# VENTA DE VEHÍCULOS EN AUSTRALIA

## PROYECTO FINAL

**Andrés Felipe Escobar Mosquera**

**Comisión 49175**
**Data Science**

# INDICE

- Contexto
- Objetivo
- Resumen de Metadata
- Preguntas de Interés
- EDA
- Insights
- Modelado de Machine Learning
- Conclusiones

# CONTEXTO

Establecer el precio de venta de un vehículo para que sea más competitivo va a depender de las características propias del mismo, pero también de las opciones que ofrece el mercado automotriz.

El Dataset de este estudio contiene la información más reciente sobre los precios de vehículos en Australia para el año 2023 incluyendo, además diversas características importantes de los vehículos.

Vamos a estudiar estas características, identificando cuáles son las que más influyen para determinar el Precio del vehículo (Variable Target).

## Audiencia

El análisis que se va a desarrollar es de utilidad para todas las partes interesadas (clientes, vendedores, fabricantes, entre otros.), que se ven involucradas en el mercado automotriz.

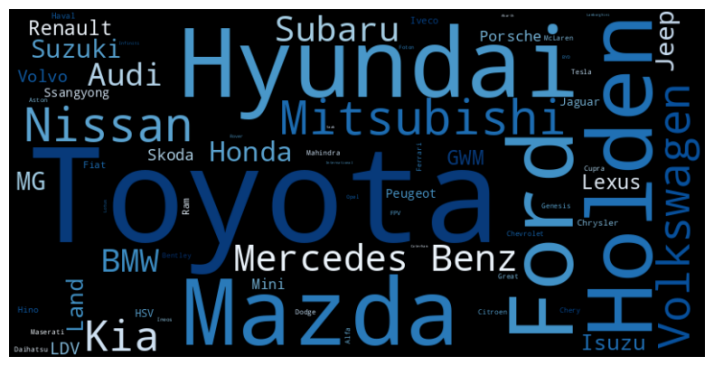
# OBJETIVO

Definir un modelo de Machine Learning que permita predecir el precio de venta de un vehículo con base a las variables que presenta. Para ello, se utilizarán algoritmos de aprendizaje supervisado, más precisamente, algoritmos de regresión, ya que permiten predecir valores continuos como el Precio del vehículo, en este caso.

# RESUMEN DE METADATA

El Dataset fue obtenido de Kaggle. Luego del Data Wrangling presenta 16582 filas y 15 variables: float64(8) object(7).

**74** Marcas de vehículos




**11** Tipos de Cuerpo

**5** Tipos de Tracción

**8** Tipos de Combustible

**+75%** Vehículos de 4 puertas y 5 asientos


**9** Estados

**12** Cilindros en el Motor

# PREGUNTAS DE INTERÉS

1. ¿Cuántos vehículos último modelo (año 2023) están disponibles a la venta? ¿A qué año corresponde la mayor cantidad de vehículos en venta?

2. ¿Los vehículos tipo SUV son los más ofertados en el mercado automotriz de Australia? ¿en qué porcentaje?

3. ¿Qué marcas de vehículos ofrecen precios por debajo de 20.000 AUD?. Mostrar las marcas según si el vehículo es usado, nuevo o demo y la cantidad de vehículos disponible para cada una de estas tres categorías.

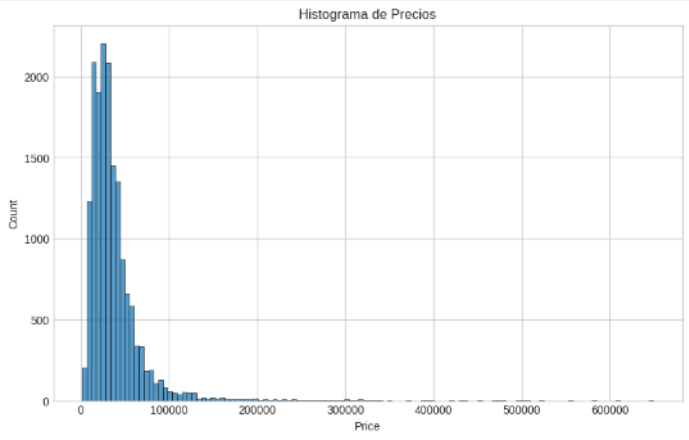
4. ¿En qué estado está disponible el vehículo eléctrico más económico?¿Cómo es la distribución de precios de vehículos eléctricos y cuantos hay disponibles según el estado?

5. De los vehículos SUV disponibles, ¿qué marcas presentan un consumo de combustible menor a 6 L/ 100 Km, sin tener en cuenta los vehículos eléctricos?. Compararlos según sean usados, nuevos o demo.

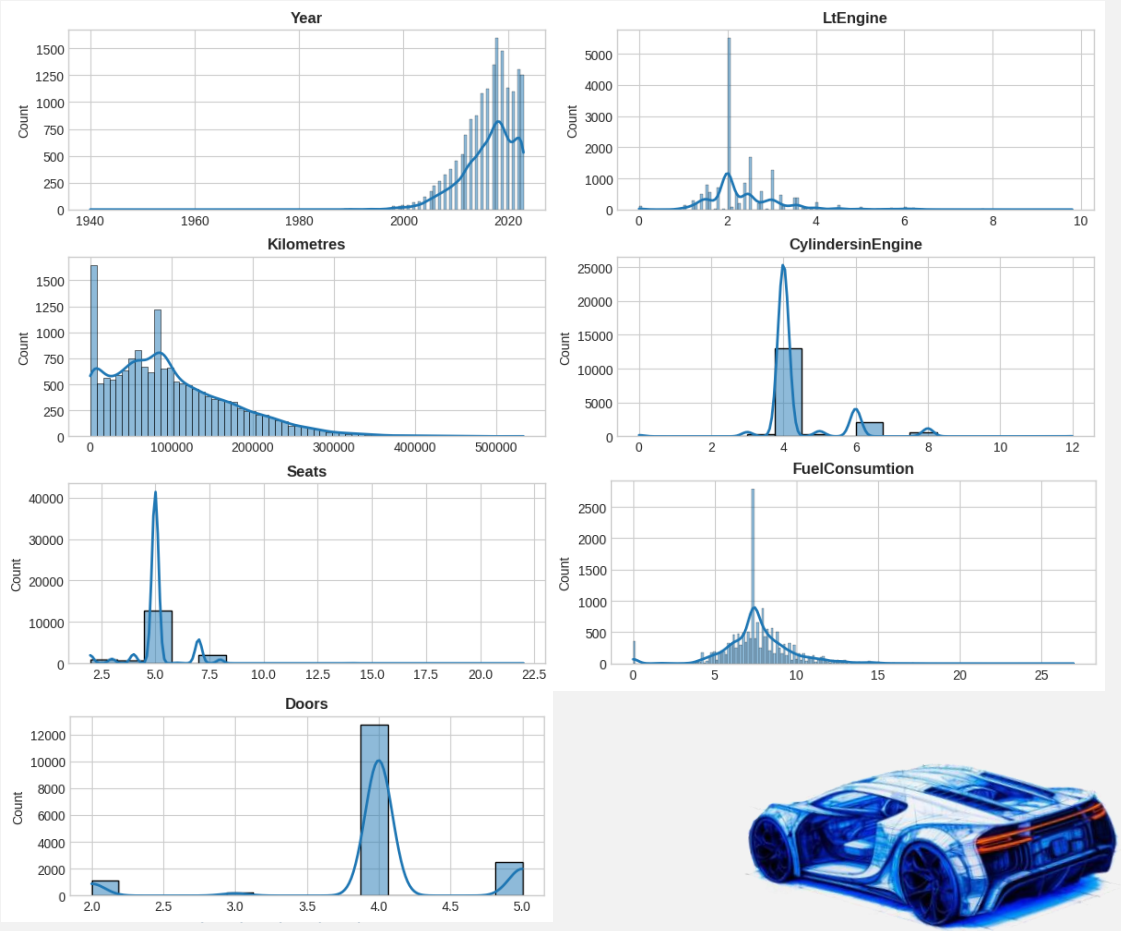
# EXPLORATION DATA ANALYSIS (EDA)

