INTELIGENCIA DE NEGOCIOS APLICADA AL SECTOR AUTOMOVILÍSTICO

Trabajo Final de Grado

Grado en Administración y Dirección de Empresas.

Alumno:

D. Jorge Rueda Soler

Tutor:

D. Francisco Guijarro Martínez

Curso académico 2020-2021

VALENCIA, JULIO DE 2021

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ILUSTRACIONES

1. Introducción

1.1 Resumen

La motivación principal de este trabajo es informar a los consumidores de cuáles son los aspectos más relevantes a la hora de comprar un coche para que puedan elegir mejor en función de las conclusiones de este estudio.

Todo este trabajo ha sido realizado con el software R Studio, el código de programación estará oculto por razones visuales, no obstante, se incluye en los anexos del trabajo.

Para lograr los diferentes objetivos, se realizan diferentes secciones, cada una acorde a los distintos objetivos del trabajo, los cuáles serán desarrollados en el siguiente punto.

En primer lugar, se determina la tipología de variables, es decir, identificar cuáles son numéricas y cuáles de ellas son categóricas. La principal diferencia entre estos tipos de variables es que las categóricas identifican la clase de la observación, por ejemplo, una persona puede ser hombre o mujer, mientras que las numéricas determinan un valor numérico de una variable, por ejemplo, la altura de una persona.

Se realiza un análisis exploratorio de las variables con el fin de conocer mejor la base de datos. Una vez hecho este análisis, se comprueba que no haya variables constantes en las observaciones, porque como el propio nombre indica son variables, si fueran constantes habría que eliminarlas.

También serán eliminadas aquellas que presenten un exceso de valores inconsistentes. Se define como valor inconsistente a la observación que tiene un valor que está muy distante al resto y que no es lógico.

Además, serán eliminadas aquellas variables u observaciones que tengan un porcentaje de valores faltantes (NA: Not Available) superior al 20%. Este porcentaje se toma como referencia, ya que supondría que una de cada cinco observaciones tendría un NA en esa variable o que la observación tendría una de cada cinco variables en NA. Como ya se ha comentado, se toma como referencia, pero hay que tener en cuenta que cada base de datos es distinta y a pesar de que este porcentaje se utilice como referencia, siempre dependerá de la persona que trata la base de datos (más o menos estricta) y de la base datos (número de observaciones y de variables).

Tras esto, con la librería Van Buuren and Groothuis-Oudshoorn (2011), se imputarán dichos valores faltantes en caso de que no superen el porcentaje marcado. Imputar significa estimar qué valor tendría la observación faltante basándose en el resto de las características que tiene esa observación. Esas características las compartirá con otras observaciones de la base de datos, y de esta forma se podrá estimar. Existen también procesos alternativos para las variables numéricas como la sustitución del dato faltante por la media de las observaciones con datos.

Finalmente, tras haber preparado la base de datos, se realizan relaciones entre variables de las cuáles se pueda obtener información relevante.

Al acabar estos primeros pasos, se realiza un método de aprendizaje no supervisado, el cual tiene como fin principal ver las relaciones entre todas las variables numéricas. El método escogido es el Análisis de Componentes Principales (PCA). De esta forma se consigue eliminar el posible efecto de correlación que pudieran tener las variables entre sí. Este efecto de correlación se verá más adelante en estudios individualizados entre variables.

A continuación, se realizarán diversos métodos de aprendizaje supervisado para ver que variables son las más influyentes a la hora de predecir si un coche estará por encima de 30.000$. Este valor es la mediana del precio en esta base de datos. Se escoge este valor ya que 30.000 $ es un precio considerable para un coche y de esta forma se divide en dos la base de datos, dividendo justo por la mitad de las observaciones. Se podría dividir por la media, pero valores extremos en el precio (coches de más de 2.000.000 $) hacen que este valor sea elevado y por tanto haya más cantidad de coches por debajo de la media. Con el uso de la mediana se evita este efecto y se busca predecir qué características tendrá un coche para estar en un grupo u otro.

La principal diferencia entre estos métodos es el objetivo, el aprendizaje supervisado busca predecir, mientras que el aprendizaje no supervisado busca obtener información útil a partir de la base de datos.

1.2 Objetivo

El objetivo principal del trabajo realizado sobre la base de datos de los coches es estudiar qué variables son las que más afectan a aspectos tangibles tales como el consumo o precio a la par que estudiar la relación y el comportamiento entre las distintas variables que presenta un coche.

El objetivo es aparentemente general, pero hay que recalcar que se realizarán las relaciones entre variables que puedan aportar información relevante al consumidor, que es el objetivo principal.

Además, se busca realizar una caracterización de los vehículos en el mercado americano, analizando la evolución de sus características y las preferencias de los consumidores en los últimos años, para finalmente, poder determinar qué características fijan el precio de venta de los vehículos.

Para todo esto, se ha obtenido una base de datos a través de la página web Kaggle. Se trata de un espacio web que permite poner en contacto a científicos e ingenieros de datos para explorar y crear modelos. La base de datos obtenida cuenta con 11.914 observaciones y 16 variables. Kaggle. (21 de diciembre de 2016). Car Features and MSRP. *https://www.kaggle.com/CooperUnion/cardataset*

Con esta base de datos, se busca extrapolar a la actualidad y a todos los modelos de coches que existen actualmente. No obstante, esta muestra tiene alguna limitación, por ejemplo, presenta más coches fabricados recientemente que antiguos, por lo que las conclusiones que se obtengan respecto a determinadas características podrían no ser exactas.

2. Preparación de la base de datos

2.1 Resumen de los tipos de variables

Las variables de la base de datos son:

- Make: Define al fabricante del coche (categórica).
- Model: Identifica el modelo de vehículo (categórica).
- Year: Año de matriculación del automóvil (numérica).
- Engine.Fuel.Type: Tipo de combustible que utiliza el vehículo (categórica).
- Engine.HP: Potencia (CV) del coche (numérica).
- Engine.Cylinders: Número de cilindros que tiene el automóvil (numérica).
- Transmission.Type: Tipo de cambio de marcha del coche (categórica).
- Driven_Wheels: Tipo de tracción de las ruedas (categórica).
- Number.of.Doors: Número de puertas (numérica).
- Vehicle.Size: Tamaño del vehículo (categórica).
- Vehicle.Style: Tipo de coche (categórica).
- highway.MPG: Consumo del coche en carretera (numérica).
- city.mpg: Consumo del coche en ciudad (numérica).
- Popularity: Popularidad del vehículo (numérica).
- MSRP: Precio del coche (numérica). Expresada en dólares.

Se realiza un resumen de todas las variables, diferenciándolas entre numéricas y categóricas. Se excluye de este resumen a la variable “Model” puesto que existen 915 modelos de diferentes fabricantes, lo cual hace una lista enorme y no aporta información significativa.

Es importante remarcar que, para las variables de consumo, se trata de un consumo medido en Miles per Gallon, lo que indica cuantas millas se recorren con un galón. Por lo que cuanto más alto sea este valor, menor será el consumo del vehículo.

Sucede lo contrario con el sistema que se utiliza en España, donde el consumo se mide en litros por kilómetros, por lo que, en España, cuanto menor sea el valor, menor consumo de combustible, y en Estados Unidos, cuanto más alto sea el valor, menor consumo.

```
      Year          Engine.HP      Engine.Cylinders    Number.of.Doors
 Min.   :1990   Min.   : 55.0   Min.   : 0.000   Min.   :2.000
 1st Qu.:2007   1st Qu.: 170.0  1st Qu.: 4.000   1st Qu.:2.000
 Median :2015   Median : 227.0  Median : 6.000   Median :4.000
 Mean   :2010   Mean   : 249.4  Mean   : 5.629   Mean   :3.403
 3rd Qu.:2016   3rd Qu.: 300.0  3rd Qu.: 6.000   3rd Qu.:4.000
 Max.   :2017   Max.   :1001.0  Max.   :16.000   Max.   :4.000
                NA's   :69      NA's   :30       NA's   :6
```

highway.MPG	city.mpg	Popularity	MSRP
Min. : 12.00	Min. : 7.00	Min. : 2	Min. : 2000 $
1st Qu.: 22.00	1st Qu.: 16.00	1st Qu.: 549	1st Qu.: 21.000 $
Median : 26.00	Median : 18.00	Median :1385	Median : 29.995 $
Mean : 26.64	Mean : 19.73	Mean :1555	Mean : 40.595 $
3rd Qu : 30.00	3rd Qu : 22.00	3rd Qu :2009	3rd Qu : 42.231 $
Max. :354.00	Max. :137.00	Max. :5657	Max. : 2.065.902 $

Tabla 1: Resumen de variables numéricas
Fuente: Elaboración propia

Make

Acura	Alfa Romeo	Aston Martin	Audi	Bentley	BMW	Bugatti	Buick	Cadillac	Chevrolet
252	5	93	328	74	334	3	196	397	1123
Chrysler	Dodge	Ferrari	FIAT	Ford	Genesis	GMC	Honda	HUMMER	Hyundai
187	626	69	62	881	3	515	449	17	303
Infiniti	Kia	Lamborghini	Land Rover	Lexus	Lincoln	Lotus	Maserati	Maybach	Mazda
330	231	52	143	202	164	29	58	16	423
McLaren	Mercedes	Mitsubishi	Nissan	Oldsmobile	Plymouth	Pontiac	Porsche	Rolls-Royce	
5	353	213	558	150	82	186	136	31	
Scion	Spyker	Subaru	Suzuki	Tesla	Toyota	Volkswagen	Volvo	Saab	
60	3	256	351	18	746	809	281	111	

Engine.Fuel.Type

diésel	electric	flex-fuel	gas natural	gasolina
154	66	985	2	10707

Transmission.Type

AUTOMATED_MANUAL	AUTOMATIC	DIRECT_DRIVE	MANUAL	UNKNOWN
626	8266	68	2935	19

Driven_Wheels

four wheel drive	front wheel drive	rear wheel drive
3756	4787	3371

Market.Category

Crossover	Crossover,Diesel	Crossover,Exotic,Luxury,High-Performance
1110	7	1
Crossover,Exotic,Luxury,Performance	Crossover,Factory Tuner,Luxury,High-Performance	
1	26	
Crossover,Factory Tuner,Luxury,Performance	Crossover,Factory Tuner,Performance	
5	4	
Crossover,Flex Fuel	Crossover,Flex Fuel,Luxury	Crossover,Flex Fuel,Luxury,Performance
64	10	6
Crossover,Flex Fuel,Performance	Crossover,Hatchback	Crossover,Hatchback,Factory Tuner,Performance
6	72	6
Crossover,Hatchback,Luxury	Crossover,Hatchback,Performance	Crossover,Hybrid
7	6	42
Crossover,Luxury	Crossover,Luxury,Diesel	Crossover,Luxury,High-Performance
410	34	9
Crossover,Luxury,Hybrid	Crossover,Luxury,Performance	Crossover,Luxury,Performance,Hybrid
24	113	2
Crossover,Performance	Diesel	Diesel,Luxury
69	84	51
Exotic,Factory Tuner,High-Performance	Exotic,Factory Tuner,Luxury,High-Performance	
21	52	
Exotic,Factory Tuner,Luxury,Performance	Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	
3	13	

Exotic,Flex Fuel,Luxury,High-Performance	Exotic,High-Performance	
11	261	
Exotic,Luxury	Exotic,Luxury,High-Performance	Exotic,Luxury,High-Performance,Hybrid
12	79	1
Exotic,Luxury,Performance	Exotic,Performance	Factory Tuner,High-Performance
36	10	106
Factory Tuner,Luxury	Factory Tuner,Luxury,High-Performance	Factory Tuner,Luxury,Performance
2	215	31
Factory Tuner,Performance	Flex Fuel	Flex Fuel,Diesel
92	872	16
Flex Fuel,Factory Tuner,Luxury,High-Performance	Flex Fuel,Hybrid	Flex Fuel,Luxury
1	2	39
Flex Fuel,Luxury,High-Performance	Flex Fuel,Luxury,Performance	Flex Fuel,Performance
33	28	87
Flex Fuel,Performance,Hybrid	Hatchback	Hatchback,Diesel
2	641	14
Hatchback,Factory Tuner,High-Performance	Hatchback,Factory Tuner,Luxury,Performance	
13	9	
Hatchback,Factory Tuner,Performance	Hatchback,Flex Fuel	Hatchback,Hybrid
22	7	72
Hatchback,Luxury	Hatchback,Luxury,Hybrid	Hatchback,Luxury,Performance
46	3	38
Hatchback,Performance	High-Performance	Hybrid	Luxury
252	199	123	855
Luxury,High-Performance	Luxury,High-Performance,Hybrid	Luxury,Hybrid
334	12	52
Luxury,Performance	Luxury,Performance,Hybrid	Performance
673	11	601
Performance,Hybrid	**N/A**	
1	**3742**	

Vehicle.Size

Compact	Large	Midsize
4764	2777	4373

Vehicle.Style

2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible
506	138	702	2488	71	95	793
Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	
29	1211	681	623	417	128	
Regular Cab Pickup	Sedan	Wagon				
392	3048	592				

Tabla 2: Resumen de variables categóricas
Fuente: Elaboración propia

Se observa que los NA dentro de Market Category son 3742 casos, lo que supone un 31.41 %, es decir, más del 20% de valores faltantes en una variable, este porcentaje es muy elevado, por lo que la eliminamos esta variable. Además, no va a aportar ninguna información relevante, puesto que tiene variables que aportan información similar ("Vehicle.Size", "Vehicle.Style") y tiene muchas categorías similares entre ellas. Si se tratara de una variable vital en el estudio, se evitaría esta acción, pero no es el caso.

Además, con este resumen se comprueba que ninguna variable, tanto categórica como numérica, de las que aparecen en la base de datos sea constante, puesto que si lo fuera, se debería eliminar. Al variar todas, no se elimina ninguna más por el momento.

3. Valores inconsistentes o anómalos

3.1. Valores inconsistentes en variables numéricas

Se genera un diagrama de caja, en el cual se representan todas las variables numéricas con el objetivo de comprobar si existen valores anómalos.

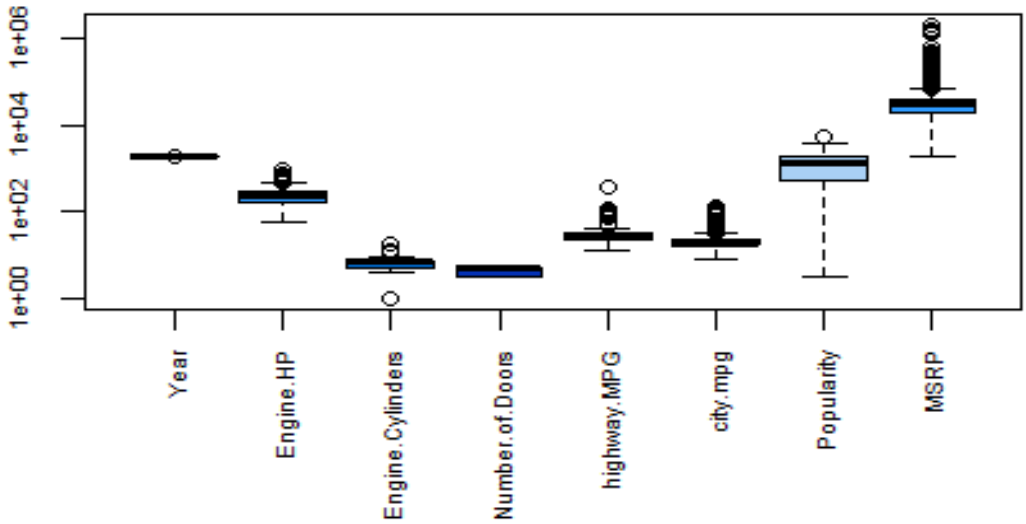

Gráfico 1: Diagrama de caja de variables numéricas
Fuente: Elaboración propia

A la vista del diagrama de caja, se decide analizar una por una las variables de consumo de combustible ("highway.MPG" y "city.mpg"), potencia del motor y año de matriculación para estudiar más en profundidad estas variables y asegurarnos de que los outliers presentados en el gráfico no sean valores inconsistentes. Un outlier es una observación cuyo valor es muy distante al resto.

No se profundiza en los outliers del precio de venta, ya que, al ser libre, cada vendedor puede poner el que considere adecuado, por lo que no tiene sentido analizar si es inconsistente.

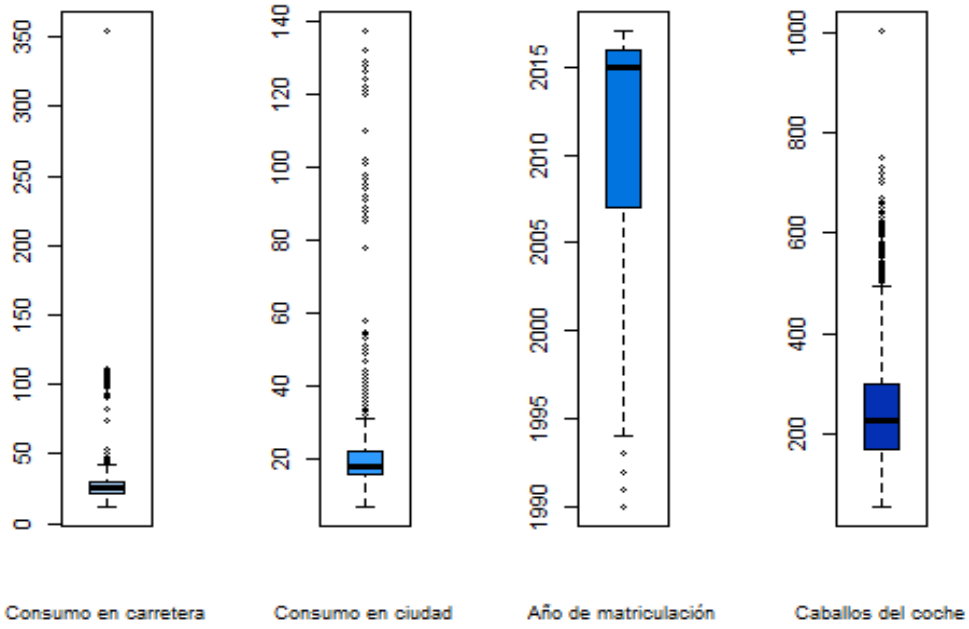

Gráfico 2: Diagrama de caja de variables numéricas seleccionadas
Fuente: Elaboración propia

Al analizar todas estas variables, llama la atención que en la variable de consumo de combustible en carretera (highway.MPG) haya un valor máximo tan distante al resto.

