valor de los automóviles en el mercado estadounidense. Especificamente deberemos predecir los precios a partir de variables ya conocidas.

A nuestra izquierda geo-referenciamos los países involucrados. Y más precisamente la ubicación de nuestras oficinas.

***Columnas:***

- **car_ID:** numeracion secuencial de los autos
- **CarName:** marca/modelo
- **fueltype:** tipo de combustible
- **aspiration:** tipo de aspiracion (standard/turbo)
- **doornumber:** numero de puertas
- **carbody:** tipo de auto
- **drivewheel:** traccion (delantera/trasera/4x4)
- **enginelocation:** ubicacion del motor
- **wheelbase:** distancia entre ejes
- **carlength:** largo del auto
- **carwidth:** ancho del auto
- **carheight:** altura del auto
- **curbweight:** peso en vacio
- **enginetype:** tipo de motor
- **cylindernumber:** numero de cilindros
- **enginesize:** cilindrada del motor
- **fuelsystem:** sistema de combustible
- **boreratio:** diametro del cilindro
- **stroke:** stroke
- **compressionratio:** ratio de compresion
- **horsepower:** potencia
- **peakrpm:** revoluciones por minuto a maxima potencia
- **citympg:** millas por galon en ciudad
- **highwaympg:** millas por galon en autopista
- **price:** precio del automovil

## Descripción de la temática de los datos:

_Con base en varias encuestas de mercado, hemos recopilado un gran conjunto de datos de diferentes tipos de automóviles en el mercado estadounidense.

Los mismos están disponibles en:

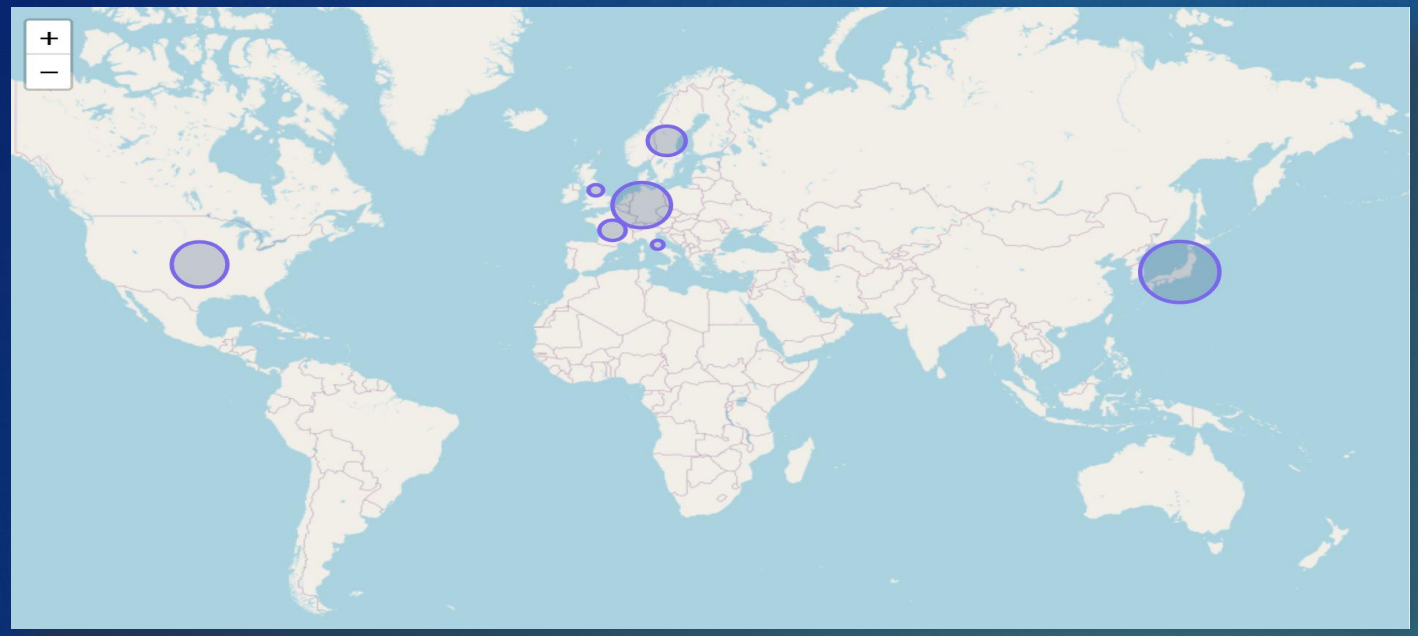
www.kaggle.com

Acerca del conjunto de datos:

_Consta de 26 columnas por 205 filas

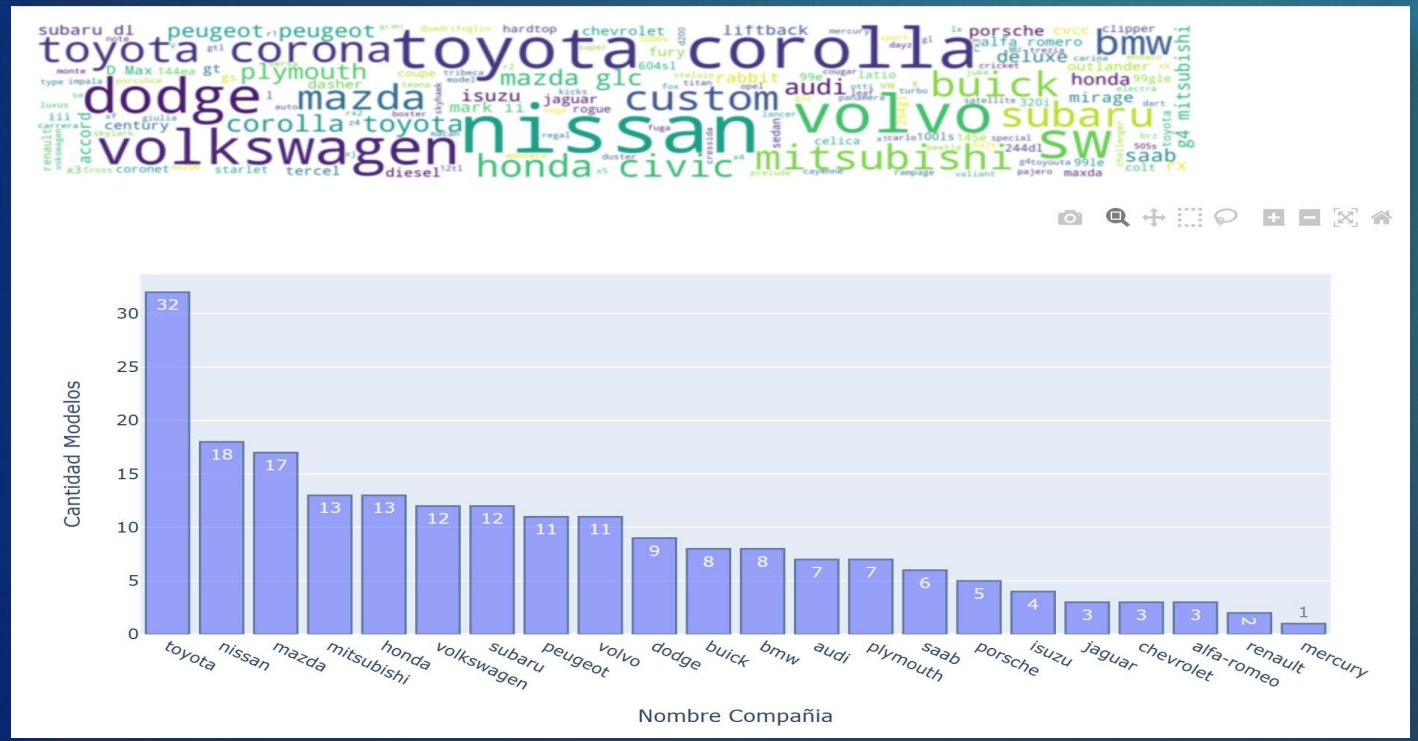
_Iniciamos nuestro análisis del DataFrame a través de los siguientes pasos:

_Identificar los datos faltantes
_Trabajar los datos faltantes
_Asignar formato de datos correcto



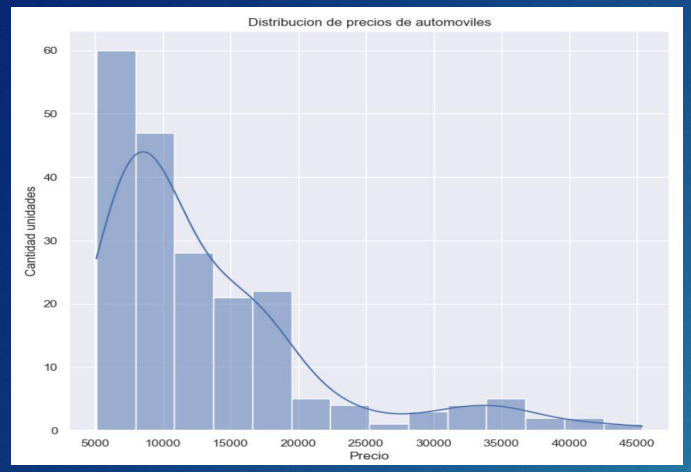
_Geo-referenciamos las compañías automotrices de acuerdo a su país. El tamaño de la burbuja las cuantifica.

Tip: Las compañías japonesas son las más numerosas, con 7 en total.