## UNIVARIADO

### Variables Numéricas

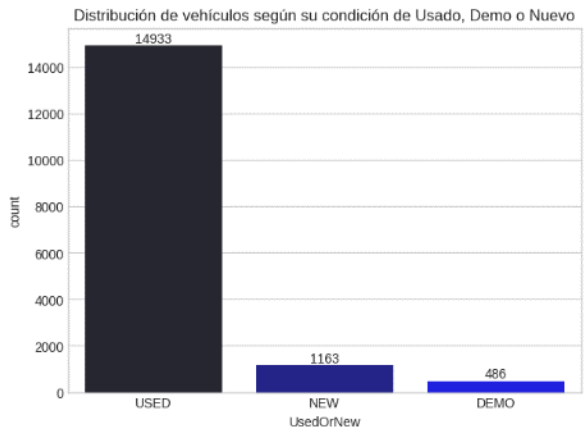


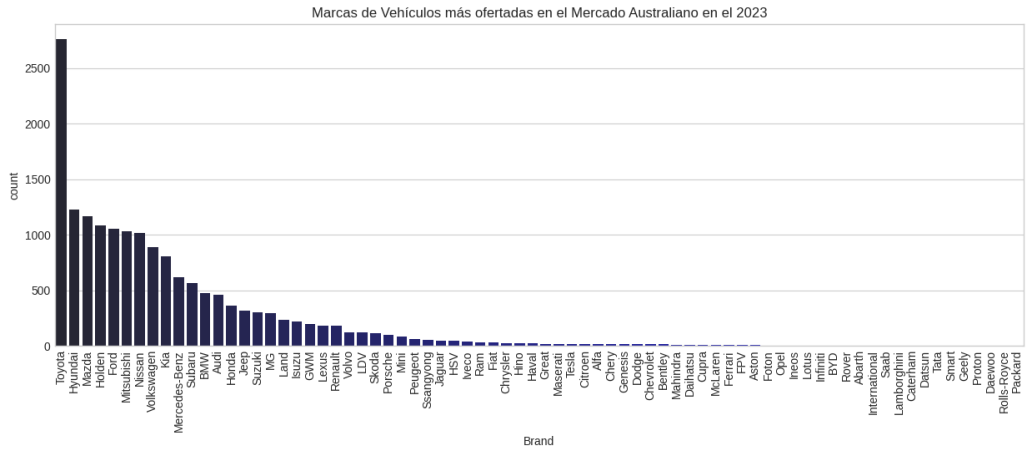

### Variable Target: Precio



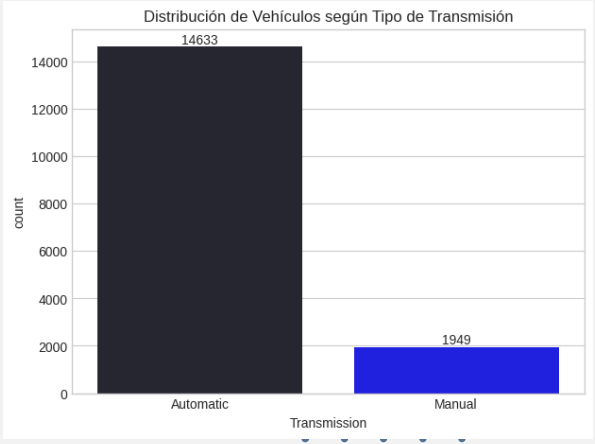
# EXPLORATION DATA ANALYSIS (EDA)