Es lógico que en carretera los coches consuman menos, de hecho, la media de consumo de combustible en carretera es mayor a la media de consumo en ciudad (más kilómetros recorridos con un galón). Sin embargo, el punto máximo está muy alejado, por lo que se profundiza el análisis para comprobar si se trata de un dato anómalo, ya que se trata de un consumo mayor de 350 MPG, más del doble que el coche con el mínimo consumo en ciudad. Además, se comprueba si es razonable que haya observaciones con 1000 CV.

Primero, se identifica el vehículo con un consumo anómalo:

```
Make: Audi   Model: A6   Year: 2017
```

Media de consumo en carretera de Audi A6:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
27.00	30.00	32.00	46.35	35.00	354.00

Media de consumo en carretera de Audi A6 corregida:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
27.00	30.00	32.00	32.50	35.00	38.00

Tabla 3: Consumo en carretera de Audi A6
Fuente: Elaboración propia

Se trata de un Audi A6, cuyo consumo (354 MPG) es 10 veces menor (en la medición americana) al de la media de los otros Audi A6, por lo que se divide entre 10 este consumo, ya que se considera un error de decimales.

Por otro lado, se analiza la observación con 1.000 CV como ya se había comentado antes.

Primero, se identifica el vehículo:

```
Make: Bugatti   Model: Veyron 16.4   Year: 2008
```

CV de Bugatti Veyron 16.4:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1001	1001	1001	1001	1001	1001

Tabla 4: Resumen de CV de Bugatti Veyron 16.4
Fuente: Elaboración propia

Se contabilizan los Bugatti Veyron 16.4 para asegurarse de que no se trata de una única observación aislada con valores anómalos.

Contabilización de Bugatti Veyron 16.4:

```
FALSE: 11911
TRUE:  3
```

Hay 3 Bugatti Veyron 16.4, los 3 tienen 1001 CV. Se trata de un coche de lujo, además de existir más de una observación, por lo que no se trata de un valor anómalo.

3.2. Valores faltantes por variable

Se realiza un gráfico en el cual se muestra el porcentaje de valores faltantes que presentan las variables de la base de datos.

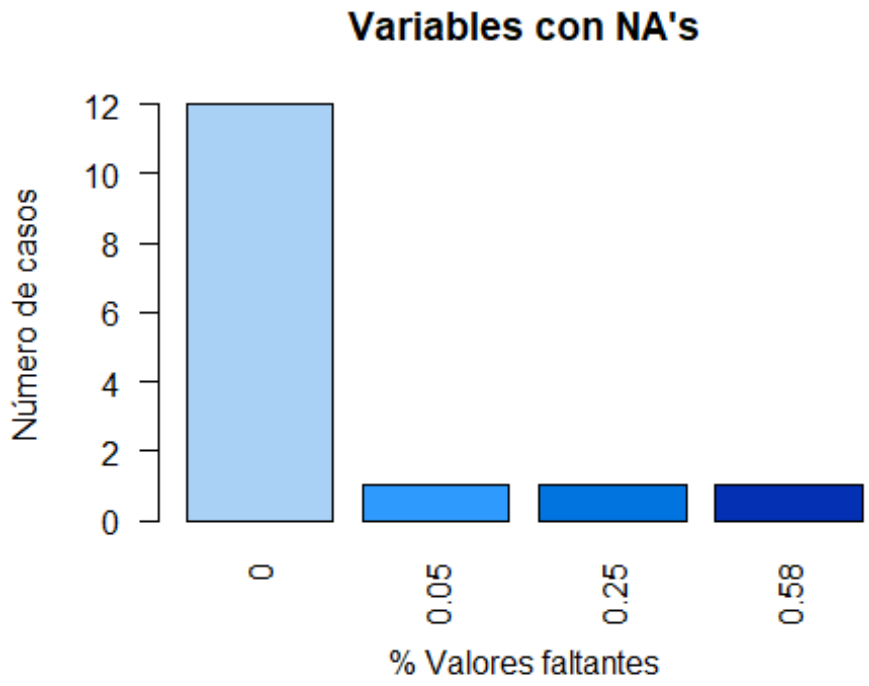

Gráfico 3: Porcentaje de valores faltantes por variable
Fuente: Elaboración propia

En el gráfico de barras se observa que no hay variables que superen el 20% de valores faltantes, siendo este el valor que se considera como referencia de porcentaje máximo aceptable de valores faltantes en una variable. Existen 3 variables con NA's, pero no llegan ni al 1% de observaciones con datos faltantes, por lo que más adelante se imputarán con la librería van Buuren and Groothuis-Oudshoorn (2011). En caso de que superaran el umbral fijado, se identificarían y se evaluaría si es correcto eliminarla o puede aportar información a pesar de tantos datos faltantes.

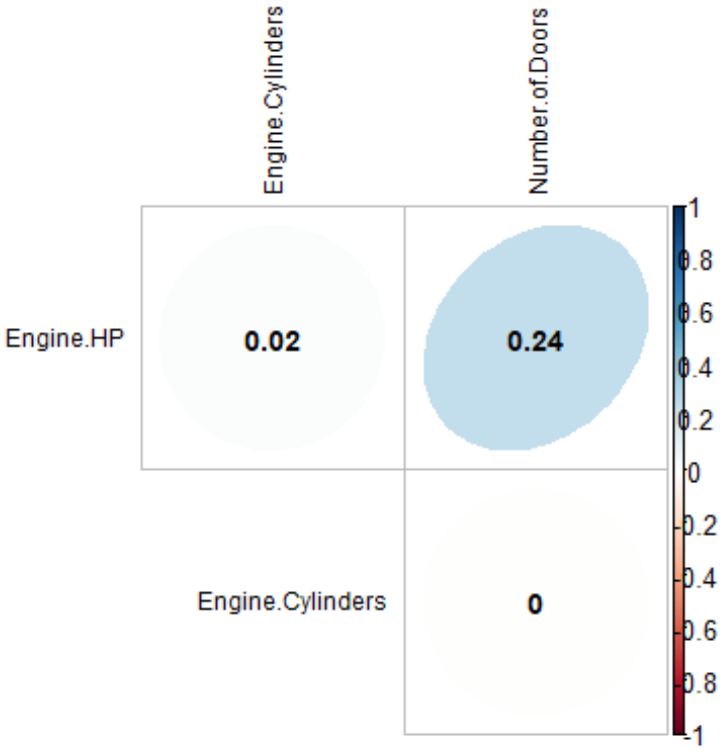

Gráfico 4: Correlación entre datos faltantes y variables
Fuente: Elaboración propia

Utilizando la librería Wei and Simko (2017) se crea una matriz de correlación de NA's con el objetivo de estudiar si existe aleatoriedad en el patrón de datos faltantes, es decir, si cuando falta una respuesta en una variable, suele faltar la respuesta a otra variable. Valores cercanos a 1 o -1 indicarían correlación entre los NA's.

En este caso, las correlaciones entre los valores faltantes de las variables son muy bajas, por lo que se puede deducir que hay aleatoriedad en el patrón, por lo tanto, no se debe de tomar ninguna medida.

3.3. Valores faltantes por caso

Se realiza un gráfico en el cual se muestra el porcentaje de valores faltantes que presentan las observaciones de la base de datos.

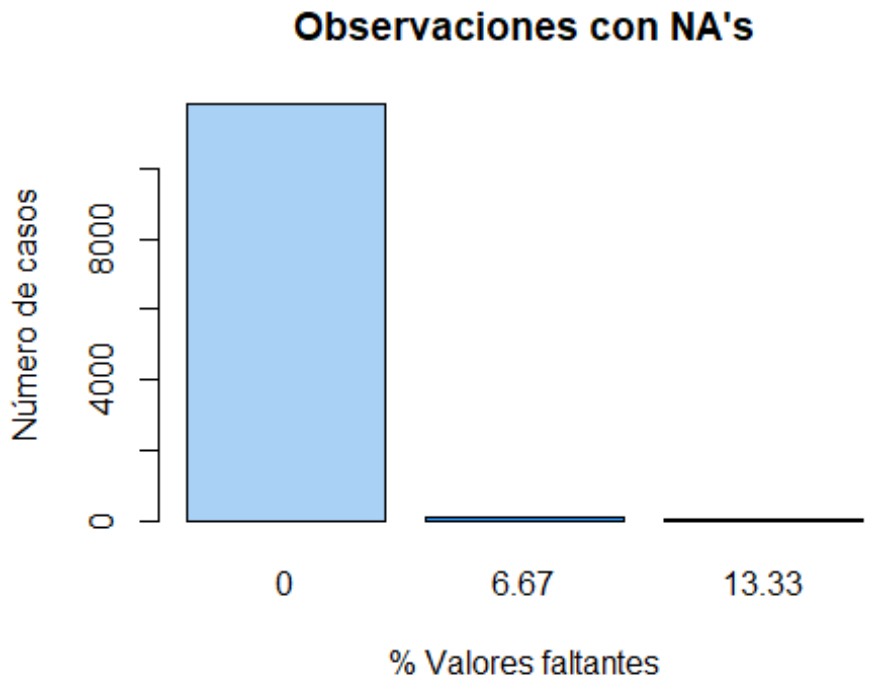

Gráfico 5: Porcentaje de valores faltantes por caso
Fuente: Elaboración propia

Se trata del mismo procedimiento realizado en el punto anterior, ya que se busca que ninguna observación tenga más del 20% de variables faltantes. En este caso tampoco existe ninguna observación con más del máximo porcentaje de datos faltantes que se ha establecido por lo que no se elimina ninguna observación.

3.4. Imputación de valores faltantes

Se imputan los valores faltantes con la librería van Buuren and Groothuis-Oudshoorn (2011). Como solo hay NA's en las variables numéricas, no hace falta que se pasen a factor las variables categóricas, ya que está librería no imputa los valores faltantes en caso de que las variables categóricas no fueran factor. También se imputan los valores de la variable "Transmission.Type", ya que tiene un tipo de transmisión "UNKNOWN" la cual no es detectada como NA por el R Studio. Tampoco sería eliminada ya que son 19 observaciones con UNKNOWN dentro de las 11.914 observaciones que presenta la base de datos.

A continuación, se realiza la comprobación para asegurarse de que no haya NA's en las variables que presentaban estos datos faltantes.

CV del motor:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  55.0   170.0   225.0   249.1   300.0    1001
```

Número de cilindros:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00    4.00    6.00    5.62    6.00   16.00
```

Número de puertas:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.00    2.00    4.00    3.40    4.00    4.00
```

Tipo de cambio de marcha:

```
AUTOMATED_MANUAL        AUTOMATIC       DIREC_DRIVE           MANUAL
             626             8272               68             2948
```

Tabla 5: Comprobación de variables con datos faltantes
Fuente: Elaboración propia

Finalmente, la base de datos ya está tratada y se puede comenzar a trabajar con ella.

4. Correlaciones

4.1 Correlaciones entre las variables numéricas

En esta sección se analizan las correlaciones lineales entre las variables. En la tabla 6 se presentan los coeficientes de correlación de las variables numéricas y en la tabla 7 su valor p para comprobar la significatividad de este coeficiente.

	Year	Engine.HP	Cylinders	Nº de puertas	highway.MPG	city.mpg	Popularity	MSRP
Year	1.00	0.35	-0.04	0.28	0.27	0.20	0.07	0.23
Engine.HP	0.35	1.00	0.77	-0.09	-0.40	-0.38	0.03	0.66
Cylinders	-0.04	0.77	1.00	-0.15	-0.66	-0.59	0.04	0.53
Nº de puertas	0.28	-0.09	-0.15	1.00	0.15	0.14	-0.08	-0.11
highway.MPG	0.27	-0.40	-0.66	0.15	1.00	0.94	-0.03	-0.17
city.mpg	0.20	-0.38	-0.59	0.14	0.94	1.00	0.00	-0.16
Popularity	0.07	0.03	0.04	-0.08	-0.03	0.00	1.00	-0.05
MSRP	0.23	0.66	0.53	-0.11	-0.17	-0.16	-0.05	1.00

Tabla 6: Correlación entre variables numéricas
Fuente: Elaboración propia

	Year	Engine.HP	Cylinders	Nº de puertas	highway.MPG	city.mpg	Popularity	MSRP
Year		0.00	0.00	0.00	0.00	0.00	0.00	0.00
Engine.HP	0.00		0.00	0.00	0.00	0.00	0.0001	0.00
Cylinders	0.00	0.00		0.00	0.00	0.00	0.00	0.00
Nº de puertas	0.00	0.00	0.00		0.00	0.00	0.00	0.00
highway.MPG	0.00	0.00	0.00	0.00		0.00	0.0049	0.00
city.mpg	0. 00	0.00	0.00	0.00	0.00		0.7255	0.00
Popularity	0.00	0.0001	0.00	0.00	0.0049	0.7255		0.00
MSRP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Tabla 7: Valor p de las correlaciones numéricas
Fuente: Elaboración propia

Todas las correlaciones, a excepción de la correlación entre consumo en ciudad y popularidad, tienen un valor igual o muy cercano a 0 (tabla 7), lo que significa que son significativas. En el caso de popularidad y consumo en ciudad la correlación no es significativa puesto que el valor de correlación es 0.

Respecto a las variables numéricas, se observa que las que más influyen en la determinación del precio son los caballos del motor y los cilindros. Además, lo hacen de forma positiva, es decir, cuantos más, más precio. De esta misma forma actúa la variable de los años, pero hay que tener en cuenta que es una variable en sentido ascendente (desde 1990 hasta 2017), lo que significa que cuanto más año (fecha de matriculación más reciente), más precio del coche.

Sin embargo, hay otras variables que actúan en el sentido contrario, se trata del consumo en ciudad, consumo en carretera, popularidad y número de puertas. Esto significa que, cuantos menos kilómetros recorre con un galón, (independientemente de si es en ciudad o en carretera) más precio. La variable de número de puertas actúa en el mismo sentido que el consumo, cuantas menos puertas, mayor será el precio.

Estas correlaciones hacen pensar que el análisis está afectado en gran medida por los coches deportivos, que tienen un motor mejor (correlación positiva con el precio) y suelen tener dos puertas y un consumo alto.

A priori un coche con cuatro puertas es más cómodo que otro que tenga dos y, por tanto, debería ser algo más caro el de cuatro puertas. Más adelante se realizará un análisis diferenciando por estilo de coche, para comprobar si esta hipótesis es cierta.

La popularidad tiene prácticamente un efecto nulo, ya que su valor es casi 0. No obstante, para asegurarse de que no tiene un efecto significativo en el precio, más adelante se analiza su posible impacto en el precio clasificando por rangos de popularidad.

Por otro lado, se comprueba la correlación positiva y cercana a 1 entre consumo en ciudad y consumo en carretera, es decir, que cuanto más consuma en ciudad, más consumirá en carretera. Además, destaca la correlación negativa de los consumos con los caballos y el número de cilindros, lo que significa que cuanto más caballos y número de cilindros, mayor será el consumo (recorre menos kilómetros con un galón).

Llama la atención la correlación positiva entre año y consumo de combustible (tanto en carretera como en ciudad), lo que significa que los coches modernos consumen menos que los antiguos, lo que sería fruto de una mejora en la fabricación de los vehículos y el uso de combustibles alternativos con bajos consumos (coches eléctricos). Además, hay una correlación positiva con el número de puertas, por lo que los modernos tienden a tener cuatro puertas en lugar de dos. Esto también será comprobado más adelante.

Resalta que cuanto más número de puertas, menor será el consumo de carburante tanto en carretera como en ciudad, sin embargo, menor será la potencia del motor y menor número de cilindros tendrá. Este dato da fuerza a la hipótesis sobre el efecto de los coches deportivos en esta base de datos.

Por último, para comprobar la relación del fabricante con el precio (variable categórica no incluida en este análisis), se realiza un test de chi cuadrado y se comprueba la significación estadística de esa relación, siendo la hipótesis nula que el precio entre fabricantes es igual, es decir, todos tienen el mismo valor, y siendo la hipótesis alternativa que los precios entre fabricantes son distintos

Pearson's Chi-squared test*:* Data: MSRP and Make

X-squared = 380475, df = 284256, **p-value < 2.2e-16**

El valor p es 0, es decir, la probabilidad de que la hipótesis nula fuera cierta asumiendo un grado de confianza del 95% es 0, por lo que el fabricante importa a la hora de determinar el precio.

4.2 Relación entre tamaño y consumo

Como ya ha sido comentado anteriormente, cuanto más consumo tenga en ciudad el vehículo, mayor será el consumo en carretera, por lo que ahora, se busca comprobar de qué forma afecta el tamaño del coche al consumo. Es lógico pensar que los grandes al mover más peso deberían consumir más. Además, en la siguiente tabla se comprueba el precio medio según el tamaño del vehículo.

Tamaño	Precio medio	Consumo ciudad (MPG)	Consumo carretera (MPG)
Compact	34.275 $	22.2	28.9
Large	53.891 $	16.1	22.4
Midsize	39.036 $	19.4	26.7

Tabla 8: Precio y consumo según tamaño
Fuente: Elaboración propia

Se aprecia que los más caros son los grandes, además de ser los que más consumen (consumo medido en millas por galón). Sucede lo contrario en los pequeños (Compacts), que son los más baratos y los que menor consumo tienen tanto en ciudad como en carretera. Esto viene a corroborar la matriz de correlación (sección 4.1), que indicaba una relación inversa entre el precio del vehículo con el consumo del mismo (cuanto más precio, menos millas recorridas con un galón, lo que significa mayor consumo).

4.3 Relación entre precio y popularidad

Se crea una variable con el intervalo en el que está el valor de la popularidad, ya que esta variable se presenta en valor numérico. A continuación, se genera una tabla entre popularidad y fabricante. Es importante remarcar que la popularidad se refiere a si es popular entre el público, en este caso, el público se trata de ciudadanos estadounidenses, por lo que puede darse el caso de que haya fabricantes conocidos en España que tengan poca popularidad. Los rangos son:

- Muy baja: [0,400]
- Baja: [401,800]
- Media: [801,1200]
- Media-alta: [1201,1600]
- Alta: [1601,2100]
- Muy alta: [2101,6000]

	Muy baja	Baja	Media	Media-alta	Alta	Muy alta
Acura	252	0	0	0	0	0
Alfa Romeo	5	0	0	0	0	0
Aston Martin	93	0	0	0	0	0
Audi	0	0	0	0	0	328
Bentley	0	74	0	0	0	0
BMW	0	0	0	0	0	334
Bugatti	0	0	3	0	0	0
Buick	196	0	0	0	0	0
Cadillac	0	0	0	0	397	0
Chevrolet	0	0	0	1123	0	0
Chrysler	0	0	187	0	0	0
Dodge	0	0	0	0	626	0
Ferrari	0	0	0	0	0	69
FIAT	0	0	62	0	0	0
Ford	0	0	0	0	0	881
Genesis	3	0	0	0	0	0
GMC	0	515	0	0	0	0
Honda	0	0	0	0	0	449
HUMMER	17	0	0	0	0	0
Hyundai	0	0	0	303	0	0
Infiniti	330	0	0	0	0	0
Kia	0	0	0	0	231	0
Lamborghini	0	0	52	0	0	0
Land Rover	143	0	0	0	0	0
Lexus	0	202	0	0	0	0
Lincoln	164	0	0	0	0	0
Lotus	0	29	0	0	0	0
Maserati	58	0	0	0	0	0
Maybach	16	0	0	0	0	0
Mazda	0	423	0	0	0	0
McLaren	0	5	0	0	0	0
Mercedes-Benz	0	353	0	0	0	0
Mitsubishi	0	213	0	0	0	0
Nissan	0	0	0	0	558	0
Oldsmobile	150	0	0	0	0	0

Plymouth	0	82	0	0	0	0
Pontiac	186	0	0	0	0	0
Porsche	0	0	0	0	136	0
Rolls-Royce	31	0	0	0	0	0
Saab	111	0	0	0	0	0
Scion	60	0	0	0	0	0
Spyker	3	0	0	0	0	0
Subaru	0	256	0	0	0	0
Suzuki	0	351	0	0	0	0
Tesla	0	0	0	18	0	0
Toyota	0	0	0	0	746	0
Volkswagen	0	0	809	0	0	0
Volvo	0	0	281	0	0	0

Tabla 9: Nivel de popularidad por fabricante
Fuente: Elaboración propia

Se observa que todos los fabricantes se sitúan en un rango de popularidad, lo que quiere decir que no depende del modelo, si no del fabricante, ya que todos tienen todos sus automóviles en un mismo rango.

Es lógico que el fabricante afecte al precio ya que es quien lo fija, pero ¿la percepción entre el público afecta? Se realiza una tabla de precios con la popularidad de los vehículos.

Popularidad	Precio medio
Muy baja	55.534 $
Baja	39.127 $
Media	42.805 $
Media-alta	28.272 $
Alta	34.749 $
Muy alta	43.979 $

Tabla 10: Precio medio según popularidad
Fuente: Elaboración propia

Se observa que a priori la popularidad no afecta al precio medio del coche. Esto es debido a que marcas de lujo como Maserati o Rolls-Royce aparecen en el rango de popularidad “Muy baja”, ya que es lógico que la mayoría de los ciudadanos no puedan adquirir estos vehículos y por tanto no sean tan populares como otros que son más asequibles.

4.4 Relación entre potencia de motor y estilo de coche

Estilo de vehículo	Media de CV	Precio medio	Consumo ciudad (MPG)
Coupe	329.	76.248 $	17.8
Convertible	313.	84.224 $	18.1
Crew Cab Pickup	301.	37.220 $	15.6
4dr SUV	263.	40.422 $	18.5
Passenger Van	255.	29.015 $	12.5
Extended Cab Pickup	242.	22.489 $	15.3
Sedan	240.	39.271 $	21.7
Cargo Van	214.	15.280 $	12.7
Regular Cab Pickup	213.	15.954 $	15.7
Passenger Minivan	213.	25.621 $	17.0
Wagon	200.	25.558 $	21.6
2dr SUV	182.	10.115 $	14.5
Cargo Minivan	169.	20.921 $	18.8
4dr Hatchback	162.	22.421 $	31.7
2dr Hatchback	160.	16.868 $	23.9
Convertible SUV	135.	17.424 $	20.2

Tabla 11: Potencia, consumo y precio según estilo de coche
Fuente: Elaboración propia

Se comprueba que los coches que más CV y precio son los descapotables y los coupés, lo que hace que los coches de dos puertas tengan un mayor precio de media, no por el hecho de tener 2 puertas, si no porque tienen mejor motor. No obstante, su consumo es menor que coches grandes como es el caso de “Crew Cab Pickup” o furgonetas como “Cargo Van”. El hecho de que el coche pueda ser descapotable hace que el precio sea 8.000$ mayor, aun teniendo un motor menos potente. Por lo que se puede concluir que el estilo del coche influye a la hora de determinar el precio.

4.5 Relación entre número de puertas y antigüedad del coche

Como ya ha sido comentado, de media, los coches con dos puertas tienen mejores motores y por tanto de ahí que sea superior el precio. Se corrobora en la siguiente tabla:

Número de puertas	Media.de.CV	Media de cilindros	Año de fabricación medio
2	263	6.03	2007
4	243	5.45	2012

Tabla 12: Potencia y antigüedad según número de puertas
Fuente: Elaboración propia

Además, tienen una fecha de matriculación más reciente los coches con 4 puertas, pero ¿existe una tendencia de fabricación de coches de dos puertas? Para ello, se realiza un gráfico que permita ver la evolución a lo largo de los años de la fabricación de coches de 4 y 2 puertas.

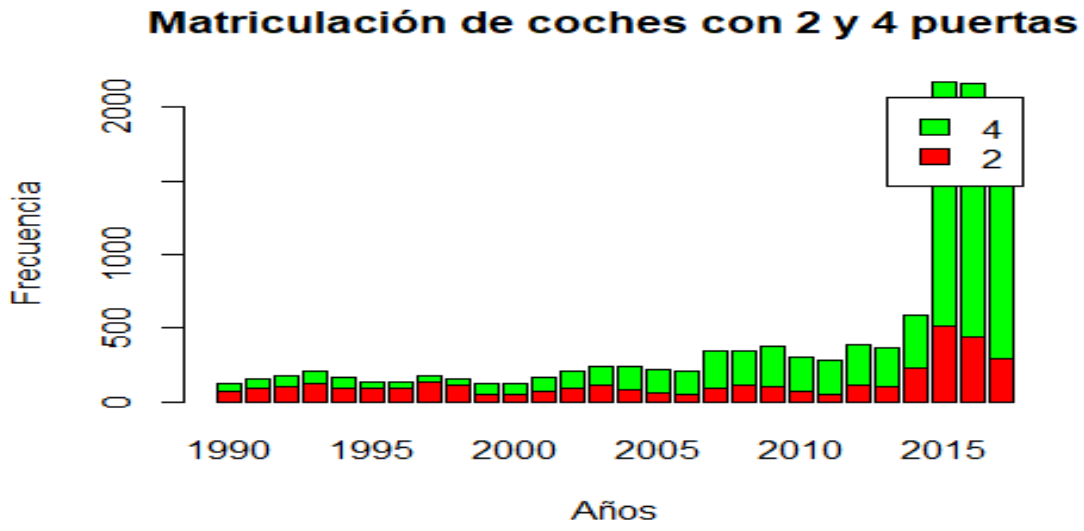

Gráfico 6: Evolución de coches con 2 y 4 puertas
Fuente: Elaboración propia

En el gráfico de barras se aprecia que en años inferiores al 2000, se matriculaban más coches con 2 puertas que con 4. Tras este periodo comienza un nuevo ciclo en el que, aparentemente, predomina la matriculación de coches con 4 puertas, consolidándose en el año 2015, donde los coches de 4 puertas triplican en número a los de 2 puertas.

Esta tendencia sigue creciendo, por lo que se podría extrapolar a la actualidad, siempre teniendo en cuenta que la base de datos presenta más coches de matriculación reciente, lo cual es una limitación. Habría que conocer la producción total de coches en Estados Unidos y realizar una evolución de la fabricación de coches de 2 y 4 puertas para poder confirmar esta tendencia con total seguridad.

4.6 Relación entre precio y tipo de cambio

Para comprobar que el tipo de cambio afecta al precio se realiza la siguiente tabla:

Tipo de transmisión	Precio medio	Año medio	Consumo carretera (MPG)
AUTOMATED_MANUAL	99.508 $	2014	28.7
AUTOMATIC	41.084 $	2012	25.8
DIREC_DRIVE	47.351 $	2015	98.0
MANUAL	26.555 $	2006	26.9

Tabla 13: Precio, antigüedad y consumo según tipo de cambio
Fuente: Elaboración propia

Aparentemente, el tipo de transmisión afecta al precio del automóvil. La base de datos diferencia entre tipo de cambio manual, automático, automático-manual (tipo de cambio automático con levas para poder conducir con transmisión manual si se desea) y el direct drive, también llamado caja de cambios de variación continua. Este sistema se diferencia por cuestiones mecánicas, ya que lo común es que, para transportar el giro del motor a las ruedas, se utilicen diversos engranajes. Este último sistema no sigue este mecanismo.

Para asegurar que la relación entre tipo de cambio y precio es significativa se realiza un test de chi cuadrado:

Pearson's Chi-squared test Data: MSRP and Transmission.Type

X-squared = 22634, df = 18144, **p-value < 2.2e-16**

El test da un valor p igual a 0, por lo que se comprueba la relación es significativa. Tras conocer la significatividad de la relación entre variables se procede a interpretar la tabla.

En la tabla se observa que los coches manuales son los coches más baratos de media, también es importante recalcar que este valor está afectado porque el año de matriculación fue más temprano que el resto, por lo que se realizará un análisis de componentes principales para ver en que grado afecta el tipo de cambio y el año de matriculación.

4.7 Relación entre motor y consumo.

Nº de cilindros	Potencia media(CV)	Consumo en ciudad (MPG)	Precio medio
0	163.	113.	47.914 $
3	77.4	33.5	10.992 $
4	173.	23.8	23.809 $
5	193.	19.2	20.819 $
6	261.	17.1	34.193 $
8	367.	14.1	61.487 $
10	578.	12.5	184.124 $
12	549.	11.2	285.178 $
16	1001	8	1.757.224 $

Tabla 14: CV, consumo y precio según número de cilindros
Fuente: Elaboración propia

Se realizará un estudio aparte de los coches cuyo motor tenga 0 cilindros, ya que aparentemente tienen un comportamiento anómalo.

Se observa que los que tienen 3 o más cilindros van en sentido ascendente en CV medios. Por lo que cuantos más cilindros tenga el motor más potencia tendrá el mismo. Sin embargo, el consumo va en sentido descendente, lo que significa que cuanta más potencia en el motor (y más cilindros), menor son los kilómetros que recorre con un mismo galón de gasolina, por lo que mayor es el consumo de combustible. Además, el precio medio aumenta según aumenta el número de cilindros.

En lo que respecta a los motores con 0 cilindros, no sigue este patrón de comportamiento que se ha mencionado anteriormente, por lo que se realiza un estudio aparte, para ver que coches son y que características tienen.

Make	Tipo de combustible	Tipo de cambio	Consumo en carretera	Consumo en ciudad
FIAT	electric	DIREC_DRIVE	108	122
FIAT	electric	DIREC_DRIVE	103	121
