17 DE MAYO DE 2022

ANÁLISIS DEL MERCADO DE VEHÍCULOS USADOS

SAÚL ARRANZ HERERO

THE BRIDGE

Análisis del mercado de vehículos usados

Introducción

En los últimos meses/años el mercado de vehículos de segunda mano ha tenido gran importancia debido a la crisis de componentes que ha afectado a la fabricación de nuevos vehículos, así como la incertidumbre creada sobre cuál puede ser el combustible del futuro o las diferentes crisis que hemos afrontado (pandemia, guerra Ucrania), todo esto hace dudar a la hora de comprar un vehículo tanto nuevo como de segunda mano. Por esta razón vamos a realizar un estudio sobre el precio de los vehículos de segunda mano en función de diferentes variables como puede ser, Combustible, kilómetros, Marca, Vendedor, Comunidad Autónoma...

En la plataforma de datos datamarket.es podemos encontrar los datos de los vehículos de segunda mano a la venta de las principales plataformas.

Los datos utilizados en este EDA se han obtenido de la base de datos (https://datamarket.es/#coches-de-segunda-mano-dataset).

El conjunto de datos consta de marcas, modelos y versiones de gran cantidad de vehículos, así como potencia, kilómetros, año, precio, vendedor...

Tabla de contenido

1. Business case & Data Collection

Hipótesis.

- Los vehículos con más km son los más baratos
- La marca de coches más barata es Dacia
- Los vehículos híbridos son más caros que los Diésel
- la media de precios es la misma para todas las comunidades autónomas
- Los vehículos vendidos por profesionales son más caros que los vendidos por particulares
- Los vehículos más potentes son los más caros

Requerimientos de los datos y Calidad

Los datos que encontramos en el dataset son suficientes y nos aportan la información que necesitamos para poder confirmar o descartar las hipótesis planteadas.

El dataset no tiene excesivos missing por lo que incluso teniendo que descartarlos tendríamos una gran cantidad de datos para poder llegar a las conclusiones.

Información necesaria:

- Precio
- Kilómetros
- Marca
- Año
- Potencia

Esas serían las principales variables de las que depende el precio del vehículo.

2.Data Understanding

Tabla de variables

Las columnas que tiene el dataset:

- url: identificador del vehículo a la venta
- company: identificador de la compañía
- make: marca del vehículo a la venta
- model: modelo del vehículo a la venta
- version: modelo del vehículo a la venta
- price: precio del vehículo a la venta
- price_financed: precio del vehículo a la venta
- fuel: combustible del vehículo a la venta
- year: anyo de compra (nuevo) del vehículo a la venta
- kms: kilómetros del vehículo a la venta
- power: potencia del vehículo a la venta
- doors: nº de puertas del vehículo a la venta
- shift: tipo de cambio del vehículo a la venta
- color: color del vehículo a la venta
- photos: nº de fotos del vehículo a la venta
- is_prfessional: vendedor profesional (1) o particular (0)
- dealer: empresa que vende el vehículo
- province: provincia en la que se vende el vehículo
- country: país en la que se vende el vehículo
- publish_date: fecha de publicación del anuncio de la venta
- insert_date: fecha en la que se incluye el vehículo a la lista

3.Data Cleaning

Eliminar/Crear columnas

Antes de borrar las columnas que no queremos para el análisis, vamos a verificar que no tenemos elementos duplicados. Para verificar los elementos duplicados tenemos que hacerlo antes de borrar las columnas "url" y "company" ya que el resto de columnas puede que sí que se repitan aunque no sea el mismo vehículo (un vehículo puede tener el mismo modelo, versión y kilómetros que otro vehículo sin ser el mismo), con las columnas "url" y "company" nos aseguramos que no son el mismo vehículo.

Vamos a crear una nueva columna con la comunidad autónoma en la que se vende el vehículo, lo haremos a partir de la columna provincia, posteriormente eliminaremos la columna provincia. Esto nos permitirá tener menos categorías que nos dividan el territorio nacional.

Ahora ya podemos eliminar las columnas que no son necesarias por el momento, esas columnas son:

- url, company, doors, shift, color, photos, province, country, publish_date, insert_date

Una vez que tenemos las variables que nos interesan, renombramos las columnas con su traducción al español.

Missings

Tenemos valores nulos en varias de las columnas, trataremos estos valores columna a columna:

- Valores nulos de “Marca”:

  En este caso buscando en internet podemos comprobar que ese modelo y versión corresponde a un vehículo de la marca ``Invica Electric`` por lo que asignamos esa marca a estos 2 valores missing

- Valores nulos de “Modelo”:

  Para rellenar los nulos de la variable modelo, vamos a comparar la versión de dichos nulos con la versión de otros que sí que tengan el modelo asociado y si alguna de las palabras contenidas en 'versión' coincide con el valor de modelo, podremos asignar ese valor a los modelos nulos. Para realizar esto operación antes tenemos que quitar los acentos de la columna modelo.

  Se han realizado 2 funciones para completar esta columna, una para quitar acentos y otra para buscar y reemplazar los valores en “Modelo”.

- Valores nulos de la variable Potencia:

  Para rellenar los valores nulos de la variable Potencia en primer lugar listaremos los valores de Modelo y Combustible de los valores nulos de potencia, con esos valores de modelo y combustible calcularemos la media de las potencias de los vehículos de los que sí que tenemos ese dato y después asignaremos el valor de la media sobre los vehículos con potencia nula.

  También creamos una función para realizar esta operación.

  Algunos valores de potencia no se han podido completar ya que no había datos de ese modelo y combustible con potencia, por esto comprobamos cuantos valores de

potencia siguen quedando nulos y los eliminamos ya que no tenemos manera de rellenarlos con cierta fiabilidad.

Después de quitar los valores nulos de potencia ya solo quedarían 2 datos nulos en Combustible, 2 más en Anyo y 6 más en Comunidad_Autonoma que anteriormente denominamos como desconocida. Eliminamos estos datos ya que no es posible rellenarlos de forma fiable.