## UNIVARIADO



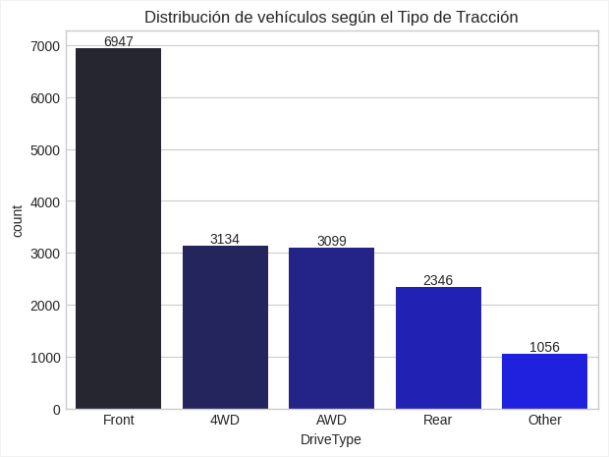

## Variables Categóricas

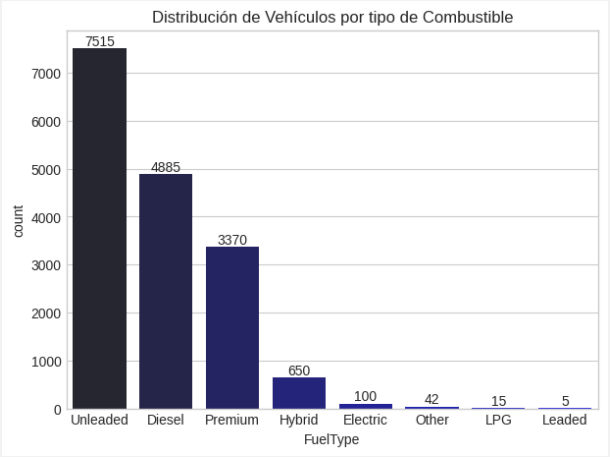


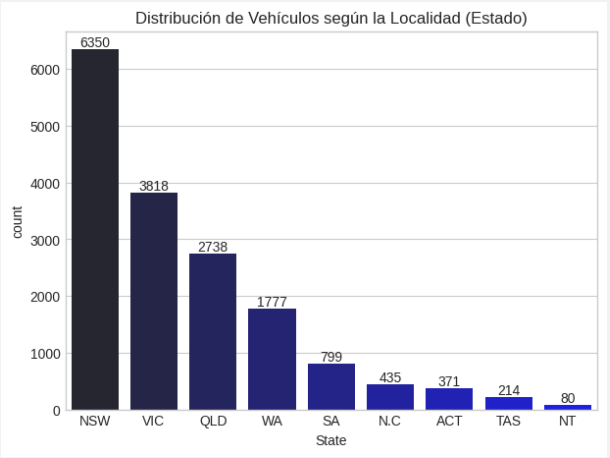

# EXPLORATION DATA ANALYSIS (EDA)

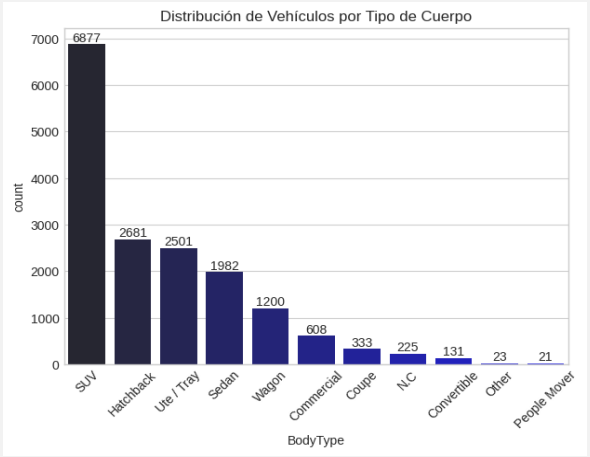
## UNIVARIADO

### Variables Categóricas



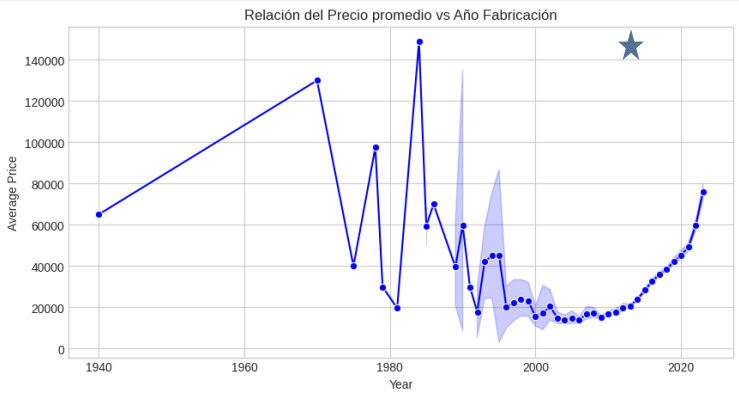



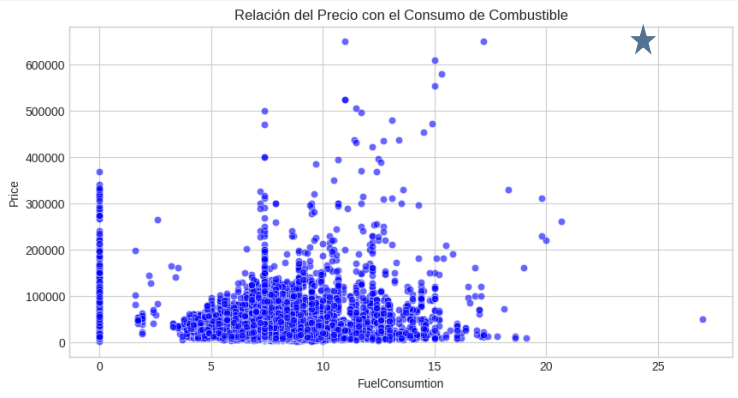



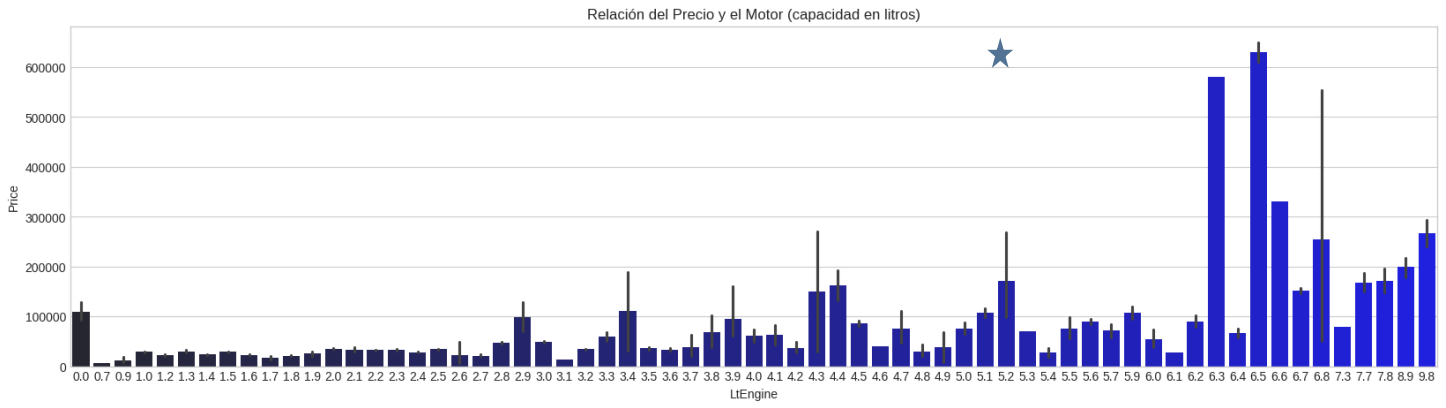

# EXPLORATION DATA ANALYSIS (EDA)

## BIVARIADO

### Relación del Precio con las Variables Numéricas



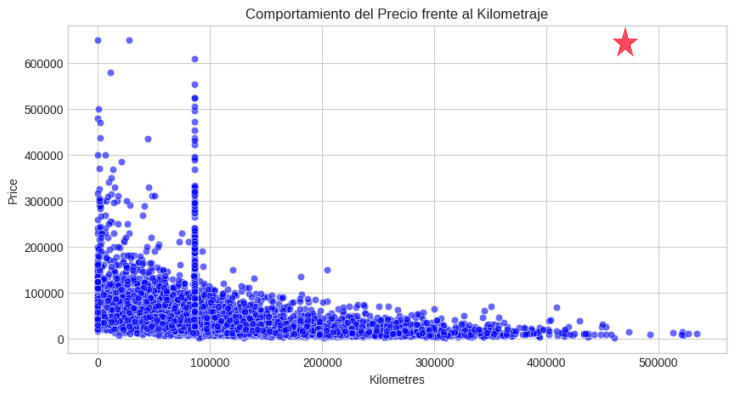

★ Se cumple la Hipótesis

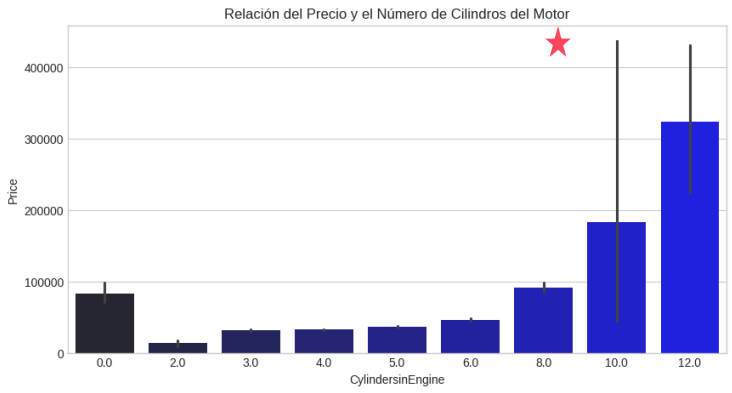
★ No se cumple la Hipótesis

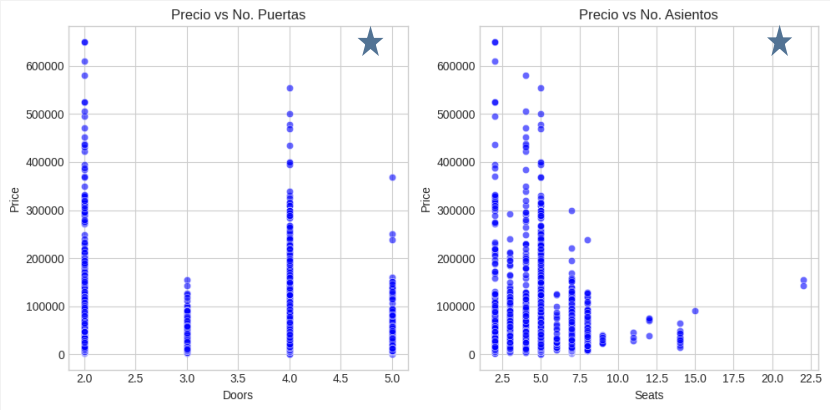
# EXPLORATION DATA ANALYSIS (EDA)