FIAT	electric	DIREC_DRIVE	103	121
Mercedes-Benz	electric	DIREC_DRIVE	82	85
Mercedes-Benz	electric	DIREC_DRIVE	82	85
Mercedes-Benz	electric	DIREC_DRIVE	82	85
Chevrolet	electric	DIREC_DRIVE	110	128
Chevrolet	electric	DIREC_DRIVE	110	128
Volkswagen	electric	DIREC_DRIVE	105	126
Volkswagen	electric	DIREC_DRIVE	105	126
Volkswagen	electric	DIREC_DRIVE	105	126
Volkswagen	electric	DIREC_DRIVE	105	126
Honda	electric	DIREC_DRIVE	105	132
Honda	electric	DIREC_DRIVE	105	132
Ford	electric	DIREC_DRIVE	99	110
Ford	electric	DIREC_DRIVE	99	110
Ford	electric	DIREC_DRIVE	99	110
Mitsubishi	electric	DIREC_DRIVE	99	126
Mitsubishi	electric	DIREC_DRIVE	99	126
Mitsubishi	electric	DIREC_DRIVE	102	121
BMW	electric	DIREC_DRIVE	111	137

BMW	electric	DIREC_DRIVE	111	137
BMW	electric	DIREC_DRIVE	111	137
BMW	electric	DIREC_DRIVE	106	129
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	126
Nissan	electric	DIREC_DRIVE	101	124
Nissan	electric	DIREC_DRIVE	101	124
Tesla	electric	DIREC_DRIVE	90	88
Tesla	electric	DIREC_DRIVE	97	94
Tesla	electric	DIREC_DRIVE	94	86
Tesla	electric	DIREC_DRIVE	90	88
Tesla	electric	DIREC_DRIVE	97	94
Tesla	electric	DIREC_DRIVE	102	101
Tesla	electric	DIREC_DRIVE	106	95
Tesla	electric	DIREC_DRIVE	98	89
Tesla	electric	DIREC_DRIVE	90	88
Tesla	electric	DIREC_DRIVE	105	102
Tesla	electric	DIREC_DRIVE	101	98
Tesla	electric	DIREC_DRIVE	105	92
Tesla	electric	DIREC_DRIVE	100	97
Tesla	electric	DIREC_DRIVE	107	101
Tesla	electric	DIREC_DRIVE	102	101
Tesla	electric	DIREC_DRIVE	107	101
Tesla	electric	DIREC_DRIVE	100	91
Tesla	electric	DIREC_DRIVE	90	88
Toyota	electric	DIREC_DRIVE	74	78
Toyota	electric	DIREC_DRIVE	74	78
Kia	electric	DIREC_DRIVE	92	120
Kia	electric	DIREC_DRIVE	92	120
Kia	electric	DIREC_DRIVE	92	120
Kia	electric	DIREC_DRIVE	92	120
Kia	electric	DIREC_DRIVE	92	120
Chevrolet	electric	DIREC_DRIVE	109	128
Chevrolet	electric	DIREC_DRIVE	109	128
Chevrolet	electric	DIREC_DRIVE	109	128
Chevrolet	electric	DIREC_DRIVE	109	128
Chevrolet	electric	DIREC_DRIVE	109	128
Chevrolet	electric	DIREC_DRIVE	109	128

Tabla 15: Identificación de coches con 0 cilindros
Fuente: Elaboración propia

Resalta el hecho de que todos son eléctricos y tienen un tipo de cambio llamado “DIREC_DRIVE”, el cual ya ha sido explicado. El hecho que explica que el consumo de estos coches sea tan escaso es que son eléctricos. Este consumo es cerca de 4 veces menor al consumo del resto de vehículos. Estos coches tienen un consumo tan escaso porque al ser eléctricos, consumen principalmente energía eléctrica y no combustibles convencionales.

Para corroborar este hecho, se realiza la siguiente sección en la que se aísla el consumo medio en carretera y ciudad según el tipo de combustible.

Al mismo tiempo, surge la duda de si todos los eléctricos tienen 0 cilindros y ese tipo de cambio.

Número de cilindros	Frecuencia
0	65
3	1

Tabla 16: Número de cilindros en coches eléctricos
Fuente: Elaboración propia

Existe un coche eléctrico que tiene 3 cilindros, pero lo frecuente es que tengan 0.

Tipo de combustible	Frecuencia
Eléctrico	66
Gasolina	2

Tabla 17: Tipo de combustible en tipo de cambio “Direc_drive”
Fuente: Elaboración propia

En esta tabla también se comprueba que los coches cuyo tipo de cambio es “DIREC_DRIVE” suelen ser eléctricos. Por lo que la conclusión es que los coches eléctricos suelen tener el tipo de cambio “DIREC_DRIVE” y 0 cilindros.

4.8 Relación entre tipo de combustible y consumo.

Para observar las diferencias de consumo según tipo de combustible utilizado por el vehículo, se realiza la siguiente tabla:

Tipo de carburante	Consumo en ciudad	Consumo en carretera
eléctrico	113	99.6
gas natural	27	38
diésel	26.4	36.6
gasolina	19.4	26.4
flex-fuel	16.0	22.6

Tabla 18: Consumo en función de combustible
Fuente: Elaboración propia

Como se puede observar en la anterior tabla, el coche eléctrico es el que menor consumo tiene, aproximadamente 4 veces menos consumo en ciudad que el segundo tipo de carburante que menos consume, que es el gas natural. No obstante, este carburante solo lo presentan 2 coches en toda la base de datos, y no es muy común observarlo.

Es importante remarcar que el coche eléctrico consume batería, por lo que la distancia que recorre el vehículo se mide con lo que sería el equivalente a un galón de combustible. Es decir, si con un galón de combustible un coche convencional recorre 25 millas, un coche eléctrico recorre con lo que sería un galón de batería (en lugar de estar medido en litros estará medido en kWh) 100 millas.

El sistema flex-fuel se trata de un motor que está diseñado para funcionar con el uso de dos combustibles, por ejemplo, gasolina y etanol.

4.9 Relación entre tipo de tracción y consumo

En esta sección se busca ver el impacto que tiene el tipo de tracción en los coches, centrándose en consumo y precio. Para ello, se realiza la siguiente tabla:

Tipo de tracción	Consumo en ciudad	Precio medio
Ruedas delanteras	23.7	23.057 $
Cuatro ruedas	17.6	49.960 $
Ruedas traseras	16.5	55.065 $

Tabla 19: Consumo y precio según tipo de tracción
Fuente: Elaboración propia

Los coches que mayor consumo presentan son los coches con tracción trasera, además de tener el precio más alto. Por el contrario, los coches con tracción en las ruedas delanteras son los más baratos y los que menos consumen. Los coches con tracción en las cuatro ruedas son similares a los coches con tracción trasera, pero con menor consumo y precio.

Clasificándolos según el tamaño:

	Compact	Large	Midsize
Cuatro ruedas	1053	1175	1528
Ruedas delanteras	2491	389	1907
Ruedas traseras	1220	1213	938

Tabla 20: Número de coches según tamaño y tracción
Fuente: Elaboración propia

Los pequeños tienden a tener la tracción en las ruedas delanteras mientras que los grandes suelen tenerlas en las cuatro ruedas o en las traseras. Los coches medianos se reparten más, pero con cierto predominio en las ruedas delanteras.

5. Evolución temporal de las variables de interés

Tras conocer mejor la base de datos y la caracterización de los coches, se busca conocer la evolución del precio de los vehículos, el consumo de estos, los caballos y sus cilindros. Para ello, se obtiene la media de cada variable para cada año, es decir, la media del precio de los coches de 1990, la media del precio de los coches de 1991…

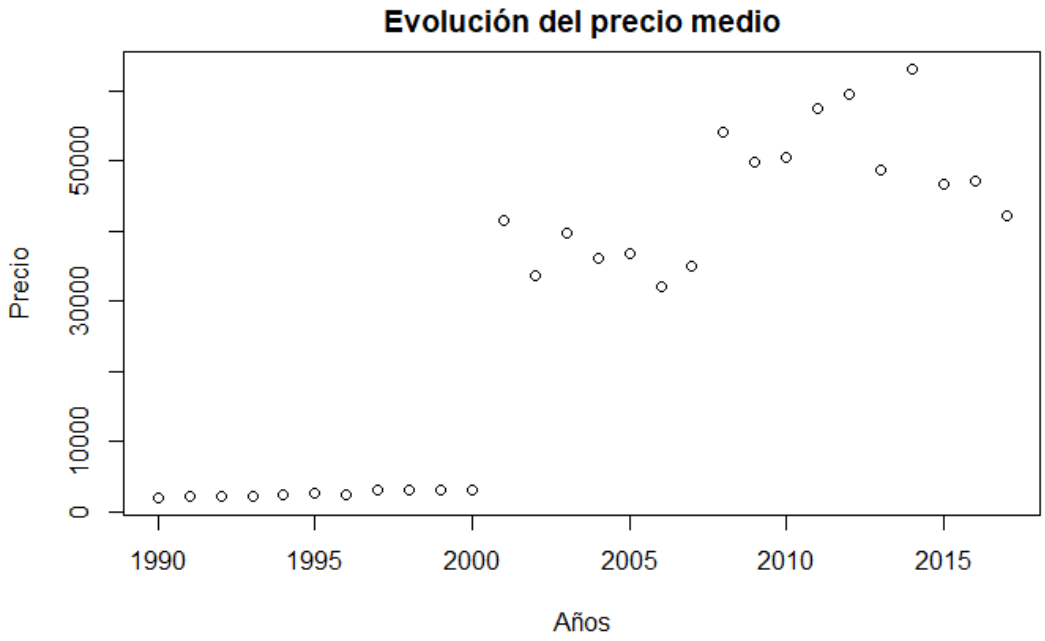

Gráfico 7: Evolución del precio medio de los coches
Fuente: Elaboración propia

En este primer gráfico se observa el precio medio de los coches en función del año de matriculación. No se puede determinar un patrón en el precio de los vehículos ya que el precio de los coches antiguos está muy afectado por el año en el que se obtiene la base de datos (2017).

Para poder conocer con exactitud si existe un aumento en el precio medio de los vehículos como consecuencia, por ejemplo, de la inflación, habría que conocer el precio exacto del vehículo en el año de matriculación.

Sin embargo, a la vista del gráfico si se puede observar la gran diferencia que existe en el precio de los coches matriculados después del 2000. Con una diferencia de 1 o 2 años, presentan una diferencia en el precio medio de más de 20.000$.

No obstante, otras variables técnicas sí que pueden ser estudiadas con mayor precisión y encontrar, si existe un patrón a lo largo del tiempo. Entre ellas, la media de los caballos de los vehículos.

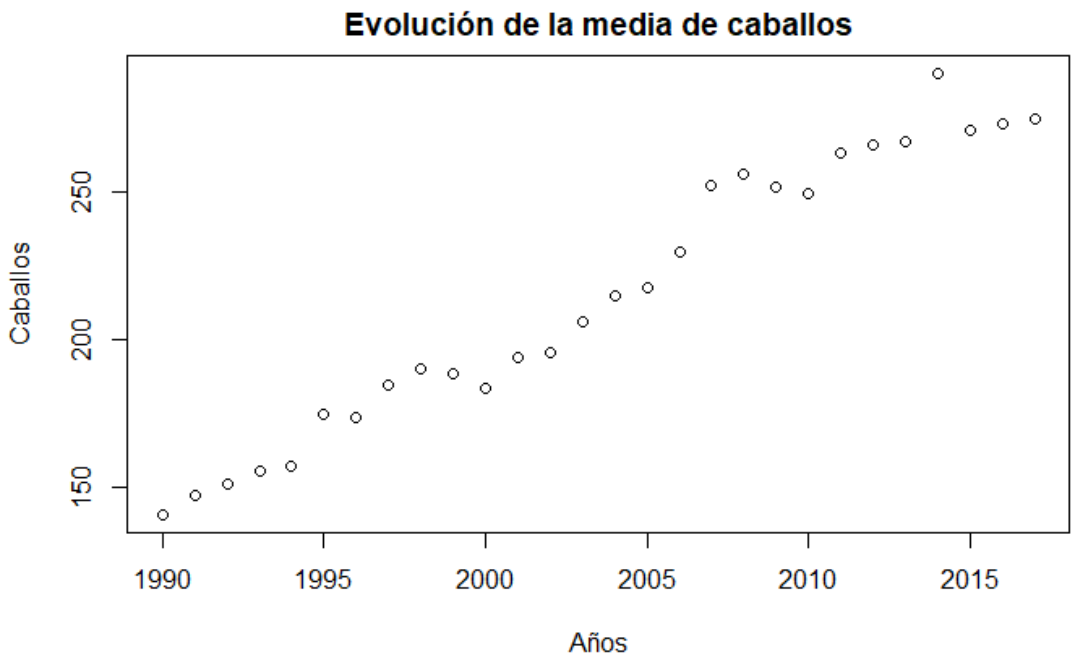

Gráfico 8: Evolución de la media de caballos
Fuente: Elaboración propia

En este caso, se puede apreciar como los caballos de los coches van en sentido ascendente, denotando una mejora técnica de los vehículos. La línea tiene ligeras desviaciones, como es el caso del año 2014, que presenta vehículos con una mayor media de caballos debido a la presencia de más coches lujosos.

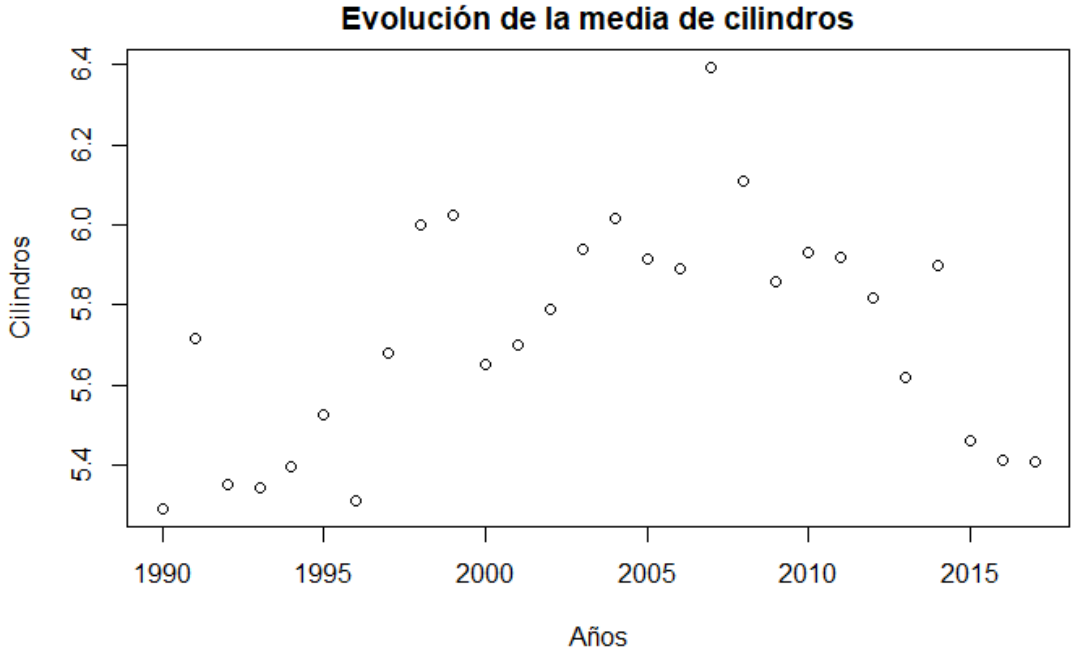

Gráfico 9: Evolución de la media de cilindros
Fuente: Elaboración propia

En este caso, los cilindros no siguen un patrón claro, y se presentan dispersos. Es curioso, ya que los cilindros están correlacionados con la cantidad de caballos del coche, por lo que, a más caballos, más cilindros. Sin embargo, al observar la siguiente tabla se descubre el porqué:

Nº	Year	Engine.Fuel.Type	Nº	Year	Engine.Fuel.Type
1	2015	electric	34	2016	electric
2	2016	electric	35	2014	electric
3	2017	electric	36	2014	electric
4	2015	electric	37	2014	electric
5	2016	electric	38	2014	electric
6	2017	electric	39	2015	electric
7	2017	electric	40	2015	electric
8	2017	electric	41	2015	electric
9	2015	electric	42	2015	electric
10	2015	electric	43	2015	electric
11	2016	electric	44	2016	electric
12	2016	electric	45	2016	electric
13	2013	electric	46	2016	electric
14	2014	electric	47	2016	electric
15	2015	electric	48	2016	electric
16	2016	electric	49	2016	electric
17	2017	electric	50	2016	electric
18	2014	electric	51	2016	electric
19	2016	electric	52	2016	electric
20	2017	electric	53	2012	electric
21	2015	electric	54	2013	electric
22	2016	electric	55	2014	electric
23	2017	electric	56	2015	electric
24	2017	electric	57	2015	electric
25	2014	electric	58	2016	electric
26	2014	electric	59	2016	electric
27	2014	electric	60	2016	electric
28	2015	electric	61	2014	electric
29	2015	electric	62	2014	electric
30	2015	electric	63	2015	electric
31	2015	electric	64	2015	electric
32	2016	electric	65	2016	electric
33	2016	electric	66	2016	electric

Tabla 21: Recuento de todos los coches eléctricos
Fuente: Elaboración propia

Los coches eléctricos aparecen en la base de datos con el año de matriculación más tardío en 2014, que es el punto donde decae el número medio de cilindros, ya que hasta 2010 presenta cierta tendencia al alza.

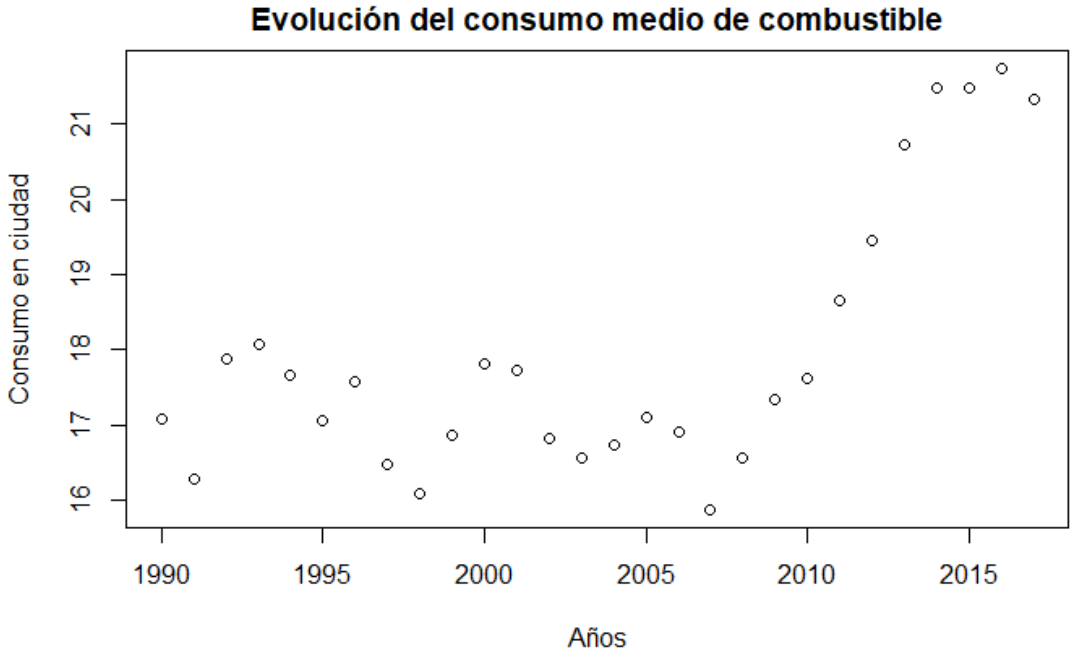

Gráfico 10: Evolución del consumo medio de combustible
Fuente: Elaboración propia

Por último, se representa el consumo medio de combustible en ciudad. No se representa el consumo en carretera ya que ambos siguen la misma línea y la gráfica sería muy similar.

Desde 2006, los coches de la base de datos reducen su consumo, fortaleciendo así la conclusión de la sección 4.1, que indicaba una correlación positiva entre año y consumo de combustible. A pesar de que la línea sea ascendente, no hay que olvidar que se está midiendo el consumo en el formato americano, el cual es Millas por Galón, esto significa que los coches cada vez recorren más millas con un mismo galón. Esto es debido a la mejora en la fabricación y a la mejora de los combustibles.

6. Análisis de componentes principales (PCA)

6.1 Formulación del modelo

Con este análisis se reduce el impacto oculto que puedan tener otras variables a la hora de estudiar la relación entre dos variables, como es el caso del número de puertas y precio (afectado por el año de matriculación) o tipo de cambio y precio (también afectado por el año de matriculación). Este modelo se formula mediante el uso de la librería Lê, Josse, and Husson (2008).

El PCA solo admite variables numéricas, por lo que para poder incluir todas las variables se han pasado las variables categóricas a numéricas, por ejemplo, en el caso del tamaño del vehículo están los valores 1 (Compact), 2 (Midsize) y 3 (Large). Están todas las variables a excepción del fabricante, que tiene 48 niveles y resulta difícil identificar un fabricante con un número y estilo de vehículo que tiene 16 niveles, pero no están excluidas del modelo, ya que se tienen en cuenta como complemento para formularlo.

Entre los diferentes valores asignados a las variables categóricas están:

- Tipo de carburante: 1- Diésel, 2- Eléctrico, 3- Flex-fuel, 4- Gas natural, 5- Gasolina
- Tipo de cambio: 1- Automated manual, 2- Automatic, 3- Direct drive, 4- Manual
- Tipo de tracción: 1- Trasera, 2- Delantera, 3- Cuatro ruedas
- Tamaño del vehículo: 1- Compact, 2- Midsize, 3- Large

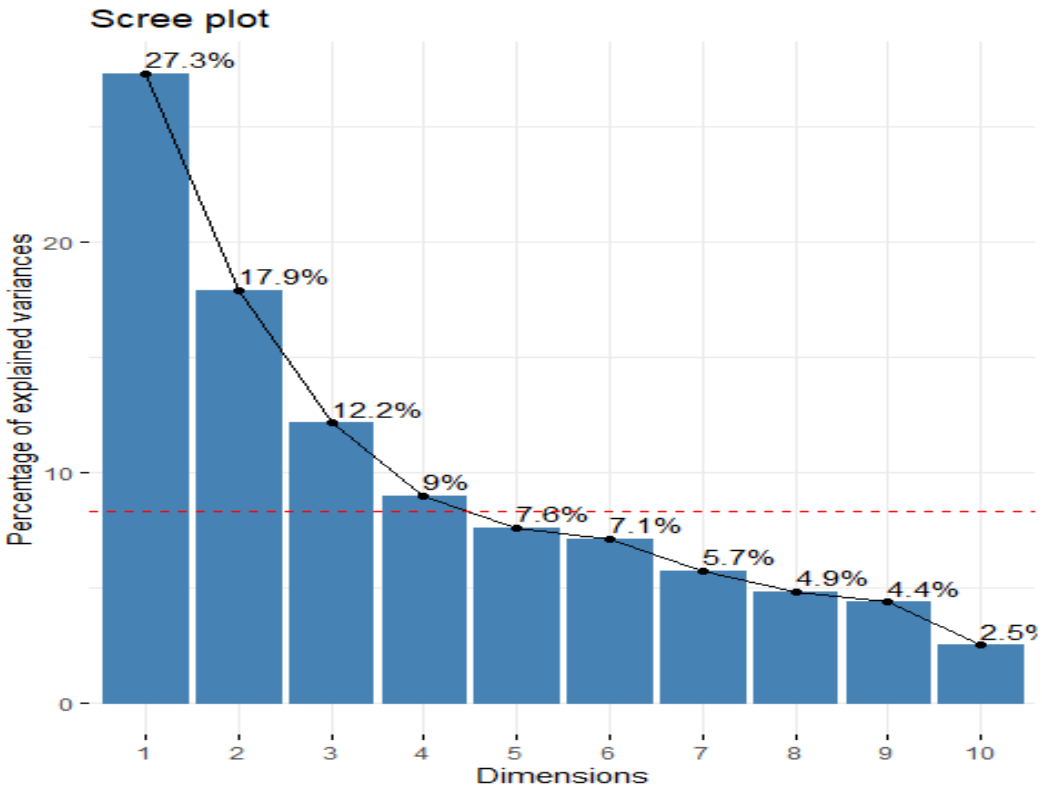

Gráfico 11: Dimensiones del PCA
Fuente: Elaboración propia

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.2767204	27.306003	27.30600
Dim.2	2.1502576	17.918813	45.22482
Dim.3	1.4583689	12.153074	57.37789
Dim.4	1.0814527	9.012106	66.39000
Dim.5	0.9128045	7.606704	73.99670
Dim.6	0.8550637	7.125531	81.12223

Tabla 22: Porcentaje de explicación de cada dimensión en PCA
Fuente: Elaboración propia

Se escogen 6 dimensiones, que explican un 81,12% de la variabilidad del modelo. Una dimensión está compuesta por distintas variables, de forma que es más fácil explicar la base de datos ya que la simplifica al englobar varias variables en una dimensión. Con este porcentaje de variabilidad del modelo se está cerca del 100%, y a partir de la dimensión 6 los valores de explicación del modelo son pequeños.

6.2 Validación del modelo PCA

Tras la formulación del modelo, hay que realizar su validación. Para ello, se recurre al Test de Hotelling. Este test es una medida de distancia desde la proyección de una observación al centro del modelo. A continuación, se representa un gráfico con los valores del T^2 y el límite de confianza al 95% en naranja. Las observaciones que caigan fuera del límite naranja son consideradas extremas.

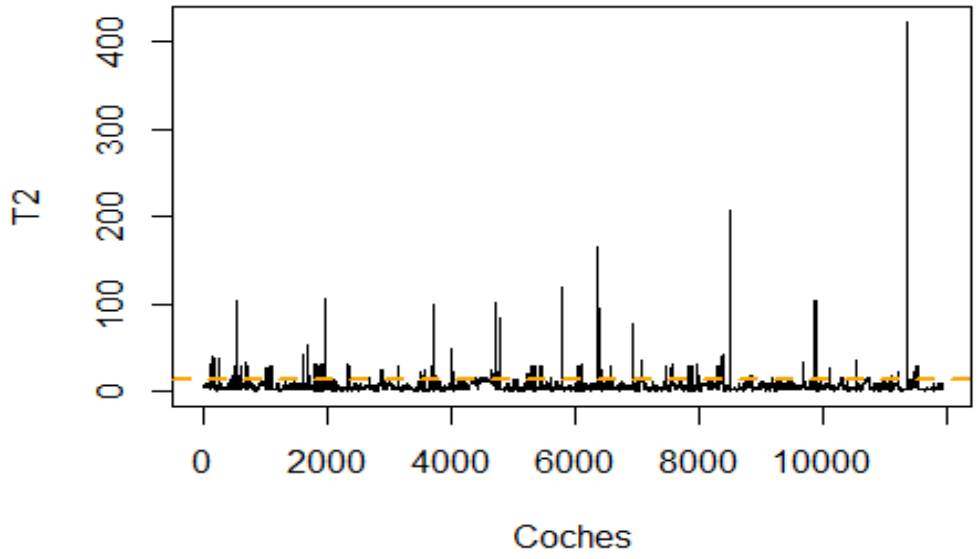

Gráfico 12: Control de observaciones anómalas
Fuente: Elaboración propia

En el gráfico apenas se observa cuáles superan la línea naranja, por lo que se realiza el cálculo numéricamente.

Observaciones anómalas: 862

En este caso, se trata de 862 observaciones que superan el límite del 95% de confianza. Sin embargo, se esperan 596 falsos positivos, es decir, el 5% de las 11914 observaciones que componen este estudio, y en este caso, al encontrar 862 casos que superan el error, habría que eliminar las 266 observaciones con mayor T^2.

No obstante, al tratarse de un número tan escaso dentro de una base de datos con tantas observaciones y siendo un poco flexible, no se eliminan, puesto que se trata de 2% de observaciones anómalas y su impacto es prácticamente nulo.

6.3 Interpretación del modelo PCA

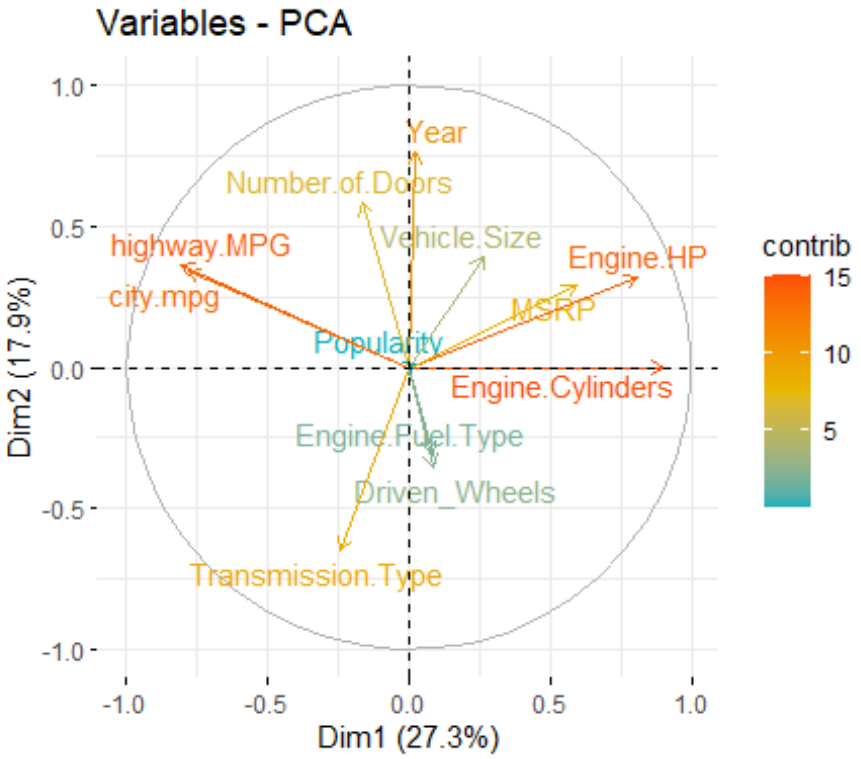

Gráfico 13: Dimensión 1 y 2 PCA
Fuente: Elaboración propia

La dimensión 1 (eje horizontal) del modelo está explicada por la potencia del motor, el número de cilindros, el consumo de combustible y el precio.

En el gráfico, se comprueba lo visto en la sección 4.1, en la que se comprobaba que cuánto más CV y cilindros más precio y más consumo de combustible. El consumo está en sentido contrario debido a que la medida de consumo es Miles per Gallon, es decir, menos millas recorridas con un galón, o lo que es lo mismo, mayor consumo). El tamaño del vehículo tiene un papel secundario y el año no tiene ningún efecto en esta dimensión.

La dimensión 2 (eje vertical) está explicada principalmente por el año, que va en el mismo sentido que el número de puertas, es decir, cuanto más año (más tardía la fecha de matriculación), más puertas, confirmando así la posible tendencia a la fabricación de coches con 4 puertas. En el sentido contrario aparece el tipo de transmisión, lo que indica que cuanto más año, menos valor en TransmissionType. Esta variable tiene actualmente los valores 1 y 2 para el tipo de cambio "AUTOMATIC" y "AUTOMATED_MANUAL" respectivamente, siendo el 4 el "MANUAL", por lo que, los coches más nuevos, tienden a tener un tipo de cambio automático, ya que valores altos en el año indican valores bajos (4) en el tipo de cambio (manual) y viceversa.