Por último, ya solo nos faltaría eliminar las columnas Precio_financiado, Modelo y Version (ya no son necesarias) y cambiar el tipo de dato de las variables "'Anyo'" y "Vendedor_profesional" que están como numéricas pero son variables categóricas.

4.Analysis

Analysis univariante

Variables numércias

Precio.

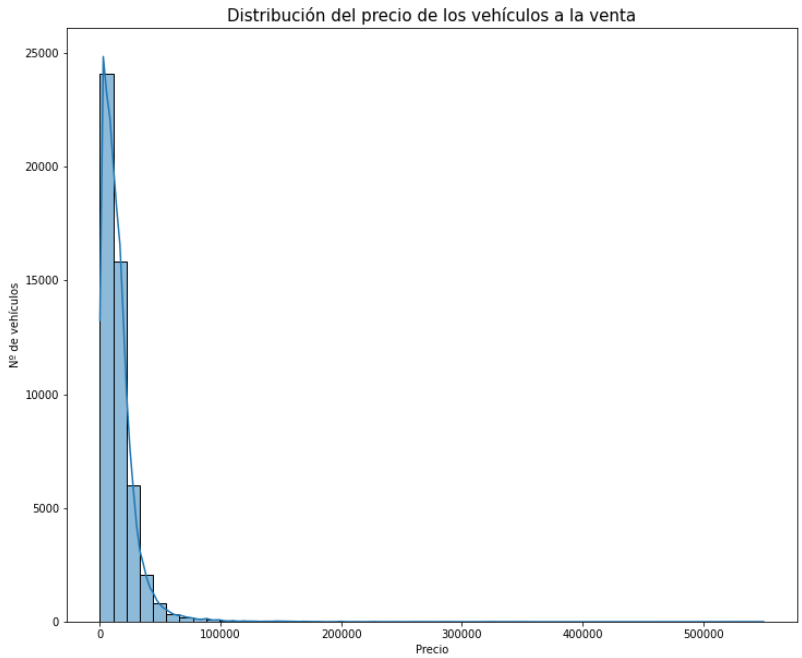

Ilustración 1

En la ilustración 1 vemos la distribución de la variable “Precio”. La mayor parte de los precios se encuentran entre 0 y 50000€, hay unos pocos vehículos con un precio muy elevado, esto hace que los precios se concentren en la primera parte de la gráfica. Los precios no siguen una distribución normal

Potencia

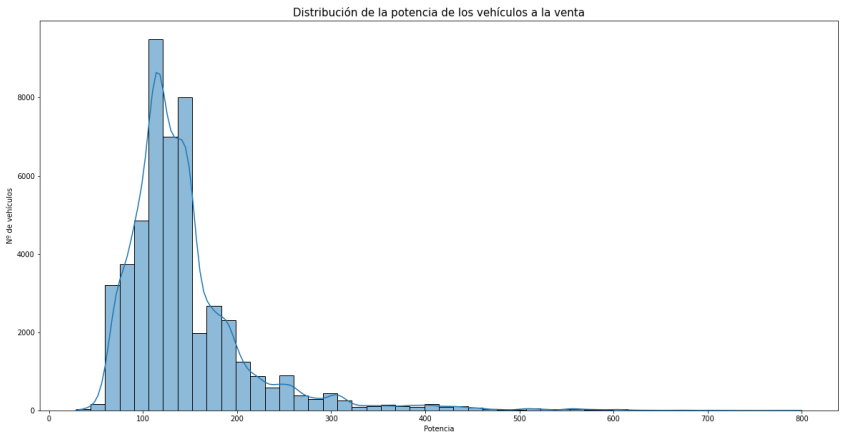

Ilustración 2

Kilómetros

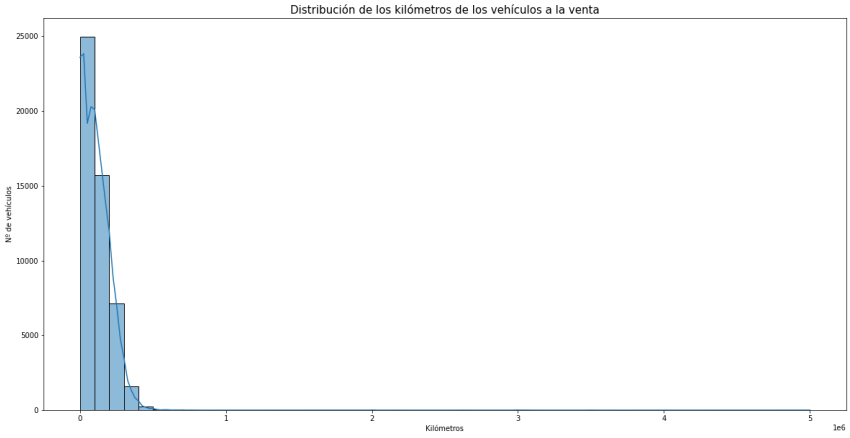

Ilustración 3

Las variables numéricas tienen muchos valores outliers por lo que los vamos a eliminar para poder realizar un mejor análisis ya que lo que nos interesa es hacer un análisis general sin tener en cuenta vehículos exclusivos como pueden ser los que tienen mucha

potencia o los que cuestan mucho dinero, o vehículos muy usados como son los que tienen muchos kilómetros,

Para eliminar los valores outliers se ha creado una función con el siguiente criterio:

* Sean superiores a Q_3 + 1.5 x RI

* Sean inferiores a Q_1 - 1.5 x RI

Quedan los límites que aparecen en la tabla:

	Precio	Kilometros	Potencia
num_outliers	2195	475	3836
valor_min	-16085	-165750	32.5
valor_max	41475	377850	236.5

Tabla 1

Una vez que hemos quitado los outliers, volvemos a mostrar las distribuciones de las variables.