## BIVARIADO

### Relación del Precio con las Variables Numéricas



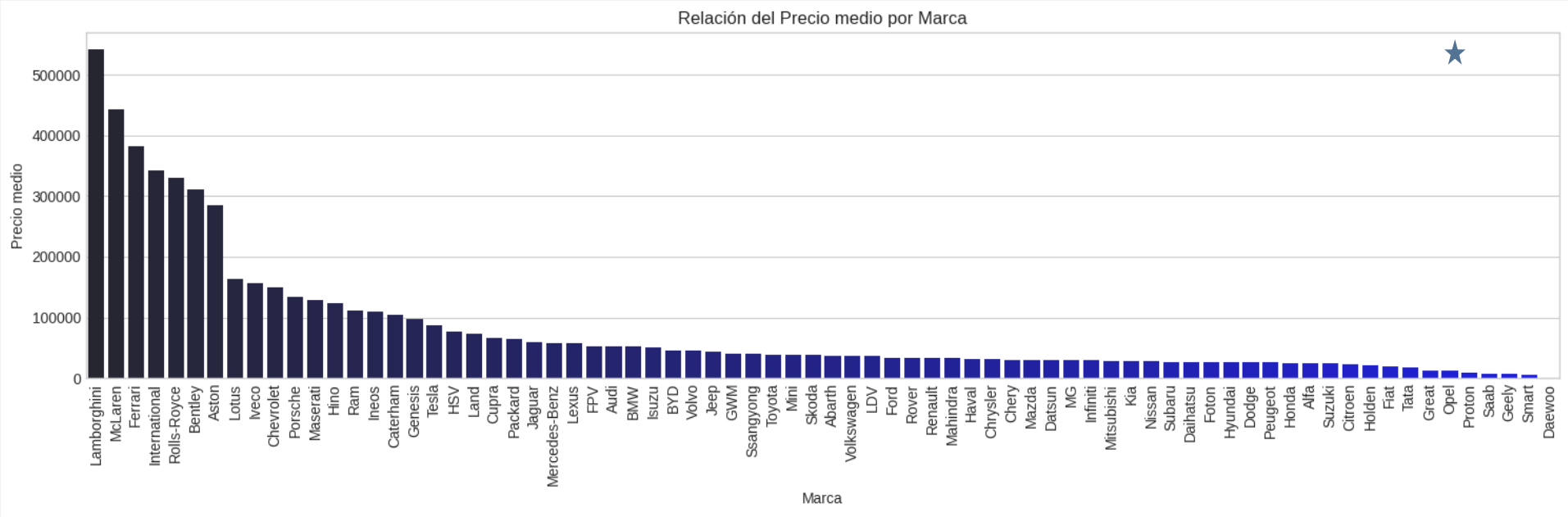

★ Se cumple la Hipótesis

★ No se cumple la Hipótesis

# EXPLORATION DATA ANALYSIS (EDA)

## BIVARIADO

### Relación del Precio con las Variables Categóricas




★ Se cumple la Hipótesis

★ No se cumple la Hipótesis

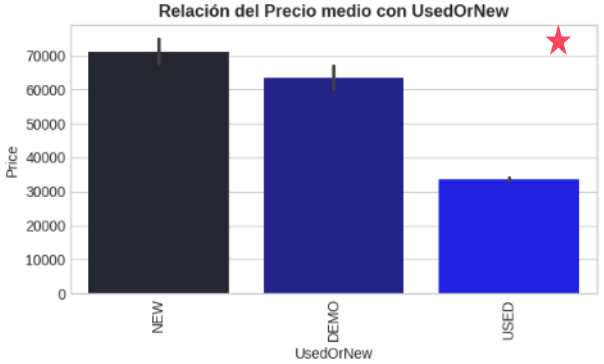
# EXPLORATION DATA ANALYSIS (EDA)

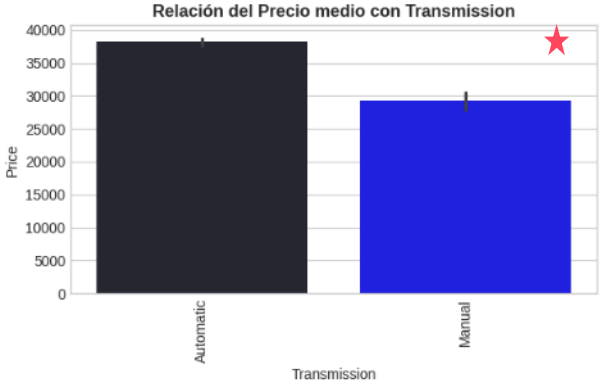
## BIVARIADO

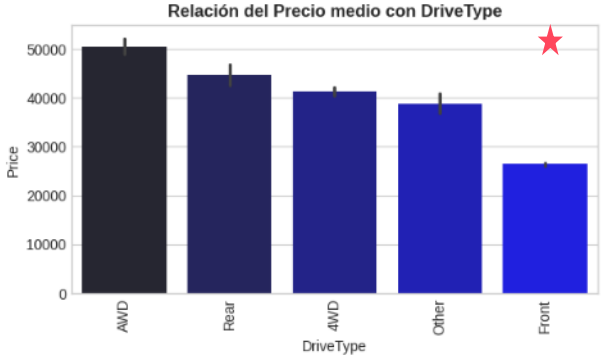
★ Se cumple la Hipótesis

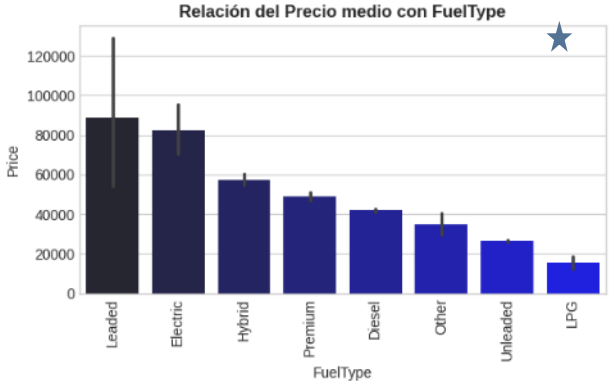
★ No se cumple la Hipótesis

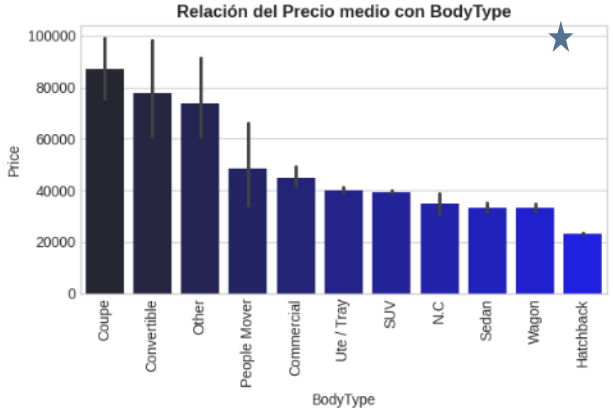
### Relación del Precio con las Variables Categóricas