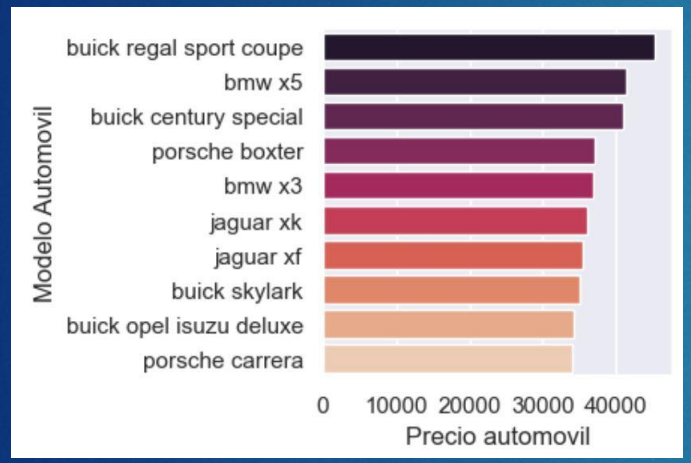

_Visualizamos la cantidad de modelos por nombre de compañía / modelos más comunes.

Tip: Japón es el país con mayor cantidad de modelos, 109 en total.



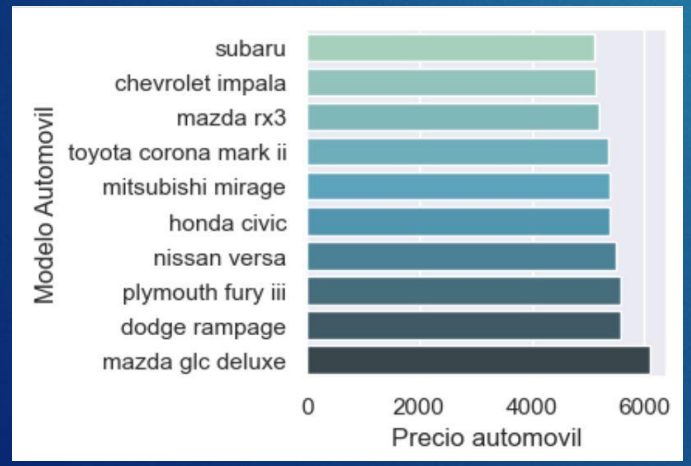

_Detalle de la distribución de precios de los automóviles

Tip: el precio más alto es $45.400; y el más bajo $5.118




_Detalle de los 10 autos más caros

Tip: Buick y BMW son los más caros




_Detalle de los 10 autos más baratos

Tip: Subaru y Chevrolet son los más baratos

_*Visualizando las variables numericas*

- wheelbase
- carlength
- carwidth
- carheight
- curbweight
- enginesize
- boreratio
- stroke
- compressionratio
- horsepower
- peakrpm
- citympg
- highwaympg
- car_price

| | Correlacion / Precio |
|---|---|
| **enginesize** | 0.874145 |
| **curbweight** | 0.835305 |
| **horsepower** | 0.808138 |
| **carwidth** | 0.759325 |
| **carlength** | 0.682921 |
| **wheelbase** | 0.577816 |
| **boreratio** | 0.553174 |
| **carheight** | 0.119337 |
| **stroke** | 0.079443 |
| **compressionratio** | 0.067984 |
| **symboling** | -0.079978 |
| **peakrpm** | -0.085268 |
| **car_ID** | -0.109093 |
| **citympg** | -0.685752 |
| **highwaympg** | -0.697600 |

