Análisis Exploratorio de Datos

Carlos I. Cabruja Rodil

The Bridge

Data Science | Part- Time

INTRODUCCIÓN

Acabas de entrar en el departamento de Analytics y Business Intelligence de una cadena de concesionarios de coches de segunda mano que se llama Broom Broom Cars S.L. La empresa está situada en Gran Bretaña y tú eres responsable de maximizar la venta de coches a las afueras del sur de Londres.

Los cargos directivos de la empresa están muy emocionados con tu incorporación ya que, por fin, van a poder apoyar sus decisiones en datos. Por este motivo, se te ha concedido un una lista con todos los coches vendidos de los concesionarios de la zona durante el año 2020 y 2021.

El objetivo del reto es generar un dashboard o presentación que permita al equipo directivo ver cuáles son los coches que han tenido más éxito durante 2020 y 2021, ver cuál es la tendencia y dar tu recomendación sobre los coches que Broom Broom Cars S.L. tiene que intentar obtener para sus concesionarios durante el año 2022.

Las variables del dataset son las siguientes:

- **Model:** Modelo del coche
- **Price:** Precio de venta
- **Transmission:** Tipo de transmisión
- **Mileage:** Millas que había recorrido el coche antes de la venta
- **Fuel Type:** Tipo de combustible
- **Tax:** Impuestos de venta
- **Miles Per Gallon (mpg):** Millas consumidas por cada galón de combustible
- **Engine Size:** Tamaño del motor
- **Maker:** Fabricante

HIPÓTESIS

¿Qué coches o tipo de coches tenemos que comprar en nuestro concesionario para así generar más ingresos durante el año 2022?

- ¿Cuáles son las marcas de coches que generan más ingresos?
- ¿Cuáles son los coches que generan más ingresos (tanto por coche como en conjunto)?
- ¿Hay dependencias entre el tamaño del motor y el precio de venta?
- ¿Es importante fijarse en los impuestos a pagar por el coche?
- ¿Qué modelo es el que le gusta más a la gente? ¿Y el que ha generado más ingresos de manera global?
- ¿Vale la pena targetear coches de transmisión híbrida? ¿Qué transmisión es la mejor? ¿Y los consumidores se fijan en el mpg?
- ¿El milage hace bajar el precio de los coches exponencialmente, linealmente o logarítmicamente?

PROCEDIMIENTO

1. **Cargado y Limpieza de Datos:**

   Primero se cargó la lista suministrada por el concesionario y vimos que contenía filas repetidas, lo cual se interpretó como cantidad de coches que se vendieron en exactamente las mismas características y con eso se creó la columna *“ventas”* que representa el precio del vehículo por la cantidad de veces que aparecía en la lista.

   En el proceso de visitado de los datos se descubrió que una porción de los datos la mayoría compuesto por la marca Audi y la gama de modelos “Q” tenían un tamaño de motor 0. Se planteó en un futuro utilizar el resto de variables suministradas en la lista para rellenar este porcentaje de datos que de momento son de tamaño de motor desconocido.

Lo siguiente que se pensó para mejorar la investigación fue crear una columna del tamaño de motor que esté compuesta por los tamaños de motor agrupados a como:

- **Motor pequeño** = 1
- **Motor mediano** = 1 - 2
- **Motor grande** = 2 - 3
- **Motor muy grande** > 3

Una vez se consideró que nuestra tabla estaba completa para el análisis, se estudió la posibilidad de la presencia de outliers en la lista (ver Figura 1), y vemos que el concesionario, tiene coches con un **precio** muy alejado de su rango de precio de venta, los **mpg** se podría estudiar a través del tamaño del motor, ya que sería lógico pensar que motores muy grandes consumen más combustible por milla. También vemos que tienen coches que han hecho mucho **milage** antes de su venta y que se produjo una **venta** extraordinaria de un modelo de coche que generó 250.000.

Sin embargo, se estudió qué porcentaje representan estas observaciones en la lista de coches y resultó ser de un **4,16%** por lo que como representa muy poco de nuestro dataset continuamos nuestro análisis sin esas observaciones, ya que así las medias de nuestras columnas no se encuentra perturbada por estas observaciones.

2. **Análisis univariante** (ver Figura 2 y Figura 3)**:**

   La mayoría de los coches que se venden son:

   - De transmisión semi-automática.
   - Los que el tipo de combustible es gasolina.
   - Los que tienen un impuesto de 145.
   - Los de la marca Volkswagen.
   - Los de motores medianos (1 - 2 litros).
   - Los que tienen un precio entre 10.000 y 40.000.
   - Los que tengan un promedio de **mpg** entre 30 y 60
   - Los que no superan las 2000 millas de recorrido

   También vemos que la mayoría de los modelos vendidos en el año anterior no superan los 50.000 en ventas, y que nuestras variables numéricas parecen no estar

distribuidas de manera normal, por lo que nos valdremos de pruebas no paramétricas para realizar nuestro análisis.