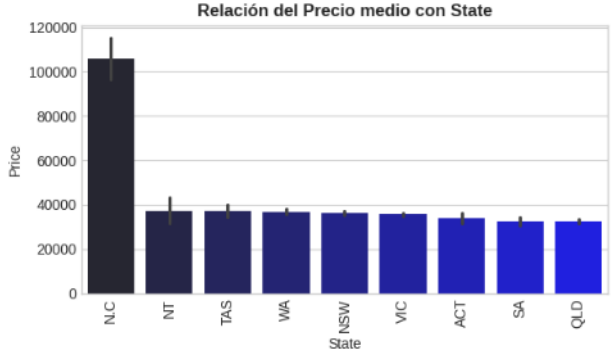


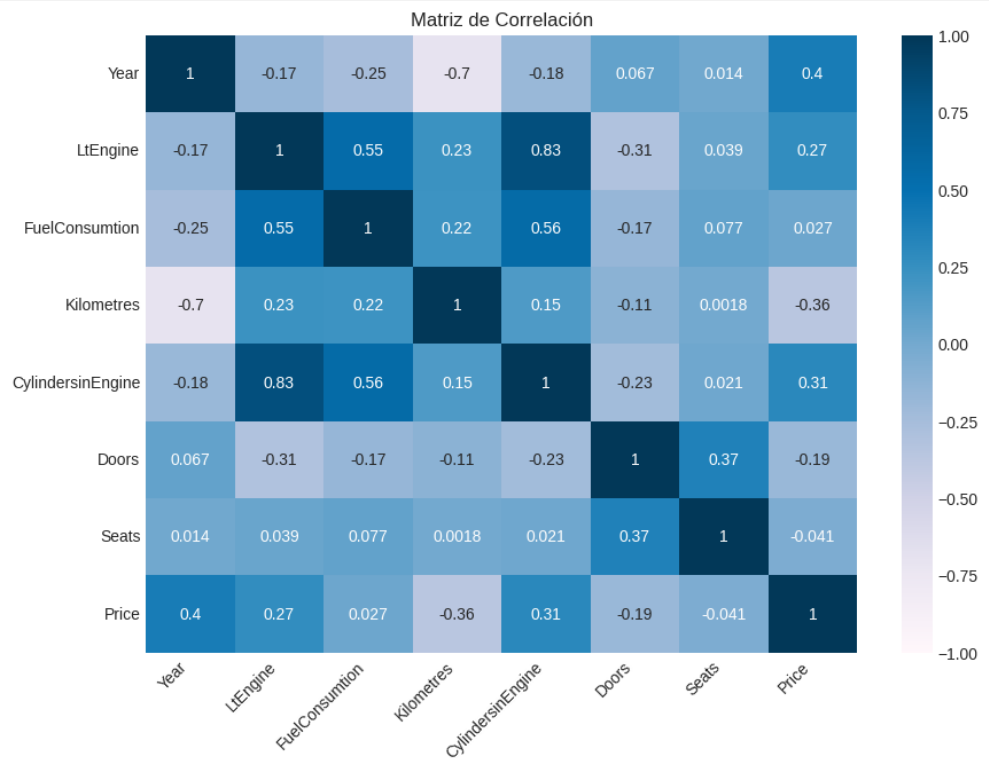



# EXPLORATION DATA ANALYSIS (EDA)

## MULTIVARIADO




**Variables con Corr (+)**

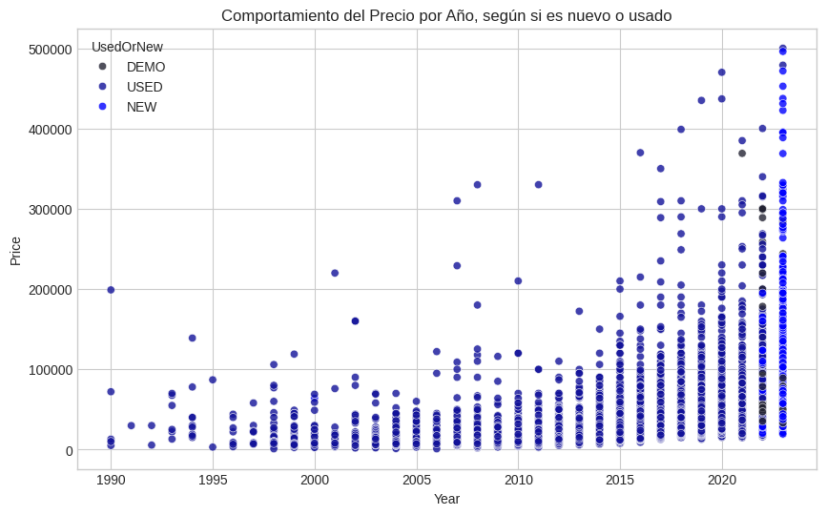
| | |
|---|---|
| Year | 0.40 |
| CylindersinEngine | 0.31 |
| LtEngine | 0.27 |
| FuelConsumtion | 0.03 |