Precio sin outliers

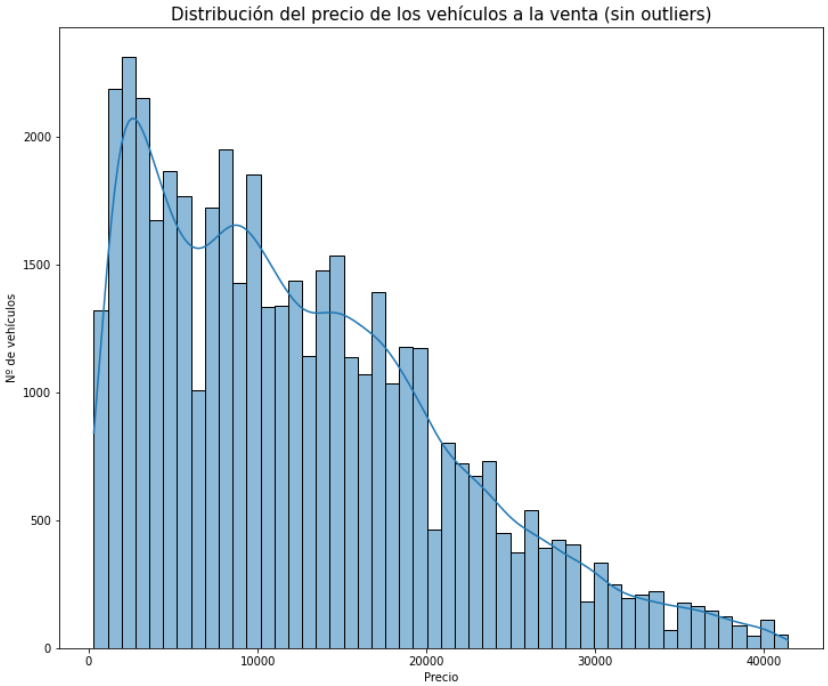

Ilustración 4

En la ilustración 4 podemos ver que aun quitando los outliers, la distribución de los precios sigue sin ser una distribución normal, en cualquier caso, realizamos un normmaltest de scipy para confirmarlo de manera estadística. El test shapiro no sería adecuado ya que no es preciso para muestras superiores a 5000 observaciones. El p_valor sale igual a 0 por lo que se descarta la hipótesis nula