3. **Análisis bivariante:**

Para el análisis bivariante, nos preguntamos ya que vimos que el rango de precios es entre 10.000 y 40.000, cuales son las marcas de coches que tienen modelos de coches que se vendan en rango ya que son las marcas que tienen los precios que los consumidores parecen desear, por lo que se hizo un gráfico cajas que compara todas las marcas dadas en la lista, y al lado se hizo un gráfico a manera de curiosidad que con las observaciones antes mencionadas como outliers a fin de saber más detalles sobre esos datos (ver Figura 4).

Aquí vimos que la marca de coches vendió modelos a un precio incluso fuera del precio promedio de la misma marca fue los coches Audi, vemos también que las marcas que están un poco fuera del rango del precio más vendido son los BMW y audi, y la marca Volkswagen tiene un rango de precios que cubre desde los 10.000 a los 40.000 lo que explica que sea la marca más vendida, ya que tiene el rango preferido por los consumidores.

Después nos preguntamos cuáles son las marcas que generaron más en ventas, y hicimos otro gráfico para comparar las ventas a través de las marcas (ver Figura 5) también representando los outliers visto en ventas, y vemos que hubo un modelo que generó una extraordinaria venta en la marca Volkswagen que fue el modelo Passat, ahora sin contar específicamente con esa venta, vemos que la marca BMW y Audi tiene más ingresos en ventas consideradas por sí mismas que la marca Volkswagen, que tiene sentido ya que tienden a tener modelos de mayor precio que Volkswagen, pero aquí nos preguntamos si ¿La media de las ventas producidas por estas 3 marcas son distintas o genera lo mismo en ventas en promedio? por lo que primero se comprueba si realmente nuestra columna de ventas no está distribuida de forma normal haciendo una prueba de normalidad de D'Agostino que es la más utilizada para muestras mayores a 50 observaciones y la prueba tiene como resultado que no está distribuida de manera normal. Por lo que nos valdremos de una prueba de diferencia de medias no paramétrica conocida como la prueba de Kruskal-Wallis, que solo tenemos que asumir que los datos provienen de una misma distribución.

Una vez realizada la prueba entre las medias de los 3 modelos, tenemos como resultado que las medias de ventas producidas entre los modelos BMW y Audi son estadísticamente iguales, y estas dos anteriores son distintas al promedio de ventas de Volkswagen por lo que, a pesar de ser la marca preferida por los consumidores, con los BMW y Audi vendidos en el año 2020-2021 hizo más dinero en **ventas** que los Volkswagen vendidos (Sin contar con la venta extraordinaria de los Passat). (ver Tabla 1).

4. **Análisis multivariante:**

Después de conocer las marcas de coches que generaron más ingresos en el periodo 2020-2021, vamos a estudiar los modelos de esas marcas que generaron más ingresos (ver Figura 6).

El gráfico nos muestra que de los **BMW**, los modelos **Z4** fueron los que más dinero en ventas generó, pero vemos que fue por una venta que se produjo en más de 70.000, pero que en promedio los modelos **Z4** generan entre 30.000 y 40.000 en ventas. Pero el claro ganador de los modelos **BMW** fue el **X5** que produjo en ventas entre los 45.000 y 65.000 y en el caso de los **Audi**, si el claro ganador es el modelo **Q5**, generando ventas entre los 35.000 y 60.000, después pudiésemos decir que el modelo **A5** es el segundo mejor vendido generando ventas entre los 45.000 y 65.000. De igual forma vamos a completar la idea que nos podamos hacer, viendo cuáles fueron los modelos que se vendieron más, que no quiere decir que sean los que generan más ingresos (ver Figura 7).

En este gráfico vemos que el **Z4** y el **X5** a pesar de ser los modelos que más ingresos generaron, no son los modelos **BMW** que más se venden, sino que son los modelos **Serie**, siendo **3 Series** el más vendido, así que nos pudiésemos preguntar si es la media de los ingresos producidos por **3 Series** distinta a la media de ingresos producidos por el modelo **X5** que es el coche donde los datos están en el rango de ingresos más altos.

Y en el caso de los **Audi** el **Q3** es el más vendido pero no muy distinto del grupo **Q5**, **A5**, **A1**, **A4**, por lo que el **Q5** sigue siendo nuestro candidato a ganador con un rango de ingresos bastante amplio.

Pues ya que conocíamos cuáles fueron los modelos que más se vendieron y los modelos que más ingresos generaron, podremos valernos de la prueba estadística de comparación de medias antes mencionada para responder las siguientes preguntas:

- *¿Es la media de ingresos del modelo 3 Series de BMW distinta a la media de ingresos del modelo X5?*
- *¿Es la media de ingresos del modelo Q5 de Audi distinta a la media de ingresos del modelo Q3?*

Y las pruebas dieron como resultado que tienen medias de ingreso estadísticamente distintas, siendo:

- La media de ventas de BMW 3 Series es: 34327.965986394556 con 147 ventas
- La media de ventas de BMW X5 es: 57169.6875 con 16 ventas
- La media de ventas de Q5 es: 43003.614457831325 con 83 ventas
- La media de ventas de Q3 es: 34021.15267175573 con 131 ventas