Para comprobar estas afirmaciones se generan dos gráficos en los que se puede comprobar la aportación de cada una de las variables a las dos dimensiones.

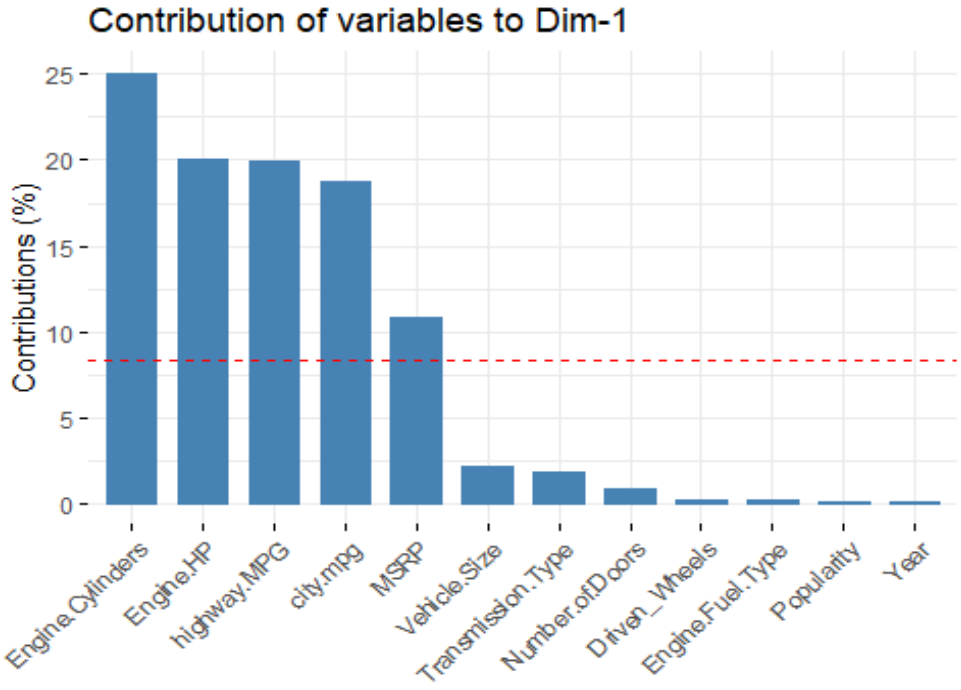

Gráfico 14: Contribución de variables a la dimensión 1
Fuente: Elaboración propia

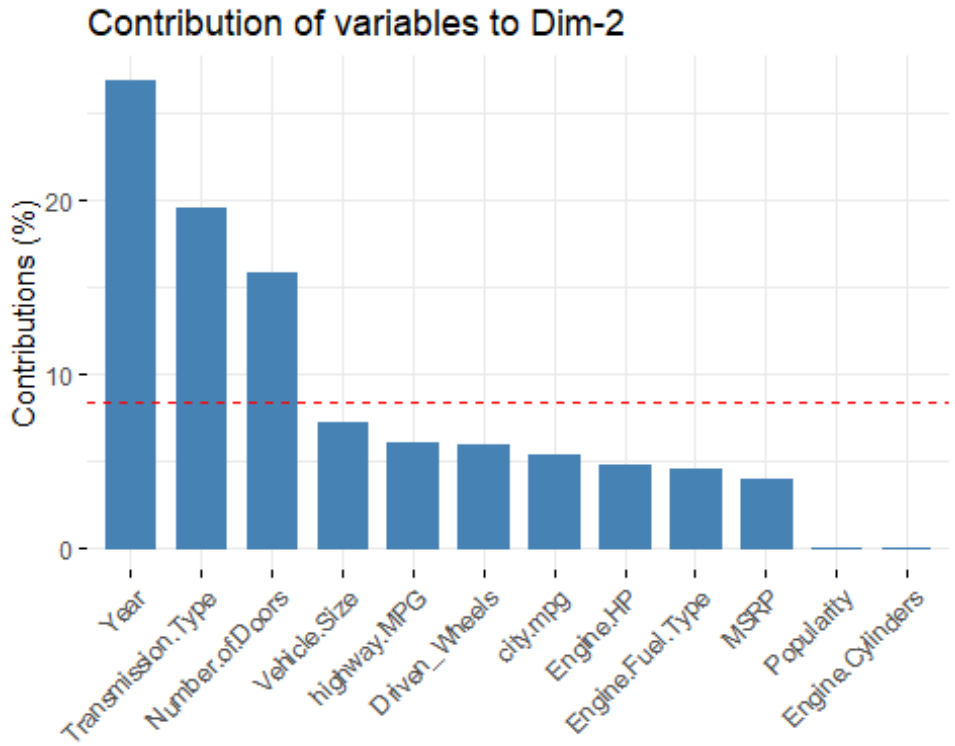

Gráfico 15: Contribución de variables a la dimensión 2
Fuente: Elaboración propia

Tras hacer esta comprobación, se procede a explicar las dimensiones 3 y 4.

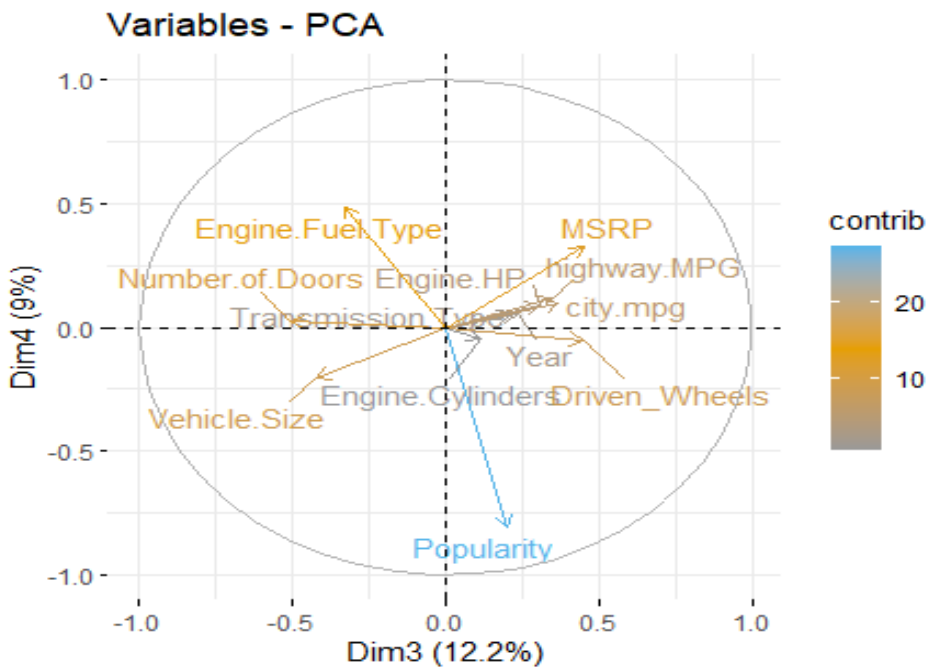

Gráfico 16: Dimensión 3 y 4 PCA
Fuente: Elaboración propia

La dimensión 3 está explicada principalmente por el número de puertas, tamaño de vehículo, consumo de combustible, precio y tipo de tracción.

Cuanto menos tamaño del vehículo, menor número de puertas, más precio y menos consumo, ya que a pesar de ir en sentido contrario significa que serán menos millas recorridas con un galón de combustible. Además, en estos casos el tipo de tracción será en las cuatro ruedas o tracción delantera (valores 2 y 3).

En esta dimensión están agrupados los vehículos pequeños deportivos, los cuales tienen tendencia a tener la tracción delantera o de cuatro ruedas, teniendo menos puertas que el resto y un mayor precio. Además, se explica el consumo menor ya que esta base de datos incluye furgonetas, todoterrenos y otros tipos de coche grandes, con un gran consumo de combustible.

La dimensión 4 está explicada casi en su totalidad por popularidad y tipo de combustible, pero afecta un poco el precio.

De esta dimensión se puede extraer que cuanta más popularidad, menos valor tiene la variable de tipo de combustible, es decir, será o diésel, o eléctrico, incluso flex-fuel (valores 1,2 y 3 respectivamente), pero no gasolina (valor 5). Por lo que los fabricantes que no producen vehículos a gasolina son más populares. Además, el precio va en el mismo sentido que el tipo de combustible, por lo que significa que los coches que utilizan gasolina como combustible, tienden a ser más caros.

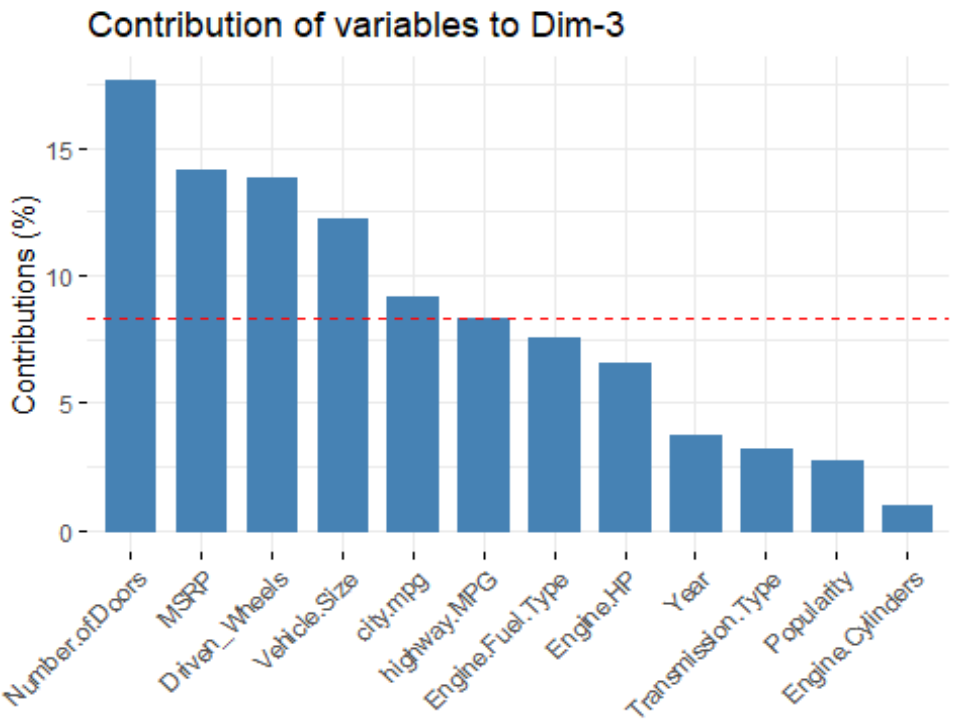

Gráfico 17: Contribución de variables a la dimensión 3
Fuente: Elaboración propia

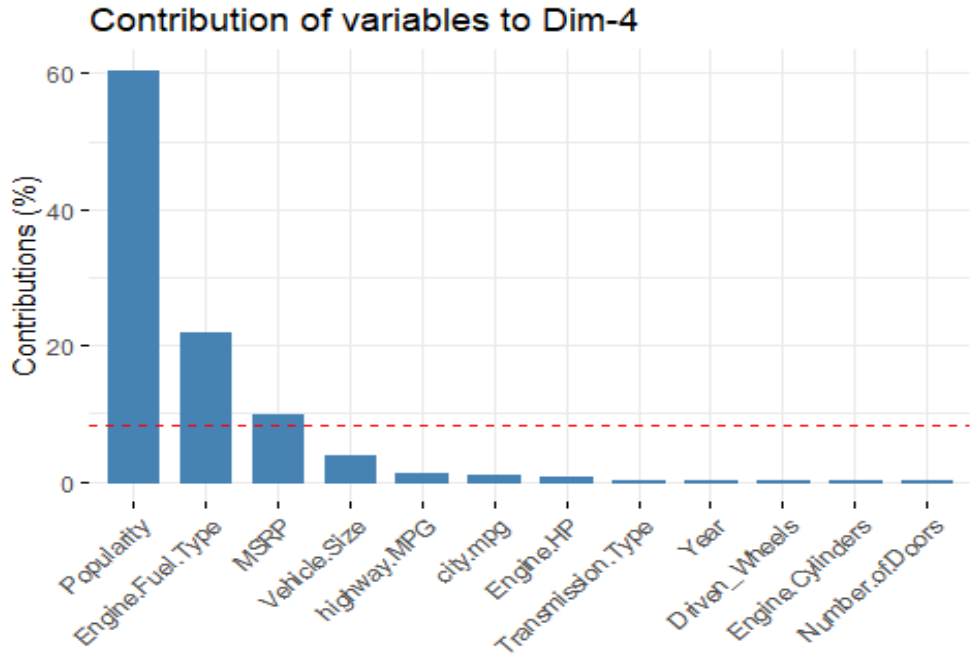

Gráfico 18: Contribución de variables a la dimensión 4
Fuente: Elaboración propia

Tras confirmar la contribución de cada variable a la dimensión 3 y 4, se genera el gráfico de las dimensiones 5 y 6

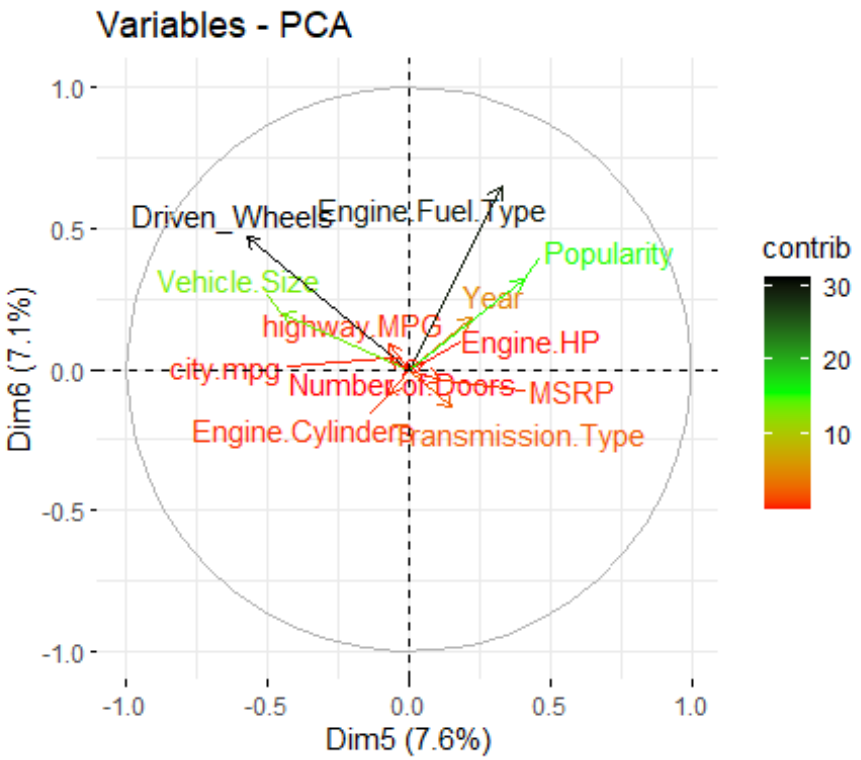

Gráfico 19: Dimensión 5 y 6 PCA
Fuente: Elaboración propia

Respecto a la dimensión 5, se observa que está explicada por el tipo de tracción, el tamaño del vehículo, la popularidad y el tipo de combustible.

Cuanto más grande el vehículo, menos popularidad tendrá. Además, el vehículo tendrá tracción en las cuatro ruedas o delantera (máximos valores otorgados en su categoría). El coche será diésel o eléctrico (valores 1 y 2 respectivamente).

Por último, la dimensión 6, que está explicada por el tipo de combustible y el tipo de tracción, lo que significa que, a más valor de tipo de combustible, más valor de tipo de tracción, lo que significa que los coches que son gasolina (máximo valor), tendrán tracción en las cuatro ruedas, y que los coches diésel o eléctricos (valores más bajos), tendrán tracción trasera.

7. Modelos de predicción

7.1 Preparación de la base de datos

En este apartado se llevan a cabo distintos modelos supervisados de clasificación con el objetivo de predecir si un coche tendrá un precio mayor a 30.000 $ (mediana), o será inferior. La mediana es el valor donde se encuentran el 50% de las observaciones, es decir, la posición central de todos los valores, en este caso, para la variable precio. Se busca predecir esto, ya que 30.000 $ es un precio considerable para un coche y de esta forma se divide en dos la base de datos, dividendo justo por la mitad de las observaciones. Se podría dividir por la media, pero valores extremos en el precio (coches de más de 2.000.000 $) hacen que este valor sea elevado y por tanto haya más cantidad de coches por debajo de la media. Con el uso de la mediana se evita este efecto y se busca predecir qué características tendrá un coche para estar en un grupo u otro.

En primer lugar, se elimina la variable precio, puesto que sería determinante en todos los modelos (si es mayor que 30.000 $ predecirá que estará por encima de la mediana) y a continuación se divide la base de datos en dos grupos diferentes, uno para elaborar los modelos, siendo este el grupo de entrenamiento, y otro para validarlos. Finalmente, se evalúan los modelos con la curva ROC y el coeficiente AUC utilizando la librería Sing et al. (2005).

Una vez dividida la base de datos en dos subconjuntos (el de entrenamiento con un 60% de las observaciones y el de validación con un 40%), se procede a realizar los distintos modelos de predicción.

Los modelos utilizados serán

1. Regresión general
2. Árbol de partición
3. Random Forest
4. Naive Bayes
5. Naive Bayes Laplace
6. Soporte vectorial (SVP)
7. Soporte vectorial (SVM)
8. Vecino más próximo

7.2 Construcción de los modelos

7.2.1 Modelo de regresión general

a) Construcción del modelo con los datos de entrenamiento

Con el conjunto de observaciones de las que se dispone, el objetivo es generar un modelo que pueda predecir nuevos casos. El modelo genera un coeficiente para cada variable de forma que según su valor numérico y su significatividad se observa que variables y que niveles dentro de la misma son más significativas a la hora de predecir si un coche será superior a 30.000 $.

El modelo no se muestra puesto que al haber 48 niveles en fabricante (48 fabricantes distintos) es demasiado extenso. Por lo que simplemente se calcula su porcentaje de acierto en la predicción.

b) Evaluación del modelo

Porcentaje de acierto: 0.9769476

Este modelo presenta un porcentaje de acierto del 97.69%,

7.2.2 Árbol de partición

a) Construcción del modelo con los datos de entrenamiento

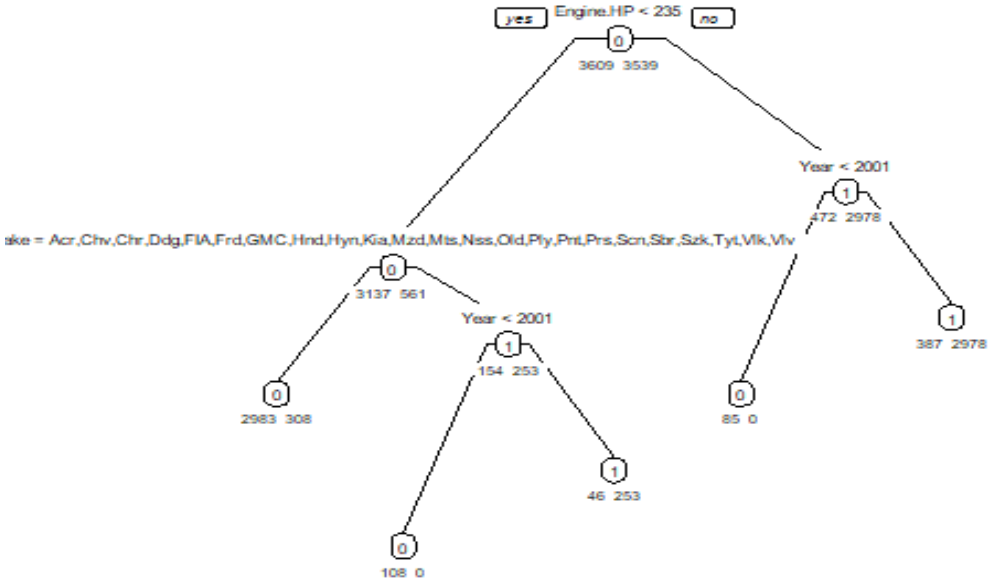

Ilustración 1: árbol de partición con 4 divisiones
Fuente: Elaboración propia

Classification tree:
Variables actually used in tree construction: Engine.HP, Make, Year

Root node error: 3539/7148 = 0.4951

n= 7148

CP	nsplit	rel error	xerror	xstd
0.708110	0	1.00000	1.02063	0.0119442
0.029246	1	0.29189	0.29556	0.0084436
0.024018	3	0.23340	0.24188	0.0077564
0.010000	4	0.20938	0.21814	0.0074150

Tabla 23: Error según número de particiones
Fuente: Elaboración propia

Mediante el uso de las librerías Therneau and Atkinson (2019) se crea el árbol de partición. Este árbol divide en función de las características de la base de datos para que los elementos de los subconjuntos creados sean lo más parecido posible entre ellos. El mejor árbol es que tiene menor x-error, en este caso es el de 4 “splits” o divisiones y, por tanto, no se poda.

b) Evaluación del modelo

Porcentaje de acierto: 0.9039998

7.2.3 Random forest

a) Construcción del modelo con los datos de entrenamiento

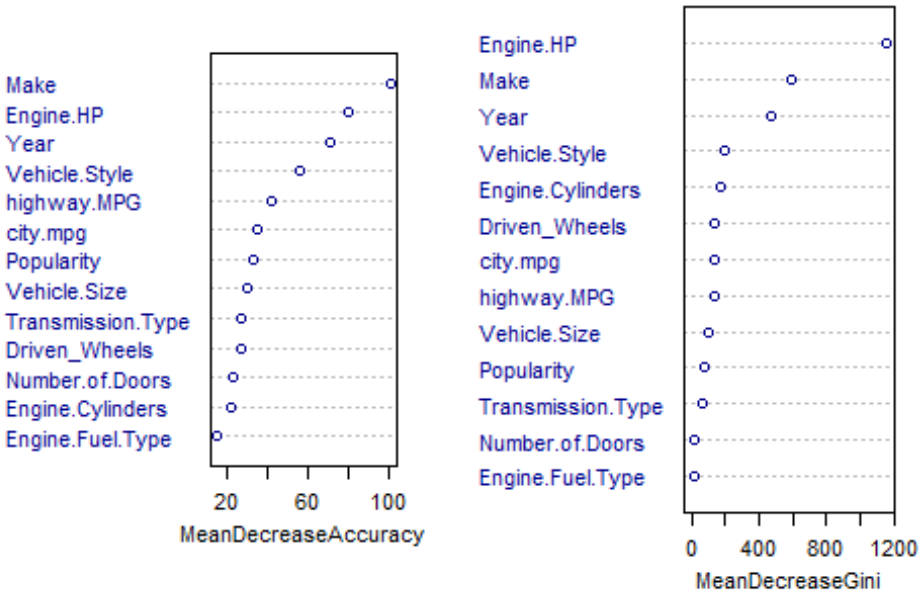

Ilustración 2: Variables significativas según random forest
Fuente: Elaboración propia

El software Liaw and Wiener (2002) permite optimizar el estudio en base a árboles creando por defecto 500. Para desarrollar el modelo se establece 3 como el parámetro "mtry", que es el número de variables elegidas al azar en cada split. Este número se escoge ya que es el número inferior a 3.74, es decir, la raíz cuadrada del número de variables (14).

Al crear el modelo se puede observar que las variables fabricante, potencia del motor y el año de matriculación son determinantes para predecir si el coche tendrá un precio superior a 30.000 $. El orden de estas dependerá del criterio considerado que uno elija. Si se quiere ser más preciso a la hora de predecir la clase (MeanDecreaseAccuracy) o si se quiere ordenar por importancia de las variables en el modelo (MeanGiniDecrease).

b) Evaluación del modelo

Porcentaje de acierto: 0.9804529

El random forest presenta más porcentaje de acierto que el árbol de partición

7.2.4 Naive Bayes

a) Construcción del modelo con los datos de entrenamiento

Utilizando el software Meyer et al. (2020) se crea el modelo de NaiveBayes. Este modelo es un modelo probabilístico basado en la teoría de Bayes. Su fórmula matemática es:

$$P(C/X1, X2, \ldots, Xp) = \frac{P(X1, X2, \ldots, Xp/C) * P(C)}{P(X1, X2, \ldots, Xp)}$$

C es la clase y X son los valores predictivos.

Tras su formulación, se evalúa su porcentaje de acierto en la predicción.

b) Evaluación del modelo creado

Porcentaje de acierto: 0.9546781

7.2.5 Máquinas de soporte vectorial

a) Construcción del modelo con los datos de entrenamiento

Utilizando la librería Meyer et al. (2020) se obtiene el modelo SVM y Karatzoglou et al. (2004) el modelo SVP. Estos modelos son métodos de aprendizaje supervisados empleados para la clasificación de los datos. Las máquinas de soporte vectorial buscan un hiperplano donde se separa de forma óptima los puntos de una clase de los puntos de otra clase.

b) Evaluación del modelo creado

Porcentaje de acierto SVM: 0.9826459

Porcentaje de acierto SVP: 0.9800656

7.2.6 Vecino más próximo

Para poder llevar a cabo el método del vecino más próximo se necesita primero pasar las variables categóricas a numéricas. Una vez pasadas, se genera el modelo. Este modelo consiste en predecir el valor de una observación en función de las observaciones más cercanas. Es un modelo que utiliza distancias, es decir, cuanta más distancia haya entre observaciones, mayor serán las diferencias entre ellas.

a) Determinación de las clases más cercanas

Mediante el software Venables and Ripley (2002) se construye el modelo, determinando la clase más cercana.

b) Predicción en base al modelo creado

Porcentaje de acierto: 0.502962

7.3 Curvas ROC de los modelos

Una vez ya han sido creado todos los modelos, se compara cómo de buenas son las predicciones creadas en base a ellos con sus respectivas curvas ROC. La curva ROC mide el área que hay por debajo de la curva, cuanto más cercano a 1 mejor será el modelo para predecir.

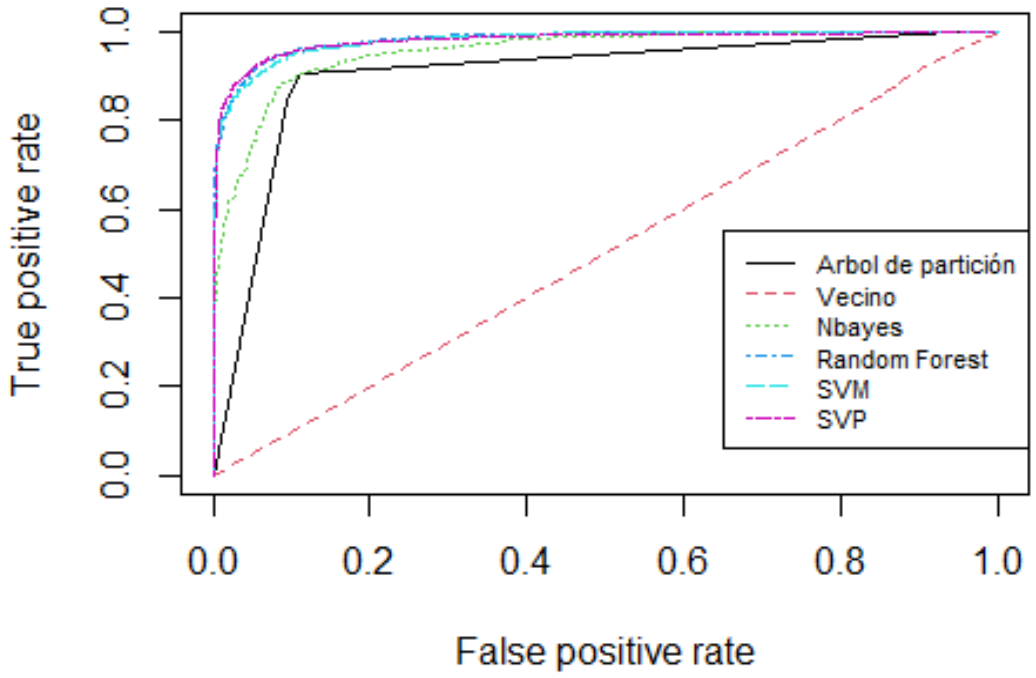

Gráfico 20: Curvas ROC de los modelos
Fuente: Elaboración propia

Comparando las curvas ROC de cada uno de los modelos se obtiene que los creados con el método de la máquina de soporte vectorial y el random forest son aquellos con mejor precisión. Siendo el área debajo de la curva ROC del modelo SVP 0.9800656 y el del SVM 0.9826459. El valor del área debajo de la curva con el random forest es 0.9804529.

Al ser el random forest el siguiente segundo mejor modelo, y siendo su interpretabilidad mayor se opta por analizar este modelo.

8. Conclusión

Como ya ha sido comentado en el apartado correspondiente al random forest, las variables más significativas para determinar si un coche tiene un precio superior a 30.000$ serán fabricante, potencia del motor y año de matriculación.

Además, respecto a las variables numéricas más influyentes en el precio son CV del motor, número de cilindros y año de matriculación. En ese mismo apartado, se realiza un test de chi cuadrado para conocer si el fabricante influye en el precio, con un resultado positivo a esta pregunta, por lo que, los consumidores deberían tener en cuenta todas estas variables mencionadas anteriormente.

Estas conclusiones no serían novedosas, puesto que la mayoría de los consumidores (grupo en el que me incluyo), a priori dirían que estas variables son las más relevantes. Sin embargo, otras conclusiones más interesantes pueden ser extraídas.

En primer lugar, existe una tendencia a fabricar coches con cuatro puertas en la actualidad, además de que estos coches más modernos tienen consumos menores.

La popularidad o percepción que se tiene sobre la marca incluye a todos los modelos disponibles, es decir, en la mente del consumidor se posiciona la marca, no separa entre modelos. No obstante, este factor influye poco a la hora de determinar cualquier variable, ya que sus valores de correlación con el resto de variables son prácticamente nulos.

El tamaño afecta también al consumo y al precio, siendo los grandes los que más consumen y los más caros. Sus opuestos (los pequeños), presentan lo contrario, precios más baratos y consumos menores. No obstante, el estilo de coche también afecta, siendo un claro ejemplo los descapotables y los coches deportivos, siendo estos últimos 8.000$ más baratos de media, a pesar de tener más caballos.

Por otro lado, la tracción afecta tanto al consumo como al precio, ya que la tracción trasera está presente en los coches más caros y con más consumo. Sucede lo contrario en la tracción en las 4 ruedas, que son los más baratos y los que menos consumen.

También existe una tendencia a la fabricación de coches automáticos, ya que los manuales presentan una media de año de matriculación más antiguo que los automáticos.

Además, los coches que son gasolina tendrán más posibilidades de tener tracción en las cuatro ruedas, y los coches diésel o eléctricos tendrán más propensión a tener tracción trasera.

Por otro lado, los coches eléctricos presentan casi siempre el tipo de cambio “direc_drive” y 0 cilindros en el motor, además de tener un consumo casi 5 veces menor al resto de coches.

9. Bibliografía

Car Features and MSRP. (2016, 21 diciembre). Kaggle. *https://www.kaggle.com/CooperUnion/cardataset*

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik, and Achim Zeileis. 2004. "Kernlab – an S4 Package for Kernel Methods in R." Journal of Statistical Software 11 (9): 1–20. *http://www.jstatsoft.org/v11/i09/*.

Lê, Sébastien, Julie Josse, and François Husson. 2008. "FactoMineR: A Package for Multivariate Analysis." Journal of Statistical Software 25 (1): 1–18. *https://doi.org/10.18637/jss.v025.i01*.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." R News 2 (3): 18–22. *https://CRAN.R-project.org/doc/Rnews/*.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2020. E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), Tu Wien. *https://CRAN.R-project.org/package=e1071*.

Milborrow, Stephen. 2020. Rpart.plot: Plot 'Rpart' Models: An Enhanced Version of 'Plot.rpart'. https: //CRAN.R-project.org/package=rpart.plot.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. *https://www.R-project.org/*.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. "ROCR: Visualizing Classifier Performance in R." Bioinformatics 21 (20): 7881. *http://rocr.bioinf.mpi-sb.mpg.de*.

Therneau, Terry, and Beth Atkinson. 2019. Rpart: Recursive Partitioning and Regression Trees. *https: //CRAN.R-project.org/package=rpart*.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." Journal of Statistical Software 45 (3): 1–67. *https://www.jstatsoft.org/v45/i03/*.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Fourth. New York: Springer. *http://www.stats.ox.ac.uk/pub/MASS4*.

Wei, Taiyun, and Viliam Simko. 2017. R Package "Corrplot": Visualization of a Correlation Matrix. *https://github.com/taiyun/corrplot*.

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2021). Hmisc: Harrell Miscellaneous. *https://CRAN.R-project.org/package=Hmisc*

RStudio Team. 2016. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, Inc. *http://www.rstudio.com/*

Torgo, L. 2016. Data Mining with R, learning with case studies, 2nd edition. Chapman; Hall/CRC. *http://ltorgo.github.io/DMwR2.*

ANEXOS

ANEXO 1: CÓDIGO UTILIZADO EN R STUDIO

a) Introducción