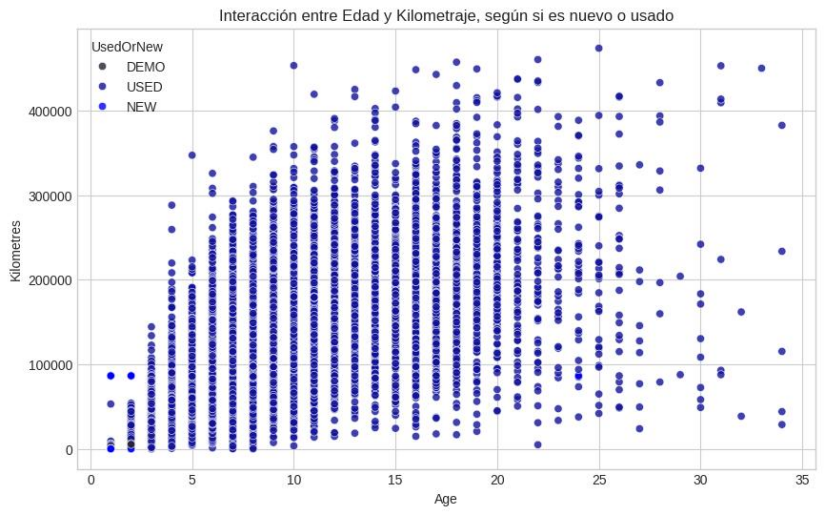
**Variables con Corr (-)**

| | |
|---|---|
| Kilometres | -0.36 |
| Doors | -0.19 |
| Seats | -0.04 |

# EXPLORATION DATA ANALYSIS (EDA)

## MULTIVARIADO

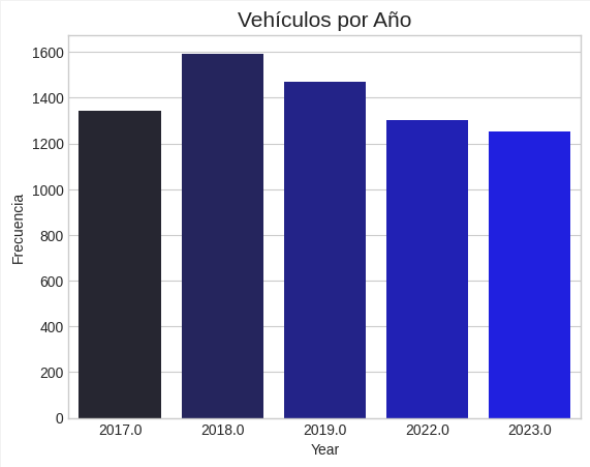



# EXPLORATION DATA ANALYSIS (EDA)

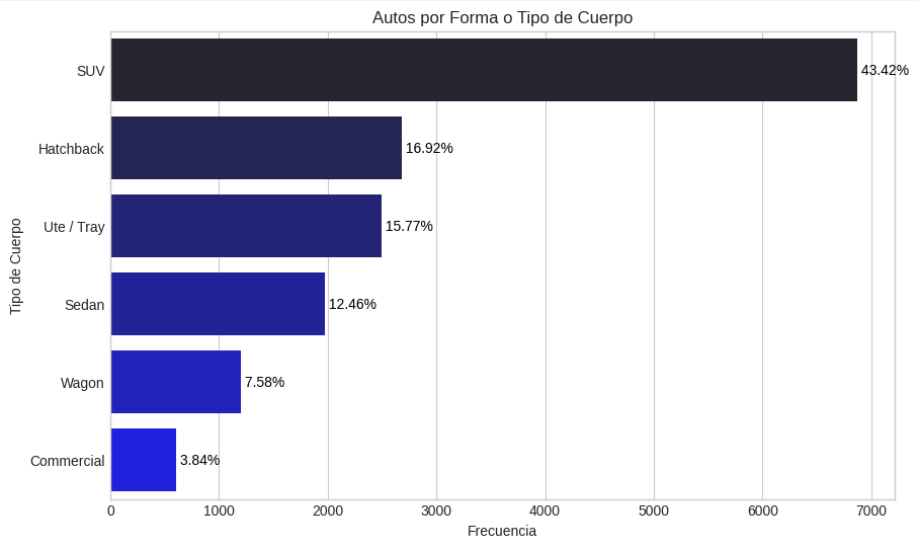
## RESPUESTA A PREGUNTAS

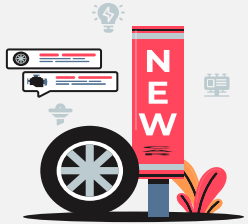
1.




| | Year | Frecuencia |
|---|---|---|
| 0 | 2018.0 | 1594 |
| 1 | 2019.0 | 1472 |
| 2 | 2017.0 | 1344 |
| 3 | 2022.0 | 1303 |
| 4 | 2023.0 | 1251 |

2.

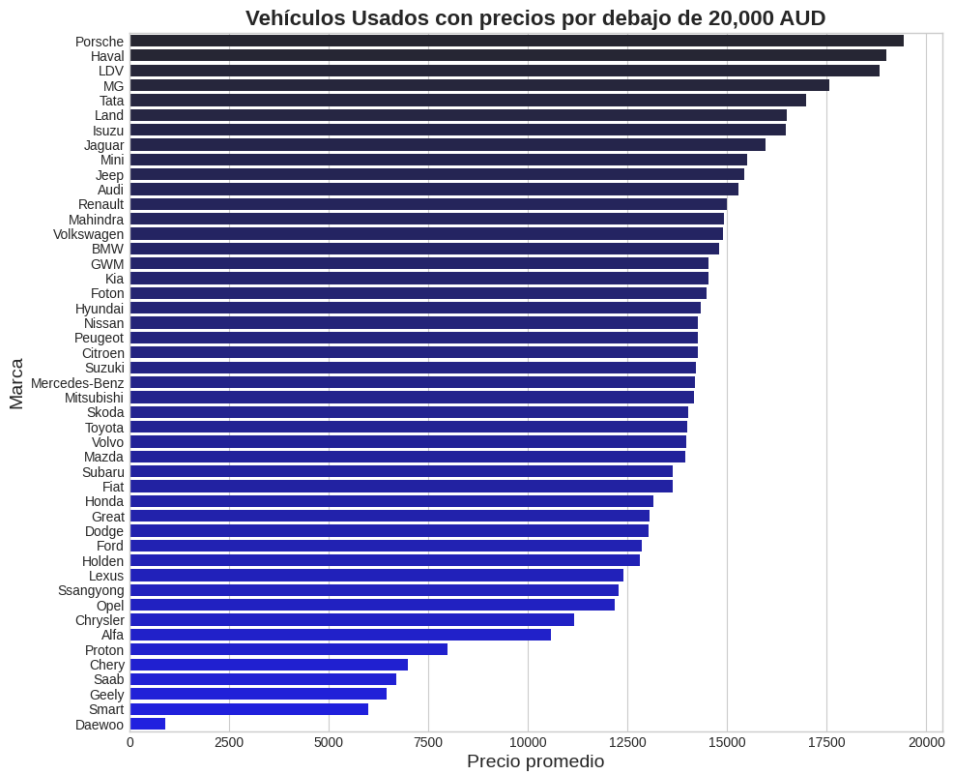


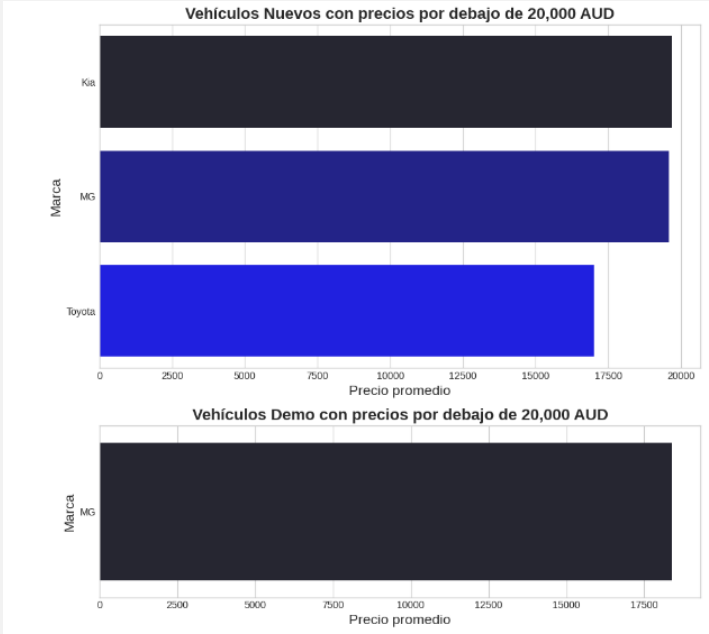



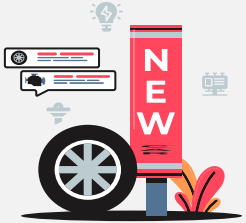
# EXPLORATION DATA ANALYSIS (EDA)

## RESPUESTA A PREGUNTAS

3.