5. **Contraste de hipótesis:**

Ahora que ya tenemos una idea de cuales son las marcas y los modelos que más rentables para el concesionario, vamos a responder el resto de preguntas, como:

- ¿Hay dependencias entre el tamaño del motor y el precio de venta? (ver Figura 8)

El gráfico nos invita a pensar que si hay una dependencia entre el tamaño del motor y el precio pero para confirmar vamos a hacer una prueba de diferencia de medias, antes se estudió la variable de precio a ver si tiene una distribución normal con una prueba d'Agostino y nos dió como resultado que no se distribuye normalmente, entonces continuaremos usando la prueba no paramétrica que hemos utilizado durante el análisis, y la prueba dió como resultado que con un 95% de confianza las medias de los precios según el tamaño del motor son distintas.

- ¿Vale la pena targetear a los coches con transmisión híbrida?

Ya en los anteriores gráficos vimos que son los coches que más se venden, pero vamos a hacer una prueba para diferenciar la media de los ingresos producidos por los coches dividido entre las transmisiones y dió como resultado que si son distintas, entonces podemos concluir que si que vale la pena targetear los coches de transmisión híbrida

- ¿Se fijan los consumidores en las millas por galón? (ver Figura 9)

Si los consumidores se fijan en las millas por galón de un coche podríamos ver una relación entre la columna de nuestro dataset mpg y ventas, y resulta que existe una correlación inversa del 55% de las millas por galón con las ventas, es decir que a medida que aumenta el consumo de las millas por galón, los ingresos producidos por esos coches disminuyen. Y que los consumidores parecen preferir los coches que están entre 40 y 60 millas por galón, pero cabe destacar que, los coches que están entre 30 y 40 millas por galón fueron los que produjeron más ingresos.

- ¿El milage hace bajar el precio de los coches exponencialmente, linialmente o logarítmicamente?

  Para responder una pregunta se hizo los 3 modelos y se estudió el error medio:

  - El error medio de lineal es: 12085.775516758324
  - El error medio de exponencial es: -1818.1094387110456
  - El error medio de logarítmica es: 7.976253697767149e-12

Y vemos que el modelo que más se ajusta a la relación que existe entre milage y precio es el logaritmo, y la diferencia entre los números es tal que no haría falta hacer una prueba estadística. Por lo que ya si el coche tiene muchas millas recorridas, el precio no varía tanto como cuando no tiene millas recorridas.

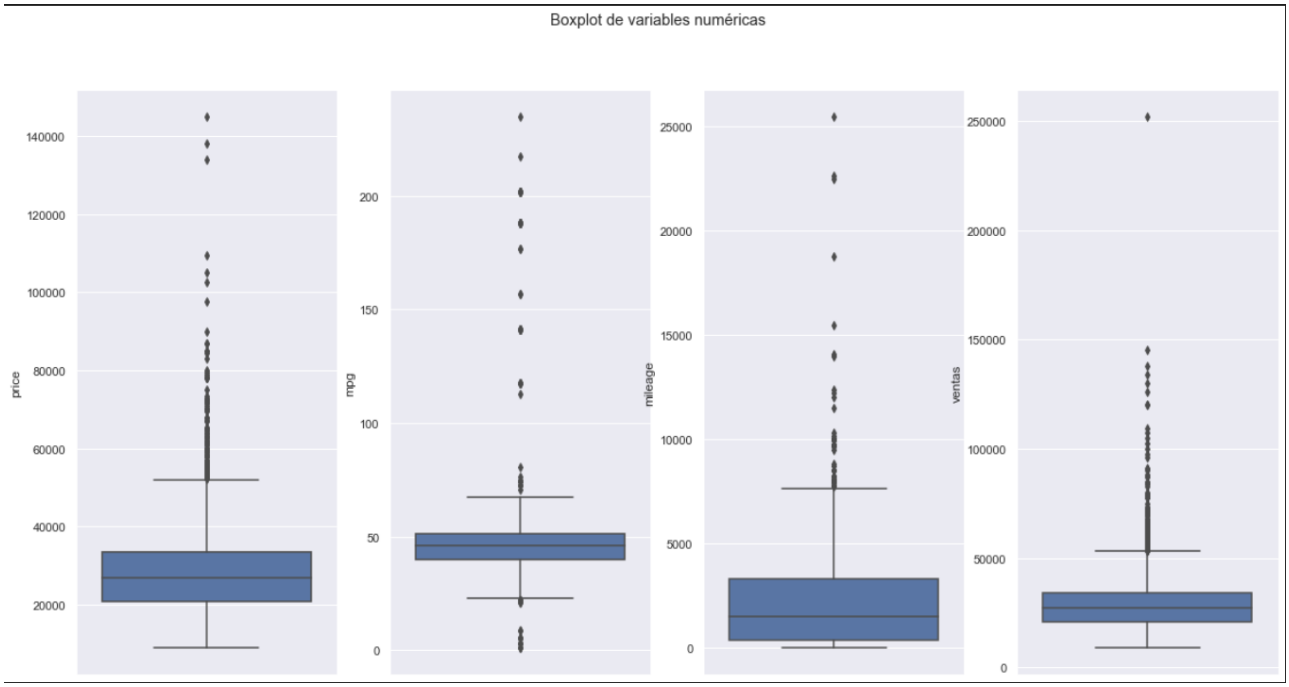
CONCLUSIÓN

¿Qué coches o tipo de coches tenemos que comprar en nuestro concesionario para así generar más ingresos durante el año 2022?