````
# 1. Introducción

```{r setup, include=FALSE}
library(corrplot)
library(Hmisc)
library(dplyr)
library(readr)
library(mice)
library(cluster)
library(NbClust)
library(clValid)
library(knitr)
library(FactoMineR)
library(factoextra)
library(knitr)
library(ggplot2)
library(rpart)
library(rpart.plot)
library(caret)
library(randomForest)
library(e1071)
library(class)
library(ROCR)
library(kernlab)
```
````

b) Preparación de la base de datos

2. Preparación de la base de datos

```{r bibs, include= FALSE}
tfg <- read.csv("data.csv")
tfg <- as.data.frame(tfg)
descTfg = data.frame("variable" = colnames(tfg),
                "tipo" = c("categorical", "model", "numerical",
                        "categorical",rep("numerical",2),
                        rep("categorical",2),"numerical",rep("categorical",3),
                        rep("numerical",4)),
                        stringsAsFactors = FALSE)
rownames(descTfg) = descTfg$variable
tfg$Make= as.factor(tfg$Make)
tfg$Model= factor(tfg$Model)
tfg$Transmission.Type=factor(tfg$Transmission.Type)
tfg$Market.Category=factor(tfg$Market.Category)
tfg$Vehicle.Style=factor(tfg$Vehicle.Style)
tfg$Driven_Wheels[tfg$Driven_Wheels== "all wheel drive"] = "four wheel drive"
tfg$Driven_Wheels= factor(tfg$Driven_Wheels,
    labels= c("rear wheel drive","front wheel drive", "four wheel drive"),
    levels =c("rear wheel drive","front wheel drive", "four wheel drive"))
tfg$Vehicle.Size= factor(tfg$Vehicle.Size,
    labels=c("Compact","Midsize","Large"),
    levels= c("Compact","Midsize","Large"))
tfg[11322,4]= "regular unleaded"
tfg[11323,4]= "regular unleaded"
tfg[11324,4]= "regular unleaded"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "flex-fuel (premium unleaded
recommended/E85)"] = "flex-fuel"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "flex-fuel (premium unleaded
required/E85)"] = "flex-fuel"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "flex-fuel (unleaded/E85)"] = "flex-fuel"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "flex-fuel (unleaded/natural gas)"] =
"flex-fuel"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "premium unleaded (recommended)"]
= "gasolina"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "premium unleaded (required)"] =
"gasolina"
tfg$Engine.Fuel.Type[tfg$Engine.Fuel.Type== "regular unleaded"] = "gasolina"
tfg$Engine.Fuel.Type= factor(tfg$Engine.Fuel.Type,
     labels= c("diésel","electric", "flex-fuel", "gas natural", "gasolina"),
     levels =c("diesel","electric", "flex-fuel", "natural gas","gasolina"))
```

2.1 Resumen de los tipos de variables