$$H_0 = \text{La distribución es normal}$$

$$H_1 = \text{La distribución no es normal}$$

Potencia sin outliers

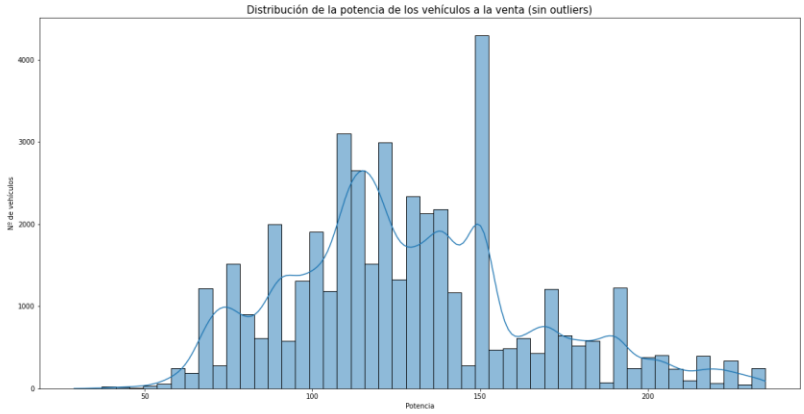

Ilustración 5

Kilómetros sin outliers

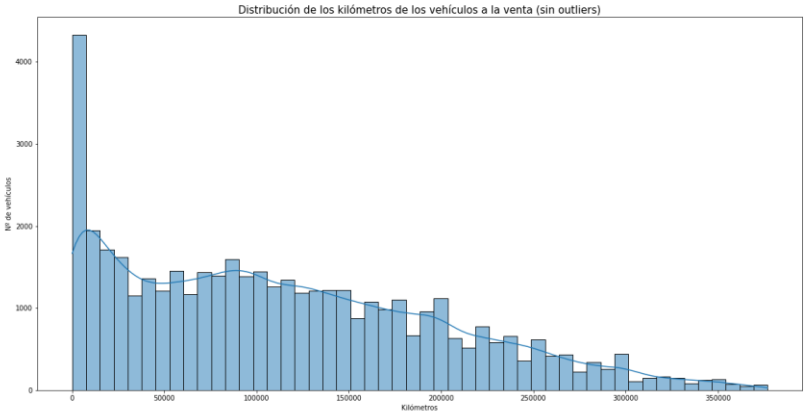

Ilustración 6

Variables Categóricas

Marca

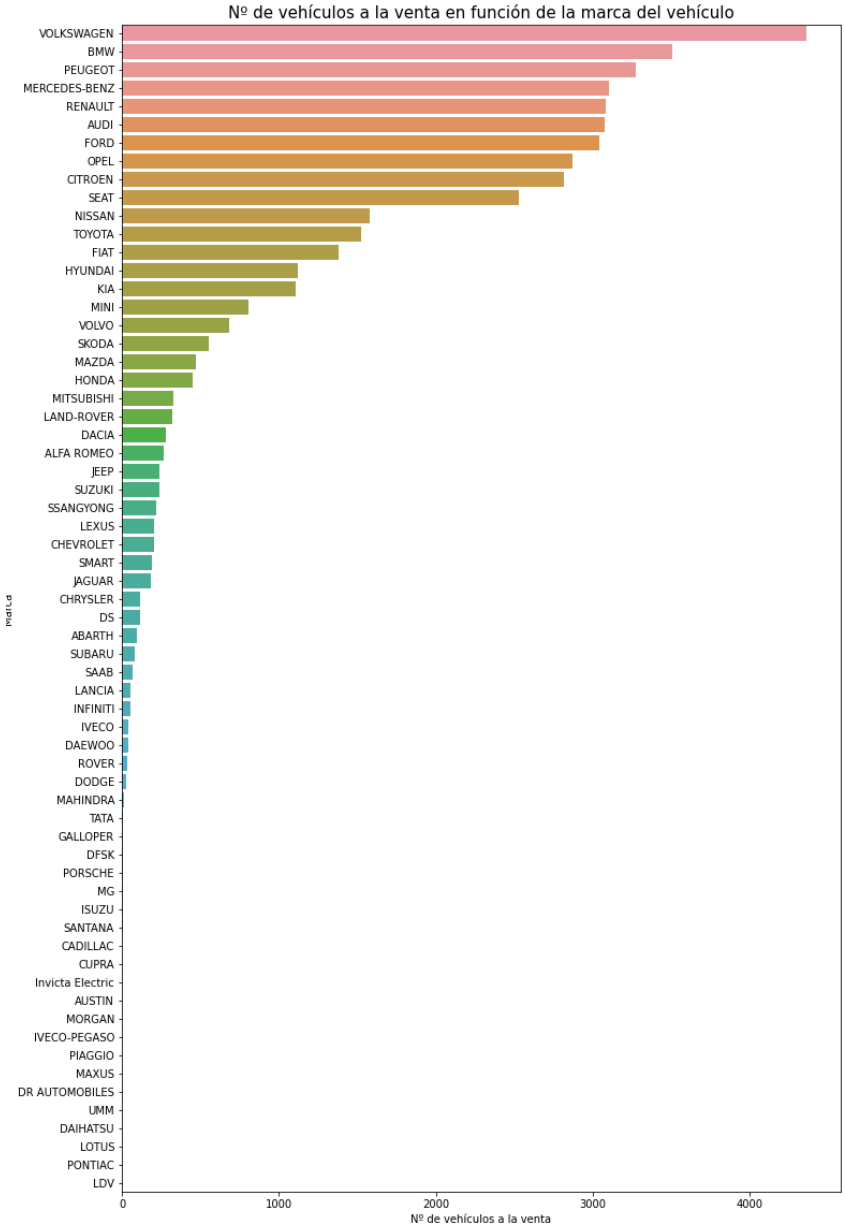

Ilustración 7

En la ilustración 7 se ve claramente un salto entre las 10 marcas con más vehículos a la venta y el resto de las marcas. Hay muchas marcas de vehículos que tienen muy pocos coches a la venta

Combustible

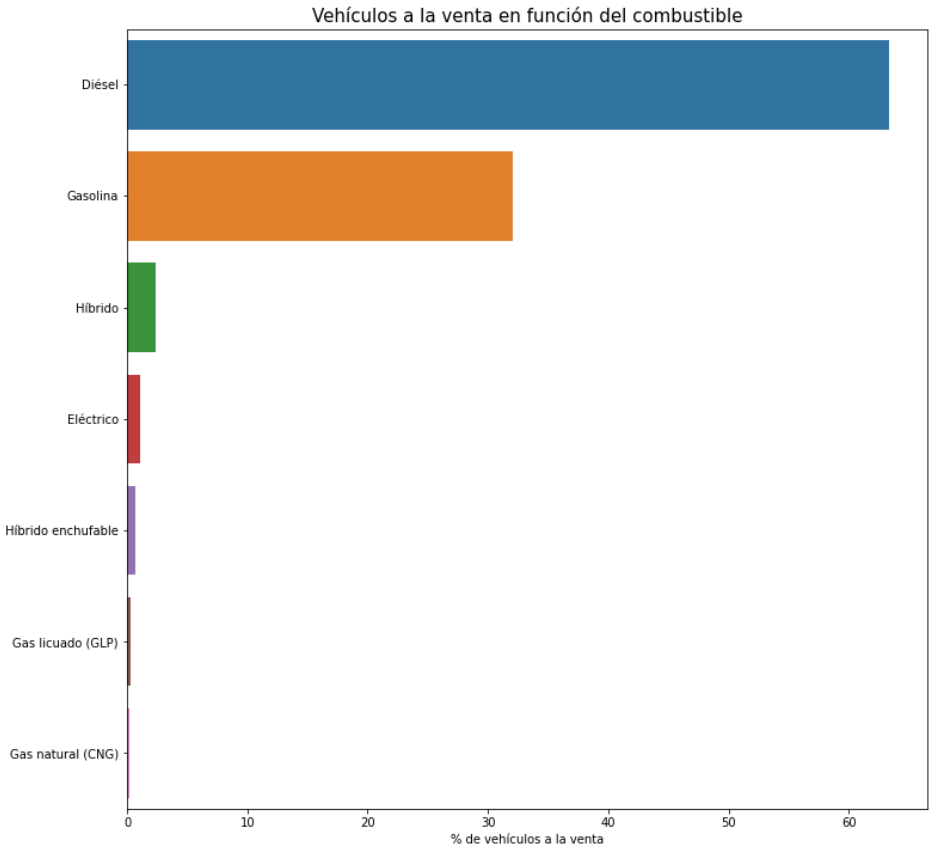

Ilustración 8

En la ilustración 8 mostramos el grafico en % para ver el reparto. En el mercado de vehículos de 2ª mano, los vehículos Diésel y Gasolina siguen siendo los grandes protagonistas, es lógico ya que el "boom" de vehículos híbridos y eléctricos ha sido en los últimos 3 años, por lo que todavía hay pocos vehículos de 2ª mano.