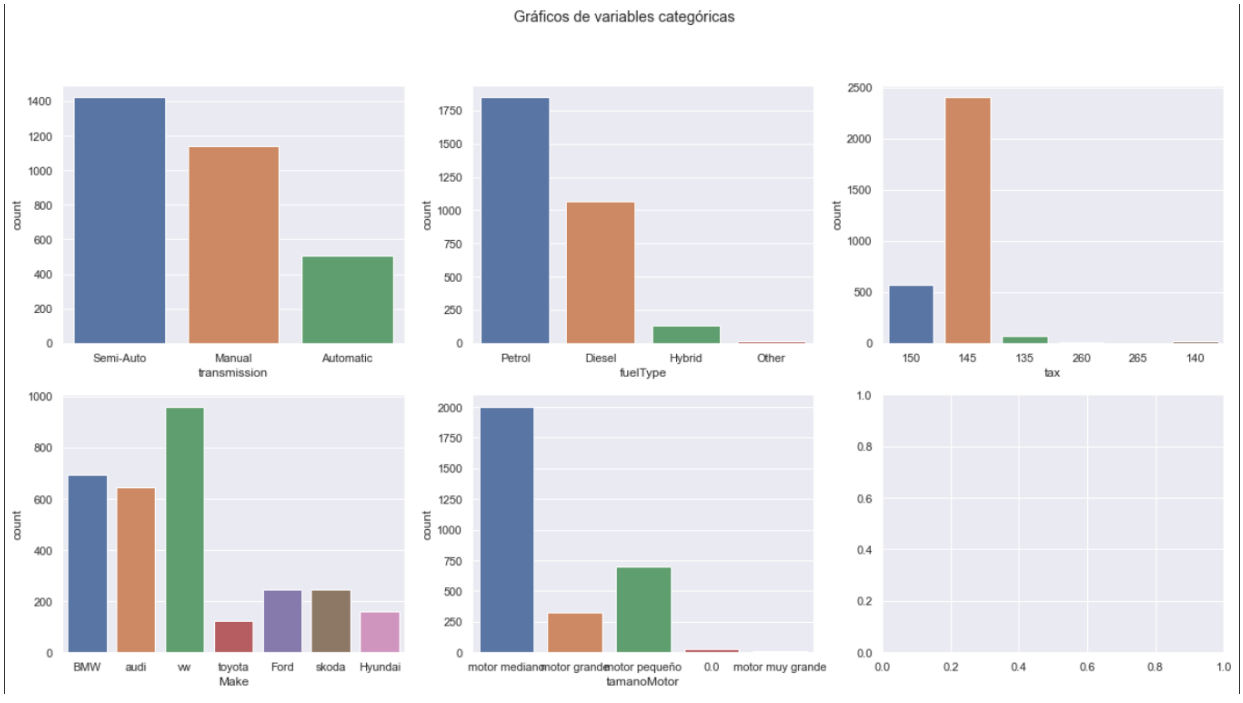
- ➔ **Top 3 más vendidos**: BMW Series 3, Audi Q3, Volkswagen Golf.
- ➔ **Top 3 más ingresos producidos**: BMW X5, Audi Q5, Volkswagen Passat.
- ➔ Existe una dependencia entre los precios y el tamaño del motor, por lo que si el tamaño del motor es mayor, el precio es mayor. Por ende, los coches con un tamaño de motor mediano tienen una mayor probabilidad de ser vendidos ya que entran en el rango de precios más vendido.
- ➔ Es importante los impuestos a pagar ya que la mayoría de las ventas se concentraron en el impuesto de 145
- ➔ El modelo que le gusta más a la gente es el Volkswagen Golf
- ➔ El modelo que genera más ingresos es el BMW X5
- ➔ Los coches de transmisión híbrida son los que más se venden, pero además de ser los más vendidos, también son los que generan más ingresos.
- ➔ No existe una verdadera correlación entre las ventas y el consumo de las millas por galón, pero existe una correlación inversa, es decir, que a medida que aumenta el consumo de las millas por galón, los ingresos producidos por esos coches disminuyen.
- ➔ Las millas recorridas por los coches hace bajar el precio de los coches logarítmicamente.