```
tfg$Number.of.Doors[tfg$Number.of.Doors==3]=2
#Numericas
summary(tfg[,descTfg$variable[descTfg$tipo == "numerical"]])
#Categoricas
apply(tfg[,descTfg$variable[descTfg$tipo == "categorical"]], 2, table,
    useNA = "i")
```

```{r eliminado market, include= FALSE}
#Eliminación de Market Category
tfg= tfg[,-10]
descTfg = descTfg[colnames(tfg),]
```

c) Valores inconsistentes o anómalos

3. Valores inconsistentes o anómalos

3.1. Valores inconsistentes en variables numéricas

```{R VARIN, echo= FALSE}
par(mar = c(9,4,4,2))
boxplot(tfg[,descTfg$variable[descTfg$tipo == "numerical"]]+1,
      log = "y", las = 3, cex.axis=0.62, col = c("#A9D0F5", "#2E9AFE",
                                    "#0174DF", "#0431B4", "#08088A", "#210B61"))
```

```{r numericas, echo= FALSE}
par(mfrow = c(1,4))
boxplot(tfg$highway.MPG, xlab = "Consumo en carretera",
            cex.lab=0.85, col = c("#A9D0F5"))
boxplot(tfg$city.mpg, xlab = "Consumo en ciudad",
            cex.lab=0.85, col = c( "#2E9AFE"))
boxplot(tfg$Year, xlab = "Año de matriculación",
            cex.lab=0.85, col = c( "#0174DF"))
boxplot(tfg$Engine.HP, xlab = "Caballos del coche",
            cex.lab= 0.85, col = c( "#0431B4"))
```

```{r elimin A6, echo= FALSE}
#Máximo consumo en carretera
head(tfg[1120,1:3])
summary(tfg$highway.MPG[tfg$Model=="A6"])
tfg[1120,12]= tfg[1120,12]/10
summary(tfg$highway.MPG[tfg$Model=="A6"])
```
```{r caballos, echo= FALSE}
head(tfg[11363,1:3])
summary(tfg$Engine.HP[tfg$Model=="Veyron 16.4"])
#Cuantos Bugatti Veyron 16.4
tfg%>%
  count(tfg$Model=="Veyron 16.4")
```

3.2. Valores faltantes por variable

```{r NAvars, echo= FALSE}
numNA.V = apply(tfg, 2, function(x) sum(is.na(x)))
percNA.V = round(100*apply(tfg, 2, function(x) mean(is.na(x))), 2)
tablaNA.V = data.frame("Variable" = colnames(tfg), numNA.V, percNA.V)
barplot(table(tablaNA.V$percNA.V), xlab = "% Valores faltantes",
        ylab = "Número de casos",main = "Variables con NA's",
         col = c("#A9D0F5", "#2E9AFE", "#0174DF", "#0431B4", "#08088A",
             "#210B61"), las = 2)
```

```{r Na continuacion, echo= FALSE}
#Correlación NAs entre variables
#correlaciones dicotomizadas
par(mar=c(1,1,1,4))
faltantes = is.na(tfg)
corre = cor(faltantes[,tablaNA.V$numNA.V > 0])
corrplot(corre, is.corr = TRUE, method = "ellipse", diag = FALSE,
            addCoef.col = 1, tl.col = 1, type = "upper", number.cex =0.9, tl.cex = 0.75)
```

3.3. Valores faltantes por caso

```{R OBS, echo=FALSE}
numNA.O = apply(tfg, 1, function(x) sum(is.na(x)))
percNA.O = round(100*apply(tfg, 1, function(x) mean(is.na(x))), 2)
tablaNA.O = data.frame(numNA.O, percNA.O)
barplot(table(tablaNA.O$percNA), xlab = "% Valores faltantes",
          ylab = "Número de casos",
          main = "Observaciones con NA's", col = c("#A9D0F5", "#2E9AFE",
         "#0174DF","#0431B4", "#08088A", "#210B61"))
```

3.4. Imputación de valores faltantes

```{r imputacion mice, include= FALSE}
tfg= tfg[,-2]
descTfg = descTfg[colnames(tfg),]
tfg$Make= as.character(tfg$Make)
tfg$Engine.Fuel.Type= as.character(tfg$Engine.Fuel.Type)
tfg$Driven_Wheels=as.character(tfg$Driven_Wheels)
tfg$Vehicle.Size= as.character(tfg$Vehicle.Size)
tfg$Vehicle.Style=as.character(tfg$Vehicle.Style)
tfg$Transmission.Type[tfg$Transmission.Type=="UNKNOWN"]= "NA"
tfg$Transmission.Type= factor(tfg$Transmission.Type, labels= c(1,2,3,4), levels =
c("AUTOMATED_MANUAL","AUTOMATIC","DIRECT_DRIVE","MANUAL"))
```

```
set.seed(1)
tfg.IMP = mice(tfg, seed = 123, m = 10, print = FALSE, method = NULL)
tfg.IMP = complete(tfg.IMP)
tfg.IMP$Make= as.factor(tfg$Make)
tfg.IMP$Engine.Fuel.Type= factor(tfg$Engine.Fuel.Type)
tfg.IMP$Driven_Wheels=factor(tfg$Driven_Wheels)
tfg.IMP$Vehicle.Size= factor(tfg$Vehicle.Size)
tfg.IMP$Vehicle.Style=factor(tfg$Vehicle.Style)
tfg.IMP$Transmission.Type= factor(tfg.IMP$Transmission.Type, labels=
c("AUTOMATED_MANUAL","AUTOMATIC","DIREC_DRIVE","MANUAL"), levels =
c(1,2,3,4))
```

```{r comprobamos , echo= FALSE}
summary(tfg.IMP$Engine.HP)
summary(tfg.IMP$Engine.Cylinders)
summary(tfg.IMP$Number.of.Doors)
summary(tfg.IMP$Transmission.Type)
```

d) Correlaciones

```
# 4. Correlaciones

## 4.1 Correlaciones entre las variables numéricas

```{r correlaciones num, echo= FALSE}
tfg.IMP.NUM= tfg.IMP[,descTfg$variable[descTfg$tipo=="numerical"]]
rcorr(as.matrix(tfg.IMP.NUM))
```

```{r chi test, echo=FALSE, warning=FALSE}
chisq.test(tfg.IMP$MSRP,tfg.IMP$Make)
```

4.2 Relación entre tamaño y consumo

```{r tipo cambio, echo= FALSE}
tfg.IMP%>%
  group_by(Vehicle.Size) %>%
  summarise(precio.medio= mean(MSRP), ciudad.consumo= mean(city.mpg),
carretera.consumo= mean(highway.MPG))
```

4.3 Relación entre precio y popularidad

```{r correlaciones cate, echo= FALSE}
tfg.IMP$intervalo= cut(tfg.IMP$MSRP, breaks =
c(0,12000,22000,35000,50000,100000,3000000),labels =
c(2~12,12~22,22~35,35~50,50~100,">100"))
tfg.IMP$popu= cut(tfg.IMP$Popularity, breaks =
c(0,400,800,1200,1600,2100,6000),labels = c("Muy baja","Baja","Media","Media-
alta","Alta","Muy alta"))
table(tfg.IMP$Make, tfg.IMP$popu)
```

```{r popularidad precio, echo=FALSE}
tfg.IMP%>%
  group_by(popu) %>%
  summarise(precio.medio= mean(MSRP))
tfg.IMP= tfg.IMP[,-15]
tfg.IMP= tfg.IMP[,-15]
```


4.4 Relación entre potencia de motor y estilo de coche
```

```{r potencia y estilo, echo = FALSE}
tfg.IMP%>%
 group_by(Vehicle.Style)%>%
    summarise(Media.de.CV= mean(Engine.HP), precio.medio= mean(MSRP),
  consumo.ciudad= mean(city.mpg))%>%
 dplyr::arrange(desc(Media.de.CV))
```

4.5 Relación entre número de puertas y antigüedad del coche

```{r potencia y puertas, echo= FALSE}
tfg.IMP$Number.of.DoorsF= factor(tfg.IMP$Number.of.Doors, labels= c(2,4), levels =
c(2,4))
tfg.IMP%>%
   group_by(Number.of.DoorsF)%>%
    summarise(Media.de.CV= mean(Engine.HP), media.cilindros=
  mean(Engine.Cylinders))
```

```{r potencia filtrados coupe, echo= FALSE}
barplot(table(tfg.IMP$Number.of.DoorsF, tfg.IMP$Year), main= "Matriculación de
coches con 2 y 4 puertas", xlab= "Años", ylab= "Frecuencia", legend =
rownames(table(tfg.IMP$Number.of.DoorsF, tfg.IMP$Year)), col= c("red", "green"))
```

```{r potencia filtrados descap, echo= FALSE}
tfg.IMP%>%
 group_by(Number.of.DoorsF)%>%
 summarise(año.matriculación.medio= mean(Year))
```

4.6 Relación entre precio y tipo de cambio

```{r tamaño , echo= FALSE}
tfg.IMP%>%
 group_by(Transmission.Type) %>%
    summarise(precio.medio= mean(MSRP), Año.medio= mean(Year),
  carretera.consumo= mean(highway.MPG))
```

```{r chisqutestqasa, echo= FALSE}
  chisq.test(tfg.IMP$MSRP,tfg.IMP$Transmission.Type)
```

4.7 Relación entre motor y consumo.

```{r consumo, echo= FALSE}
tfg.IMP$Engine.CylindersF= factor(tfg.IMP$Engine.Cylinders, labels=
c(0,3,4,5,6,8,10,12,16), levels = c(0,3,4,5,6,8,10,12,16))
tfg.IMP%>%
  group_by(Engine.CylindersF) %>%
     summarise(Potencia.media=mean(Engine.HP), ciudad.consumo= mean(city.mpg),
   precio.medio= mean(MSRP))
```

```{r consumo 0 jeje, echo= FALSE}
tfg.IMP%>%
  filter(Engine.CylindersF==0)%>%
  select(Make,Engine.Fuel.Type,Transmission.Type,highway.MPG, city.mpg)
tfg.IMP= tfg.IMP[,-15]
tfg.IMP= tfg.IMP[,-15]
```

```{r prueba 1540510, echo= FALSE}
tfg.IMP%>%
  filter(Engine.Fuel.Type=="electric")%>%
  count(Engine.Cylinders)
```

```{r dasdadsa, echo = FALSE}
tfg.IMP%>%
  filter(Transmission.Type=="DIREC_DRIVE")%>%
  count(Engine.Fuel.Type)
```

4.8 Relación entre tipo de combustible y consumo.