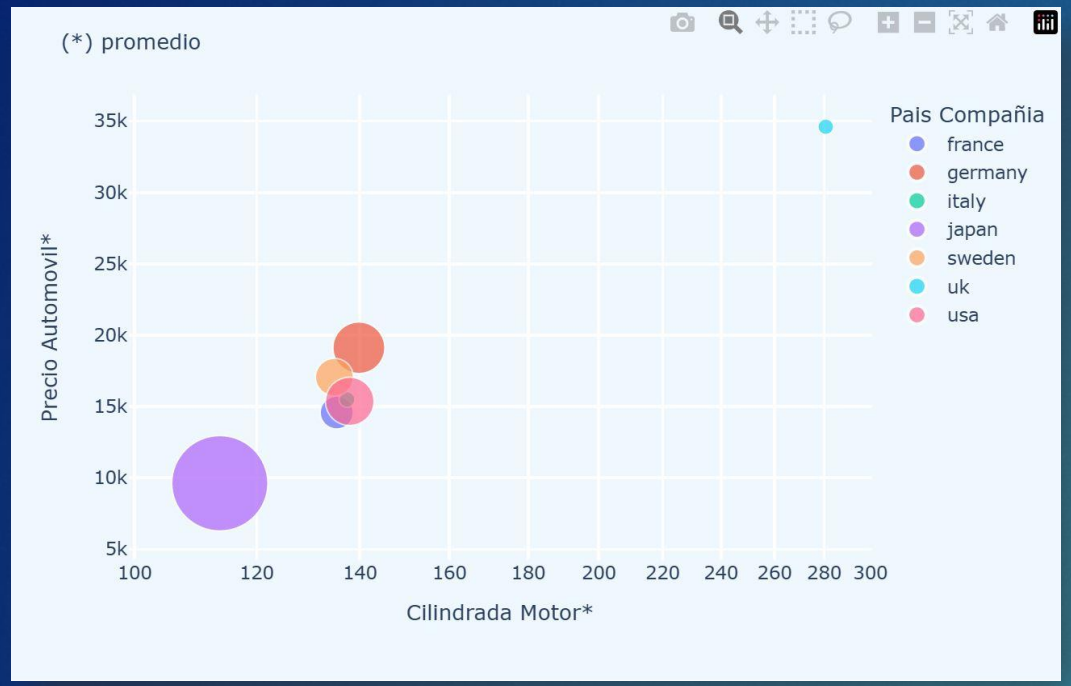
**Correlaciones:**

En principio vamos a estudiar las variables numéricas , ya que de esta forma entenderemos cuales son las variables que mejor explican el precio. A esto lo llamamos correlación:

Correlación positiva: cuando una variable aumenta de valor la otra variable también aumenta. Para que sea positiva, el valor del coeficiente de correlación debe estar entre 0 (no incluido) y 1 (incluido)

Correlación negativa: cuando una variable aumenta la otra disminuye, y al revés, si una variable disminuye la otra aumenta. El valor del coeficiente de correlación está entre -1 (incluido) y 0 (no incluido).

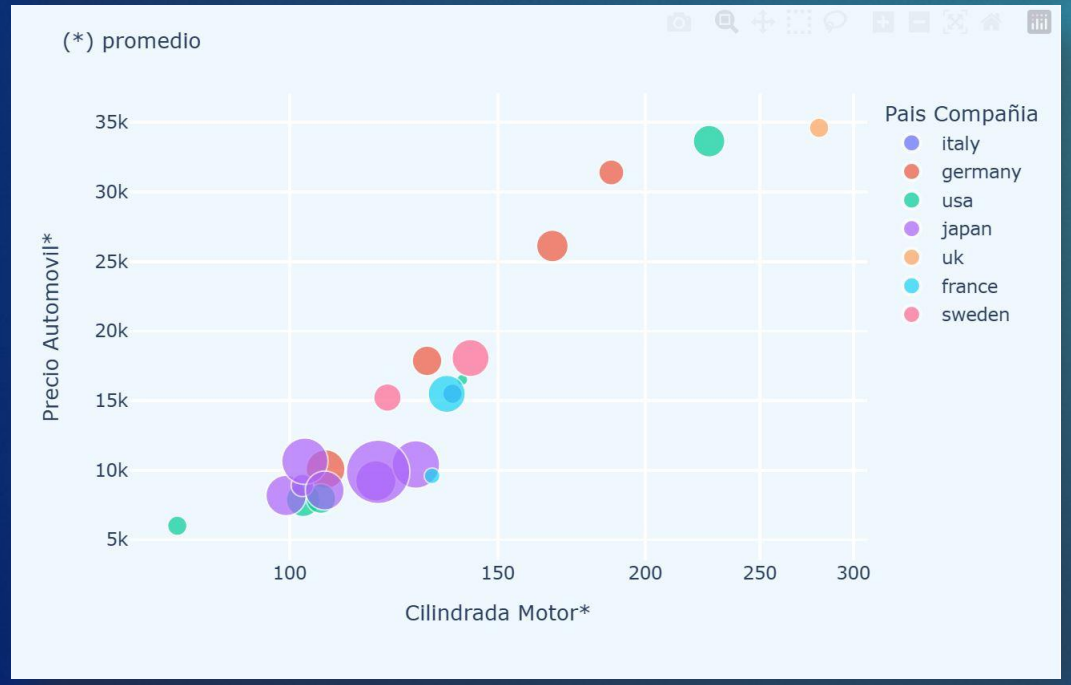
Para nuestro estudio vamos a seleccionar la correlación positiva y negativa más fuertes . Es decir “enginesize” (Cilindrada) y “highwaympg” (Autopista MPG)




## Cilindrada Motor vs Precio:

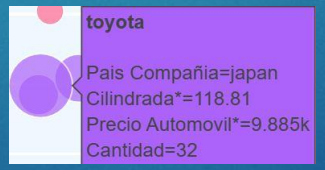
Las compañías están agrupadas por país (se obtuvo un promedio país).
Cada burbuja contiene información del promedio de marcas por país:


Tip: las companies japonesas son en promedio las de menor cilindrada

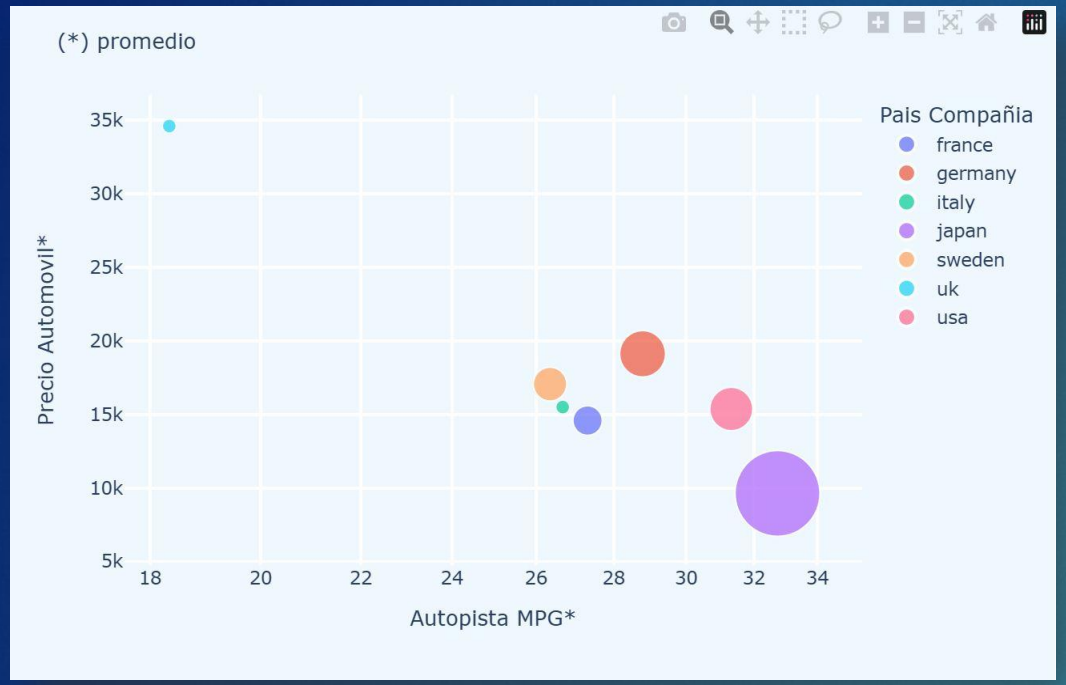



## Cilindrada Motor vs Precio:

Las compañías se muestran en forma individual (se obtuvo un promedio marca).
Cada burbuja contiene información del promedio por marca:

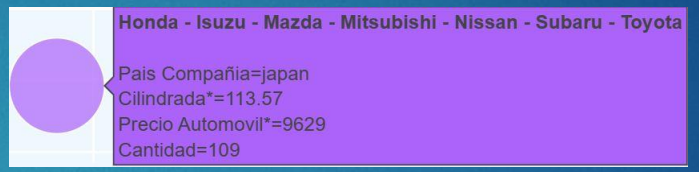



Tip: Chevrolet tiene el auto con menor cilindrada

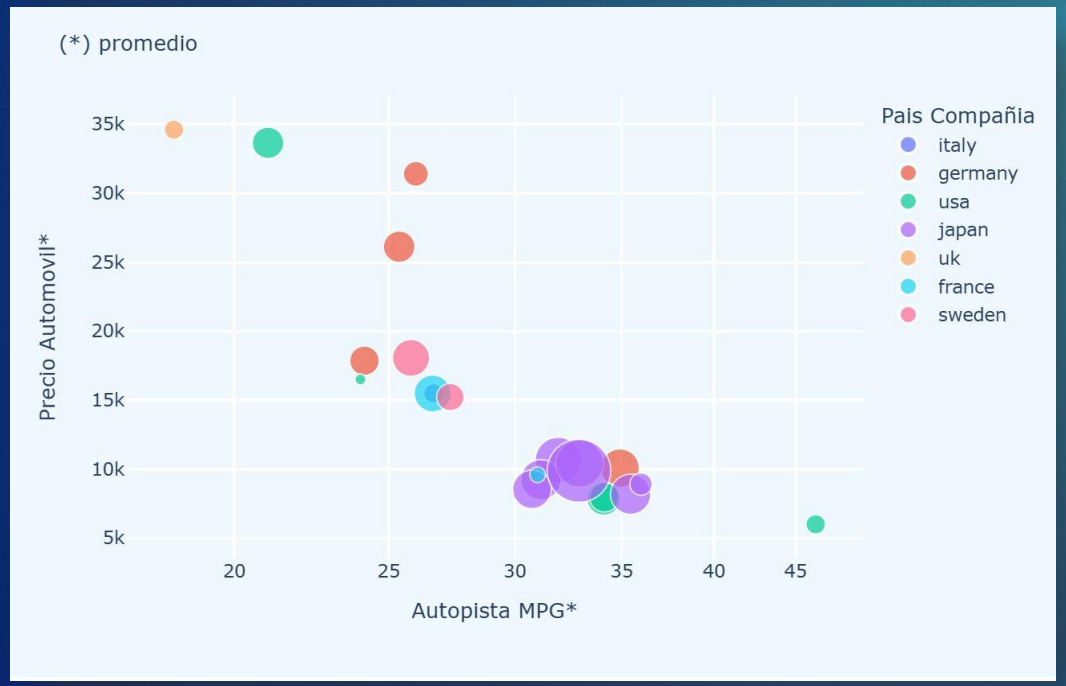



## Autopista MPG vs Precio:

Las compañías están agrupadas por país (se obtuvo un promedio país).
Cada burbuja contiene información del promedio de marcas por país:

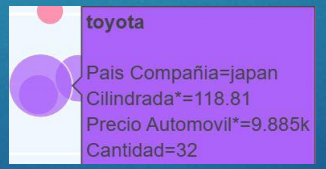



Tip: los autos japoneses, en promedio, rinden mayor cantidad de MPG

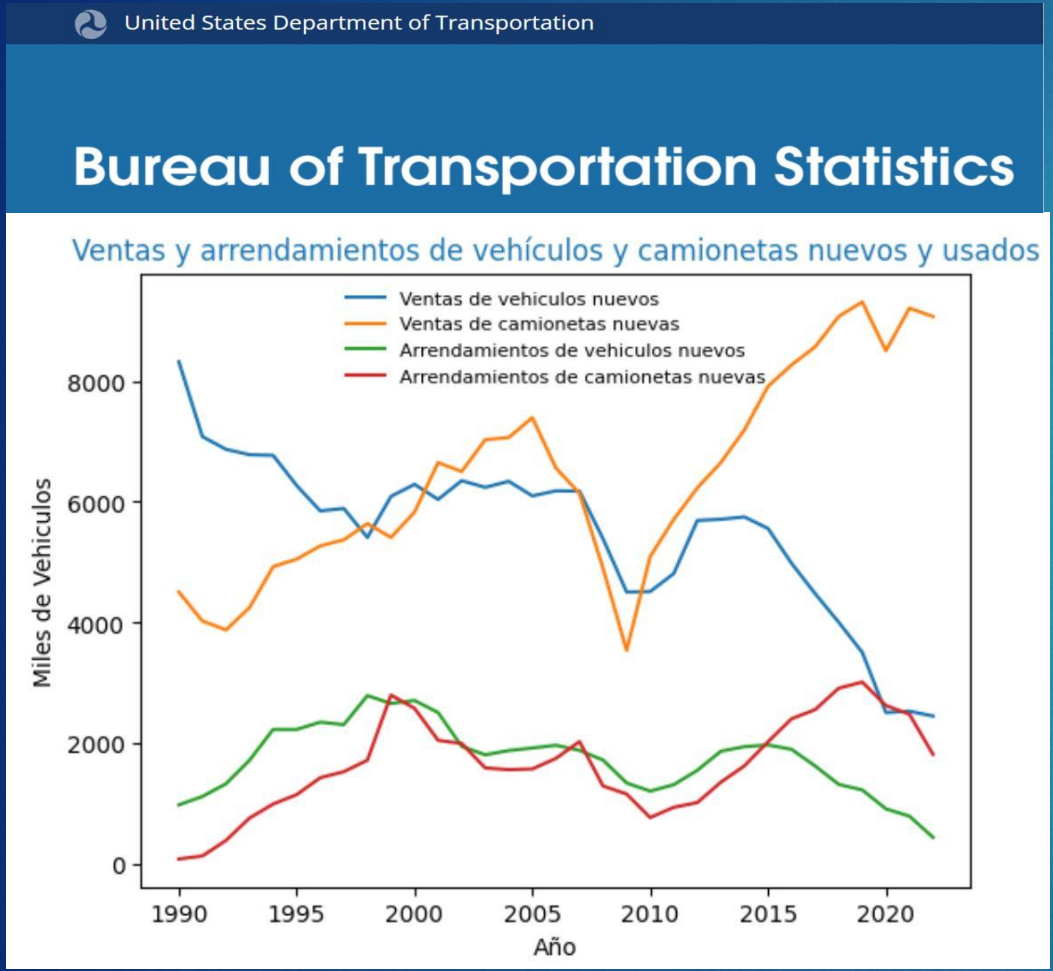



## Autopista MPG vs Precio:

Las compañías se muestran en forma individual (se obtuvo un promedio marca).
Cada burbuja contiene información del promedio por marca:



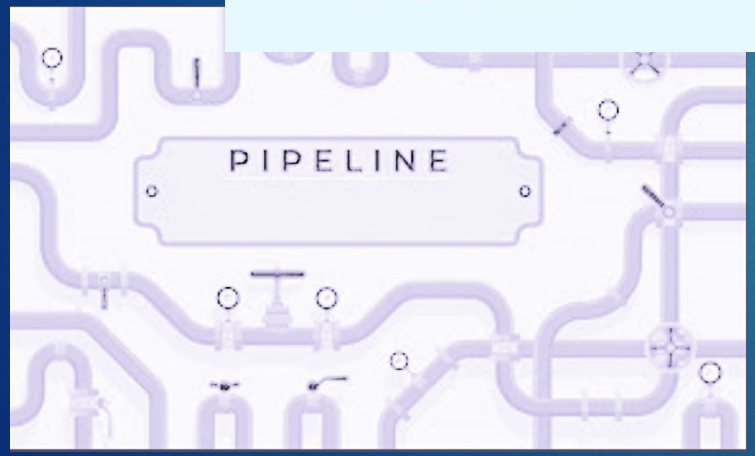

Tip: Chevrolet es el auto que rinde mayor cantidad de MPG




**Potenciando nuestro análisis:**

La información de La Oficina de Estadísticas de Transporte (BTS), nos permitirá tener una visión crítica y nos guiará en nuestra hipótesis previa.

Entender este grafico nos ayudara en la toma de decisiones, no solo en la predicción de precios, sino también en la evaluación del mercado a futuro . Siendo una clara señal que nuestro cliente deberá incorporar las camionetas si no quiere perder participación en el mercado automotor, ya que a través de los años la cuota de los automóviles ha ido decreciendo constantemente.


## Etapas de un problema de machine learning

Definir el problema: ¿Qué se pretende predecir? ¿De qué datos se dispone? o ¿Qué datos es necesario conseguir?

Explorar y entender los datos que se van a emplear.

Métrica de éxito: definir una forma apropiada de cuantificar los resultados obtenidos.

A fin de conseguir nuestro objetivo, deberemos:

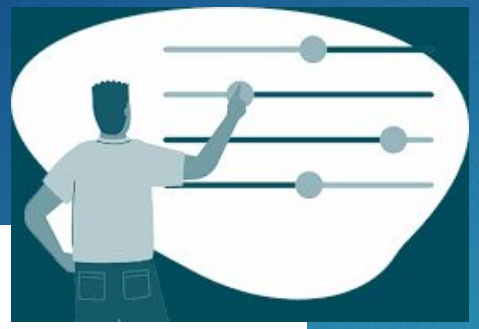
Preparar la estrategia para evaluar el modelo: separar las observaciones en un conjunto de entrenamiento, un conjunto de validación (o validación cruzada) y un conjunto de test. Es muy importante asegurar que ninguna información de test participa en el proceso de entrenamiento del modelo.

Preprocesar los datos: aplicar las transformaciones necesarias para que los datos puedan ser interpretados por el algoritmo de machine learning.

Gradualmente, mejorar el modelo incorporando-creando nuevas variables u optimizando los hiperparámetros (datos definidos por el usuario).

Evaluar la capacidad del modelo final con el conjunto de test

Entrenar el modelo final con todos los datos disponibles.