'Anyo

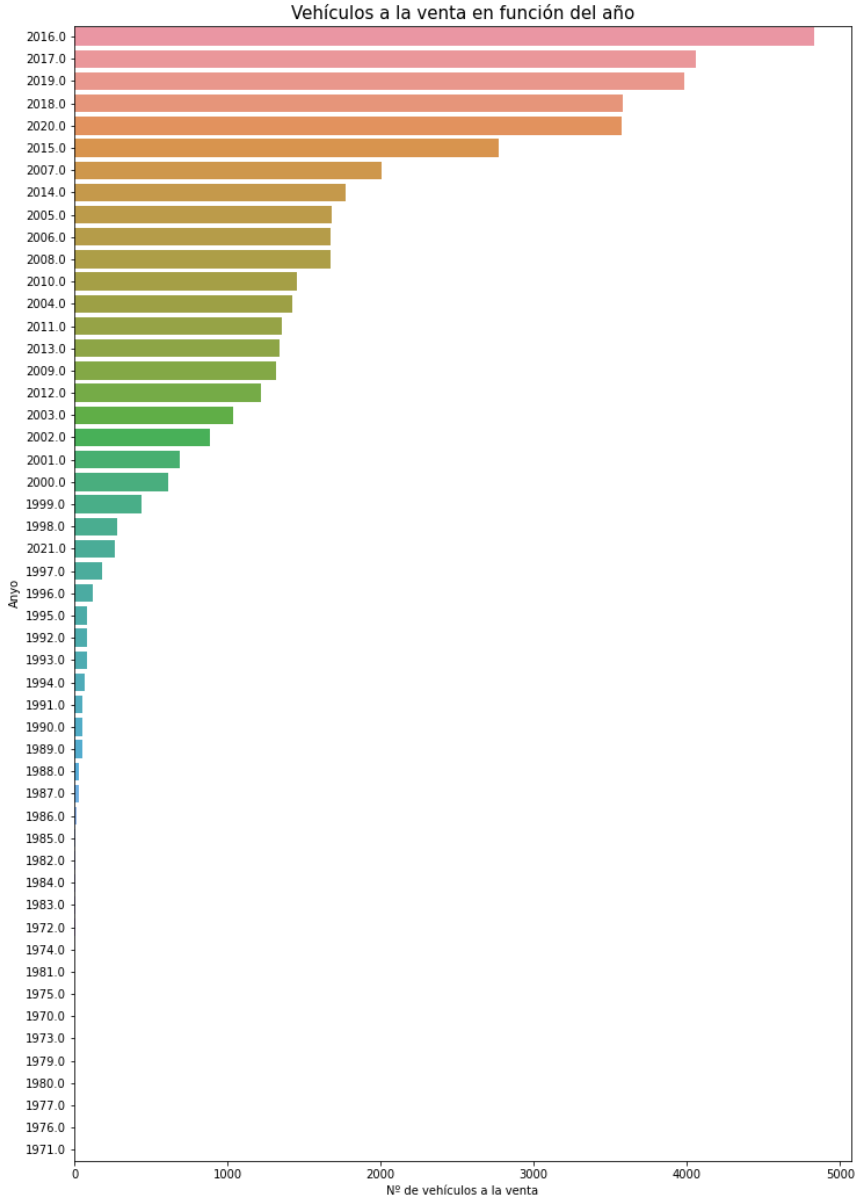

Ilustración 9

Los vehículos que más se venden son los que tienen una antigüedad de entre 2 y 7 años

Comunidad Autónoma

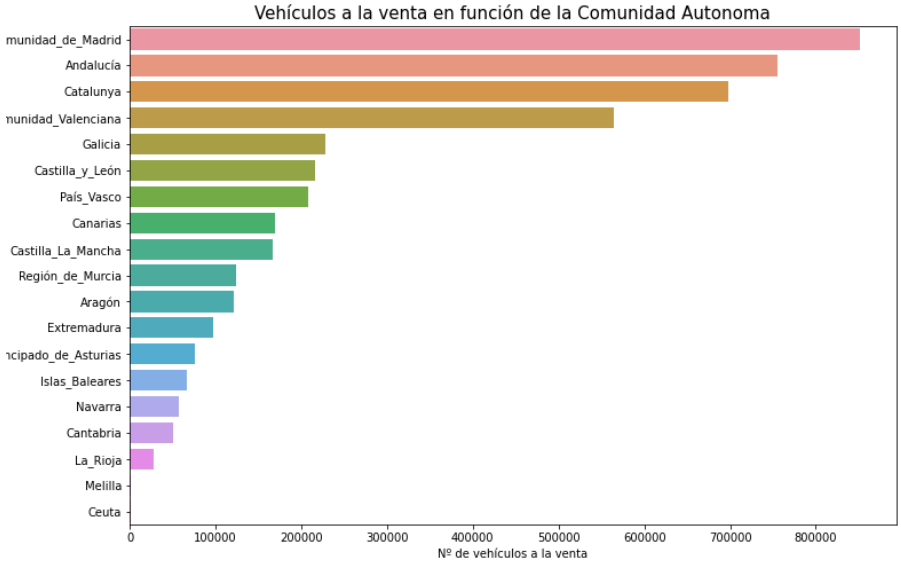

Ilustración 10

Sobre la ilustración 10 se ve claro como las comunidades con más habitantes son las que más vehículos tienen a la venta.

En la gráfica de comunidades autónomas vemos como Melilla y Ceuta no tienen prácticamente vehículos a la venta comparados con el resto de las comunidades. Como en el análisis vamos a tener en cuenta las comunidades, vamos a eliminar estas 2 ciudades autónomas que aportan muy pocos datos al global.

Haremos lo mismo con los vehículos de combustible gas, ya que como hemos visto en la gráfica de combustible aportan muy pocos datos al global

Analysis bivariante

Mapa calor variables numéricas

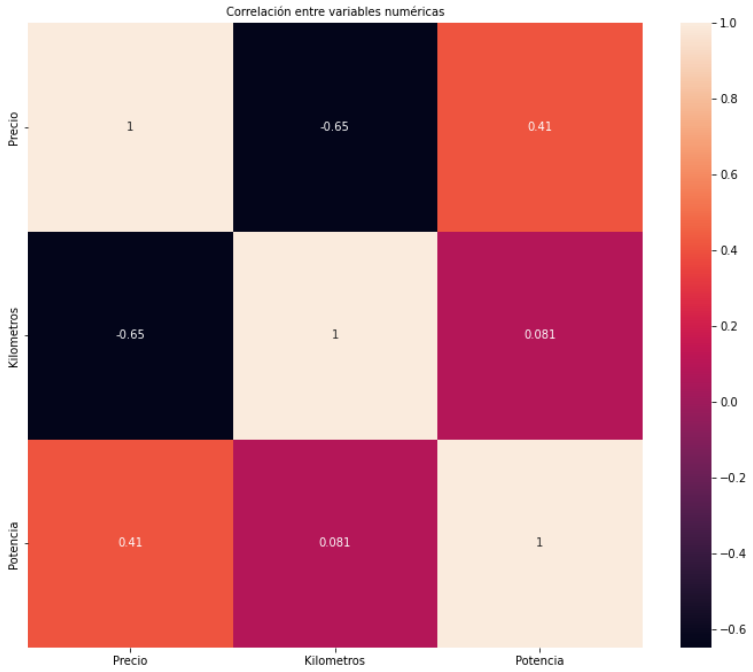

Ilustración 11

Se ve cierta relación entre los kilómetros y el precio (a más kilómetros, menor precio).