```{r electrricos, echo= FALSE}
tfg.IMP%>%
  group_by(Engine.Fuel.Type)%>%
  summarise(ciudad.consumo= mean(city.mpg), carretera.consumo=
mean(highway.MPG))%>%
  arrange(desc(ciudad.consumo))
```

4.9 Relación entre tipo de tracción y consumo

```{r tracción consumo husdad, echo= FALSE}
tfg.IMP%>%
  group_by(Driven_Wheels)%>%
  summarise(ciudad.consumo= mean(city.mpg), precio.medio= mean(MSRP))%>%
  arrange(desc(ciudad.consumo))
```

```{r probandoq wesad, echo= FALSE}
table(tfg.IMP$Driven_Wheels, tfg.IMP$Vehicle.Size)
```

e) Evolución temporal de las variables de interés

#5 Evolución temporal de las variables de interés

```{R EVOLUCIÓN TEMPORAL asd, ECHO= FALSE}
#caballos
c1990= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1990"])
c1991= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1991"])
c1992= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1992"])
c1993= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1993"])
c1994= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1994"])
c1995= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1995"])
c1996= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1996"])
c1997= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1997"])
c1998= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1998"])
c1999= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="1999"])
c2000= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2000"])
c2001= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2001"])
c2002= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2002"])
c2003= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2003"])
c2004= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2004"])
c2005= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2005"])
c2006= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2006"])
c2007= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2007"])
c2008= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2008"])
c2009= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2009"])
c2010= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2010"])
c2011= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2011"])
c2012= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2012"])
c2013= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2013"])
c2014= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2014"])
c2015= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2015"])
c2016= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2016"])
c2017= mean(tfg.IMP$Engine.HP[tfg.IMP$Year=="2017"])
caballos=
c(c1990,c1991,c1992,c1993,c1994,c1995,c1996,c1997,c1998,c1999,c2000,c2001,c20
02,c2003,c2004,c2005,c2006,c2007,
c2008,c2009,c2010,c2011,c2012,c2013,c2014,c2015,c2016,c2017)
tiempo=
c(1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004
,2005,2006,2007, 2008,2009,2010,2011,2012,2013,2014,2015,2016,2017)
grafica= data.frame(tiempo,caballos)
plot(grafica, main= "Evolución de la media de caballos", xlab= "Años",
ylab="Caballos")
```

```{r numero de cilindros, echo= FALSE}
#cilindros
c1990= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1990"])
```

```
c1991= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1991"])
c1992= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1992"])
c1993= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1993"])
c1994= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1994"])
c1995= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1995"])
c1996= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1996"])
c1997= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1997"])
c1998= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1998"])
c1999= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="1999"])
c2000= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2000"])
c2001= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2001"])
c2002= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2002"])
c2003= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2003"])
c2004= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2004"])
c2005= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2005"])
c2006= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2006"])
c2007= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2007"])
c2008= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2008"])
c2009= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2009"])
c2010= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2010"])
c2011= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2011"])
c2012= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2012"])
c2013= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2013"])
c2014= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2014"])
c2015= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2015"])
c2016= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2016"])
c2017= mean(tfg.IMP$Engine.Cylinders[tfg.IMP$Year=="2017"])
cilindros=
c(c1990,c1991,c1992,c1993,c1994,c1995,c1996,c1997,c1998,c1999,c2000,c2001,c20
02,c2003,c2004,c2005,c2006,c2007,
c2008,c2009,c2010,c2011,c2012,c2013,c2014,c2015,c2016,c2017)

grafica2= data.frame(tiempo,cilindros)
plot(grafica2, main= "Evolución de la media de cilindros", xlab= "Años",
ylab="Cilindros")
tfg.IMP%>%
  dplyr::filter(tfg.IMP$Engine.Fuel.Type=="electric")%>%
  select(Year,Engine.Fuel.Type) ```
```{r numero de cilindroseqeasd, echo= FALSE}
#consumo
c1990= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1990"])
c1991= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1991"])
c1992= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1992"])
c1993= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1993"])
c1994= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1994"])
c1995= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1995"])
c1996= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1996"])
```
```

```
c1997= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1997"])
c1998= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1998"])
c1999= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="1999"])
c2000= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2000"])
c2001= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2001"])
c2002= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2002"])
c2003= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2003"])
c2004= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2004"])
c2005= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2005"])
c2006= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2006"])
c2007= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2007"])
c2008= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2008"])
c2009= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2009"])
c2010= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2010"])
c2011= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2011"])
c2012= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2012"])
c2013= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2013"])
c2014= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2014"])
c2015= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2015"])
c2016= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2016"])
c2017= mean(tfg.IMP$city.mpg[tfg.IMP$Year=="2017"])
consumo=
c(c1990,c1991,c1992,c1993,c1994,c1995,c1996,c1997,c1998,c1999,c2000,c2001,c20
02,c2003,c2004,c2005,c2006,c2007,
c2008,c2009,c2010,c2011,c2012,c2013,c2014,c2015,c2016,c2017)
grafica3= data.frame(tiempo,consumo)
plot(grafica3, main= "Evolución del consumo medio de combustible", xlab= "Años",
ylab="Consumo en ciudad")
```

f) Análisis de componentes principales (PCA)

6. PCA

6.1 Formulación del modelo

```{r pca tfg qeasimp, echo= FALSE, fig.height=6}
tfg.IMP$Transmission.Type= as.numeric(tfg.IMP$Transmission.Type)
tfg.IMP$Driven_Wheels= as.numeric(tfg.IMP$Driven_Wheels)
tfg.IMP$Engine.Fuel.Type= as.numeric(tfg.IMP$Engine.Fuel.Type)
tfg.IMP$Vehicle.Size= as.numeric(tfg.IMP$Vehicle.Size)
tfg.IMP$Engine.Fuel.Type= as.numeric(tfg.IMP$Engine.Fuel.Type)
descPCA = data.frame("variable" = colnames(tfg.IMP),
                "tipo" = c ("categorical",rep("numerical",8),
                        "categorical","numerical","numerical","numerical",
                        stringsAsFactors = FALSE))
res.pca = PCA(tfg.IMP, scale.unit = TRUE, graph = FALSE, ncp = 6, quali.sup =
which(descPCA$tipo == "categorical"))
eig.val <- get_eigenvalue(res.pca)
VPmedio = 100 * (1/nrow(eig.val))
fviz_eig(res.pca, addlabels = TRUE, number.cex= 0.55) +
geom_hline(yintercept=VPmedio, linetype=2, color="red", cex= 0.55)
kable(eig.val[1:6,])
```

6.2 Validación del modelo PCA

```{r test hotel, echo = FALSE}
K = 6
misScores = res.pca$ind$coord[,1:K]
miT2 = colSums(t(misScores**2) / eig.val[1:K])
I = nrow(tfg.IMP)
F95 = K*(I**2 - 1)/(I*(I - K)) * qf(0.95, K, I-K)
plot(1:length(miT2), miT2, type = "l", xlab = "Coches", ylab = "T2")
abline(h = F95, col = "orange", lty = 2, lwd = 2)
anomalas = which(miT2 > F95)
length(anomalas)
```

6.3 Interpretación del modelo PCA

```{r primer gráfico, echo=FALSE}
fviz_pca_var(res.pca, axes = c(1,2), repel = TRUE, col.var = "contrib",
        gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
```

```{r contribuciohgohkjn, echo= FALSE}
fviz_contrib(res.pca, choice = "var", axes = 1)
fviz_contrib(res.pca, choice = "var", axes = 2)
```

```{r dimension 3 y 4, echo= FALSE}
fviz_pca_var(res.pca, axes = c(3,4), repel = TRUE, col.var = "contrib",
        gradient.cols = c("#999999", "#E69F00", "#56B4E9"))
```

```{r contribucion, echo= FALSE}
fviz_contrib(res.pca, choice = "var", axes = 3)
fviz_contrib(res.pca, choice = "var", axes = 4)
```

```{r dimension 5 y 6, echo= FALSE}
fviz_pca_var(res.pca, axes = c(5,6), repel = TRUE, col.var = "contrib",
        gradient.cols = c("Red", "Green", "Black"))
```

g) Modelos de predicción

7. Modelos de predicción

6.1 Preparación de la base de datos

```{r part.s, echo= FALSE}
tfg.IMP$Make= as.factor(tfg$Make)
tfg.IMP$Engine.Fuel.Type= factor(tfg$Engine.Fuel.Type)
tfg.IMP$Driven_Wheels=factor(tfg$Driven_Wheels)
tfg.IMP$Vehicle.Size= factor(tfg$Vehicle.Size)
tfg.IMP$Vehicle.Style=factor(tfg$Vehicle.Style)
tfg.IMP$Transmission.Type= factor(tfg.IMP$Transmission.Type, labels=
c("AUTOMATED_MANUAL","AUTOMATIC","DIREC_DRIVE","MANUAL"), levels =
c(1,2,3,4))

#particion
tfg.IMP$mediana= cut(tfg.IMP$MSRP, breaks = c(0,30000,3000000),labels = c(0,1))
tfg.IMPN= tfg.IMP
tfg.IMPN= tfg.IMPN[,-14]
set.seed(12)
tfg.IMPN <- as.data.frame(tfg.IMPN)
train.index <- sample(c(1:dim(tfg.IMPN)[1]), dim(tfg.IMPN)[1]*0.6)
train.tfg <- tfg.IMPN[train.index, ]
valid.tfg <- tfg.IMPN[-train.index, ]
#balanceado
set.seed(12)
s<-sample(rownames(train.tfg[train.tfg$mediana==1,]))
train.bal<-rbind(train.tfg[train.tfg$mediana==0,], train.tfg[s,])
train.bal$mediana <- as.numeric(train.bal$mediana)
train.bal$mediana[train.bal$mediana == 1 ]<- 0
train.bal$mediana[train.bal$mediana == 2 ]<- 1
```

7.2 Construcción de los modelos

7.2.1 Modelo de regresión general

a) Construcción del modelo con los datos de entrenamiento

```{r consreg, include= FALSE, warning= FALSE}
reg <- glm(mediana ~ ., data = train.bal, family = "binomial")
final.reg <- step(reg)
```

```{r sumreg, include= FALSE}
options(scipen=999)
summary(final.reg)
```

b) Evaluación del modelo

```{r predreg, echo= FALSE}
pred.logit <- predict(final.reg, valid.tfg, type = "response")
pred.reg <- prediction(pred.logit, valid.tfg[,14])
#curva roc
perf.reg<-performance(pred.reg,"tpr","fpr")
#auc
perf.auc.reg <- performance(pred.reg,"auc")
perf.auc.reg@y.values
```

7.2.2 Árbol de partición

a) Construcción del modelo con los datos de entrenamiento

```{r consarb, echo= FALSE}
#Modelo
default <- rpart(factor(mediana)~., data = train.bal, method = "class")
#Representación del árbol
prp(default, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10)
printcp(default)
```

b) Evaluación del modelo

```{r predarb, echo= FALSE}
pred.tree <- prediction(predict(default, valid.tfg[,-14])[,2],
                        valid.tfg[,14])
#curva roc
perf.tree<-performance(pred.tree,"tpr","fpr")
#auc
perf.auc.tree <- performance(pred.tree,"auc")
perf.auc.tree@y.values
```

7.2.3 Random Forest

a) Construcción del modelo con los datos de entrenamiento

```{r consrf, echo= FALSE}
rf <- randomForest(factor(mediana) ~ ., data = train.bal, mtry= 3,
                   method="class", importance= T)
varImpPlot(rf, main = "", col="dark blue",cex= 0.7)
```

b) Evaluación del modelo

```{r predrf, echo= FALSE}
pred.rf.1 <- predict(rf, valid.tfg, type="prob")
pred.rf <- prediction(pred.rf.1[,2], valid.tfg$mediana)
#ROC
perf.rf<-performance(pred.rf,"tpr","fpr")
#AUC
perf.auc.rf<- performance(pred.rf,"auc")
perf.auc.rf@y.values
```

7.2.4 Naive Bayes

a) Construcción del modelo con los datos de entrenamiento

```{r consnb, echo= FALSE}
# NB
nb<- naiveBayes(factor(mediana)~ ., data = train.bal)
```
b) Evaluación del modelo creado

```{r prednb, echo= FALSE}
# NB
pred.nbayes <- prediction(predict(nb, valid.tfg[,-14], type="raw")[,2],
                    valid.tfg[,14])
#curva roc
perf.nbayes<-performance(pred.nbayes,"tpr","fpr")
#auc
perf.auc.nb <- performance(pred.nbayes,"auc")
perf.auc.nb@y.values
```
7.2.5 Máquinas de soporte vectorial

a) Construccion del modelo con los datos de entrenamiento

```{r conssv, echo= FALSE}
#SVM
svm <- svm(factor(mediana)~ ., data = train.bal, method = "C-classification",
        kernel = "radial",cost = 10, gamma = 0.1, probability=TRUE)


#SVP
svp <- ksvm(factor(mediana)~ ., data=train.bal, type = "C-svc",
         kernel = "rbfdot",kpar = "automatic", prob.model= TRUE)

```

b) Evaluación del modelo creado

```{r predsv, echo= FALSE}
#SVM
pred.svm <- prediction(attr(predict(svm, valid.tfg[,-14], probability=TRUE),
                    "probabilities")[,2], valid.tfg[,14])
##ROC
perf.svm<-performance(pred.svm,"tpr","fpr")
##AUC
perf.auc.svm <- performance(pred.svm,"auc")
perf.auc.svm@y.values

#SVP
pred.svp<-prediction(predict(svp, valid.tfg[,-14], type="probabilities")[,2],
              valid.tfg[,14])
##ROC
perf.svp<-performance(pred.svp,"tpr","fpr")
##AUC
perf.auc.svp<- performance(pred.svp,"auc")
perf.auc.svp@y.values
```

7.2.6 Vecino más próximo
```{r varnums, echo=FALSE}
# BBDD entrenamiento
train.bal$Make <- as.numeric(train.bal$Make)
train.bal$Engine.Fuel.Type <- as.numeric(train.bal$Engine.Fuel.Type)
train.bal$Transmission.Type <- as.numeric(train.bal$Transmission.Type)
train.bal$Driven_Wheels<- as.numeric(train.bal$Driven_Wheels)
train.bal$Vehicle.Size<- as.numeric(train.bal$Vehicle.Size)
train.bal$Vehicle.Style<- as.numeric(train.bal$Vehicle.Style)
# BBDD validacion
valid.tfg$Make <- as.numeric(valid.tfg$Make)
valid.tfg$Engine.Fuel.Type <- as.numeric(valid.tfg$Engine.Fuel.Type)
valid.tfg$Transmission.Type <- as.numeric(valid.tfg$Transmission.Type)
valid.tfg$Driven_Wheels<- as.numeric(valid.tfg$Driven_Wheels)
valid.tfg$Vehicle.Size<- as.numeric(valid.tfg$Vehicle.Size)
valid.tfg$Vehicle.Style<- as.numeric(valid.tfg$Vehicle.Style)
```

a) Determinación de las clases más cercanas

```{r consVP, echo= FALSE}
vp<-knn(train.bal[,-14], valid.tfg[,-14], factor(train.bal$mediana),
      k = 3, prob = TRUE)
```

b) Predicción en base al modelo creado

```{r predVP, echo= FALSE}
pred.knn <- prediction(attr(vp,"prob"),valid.tfg[,14])
#ROC
perf.knn<-performance(pred.knn,"tpr","fpr")
#AUC
perf.auc.knn <- performance(pred.knn,"auc")
perf.auc.knn@y.values
```

7.3 Curvas ROC de los modelos

```{r ROC,echo= FALSE}
plot(perf.tree) # arbol de partición
lines(perf.knn@x.values[[1]],perf.knn@y.values[[1]], lty=2, col=2)#vecino
lines(perf.nbayes@x.values[[1]],perf.nbayes@y.values[[1]], lty=3, col=3)#nbayes
lines(perf.svm@x.values[[1]],perf.svm@y.values[[1]], lty=4, col=4)#svm
lines(perf.svp@x.values[[1]],perf.svp@y.values[[1]], lty=5, col=5)#svp
lines(perf.rf@x.values[[1]],perf.rf@y.values[[1]], lty=6, col=6)#random
legend(0.65,0.55,c("Arbol de partición","Vecino" ,"Nbayes","Random
Forest","SVM","SVP"),
     lty=1:6, col=1:6, cex = 0.7 )
```

ANEXO 2: OBJETIVOS DE DESARROLLO SOSTENIBLE

Reflexión sobre la relación del TFG con los ODS en general y con los ODS más relacionados.

El TFG trata sobre el sector automovilístico, uno de los grandes focos de contaminación mundial y de emisión de CO2. Es por eso por lo que el Objetivo de Desarrollo Sostenible principal es el de producción y consumo responsable. Este objetivo tiene como fin principal el producir más contaminando menos.

El sector del automóvil tiene grandes desafíos por delante en este sentido, ya que con una mayor inversión en la industria este objetivo es alcanzable a largo plazo. De esta forma, también se conseguiría una industrialización sostenible y competitiva que pueda usar los recursos de la manera más eficiente posible. Con esta inversión, aparte de reducir el impacto medioambiental, se produciría más eficientemente, lo que también reportaría beneficio económico para la industria debido a costes más bajos. Además, podría generarse una disminución en los precios que beneficiaría al consumidor.

Por otro lado, existen acciones gubernamentales que afectan a este sector. Como es el caso de incentivos para aquellos compradores que adquieran un vehículo eléctrico o híbrido en España. Más en concreto existe el Programa de Incentivos a la Movilidad Eficiente y Sostenible (MOVES III), que llega a financiar hasta 9.000 euros a particulares y autónomos la compra de furgonetas, o hasta 7.000 euros a turismos si el particular achatarra un vehículo de más de siete años. En caso de no aportar ningún vehículo la financiación alcanza los 4.500 euros para los turismos. La Moncloa. (13 de abril de 2021). El Gobierno aprueba ayudas directas para adquirir vehículos eléctricos o híbridos enchufables e instalar infraestructuras de recarga. *https://www.lamoncloa.gob.es/consejodeministros/resumenes/Paginas/2021/130421-cministros.aspx*

También existen otras acciones gubernamentales para fomentar el uso de vehículos eléctricos como el aumento de los impuestos a los combustibles tradicionales. De esta manera aumenta el precio del combustible y hace valorar al consumidor la opción de adquirir un coche eléctrico.

Estas acciones fomentan el uso de fuentes de energía renovables y limpias, al mismo tiempo que reducen el uso de fuentes de energía convencionales, limitadas y más contaminantes, como es el caso del petróleo. De esta forma, se consigue el séptimo objetivo, relacionado con el uso de una energía asequible y no contaminante.

Este objetivo pasa principalmente por el aumento de la producción de una energía sostenible, este aumento es fomentado por una mayor demanda de coches que utilicen energía limpia como combustible. Y es que ya se puede apreciar como cada vez son más las marcas de coches que tienen disponible en su catálogo vehículos eléctricos.

Con estas acciones, se conseguiría un menor impacto en la fabricación de los coches, así como el uso de energías más limpias como combustible, consiguiendo así el objetivo número 13: Acción por el clima.

El cambio climático es uno de los desafíos más importantes en la actualidad y las emisiones de CO2 son el principal causante del efecto invernadero. En este sentido no solo la industria del automóvil, si no todas, deben de adoptar políticas dirigidas a reducir el impacto medioambiental. Es deber de la sociedad exigir acciones gubernamentales que protejan al planeta y consecuentemente a los ciudadanos.