ANEXO

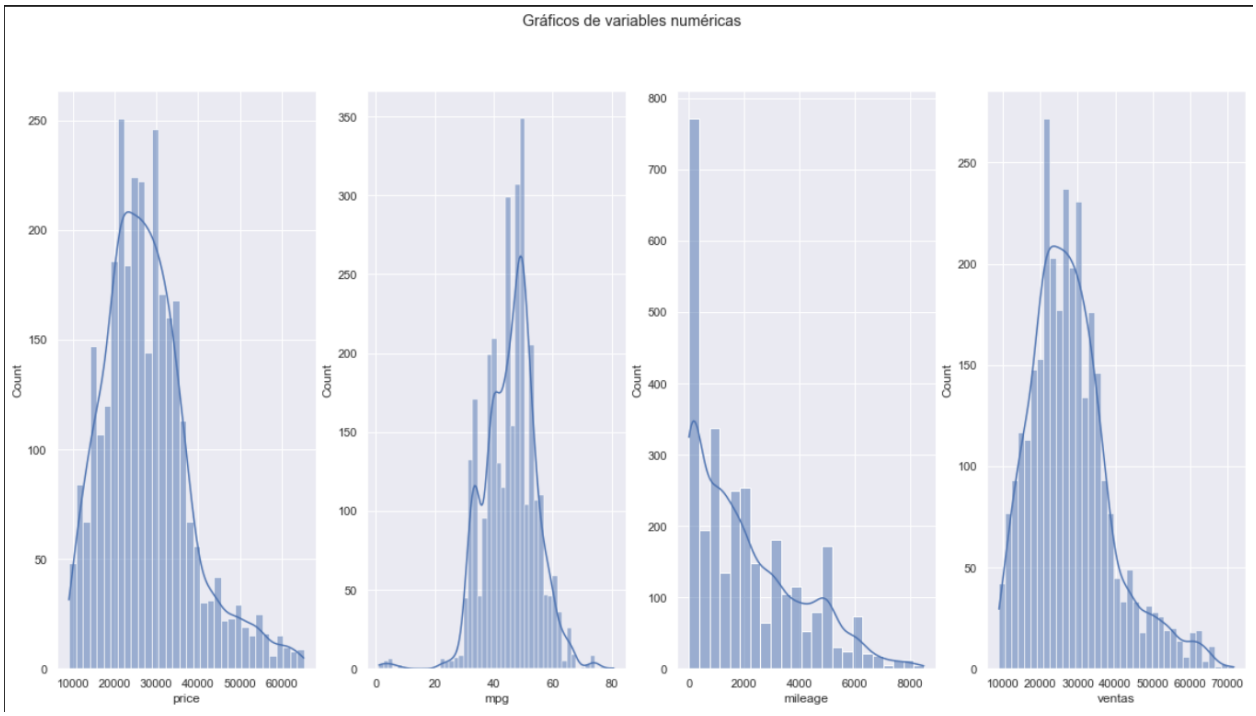
- Figura 1:



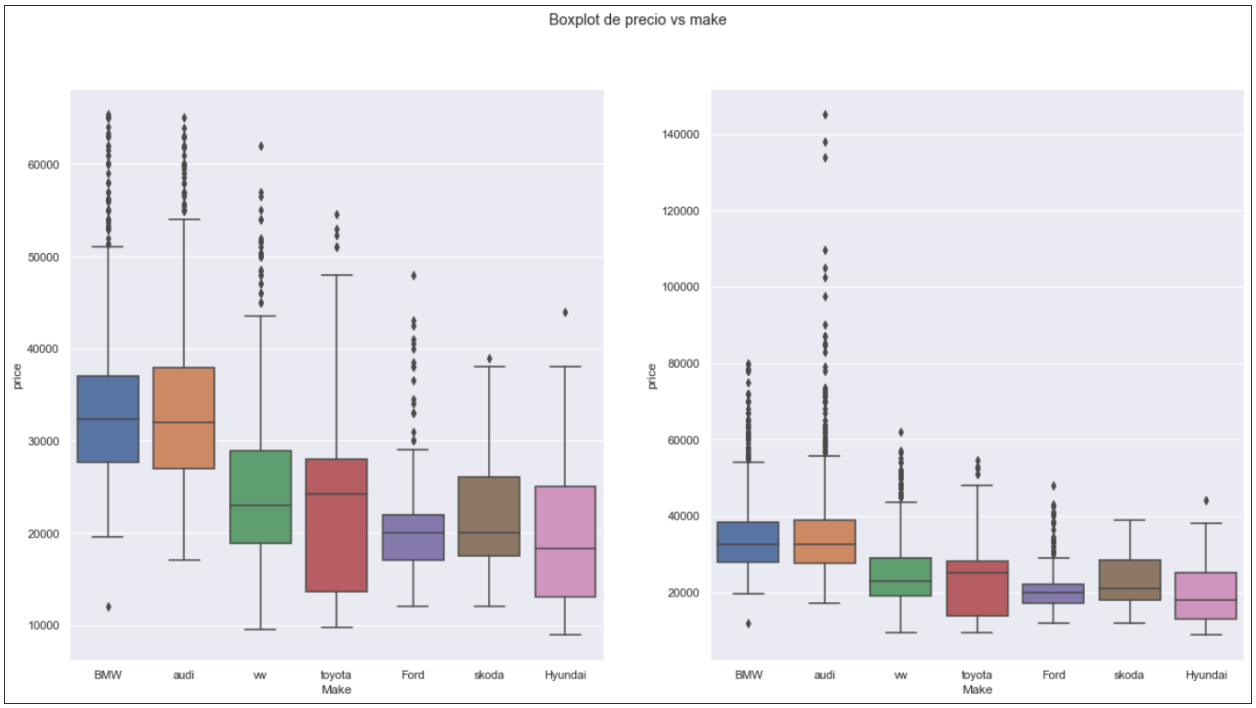
- Figura 2:



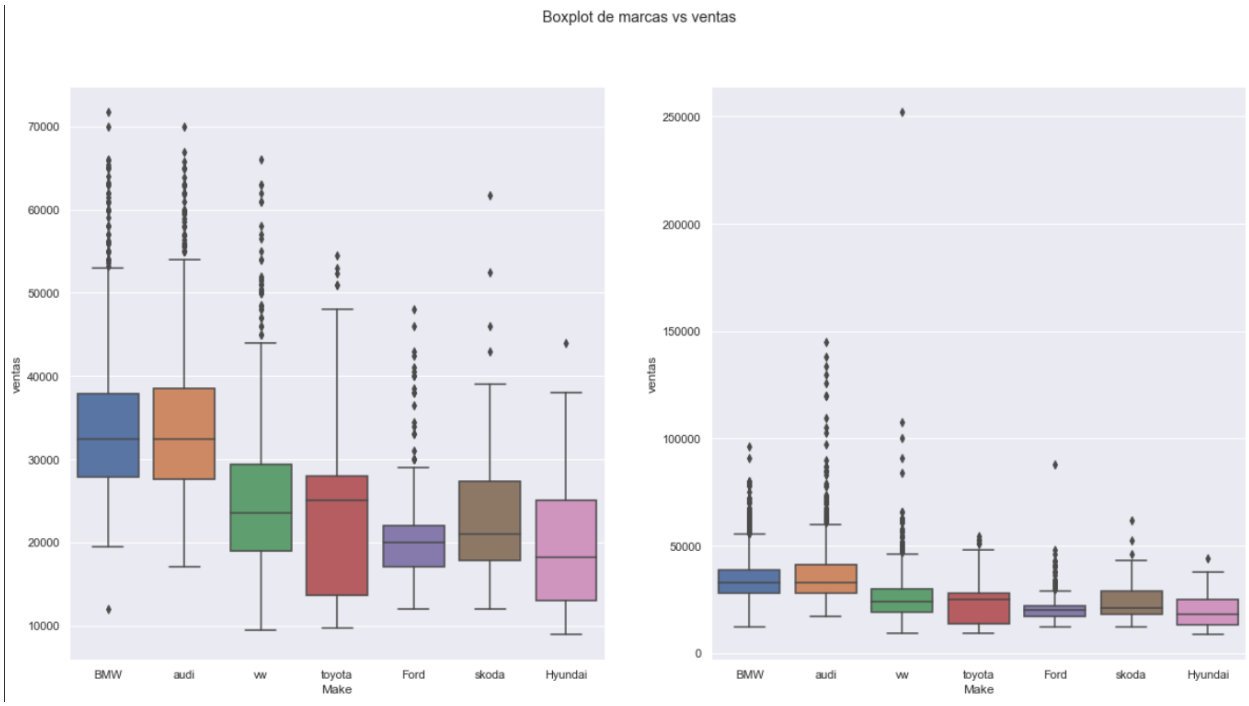
- Figura 3:



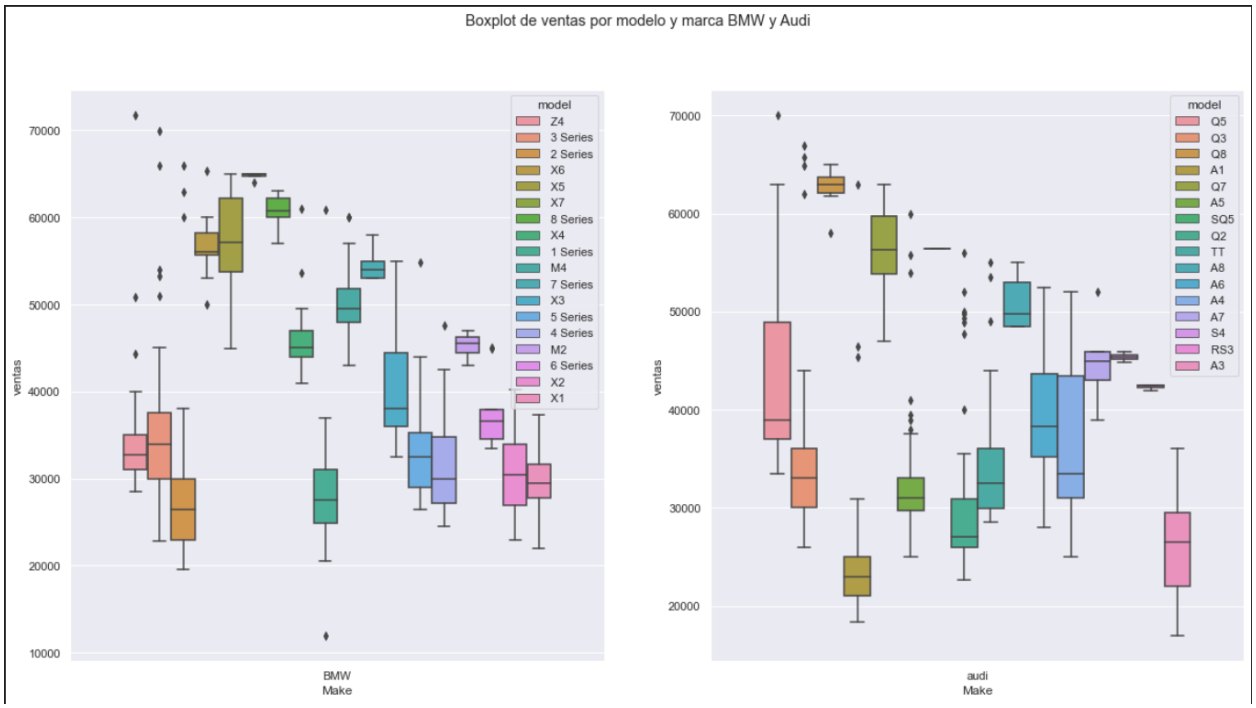
- Figura 4:



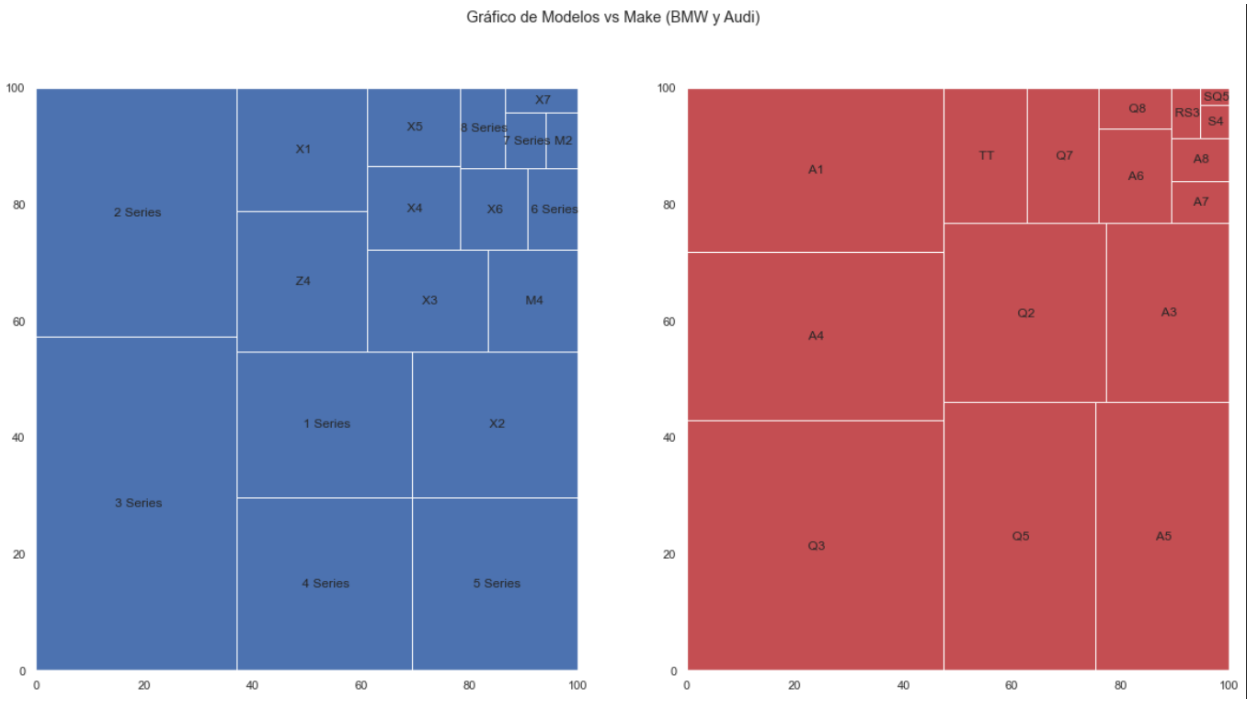
- Figura 5:



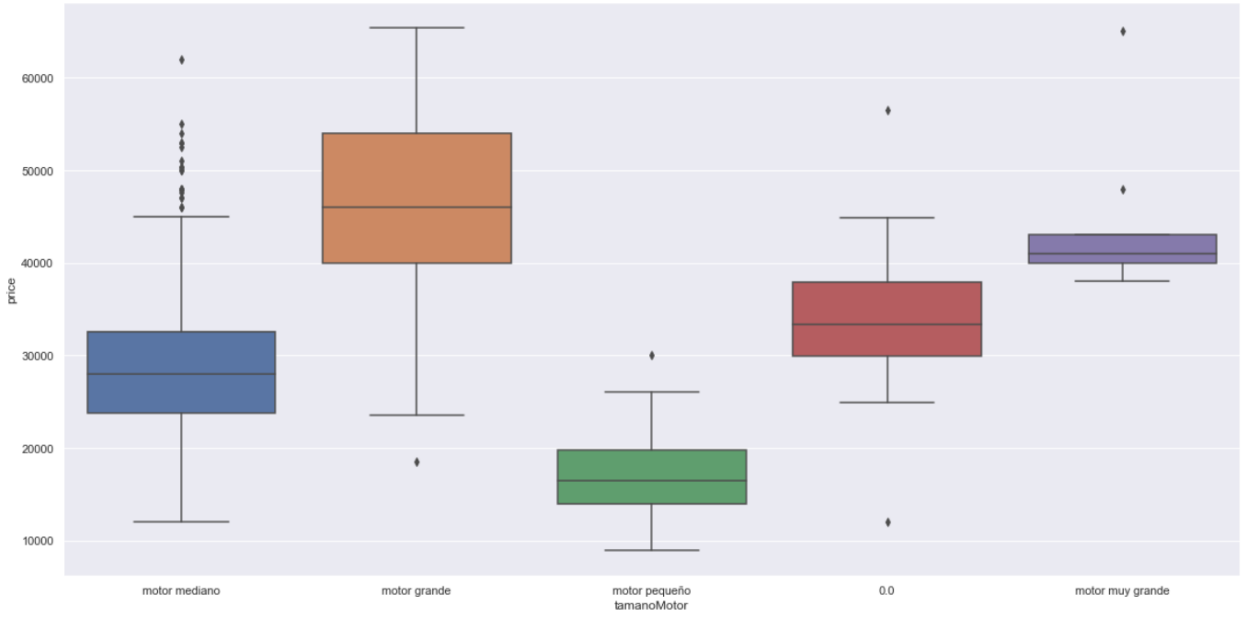
- Figura 6:



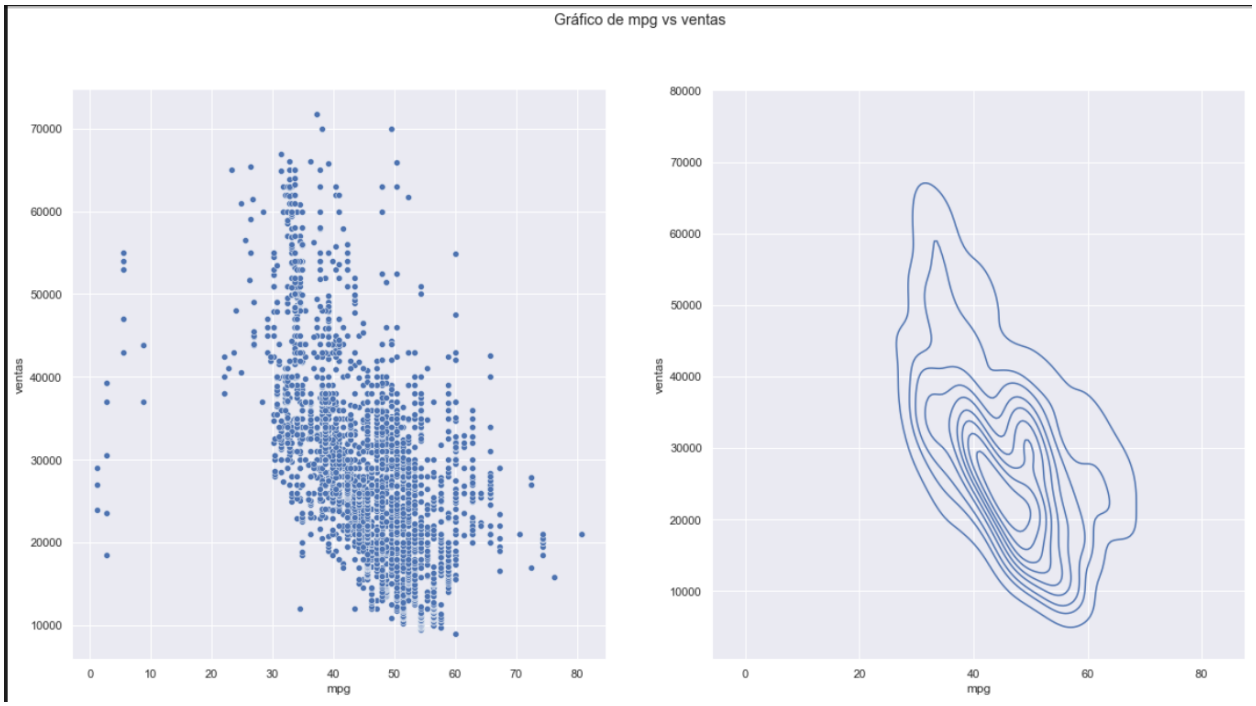
- Figura 7:



- Figura 8:



- Figura 9:



- Tabla 1:

Make	Media
BMW	34598.776012
audi	34485.576087
vw	24711.264859
toyota	23098.745902
skoda	22817.089796
Ford	20838.946939
Hyundai	19788.474684