También se ve cierta relación entre el Precio y la potencia, pero en menor medida

Kilómetros vs Precio

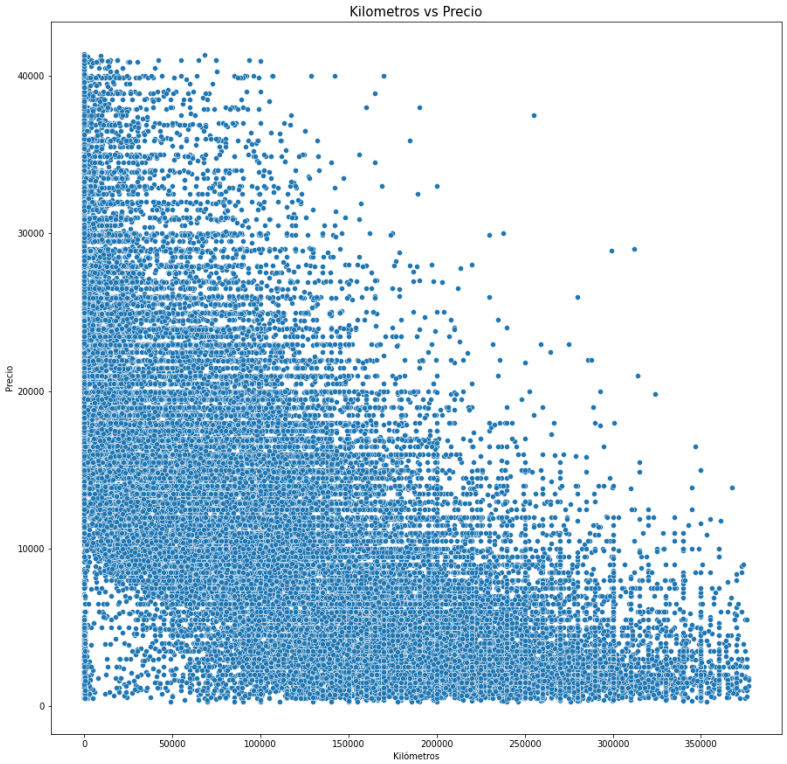

Ilustración 12

Como ya hemos visto en el mapa de calor, a medida que aumentan los kilómetros disminuye el precio con bastante relación. los puntos muestran mucha amplitud de precios para los mismos kilómetros, más adelante investigaremos a qué puede ser debido.

Combustible vs precio

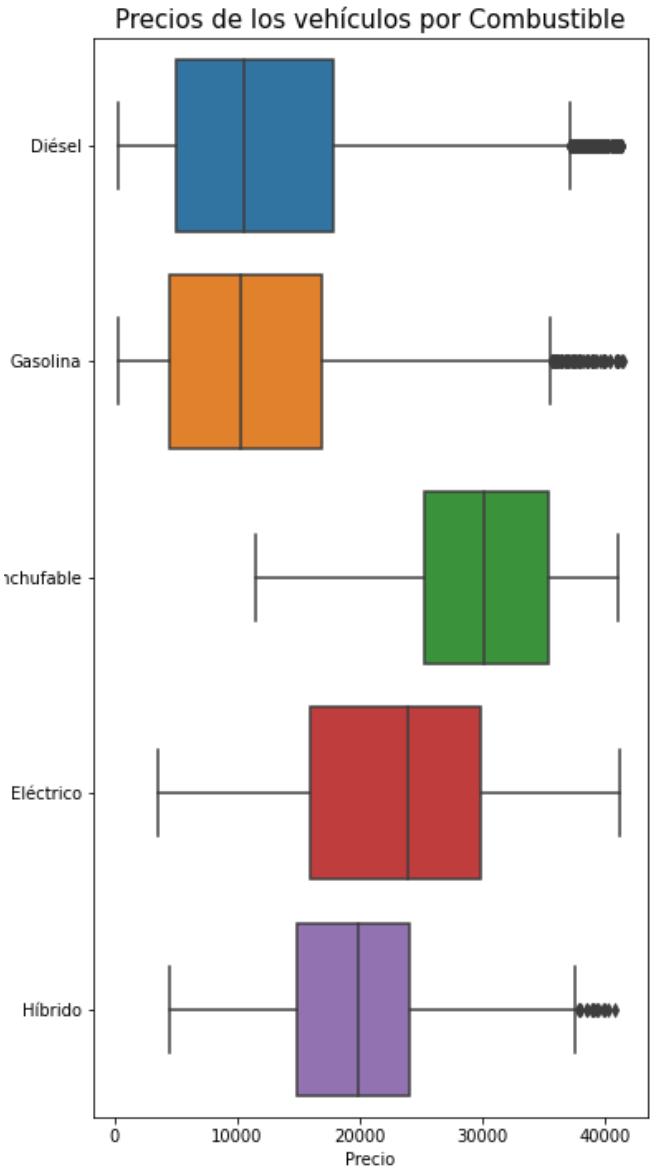

Ilustración 13

De manera general, los coches híbridos y eléctricos son más caros que los diésel o gasolina, más adelante también investigaremos cual puede ser la causa.