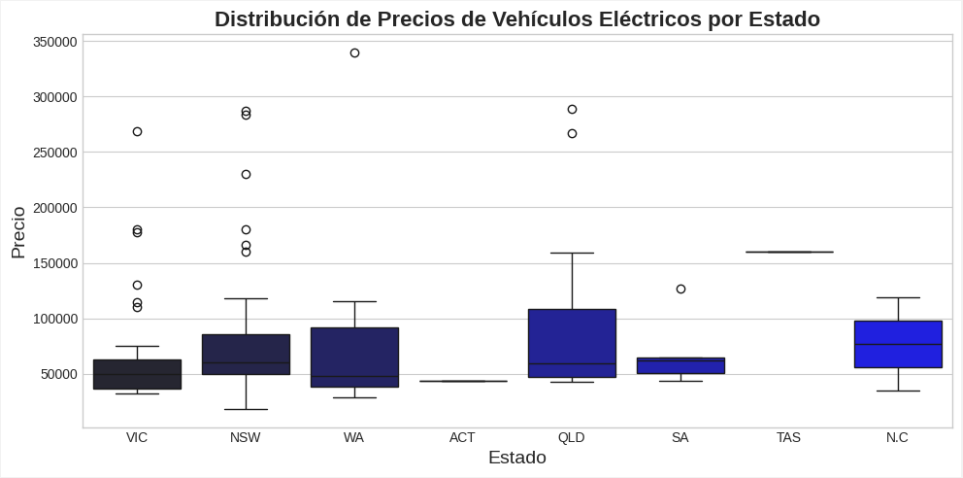
# EXPLORATION DATA ANALYSIS (EDA)

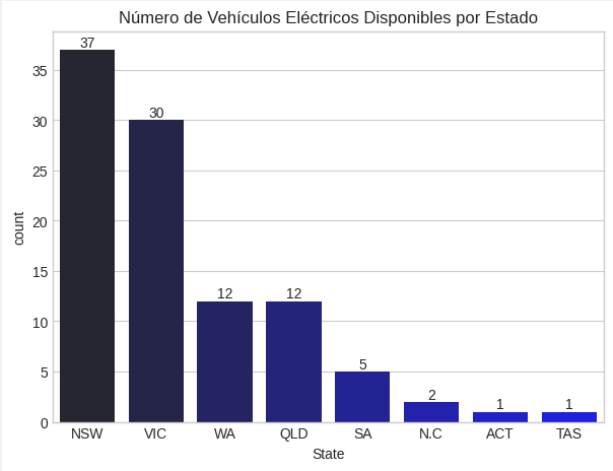
## RESPUESTA A PREGUNTAS

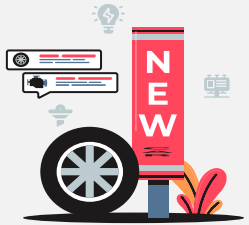
4.

El vehículo eléctrico más económico es:

| | |
|---|---|
| Brand | Nissan |
| Year | 2014.0 |
| UsedOrNew | Used |
| Transmission | Automatic |
| LtEngine | 0.0 |
| DriveType | Other |
| FuelType | Electric |
| FuelConsumtion | 7.4 |
| Kilometres | 63007.0 |
| CylindersinEngine | 4.0 |
| BodyType | Hatchback |
| Doors | 4.0 |
| Seats | 5.0 |
| Price | 17995.0 |
| State | NSW |
| Age | 10.0 |



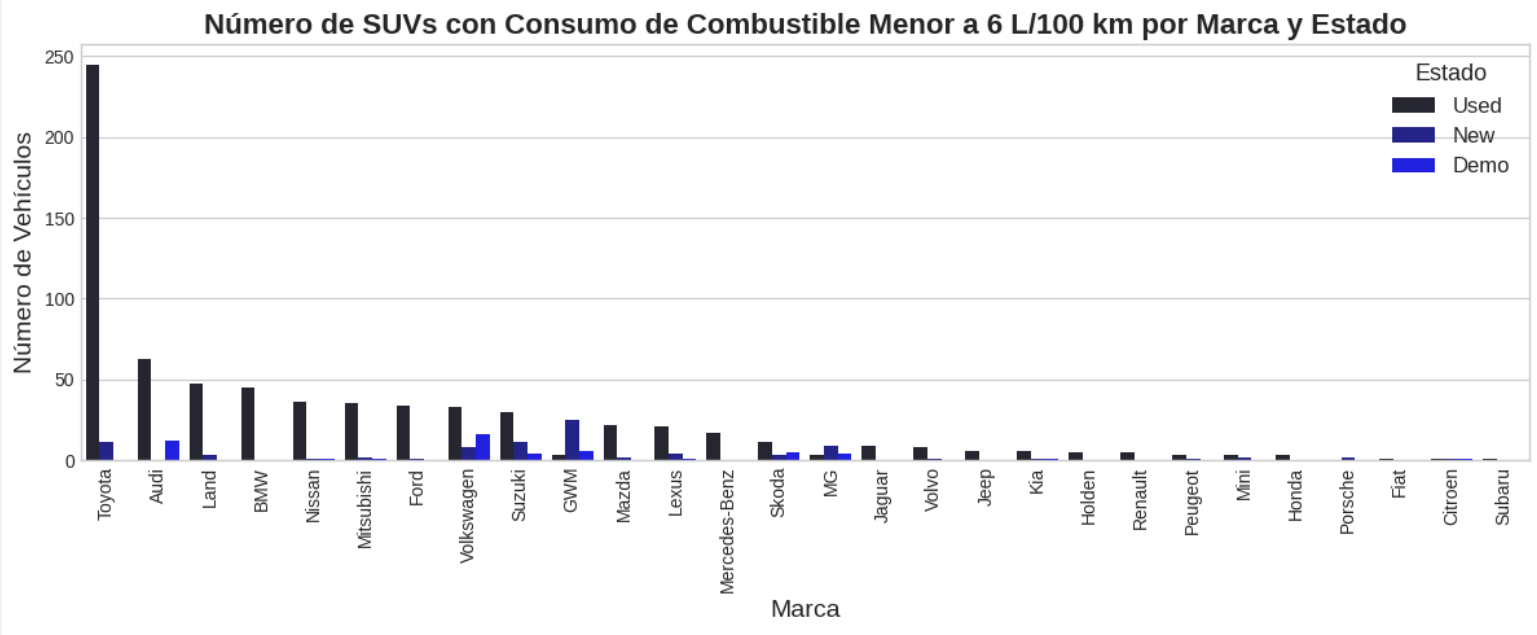



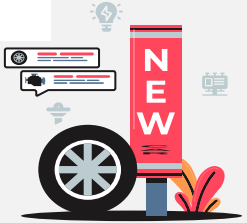



# EXPLORATION DATA ANALYSIS (EDA)

## RESPUESTA A PREGUNTAS

5.



# INSIGHTS

Las variables que más influyen en el precio de un vehículo son: el año de fabricación (años más recientes, vehículos más costosos), tamaño del motor (aumenta el precio a mayor tamaño de motor y mayor número de cilindros), Kilometraje (a menor kilometraje, mayor costo del vehículo) que sea nuevo y de transmisión automática (aumenta el costo respectivamente).