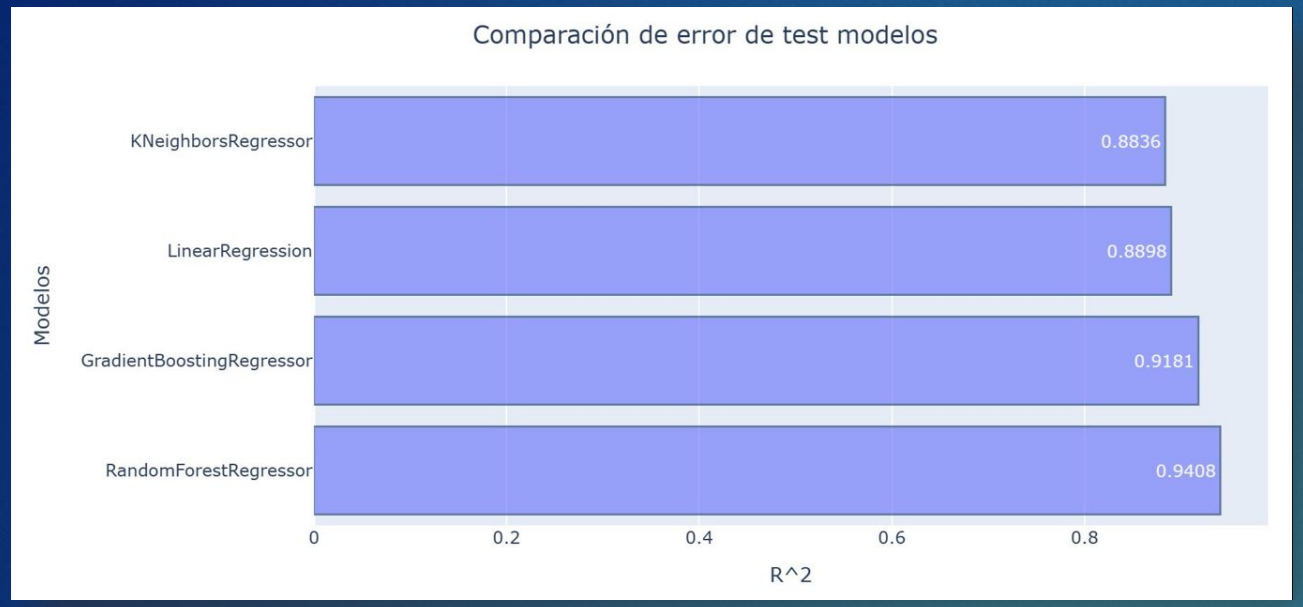


**Ajustes:**

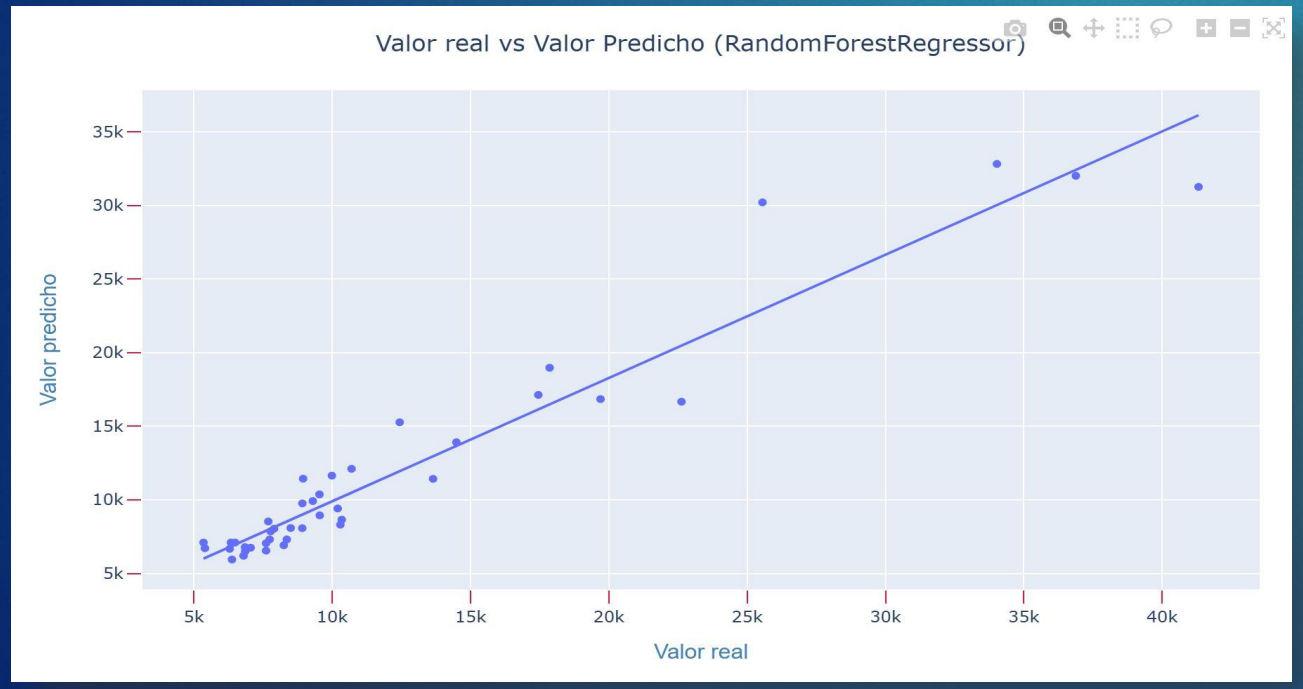
Durante el proceso de análisis hemos realizado diferentes transformaciones sobre el dataset, sin embargo hemos visto mejoras poco significativas en las métricas resultantes.

Asimismo, no es el propósito de esta presentación detallarlas una a una. Sino centrarnos en aquellas que presentaron  los mejores valores

Las detallamos a continuación




Comparamos las mejores métricas obtenidas en cada uno de nuestros modelos




Visualizamos rápidamente que tan bien predice nuestro mejor modelo


**Conclusión final:**

La ciencia de datos es mucho más que conocer los últimos métodos y herramientas disponibles.

Primero comprenda su problema. Evalúe sus datos de acuerdo con su objetivo.

Visualizar datos es clave. Ayúdese con gráficos para comprender las métricas.

Finalmente hay que entender que la ciencia de datos real requiere prueba y error. Debe validar constantemente la entrada y la salida.

Tips:

La validación cruzada nos dará métricas más robustas, pues al final el modelo terminará siendo entrenado y validado con la totalidad de los datos.

Los pipelines evitaran que se filtre información al conjunto de test.

Los modelos de ensembling, a través del tuneo de hiperparametros, permitirán obtener mejores métricas.

# MUCHAS GRACIAS!