Comunidad Autónoma vs Precio

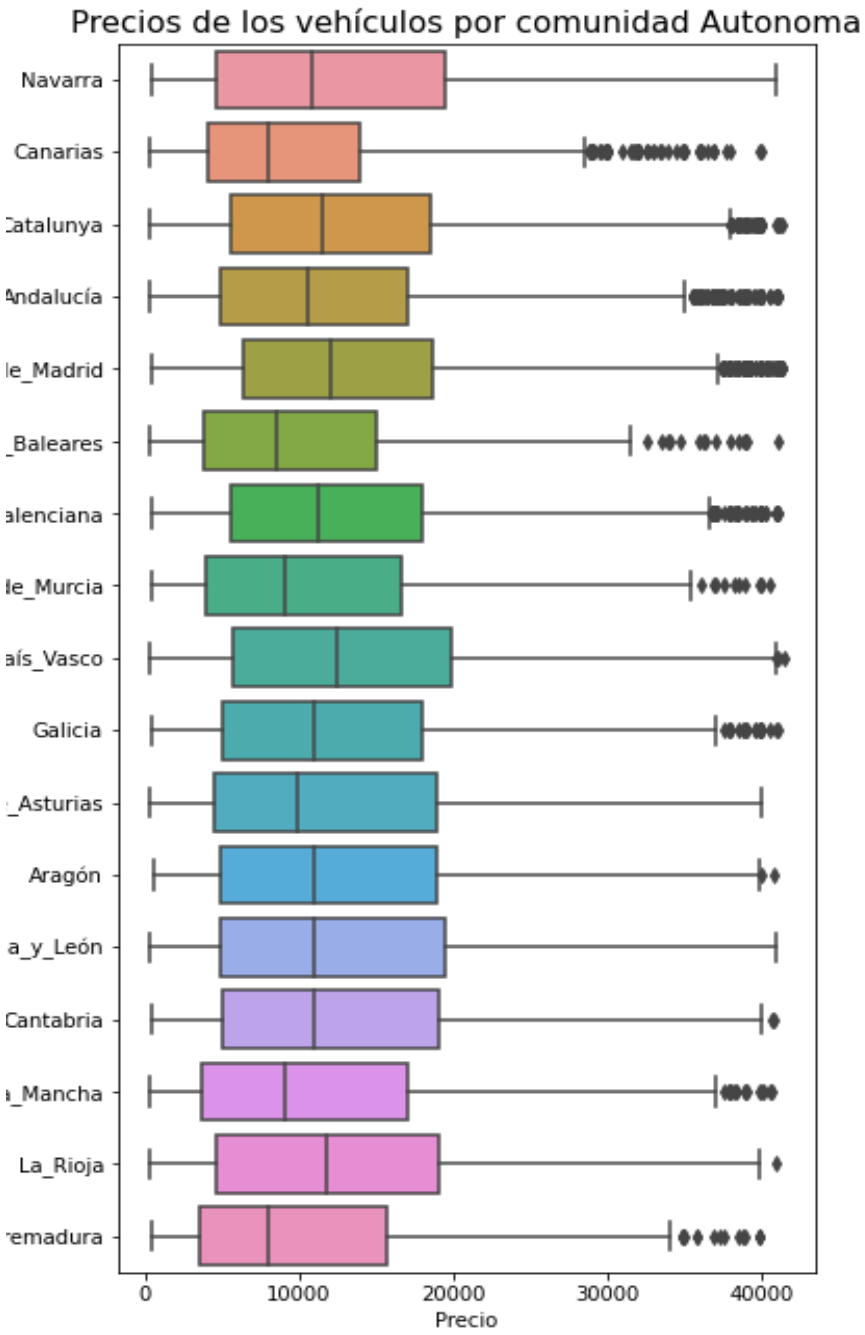

Ilustración 14

Sobre la gráfica, la mayor parte de las comunidades tiene una media de precios similar

Vendedor Profesional vs Precio

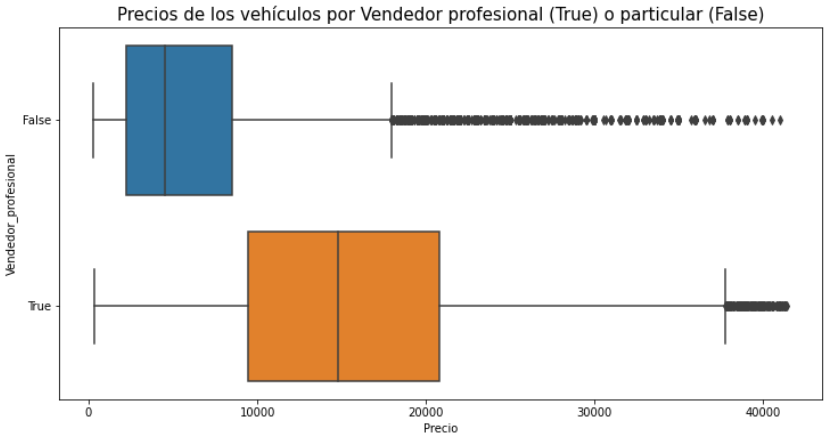

Ilustración 15

Los precios de los vehículos vendidos por profesionales son mayor que los precios de los vehículos vendidos por particulares

5.Resultados

Contraste de hipótesis

Los vehículos con más kilómetros son los más baratos

Como hemos visto en la ilustración 12, relación entre Precio y Kilómetros, podemos decir que sí, los vehículos con más kilómetros son los más baratos. Si entramos más en el detalle y mostramos esa relación en función del combustible o de la marca, se observa como todas siguen la misma tendencia con diferentes rangos de precio (acudir al notebook para ver las gráficas), lo que provoca que sobre la ilustración 12 tengamos mucho rango de precio par los mismos kilómetros.