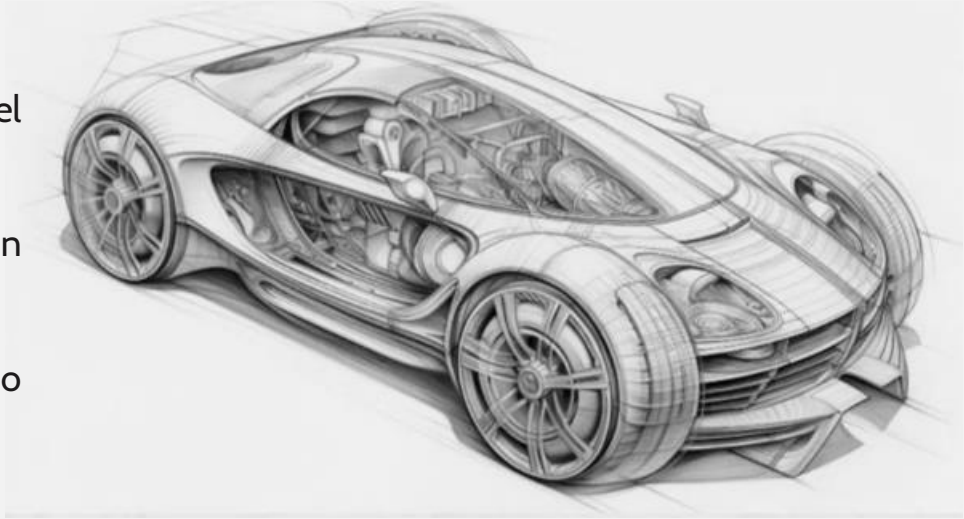
El precio promedio del 99.99% de los vehículos de este estudio está por debajo de los 200,000 AUD.

Las marcas de lujo como Lamborghini y Ferrari presentan los precios más elevados dentro de los vehículos de este estudio.

Los vehículos tipo SUV son los más ofertados dentro del dataset.

Los vehículos que más consumen combustible suelen tener precios más elevados.

Los vehículos más compactos presentan un mayor costo en relación con los demás.

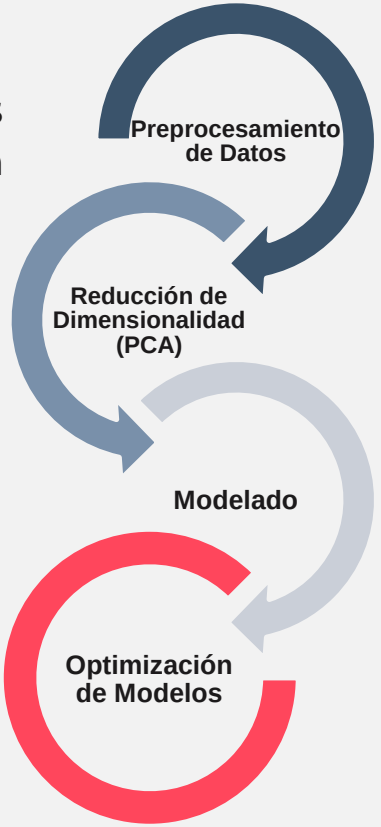


# MODELADO DE ML

A continuación, se presenta un resumen con los resultados de las métricas obtenidas para cada uno de los modelos de ML, con una división de 80/20 para Train/Test:

| | Model | MSE_Test | RMSE_Test | MAE_Test | MAPE_Test | RAE_Test | R2_Test | R2_Ajustado_Test |
|---|---|---|---|---|---|---|---|---|
| 0 | LinearRegression | 6.460627e+08 | 25417.763638 | 12664.551639 | 46.658092 | 0.650003 | 0.464662 | 0.461405 |
| 1 | RandomForestRegressor | 2.518510e+08 | 15869.814284 | 6724.012837 | 18.690433 | 0.345107 | 0.791312 | 0.790042 |
| 2 | SVR | 1.248934e+09 | 35340.262482 | 17854.404173 | 60.289879 | 0.916371 | -0.034887 | -0.041184 |
| 3 | GradientBoostingRegressor | 3.183198e+08 | 17841.518271 | 8569.448927 | 25.392567 | 0.439824 | 0.736235 | 0.734630 |
| 4 | XGBoost | 2.567954e+08 | 16024.838135 | 7056.245174 | 20.429313 | 0.362159 | 0.787215 | 0.785920 |

Los mejores modelos en términos de precisión y rendimiento son Random Forest Regressor y XGBoost; aunque ambos modelos parecen estar ligeramente sobre-ajustados.



# MODELADO DE ML

**OPTIMIZACIÓN DE MODELOS :** Se decide utilizar Randomized Search y Grid Search

| Model | MSE_Test | RMSE_Test | MAE_Test | MAPE_Test | RAE_Test | R2_Test | R2_Ajustado_Test |
|---|---|---|---|---|---|---|---|
| XGBoost | 2.567954e+08 | 16024.838135 | 7056.245174 | 20.429313 | 0.362159 | 0.787215 | 0.785920 |
| RandomForestRegressor | 2.518510e+08 | 15869.814284 | 6724.012837 | 18.690433 | 0.345107 | 0.791312 | 0.790042 |

Resultados antes de la Optimización

Resultados después de la Optimización

| Model | MSE_Test | RMSE_Test | MAE_Test | MAPE_Test | RAE_Test | R2_Test |
|---|---|---|---|---|---|---|
| RandomForestRegressor | 2.594782e+08 | 16108.327910 | 6706.553821 | 18.379927 | 0.344211 | 0.784992 |
| XGBoost | 2.294712e+08 | 15148.305874 | 6657.431468 | 18.797800 | 0.341690 | 0.809857 |

El modelo de **XGBoost** es el que logra obtener mejores métricas luego de realizar la optimización de hiperparámetros, por tanto, es el elegido.

**XGBoost logra explicar aproximadamente el 80,99% de la variabilidad en los datos de prueba.**

# CONCLUSIONES

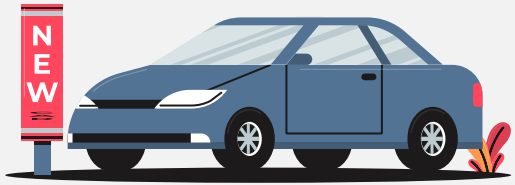
- A través del EDA pudimos conocer el comportamiento univariado de nuestra variable objetivo (Precio), pero también su relación frente a cada una de las demás categorías.

- En general, podemos evidenciar que el año de fabricación incrementa el precio del vehículo cuando es más reciente; si el tamaño del motor aumenta, en general también el precio; y entre menor sea el kilometraje, el precio del vehículo tiende a ser mayor.

- El uso de PCA ayudó a reducir la dimensionalidad de los datos sin perder mucha información, lo que facilitó el entrenamiento de los modelos y potencialmente mejoró su rendimiento al eliminar el ruido y las características redundantes.

- La optimización de hiperparámetros ayudó a mejorar significativamente las métricas de rendimiento, especialmente para el modelo XGBoost.

- El modelo XGBoost es el modelo preferido porque es el que obtuvo mejores métricas en el conjunto de prueba y de entrenamiento luego de la optimización.



# GRACIAS