Marca de coches más barata es Dacia

La marca más barata sería Daewoo con un precio medio de 1264€ por lo que rechazamos la hipótesis. La marca Dacia no es de las marcas que menos coches tiene a la

venta y no tiene vehículos tan antiguos (como Daewoo y otras marcas) por lo que no ocupa ni los primeros puestos en cuanto a marca barata

Los vehículos híbridos son más caros que los diésel

Como hemos visto en el análisis bivariante, los vehículos híbridos son de manera genera más caros que los vehículos diésel pero en este caso si mostramos los kilómetros que tienen de media los vehículos Diésel son mucho mayores que los que tienen los híbridos y como hemos visto justo antes, esos kilómetros sí que afectan al precio

En la siguiente tabla podemos ver la media de kilómetros de los vehículos por tipo de combustible

Combustible	Kilometros
Diésel	135113.12
Eléctrico	16975.347
Gasolina	83224.27
Híbrido	51100.835
Híbrido enchufable	28777.943

Tabla 2

Para poder hacer una comparativa más justa, vamos a filtrar cual es el rango de kilómetros y potencia sobre los que encontramos la mayor parte de los vehículos híbridos y realizaremos la comparación hibrido-Diesel sobre ese rango

Con los nuevos datos, las distribuciones de precio por combustibles son mucho más parecidas. Para poder confirmar que estadísticamente no existe igualdad en las medias de precios aplicaremos el test Anova sobre el precio en las diferentes categorías de combustible.

Como hemos visto en el análisis univariante, la variable precio, no sigue una distribución normal por lo que no podemos aplicar Anova directamente, lo aplicaremos por medio del teorema central del límite

Se ha creado una función que aplica el teorema tomando 100 muestras de la media de 30 valores y sobre esa nueva variable se aplica Anova (oneway de scipy), la función se llama ``teorema_central_limite`` y compara las medas una a una todas las categorías de la variable.

Nos queda:

	Diésel	Gasolina	Híbrido	Eléctrico	Híbrido enchufable
Diésel	X	X	X	X	X
Gasolina	Distintas	X	X	X	X
Híbrido	Distintas	Distintas	X	X	X
Eléctrico	Distintas	Distintas	Distintas	X	X
Híbrido enchufable	Distintas	Distintas	Distintas	Distintas	X

Tabla 3

En la tabla 3 podemos ver ninguna de las medias de precios por combustible son iguales estadísticamente.

Combustible	Precio
Diésel	18687.179
Eléctrico	22558.333
Gasolina	14941.819
Híbrido	20891.95
Híbrido enchufable	28891.649

Tabla 4

Con los valores de la tabla podemos confirmar la hipótesis de la media de precios de los vehículos híbridos es mayor que la de los vehículos de diésel.

Para dar más información sobre las posibles causas de las diferencias en precio (utilizando los datos que nos ofrece el dataset) vamos a realizar el test chi2 sobre las variables categóricas del df. Como hemos visto en el análisis bivariante, el precio es distinto en función de la Marca, del vendedor, del año.... si el test chi2 nos confirma que las variables categóricas no son independientes, podríamos decir que ese reparto de las categorías en los distintos combustibles no va a ser el mismo, por lo que no estamos repartiendo el precio de la misma manera en todos los combustibles. Por supuesto habrá más causas (como precio nuevo del

vehículo, estado general, valoración personal) que hagan que el precio pueda ser distinto en función del combustible

$$El\ test\ de\ \chi\text{^}2\ se\ bassa\ en\ la\ fórmula$$

$$\sum \frac{(observaciones_i - Esperado_i)^2}{Esperado_i}$$

De esta manera comprobamos que todas las variables categóricas son dependientes lo que, en este caso, quiere decir que no hay un reparto equitativo del número vehículos a la venta entre las diferentes categorías en función del combustible.

En el notebook hay un ejemplo para ver el reparto de las marcas en función del combustible.

Media de precios igual para todas las comunidades autónomas

Al igual que en la hipótesis anterior, en este caso utilizaremos el teorema central del límite para comparar las medias de precios de las comunidades autónomas.

En el notebook están la tabla indicando que tenemos medías iguales estadísticamente en algunas comunidades y distintas en otras comunidades.

Podemos decir que no tenemos las mismas medias de precios en todas las comunidades por lo que se descarta la hipótesis de las medias son iguales para todas las comunidades Autónomas

Los vehículos vendidos por profesionales son más caros que los vendidos por particulares

Por lo que hemos visto en el análisis bivariante, podemos confirmar la hipótesis ya que los vehículos vendidos por profesionales son más caros que los vendidos por particulares

Vendedor_profesional	Precio
False	11633.3169
True	18238.42884

Tabla 5

En la siguiente tabla podemos ver como el reparto desigual de los Combustibles en los tipos de vendedor puede hacer que los precios sean distintos

Combustible	Diésel	Eléctrico	Gasolina	Híbrido	Híbrido enchufable	All
Vendedor_profesional						
False	10635	23	4886	92	21	15657
True	17732	464	9494	973	277	28940
All	28367	487	14380	1065	298	44597

Tabla 6

Se puede ver en la tabla 6 que hay más proporción de vehículos Híbridos, Eléctricos e Híbridos enchufables vendidos por profesionales y como ya hemos visto antes, estos vehículos son más caros, lo que podría ser una de las causas que explique esta diferencia de precios.

Los vehículos más potentes son los más caros

En el notebook podemos ver las gráficas de la relación entre Potencia y precio en función del combustible y de las 10 marcas más vendidas

En este caso las gráficas no son tan claras como en ocasiones anteriores, pero se puede ver como la tendencia es aumentar el precio según aumenta la potencia.

Resumen de las hipótesis

- Los vehículos con más km son los más baratos --> Confirmada

- La marca de coches más barata es Dacia --> Rechazada

- Los vehículos híbridos son más caros que los Diésel --> Confirmada

- la media de precios es la misma para todas las comunidades autónomas --> Rechazada

- Los vehículos vendidos por profesionales son más caros que los vendidos por particulares --> Confirmada

- Los vehículos más potentes son los más caros --> Rechazada

6.Bibliografia

https://pandas.pydata.org/

From data to Viz | Find the graphic you need (data-to-viz.com)

Coches de segunda mano | Kaggle

3.10.4 Documentation (python.org)

Invicta Electric | Vehiculos 100% eléctricos para disfrutar en la ciudad